Chapter 2
Multiple Linear Regression

This chapter considers several estimators for the multiple linear regression model. Large sample theory is give for p fixed, but the prediction intervals can have $p > n$. Some testing for the OPLS and MMLE estimators can also have $p > n$.

Definition 2.1. For an important class of regression models, **regression** is the study of the conditional distribution $Y|\boldsymbol{A}\boldsymbol{x}$ of the response variable Y given $\boldsymbol{A}\boldsymbol{x}$, where the vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$ and $\boldsymbol{A}$ is a $k \times p$ constant matrix of full rank k with $1 \leq k \leq p$.

Remark 2.1. If $\boldsymbol{A} = \boldsymbol{I}_p$, then $Y|\boldsymbol{A}\boldsymbol{x} = Y|\boldsymbol{x}$. If $\boldsymbol{\beta}$ is a $p \times 1$ coefficient vector and $\boldsymbol{A} = \boldsymbol{\beta}^T$, then $Y|\boldsymbol{A}\boldsymbol{x} = Y|\boldsymbol{\beta}^T\boldsymbol{x} = Y|\boldsymbol{x}^T\boldsymbol{\beta}$.

Definition 2.2. A **quantitative variable** takes on numerical values while a **qualitative variable** takes on categorical values.

Remark 2.2. The literature often claims that $Y|\boldsymbol{x} = Y|\boldsymbol{\beta}^T\boldsymbol{x}$. This claim is often much too strong.

Notation. Often the conditioning and the index i will be suppressed. For example, the *multiple linear regression model*

$$Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{2.1}$$

for $i = 1, ..., n$ where $\boldsymbol{\beta}$ is a $p \times 1$ unknown vector of parameters, and e_i is a random error. This model could be written $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. More accurately, $Y|\boldsymbol{\beta}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{\beta} + e$, but the conditioning on $\boldsymbol{\beta}^T\boldsymbol{x}$ will often be suppressed. Often the errors $e_1, ..., e_n$ are **iid** (independent and identically distributed). Often the distribution of the errors is unknown, but often it is assumed that the iid e_i's come from a distribution that is known except for a scale parameter. For example, the e_i's might be iid from a normal (Gaussian) distribution with *mean* 0 and unknown *standard deviation* σ. For this Gaussian model, estimation of $\boldsymbol{\beta}$ and σ is important for inference and for predicting a new future value of the response variable Y_f given a new vector of predictors $\boldsymbol{x}_f$.

2.1 The MLR Model

For **multiple linear regression (MLR)**, it is usually useful to have a constant in the model. Sometimes it is convenient to use $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ where $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$ and the constant is β_1. Sometimes it is convenient to separate the constant from the nontrivial predictors and use $Y|(\alpha + \boldsymbol{\beta}^T\boldsymbol{x})$ where α is the constant. We could also use $\boldsymbol{\beta}^T = (\beta_1, \boldsymbol{\beta}_2^T)$ where β_1 is the intercept and the slopes vector $\boldsymbol{\beta}_2 = (\beta_2, ..., \beta_p)^T$, and $\boldsymbol{x}_i^T = (1, \boldsymbol{u}_i^T)$ where the nontrivial predictors $\boldsymbol{u}_i = (x_{i2}, ..., x_{ip})^T$. Hence we get the following two MLR models. The first model is often used in the theory of linear models, while the second model is often useful for Statistical Learning, MLR with heterogeneity, and high dimensional statistics.

Definition 2.3. Suppose that the response variable Y and at least one predictor variable x_i are quantitative.
a) Let the **MLR model 1** be

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{2.2}$$

for $i = 1, ..., n$. Here n is the sample size and the random variable e_i is the ith error. Assume that the e_i are iid with expected value $E(e_i) = 0$ and variance $V(e_i) = \sigma^2$. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times p$ matrix of predictors, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors.

b) Let the **MLR model 2** be

$$Y_i = \alpha + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{2.3}$$

for $i = 1, ..., n$. For this model, we may use $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T$ with $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$.

In matrix notation, suppose the n equations are

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{2.4}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X} = [\boldsymbol{v}_1, \boldsymbol{v}_2, ..., \boldsymbol{v}_p]$ is an $n \times p$ matrix of predictors with ith column $\boldsymbol{v}_i$ corresponding to the ith predictor, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown coefficients, and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Equivalently,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,p} \\ x_{2,1} & x_{2,2} & \dots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \dots & x_{n,p} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}. \tag{2.5}$$

For MLR model 1, the first column of $\boldsymbol{X}$ is $\boldsymbol{v}_1 = \mathbf{1}$, the $n \times 1$ vector of ones. The ith **case** $(\boldsymbol{x}_i^T, Y_i)^T = (x_{i1}, x_{i2}, ..., x_{ip}, Y_i)^T$ corresponds to the ith row $\boldsymbol{x}_i^T$ of $\boldsymbol{X}$ and the ith element of $\boldsymbol{Y}$ (if $x_{i1} \equiv 1$, then x_{i1} could be omitted). In the MLR model $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, the Y and e are random variables, but we only have observed values Y_i and $\boldsymbol{x}_i$. MLR is used to estimate the unknown parameters $\boldsymbol{\beta}$ and σ^2.

Definition 2.4. The **constant variance MLR model** uses the assumption that the errors $e_1, ..., e_n$ are iid with mean $E(e_i) = 0$ and variance $\text{VAR}(e_i) = \sigma^2 < \infty$. Also assume that the errors are independent of the predictor variables $\boldsymbol{x}_i$. The predictor variables $\boldsymbol{x}_i$ are assumed to be fixed and measured without error. The cases $(\boldsymbol{x}_i^T, Y_i)^T$ are independent for $i = 1, ..., n$.

If the predictor variables are random variables, then the above MLR model is conditional on the observed values of the $\boldsymbol{x}_i$. That is, observe the $\boldsymbol{x}_i$ and then act as if the observed $\boldsymbol{x}_i$ are fixed.

Definition 2.5. The **unimodal MLR model** has the same assumptions as the constant variance MLR model, as well as the assumption that the zero mean constant variance errors $e_1, ..., e_n$ are iid from a unimodal distribution that is not highly skewed. Note that $E(e_i) = 0$ and $V(e_i) = \sigma^2 < \infty$.

Definition 2.6. The *normal MLR model* or **Gaussian MLR model** has the same assumptions as the unimodal MLR model but adds the assumption that the errors $e_1, ..., e_n$ are iid $N(0, \sigma^2)$ random variables. That is, the e_i are iid normal random variables with zero mean and variance σ^2.

The unknown coefficients for the above 3 models are usually estimated using (ordinary) least squares (OLS).

Notation. The symbol $A \equiv B = f(c)$ means that A and B are equivalent and equal, and that $f(c)$ is the formula used to compute A and B.

Definition 2.7. Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b} = x_{i,1}b_1 + \cdots + x_{i,p}b_p.$$

The vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. Thus ith residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1}b_1 - \cdots - x_{i,p}b_p$.

2.1.1 OLS Theory

Ordinary least squares (OLS) large sample theory will be useful. Let $\boldsymbol{X} = (\mathbf{1} \;\; \boldsymbol{X}_1)$. For model (2.2), the ith row of $\boldsymbol{X}$ is $(1, x_{i,2}, ..., x_{i,p})$ while for model (2.3), the ith row of $\boldsymbol{X}$ is $(1, x_{i,1}, ..., x_{i,p})$, and $\boldsymbol{Y} = \alpha\mathbf{1} + \boldsymbol{X}_1\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}$.

Definition 2.8. Using the above notation for MLR model 2 given by Equation (2.3), let $\boldsymbol{x}_i^T = (x_{i1}, ..., x_{ip})$, let α be the intercept, and let the slopes vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$. Let the population covariance matrices

$$\mathrm{Cov}(\boldsymbol{x}) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T] = \boldsymbol{\Sigma}_{\boldsymbol{x}}, \;\; \text{and}$$

$$\mathrm{Cov}(\boldsymbol{x}, Y) = E[(\boldsymbol{x} - E(\boldsymbol{x}))(Y - E(Y))] = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}.$$

If the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population where $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exists and $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ is nonsingular, then the population coefficients from an OLS regression of Y on $\boldsymbol{x}$ (even if a linear model does not hold) are

$$\alpha = \alpha_{OLS} = E(Y) - \boldsymbol{\beta}^T E(\boldsymbol{x}) \;\; \text{and} \;\; \boldsymbol{\beta} = \boldsymbol{\beta}_{\mathrm{OLS}} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}\mathrm{Y}}.$$

Definition 2.9. Let the sample covariance matrices be

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T \;\; \text{and} \;\; \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}).$$

Let the method of moments estimators be $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^T$ and

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i Y_i - \overline{\boldsymbol{x}}\,\overline{Y}.$$

The method of moment estimators are often called the maximum likelihood estimators, but are the MLE if the $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from a multivariate normal distribution, a very strong assumption. In Theorem 2.1, note that $\boldsymbol{D} = \boldsymbol{X}_1^T\boldsymbol{X}_1 - n\overline{\boldsymbol{x}}\,\overline{\boldsymbol{x}}^T = (n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$.

Theorem 2.1: Seber and Lee (2003, p. 106). Let $\boldsymbol{X} = (\mathbf{1} \;\; \boldsymbol{X}_1)$. Then $\boldsymbol{X}^T\boldsymbol{Y} = \begin{pmatrix} n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y} \end{pmatrix} = \begin{pmatrix} n\overline{Y} \\ \sum_{i=1}^{n}\boldsymbol{x}_i Y_i \end{pmatrix}$, $\boldsymbol{X}^T\boldsymbol{X} = \begin{pmatrix} n & n\overline{\boldsymbol{x}}^T \\ n\overline{\boldsymbol{x}} & \boldsymbol{X}_1^T\boldsymbol{X}_1 \end{pmatrix}$,

$$\text{and} \;\; (\boldsymbol{X}^T\boldsymbol{X})^{-1} = \begin{pmatrix} \frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1} \end{pmatrix}$$

where the $p \times p$ matrix $\boldsymbol{D}^{-1} = [(n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}]^{-1} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/(n-1)$.

Under model (2.3), $\hat{\boldsymbol{\phi}} = \hat{\boldsymbol{\phi}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$.

Theorem 2.2: Second way to compute $\hat{\boldsymbol{\phi}}$:
a) If $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}$ exists, then $\hat{\alpha} = \overline{Y} - \hat{\boldsymbol{\beta}}^T\overline{\boldsymbol{x}}$ and

$$\hat{\boldsymbol{\beta}} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}.$$

b) Suppose that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid random vectors such that σ_Y^2, $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$, and $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ exist. Then $\hat{\alpha} \xrightarrow{P} \alpha$ and

$$\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta} \text{ as } \mathrm{n} \to \infty$$

where α and $\boldsymbol{\beta}$ are given by Definition 2.8.

Proof. Note that

$$\boldsymbol{Y}^T\boldsymbol{X}_1 = (Y_1 \cdots Y_n)\begin{bmatrix}\boldsymbol{x}_1^T \\ \vdots \\ \boldsymbol{x}_n^T\end{bmatrix} = \sum_{i=1}^n Y_i\boldsymbol{x}_i^T$$

and

$$\boldsymbol{X}_1^T\boldsymbol{Y} = [\boldsymbol{x}_1 \cdots \boldsymbol{x}_n]\begin{bmatrix}Y_1 \\ \vdots \\ Y_n\end{bmatrix} = \sum_{i=1}^n \boldsymbol{x}_i Y_i.$$

So

$$\begin{bmatrix}\hat{\alpha} \\ \hat{\boldsymbol{\beta}}\end{bmatrix} = \begin{bmatrix}\frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{1}^T \\ \boldsymbol{X}_1^T\end{bmatrix}\boldsymbol{Y} =$$

$$\begin{bmatrix}\frac{1}{n} + \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & -\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} \\ -\boldsymbol{D}^{-1}\overline{\boldsymbol{x}} & \boldsymbol{D}^{-1}\end{bmatrix}\begin{bmatrix}n\overline{Y} \\ \boldsymbol{X}_1^T\boldsymbol{Y}\end{bmatrix}.$$

Thus $\hat{\boldsymbol{\beta}} = -n\boldsymbol{D}^{-1}\overline{\boldsymbol{x}}\,\overline{Y} + \boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \boldsymbol{D}^{-1}(\boldsymbol{X}_1^T\boldsymbol{Y} - n\overline{\boldsymbol{x}}\,\overline{Y}) =$

$$\boldsymbol{D}^{-1}\left[\sum_{i=1}^n \boldsymbol{u}_i Y_i - n\overline{\boldsymbol{x}}\,\overline{Y}\right] = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}}{n-1}n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \frac{n}{n-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}. \text{ Then}$$

$\hat{\alpha} = \overline{Y} + n\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\overline{\boldsymbol{x}}\,\overline{Y} - \overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1}\boldsymbol{X}_1^T\boldsymbol{Y} = \overline{Y} + [n\overline{Y}\overline{\boldsymbol{x}}^T\boldsymbol{D}^{-1} - \boldsymbol{Y}^T\boldsymbol{X}_1\boldsymbol{D}^{-1}]\overline{\boldsymbol{x}}$
$= \overline{Y} - \hat{\boldsymbol{\beta}}^T\overline{\boldsymbol{x}}$. The convergence in probability results hold since sample means and sample covariance matrices are consistent estimators of the population means and population covariance matrices. □

Remark 2.3. It is important to note that the convergence in probability results are for iid $(Y_i, \boldsymbol{x}_i^T)^T$ with second moments and nonsingular $\boldsymbol{\Sigma}_{\boldsymbol{x}}$: a linear model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ does not need to hold. When the linear model does hold, the second method for computing $\hat{\boldsymbol{\beta}}$ is still valid even if $\boldsymbol{X}$ is a

constant matrix, and $\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{\beta}$ by Theorem 2.3 b). From Theorem 2.3,

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1} = \hat{\boldsymbol{V}} = \begin{pmatrix} \hat{\boldsymbol{V}}_{11} & \hat{\boldsymbol{V}}_{12} \\ \hat{\boldsymbol{V}}_{21} & \hat{\boldsymbol{V}}_{22} \end{pmatrix} \xrightarrow{P} \boldsymbol{V} = \begin{pmatrix} \boldsymbol{V}_{11} & \boldsymbol{V}_{12} \\ \boldsymbol{V}_{21} & \boldsymbol{V}_{22} \end{pmatrix}.$$

Thus $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1} \xrightarrow{P} \boldsymbol{V}_{22}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{V}_{22}^{-1}$. Note that for Theorem 2.3 b) with iid cases and $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$,

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V} = \begin{bmatrix} 1 + \boldsymbol{\mu}_{\boldsymbol{x}}^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}} & -\boldsymbol{\mu}_{\boldsymbol{x}}^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \\ -\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\mu}_{\boldsymbol{x}} & \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \end{bmatrix}.$$

Definition 2.10. For OLS and MLR model 1 from Definition 2.3, $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$. Let the *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Then $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_{OLS} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. The ith leverage $h_i = \boldsymbol{H}_{ii}$ = the ith diagonal element of $\boldsymbol{H}$.

There are many large sample theory results for ordinary least squares. For Theorem 2.3, see, for example, Sen and Singer (1993, p. 280). Theorem 2.3 is analogous to the central limit theorem and the theory for the t–interval for μ based on $\overline{Y}$ and the sample standard deviation (SD) S_Y. If the data $Y_1, ..., Y_n$ are iid with mean 0 and variance σ^2, then $\overline{Y}$ is asymptotically normal and the t–interval will perform well if the sample size is large enough. The results below suggests that the OLS estimators $\hat{Y}_i$ and $\hat{\boldsymbol{\beta}}$ are good if the sample size is large enough. The condition $\max h_i \to 0$ in probability usually holds if the researcher picked the design matrix $\boldsymbol{X}$ or if the $\boldsymbol{x}_i$ are iid random vectors from a well behaved population. Outliers can cause the condition to fail. Theorem 2.3 a) implies that $\hat{\boldsymbol{\beta}} \approx N_p[\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}]$. For Theorem 2.3 a), $\text{rank}(\boldsymbol{X}) = p$ since $\boldsymbol{X}^T\boldsymbol{X}$ is nonsingular. For Theorem 2.3 b), $\text{rank}(\boldsymbol{X}) = p + 1$.

Theorem 2.3, OLS CLTs. Consider the MLR model and assume that the zero mean errors are iid with $E(e_i) = 0$ and $\text{VAR}(e_i) = \sigma^2$. If the $\boldsymbol{x}_i$ are random vectors, assume that the cases $(\boldsymbol{x}_i, Y_i)$ are independent, and that the $\boldsymbol{e}_i$ and $\boldsymbol{x}_i$ are independent. Also assume that $\max_i(h_1, ..., h_n) \to 0$ and

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{V}^{-1}$$

as $n \to \infty$ where the convergence is in probability if the $\boldsymbol{x}_i$ are random vectors (instead of nonstochastic constant vectors).

a) For Equation (2.2), the OLS estimator $\hat{\boldsymbol{\beta}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{V}). \tag{2.6}$$

Equivalently,

$$(\boldsymbol{X}^T\boldsymbol{X})^{1/2}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{I}_p). \tag{2.7}$$

b) For Equation (2.3), the OLS estimator $\hat{\boldsymbol{\phi}}$ satisfies

$$\sqrt{n}(\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}) \xrightarrow{D} N_{p+1}(\boldsymbol{0}, \sigma^2\ \boldsymbol{V}). \tag{2.8}$$

c) Suppose the cases $(\boldsymbol{x}_i, Y_i)$ are iid from some population and the Equation (2.3) MLR model $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ holds. Assume that $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}$ and $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ exist. Then Equation (2.8) holds and

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\ \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}) \tag{2.9}$$

where $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$.

Remark 2.4. I) Consider Theorem 2.3. For a) and b), the theory acts as if the $\boldsymbol{x}_i$ are constant even if the $\boldsymbol{x}_i$ are random vectors. The literature says the $\boldsymbol{x}_i$ can be constants, or condition on $\boldsymbol{x}_i$ if the $\boldsymbol{x}_i$ are random vectors. The main assumptions for a) and b) are that the errors are iid with second moments and that $n(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ is well behaved. The strong assumptions for c) are much stronger than those for a) and b), but the assumption of iid cases is often reasonable if the cases come from some population.
II) Suppose $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ where the e_i are iid. Then $\hat{\boldsymbol{\beta}}_{OLS} \approx N_p(\boldsymbol{\beta}, MSE\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{-1}/n)$ even if the cases are not iid, and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{V}_{22}^{-1}$, where $\boldsymbol{V}_{22}^{-1}$ is not necessarily equal to $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, by Remark 2.3. Thus
$(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})^T\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})/MSE \xrightarrow{D} \chi_p^2$ as $n \to \infty$. This result is useful since no matrix inversion is required.

Remark 2.5. Consider MLR model (2.3). Let $\boldsymbol{w}_i = \boldsymbol{A}_n\boldsymbol{x}_i$ for $i = 1, ..., n$ where $\boldsymbol{A}_n$ is a full rank $k \times p$ matrix with $1 \le k \le p$.

a) Let $\boldsymbol{\Sigma}^*$ be $\hat{\boldsymbol{\Sigma}}$ or $\tilde{\boldsymbol{\Sigma}}$. Then $\boldsymbol{\Sigma}_{\boldsymbol{w}}^* = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}}^*\boldsymbol{A}_n^T$ and $\boldsymbol{\Sigma}_{\boldsymbol{w}Y}^* = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^*$.

b) If $\boldsymbol{A}_n$ is a constant matrix, then $\boldsymbol{\Sigma}_{\boldsymbol{w}} = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}_n^T$ and
$\boldsymbol{\Sigma}_{\boldsymbol{w}Y} = \boldsymbol{A}_n\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$.

c) Let $\hat{\boldsymbol{\beta}}(\boldsymbol{u}, Y)$ and $\boldsymbol{\beta}(\boldsymbol{u}, Y)$ be the estimator and parameter from the OLS regression of Y on $\boldsymbol{u}$. The constant parameter vector should not depend on n. Suppose the cases are iid and $\boldsymbol{A}$ is a constant matrix that does not depend on n. By Theorem 2.2, $\hat{\boldsymbol{\beta}}(\boldsymbol{w}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Y} = [\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\boldsymbol{A}_n]^{-1}\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = [\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\boldsymbol{A}_n]^{-1}\boldsymbol{A}_n\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}(\boldsymbol{x}, Y)$. If $\boldsymbol{A}_n \xrightarrow{P} \boldsymbol{A}$, $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{x}}$, and $\hat{\boldsymbol{\beta}}(\boldsymbol{x}, Y) \xrightarrow{P} \boldsymbol{\beta}(\boldsymbol{x}, Y)$, then $\hat{\boldsymbol{\beta}}(\boldsymbol{w}, Y) \xrightarrow{P} \boldsymbol{\beta}(\boldsymbol{w}, Y) = [\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}]^{-1}\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}(\boldsymbol{x}, Y)$.

A problem with OLS, is that $\boldsymbol{V}$ generally can't be estimated if $p > n$ since typically $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ does not exist. If $p > n$, using $\hat{\boldsymbol{\phi}} = (\boldsymbol{X}^T\boldsymbol{X})^-\boldsymbol{X}^T\boldsymbol{Y}$ is a poor estimator that interpolates the data, where $\boldsymbol{A}^-$ is a generalized inverse of $\boldsymbol{A}$. Often the software will not compute $\hat{\boldsymbol{\phi}}$ if $p > n$.

2.2 Statistical Learning Methods for MLR

There are many MLR methods, including OLS for the full model, forward selection with OLS, the marginal maximum likelihood estimator (MMLE), elastic net, principal components regression (PCR), partial least squares (PLS), lasso, lasso variable selection, and ridge regression (RR). For the last six methods, it is often convenient to use centered or scaled data. Suppose U has observed values $U_1, ..., U_n$. For example, if $U_i = Y_i$ then U corresponds to the response variable Y. The observed values of a random variable V are *centered* if their sample mean is 0. The centered values of U are $V_i = U_i - \overline{U}$ for $i = 1, ..., n$. Let g be an integer near 0. If the sample variance of the U_i is

$$\hat{\sigma}_g^2 = \frac{1}{n-g}\sum_{i=1}^{n}(U_i - \overline{U})^2,$$

then the sample standard deviation of U_i is $\hat{\sigma}_g$. If the values of U_i are not all the same, then $\hat{\sigma}_g > 0$, and the standardized values of the U_i are

$$W_i = \frac{U_i - \overline{U}}{\hat{\sigma}_g}.$$

Typically $g = 1$ or $g = 0$ are used: $g = 1$ gives an unbiased estimator of σ^2 while $g = 0$ gives the method of moments estimator. Note that the standardized values are centered, $\overline{W} = 0$, and the sample variance of the standardized values

$$\frac{1}{n-g}\sum_{i=1}^{n} W_i^2 = 1. \tag{2.10}$$

Remark 2.6. Let $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Let $\boldsymbol{w}_i^T = (w_{i,1}, ..., w_{i,p})$ be the standardized vector of nontrivial predictors for the ith case. Since the standardized predictors are also centered, $\overline{\boldsymbol{w}} = \boldsymbol{0}$. Let the $n \times p$ matrix of standardized nontrivial predictors $\boldsymbol{W}_g = (W_{ij})$ when the predictors are standardized using $\hat{\sigma}_g$. Then the ith row of $\boldsymbol{W}_g$ is $\boldsymbol{w}_i^T$. Thus, $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n-g$ for $j = 1, ..., p$. Hence

$$W_{ij} = \frac{x_{i,j} - \overline{x}_j}{\hat{\sigma}_j} \quad \text{where} \quad \hat{\sigma}_{\mathrm{j}}^2 = \frac{1}{\mathrm{n-g}}\sum_{\mathrm{i=1}}^{\mathrm{n}}(\mathrm{x_{i,j}} - \overline{\mathrm{x}}_{\mathrm{j}})^2$$

is $\hat{\sigma}_g$ for the jth variable x_j. Then the sample covariance matrix of the $\boldsymbol{w}_i$ is the sample correlation matrix of the $\boldsymbol{x}_i$:

$$\hat{\boldsymbol{\rho}}_{\boldsymbol{x}} = \boldsymbol{R}_{\boldsymbol{x}} = (r_{ij}) = \frac{\boldsymbol{W}_g^T\boldsymbol{W}_g}{n-g}$$

where r_{ij} is the sample correlation of x_i and x_j. Thus the sample correlation matrix $\boldsymbol{R_x}$ does not depend on g. Let $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$. Since the R software tends to use $g = 0$, let $\boldsymbol{W} = \boldsymbol{W}_0$. Note that $n \times p$ matrix $\boldsymbol{W}$ does not include a vector $\mathbf{1}$ of ones. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon} \tag{2.11}$$

where $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)^T$. The vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Remark 2.7. i) Interest is in model (2.3): estimate $\hat{Y}_f$ and $\hat{\boldsymbol{\beta}}$. For many regression estimators, a method is needed so that everyone who uses the same units of measurements for the predictors and Y gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. Equation (2.11) is a commonly used method for achieving this goal. Suppose $g = 0$. The method of moments estimator of the variance σ_w^2 is

$$\hat{\sigma}^2_{g=0} = S_M^2 = \frac{1}{n}\sum_{i=1}^{n}(w_i - \overline{w})^2.$$

When data x_i are standardized to have $\overline{w} = 0$ and $S_M^2 = 1$, the standardized data w_i has no units. ii) Hence the estimators $\hat{\boldsymbol{Z}}$ and $\hat{\boldsymbol{\eta}}$ do not depend on the units of measurement of the x_i if standardized data and Equation (2.11) are used. Linear combinations of the $\boldsymbol{w}_i$ are linear combinations of the $\boldsymbol{x}_i$. Thus the estimators $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}$ are obtained using $\hat{\boldsymbol{Z}}$, $\hat{\boldsymbol{\eta}}$, and $\overline{\boldsymbol{Y}}$. The linear transformation to obtain $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$ from $(\hat{\boldsymbol{Z}}, \hat{\boldsymbol{\eta}})$ is unique for a given set of units of measurements for the x_i and Y. Hence everyone using the same units of measurements gets the same $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. iii) Also, since $\overline{W}_j = 0$ and $S^2_{M,j} = 1$, the standardized predictor variables have similar spread, and the magnitude of $\hat{\eta}_i$ is a measure of the importance of the predictor variable W_j for predicting Y.

Definition 2.11. Consider model (2.2): $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$. If $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, where the $n \times q$ matrix $\boldsymbol{W}$ has full rank $q = p - 1$, then the *OLS estimator*

$$\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

minimizes the OLS criterion $Q_{OLS}(\boldsymbol{\eta}) = \boldsymbol{r}(\boldsymbol{\eta})^T\boldsymbol{r}(\boldsymbol{\eta})$ over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$. The vector of *predicted* or *fitted values* $\widehat{\boldsymbol{Z}}_{OLS} = \boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{H}\boldsymbol{Z}$ where $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T$. The vector of residuals $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{W}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}$.

For model (2.2): $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, let $\boldsymbol{x} = (1\ \boldsymbol{u})^T$, and let $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. Assume that the sample correlation matrix

$$\boldsymbol{R_u} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}. \tag{2.12}$$

Note that $\boldsymbol{V}^{-1} = \boldsymbol{\rho}_{\boldsymbol{u}}$, the population correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$, if the $\boldsymbol{u}_i$ are a random sample from a population. Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,\ldots,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Olive (2024) examines whether the OLS estimator satisfies

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{2.13}$$

Remark 2.8. Variable selection is the search for a subset of predictor variables that can be deleted without important loss of information if n/p is large (and the search for a useful subset of predictors if n/p is not large). Refer to Chapter 1: Remark 1.1 for variable selection and Equation (1.1) where

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S. \tag{2.14}$$

Let p be the number of predictors in the full model, including a constant. Let $q = p - 1$ be the number of nontrivial predictors in the full model. Let $a = a_I$ be the number of predictors in the submodel I, including a constant. Let $k = k_I = a_I - 1$ be the number of nontrivial predictors in the submodel. For submodel I, think of I as indexing the predictors in the model, including the constant. Let A index the nontrivial predictors in the model. Hence I adds the constant (trivial predictor) to the collection of nontrivial predictors in A. In Equation (2.14), there is a "true submodel" $\boldsymbol{Y} = \boldsymbol{X}_S\boldsymbol{\beta}_S + \boldsymbol{e}$ where all of the elements of $\boldsymbol{\beta}_S$ are nonzero but all of the elements of $\boldsymbol{\beta}$ that are not elements of $\boldsymbol{\beta}_S$ are zero. Then $a = a_S$ is the number of predictors in that submodel, including a constant, and $k = k_S$ is the number of active predictors = number of nonnoise variables = number of nontrivial predictors in the true model $S = I_S$. Then there are $p - a$ noise variables (x_i that have coefficient $\beta_i = 0$) in the full model. The true model is generally only known in simulations. For Equation (2.14), we also assume that if $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$, then $S \subseteq I$. Hence S is the unique smallest subset of predictors such that $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$.

Model selection generates M models. Then a hopefully good model is selected from these M models. Variable selection is a special case of model selection. Many methods for variable and model selection have been suggested for the MLR model. We will consider several R functions including i) forward selection computed with the `regsubsets` function from the `leaps` library, ii) principal components regression (PCR) with the `pcr` function from the `pls` library, iii) partial least squares (PLS) with the `plsr` function from the `pls` library, iv) ridge regression with the `cv.glmnet` or `glmnet` function from the `glmnet` library, v) lasso with the `cv.glmnet` or `glmnet` function from the `glmnet` library, and vi) lasso variable selection which is OLS applied to the lasso active set (nontrivial predictors with nonzero coefficients) and a constant. See Sections 2.3–2.12 and James et al. (2013, ch. 6).

These six methods produce M models and use a criterion to select the final model (e.g. C_p or 10-fold cross validation (CV)). See Section 2.14. The

number of models M depends on the method. Often one of the models is the full model (2.3) that uses all $p-1$ nontrivial predictors. The full model is (approximately) fit with (ordinary) least squares. For one of the M models, some of the methods use $\hat{\boldsymbol{\eta}} = \mathbf{0}$ and fit the model $Y_i = \beta_1 + e_i$ with $\hat{Y}_i \equiv \overline{Y}$ that uses none of the nontrivial predictors. Forward selection, PCR, and PLS use variables $v_1 = 1$ (the constant or trivial predictor) and $v_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ that are linear combinations of the predictors for $j = 2, ..., p$. Model I_i uses variables $v_1, v_2, ..., v_i$ for $i = 1, ..., M$ where $M \leq p$ and often $M \leq \min(p, n/10)$. Then M models I_i are used. (For forward selection and PCR, OLS is used to regress Y (or Z) on $v_1, ..., v_i$.) Then a criterion chooses the final submodel I_d from candidates $I_1, ..., I_M$.

Overfitting or "fitting noise" occurs when there is not enough data to estimate the $p \times 1$ vector $\boldsymbol{\beta}$ well with the estimation method, such as OLS. The OLS model is overfitting if $n < 5p$. When $n < p$, $\boldsymbol{X}^T\boldsymbol{X}$ is usually not invertible, but if $n = p$, then $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{I}_n\boldsymbol{Y} = \boldsymbol{Y}$ regardless of how bad the predictors are. If $n < p$, then the OLS program fails or $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$: the fitted regression plane interpolates the training data response variables $Y_1, ..., Y_n$. The following rule of thumb is useful for many regression methods. Note that $d = p$ for the full OLS model.

Rule of thumb 2.1. We want $n \geq 10d$ to avoid overfitting. Occasionally n as low as $5d$ is used, but models with $n < 5d$ are overfitting.

Remark 2.9. Use $\boldsymbol{Z}_n \sim AN_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_r(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let a be a constant, let $\boldsymbol{A}$ be a $k \times r$ constant matrix (often with full rank $k \leq r$), and let $\boldsymbol{c}$ be a $k \times 1$ constant vector. If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_r(\mathbf{0}, \boldsymbol{V})$, then $a\boldsymbol{Z}_n = a\boldsymbol{I}_r\boldsymbol{Z}_n$ with $\boldsymbol{A} = a\boldsymbol{I}_r$,

$$a\boldsymbol{Z}_n \sim AN_r\left(a\boldsymbol{\mu}_n, a^2\boldsymbol{\Sigma}_n\right), \quad \text{and} \quad \boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\mu}_n + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}_n\boldsymbol{A}^T\right),$$

$$\hat{\boldsymbol{\theta}}_n \sim AN_r\left(\boldsymbol{\theta}, \frac{\boldsymbol{V}}{n}\right), \quad \text{and} \quad \boldsymbol{A}\hat{\boldsymbol{\theta}}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{c}, \frac{\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T}{n}\right).$$

Theorem 2.3 gives the large sample theory for the OLS full model. Then $\hat{\boldsymbol{\beta}} \approx N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}))$ or $\hat{\boldsymbol{\beta}} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1}))$.

When minimizing or maximizing a real valued function $Q(\boldsymbol{\eta})$ of the $k \times 1$ vector $\boldsymbol{\eta}$, the solution $\hat{\boldsymbol{\eta}}$ is found by setting the gradient of $Q(\boldsymbol{\eta})$ equal to $\mathbf{0}$. The following definition and lemma follow Graybill (1983, pp. 351-352) closely. Maximum likelihood estimators are examples of estimating equations. There is a vector of parameters $\boldsymbol{\eta}$, and the gradient of the log likelihood function $\log L(\boldsymbol{\eta})$ is set to zero. The solution $\hat{\boldsymbol{\eta}}$ is the MLE, an estimator of the parameter vector $\boldsymbol{\eta}$, but in the log likelihood, $\boldsymbol{\eta}$ is a dummy variable vector, not the fixed unknown parameter vector.

Definition 2.12. Let $Q(\boldsymbol{\eta})$ be a real valued function of the $k \times 1$ vector $\boldsymbol{\eta}$. The gradient of $Q(\boldsymbol{\eta})$ is the $k \times 1$ vector

$$\bigtriangledown Q = \bigtriangledown Q(\boldsymbol{\eta}) = \frac{\partial Q}{\partial \boldsymbol{\eta}} = \frac{\partial Q(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \frac{\partial}{\partial \eta_1} Q(\boldsymbol{\eta}) \\ \frac{\partial}{\partial \eta_2} Q(\boldsymbol{\eta}) \\ \vdots \\ \frac{\partial}{\partial \eta_k} Q(\boldsymbol{\eta}) \end{bmatrix}.$$

Suppose there is a model with unknown parameter vector $\boldsymbol{\eta}$. A set of *estimating equations* $f(\boldsymbol{\eta})$ is used to maximize or minimize $Q(\boldsymbol{\eta})$ where $\boldsymbol{\eta}$ is a dummy variable vector.

Often $f(\boldsymbol{\eta}) = \bigtriangledown Q$, and we solve $f(\boldsymbol{\eta}) = \bigtriangledown Q \stackrel{set}{=} \mathbf{0}$ for the solution $\hat{\boldsymbol{\eta}}$, and $f : \mathbb{R}^k \to \mathbb{R}^k$. Note that $\hat{\boldsymbol{\eta}}$ is an estimator of the unknown parameter vector $\boldsymbol{\eta}$ in the model, but $\boldsymbol{\eta}$ is a dummy variable in $Q(\boldsymbol{\eta})$. Hence we could use $Q(\boldsymbol{b})$ instead of $Q(\boldsymbol{\eta})$, but the solution of the estimating equations would still be $\hat{\boldsymbol{b}} = \hat{\boldsymbol{\eta}}$.

As a mnemonic (memory aid) for the following theorem, note that the derivative $\frac{d}{dx} ax = \frac{d}{dx} xa = a$ and $\frac{d}{dx} ax^2 = \frac{d}{dx} xax = 2ax$.

Theorem 2.4. a) If $Q(\boldsymbol{\eta}) = \boldsymbol{a}^T\boldsymbol{\eta} = \boldsymbol{\eta}^T\boldsymbol{a}$ for some $k \times 1$ constant vector $\boldsymbol{a}$, then $\bigtriangledown Q = \boldsymbol{a}$.

b) Let $\boldsymbol{A}$ be a symmetric matrix. If $Q(\boldsymbol{\eta}) = \boldsymbol{\eta}^T\boldsymbol{A}\boldsymbol{\eta}$ for some $k \times k$ constant matrix $\boldsymbol{A}$, then $\bigtriangledown Q = 2\boldsymbol{A}\boldsymbol{\eta}$.

c) If $Q(\boldsymbol{\eta}) = \sum_{i=1}^{k} |\eta_i| = \|\boldsymbol{\eta}\|_1$, then $\bigtriangledown Q = \boldsymbol{s} = \boldsymbol{s}\boldsymbol{\eta}$ where $s_i = \text{sign}(\eta_i)$ where $\text{sign}(\eta_i) = 1$ if $\eta_i > 0$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 0$. This gradient is only defined for $\boldsymbol{\eta}$ where none of the k values of η_i are equal to 0.

Example 2.1. If $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, then the OLS estimator minimizes $Q(\boldsymbol{\eta}) = \|\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}\|_2^2 = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) = \boldsymbol{Z}^T\boldsymbol{Z} - 2\boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\eta}^T(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta}$. Using Theorem 2.4 with $\boldsymbol{a}^T = \boldsymbol{Z}^T\boldsymbol{W}$ and $\boldsymbol{A} = \boldsymbol{W}^T\boldsymbol{W}$ shows that $\bigtriangledown Q = -2\boldsymbol{W}^T\boldsymbol{Z} + 2(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta}$. Let $\bigtriangledown Q(\hat{\boldsymbol{\eta}})$ denote the gradient evaluated at $\hat{\boldsymbol{\eta}}$. Then the OLS estimator satisfies the normal equations $(\boldsymbol{W}^T\boldsymbol{W})\hat{\boldsymbol{\eta}} = \boldsymbol{W}^T\boldsymbol{Z}$.

Example 2.2. The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. We will study the relationship between $Y =$ the *number of women married to civilians* in the district with the predictors x_1 = constant, $x_2 = pop =$ the *population of the district in 1843*, $x_3 = mmen$ = the *number of married civilian men* in the district, $x_4 = mmilmen =$ the *number of married men in the military* in the district, and $x_5 = milwmn =$ the *number of women married to husbands in the military* in the district. Sometimes the person conducting the survey would not count a spouse if the spouse was not at home. Hence Y is highly correlated but not equal to

x_3. Similarly, x_4 and x_5 are highly correlated but not equal. We expect that $Y = x_3 + e$ is a good model, but $n/p = 5.2$ is small. See the following output.

```
source("http://parker.ad.siu.edu/Olive/hdpack.txt")
source("http://parker.ad.siu.edu/Olive/hddata.txt")
x <- marry[,-3]; Y <- marry[,3]; out<-lsfit(x,Y)
ls.print(out)
Residual Standard Error=392.8709
R-Square=0.9999, p-value=0
F-statistic (df=4, 21)=67863.03
          Estimate  Std.Err t-value Pr(>|t|)
Intercept 242.3910 263.7263  0.9191   0.3685
pop         0.0004   0.0031  0.1130   0.9111
mmen        0.9995   0.0173 57.6490   0.0000
mmilmen    -0.2328   2.6928 -0.0864   0.9319
milwmn      0.1531   2.8231  0.0542   0.9572
res<-out$res
yhat<-Y-res #d = 5 predictors used including x_1
AERplot2(yhat,Y,res=res,d=5)
#response plot with 90% pointwise PIs
$respi #90% PI for a future residual
[1] -950.4811 1445.2584 #90% PI length = 2395.74
```

2.3 Forward Selection

Forward selection is a variable selection method where model I_j uses j predictors $x_1^*, ..., x_j^*$ including the constant $x_1^* \equiv 1$. If n/p is not large, instead of forming p submodels $I_1, ..., I_p$, form the sequence of M submodels $I_1, ..., I_M$ where $M = \min(\lceil n/J \rceil, p)$ for some positive integer J such as $J = 5, 10$, or 20. Here $\lceil x \rceil$ is the smallest integer $\geq x$, e.g., $\lceil 7.7 \rceil = 8$. Then for each submodel I_j, OLS is used to regress Y on 1, $x_2^*, ..., x_j^*$. Then a criterion chooses which model I_d from candidates $I_1, ..., I_M$ is to be used as the final submodel.

Let criteria $C_S(I)$ have the form

$$C_S(I) = SSE(I) + aK_n\hat{\sigma}^2.$$

These criteria need a good estimator of σ^2 and n/p large. See Shibata (1984). The criterion $C_p(I) = AIC_S(I)$ uses $K_n = 2$ while the $BIC_S(I)$ criterion uses $K_n = \log(n)$. See Jones (1946) and Mallows (1973) for C_p. It can be shown that $C_p(I) = AIC_S(I)$ is equivalent to the $C_P(I)$ criterion of Definition 2.27. Typically $\hat{\sigma}^2$ is the OLS full model MSE when n/p is large.

The following criteria also need n/p large. AIC is due to Akaike (1973), AIC_C is due to Hurvich and Tsai (1989), and BIC to Schwarz (1978) and

Akaike (1977, 1978). Also see Burnham and Anderson (2004).

$$AIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + 2a,$$

$$AIC_C(I) = n \log\left(\frac{SSE(I)}{n}\right) + \frac{2a(a+1)}{n-a-1},$$

$$\text{and } BIC(I) = n \log\left(\frac{SSE(I)}{n}\right) + a \log(n).$$

Suppose the selected model is I_d, and $\boldsymbol{\beta}_{I_d}$ is $a_d \times 1$. Forward selection with C_p and AIC often gives useful results if $n \geq 5p$ and if $n \geq 10a_d$. For $p < n < 5p$, forward selection with C_p and AIC tends to pick the full model (which overfits since $n < 5p$) too often, especially if $\hat{\sigma}^2 = MSE$. The Hurvich and Tsai (1989, 1991) AIC_C criterion can be useful if $n \geq \max(2p, 10a_d)$.

The EBIC criterion given in Luo and Chen (2013) may be useful when n/p is not large. Let $0 \leq \gamma \leq 1$ and $|I| = a \leq \min(n, p)$ if $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$. We may use $a \leq \min(n/5, p)$. Then $EBIC(I) =$

$$n \log\left(\frac{SSE(I)}{n}\right) + a \log(n) + 2\gamma \log\left[\binom{p}{a}\right] = BIC(I) + 2\gamma \log\left[\binom{p}{a}\right].$$

This criterion can give good results if $p = p_n = O(n^k)$ and $\gamma > 1 - 1/(2k)$. Hence we will use $\gamma = 1$. Then minimizing $EBIC(I)$ is equivalent to minimizing $BIC(I) - 2\log[(p-a)!] - 2\log(a!)$ since $\log(p!)$ is a constant.

The above criteria can be applied to forward selection and lasso variable selection. The C_p criterion can also be applied to lasso. See Efron and Hastie (2016, pp. 221, 231).

Remark 2.10. Suppose n/J is an integer. If $p \leq n/J$, then forward selection fits $(p-1) + (p-2) + \cdots + 2 + 1 = p(p-1)/2 \approx p^2/2$ models, where $p - i$ models are fit at step i for $i = 1, ..., (p-1)$. If $n/J < p$, then forward selection uses $(n/J) - 1$ steps and fits $\approx (p-1) + (p-2) + \cdots + (p-(n/J)+1) = p((n/J) - 1) - (1 + 2 + \cdots + ((n/J) - 1)) =$

$$p(\frac{n}{J} - 1) - \frac{\frac{n}{J}(\frac{n}{J} - 1)}{2} \approx \frac{n}{J} \ \frac{(2p - \frac{n}{J})}{2}$$

models. Thus forward selection can be slow if n and p are both large, although the R package `leaps` uses a branch and bound algorithm that likely eliminates many of the possible fits. Note that after step i, the model has $i + 1$ predictors, including the constant.

The R function `regsubsets` can be used for forward selection if $p < n$, and if $p \geq n$ if the maximum number of variables is less than n. Then warning messages are common. Some R code is shown below.

```
#regsubsets works if p < n, e.g. p = n-1, and works
```

```
#if p > n with warnings if nvmax is small enough
set.seed(13)
n<-100
p<-200
k<-19 #the first 19 nontrivial predictors are active
J<-5
q <- p-1
b <- 0 * 1:q
b[1:k] <- 1 #beta = (1, 1, ..., 1, 0, 0, ..., 0)^T
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + x %*% b + rnorm(n)
nc <- ceiling(n/J)-1 #the constant will also be used
nc <- min(nc,q)
nc <- max(nc,1) #nc is the maximum number of
#nontrivial predictors used by forward selection
pp <- nc+1  #d = pp is used for PI (2.14)
vars <- as.vector(1:(p-1))
temp<-regsubsets(x,y,nvmax=nc,method="forward")
out<-summary(temp)
num <- length(out$cp)
mod <- out$which[num,] #use the last model
#do not need the constant in vin
vin <- vars[mod[-1]]
```

```
out$rss
 [1] 1496.49625 1342.95915 1214.93174 1068.56668
     973.36395  855.15436  745.35007  690.03901
     638.40677  590.97644  542.89273  503.68666
     467.69423  420.94132  391.41961  328.62016
     242.66311  178.77573   79.91771
out$bic
 [1]   -9.4032  -15.6232  -21.0367  -29.2685
      -33.9949  -42.3374  -51.4750  -54.5804
      -57.7525  -60.8673  -64.7485  -67.6391
      -70.4479  -76.3748  -79.0410  -91.9236
     -117.6413 -143.5903 -219.498595
tem <- lsfit(x[,1:19],y) #last model used the
sum(tem$resid^2)         #first 19 predictors
[1] 79.91771             #SSE(I) = RSS(I)
n*log(out$rss[19]/n) + 20*log(n)
[1] 69.68613             #BIC(I)
for(i in 1:19)   #a formula for BIC(I)
print( n*log(out$rss[i]/n) + (i+1)*log(n) )
bic <- c(279.7815, 273.5616, 268.1480, 259.9162,
255.1898, 246.8474, 237.7097, 234.6043, 231.4322,
228.3175, 224.4362, 221.5456, 218.7368, 212.8099,
```

```
210.1437, 197.2611, 171.5435, 145.5944,  69.6861)
tem<-lsfit(bic,out$bic)
tem$coef
   Intercept            X
-289.1846831    0.9999998 #bic - 289.1847 = out$bic
xx <- 1:min(length(out$bic),p-1)+1
ebic <- out$bic+2*log(dbinom(x=xx,size=p,prob=0.5))
#actually EBIC(I) - 2 p log(2).
```

Example 2.2, continued. The output below shows results from forward selection for the marry data. The minimum C_p model I_{min} uses a constant and *mmem*. The forward selection PIs are shorter than the OLS full model PIs.

```
library(leaps);Y <- marry[,3]; X <- marry[,-3]
temp<-regsubsets(X,Y,method="forward")
out<-summary(temp)
Selection Algorithm: forward
         pop mmen mmilmen milwmn
1  ( 1 ) " " "*"  " "     " "
2  ( 1 ) " " "*"  "*"     " "
3  ( 1 ) "*" "*"  "*"     " "
4  ( 1 ) "*" "*"  "*"     "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
#mmen and a constant = Imin
mincp <- out$which[out$cp==min(out$cp),]
#do not need the constant in vin
vin <- vars[mincp[-1]]
sub <- lsfit(X[,vin],Y)
ls.print(sub)
Residual Standard Error=369.0087
R-Square=0.9999
F-statistic (df=1, 24)=307694.4
          Estimate  Std.Err  t-value Pr(>|t|)
Intercept 241.5445 190.7426   1.2663   0.2175
X           1.0010   0.0018 554.7021   0.0000
res<-sub$res
yhat<-Y-res #d = 2 predictors used including x_1
AERplot2(yhat,Y,res=res,d=2)
#response plot with 90% pointwise PIs
$respi   #90% PI for a future residual
[1] -778.2763 1336.4416 #length 2114.72
```

Consider forward selection where $\boldsymbol{x}_I$ is $a \times 1$. Underfitting occurs if S is not a subset of I so $\boldsymbol{x}_I$ is missing important predictors. A special case

of underfitting is $d = a < a_S$. Overfitting for forward selection occurs if i) $n < 5a$ so there is not enough data to estimate the a parameters in $\boldsymbol{\beta}_I$ well, or ii) $S \subseteq I$ but $S \neq I$. Overfitting is serious if $n < 5a$, but "not much of a problem" if $n > Jp$ where $J = 10$ or 20 for many data sets. Underfitting is a serious problem for estimating the full model $\boldsymbol{\beta}$. Let $Y_i = \boldsymbol{x}_{I,i}^T \boldsymbol{\beta}_I + e_{I,i}$. Then $V(e_{I,i})$ may not be a constant σ^2: $V(e_{I,i})$ could depend on case i, and the model may no longer be linear. Check model I with response and residual plots.

Forward selection is a *shrinkage* method: p models are produced and except for the full model, some $|\hat{\beta}_i|$ are shrunk to 0. Lasso and ridge regression are also shrinkage methods. Ridge regression is a shrinkage method, but $|\hat{\beta}_i|$ is not shrunk to 0. Shrinkage methods that shrink $\hat{\beta}_i$ to 0 are also variable selection methods. See Sections 2.6, 2.7, and 2.8.

Definition 2.13. A fitted or population regression model is *sparse* if a of the predictors are active (have nonzero $\hat{\beta}_i$ or β_i) where $n \geq Ja$ with $J \geq 10$. Otherwise the model is *nonsparse*. A high dimensional population regression model is *abundant* or *dense* if the regression information is spread out among the p predictors (nearly all of the predictors are active). Hence an abundant model is a nonsparse model.

Suppose the population model has $\boldsymbol{\beta}_S$ an $a_S \times 1$ vector, including a constant. Then $a = a_S - 1$ for the population model. Note that $a = a_S$ if the model does not include a constant. See Equation (2.14).

2.4 Principal Components Regression

Some notation for eigenvalues, eigenvectors, orthonormal eigenvectors, positive definite matrices, and positive semidefinite matrices will be useful before defining principal components regression, which is also called principal component regression.

Notation: Recall that a square symmetric $p \times p$ matrix $\boldsymbol{A}$ has an *eigenvalue* λ with corresponding *eigenvector* $\boldsymbol{x} \neq \boldsymbol{0}$ if

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{2.15}$$

The eigenvalues of $\boldsymbol{A}$ are real since $\boldsymbol{A}$ is symmetric. Note that if constant $c \neq 0$ and $\boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$, then $c\ \boldsymbol{x}$ is an eigenvector of $\boldsymbol{A}$. Let $\boldsymbol{e}$ be an eigenvector of $\boldsymbol{A}$ with unit length $\|\boldsymbol{e}\|_2 = \sqrt{\boldsymbol{e}^T \boldsymbol{e}} = 1$. Then $\boldsymbol{e}$ and $-\boldsymbol{e}$ are eigenvectors with unit length, and $\boldsymbol{A}$ has p eigenvalue eigenvector pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$. Since $\boldsymbol{A}$ is symmetric, the eigenvectors are chosen such that the $\boldsymbol{e}_i$ are *orthonormal*: $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ for $i \neq j$. The symmetric matrix $\boldsymbol{A}$ is *positive definite* iff all of its eigenvalues are

positive, and *positive semidefinite* iff all of its eigenvalues are nonnegative. If $\boldsymbol{A}$ is positive semidefinite, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. If $\boldsymbol{A}$ is positive definite, then $\lambda_p > 0$.

Theorem 2.5. Let $\boldsymbol{A}$ be a $p \times p$ symmetric matrix with eigenvector eigenvalue pairs $(\lambda_1, \boldsymbol{e}_1), (\lambda_2, \boldsymbol{e}_2), ..., (\lambda_p, \boldsymbol{e}_p)$ where $\boldsymbol{e}_i^T \boldsymbol{e}_i = 1$ and $\boldsymbol{e}_i^T \boldsymbol{e}_j = 0$ if $i \neq j$ for $i = 1, ..., p$. Then the *spectral decomposition* of $\boldsymbol{A}$ is

$$\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T = \lambda_1 \boldsymbol{e}_1 \boldsymbol{e}_1^T + \cdots + \lambda_p \boldsymbol{e}_p \boldsymbol{e}_p^T.$$

Using the same notation as Johnson and Wichern (1988, pp. 50-51), let $\boldsymbol{P} = [\boldsymbol{e}_1 \ \boldsymbol{e}_2 \ \cdots \ \boldsymbol{e}_p]$ be the $p \times p$ orthogonal matrix with ith column $\boldsymbol{e}_i$. Then $\boldsymbol{P}\boldsymbol{P}^T = \boldsymbol{P}^T\boldsymbol{P} = \boldsymbol{I}$. Let $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, ..., \lambda_p)$ and let $\boldsymbol{\Lambda}^{1/2} = \text{diag}(\sqrt{\lambda_1}, ..., \sqrt{\lambda_p})$. If $\boldsymbol{A}$ is a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$, then $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}^T$ and

$$\boldsymbol{A}^{-1} = \boldsymbol{P}\boldsymbol{\Lambda}^{-1}\boldsymbol{P}^T = \sum_{i=1}^{p} \frac{1}{\lambda_i} \boldsymbol{e}_i \boldsymbol{e}_i^T.$$

Theorem 2.6. Let $\boldsymbol{A}$ be a positive definite $p \times p$ symmetric matrix with spectral decomposition $\boldsymbol{A} = \sum_{i=1}^{p} \lambda_i \boldsymbol{e}_i \boldsymbol{e}_i^T$. The *square root matrix* $\boldsymbol{A}^{1/2} = \boldsymbol{P}\boldsymbol{\Lambda}^{1/2}\boldsymbol{P}^T$ is a positive definite symmetric matrix such that $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Let $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Consider the correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}$ of the p nontrivial predictors $x_1, ..., x_p$. Suppose $\boldsymbol{R}_{\boldsymbol{x}}$ has eigenvalue eigenvector pairs $(\hat{\lambda}_1, \hat{\boldsymbol{e}}_1), ..., (\hat{\lambda}_K, \hat{\boldsymbol{e}}_K)$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_K \geq 0$ where $K = \min(n, p)$. Then $\boldsymbol{R}_{\boldsymbol{x}}\hat{\boldsymbol{e}}_i = \hat{\lambda}_i \hat{\boldsymbol{e}}_i$ for $i = 1, ..., K$. Since $\boldsymbol{R}_{\boldsymbol{x}}$ is a symmetric positive semidefinite matrix, the $\hat{\lambda}_i$ are real and nonnegative.

The eigenvectors $\hat{\boldsymbol{e}}_i$ are *orthonormal*: $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_i = 1$ and $\hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \neq j$. If the eigenvalues are unique, then $\hat{\boldsymbol{e}}_i$ and $-\hat{\boldsymbol{e}}_i$ are the only orthonormal eigenvectors corresponding to $\hat{\lambda}_i$. For example, the eigenvalue eigenvector pairs can be found using the singular value decomposition of the matrix $\boldsymbol{W}_g/\sqrt{n-g}$ where $\boldsymbol{W}_g$ is the data matrix of standardized cases: the ith row of $\boldsymbol{W}_g$ is $\boldsymbol{w}_i^T$, the sample covariance matrix

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{\boldsymbol{W}_g^T \boldsymbol{W}_g}{n-g} = \frac{1}{n-g} \sum_{i=1}^{n} (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \frac{1}{n-g} \sum_{i=1}^{n} \boldsymbol{w}_i \boldsymbol{w}_i^T = \boldsymbol{R}_{\boldsymbol{x}},$$

and usually $g = 0$ or $g = 1$. If $n > K = p$, then the *spectral decomposition* of $\boldsymbol{R}_{\boldsymbol{x}}$ is

$$\boldsymbol{R}_{\boldsymbol{x}} = \sum_{i=1}^{p} \hat{\lambda}_i \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^T = \hat{\lambda}_1 \hat{\boldsymbol{e}}_1 \hat{\boldsymbol{e}}_1^T + \cdots + \hat{\lambda}_p \hat{\boldsymbol{e}}_p \hat{\boldsymbol{e}}_p^T,$$

and $\sum_{i=1}^{p} \hat{\lambda}_i = p$.

Let $\boldsymbol{w}_1, ..., \boldsymbol{w}_n$ denote the n standardized cases of nontrivial predictors. See Remark 2.6. Then the K *principal components* corresponding to the jth case $\boldsymbol{w}_j$ are $P_{j1} = \hat{\boldsymbol{e}}_1^T \boldsymbol{w}_j$, ..., $P_{jK} = \hat{\boldsymbol{e}}_K^T \boldsymbol{w}_j$. Let the transformed case, that uses K principal components, corresponding to $\boldsymbol{w}_j$ be $\boldsymbol{v}_j = (P_{j1}, ..., P_{jK})^T$. Following Hastie et al. (2009, p. 66), the ith eigenvector $\hat{\boldsymbol{e}}_i$ is known as the ith *principal component direction* or *Karhunen Loeve direction* of $\boldsymbol{W}_g$.

Principal components have a nice geometric interpretation if $n > K = p$. If $n > K$ and $\boldsymbol{R_x}$ is nonsingular, then the hyperellipsoid

$$\{\boldsymbol{w} | D_{\boldsymbol{w}}^2(\boldsymbol{0}, \boldsymbol{R_x}) \leq h^2\} = \{\boldsymbol{w} : \boldsymbol{w}^T \boldsymbol{R_x}^{-1} \boldsymbol{w} \leq h^2\}$$

is centered at $\boldsymbol{0}$. The volume of the hyperellipsoid is

$$\frac{2\pi^{K/2}}{K\Gamma(K/2)} |\boldsymbol{R_x}|^{1/2} h^K.$$

Then points at squared distance $\boldsymbol{w}^T \boldsymbol{R_x}^{-1} \boldsymbol{w} = h^2$ from the origin lie on the hyperellipsoid centered at the origin whose axes are given by the eigenvectors $\hat{\boldsymbol{e}}_i$ where the half length in the direction of $\hat{\boldsymbol{e}}_i$ is $h\sqrt{\hat{\lambda}_i}$. Let $j = 1, ..., n$. Then the first principal component P_{j1} is obtained by projecting the $\boldsymbol{w}_j$ on the (longest) major axis of the hyperellipsoid, the second principal component P_{j2} is obtained by projecting the $\boldsymbol{w}_j$ on the next longest axis of the hyperellipsoid, ..., and the (p)th principal component $P_{j,p}$ is obtained by projecting the $\boldsymbol{w}_j$ on the (shortest) minor axis of the hyperellipsoid. Examine Figure 2.3 for two ellipsoids with 2 nontrivial predictors. The axes of the hyperellipsoid are a rotation of the usual axes about the origin.

Let the random variable V_i correspond to the ith principal component, and let the ith principal component vector $\boldsymbol{c}_i = (P_{1i}, ..., P_{ni})^T = (V_{1i}, ..., V_{ni})^T$ be the observed data for V_i. Let $g = 1$. Then the sample mean

$$\overline{V}_i = \frac{1}{n} \sum_{k=1}^{n} V_{ki} = \frac{1}{n} \sum_{k=1}^{n} \hat{\boldsymbol{e}}_i^T \boldsymbol{w}_k = \hat{\boldsymbol{e}}_i^T \overline{\boldsymbol{w}} = \hat{\boldsymbol{e}}_i^T \boldsymbol{0} = 0,$$

and the sample covariance of V_i and V_j is $Cov(V_i, V_j) =$

$$\frac{1}{n} \sum_{k=1}^{n} (V_{ki} - \overline{V}_i)(V_{kj} - \overline{V}_j) = \frac{1}{n} \sum_{k=1}^{n} \hat{\boldsymbol{e}}_i^T \boldsymbol{w}_k \boldsymbol{w}_k^T \hat{\boldsymbol{e}}_j = \hat{\boldsymbol{e}}_i^T \boldsymbol{R_x} \hat{\boldsymbol{e}}_j$$

$= \hat{\lambda}_j \hat{\boldsymbol{e}}_i^T \hat{\boldsymbol{e}}_j = 0$ for $i \neq j$ since the sample covariance matrix of the standardized data is

$$\frac{1}{n} \sum_{k=1}^{n} \boldsymbol{w}_k \boldsymbol{w}_k^T = \boldsymbol{R_x}$$

and $\boldsymbol{R_x} \hat{\boldsymbol{e}}_j = \hat{\lambda}_j \hat{\boldsymbol{e}}_j$. Hence V_i and V_j are uncorrelated.

In the following definition, note that $\boldsymbol{c}_i^T\boldsymbol{c}_j = \hat{\boldsymbol{e}}_i^T\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{e}}_j = n\hat{\boldsymbol{e}}_i\boldsymbol{R}_{\boldsymbol{x}}\hat{\boldsymbol{e}}_j = n\lambda_j\hat{\boldsymbol{e}}_i^T\hat{\boldsymbol{e}}_j = 0$ for $i \neq j$. Thus $\boldsymbol{c}_i$ and $\boldsymbol{c}_j$ are orthogonal: $\boldsymbol{c}_i \perp \boldsymbol{c}_j$ for $i \neq j$. Also, $\boldsymbol{c}_i^T\boldsymbol{1} = (\sum_{k=1}^n \boldsymbol{w}_k)\hat{\boldsymbol{e}}_i = \boldsymbol{0}^T\hat{\boldsymbol{e}}_i = 0$ since the standardized predictor variables sum to 0. The ith principle component vector $\boldsymbol{c}_i$ corresponds to the derived predictor V_i, for $i = 1, ..., p-1$.

Definition 2.14. Consider the standardized model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$ where $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Let

$$\boldsymbol{v}_i = \hat{\boldsymbol{A}}_{k,n}\boldsymbol{w}_i = \begin{pmatrix} \boldsymbol{w}_i^T\hat{\boldsymbol{e}}_1 \\ \vdots \\ \boldsymbol{w}_i^T\hat{\boldsymbol{e}}_k \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{e}}_1^T\boldsymbol{w}_i \\ \vdots \\ \hat{\boldsymbol{e}}_k^T\boldsymbol{w}_i^T \end{pmatrix} \text{ where } \hat{\boldsymbol{A}}_{\mathrm{k,n}} = \begin{pmatrix} \hat{\boldsymbol{e}}_1^T \\ \vdots \\ \hat{\boldsymbol{e}}_k^T \end{pmatrix}.$$

Let

$$\boldsymbol{c}_i = \boldsymbol{W}\hat{\boldsymbol{e}}_i = \begin{pmatrix} \boldsymbol{w}_1^T\hat{\boldsymbol{e}}_i \\ \vdots \\ \boldsymbol{w}_n^T\hat{\boldsymbol{e}}_i \end{pmatrix}$$

be the *ith principle component vector* for $i = 1, ..., p$. Principal components regression (PCR) uses OLS regression on the principal component vectors of the correlation matrix $\boldsymbol{R}_{\boldsymbol{x}}$. Hence PCR uses linear combinations of the standardized data as predictors. Let

$$\boldsymbol{V}_k = (\boldsymbol{c}_1, ..., \boldsymbol{c}_k) = \begin{pmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_n^T \end{pmatrix} = \boldsymbol{W}\hat{\boldsymbol{A}}_{k,n}^T$$

for $k = 1, ..., p$. Let the working OLS model

$$\boldsymbol{Z} = \boldsymbol{V}_k\boldsymbol{\gamma}_k + \boldsymbol{\epsilon} = \boldsymbol{W}\boldsymbol{\beta}_{kPCR} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}$ depends on the model. Then $\hat{\boldsymbol{\beta}}_{kPCR}$ is the k-component PCR estimator for $k = 1, ..., p$. The model selection estimator chooses one of the k-component estimators, e.g. using a holdout sample or cross validation, and will be denoted by $\hat{\boldsymbol{\beta}}_{MSPCR}$.

Remark 2.11. a) The set of $p \times 1$ vectors $\{(1, 0, ..., 0)^T, (0, 1, 0, ..., 0)^T, (0, ...0, 1)^T\}$ is the standard basis for $\mathbb{R}^p$. The set of vectors $\{\hat{\boldsymbol{e}}_1, ..., \hat{\boldsymbol{e}}_p\}$ is also a basis for $\mathbb{R}^p$.

b) Let $\hat{\boldsymbol{\gamma}}_k = (\hat{\gamma}_1, ..., \hat{\gamma}_k)^T$. Since the columns of $\boldsymbol{V}_k$ are orthogonal, $\boldsymbol{c}_i \perp \boldsymbol{c}_j$ for $i \neq j$,

$$\hat{\gamma}_i = \frac{\boldsymbol{c}_i^T\boldsymbol{Z}}{\boldsymbol{c}_i^T\boldsymbol{c}_i} = \frac{\boldsymbol{c}_i^T\boldsymbol{Y}}{\boldsymbol{c}_i^T\boldsymbol{c}_i}.$$

c) Since $\hat{\boldsymbol{Z}} = \boldsymbol{V}_k\hat{\boldsymbol{\gamma}}_k + \boldsymbol{r} = \boldsymbol{W}\hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k + \boldsymbol{r} = \boldsymbol{W}\hat{\boldsymbol{\beta}}_{kPCR} + \boldsymbol{r}$, where $\hat{\boldsymbol{\beta}}_{kPCR} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k$. By Remark 2.5,

$$\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}Z} = [\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Z} =$$

$$[\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{w}, Z).$$

Thus

$$\hat{\boldsymbol{\beta}}_{kPCR} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{A}}_{k,n}^T[\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{w}, Z).$$

Note that $\hat{\boldsymbol{\beta}}_{pPCR} = \hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{w}, Z)$.

d) Let $\boldsymbol{e}_i = \boldsymbol{e}_i(\hat{\boldsymbol{\rho}}_{\boldsymbol{x}})$ be the ith eigenvector of the population correlation matrix $\hat{\boldsymbol{\rho}}_{\boldsymbol{x}}$ of the $\boldsymbol{x}$, and let

$$\boldsymbol{A}_k = \begin{pmatrix} \boldsymbol{e}_1^T \\ \vdots \\ \boldsymbol{e}_i^T \end{pmatrix}.$$

It is possible that $\hat{\boldsymbol{e}}_{i,n}$ is arbitrarily close to $\boldsymbol{e}_i$ for some values of n and arbitrarily close to $-\boldsymbol{e}_i$ for other values of n so that $\hat{\boldsymbol{e}}_i \equiv \hat{\boldsymbol{e}}_{i,n}$ oscillates and does not converge in probability to either $\boldsymbol{e}_i$ or $-\boldsymbol{e}_i$. Hence we can not say that the ith eigenvector $\hat{\boldsymbol{e}}_i = \hat{\boldsymbol{e}}_{i,n} \xrightarrow{P} \boldsymbol{e}_i$ or that $\boldsymbol{A}_{k,n} \xrightarrow{P} \boldsymbol{A}_k$. If $\hat{\boldsymbol{\Sigma}} \xrightarrow{P} c\boldsymbol{\Sigma}$ for some constant $c > 0$, and if the eigenvalues $\lambda_1 > \cdots > \lambda_p > 0$ of $\boldsymbol{\Sigma}$ are unique, then the absolute value of the correlation of $\hat{\boldsymbol{e}}_j$ with $\boldsymbol{e}_j$ converges to 1 in probability: $|\text{corr}(\hat{\boldsymbol{e}}_j, \boldsymbol{e}_j)| \xrightarrow{P} 1$. See Olive (2017b, p. 190). Let $\boldsymbol{\gamma}_k$ be the population vector from the OLS regression on the principal component vectors of the population correlation matrix $\boldsymbol{\rho}_{\boldsymbol{x}}$. Then $\boldsymbol{\gamma}_k$ and $\boldsymbol{A}_k$ are not unique since columns of $\boldsymbol{A}_k$ and elements of $\boldsymbol{\gamma}_k$ can be multiplied by -1 (an orthonormal eigenvector can be $\boldsymbol{e}_i$ or $-\boldsymbol{e}_i$), but if a column $\boldsymbol{e}_j$ of A_k is multiplied by -1 then the jth element of $\gamma_{k,j}$ is multiplied by -1 so $\boldsymbol{A}_k^T\boldsymbol{\gamma}_k$ is unique. Thus $\hat{\boldsymbol{A}}_{k,n}^T\hat{\gamma}_k \xrightarrow{P} \boldsymbol{A}_k^T\boldsymbol{\gamma}_k$. Let $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \xrightarrow{P} \boldsymbol{\rho}_{\boldsymbol{u}}$. Then

$$\boldsymbol{\beta}_{kPCR} = \boldsymbol{A}_k^T\boldsymbol{\phi}_k = \boldsymbol{A}_k^T[\boldsymbol{A}_k\boldsymbol{\rho}_{\boldsymbol{x}}\boldsymbol{A}_k^T]^{-1}\boldsymbol{A}_k\boldsymbol{\rho}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Z).$$

See Helland and Almøy (1994).

e) In general, $\hat{\boldsymbol{\beta}}_{kPCR}$ estimates $\boldsymbol{\beta}_{kPCR} \neq \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Z)$ unless $k = p$. Using standardized predictors and estimated eigenvectors likely causes problems for finding a CLT, as in Remark 2.6.

f) Generally there is no reason why the "predictors" should be ranked from best to worst by $V_1, V_2, ..., V_k$. For example, the last few principal component vectors (and a constant) could be much better for prediction than the other principal component vectors. See Jolliffe (1983) and Cook and Forzani (2008).

g) Suppose $\sum_{i=1}^{J} \hat{\lambda}_i \geq q(p)$ where $0.5 \leq q \leq 1$, e.g. $q = 0.8$ where J is a lot smaller than p. Then the J predictors $V_1, ..., V_J$ capture much of the information of the standardized nontrivial predictors $w_1, ..., w_p$. Then regressing Y on $1, V_1, ..., V_J$ may be competitive with regressing Y on $w_1, ..., w_p$. PCR is equivalent to OLS on the full model when Y is regressed on a constant and all $K = p$ of the principal components. PCR can also be useful if $\boldsymbol{X}$ is singular or nearly singular (ill conditioned).

Example 2.2, continued. The PCR output below shows results for the marry data where 10-fold CV was used. The OLS full model was selected.

```
library(pls); y <- marry[,3]; x <- marry[,-3]
z <- as.data.frame(cbind(y,x))
out<-pcr(y~.,data=z,scale=T,validation="CV")
tem<-MSEP(out)
tem
   (Int)     1 comps   2 comps 3 comps 4 comps
CV 1.743e+09 449479706 8181251 371775   197132
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
nc <-max(which.min(cvmse)-1,1)
res <- out$residuals[,,nc]
yhat<-y-res #d = 5 predictors used including constant
AERplot2(yhat,y,res=res,d=5)
#response plot with 90% pointwise PIs
$respi #90% PI same as OLS full model
-950.4811 1445.2584 #PI length = 2395.74
```

Several statistical methods can be computed using an $n \times n$ matrix or a $p \times p$ matrix, depending on whether n or p is smaller. The remainder of this section shows the computations for principle components analysis (PCA), which is used for principle components regression.

Suppose $\boldsymbol{W}$ is the standardized $n \times p$ data matrix and $\boldsymbol{T} = \boldsymbol{W}_g/\sqrt{n-g}$. If $n < p$, then the correlation matrix $\boldsymbol{R} = \boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{W}_g^T\boldsymbol{W}_g/(n-g)$ does not have full rank. By singular value decomposition (SVD) theory, the SVD of $\boldsymbol{T}$ is $\boldsymbol{T} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ where the positive singular values σ_i are square roots of the positive eigenvalues of both $\boldsymbol{T}^T\boldsymbol{T}$ and of $\boldsymbol{T}\boldsymbol{T}^T$. (The singular values are **not** standard deviations.) Also $\boldsymbol{V} = (\hat{\boldsymbol{e}}_1\ \hat{\boldsymbol{e}}_2\ \cdots\ \hat{\boldsymbol{e}}_p)$, and $\boldsymbol{T}^T\boldsymbol{T}\hat{\boldsymbol{e}}_i = \sigma_i^2\hat{\boldsymbol{e}}_i$. Hence classical principal component analysis on the standardized data can be done using $\hat{\boldsymbol{e}}_i$ and $\hat{\lambda}_i = \sigma_i^2$. The SVD of $\boldsymbol{T}^T$ is $\boldsymbol{T}^T = \boldsymbol{V}\boldsymbol{\Lambda}^T\boldsymbol{U}^T$, and

$$\boldsymbol{T}\boldsymbol{T}^T = \frac{1}{n-g}\begin{bmatrix} \boldsymbol{w}_1^T\boldsymbol{w}_1 & \boldsymbol{w}_1^T\boldsymbol{w}_2 & \dots & \boldsymbol{w}_1^T\boldsymbol{w}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}_n^T\boldsymbol{w}_1 & \boldsymbol{w}_n^T\boldsymbol{w}_2 & \dots & \boldsymbol{w}_n^T\boldsymbol{w}_n \end{bmatrix}$$

which is the matrix of scalar products divided by n. Similarly, if $\boldsymbol{W}_c$ is the centered data matrix (subtract the means), then $\boldsymbol{T}_c = \boldsymbol{W}_c/\sqrt{n-g}$, and the

covariance matrix $\boldsymbol{S} = \boldsymbol{T}_c^T \boldsymbol{T}_c = \boldsymbol{W}_c^T \boldsymbol{W}_c/(n-g)$. For more information about the SVD, see Datta (1995, pp. 552-556) and Fogel et al. (2013).

The following output shows how to do classical PCA with $\boldsymbol{S}$ on a data set using the SVD and $g = 1$. The eigenvectors agree up to sign.

```
x<-cbind(buxx,buxy) # data matrix
mn <- apply(x,2,mean) #sample mean
J <- 0*1:87 + 1  # vector of n ones, n = 87
J <- J%*%t(J)/87 #J%*%x has rows = mn
zc <- x-J%*%x #centered x
yc <- zc/sqrt(87-1) #t(yc) %*% yc = cov(x)
svd(yc)$v              #right eigenvectors of Yc
          [,1]     [,2]     [,3]     [,4]    [,5]
[1,]  0.653883  0.75596 -0.01173  0.00988  0.0268
[2,] -0.001366  0.03980  0.06800 -0.42534 -0.9016
[3,] -0.000489 -0.01276 -0.99161 -0.12775 -0.0151
[4,] -0.000714  0.00251 -0.10890  0.89588 -0.4308
[5,] -0.756594  0.65327 -0.00952  0.00854  0.0252
> svd(t(yc))$u        #left eigenvectors of Yc^T
          [,1]     [,2]     [,3]     [,4]    [,5]
[1,] -0.653883 -0.75596  0.01173 -0.00988 -0.0268
[2,]  0.001366 -0.03980 -0.06800  0.42534  0.9016
[3,]  0.000489  0.01276  0.99161  0.12775  0.0151
[4,]  0.000714 -0.00251  0.10890 -0.89588  0.4308
[5,]  0.756594 -0.65327  0.00952 -0.00854 -0.0252
> prcomp(x)
Standard deviations:
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
Rotation:
               PC1      PC2      PC3      PC4     PC5
len       0.653883  0.75596 -0.01173  0.00988  0.0268
nasal    -0.001366  0.03980  0.06800 -0.42534 -0.9016
bigonal  -0.000489 -0.01276 -0.99161 -0.12775 -0.0151
cephalic -0.000714  0.00251 -0.10890  0.89588 -0.4308
buxy     -0.756594  0.65327 -0.00952  0.00854  0.0252
svd(yc)$d        #singular values = sqrt(eigenvalues)
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
svd(t(yc))$d     #singular values = sqrt(eigenvalues)
[1] 523.70760  42.50435   6.06073   4.39067   3.80398
```

Although PCA can be done if $p > n$, in general need p fixed for the sample eigenvector to be a good estimator of a population eigenvector.

2.5 Partial Least Squares

Consider the MLR model $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i = \alpha + x_{i,1}\boldsymbol{\beta}_1 + \cdots + x_{i,p}\boldsymbol{\beta}_p + e_i$ for $i = 1, ..., n$. Principal components regression (PCR) and partial least squares (PLS) models use p linear combinations $\boldsymbol{\eta}_1^T\boldsymbol{x}, ..., \boldsymbol{\eta}_p^T\boldsymbol{x}$. Then there are p conditional distributions

$$\begin{gathered} Y|\boldsymbol{\eta}_1^T\boldsymbol{x} \\ Y|(\boldsymbol{\eta}_1^T\boldsymbol{x}, \boldsymbol{\eta}_2^T\boldsymbol{x}) \\ \vdots \\ Y|(\boldsymbol{\eta}_1^T\boldsymbol{x}, \boldsymbol{\eta}_2^T\boldsymbol{x}, ..., \boldsymbol{\eta}_p^T\boldsymbol{x}). \end{gathered}$$

Estimating the $\boldsymbol{\eta}_j$ and performing the ordinary least squares (OLS) regression of Y on $(\hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}, \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x}, ..., \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x})$ and a constant gives the k-component estimator, e.g. the k-component PLS estimator $\hat{\boldsymbol{\beta}}_{kPLS}$ or the k-component PCR estimator, for $k = 1, ..., J$ where $J \leq p$ and the p-component estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. Denote the one component PLS (OPLS) estimator by $\hat{\boldsymbol{\beta}}_{OPLS}$. The model selection estimator chooses one of the k-component estimators, e.g. using a holdout sample or cross validation, and will be denoted by $\hat{\boldsymbol{\beta}}_{MSPLS}$. For the OPLS estimator, $\boldsymbol{\eta}_1 = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ and $\hat{\boldsymbol{\eta}}_1 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. See Sections 2.10 and 2.11 for more on the OPLS estimator.

Remark 2.12. Olive and Zhang (2024) showed that $\hat{\boldsymbol{\beta}}_{kPLS}$ estimates $\boldsymbol{\beta}_{kPLS}$, and in general, $\boldsymbol{\beta}_{kPLS} \neq \boldsymbol{\beta}_{OLS}$ for $k < p$. In particular, $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$ except under very strong regularity conditions. The PLS literature incorrectly suggests that $\boldsymbol{\beta}_{kPLS} = \boldsymbol{\beta}_{OLS}$, under mild regularity conditions, for $1 \leq k < p$ if p is fixed. Also see Chun and Keleş (2010), Cook (2018), Cook et al. (2013), and Cook and Forzani (2018, 2019, 2024).

There are several ways to compute k-component partial least squares (PLS) estimators for multiple linear regression. A simple way is to do the OLS regression on (a constant and) $W_1, ..., W_k$ where $W_j = \hat{\boldsymbol{\eta}}_j^T\boldsymbol{x}$ and $\hat{\boldsymbol{\eta}}_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{j-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$, and $k \leq min(n-2, p)$. Then the one component PLS estimator is OPLS: $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\boldsymbol{\beta}}_{1PLS}$ with $k = 1$, and $\hat{\boldsymbol{\beta}}_{OLS} = \hat{\boldsymbol{\beta}}_{pPLS}$ with $k = p$ if $n > p + 1$. The 3-component PLS estimator regresses Y on (a constant and) $W_1 = \hat{\boldsymbol{\eta}}_1^T\boldsymbol{x} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}$, $W_2 = \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x} = [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}]^T\boldsymbol{x}$, and $W_3 = \hat{\boldsymbol{\eta}}_3^T\boldsymbol{x} = [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}]^T\boldsymbol{x}$. Let $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}_{kPLS} + \epsilon$ be a working model. From Naik and Tsai (2000), Helland and Almøy (1994), and Helland (1990), let $\hat{\boldsymbol{A}}_{k,n}^T = [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}, ..., \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{k-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}]$. Let $\boldsymbol{w} = \hat{\boldsymbol{A}}_{k,n}\boldsymbol{x}$ with $Y = \alpha + \boldsymbol{w}^T\boldsymbol{\gamma}_k + \epsilon$ the working model so $\hat{\boldsymbol{\beta}}_{kPLS} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k$. Then $\hat{\boldsymbol{\beta}}_{kPLS} =$

$$\hat{\boldsymbol{A}}_{k,n}^T[\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \hat{\boldsymbol{A}}_{k,n}^T[\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T]^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}, Y).$$

Example 2.2, continued. The PLS output below shows results for the marry data where 10-fold CV was used. The OLS full model was selected. The Mevik et al. (2015) `pls` library is useful for computing PLS and PCR.

```
library(pls); y <- marry[,3]; x <- marry[,-3]
z <- as.data.frame(cbind(y,x))
out<-plsr(y~.,data=z,scale=T,validation="CV")
tem<-MSEP(out)
tem
   (Int)     1 comps   2 comps 3 comps 4 comps
CV 1.743e+09 256433719 6301482 249366  206508
cvmse<-tem$val[,,1:(out$ncomp+1)][1,]
nc <-max(which.min(cvmse)-1,1)
res <- out$residuals[,,nc]
yhat<-y-res #d = 5 predictors used including constant
AERplot2(yhat,y,res=res,d=5)
$respi  #90% PI same as OLS full model
-950.4811 1445.2584  #PI length = 2395.74
```

There are some other equivalent ways to formulate PLS. The following formulation shows that PLS seeks PLS directions that are correlated with Y. Note that PCR components are formed without using Y. Let $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}_{kPLS} + \epsilon$ be a working model. Let $\boldsymbol{X} = (\boldsymbol{1}\ \boldsymbol{X}_1)$. An equivalent way to formulate PLS is to form $\boldsymbol{b}_j$ iteratively where $\boldsymbol{b}_k = \arg\max_{\boldsymbol{b}}\{[corr(\boldsymbol{Y}, \boldsymbol{X}_1\boldsymbol{b})]^2 V(\boldsymbol{X}_1\boldsymbol{b})\}$ subject to $\boldsymbol{b}^T\boldsymbol{b} = 1$ and $\boldsymbol{b}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{b}_j = 0$ for $j = 1, ..., k-1$. Let the $\hat{\boldsymbol{b}}_j$ be the estimates of $\boldsymbol{b}_j$, and perform the OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}_1\hat{\boldsymbol{C}}_{k,n}$ and a constant where $\hat{\boldsymbol{C}}_{k,n} = [\hat{\boldsymbol{b}}_1, ..., \hat{\boldsymbol{b}}_k]$ to find $\hat{\boldsymbol{\gamma}}_k$. Then $\hat{\boldsymbol{\beta}}_{kPLS} = \hat{\boldsymbol{C}}_{k,n}\hat{\boldsymbol{\gamma}}_k$.

Here is another way to formulate PLS. Let $\boldsymbol{X}_c$ be the matrix of centered predictors (subtract the sample mean from each predictor) so that $\boldsymbol{D} = \boldsymbol{X}_c^T\boldsymbol{X}_x = (n-1)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}$ and let $\boldsymbol{Z}$ be the vector of centered response variables. Let $\boldsymbol{d} = \boldsymbol{X}_c^T\boldsymbol{Z} = (n-1)\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. An equivalent way to compute the k-component PLS estimator is to find unit vectors $\hat{\boldsymbol{\eta}}_1, ..., \hat{\boldsymbol{\eta}}_k$ and perform the OLS regression of Y on a constant and the $U_i = \hat{\boldsymbol{\eta}}_i^T\boldsymbol{x}$ for $i = 1, ..., k$. Following Brown (1993, pp. 71-72), first maximize $(\boldsymbol{c}^T\boldsymbol{d})^2$ subject to the constraint $\boldsymbol{c}^T\boldsymbol{c} = \|\boldsymbol{c}\|^2 = 1$. The maximum occurs at $\boldsymbol{c}_1 = \hat{\boldsymbol{\eta}}_1 = \boldsymbol{d}/\|\boldsymbol{d}\| = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}/\|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}\| = \hat{\boldsymbol{\eta}}_{OPLS}/\|\hat{\boldsymbol{\eta}}_{OPLS}\|$. Then $\boldsymbol{c}_2 = \hat{\boldsymbol{\eta}}_2$ is found by maximizing $(\boldsymbol{c}^T\boldsymbol{d})^2$ subject to both $\|\boldsymbol{c}\| = 1$ and $\boldsymbol{c}^T\boldsymbol{D}\boldsymbol{c}_1 = 0$ (called $\boldsymbol{D}$-norm orthogonalization) to get $\boldsymbol{c}_2 = \hat{\boldsymbol{\eta}}_2$. Continue in this way to get the remaining vectors $\boldsymbol{c}_3, ..., \boldsymbol{c}_k$.

2.6 Ridge Regression

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Ridge regression often uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Note that in Definition 2.16, $\lambda_{1,n}$ is a tuning parameter, not an eigenvalue. The residuals $\boldsymbol{r} = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. Refer to Definition 2.11 for the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$.

Definition 2.15. Consider the MLR model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. Let $\boldsymbol{b}$ be a $(p-1) \times 1$ vector. Then the fitted value $\hat{Z}_i(\boldsymbol{b}) = \boldsymbol{w}_i^T\boldsymbol{b}$ and the residual $r_i(\boldsymbol{b}) = Z_i - \hat{Z}_i(\boldsymbol{b})$. The vector of fitted values $\hat{\boldsymbol{Z}}(\boldsymbol{b}) = \boldsymbol{W}\boldsymbol{b}$ and the vector of residuals $\boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}(\boldsymbol{b})$.

Definition 2.16. a) Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. The *ridge regression estimator* $\hat{\boldsymbol{\eta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}\eta_i^2 \tag{2.16}$$

over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$ where $\lambda_{1,n} \geq 0$ and $a > 0$ are known constants with $a = 1, 2, n$, and $2n$ common. Then

$$\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}. \tag{2.17}$$

The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS}$. The ridge regression vector of fitted values is $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_R = \boldsymbol{W}\hat{\boldsymbol{\eta}}_R$, and the ridge regression vector of residuals $\boldsymbol{r}_R = \boldsymbol{r}(\hat{\boldsymbol{\eta}}_R) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}_R$. The estimator is said to be *regularized* if $\lambda_{1,n} > 0$. Obtain $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}_R$ using $\hat{\boldsymbol{\eta}}_R$, $\hat{\boldsymbol{Z}}$, and $\overline{\boldsymbol{Y}}$.

b) Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Let $\lambda \geq 0$ be a constant. One *ridge regression estimator* $\hat{\boldsymbol{\beta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p}\beta_i^2 \tag{2.18}$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$. Then

$$\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y}. \tag{2.19}$$

The residual sum of squares $RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. The ridge regression vector of fitted values is $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_R = \boldsymbol{X}\hat{\boldsymbol{\beta}}_R$, and the ridge regression vector of residuals $\boldsymbol{r}_R = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_R) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_R$.

c) Another *ridge regression estimator* $\tilde{\boldsymbol{\beta}}_{RR}$ minimizes the *ridge regression criterion*

$$Q_{RR}(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=2}^{p}\beta_i^2$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$.

The estimators b) and c) agree when a) is used. Using a vector of parameters $\boldsymbol{\eta}$ and a dummy vector $\boldsymbol{\eta}$ in Q_R is common for minimizing a criterion $Q(\boldsymbol{\eta})$, often with estimating equations. See the paragraphs above and below Definition 2.12. We could also write

$$Q_R(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\boldsymbol{b}^T\boldsymbol{b}$$

where the minimization is over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$. Note that $\sum_{i=1}^{p-1}\eta_i^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \|\boldsymbol{\eta}\|_2^2$. The literature often uses $\lambda_a = \lambda = \lambda_{1,n}/a$.

Note that $\lambda_{1,n}\boldsymbol{b}^T\boldsymbol{b} = \lambda_{1,n}\sum_{i=1}^{p-1}b_i^2$. Each coefficient b_i is penalized equally by $\lambda_{1,n}$. Hence using standardized nontrivial predictors makes sense so that if η_i is large in magnitude, then the standardized variable w_i is important.

Remark 2.13. i) If $\lambda_{1,n} = 0$, the ridge regression estimator becomes the OLS full model estimator: $\hat{\boldsymbol{\eta}}_R = \hat{\boldsymbol{\eta}}_{OLS}$.

ii) If $\lambda_{1,n} > 0$, then $\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}$ is nonsingular. Hence $\hat{\boldsymbol{\eta}}_R$ exists even if $\boldsymbol{X}$ and $\boldsymbol{W}$ are singular or ill conditioned, or if $p > n$.

iii) Following Hastie et al. (2009, p. 96), let the augmented matrix $\boldsymbol{W}_A$ and the augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_{\mathrm{A}} = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{0}$ is the $(p-1) \times 1$ zero vector. For $\lambda_{1,n} > 0$, the OLS estimator from regressing $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$ is

$$\hat{\boldsymbol{\eta}}_A = (\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z}_A = \hat{\boldsymbol{\eta}}_R$$

since $\boldsymbol{W}_A^T\boldsymbol{Z}_A = \boldsymbol{W}^T\boldsymbol{Z}$ and

$$\boldsymbol{W}_A^T\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W}^T & \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix}\begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_{1,n}}\ \boldsymbol{I}_{p-1} \end{pmatrix} = \boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\ \boldsymbol{I}_{p-1}.$$

iv) A simple way to regularize a regression estimator, such as the L_1 estimator, is to compute that estimator from regressing $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$.

Remark 2.13 iii) is interesting. Note that for $\lambda_{1,n} > 0$, the $(n+p-1)\times(p-1)$ matrix $\boldsymbol{W}_A$ has full rank $p-1$. The augmented OLS model consists of adding $p-1$ pseudo-cases $(\boldsymbol{w}_{n+1}^T, Z_{n+1})^T, ..., (\boldsymbol{w}_{n+p-1}^T, Z_{n+p-1})^T$ where $Z_j = 0$ and

$\boldsymbol{w}_j = (0, ..., \sqrt{\lambda_{1,n}}, 0, ..., 0)^T$ for $j = n+1, ..., n+p-1$ where the nonzero entry is in the kth position if $j = n + k$. For centered response and standardized nontrivial predictors, the population OLS regression fit runs through the origin $(\boldsymbol{w}^T, Z)^T = (\boldsymbol{0}^T, 0)^T$. Hence for $\lambda_{1,n} = 0$, the augmented OLS model adds $p - 1$ typical cases at the origin. If $\lambda_{1,n}$ is not large, then the pseudo-data can still be regarded as typical cases. If $\lambda_{1,n}$ is large, the pseudo-data act as w–outliers (outliers in the standardized predictor variables), and the OLS slopes go to zero as $\lambda_{1,n}$ gets large, making $\hat{\boldsymbol{Z}} \approx \boldsymbol{0}$ so $\hat{\boldsymbol{Y}} \approx \overline{\boldsymbol{Y}}$.

To prove Remark 2.13 ii), let $(\psi, \boldsymbol{g})$ be an eigenvalue eigenvector pair of $\boldsymbol{W}^T\boldsymbol{W} = n\boldsymbol{R}_{\boldsymbol{u}}$. Then $[\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1}]\boldsymbol{g} = (\psi + \lambda_{1,n})\boldsymbol{g}$, and $(\psi + \lambda_{1,n}, \boldsymbol{g})$ is an eigenvalue eigenvector pair of $\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1} > 0$ provided $\lambda_{1,n} > 0$.

The degrees of freedom for a ridge regression with known $\lambda_{1,n}$ is also interesting and will be found in the next paragraph. The sample correlation matrix of the nontrivial predictors

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{1}{n-g}\boldsymbol{W}_g^T\boldsymbol{W}_g$$

where we will use $g = 0$ and $\boldsymbol{W} = \boldsymbol{W}_0$. Then $\boldsymbol{W}^T\boldsymbol{W} = n\boldsymbol{R}_{\boldsymbol{u}}$. By singular value decomposition (SVD) theory, the SVD of $\boldsymbol{W}$ is $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{V}^T$ where the positive singular values σ_i are square roots of the positive eigenvalues of both $\boldsymbol{W}^T\boldsymbol{W}$ and of $\boldsymbol{W}\boldsymbol{W}^T$. Also $\boldsymbol{V} = (\hat{\boldsymbol{e}}_1\ \hat{\boldsymbol{e}}_2\ \cdots\ \hat{\boldsymbol{e}}_p)$, and $\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{e}}_i = \sigma_i^2\hat{\boldsymbol{e}}_i$. Hence $\hat{\lambda}_i = \sigma_i^2$ where $\hat{\lambda}_i = \hat{\lambda}_i(\boldsymbol{W}^T\boldsymbol{W})$ is the ith eigenvalue of $\boldsymbol{W}^T\boldsymbol{W}$, and $\hat{\boldsymbol{e}}_i$ is the ith orthonormal eigenvector of $\boldsymbol{R}_{\boldsymbol{u}}$ and of $\boldsymbol{W}^T\boldsymbol{W}$. The SVD of $\boldsymbol{W}^T$ is $\boldsymbol{W}^T = \boldsymbol{V}\boldsymbol{\Lambda}^T\boldsymbol{U}^T$, and the *Gram matrix*

$$\boldsymbol{W}\boldsymbol{W}^T = \begin{bmatrix} \boldsymbol{w}_1^T\boldsymbol{w}_1 & \boldsymbol{w}_1^T\boldsymbol{w}_2 & \ldots & \boldsymbol{w}_1^T\boldsymbol{w}_n \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{w}_n^T\boldsymbol{w}_1 & \boldsymbol{w}_n^T\boldsymbol{w}_2 & \ldots & \boldsymbol{w}_n^T\boldsymbol{w}_n \end{bmatrix}$$

which is the matrix of scalar products. **Warning:** Note that σ_i is the ith singular value of $\boldsymbol{W}$, not the standard deviation of w_i.

Following Hastie et al. (2009, p. 68), if $\hat{\lambda}_i = \hat{\lambda}_i(\boldsymbol{W}^T\boldsymbol{W})$ is the ith eigenvalue of $\boldsymbol{W}^T\boldsymbol{W}$ where $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_{p-1}$, then the (effective) degrees of freedom for the ridge regression of $\boldsymbol{Z}$ on $\boldsymbol{W}$ with known $\lambda_{1,n}$ is $df(\lambda_{1,n}) =$

$$tr[\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T] = \sum_{i=1}^{p-1}\frac{\sigma_i^2}{\sigma_i^2 + \lambda_{1,n}} = \sum_{i=1}^{p-1}\frac{\hat{\lambda}_i}{\hat{\lambda}_i + \lambda_{1,n}} \quad (2.20)$$

where the trace of a square $(p-1) \times (p-1)$ matrix $\boldsymbol{A} = (a_{ij})$ is $tr(\boldsymbol{A}) = \sum_{i=1}^{p-1} a_{ii} = \sum_{i=1}^{p-1} \hat{\lambda}_i(\boldsymbol{A})$. Note that the trace of $\boldsymbol{A}$ is the sum of the diagonal elements of $\boldsymbol{A}$ = the sum of the eigenvalues of $\boldsymbol{A}$.

Note that $0 \leq df(\lambda_{1,n}) \leq p-1$ where $df(\lambda_{1,n}) = p-1$ if $\lambda_{1,n} = 0$ and $df(\lambda_{1,n}) \to 0$ as $\lambda_{1,n} \to \infty$. The R code below illustrates how to compute ridge regression degrees of freedom.

```
set.seed(13)
n<-100; q<-3  #q = p-1
b <- 0 * 1:q + 1
u <- matrix(rnorm(n * q), nrow = n, ncol = q)
y <- 1 + u %*% b + rnorm(n) #make MLR model
w1 <- scale(u) #t(w1) %*% w1 = (n-1) R = (n-1)*cor(u)
w <- sqrt(n/(n-1))*w1   #t(w) %*% w = n R = n cor(u)
t(w) %*% w/n
             [,1]         [,2]         [,3]
[1,]  1.00000000 -0.04826094 -0.06726636
[2,] -0.04826094  1.00000000 -0.12426268
[3,] -0.06726636 -0.12426268  1.00000000
cor(u) #same as above
rs <- t(w)%*%w #scaled correlation matrix n R
svs <-svd(w)$d  #singular values of w
lambda <- 0
d <- sum(svs^2/(svs^2+lambda))
#effective df for ridge regression using w
d
[1] 3  #= q = p-1
112.60792 103.88089  83.51119
svs^2 #as above
uu<-scale(u,scale=F) #centered but not scaled
svs <-svd(uu)$d #singular values of uu
svs^2
[1] 135.78205 108.85903  85.83395
d <- sum(svs^2/(svs^2+lambda))
#effective df for ridge regression using uu
#d is again 3 if lambda = 0
```

In general, if $\hat{\boldsymbol{Z}} = \boldsymbol{H}_\lambda \boldsymbol{Z}$, then $df(\hat{\boldsymbol{Z}}) = tr(\boldsymbol{H}_\lambda)$ where $\boldsymbol{H}_\lambda$ is a $(p-1) \times (p-1)$ "hat matrix." For computing $\hat{\boldsymbol{Y}}$, $df(\hat{\boldsymbol{Y}}) = df(\hat{\boldsymbol{Z}}) + 1$ since a constant $\hat{\boldsymbol{\beta}}_1$ also needs to be estimated. These formulas for degrees of freedom assume that λ is known before fitting the model. The formulas do not give the model degrees of freedom if $\hat{\lambda}$ is selected from M values $\lambda_1, ..., \lambda_M$ using a criterion such as k-fold cross validation.

Suppose the ridge regression criterion is written, using $a = 2n$, as

$$Q_{R,n}(\boldsymbol{b}) = \frac{1}{2n}\boldsymbol{r}(\boldsymbol{b})^T \boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n}\boldsymbol{b}^T\boldsymbol{b}, \tag{2.21}$$

as in Hastie et al. (2015, p. 10). Then $\lambda_{2n} = \lambda_{1,n}/(2n)$ using the $\lambda_{1,n}$ from (2.16).

The following remark is interesting if $\lambda_{1,n}$ and p are fixed. However, $\hat{\lambda}_{1,n}$ is usually used, for example, after 10-fold cross validation. The fact that $\hat{\boldsymbol{\beta}}_R = \boldsymbol{A}_{n,\lambda}\hat{\boldsymbol{\beta}}_{OLS}$ appears in Efron and Hastie (2016, p. 98), and Marquardt and Snee (1975). See Theorem 2.7 for the ridge regression central limit theorem.

Remark 2.14. Ridge regression has a simple relationship with OLS if $n > p$ and $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$ exists. Then $\hat{\boldsymbol{\beta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{A}_{n,\lambda}\hat{\boldsymbol{\beta}}_{OLS}$ where $\boldsymbol{A}_{n,\lambda} \equiv \boldsymbol{A}_n = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}$. By the OLS CLT Equation (2.6) with $\hat{\boldsymbol{V}}/n = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$, a normal approximation for OLS is

$$\hat{\boldsymbol{\beta}}_{OLS} \sim AN_p(\boldsymbol{\beta}, MSE\ (\boldsymbol{X}^T\boldsymbol{X})^{-1}).$$

Hence a normal approximation for ridge regression is

$$\hat{\boldsymbol{\beta}}_R \sim AN_p(\boldsymbol{A}_n\boldsymbol{\beta}, MSE\ \boldsymbol{A}_n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{A}_n^T) \sim$$

$$AN_p[\boldsymbol{A}_n\boldsymbol{\beta}, MSE\ (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}].$$

If Equation (2.6) holds and $\lambda_{1,n}/n \to 0$ as $n \to \infty$, then $\boldsymbol{A}_n \xrightarrow{P} \boldsymbol{I}_p$.

Remark 2.15. The ridge regression criterion from Definition 2.16 can also be defined by

$$Q_R(\boldsymbol{\eta}) = \|\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}\|_2^2 + \lambda_{1,n}\boldsymbol{\eta}^T\boldsymbol{\eta}. \tag{2.22}$$

Then by Theorem 2.4, the gradient $\bigtriangledown Q_R = -2\boldsymbol{W}^T\boldsymbol{Z} + 2(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta} + 2\lambda_{1,n}\boldsymbol{\eta}$. Cancelling constants and evaluating the gradient at $\hat{\boldsymbol{\eta}}_R$ gives the score equations

$$-\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_R) + \lambda_{1,n}\hat{\boldsymbol{\eta}}_R = \boldsymbol{0}. \tag{2.23}$$

Following Efron and Hastie (2016, pp. 381-382, 392), this means $\hat{\boldsymbol{\eta}}_R = \boldsymbol{W}^T\boldsymbol{a}$ for some $n \times 1$ vector $\boldsymbol{a}$. Hence $-\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{W}^T\boldsymbol{a}) + \lambda_{1,n}\boldsymbol{W}^T\boldsymbol{a} = \boldsymbol{0}$, or

$$\boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{W}^T + \lambda_{1,n}\boldsymbol{I}_n)]\boldsymbol{a} = \boldsymbol{W}^T\boldsymbol{Z}$$

which has solution $\boldsymbol{a} = (\boldsymbol{W}\boldsymbol{W}^T + \lambda_{1,n}\boldsymbol{I}_n)^{-1}\boldsymbol{Z}$. Hence

$$\hat{\boldsymbol{\eta}}_R = \boldsymbol{W}^T\boldsymbol{a} = \boldsymbol{W}^T(\boldsymbol{W}\boldsymbol{W}^T + \lambda_{1,n}\boldsymbol{I}_n)^{-1}\boldsymbol{Z} = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}.$$

Using the $n \times n$ matrix $\boldsymbol{W}\boldsymbol{W}^T$ is computationally efficient if $p > n$ while using the $p \times p$ matrix $\boldsymbol{W}^T\boldsymbol{W}$ is computationally efficient if $n > p$. If $\boldsymbol{A}$ is $k \times k$, then computing $\boldsymbol{A}^{-1}$ has $O(k^3)$ complexity.

The following identity from Gunst and Mason (1980, p. 342) is useful for ridge regression inference: $\hat{\boldsymbol{\eta}}_R = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{Y}$

$$= (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

$$= (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{A}_n\hat{\boldsymbol{\beta}}_{OLS} =$$

$$[\boldsymbol{I}_p - \lambda_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}]\hat{\boldsymbol{\beta}}_{OLS} = \boldsymbol{B}_n\hat{\boldsymbol{\beta}}_{OLS} =$$

$$\hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_{1n}}{n}n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\hat{\boldsymbol{\beta}}_{OLS}$$

since $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$, where $\boldsymbol{A}_n = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}(\boldsymbol{X}^T\boldsymbol{X}) = \boldsymbol{B}_n$ $= \boldsymbol{I}_p - \lambda_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}$. See Problem 2.3. Assume

$$\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \to \boldsymbol{V}^{-1}$$

as $n \to \infty$. If $\lambda_{1,n}/n \to 0$ then

$$\frac{\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p}{n} \xrightarrow{P} \boldsymbol{V}^{-1}, \quad \text{and} \quad \mathrm{n}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda_{1,\mathrm{n}}\boldsymbol{I}_{\mathrm{p}})^{-1} \xrightarrow{\mathrm{P}} \boldsymbol{V}.$$

Note that

$$\boldsymbol{A}_n = \boldsymbol{A}_{n,\lambda} = \left(\frac{\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p}{n}\right)^{-1}\frac{\boldsymbol{X}^T\boldsymbol{X}}{n} \xrightarrow{P} \boldsymbol{V}\,\boldsymbol{V}^{-1} = \boldsymbol{I}_p$$

if $\lambda_{1,n}/n \to 0$ since matrix inversion is a continuous function of a positive definite matrix. See, for example, Bhatia et al. (1990), Stewart (1969), and Severini (2005, pp. 348-349).

For model selection, the M values of $\lambda = \lambda_{1,n}$ are denoted by $\lambda_1, \lambda_2, ..., \lambda_M$ where $\lambda_i = \lambda_{1,n,i}$ depends on n for $i = 1, ..., M$. If λ_s corresponds to the model selected, then $\hat{\lambda}_{1,n} = \lambda_s$. The following theorem shows that ridge regression and the OLS full model are asymptotically equivalent if $\hat{\lambda}_{1,n} = o_P(n^{1/2})$ so $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$.

Theorem 2.7, RR CLT (Ridge Regression Central Limit Theorem. Assume p is fixed and that the conditions of the OLS CLT Theorem Equation (2.6) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(-\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}).$$

Proof: If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, then by the above Gunst and Mason (1980) identity,

$$\hat{\boldsymbol{\beta}}_R = [\boldsymbol{I}_p - \hat{\lambda}_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}_{1,n}\boldsymbol{I}_p)^{-1}]\hat{\boldsymbol{\beta}}_{OLS}.$$

Hence

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_R - \hat{\boldsymbol{\beta}}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - \sqrt{n}\frac{\hat{\lambda}_{1,n}}{n} n(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}_{1,n}\boldsymbol{I}_p)^{-1}\hat{\boldsymbol{\beta}}_{OLS}$$

$$\xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \tau\boldsymbol{V}\boldsymbol{\beta} \sim N_p(-\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}). \quad \square$$

For p fixed, Knight and Fu (2000) note i) that $\hat{\boldsymbol{\beta}}_R$ is a consistent estimator of $\boldsymbol{\beta}$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, ii) OLS and ridge regression are asymptotically equivalent if $\lambda_{1,n}/\sqrt{n} \to 0$ as $n \to \infty$, iii) ridge regression is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded), and iv) if $\lambda_{1,n}/\sqrt{n} \to \tau \geq 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) \xrightarrow{D} N_p(-\tau\boldsymbol{V}\boldsymbol{\beta}, \sigma^2\boldsymbol{V}).$$

Hence the bias can be considerable if $\tau \neq 0$. If $\tau = 0$, then OLS and ridge regression have the same limiting distribution.

Even if p is fixed, there are several problems with ridge regression inference if $\hat{\lambda}_{1,n}$ is selected, e.g. after 10-fold cross validation. For OLS forward selection, the probability that the model I_{min} underfits goes to zero, and each model with $S \subseteq I$ produced a $\sqrt{n}$ consistent estimator $\hat{\boldsymbol{\beta}}_{I,0}$ of $\boldsymbol{\beta}$. Ridge regression with 10-fold CV often shrinks $\hat{\boldsymbol{\beta}}_R$ too much if both i) the number of population active predictors $k_S = a_S - 1$ in Equation (2.14) and Remark 2.5 is greater than about 20, and ii) the predictors are highly correlated. If p is fixed and $\lambda_{1,n} = o_P(\sqrt{n})$, then the OLS full model and ridge regression are asymptotically equivalent, but much larger sample sizes may be needed for the normal approximation to be good for ridge regression since the ridge regression estimator can have large bias for moderate n. Ten fold CV does not appear to guarantee that $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$ or $\hat{\lambda}_{1,n}/n \xrightarrow{P} 0$.

Ridge regression can be a lot better than the OLS full model if i) $\boldsymbol{X}^T\boldsymbol{X}$ is singular or ill conditioned or ii) n/p is small. Ridge regression can be much faster than forward selection if $M = 100$ and n and p are large.

Roughly speaking, the biased estimation of the ridge regression estimator can make the MSE of $\hat{\boldsymbol{\beta}}_R$ or $\hat{\boldsymbol{\eta}}_R$ less than that of $\hat{\boldsymbol{\beta}}_{OLS}$ or $\hat{\boldsymbol{\eta}}_{OLS}$, but the large sample inference may need larger n for ridge regression than for OLS. However, the large sample theory has $n >> p$. We will try to use prediction intervals to compare OLS, forward selection, ridge regression, and lasso for data sets where $p > n$. See Sections 2.1, 2.3, 2.6, 2.7, and 2.13.

Warning. The R functions `glmnet` and `cv.glmnet` do ridge regression using Definition 2.16 c).

Example 2.2, continued. The ridge regression output below shows results for the marry data where 10-fold CV was used. A grid of 100 λ values was used, and $\lambda_0 > 0$ was selected. A problem with getting the false degrees of

freedom d for ridge regression is that it is not clear that $\lambda = \lambda_{1,n}/(2n)$. We need to know the relationship between λ and $\lambda_{1,n}$ in order to compute d. It seems unlikely that $d \approx 1$ if λ_0 is selected.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
out<-cv.glmnet(x,y,alpha=0)
lam <- out$lambda.min #value of lambda that minimizes
#the 10-fold CV criterion
yhat <- predict(out,s=lam,newx=x)
res <- y - yhat
n <- length(y)
w1 <- scale(x)
w <- sqrt(n/(n-1))*w1   #t(w) %*% w = n R_u, u = x
diag(t(w)%*%w)
    pop    mmen mmilmen  milwmn
     26      26      26      26
#sum w_i^2 = n = 26 for i = 1, 2, 3, and 4
svs <- svd(w)$d  #singular values of w,
pp <- 1 + sum(svs^2/(svs^2+2*n*lam))  #approx 1
# d for ridge regression if lam = lam_{1,n}/(2n)
AERplot2(yhat,y,res=res,d=pp)
$respi #90% PI for a future residual
[1] -5482.316 14854.268 #length = 20336.584
#try to reproduce the fitted values
z <- y - mean(y)
q<-dim(w)[2]
I <- diag(q)
M<- w%*%solve(t(w)%*%w + lam*I/(2*n))%*%t(w)
fit <- M%*%z + mean(y)
plot(fit,yhat) #they are not the same
max(abs(fit-yhat))
[1] 46789.11
M<- w%*%solve(t(w)%*%w + lam*I/(1547.1741))%*%t(w)
fit <- M%*%z + mean(y)
max(abs(fit-yhat)) #close
[1] 8.484979
```

2.7 Lasso

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{e}$. Lasso often uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta}+\boldsymbol{\epsilon}$ as described in Section 2.2. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. The residuals $\boldsymbol{r} = \boldsymbol{r}(\hat{\boldsymbol{\beta}}_L) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$. Recall that $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$.

Definition 2.17. a) Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{\epsilon}$. The *lasso estimator* $\hat{\boldsymbol{\eta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i| \tag{2.24}$$

over all vectors $\boldsymbol{\eta} \in \mathbb{R}^{p-1}$ where $\lambda_{1,n} \geq 0$ and $a > 0$ are known constants with $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$ if $\boldsymbol{W}$ has full rank $p-1$. The lasso vector of fitted values is $\hat{\boldsymbol{Z}} = \hat{\boldsymbol{Z}}_L = \boldsymbol{W}\hat{\boldsymbol{\eta}}_L$, and the lasso vector of residuals $\boldsymbol{r}(\hat{\boldsymbol{\eta}}_L) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}_L$. The estimator is said to be *regularized* if $\lambda_{1,n} > 0$. Obtain $\hat{\boldsymbol{Y}}$ and $\hat{\boldsymbol{\beta}}_L$ using $\hat{\boldsymbol{\eta}}_L$, $\hat{\boldsymbol{Z}}$, and $\overline{\boldsymbol{Y}}$.

b) The *lasso estimator* $\hat{\boldsymbol{\beta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\beta}) = \frac{1}{a}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \frac{\lambda_{1,n}}{a}\sum_{i=2}^{p}|\beta_i| \tag{2.25}$$

over all vectors $\boldsymbol{\beta} \in \mathbb{R}^p$. The residual sum of squares $RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$ if $\boldsymbol{X}$ has full rank p. The lasso vector of fitted values is $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Y}}_L = \boldsymbol{X}\hat{\boldsymbol{\beta}}_L$, and the lasso vector of residuals $\boldsymbol{r}(\hat{\boldsymbol{\beta}}_L) = \boldsymbol{Y} - \hat{\boldsymbol{Y}}_L$.

Using a vector of parameters $\boldsymbol{\eta}$ and a dummy vector $\boldsymbol{\eta}$ in Q_L is common for minimizing a criterion $Q(\boldsymbol{\eta})$, often with estimating equations. See the paragraphs above and below Definition 2.12. We could also write

$$Q_L(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\sum_{j=1}^{p-1}|b_j|, \tag{2.26}$$

where the minimization is over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$. The literature often uses $\lambda_a = \lambda = \lambda_{1,n}/a$.

For fixed $\lambda_{1,n}$, the lasso optimization problem is convex. Hence fast algorithms exist. As $\lambda_{1,n}$ increases, some of the $\hat{\eta}_i = 0$. If $\lambda_{1,n}$ is large enough, then $\hat{\boldsymbol{\eta}}_L = \boldsymbol{0}$ and $\hat{Y}_i = \overline{Y}$ for $i = 1, ..., n$. If none of the elements $\hat{\eta}_i$ of $\hat{\boldsymbol{\eta}}_L$ are zero, then $\hat{\boldsymbol{\eta}}_L$ can be found, in principle, by setting the partial derivatives of $Q_L(\boldsymbol{\eta})$ to 0. Potential minimizers also occur at values of $\boldsymbol{\eta}$ where not all of the partial derivatives exist. An analogy is finding the minimizer of a real valued function of one variable $h(x)$. Possible values for the minimizer include values of x_c satisfying $h'(x_c) = 0$, and values x_c where the derivative does not exist. Typically some of the elements $\hat{\eta}_i$ of $\hat{\boldsymbol{\eta}}_L$ that minimizes $Q_L(\boldsymbol{\eta})$ are zero, and differentiating does not work.

The following identity from Efron and Hastie (2016, p. 308), for example, is useful for inference for the lasso estimator $\hat{\boldsymbol{\eta}}_L$:

$$\frac{-1}{n}\boldsymbol{X}^T(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_L)+\frac{\lambda_{1,n}}{2n}\boldsymbol{s}_n=\boldsymbol{0} \text{ or } -\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\mathrm{L}})+\frac{\lambda_{1,\mathrm{n}}}{2}\boldsymbol{s}_{\mathrm{n}}=\boldsymbol{0}$$

where $s_{in} \in [-1,1]$ and $s_{in} = \text{sign}(\hat{\beta}_{i,L})$ if $\hat{\beta}_{i,L} \neq 0$. Here $\text{sign}(\beta_i) = 1$ if $\beta_i > 0$ and $\text{sign}(\beta_i) = -1$ if $\beta_i < 0$. Note that $\boldsymbol{s}_n = \boldsymbol{s}_{n,\hat{\boldsymbol{\beta}}_L}$ depends on $\hat{\boldsymbol{\beta}}_L$. Thus $\hat{\boldsymbol{\beta}}_L$

$$= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\, \boldsymbol{s}_n = \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\, \boldsymbol{s}_n.$$

If none of the elements of $\boldsymbol{\beta}$ are zero, and if $\hat{\boldsymbol{\beta}}_L$ is a consistent estimator of $\boldsymbol{\beta}$, then $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$. If $\lambda_{1,n}/\sqrt{n} \to 0$, then OLS and lasso are asymptotically equivalent even if $\boldsymbol{s}_n$ does not converge to a vector $\boldsymbol{s}$ as $n \to \infty$ since $\boldsymbol{s}_n$ is bounded. For model selection, the M values of λ are denoted by $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$ where $\lambda_i = \lambda_{1,n,i}$ depends on n for $i = 1, ..., M$. Also, λ_M is the smallest value of λ such that $\hat{\boldsymbol{\beta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\beta}}_{\lambda_i} \neq \boldsymbol{0}$ for $i < M$. If λ_s corresponds to the model selected, then $\hat{\lambda}_{1,n} = \lambda_s$. The following theorem shows that lasso and the OLS full model are asymptotically equivalent if $\hat{\lambda}_{1,n} = o_P(n^{1/2})$ so $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$: thus $\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \hat{\boldsymbol{\beta}}_{OLS}) = o_p(1)$.

Theorem 2.8, Lasso CLT. Assume p is fixed and that the conditions of the OLS CLT Theorem Equation (2.6) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) \xrightarrow{D} N_p\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right).$$

Proof. If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_L - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_L - \hat{\boldsymbol{\beta}}_{OLS} + \hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) =$$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - \sqrt{n}\frac{\lambda_{1,n}}{2n}n(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{s}_n \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \frac{\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_p\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right)$$

since under the OLS CLT, $n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V}$.

Part a) does not need $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s}$ as $n \to \infty$, since $\boldsymbol{s}_n$ is bounded. □

Suppose p is fixed. Knight and Fu (2000) note i) that $\hat{\boldsymbol{\beta}}_L$ is a consistent estimator of $\boldsymbol{\eta}$ if $\lambda_{1,n} = o(n)$ so $\lambda_{1,n}/n \to 0$ as $n \to \infty$, ii) OLS and lasso are asymptotically equivalent if $\lambda_{1,n} \to \infty$ too slowly as $n \to \infty$ (e.g. if $\lambda_{1,n} = \lambda$ is fixed), iii) lasso is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded). Note that Theorem 2.8 shows that OLS and lasso are asymptotically equivalent if $\lambda_{1,n}/\sqrt{n} \to 0$ as $n \to 0$.

In the literature, the criterion often uses $\lambda_a = \lambda_{1,n}/a$:

$$Q_{L,a}(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_a \sum_{j=1}^{p-1} |b_j|.$$

The values $a = 1,\ 2$, and $2n$ are common. Following Hastie et al. (2015, pp. 9, 17, 19) for the next two paragraphs, it is convenient to use $a = 2n$:

$$Q_{L,2n}(\boldsymbol{b}) = \frac{1}{2n}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n} \sum_{j=1}^{p-1} |b_j|, \tag{2.27}$$

where the Z_i are centered and the w_j are standardized using $g = 0$ so $\overline{w}_j = 0$ and $n\hat{\sigma}_j^2 = \sum_{i=1}^n w_{i,j}^2 = n$. Then $\lambda = \lambda_{2n} = \lambda_{1,n}/(2n)$ in Equation (2.25). For model selection, the M values of λ are denoted by $0 \leq \lambda_{2n,1} < \lambda_{2n,2} < \cdots < \lambda_{2n,M}$ where $\hat{\boldsymbol{\eta}}_\lambda = \boldsymbol{0}$ iff $\lambda \geq \lambda_{2n,M}$ and

$$\lambda_{2n,max} = \lambda_{2n,M} = \max_j \left| \frac{1}{n} \boldsymbol{s}_j^T \boldsymbol{Z} \right|$$

and $\boldsymbol{s}_j$ is the jth column of $\boldsymbol{W}$ corresponding to the jth standardized nontrivial predictor W_j. In terms of the $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_M$, used above Theorem 2.8, we have $\lambda_i = \lambda_{1,n,i} = 2n\lambda_{2n,i}$ and

$$\lambda_M = 2n\lambda_{2n,M} = 2 \max_j \left| \boldsymbol{s}_j^T \boldsymbol{Z} \right|.$$

For model selection we let I denote the index set of the predictors in the fitted model including the constant. The set A defined below is the index set without the constant.

Definition 2.18. The *active set* A is the index set of the nontrivial predictors in the fitted model: the predictors with nonzero $\hat{\eta}_i$.

Suppose that there are k active nontrivial predictors. Then for lasso, $k \leq n$. Let the $n \times k$ matrix $\boldsymbol{W}_A$ correspond to the standardized active predictors. If the columns of $\boldsymbol{W}_A$ are in general position, then the lasso vector of fitted

values

$$\hat{\boldsymbol{Z}}_L = \boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z} - n\lambda_{2n}\boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{s}_A$$

where $\boldsymbol{s}_A$ is the vector of signs of the active lasso coefficients. Here we are using the λ_{2n} of (2.27), and $n\lambda_{2n} = \lambda_{1,n}/2$. We could replace n λ_{2n} by λ_2 if we used $a = 2$ in the criterion

$$Q_{L,2}(\boldsymbol{b}) = \frac{1}{2}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_2\sum_{j=1}^{p-1}|b_j|. \tag{2.28}$$

See, for example, Tibshirani (2015). Note that $\boldsymbol{W}_A(\boldsymbol{W}_A^T\boldsymbol{W}_A)^{-1}\boldsymbol{W}_A^T\boldsymbol{Z}$ is the vector of OLS fitted values from regressing $\boldsymbol{Z}$ on $\boldsymbol{W}_A$ without an intercept.

Example 2.2, continued. The lasso output below shows results for the marry data where 10-fold CV was used. A grid of 38 λ values was used, and $\lambda_0 > 0$ was selected.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
out<-cv.glmnet(x,y)
lam <- out$lambda.min #value of lambda that minimizes
#the 10-fold CV criterion
yhat <- predict(out,s=lam,newx=x)
res <- y - yhat
pp <- out$nzero[out$lambda==lam] + 1 #d for lasso
AERplot2(yhat,y,res=res,d=pp)
$respi #90% PI for a future residual
-4102.672  4379.951  #length = 8482.62
```

There are some problems with lasso. i) Lasso large sample theory is worse or as good as that of the OLS full model if n/p is large. ii) Ten fold CV does not appear to guarantee that $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$ or $\hat{\lambda}_{1,n}/n \xrightarrow{P} 0$. iii) Lasso often shrinks $\hat{\boldsymbol{\beta}}$ too much if $a_S \geq 20$ and the predictors are highly correlated. iv) Ridge regression can be better than lasso if $a_S > n$.

Lasso can be a lot better than the OLS full model if i) $\boldsymbol{X}^T\boldsymbol{X}$ is singular or ill conditioned or ii) n/p is small. iii) For lasso, $M = M(lasso)$ is often near 100. Let $J \geq 5$. If n/J and p are both a lot larger than $M(lasso)$, then lasso can be considerably faster than forward selection, PLS, and PCR if $M = M(lasso) = 100$ and $M = M(F) = \min(\lceil n/J \rceil, p)$ where F stands for forward selection, PLS, or PCR. iv) The number of nonzero coefficients in $\hat{\boldsymbol{\eta}}_L \leq n$ even if $p > n$. This property of lasso can be useful if $p >> n$ and the population model is sparse.

2.8 Lasso Variable Selection

Lasso variable selection applies OLS on a constant and the k active predictors that have nonzero lasso $\hat{\eta}_i$ (model $I = I_{min}$). Lasso variable selection is called relaxed lasso by Hastie et al. (2015, p. 12), and the relaxed lasso estimator with $\phi = 0$ by Meinshausen (2007). The method is also called OLS-post lasso and post model selection OLS.

Theory for lasso variable selection was given in Pelawa Watagoda and Olive (2021b) and Rathnayake and Olive (2023). Lasso variable selection will often be better than lasso when the model is sparse or if $n \geq 10(k+1)$. Lasso can be better than lasso variable selection if $(\boldsymbol{X}_I^T \boldsymbol{X}_I)$ is ill conditioned or if $n/(k+1) < 10$. Lasso variable selection used a grid of K λ_i values for $i = 1, ..., K$ where $\lambda_1 < \lambda_2 < \cdots < \lambda_K$. If $K = 100$, then lasso variable selection can be much faster than forward selection if p is large. If n/p is not large, using $K > 100$ is likely a good idea due to the multitude of MLR models result. See Section 2.16. When p is fixed, $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ does not do variable selection well. For variable selection, want $\hat{\lambda}_{1,n}/\sqrt{n} \to \infty$, but $\hat{\lambda}_{1,n}/n \to 0$. See Fan and Li (2001). Let $\lambda_1 = 2n\lambda$. Guan and Tibshirani (2020) (and likely `glmnet`) use $\lambda < Cn^{-1/4}$ for some large constant C. Hence $\lambda_{1,n} = \lambda_1 \propto n^{3/4}$, and the consistency rate of the lasso algorithm is as best $n^{1/4}$, but variable selection lasso has the $\sqrt{n}$ rate (if λ_k is selected by lasso, make $\hat{\lambda} = \min(\lambda_k, n/log(n)$ so that $\hat{\lambda}/n \to 0$ as $n \to \infty$.)

Suppose the $n \times q$ matrix x has the $q = p - 1$ nontrivial predictors. The following R code gives some output for a lasso estimator and then the corresponding lasso variable selection estimator.

```
library(glmnet)
y <- marry[,3]
x <- marry[,-3]
out<-glmnet(x,y,dfmax=2)  #Use 2 for illustration:
#often dfmax approx min(n/J,p) for some J >= 5.
lam<-out$lambda[length(out$lambda)]
yhat <- predict(out,s=lam,newx=x)
#lasso with smallest lambda in grid such that df = 2
lcoef <- predict(out,type="coefficients",s=lam)
as.vector(lcoef) #first term is the intercept
#3.000397e+03 1.800342e-03 9.618035e-01 0.0 0.0
res <- y - yhat
AERplot(yhat,y,res,d=3,alph=1) #lasso response plot
##lasso variable selection =
#OLS on lasso active predictors and a constant
vars <- 1:dim(x)[2]
lcoef<-as.vector(lcoef)[-1] #don't need an intercept
vin <- vars[lcoef>0] #the lasso active set
vin
```

```
#1  2  since predictors 1 and 2 are active
sub <- lsfit(x[,vin],y) #lasso variable selection
sub$coef
#  Intercept          pop         mmen
#2.380912e+02 6.556895e-05 1.000603e+00
# 238.091     6.556895e-05 1.0006
res <- sub$resid
yhat <- y - res
AERplot(yhat,y,res,d=3,alph=1) #response plot
```

Example 2.2, continued. The lasso variable selection output below shows results for the marry data where 10-fold CV was used to choose the lasso estimator. Then lasso variable selection is OLS applied to the active variables with nonzero lasso coefficients and a constant. A grid of 38 λ values was used, and $\lambda_1 > 0$ was selected. The OLS SE, t statistic and pvalue are generally not valid for lasso variable selection by Remark 2.5 and Theorem 2.4.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
out<-cv.glmnet(x,y)
lam <- out$lambda.min #value of lambda that minimizes
#the 10-fold CV criterion
pp <- out$nzero[out$lambda==lam] + 1
#d for lasso variable selection
#get lasso variable selection
lcoef <- predict(out,type="coefficients",s=lam)
lcoef<-as.vector(lcoef)[-1]
vin <- vars[lcoef!=0]
sub <- lsfit(x[,vin],y)
ls.print(sub)
Residual Standard Error=376.9412
R-Square=0.9999
F-statistic (df=2, 23)=147440.1
          Estimate  Std.Err t-value Pr(>|t|)58
Intercept 238.0912 248.8616  0.9567   0.3487
pop         0.0001   0.0029  0.0223   0.9824
mmen        1.0006   0.0164 60.9878   0.0000
res <- sub$resid
yhat <- y - res
AERplot2(yhat,y,res=res,d=pp)
$respi #90% PI for a future residual
-822.759 1403.771  #length = 2226.53
```

To summarize Example 2.2, forward selection selected the model with the minimum C_p while the other methods used 10-fold CV. PLS and PCR used the OLS full model with PI length 2395.74, forward selection used a constant and *mmen* with PI length 2114.72, ridge regression had PI length

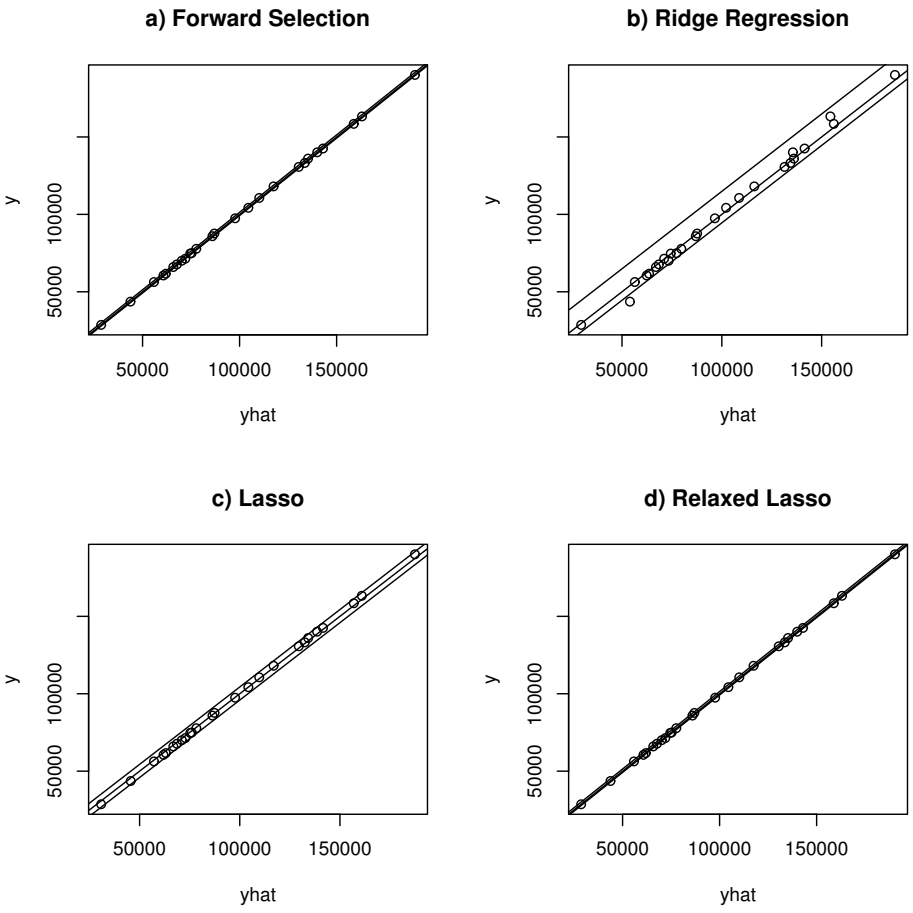

Fig. 2.1 Marry Data Response Plots

20336.58, lasso and lasso variable selection used a constant, *mmen*, and *pop* with lasso PI length 8482.62 and lasso variable selection PI length 2226.53. A PI from Section 2.13 was used. Figure 2.1 shows the response plots for forward selection, ridge regression, lasso, and lasso variable selection (labeled relaxed lasso). The plots for PLS=PCR=OLS full model were similar to those of forward selection and lasso variable selection. The plots suggest that the MLR model is appropriate since the plotted points scatter about the identity line. The 90% pointwise prediction bands are also shown, and consist of two lines parallel to the identity line. These bands are very narrow in Figure 2.1 a) and d).

2.9 The Elastic Net

Following Hastie et al. (2015, p. 57), let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_S^T)^T$, let $\lambda_{1,n} \geq 0$, and let $\alpha \in [0, 1]$. Let

$$RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2.$$

For a $k \times 1$ vector $\boldsymbol{\eta}$, the squared (Euclidean) L_2 norm $\|\boldsymbol{\eta}\|_2^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \sum_{i=1}^k \eta_i^2$ and the L_1 norm $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^k |\eta_i|$.

Definition 2.19. The *elastic net* estimator $\hat{\boldsymbol{\beta}}_{EN}$ minimizes the criterion

$$Q_{EN}(\boldsymbol{\beta}) = \frac{1}{2}RSS(\boldsymbol{\beta}) + \lambda_{1,n}\left[\frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}_S\|_2^2 + \alpha\|\boldsymbol{\beta}_S\|_1\right], \text{ or} \tag{2.29}$$

$$Q_2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}_S\|_2^2 + \lambda_2\|\boldsymbol{\beta}_S\|_1 \tag{2.30}$$

where $0 \leq \alpha \leq 1$, $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$.

Note that $\alpha = 1$ corresponds to lasso (using $\lambda_{a=0.5}$), and $\alpha = 0$ corresponds to ridge regression estimator of Definition 2.16 c), which is not the usual ridge regression estimator. For $\alpha < 1$ and $\lambda_{1,n} > 0$, the optimization problem is *strictly convex* with a unique solution. The elastic net is due to Zou and Hastie (2005). It has been observed that the elastic net can have much better prediction accuracy than lasso when the predictors are highly correlated.

As with lasso, it is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\mathbf{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W}$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{2.31}$$

where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Ridge regression can be computed using OLS on augmented matrices. Similarly, the elastic net can be computed using lasso on augmented matrices. Let the elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimize

$$Q_{EN}(\boldsymbol{\eta}) = RSS_W(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{2.32}$$

where $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$. Let the $(n+p-1) \times (p-1)$ augmented matrix $\boldsymbol{W}_A$ and the $(n+p-1) \times 1$ augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{W}_A = \begin{pmatrix} \boldsymbol{W} \\ \sqrt{\lambda_1}\,\boldsymbol{I}_{p-1} \end{pmatrix}, \text{ and } \boldsymbol{Z}_A = \begin{pmatrix} \boldsymbol{Z} \\ \mathbf{0} \end{pmatrix},$$

where $\mathbf{0}$ is the $(p-1) \times 1$ zero vector. Let $RSS_A(\boldsymbol{\eta}) = \|\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}\|_2^2$. Then $\hat{\boldsymbol{\eta}}_{EN}$ can be obtained from the lasso of $\boldsymbol{Z}_A$ on $\boldsymbol{W}_A$: that is, $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_L(\boldsymbol{\eta}) = RSS_A(\boldsymbol{\eta}) + \lambda_2 \|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta}). \tag{2.33}$$

Proof: We need to show that $Q_L(\boldsymbol{\eta}) = Q_{EN}(\boldsymbol{\eta})$. Note that $\boldsymbol{Z}_A^T \boldsymbol{Z}_A = \boldsymbol{Z}^T \boldsymbol{Z}$,

$$\boldsymbol{W}_A \ \boldsymbol{\eta} = \begin{pmatrix} \boldsymbol{W}\boldsymbol{\eta} \\ \sqrt{\lambda_1}\ \boldsymbol{\eta} \end{pmatrix},$$

and $\boldsymbol{Z}_A^T \boldsymbol{W}_A \ \boldsymbol{\eta} = \boldsymbol{Z}^T \boldsymbol{W}\boldsymbol{\eta}$. Then

$$RSS_A(\boldsymbol{\eta}) = \|\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}\|_2^2 = (\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta})^T(\boldsymbol{Z}_A - \boldsymbol{W}_A\boldsymbol{\eta}) =$$

$$\boldsymbol{Z}_A^T\boldsymbol{Z}_A - \boldsymbol{Z}_A^T\boldsymbol{W}_A\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}_A^T\boldsymbol{Z}_A + \boldsymbol{\eta}^T\boldsymbol{W}_A^T\boldsymbol{W}_A\boldsymbol{\eta} =$$

$$\boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{Z} + \begin{pmatrix} \boldsymbol{\eta}^T\boldsymbol{W}^T & \sqrt{\lambda_1} & \boldsymbol{\eta}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{W}\boldsymbol{\eta} \\ \sqrt{\lambda_1}\ \boldsymbol{\eta} \end{pmatrix}.$$

Thus

$$Q_L(\boldsymbol{\eta}) = \boldsymbol{Z}^T\boldsymbol{Z} - \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{\eta} - \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{Z} + \boldsymbol{\eta}^T\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{\eta} + \lambda_1\boldsymbol{\eta}^T\boldsymbol{\eta} + \lambda_2\|\boldsymbol{\eta}\|_1 =$$

$$RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta}). \ \square$$

Remark 2.16. i) You could compute the elastic net estimator using a grid of 100 $\lambda_{1,n}$ values and a grid of $J \geq 10$ α values, which would take about $J \geq 10$ times as long to compute as lasso. The above equivalent lasso problem (2.30) still needs a grid of $\lambda_1 = (1-\alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ values. Often $J = 11$, 21, 51, or 101. The elastic net estimator tends to be computed with fast methods for optimizing convex problems, such as coordinate descent. ii) Like lasso and ridge regression, the elastic net estimator is asymptotically equivalent to the OLS full model if p is fixed and $\hat{\lambda}_{1,n} = o_P(\sqrt{n})$, but behaves worse than the OLS full model otherwise. See Theorem 2.9. iii) For prediction intervals, let d be the number of nonzero coefficients from the equivalent augmented lasso problem (2.33). Alternatively, use d_2 with $d \approx d_2 = tr[\boldsymbol{W}_{AS}(\boldsymbol{W}_{AS}^T\boldsymbol{W}_{AS} + \lambda_{2,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}_{AS}^T]$ where $\boldsymbol{W}_{AS}$ corresponds to the active set (not the augmented matrix). See Tibshirani and Taylor (2012, p. 1214). Again $\lambda_{2,n}$ may not be the λ_2 given by the software. iv) The number of nonzero lasso components (not including the constant) is at most $\min(n, p-1)$. Elastic net tends to do variable selection, but the number of nonzero components can equal $p-1$ (make the elastic net equal to ridge regression). Note that the number of nonzero components in the augmented lasso problem (2.33) is at most $\min(n+p-1, p-1) = p-1$. vi) The elastic net can be computed with `glmnet`, and there is an R package `elasticnet`. vii) For fixed $\alpha > 0$, we could get λ_M for elastic net from the equivalent lasso problem. For ridge regression, we could use the λ_M for an α near 0.

Since lasso uses at most $\min(n, p-1)$ nontrivial predictors, elastic net and ridge regression can perform better than lasso if the true number of active

nontrivial predictors $a_S > \min(n, p-1)$. For example, suppose $n = 1000$, $p = 5000$, and $a_S = 1500$.

The following theorem is probably for the elastic net estimator that uses the usual ridge regression estimator of Definition 2.16 b), rather that the ridge regression estimator of Definition 2.16 c). Hence Equation (2.30) would need to be modified. Following Jia and Yu (2010), by standard Karush-Kuhn-Tucker (KKT) conditions for convex optimality for the "modified Equation (2.30)," $\hat{\boldsymbol{\beta}}_{EN}$ is optimal if

$$2\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_{EN} - 2\boldsymbol{X}^T\boldsymbol{Y} + 2\lambda_1\hat{\boldsymbol{\beta}}_{EN} + \lambda_2\boldsymbol{s}_n = \boldsymbol{0}, \quad \text{or}$$

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)\hat{\boldsymbol{\beta}}_{EN} = \boldsymbol{X}^T\boldsymbol{Y} - \frac{\lambda_2}{2}\boldsymbol{s}_n, \quad \text{or}$$

$$\hat{\boldsymbol{\beta}}_{EN} = \hat{\boldsymbol{\beta}}_R - n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{2.34}$$

Hence

$$\hat{\boldsymbol{\beta}}_{EN} = \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_1}{n}\, n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\, \hat{\boldsymbol{\beta}}_{OLS} - \frac{\lambda_2}{2n}\, n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\, \boldsymbol{s}_n$$

$$= \hat{\boldsymbol{\beta}}_{OLS} - n(\boldsymbol{X}^T\boldsymbol{X} + \lambda_1\boldsymbol{I}_p)^{-1}\, [\frac{\lambda_1}{n}\hat{\boldsymbol{\beta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n].$$

Note that if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau$ and $\hat{\alpha} \xrightarrow{P} \psi$, then $\hat{\lambda}_1/\sqrt{n} \xrightarrow{P} (1-\psi)\tau$ and $\hat{\lambda}_2/\sqrt{n} \xrightarrow{P} 2\psi\tau$. The following theorem shows elastic net is asymptotically equivalent to the OLS full model if $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$. Note that we get the RR CLT if $\psi = 0$ and the lasso CLT (using $2\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 2\tau$) if $\psi = 1$. Under these conditions,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) - n(\boldsymbol{X}^T\boldsymbol{X} + \hat{\lambda}_1\boldsymbol{I}_p)^{-1}\, [\frac{\hat{\lambda}_1}{\sqrt{n}}\hat{\boldsymbol{\beta}}_{OLS} + \frac{\hat{\lambda}_2}{2\sqrt{n}}\boldsymbol{s}_n].$$

The following theorem is due to Slawski et al. (2010), and summarized in Pelawa Watagoda and Olive (2021b).

Theorem 2.9, Elastic Net CLT. Assume p is fixed and that the conditions of the OLS CLT Equation (2.6) hold for the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$.

a) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

b) If $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0,1]$, and $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\beta}}$, then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) \xrightarrow{D} N_p\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\beta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

Proof. By the above remarks and the RR CLT Theorem 2.7,

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \hat{\boldsymbol{\beta}}_R + \hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}) + \sqrt{n}(\hat{\boldsymbol{\beta}}_{EN} - \hat{\boldsymbol{\beta}}_R)$$

$$\xrightarrow{D} N_p\left(-(1-\psi)\tau \boldsymbol{V}\boldsymbol{\beta}, \sigma^2 \boldsymbol{V}\right) \quad - \quad \frac{2\psi\tau}{2}\boldsymbol{V}\boldsymbol{s}$$

$$\sim N_p\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\beta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

The mean of the normal distribution is $\mathbf{0}$ under a) since $\hat{\alpha}$ and $\boldsymbol{s}_n$ are bounded. □

Example 2.2, continued. The `slpack` function `enet` does elastic net using 10-fold CV and a grid of α values $\{0, 1/am, 2/am, ..., am/am = 1\}$. The default uses $am = 10$. The default chose lasso with $alph = 1$. The function also makes a response plot, but does not add the lines for the pointwise prediction intervals since the false degrees of freedom d is not computed.

```
library(glmnet); y <- marry[,3]; x <- marry[,-3]
tem <- enet(x,y)
tem$alph
[1] 1  #elastic net was lasso
tem<-enet(x,y,am=100)
tem$alph
[1] 0.97 #elastic net was not lasso with a finer grid
```

The *elastic net variable selection* estimator applies OLS to a constant and the active predictors that have nonzero elastic net $\hat{\eta}_i$. Hence elastic net is used as a variable selection method. Let $\boldsymbol{X}_A$ denote the matrix with a column of ones and the unstandardized active nontrivial predictors. Hence the elastic net variable selection estimator is $\hat{\boldsymbol{\beta}}_{ENV} = (\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^T\boldsymbol{Y}$, and elastic net variable selection is an alternative to forward selection. Let k be the number of active (nontrivial) predictors so $\hat{\boldsymbol{\beta}}_{ENV}$ is $(k+1)\times 1$. Let I_{min} correspond to the elastic net variable selection estimator and $\hat{\boldsymbol{\beta}}_{ENV,0} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ to the zero padded elastic net variable selection estimator. When p is fixed, $\hat{\boldsymbol{\beta}}_{ENV,0}$ is $\sqrt{n}$ consistent when elastic net is consistent, with the limiting distribution for $\hat{\boldsymbol{\beta}}_{ENV,0}$ given by Rathnayake and Olive (2023). Elastic net variable selection will often be better than elastic net when the model is sparse or if $n \geq 10(k+1)$. The elastic net can be better than elastic net variable selection if $(\boldsymbol{X}_A^T\boldsymbol{X}_A)$ is ill conditioned or if $n/(k+1) < 10$.

2.10 OPLS

Cook, Helland, and Su (2013) showed that the OPLS estimator $\hat{\boldsymbol{\beta}}_{OPLS}$ estimates $\boldsymbol{\beta}_{OPLS}$, and that the OPLS estimator can be computed from the OLS simple linear regression (SLR) of Y on $W = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}$, giving

$\hat{Y} = \hat{\alpha}_{OPLS} + \hat{\lambda} W = \hat{\alpha}_{OPLS} + \hat{\boldsymbol{\beta}}_{OPLS}^T \boldsymbol{x}$. Also see Basa et al. (2024) and Wold (1975).

Definition 2.20. The *one component partial least squares (OPLS) estimator* $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\lambda} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ estimates $\lambda \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$ where

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}} \text{ and } \hat{\lambda} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}} \tag{2.35}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \mathbf{0}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$, then $\boldsymbol{\beta}_{OPLS} = \mathbf{0}$.

The following Olive and Zhang (2024) theorem gives some large sample theory for $\hat{\boldsymbol{\eta}} = \widehat{\mathrm{Cov}}(\boldsymbol{x}, Y)$. This theory needs $\boldsymbol{\eta} = \boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ to exist for $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ to be a consistent estimator of $\boldsymbol{\eta}$. Let $\boldsymbol{x}_i = (x_{i1}, ..., x_{ip})^T$ and let $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ be defined below where

$$\mathrm{Cov}(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{w}} = E[(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})^T (Y_i - \mu_Y)^2)] - \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T.$$

Then the low order moments are needed for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ to be a consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{w}}$. The theory uses milder regularity conditions than the theory in the previous literature. The theory can be used for testing, including some high dimensional tests for low dimensional quantities such as $H_O : \beta_i = 0$ or $H_0 : \beta_i - \beta_j = 0$. These tests depended on iid cases, but not on linearity or the constant variance assumption. Data splitting uses model selection (variable selection is a special case) to reduce the high dimensional problem to a low dimensional problem. Olive et al. (2024) gave alternative proofs, and showed that the results hold for multiple linear regression with heterogeneity.

Theorem 2.10. Assume the cases $(\boldsymbol{x}_i^T, Y_i)^T$ are iid. Assume $E(x_{ij}^k \; Y_i^m)$ exist for $j = 1, ..., p$ and $k, m = 0, 1, 2$. Let $\boldsymbol{\mu}_{\boldsymbol{x}} = E(\boldsymbol{x})$ and $\mu_Y = E(Y)$. Let $\boldsymbol{w}_i = (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)$ with sample mean $\overline{\boldsymbol{w}}_n$. Let $\boldsymbol{\eta} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Then a)

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \; \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \tag{2.36}$$

$$\text{and } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}).$$

b) Let $\boldsymbol{z}_i = \boldsymbol{x}_i(Y_i - \overline{Y}_n)$ and $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(Y_i - \overline{Y}_n)$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + O_P(n^{-1/2}) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$. Hence $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + O_P(n^{-1/2}) = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$.
c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$, assume $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \mathbf{0}$ is true, and assume $\hat{\lambda} \xrightarrow{P} \lambda \neq 0$. Then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_k(\mathbf{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{A}^T). \tag{2.37}$$

Proof. a) Note that $\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$ by the multivariate central limit theorem since the $\boldsymbol{w}_i$ are iid with $E(\boldsymbol{w}_i) = \boldsymbol{\eta} = \mathrm{Cov}(\boldsymbol{x}, Y)$ and

$\mathrm{Cov}(\boldsymbol{w}) = \boldsymbol{\Sigma_w}$. Now $n\tilde{\boldsymbol{\eta}}_n =$

$$\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu_x} + \boldsymbol{\mu_x} - \overline{\boldsymbol{x}})(Y_i - \mu_Y + \mu_Y - \overline{Y}) = \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu_x})(Y_i - \mu_Y)$$

$$+ \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu_x})(\mu_Y - \overline{Y}) + (\boldsymbol{\mu_x} - \overline{\boldsymbol{x}})\sum_i (Y_i - \mu_Y) + n(\boldsymbol{\mu_x} - \overline{\boldsymbol{x}})(\mu_Y - \overline{Y})$$

$$= \sum_i \boldsymbol{w}_i - n\boldsymbol{a}_n - n\boldsymbol{a}_n + n\boldsymbol{a}_n = \sum_i \boldsymbol{w}_i - n(\boldsymbol{\mu_x} - \overline{\boldsymbol{x}})(\mu_Y - \overline{Y}).$$

$$\text{Thus } \sqrt{\mathrm{n}}\tilde{\boldsymbol{\eta}}_{\mathrm{n}} = \sqrt{\mathrm{n}}\frac{1}{\mathrm{n}}\sum_{\mathrm{i}} \boldsymbol{w}_{\mathrm{i}} - \frac{\sqrt{\mathrm{n}}(\overline{\boldsymbol{x}} - \boldsymbol{\mu_x})\sqrt{\mathrm{n}}(\overline{\mathrm{Y}} - \mu_{\mathrm{Y}})}{\sqrt{\mathrm{n}}} = \sqrt{\mathrm{n}}\,\overline{\boldsymbol{w}}_{\mathrm{n}} + \mathrm{o_P}(1).$$

$$\text{Hence } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) = \sqrt{\mathrm{n}}(\overline{\boldsymbol{w}}_{\mathrm{n}} - \boldsymbol{\eta}) + \mathrm{o_P}(1).$$

$$\text{Thus } \sqrt{\mathrm{n}}(\tilde{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{\Sigma_w})$$

by Slutsky's theorem. Now

$$\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = \sqrt{n}\left(\frac{n}{n-1}\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}\right) = \sqrt{n}\left(\frac{n}{n-1}\tilde{\boldsymbol{\eta}} - \frac{n}{n-1}\boldsymbol{\eta} + \frac{n}{n-1}\boldsymbol{\eta} - \boldsymbol{\eta}\right)$$

$$= \sqrt{n}\frac{n}{n-1}(\tilde{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \sqrt{n}\left(\frac{\boldsymbol{\eta}}{n-1}\right).$$

$$\text{Thus } \sqrt{\mathrm{n}}(\hat{\boldsymbol{\eta}}_{\mathrm{n}} - \boldsymbol{\eta}) \xrightarrow{\mathrm{D}} \mathrm{N_p}(\boldsymbol{0}, \boldsymbol{\Sigma_w}).$$

b) See Olive et al. (2024).

c) If H_0 is true, then $\boldsymbol{A\eta} = \boldsymbol{0}$, and

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) = \sqrt{n}\boldsymbol{A}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \hat{\lambda}\boldsymbol{\eta} + \hat{\lambda}\boldsymbol{\eta} - \boldsymbol{\beta}_{OPLS}) =$$

$$\hat{\lambda}\boldsymbol{A}\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) + \boldsymbol{A}\sqrt{n}(\hat{\lambda} - \lambda)\boldsymbol{\eta} = \boldsymbol{Z}_n + \boldsymbol{b}_n \xrightarrow{D} N_k(\boldsymbol{0}, \lambda^2 \boldsymbol{A\Sigma_w A}^T)$$

since $\boldsymbol{b}_n = \boldsymbol{0}$ when H_0 is true. □

In Theorems 2.10 and 2.11, the scalars λ and $\hat{\lambda}$ are given by Equation (2.35), $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)^T$, and $\boldsymbol{\Sigma_\eta} = \boldsymbol{\Sigma_w}$. Results from Su and Cook (2012) and Olive et al. (2024), for example, show that elements of a sample covariance matrix can be stacked to get large sample theory. Then $\hat{\lambda}$ and $\hat{\boldsymbol{\eta}}$ can be stacked as in Theorem 2.11 by the multivariate delta method. Theorem 2.10 c) and Theorem 2.11 c) are equivalent with different notation. Currently $\boldsymbol{\Sigma}$ from Theorem 2.11 is difficult to estimate.

Theorem 2.11. Assume

$$\sqrt{n}\left(\begin{pmatrix}\hat{\lambda}\\ \hat{\boldsymbol{\eta}}\end{pmatrix} - \begin{pmatrix}\lambda\\ \boldsymbol{\eta}\end{pmatrix}\right) \xrightarrow{D} N_{p+1}\left(\begin{pmatrix}0\\ \boldsymbol{0}\end{pmatrix}, \begin{pmatrix}\Sigma_\lambda & \boldsymbol{\Sigma}_{\lambda\boldsymbol{\eta}}\\ \boldsymbol{\Sigma}_{\boldsymbol{\eta}\lambda} & \boldsymbol{\Sigma_\eta}\end{pmatrix}\right) \sim N_{p+1}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

a) $\sqrt{n}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\eta}})$.

b) $\sqrt{n}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \lambda\boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_p\left(\mathbf{0}, \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T\right)$ with $\boldsymbol{D} = [\boldsymbol{\eta} \;\; \lambda \boldsymbol{I}_p]$ where $\boldsymbol{I}_p$ is the $p \times p$ identity matrix.

c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \leq p$ and $\boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \mathbf{0} = \boldsymbol{A}\boldsymbol{\eta}$. Then

$$\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{OPLS} - \mathbf{0}) \xrightarrow{D} N_k\left(\mathbf{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{A}^T\right).$$

Proof. a) Follows by Equation (2.36) or since joint convergence in distribution implies marginal convergence in distribution.

b) Follows by the Multivariate Delta Method with

$$\boldsymbol{g}\begin{pmatrix}\lambda \\ \boldsymbol{\eta}\end{pmatrix} = \lambda\boldsymbol{\eta} =$$

$(\lambda\eta_1, ..., \lambda\eta_p)^T$, and the Jacobian matrix of partial derivatives $\boldsymbol{D} = \boldsymbol{D}_{\boldsymbol{g}}$.

$$\text{c) By b), } \sqrt{\mathrm{n}}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{\mathrm{OPLS}} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{\mathrm{D}} \mathrm{N_k}\left(\mathbf{0}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^{\mathrm{T}}\boldsymbol{A}^{\mathrm{T}}\right),$$

but $\boldsymbol{A}\boldsymbol{D} = [\mathbf{0} \;\; \lambda\boldsymbol{A}]$. Hence $\boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T\boldsymbol{A}^T = \lambda^2\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{A}^T$. □

Some additional useful OPLS and OLS formulas are derived next if the cases are iid. Let $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Then $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \mathrm{Cov}(\boldsymbol{x}, Y) = \mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$. Since $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}$,

$$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}, \; \boldsymbol{\beta}_{OPLS} = \lambda\mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta}_{OLS}, \text{ and}$$

$$\boldsymbol{\beta}_{OLS} = \frac{1}{\lambda}[\mathrm{Cov}(\boldsymbol{x})]^{-1}\boldsymbol{\beta}_{OPLS}.$$

Chun and Keleş (2010) suggested that $\hat{\boldsymbol{\beta}}_{OPLS}$ only estimates $\boldsymbol{\beta}_{OLS}$ under very strong regularity conditions. For iid cases, Cook and Forzani (2018, 2019) showed that the regularity condition is $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, in which case $\sqrt{n}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OLS}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{C})$. Cook and Forzani (2018, 2019) also showed that under very strong regularity conditions for high dimensions, $\hat{\boldsymbol{\beta}}_{OPLS}$ is a consistent estimator of $\boldsymbol{\beta}_{OLS}$. Also see Basa et al. (2024).

In the literature, there is a tendency (perhaps a common Statistical paradigm) to assume that if the estimated model fits the data well, then the model corresponding to the estimator is the model for $Y|\boldsymbol{x}$. For example, in much of the OPLS literature, an assumption is $Y|\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$. Then $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$ by the OLS CLT, and the results in Table 2.1 hold.

The above tendency leads to problems that have perhaps not often been observed in the literature. To see some problems, consider multiple linear regression with $\mathrm{Cov}(\boldsymbol{x}) = diag(1, 2, ..., p)$. First consider OPLS with $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS}$. Then at most one element of $\mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is nonzero since

Table 2.1 OPLS Results

General	$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\beta}_{OPLS}$
$\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{1}{\lambda}[Cov(\boldsymbol{x})]^{-1}\boldsymbol{\beta}_{OPLS}$	$\boldsymbol{\beta}_{OLS}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$
$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda Cov(\boldsymbol{x})\boldsymbol{\beta}_{OLS}$	$\boldsymbol{\beta}_{OPLS}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$
$\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = Cov(\boldsymbol{x})\boldsymbol{\beta}_{OLS}$	$\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$
$\hat{\boldsymbol{\beta}}_{kPLS}$ estimates $\boldsymbol{\beta}_{kPLS}$	$\hat{\boldsymbol{\beta}}_{kPLS}$ estimates $\boldsymbol{\beta}_{OLS}$

$\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is an eigenvector of $\text{Cov}(\boldsymbol{x})$. Hence at most one predictor is correlated with Y, regardless of the value of p. This restriction is too strong.

If the cases are iid from a multivariate normal distribution, then $Y|\boldsymbol{x} = \alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} + e$ and $Y|\boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$ are both linear models by Section 2.16 where e depends on the model. Since $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$ forces $\boldsymbol{\beta}_{OLS}$ to be an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, if $\boldsymbol{\beta}_{OLS}$ is not an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$. For a computational example, let $\boldsymbol{x} \sim N_p(\boldsymbol{0}, diag(1,2,3,4))$ with $\boldsymbol{\Sigma}_{\boldsymbol{x}} = diag(1,2,3,4)$, and let the population generating model be $Y_i = x_{i1} + x_{i2} + e_i$ for $i = 1, ..., n$ where the e_i are iid $N(0,1)$ and independent of the $\boldsymbol{x}_i$. Then $\alpha = 0$ and $\boldsymbol{\beta} = (1,1,0,0)^T$. Hence $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta} = (1,1,0,0)^T$, $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS} = (1,2,0,0)^T$, and

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x},Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}} = 5/9.$$

Thus $\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS} = (5/9, 10/9, 0, 0)^T \neq \boldsymbol{\beta}_{OLS}$.

Thus OLS and OPLS usually give different valid population multiple linear regression models with $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$. However, the model $Y|\boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$ is often a useful multiple linear regression model with large sample theory given by Theorem 2.11. The claims in the OPLS literature that $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS} =$ an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ under mild regularity conditions are incorrect. See, for example, Basa et al. (2024), Cook and Forzani (2018, 2019, 2024), and Cook, Helland and Su (2013). The regularity conditions for $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS}$ are very strong. In the OLS literature $\boldsymbol{\beta}_{OLS}$ can be any vector in $\mathbb{R}^p$. If $\boldsymbol{\beta}_{OLS}$, $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, and $\boldsymbol{\beta}_{OPLS}$ were restricted to be eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then the OLS and OPLS estimators would often not fit the data well.

2.11 The MMLE

The marginal maximum likelihood estimator (MMLE or marginal least squares estimator) is due to Fan and Lv (2008) and Fan and Song (2010). This estimator computes the marginal regression of Y on x_i resulting in the estimator $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M})$ for $i = 1, ..., p$. Then $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_{1,M}, ..., \hat{\beta}_{p,M})^T$.

For multiple linear regression, the marginal estimators are the simple linear regression (SLR) estimators, and $(\hat{\alpha}_{i,M}, \hat{\beta}_{i,M}) = (\hat{\alpha}_{i,SLR}, \hat{\beta}_{i,SLR})$. Hence

$$\hat{\boldsymbol{\beta}}_{MMLE} = [diag(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})]^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}.$$

If the $\boldsymbol{t}_i$ are the predictors are scaled or standardized to have unit sample variances, then

$$\hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{t}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}Y} = \boldsymbol{I}^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t}Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{t}, Y) \tag{2.38}$$

where $(\boldsymbol{t}, Y)$ denotes that Y was regressed on $\boldsymbol{t}$, and $\boldsymbol{I}$ is the $p \times p$ identity matrix. Olive et al. (2024) gave some large sample theory for the MMLE.

The MMLE is also used for variable selection. For example, standardize the predictors and take the $K - 1$ variables corresponding to the largest $|\hat{\beta}_i|$ where $\hat{\boldsymbol{\beta}}_{MMLE} = (\hat{\beta}_1, ..., \hat{\boldsymbol{\beta}}_p)^T$. Then perform the regression on these variables (perhaps not standardized) and a constant. This variable selection method is useful for very large p since the method is fast, but the selected predictors are often highly correlated. Hence it may be useful to perform lasso variable selection or forward selection using the variables selected by MMLE variable selection. Choosing K near $\min(n/J, p)$ for $J = 1, 5$ or 10 may be useful.

MMLE variable selection can also be useful when the predictors are orthogonal. See Goh and Dey (2019) for references. This result may be useful for PCR, PLS, and wavelets.

2.12 k-Component Regression Estimators

Consider the MLR model $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. The k-component regression estimators, such as PCR and PLS, use p linear combinations $\boldsymbol{\eta}_1^T\boldsymbol{x}, ..., \boldsymbol{\eta}_p^T\boldsymbol{x}$. Then there are p conditional distributions

$$\begin{gathered} Y|\boldsymbol{\eta}_1^T\boldsymbol{x} \\ Y|(\boldsymbol{\eta}_1^T\boldsymbol{x}, \boldsymbol{\eta}_2^T\boldsymbol{x}) \\ \vdots \\ Y|(\boldsymbol{\eta}_1^T\boldsymbol{x}, \boldsymbol{\eta}_2^T\boldsymbol{x}, ..., \boldsymbol{\eta}_p^T\boldsymbol{x}). \end{gathered}$$

Estimating the $\boldsymbol{\eta}_i$ and performing the ordinary least squares (OLS) regression of Y on $(\hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}, \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x}, ..., \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x})$ gives the k-component estimator, e.g. the k-component PLS estimator $\hat{\boldsymbol{\beta}}_{kPLS}$ or the k-component PCR estimator, for $k = 1, ..., J$ where $J \leq p$ and the p-component estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$.

Definition 2.21. Consider the MLR model $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Let $\boldsymbol{X} = (\boldsymbol{1} \quad \boldsymbol{X}_1)$. Let

$$\boldsymbol{v}_i = \hat{\boldsymbol{A}}_{k,n}\boldsymbol{x}_i = \begin{pmatrix} \boldsymbol{x}_i^T\hat{\boldsymbol{\eta}}_1 \\ \vdots \\ \boldsymbol{x}_i^T\hat{\boldsymbol{\eta}}_k \end{pmatrix} = \begin{pmatrix} \hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}_i \\ \vdots \\ \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x}_i \end{pmatrix} \text{ where } \hat{\boldsymbol{A}}_{\mathrm{k,n}} = \begin{pmatrix} \hat{\boldsymbol{\eta}}_1^T \\ \vdots \\ \hat{\boldsymbol{\eta}}_k^T \end{pmatrix}.$$

Let

$$\boldsymbol{c}_i = \boldsymbol{X}_1\hat{\boldsymbol{\eta}}_i = \begin{pmatrix} \boldsymbol{x}_1^T\hat{\boldsymbol{\eta}}_i \\ \vdots \\ \boldsymbol{x}_n^T\hat{\boldsymbol{\eta}}_i \end{pmatrix}$$

be the *ith component vector* for $i = 1, ..., p$. Let

$$\boldsymbol{V}_k = (\boldsymbol{c}_1, ..., \boldsymbol{c}_k) = \begin{pmatrix} \boldsymbol{v}_1^T \\ \vdots \\ \boldsymbol{v}_n^T \end{pmatrix} = \boldsymbol{X}_1\hat{\boldsymbol{A}}_{k,n}^T$$

for $k = 1, ..., p$. Let the working OLS model

$$\boldsymbol{Y} = \alpha_k\boldsymbol{1} + \boldsymbol{V}_k\boldsymbol{\gamma}_k + \boldsymbol{\epsilon}$$

where $\boldsymbol{\epsilon}$ depends on the model. Then $\hat{\boldsymbol{\beta}}_{kE} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k$ is the k-component estimator for $k = 1, ..., p$. The model selection estimator chooses one of the k-component estimators, e.g. using a holdout sample or cross validation, and will be denoted by $\hat{\boldsymbol{\beta}}_{MS,E}$.

The OLS regression of Y on $\boldsymbol{w} = \hat{\boldsymbol{A}}_{k,n}\boldsymbol{x}$ gives

$$\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w},Y} = (\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T)^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}.$$

Thus

$$\hat{\boldsymbol{\beta}}_{kE} = \hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k = \hat{\boldsymbol{A}}_{k,n}^T(\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T)^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y} = \hat{\boldsymbol{\Lambda}}_k\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$$

$$= \hat{\boldsymbol{A}}_{k,n}^T(\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{A}}_{k,n}^T)^{-1}\hat{\boldsymbol{A}}_{k,n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}, Y) = \hat{\boldsymbol{\Lambda}}_k\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}, Y).$$

If $\hat{\boldsymbol{\eta}}_i \xrightarrow{P} \boldsymbol{\eta}_i$, and

$$\hat{\boldsymbol{A}}_{k,n} \xrightarrow{P} \boldsymbol{A}_k = \begin{pmatrix} \boldsymbol{\eta}_1^T \\ \vdots \\ \boldsymbol{\eta}_k^T \end{pmatrix},$$

then

$$\hat{\boldsymbol{\beta}}_{kE} \xrightarrow{P} \boldsymbol{\beta}_{kE} = \boldsymbol{A}_k^T(\boldsymbol{A}_k\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{A}_k^T)^{-1}\boldsymbol{A}_k\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y) = \boldsymbol{\Lambda}_k\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y).$$

This convergence can also occur if $\hat{\boldsymbol{\eta}}_i = \hat{\boldsymbol{e}}_i$ are orthonormal eigenvectors such that $\hat{\boldsymbol{A}}_{k,n}^T\hat{\boldsymbol{\gamma}}_k \xrightarrow{P} \boldsymbol{A}_k^T\boldsymbol{\gamma}_k$, which happened for PCR.

The regularity conditions for $\boldsymbol{\beta}_{kE} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$ tend to be very strong, at least for k near 1. Note that $\boldsymbol{\beta}_{pE} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$ if the inverse matrices exist (and if $p = 1$), and $\boldsymbol{\beta}_{kE} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$ if $\boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y) = \mathbf{0}$. Suppose $\boldsymbol{\beta}_{OLS} = \sum_{j=1}^{m} c_{i_j} \boldsymbol{\eta}_{i_j}$ for some m where $1 \leq m \leq p$ and the $c_{i_j} \neq 0$. If k is large enough to include the m $\boldsymbol{\eta}_{i_j}$, then $\boldsymbol{\beta}_{kE} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$. This regularity condition becomes weaker as m increases, and $\boldsymbol{\beta}_{kE}$ can become very highly correlated with $\boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$ as k increases.

In the high dimensional setting, the regularity conditions for $\hat{\boldsymbol{\eta}}_i \xrightarrow{P} \boldsymbol{\eta}_i$ tend to be very strong.

2.13 Prediction Intervals

This section will use the prediction intervals applied to the MLR model with $\hat{Y} = \boldsymbol{x}_I^T \hat{\boldsymbol{\beta}}_I$ and I corresponds to the predictors used by the MLR method. We will use the six methods forward selection with OLS, PCR, PLS, lasso, lasso variable selection, and ridge regression. The number of components for PLS and PCR will be selected using cross validation, hence the model selction versions of PLS and PCR are used. When $p > n$, results from Hastie et al. (2015, pp. 20, 296, ch. 6, ch. 11) and Luo and Chen (2013) suggest that lasso, lasso variable selection, and forward selection with EBIC can perform well for sparse models: the subset S in Equation (2.14) and Remark 2.8 has a_S small.

Notation: $P(A_n)$ is "eventually bounded below" by $1 - \delta$ if $P(A_n)$ gets arbitrarily close to or higher than $1 - \delta$ as $n \to \infty$. Hence $P(A_n) > 1 - \delta - \epsilon$ for any $\epsilon > 0$ if n is large enough. If $P(A_n) \to 1 - \delta$ as $n \to \infty$, then $P(A_n)$ is eventually bounded below by $1 - \delta$. The actual coverage is $1 - \gamma_n = P(Y_f \in [L_n, U_n])$, the nominal coverage is $1 - \delta$ where $0 < \delta < 1$. The 90% and 95% large sample prediction intervals and prediction regions are common.

Definition 2.22. Consider predicting a future test value Y_f given a $p \times 1$ vector of predictors $\boldsymbol{x}_f$ and training data $(Y_1, \boldsymbol{x}_1), ..., (Y_n, \boldsymbol{x}_n)$. A large sample $100(1 - \delta)\%$ *prediction interval* (PI) for Y_f has the form $[\hat{L}_n, \hat{U}_n]$ where $P(\hat{L}_n \leq Y_f \leq \hat{U}_n)$ is eventually bounded below by $1 - \delta$ as the sample size $n \to \infty$. A large sample $100(1 - \delta)\%$ PI is *asymptotically optimal* if it has the shortest asymptotic length: the length of $[\hat{L}_n, \hat{U}_n]$ converges to $U_s - L_s$ as $n \to \infty$ where $[L_s, U_s]$ is the population shorth: the shortest interval covering at least $100(1 - \delta)\%$ of the mass.

If $Y_f | \boldsymbol{x}_f$ has a pdf, we often want $P(\hat{L}_n \leq Y_f \leq \hat{U}_n) \to 1 - \delta$ as $n \to \infty$. The interpretation of a 100 $(1 - \delta)\%$ PI for a random variable Y_f is similar to that of a confidence interval (CI). Collect data, then form the PI, and repeat for a total of k times where the k trials are independent from the same population. If Y_{fi} is the ith random variable and PI_i is the ith PI,

then the probability that $Y_{fi} \in PI_i$ for j of the PIs approximately follows a binomial$(k, \rho = 1 - \delta)$ distribution. Hence if 100 95% PIs are made, $\rho = 0.95$ and $Y_{fi} \in PI_i$ happens about 95 times.

There are two big differences between CIs and PIs. First, the length of the CI goes to 0 as the sample size n goes to ∞ while the length of the PI converges to some nonzero number J, say. Secondly, many confidence intervals work well for large classes of distributions while many prediction intervals assume that the distribution of the data is known up to some unknown parameters. Usually the $N(\mu, \sigma^2)$ distribution is assumed, and the parametric PI may not perform well if the normality assumption is violated. This section will describe three nonparametric PIs for the multiple linear regression model, $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e$, that work well for a large class of unknown zero mean error distributions.

Consider the location model, $Y_i = \mu + e_i$, where $Y_1, ..., Y_n, Y_f$ are iid, and there are no vectors of predictors $\boldsymbol{x}_i$ and $\boldsymbol{x}_f$. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the iid training data $Y_1, ..., Y_n$. Then the unknown future value Y_f is the test data.

Remark 2.17. Confidence intervals, prediction intervals, confidence regions, and prediction regions should used closed sets not open sets. The closed sets have the same volume as the open sets, but have coverage at least as high as the open sets with weaker regularity conditions. In particular, confidence and prediction intervals should be closed intervals, not open intervals.

In the following theorem, if the open interval $(Y_{(k_1)}, Y_{(k_2)})$ was used, we would need to add the regularity condition that $Y_{\delta/2}$ and $Y_{1-\delta/2}$ are continuity points of $F_Y(y)$.

Theorem 2.12. Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. Let $k_1 = \lceil n\delta/2 \rceil$ and $k_2 = \lceil n(1-\delta/2) \rceil$ where $0 < \delta < 1$. The large sample $100(1-\delta)\%$ *percentile prediction interval* for Y_f is

$$[Y_{(k_1)}, Y_{(k_2)}]. \tag{2.39}$$

The shorth(c) estimator of the population shorth is useful for making asymptotically optimal prediction intervals. For the uniform distribution, the population shorth is not unique. Of course the length of the population shorth is unique. For a large sample $100(1-\delta)\%$ PI, the nominal coverage is $100(1-\delta)\%$. Undercoverage occurs if the actual coverage is below the nominal coverage. For example, if the actual coverage is 0.93 for a large sample 95% PI, than the undercoverage is 0.02.

Definition 2.23. Let the shortest closed interval containing at least c of the $Y_1, ..., Y_n$ be

$$\text{shorth(c)} = [\mathrm{Y_{(s)}}, \mathrm{Y_{(s+c-1)}}]. \tag{2.40}$$

Theorem 2.13, Frey (2013). Let $Y_1, ..., Y_n$ be iid. Let

$$k_n = \lceil n(1-\delta) \rceil. \tag{2.41}$$

For large $n\delta$ and iid data, the large sample $100(1-\delta)\%$ shorth(k_n) prediction interval has maximum undercoverage $\approx 1.12\sqrt{\delta/n}$. The maximum undercoverage occurs for the family of uniform $U(\theta_1, \theta_2)$ distributions.

Theorem 2.14, Frey (2013). Let $Y_1, ..., Y_n, Y_f$ be iid. Let $Y_{(1)} \leq Y_{(2)} \leq \cdots \leq Y_{(n)}$ be the order statistics of the training data. The large sample $100(1-\delta)\%$ shorth(c) prediction interval for Y_f is

$$[Y_{(s)}, Y_{(s+c-1)}] \text{ where } \mathrm{c} = \min(\mathrm{n}, \lceil \mathrm{n}[1 - \delta + 1.12\sqrt{\delta/\mathrm{n}}\ ]\ \rceil). \tag{2.42}$$

A problem with the prediction intervals that cover $\approx 100(1-\delta)\%$ of the training data cases Y_i (such as (2.40) using $c = k_n$ given by (2.41)), is that they have coverage lower than the nominal coverage of $1-\delta$ for moderate n. This result is not surprising since empirically statistical methods perform worse on test data. For iid data, Frey (2013) used (2.42) to correct for undercoverage.

Remark 2.18. a) The shorth PI (2.42) often has good coverage for $n \geq 50$ and $0.05 \leq \delta \leq 0.1$, but the convergence of $U_n - L_n$ to the population shorth length $U_s - L_s$ can be quite slow. Under regularity conditions, Grübel (1982) showed that for iid data, the length and center the shorth(k_n) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval, respectively. The correction factor also increases the length. For a unimodal and symmetric error distribution, the nonparametric percentile PI (2.39) and the shorth PI (2.42) are asymptotically equivalent, but PI (2.39) can be the shorter. b) The percentile PI (2.39) can be much longer than the shorth PI (2.42) if the data distribution is skewed.

Example 2.3. Given below were votes for preseason 1A basketball poll from Nov. 22, 2011 WSIL News where the 778 was a typo: the actual value was 78. As shown below, finding shorth(3) from the ordered data is simple. If the outlier was corrected, shorth(3) = [76,78].

```
111   89   778   78   76

order data: 76 78 89 111 778

                  13 = 89 - 76

                      33 = 111 - 78

                         689 = 778 - 89
shorth(3) = [76,89]
```

Many things can go wrong with prediction. It is assumed that the test data follows the same MLR model as the training data. Population drift is a common reason why the above assumption, which assumes that the various distributions involved do not change over time, is violated. Population drift occurs when the population distribution does change over time.

A second thing that can go wrong is that the training or test data set is distorted away from the population distribution. This could occur if outliers are present or if the training data set and test data set are drawn from different populations. For example, the training data set could be drawn from three hospitals, and the test data set could be drawn from two more hospitals. These two populations of three and two hospitals may differ.

A third thing that can go wrong is *extrapolation*: if $\boldsymbol{x}_f$ is added to $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$, then there is extrapolation if $\boldsymbol{x}_f$ is not like the $\boldsymbol{x}_i$, e.g. $\boldsymbol{x}_f$ is an outlier. Predictions based on extrapolation are not reliable. Check whether the Euclidean distance of $\boldsymbol{x}_f$ from the coordinatewise median MED($\boldsymbol{X}$) of the $\boldsymbol{x}_1, ..., \boldsymbol{x}_n$ satisfies $D_{\boldsymbol{x}_f}(\text{MED}(\boldsymbol{X}), \boldsymbol{I}_p) \leq \max_{i=1,...,n} D_i(\text{MED}(\boldsymbol{X}), \boldsymbol{I}_p)$. Alternatively, use the `ddplot5` function, described in Chapter 1, applied to $\boldsymbol{x}_1, ..., \boldsymbol{x}_n, \boldsymbol{x}_f$ to check whether $\boldsymbol{x}_f$ is an outlier.

When $n \geq 10p$, let the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$. Let $h_i = h_{ii}$ be the ith diagonal element of $\boldsymbol{H}$ for $i = 1, ..., n$. Then h_i is called the ith **leverage** and $h_i = \boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i$. Then the leverage of $\boldsymbol{x}_f$ is $h_f = \boldsymbol{x}_f^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_f$. Then a rule of thumb is that extrapolation occurs if $h_f > \max(h_1, ..., h_n)$. This rule works best if the predictors are linearly related in that a plot of x_i versus x_j should not have any strong nonlinearities. If there are strong nonlinearities among the predictors, then $\boldsymbol{x}_f$ could be far from the $\boldsymbol{x}_i$ but still have $h_f < \max(h_1, ..., h_n)$. If the regression method, such as lasso or forward selection, uses a set I of a predictors, including a constant, where $n \geq 10a$, the above rule of thumb could be used for extrapolation where $\boldsymbol{x}_f$, $\boldsymbol{x}_i$, and $\boldsymbol{X}$ are replaced by $\boldsymbol{x}_{I,f}$, $\boldsymbol{x}_{I,i}$, and $\boldsymbol{X}_I$.

Prediction intervals based on the shorth of the residuals need a correction factor for good coverage since the residuals tend to underestimate the errors in magnitude. With the exception of ridge regression, let d be the number of "variables" used by the method. For MLR, forward selection, lasso, and lasso variable selection use variables $x_1^*, ..., x_d^*$ while PCR and PLS use variables that are linear combinations of the predictors $V_j = \boldsymbol{\gamma}_j^T\boldsymbol{x}$ for $j = 1, ..., d$. We want $n \geq 10d$ so that the model does not overfit. (We could let $d = j$ if j is the degrees of freedom of the selected model if that model was chosen in advance without model or variable selection. Hence $d = j$ is not the model degrees of freedom if model selection was used.) See Hong et al. (2018) for why classical prediction intervals after variable selection fail to work.

Pelawa Watagoda and Olive (2021b) gave two prediction intervals that can be useful even if n/p is not large. These PIs will be defined below. If the OLS model I has d predictors, and $S \subseteq I$, then

$$E(MSE(I)) = E\left(\sum_{i=1}^{n} \frac{r_i^2}{n-d}\right) = \sigma^2 = E\left(\sum_{i=1}^{n} \frac{e_i^2}{n}\right)$$

and $MSE(I)$ is a $\sqrt{n}$ consistent estimator of σ^2 for many error distributions by Su and Cook (2012). Also see Freedman (1981). For a wide range of regression models, extrapolation occurs if the leverage $h_f = \boldsymbol{x}_{I,f}^T(\boldsymbol{X}_I^T\boldsymbol{X}_I)^{-1}\boldsymbol{x}_{I,f} > 2d/n$: if $\boldsymbol{x}_{I,f}$ is too far from the data $\boldsymbol{x}_{I,1}, ..., \boldsymbol{x}_{I,n}$, then the model may not hold and prediction can be arbitrarily bad. These results suggests that

$$\sqrt{\frac{n}{n-d}}\sqrt{(1+h_f)} \;\; r_i \approx \sqrt{\frac{n+2d}{n-d}} \;\; r_i \approx e_i.$$

In simulations for prediction intervals and prediction regions with $n = 20d$, the maximum simulated undercoverage was near 5% if q_n in (2.43) is changed to $q_n = 1 - \delta$.

Next we give the correction factor and the first prediction interval. Let $q_n = \min(1 - \delta + 0.05, 1 - \delta + d/n)$ for $\delta > 0.1$ and

$$q_n = \min(1 - \delta/2, 1 - \delta + 10\delta d/n), \quad \text{otherwise.} \tag{2.43}$$

If $1 - \delta < 0.999$ and $q_n < 1 - \delta + 0.001$, set $q_n = 1 - \delta$. Let

$$c = \lceil nq_n \rceil, \tag{2.44}$$

and let

$$b_n = \left(1 + \frac{15}{n}\right)\sqrt{\frac{n+2d}{n-d}} \tag{2.45}$$

if $d \leq 8n/9$, and

$$b_n = 5\left(1 + \frac{15}{n}\right),$$

otherwise. As d gets close to n, the model overfits and the coverage will be less than the nominal. The piecewise formula for b_n allows the prediction interval to be computed even if $d \geq n$.

Definition 2.24. Compute the shorth(c) of the residuals $= [r_{(s)}, r_{(s+c-1)}] = [\tilde{\xi}_{\delta_1}, \tilde{\xi}_{1-\delta_2}]$. Then a 100 $(1 - \delta)$% large sample PI for Y_f is

$$[\hat{Y}_f + b_n\tilde{\xi}_{\delta_1}, \hat{Y}_f + b_n\tilde{\xi}_{1-\delta_2}]. \tag{2.46}$$

The second PI randomly divides the data into two half sets H and V where H has $n_H = \lceil n/2 \rceil$ of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. The estimator $\hat{m}_H(\boldsymbol{x}) = \hat{\boldsymbol{\beta}}_{IH}^T\boldsymbol{x}$ is computed using the training data set H. Then the validation residuals $v_j = Y_{i_j} - \hat{m}_H(\boldsymbol{x}_{i_j})$ are computed for the $j = 1, ..., n_V$ cases in the validation set V. Find the Frey PI $[v_{(s)}, v_{(s+c-1)}]$

of the validation residuals (replacing n in (2.42) by $n_V = n - n_H$). Let $\hat{Y}_{fH} = \hat{m}_H(\boldsymbol{x}_f) = \hat{\boldsymbol{\beta}}_{IH}^T \boldsymbol{x}_f$.

Definition 2.25. Then a $100(1-\delta)\%$ large sample PI for Y_f is

$$[\hat{Y}_{fH} + v_{(s)}, \hat{Y}_{fH} + v_{(s+c-1)}]. \tag{2.47}$$

Remark 2.19. Note that correction factors $b_n \to 1$ are used in large sample confidence intervals and tests if the limiting distribution is N(0,1) or χ^2_p, but a t_{d_n} or pF_{p,d_n} cutoff is used: $t_{d_n,1-\delta}/z_{1-\delta} \to 1$ and $pF_{p,d_n,1-\delta}/\chi^2_{p,1-\delta} \to 1$ if $d_n \to \infty$ as $n \to \infty$. Using correction factors for large sample confidence intervals, tests, prediction intervals, prediction regions, and bootstrap confidence regions improves the performance for moderate sample size n.

Remark 2.20. For a good fitting model, residuals r_i tend to be smaller in magnitude than the errors e_i, while validation residuals v_i tend to be larger in magnitude than the e_i. Thus the Frey correction factor can be used for PI (2.47) while PI (2.46) needs a stronger correction factor.

A sufficient condition for (2.46) and (2.47) to be large sample PIs, is that the residuals need to be consistent estimators of the iid errors e_i and $\hat{\boldsymbol{\beta}}_I$ needs to be a consistent estimator $\boldsymbol{\beta}_I$ where $Y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}_I + e_i$ is a valid MLR model and the iid e_i depend on I. This regularity condition tends to roughly hold when $n >> p$, but the regularity condition is often much too strong if $p > n$.

Another regularity condition for PI (2.47) is that the cases are iid. This assumption is strong but sometimes holds. Then we can motivate PI (2.47) by modifying the justification for the Lei et al. (2018) split conformal prediction interval

$$[\hat{m}_H(\boldsymbol{x}_f) - a_q, \hat{m}_H(\boldsymbol{x}_f) + a_q] \tag{2.48}$$

where a_q is the $100(1-\delta)$th quantile of the absolute validation residuals. PI (2.47) is a modification of the split conformal PI that is asymptotically optimal. Suppose $(Y_i, \boldsymbol{x}_i)$ are iid for $i = 1, ..., n, n+1$ where $(Y_f, \boldsymbol{x}_f) = (Y_{n+1}, \boldsymbol{x}_{n+1})$. Compute $\hat{m}_H(\boldsymbol{x})$ from the cases in H. For example, get $\hat{\boldsymbol{\beta}}_H$ from the cases in H. Consider the validation residuals v_i for $i = 1, ..., n_V$ and the validation residual v_{n_V+1} for case $(Y_f, \boldsymbol{x}_f)$. Since these $n_V + 1$ cases are iid, the probability that v_t has rank j for $j = 1, ..., n_V + 1$ is $1/(n_V + 1)$ for each t, i.e., the ranks follow the discrete uniform distribution. Let $t = n_V + 1$ and let the $v_{(j)}$ be the ordered residuals using $j = 1, ..., n_V$. That is, get the order statistics without using the unknown validation residual v_{n_V+1}. Then $v_{(i)}$ has rank i if $v_{(i)} < v_{n_V+1}$ but rank $i+1$ if $v_{(i)} > v_{n_V+1}$. Thus

$$P(Y_f \in [\hat{m}_H(\boldsymbol{x}_f) + v_{(k)}, \hat{m}_H(\boldsymbol{x}_f) + v_{(k+b-1)}]) = P(v_{(k)} \le v_{n_V+1} \le v_{(k+b-1)}) \ge$$

$P(v_{n_V+1}$ has rank between $k+1$ and $k+b-1$ and there are no tied ranks) $\ge (b-1)/(n_V+1) \approx 1-\delta$ if $b = \lceil (n_V+1)(1-\delta) \rceil + 1$ and $k + b - 1 \le n_V$.

This probability statement holds for a fixed k such as $k = \lceil n_V \ \delta/2 \rceil$. The statement is not true when the shorth(b) estimator is used since the shortest interval using $k = s$ can have s change with the data set. That is, s is not fixed. Hence if PI's were made from J independent data sets, the PI's with fixed k would contain Y_f about $J(1-\delta)$ times, but this value would be smaller for the shorth(b) prediction intervals where s can change with the data set. The above argument works if the estimator $\hat{m}(\boldsymbol{x})$ is "symmetric in the data," which is satisfied for multiple linear regression estimators.

Prediction intervals (2.46), (2.47), and (2.48) can be used to compare different MLR methods such as PLS and lasso variable selection. In the simulations, none of these three prediction intervals dominates the other two. Recall that $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector in (2.14). If a good fitting method, such as lasso or forward selection with EBIC, is used, and $1.5a_S \leq n \leq 5a_S$, then PI (2.46) can be much shorter than PIs (2.47) and (2.48). For n/d large, PIs (2.46) and (2.47) can be shorter than PI (2.48) if the error distribution is not unimodal and symmetric; however, PI (2.48) is often shorter if n/d is not large since the sample shorth converges to the population shorth rather slowly. Grübel (1982) shows that for iid data, the length and center the shorth(k_n) interval are $\sqrt{n}$ consistent and $n^{1/3}$ consistent estimators of the length and center of the population shorth interval. For a unimodal and symmetric error distribution, the three PIs are asymptotically equivalent (with p fixed and $n \to \infty$), but PI (2.48) can be the shortest PI due to different correction factors.

If the estimator is poor, the split conformal PI (2.48) and PI (2.47) can have coverage closer to the nominal coverage than PI (2.46). For example, if $\hat{m}$ interpolates the data and $\hat{m}_H$ interpolates the training data from H, then the validation residuals will be huge. Hence PI (2.48) will be long compared to PI (2.46).

Asymptotically optimal PIs estimate the population shorth of the zero mean error distribution. Hence PIs that use the shorth of the residuals, such as PIs (2.46) and (2.47), may be the only easily computed asymptotically optimal PIs for a wide range of consistent estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ for the multiple linear regression model. If the error distribution is $e \sim EXP(1) - 1$, then the asymptotic length of the 95% PI (2.46) or (2.47) is 2.966 while that of the split conformal PI is $2(1.966) = 3.992$. For more about these PIs applied to MLR models, Pelawa Watagoda and Olive (2021b).

For the simulation from Pelawa Watagoda and Olive (2021b), we used several R functions including forward selection (FS) as computed with the `regsubsets` function from the `leaps` library, (model selection) principal components regression (PCR) with the `pcr` function and (model selection) partial least squares (PLS) with the `plsr` function from the `pls` library, and ridge regression (RR, see Definition 2.16 c)) and lasso with the `cv.glmnet` function from the `glmnet` library. Lasso variable selection (LVS) was applied to the selected lasso model.

Let $\boldsymbol{x} = (1 \ \boldsymbol{u}^T)^T$ where $\boldsymbol{u}$ is the $(p-1) \times 1$ vector of nontrivial predictors. In the simulations, for $i = 1, ..., n$, we generated $\boldsymbol{w}_i \sim N_{p-1}(\boldsymbol{0}, \boldsymbol{I})$ where the

Table 2.2 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

n	p	ψ	k		FS	lasso	LVS	RR	PLS	PCR
100	20	0	1	cov	0.9644	0.9750	0.9666	0.9560	0.9438	0.9772
				len	4.4490	4.8245	4.6873	4.5723	4.4149	5.5647
100	40	0	1	cov	0.9654	0.9774	0.9588	0.9274	0.8810	0.9882
				len	4.4294	4.8889	4.6226	4.4291	4.0202	7.3393
100	200	0	1	cov	0.9648	0.9764	0.9268	0.9584	0.6616	0.9922
				len	4.4268	4.9762	4.2748	6.1612	2.7695	12.412
100	50	0	49	cov	0.8996	0.9719	0.9736	0.9820	0.8448	1.0000
				len	22.067	6.8345	6.8092	7.7234	4.2141	38.904
200	20	0	19	cov	0.9788	0.9766	0.9788	0.9792	0.9550	0.9786
				len	4.9613	4.9636	4.9613	5.0458	4.3211	4.9610
200	40	0	19	cov	0.9742	0.9762	0.9740	0.9738	0.9324	0.9792
				len	4.9285	5.2205	5.1146	5.2103	4.2152	5.3616
200	200	0	19	cov	0.9728	0.9778	0.9098	0.9956	0.3500	1.0000
				len	4.8835	5.7714	4.5465	22.351	2.1451	51.896
400	20	0.9	19	cov	0.9664	0.9748	0.9604	0.9726	0.9554	0.9536
				len	4.5121	10.609	4.5619	10.663	4.0017	3.9771
400	40	0.9	19	cov	0.9674	0.9608	0.9518	0.9578	0.9482	0.9646
				len	4.5682	14.670	4.8656	14.481	4.0070	4.3797
400	400	0.9	19	cov	0.9348	0.9636	0.9556	0.9632	0.9462	0.9478
				len	4.3687	47.361	4.8530	48.021	4.2914	4.4764
400	400	0	399	cov	0.9486	0.8508	0.5704	1.0000	0.0948	1.0000
				len	78.411	37.541	20.408	244.28	1.1749	305.93
400	800	0.9	19	cov	0.9268	0.9652	0.9542	0.9672	0.9438	0.9554
				len	4.3427	67.294	4.7803	66.577	4.2965	4.6533

$m = p - 1$ elements of the vector $\boldsymbol{w}_i$ are iid N(0,1). Let the $m \times m$ matrix $\boldsymbol{A} = (a_{ij})$ with $a_{ii} = 1$ and $a_{ij} = \psi$ where $0 \leq \psi < 1$ for $i \neq j$. Then the vector $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\mathrm{Cov}(\boldsymbol{u}_i) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1+(m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi+(m-2)\psi^2]$. Hence the correlations are $cor(x_i, x_j) = \rho = (2\psi+(m-2)\psi^2)/(1+(m-1)\psi^2)$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let $Y_i = 1+1x_{i,2}+ \cdots + 1x_{i,k+1} + e_i$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (1, .., 1, 0, ..., 0)^T$ with $k+1$ ones and $p-k-1$ zeros. The zero mean errors e_i were iid from five distributions: i) N(0,1), ii) t_3, iii) EXP(1) - 1, iv) uniform$(-1, 1)$, and v) 0.9 N(0,1) + 0.1 N(0,100). Normal distributions usually appear in simulations, and the uniform distribution is the distribution where the shorth undercoverage is maximized by Frey (2013). Distributions ii) and v) have heavy tails, and distribution iii) is not symmetric.

The population shorth 95% PI lengths estimated by the asymptotically optimal 95% PIs are i) $3.92 = 2(1.96)$, ii) 6.365, iii) 2.996, iv) $1.90 = 2(0.95)$, and v) 13.490. The split conformal PI (2.48) is not asymptotically optimal for iii), and for iii) PI (2.48) has asymptotic length $2(1.966) = 3.992$. The simulation used 5000 runs, so an observed coverage in [0.94, 0.96] gives no

reason to doubt that the PI has the nominal coverage of 0.95. The simulation used $p = 20, 40, 50, n$, or $2n$; $\psi = 0, 1/\sqrt{p}$, or 0.9; and $k = 1, 19$, or $p-1$. The OLS full model fails when $p = n$ and $p = 2n$, where regularity conditions for consistent estimators are strong. The values $k = 1$ and $k = 19$ are sparse models where lasso, lasso variable selection, and forward selection with EBIC can perform well when n/p is not large. If $k = p-1$ and $p \geq n$, then the model is dense. When $\psi = 0$, the predictors are uncorrelated, when $\psi = 1/\sqrt{p}$, the correlation goes to 0.5 as p increases and the predictors are moderately correlated. For $\psi = 0.9$, the predictors are highly correlated with 1 dominant principal component, a setting favorable for PLS and PCR. The simulated data sets are rather small since the some of the R estimators are rather slow.

The simulations were done in *R*. See R Core Team (2020). The results were similar for all five error distributions, and we show some results for the normal and shifted exponential distributions. Tables 2.2 and 2.3 show some simulation results for PI (2.46) where forward selection used C_p for $n \geq 10p$ and EBIC for $n < 10p$. The other methods minimized 10-fold CV. For forward selection, the maximum number of variables used was approximately $\min(\lceil n/5 \rceil, p)$. Ridge regression used the same d that was used for lasso.

For $n \geq 5p$, coverages tended to be near or higher than the nominal value of 0.95. The average PI length was often near 1.3 times the asymptotically optimal length for $n = 10p$ and close to the optimal length for $n = 100p$. C_p and EBIC produced good PIs for forward selection, and 10-fold CV produced good PIs for PCR and PLS. For lasso and ridge regression, 10-fold CV produced good PIs if $\psi = 0$ or if k was small, but if both $k \geq 19$ and $\psi \geq 0.5$, then 10-fold CV tended to shrink too much and the PI lengths were often too long. Lasso variable selection was good for $n/p \geq 5$. (For MLR, the lasso estimator $\hat{\boldsymbol{\beta}}_{I,0}$ is a consistent estimator of $\boldsymbol{\beta}$ if p is fixed, $\hat{\lambda}_{1,n}/n \to 0$, and $n \to \infty$, which requires $P(S \subseteq I) \to 1$ as $n \to \infty$.)

For n/p not large, good performance needed stronger regularity conditions, and all six methods can have problems. PLS tended to have severe undercoverage with small average length, but sometimes performed well for $\psi = 0.9$. The PCR length was often too long for $\psi = 0$. If there was $k = 1$ active population predictor, then forward selection with EBIC, lasso, and lasso variable selection often performed well. For $k = 19$, forward selection with EBIC often performed well, as did lasso and lasso variable selection for $\psi = 0$. (Good performance can occur if $\hat{\boldsymbol{\beta}}_I$ is a good estimator of $\boldsymbol{\beta}_I$ and $Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + e$ where the errors e depend on I.) For dense models with $k = p - 1$ and n/p not large, there was often undercoverage. Here forward selection would use about $n/5$ variables. Let $d - 1$ be the number of active nontrivial predictors in the selected model. For $N(0, 1)$ errors, $\psi = 0$, and $d < k$, an asymptotic population 95% PI has length $3.92\sqrt{k - d + 1}$. Note that when the $(Y_i, \boldsymbol{u}_i^T)^T$ follow a multivariate normal distribution, every subset follows a multiple linear regression model. EBIC occasionally had undercoverage, especially for $k = 19$ or $p - 1$, which was usually more severe for $\psi = 0.9$ or $1/\sqrt{p}$.

Table 2.3 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1)-1$

n	p	ψ	k		FS	lasso	LVS	RR	PLS	PCR
100	20	0	1	cov	0.9622	0.9728	0.9648	0.9544	0.9460	0.9724
				len	3.7909	4.4344	4.3865	4.4375	4.2818	5.5065
2000	20	0	1	cov	0.9506	0.9502	0.9500	0.9488	0.9486	0.9542
				len	3.1631	3.1199	3.1444	3.2380	3.1960	3.3220
200	20	0.9	1	cov	0.9588	0.9666	0.9664	0.9666	0.9556	0.9612
				len	3.7985	3.6785	3.7002	3.7491	3.5049	3.7844
200	20	0.9	19	cov	0.9704	0.9760	0.9706	0.9784	0.9578	0.9592
				len	4.6128	12.1188	4.8732	12.0363	3.3929	3.7374
200	200	0.9	19	cov	0.9338	0.9750	0.9564	0.9740	0.9440	0.9596
				len	4.6271	37.3888	5.1167	56.2609	4.0550	4.6994
400	40	0.9	19	cov	0.9678	0.9654	0.9492	0.9624	0.9426	0.9574
				len	4.3433	14.7390	4.7625	14.6602	3.6229	4.1045

Table 2.4 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

n,p,ψ,k		FS	CFS	LVS	CLVS	Lasso	CL	RR	CRR
200,20, 0,19	cov	0.9574	0.9446	0.9522	0.9420	0.9538	0.9382	0.9542	0.9430
	len	4.6519	4.3003	4.6375	4.2888	4.6547	4.2964	4.7215	4.3569
200,40,0,19	cov	0.9564	0.9412	0.9524	0.9440	0.9550	0.9406	0.9548	0.9404
	len	4.9188	4.5426	5.2665	4.8637	5.1073	4.7193	5.3481	4.9348
200,200, 0,19	cov	0.9488	0.9320	0.9548	0.9392	0.9480	0.9380	0.9536	0.9394
	len	7.0096	6.4739	5.1671	4.7698	31.1417	28.7921	47.9315	44.3321
400,20,0.9,19	cov	0.9498	0.9406	0.9488	0.9438	0.9524	0.9426	0.9550	0.9426
	len	4.4153	4.1981	4.5849	4.3591	9.4405	8.9728	9.2546	8.8054
400,40,0.9,19	cov	0.9504	0.9404	0.9476	0.9388	0.9496	0.9400	0.9470	0.9410
	len	4.7796	4.5423	4.9704	4.7292	13.3756	12.7209	12.9560	12.3118
400,400,0.9,19	cov	0.9480	0.9398	0.9554	0.9444	0.9506	0.9422	0.9506	0.9408
	len	5.2736	5.0131	4.9764	4.7296	43.5032	41.3620	42.6686	40.5578
400,800,0.9,19	cov	0.9550	0.9474	0.9522	0.9412	0.9550	0.9450	0.9550	0.9446
	len	5.3626	5.0943	4.9382	4.6904	60.9247	57.8783	60.3589	57.3323

Tables 2.4 and 2.5 show some results for PIs (2.47) and (2.48). Here forward selection using the minimum C_p model if $n_H > 10p$ and EBIC otherwise. The coverage was very good. Labels such as CFS and CLVS used PI (2.48). For lasso variable selection, the program sometimes failed to run for 5000 runs, e.g., if the number of variables selected $d = n_H$. In Table 2.4, PIs (2.47) and (2.48) are asymptotically equivalent if p is fixed, but PI (2.48) had shorter lengths for moderate n. In Table 2.5, PI (2.47) is shorter than PI (2.48) asymptotically, but for moderate n, PI (2.48) was often shorter.

Table 2.6 shows some results for PIs (2.46) and (2.47) for lasso and ridge regression. The header lasso indicates PI (2.46) was used while vlasso indicates that PI (2.47) was used. PI (2.47) tended to work better when the fit

Table 2.5 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1) - 1$

n,p,ψ,k		FS	CFS	LVS	CLVS	Lasso	CL	RR	CRR
200,20,0,1	cov	0.9596	0.9504	0.9588	0.9374	0.9604	0.9432	0.9574	0.9438
	len	4.6055	4.2617	4.5984	4.2302	4.5899	4.2301	4.6807	4.2863
2000,20,0,1	cov	0.9560	0.9508	0.9530	0.9464	0.9544	0.9462	0.9530	0.9462
	len	3.3469	3.9899	3.3240	3.9849	3.2709	3.9786	3.4307	3.9943
200,20,0.9,1	cov	0.9564	0.9402	0.9584	0.9362	0.9634	0.9412	0.9638	0.9418
	len	3.9184	3.8957	3.8765	3.8660	3.8406	3.8483	3.8467	3.8509
200,20,0.9,19	cov	0.9630	0.9448	0.9510	0.9368	0.9554	0.9430	0.9572	0.9420
	len	5.0543	4.6022	4.8139	4.3841	9.8640	9.0748	9.5218	8.7366
200,200,0.9,19	cov	0.9570	0.9434	0.9588	0.9418	0.9552	0.9392	0.9544	0.9394
	len	5.8095	5.2561	5.2366	4.7292	31.1920	28.8602	47.9229	44.3251
400,40,0.9,19	cov	0.9476	0.9402	0.9494	0.9416	0.9584	0.9496	0.9562	0.9466
	len	4.6992	4.4750	4.9314	4.6703	13.4070	12.7442	13.0579	12.4015

was poor while PI (2.46) was better for $n = 2p$ and $k = p - 1$. The PIs are asymptotically equivalent for consistent estimators.

Table 2.6 PIs (2.46) and (2.47): Simulated Large Sample 95% PI Coverages and Lengths

n	p	ψ	k		dist	lasso	vlasso	RR	vRR
100	20	0	1	cov	N(0,1)	0.9750	0.9632	0.9564	0.9606
				len		4.8245	4.7831	4.5741	5.3277
100	20	0	1	cov	EXP(1)−1	0.9728	0.9582	0.9546	0.9612
				len		4.4345	5.0089	4.4384	5.6692
100	50	0	49	cov	N(0,1)	0.9714	0.9606	0.9822	0.9618
				len		6.8345	22.3265	7.7229	27.7275
100	50	0	49	cov	EXP(1)−1	0.9716	0.9618	0.9814	0.9608
				len		6.9460	22.4097	7.8316	27.8306
400	400	0	399	cov	N(0,1)	0.8508	0.9518	1.0000	0.9548
				len		37.5418	78.0652	244.1004	69.5812
400	400	0	399	cov	EXP(1)−1	0.8446	0.9586	1.0000	0.9558
				len		37.5185	78.0564	243.7929	69.5474

2.14 Cross Validation

For MLR variable selection there are many methods for choosing the final submodel, including AIC, BIC, C_p, and EBIC. Variable selection is a special

case of model selection where there are M models and a final model needs to be chosen. Cross validation is a common criterion for model selection.

Definition 2.26. For *k-fold cross validation* (k-fold CV), randomly divide the training data into k groups or folds of approximately equal size $n_j \approx n/k$ for $j = 1, ..., k$. Leave out the first fold, fit the statistical method to the $k-1$ remaining folds, and then compute some criterion for the first fold. Repeat for folds 2, ..., k.

Following James et al. (2013, p. 181), if the statistical method is an MLR method, we often compute $\hat{Y}_i(j)$ for each Y_i in the fold j left out. Then

$$MSE_j = \frac{1}{n_j} \sum_{i=1}^{n_j} (Y_i - \hat{Y}_i(j))^2,$$

and the overall criterion is

$$CV_{(k)} = \frac{1}{k} \sum_{j=1}^{k} MSE_j.$$

Note that if each $n_j = n/k$, then

$$CV_{(k)} = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i(j))^2.$$

Then $CV_{(k)} \equiv CV_{(k)}(I_i)$ is computed for $i = 1, ..., M$, and the model I_c with the smallest $CV_{(k)}(I_i)$ is selected.

Assume that model (2.1) holds: $\boldsymbol{Y} = \boldsymbol{x}^T\boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{e}$ where $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector. Suppose p is fixed and $n \to \infty$. If $\hat{\boldsymbol{\beta}}_I$ is $a \times 1$, form the $p \times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then Section 2.17 shows that $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ under mild regularity conditions. Note that if $a_S = p$, then $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is asymptotically equivalent to the OLS full model $\hat{\boldsymbol{\beta}}$ (since S is equal to the full model).

Choosing folds for k-fold cross validation is similar to randomly allocating cases to treatment groups. The following code is useful for a simulation. It makes copies of 1 to k in a vector of length n called *tfolds*. The sample command makes a permutation of tfolds to get the *folds*. The lengths of the k folds differ by at most 1.

```
n<-26
k<-5
J<-as.integer(n/k)+1
tfolds<-rep(1:k,J)
tfolds<-tfolds[1:n] #can pass tfolds to a loop
```

```
folds<-sample(tfolds)
folds
4 2 3 5 3 3 1 5 2 2 5 1 2 1 3 4 2 1 5 5 1 4 1 4 4 3
```

Example 2.2, continued. The *slpack* function `pifold` uses k-fold CV to get the coverage and average PI lengths. We used 5-fold CV with coverage and average 95% PI length to compare the forward selection models. All 4 models had coverage 1, but the average 95% PI lengths were 2591.243, 2741.154, 2902.628, and 2972.963 for the models with 2 to 5 predictors. See the following R code.

```
y <- marry[,3]; x <- marry[,-3]
x1 <- x[,2]
x2 <- x[,c(2,3)]
x3 <- x[,c(1,2,3)]
pifold(x1,y) #nominal 95% PI
$cov
[1] 1
$alen
[1] 2591.243
pifold(x2,y)
$cov
[1] 1
$alen
[1] 2741.154
pifold(x3,y)
$cov
[1] 1
$alen
[1] 2902.628
pifold(x,y)
$cov
[1] 1
$alen
[1] 2972.963
#Validation PIs for submodels: the sample size is
#likely too small and the validation PI is formed
#from the validation set.
n<-dim(x)[1]
nH <- ceiling(n/2)
indx<-1:n
perm <- sample(indx,n)
H <- perm[1:nH]
vpilen(x1,y,H) #13/13 were in the validation PI
$cov
[1] 1.0
```

```
$len
[1] 116675.4
vpilen(x2,y,H)
$cov
[1] 1.0
$len
[1] 116679.8
vpilen(x3,y,H)
$cov
[1] 1.0
$len
[1] 116312.5
vpilen(x,y,H)
$cov
[1] 1.0
$len  #shortest length
[1] 116270.7
```

Some more code is below.

```
n <- 100
p <- 4
k <- 1
q <- p-1
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
b <- 0 * 1:q
b[1:k] <- 1
y <- 1 + x %*% b + rnorm(n)
x1 <- x[,1]
x2 <- x[,c(1,2)]
x3 <- x[,c(1,2,3)]
pifold(x1,y)
$cov
[1] 0.96
$alen
[1] 4.2884
pifold(x2,y)
$cov
[1] 0.98
$alen
[1] 4.625284
pifold(x3,y)
$cov
[1] 0.98
$alen
[1] 4.783187
```

```
pifold(x,y)
$cov
[1] 0.98
$alen
[1] 4.713151

n <- 10000
p <- 4
k <- 1
q <- p-1
x <- matrix(rnorm(n * q), nrow = n, ncol = q)
b <- 0 * 1:q
b[1:k] <- 1
y <- 1 + x %*% b + rnorm(n)
x1 <- x[,1]
x2 <- x[,c(1,2)]
x3 <- x[,c(1,2,3)]
pifold(x1,y)
$cov
[1] 0.9491
$alen
[1] 3.96021
pifold(x2,y)
$cov
[1] 0.9501
$alen
[1] 3.962338
pifold(x3,y)
$cov
[1] 0.9492
$alen
[1] 3.963305
pifold(x,y)
$cov
[1] 0.9498
$alen
[1] 3.96203
```

2.15 Data Splitting

Remark 2.21. a) When $p > n$, the fitted model should do better than i) interpolating the data or ii) discarding all of the predictors and using the location model of Section 1.4.1 for inference. If $p > n$, forward selection, lasso,

lasso variable selection, elastic net, and elastic net variable selection can be useful for several regression models. Ridge regression, partial least squares, and principal components regression can also be computed for multiple linear regression. Section 2.13 gives prediction intervals.

b) One of the **biggest errors in regression** is to use the response variable to build the regression model using all n cases, and then do inference as if the built model was selected without using the response, e.g., selected before gathering data. Using the response variable to build the model is called *data snooping*, then inference is generally no longer valid, and the model built from data snooping tends to fit the data too well. In particular, do not use data snooping and then use variable selection or cross validation. See Hastie et al (2009, p. 245) and Olive (2017a, pp. 85-89).

c) Building a regression model from data is one of the most challenging regression problems. The "final full model" will have response variable $Y = t(Z)$, a constant x_1, and predictor variables $x_2 = t_2(w_2, ..., w_r), ..., x_p = t_p(w_2, ..., w_r)$ where the initial data consists of $Z, w_2, ..., w_r$. Choosing $t, t_2, ..., t_p$ so that the final full model is a useful regression approximation to the data can be difficult.

d) As a rule of thumb, if strong nonlinearities are apparent in the predictors $w_2, ..., w_p$, it is often useful to remove the nonlinearities by transforming the predictors using power transformations. When p is large, a scatterplot matrix of $w_2, ..., w_p$ can not be made, but the log rule of Section 1.2 can be useful. Plots from Chapter 1, such as the DD plot, can also be useful. A scatterplot matrix of the w_i is an array of scatterplots of w_i versus w_j. A scatterplot is a plot of w_i versus w_j.

Data splitting divides the data into two parts. The first part can use the response variable to build the model, then the second part can be used for inference. This avoids the Remark 2.21 b) error since the model is not built using all n cases.

A common method for data splitting randomly divides the data set into two half sets: the training set H and the validation set V. For the data in H, fit the model selection method, e.g. forward selection or lasso, to get model I with a predictors. Use this model as the full model for the set V: use the standard OLS inference from regressing the response on the predictors found from the set H. This method can be inefficient if $n \geq 10p$, but is useful for a sparse model if $n \leq 5p$, if the probability that the model underfits goes to zero, and if $n \geq 20a$. A model is sparse if the number of predictors with nonzero coefficients is small.

For lasso, the active set I of a predictors from the data in training set H is found, and data splitting estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{I,D}$ computed from the validation data in set V. This estimator is not the lasso variable selection estimator. The estimator $\hat{\boldsymbol{\beta}}_{I,D}$ has the same large sample theory as if I was chosen before obtaining the data.

If n/p is not large, data splitting is useful for many regression models when the n cases are independent, including multiple linear regression, multivariate linear regression where there are $m \geq 2$ response variables, generalized linear models (GLMs), the Cox (1972) proportional hazards regression model, and parametric survival regression models.

Consider a regression model with response variable Y and a $p \times 1$ vector of predictors $\boldsymbol{x}$. This model is the full model. Suppose the n cases are independent. To perform data splitting, randomly divide the data into two sets H and V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. Find a model I, possibly with data snooping or model selection, using the data in the training set H. Use the model I as the full model to perform inference using the data in the validation set V. That is, regress Y_V on $\boldsymbol{X}_{V,I}$ and perform the usual inference for the model using the $j = 1, ..., n_V$ cases in the validation set V. If $\boldsymbol{\beta}_I$ uses a predictors, we want $n_V \geq 10a$ and we want $(Y_V, \boldsymbol{X}_{V,I})$ to follow a regression model, e.g. $Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + e$ where e depends on I.

In the literature, often $n_H \approx \lceil n/2 \rceil$. For model selection, use the training set data to fit the model selection method, e.g. forward selection or lasso, to get the a predictors. On the validation set, use the standard regression inference from regressing the response on the predictors found from the training set data. This method can be inefficient if $n \geq 10p$, but is useful for a sparse model if $n \leq 5p$, if the probability that the model underfits goes to zero, and if $n \geq 20a$.

The method is simple, use one half set to get the predictors, then fit the regression model, such as a GLM or OLS, to the validation half set $(\boldsymbol{Y}_V, \boldsymbol{X}_{V,I})$. The regression model needs to hold for $(\boldsymbol{Y}_V, \boldsymbol{X}_{V,I})$ and we want $n_V \geq 10a$ if I uses a predictors. The regression model can hold if $S \subseteq I$ and the model is sparse. Let $\boldsymbol{x} = (\boldsymbol{x}_1, ..., \boldsymbol{x}_p)^T$ where $\boldsymbol{x}_1$ is a constant. If $(Y, \boldsymbol{x}_2, ..., \boldsymbol{x}_p)^T$ follows a multivariate normal distribution, then $(Y, \boldsymbol{x}_I)$ follows a multiple linear regression model for every I. Hence the full model need not be sparse, although the selected model may be suboptimal.

Of course other sample sizes than half sets could be used. For example if $n = 1000p$, use $n = 10p$ for the training set and $n = 990p$ for the validation set.

Remark 2.22. i) One use of data splitting is to try to transform the $p \geq n$ problem into an $n \geq 10k$ problem. Thus this method needs the fitted model I to be sparse. For MLR, check that $Y = \boldsymbol{x}_I^T \boldsymbol{\beta}_I + e_I$ with response and residual plots. If $\boldsymbol{\beta}_I$ is $k \times 1$, we want $n \geq 10k$ and $V(e_{I,i}) = \sigma_I^2$ to be small. Note that data splitting does not need a sparse population model with $S \subseteq I$ and $a_S \leq k$. For multiple linear regression, data splitting can work if $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, since then all subsets I satisfy an MLR model: $Y_i = \boldsymbol{x}_{I,i}^T \boldsymbol{\beta}_I + \boldsymbol{e}_{I,i}$. See Section 2.16. The above multivariate normal assumption for MLR rarely hold, but if several predictors satisfy a simple linear regression model with Y, then those predictors often satisfy an MLR with Y.

ii) Data splitting can be tricky for lasso, ridge regression, and elastic net if the sample sizes of the training and validation sets differ. Roughly set $\lambda_{1,n_1}/(2n_1) = \lambda_{2,n_2}/(2n_2)$. Data splitting is much easier for variable selection methods such as forward selection, lasso variable selection, and elastic net variable selection. Find the variables $x_1^*, ..., x_k^*$ indexed by I from the training set, and use model I as the full model for the validation set.

iii) Another use of data splitting is that data snooping can be used on the training set H: use the model I found from H as the full model for the validation set V.

2.16 The Multitude of MLR Models

There are often a multitude of population regression models that are estimating different population parameters. Note that when j predictors each satisfy a marginal regression model with the response Y (such as simple linear regression), then subsets of those j predictors will often satisfy a regression model with the response Y (such as multiple linear regression).

This chapter showed that OPLS and OLS typically estimate different quantities. There are often a multitude of valid MLR models. For example, if the cases $(Y_i \ \boldsymbol{x}_i^T)^T$ are iid from a nonsingular multivariate normal distribution, then $Y|\boldsymbol{\eta}^T\boldsymbol{x}$ satisfies a MLR model for any linear combination $\boldsymbol{\eta}^T\boldsymbol{x}$. See Olive and Zhang (2023). Under multivariate normality, it is known that $Y|\boldsymbol{x}_I$ follows a multiple linear regression model where $\boldsymbol{x}_I = (x_{i1}, ..., x_{ik})^T$ is a vector corresponding to a subset of the predictors. Theorem 2.15 b) gives a similar result for every linear combination of the predictors $\boldsymbol{\eta}^T\boldsymbol{x}$, including sparse and nonsparse models. Much of Theorem 2.15 b) can also be shown by performing the population SLR of Y on $\boldsymbol{\eta}^T\boldsymbol{x}$, but linearity may fail to hold if multivariate normality does not hold. Note that data sets where the cases are iid from a multivariate normal distribution are rather uncommon. Let $\Sigma_Y = \sigma_Y^2$.

Theorem 2.15. Suppose the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from a multivariate normal distribution:

$$\begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} \sim N_{p+1}\left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\mu_x} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{Y\boldsymbol{x}} \\ \boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\Sigma_x} \end{pmatrix} \right).$$

a) Then $Y|\boldsymbol{x} \sim Y|(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x}) \sim N(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x}, \sigma^2)$ follows a multiple linear regression model.

b) So does $Y|\boldsymbol{\eta}^T\boldsymbol{x} \sim N(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}, \sigma_O^2)$ where $\alpha_O = \mu_Y - \boldsymbol{\beta}_O^T\boldsymbol{\mu_x}$, $\boldsymbol{\beta}_O = \lambda\boldsymbol{\eta}$, $\sigma_O^2 = \Sigma_Y - \boldsymbol{\beta}_O^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, and

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\eta}}{\boldsymbol{\eta}^T\boldsymbol{\Sigma_x}\boldsymbol{\eta}}.$$

c) So does $Y|\boldsymbol{A}\boldsymbol{x}$ where $\boldsymbol{A}$ is a full rank $k \times p$ constant matrix with $k \leq p$.

Proof. a) is a special case of c) with $\boldsymbol{A} = \boldsymbol{I}_p$, and see Remark 1.5.
b)

$$\begin{pmatrix} 1 & \boldsymbol{0}^T \\ 0 & \boldsymbol{\eta}^T \end{pmatrix} \begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} Y \\ \boldsymbol{\eta}^T \boldsymbol{x} \end{pmatrix}$$

$$\sim N_2 \left(\begin{pmatrix} \mu_Y \\ \boldsymbol{\eta}^T \boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta} \\ \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta} \end{pmatrix} \right).$$

Hence $W = Y | \boldsymbol{\eta}^T \boldsymbol{x} \sim N(\mu_W, \sigma_W^2)$ where

$$\mu_W = \mu_Y + \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta}}{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta}} (\boldsymbol{\eta}^T \boldsymbol{x} - \boldsymbol{\eta}^T \boldsymbol{\mu}_{\boldsymbol{x}}) = \mu_Y - \lambda \boldsymbol{\eta}^T \boldsymbol{\mu}_{\boldsymbol{x}} + \lambda \boldsymbol{\eta}^T \boldsymbol{x},$$

and

$$\sigma_W^2 = \sigma_O^2 = \sigma_Y^2 - \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta} \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta}} = \sigma_Y^2 - \frac{(\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\eta})^2}{\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\eta}} = \sigma_Y^2 - \lambda \boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}.$$

c)

$$\begin{pmatrix} 1 & \boldsymbol{0}^T \\ \boldsymbol{0} & \boldsymbol{A} \end{pmatrix} \begin{pmatrix} Y \\ \boldsymbol{x} \end{pmatrix} = \begin{pmatrix} Y \\ \boldsymbol{A}\boldsymbol{x} \end{pmatrix}$$

$$\sim N_{q+1} \left(\begin{pmatrix} \mu_Y \\ \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{x}} \end{pmatrix}, \begin{pmatrix} \Sigma_Y & \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{A}^T \\ \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} & \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{A}^T \end{pmatrix} \right).$$

Let $\boldsymbol{w} = \boldsymbol{A}\boldsymbol{x}$. Then $E(Y|\boldsymbol{w}) = \mu_Y + \boldsymbol{\Sigma}_{Y\boldsymbol{w}} \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} (\boldsymbol{w} - \boldsymbol{\mu}_{\boldsymbol{w}})$
$= \mu_Y - \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{\mu}_{\boldsymbol{w}} + \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{w} = \alpha_{OLS}(\boldsymbol{w}, Y) + \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{A}\boldsymbol{x}$
where $(\boldsymbol{w}, Y)$ indicates a population OLS regression of Y on $\boldsymbol{w}$. Thus

$$\boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{\Sigma}_{Y\boldsymbol{w}}^T = \boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{w}Y} = (\boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{A}^T)^{-1} \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{x}Y},$$

and

$$\alpha_{OLS}(\boldsymbol{w}Y) = \boldsymbol{\mu}_Y - \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{\mu}_{\boldsymbol{w}} = \boldsymbol{\mu}_Y - \boldsymbol{\beta}_{OLS}(\boldsymbol{w}, Y)^T \boldsymbol{A}\boldsymbol{\mu}_{\boldsymbol{x}}.$$

□

Note that $\sigma_O^2 < \sigma_Y^2 = \Sigma_Y$ unless $\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = 0$. If $\boldsymbol{\eta} = \boldsymbol{\beta}_{OLS}$, then $\lambda = 1$ and $\sigma_O^2 = \sigma_Y^2 - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. The population quantity estimated by the one component partial least squares estimator corresponds to $\boldsymbol{\eta} = \mathrm{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Note that b) is a special case of c) with $\boldsymbol{A} = \boldsymbol{\eta}^T$.

Since the Weibull regression model is a proportional hazards regression model for Y and a multiple linear regression model for $\log(Y)$, there can be many linear combinations that result in a proportional hazards model. For Poisson regression, $\log(Y + 1)$ often has a weighted least squares relationship with the predictors used for minimum chi-square estimators. See Agresti (2002, pp. 611-612) and Olive (2013). Hence often many linear combinations will result in a Poisson regression model.

2.17 Variable Selection Theory

From Section 1.1, a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{2.49}$$

where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p-a_S)\times 1$ vector. Given that $\boldsymbol{x}_S$ is in the model, $\boldsymbol{\beta}_E = \boldsymbol{0}$ and E denotes the subset of terms that can be eliminated given that the subset S is in the model. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). Suppose that S is a subset of I and that model (2.49) holds. Then

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_I^T\boldsymbol{\beta}_I + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I.$$

Thus $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. The model using $\boldsymbol{x}^T\boldsymbol{\beta}$ is the *full model*. The full model uses all of the predictors with $\boldsymbol{\beta}_F = \boldsymbol{\beta}$.

For multiple linear regression, if the candidate model of $\boldsymbol{x}_I$ has k terms (including the constant), then the partial F statistic for testing whether the $p-k$ predictor variables in $\boldsymbol{x}_O$ can be deleted is

$$F_I = \frac{SSE(I) - SSE}{(n-k)-(n-p)} / \frac{SSE}{n-p} = \frac{n-p}{p-k}\left[\frac{SSE(I)}{SSE} - 1\right]$$

where SSE is the error sum of squares from the full model, and SSE(I) is the error sum of squares from the candidate submodel. An important criterion for variable selection is the C_p criterion.

Definition 2.27.

$$C_p(I) = \frac{SSE(I)}{MSE} + 2k - n = (p-k)(F_I - 1) + k$$

where MSE is the error mean square for the full model.

Note that when $H_0 : \boldsymbol{\beta}_O = \boldsymbol{0}$ is true, $(p-k)(F_I-1)+k \xrightarrow{D} \chi^2_{p-k} + 2k - p$ for a large class of iid error distributions. Minimizing $C_p(I)$ is equivalent to minimizing $MSE\ [C_p(I)] = SSE(I) + (2k-n)MSE = \boldsymbol{r}^T(I)\boldsymbol{r}(I) + (2k-n)MSE$. The following theorem helps explain why C_p is a useful criterion and suggests that for subsets I with k terms, submodels with $C_p(I) \leq \min(2k, p)$ are especially interesting. Denote the residuals and fitted values from the *full model* by $r_i = Y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} = Y_i - \hat{Y}_i$ and $\hat{Y}_i = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$ respectively. Similarly, let $\hat{\boldsymbol{\beta}}_I$ be the estimate of $\boldsymbol{\beta}_I$ obtained from the regression of Y on $\boldsymbol{x}_I$ and denote the corresponding residuals and fitted values by $r_{I,i} = Y_i - \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ and $\hat{Y}_{I,i} = \boldsymbol{x}_{I,i}^T\hat{\boldsymbol{\beta}}_I$ where $i = 1, ..., n$.

Theorem 2.16. Suppose that a numerical variable selection method suggests several submodels with k predictors, including a constant, where $2 \leq k \leq p$.

a) The model I that minimizes $C_p(I)$ maximizes $\mathrm{corr}(r, r_I)$.

b) $C_p(I) \leq 2k$ implies that $\mathrm{corr}(\mathrm{r}, \mathrm{r_I}) \geq \sqrt{1 - \frac{\mathrm{p}}{\mathrm{n}}}$.

c) As $\mathrm{corr}(r, r_I) \to 1$,

$$\mathrm{corr}(\boldsymbol{x}^{\mathrm{T}}\hat{\boldsymbol{\beta}}, \boldsymbol{x}_{\mathrm{I}}^{\mathrm{T}}\hat{\boldsymbol{\beta}}_{\mathrm{I}}) = \mathrm{corr}(\mathrm{ESP}, \mathrm{ESP(I)}) = \mathrm{corr}(\hat{\mathrm{Y}}, \hat{\mathrm{Y}}_{\mathrm{I}}) \to 1.$$

Proof. These results are a corollary of Theorem 2.17 below. □

Consider plotting w on the horizontal axis versus z on the vertical axis. The response plot is the plot of $\hat{Y}$ versus Y, and an important residual plot is the plot of $\hat{Y}$ versus r.

Theorem 2.17. Suppose that every submodel contains a constant and that $\boldsymbol{X}$ is a full rank matrix.

Response Plot: i) If $w = \hat{Y}_I$ and $z = Y$ then the OLS line is the identity line.

ii) If $w = Y$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\mathrm{corr}(Y, \hat{Y}_I)]^2 = R^2(I)$ and intercept $a = \overline{Y}(1 - R^2(I))$ where $\overline{Y} = \sum_{i=1}^{n} Y_i/n$ and $R^2(I)$ is the coefficient of multiple determination from the candidate model.

FF or EE Plot: iii) If $w = \hat{Y}_I$ and $z = \hat{Y}$ then the OLS line is the identity line. Note that $ESP(I) = \hat{Y}_I$ and $ESP = \hat{Y}$.

iv) If $w = \hat{Y}$ and $z = \hat{Y}_I$ then the OLS line has slope $b = [\mathrm{corr}(\hat{Y}, \hat{Y}_I)]^2 = SSR(I)/SSR$ and intercept $a = \overline{Y}[1 - (SSR(I)/SSR)]$ where SSR is the regression sum of squares.

RR Plot: v) If $w = r$ and $z = r_I$ then the OLS line is the identity line.

vi) If $w = r_I$ and $z = r$ then $a = 0$ and the OLS slope $b = [\mathrm{corr}(r, r_I)]^2$ and

$$\mathrm{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}} = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}.$$

Proof: Recall that $\boldsymbol{H}$ and $\boldsymbol{H}_I$ are symmetric idempotent matrices and that $\boldsymbol{H}\boldsymbol{H}_I = \boldsymbol{H}_I$. The mean of OLS fitted values is equal to $\overline{Y}$ and the mean of OLS residuals is equal to 0. If the OLS line from regressing z on w is $\hat{z} = a + bw$, then $a = \overline{z} - b\overline{w}$ and

$$b = \frac{\sum(w_i - \overline{w})(z_i - \overline{z})}{\sum(w_i - \overline{w})^2} = \frac{SD(z)}{SD(w)}\mathrm{corr}(z, w).$$

Also recall that the OLS line passes through the means of the two variables $(\overline{w}, \overline{z})$.

(*) Notice that the OLS slope from regressing z on w is equal to one if and only if the OLS slope from regressing w on z is equal to $[\text{corr}(z, w)]^2$.

i) The slope $b = 1$ if $\sum \hat{Y}_{I,i} Y_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}_I^T \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

ii) By (*), the slope

$$b = [\text{corr}(Y, \hat{Y}_I)]^2 = R^2(I) = \frac{\sum (\hat{Y}_{I,i} - \overline{Y})^2}{\sum (Y_i - \overline{Y})^2} = SSR(I)/SSTO.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

iii) The slope $b = 1$ if $\sum \hat{Y}_{I,i} \hat{Y}_i = \sum \hat{Y}_{I,i}^2$. This equality holds since $\hat{\boldsymbol{Y}}^T \hat{\boldsymbol{Y}}_I = \boldsymbol{Y}^T \boldsymbol{H} \boldsymbol{H}_I \boldsymbol{Y} = \boldsymbol{Y}^T \boldsymbol{H}_I \boldsymbol{Y} = \hat{\boldsymbol{Y}}_I^T \hat{\boldsymbol{Y}}_I$. Since $b = 1$, $a = \overline{Y} - \overline{Y} = 0$.

iv) From iii),

$$1 = \frac{SD(\hat{Y})}{SD(\hat{Y}_I)} [\text{corr}(\hat{Y}, \hat{Y}_I)].$$

Hence

$$\text{corr}(\hat{Y}, \hat{Y}_I) = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})}$$

and the slope

$$b = \frac{SD(\hat{Y}_I)}{SD(\hat{Y})} \text{corr}(\hat{Y}, \hat{Y}_I) = [\text{corr}(\hat{Y}, \hat{Y}_I)]^2.$$

Also the slope

$$b = \frac{\sum (\hat{Y}_{I,i} - \overline{Y})^2}{\sum (\hat{Y}_i - \overline{Y})^2} = SSR(I)/SSR.$$

The result follows since $a = \overline{Y} - b\overline{Y}$.

v) The OLS line passes through the origin. Hence $a = 0$. The slope $b = \boldsymbol{r}^T \boldsymbol{r}_I / \boldsymbol{r}^T \boldsymbol{r}$. Since $\boldsymbol{r}^T \boldsymbol{r}_I = \boldsymbol{Y}^T (\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I) \boldsymbol{Y}$ and $(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H}_I) = \boldsymbol{I} - \boldsymbol{H}$, the numerator $\boldsymbol{r}^T \boldsymbol{r}_I = \boldsymbol{r}^T \boldsymbol{r}$ and $b = 1$.

vi) Again $a = 0$ since the OLS line passes through the origin. From v),

$$1 = \sqrt{\frac{SSE(I)}{SSE}} [\text{corr}(r, r_I)].$$

Hence

$$\text{corr}(r, r_I) = \sqrt{\frac{SSE}{SSE(I)}}$$

and the slope

$$b = \sqrt{\frac{SSE}{SSE(I)}}[\text{corr}(r, r_I)] = [\text{corr}(r, r_I)]^2.$$

Algebra shows that

$$\text{corr}(r, r_I) = \sqrt{\frac{n-p}{C_p(I) + n - 2k}} = \sqrt{\frac{n-p}{(p-k)F_I + n - p}}. \quad \square$$

Remark 2.23. a) Let I_{min} be the model than minimizes $C_p(I)$ among the models I generated from the variable selection method such as forward selection. Assuming the full model I_p is one of the models generated, then $C_p(I_{min}) \leq C_p(I_p) = p$, and $\text{corr}(r, r_{I_{min}}) \to 1$ as $n \to \infty$ by Theorem 2.17 vi). Referring to Equation (2.49), if $P(S \subseteq I_{min})$ does not go to 1 as $n \to \infty$, then the above correlation would not go to one. Hence $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. This result is due to Rathnayake and Olive (2023).

b) If none of the $\beta_i = 0$, then $S = F$, the full model. An assumption that some of the β_i are exactly equal to zero may be very strong, but c) and d) suggest that variable selection criterion still select models I that may be as good or better than the full model when $n \geq Jp$ with $J \geq 10$. Also note that Equation (2.49) does not assume that $\boldsymbol{\beta}_E = \mathbf{0}$ if $S = F$, since then E is the empty set, and $\boldsymbol{x} = \boldsymbol{x}_S = \boldsymbol{x}_F$ with $\boldsymbol{\beta} = \boldsymbol{\beta}_S = \boldsymbol{\beta}_F$. For more on the assumption $H_0 : \beta_i = 0$, see, for example, Gelman and Carlin (2017), Nester (1996), and Tukey (1991).

c) If some of the nonzero β_i are very small, then n may need to be very large before $P(S \subseteq I_{min})$ is close to 1. However, by Theorem 2.16, the C_p criterion often picks model $I = I_{min}$ such that the residuals and fitted values from model I are highly correlated with those of the full model F. If $n \geq 10p$, then $C_p(I_{min}) \leq C_p(F) = p$, and thus $\text{corr}(r, r_I) \geq \sqrt{1 - \frac{p}{10p}} \geq \sqrt{0.9} = 0.948$.

d) By Section 2.16, there is often a multitude of good MLR models, and variable selection criterion such as C_p, AIC, and BIC tend to produce a model $I = I_{min}$ such that the residuals and fitted values from model I are highly correlated with those of the full model F.

2.17.1 Variable Selection Theory in Low Dimensions

Large sample theory is often tractable if the optimization problem is convex. The optimization problem for variable selection is not convex, so new tools are needed. Tibshirani et al. (2018) and Leeb and Pötscher (2006, 2008) note that we can not find the limiting distribution of $\boldsymbol{Z}_n = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{I_{min}} - \boldsymbol{\beta}_I)$

after variable selection. One reason is that with positive probability, $\hat{\boldsymbol{\beta}}_{I_{min}}$ does not have the same dimension as $\boldsymbol{\beta}_I$ if AIC or C_p is used. Hence $\boldsymbol{Z}_n$ is not defined with positive probability.

2.17.2 Some Variable Selection Estimators

Consider 1D regression models that study the conditional distribution $Y|\boldsymbol{x}^T\boldsymbol{\beta}$ of the response variable Y given $\boldsymbol{x}^T\boldsymbol{\beta}$ where $\boldsymbol{x}$ is the $p\times 1$ vector of predictors. Many important regression models are special cases, including multiple linear regression, the Nelder and Wedderburn (1972) generalized linear models (GLMs), and the Cox (1972) proportional hazards regression model. Forward selection or backward elimination with the Akaike (1973) AIC criterion or Schwarz (1978) BIC criterion are often used for variable selection.

Sparse regression methods can also be used for variable selection even if n/p is not large: the regression submodel, such as a Nelder and Wedderburn (1972) generalized linear model (GLM), uses the predictors that had nonzero sparse regression estimated coefficients. These methods include least angle regression, lasso, relaxed lasso, elastic net, and sparse regression by projection. Least angle regression variable selection is the LARS-OLS hybrid estimator of Efron et al. (2004, p. 421). Lasso variable selection is called relaxed lasso by Hastie, Tibshirani, and Wainwright (2015, p. 12), and the relaxed lasso estimator with $\phi = 0$ by Meinshausen (2007, p. 376). Also see Fan and Li (2001), Friedman et al. (2007), Friedman, Hastie, and Tibshirani (2010), Qi et al. (2015), Simon et al. (2011), Tibshirani (1996), and Zou and Hastie (2005). The Meinshausen (2007) relaxed lasso estimator fits lasso with penalty λ_n to get a subset of variables with nonzero coefficients, and then fits lasso with a smaller penalty ϕ_n to this subset of variables where n is the sample size.

Let I_{min} correspond to the set of predictors selected by a variable selection method such as forward selection or lasso variable selection. If $\hat{\boldsymbol{\beta}}_I$ is $a\times 1$, use zero padding to form the $p\times 1$ vector $\hat{\boldsymbol{\beta}}_{I,0}$ from $\hat{\boldsymbol{\beta}}_I$ by adding 0s corresponding to the omitted variables. For example, if $p = 4$ and $\hat{\boldsymbol{\beta}}_{I_{min}} = (\hat{\beta}_1, \hat{\beta}_3)^T$, then the observed variable selection estimator $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0} = (\hat{\beta}_1, 0, \hat{\beta}_3, 0)^T$. As a statistic, $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets, e.g. $J = 2^p - 1$.

The large sample theory for $\hat{\boldsymbol{\beta}}_{MIX}$, defined below, is useful for explaining the large sample theory of $\hat{\boldsymbol{\beta}}_{VS}$. Review Section 1.6 for mixture distributions.

Definition 2.28. The *variable selection estimator* $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_{min},0}$, and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities $\pi_{kn} = P(I_{min} = I_k)$ for $k = 1, ..., J$ where there are J subsets.

Definition 2.29. Let $\hat{\boldsymbol{\beta}}_{MIX}$ be a random vector with a mixture distribution of the $\hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities equal to π_{kn}. Hence $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with same probabilities π_{kn} of the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$, but the I_k are randomly selected.

2.17.3 Large Sample Theory for Variable Selection Estimators

Theorems 2.18 and 2.19 in this subsection are due to Rathnayake and Olive (2023), and generalize the Pelawa Watagoda and Olive (2021b) theory for multiple linear regression to many other models. The theory assumes that there is a "true model" S and that at least one subset I is considered such that $S \subseteq I$. For example, with forward selection and backward elimination, the theory assumes that the full model contains S. The theory does not hold if the true model S is not a subset of any of the considered models. For example, S could contain some interactions that were not included in the "full" model. Checking that the full model is good is important.

Assume p is fixed. Suppose model (2.49) holds, and that if $S \subseteq I_j$ where the dimension of I_j is a_j, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j} - \boldsymbol{\beta}_{I_j}) \xrightarrow{D} N_{a_j}(\mathbf{0}, \boldsymbol{V}_j)$ where $\boldsymbol{V}_j$ is the covariance matrix of the asymptotic multivariate normal distribution. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{V}_{j,0}) \tag{2.50}$$

where $\boldsymbol{V}_{j,0}$ adds columns and rows of zeros corresponding to the x_i not in I_j, and $\boldsymbol{V}_{j,0}$ is singular unless I_j corresponds to the full model. This large sample theory holds for many models, including multiple linear regression fit by least squares (OLS), GLMs fit by maximum likelihood, and Cox regression fit by maximum partial likelihood. See, for example, Sen and Singer (1993, pp. 280, 309).

The first assumption in Theorem 2.18 is $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Then the variable selection estimator corresponding to I_{min} underfits with probability going to zero, and the assumption holds under regularity conditions if BIC or AIC is used for many parametric regression models such as GLMs. See Charkhi and Claeskens (2018) and Claeskens and Hjort (2008, pp. 70, 101, 102, 114, 232). This assumption is a necessary condition for a variable selection estimator to be a consistent estimator. See Zhao and Yu (2006). Thus if a sparse estimator that does variable selection is a consistent estimator of $\boldsymbol{\beta}$, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Hence Theorem 2.18c) proves that the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}$ if lasso and elastic net are consistent. Also see Theorem 2.19. The assumption on $\boldsymbol{u}_{jn}$ in Theorem 2.18 is reasonable by (2.50) since $S \subseteq I_j$ for each π_j, and since $\hat{\boldsymbol{\beta}}_{MIX}$ uses random selection.

Consider the assumption $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$ for multiple linear regression. Charkhi and Claeskens (2018) proved the assumption holds for AIC for a wide variety of error distributions. Shao (1993) gave similar results for AIC, BIC, and C_p. Also see Remark 2.23 a). The assumption holds for lasso variable selection and elastic net variable selection provided that $\hat{\lambda}_n/n \to 0$ as $n \to \infty$ so lasso and elastic net are consistent estimators. Here $\hat{\lambda}_n$ is the shrinkage penalty parameter selected after k-fold cross validation. See Theorems 2.8, 2.9, Pelawa Watogoda and Olive (2021b) and Knight and Fu (2000).

Theorem 2.18 a) proves that $\boldsymbol{u}$ is a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma_u} = \sum_j \pi_j \boldsymbol{V}_{j,0}$. Some of the submodels I_k will have $\pi_k = 0$. For example, since the probability of underfitting goes to zero, every submodel I_k that underfits has $\pi_k = 0$. Hence $S \subseteq I_j$ corresponding to the $\pi_j > 0$. If $\pi_d = 1$, then submodel I_d is picked with probability going to 1 as $n \to \infty$, and I_d is the only submodel with a positive π_k. Often $\pi_d = \pi_S$ in the literature. For $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX}$ with $\boldsymbol{\theta} = \boldsymbol{A\beta}$, we have $\sqrt{n}(T_n - \boldsymbol{\theta}) \xrightarrow{D} \boldsymbol{v}$ by (2.52) where $E(\boldsymbol{v}) = \boldsymbol{0}$, and $\boldsymbol{\Sigma_v} = \sum_j \pi_j \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T$.

Theorem 2.18. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{MIX} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{u}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u}_j \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{j,0})$. a) Then

$$\boldsymbol{u}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \tag{2.51}$$

where the cdf of $\boldsymbol{u}$ is $F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$. Thus $\boldsymbol{u}$ has a mixture distribution of the $\boldsymbol{u}_j$ with probabilities π_j, $E(\boldsymbol{u}) = \boldsymbol{0}$, and $\mathrm{Cov}(\boldsymbol{u}) = \boldsymbol{\Sigma_u} = \sum_j \pi_j \boldsymbol{V}_{j,0}$.

b) Let $\boldsymbol{A}$ be a $g \times p$ full rank matrix with $1 \leq g \leq p$. Then

$$\boldsymbol{v}_n = \boldsymbol{A}\boldsymbol{u}_n = \sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} = \boldsymbol{v} \tag{2.52}$$

where $\boldsymbol{v}$ has a mixture distribution of the $\boldsymbol{v}_j = \boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T)$ with probabilities π_j.

c) The estimator $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$: $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) = O_P(1)$.

d) If $\pi_d = 1$, then $\sqrt{n}(\hat{\boldsymbol{\beta}}_{SEL} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{u} \sim N_p(\boldsymbol{0}, \boldsymbol{V}_{d,0})$ where SEL is VS or MIX.

Proof. a) Since $\boldsymbol{u}_n$ has a mixture distribution of the $\boldsymbol{u}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{u}_n$ is $F_{\boldsymbol{u}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{u}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{u}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{u}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{u}_j}(\boldsymbol{t})$ as $n \to \infty$.

b) Since $\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{u}$, then $\boldsymbol{A}\boldsymbol{u}_n \xrightarrow{D} \boldsymbol{A}\boldsymbol{u}$.

c) The result follows since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959).

d) If $\pi_d = 1$, there is no selection bias, asymptotically. The result also follows by Pötscher (1991, Lemma 1). □

The following subscript notation is useful. Subscripts before the MIX are used for subsets of $\hat{\boldsymbol{\beta}}_{MIX} = (\hat{\beta}_1, ..., \hat{\beta}_p)^T$. Let $\hat{\boldsymbol{\beta}}_{i,MIX} = \hat{\beta}_i$. Similarly, if $I = \{i_1, ..., i_a\}$, then $\hat{\boldsymbol{\beta}}_{I,MIX} = (\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_a})^T$. Subscripts after MIX denote the ith vector from a sample $\hat{\boldsymbol{\beta}}_{MIX,1}, ..., \hat{\boldsymbol{\beta}}_{MIX,B}$. Similar notation is used for other estimators such as $\hat{\boldsymbol{\beta}}_{VS}$. The subscript 0 is still used for zero padding. We may use $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{FULL}$ to denote the full model.

Typically the mixture distribution is not asymptotically normal unless a $\pi_d = 1$ (e.g. if S is the full model F), or if for each π_j, $\boldsymbol{A}\boldsymbol{u}_j \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{V}_{j,0}\boldsymbol{A}^T) = N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. Then $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{A}\boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{A}\boldsymbol{u} \sim N_g(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T)$. This special case occurs for $\hat{\boldsymbol{\beta}}_{S,MIX}$ if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V})$ where the asymptotic covariance matrix $\boldsymbol{V}$ is diagonal and nonsingular. Then $\hat{\boldsymbol{\beta}}_{S,MIX}$ and $\hat{\boldsymbol{\beta}}_{S,FULL}$ have the same multivariate normal limiting distribution. For several criteria, this result should hold for $\hat{\boldsymbol{\beta}}_{VS}$ since asymptotically, $\sqrt{n}(\boldsymbol{A}\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{A}\boldsymbol{\beta})$ is selecting from the $\boldsymbol{A}\boldsymbol{u}_j$ which have the same distribution. In the simulations when $\boldsymbol{V}$ is diagonal, the confidence regions applied to $\boldsymbol{A}\hat{\boldsymbol{\beta}}^*_{SEL} = \boldsymbol{B}\hat{\boldsymbol{\beta}}^*_{S,SEL}$ had similar volume and cutoffs where SEL is MIX, VS, or $FULL$.

Theorem 2.18 can be used to justify prediction intervals after variable selection. See Pelawa Watagoda and Olive (2021b) and Olive, Rathnayake, and Haile (2022). Theorem 2.18 d) is useful for *variable selection consistency* and the *oracle property* where $\pi_d = \pi_S = 1$ if $P(I_{min} = S) \to 1$ as $n \to \infty$. See Claeskens and Hjort (2008, pp. 101-114) and Fan and Li (2001) for references. A necessary condition for $P(I_{min} = S) \to 1$ is that S is one of the models considered with probability going to one. This condition holds under very strong regularity conditions for fast methods if $S \neq F$. See Wieczorek and Lei (2022) for forward selection and Hastie, Tibshirani, and Wainwright (2015, pp. 295-302) for lasso, where the predictors need a "near orthogonality" condition.

Remark 2.24. If $A_1, A_2, ..., A_k$ are pairwise disjoint and if $\cup_{i=1}^k A_i = S$, then the collection of sets $A_1, A_2, ..., A_k$ is a *partition* of S. Then the *Law of Total Probability* states that if $A_1, A_2, ..., A_k$ form a partition of S such that $P(A_i) > 0$ for $i = 1, ..., k$, then

$$P(B) = \sum_{j=1}^{k} P(B \cap A_j) = \sum_{j=1}^{k} P(B|A_j)P(A_j).$$

Let sets $A_{k+1}, ..., A_m$ satisfy $P(A_i) = 0$ for $i = k+1, ..., m$. Define $P(B|A_j) = 0$ if $P(A_j) = 0$. Then a Generalized Law of Total Probability is

$$P(B) = \sum_{j=1}^{m} P(B \cap A_j) = \sum_{j=1}^{m} P(B|A_j)P(A_j),$$

and will be used in the proof of the result in the following paragraph.

Pötscher (1991) used the conditional distribution of $\hat{\boldsymbol{\beta}}_{VS}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$ to find the distribution of $\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$. Let $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ be a random vector from the conditional distribution $\hat{\boldsymbol{\beta}}_{I_k,0}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})$. Let $\boldsymbol{w}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) \sim \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta})$. Denote $F_{\boldsymbol{z}}(\boldsymbol{t}) = P(z_1 \le t_1, ..., z_p \le t_p)$ by $P(\boldsymbol{z} \le \boldsymbol{t})$. Then Pötscher (1991) and Pelawa Watagoda and Olive (2021b) show

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}.$$

Hence $\hat{\boldsymbol{\beta}}_{VS}$ has a mixture distribution of the $\hat{\boldsymbol{\beta}}^C_{I_k,0}$ with probabilities π_{kn}, and $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}.

Proof: Let $W = W_{VS} = k$ if $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ where $P(W_{VS} = k) = \pi_{kn}$ for $k = 1, ..., J$. Then $(\hat{\boldsymbol{\beta}}_{VS:n}, W_{VS:n}) = (\hat{\boldsymbol{\beta}}_{VS}, W_{VS})$ has a joint distribution where the sample size n is usually suppressed. Note that $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_W,0}$. Then by Remark 2.24,

$$F_{\boldsymbol{w}_n}(\boldsymbol{t}) = P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}] =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]P(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}) =$$

$$\sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}|(\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0})]\pi_{kn}$$

$$= \sum_{k=1}^{J} P[n^{1/2}(\hat{\boldsymbol{\beta}}^C_{I_k,0} - \boldsymbol{\beta}) \le \boldsymbol{t}]\pi_{kn} = \sum_{k=1}^{J} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t})\pi_{kn}. \quad \square$$

Charkhi and Claeskens (2018) showed that $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$ if $S \subseteq I_j$ for the maximum likelihood estimator (MLE) with AIC, and gave a forward selection example. They claim that $\boldsymbol{w}_j$ is a multivariate truncated normal distribution (where no truncation is possible) that is symmetric about $\boldsymbol{0}$. Hence $E(\boldsymbol{w}_j) = 0$, and $\text{Cov}(\boldsymbol{w}_j) = \boldsymbol{\Sigma}_j$ exits. Note that both $\sqrt{n}(\hat{\boldsymbol{\beta}}_{MIX} - \boldsymbol{\beta})$ and $\sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta})$ are selecting from the $\boldsymbol{u}_{kn} = \sqrt{n}(\hat{\boldsymbol{\beta}}_{I_k,0} - \boldsymbol{\beta})$ and asymptotically from the $\boldsymbol{u}_j$. The random selection for $\hat{\boldsymbol{\beta}}_{MIX}$ does not change the distribution of $\boldsymbol{u}_{jn}$, but selection bias does change the distribution of the selected $\boldsymbol{u}_{jn}$ and $\boldsymbol{u}_j$ to that of $\boldsymbol{w}_{jn}$ and $\boldsymbol{w}_j$. The assumption that $\boldsymbol{w}_{jn} \xrightarrow{D} \boldsymbol{w}_j$ may not be mild. The proof for Equation (2.53) is the same as that for (2.51).

Theorem 2.19 proves that $\boldsymbol{w}$ is a mixture distribution of the $\boldsymbol{w}_j$ with probabilities π_j.

Theorem 2.19. Assume $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and let $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{I_k,0}$ with probabilities π_{kn} where $\pi_{kn} \to \pi_k$ as $n \to \infty$. Denote the positive π_k by π_j. Assume $\boldsymbol{w}_{jn} = \sqrt{n}(\hat{\boldsymbol{\beta}}^C_{I_j,0} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w}_j$. Then

$$\boldsymbol{w}_n = \sqrt{n}(\hat{\boldsymbol{\beta}}_{VS} - \boldsymbol{\beta}) \xrightarrow{D} \boldsymbol{w} \tag{2.53}$$

where the cdf of $\boldsymbol{w}$ is $F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$.

Proof. Since $\boldsymbol{w}_n$ has a mixture distribution of the $\boldsymbol{w}_{kn}$ with probabilities π_{kn}, the cdf of $\boldsymbol{w}_n$ is $F_{\boldsymbol{w}_n}(\boldsymbol{t}) = \sum_k \pi_{kn} F_{\boldsymbol{w}_{kn}}(\boldsymbol{t}) \to F_{\boldsymbol{w}}(\boldsymbol{t}) = \sum_j \pi_j F_{\boldsymbol{w}_j}(\boldsymbol{t})$ at continuity points of the $F_{\boldsymbol{w}_j}(\boldsymbol{t})$ as $n \to \infty$. □

Remark 2.25. a) If $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, then $\hat{\boldsymbol{\beta}}_{VS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}$ since selecting from a finite number J of $\sqrt{n}$ consistent estimators (even on a set that goes to one in probability) results in a $\sqrt{n}$ consistent estimator by Pratt (1959). By both this result and Theorems 2.18 and 2.19, the lasso variable selection and elastic net variable selection estimators are $\sqrt{n}$ consistent if lasso and elastic net are consistent.

b) If the data is not simulated, then having some $\beta_i = 0$ may not be reasonable. Then $S = F$ and Theorem 2.19 proves that $\hat{\boldsymbol{\beta}}_{VS}$ and $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_F$ are asymptotically equivalent. Also see Remark 2.23.

Remark 2.26. Another variable selection model is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_{S_i}^T \boldsymbol{\beta}_{S_i}$ for $i = 1, ..., K$. Then submodel I underfits if no $S_i \subseteq I$. A necessary condition for an estimator to be consistent is P(no $S_i \subseteq I_{min}$) $\to 0$ as $n \to \infty$. By Remark 2.23, the above probability holds if C_p is used. Then in Theorem 2.19, we can replace $P(S \subseteq I_{min}) \to 1$ by P(no $S_i \subseteq I_{min}$) $\to 0$ as $n \to \infty$.

Example 2.4. This is an example where the $\pi_{kn} \to \pi_k$ as $n \to \infty$. Assume $S \subseteq I$ where I has a predictors, including a constant. Then for a wide variety of iid error distributions, $F_I \xrightarrow{D} X/(p-a)$ where $X \sim \chi^2_{p-a}$. Let F denote the full model, and let $S = I = I_i$ be the model that deletes predictor x_i with $a = p-1$. Then from Definition 2.27, $C_p(I) \xrightarrow{D} X + p - 2$ where $X \sim \chi^2_1$. Let F denote the full model and consider all subsets variable selection with C_p. Since only S and F do not underfit, only π_S and π_F are positive. Since $C_p(F) = p$, $I = S$ is selected if $C_p(I) < p$. Hence $\pi_S = P(\chi^2_1 + p - 2 < p) = P(\chi^2_1 < 2) = 0.8427$, and $\pi_F = 1 - \pi_S = 0.1573$. This result also holds for backward elimination since the probability that x_i will be the first predictor deleted goes to 1 as $n \to \infty$ because $C_p(I_i) = C_p(S)$ is bounded in probability while $C_p(I_j)$ diverges as $n \to \infty$ for $j \neq i$. For forward selection with correlated predictors, expect that $\pi_S < P(\chi^2_1 < 2)$, and hence $\pi_F > 1 - P(\chi^2_1 < 2)$.

For the *R* code below, $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ is a $p \times 1$ vector with $k+1$ ones and $p - k + 1$ zeroes. Hence $k = p - 2$ deletes the predictor x_p. The function `belimsim` generates 1000 data sets, performs backward elimination, and finds the proportion of time the full model was selected, which was $0.158 \approx 0.1573$.

```
belimsim(n=100,p=5,k=3,nruns=1000)
$fullprop
[1] 0.158
```

2.17.4 Variable Selection Theory in High Dimensions

Remark 2.27. a) When $\sqrt{n}$ consistent estimators are used,

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2 = \|\hat{\boldsymbol{\beta}}_F - \boldsymbol{\beta}_F\|^2 = \sum_{i=1}^{n} (\hat{\beta}_i - \beta_i)^2 \propto \frac{p}{n}. \tag{2.54}$$

In low dimensions where p is fixed, $p/n \to 0$ as $n \to \infty$ and $\hat{\boldsymbol{\beta}}$ is a consistent estimator. In high dimensions, $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ tends to not be close to 0. For example, if $p = p_n = n^{\tau+1}$, then $p_n/n = n^{\tau}$ which tends to be large if n is large and $\tau > 1$. Hence in high dimensions, it is difficult to get a good estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta} = \boldsymbol{\beta}_F$ for the full model that uses all p predictors $x_1, ..., x_p$.

b) When $n/p \to 0$ as $n \to \infty$, consistent estimators of $\boldsymbol{\beta}_F$ generally cannot be found unless the model has a simplifying structure. A sparse population model is one such structure. Let model I be the model selected by a procedure such as lasso. For Equation (2.49), assume that $\boldsymbol{\beta}_S$ is $a_S \times 1$, $\boldsymbol{\beta}_I$ is $k \times 1$, $S \subseteq I$, $n \geq Jk$ with $J > 1$ and preferably $J \geq 10$, and $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta} = \boldsymbol{\beta}_F$. If a $\sqrt{n}$ consistent estimator is used, then

$$\|\hat{\boldsymbol{\beta}}_{I,0} - \boldsymbol{\beta}_F\|^2 = \|\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I\|^2 = \sum_{i=1}^{k} (\hat{\beta}_{iI} - \beta_{iI})^2 \propto k/n$$

which can be small. This "bet on sparsity principle" requires that a large percentage of the $\beta_i = 0$ and that the method selects I such that $S \subseteq I$ with high probability where k/n is small. The assumptions $S \subseteq I$ and $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta}_F$ may be very strong. There is a large literature on "sparsity bounds." See Giraud (2022) and Wainwright (2019) for references.

We can also consider sparse fitted models $\hat{\boldsymbol{\beta}}_I$ that use k predictors with $n \geq Jk$ with $J \geq 5$. With the sparse fitted model, we are not necessarily assuming that i) $S \subseteq I$, that ii) $S \neq F$, or that iii) $\boldsymbol{\beta}_{I,0} = \boldsymbol{\beta}_F$. We can also use data splitting with $n_H \geq Jk$ with $J \geq 5$. Check that the selected model is reasonable, using response plots if possible.

2.17.5 Sparse Models

For multiple linear regression with $p > n$, results from Hastie et al. (2015, pp. 20, 296, ch. 6, ch. 11) and Luo and Chen (2013) suggest that lasso, lasso variable selection, and forward selection with EBIC can perform well for sparse models. Least angle regression, elastic net, and elastic net variable selection can also be useful.

Suppose the selected model is I_d, and $\boldsymbol{\beta}_{I_d}$ is $a_d \times 1$. For multiple linear regression, forward selection with C_p and AIC often gives useful results if $n \geq 5p$ and if the final model I has $n \geq 10a_d$. For $p < n < 5p$, forward selection with C_p and AIC tends to pick the full model (which overfits since $n < 5p$) too often, especially if $\hat{\sigma}^2 = MSE$. The Hurvich and Tsai (1989) AIC_C criterion can be useful for MLR and time series if $n \geq \max(2p, 10a_d)$. If $n \geq 5p$, AIC and BIC are useful for many regression models, and forward selection with EBIC can be used for some models if n/p is small. See Section 2.1 and Chen and Chen (2008).

2.18 Summary

1) The MLR model is $Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. This model is also called the **full model**. In matrix notation, these n equations become $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Note that $x_{i,1} \equiv 1$.

2) The ordinary least squares OLS full model estimator $\hat{\boldsymbol{\beta}}_{OLS}$ minimizes $Q_{OLS}(\boldsymbol{\beta}) = \sum_{i=1}^n r_i^2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. In the estimating equations $Q_{OLS}(\boldsymbol{\beta})$, the vector $\boldsymbol{\beta}$ is a dummy variable. The minimizer $\hat{\boldsymbol{\beta}}_{OLS}$ estimates the parameter vector $\boldsymbol{\beta}$ for the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Note that $\hat{\boldsymbol{\beta}}_{OLS} \sim AN_p(\boldsymbol{\beta}, MSE(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

3) Given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *predicted values* or *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$. Thus the ith fitted value

$$\hat{Y}_i \equiv \hat{Y}_i(\boldsymbol{b}) = \boldsymbol{x}_i^T\boldsymbol{b} = x_{i,1}b_1 + \cdots + x_{i,p}b_p.$$

The vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. Thus ith residual $r_i \equiv r_i(\boldsymbol{b}) = Y_i - \hat{Y}_i(\boldsymbol{b}) = Y_i - x_{i,1}b_1 - \cdots - x_{i,p}b_p$. A *response plot* for MLR is a plot of $\hat{Y}_i$ versus Y_i. A *residual plot* is a plot of $\hat{Y}_i$ versus r_i. If the e_i are iid from a unimodal distribution that is not highly skewed, the plotted points should scatter about the identity line and the $r = 0$ line.

	Label	coef	SE	shorth 95% CI for β_i
4)	Constant=intercept= x_1	$\hat{\beta}_1$	$SE(\hat{\beta}_1)$	$[\hat{L}_1, \hat{U}_1]$
	x_2	$\hat{\beta}_2$	$SE(\hat{\beta}_2)$	$[\hat{L}_2, \hat{U}_2]$
	$\vdots$			
	x_p	$\hat{\beta}_p$	$SE(\hat{\beta}_p)$	$[\hat{L}_p, \hat{U}_p]$

The classical OLS large sample 95% CI for β_i is $\hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$. Consider testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If $0 \in$ CI for β_i, then fail to reject H_0, and conclude x_i is not needed in the MLR model given the other predictors are in the model. If $0 \notin$ CI for β_i, then reject H_0, and conclude x_i is needed in the MLR model.

5) Let $\boldsymbol{x}_i^T = (1 \;\; \boldsymbol{u}_i^T)$. It is often convenient to use the centered response $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W} = (W_{ij})$. For $j = 1, ..., p-1$, let W_{ij} denote the $(j+1)$th variable standardized so that $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n$. Then the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T \boldsymbol{W}}{n}.$$

Then regression through the origin is used for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$ where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$. Thus the centered response $Z_i = Y_i - \overline{Y}$ and $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Then $\hat{\boldsymbol{\eta}}$ does not depend on the units of measurement of the predictors. Linear combinations of the $\boldsymbol{u}_i$ can be written as linear combinations of the $\boldsymbol{x}_i$, hence $\hat{\boldsymbol{\beta}}$ can be found from $\hat{\boldsymbol{\eta}}$.

6) A *model for variable selection* is $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{x} = (\boldsymbol{x}_S^T, \boldsymbol{x}_E^T)^T$, $\boldsymbol{x}_S$ is an $a_S \times 1$ vector, and $\boldsymbol{x}_E$ is a $(p - a_S) \times 1$ vector. Let $\boldsymbol{x}_I$ be the vector of a terms from a candidate subset indexed by I, and let $\boldsymbol{x}_O$ be the vector of the remaining predictors (out of the candidate submodel). If $S \subseteq I$, then $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_{I/S}^T\boldsymbol{\beta}_{(I/S)} + \boldsymbol{x}_O^T\boldsymbol{0} = \boldsymbol{x}_I^T\boldsymbol{\beta}_I$ where $\boldsymbol{x}_{I/S}$ denotes the predictors in I that are not in S. Since this is true regardless of the values of the predictors, $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. Note that $\boldsymbol{\beta}_E = \boldsymbol{0}$. Let $k_S = a_S - 1 =$ the number of population active nontrivial predictors. Then $k = a - 1$ is the number of active predictors in the candidate submodel I.

7) Let $Q(\boldsymbol{\eta})$ be a real valued function of the $k \times 1$ vector $\boldsymbol{\eta}$. The gradient of $Q(\boldsymbol{\eta})$ is the $k \times 1$ vector

$$\bigtriangledown Q = \bigtriangledown Q(\boldsymbol{\eta}) = \frac{\partial Q}{\partial \boldsymbol{\eta}} = \frac{\partial Q(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \frac{\partial}{\partial \eta_1} Q(\boldsymbol{\eta}) \\ \frac{\partial}{\partial \eta_2} Q(\boldsymbol{\eta}) \\ \vdots \\ \frac{\partial}{\partial \eta_k} Q(\boldsymbol{\eta}) \end{bmatrix}.$$

Suppose there is a model with unknown parameter vector $\boldsymbol{\eta}$. A set of *estimating equations* $f(\boldsymbol{\eta})$ is minimized or maximized where $\boldsymbol{\eta}$ is a dummy variable vector in the function $f : \mathbb{R}^k \to \mathbb{R}^k$.

8) As a mnemonic (memory aid) for the following results, note that the derivative $\frac{d}{dx}ax = \frac{d}{dx}xa = a$ and $\frac{d}{dx}ax^2 = \frac{d}{dx}xax = 2ax$.

a) If $Q(\boldsymbol{\eta}) = \boldsymbol{a}^T\boldsymbol{\eta} = \boldsymbol{\eta}^T\boldsymbol{a}$ for some $k \times 1$ constant vector $\boldsymbol{a}$, then $\bigtriangledown Q = \boldsymbol{a}$.

b) If $Q(\boldsymbol{\eta}) = \boldsymbol{\eta}^T\boldsymbol{A}\boldsymbol{\eta}$ for some $k \times k$ constant matrix $\boldsymbol{A}$, then $\bigtriangledown Q = 2\boldsymbol{A}\boldsymbol{\eta}$.

c) If $Q(\boldsymbol{\eta}) = \sum_{i=1}^k |\eta_i| = \|\boldsymbol{\eta}\|_1$, then $\bigtriangledown Q = \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$ where $s_i = \text{sign}(\eta_i)$ where $\text{sign}(\eta_i) = 1$ if $\eta_i > 0$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 0$. This gradient is only defined for $\boldsymbol{\eta}$ where none of the k values of η_i are equal to 0.

9) Forward selection with OLS generates a sequence of M models $I_1, ..., I_M$ where I_j uses j predictors $x_1^* \equiv 1, x_2^*, ..., x_M^*$. Often $M = \min(\lceil n/J \rceil, p)$ where J is a positive integer such as $J = 5$.

10) For the model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, methods such as forward selection, PCR, PLS, ridge regression, lasso variable selection, and lasso each generate M fitted models $I_1, ..., I_M$, where M depends on the method. For forward selection the simulation used C_p for $n \geq 10p$ and EBIC for $n < 10p$. The other methods minimized 10-fold CV. For forward selection, the maximum number of variables used was approximately $\min(\lceil n/5 \rceil, p)$.

11) Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1} |\eta_i|^j \tag{2.55}$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Then $j = 2$ corresponds to ridge regression $\hat{\boldsymbol{\eta}}_R$, $j = 1$ corresponds to lasso $\hat{\boldsymbol{\eta}}_L$, and $a = 1, 2, n$, and $2n$ are common. The residual sum of squares $RSS_W(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})$, and $\lambda_{1,n} = 0$ corresponds to the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Note that for a $k \times 1$ vector $\boldsymbol{\eta}$, the squared (Euclidean) L_2 norm $\|\boldsymbol{\eta}\|_2^2 = \boldsymbol{\eta}^T\boldsymbol{\eta} = \sum_{i=1}^k \eta_i^2$ and the L_1 norm $\|\boldsymbol{\eta}\|_1 = \sum_{i=1}^k |\eta_i|$.

Lasso and ridge regression have a parameter λ. When $\lambda = 0$, the OLS full model is used. Otherwise, the centered response and scaled nontrivial predictors are used with $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. See 5). These methods also use a maximum value λ_M of λ and a grid of M λ values $0 \leq \lambda_1 < \lambda_2 < \cdots < \lambda_{M-1} < \lambda_M$ where often $\lambda_1 = 0$. For lasso, λ_M is the smallest value of λ such that $\hat{\boldsymbol{\eta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \boldsymbol{0}$ for $i < M$.

12) The elastic net estimator $\hat{\boldsymbol{\eta}}_{EN}$ minimizes

$$Q_{EN}(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{2.56}$$

where $\lambda_1 = (1 - \alpha)\lambda_{1,n}$ and $\lambda_2 = 2\alpha\lambda_{1,n}$ with $0 \leq \alpha \leq 1$.

13) Use $\boldsymbol{Z}_n \sim AN_g(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$ to indicate that a normal approximation is used: $\boldsymbol{Z}_n \approx N_g(\boldsymbol{\mu}_n, \boldsymbol{\Sigma}_n)$. Let a be a constant, let $\boldsymbol{A}$ be a $k \times g$ constant

matrix, and let $\boldsymbol{c}$ be a $k \times 1$ constant vector. If $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_g(\boldsymbol{0}, \boldsymbol{V})$, then $a\boldsymbol{Z}_n = a\boldsymbol{I}_g \boldsymbol{Z}_n$ with $\boldsymbol{A} = a\boldsymbol{I}_g$,

$$a\boldsymbol{Z}_n \sim AN_g\left(a\boldsymbol{\mu}_n, a^2\boldsymbol{\Sigma}_n\right), \quad \text{and} \quad \boldsymbol{A}\boldsymbol{Z}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\mu}_n + \boldsymbol{c}, \boldsymbol{A}\boldsymbol{\Sigma}_n\boldsymbol{A}^T\right),$$

$$\hat{\boldsymbol{\theta}}_n \sim AN_g\left(\boldsymbol{\theta}, \frac{\boldsymbol{V}}{n}\right), \quad \text{and} \quad \boldsymbol{A}\hat{\boldsymbol{\theta}}_n + \boldsymbol{c} \sim AN_k\left(\boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{c}, \frac{\boldsymbol{A}\boldsymbol{V}\boldsymbol{A}^T}{n}\right).$$

14) Assume $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$. Let $\boldsymbol{s}_n = (s_{1n}, ..., s_{p-1,n})^T$ where $s_{in} \in [-1, 1]$ and $s_{in} = \text{sign}(\hat{\eta}_i)$ if $\hat{\eta}_i \neq 0$. Here $\text{sign}(\eta_i) = 1$ if $\eta_i > 1$ and $\text{sign}(\eta_i) = -1$ if $\eta_i < 1$. Then

i) $\hat{\boldsymbol{\eta}}_R = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1n}}{n} n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$.

ii) $\hat{\boldsymbol{\eta}}_L = \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda_{1,n}}{2n}\, n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\, \boldsymbol{s}_n$.

iii) $\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_{OLS} - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\, \left[\frac{\lambda_1}{n}\hat{\boldsymbol{\eta}}_{OLS} + \frac{\lambda_2}{2n}\boldsymbol{s}_n\right]$.

15) Assume that the sample correlation matrix $\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}$. Let $\boldsymbol{H} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = (h_{ij})$, and assume that $\max_{i=1,...,n} h_{ii} \xrightarrow{P} 0$ as $n \to \infty$. Let $\hat{\boldsymbol{\eta}}_A$ be $\hat{\boldsymbol{\eta}}_{EN}$, $\hat{\boldsymbol{\eta}}_L$, or $\hat{\boldsymbol{\eta}}_R$. Let p be fixed.

i) LS CLT: $\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V})$.

ii) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_A - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}).$$

iii) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, $\hat{\alpha} \xrightarrow{P} \psi \in [0, 1]$, *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{EN} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(-\boldsymbol{V}[(1-\psi)\tau\boldsymbol{\eta} + \psi\tau\boldsymbol{s}], \sigma^2\boldsymbol{V}\right).$$

iv) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_R - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}).$$

v) *If* $\hat{\lambda}_{1,n}/\sqrt{n} \xrightarrow{P} \tau \geq 0$ *and* $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right).$$

ii) and v) are the Lasso CLT, ii) and iv) are the RR CLT, and ii) and iii) are the EN CLT.

16) Under the conditions of 15), lasso variable selection and elastic net variable selection are $\sqrt{n}$ consistent under much milder conditions than lasso

and elastic net, since the variable selection estimators are $\sqrt{n}$ consistent when lasso and elastic net are consistent. Let I_{min} correspond to the predictors chosen by lasso, elastic net, or forward selection, including a constant. Let $\hat{\boldsymbol{\beta}}_{I_{min}}$ be the OLS estimator applied to these predictors, let $\hat{\boldsymbol{\beta}}_{I_{min},0}$ be the zero padded estimator. The large sample theory for $\hat{\boldsymbol{\beta}}_{I_{min},0}$ (from forward selection, lasso variable selection, and elastic net variable selection) is given by Theorem 2.4. Note that the large sample theory for the estimators $\hat{\boldsymbol{\beta}}$ is given for $p \times 1$ vectors. The theory for $\hat{\boldsymbol{\eta}}$ is given for $(p-1) \times 1$ vectors In particular, the theory for lasso and elastic net does not cast away the $\hat{\eta}_i = 0$.

17) Under Equation (2.1) with p fixed, if lasso or elastic net are consistent, then $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$. Hence when lasso and elastic net do variable selection, they are often not $\sqrt{n}$ consistent.

18) Refer to 6). a) The *OLS full model* tends to be useful if $n \geq 10p$ with large sample theory better than that of lasso, ridge regression, and elastic net. Testing is easier and the Olive (2007) PI tailored to the OLS full model will work better for smaller sample sizes than PI (2.14) if $n \geq 10p$. If $n \geq 10p$ but $\boldsymbol{X}^T\boldsymbol{X}$ is singular or ill conditioned, other methods can perform better.

Forward selection, lasso variable selection, and elastic net variable selection are competitive with the OLS full model even when $n \geq 10p$ and $\boldsymbol{X}^T\boldsymbol{X}$ is well conditioned. If $n \leq p$ then OLS interpolates the data and is a poor method. If $n = Jp$, then as J decreases from 10 to 1, other methods become competitive.

b) If $n \geq 10p$ and $k_S < p-1$, then *forward selection* can give more precise inference than the OLS full model. When n/p is small, the PI (2.14) for forward selection can perform well if n/k_S is large. Forward selection can be worse than ridge regression or elastic net if $k_S > \min(n/J, p)$. Forward selection can be too slow if both n and p are large. Forward selection, lasso variable selection, and elastic net variable selection tend to be bad if $(\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}$ is ill conditioned where $A = I_{min}$.

c) If $n \geq 10p$, *lasso* can be better than the OLS full model if $\boldsymbol{X}^T\boldsymbol{X}$ is ill conditioned. Lasso seems to perform best if k_S is not much larger than 10 or if the nontrivial predictors are orthogonal or uncorrelated. Lasso can be outperformed by ridge regression or elastic net if $k_S > \min(n, p-1)$.

d) If $n \geq 10p$ *ridge regression* and *elastic net* can be better than the OLS full model if $\boldsymbol{X}^T\boldsymbol{X}$ is ill conditioned. Ridge regression (and likely elastic net) seems to perform best if k_S is not much larger than 10 or if the nontrivial predictors are orthogonal or uncorrelated. Ridge regression and elastic net can outperform lasso if $k_S > \min(n, p-1)$.

e) The *PLS* PI (2.14) can perform well if $n \geq 10p$ if some of the other five methods used in the simulations start to perform well when $n \geq 5p$. PLS may or may not be inconsistent if n/p is not large. Ridge regression tends to be inconsistent unless $P(d \to p) \to 1$ so that ridge regression is asymptotically equivalent to the OLS full model.

19) Under strong regularity conditions, lasso and lasso variable selection with k–fold CV, and forward selection with EBIC can perform well even if n/p is small. So PI (2.14) can be useful when n/p is small.

20) Using the response variable to build a model is known as data snooping, and invalidates inference if data snooping is used on the entire data set of n cases.

21) Suppose $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$ where $\boldsymbol{\beta}_S$ is an $a_S \times 1$ vector. A regression model is sparse if a_S is small. We want $n \geq 10a_S$.

22) Assume the cases are independent. To perform data splitting, randomly divide the data into two half sets H and V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. Build the model, possibly with data snooping, or perform variable selection to Find a model I, possibly with data snooping or model selection, using the data in the training set H. Use the model I as the full model to perform inference using the data in the validation set V.

2.19 Complements

Good references for forward selection, PCR, PLS, ridge regression, and lasso are Hastie et al. (2009, 2015), James et al. (2013), and Pelawa Watagoda and Olive (2021b). Also see Efron and Hastie (2016). An early reference for forward selection is Efroymson (1960). Under strong regularity conditions, Gunst and Mason (1980, ch. 10) covers inference for ridge regression (and a modified version of PCR) when the iid errors $e_i \sim N(0, \sigma^2)$.

Xu et al. (2011) notes that sparse algorithms are not stable. Belsley (1984) shows that centering can mask ill conditioning of $\boldsymbol{X}$.

Classical principal component analysis based on the correlation matrix can be done using the singular value decomposition (SVD) of the scaled matrix $\boldsymbol{W}_S = \boldsymbol{W}_g/\sqrt{n-1}$ using $\hat{\boldsymbol{e}}_i$ and $\hat{\lambda}_i = \sigma_i^2$ where $\hat{\lambda}_i = \hat{\lambda}_i(\boldsymbol{W}_S^T\boldsymbol{W}_S)$ is the ith eigenvalue of $\boldsymbol{W}_S^T\boldsymbol{W}_S$. Here the scaling is using $g = 1$. For more information about the SVD, see Datta (1995, pp. 552-556) and Fogel et al. (2013).

Variable Selection and Post-Selection Inference:

There is massive literature on variable selection and a fairly large literature for inference after variable selection. See, for example, Bertsimas et al. (2016), Fan and Lv (2010), Ferrari and Yang (2015), Fithian et al. (2014), Hjort and Claeskins (2003), Knight and Fu (2000), Leeb and Pötscher (2005, 2006), Lockhart et al. (2014), Qi et al. (2015), and Tibshirani et al. (2016).

For post-selection inference, the methods in the literature are often for multiple linear regression assuming normality (an assumption that is too strong), or are asymptotically equivalent to using the full model, or find a quantity to test that is not $\boldsymbol{A}\boldsymbol{\beta}$. Typically the methods have not been shown to perform better than data splitting. See Ewald and Schneider (2018). Leeb et al. (2015) suggests that the Berk et al. (2013) method does not really work.

Kivaranovic and Leeb (2021) show that E(CI length) tends to be infinity for a method proposed by Lee et al. (2016). Also see Lu et al. (2017), and Tibshirani et al. (2016).

Warning: For $n < 5p$, validate sparse fitted models with response and residual plots. PIs can also help.

High Dimensional Testing and Confidence Intervals:

As of 2023, testing sparse fitted models with data splitting and the tests of Olive and Zhang (2023) appear to be backed by theory under reasonable regularity conditions. Assuming that $(Y_i, \boldsymbol{x}_i^T)^T$ are iid $N_{p+1}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is not a reasonable regularity conditions. For data splitting, forward selection with EBIC, lasso variable selection, and MMLE variable selection can be useful. Chetverikov, Liao and Chernozhukov (2022) show that k-fold CV with lasso often picks an MLR model good for prediction.

Also see Basa et al. (2022), Dezeure et al. (2015), Javanmard and Montanari (2014), Rinaldo, Wasserman, and G'Sell (2019), van de Geer et al. (2014), and Zhang and Cheng (2017). Fan and Lv (2010) gave large sample theory for some methods if $p = o(n^{1/5})$. The method of Ning and Liu (2017) needs a log likelihood.

Full OLS Model: A sufficient condition for $\hat{\boldsymbol{\beta}}_{OLS}$ to be a consistent estimator of $\boldsymbol{\beta}$ is $\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \mathbf{0}$ as $n \to \infty$. See Lai et al. (1979). For more OLS large sample theory, see Eicker (1963) and White (1984).

Forward Selection: See Olive and Hawkins (2005), Pelawa Watagoda and Olive (2021ab), and Rathnayake and Olive (2023).

The Oracle Property:

The oracle property says $P(I_{min} = S) \to 1$ as $n \to \infty$. A necessary condition for the oracle property is that S is in the search path with probability going to 1 as $n \to \infty$. For "fast methods" like lasso and forward selection, this requires the predictors to be nearly orthogonal. Hence *the regularity conditions for the oracle property are much too strong* if the predictors are moderately or highly correlated. The oracle property may be useful for wavelets and PCR. See Su (2018), Su, Bogdan, and Candés (2017), and Wieczorek and Lei (2022).

Principal Components Regression: Principal components are Karhunen Loeve directions of centered X. See Hastie et al. (2009, p. 66). A useful PCR paper is Cook and Forzani (2008).

Partial Least Squares: An important PLS paper is Wold (1975). Also see Wold (1985, 2006). Olive and Zhang (2023) showed $\hat{\boldsymbol{\beta}}_{OPLS}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\beta}_{OPLS}$ if the cases $(\boldsymbol{x}_i, Y_i)$ are iid with a few moments, p is fixed, and $n \to \infty$. Olive and Zhang (2023) also suggested that much of the theory for OPLS and PLS appears to be incorrect, except under regularity conditions that are much too strong. See, for example, Basa, et al. (2022), Cook et al. (2013), Cook (2018), Cook and Forzani (2018, 2019), Cook and Su (2016), and Chun and Keleş (2010). Denham (1997) suggested a PI for PLS that assumes the number of components is selected in advance.

Much of the PLS literature claims that if the cases are iid, then under mild conditions, $\hat{\boldsymbol{\beta}}_{OPLS}$, $\hat{\boldsymbol{\beta}}_{kPLS}$, and $\hat{\boldsymbol{\beta}}_{MSPLS}$ estimate $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. See for example, Basa et al. (2024) and Cook and Forzani (2024). However, they use a very strong regularity condition:

$$Y|\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T \boldsymbol{x} + e. \tag{2.57}$$

When $Y|\boldsymbol{x} = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$, then under mild regularity conditions, $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$. Hence regularity condition (2.46) and iid cases forces $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$. Thus regularity condition (2.46) forces $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ and $\boldsymbol{\beta}_{OLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ to be eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ if $\lambda \neq 0$. Hence $\boldsymbol{\beta}_{OLS}^T\boldsymbol{x}$ is equivalent (up to a positive constant multiplier) to the population principal component regression (PCR) component $\boldsymbol{\eta}_j^T\boldsymbol{x}$ that is most correlated with Y, where $\boldsymbol{\eta}_j$ is one of the eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$.

Ridge Regression: An important ridge regression paper is Hoerl and Kennard (1970). Also see Gruber (1998). Ridge regression is known as Tikhonov regularization in the numerical analysis literature.

Lasso: Lasso was introduced by Tibshirani (1996). Efron et al. (2004) and Tibshirani et al. (2012) are important papers. Su et al. (2017) note some problems with lasso. If n/p is large, see Knight and Fu (2000) for the residual bootstrap with OLS full model residuals. Camponovo (2015) suggested that the nonparametric bootstrap does not work for lasso. Chatterjee and Lahiri (2011) stated that the residual bootstrap with lasso does not work. Hall et al. (2009) stated that the residual bootstrap with OLS full model residuals does not work, but the m out of n residual bootstrap with OLS full model residuals does work. Rejchel (2016) gave a good review of lasso theory. Fan and Lv (2010) reviewed large sample theory for some alternative methods. See Lockhart et al. (2014) for a partial remedy for hypothesis testing with lasso. The Ning and Liu (2017) method needs a log likelihood. Knight and Fu (2000) gave theory for fixed p.

Regularity conditions for testing are strong. Often lasso tests assume that Y and the nontrivial predictors follow a multivariate normal (MVN) distribution. For the MVN distribution, the MLR model tends to be dense not sparse if n/p is small.

lasso variable selection:

Applying OLS on a constant and the k nontrivial predictors that have nonzero lasso $\hat{\eta}_i$ is called *lasso variable selection*. We want $n \geq 10(k+1)$. If $\lambda_1 = 0$, a variant of lasso variable selection computes the OLS submodel for the subset corresponding to λ_i for $i = 1, ..., M$. If C_p is used, then this variant has large sample theory given by Theorem 2.4.

Lasso can also be used for other estimators, such as generalized linear models (GLMs). Then lasso variable selection is the "classical estimator," such as a GLM, applied to the lasso active set. For prediction, lasso variable selection is often better than lasso, but sometimes lasso is better.

See Meinshausen (2007) for the relaxed lasso method with R package `relaxo` for MLR: apply lasso with penalty λ to get a subset of variables with nonzero coefficients. Then reduce the shrinkage of the nonzero elements by applying lasso again to the nonzero coefficients but with a smaller penalty ϕ. This two stage estimator could be used for other estimators. Lasso variable selection corresponds to the limit as $\phi \to 0$.

Dense Regression or Abundant Regression: occurs when most of the predictors contribute to the regression. Hence the regression is not sparse. See Cook et al. (2013).

Other Methods: Consider the MLR model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Let $\lambda \geq 0$ be a constant and let $q \geq 0$. The *estimator* $\hat{\boldsymbol{\eta}}_q$ minimizes the *criterion*

$$Q_q(\boldsymbol{b}) = \boldsymbol{r}(\boldsymbol{b})^T \boldsymbol{r}(\boldsymbol{b}) + \lambda \sum_{j=1}^{p-1} |b_i|^q, \tag{2.58}$$

over all vectors $\boldsymbol{b} \in \mathbb{R}^{p-1}$ where we take $0^0 = 0$. Then $q = 1$ corresponds to lasso and $q = 2$ corresponds to ridge regression. If $q = 0$, the penalty $\lambda \sum_{j=1}^{p-1} |b_i|^0 = \lambda k$ where k is the number of nonzero components of $\boldsymbol{b}$. Hence the $q = 0$ estimator is often called the "best subset" estimator. See Frank and Friedman (1993). For fixed p, large sample theory is given by Knight and Fu (2000). Following Hastie et al. (2009, p. 72), the optimization problem is convex if $q \geq 1$ and λ is fixed.

Suppose model I_k contains k predictors including a constant. For multiple linear regression, the forward selection algorithm in Chapter 4 adds a predictor x_{k+1}^* that minimizes the residual sum of squares, while the Pati et al. (1993) "orthogonal matching pursuit algorithm" uses predictors (scaled to have unit norm: $\boldsymbol{x}_i^T \boldsymbol{x}_i = 1$ for the nontrivial predictors), and adds the scaled predictor x_{k+1}^* that maximizes $|\boldsymbol{x}_{k+1}^{*T} \boldsymbol{r}_k|$ where the maximization is over variables not yet selected and the $\boldsymbol{r}_k$ are the OLS residuals from regressing Y on $\boldsymbol{X}_{I_k}^*$. Fan and Li (2001) and Candes and Tao (2007) gave competitors to lasso. Some fast methods seem similar to the first PLS component.

If $n \leq 400$ and $p \leq 3000$, Bertsimas et al. (2016) give a fast "all subsets" variable selection method. Lin et al. (2012) claim to have a very fast method for variable selection. Lee and Taylor (2014) suggest the marginal screening algorithm: let $\boldsymbol{W}$ be the matrix of standardized nontrivial predictors. Compute $\boldsymbol{W}^T \boldsymbol{Y} = (c_1, ..., c_{p-1})^T$ and select the J variables corresponding to the J largest $|c_i|$. These are the J standardized variables with the largest absolute correlations with Y. Then do an OLS regression of Y on these J variables and a constant. A slower algorithm somewhat similar but much slower than the Lin et al. (2012) algorithm follows. Let a constant x_1 be in the model, and let $\boldsymbol{W} = [\boldsymbol{a}_1, ..., \boldsymbol{a}_{p-1}]$ and $\boldsymbol{r} = \boldsymbol{Y} - \overline{Y}$. Compute $\boldsymbol{W}^T \boldsymbol{r}$ and let x_2^* correspond to the variable with the largest absolute entry. Remove the corresponding $\boldsymbol{a}_j$ from $\boldsymbol{W}$ to get $\boldsymbol{W}_1$. Let $\boldsymbol{r}_1$ be the OLS residuals from regressing Y on x_1 and x_2^*. Compute $\boldsymbol{W}^T \boldsymbol{r}_1$ and let x_3^* correspond to the variable with the

largest absolute entry. Continue in this manner to get $x_1, x_2^*, ..., x_J^*$ where $J = min(p, \lceil n/5 \rceil)$. Like forward selection, evaluate the $J-1$ models I_j containing the first j predictors $x_1, x_2^*, ..., x_J^*$ for $j = 2, ..., J$ with a criterion such as C_p.

Following Sun and Zhang (2012), let (2.6) hold and let $Q(\boldsymbol{\eta}) = \frac{1}{2n}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \lambda^2 \sum_{i=1}^{p-1} \rho\left(\frac{|\eta_i|}{\lambda}\right)$ where ρ is scaled such that the derivative $\rho'(0+) = 1$. As for lasso and elastic net, let $s_j = sgn(\hat{\eta}_j)$ where $s_j \in [-1, 1]$ if $\hat{\eta}_j = 0$. Let $\rho_j' = \rho'(|\hat{\eta}_j|/\lambda)$ if $\hat{\eta}_j \neq 0$, and $\rho_j' = 1$ if $\hat{\eta}_j = 0$. Then $\hat{\boldsymbol{\eta}}$ is a critical point of $Q(\boldsymbol{\eta})$ iff $\boldsymbol{w}_j^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}) = n\lambda s_j \rho_j'$ for $j = 1, ..., n$. If ρ is convex, then these conditions are the KKT conditions. Let $d_j = s_j \rho_j'$. Then $\boldsymbol{W}^T\boldsymbol{Z} - \boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}} = n\lambda \boldsymbol{d}$, and $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_{OLS} - n\lambda(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{d}$. If the d_j are bounded, then $\hat{\boldsymbol{\eta}}$ is consistent if $\lambda \to 0$ as $n \to \infty$, and $\hat{\boldsymbol{\eta}}$ is asymptotically equivalent to $\hat{\boldsymbol{\eta}}_{OLS}$ if $n^{1/2}\lambda \to 0$. Note that $\rho(t) = t$ for $t > 0$ gives lasso with $\lambda = \lambda_{1,n}/(2n)$.

Gao and Huang (2010) give theory for a LAD–lasso estimator, and Qi et al. (2015) is an interesting lasso competitor.

Multivariate linear regression has $m \geq 2$ response variables. See Olive (2017ab: ch. 12). PLS also works if $m \geq 1$, and methods like ridge regression and lasso can also be extended to multivariate linear regression. See, for example, Haitovsky (1987) and Obozinski et al. (2011). Sparse envelope models are given in Su et al. (2016).

Model Building:

When the entire data set is used to build a model with the response variable, the inference tends to be invalid, and cross validation should not be used to check the model. See Hastie et al. (2009, p. 245). In order for the inference and cross validation to be useful, the response variable and the predictors for the regression should be chosen before looking at the response variable. Predictor transformations can be done as long as the response variable is not used to choose the transformation. You can do model building on the test set, and then inference for the chosen (built) model as the full model with the validation set, provided this model follows the regression model used for inference (e.g. multiple linear regression or a GLM). This process is difficult to simulate.

AIC and BIC Type Criterion:

Olive and Hawkins (2005) and Burnham and Anderson (2004) are useful reference when p is fixed. Some interesting theory for AIC appears in Zhang (1992). Zheng and Loh (1995) show that BIC_S can work if $p = p_n = o(\log(n))$ and there is a consistent estimator of σ^2. For the C_p criterion, see Jones (1946) and Mallows (1973).

AIC and BIC type criterion and variable selection for high dimensional regression are discussed in Chen and Chen (2008), Fan and Lv (2010), Fujikoshi et al. (2014), and Luo and Chen (2013). Wang (2009) suggests using

$$WBIC(I) = \log[SSE(I)/n] + n^{-1}|I|[\log(n) + 2\log(p)].$$

See Bogdan et al. (2004), Cho and Fryzlewicz (2012), and Kim et al. (2012). Luo and Chen (2013) state that $WBIC(I)$ needs $p/n^a < 1$ for some $0 < a < 1$.

If n/p is large and one of the models being considered is the true model S (shown to occur with probability going to one only under very strong assumptions by Wieczorek and Lei (2021)), then BIC tends to outperform AIC. If none of the models being considered is the true model, then AIC tends to outperform BIC. See Yang (2003).

Robust Versions: Hastie et al. (2015, pp. 26-27) discuss some modifications of lasso that are robust to certain types of outliers. Robust methods for forward selection and LARS are given by Uraibi et al. (2017, 2019) that need $n >> p$. If n is not much larger than p, then Hoffman et al. (2015) have a robust Partial Least Squares–Lasso type estimator that uses a clever weighting scheme.

A simple method to make an MLR method robust to certain types of outliers is to find the *covmb2* set B of Chapter 1 applied to the quantitative predictors. Then use the MLR method (such as elastic net, lasso, PLS, PCR, ridge regression, or forward selection) applied to the cases corresponding to the $\boldsymbol{x}_j$ in B. Make a response and residual plot, based on the robust estimator $\hat{\boldsymbol{\beta}}_B$, using all n cases.

Prediction Intervals:

Lei et al. (2018) and Wasserman (2014) suggested prediction intervals for estimators such as lasso. The method has interesting theory if the $(\boldsymbol{x}_i, Y_i)$ are iid from some population. Also see Butler and Rothman (1980) and Steinberger and Leeb (2023).

Let p be fixed, d be for PI (2.14), and $n \to \infty$. For elastic net, forward selection, PCR, PLS, ridge regression, lasso variable selection, and lasso, if $P(d \to p) \to 1$ as $n \to \infty$ then the seven methods are asymptotically equivalent to the OLS full model, and the PI (2.14) is asymptotically optimal on a large class of iid unimodal zero mean error distributions. The asymptotic optimality holds since the sample quantile of the OLS full model residuals are consistent estimators of the population quantiles of the unimodal error distribution for a large class of distributions. Note that $d \xrightarrow{P} p$ if $P(\hat{\lambda}_{1n} \to 0) \to 1$ for elastic net, lasso, and ridge regression, and $d \xrightarrow{P} p$ if the number $d-1$ of components ($\boldsymbol{\gamma}_j^T \boldsymbol{x}$ or $\boldsymbol{\gamma}_j^T \boldsymbol{w}$) used by the method satisfies $P(d-1 \to p-1) \to 1$. Consistent estimators $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ also produce residuals such that the sample quantiles of the residuals are consistent estimators of quantiles of the error distribution. See Remark 2.21, Olive and Hawkins (2003), and Rousseeuw and Leroy (1987, p. 128).

Degrees of Freedom:

A formula for the model degrees of freedom *df* tend to be given for a model when there is no model selection or variable selection. For many estimators,

the degrees of freedom is not known if model selection is used. A d for PI (2.14) is often obtained by plugging in the degrees of freedom formula as if model selection did not occur. Then the resulting d is rarely an actual degrees of freedom. As an example, if $\hat{\boldsymbol{Y}} = \boldsymbol{H}_\lambda \boldsymbol{Y}$, then often $df = trace(\boldsymbol{H}_\lambda)$ if λ is selected before examining the data. If model selection is used to pick $\hat{\lambda}$, then $d = trace(\boldsymbol{H}_{\hat{\lambda}})$ is not the model degrees of freedom.

2.20 Problems

2.1. For ridge regression, suppose $\boldsymbol{V} = \boldsymbol{\rho}_{\boldsymbol{u}}^{-1}$. Show that if p/n and $\lambda/n = \lambda_{1,n}/n$ are both small, then

$$\hat{\boldsymbol{\eta}}_R \approx \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda}{n}\boldsymbol{V}\hat{\boldsymbol{\eta}}_{OLS}.$$

2.2. Consider choosing $\hat{\boldsymbol{\eta}}$ to minimize the criterion

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta})^T(\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i|^j$$

where $\lambda_{1,n} \geq 0$, $a > 0$, and $j > 0$ are known constants. Consider the regression methods OLS, forward selection, lasso, PLS, PCR, ridge regression, and lasso variable selection.
a) Which method corresponds to $j = 1$?
b) Which method corresponds to $j = 2$?
c) Which method corresponds to $\lambda_{1,n} = 0$?

2.3. a) For ridge regression, let $\boldsymbol{A}_n = (\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}$ and $\boldsymbol{B}_n = [\boldsymbol{I}_p - \lambda_{1,n}(\boldsymbol{X}^T\boldsymbol{X} + \lambda_{1,n}\boldsymbol{I}_p)^{-1}]$. Show $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

b) For ridge regression, let $\boldsymbol{A}_n = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}$ and $\boldsymbol{B}_n = [\boldsymbol{I}_{p-1} - \lambda_{1,n}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}]$. Show $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

2.4. Suppose $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ where $\boldsymbol{H}$ is an $n \times n$ hat matrix. Then the degrees of freedom $df(\hat{\boldsymbol{Y}}) = tr(\boldsymbol{H}) =$ sum of the diagonal elements of $\boldsymbol{H}$. An estimator with low degrees of freedom is inflexible while an estimator with high degrees of freedom is flexible. If the degrees of freedom is too low, the estimator tends to underfit while if the degrees of freedom is to high, the estimator tends to overfit.

a) Find $df(\hat{\boldsymbol{Y}})$ if $\hat{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$ which uses $\boldsymbol{H} = (h_{ij})$ where $h_{ij} \equiv 1/n$ for all i and j. This inflexible estimator uses the sample mean $\overline{Y}$ of the response variable as $\hat{Y}_i$ for $i = 1, ..., n$.

b) Find $df(\hat{\boldsymbol{Y}})$ if $\hat{\boldsymbol{Y}} = \boldsymbol{Y} = \boldsymbol{I}_n\boldsymbol{Y}$ which uses $\boldsymbol{H} = \boldsymbol{I}_n$ where $h_{ii} = 1$. This bad flexible estimator interpolates the response variable.

2.5. Suppose $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$, $\hat{\boldsymbol{Z}} = \boldsymbol{W}\hat{\boldsymbol{\eta}}$, $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$, and $\hat{\boldsymbol{Y}} = \hat{\boldsymbol{Z}} + \overline{\boldsymbol{Y}}$. Let the $n \times p$ matrix $\boldsymbol{W}_1 = [\boldsymbol{1} \quad \boldsymbol{W}]$ and the $p \times 1$ vector $\hat{\boldsymbol{\eta}}_1 = (\overline{Y} \quad \hat{\boldsymbol{\eta}}^T)^T$ where the scalar $\overline{Y}$ is the sample mean of the response variable. Show $\hat{\boldsymbol{Y}} = \boldsymbol{W}_1\hat{\boldsymbol{\eta}}_1$.

2.6. Let $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{G} = (G_{ij})$. For $j = 1, ..., p-1$, let G_{ij} denote the $(j+1)$th variable standardized so that $\sum_{i=1}^n G_{ij} = 0$ and $\sum_{i=1}^n G_{ij}^2 = 1$. Note that the sample correlation matrix of the nontrivial predictors $\boldsymbol{u}_i$ is $\boldsymbol{R}_{\boldsymbol{u}} = \boldsymbol{G}^T\boldsymbol{G}$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{G}\boldsymbol{\eta} + \boldsymbol{e} \tag{2.59}$$

where the vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$. The standardization differs from that used for earlier regression models since $\sum_{i=1}^n G_{ij}^2 = 1 \neq n = \sum_{i=1}^n W_{ij}^2$. Note that

$$\boldsymbol{G} = \frac{1}{\sqrt{n}}\boldsymbol{W}.$$

Following Zou and Hastie (2005), the *naive elastic net* $\hat{\boldsymbol{\eta}}_N$ estimator is the minimizer of

$$Q_N(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_2^*\|\boldsymbol{\eta}\|_2^2 + \lambda_1^*\|\boldsymbol{\eta}\|_1 \tag{2.60}$$

where $\lambda_i^* \geq 0$. The term "naive" is used because the elastic net estimator is better. Let $\tau = \dfrac{\lambda_2^*}{\lambda_1^* + \lambda_2^*}, \gamma = \dfrac{\lambda_1^*}{\sqrt{1+\lambda_2^*}}$, and $\boldsymbol{\eta}_A = \sqrt{1+\lambda_2^*}\ \ \boldsymbol{\eta}$. Let the $(n+p-1) \times (p-1)$ augmented matrix $\boldsymbol{G}_A$ and the $(n+p-1) \times 1$ augmented response vector $\boldsymbol{Z}_A$ be defined by

$$\boldsymbol{G}_A = \begin{pmatrix} \boldsymbol{G} \\ \sqrt{\lambda_2^*}\ \boldsymbol{I}_{p-1} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{Z}_{\mathrm{A}} = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{0} \end{pmatrix},$$

where $\boldsymbol{0}$ is the $(p-1) \times 1$ zero vector. Let $\hat{\boldsymbol{\eta}}_A = \sqrt{1+\lambda_2^*}\ \ \hat{\boldsymbol{\eta}}$ be obtained from the lasso of $\boldsymbol{Z}_A$ on $\boldsymbol{G}_A$: that is $\hat{\boldsymbol{\eta}}_A$ minimizes

$$Q_N(\boldsymbol{\eta}_A) = \|\boldsymbol{Z}_A - \boldsymbol{G}_A\boldsymbol{\eta}_A\|_2^2 + \gamma\|\boldsymbol{\eta}_A\|_1 = Q_N(\boldsymbol{\eta}).$$

Prove $Q_N(\boldsymbol{\eta}_A) = Q_N(\boldsymbol{\eta})$.
(Then

$$\hat{\boldsymbol{\eta}}_N = \frac{1}{\sqrt{1+\lambda_2^*}}\hat{\boldsymbol{\eta}}_A \text{ and } \hat{\boldsymbol{\eta}}_{EN} = \sqrt{1+\lambda_2^*}\ \hat{\boldsymbol{\eta}}_A = (1+\lambda_2^*)\hat{\boldsymbol{\eta}}_N.$$

The above elastic net estimator minimizes the criterion

$$Q_G(\boldsymbol{\eta}) = \frac{\boldsymbol{\eta}^T \boldsymbol{G}^T \boldsymbol{G}\boldsymbol{\eta}}{1+\lambda_2^*} - 2\boldsymbol{Z}^T\boldsymbol{G}\boldsymbol{\eta} + \frac{\lambda_2^*}{1+\lambda_2^*}\|\boldsymbol{\eta}\|_2^2 + \lambda_1^*\|\boldsymbol{\eta}\|_1,$$

and hence is not the elastic net estimator corresponding to Equation (3.22).)

2.7. Let $\boldsymbol{\beta} = (\beta_1, \boldsymbol{\beta}_S^T)^T$. Consider choosing $\hat{\boldsymbol{\beta}}$ to minimize the criterion

$$Q(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}_S\|_2^2 + \lambda_2\|\boldsymbol{\beta}_S\|_1$$

where $\lambda_i \geq 0$ for $i = 1, 2$.
a) Which values of λ_1 and λ_2 correspond to ridge regression?
b) Which values of λ_1 and λ_2 correspond to lasso?
c) Which values of λ_1 and λ_2 correspond to elastic net?
d) Which values of λ_1 and λ_2 correspond to the OLS full model?

2.8. For the output below, an asterisk means the variable is in the model. All models have a constant, so model 1 contains a constant and mmen.

a) List the variables, including a constant, that models 2, 3, and 4 contain.

b) The term out$cp lists the C_p criterion. Which model (1, 2, 3, or 4) is the minimum C_p model I_{min}?

c) Suppose $\hat{\boldsymbol{\beta}}_{I_{min}} = (241.5445, 1.001)^T$. What is $\hat{\boldsymbol{\beta}}_{I_{min},0}$?

```
Selection Algorithm: forward #output for Problem 3.8
          pop mmen mmilmen milwmn
1   ( 1 ) " "  "*"   " "      " "
2   ( 1 ) " "  "*"   "*"      " "
3   ( 1 ) "*"  "*"   "*"      " "
4   ( 1 ) "*"  "*"   "*"      "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
```

2.9. Tremearne (1911) presents a data set of about 17 measurements on 112 people of Hausa nationality. We used $Y =$ *height*. Along with a constant $x_{i,1} \equiv 1$, the five additional predictor variables used were $x_{i,2} =$ *height when sitting*, $x_{i,3} =$ *height when kneeling*, $x_{i,4} =$ *head length*, $x_{i,5} =$ *nasal breadth*, and $x_{i,6} =$ *span* (perhaps from left hand to right hand). The output below is for the OLS full model.

```
            Estimate Std.Err 95% shorth CI
Intercept -77.0042 65.2956 [-208.864,55.051]
X2          0.0156  0.0992 [-0.177,  0.217]
X3          1.1553  0.0832 [ 0.983,  1.312]
X4          0.2186  0.3180 [-0.378,  0.805]
X5          0.2660  0.6615 [-1.038,  1.637]
X6          0.1396  0.0385 [0.0575,  0.217]
```

a) Give the shorth 95% CI for β_2.

b) Compute the standard 95% CI for β_2.

c) Which variables, if any, are needed in the MLR model given that the other variables are in the model?

Now we use forward selection and I_{min} is the minimum C_p model.

```
           Estimate Std.Err 95% shorth CI
Intercept -42.4846 51.2863 [-192.281, 52.492]
X2          0               [   0.000,  0.268]
X3          1.1707  0.0598  [   0.992,  1.289]
X4          0               [   0.000,  0.840]
X5          0               [   0.000,  1.916]
X6          0.1467  0.0368  [  0.0747,  0.215]
  (Intercept)     a    b     c     d     e
1        TRUE FALSE TRUE FALSE FALSE FALSE
2        TRUE FALSE TRUE FALSE FALSE  TRUE
3        TRUE FALSE TRUE  TRUE FALSE  TRUE
4        TRUE FALSE TRUE  TRUE  TRUE  TRUE
5        TRUE  TRUE TRUE  TRUE  TRUE  TRUE
> tem2$cp
[1] 14.389492  0.792566  2.189839  4.024738  6.000000
```

d) What is the value of $C_p(I_{min})$ and what is $\hat{\boldsymbol{\beta}}_{I_{min},0}$?

e) Which variables, if any, are needed in the MLR model given that the other variables are in the model?

f) List the variables, including a constant, that model 3 contains.

2.10. Table 2.7 below shows simulation results for bootstrapping OLS (reg) and forward selection (vs) with C_p when $\boldsymbol{\beta} = (1, 1, 0, 0, 0)^T$. The β_i columns give coverage = the proportion of CIs that contained β_i and the average length of the CI. The test is for $H_0 : (\beta_3, \beta_4, \beta_5)^T = \mathbf{0}$ and H_0 is true. The "coverage" is the proportion of times the prediction region method bootstrap test failed to reject H_0. Since 1000 runs were used, a cov in [0.93,0.97] is reasonable for a nominal value of 0.95. Output is given for three different error distributions. If the coverage for both methods ≥ 0.93, the method with the shorter average CI length was more precise. (If one method had coverage ≥ 0.93 and the other had coverage < 0.93, we will say the method with coverage ≥ 0.93 was more precise.)

a) For β_3, β_4, and β_5, which method, forward selection or the OLS full model, was more precise?

b) The test "length" is the average length of the interval $[0, D_{(U_B)}] = D_{(U_B)}$ where the test fails to reject H_0 if $D_{\mathbf{0}} \leq D_{(U_B)}$. The OLS full model is asymptotically normal, and hence for large enough n and B the reg len row for the test column should be near $\sqrt{\chi^2_{3,0.95}} = 2.795$.

Were the three values in the test column for reg within 0.1 of 2.795?

2.11. Suppose the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, and the regression method fits $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$. Suppose $\hat{Z} = 245.63$ and $\overline{Y} = 105.37$. What is $\hat{Y}$?

Table 2.7 Bootstrapping Forward Selection, $n = 100, p = 5, \psi = 0, B = 1000$

	β_1	β_2	β_3	β_4	β_5	test
reg cov	0.95	0.93	0.93	0.93	0.94	0.93
len	0.658	0.672	0.673	0.674	0.674	2.861
vs cov	0.95	0.94	0.998	0.998	0.999	0.993
len	0.661	0.679	0.546	0.548	0.544	3.11
reg cov	0.96	0.93	0.94	0.96	0.93	0.94
len	0.229	0.230	0.229	0.231	0.230	2.787
vs cov	0.95	0.94	0.999	0.997	0.999	0.995
len	0.228	0.229	0.185	0.187	0.186	3.056
reg cov	0.94	0.94	0.95	0.94	0.94	0.93
len	0.393	0.398	0.399	0.399	0.398	2.839
vs cov	0.94	0.95	0.997	0.997	0.996	0.990
len	0.392	0.400	0.320	0.322	0.321	3.077

2.12. To get a large sample 90% PI for a future value Y_f of the response variable, find a large sample 90% PI for a future residual and add $\hat{Y}_f$ to the endpoints of the of that PI. Suppose forward selection is used and the large sample 90% PI for a future residual is $[-778.28, 1336.44]$. What is the large sample 90% PI for Y_f if $\hat{\boldsymbol{\beta}}_{I_{min}} = (241.545, 1.001)^T$ used a constant and the predictor *mmen* with corresponding $\boldsymbol{x}_{I_{min},f} = (1, 75000)^T$?

2.13. Table 2.8 below shows simulation results for bootstrapping OLS (reg), lasso, and ridge regression (RR) with 10-fold CV when $\boldsymbol{\beta} = (1, 1, 0, 0)^T$. The β_i columns give coverage = the proportion of CIs that contained β_i and the average length of the CI. The test is for $H_0 : (\beta_3, \beta_4)^T = \mathbf{0}$ and H_0 is true. The "coverage" is the proportion of times the prediction region method bootstrap test failed to reject H_0. OLS used 1000 runs while 100 runs were used for lasso and ridge regression. Since 100 runs were used, a cov in [0.89, 1] is reasonable for a nominal value of 0.95. If the coverage for both methods ≥ 0.89, the method with the shorter average CI length was more precise. (If one method had coverage ≥ 0.89 and the other had coverage < 0.89, we will say the method with coverage ≥ 0.89 was more precise.) The results for the lasso test were omitted since sometimes $\boldsymbol{S}_T^*$ was singular. (Lengths for the test column are not comparable unless the statistics have the same asymptotic distribution.)

a) For β_3 and β_4 which method, ridge regression or the OLS full model, was better?

b) For β_3 and β_4 which method, lasso or the OLS full model, was more precise?

2.14. Suppose $n = 15$ and 5-fold CV is used. Suppose observations are measured for the following people. Use the output below to determine which people are in the first fold.

```
folds: 4  3  4  2  1  4  3  5  2  2  3  1  5  5  1
```

Table 2.8 Bootstrapping lasso and RR, $n = 100, \psi = 0.9, p = 4, B = 250$

		β_1	β_2	β_3	β_4	test
reg	cov	0.942	0.951	0.949	0.943	0.943
	len	0.658	5.447	5.444	5.438	2.490
RR	cov	0.97	0.02	0.11	0.10	0.05
	len	0.681	0.329	0.334	0.334	2.546
reg	cov	0.947	0.955	0.950	0.951	0.952
	len	0.658	5.511	5.497	5.500	2.491
lasso	cov	0.93	0.91	0.92	0.99	
	len	0.698	3.765	3.922	3.803	

1) Athapattu, 2) Azizi, 3) Cralley 4) Gallage, 5) Godbold, 6) Gunawardana, 7) Houmadi, 8) Mahappu, 9) Pathiravasan, 10) Rajapaksha, 11) Ranaweera, 12) Safari, 13) Senarathna, 14) Thakur, 15) Ziedzor

2.15. Table 2.9 below shows simulation results for a large sample 95% prediction interval. Since 5000 runs were used, a cov in [0.94, 0.96] is reasonable for a nominal value of 0.95. If the coverage for a method ≥ 0.94, the method with the shorter average PI length was more precise. Ignore methods with cov < 0.94. The MLR model had $\boldsymbol{\beta} = (1, 1, ..., 1, 0, ..., 0)^T$ where the first $k+1$ coefficients were equal to 1. If $\psi = 0$ then the nontrivial predictors were uncorrelated, but highly correlated if $\psi = 0.9$.

Table 2.9 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0, 1)$

n	p	ψ	k		FS	lasso	RL	RR	PLS	PCR
100	40	0	1	cov	0.9654	0.9774	0.9588	0.9274	0.8810	0.9882
				len	4.4294	4.8889	4.6226	4.4291	4.0202	7.3393
400	400	0.9	19	cov	0.9348	0.9636	0.9556	0.9632	0.9462	0.9478
				len	4.3687	47.361	4.8530	48.021	4.2914	4.4764

a) Which method was most precise, given cov ≥ 0.94, when $n = 100$?
b) Which method was most precise, given cov ≥ 0.94, when $n = 400$?

2.16. When doing a PI or CI simulation for a nominal $100(1-\delta)\% = 95\%$ interval, there are m runs. For each run, a data set and interval are generated, and for the ith run $Y_i = 1$ if μ or Y_f is in the interval, and $Y_i = 0$, otherwise. Hence the Y_i are iid Bernoulli$(1 - \delta_n)$ random variables where $1 - \delta_n$ is the true probability (true coverage) that the interval will contain μ or Y_f. The observed coverage (= coverage) in the simulation is $\overline{Y} = \sum_i Y_i/m$. The variance $V(\overline{Y}) = \sigma^2/m$ where $\sigma^2 = (1 - \delta_n)\delta_n \approx (1 - \delta)\delta \approx (0.95)0.05$ if $\delta_n \approx \delta = 0.05$. Hence

$$SD(\overline{Y}) \approx \sqrt{\frac{0.95(0.05)}{m}}.$$

If the (observed) coverage is within $0.95 \pm kSD(\overline{Y})$ the integer k is near 3, then there is no reason to doubt that the actual coverage $1 - \delta_n$ differs from the nominal coverage $1-\delta = 0.95$ if $m \geq 1000$ (and as a crude benchmark, for $m \geq 100$). In the simulation, the length of each interval is computed, and the average length is computed. For intervals with coverage $\geq 0.95 - kSD(\overline{Y})$, intervals with shorter average length are better (have more precision).

a) If $m = 5000$ what is 3 SD$(\overline{Y})$, using the above approximation? Your answer should be close to 0.01.

b) If $m = 1000$ what is 3 SD$(\overline{Y})$, using the above approximation?

R Problem

Use the command *source("G:/slpack.txt")* **to download the functions** and the command *source("G:/sldata.txt")* **to download the data. See Preface or Section 11.1.** Typing the name of the `slpack` function, e.g. *vsbootsim3*, will display the code for the function. Use the `args` command, e.g. *args(vsbootsim3)*, to display the needed arguments for the function. For the following problem, the R command can be copied and pasted from (http://parker.ad.siu.edu/Olive/slrhw.txt) into R.

2.17. The R program generates data satisfying the MLR model

$$Y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4)^T = (1, 1, 0, 0)$.

a) Copy and paste the commands for this part into R. The output gives $\hat{\boldsymbol{\beta}}_{OLS}$ for the OLS full model. Give $\hat{\boldsymbol{\beta}}_{OLS}$. Is $\hat{\boldsymbol{\beta}}_{OLS}$ close to $\boldsymbol{\beta} = 1, 1, 0, 0)^T$?

b) The commands for this part bootstrap the OLS full model using the residual bootstrap. Copy and paste the output into *Word*. The output shows $T_j^* = \hat{\boldsymbol{\beta}}_j^*$ for $j = 1, ..., 5$.

c) $B = 1000$ T_j^* were generated. The commands for this part compute the sample mean $\overline{T}^*$ of the T_j^*. Copy and paste the output into *Word*. Is $\overline{T}^*$ close to $\hat{\boldsymbol{\beta}}_{OLS}$ found in a)?

d) The commands for this part bootstrap the forward selection using the residual bootstrap. Copy and paste the output into *Word*. The output shows $T_j^* = \hat{\boldsymbol{\beta}}^*_{I_{min},0,j}$ for $j = 1, ..., 5$. The last two variables may have a few 0s.

e) $B = 1000$ T_j^* were generated. The commands for this part compute the sample mean $\overline{T}^*$ of the T_j^* where T_j^* is as in d). Copy and paste the output into *Word*. Is $\overline{T}^*$ close to $\boldsymbol{\beta} = (1, 1, 0, 0)$?

2.18. This simulation is similar to that used to form Table 2.2, but 1000 runs are used so coverage in [0.93,0.97] suggests that the actual coverage is close to the nominal coverage of 0.95.

The model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e$ where $\boldsymbol{\beta}_S = (\beta_1, \beta_2, ..., \beta_{k+1})^T = (\beta_1, \beta_2)^T$ and $k = 1$ is the number of active nontrivial predictors in the population model. The output for *test* tests $H_0 : (\beta_{k+2}, ..., \beta_p)^T = (\beta_3, ..., \beta_p)^T = \mathbf{0}$

and H_0 is true. The output gives the proportion of times the prediction region method bootstrap test fails to reject H_0. The nominal proportion is 0.95.

After getting your output, make a table similar to Table 2.2 with 4 lines. If your $p = 5$ then you need to add a column for β_5. Two lines are for reg (the OLS full model) and two lines are for vs (forward selection with I_{min}). The β_i columns give the coverage and lengths of the 95% CIs for β_i. If the coverage ≥ 0.93, then the shorter CI length is more precise. Were the CIs for forward selection more precise than the CIs for the OLS full model for β_3 and β_4?

To get the output, copy and paste the source commands from (http://parker.ad.siu.edu/Olive/slrhw.txt) into *R*. Copy and past the library command for this problem into *R*.

If you are person j then copy and paste the *R* code for person j for this problem into *R*.

2.19. This problem is like Problem 3.19, but ridge regression is used instead of forward selection. This simulation is similar to that used to form Table 2.2, but 100 runs are used so coverage in [0.89,1.0] suggests that the actual coverage is close to the nominal coverage of 0.95.

The model is $Y = \boldsymbol{x}^T\boldsymbol{\beta} + e = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + e$ where $\boldsymbol{\beta}_S = (\beta_1, \beta_2, ..., \beta_{k+1})^T = (\beta_1, \beta_2)^T$ and $k = 1$ is the number of active nontrivial predictors in the population model. The output for *test* tests $H_0 : (\beta_{k+2}, ..., \beta_p)^T = (\beta_3, ..., \beta_p)^T = \mathbf{0}$ and H_0 is true. The output gives the proportion of times the prediction region method bootstrap test fails to reject H_0. The nominal proportion is 0.95.

After getting your output, make a table similar to Table 2.2 with 4 lines. If your $p = 5$ then you need to add a column for β_5. Two lines are for reg (the OLS full model) and two lines are for ridge regression (with 10 fold CV). The β_i columns give the coverage and lengths of the 95% CIs for β_i. If the coverage ≥ 0.89, then the shorter CI length is more precise. Were the CIs for ridge regression more precise than the CIs for the OLS full model for β_3 and β_4?

To get the output, copy and paste the source commands from (http://parker.ad.siu.edu/Olive/slrhw.txt) into *R*. Copy and past the library command for this problem into *R*.

If you are person j then copy and paste the *R* code for person j for this problem into *R*.

2.20. This is like Problem 2.19, except lasso is used. If you are person j in Problem 2.19, then copy and paste the *R* code for person j for this problem into *R*. Make a table with 4 lines: two for OLS and 2 for lasso. Were the CIs for lasso more precise than the CIs for the OLS full model for β_3 and β_4?