Chapter 5
Statistical Learning Alternatives to OLS

This chapter considers several alternatives to OLS for the multiple linear regression model. Large sample theory is give for p fixed, but the prediction intervals can have $p > n$.

5.1 The MLR Model

From Definition 1.9, the multiple linear regression (MLR) model is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + e_i \tag{5.1}$$

for $i = 1, ..., n$. This model is also called the **full model**. Here n is the sample size and the random variable e_i is the ith error. Assume that the e_i are iid with 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.

There are many methods for estimating $\boldsymbol{\beta}$, including (ordinary) least squares (OLS) for the full model, forward selection with OLS, 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 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

$$\frac{1}{n-g}\sum_{i=1}^{n} W_i^2 = 1. \tag{5.2}$$

Remark 5.1. Let the nontrivial predictors $\boldsymbol{u}_i^T = (x_{i,2}, ..., x_{i,p}) = (u_{i,1}, ..., u_{i,p-1})$. Then $\boldsymbol{x}_i = (1, \boldsymbol{u}_i^T)^T$. Let the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W}_g = (W_{ij})$ when the predictors are standardized using $\hat{\sigma}_g$. Thus, $\sum_{i=1}^n W_{ij} = 0$ and $\sum_{i=1}^n W_{ij}^2 = n - g$ for $j = 1, ..., p-1$. Hence

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

is $\hat{\sigma}_g$ for the $(j+1)$th variable x_{j+1}. Let $\boldsymbol{w}_i^T = (w_{i,1}, ..., w_{i,p-1})$ be the standardized vector of nontrivial predictors for the ith case. Since the standardized data are also centered, $\overline{\boldsymbol{w}} = \boldsymbol{0}$. Then the sample covariance matrix of the $\boldsymbol{w}_i$ is the sample correlation matrix of the $\boldsymbol{u}_i$:

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

where r_{ij} is the sample correlation of $u_i = x_{i+1}$ and $u_j = x_{j+1}$. Thus the sample correlation matrix $\boldsymbol{R}_{\boldsymbol{u}}$ does not depend on g. Let $\boldsymbol{Z} = \boldsymbol{Y} - \overline{\boldsymbol{Y}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$. Since the *R* software tends to use $g = 0$, let $\boldsymbol{W} = \boldsymbol{W}_0$. Note that $n \times (p-1)$ matrix $\boldsymbol{W}$ does not include a vector $\boldsymbol{1}$ of ones. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{5.3}$$

where $\boldsymbol{Z} = (Z_1, ..., Z_n)^T$ and $\boldsymbol{\eta} = (\eta_1, ..., \eta_{p-1})^T$. The vector of fitted values $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$.

Remark 5.2. i) Interest is in model (5.1): 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 $(\hat{\boldsymbol{Y}}, \hat{\boldsymbol{\beta}})$. Also, see Remark 7.7. Equation (5.3) is a commonly used method for achieving this goal. Suppose $g = 0$. The method of moments estimator of the variance σ_w^2 is

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

When data x_i are standardized to have $\overline{w} = 0$ and $S^2_M = 1$, 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 (5.3) are used. Linear combinations of the $\boldsymbol{w}_i$ are linear combinations of the $\boldsymbol{u}_i$, which are linear combinations of the $\boldsymbol{x}_i$. (Note that $\boldsymbol{\gamma}^T\boldsymbol{u} = (0\ \boldsymbol{\gamma}^T)\ \boldsymbol{x}$.) 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.

Remark 5.3. Let $\hat{\sigma}_j$ be the sample standard deviation of variable x_j (often with $g = 0$) for $j = 2, ..., p$. Let $\hat{Y}_i = \hat{\beta}_1 + x_{i,2}\hat{\beta}_2 + \cdots + x_{i,p}\hat{\beta}_p = \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}$. If standardized nontrivial predictors are used, then

$$\hat{Y}_i = \hat{\gamma} + w_{i,1}\hat{\eta}_1 + \cdots + w_{i,p-1}\hat{\eta}_{p-1} = \hat{\gamma} + \frac{x_{i,2} - \overline{x}_2}{\hat{\sigma}_2}\hat{\eta}_1 + \cdots + \frac{x_{i,p} - \overline{x}_p}{\hat{\sigma}_p}\hat{\eta}_{p-1}$$

$$= \hat{\gamma} + \boldsymbol{w}_i^T\hat{\boldsymbol{\eta}} = \hat{\gamma} + \hat{Z}_i \tag{5.4}$$

where

$$\hat{\eta}_j = \hat{\sigma}_{j+1}\hat{\beta}_{j+1} \tag{5.5}$$

for $j = 1, ..., p-1$. Often $\hat{\gamma} = \overline{Y}$ so that $\hat{Y}_i = \overline{Y}$ if $x_{i,j} = \overline{x}_j$ for $j = 2, ..., p$. Then $\hat{\boldsymbol{Y}} = \overline{\boldsymbol{Y}} + \hat{\boldsymbol{Z}}$ where $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$. Note that

$$\hat{\gamma} = \hat{\beta}_1 + \frac{\overline{x}_2}{\hat{\sigma}_2}\hat{\eta}_1 + \cdots + \frac{\overline{x}_p}{\hat{\sigma}_p}\hat{\eta}_{p-1}.$$

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.

Most regression methods attempt to find an estimate $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ which minimizes some criterion function $Q(\boldsymbol{b})$ of the residuals. As in Definition 1.13, given an estimate $\boldsymbol{b}$ of $\boldsymbol{\beta}$, the corresponding vector of *fitted values* is $\widehat{\boldsymbol{Y}} \equiv \widehat{\boldsymbol{Y}}(\boldsymbol{b}) = \boldsymbol{X}\boldsymbol{b}$, and the vector of *residuals* is $\boldsymbol{r} \equiv \boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}(\boldsymbol{b})$. See Definition 1.14 for the OLS model for $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. The following model is useful for the centered response and standardized nontrivial predictors, or if $\boldsymbol{Z} = \boldsymbol{Y}$, $\boldsymbol{W} = \boldsymbol{X}_I$, and $\boldsymbol{\eta} = \boldsymbol{\beta}_I$ corresponds to a submodel I.

Definition 5.1. 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 $\boldsymbol{r} = \boldsymbol{r}(\boldsymbol{Z}, \boldsymbol{W}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Z}$.

Assume that the sample correlation matrix

$$\boldsymbol{R}_{\boldsymbol{u}} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}. \tag{5.6}$$

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 $n \to \infty$. Then by Theorem 2.26 (the LS CLT), the OLS estimator satisfies

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}). \tag{5.7}$$

Remark 5.4: 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 4 for variable selection and Equation (4.1) where $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. 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 (4.1), 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 (4.1), 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 $\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S$. Two alternative variable selection models were given by Remark 4.24.

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) relaxed lasso which is OLS applied to the lasso active set (nontrivial predictors with nonzero coefficients) and a constant. See Sections 5.2–5.7 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 5.10. The number of models M depends on the method. Often one of the models is the full model (5.1) 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 $\hat{\boldsymbol{\eta}} = \boldsymbol{0}$ 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 $I_1, ..., I_M$.

Remark 5.5. Prediction interval (4.14) used a number d that was often the number of predictors in the selected model. For forward selection, PCR, PLS, lasso, and relaxed lasso, let d be the number of predictors $v_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ in the final model (with nonzero coefficients), including a constant v_1. For forward selection, lasso, and relaxed lasso, v_j corresponds to a single nontrivial predictor, say $v_j = x_j^* = x_{k_j}$. Another method for obtaining d is to 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.

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 $n > p$, $\boldsymbol{X}$ is not invertible, but if $n = p$, then $\hat{\boldsymbol{Y}} = \boldsymbol{HY} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y} = \boldsymbol{I}_n\boldsymbol{Y} = \boldsymbol{Y}$ 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 5.1. We want $n \geq 10d$ to avoid overfitting. Occasionally n as low as $5d$ is used, but models with $n < 5d$ are overfitting.

Remark 5.6. 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 $\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \xrightarrow{D} N_r(\boldsymbol{0}, \boldsymbol{V})$, 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 2.26 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 $\boldsymbol{0}$. 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 5.2. 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

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

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 $f(\boldsymbol{\eta}) = \triangledown Q$, and we solve $f(\boldsymbol{\eta}) = \triangledown Q \overset{set}{=} \boldsymbol{0}$ 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 5.1. 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 $\triangledown Q = \boldsymbol{a}$.

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

c) If $Q(\boldsymbol{\eta}) = \sum_{i=1}^k |\eta_i| = \|\boldsymbol{\eta}\|_1$, then $\triangledown Q = \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$ 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 5.1. 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 5.1 with $\boldsymbol{a}^T = \boldsymbol{Z}^T\boldsymbol{W}$ and $\boldsymbol{A} = \boldsymbol{W}^T\boldsymbol{W}$ shows that $\triangledown Q = -2\boldsymbol{W}^T\boldsymbol{Z} + 2(\boldsymbol{W}^T\boldsymbol{W})\boldsymbol{\eta}$. Let $\triangledown Q(\hat{\boldsymbol{\eta}})$ 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 5.2. 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.

```
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
```

5.2 Forward Selection

Variable selection methods such as forward selection were covered in Chapter 4 where model I_j uses j predictors $x_1^*, ..., x_j^*$ including the constant $x_1^* \equiv 1$. If n/p is not large, forward selection can be done as in Chapter 4 except 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.

Remark 5.7. 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 (4.14)
```

```
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$residˆ2)         #first 19 predictors
[1] 79.91771             #SSE(I) = RSS(I)
n*log(out$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 5.2, 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. 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 5.5, 5.6, and 5.8.

Definition 5.3. Suppose the population MLR model has $\boldsymbol{\beta}_S$ an $a_S \times 1$ vector. The population MLR model is *sparse* if a_S is small. The population MLR model is *dense* or abundant if $n/a_S < J$ where $J = 5$ or $J = 10$, say.

The fitted model $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ is *sparse* if $d =$ number of nonzero coefficients is small. The fitted model is *dense* if $n/d < J$ where $J = 5$ or $J = 10$.

5.3 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

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{x}. \tag{5.8}$$

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 5.2. 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 5.3. 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 $\boldsymbol{A}^{1/2}\boldsymbol{A}^{1/2} = \boldsymbol{A}$.

Principal components regression (PCR) uses OLS regression on the principal components of the correlation matrix $\boldsymbol{R_u}$ of the $p-1$ nontrivial predictors $u_1 = x_2, ..., u_{p-1} = x_p$. Suppose $\boldsymbol{R_u}$ 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 $K = \min(n, p-1)$. Then $\boldsymbol{R_u}\hat{\boldsymbol{e}}_i = \hat{\lambda}_i \hat{\boldsymbol{e}}_i$ for $i = 1, ..., K$. Since $\boldsymbol{R_u}$ 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 matrix of the standardized nontrivial predictors $\boldsymbol{w}_i$, the sample covariance matrix

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

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

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

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

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

Principal components have a nice geometric interpretation if $n > K = p-1$. If $n > K$ and $\boldsymbol{R_u}$ is nonsingular, then the hyperellipsoid

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

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

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

Then points at squared distance $\boldsymbol{w}^T\boldsymbol{R_u}^{-1}\boldsymbol{w} = h^2$ 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 $(p-1)$th principal component $P_{j,p-1}$ is obtained by projecting the $\boldsymbol{w}_j$ on the (shortest) minor axis of the hyperellipsoid. Examine Figure 4.3 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 $(P_{1i}, ..., P_{ni})^T = (V_{1i}, ..., V_{ni})^T$ 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 $Cov(V_i, V_j) =$

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

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

$$\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{w}_k \boldsymbol{w}_k^T = \boldsymbol{R}_{\boldsymbol{u}}$$

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

PCR uses linear combinations of the standardized data as predictors. Let $V_j = \hat{\boldsymbol{e}}_j^T \boldsymbol{w}$ for $j = 1, ..., K$. Let model J_i contain $V_1, ..., V_i$. Then for model J_i, use OLS regression of $Z = Y - \overline{Y}$ on $V_1, ..., V_i$ with $\hat{Y} = \hat{Z} + \overline{Y}$. Since linear combinations of $\boldsymbol{w}$ are linear combinations of $\boldsymbol{x}$, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{PCR,I_j}$ where the model I_j uses a constant and the first $j-1$ PCR components.

Notation: Just as we use x_i or X_i to denote the ith predictor, we will use v_j or V_j to denote predictors that are linear combinations of the original predictors: e.g. $v_j = V_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ or $v_j = V_j = \boldsymbol{\gamma}_j^T \boldsymbol{u}$.

Remark 5.8. The set of $(p-1) \times 1$ vectors $\{(1, 0, ..., 0)^T, (0, 1, 0, ..., 0)^T, (0, ...0, 1)^T\}$ is the standard basis for $\mathbb{R}^{p-1}$. The set of vectors $\{\hat{\boldsymbol{e}}_1, ..., \hat{\boldsymbol{e}}_{p-1}\}$ is also a basis for $\mathbb{R}^{p-1}$. For PCR and some constants θ_i, $\sum_{i=1}^{j} \theta_i \hat{\boldsymbol{e}}_j^T \boldsymbol{w} = \sum_{i=1}^{p-1} \eta_i w_i$ if $j = p-1$, but not if $j < p-1$ in general. Hence PCR tends to give inconsistent estimators unless P$(j = p-1)$ = P(PCR uses the OLS full model) goes to one.

There are at least two problems with PCR. i) In general, $\hat{\boldsymbol{\beta}}_{PCR,I_j}$ is an inconsistent estimator of $\hat{\boldsymbol{\beta}}$ unless $P(j \to p-1) = P(\hat{\boldsymbol{\beta}}_{PCR,I_j} \to \hat{\boldsymbol{\beta}}_{OLS}) \to 1$

as $n \to \infty$. ii) 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 components (and a constant) could be much better for prediction than the other principal components. See Jolliffe (1983) and Cook and Forzani (2008). If $n \geq 10p$, often PCR needs to use all $p-1$ components (i.e., PCR = OLS full model) to be competitive with other regression models. Performing OLS forward selection or lasso on $V_1, ..., V_K$ may be more effective. There is one exception. Suppose $\sum_{i=1}^{J} \hat{\lambda}_i \geq q(p-1)$ where $0.5 \leq q \leq 1$, e.g. $q = 0.8$ where J is a lot smaller than $p-1$. Then the J predictors $V_1, ..., V_J$ capture much of the information of the standardized nontrivial predictors $w_1, ..., w_{p-1}$. Then regressing Y on $1, V_1, ..., V_J$ may be competitive with regressing Y on $1, w_1, ..., w_{p-1}$. PCR is equivalent to OLS on the full model when Y is regressed on a constant and all K of the principal components. PCR can also be useful if $\boldsymbol{X}$ is singular or nearly singular (ill conditioned).

Example 5.2, 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
```

5.4 Partial Least Squares

Partial least squares (PLS) uses variables $v_1 = 1$ (the constant or trivial predictor) and "PLS components" $v_j = \boldsymbol{\gamma}_j^T \boldsymbol{x}$ for $j = 2, ..., p$. Next let the response Y be used with the standardized predictors W_j. Let the "PLS components" $V_j = \hat{\boldsymbol{g}}_j^T \boldsymbol{w}$. Let model J_i contain $V_1, ..., V_i$. Often k–fold cross validation is used to pick the PLS model from $J_1, ..., J_M$. PLS seeks directions $\hat{\boldsymbol{g}}_j$ such that the PLS components V_j are highly correlated with Y, subject to being uncorrelated with other PLS components V_i for $i \neq j$. Note that PCR components are formed without using Y.

Remark 5.9. PLS may or may not give a consistent estimator of $\boldsymbol{\beta}$ if p/n does not go to zero: rather strong regularity conditions have been used to prove consistency or inconsistency if p/n does not go to zero. See Chun and Keleş (2010), Cook (2018), Cook et al. (2013), and Cook and Forzani (2018, 2019).

Following Hastie et al. (2009, pp. 80-81), let $\boldsymbol{W} = [\boldsymbol{s}_1, ..., \boldsymbol{s}_{p-1}]$ so $\boldsymbol{s}_j$ is the vector corresponding to the standardized jth nontrivial predictor. Let $\hat{g}_{1i} = \boldsymbol{s}_j^T \boldsymbol{Y}$ be n times the least squares coefficient from regressing Y on $\boldsymbol{s}_i$. Then the first PLS direction $\hat{\boldsymbol{g}}_1 = (\hat{g}_{11}, ..., \hat{g}_{1,p-1})^T$. Note that $\boldsymbol{W}\hat{\boldsymbol{g}}_i = (V_{i1}, ..., V_{in})^T = \boldsymbol{p}_i$ is the ith PLS component. This process is repeated using matrices $\boldsymbol{W}^k = [\boldsymbol{s}_1^k, ..., \boldsymbol{s}_{p-1}^k]$ where $\boldsymbol{W}^0 = \boldsymbol{W}$ and $\boldsymbol{W}^k$ is orthogonalized with respect to $\boldsymbol{p}_k$ for $k = 1, ..., p-2$. So $\boldsymbol{s}_j^k = \boldsymbol{s}_j^{k-1} - [\boldsymbol{p}_k^T \boldsymbol{s}_j^{k-1} / (\boldsymbol{p}_k^T \boldsymbol{p}_k)]\boldsymbol{p}_k$ for $j = 1, ..., p-1$. If the PLS model I_i uses a constant and PLS components $V_1, ..., V_{i-1}$, let $\hat{\boldsymbol{Y}}_{I_i}$ be the predicted values from the PLS model using I_i. Then $\hat{\boldsymbol{Y}}_{I_i} = \hat{\boldsymbol{Y}}_{I_{i-1}} + \hat{\theta}_i \boldsymbol{p}_i$ where $\hat{\boldsymbol{Y}}_{I_0} = \overline{Y}\boldsymbol{1}$ and $\hat{\theta}_i = \boldsymbol{p}_i^T \boldsymbol{Y} / (\boldsymbol{p}_i^T \boldsymbol{p}_i)$. Since linear combinations of $\boldsymbol{w}$ are linear combinations of $\boldsymbol{x}$, $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{PLS,I_j}$ where I_j uses a constant and the first $j-1$ PLS components. If $j = p$, then the PLS model I_p is the OLS full model.

Example 5.2, continued. The PLS 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<-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
```

The Mevik et al. (2015) `pls` library is useful for computing PLS and PCR.

5.5 Ridge Regression

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$. Ridge regression uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model

$\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. Then $\hat{Y}_i = \hat{Z}_i + \overline{Y}$. Note that in Definition 5.5, $\lambda_{1,n}$ 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 5.1 for the OLS estimator $\hat{\boldsymbol{\eta}}_{OLS} = (\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$.

Definition 5.4. Consider the MLR model $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. 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 $\hat{\boldsymbol{Z}}(\boldsymbol{b}) = \boldsymbol{Wb}$ and the vector of residuals $\boldsymbol{r}(\boldsymbol{b}) = \boldsymbol{Z} - \hat{\boldsymbol{Z}}(\boldsymbol{b})$.

Definition 5.5. Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. Let $\lambda \geq 0$ be a constant. The *ridge regression estimator* $\hat{\boldsymbol{\eta}}_R$ minimizes the *ridge regression criterion*

$$Q_R(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}\eta_i^2 \tag{5.9}$$

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

$$\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}. \tag{5.10}$$

The residual sum of squares $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta})$, 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 $\overline{\boldsymbol{Y}}$.

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 5.2. 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 5.10. 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

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

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 5.10 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 5.10 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*

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

which is the matrix of scalar products. **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 $df(\lambda_{1,n}) =$

$$tr[\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T] = \sum_{i=1}^{p-1} \frac{\sigma_i^2}{\sigma_i^2 + \lambda_{1,n}} = \sum_{i=1}^{p-1} \frac{\hat{\lambda}_i}{\hat{\lambda}_i + \lambda_{1,n}} \tag{5.11}$$

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

$$Q_{R,n}(\boldsymbol{b}) = \frac{1}{2n} \boldsymbol{r}(\boldsymbol{b})^T \boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n} \boldsymbol{b}^T \boldsymbol{b}, \tag{5.12}$$

as in Hastie et al. (2015, p. 10). Then $\lambda_{2n} = \lambda_{1,n}/(2n)$ using the $\lambda_{1,n}$ from (5.9).

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 $\hat{\boldsymbol{\eta}}_R = \boldsymbol{A}_{n,\lambda} \hat{\boldsymbol{\eta}}_{OLS}$ appears in Efron and Hastie (2016, p. 98), and Marquardt and Snee (1975). See Theorem 5.4 for the ridge regression central limit theorem.

Remark 5.11. Ridge regression has a simple relationship with OLS if $n > p$ and $(\boldsymbol{W}^T\boldsymbol{W})^{-1}$ exists. Then $\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z} = (\boldsymbol{W}^T\boldsymbol{W}+\lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}(\boldsymbol{W}^T\boldsymbol{W})(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z} = \boldsymbol{A}_{n,\lambda}\hat{\boldsymbol{\eta}}_{OLS}$ where $\boldsymbol{A}_{n,\lambda} \equiv \boldsymbol{A}_n = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}$. By the LS CLT Equation (5.7) with $\hat{\boldsymbol{V}}/n = (\boldsymbol{W}^T\boldsymbol{W})^{-1}$, a normal approximation for OLS is

$$\hat{\boldsymbol{\eta}}_{OLS} \sim AN_{n-p}(\boldsymbol{\eta}, MSE\ (\boldsymbol{W}^T\boldsymbol{W})^{-1}).$$

Hence a normal approximation for ridge regression is

$$\hat{\boldsymbol{\eta}}_R \sim AN_{p-1}(\boldsymbol{A}_n\boldsymbol{\eta}, MSE\ \boldsymbol{A}_n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{A}_n^T) \sim$$

$$AN_{p-1}[\boldsymbol{A}_n\boldsymbol{\eta}, MSE\ (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}(\boldsymbol{W}^T\boldsymbol{W})(\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}].$$

If Equation (5.7) holds and $\lambda_{1,n}/n \to 0$ as $n \to \infty$, then $\boldsymbol{A}_n \overset{P}{\to} \boldsymbol{I}_{p-1}$.

Remark 5.12. The ridge regression criterion from Definition 5.5 can also be defined by

$$Q_R(\boldsymbol{\eta}) = \|\boldsymbol{Z} - \boldsymbol{W}\boldsymbol{\eta}\|_2^2 + \lambda_{1,n}\boldsymbol{\eta}^T\boldsymbol{\eta}. \tag{5.13}$$

Then by Theorem 5.1, 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

$$-\boldsymbol{W}^T(\boldsymbol{Z} - \boldsymbol{W}\hat{\boldsymbol{\eta}}_R) + \lambda_{1,n}\hat{\boldsymbol{\eta}}_R = \boldsymbol{0}. \tag{5.14}$$

Following Hastie and Efron (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: $\hat{\boldsymbol{\eta}}_R = (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T\boldsymbol{Z}$$

$$= (\boldsymbol{W}^T\boldsymbol{W} + \lambda_{1,n}\boldsymbol{I}_{p-1})^{-1}\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{OLS} = \boldsymbol{A}_n\hat{\boldsymbol{\eta}}_{OLS} =$$

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

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

since $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$. See Problem 5.3. Assume Equation (5.6) holds. If $\lambda_{1,n}/n \to 0$ then

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

Note that

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

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 5.4, RR CLT (Ridge Regression Central Limit Theorem. Assume p is fixed and that the conditions of the LS CLT Theorem Equation (5.7) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

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

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

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

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

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

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

Hence

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - \sqrt{n}\frac{\hat{\lambda}_{1,n}}{n} n(\boldsymbol{W}^T\boldsymbol{W} + \hat{\lambda}_{1,n}\boldsymbol{I}_{p-1})^{-1}\hat{\boldsymbol{\eta}}_{OLS}$$

$$\xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \tau\boldsymbol{V}\boldsymbol{\eta} \sim N_{p-1}(-\tau\boldsymbol{V}\boldsymbol{\eta}, \sigma^2\boldsymbol{V}). \quad \square$$

For p fixed, Knight and Fu (2000) note i) that $\hat{\boldsymbol{\eta}}_R$ 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 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 $\boldsymbol{\eta}$ 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

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

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 (4.1) and Remark 5.4 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 5.9, 5.10, 5.11, and 5.12.

Warning. Although the *R* functions `glmnet` and `cv.glmnet` appear to do ridge regression, getting the fitted values, $\hat{\lambda}_{1,n}$, and degrees of freedom to match up with the formulas of this section can be difficult.

Example 5.2, 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
```

5.6 Lasso

Consider the MLR model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$. Lasso uses the centered response $Z_i = Y_i - \overline{Y}$ and standardized nontrivial predictors in the model $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$ as described in Remark 5.1. 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 5.6. Consider fitting the MLR model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$ using $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. The *lasso estimator* $\hat{\boldsymbol{\eta}}_L$ minimizes the *lasso criterion*

$$Q_L(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i| \tag{5.15}$$

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 $RSS(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta})$, 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 $\overline{\boldsymbol{Y}}$.

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 5.2. We could also write

$$Q_L(\boldsymbol{b}) = \frac{1}{a}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \frac{\lambda_{1,n}}{a}\sum_{j=1}^{p-1}|b_j|, \tag{5.16}$$

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 $\hat{\boldsymbol{\eta}}_L$:

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

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

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

If none of the elements of $\boldsymbol{\eta}$ are zero, and if $\hat{\boldsymbol{\eta}}_L$ is a consistent estimator of $\boldsymbol{\eta}$, then $\boldsymbol{s}_n \xrightarrow{P} \boldsymbol{s} = \boldsymbol{s}_{\boldsymbol{\eta}}$. 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 $\hat{\boldsymbol{\eta}}_{\lambda_M} = \boldsymbol{0}$. Hence $\hat{\boldsymbol{\eta}}_{\lambda_i} \neq \boldsymbol{0}$ 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 $\sqrt{n}(\hat{\boldsymbol{\eta}}_L - \hat{\boldsymbol{\eta}}_{OLS}) = o_p(1)$.

Theorem 5.5, Lasso CLT. Assume p is fixed and that the conditions of the LS CLT Theorem Equation (5.7) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

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

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

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

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

Proof. 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

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

$$\sqrt{n}(\hat{\boldsymbol{\eta}}_{OLS} - \boldsymbol{\eta}) - \sqrt{n}\frac{\lambda_{1,n}}{2n} n(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{s}_n \xrightarrow{D} N_{p-1}(\boldsymbol{0}, \sigma^2\boldsymbol{V}) - \frac{\tau}{2}\boldsymbol{V}\boldsymbol{s}$$
$$\sim N_{p-1}\left(\frac{-\tau}{2}\boldsymbol{V}\boldsymbol{s}, \sigma^2\boldsymbol{V}\right)$$

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

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 $\hat{\boldsymbol{\eta}}_L$ 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 $\boldsymbol{\eta}$ if $\lambda_{1,n} = O(\sqrt{n})$ (so $\lambda_{1,n}/\sqrt{n}$ is bounded). Note that Theorem 5.5 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 $a = 2n$:

$$Q_{L,2n}(\boldsymbol{b}) = \frac{1}{2n}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_{2n} \sum_{j=1}^{p-1} |b_j|, \tag{5.17}$$

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 (5.15). For model selection, the M values of λ are denoted by $0 \le \lambda_{2n,1} < \lambda_{2n,2} < \cdots < \lambda_{2n,M}$ where $\hat{\boldsymbol{\eta}}_\lambda = \boldsymbol{0}$ iff $\lambda \ge \lambda_{2n,M}$ 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 $0 \le \lambda_1 < \lambda_2 < \cdots < \lambda_M$, used above Theorem 5.5, 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 5.7. 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 (5.17), and $n\lambda_{2n} = \lambda_{1,n}/2$. We could replace $n\ \lambda_{2n}$ by λ_2 if we used $a = 2$ in the criterion

$$Q_{L,2}(\boldsymbol{b}) = \frac{1}{2}\boldsymbol{r}(\boldsymbol{b})^T\boldsymbol{r}(\boldsymbol{b}) + \lambda_2 \sum_{j=1}^{p-1} |b_j|. \tag{5.18}$$

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 5.2, 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.

5.7 Lasso Variable Selection

Lasso variable selection applies OLS on a constant and the active predictors that have nonzero lasso $\hat{\eta}_i$. The method is called relaxed lasso by Hastie et al. (2015, p. 12), and the relaxed lasso ($\phi = 0$) estimator by Meinshausen (2007). The method is also called OLS-post lasso and post model selection OLS. Let $\boldsymbol{X}_A$ denote the matrix with a column of ones and the unstandardized active nontrivial predictors. Hence the lasso variable selection estimator is $\hat{\boldsymbol{\beta}}_{LVS} = (\boldsymbol{X}_A^T \boldsymbol{X}_A)^{-1} \boldsymbol{X}_A^T \boldsymbol{Y}$, and lasso variable selection is an alternative to forward selection. Let k be the number of active (nontrivial) predictors so $\hat{\boldsymbol{\beta}}_{VLS}$ is $(k+1) \times 1$.

Let I_{min} correspond to the lasso variable selection estimator and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{LVS,0} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ to the zero padded lasso variable selection estimator. Then by Remark 4.5 where p is fixed, $\hat{\boldsymbol{\beta}}_{LVS,0}$ is $\sqrt{n}$ consistent when lasso is consistent, with the limiting distribution for $\hat{\boldsymbol{\beta}}_{LVS,0}$ given by Theorem 4.4. Hence, relaxed lasso can be bootstrapped with the same methods used for forward selection in Chapter 4. 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 $(\boldsymbol{X}_A^T \boldsymbol{X}_A)$ is ill conditioned or if $n/(k+1) < 10$. Also see Pelawa Watagoda and Olive (2020) and Rathnayake and Olive (2020).

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 relaxed lasso 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
##relaxed lasso =
#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 5.2, 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 $\lambda_0 > 0$ was selected. The OLS SE, t statistic and pvalue are generally not valid for relaxed lasso by Remark 4.5 and Theorem 4.4.

```
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 5.2, forward selection selected the model with the minimum C_p while the other methods used 10-fold CV. PLS and PCR used

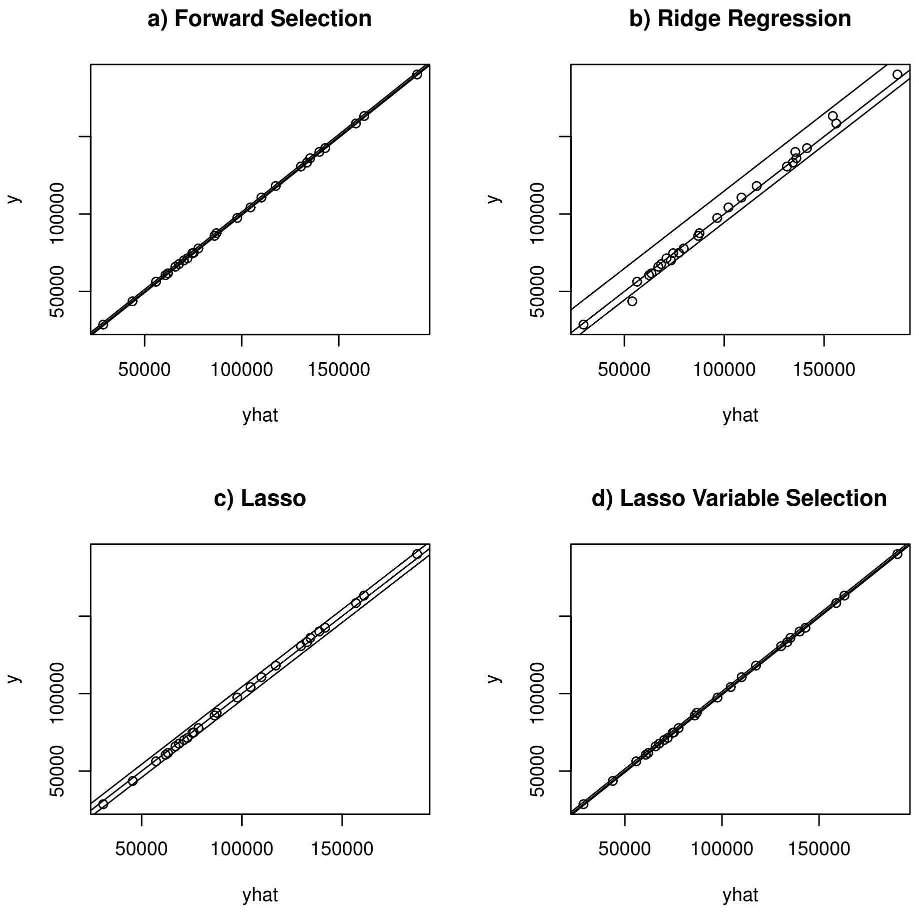

Fig. 5.1 Marry Data Response Plots

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 20336.58, lasso and lasso variable selection used a constant, *mmen*, and *pop* with lasso PI length 8482.62 and relaxed lasso PI length 2226.53. PI (4.14) was used. Figure 5.1 shows the response plots for forward selection, ridge regression, lasso, and lasso variable selection. 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 5.1 a) and d).

5.8 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 5.8. The *elastic net* estimator $\hat{\boldsymbol{\beta}}_{EN}$ minimizes the criterion

$$Q_{EN}(\boldsymbol{\beta}) = \frac{1}{2}RSS(\boldsymbol{\beta}) + \lambda_{1,n}\left[\frac{1}{2}(1-\alpha)\|\boldsymbol{\beta}_S\|_2^2 + \alpha\|\boldsymbol{\beta}_S\|_1\right], \text{ or} \tag{5.19}$$

$$Q_2(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta}) + \lambda_1\|\boldsymbol{\beta}_S\|_2^2 + \lambda_2\|\boldsymbol{\beta}_S\|_1 \tag{5.20}$$

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. 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 $\overline{\boldsymbol{Y}} = \overline{Y}\boldsymbol{1}$, and the $n \times (p-1)$ matrix of standardized nontrivial predictors $\boldsymbol{W}$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e} \tag{5.21}$$

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

$$Q_{EN}(\boldsymbol{\eta}) = RSS_W(\boldsymbol{\eta}) + \lambda_1\|\boldsymbol{\eta}\|_2^2 + \lambda_2\|\boldsymbol{\eta}\|_1 \tag{5.22}$$

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

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

where $\boldsymbol{0}$ 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

$$Q_L(\boldsymbol{\eta}) = RSS_A(\boldsymbol{\eta}) + \lambda_2 \|\boldsymbol{\eta}\|_1 = Q_{EN}(\boldsymbol{\eta}). \tag{5.23}$$

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 5.13. 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 (5.23) 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 5.6. iii) For prediction intervals, let d be the number of nonzero coefficients from the equivalent augmented lasso problem (5.23). 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 (5.23) 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 $a_S = 1500$.

Following Jia and Yu (2010), by standard Karush-Kuhn-Tucker (KKT) conditions for convex optimality for Equation (5.20), $\hat{\boldsymbol{\eta}}_{EN}$ is optimal if

$$2\boldsymbol{W}^T\boldsymbol{W}\hat{\boldsymbol{\eta}}_{EN} - 2\boldsymbol{W}^T\boldsymbol{Z} + 2\lambda_1\hat{\boldsymbol{\eta}}_{EN} + \lambda_2\boldsymbol{s}_n = 0, \quad \text{or}$$

$$(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})\hat{\boldsymbol{\eta}}_{EN} = \boldsymbol{W}^T\boldsymbol{Z} - \frac{\lambda_2}{2}\boldsymbol{s}_n, \quad \text{or}$$

$$\hat{\boldsymbol{\eta}}_{EN} = \hat{\boldsymbol{\eta}}_R - n(\boldsymbol{W}^T\boldsymbol{W} + \lambda_1\boldsymbol{I}_{p-1})^{-1}\frac{\lambda_2}{2n}\boldsymbol{s}_n. \tag{5.24}$$

Hence

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

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

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,

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

The following theorem is due to Slawski et al. (2010), and summarized in Pelawa Watagoda and Olive (2020).

Theorem 5.6, Elastic Net CLT. Assume p is fixed and that the conditions of the LS CLT Equation (5.7) hold for the model $\boldsymbol{Z} = \boldsymbol{W}\boldsymbol{\eta} + \boldsymbol{e}$.

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

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

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

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

Proof. By the above remarks and the RR CLT Theorem 5.4,

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

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

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

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

Example 5.2, 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 relaxed elastic net estimator is $\hat{\boldsymbol{\beta}}_{RL} = (\boldsymbol{X}_A^T\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^T\boldsymbol{Y}$, and relaxed elastic net is an alternative to forward selection. Let k be the number of active (nontrivial) predictors so $\hat{\boldsymbol{\beta}}_{REN}$ is $(k+1)\times 1$. Let I_{min} correspond to the elastic net variable selection estimator and $\hat{\boldsymbol{\beta}}_{VS} = \hat{\boldsymbol{\beta}}_{ENVS,0} = \hat{\boldsymbol{\beta}}_{I_{min},0}$ to the zero padded relaxed elastic net estimator. Then by Remark 4.5 where p is fixed, $\hat{\boldsymbol{\beta}}_{ENVS,0}$ is $\sqrt{n}$ consistent when elastic net is consistent, with the limiting distribution for $\hat{\boldsymbol{\beta}}_{REN,0}$ given by Theorem 4.4. Hence, relaxed elastic net can be bootstrapped with the same methods used for forward selection in Chapter 4. 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 $n/(k+1) < 10$. Also see Olive (2019) and Rathnayake and Olive (2020).

5.9 Prediction Intervals

This section will use the prediction intervals from Section 4.3 applied to the MLR model with $\hat{m}(\boldsymbol{x}) = \boldsymbol{x}_I^T\hat{\boldsymbol{\beta}}_I$ and I corresponds to the predictors used by the MLR method. We will use the six methods forward selection with OLS, PCR, PLS, lasso, relaxed lasso, and ridge regression. When $p > n$, results from Hastie et al. (2015, pp. 20, 296, ch. 6, ch. 11) and Luo and Chen (2013) suggest that lasso, relaxed lasso, and forward selection with EBIC can perform well for sparse models: the subset S in Equation (4.1) and Remark 5.4 has a_S small.

Consider d for the prediction interval (4.14). As in Chapter 4, with the exception of ridge regression, let d be the number of "variables" used by the method, including a constant. Hence for lasso, relaxed lasso, and forward selection, $d-1$ is the number of active predictors while $d-1$ is the number of "components" used by PCR and PLS.

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 7, 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 $\boldsymbol{X}_I$.

For the simulation from Pelawa Watagoda and Olive (2019b), we used several R functions including forward selection (FS) as computed with the `regsubsets` function from the `leaps` library, principal components regression (PCR) with the `pcr` function and partial least squares (PLS) with the `plsr` function from the `pls` library, and ridge regression (RR) and lasso with the `cv.glmnet` function from the `glmnet` library. Relaxed lasso (RL) 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 5.1 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	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 $\text{Cov}(\boldsymbol{u}_i) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ 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 (4.16) is not asymptotically optimal for iii), and for iii) PI (4.16) 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, relaxed lasso, 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 (2016). The results were similar for all five error distributions, and we show some results for the normal and shifted exponential distributions. Tables 5.1 and 5.2 show some simulation results for PI (4.14) 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 did appear to select $S \subseteq I_{min}$ since relaxed lasso was good.

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 relaxed lasso often performed well. For $k = 19$, forward selection with EBIC often performed well, as did lasso and relaxed lasso for $\psi = 0$. 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 $1/\sqrt{p}$.

Tables 5.3 and 5.4 show some results for PIs (4.15) and (4.16). 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 CRL used PI (4.16). For relaxed lasso, the program sometimes failed to run for 5000 runs, e.g., if the

Table 5.2 Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1)-1$

n	p	ψ	k		FS	lasso	RL	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 5.3 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim N(0,1)$

n,p,ψ,k		FS	CFS	RL	CRL	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

number of variables selected $d = n_H$. In Table 5.3, PIs (4.15) and (4.16) are asymptotically equivalent, but PI (4.16) had shorter lengths for moderate n. In Table 5.4, PI (4.15) is shorter than PI (4.16) asymptotically, but for moderate n, PI (4.16) was often shorter.

Table 5.5 shows some results for PIs (4.14) and (4.15) for lasso and ridge regression. The header lasso indicates PI (4.14) was used while vlasso indicates that PI (4.15) was used. PI (4.15) tended to work better when the fit was poor while PI (4.14) was better for $n = 2p$ and $k = p - 1$. The PIs are asymptotically equivalent for consistent estimators.

Table 5.4 Validation Residuals: Simulated Large Sample 95% PI Coverages and Lengths, $e_i \sim EXP(1) - 1$

n,p,ψ,k		FS	CFS	RL	CRL	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

Table 5.5 PIs (4.14) and (4.15): 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

5.10 Cross Validation

For MLR variable selection there are many methods for choosing the final submodel, including AIC, BIC, C_p, and EBIC. See Section 4.1. Variable selection is a special case of model selection where there are M models a a final model needs to be chosen. Cross validation is a common criterion for model selection.

Definition 5.9. 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 (4.1) 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 Theorem 4.4 and Remark 4.5 showed 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 5.2, continued. The *linmodpack* 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
```

5.11 Hypothesis Testing After Model Selection, n/p Large

Section 4.6 showed how to use the bootstrap for hypothesis test $H_0 : \boldsymbol{\theta} = \boldsymbol{A\beta} = \boldsymbol{\theta}_0$ versus $H_1 : \boldsymbol{\theta} = \boldsymbol{A\beta} \neq \boldsymbol{\theta}_0$ with the statistic $T_n = \boldsymbol{A}\hat{\boldsymbol{\beta}}_{I_{min},0}$ where $\hat{\boldsymbol{\beta}}_{I_{min},0}$ is the zero padded OLS estimator computed from the variables corresponding to I_{min}. The theory needs $P(S \subseteq I_{min}) \to 1$ as $n \to \infty$, and hence applies to OLS variable selection with AIC, BIC, and C_p, and to relaxed lasso and relaxed elastic net if lasso and elastic net are consistent.

Assume $n \geq 20p$ and that the error distribution is unimodal and not highly skewed. The response plot and residual plot are plots with $\hat{Y} = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ on the horizontal axis and Y or r on the vertical axis, respectively. Then the plotted points in these plots should scatter in roughly even bands about the identity line (with unit slope and zero intercept) and the $r = 0$ line, respectively. See Figure 1.1. If the plots for the OLS full model suggest that the error distribution is skewed or multimodal, then much larger sample sizes may be needed.

Let p be fixed. Then lasso is asymptotically equivalent to OLS if $\hat{\lambda}_{1n}/\sqrt{n} \to 0$, and hence should not have any $\hat{\beta}_i = 0$, asymptotically. If $a_S < p$, then lasso tends not be $\sqrt{n}$ consistent if lasso selects S with high probability by Ewald and Schneider (2018), but then relaxed lasso tends to be $\sqrt{n}$ consistent. If $\hat{\lambda}_{1n}/n \to 0$, then lasso is consistent so $P(S \subseteq I) \to 1$ as $n \to \infty$. Hence often if lasso has more than one $\hat{\boldsymbol{\beta}}_i = 0$, then lasso is not $\sqrt{n}$ consistent.

Suppose we use the residual bootstrap where $\boldsymbol{Y}^* = \boldsymbol{X}\hat{\boldsymbol{\beta}}_{OLS} + \boldsymbol{r}^W$ follows a standard linear model where the elements r_i^W of $\boldsymbol{r}^W$ are iid from the empirical distribution of the OLS full model residuals r_i. In Section 4.6 we used forward selection when regressing Y^* on $\boldsymbol{X}$, but we could use lasso or ridge regression instead. Since these estimators are consistent if $\hat{\lambda}_{1n}/n \to 0$ as $n \to \infty$, we expect $\hat{\boldsymbol{\beta}}_L^*$ and $\hat{\boldsymbol{\beta}}_R^*$ to be centered at $\hat{\boldsymbol{\beta}}_{OLS}$. If the variabliity of the $\hat{\boldsymbol{\beta}}^*$ is similar to or greater than that of $\hat{\boldsymbol{\beta}}_{OLS}$, then by the geometric argument Theorem 4.5, we might get simulated coverage close to or higher than the nominal. If lasso or ridge regression shrink $\hat{\boldsymbol{\beta}}^*$ too much, then the coverage could be bad. In limited simulations, the prediction region method only simulated well for ridge regression with $\psi = 0$. Results from Ewald and Schneider (2018, p. 1365) suggest that the lasso confidence region volume is greater than OLS confidence region volume when lasso uses $\lambda_{1n} = \sqrt{n}/2$.

A small simulation was done for confidence intervals and confidence regions, using the same type of data as for the variable selection simulation in Section 4.6 and the prediction interval simulation in Section 5.9, with $B = \max(1000, n, 20p)$ and 5000 runs. The regression model used $\boldsymbol{\beta} = (1, 1, 0, 0)^T$ with $n = 100$ and $p = 4$. When $\psi = 0$, the design matrix $\boldsymbol{X}$ consisted of iid N(0,1) random variables. See Table 5.6 which was taken from Pelawa Watagoda (2017). The residual bootstrap was used. Types 1)–5) correspond to types i)–v), and the ϵ value only applies to the type 5) error distribution. The function `lassobootsim3` uses the prediction region method for lasso and ridge regression. The function `lassobootsim4` can be used to simulate confidence intervals for the β_i is $\boldsymbol{S}_T^*$ is singular for lasso. The test was for $H_0 : (\beta_3, \beta_4)^T = (0, 0)^T$.

5.12 Data Splitting

A common method for data splitting randomly divides the data set into two half sets. On the first half set, fit the model selection method, e.g. forward

Table 5.6 Bootstrapping Lasso, $\psi = 0$

n	ϵ	type		β_1	β_2	β_3	β_4	test
100		1	cov	0.9440	0.9376	0.9910	0.9946	0.9790
			len	0.4143	0.4470	0.3759	0.3763	2.6444
		2	cov	0.9468	0.9428	0.9946	0.9944	0.9816
			len	0.6870	0.7565	0.6238	0.6226	2.6832
		3	cov	0.9418	0.9408	0.9930	0.9948	0.9840
			len	0.4110	0.4506	0.3743	0.3746	2.6684
		4	cov	0.9468	0.9370	0.9938	0.9948	0.9838
			len	0.2392	0.2578	0.2151	0.2153	2.6454
	0.5	5	cov	0.9438	0.9344	0.9988	0.9970	0.9924
			len	2.9380	2.5042	2.4912	2.4715	2.8536
	0.9	5	cov	0.9506	0.9290	0.9974	0.9976	0.9956
			len	3.9180	3.2760	3.7356	3.2739	2.8836

selection or lasso, to get the a predictors. Use this model as the full model for the second half set: use the standard OLS inference from regressing the response on the predictors found from the first half set. 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 from the first half set (training data) is found, and data splitting estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{I,D}$ computed from the second half set (test data). This estimator is not the relaxed lasso estimator. The estimator $\hat{\boldsymbol{\beta}}_{I,D}$ has the same large sample theory as if I was chosen before obtaining the data.

5.13 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.

4)

Label	coef	SE	shorth 95% CI for β_i
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

$$\boldsymbol{R_u} = \frac{\boldsymbol{W}^T\boldsymbol{W}}{n}.$$

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 $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, methods such as forward selection, PCR, PLS, ridge regression, relaxed lasso, 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

$$Q(\boldsymbol{\eta}) = \frac{1}{a}(\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta}) + \frac{\lambda_{1,n}}{a}\sum_{i=1}^{p-1}|\eta_i|^j \tag{5.25}$$

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 $RSS_W(\boldsymbol{\eta}) = (\boldsymbol{Z} - \boldsymbol{W\eta})^T(\boldsymbol{Z} - \boldsymbol{W\eta})$, 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 $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. 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

$$Q_{EN}(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_1 \|\boldsymbol{\eta}\|_2^2 + \lambda_2 \|\boldsymbol{\eta}\|_1 \tag{5.26}$$

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

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

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

15) Assume that the sample correlation matrix $\boldsymbol{R}_{\boldsymbol{u}} = \dfrac{\boldsymbol{W}^T\boldsymbol{W}}{n} \xrightarrow{P} \boldsymbol{V}^{-1}$. 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), relaxed lasso = VS-lasso and relaxed elastic net = VS-elastic net are $\sqrt{n}$ consistent under much milder conditions than lasso and elastic net, since the relaxed 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, relaxed lasso, and relaxed elastic net) is given by Theorem 4.4. 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 (4.1) 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 (4.14) 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, relaxed lasso, and relaxed elastic net 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 (4.14) 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, relaxed lasso, and relaxed elastic net 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 (4.14) 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 relaxed lasso with k–fold CV, and forward selection with EBIC can perform well even if n/p is small. So PI (4.14) can be useful when n/p is small.

5.14 Complements

Good references for forward selection, PCR, PLS, ridge regression, and lasso are Hastie et al. (2009, 2015), James et al. (2013), Olive (2019), Pelawa Watagoda (2017) and Pelawa Watagoda and Olive (2019b). 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).

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), Lee et al. (2016), 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, or are asymptotically equivalent to using the full model, or find a quantity to test that is not $\boldsymbol{A\beta}$. Typically the methods have not been shown to perform better than data splitting. See Ewald and Schneider (2018). When n/p is not large, inference is currently much more difficult. Under strong regularity conditions, lasso and forward selection with EBIC can work well. Leeb et al. (2015) suggests that the Berk et al. (2013) method does not really work. Also see Dezeure et al. (2015), Javanmard and Montanari (2014), Lu et al. (2017), Tibshirani et al. (2016), van de Geer et al. (2014), and Zhang and Cheng (2017). Fan and Lv (2010) gave large sample theory for some methods if $p = o(n^{1/5})$. See Tibshirani et al. (2016) for an *R* package.

Warning: For $n < 5p$, every estimator is unreliable, to my knowledge. Regularity conditions for consistency are strong if they exist. For example, PLS is sometimes inconsistent and sometimes $\sqrt{n}$ consistent. Validating the MLR estimator with PIs can help. Also make response and residual plots.

Full OLS Model: A sufficient condition for $\hat{\boldsymbol{\beta}}_{OLS}$ to be a consistent estimator of $\boldsymbol{\beta}$ is $\text{Cov}(\hat{\boldsymbol{\beta}}_{OLS}) = \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \to \boldsymbol{0}$ as $n \to \infty$. See Lai et al. (1979).

Forward Selection: See Olive and Hawkins (2005), Pelawa Watagoda and Olive (2019ab), and Rathnayake and Olive (2019).

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: PLS was introduced by Wold (1975). Also see Wold (1985, 2006). Two useful papers are Cook et al. (2013) and Cook and Su (2016). PLS tends to be $\sqrt{n}$ consistent if p is fixed and $n \to \infty$. If $p > n$, under two sets of strong regularity conditions, PLS can be $\sqrt{n}$ consistent or inconsistent. See Chun and Keleş (2010), Cook (2018), Cook and Forzani (2018, 2019), and Cook et al. (2013). Denham (1997) suggested a PI for PLS that assumes the number of components is selected in advance.

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. In other words, use lasso variable selection as a variable selection method. 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 $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. Let $\lambda \geq 0$ be a constant and let $q \geq 0$. The *estimator* $\hat{\boldsymbol{\eta}}_q$ minimizes the *criterion*

$$Q_q(\boldsymbol{b}) = \boldsymbol{r}(\boldsymbol{b})^T \boldsymbol{r}(\boldsymbol{b}) + \lambda \sum_{j=1}^{p-1} |b_i|^q, \tag{5.27}$$

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.

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 (5.6) 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).

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 (1992ab). 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 (2018)), 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 7 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). Steinberger and Leeb (2016) used leave-one-out residuals, but delete the upper and lower 2.5% of the residuals to make a 95% PI. Hence the PI will have undercoverage and the shorth PI will tend to be shorter when the error distribution is not symmetric.

Let p be fixed, d be for PI (4.14), and $n \to \infty$. For elastic net, forward selection, PCR, PLS, ridge regression, relaxed lasso, 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 (4.14) 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 4.21, 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 (4.15) 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.

5.15 Problems

5.1. 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

$$\hat{\boldsymbol{\eta}}_R \approx \hat{\boldsymbol{\eta}}_{OLS} - \frac{\lambda}{n} \boldsymbol{V} \hat{\boldsymbol{\eta}}_{OLS}.$$

5.2. 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 relaxed lasso.
a) Which method corresponds to $j = 1$?
b) Which method corresponds to $j = 2$?
c) Which method corresponds to $\lambda_{1,n} = 0$?

5.3. 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 $\boldsymbol{A}_n - \boldsymbol{B}_n = \boldsymbol{0}$.

5.4. 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.

5.5. 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 $\boldsymbol{W}_1 = [\boldsymbol{1} \;\; \boldsymbol{W}]$ and the $p \times 1$ vector $\hat{\boldsymbol{\eta}}_1 = (\overline{Y} \;\; \hat{\boldsymbol{\eta}}^T)^T$ where the scalar $\overline{Y}$ is the sample mean of the response variable. Show $\hat{\boldsymbol{Y}} = \boldsymbol{W}_1\hat{\boldsymbol{\eta}}_1$.

5.6. 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

$\boldsymbol{R_u} = \boldsymbol{G}^T\boldsymbol{G}$. Then regression through the origin is used for the model

$$\boldsymbol{Z} = \boldsymbol{G}\boldsymbol{\eta} + \boldsymbol{e} \tag{5.28}$$

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 (see Remark 5.1), 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

$$Q_N(\boldsymbol{\eta}) = RSS(\boldsymbol{\eta}) + \lambda_2^*\|\boldsymbol{\eta}\|_2^2 + \lambda_1^*\|\boldsymbol{\eta}\|_1 \tag{5.29}$$

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

$$\boldsymbol{G}_A = \begin{pmatrix} \boldsymbol{G} \\ \sqrt{\lambda_2^*}\ \ \boldsymbol{I}_{p-1} \end{pmatrix}, \quad \text{and} \quad \boldsymbol{Z}_{\text{A}} = \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{0} \end{pmatrix},$$

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 (5.22).)

5.7. 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?

5.8. 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 5.8
         pop mmen mmilmen milwmn
1  ( 1 ) " " "*"  " "     " "
2  ( 1 ) " " "*"  "*"     " "
3  ( 1 ) "*" "*"  "*"     " "
4  ( 1 ) "*" "*"  "*"     "*"
out$cp
[1] -0.8268967  1.0151462  3.0029429  5.0000000
```

5.9. Consider the output for Example 4.7 for the OLS full model. The column *resboot* gives the large sample 95% CI for β_i using the shorth applied to the $\hat{\beta}^*_{ij}$ for $j = 1, ..., B$ using the residual bootstrap. The standard large sample 95% CI for β_i is $\hat{\beta}_i \pm 1.96 SE(\hat{\beta}_i)$. Hence for β_2 corresponding to L, the standard large sample 95% CI is $-0.001 \pm 1.96(0.002) = -0.001 \pm 0.00392 = [-0.00492, 0.00292]$ while the shorth 95% CI is $[-0.005, 0.004]$.

a) Compute the standard 95% CIs for β_i corresponding to W, H, and S. Also write down the shorth 95% CI. Are the standard and shorth 95% CIs fairly close?

b) Consider testing $H_0 : \beta_i = 0$ versus $H_A : \beta_i \neq 0$. If the corresponding 95% CI for β_i does not contain 0, then reject H_0 and conclude that the predictor variable X_i is needed in the MLR model. If 0 is in the CI then fail to reject H_0 and conclude that the predictor variable X_i is not needed in the MLR model given that the other predictors are in the MLR model.

Which variables, if any, are needed in the MLR model? Use the standard CI if the shorth CI gives a different result. The nontrivial predictor variables are L, W, H, and S.

5.10. 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.

5.11. Table 5.7 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?

Table 5.7 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

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 $D_{\mathbf{0}} \le D_{(U_B)}$. 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?

5.12. Suppose the MLR model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{e}$, and the regression method fits $\boldsymbol{Z} = \boldsymbol{W\eta} + \boldsymbol{e}$. Suppose $\hat{Z} = 245.63$ and $\overline{Y} = 105.37$. What is $\hat{Y}$?

5.13. 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 $\boldsymbol{x}_{I_{min},f} = (1, 75000)^T$?

5.14. Table 5.8 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.)

Table 5.8 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	

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?

5.15. 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
```

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

5.16. Table 5.9 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 5.9 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 $n = 400$?

5.17. 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:/linmodpack.txt")* **to download the functions** and the command *source("G:/linmoddata.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/linmodrhw.txt) into *R*.

5.18. 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 $\boldsymbol{\beta} = (1, 1, 0, 0)$?

5.19. This simulation is similar to that used to form Table 4.2, 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 $H_0 : (\beta_{k+2}, ..., \beta_p)^T = (\beta_3, ..., \beta_p)^T = \boldsymbol{0}$ 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 4.2 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/linmodrhw.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*.

5.20. This problem is like Problem 5.19, but ridge regression is used instead of forward selection. This simulation is similar to that used to form Table 4.2, 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 $H_0 : (\beta_{k+2}, ..., \beta_p)^T = (\beta_3, ..., \beta_p)^T = \boldsymbol{0}$ 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 4.2 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/linmodrhw.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*.

5.21. This is like Problem 5.20, except lasso is used. If you are person j in Problem 5.20, 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?