One Component Partial Least Squares, High Dimensional Regression, Data Splitting, and the Multitude of Models

David J. Olive and Lingling Zhang*
Southern Illinois University

February 9, 2023

Abstract

This paper gives large sample theory for the one component partial least squares (OPLS) estimator, including some hypothesis tests for high dimensional data, under much weaker conditions than those in the literature. Simple theory is also given for some data splitting estimators and the marginal maximum likelihood estimators. It is shown that lasso, OPLS, and ordinary least squares often estimate different population multiple linear regression models. The paper also proves that there are often many valid population models for regression methods such as binary regression. The above theory is used to correct some common errors in the literature.

KEY WORDS: Cox Proportional Hazards Regression, Dimension Reduction, Elastic Net, GLM, Lasso, MMLE, OLS, PCR, PLS, Prediction, Sparse Regression, Variable Selection.

1 Introduction

This section reviews regression models, including variable selection and data splitting. Many regression models have a response variable Y that is independent of the $p \times 1$ vector of predictors $\boldsymbol{x} = (x_1, ..., x_p)^T$ given $\boldsymbol{x}^T\boldsymbol{\beta}$, written $Y \perp\!\!\!\perp \boldsymbol{x} | \boldsymbol{x}^T\boldsymbol{\beta}$. Then there are n cases $(Y_i, \boldsymbol{x}_i^T)^T$, and the sufficient predictor $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$. For the regression models, the conditioning and subscripts, such as i, will often be suppressed. The multiple linear regression model is $Y|\boldsymbol{x}^T\boldsymbol{\beta} = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$ or $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$. Consider a parametric regression model $Y|\boldsymbol{x}^T\boldsymbol{\beta} \sim D(\alpha + \boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{\gamma})$ where D is a parametric distribution that depends on $\boldsymbol{x}$ only through $\boldsymbol{x}^T\boldsymbol{\beta}$, and $\boldsymbol{\gamma}$ is a $q \times 1$ vector of parameters. Three examples follow. The *binomial logistic regression model* is

*David J. Olive is a Professor and Lingling Zhang is a recent Ph.D., School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901, USA.

$Y_i \sim$ binomial $\left(\mathrm{m_i}, \rho(\mathrm{SP}) = \dfrac{\mathrm{e^{SP}}}{1+\mathrm{e^{SP}}}\right)$. The binary logistic regression model has $m_i \equiv 1$ for $i = 1, ..., n$. A useful *Poisson regression model* is $Y \sim$ Poisson $\left(\mathrm{e^{SP}}\right)$. For $\alpha = 0$, the *Weibull proportional hazards regression model* is

$$Y|SP \sim W(\gamma = 1/\sigma, \lambda_0 \exp(SP))$$

where Y has a Weibull $W(\gamma, \lambda)$ distribution if the probability density function of Y is

$$f(y) = \lambda\gamma y^{\gamma-1} \exp[-\lambda y^{\gamma}] \quad \text{for} \quad \mathrm{y} > 0.$$

Variable selection estimators include forward selection or backward elimination when $n \geq 10p$. When n/p is not large, the Chen and Chen (2008) EBIC criterion with forward selection can be useful. Sparse regression methods can be also be used for variable selection even if n/p is not large: the regression submodel, such as a Nelder and Wedderburn (1972) generalized linear model (GLM), uses the predictors that had nonzero sparse regression estimated coefficients. These methods include least angle regression, lasso, relaxed lasso, elastic net, and sparse regression by projection. See Efron, Hastie, Johnstone, and Tibshirani (2004, p. 421), Friedman et al. (2007), Friedman, Hastie, and Tibshirani (2010), Meinshausen (2007, p. 376), Qi, Luo, Carroll, and Zhao (2015), Simon, Friedman, Hastie, and Tibshirani (2011), Tibshirani (1996), and Zou and Hastie (2005).

Following Olive and Hawkins (2005), a *model for variable selection* can be described by

$$\boldsymbol{x}^T\boldsymbol{\beta} = \boldsymbol{x}_S^T\boldsymbol{\beta}_S + \boldsymbol{x}_E^T\boldsymbol{\beta}_E = \boldsymbol{x}_S^T\boldsymbol{\beta}_S \tag{1}$$

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

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

Thus $\boldsymbol{\beta}_O = \boldsymbol{0}$ if $S \subseteq I$. The model using $\boldsymbol{x}^T\boldsymbol{\beta}$ is the full model.

To clarify notation, suppose $p = 3$, a constant α is always in the model, and $\boldsymbol{\beta} = (\beta_1, 0, 0)^T$. Then the $J = 2^p = 8$ possible subsets of $\{1, 2, ..., p\}$ are $I_1 = \emptyset$, $S = I_2 = \{1\}$, $I_3 = \{2\}$, $I_4 = \{3\}$, $I_5 = \{1, 2\}$, $I_6 = \{1, 3\}$, $I_7 = \{2, 3\}$, and $I_8 = \{1, 2, 3\}$. There are $2^{p-a_S} = 4$ subsets I_2, I_5, I_6, and I_8 such that $S \subseteq I_j$. Let $\hat{\boldsymbol{\beta}}_{I_7} = (\hat{\beta}_2, \hat{\beta}_3)^T$ and $\boldsymbol{x}_{I_7} = (x_2, x_3)^T$.

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

Theory for the variable selection estimator $\hat{\boldsymbol{\beta}}_{VS}$ is complicated. See Pelawa Watagoda and Olive (2021) for multiple linear regression, and Rathnayake and Olive (2021) for models such as GLMs and Cox (1972) proportional hazards regression. For fixed p, these two papers showed that $\hat{\boldsymbol{\beta}}_{VS}$ is $\sqrt{n}$ consistent with a complicated nonnormal limiting distribution.

Principal components regression (PCR) and partial least squares (PLS) models use p linear combinations $\boldsymbol{\eta}_1^T\boldsymbol{x}, ..., \boldsymbol{\eta}_p^T\boldsymbol{x}$. Then there are p conditional distributions

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

Estimating the $\boldsymbol{\eta}_i$ and performing the ordinary least squares (OLS) regression of Y on $(\hat{\boldsymbol{\eta}}_1^T\boldsymbol{x}, \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x}, ..., \hat{\boldsymbol{\eta}}_k^T\boldsymbol{x})$ gives the k-component estimator, e.g. the k-component PLS estimator $\hat{\boldsymbol{\beta}}_{kPLS}$ or the k-component PCR estimator, for $k = 1, ..., J$ where $J \leq p$ and the p-component estimator is the OLS estimator $\hat{\boldsymbol{\beta}}_{OLS}$. Denote the one component PLS (OPLS) estimator by $\hat{\boldsymbol{\beta}}_{OPLS}$. The model selection estimator chooses one of the k-component estimators, e.g. using a holdout sample or cross validation, and will be denoted by $\hat{\boldsymbol{\beta}}_{MSPLS}$. See Cook (2018), James et al. (2021), and Wold (1975) for more on these and related estimators.

For estimation with OLS, let the covariance matrix of $\boldsymbol{x}$ be $\text{Cov}(\boldsymbol{x}) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = E[(\boldsymbol{x} - E(\boldsymbol{x}))(\boldsymbol{x} - E(\boldsymbol{x}))^T = E(\boldsymbol{x}\boldsymbol{x}^T) - E(\boldsymbol{x})E(\boldsymbol{x}^T)$ and $\boldsymbol{\eta} = \text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = E[(\boldsymbol{x} - E(\boldsymbol{X})(Y - E(Y))] = E(\boldsymbol{x}Y) - E(\boldsymbol{x})E(Y) =$
$E[(\boldsymbol{x} - E(\boldsymbol{x}))Y] = E[\boldsymbol{x}(Y - E(Y))]$. Let

$$\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\eta}}_n = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} = \boldsymbol{S}_{\boldsymbol{x}Y} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y})$$

and

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

Then the OLS estimators are $\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T\overline{\boldsymbol{x}}$ and

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

For a multiple linear regression model with independent, identically distributed (iid) cases, $\hat{\boldsymbol{\beta}}_{OLS}$ is a consistent estimator of $\boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ under mild regularity conditions, while $\hat{\alpha}_{OLS}$ is a consistent estimator of $E(Y) - \boldsymbol{\beta}_{OLS}^T E(\boldsymbol{x})$.

Cook, Helland, and Su (2013) showed that $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\lambda}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ estimates $\lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$ where

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}} \quad \text{and} \quad \hat{\lambda} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}} \tag{2}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \mathbf{0}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \mathbf{0}$, then $\boldsymbol{\beta}_{OPLS} = \mathbf{0}$. Let $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. Large sample theory for OPLS is given in Section 2, and see Section 3.1 for earlier theory.

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

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

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

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

where $(\boldsymbol{w}, Y)$ denotes that Y was regressed on $\boldsymbol{w}$, and $\boldsymbol{I}$ is the $p \times p$ identity matrix. See, for example, James et al. (2021, p. 260).

Data splitting divides the training data set of n cases into two sets: H and the validation set V where H has n_H of the cases and V has the remaining $n_V = n - n_H$ cases $i_1, ..., i_{n_V}$. An application of data splitting is to use a variable selection method, such as forward selection or lasso, on H to get submodel I_{min} with a predictors, then fit the selected model to the cases in the validation set V using standard inference. See, for example, Rinaldo et al. (2019).

High dimensional regression has n/p small. A fitted or population regression model is sparse if a of the predictors are active (have nonzero $\hat{\beta}_i$ or β_i) where $n \geq Ja$ with $J \geq 10$. Otherwise the model is nonsparse. A high dimensional population regression model is abundant or dense if the regression information is spread out among the p predictors (nearly all of the predictors are active). Hence an abundant model is a nonsparse model. See Cook, Forzani, and Rothman (2013).

Section 2 gives the large sample theory for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ and OPLS. Section 3 proves that there are a multitude of regression models and gives more theory for regression estimators. Section 4 explains a sequential data splitting method that was used for Section 5. The simulation in Section 5 shows that lasso with k-fold cross validation often selects models that are not the population generating model, but which are useful for prediction.

2 Large Sample Theory and Testing

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

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

Then the low order moments are needed for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$ to be a consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ where

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{z}_i - \overline{\boldsymbol{z}}_n)(\boldsymbol{z}_i - \overline{\boldsymbol{z}}_n)^T,$$

and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \frac{n-1}{n}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}$.

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

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}),\ \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \tag{3}$$

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

b) Let $\boldsymbol{z}_i = \boldsymbol{x}_i(Y_i - \overline{Y}_n)$ and $\boldsymbol{v}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}}_n)(Y_i - \overline{Y}_n)$. Then $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$. Hence $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}}$.

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

$$n\tilde{\boldsymbol{\eta}}_n = \sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}} + \boldsymbol{\mu}_{\boldsymbol{x}} - \overline{\boldsymbol{x}})(Y_i - \mu_Y + \mu_Y - \overline{Y})^T = \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \mu_Y)^T$$

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

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

Thus

$$\sqrt{n}\tilde{\boldsymbol{\eta}}_n = \sqrt{n}\frac{1}{n}\sum_i \boldsymbol{w}_i - \frac{\sqrt{n}(\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})\sqrt{n}(\overline{Y} - \mu_Y)}{\sqrt{n}} = \sqrt{n}\ \overline{\boldsymbol{w}}_n + o_P(1).$$

Hence

$$\sqrt{n}(\tilde{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) = \sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) + o_P(1).$$

Thus

$$\sqrt{n}(\tilde{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}})$$

by Slutsky's theorem. Now

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

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

Thus

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

Now

$$\sum_i \boldsymbol{w}_i = \sum_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}} + \overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \overline{Y} + \overline{Y} - \mu_Y) = \sum_i (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(Y_i - \overline{Y}) +$$

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

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

Hence

$$\sum_i (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \sum_i [(\boldsymbol{z}_i + \boldsymbol{a}_n - (\overline{\boldsymbol{z}}_n + \boldsymbol{a}_n))(\boldsymbol{z}_i - \overline{\boldsymbol{z}}_n)^T] =$$

$$\sum_i (\boldsymbol{z}_i - \overline{\boldsymbol{z}})(\boldsymbol{z}_i - \overline{\boldsymbol{z}})^T.$$

Thus

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \frac{1}{n-1} \sum_{i=1}^{n} (\boldsymbol{z}_i - \overline{\boldsymbol{z}})(\boldsymbol{z}_i - \overline{\boldsymbol{z}})^T$$

and

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} = \frac{n-1}{n} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}. \ \square$$

In the following theorem, the scalars λ and $\hat{\lambda}$ are given by Equation (2), $\boldsymbol{\eta} = (\eta_1, ..., \eta_p)^T$, and $\boldsymbol{\Sigma}_{\boldsymbol{\eta}} = \boldsymbol{\Sigma}_{\boldsymbol{w}}$. For c), note that an estimator of $\lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{\eta}}\boldsymbol{A}^T$ is $\hat{\lambda}^2 \boldsymbol{A}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}\boldsymbol{A}^T$. The quantity $\boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T$ is difficult to estimate, although the nonparametric bootstrap, that draws cases with replacement, may be useful if n/p is large. Note that

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

Results from Su and Cook (2012), for example, show that elements of a sample covariance matrix can be stacked to get large sample theory. Then $\hat{\lambda}$ and $\hat{\boldsymbol{\eta}}$ can be stacked as in Theorem 2 by the multivariate delta method. This assumption in Theorem 2 may or may not be strong, but Theorem 1 can also be used for testing.

THEOREM 2. *Assume*

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

$$\sim N_{p+1}(\boldsymbol{0}, \boldsymbol{\Sigma}).$$

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

b)

$$\sqrt{n}(\hat{\lambda}\hat{\boldsymbol{\eta}} - \lambda\boldsymbol{\eta}) = \sqrt{n}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_p\left(\boldsymbol{0}, \boldsymbol{D}\boldsymbol{\Sigma}\boldsymbol{D}^T\right)$$

with $\boldsymbol{D} = [\boldsymbol{\eta} \ \ \lambda \boldsymbol{I}_p]$ *where* $\boldsymbol{I}_p$ *is the* $p \times p$ *identity matrix.*

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

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

PROOF. a) Follows by Equation (3) or since joint convergence in distribution implies marginal convergence in distribution.

b) Follows by the Multivariate Delta Method with

$$\boldsymbol{g}\left(\begin{array}{c}\lambda\\ \boldsymbol{\eta}\end{array}\right) = \lambda\boldsymbol{\eta} =$$

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

c) By b),

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

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

REMARK 1: Notice that Theorems 1 and 2 depend on the theory of both the sample covariance vector and the sample covariance matrix, not on any other model such as linearity. It is possible that $Y|\boldsymbol{x}$ does not follow a linear model, but $Y|\boldsymbol{\beta}_E^T\boldsymbol{x}$ does follow a linear model. If the population generating model $Y = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$ is a linear model, then $Y|\boldsymbol{x} = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$ is a linear model. Suppose the cases are iid, and the predictors have nonsingular covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{x}}$. Suppose a linear model holds with $Y|\boldsymbol{x} = \alpha + \boldsymbol{\beta}^T\boldsymbol{x} + e$. If the iid errors e are independent of the predictors $\boldsymbol{x}$, then under mild conditions, linearity implies that $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$ and that the covariance structure is $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}$. Suppose $(\hat{\alpha}_E, \hat{\boldsymbol{\beta}}_E)$ estimates $(\alpha_E, \boldsymbol{\beta}_E)$. If $Y|\boldsymbol{x} = \alpha_E + \boldsymbol{\beta}_E^T\boldsymbol{x} + e$, then by the above discussion, $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_E$.

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

$$E(\boldsymbol{x}\boldsymbol{x}^T\boldsymbol{\beta}) - E(\boldsymbol{x})E(\boldsymbol{x}^T)\boldsymbol{\beta}. \text{ Then}$$

$$\boldsymbol{\beta}_{OPLS} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x},Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x},Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x},Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x},Y}}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} =$$
$$\lambda\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}.$$

Since $\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}$,

$$\boldsymbol{\beta}_{OPLS} = \lambda\mathrm{Cov}(\boldsymbol{x})\boldsymbol{\beta}_{OLS} = \frac{\boldsymbol{\beta}^T[\mathrm{Cov}(\boldsymbol{x})]^2\boldsymbol{\beta}}{\boldsymbol{\beta}^T[\mathrm{Cov}(\boldsymbol{x})]^3\boldsymbol{\beta}}\ \mathrm{Cov}(\boldsymbol{x})\ \boldsymbol{\beta},$$

and

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

2.1 High Dimensional Tests

The following simple testing method reduces a possibly high dimensional problem to a low dimensional problem. Testing $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{0}$ is equivalent to testing $H_0 : \boldsymbol{A}\boldsymbol{\eta} = \boldsymbol{0}$ versus $H_1 : \boldsymbol{A}\boldsymbol{\eta} \neq \boldsymbol{0}$ where $\boldsymbol{A}$ is a $k \times p$ constant matrix. Let $\mathrm{Cov}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}) = \mathrm{Cov}(\hat{\boldsymbol{\eta}}) = \boldsymbol{\Sigma}_{\boldsymbol{W}}$ be the asymptotic covariance matrix of $\hat{\boldsymbol{\eta}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$. In high dimensions where $n < 5p$, we can't get a good nonsingular estimator of $\mathrm{Cov}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y})$, but we can get good nonsingular estimators of $\mathrm{Cov}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}) = \mathrm{Cov}((\hat{\eta}_{i1}, ..., \hat{\eta}_{ik})^T)$ with

$\boldsymbol{u} = (x_{i1}, ..., x_{ik})^T$ where $n \geq Jk$ with $J \geq 10$. (Values of J much larger than 10 may be needed if some of the k predictors and/or Y are skewed.) Simply apply Theorem 1 to the predictors $\boldsymbol{u}$ used in the hypothesis test, and thus use the sample covariance matrix of the vectors $\boldsymbol{u}_i(Y_i - \overline{Y})$. Hence we can test hypotheses like $H_0 : \beta_i - \beta_j = 0$. In particular, testing $H_0 : \beta_i = 0$ is equivalent to testing $H_0 : \eta_i = \sigma_{x_i,Y} = 0$ where $\sigma_{x_i,Y} = \mathrm{Cov}(x_i, Y)$.

Note that the tests with $\hat{\boldsymbol{\eta}}$ using k predictors x_{ij} do not depend on other predictors, including important predictors that were left out of the model (underfitting). Hence the tests can have considerable resistance to underfitting and overfitting. The tests also have some resistance to measurement error: assume that $(\boldsymbol{x}_i^T, \boldsymbol{u}_i^T, v_i, Y_i)^T$ are iid but $\boldsymbol{w}_i = \boldsymbol{x}_i + \boldsymbol{u}_i$ and $Z_i = Y_i + v_i$ are observed instead of $(\boldsymbol{x}_i, Y_i)$. Then $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{w}, Z)$ estimates $\boldsymbol{\Sigma}_{\boldsymbol{w}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{w}Z}$, while $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}Z}$ estimates $\mathrm{Cov}(\boldsymbol{x}, Y)$ if $\mathrm{Cov}(\boldsymbol{x}, v) + \mathrm{Cov}(\boldsymbol{u}, Y) + \mathrm{Cov}(\boldsymbol{u}, v) = \boldsymbol{0}$, which occurs, for example, if $\boldsymbol{x} \perp\!\!\!\perp v$, $\boldsymbol{u} \perp\!\!\!\perp Y$, and $\boldsymbol{u} \perp\!\!\!\perp v$.

The tests with $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\lambda}\hat{\boldsymbol{\eta}}$ and k predictor variables may not be as good as the tests with $\hat{\boldsymbol{\eta}}$ since $\hat{\lambda}$ needs to be a good estimator of λ. Note that $\hat{\lambda}$ can be a good estimator if $\hat{\boldsymbol{\eta}}^T\boldsymbol{x}$ is a good estimator of $\boldsymbol{\eta}^T\boldsymbol{x}$.

3 The Multitude of Models

This section shows that there are often a multitude of population regression models that are estimating different population parameters. Note that when j predictors each satisfy a marginal regression model with the response Y (such as simple linear regression), then subsets of those j predictors will often satisfy a regression model with the response Y (such as multiple linear regression). Under multivariate normality, it is known that $Y|\boldsymbol{x}_I$ follows a multiple linear regression model where $\boldsymbol{x}_I = (x_{i1}, ..., x_{ik})^T$ is a vector corresponding to a subset of the predictors. Theorem 3a) gives a similar result for every linear combination of the predictors $\boldsymbol{\eta}^T\boldsymbol{x}$, including sparse and nonsparse models. Much of Theorem 3a) can also be shown by performing the population SLR of Y on $\boldsymbol{\eta}^T\boldsymbol{x}$, but linearity may fail to hold if multivariate normality does not hold. Note that data sets where the cases are iid from a multivariate normal distribution are rather uncommon. Let $\Sigma_Y = \sigma_Y^2$.

THEOREM 3. *Suppose the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some distribution.*
a) If the joint distribution of $(Y, \boldsymbol{x}^T)^T$ is multivariate normal,

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

then $Y|\boldsymbol{x} \sim Y|(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x}) \sim N(\alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x}, \sigma^2)$ follows a multiple linear regression model, but so does $Y|\boldsymbol{\eta}^T\boldsymbol{x} \sim N(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}, \sigma_O^2)$ where $\alpha_O = \mu_Y - \boldsymbol{\beta}_O^T\boldsymbol{\mu}_{\boldsymbol{x}}$, $\boldsymbol{\beta}_O = \lambda\boldsymbol{\eta}$, $\sigma_O^2 = \Sigma_Y - \boldsymbol{\beta}_O^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, and

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

b) If the response Y is binary, then $Y|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}) \sim$ binomial$(m = 1, \rho(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}))$ where $E[Y|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})] = \rho(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x}) = P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T\boldsymbol{x})]$. Hence every linear combination of the predictors satisfies a binary regression model.

PROOF. a)

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

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

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

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

and

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

b) $E[Y|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})] = 0P[Y = 0|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})] + 1P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})] = P[Y = 1|(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})] = \rho(\alpha_O + \boldsymbol{\beta}_O^T \boldsymbol{x})$. □

For multiple linear regression, note that $\sigma_O^2 < \sigma_Y^2 = \Sigma_Y$ unless $\boldsymbol{\eta}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = 0$. If $\boldsymbol{\eta} = \boldsymbol{\beta}_{OLS}$, then $\lambda = 1$ and $\sigma_O^2 = \sigma_Y^2 - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. The population quantity estimated by the one component partial least squares estimator corresponds to $\boldsymbol{\eta} = \text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$.

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

3.1 Consequences

Although Theorems 1–3 have simple proofs, the theorems have important consequences. One consequence is the testing theory in Section 2.1. Another consequence is a better understanding of why regression models work. The discussion below also applies to multivariate regression models with m response variables of the form $\boldsymbol{y}|(\boldsymbol{\eta}_1^T \boldsymbol{x}, ..., \boldsymbol{\eta}_k^T \boldsymbol{x})$ where $\boldsymbol{y} = (Y_1, ..., Y_m)^T$. Then three important population models are i) the generating model that generated the data $(Y, \boldsymbol{x}^T)^T$, ii) the model $Y|\boldsymbol{x} \sim Y|\boldsymbol{\beta}_E^T \boldsymbol{x}$, and iii) the model $Y|\boldsymbol{\beta}_E^T \boldsymbol{x}$. In simulations, these three models are often the same (suppressing the distribution of $\boldsymbol{x}$). For example, if the generating model is $Y|\boldsymbol{x} = \alpha_E + \boldsymbol{\beta}_E^T \boldsymbol{x} + e$, then $\boldsymbol{\beta}_E = \boldsymbol{\beta}_{OLS}$ under mild regularity conditions. For the Poisson regression GLM, the generating model might be $Y|\boldsymbol{x} \sim \text{Poisson}[\exp(\alpha_{GLM} + \boldsymbol{\beta}_{GLM}^T \boldsymbol{x})]$ where $\hat{\boldsymbol{\beta}}_{GLM}$ is the maximum likelihood estimator.

Joint and conditional distributions tend to have more information than the marginal distribution of Y. Suppose the $(Y_i, \boldsymbol{x}_i^T, \boldsymbol{w}_i^T)^T = (Y_i, \boldsymbol{z}_i^T)^T$ are iid from a multivariate normal distribution with population generating model $Y|\boldsymbol{z} = \alpha + \boldsymbol{\beta}^T \boldsymbol{z} + e$ and that the $(Y_i, \boldsymbol{x}_i^T)^T$ are as in Theorem 3a). Then $Y|\boldsymbol{x} \sim Y|\boldsymbol{\beta}_{OLS}^T \boldsymbol{x} = \alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T \boldsymbol{x} + e$ while $Y|\boldsymbol{\beta}_{OPLS}^T \boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T \boldsymbol{x} + e$ where the distribution of e depends on the model. If

$Y|\boldsymbol{z}$ is a dense model, then the $Y|\boldsymbol{z}$ model may be too difficult to estimate, but the OLS and OPLS models that use $\boldsymbol{x}$ can be useful.

As another example, consider binary regression. If function $\boldsymbol{g} : \mathbb{R}^p \to \mathbb{R}^k$, then $Y|\boldsymbol{g}(\boldsymbol{x}) \sim$ binomial$(m = 1, \rho[\boldsymbol{g}(\boldsymbol{x})])$ where $E(Y|\boldsymbol{g}(\boldsymbol{x})) = \rho[\boldsymbol{g}(\boldsymbol{x})] = P(Y = 1|\boldsymbol{g}(\boldsymbol{x}))$. This result means that $Y|\boldsymbol{x}$ and the population generating model tend to be unknown for binary regression. However, if SP $= \alpha_{LR} + \boldsymbol{\beta}_{LR}^T \boldsymbol{x}$ and

$$Y|\boldsymbol{\beta}_{LR}^T \boldsymbol{x} \approx \text{binomial}\left(1, \rho(\text{SP}) = \frac{\text{e}^{\text{SP}}}{1 + \text{e}^{\text{SP}}}\right),$$

then the binomial logistic regression model tends to be useful.

Data splitting: To help understand data splitting when the cases in H are randomly selected, let I denote the predictors selected using H, possibly after variable selection or after looking at the data and building the model. Let $\hat{\boldsymbol{\beta}}_E(\boldsymbol{x}_I, Y)$ be the estimator obtained by regressing Y on $\boldsymbol{x}_I$ using the cases in V. Then $\hat{\boldsymbol{\beta}}_E(\boldsymbol{x}_I, Y)$ estimates $\boldsymbol{\beta}_I = \boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$. For example, if the cases are iid with enough low order moments, then $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{x}_I, Y)$ estimates $\boldsymbol{\beta}_I = \boldsymbol{\Sigma}_{\boldsymbol{x}_I}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}_I, Y}$ while $\hat{\boldsymbol{\beta}}_{OPLS}(\boldsymbol{x}_I, Y)$ estimates $\boldsymbol{\beta}_I = \lambda_I \boldsymbol{\Sigma}_{\boldsymbol{x}_I, Y}$. If the model is sparse, check the fitted model with the same checks used for low dimensional data. For data splitting in low dimensions, if the full model is good, then often model (1) works well in that we can eliminate predictors and often do nearly as well or better than the full model. In high dimensions, we often do not know if the full model, that regresses Y on $\boldsymbol{x}$, is good. The data splitting and high dimensional regression literature often claims that $\boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y) = \boldsymbol{\beta}_E(\boldsymbol{x}, Y)$. For example, $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}(\boldsymbol{x}, Y)$, or model (1) holds with $S \subseteq I_{min}$ and $\boldsymbol{\beta}_{I_{min}}$ a $k \times 1$ vector with $a_S \leq k \leq n/10$. While these claims can be true, the regularity conditions often become too strong as $n/p \to 0$.

MMLE and the oracle property: The MMLE is interesting since if each predictor satisfies a marginal model, then the marginal model theory can be used to find a confidence interval for β_i for $i = 1, ..., p$ where β_i is the ith component of $\boldsymbol{\beta}_{MMLE}$. For high dimensional multiple linear regression, the above regularity condition is weaker than the common assumption that the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from a multivariate normal distribution. For multiple linear regression, let $\boldsymbol{V} = diag(\boldsymbol{\Sigma}_{\boldsymbol{x}}) = diag(\sigma_1^2, ...\sigma_p^2)$. For iid cases, $\boldsymbol{\beta}_{MMLE} = \boldsymbol{V}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}, Y} = \boldsymbol{V}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\beta}_{OLS}$, and $\boldsymbol{\beta}_{MMLE} = \boldsymbol{\beta}_{OLS}$ if $\boldsymbol{\beta}_{OLS} = \boldsymbol{0}$, or if $(\boldsymbol{V}^{-1} - \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}) \boldsymbol{\Sigma}_{\boldsymbol{x}, Y} = \boldsymbol{0}$, or if $\boldsymbol{\beta}_{OLS}$ is an eigenvector of $\boldsymbol{V}^{-1} \boldsymbol{\Sigma}_{\boldsymbol{x}}$ with eigenvalue 1.

For standardized predictors, let s_j and σ_j be the sample and population standard deviations of x_j. Let $\boldsymbol{w}_i = \hat{\boldsymbol{D}} \boldsymbol{x}_i = diag(1/s_1, ...1/s_p) \boldsymbol{x}_i$ and $\boldsymbol{u}_i = \boldsymbol{D} \boldsymbol{x}_i = diag(1/\sigma_1, .., 1/\sigma_p) \boldsymbol{x}_i$. Note that $\sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}, Y} - \boldsymbol{\Sigma}_{\boldsymbol{u}, Y}) = \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}, Y} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}, Y}) + \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}, Y} - \boldsymbol{\Sigma}_{\boldsymbol{u}, Y}) = O_P(1) + \sqrt{n}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}, Y} - \boldsymbol{\Sigma}_{\boldsymbol{u}, Y})$ under mild regularity conditions for iid cases. Hence $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}, Y}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{u}, Y}$ that is not asymptotically equivalent to $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}, Y}$ unless $\boldsymbol{\Sigma}_{\boldsymbol{x}, Y} = \boldsymbol{0}$. The algebra given in the following theorem proves the theorem. Note that $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is the correlation matrix of $\boldsymbol{x}$.

THEOREM 4. *Consider the MMLE for multiple linear regression. Suppose the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid from some distribution. Let $\boldsymbol{w}_i$ be the standardized predictors and assume $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}, Y} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{u}, Y}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{u}}$ where the $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}}$ are nonsingular for large enough n and $\boldsymbol{\Sigma}_{\boldsymbol{u}}$ is nonsingular.*

$$a) \hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{w}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}, Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{w}, Y) \xrightarrow{P} \boldsymbol{\Sigma}_{\boldsymbol{u}, Y} =$$

$$\boldsymbol{\eta}_{OPLS}(\boldsymbol{u}, Y) = \boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma_u}[\boldsymbol{\Sigma_u}]^{-1}\boldsymbol{\Sigma}_{\boldsymbol{u},Y} = \boldsymbol{\Sigma_u}\boldsymbol{\beta}_{OLS}(\boldsymbol{u}, Y).$$

b) Let $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}(\boldsymbol{u}, Y)$. *Then* $\boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma_u}\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OLS}$ *if* $\boldsymbol{\beta}_{OLS} = \mathbf{0}$ *or if* $\boldsymbol{\beta}_{OLS}$ *is an eigenvector of* $\boldsymbol{\Sigma_u}$ *with eigenvalue = 1.*

The oracle property for model selection, including variable selection, is $P(I_{min} = S) \to 1$ as $n \to \infty$ for model (1). For this property to hold, S needs to be one of the subsets considered by the model selection method with probability going to 1 as $n \to \infty$. For fixed p and "fast" estimators such as lasso and forward selection, the oracle property tends to hold if the predictors are nearly orthogonal. See Wieczorek and Lei (2022) for references. The MMLE can be used for variable selection with OLS by taking the k predictors with the largest $|\hat{\beta}_{j,MMLE}|$. The oracle property for the MMLE tends not to hold for correlated predictors by Theorem 4.

MMLE variable selection often gives a useful suboptimal submodel since predictors that satisfy a marginal regression model with the response Y (such as SLR) will often satisfy a regression model with the response Y (such as multiple linear regression). Using the MMLE to find $k \approx 0.9n$ predictors I_1, and then using lasso or forward selection on these predictors to eliminate redundant predictors (keeping predictors I) may be effective.

OPLS and OLS: Chun and Keleş (2010) suggested that $\hat{\boldsymbol{\beta}}_{OPLS}$ only estimates $\boldsymbol{\beta}_{OLS}$ under very strong regularity conditions. Cook and Forzani (2018, 2019) showed that the regularity condition is $\boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \lambda\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, in which case $\sqrt{n}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OLS}) \xrightarrow{D} N_p(\mathbf{0}, \boldsymbol{C})$. Cook and Forzani (2018, 2019) also showed that under very strong regularity conditions for high dimensions, $\hat{\boldsymbol{\beta}}_{OPLS}$ is a consistent estimator of $\boldsymbol{\beta}_{OLS}$. Also see Basa et al. (2022).

In the literature, there is a tendency (perhaps a common Statistical paradigm) to assume that if the estimated model fits the data well, then the model corresponding to the estimator corresponds to the model for $Y|\boldsymbol{x}$. For example, in much of the OPLS literature, an assumption is $Y|\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$. Then $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$ by the Remark 1 in Section 2, and the results in Table 1 hold.

Table 1: OPLS Results Under Theorem 1 Assumptions

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

The above tendency leads to problems that have perhaps not yet been observed in the literature. To see some problems, consider multiple linear regression with $\text{Cov}(\boldsymbol{x}) = diag(1, 2, ..., p)$. First consider OPLS with $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS}$. Then at most one element of $\text{Cov}(\boldsymbol{x}, Y) = \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is nonzero since $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$ is an eigenvector of $\text{Cov}(\boldsymbol{x})$. Hence at most one predictor is correlated with Y, regardless of the value of p. This restriction is too strong.

If the cases are iid from a multivariate normal distribution, then $Y|\boldsymbol{x} = \alpha_{OLS} + \boldsymbol{\beta}_{OLS}^T\boldsymbol{x} + e$ and $Y|\boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T\boldsymbol{x} + e$ are both linear models by Theorem 3

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

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

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

Thus OLS and OPLS usually give different valid population multiple linear regression models with $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$. However, model iii) $Y|\boldsymbol{\beta}_{OPLS}^T \boldsymbol{x} = \alpha_{OPLS} + \boldsymbol{\beta}_{OPLS}^T \boldsymbol{x} + e$ is often a useful multiple linear regression model with large sample theory given in Section 2. The claims in the OPLS literature that $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS} =$ an eigenvector of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$ under mild regularity conditions are incorrect. See, for example, Basa et al. (2022), Cook and Forzani (2018, 2019), and Cook, Helland and Su (2013). The regularity conditions for $\boldsymbol{\beta}_{OLS} = \boldsymbol{\beta}_{OPLS}$ are very strong. In the OLS literature $\boldsymbol{\beta}_{OLS}$ can be any vector in $\mathbb{R}^p$. If $\boldsymbol{\beta}_{OLS}$, $\boldsymbol{\Sigma}_{\boldsymbol{x},Y}$, and $\boldsymbol{\beta}_{OPLS}$ were restricted to be eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{x}}$, then the OLS and OPLS estimators would often not fit the data well.

The k component partial least squares estimator $\hat{\boldsymbol{\beta}}_{kPLS}$ estimates $\boldsymbol{\beta}_{kPLS}$ for $k = 1, ..., p$. Cook, Helland, and Su (2013) showed that if $\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$, then $\boldsymbol{\beta}_{kPLS} = \boldsymbol{\beta}_{OLS}$ for $k = 1, ..., p$. Note that $\boldsymbol{\beta}_{1PLS} = \boldsymbol{\beta}_{OPLS}$ and $\boldsymbol{\beta}_{pPLS} = \boldsymbol{\beta}_{OLS}$. Typically $\boldsymbol{\beta}_{OPLS} \neq \boldsymbol{\beta}_{OLS}$. Thus it is possible that all p vectors $\boldsymbol{\beta}_{kPLS}$ differ. In limited simulations with p fixed and the OLS estimator well conditioned, the model selection PLS estimator $\hat{\boldsymbol{\beta}}_{MSPLS}$ did appear to estimate $\boldsymbol{\beta}_{OLS}$.

The Bet on Sparsity Principle: Hastie, Tibshirani, and Wainwright (2015, p. 2) state that the "bet on sparsity principle" is *use a procedure that does well in sparse problems, since no procedure does well in dense problems.* Here the dense (or abundant) problem refers to the population generating model. Estimating the optimal population generating model or the model $Y|\boldsymbol{x}$ may be too difficult for a given dense problem, but many suboptimal models, including sparse fitted models, may be useful. For regression models with iid cases, the $Y_1, ..., Y_n$ are iid, and the useful suboptimal *null model* omits all of the predictors. For high dimensional data, a reasonable goal is to find a regression model that greatly outperforms the null model.

Next, consider sparse high dimensional estimators with $\boldsymbol{\beta}_E = \boldsymbol{\beta}_{OLS}$, such as E=lasso. Suppose model (1) holds with iid cases, $\text{Cov}(\boldsymbol{x}) = diag(1,2,...,p)$, and $n \geq 10a_S$. Hence $p - a_S$ of the elements of $\boldsymbol{\beta}_{OLS}$ are equal to zero. Then $\text{Cov}(\boldsymbol{x}, Y) = \text{Cov}(\boldsymbol{x})\boldsymbol{\beta}_{OLS}$, and at least 90% of the predictors are uncorrelated with Y. If $p > 100n$, then for lasso, at least 99% of the predictors are uncorrelated with Y since lasso uses at most $a = n$ predictors. Hence for sparse models, often $\boldsymbol{\beta}_E \neq \boldsymbol{\beta}_{OLS}$ for high dimensional data. However, if data splitting with lasso variable selection is used to find model I, model iii) $Y|\boldsymbol{\beta}_I^T \boldsymbol{x}$ will often be useful. Rathnayake and Olive (2021) proved that for fixed p and model (1), lasso and elastic net variable selection estimators are $\sqrt{n}$ consistent estimators of $\boldsymbol{\beta}_{OLS}$ if lasso and elastic net are consistent estimators of $\boldsymbol{\beta}_{OLS}$.

Theorem 3 showed that sparse fitted models can do well in dense problems. The multitude of models result also helps explain why sparse fitted models can be useful even when the population generating model is not sparse. Sparse variable selection models are interesting, since data splitting can be used for testing and confidence regions, and the submodel can often be checked with plots. The sparse fitted lasso model I can be more useful than the sparse lasso variable selection model if that model is ill conditioned. For example for multiple linear regression, if $(\boldsymbol{X}_I^T \boldsymbol{X}_I)^{-1}$ is ill conditioned.

Fitted Models Tend to Be Suboptimal: Typically, the population generating model i) is not needed to be known and $Y|\boldsymbol{x}$ is not needed to be known for regression model iii) to be useful. The above result is useful for explaining why regression models work. Suppose that the population model was generated with response variable Z and predictors $w_1, ..., w_k$, but the fitted model uses response variable Y and predictors $x_1, ..., x_p$. The fitted model could be missing predictors or have many unnecessary predictors, and Y and the x_i could be measured on the wrong scale, for example perhaps $Y = \exp(Z)$ is used when $Z = \log(Y)$ should be used. However, the fitted model can still approximate the observed data (the data actually used) well. For example, if $Y|\boldsymbol{\beta}_{GLM}^T \boldsymbol{x} \approx D(\alpha_{GLM} + \boldsymbol{x}^T \boldsymbol{\beta}_{GLM}, \boldsymbol{\gamma}_{GLM})$, then the GLM can still be useful. Since Z and the w_j in the population generating model are unknown, we have over one hundred years evidence that low dimensional regression models are often useful, even if the population generating model is dense. We can often get a near optimal model on the data $(Y_i, \boldsymbol{x}_i)$ actually used if $n \geq Kp$ with $K \geq 10$. If $p > n$, obtaining a near optimal model on the data actually used may be too difficult.

On the other hand, fitted models tend to be suboptimal compared to what could be done if the variables in the population generating model were known. In particular, leaving out important predictors can result in either a suboptimal model that fits the data well, or in a suboptimal model that does not fit the data well. For example, if an estimated model without interactions fits the data well, then that model can be a useful suboptimal model even if the population generating model contains interactions. This explanation (for why pairwise and higher interactions often do not need to be included in the fitted model) may be more compelling than the explanation that the sample size n is too small to include interactions. Suppose $\hat{\boldsymbol{\beta}}_E$ is a good estimator of $\boldsymbol{\beta}_E$. Then a useful check on the fitted model is the response plot of $\hat{\alpha}_E + \hat{\boldsymbol{\beta}}_E^T \boldsymbol{x}$ versus Y on the vertical axis. In high dimensions, if variable selection and data splitting results in a sparse model I, then the response plot can be made for the data in the validation set V. Residual plots are also often useful for checking the fitted model. See Olive (2013) for more information about the response plot.

The suboptimality of models means that there is potential for better models and Statistical methods to be found. In particular, OPLS, PLS, lasso variable selection, and envelopes models (see Cook (2018)) appear to be less suboptimal than many competitors. Using model iii) (instead of models i) and ii)) *greatly increases the scope of data splitting, sparse fitted models, and PLS.* Consider estimators that fit J models, for example PLS, PCR, and lasso with a grid. The multitude of models result suggests that using a holdout sample or k-fold cross validation may work better than other methods for selecting a model selection estimator since one or several of the J models could be useful

for prediction. Also see Chetverikov, Liao, and Chernozhukov (2022). The following two sections help illustrate the ideas of this paragraph.

4 Sequential Data Splitting

The sequential data splitting algorithm is simple. Let $\lfloor x \rfloor$ be the integer part of x, e.g. $\lfloor 7.7 \rfloor = 7$. Denote the ceiling function by $\lceil x \rceil$, e.g. $\lceil 7.7 \rceil = 8$. Initially, randomly divide the data set into two sets: H_1 with $n_1 \leq n/2$ cases and V_1 with $n - n_1$ cases. Apply lasso on H_1 to get a set of a_1 predictors, including a constant if a constant is in the model. If $n_1 \geq 10a_1$, set $H = H_1$ and $V = V_1$. Otherwise, randomly select n_1 cases from V_1 to add to H_1 to form H_2. Let V_2 have the remaining cases from V_1. Apply lasso on H_2 to get a set of a_2 predictors. If $n_2 \geq 10a_2$, set $H = H_2$ and $V = V_2$. Continue in this manner, forming sets $(H_1, V_1), (H_2, V_2), ..., (H_d, V_d)$ where H_i has $n_i = in_1$. Stop when $n_d \geq 10a_d$ or $n_{d+1} > \lfloor (n - J)/2 \rfloor$ where $J = 5$ was often used in the simulations. For the second case, use $n_d = \lfloor (n - J)/2 \rfloor$. Then $H = H_d$ and $V = V_d$. Use the model I_d with a_d predictors as the full model for inference with the data in $V = V_d$.

Lasso uses up to n_d active predictors and a constant. If J is an integer between 0 and 5, set $n_1 = \max(1, \lfloor (n - J)/2 \rfloor)$ if $n < 40$. Otherwise, we often used $n_1 = 30$, but changed n_1 to $\lfloor n/2000 \rfloor$ if initially $\lfloor n/(2n_1) \rfloor > 1000$. If $n >> p$, let $n_1 = Kp$ with K a positive integer, such as $K = 10$ or $K = 20$, or use $n_1 \approx Kp \approx n/(2M)$ with $M = \lceil n/(2Kp) \rceil$. If n/p is not large, options include $M = 10$ or $n_1 = Ka_0$ where a_0 is, for example, a guess of a lower bound for the number of active predictors.

5 EXAMPLE AND SIMULATIONS

Example. For the Johnson (2021) bodyfat multiple linear regression data set with 18 body measurements as predictors and $n = 184$, many linear combinations $\boldsymbol{\eta}^T \boldsymbol{x}$ have a linear relationship with the response Y = percentage of body fat measured from college women. Using a seed and $n_1 = 30$, lasso selected 5 predictors. With $n_2 = 60$, lasso selected 12 predictors. With $n_3 = n_d = 90$, lasso selected 8 predictors: constant, Height, BMI (body mass index), and the circumference measurements Hips, Waist, PThigh (proximal thigh), Wrist, and Knee. Then OLS on the validation set used 94 cases with $R^2 = 0.63$, and likely used more predictors than necessary. The response and residual plots looked good. All subsets selection with the C_p criterion picked three predictors with $n_1 = n_d = 30$: constant, BMI, and PThigh with $R^2 = 0.59$ for the validation set.

Next we did a small simulation study where there was often underfitting, in that a predictor that generated the model was not selected, but the prediction intervals still had good coverage with short length. *Hence the model selected by lasso was still good for prediction.* The programs give the mean n_d: the number of cases used in H, the mean a_d where a_d is the number of nonzero lasso coefficients including the constant for lasso applied to the n_d cases in H_d, and $k = a_d - 1$. The program also computed large sample Olive, Rathnayake, and Haile (2022) 95% prediction intervals (PIs) for lasso applied to all n cases (lsapi), lasso variable selection applied to all n cases (LVSpi), lasso applied

to V_d (lsplitpi), and the model selected using H applied to V_d (splitpi). The second and fourth models used OLS, a GLM, or Weibull regression applied to the n cases or the cases in V_d. The coverage and average length of the prediction intervals was given. We also computed the number of times lasso and lasso variable selection did not underfit. A value of noundfit greater than 4500 indicates that in over 90% of the 5000 runs, lasso did not underfit. The simulations used 5000 runs, and $n_1 = 30$ was used unless stated otherwise. More simulations are in Zhang (2022).

The prediction intervals were computed roughly as follows. If $Y \sim D(\boldsymbol{x}^T\boldsymbol{\beta}, \boldsymbol{\theta})$, then apply a prediction interval to a bootstrap sample of size B: $Y_1^*, ..., Y_B^*$ where the Y_i^* are iid $D(\boldsymbol{x}^T\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\theta}})$. For multiple linear regression, obtain the n_c residuals r_j and apply a prediction interval to $\hat{Y}_f + r_1, ..., \hat{Y}_f + r_{n_c}$ where $\hat{Y}_f = \boldsymbol{x}_f^T\hat{\boldsymbol{\beta}}$ and $n_c = n$ or $n = n_V$ depending on whether all n cases or data splitting was used for the prediction interval.

The full model was simulated as in Pelawa Watagoda and Olive (2021) and Olive, Rathnayake, and Haile (2022). This section and the programs use a change in notation: if $\boldsymbol{\beta}_c = (\alpha\ \boldsymbol{\beta}^T)^T$ and $\boldsymbol{w} = (1\ \boldsymbol{x}^T)^T$ in Section 1 of this paper, then the program notation is $\boldsymbol{\beta} = \boldsymbol{\beta}_c$ and $\boldsymbol{x} = \boldsymbol{w}$ are $p \times 1$ vectors, $\beta_1 = \alpha$, and $\boldsymbol{u} = \boldsymbol{x}$ is a $(p-1) \times 1$ vector. For the simulations, generating $\boldsymbol{x}^T\boldsymbol{\beta}$ is important for regression models other than multiple linear regression. For example, for binomial logistic regression, typically $-5 \leq \boldsymbol{x}^T\boldsymbol{\beta} \leq 5$ or there can be problems with the maximum likelihood estimator. 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 $m = p-1$ 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 $\boldsymbol{z}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $\mathrm{Cov}(\boldsymbol{z}_i) = \boldsymbol{\Sigma}_{\boldsymbol{z}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ 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 $cor(z_i, z_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ for $i \neq j$. Then $\sum_{j=1}^{k} z_j \sim N(0, k\sigma_{ii} + k(k-1)\sigma_{ij}) = N(0, v^2)$. For multiple linear regression, let $\boldsymbol{u} = \boldsymbol{z}$. For the other regression models, let $\boldsymbol{u} = a\boldsymbol{z}/v$. Then $cor(x_i, x_j) = \rho$ 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 $\boldsymbol{u}_i$ cluster about the line in the direction of $(1, ..., 1)^T$. Let $SP = \boldsymbol{x}^T\boldsymbol{\beta} = \beta_1 + 1x_{i,2} + \cdots + 1x_{i,k+1} \sim N(\beta_1, a^2)$ for $i = 1, ..., n$. Hence $\boldsymbol{\beta} = (\beta_1, 1, .., 1, 0, ..., 0)^T$ with β_1, k ones and $p - k - 1$ zeros. The default settings for Poisson regression use $\beta_1 = 1 = a$. The default settings for binomial regression with $m = 4$ trials use $\beta_1 = 1$ and $a = 4/3$. In the Table 3 caption, these values correspond to int=1, $a = 4/3$, and $m = 4$. The bootstrap sample for the prediction intervals had size $B = 1000$.

For the Weibull regression model, there is no constant since the constant appears in the corresponding accelerated failure time model, which is a multiple linear regression model with right censored response $\log(Y)$. The data was generated as for the Poisson and Binomial regression, but replace $\boldsymbol{u}$ by $\boldsymbol{x}$ and $p-1$ by p. Let $SP = \boldsymbol{x}_i^T\boldsymbol{\beta} = 1x_{i,1} + \cdots + 1x_{i,k} \sim N(0, a^2)$ for $i = 1, ..., n$. The simulations use $a = 1$ where $\boldsymbol{\beta} = (1, ..., 1, 0, ..., 0)^T$ with k ones and $p - k$ zeros. The right censored Weibull regression data was generated in a manner similar to Zhou (2001) with $\gamma = 1$. The caption in Table 4 gives $a = 1$. The values gam and clam in the caption control the Weibull distribution and the amount of right censoring.

Table 2: prsplit

n	p/k	psi	mnnd/mnad	lsapi	LVSpi	lsplitpi	splitpi	noundfit
100	4	0.0000	34.5757	0.9872	0.9924	0.9955	0.9864	5000
	1		2.9438	7.5355	7.7445	7.1744	7.8245	
100	4	0.8000	32.8250	0.9887	0.9837	0.9967	0.9857	4578
	1		2.4602	7.8907	7.1752	8.3510	8.2722	
100	20	0.0000	40.5895	0.9934	0.9797	0.9819	0.9800	5000
	1		3.7327	8.2845	7.8980	9.0938	7.4559	
100	20	0.6000	38.6768	0.9986	0.9909	0.9898	0.9783	3834
	1		3.3482	8.1733	8.2308	8.7990	8.2294	
100	100	0.0000	43.7107	0.9946	0.9788	0.9943	0.9695	5000
	1		5.3520	8.4802	7.7537	7.6238	7.8035	
100	100	0.3000	44.5573	0.9799	0.9819	0.9870	0.9580	4165
	1		7.1632	8.0935	7.1896	7.8179	7.6183	
100	10	0.0000	45.6543	0.9912	0.9902	0.9800	0.9745	4896
	9		8.8543	8.7440	8.0106	8.1908	7.5920	
100	20	0.0000	43.6966	0.9841	0.9733	0.9685	0.9574	2363
	19		10.6884	8.1987	7.7602	8.7657	7.9307	
1000	4	0.0000	36.8298	0.9823	0.9808	0.9883	0.9840	4993
	1		2.9703	7.5124	7.6328	8.5947	8.0170	
1000	4	0.8000	34.1201	0.9911	0.9794	0.9855	0.9792	3718
	1		2.5867	7.5125	6.9803	7.9171	6.9004	
1000	20	0.0000	56.6908	0.9869	0.9777	0.9908	0.9842	4978
	1		3.3662	7.0329	7.7289	7.9915	8.1688	
1000	20	0.5000	56.4071	0.9834	0.9915	0.9849	1.0000	3418
	1		3.9243	7.9746	8.1575	7.1918	7.2256	
1000	1000	0.0000	95.2273	0.9856	0.9908	0.9924	0.9744	4789
	1		5.2895	8.0867	8.7953	7.8825	7.8767	
1000	1000	0.1000	110.0411	0.9880	0.9884	0.9902	0.9881	3240
	1		8.8354	7.3550	8.5021	8.2190	8.1900	
1000	10	0.3160	74.3965	0.9920	0.9920	0.9899	0.9836	24
	9		7.0638	7.5158	7.9454	8.1729	8.1484	

Table 3: brsplit, int=1, a=4/3, m=4, B=1000

n	p/k	psi	mnnd/mnad	lsapi	LVSpi	lsplitpi	splitpi	noundfit
100	4	0.0000	33.9066	0.9914	0.9896	0.9910	0.9872	4996
	1		2.6168	2.6678	2.6014	2.6760	2.5832	
100	4	0.8000	32.6724	0.9894	0.9888	0.9908	0.9884	3685
	1		2.5686	2.6048	2.5800	2.6032	2.5630	
100	20	0.0000	40.9616	0.9932	0.9850	0.9904	0.9752	4991
	1		4.6866	2.8414	2.6524	2.8588	2.5972	
100	20	0.5000	41.4784	0.9934	0.9904	0.9908	0.9820	3163
	1		4.5670	2.7706	2.7130	2.7524	2.6444	
100	100	0.0000	43.4521	0.9854	0.9844	0.9945	0.9843	4987
	1		5.6745	2.6474	2.4433	2.9443	2.5556	
100	100	0.2000	45.6434	0.9948	0.9782	0.9886	0.9624	3881
	1		8.7554	2.8426	2.6696	2.8244	2.5558	
1000	4	0.0000	37.1640	0.9876	0.9862	0.9866	0.9858	4993
	1		2.6016	2.4730	2.4594	2.4752	2.4546	
1000	4	0.8000	34.2360	0.9862	0.9860	0.9870	0.9854	3718
	1		2.5336	2.4468	2.4352	2.4442	2.4312	
1000	20	0.0000	56.7660	0.9856	0.9840	0.9858	0.9824	4978
	1		3.7276	2.4988	2.4460	2.4978	2.4400	
1000	20	0.5000	56.0040	0.9874	0.9872	0.9890	0.9878	3418
	1		4.0666	2.4560	2.4374	2.4614	2.4392	
1000	1000	0.0000	95.0734	0.9902	0.9820	0.9898	0.9822	4789
	1		5.3892	2.6302	2.4922	2.6320	2.4834	
1000	10	0.4000	63.4080	0.9870	0.9856	0.9862	0.9858	4
	9		5.2826	2.5050	2.4854	2.5008	2.4890	

Table 4: PHsplit, n=100, J=5, a=1, gam=1, B=1000, clam=0.1

n	p/k	psi	mnnd/mnad	LVSpi	splitpi	noundfit
100	4	0.00	31.7646	0.9550	0.9552	4913
	1		1.8314	5.5483	5.5033	
100	4	0.80	30.7004	0.9574	0.9576	156
	1		1.6076	5.5956	5.5384	
100	20	0.00	36.3172	0.9326	0.9328	4747
	1		2.7178	5.9745	25.4093	
100	20	0.60	33.4238	0.9510	0.9512	1289
	1		2.2570	5.7760	9.8008	
100	10	0.00	39.1528	0.9518	0.9520	608
	9		4.2938	6.9368	6.6001	
100	50	0.00	35.1850	0.7750	0.7752	0
	19		2.1098	346.39	332.04	

Table 5: mlrsplit, J=5, type=3

n	p/k	psi	mnnd/mnad	lsapi	LVSpi	lsplitpi	splitpi	noundfit
100	4	0.8000	33.3354	0.9676	0.9672	0.9768	0.9764	4306
	1		2.6874	4.0502	4.0545	4.6570	4.6614	
100	20	0.6000	44.1712	0.9762	0.9730	0.9780	0.9746	4690
	1		5.4618	4.5475	4.5611	5.2808	5.2408	
100	100	0.3000	46.3812	0.9780	0.9660	0.9804	0.9518	4939
	1		8.9376	4.7967	4.6882	5.5010	5.1129	
100	10	0.0000	47.0000	0.9752	0.9752	0.9786	0.9786	5000
	9		10.0000	5.0524	5.0498	5.8915	5.8846	
100	20	0.0000	46.9966	0.9810	0.9804	0.9794	0.9804	4994
	19		19.9950	5.6460	5.6398	6.8107	6.7928	
1000	4	0.0000	38.1120	0.9556	0.9564	0.9570	0.9576	4962
	1		2.5994	3.1418	3.1398	3.1485	3.1471	
1000	4	0.8000	36.0180	0.9558	0.9564	0.9562	0.9562	4321
	1		2.6694	3.1302	3.1306	3.1333	3.1333	
1000	20	0.0000	55.7820	0.9516	0.9474	0.9514	0.9490	4915
	1		3.5346	3.2048	3.2349	3.2139	3.2449	
1000	20	0.5000	64.9500	0.9548	0.9548	0.9536	0.9528	4951
	1		4.7108	3.1909	3.1942	3.2011	3.2054	
1000	10	0.3160	120.0000	0.9574	0.9570	0.9592	0.9580	5000
	9		10.0000	3.3726	3.3244	3.4032	3.3574	
1000	1000	0.0000	82.3860	0.9558	0.9460	0.9572	0.9440	4551
	1		4.5362	3.3778	3.4472	3.4008	3.4644	

Data splitting is useful for hypothesis testing and confidence intervals. Two lines per run are shown in each table. The first line gives the average coverage of the prediction intervals while the second line gives the average length. The prediction intervals were used as a check for whether lasso was finding a useful model for prediction (coverage near 0.95) even if underfitting was present. This result could occur for at least two reasons. First, as ψ increases to 1, the predictor variables are roughly $x_i = x_j + e_{ij}$ where the error magnitude rapidly gets close to 0 as $\psi \to 1$. Hence omitting some good predictors may not be a problem for prediction. Second, for some regression models, there are many linear combinations that give a good fit. See Theorem 3.

For multiple linear regression, 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). Only distribution iii) is not symmetric. The lengths of 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.

For the regression methods, first consider $k = 1$. Then there was little underfit for $\psi = 0$. The amount of underfitting tended to increase with ψ, and to be worse with larger p. With $n = 1000$, not much more than 10% of the cases were used for H. For larger values of k, lasso often underfit, especially if $k = p - 1$ and $n/k < 10$. See Table 2 for Poisson regression, see Table 3 for Binomial regression, where with m=4, a 100% PI for Y_f is [0,4] with length 4. The nominal 95% PIs were shorter than 4 in Table 3. See Table 4 for proportional hazards regression where the prediction intervals were made for Weibull regression. Hence only two prediction intervals are given. In Table 4, sometimes the two PI lengths differed. For the last two lines of Table 4, there was serious underfitting with low PI coverage and large PI length. See Table 5 for multiple linear regression where usually the data splitting PI and PI using all n cases had similar average lengths, but there were data configurations where using all n cases can give a much smaller length and better coverage.

6 CONCLUSIONS

Regression models, such as $Y|\boldsymbol{\beta}^T\boldsymbol{x}$ or $\boldsymbol{y}|(\boldsymbol{\eta}_1^T\boldsymbol{x}, ..., \boldsymbol{\eta}_d^T\boldsymbol{x})$, tend to be useful when they fit the data (actually used) well, although the models are often suboptimal. For data splitting, the $(Y_i, \boldsymbol{x}_{i,I})$ are the data actually used on the validation set. This paper shows that nonsparse models, such as OPLS, can be useful for inference even when n/p is not large. There are many problems with assuming that the regression model estimates a population generating model. Removing this assumption greatly increases the scope of data splitting, sparse fitted models, and nonsparse dimension reduction model selection estimators such as partial least squares. In particular, sparse fitted models, like lasso, tend to give poor approximations to a nonsparse population generating model, but this paper shows that the sparse fitted model can still be useful if data splitting is used.

The multitude of models result is useful and simple. For example, it is known that k component PCR tends not to estimate $\boldsymbol{\beta}_{OLS}$ for small k, but results from this paper suggest that PCR will often give a linear model. See, for example, Agarwal et al. (2021). For fixed p, lasso in `glmnet` tends to be at best $n^{1/4}$ consistent for multiple linear regression, while large sample theory for lasso and elastic net does not appear to be available

Table 6: Regression Summary

low dimensions	data splitting: sparse I	high dimensional error
general: $\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
data splitting: $\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
lasso: $\boldsymbol{\beta}_{lasso}$	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	$\boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$
OPLS: $\boldsymbol{\beta}_{OPLS} = \lambda \boldsymbol{\Sigma}_{\boldsymbol{x},Y}$	$\boldsymbol{\beta}_{I,OPLS} = \lambda_I \boldsymbol{\Sigma}_{\boldsymbol{x}_I,Y}$	$\boldsymbol{\beta}_{OPLS} = \boldsymbol{\beta}_{OLS}$
MMLE: $\boldsymbol{\beta}_{MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u},Y}$	$\boldsymbol{\beta}_{I,MMLE} = \boldsymbol{\Sigma}_{\boldsymbol{u}_I,Y}$	$\boldsymbol{\beta}_{MMLE} = \boldsymbol{\beta}_{OLS}$
dense regression:	$\boldsymbol{\beta}_I(\boldsymbol{x}_I, Y)$	no method works

for GLMs and Cox regression. See Guan and Tibshirani (2020). For fixed p, Rathnayake and Olive (2021) have the interesting result that if the sparse estimator is consistent for $\boldsymbol{\beta}$, then the sparse variable selection estimator (that applies OLS, the GLM, or the Cox regression estimator to the predictors with nonzero coefficients) is $\sqrt{n}$ consistent for $\boldsymbol{\beta}$. Thus $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{x}, Y) = \boldsymbol{\beta}_{I,0}(\boldsymbol{x}_I, Y)$. Table 6 summarizes what the regression estimators tend to estimate in low dimensions or after data splitting with a sparse fitted model I. The third column of Table 6 gives some common errors in the high dimensional literature. Often the regularity conditions are too strong for low dimensional results to hold in high dimensions.

Tay, Narasimhan, and Hastie (2021) describe lasso for several regression models. Hastie, Tibshirani, and Tibshirani (2020) compared lasso, lasso variable selection, forward selection, and a type of best subset selection using all n cases, and concluded lasso variable selection worked best. Pelawa Watagoda and Olive (2021) also found that lasso variable selection and forward selection were among the best methods.

Simulations were done in *R*. See R Core Team (2020). The collection of Olive (2023) *R* functions *slpack*, available from (http://parker.ad.siu.edu/
Olive/slpack.txt), has some useful functions for the inference. The functions for regression data splitting are `mlrsplitsim`, `prsplit`, `brsplitsim`, and `PHsplitsim`. These functions used the Friedman et al. (2015) `glmnet` package.

Acknowledgments

The authors thank the referees for their work.

REFERENCES

Agarwal, A., Shah, D., Shen, D., and Song, D. (2021), "On Robustness of Principal Component Regression," *Journal of the American Statistical Association,* 116, 1731-1745.

Agresti, A. (2002), *Categorical Data Analysis*, 2nd ed., Wiley, Hoboken, NJ.

Basa, J., Cook, R.D., Forzani, L., and Marcos, M. (2022), "Asymptotic Distribution of One-Component Partial Least Squares Regression Estimators in High Dimensions," *The Canadian Journal of Statistics*, to appear.

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criterion for Model Selection with Large Model Spaces," *Biometrika*, 95, 759-771.

Chetverikov, D., Liao, Z., and Chernozhukov, V. (2022), "On Cross Validated Lasso in High Dimensions," *The Annals of Statistics*, 49, 1300-1317.

Chun, H., and Keleş, S. (2010), "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Predictor Selection," *Journal of the Royal Statistical Society, B,* 72, 3-25.

Cook, R.D. (2018), *An Introduction to Envelopes: Dimension Reduction for Efficient Estimation in Multivariate Statistics,* Wiley, Hoboken, NJ.

Cook, R.D., and Forzani, L. (2018), "Big Data and Partial Least Squares Prediction," *The Canadian Journal of Statistics*, 46, 62-78.

Cook, R.D., and Forzani, L. (2019), "Partial Least Squares Prediction in High-Dimensional Regression," *The Annals of Statistics,* 47, 884-908.

Cook, R.D., Forzani, L., and Rothman, A. (2013), "Prediction in Abundant High-Dimensional Linear Regression," *Electronic Journal of Statistics*, 7, 3059-3088.

Cook, R.D., Helland, I.S., and Su, Z. (2013), "Envelopes and Partial Least Squares Regression," *Journal of the Royal Statistical Society, B*, 75, 851-877.

Cox, D.R. (1972), "Regression Models and Life-Tables," *Journal of the Royal Statistical Society, B,* 34, 187-220.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," (with discussion), *The Annals of Statistics,* 32, 407-451.

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society, B*, 70, 849-911.

Fan, J., and Song, R. (2010), "Sure Independence Screening in Generalized Linear Models with np-Dimensionality," *The Annals of Statistics*, 38, 3217-3841.

Friedman, J., Hastie, T., Hoefling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization," *Annals of Applied Statistics*, 1, 302-332.

Friedman, J., Hastie, T., Simon, N., and Tibshirani, R. (2015), *glmnet: Lasso and Elastic-net Regularized Generalized Linear Models, R* Package version 2.0, (http://cran.r-project.org/package=glmnet).

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1-22.

Guan, L., and Tibshirani, R. (2020), "Post Model-Fitting Exploration via a "Next-Door" Analysis," *Canadian Journal of Statistics*, 48, 447-470.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2020), "Best Subset, Forward Stepwise or Lasso? Analysis and Recommendations Based on Extensive Comparisons," *Statistical Science*, 35, 579-592.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015), *Statistical Learning with Sparsity: the Lasso and Generalizations*, CRC Press Taylor & Francis, Boca Raton, FL.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021), *An Introduction to Statistical Learning with Applications in R*, 2nd ed., Springer, New York, NY.

Johnson, R.W. (2021), "Fitting Percentage of Body Fat to Simple Body Measurements: College Women," *Journal of Statistics and Data Science Education*, 29, 304-316.

Meinshausen, N. (2007), "Relaxed Lasso," *Computational Statistics & Data Analysis*, 52, 374-393.

Nelder, J.A., and Wedderburn, R.W.M. (1972), "Generalized Linear Models," *Journal of the Royal Statistical Society, A*, 135, 370-384.

Olive, D.J. (2013), "Plots for Generalized Additive Models," *Communications in Statistics: Theory and Methods*, 42, 2610-2628.

Olive, D.J. (2023), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

Olive, D.J., and Hawkins, D.M. (2005), "Variable Selection for 1D Regression Models," *Technometrics*, 47, 43-50.

Olive, D.J., Rathnayake, R.C., and Haile, M.G. (2022), "Prediction Intervals for GLMs, GAMs, and Some Survival Regression Models," *Communications in Statistics: Theory and Methods*, 51, 8012-8026.

Pelawa Watagoda, L.C.R., and Olive, D.J. (2021), "Comparing Six Shrinkage Estimators with Large Sample Theory and Asymptotically Optimal Prediction Intervals," *Statistical Papers*, 62, 2407-2431.

Qi, X., Luo, R., Carroll, R.J., and Zhao, H. (2015), "Sparse Regression by Projection and Sparse Discriminant Analysis," *Journal of Computational and Graphical Statistics,* 24, 416-438.

R Core Team (2020), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Rathnayake, R.C., and Olive, D.J. (2021), "Bootstrapping Some GLMs and Survival Regression Models after Variable Selection," *Communications in Statistics: Theory and Methods*, to appear.

Rinaldo, A., Wasserman, L., and G'Sell, M. (2019), "Bootstrapping and Sample Splitting for High-Dimensional, Assumption-Lean Inference," *The Annals of Statistics*, 47, 3438-3469.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2011), "Regularization Paths for Cox's Proportional Hazards Model via Coordinate Descent," *Journal of Statistical Software*, 39, 1-13.

Su, Z., and Cook, R.D. (2012), "Inner Envelopes: Efficient Estimation in Multivariate Linear Regression," *Biometrika*, 99, 687-702.

Tay, J.K., Narasimhan, B. and Hastie, T. (2021), "Elastic Net Regularization Paths for All Generalized Linear Models," *Journal of Statistical Software*, to appear.

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society, B*, 58, 267-288.

Wieczorek, J., and Lei, J. (2022), "Model-Selection Properties of Forward Selection and Sequential Cross-Validation for High-Dimensional Regression," *Canadian Journal of Statistics*, 50, 454-470.

Wold, H. (1975), "Soft Modelling by Latent Variables: the Non-Linear Partial Least Squares (NIPALS) Approach," *Journal of Applied Probability*, 12, 117-142.

Zhang, L. (2022), "Data Splitting Inference," Ph.D. Thesis, Southern Illinois University. See (http://parker.ad.siu.edu/Olive/slinglingphd.pdf).

Zhou, M. (2001), "Understanding the Cox Regression Models with Time–Change Covariates," *The American Statistician*, 55, 153-155.

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society Series, B,* 67, 301-320.