Testing with the One Component Partial Least Squares and the Marginal Maximum Likelihood Estimators

David J. Olive and Abdulaziz Alshammari [*]
Southern Illinois University

April 1, 2024

Abstract

We derive some large sample theory for the marginal maximum likelihood estimator for multiple linear regression. Then testing is considered for that estimator and the one component partial least squares estimator, including some high dimensional tests. Testing with these two estimators for the multiple linear regression model with heterogeneity and for the single index model is also considered.

KEY WORDS: Data splitting, dimension reduction, high dimensional data, lasso, single index model.

1 INTRODUCTION

This section reviews multiple linear regression models, including variable selection and data splitting. Consider a multiple linear regression model with response variable Y and predictors $\boldsymbol{x} = (x_1, ..., x_p)$. Then there are n cases $(Y_i, \boldsymbol{x}_i^T)^T$, and the sufficient predictor $SP = \alpha + \boldsymbol{x}^T\boldsymbol{\beta}$. For these regression models, the conditioning and subscripts, such as i, will often be suppressed. Ordinary least squares (OLS) is often used for the multiple linear regression (MLR) model.

Let the first multiple linear regression model be

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{1}$$

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

[*]David J. Olive is Professor, School of Mathematical & Statistical Sciences, Southern Illinois University, Carbondale, IL 62901, USA.

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.

Let the second multiple linear regression model be $Y|\boldsymbol{x}^T\boldsymbol{\beta} = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$ or $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ or

$$Y_i = \alpha + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{2}$$

for $i = 1, ..., n$. Let the e_i be as for model (1). In matrix form, this model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}, \tag{3}$$

where $\boldsymbol{Y}$ is an $n \times 1$ vector of dependent variables, $\boldsymbol{X}$ is an $n \times (p+1)$ matrix with ith row $(1, \boldsymbol{x}_i^T)$, $\boldsymbol{\phi} = (\alpha, \boldsymbol{\beta}^T)^T$ is a $(p+1) \times 1$ vector , and $\boldsymbol{e}$ is an $n \times 1$ vector of unknown errors. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix.

For estimation with ordinary least squares, 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 for model (3) are $\hat{\boldsymbol{\phi}}_{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$, $\hat{\alpha}_{OLS} = \overline{Y} - \hat{\boldsymbol{\beta}}_{OLS}^T\overline{\boldsymbol{x}}$, 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 the one component partial least squares (OPLS) estimator $\hat{\boldsymbol{\beta}}_{OPLS} = \hat{\lambda}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ estimates $\lambda\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\beta}_{OPLS}$ where

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}} \quad \text{and} \quad \hat{\lambda} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}} \tag{4}$$

for $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} \neq \boldsymbol{0}$. If $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{0}$, then $\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$. Also see Basa, Cook, Forzani, and Marcos (2022) and Wold (1975). Olive and Zhang (2024) derived the large sample theory for $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and OPLS under milder regularity conditions than those in the previous literature, where $\boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. The OPLS estimator is computed from the OLS simple linear regression of Y on $W = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}$, giving $\hat{Y} = \hat{\alpha}_{OPLS} + \hat{\lambda}W = \hat{\alpha}_{OPLS} + \hat{\boldsymbol{\beta}}_{OPLS}^T\boldsymbol{x}$.

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

$$\hat{\boldsymbol{\beta}}_{MMLE} = [diag(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})]^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}. \tag{5}$$

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

$$\hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{t}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} = \boldsymbol{I}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{t}, Y) \tag{6}$$

where $(\boldsymbol{t}, Y)$ denotes that Y was regressed on $\boldsymbol{t}$, and $\boldsymbol{I}$ is the $p \times p$ identity matrix.

Sparse regression methods can be used for variable selection even if n/p is not large: the OLS submodel 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 et al. (2004, p. 421), Meinshausen (2007, p. 376), Qi et al. (2015), Tay, Narasimhan, and Hastie (2023), Rathnayake and Olive (2023), Tibshirani (1996), and Zou and Hastie (2005).

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.

Olive and Zhang (2024) proved that there are often many valid population models for multiple linear regression, gave theory for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}$ and OPLS, gave theory for data splitting estimators, and gave some theory for the MMLE for multiple linear regression under the constant variance assumption.

Section 2 gives some large sample theory, while Section 3 considers tests of hypotheses.

2 Large Sample Theory

Olive and Zhang (2024) derived the large sample theory for $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and OPLS, including some high dimensional tests for low dimensional quantities such as $H_O : \beta_i = 0$ or $H_0 : \beta_i - \beta_j = 0$. These tests depended on iid cases, but not on linearity or the constant variance assumption. Hence the tests are useful for multiple linear regression with heterogeneity. Data splitting uses model selection (variable selection is a special case) to reduce the high dimensional problem to a low dimensional problem.

Remark 1. The following result is useful for several multiple linear regression estimators. Let $\boldsymbol{w}_i = \boldsymbol{A}_n\boldsymbol{x}_i$ for $i = 1, ..., n$ where $\boldsymbol{A}_n$ is a full rank $k \times p$ matrix with $1 \leq k \leq p$.

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

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

The following Olive and Zhang (2024) theorem gives the large sample theory for $\hat{\boldsymbol{\eta}} = \widehat{\text{Cov}}(\boldsymbol{x}, Y)$, but the proof in this paper is new. 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 $\boldsymbol{\Sigma}_{\boldsymbol{w}}$.

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)

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \ \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{w}}), \tag{7}$$

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

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

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_k(\boldsymbol{0}, \lambda^2 \boldsymbol{A}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{A}^T). \tag{8}$$

Proof. Part a) is a special case of Theorem 2.

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

$$\boldsymbol{v}_i + (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\overline{Y} - \mu_Y) + (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \overline{Y}) + (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})(\overline{Y} - \mu_Y).$$

Thus $\boldsymbol{w}_i - \overline{\boldsymbol{w}} = \boldsymbol{v}_i - \overline{\boldsymbol{v}} + \boldsymbol{a}_i$ where

$$\boldsymbol{a}_i = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\overline{Y} - \mu_Y) + (\overline{\boldsymbol{x}} - \boldsymbol{\mu}_{\boldsymbol{x}})(Y_i - \overline{Y}) = O_P(n^{-1/2}).$$

Thus

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{w}_i - \overline{\boldsymbol{w}})(\boldsymbol{w}_i - \overline{\boldsymbol{w}})^T = \frac{1}{n}\sum_{i=1}^n (\boldsymbol{v}_i - \overline{\boldsymbol{v}})(\boldsymbol{v}_i - \overline{\boldsymbol{v}})^T + O_P(n^{-1/2}) = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2}).$$

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

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

Then $\lambda\boldsymbol{A}\boldsymbol{\eta} = \boldsymbol{0}$ under H_0, and

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

For the following theorem, consider a subset of k distinct elements from $\tilde{\boldsymbol{\Sigma}}$ or from $\hat{\boldsymbol{\Sigma}}$. Stack the elements into a vector, and let each vector have the same ordering. For example, the largest subset of distinct elements corresponds to

$$vech(\tilde{\boldsymbol{\Sigma}}) = (\tilde{\sigma}_{11}, ..., \tilde{\sigma}_{1p}, \tilde{\sigma}_{22}, ..., \tilde{\sigma}_{2p}, ..., \tilde{\sigma}_{p-1,p-1}, \tilde{\sigma}_{p-1,p}, \tilde{\sigma}_{pp})^T = [\tilde{\sigma}_{jk}].$$

For random variables $x_1, ..., x_p$, use notation such as $\overline{x}_j =$ the sample mean of the x_j, $\mu_j = E(x_j)$, and $\sigma_{jk} = Cov(x_j, x_k)$. Let

$$n\ vech(\tilde{\boldsymbol{\Sigma}}) = [n\ \tilde{\sigma}_{jk}] = \sum_{i=1}^{n} [(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)].$$

For general vectors of elements, the ordering of the vectors will all be the same and be denoted vectors such as $\tilde{\boldsymbol{c}} = [\tilde{\sigma}_{jk}]$, $\boldsymbol{c} = [\sigma_{jk}]$, $\boldsymbol{z}_i = [(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)]$, and $\boldsymbol{w}_i = [(x_{ij} - \mu_j)(x_{ik} - \mu_k)]$. Let $\overline{\boldsymbol{w}}_n = \sum_{i=1}^{n} \boldsymbol{w}_i/n$ be the sample mean of the $\boldsymbol{w}_i$. Assuming that $Cov(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{w}}$ exists, then $E(\boldsymbol{w}_i) = E(\overline{\boldsymbol{w}}_n) = \boldsymbol{c}$.

The following theorem proves that sample covariance matrices are asymptotically normal. The theorem may be a special case of the Su and Cook (2012) theory for the multivariate linear regression estimator when there are no predictors. When $p = 1$, the theory gives the large sample theory for the sample variance. See Olive (2014, pp. 276-277) and Bickel and Doksum (2007, p. 279). The Olive and Zhang (2024) large sample theory for $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ is also a special case. We use $Cov(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{d}}$ to avoid confusion with the $\boldsymbol{\Sigma}_{\boldsymbol{w}}$ used in Theorems 1 and 3.

Theorem 2. Assume the cases $\boldsymbol{x}_i$ are iid and that $Cov(\boldsymbol{w}_i) = \boldsymbol{\Sigma}_{\boldsymbol{d}}$ exists. Using the above notation with $\boldsymbol{c}$ a $k \times 1$ vector,

i) $\sqrt{n}(\tilde{\boldsymbol{c}} - \boldsymbol{c}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$.

ii) $\sqrt{n}(\hat{\boldsymbol{c}} - \boldsymbol{c}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$.

iii) $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + O_P(n^{-1/2})$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + O_P(n^{-1/2})$.

Proof. Note that $\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{c}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$ by the multivariate central limit theorem. i) Then

$$n\ \tilde{\boldsymbol{c}} = \sum_i [(x_{ij} - \overline{x}_j)(x_{ik} - \overline{x}_k)] = \sum_i [(x_{ij} - \mu_j + \mu_j - \overline{x}_j)(x_{ik} - \mu_k + \mu_k - \overline{x}_k)] =$$

$$\sum_i [(x_{ij} - \mu_j)(x_{ik} - \mu_k)] + \sum_i [(x_{ij} - \mu_j)(\mu_k - \overline{x}_k)] +$$

$$\sum_i [\mu_j - \overline{x}_j)(x_{ik} - \mu_k] + \sum_i [(\mu_j - \overline{x}_j)(\mu_k - \overline{x}_k)] = \sum_i \boldsymbol{w}_i - \boldsymbol{a}_n$$

$$\text{where}\ \ \boldsymbol{a}_\text{n} = [\text{n}(\overline{\text{x}}_\text{j} - \mu_\text{j})(\overline{\text{x}}_\text{k} - \mu_\text{k})] = [\sqrt{\text{n}}(\overline{\text{x}}_\text{j} - \mu_\text{j})\sqrt{\text{n}}(\overline{\text{x}}_\text{k} - \mu_\text{k})] = \text{O}_\text{P}(1).$$

By the multivariate Slutsky's theorem,

$$\sqrt{n}(\tilde{\boldsymbol{c}} - \boldsymbol{c}) = \sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{c}) + \boldsymbol{a}_n/\sqrt{n} \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{d}})$$

since $\boldsymbol{a}_n/\sqrt{n} = o_P(1)$.

iii) $\boldsymbol{w}_i = [(x_{ij}-\mu_j)(x_{ik}-\mu_k)] = [(x_{ij}-\overline{x}_j+\overline{x}_j-\mu_j)(x_{ik}-\overline{x}_k+\overline{x}_k-\mu_k)] =$ $[(x_{ij}-\overline{x}_j)(x_{ik}-\overline{x}_k)] + [(x_{ij}-\overline{x}_j)(\overline{x}_k-\mu_k)] + [(\overline{x}_j-\mu_j)(x_{ik}-\overline{x}_k)] + [(\overline{x}_j-\mu_j)(\overline{x}_k-\mu_k)]$. Hence $\boldsymbol{w}_i - \overline{\boldsymbol{w}} = \boldsymbol{z}_i - \overline{\boldsymbol{z}} + \boldsymbol{a}_i$ where

$$\boldsymbol{a}_i = [(x_{ij}-\overline{x}_j)(\overline{x}_k-\mu_k)] + [(\overline{x}_j-\mu_j)(x_{ik}-\overline{x}_k)] = O_P(n^{-1/2}).$$

Thus

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{d}} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{w}_i-\overline{\boldsymbol{w}})(\boldsymbol{w}_i-\overline{\boldsymbol{w}})^T = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{z}_i-\overline{\boldsymbol{z}})(\boldsymbol{z}_i-\overline{\boldsymbol{z}})^T + O_P(n^{-1/2}) = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{z}} + O_P(n^{-1/2}). \ \square$$

For iid cases, $\boldsymbol{\beta}_{MMLE} = \boldsymbol{V}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x},Y} = \boldsymbol{V}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}$ where $\boldsymbol{V} = diag(\sigma_1^2,...,\sigma_p^2) = diag(\boldsymbol{\Sigma}_{\boldsymbol{x}})$. For standardized predictors, let s_j and σ_j be the sample and population standard deviations of x_j. Let $\boldsymbol{t}_i = \hat{\boldsymbol{D}}\boldsymbol{x}_i = diag(1/s_1,...,1/s_p)\boldsymbol{x}_i$ and $\boldsymbol{u}_i = \boldsymbol{D}\boldsymbol{x}_i = diag(1/\sigma_1,...,1/\sigma_p)\boldsymbol{x}_i$. Note that $\hat{\boldsymbol{V}}^{-1} = \hat{\boldsymbol{D}}^2$ and $\boldsymbol{V}^{-1} = \boldsymbol{D}^2$. Olive and Zhang (2024) proved that $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y}$ is a $\sqrt{n}$ consistent estimator of $\boldsymbol{\Sigma}_{\boldsymbol{u},Y}$. For iid cases, $\boldsymbol{\beta}_{MMLE}(\boldsymbol{t},Y) = \boldsymbol{\Sigma}_{\boldsymbol{t},Y} = \boldsymbol{\eta}_{OPLS}(\boldsymbol{t},Y)$.

By Theorems 1 and 2 with iid $\boldsymbol{x}_i$ replaced by iid $(\boldsymbol{x}_i^T, Y_i)^T$,

$$\sqrt{n}\left[\begin{pmatrix} s_1^2 \\ \vdots \\ s_p^2 \\ \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} \end{pmatrix} - \begin{pmatrix} \sigma_1^2 \\ \vdots \\ \sigma_p^2 \\ \boldsymbol{\Sigma}_{\boldsymbol{x}Y} \end{pmatrix}\right] = \sqrt{n}(\hat{\boldsymbol{c}} - \boldsymbol{c}) \xrightarrow{D} N_{2p}\left(\boldsymbol{0}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{v}} & \boldsymbol{\Sigma}_{\boldsymbol{v},\boldsymbol{w}} \\ \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{v}} & \boldsymbol{\Sigma}_{\boldsymbol{w}} \end{pmatrix}\right). \quad (9)$$

Let

$$\boldsymbol{g}(\boldsymbol{c}) = \boldsymbol{\beta}_{MMLE} = \begin{pmatrix} g_1(\boldsymbol{c}) \\ \vdots \\ g_p(\boldsymbol{c}) \end{pmatrix} = \begin{pmatrix} \sigma_{1Y}/\sigma_1^2 \\ \vdots \\ \sigma_{pY}/\sigma_p^2 \end{pmatrix}.$$

Let $\boldsymbol{D}_{\boldsymbol{g}} = (\boldsymbol{D}_1, \boldsymbol{D}_2)$ where $\boldsymbol{D}_1 = diag(-\sigma_{1Y}/\sigma_1^4, -\sigma_{2Y}/\sigma_2^4, ..., -\sigma_{pY}/\sigma_p^4)$ and $\boldsymbol{D}_2 = diag(1/\sigma_1^2, 1/\sigma_2^2, ..., 1/\sigma_p^2)$. Typically $\hat{\boldsymbol{\Sigma}}_{x_{i_j}Y} = O_P(1)$, but if $\boldsymbol{\Sigma}_{x_{i_j}Y} = 0$, then $\hat{\boldsymbol{\Sigma}}_{x_{i_j}Y} = O_P(n^{-1/2})$.

Theorem 3. Let the cases $(\boldsymbol{x}_i^T, Y_i)^T$ be iid such that Equation (9) holds. Then a)

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{MMLE} - \boldsymbol{\beta}_{MMLE}) \xrightarrow{D} N_P(\boldsymbol{0}, \boldsymbol{\Sigma}_{MMLE}) \sim N_p\left(\boldsymbol{0}, \boldsymbol{D}_{\boldsymbol{g}} \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{v}} & \boldsymbol{\Sigma}_{\boldsymbol{v},\boldsymbol{w}} \\ \boldsymbol{\Sigma}_{\boldsymbol{w},\boldsymbol{v}} & \boldsymbol{\Sigma}_{\boldsymbol{w}} \end{pmatrix} \boldsymbol{D}_{\boldsymbol{g}}^T\right).$$

Let $\boldsymbol{A}$ be a full rank $k \times p$ constant matrix such that $\boldsymbol{A\beta} = (\beta_{i_1}, ..., \beta_{i_k})^T$ with $i_1, i_2, ..., i_k$ distinct. Hence the jth row of $\boldsymbol{A}$ has a 1 in the i_jth position and zeroes elsewhere. Assume $H_0 : \boldsymbol{A\beta}_{MMLE} = \boldsymbol{0}$. Then b)

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

c) For standardized predictors, assume $H_0 : \boldsymbol{A\beta}_{MMLE}(\boldsymbol{t},Y) = \boldsymbol{A\Sigma}_{\boldsymbol{t},Y} = \boldsymbol{0}$. Then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{t},Y) - \boldsymbol{\beta}_{MMLE}(\boldsymbol{t},Y)) = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{D}\boldsymbol{A}^T).$$

Proof. Theorem 3a) holds by the multivariate delta method.

b) Note that $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{MMLE} - \boldsymbol{\beta}_{MMLE}) = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{D}}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{D}^2\boldsymbol{\Sigma}_{\boldsymbol{x}Y}) =$ $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{D}}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{D}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} + \boldsymbol{D}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{D}^2\boldsymbol{\Sigma}_{\boldsymbol{x}Y}) =$

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{D}}^2 - \boldsymbol{D}^2)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} + \sqrt{n}\boldsymbol{A}\boldsymbol{D}^2(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{\Sigma}_{\boldsymbol{x}Y})$$

where by Theorem 1,

$$\sqrt{n}\boldsymbol{A}\boldsymbol{D}^2(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} - \boldsymbol{\Sigma}_{\boldsymbol{x}Y}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{D}^2\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{D}^2\boldsymbol{A}^T).$$

Now $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{D}}^2 - \boldsymbol{D}^2)\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y} =$

$$\boldsymbol{A}\begin{pmatrix} \sqrt{n}\left(\frac{1}{s_1^2} - \frac{1}{\sigma_1^2}\right)\hat{\boldsymbol{\Sigma}}_{x_1Y} \\ \vdots \\ \sqrt{n}\left(\frac{1}{s_p^2} - \frac{1}{\sigma_p^2}\right)\hat{\boldsymbol{\Sigma}}_{x_pY} \end{pmatrix} = \begin{pmatrix} \sqrt{n}\left(\frac{1}{s_{i_1}^2} - \frac{1}{\sigma_{i_1}^2}\right)\hat{\boldsymbol{\Sigma}}_{x_{i_1}Y} \\ \vdots \\ \sqrt{n}\left(\frac{1}{s_{i_k}^2} - \frac{1}{\sigma_{i_k}^2}\right)\hat{\boldsymbol{\Sigma}}_{x_{i_k}Y} \end{pmatrix} = o_P(1)$$

if $(\Sigma_{x_{i_1}Y}, ..., \Sigma_{x_{i_k}Y})^T = \boldsymbol{0}$. Hence the result follows if H_0 is true.

c) Note that $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y}) = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y} + \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y}) =$ $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y}) + \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y})$ where by Theorem 1 and Remark 1,

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y}) = \sqrt{n}\boldsymbol{A}\boldsymbol{D}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y} - \boldsymbol{\Sigma}_{\boldsymbol{x},Y}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{D}\boldsymbol{A}^T).$$

Now $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y}) = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{D}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y} - \boldsymbol{D}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}) = \sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{D}} - \boldsymbol{D})\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y} =$

$$\boldsymbol{A}\begin{pmatrix} \sqrt{n}\left(\frac{1}{s_1} - \frac{1}{\sigma_1}\right)\hat{\boldsymbol{\Sigma}}_{x_1Y} \\ \vdots \\ \sqrt{n}\left(\frac{1}{s_p} - \frac{1}{\sigma_p}\right)\hat{\boldsymbol{\Sigma}}_{x_pY} \end{pmatrix} = \begin{pmatrix} \sqrt{n}\left(\frac{1}{s_{i_1}} - \frac{1}{\sigma_{i_1}}\right)\hat{\boldsymbol{\Sigma}}_{x_{i_1}Y} \\ \vdots \\ \sqrt{n}\left(\frac{1}{s_{i_k}} - \frac{1}{\sigma_{i_k}}\right)\hat{\boldsymbol{\Sigma}}_{x_{i_k}Y} \end{pmatrix},$$

and $\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} - \hat{\boldsymbol{\Sigma}}_{\boldsymbol{u},Y}) = o_p(1)$ if $(\Sigma_{x_{i_1}Y}, ..., \Sigma_{x_{i_k}Y})^T = \boldsymbol{0}$. Hence if H_0 is true, then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} - \boldsymbol{\Sigma}_{\boldsymbol{u},Y}) \xrightarrow{D} N_k(\boldsymbol{0}, \boldsymbol{A}\boldsymbol{D}\boldsymbol{\Sigma}_{\boldsymbol{w}}\boldsymbol{D}\boldsymbol{A}^T). \quad \square$$

3 Testing

As noted by Olive and Zhang (2024), 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 $\text{Cov}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}) = \text{Cov}(\hat{\boldsymbol{\eta}}) = \boldsymbol{\Sigma}_{\boldsymbol{w}}$ 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 $\text{Cov}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y})$, but we can get good nonsingular estimators of $\text{Cov}(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{u}Y}) = \text{Cov}((\hat{\eta}_{i1}, ..., \hat{\eta}_{ik})^T)$ 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 $\sigma_{x_i,Y} = \text{Cov}(x_i, Y)$.

Note that the tests with $\hat{\boldsymbol{\eta}}$ using k distinct predictors x_{i_j} 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 OPLS 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 $\text{Cov}(\boldsymbol{x}, Y)$ if $\text{Cov}(\boldsymbol{x}, v) + \text{Cov}(\boldsymbol{u}, Y) + \text{Cov}(\boldsymbol{u}, v) = \boldsymbol{0}$, 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 $\boldsymbol{\eta}^T\boldsymbol{x}$.

Theorem 2 can be used to test $H_0 : \boldsymbol{A}\boldsymbol{c} = \boldsymbol{0}$, which can reduce a high dimensional problem to a low dimensional problem. Suppose $n > 10k$, $p > n$, and $\boldsymbol{A}\boldsymbol{\beta} = (\beta_{i_1}, ..., \beta_{i_k})^T$ with $i_1, i_2, ..., i_k$ distinct. Then Theorem 3a) can be used since no inverse matrices are required, but the asymptotic covariance matrices of Theorem 3b) and 3c) are much easier to estimate.

4 REGRESSION WITH HETEROGENEITY

A multiple linear regression model with heterogeneity is

$$Y_i = \beta_1 + x_{i,2}\beta_2 + \cdots + x_{i,p}\beta_p + e_i \tag{10}$$

for $i = 1, ..., n$ where the e_i are independent with $E(e_i) = 0$ and $V(e_i) = \sigma_i^2$. In matrix form, this model is

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

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. Also $E(\boldsymbol{e}) = \boldsymbol{0}$ and $\text{Cov}(\boldsymbol{e}) = \boldsymbol{\Sigma}_{\boldsymbol{e}} = diag(\sigma_i^2) = diag(\sigma_1^2, ..., \sigma_n^2)$ is an $n \times n$ positive definite matrix. In Section 2, the constant variance assumption was used: $\sigma_i^2 = \sigma^2$ for all i. Hence heterogeneity means that the constant variance assumption does not hold. A common assumption is that the $e_i = \sigma_i \epsilon_i$ where the ϵ_i are independent and identically distributed (iid) with $V(\epsilon_i) = 1$. See, for example, Zhou, Cook, and Zou (2023).

Weighted least squares (WLS) would be useful if the σ_i^2 were known. Since the σ_i^2 are not known, ordinary least squares (OLS) is often used. The OLS theory for MLR with heterogeneity often assume iid cases. For the following theorem, see Romano and Wolf (2017), Freedman (1981), and White (1980).

Theorem 4. Assume $Y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ for $i = 1, ..., n$ where the cases $(Y_i, \boldsymbol{x}_i^T)^T$ are iid with "fourth moments," $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$, the $e_i = e_i(\boldsymbol{x}_i)$ are independent, $E[e_i|\boldsymbol{x}_i] = 0$,

$\boldsymbol{V}^{-1} = E[\boldsymbol{x}_i\boldsymbol{x}_i^T]$, $E[e_i^2|\boldsymbol{x}_i] = v(\boldsymbol{x}_i) = \sigma_i^2$, $Cov[\boldsymbol{e}|\boldsymbol{X}] = diag(v(\boldsymbol{x}_1), ..., v(\boldsymbol{x}_n))$ and $\boldsymbol{\Omega} = E[v(\boldsymbol{x}_i)\boldsymbol{x}_i\boldsymbol{x}_i^T] = E[e_i^2\boldsymbol{x}_i\boldsymbol{x}_i^T]$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V\Omega V}). \tag{11}$$

Remark 2. a) White (1980) showed that the iid cases assumption can be weakened. Assume the cases are independent,

$$\boldsymbol{V}_n = \frac{1}{n}\sum_{i=1}^n E[\boldsymbol{x}_i\boldsymbol{x}_i^T] \xrightarrow{P} \boldsymbol{V}^{-1},$$

and

$$\boldsymbol{\Omega}_n = \frac{1}{n}\sum_{i=1}^n E[e_i^2\boldsymbol{x}_i\boldsymbol{x}_i^T] \xrightarrow{P} \boldsymbol{\Omega}.$$

Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \boldsymbol{V\Omega V}).$$

b) Under the assumptions of Theorem 4,

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^T \xrightarrow{P} \boldsymbol{V}^{-1}.$$

Let $\boldsymbol{D} = diag(\sigma_1^2, ..., \sigma_n^2) = \boldsymbol{\Sigma_e}$ and $\hat{\boldsymbol{D}} = diag(r_1^2, ..., r_n^2)$ where r_i^2 is the ith residual from OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$. Then $\hat{\boldsymbol{D}}$ is not a consistent estimator of $\boldsymbol{D}$. The following theorem, due to White (1980), shows that $\hat{\boldsymbol{D}}$ can be used to get a consistent estimator of $\boldsymbol{\Omega}$. This result leads to the sandwich estimators.

Theorem 5. Under strong regularity conditions,

$$\frac{1}{n}(\boldsymbol{X}^T\hat{\boldsymbol{D}}\boldsymbol{X}) \xrightarrow{P} \boldsymbol{\Omega} \text{ and } \frac{1}{\mathrm{n}}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{D}\boldsymbol{X}) \xrightarrow{\mathrm{P}} \boldsymbol{\Omega}.$$

Hence

$$n(\boldsymbol{X}^T\boldsymbol{X})^{-1}(\boldsymbol{X}^T\hat{\boldsymbol{D}}\boldsymbol{X})(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V\Omega V}.$$

Now write the linear model as $Y = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$. Under iid cases, OPLS theory does not depend on whether the error variance is constant or not. Hence Theorem 1 and the Section 3 theory still applies. If the cases are iid and linearity holds (with or without heterogeneity), then under reasonable conditions, $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Hence

$$\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}, \tag{12}$$

as noted by Olive and Zhang (2024) for when the iid errors e_i had constant variance. This result is useful for simulation.

5 SINGLE INDEX MODELS

The distribution of $Y|\boldsymbol{\eta}^T\boldsymbol{x}$ follows a single index model

$$Y|\boldsymbol{\eta}^T\boldsymbol{x} = Y = m(\boldsymbol{\eta}^T\boldsymbol{x}) + e$$

where $E(Y|\boldsymbol{\eta}^T\boldsymbol{x}) = m(\boldsymbol{\eta}^T\boldsymbol{x})$, $V(Y|\boldsymbol{\eta}^T\boldsymbol{x}) = v(\boldsymbol{\eta}^T\boldsymbol{x})$, and $e = Y - m(\boldsymbol{\eta}^T\boldsymbol{x})$. Note that the error variance may not be constant. The model is called a single index model since m depends on a single linear combination $\boldsymbol{\eta}^T\boldsymbol{x}$. A multi-index model would use $m(\boldsymbol{\eta}_1^T\boldsymbol{x}, ..., \boldsymbol{\eta}_k^T\boldsymbol{x})$ where $k > 1$.

If $\boldsymbol{\eta} = \boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$ and the cases are iid, then inference for the single index model can be done using Theorem 1 and Section 3. When the cases are iid, the OPLS single index model estimators can have considerable resistance to overfitting, underfitting, heterogeneity, measurement error, highly correlated predictors, and the number of predictors.

If $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ is a good estimator of $\boldsymbol{\Sigma}_{\boldsymbol{x}Y}$, which can occur if $n \geq 10p$, then the OPLS single index model can be visualized with a response plot of $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}$ versus Y on the vertical axis with a scatterplot smoother added as a visual aid. If the variability about the scatterplot smoother is less than that about any horizontal line, then the model may be useful compared to simply doing inference on the $Y_1, ..., Y_n$ without any predictors.

If $Y|\boldsymbol{x} = m(\alpha + \boldsymbol{\beta}^T\boldsymbol{x}) + e$ and if the predictors $\boldsymbol{x}_i$ are iid from a large class of elliptically contoured distributions, then Li and Duan (1989) and Chen and Li (1998) showed that, under regularity conditions, $\boldsymbol{\beta}_{OLS} = c\boldsymbol{\beta}$. Hence $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = c\boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}$. Thus $\boldsymbol{\Sigma}_{\boldsymbol{x}Y} = d\boldsymbol{\beta}$ if $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \tau^2\boldsymbol{I}_p$ for some constant $\tau^2 > 0$. If $\boldsymbol{\beta} = \boldsymbol{\beta}_{OLS}$ in this case, then $\beta_i = 0$ implies that $Cov(x_i, Y) = 0$. The constant c is typically nonzero unless m has a lot of symmetry about the distribution of $\alpha + \boldsymbol{\beta}^T\boldsymbol{x}$. Chang and Olive (2010) considered OLS tests for these models. Simulation with $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ can be difficult if the population values of c and d are unknown.

6 EXAMPLE AND SIMULATIONS

Example. This example was used by Olive and Zhang (2024). The Hebbler (1847) data was collected from $n = 26$ districts in Prussia in 1843. Let Y = the *number of women married to civilians* in the district with a constant and predictors x_1 = the *population of the district in 1843*, x_2 = the *number of married civilian men* in the district, x_3 = the *number of married men in the military* in the district, and x_4 = 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 and x_2 are highly correlated but not equal. Similarly, x_3 and x_4 are highly correlated but not equal. Then $\hat{\boldsymbol{\beta}}_{OLS} = (0.00035, 0.9995, -0.2328, 0.1531)^T$, forward selection with OLS and the C_p criterion used $\hat{\boldsymbol{\beta}}_{I,0} = (0, 1.0010, 0, 0)^T$, lasso had $\hat{\boldsymbol{\beta}}_L = (0.0015, 0.9605, 0, 0)^T$, lasso variable selection $\hat{\boldsymbol{\beta}}_{LVS} = (0.00007, 1.006, 0, 0)^T$, $\hat{\boldsymbol{\beta}}_{MMLE} = (0.1782, 1.0010, 48.5630, 51.5513)^T$, and $\hat{\boldsymbol{\beta}}_{OPLS} = (0.1727, 0.0311, 0.00018, 0.00018)^T$. The fitted values from the MMLE estimator tend not to estimate Y. Let $W = \boldsymbol{x}^T\hat{\boldsymbol{\beta}}_{MMLE}$ and perform the simple linear regression of

Y on W to get the reweighted or scaled estimators $\hat{\alpha}_R$ and b. Then $\hat{\boldsymbol{\beta}}_R = b\hat{\boldsymbol{\beta}}_{MMLE}$. Then the fitted values $\hat{Y}_i = \hat{\alpha}_R + \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_R$ can be used for prediction. If the scaled predictors $\boldsymbol{u}$ have unit sample variances, then $\hat{\boldsymbol{\beta}}_{OPLS}(\boldsymbol{u}, Y) = \hat{\boldsymbol{\beta}}_R(\boldsymbol{u}, Y)$.

Next, we describe a small WLS simulation study that done by Rajapaksha and Olive (2024). The simulation used $\psi = 0, 0.5, 1/\sqrt{p}$, and 0.9; and $k = 1, p-2$, and $p-1$ where k and ψ are defined in the following paragraph.

Let $\boldsymbol{u} = (1\ \boldsymbol{x}^T)^T$ where $\boldsymbol{x}$ 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 independent and identically distributed (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 $\boldsymbol{x}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{x}_i) = \boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ 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 $Y_i = 1 + 1x_{i,1} + \cdots + 1x_{i,k} + e_i$ for $i = 1, ..., n$. Hence $\alpha = 1$ and $\boldsymbol{\phi} = (1, .., 1, 0, ..., 0)^T$ with $k+1$ ones and $p-k-1$ zeros.

The zero mean iid errors $\tilde{e}_i = \epsilon_i$ 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. Then wtype = 1 if $e_i = \epsilon_i$ (the WLS model is the OLS model), 2 if $e_i = |\boldsymbol{x}_i^T\boldsymbol{\beta} - 5|\epsilon_i$, 3 if $e_i = \sqrt{(}1 + 0.5x_{i2}^2)\epsilon_i$, 4 if $e_i = \exp[1 + \log(|x_{i2}|) + ... + \log(|x_{ip}|)]\epsilon_i$, 5 if $e_i = [1 + \log(|x_{i2}|) + ... + \log(|x_{ip}|)]\epsilon_i$, 6 if $e_i = [\exp([\log(|x_{i2}|) + ... + \log(|x_{ip}|)]/(p-1))]\epsilon_i$, 7 if $e_i = [[\log(|x_{i2}|) + ... + \log(|x_{ip}|)]/(p-1)]\epsilon_i$, The last four types were special cases of types suggested by Romano and Wolf (2017). For type 6, the weighting function is the geometric mean of $|x_{i2}|, ..., |x_{ip}|$. For $n = 100$ and $p = 100$ with $\psi \neq 0$, the CI lengths were too long for wtype = 4.

When $\psi = 0$ and wtype = 1, the OLS confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n} \approx 2(1.96)\sigma/10 = 0.392\sigma$ when $n = 100$ and the iid zero mean errors have variance σ^2.

The simulation computed $\boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y} = (\eta_1, ..., \eta_{p-1})^T = \boldsymbol{\Sigma}_{\boldsymbol{x}}\boldsymbol{\beta}_{OLS}$ where $\boldsymbol{\Sigma}_{\boldsymbol{x}} = \boldsymbol{A}\boldsymbol{A}^T$ is a $(p-1) \times (p-1)$ matrix. Storage problems can occur if $p > 10000$. Then the Theorem 1 large sample $100(1-\delta)$ CI is $\hat{\eta}_i \pm t_{n-1,1-\delta/2}SE(\hat{\eta}_i)$ could be computed for each η_i. If 0 is not in the confidence interval, then $H_0 : \eta_i = 0$ and $H_0 : \beta_{iE} = 0$ are both rejected for estimators E = OPLS and MMLE. In the simulations with $n = 50$ and $\psi > 0$, the maximum observed undercoverage was about $0.05 = 5\%$. Hence the program has the option to replace the cutoff $t_{n-1,1-\delta/2}$ by $t_{n-1,up}$ where $up = min(1 - \delta/2 + 0.05, 1 - \delta/2 + 2.5/n)$ if $\delta/2 > 0.1$,

$$up = min(1 - \delta/4, 1 - \delta/2 + 12.5\delta/n)$$

if $\delta/2 \leq 0.1$. If $up < 1 - \delta/2 + 0.001$, then use $up = 1 - \delta/2$. This correction factor was used in the simulations for the nominal 95% CIs, where the correction factor uses a cutoff that is between $t_{n-1,0.975}$ and the cutoff $t_{n-1,0.9875}$ that would be used for a 97.5% CI. The nominal coverage was 0.95 with $\delta = 0.05$. Observed coverage between 0.94 and 0.96 suggests coverage is close to the nominal value. Pötscher and Preinerstorfer (2023) noted that WLS tests tend to reject H_0 too often (liberal tests with undercoverage).

To summarize the $p-1$, confidence intervals, the average length of the $p-1$ confidence intervals over 5000 runs was computed. Then the minimum, mean, and maximum of the average lengths was computed. The proportion of times each confidence interval contained its population parameter was computed. These proportions were the observed coverages of the $p-1$ confidence intervals. Then the minimum observed coverage was found. The percentage of the observed coverages that were ≥ 0.9, 0.92, 0.93, 0.94, and 0.96 were also recorded.

7 CONCLUSIONS

There is a large literature for multiple linear regression models with heterogeneity. See, for example, Buja et al. (2019), Eicker (1963, 1967), Flachaire (2005), Hinkley (1977), Huber (1967), Long and Ervin (2000), MacKinnon and White (1985), Rajapaksha and Olive (2024), Romano and Wolf (2017), and White (1980). The response plot of $\hat{\boldsymbol{\phi}}_{OPLS}$ versus Y and the EE plot of $\hat{\boldsymbol{\phi}}^T_{OPLS}\boldsymbol{x}$ versus $\hat{\boldsymbol{\phi}}^T_{OLS}\boldsymbol{x}$ can be used to check whether OPLS is useful for WLS. See Olive (2013) for more on these two plots.

Tests for high dimensional covariance matrices include Chen, Zhang, and Zhong (2010), and Himeno and Yamada (2014).

Software

The *R* software was used in the simulations. See R Core Team (2020). Programs are available from the Olive (2023) collections of *R* functions *slpack.txt*, available from (http://parker.ad.siu.edu/Olive/slpack.txt). The function `OPLSplot` make the response plot and residual plot for multiple linear regression based on one component partial least squares. The function `OPLSEEplot` plots the OPLS fitted values versus the OLS fitted values. Let $up \approx 1-\alpha/2$ be the correction factor used for the confidence intervals. The function `covxycis` obtains the large sample $100(1-\alpha)\%$ confidence intervals $\approx \hat{\eta}_j \pm t_{n-1,up} SE(\hat{\eta}_j)$ for $\eta_j = \text{Cov}(x_j, Y)$ for $j = 1, ..., p$. The function `oplscis` obtains the large sample $100(1-\alpha)\%$ confidence intervals $\approx \hat{\beta}_j \pm t_{n-1,up} SE(\hat{\beta}_j)$ for $\beta_j = \lambda \text{Cov}(x_j, Y)$ for $j = 1, ..., p$. If $[L_j, U_j]$ is the confidence interval for η_j, then $[\hat{\lambda} L_j, \hat{\lambda} U_j]$ is the confidence interval for β_j. The function `oplswls` generates a weighted least squares data set of types used by the simulation, the OPLS response plot, the OLS response plot, and the plot of the OPLS fitted values versus the OLS fitted values. In the literature, simulated WLS data set often contain outliers and are often not very linear. The response plot can be used to check for these two problems. The function `oplswsim` was used for the simulation of confidence intervals for η_i. The function `rcovxy` makes the classical and three robust estimators of $\boldsymbol{\eta}$, and makes a scatterplot matrix of the four estimated sufficient predictors $\hat{\boldsymbol{\eta}}^T \boldsymbol{x}$ and Y. Only two robust estimators are made if $n \leq 2.5p$. The function `oplssim` simulated confidence intervals for η_i when the errors were iid. The function `oplssim2` simulated confidence intervals for β_i when the errors were iid.

References

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.

Bickel, P.J., and Doksum, K.A. (2007), *Mathematical Statistics: Basic Ideas and Selected Topics,* Vol. 1., 2nd ed., Updated Printing, Pearson Prentice Hall, Upper Saddle River, NJ.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., and Zhao, L. (2019), "Models as Approximations I: Consequences Illustrated with Linear Regression," *Statistical Science*, 34, 523-544.

Chang, J., and Olive, D.J. (2010), "OLS for 1D Regression Models," *Communications in Statistics: Theory and Methods*, 39, 1869-1882.

Chen, C.H., and Li, K.C. (1998), "Can SIR be as Popular as Multiple Linear Regression?," *Statistica Sinica*, 8, 289-316.

Chen, S.X., Zhang, L.X. and Zhong, P.S. (2010), "Tests for High-Dimensional Covariance Matrices, *Journal of the American Statistical Association*, 105(490):810-819.

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.

Eicker, F. (1963), "Asymptotic Normality and Consistency of the Least Squares Estimators for Families of Linear Regressions," *Annals of Mathematical Statistics*, 34, 447-456.

Eicker, F. (1967), "Limit Theorems for Regressions with Unequal and Dependent Errors," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol. I: Statistics*, eds. Le Cam, L.M., and Neyman, J., University of California Press, Berkeley, CA, 59-82.

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.

Flachaire, E. (2005), "Bootstrapping Heteroskedastic Regression Models: Wild Bootstrap vs. Pairs Bootstrap, *Computational Statistics & Data Analysis*, 49, 361-376.

Freedman, D.A. (1981), "Bootstrapping Regression Models," *The Annals of Statistics*, 9, 1218-1228.

Hebbler, B. (1847), "Statistics of Prussia," *Journal of the Royal Statistical Society, A*, 10, 154-186.

Himeno, T., and Yamada, T. (2014), "Estimations for Some Functions of Covariance Matrix in High Dimension under Non-Normality and Its Applications," *Journal of Multivariate Analysis*, 130, 27-44.

Hinkley, D.V. (1977), "Jackknifing in Unbalanced Situations," *Technometrics*, 19, 285-292.

Huber, P.J. (1967), "The Behavior of Maximum Likelihood Estimation Under Nonstandard Conditions," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability 1*, eds. LeCam, L.M., and Neyman, J., University of California Press, Berkeley, CA, 221-223.

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.

Li, K.C., and Duan, N. (1989), "Regression Analysis Under Link Violation," *The Annals of Statistics,* 17, 1009-1052.

Long, J.S., and Ervin, L.H. (2000), "Using Heteroscedasticity Consistent Standard Errors in the Linear Model," *The American Statistician*, 54, 217-224.

MacKinnon, J.G., and White, H. (1985), "Some Heteroskedasticity-Consistent Covariance Matrix Estimators with Improved Finite Sample Properties," *Journal of Econometrics*, 29, 305-325.

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

Olive, D.J. (2017), *Linear Regression*, Springer, New York, NY.

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

Olive, D.J., and Zhang, L. (2024), "One Component Partial Least Squares, High Dimensional Regression, Data Splitting, and the Multitude of Models," *Communications in Statistics: Theory and Methods*, to appear.

Pötscher, B.M., and Preinerstorfer, D. (2023), "How Reliable are Bootstrap-Based Heteroskedasticity Robust Tests?" *Econometric Theory*, 39, 789-847.

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

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

Rajapaksha, K.W.G.D.H., and Olive, D.J. (2024), "Wald Type Tests with the Wrong Dispersion Matrix," *Communications in Statistics: Theory and Methods*, 53, 2236-2251.

Rathnayake, R.C., and Olive, D.J. (2023), "Bootstrapping Some GLMs and Survival Regression Models after Variable Selection," *Communications in Statistics: Theory and Methods*, 52, 2625-2645.

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.

Romano, J.P., and Wolf, M. (2017), "Resurrecting Weighted Least Squares," *Journal of Econometrics*, 197, 1-19.

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

White, H. (1980), "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48, 817-838.

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

Zhou, L., Cook, R.D., and Zou, H. (2023), "Enveloped Huber Regression," *Journal of the American Statistical Association*, to appear.

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.