Testing Poisson Regression with the One Component Partial Least Squares Estimator

David J. Olive and Paul Quaye *
Southern Illinois University

April 9, 2024

Abstract

Poisson regression, negative binomial regression, and related regression methods are often used when the response variable is a count. A log transformation often results in a linear model with heterogeneity. Then testing can be done with the one component partial least squares estimator for multiple linear regression, including some high dimensional tests. For prediction, a simple method that uses information from several estimators, is also considered.

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

1 INTRODUCTION

This section reviews regression models where the nonnegative integer count response variable is 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. A useful *Poisson regression (PR) model* is $Y \sim \text{Poisson}\left(\mathrm{e}^{\mathrm{SP}}\right)$. This model has $E(Y|SP) = V(Y|SP) = \exp(SP)$.

Some notation is needed for the negative binomial regression model. If Y has a (generalized) negative binomial distribution, $Y \sim NB(\mu, \kappa)$, then the probability mass function (pmf) of Y is

$$P(Y = y) = \frac{\Gamma(y+\kappa)}{\Gamma(\kappa)\Gamma(y+1)}\left(\frac{\kappa}{\mu+\kappa}\right)^{\kappa}\left(1-\frac{\kappa}{\mu+\kappa}\right)^{y}$$

for $y = 0, 1, 2, ...$ where $\mu > 0$ and $\kappa > 0$. Then $E(Y) = \mu$ and $\mathrm{V}(Y) = \mu + \mu^2/\kappa$.

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

The *negative binomial regression model* states that $Y_1, ..., Y_n$ are independent random variables with

$$Y|SP \sim \text{NB}(\exp(\text{SP}), \kappa).$$

This model has $E(Y|SP) = \exp(SP)$ and

$$V(Y|SP) = \exp(SP)\left(1 + \frac{\exp(SP)}{\kappa}\right) = \exp(SP) + \tau \exp(2\ SP).$$

Following Agresti (2002, p. 560), as $\tau \equiv 1/\kappa \to 0$, it can be shown that the negative binomial regression model converges to the Poisson regression model.

The *quasi-Poisson regression model* has $E(Y|SP) = \exp(SP)$ and $V(Y|SP) = \phi \exp(SP)$ where the dispersion parameter $\phi > 0$. Note that this model and the Poisson regression model have the same conditional mean function, and the conditional variance functions are the same if $\phi = 1$.

Next, some notation is needed for the zero truncated Poisson regression model. See Olive (2017, pp. 430–431). Y has a zero truncated Poisson distribution, $Y \sim ZTP(\mu)$, if the probability mass function of Y is

$$f(y) = \frac{e^{-\mu}\ \mu^y}{(1 - e^{\mu})\ y!}$$

for $y = 1, 2, 3, ...$ where $\mu > 0$. The ZTP pmf is obtained from a Poisson distribution where $y = 0$ values are truncated, so not allowed. Now $E(Y) = \mu/(1 - e^{-\mu})$, and

$$V(Y) = \frac{\mu^2 + \mu}{1 - e^{-\mu}} - \left(\frac{\mu}{1 - e^{-\mu}}\right)^2.$$

The *zero truncated Poisson regression* model has $Y|SP \sim ZTP(\exp(SP))$. Hence the parameter $\mu(SP) = \exp(SP)$,

$$E(Y|SP) = \frac{\exp(SP)}{1 - \exp(-\exp(SP))}, \quad \text{and}$$

$$V(Y|SP) = \frac{[\exp(SP)]^2 + \exp(SP)}{1 - \exp(-\exp(SP))} - \left(\frac{\exp(SP)}{1 - \exp(-\exp(SP))}\right)^2.$$

Other alternatives include the zero truncated negative binomial regression model, the hurdle or zero inflated Poisson regression model, and the hurdle or zero inflated negative binomial regression model. See Zuur et al. (2009), Simonoff (2003), and Hilbe (2011).

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 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. For Poisson rgression, these methods include lasso and elastic net. See Friedman et al. (2007), Friedman, Hastie, and Tibshirani (2010), 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 generalized linear models. For fixed p, these two papers showed that $\hat{\boldsymbol{\beta}}_{VS}$ is $\sqrt{n}$ consistent with a complicated nonnormal limiting distribution.

Let the log transformation $Z_i = \log(Y_i)$ if $Y_i > 0$ and $Z_i = \log(0.5)$ if $Y_i = 0$. This transformation often results in a linear model with heterogeneity:

$$Z_i = \alpha_Z + \boldsymbol{x}_i^T\boldsymbol{\beta}_Z + e_i \tag{2}$$

where the e_i are independent with expected value $E(Z_i) = 0$ and variance $V(Z_i) = \sigma_i^2$. For Poisson regression, the minimum chi-square estimator is the weighted least squares estimator from the regression of Z_i on $\boldsymbol{x}_i$ with weights $w_i = e^{Z_i}$. See Agresti (2002, pp. 611–612) and Olive (2013, 2017: pp. 406–407).

Hence multiple linear regression models will be useful. Now let the response variable Y be for multiple linear regression, so Y need not be a nonnegative integer. A useful multiple linear regression model is $Y|\boldsymbol{x}^T\boldsymbol{\beta} = \alpha + \boldsymbol{x}^T\boldsymbol{\beta} + e$ or $Y_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i$ or

$$Y_i = \alpha + x_{i,1}\beta_1 + \cdots + x_{i,p}\beta_p + e_i = \alpha + \boldsymbol{x}_i^T\boldsymbol{\beta} + e_i \tag{3}$$

for $i = 1, ..., n$. 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 form, this model is

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\phi} + \boldsymbol{e}, \tag{4}$$

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 a multiple linear regression model with heterogeneity, assume model (4) holds with $E(\boldsymbol{e}) = \boldsymbol{0}$ 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. When the σ_i^2 are known, weighted least squares (WLS) is often used. Under regularity conditions, the ordinary least squares (OLS) estimator $\hat{\boldsymbol{\phi}}_{OLS} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T \boldsymbol{Y}$ can be shown to be a consistent estimator of $\boldsymbol{\phi}$. See, for example, White (1980).

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

$$\lambda = \frac{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}Y}}{\boldsymbol{\Sigma}_{\boldsymbol{x}Y}^T \boldsymbol{\Sigma}_{\boldsymbol{x}} \boldsymbol{\Sigma}_{\boldsymbol{x}Y}} \quad \text{and} \quad \hat{\lambda} = \frac{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}}{\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}^{\mathrm{T}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}\mathrm{Y}}} \tag{5}$$

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 $\boldsymbol{\eta}_{OPLS} = \boldsymbol{\Sigma}_{\boldsymbol{x}Y}$. Olive and Alshammari (2024) showed that for iid cases $(\boldsymbol{x}_i, Y_i)$, these results still hole for multiple linear regression models with heterogeneity. Thus the OPLS regression of Z_i on $\boldsymbol{x}_i$ is useful to model (2).

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, such as Poisson regression, 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

$$\hat{\boldsymbol{\beta}}_{MMLE} = [diag(\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}})]^{-1} \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x},Y}. \tag{6}$$

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

$$\hat{\boldsymbol{\beta}}_{MMLE} = \hat{\boldsymbol{\beta}}_{MMLE}(\boldsymbol{t}, Y) = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} = \boldsymbol{I}^{-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{t},Y} = \hat{\boldsymbol{\eta}}_{OPLS}(\boldsymbol{t}, Y) \tag{7}$$

where $(\boldsymbol{t}, Y)$ denotes that Y was regressed on $\boldsymbol{t}$, and $\boldsymbol{I}$ is the $p \times p$ identity matrix. Olive and Alshammari (2024) derived large sample theory for the MMLE for the multiple linear regression models, including models with heterogeneity.

If the regression model for Y depends on $\boldsymbol{x}$ only through $\alpha + \boldsymbol{\beta}^T \boldsymbol{x}$, and if the predictors $\boldsymbol{x}_i$ are independent and identically distributed (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. Results from Cameron and Trivedi (1998, p. 89) suggest that if a Poisson regression model is fit using OLS software for multiple linear regression, then a rough approximation is $\hat{\boldsymbol{\beta}}_{PR} \approx \hat{\boldsymbol{\beta}}_{OLS}/\overline{Y}$.

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, Olive and Zhang (2024) and 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.

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

2 Large Sample Theory

This section reviews the Olive and Zhang (2024) large sample theory for $\hat{\boldsymbol{\eta}}_{OPLS} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$ and OPLS for the multiple linear regression model, 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. Also see the large sample theory given in Olive and Alshammari (2024).

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

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

The following Olive and Zhang (2024) theorem gives the large sample theory for $\hat{\boldsymbol{\eta}} = \widehat{\text{Cov}}(\boldsymbol{x}, Y)$. This theory needs $\boldsymbol{\eta} = \boldsymbol{\eta}_{OPLS} = \mathbf{\Sigma}_{\boldsymbol{x},Y}$ to exist for $\hat{\boldsymbol{\eta}} = \hat{\mathbf{\Sigma}}_{\boldsymbol{x},Y}$ 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

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

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

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 $\boldsymbol{\eta} = \mathbf{\Sigma}_{\boldsymbol{x},Y}$. Then a)

$$\sqrt{n}(\overline{\boldsymbol{w}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \mathbf{\Sigma}_{\boldsymbol{w}}), \ \sqrt{n}(\hat{\boldsymbol{\eta}}_n - \boldsymbol{\eta}) \xrightarrow{D} N_p(\boldsymbol{0}, \mathbf{\Sigma}_{\boldsymbol{w}}), \tag{8}$$

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

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 $\hat{\mathbf{\Sigma}}_{\boldsymbol{w}} = \hat{\mathbf{\Sigma}}_{\boldsymbol{z}} + O_P(n^{-1/2}) = \hat{\mathbf{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$. Hence $\tilde{\mathbf{\Sigma}}_{\boldsymbol{w}} = \tilde{\mathbf{\Sigma}}_{\boldsymbol{z}} + O_P(n^{-1/2}) = \tilde{\mathbf{\Sigma}}_{\boldsymbol{v}} + O_P(n^{-1/2})$.

c) Let $\boldsymbol{A}$ be a $k \times p$ full rank constant matrix with $k \le p$, assume $H_0 : \boldsymbol{A}\boldsymbol{\beta}_{OPLS} = \boldsymbol{0}$ is true, and assume $\hat{\lambda} \xrightarrow{P} \lambda \neq 0$. Then

$$\sqrt{n}\boldsymbol{A}(\hat{\boldsymbol{\beta}}_{OPLS} - \boldsymbol{\beta}_{OPLS}) \xrightarrow{D} N_k(\boldsymbol{0}, \lambda^2 \boldsymbol{A} \mathbf{\Sigma}_{\boldsymbol{w}} \boldsymbol{A}^T). \tag{9}$$

2.1 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 $\text{Cov}(\hat{\mathbf{\Sigma}}_{\boldsymbol{x}Y}) = \text{Cov}(\hat{\boldsymbol{\eta}}) = \mathbf{\Sigma}_{\boldsymbol{w}}$ be the asymptotic covariance matrix of $\hat{\boldsymbol{\eta}} = \hat{\mathbf{\Sigma}}_{\boldsymbol{x}Y}$. In high dimensions where $n < 5p$, we can't get a good nonsingular estimator of $\text{Cov}(\hat{\mathbf{\Sigma}}_{\boldsymbol{x}Y})$, but we can get good nonsingular estimators of $\text{Cov}(\hat{\mathbf{\Sigma}}_{\boldsymbol{u}Y}) = \text{Cov}((\hat{\eta}_{i1}, ..., \hat{\eta}_{ik})^T)$ with $\boldsymbol{u} = (x_{i1}, ..., x_{ik})^T$ where $n \ge Jk$ with $J \ge 10$. (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 $\boldsymbol{\eta}^T\boldsymbol{x}$.

3 Incorporating Information from Several Regression Estimators

The theory and tests from the previous section can be applied to model (2) with Z replacing Y.

There are several ways to compute k-component partial least squares (PLS) estimators for multiple linear regression. A simple way is to do the OLS regression on $W_1, ..., W_k$ where $W_j = \hat{\boldsymbol{\eta}}_j^T\boldsymbol{x}$ and $\hat{\boldsymbol{\eta}}_j = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^{j-1}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}$, and $k < n - 1$. Then the one component PLS estimator is OPLS while the 3-component PLS estimator regresses Y on $W_1 = \hat{\boldsymbol{\eta}}_1^T\boldsymbol{x} = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}^T\boldsymbol{x}$, $W_2 = \hat{\boldsymbol{\eta}}_2^T\boldsymbol{x} = [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}]^T\boldsymbol{x}$, and $W_3 = \hat{\boldsymbol{\eta}}_3^T\boldsymbol{x} = [\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Y}]^T\boldsymbol{x}$. See Helland (1990).

This result suggests computing $W_i = \hat{\boldsymbol{\eta}}_i^T\boldsymbol{x}$ for $i = 1, ..., J$ and fit the OLS model that regresses Z on the W_i or, for example, the Poisson regression model that regresses Y on the W_i. Some interesting choices are $\hat{\boldsymbol{\eta}}_1 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}$, $\hat{\boldsymbol{\eta}}_2 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}$, $\hat{\boldsymbol{\eta}}_3 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}Z}$, $\hat{\boldsymbol{\eta}}_4 = \hat{\boldsymbol{\beta}}_L(\boldsymbol{x}, Z)$= the lasso estimator from regressing Z on $\boldsymbol{x}$, $\hat{\boldsymbol{\eta}}_5 = \hat{\boldsymbol{\beta}}_{RR}(\boldsymbol{x}, Z)$= the ridge regression estimator from regressing Z on $\boldsymbol{x}$, $\hat{\boldsymbol{\eta}}_6 = \hat{\boldsymbol{\beta}}_{LPR}(\boldsymbol{x}, Y)$= the lasso Poisson regression estimator from regressing Y on $\boldsymbol{x}$. Let $\boldsymbol{x}_I$ denote the set of variables selected using $\hat{\boldsymbol{\eta}}_4$. Then $\hat{\boldsymbol{\eta}}_7 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_I Z}$, $\hat{\boldsymbol{\eta}}_8 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_I}\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_I Z}$, $\hat{\boldsymbol{\eta}}_9 = \hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_I}^2\hat{\boldsymbol{\Sigma}}_{\boldsymbol{x}_I Z}$, $\hat{\boldsymbol{\eta}}_{10} = \hat{\boldsymbol{\beta}}_{RR}(\boldsymbol{x}_I, Z)$= the ridge regression estimator from regressing Z on $\boldsymbol{x}_I$. Other good choices can easily be obtained. For example, let $\boldsymbol{x}_G$ denote the set of variables selected using $\hat{\boldsymbol{\eta}}_6$.

4 EXAMPLE AND SIMULATIONS

5 CONCLUSIONS

The response plot of the estimated sufficient predictor $\hat{\alpha} + \boldsymbol{x}^T\hat{\boldsymbol{\beta}}$ versus Y is useful for checking many regression models. See Olive (2013) for more on plots for such models, including a plot to detect overdispersion.

Software

The *R* software was used in the simulations. See R Core Team (2020). Programs will be added to the Olive (2023) collections of *R* functions *slpack.txt*, available from (http://parker.ad.siu.edu/Olive/slpack.txt).

References

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.

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.

Cameron, A.C., and Trivedi, P.K. (1998), *Regression Analysis of Count Data*, 1st and Cambridge University Press, Cambridge, UK.

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

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.

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

Helland, I.S. (1990), "Partial Least Squares Regression and Statistical Models," *Scandanavian Journal of Statistics*, 17, 97-114.

Hilbe, J.M. (2011), *Negative Binomial Regression*, Cambridge University Press, 2nd ed., Cambridge, UK.

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

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. (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 Alshammari, A. (2024), "Testing with the One Component Partial Least Squares and the Marginal Maximum Likelihood Estimators," is at (http://parker.ad.siu.edu/Olive/pphdwls.pdf).

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.

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.

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).

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.

Simonoff, J.S. (2003), *Analyzing Categorical Data*, Springer, New York, NY.

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.

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.

Zuur, A.F., Ieno, E.N., Walker, N.J., Saveliev, A.A., and Smith, G.M. (2009), *Mixed Effects Models and Extensions in Ecology with R,* Springer, New York, NY.