OLS Testing with Predictors Scaled to Have Unit Sample Variance

David J. Olive and Sanjuka Johana Lemonge*
Southern Illinois University

December 12, 2024

Abstract

We consider hypothesis tests for the multiple linear regression model with ordinary least squares if the predictor variables have been scaled to have unit sample variance. Some tests are unchanged, but confidence intervals, confidence regions, and some tests are no longer valid.

KEY WORDS: Multiple linear regression.

1 INTRODUCTION

This section reviews multiple linear regression models. Consider a multiple linear regression model with response variable Y and predictors $\boldsymbol{x} = (x_1, ..., x_p)$ where a constant $x_1 \equiv 1$ is in the model. Then there are n cases $(Y_i, \boldsymbol{x}_i^T)^T$, and the sufficient predictor $SP = \boldsymbol{x}^T\boldsymbol{\beta}$. 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 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 $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}$ 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 the covariance matrix $\text{Cov}(\boldsymbol{e}) = \sigma^2\boldsymbol{I}_n$ where $\boldsymbol{I}_n$ is the $n \times n$ identity matrix. The OLS estimator for $\boldsymbol{\beta}$ is $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$,

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

the vector of fitted values is $\hat{\boldsymbol{Y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$, the vector of residuals is $\boldsymbol{r} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$, and $\hat{\sigma}^2 = MSE = \sum_{i=1}^n r_i^2/(n-p)$.

There are many multiple linear regression methods, and it is often convenient to use centered or scaled data. See James et al. (2021). Suppose U has observed values $U_1, ..., U_n$. Let g be an integer near 0. If the sample variance of the U_i is

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

then the sample standard deviation of U_i is $\hat{\sigma}_g$. If the values of U_i are not all the same, then $\hat{\sigma}_g > 0$. Using $g = 1$ gives an unbiased estimator s^2 of σ^2, while $g = 0$ gives the method of moments estimator.

Next consider scaling the predictors. If $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{X},\boldsymbol{Y}) + \boldsymbol{e}$, the model with scaled predictors is $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\beta}(\boldsymbol{W},\boldsymbol{Y}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\beta}(\boldsymbol{X},\boldsymbol{Y})$ denotes the population coefficients from the OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$. Here $\boldsymbol{W} = \boldsymbol{X}\hat{\boldsymbol{D}}_n$ where the $p \times p$ matrix $\hat{\boldsymbol{D}}_n = diag(1, 1/s_2, ..., 1/s_p)$ where $s_j = \hat{\sigma}_j$ for the jth predictor x_j, and $j = 2, ..., p$. Since OLS is affine equivariant and $\hat{\boldsymbol{D}}_n$ is nonsingular, $\hat{\boldsymbol{\beta}}(\boldsymbol{W},\boldsymbol{Y}) = \hat{\boldsymbol{\beta}}(\boldsymbol{X}\hat{\boldsymbol{D}}_n,\boldsymbol{Y}) = \hat{\boldsymbol{D}}_n^{-1}\hat{\boldsymbol{\beta}}(\boldsymbol{X},\boldsymbol{Y})$. Then $\boldsymbol{H}_{\boldsymbol{W}} = \boldsymbol{W}(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{W}^T = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T = \boldsymbol{H}_{\boldsymbol{X}}$, and the residuals and fitted values are the same for both models. See, for example, Olive (2017, p. 413).

Now consider centered data $Y_i - \overline{Y} = \beta_1^* + (x_{i,2} - \overline{x}_2)\beta_2 + \cdots + (x_{i,p} - \overline{x}_p)\beta_p + \epsilon_i$ or $Z_i = \beta_1^* + w_{i,2}\beta_2 + \cdots + w_{i,p}\beta_p + \epsilon_i$. Do the OLS regression. Since the sample means of the centered response and centered predictors are 0, $\hat{\beta}_1^* = 0$. In terms of the original predictors, $\hat{Y}_i = \tilde{\beta}_1 + x_{i,2}\tilde{\beta}_2 + \cdots + x_{i,p}\tilde{\beta}_p$ where $\tilde{\beta}_1 = \overline{Y} - \tilde{\beta}_2\overline{x}_2 - \cdots - \tilde{\beta}_p\overline{x}_p$. Then $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ since OLS estimators minimize the sum of squared residuals (if $\tilde{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}$, then one of the estimators has a smaller sum of squared residuals, contradicting the fact that both estimators are OLS estimators). Hence centering the response and predictors gives an equivalent method for computing $\hat{\boldsymbol{\beta}}$, and the large sample theory for the equivalent estimators is unchanged.

Often inference for the the scaled data $(\boldsymbol{W},\boldsymbol{Y})$ is done using output from OLS software. The large sample theory from Section 2 shows that confidence intervals and some hypothesis tests are no longer valid. Section 3 gives a small simulation study illustrating the results.

2 Large Sample Theory

There are many large sample theory results for ordinary least squares. The following theorem is important. See, for example, Sen and Singer (1993, p. 280). Let $\boldsymbol{H} = \boldsymbol{H}_{\boldsymbol{X}}$, and let h_i be the ith diagonal element of $\boldsymbol{H}$. Theorem 1 acts if the $\boldsymbol{x}_i$ are constant even if the $\boldsymbol{x}_i$ are random vectors. The literature says the $\boldsymbol{x}_i$ can be constants, or condition on $\boldsymbol{x}_i$ if the $\boldsymbol{x}_i$ are random vectors. Let the leverages $h_i = \boldsymbol{H}_{ii}$ be the diagonal elements of $\boldsymbol{H}$.

Theorem 1. Consider the MLR model and assume that the zero mean errors are iid with $E(e_i) = 0$ and VAR$(e_i) = \sigma^2$. If the $\boldsymbol{x}_i$ are random vectors, assume that the cases $(\boldsymbol{x}_i, Y_i)$ are independent, and that the $\boldsymbol{e}_i$ and $\boldsymbol{x}_i$ are independent. Also assume that $\max_i(h_1, ..., h_n) \to 0$ and

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

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

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

Consider testing $H_0 : \boldsymbol{L}\boldsymbol{\beta} = \boldsymbol{c}$ where $\boldsymbol{L}$ is a full rank $k \times p$ constant matrix and $\boldsymbol{c}$ is a $k \times 1$ constant vector. If H_0 is true, then by Theorem 1, $\sqrt{n}\boldsymbol{L}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) = \sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} N_k(\boldsymbol{0}, \sigma^2\ \boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^T)$. Hence $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T(\sigma^2\ \boldsymbol{L}\boldsymbol{V}\boldsymbol{L}^T)^{-1}\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \chi^2_k$ as $n \to \infty$. Let $\hat{\sigma}^2 = MSE \xrightarrow{P} \sigma^2$ and $\hat{\boldsymbol{V}} = n(\boldsymbol{X}^T\boldsymbol{X})^{-1} \xrightarrow{P} \boldsymbol{V}$ as $n \to \infty$ where convergence in probability indicates a consistent estimator. Then $\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T(\hat{\sigma}^2\ \boldsymbol{L}\hat{\boldsymbol{V}}\boldsymbol{L}^T)^{-1}\sqrt{n}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) =$

$$kF_1 = \frac{1}{MSE}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c})^T[\boldsymbol{L}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}^T]^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \xrightarrow{D} \chi^2_k \tag{3}$$

as $n \to \infty$ if H_0 is true. If H_0 is true, then an $F_{1-\alpha,k,n-p}$ cutoff can be used for $F_1 = kF_1/k$ since $kF_{k,n-p} \xrightarrow{D} \chi^2_k$ as $n \to \infty$. See Seber and Lee (2003, p. 100).

If $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta}(\boldsymbol{X}, \boldsymbol{Y}) + \boldsymbol{e}$, the model with scaled predictors is $\boldsymbol{Y} = \boldsymbol{W}\boldsymbol{\beta}(\boldsymbol{W}, \boldsymbol{Y}) + \boldsymbol{\epsilon}$ where $\boldsymbol{\beta}(\boldsymbol{X}, \boldsymbol{Y})$ denotes the population coefficients from the OLS regression of $\boldsymbol{Y}$ on $\boldsymbol{X}$. Here $\boldsymbol{W} = \boldsymbol{X}\hat{\boldsymbol{D}}_n$. As noted in Section 1, and the residuals and fitted values are the same for both models. Thus $\hat{Y} =$

$$\hat{\beta}_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p = \hat{\beta}_1 + \hat{\beta}_2 s_2 \frac{x_2}{s_2} + \cdots + \hat{\beta}_p s_p \frac{x_p}{s_p} = \hat{\beta}_1 + \hat{\beta}_2(\boldsymbol{W}, Y) w_2 + \cdots + \hat{\beta}_p(\boldsymbol{W}, Y) w_p.$$

Hence $\hat{\boldsymbol{\beta}}(\boldsymbol{W}, \boldsymbol{Y}) = (\hat{\beta}_1, \hat{\beta}_2 s_2, ..., \hat{\beta}_p s_p)^T = \hat{\boldsymbol{D}}_n^{-1}\hat{\boldsymbol{\beta}}(\boldsymbol{X}, Y)$ where $\hat{\boldsymbol{\beta}}(\boldsymbol{X}, Y) = (\hat{\beta}_1, \hat{\beta}_2, ..., \hat{\beta}_p)^T$.

For the scaled predictors, assume $\hat{\boldsymbol{D}}_n \xrightarrow{P} \boldsymbol{D} = diag(1, 1/\sigma_2, ..., 1/\sigma_p)$ where each $\sigma_i > 0$. This assumption often holds if the $\boldsymbol{x}_i$ are iid from some population. Let $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{X}, \boldsymbol{Y})$. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}(\boldsymbol{W}, \boldsymbol{Y}) - \boldsymbol{D}^{-1}\boldsymbol{\beta}) = \sqrt{n}(\hat{\boldsymbol{D}}_n^{-1}\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{D}}_n^{-1}\boldsymbol{\beta} + \hat{\boldsymbol{D}}_n^{-1}\boldsymbol{\beta} - \boldsymbol{D}^{-1}\boldsymbol{\beta})$$
$$= \sqrt{n}\hat{\boldsymbol{D}}_n^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + \sqrt{n}(\hat{\boldsymbol{D}}_n^{-1} - \boldsymbol{D}^{-1})\boldsymbol{\beta} = \boldsymbol{z}_n + \boldsymbol{b}_n$$

where $\boldsymbol{z}_n = \sqrt{n}\hat{\boldsymbol{D}}_n^{-1}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{D}^{-1}\boldsymbol{V}_{\boldsymbol{x}}\boldsymbol{D}^{-1})$ if $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{D} N_p(\boldsymbol{0}, \sigma^2\boldsymbol{V}_{\boldsymbol{x}})$. Note that $\hat{\boldsymbol{D}}_n^{-1}\hat{\boldsymbol{\beta}} \xrightarrow{P} \boldsymbol{D}^{-1}\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{W}, \boldsymbol{Y})$. Now

$$\boldsymbol{b}_n = \begin{pmatrix} 0 \\ \sqrt{n}(\hat{\sigma}_2 - \sigma_2)\beta_2 \\ \vdots \\ \sqrt{n}(\hat{\sigma}_p - \sigma_p)\beta_p \end{pmatrix} = \begin{pmatrix} 0 \\ b_{2,n} \\ \vdots \\ b_{p,n} \end{pmatrix} = O_p(1)$$

if $\sqrt{n}(\hat{\sigma}_i - \sigma_i) \xrightarrow{D} N(0, \tau_i^2)$. Then $b_{i,n} \xrightarrow{D} N(0, \beta_i^2 \tau_i^2)$ for $i = 2, ..., p$. Thus $\sqrt{n}(\hat{\boldsymbol{\beta}}(\boldsymbol{W}, \boldsymbol{Y}) - \boldsymbol{D}^{-1}\boldsymbol{\beta})$ does not converge in distribution to $\boldsymbol{z} \sim N_p(\boldsymbol{0}, \sigma^2 \boldsymbol{D}^{-1}\boldsymbol{V}_{\boldsymbol{x}}\boldsymbol{D}^{-1})$ unless $\boldsymbol{b}_n \xrightarrow{P} \boldsymbol{0}$.

Using the scaled data $(\boldsymbol{W}, Y)$ in the OLS software gives an incorrect normal approximation $\hat{\boldsymbol{\beta}}(\boldsymbol{W}, Y) \approx N_p(\boldsymbol{\beta}(\boldsymbol{W}, Y), MSE\ n\ (\boldsymbol{W}^T\boldsymbol{W})^{-1}) =$

$$N_p(\boldsymbol{D}^{-1}\boldsymbol{\beta}(\boldsymbol{X}, Y), MSE\ n\ \hat{\boldsymbol{D}}_n^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\boldsymbol{D}}_n^{-1}).$$

Hence confidence intervals, confidence regions, and many tests of hypotheses will no longer be valid. An important exception occurs for the partial F tests of the form $H_0 : \boldsymbol{L}_O\boldsymbol{\beta} = \boldsymbol{0}$ with $\boldsymbol{c} = \boldsymbol{0}$ and $\boldsymbol{L}_O$ a full rank $k \times p$ matrix where $\boldsymbol{L}_O\boldsymbol{\beta} = \boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_k})^T$ and $O = \{i_1, ..., i_k\}$. For such a test, we would like to leave the predictors $\boldsymbol{L}_O\boldsymbol{x} = \boldsymbol{x}_O = (x_{i_1}, ..., x_{i_k})^T$ out of the regression model, resulting in a reduced model. Note that the jth row of L_O has a 1 in the i_jth position, with all other entries equal to 0.

Let the ijth element of a $p \times m$ matrix $\boldsymbol{A}$ be a_{ij}. Then $\boldsymbol{A} = (a_{ij})$. Thus $\boldsymbol{L}_O\boldsymbol{A} = \boldsymbol{A}_O = (a_{i_a,j})$ where the ath row of $\boldsymbol{A}_O$ is the i_ath row of $\boldsymbol{A}$ for $a = 1, ..., k$. Similarly, if $\boldsymbol{C} = (c_{ij})$ is a $p \times p$ matrix, then

$$\boldsymbol{L}_O\boldsymbol{C}\boldsymbol{L}_O^T = \boldsymbol{C}_{OO} = \begin{pmatrix} c_{i_1,i_1} & c_{i_1,i_2} & \cdots & c_{i_1,i_k} \\ c_{i_2,i_1} & c_{i_2,i_2} & \cdots & c_{i_2,i_k} \\ \vdots & \vdots & \ldots & \vdots \\ c_{i_k,i_1} & c_{i_k,i_2} & \cdots & c_{i_k,i_k} \end{pmatrix} = (c_{i_a,i_b}).$$

Let $\boldsymbol{Q} = diag(d_1, ..., d_p)$ be a $p \times p$ diagonal matrix with diagonal elements $d_1, ..., d_p$. Let $\boldsymbol{H} = \boldsymbol{Q}\boldsymbol{A} = (h_{ij}) = (d_i a_{ij})$. Then $\boldsymbol{L}_O\boldsymbol{Q}\boldsymbol{A} = \boldsymbol{L}_O\boldsymbol{H} = \boldsymbol{H}_O = (h_{i_a,j}) = (d_{i_a}a_{i_a,j}) = \boldsymbol{Q}_{OO}\boldsymbol{A}_{OO}$. Let $\boldsymbol{B} = \boldsymbol{Q}\boldsymbol{C}\boldsymbol{Q} = (b_{ij}) = (d_i d_j c_{ij})$. Then $\boldsymbol{L}_O\boldsymbol{B}\boldsymbol{L}_O^T = \boldsymbol{B}_{OO} = (b_{i_a,i_b}) = (d_{i_a}d_{i_b}c_{i_a,i_b}) = \boldsymbol{Q}_{OO}\boldsymbol{C}_{OO}\boldsymbol{Q}_{OO}$.

Theorem 2. For the test $H_0 : \boldsymbol{L}_O\boldsymbol{\beta} = \boldsymbol{0}$, the partial F test statistics from the scaled data and the unscaled data are the same.

Proof. The result holds if

$$(\boldsymbol{L}_O\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}_O(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_O^T]^{-1}(\boldsymbol{L}_O\hat{\boldsymbol{\beta}}) = (\boldsymbol{L}_O\hat{\boldsymbol{\beta}}(\boldsymbol{W}, Y))^T[\boldsymbol{L}_O(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{L}_O^T]^{-1}(\boldsymbol{L}_O\hat{\boldsymbol{\beta}}(\boldsymbol{W}, Y)).$$

By the above remarks, $\boldsymbol{L}_O\hat{\boldsymbol{D}}_n\boldsymbol{L}_O^T = \hat{\boldsymbol{D}}_{OO} = diag(1/s_{i_1}, ..., 1/s_{i_k})$ where we define $s_1 = 1$. Let $\boldsymbol{Q} = \boldsymbol{D}_n^{-1}$ and $\boldsymbol{C} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

Then $\boldsymbol{L}_O\hat{\boldsymbol{\beta}}(\boldsymbol{W}, Y) = \boldsymbol{L}_O\hat{\boldsymbol{D}}_n^{-1}\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_O(\boldsymbol{W}, Y) = \hat{\boldsymbol{D}}_{OO}^{-1}\hat{\boldsymbol{\beta}}_O = \hat{\boldsymbol{D}}_{OO}^{-1}\boldsymbol{L}_O\hat{\boldsymbol{\beta}}$, while

$$\boldsymbol{L}_O(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{L}_O^T = \boldsymbol{L}_O(\hat{\boldsymbol{D}}_n\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{D}}_n)^{-1}\boldsymbol{L}_O^T = \boldsymbol{L}_O\hat{\boldsymbol{D}}_n^{-1}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\hat{\boldsymbol{D}}_n^{-1}\boldsymbol{L}_O^T$$

$$= \hat{\boldsymbol{D}}_{OO}^{-1}(\boldsymbol{X}^T\boldsymbol{X})_{OO}^{-1}\hat{\boldsymbol{D}}_{OO}^{-1} = \hat{\boldsymbol{D}}_{OO}^{-1}\boldsymbol{L}_O(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_O^T\hat{\boldsymbol{D}}_{OO}^{-1}.$$

Thus $(\boldsymbol{L}_O\hat{\boldsymbol{\beta}}(\boldsymbol{W}, Y))^T[\boldsymbol{L}_O(\boldsymbol{W}^T\boldsymbol{W})^{-1}\boldsymbol{L}_O^T]^{-1}(\boldsymbol{L}_O\hat{\boldsymbol{\beta}}(\boldsymbol{W}, Y)) =$

$$(\hat{\boldsymbol{D}}_{OO}^{-1}\boldsymbol{L}_O\hat{\boldsymbol{\beta}})^T[\hat{\boldsymbol{D}}_{OO}^{-1}\boldsymbol{L}_O(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_O^T\hat{\boldsymbol{D}}_{OO}^{-1}]^{-1}\hat{\boldsymbol{D}}_{OO}^{-1}\boldsymbol{L}_O\hat{\boldsymbol{\beta}} =$$

$$(\boldsymbol{L}_O\hat{\boldsymbol{\beta}})^T[\boldsymbol{L}_O(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{L}_O^T]^{-1}(\boldsymbol{L}_O\hat{\boldsymbol{\beta}}),$$

proving the theorem. $\square$

Let $\boldsymbol{x}_i^T$ be the ith row of $\boldsymbol{X}$, and let $\boldsymbol{w}_i^T$ be the ith row of $\boldsymbol{W}$. Let $\hat{\beta}_i = \hat{\beta}_i(\boldsymbol{x}, Y)$ be the ith OLS estimator of $\beta_i = \beta_i(\boldsymbol{x}, Y)$ where $(\boldsymbol{x}, Y)$ denotes that the Y were regressed on the $\boldsymbol{x}$. Similarly, $\hat{\beta}_i(\boldsymbol{w}, Y)$ is the estimator when the Y are regressed on the $\boldsymbol{w}_i$. Let $[L_{in}, U_{in}] = \hat{\beta}_i \pm t_{1-\alpha/2,n-p} SE(\hat{\beta}_i)$ be the large sample $100(1-\alpha)\%$ confidence interval CI for β_i. Let $\sigma_i^2 = Var(x_i)$ for $i = 2, ..., p$. Then $\beta_i(\boldsymbol{w}, Y) = \sigma_i \beta_i(\boldsymbol{x}, Y)$ for $i = 2, ..., p$, and the "CI" for $\beta_i(\boldsymbol{w}, Y)$ is $[s_i L_{in}, s_i U_{in}]$. This result holds since $(\boldsymbol{W}^T \boldsymbol{W})^{-1} = \hat{\boldsymbol{D}}_n^{-1} (\boldsymbol{X}^T \boldsymbol{X})^{-1} \hat{\boldsymbol{D}}_n^{-1}$. Scaling does not change the MSE, hence $SE[\hat{\beta}_i(\boldsymbol{w}, Y)] = s_i SE[\hat{\beta}_i(\boldsymbol{x}, Y)]$ for $i = 2, ..., p$ where s_i^2 is the usual unbiased estimator of σ_i^2. If $\beta_i(\boldsymbol{w}, Y) = \beta_i(\boldsymbol{x}, Y) = 0$, then $\beta_i = 0$ is in the interval $[L_{in}, U_{in}]$ if and only if $\beta_i(\boldsymbol{w}, Y) = \sigma_i \beta_i(\boldsymbol{x}, Y) = 0$ is in the "CI" $[s_i L_{in}, s_i U_{in}]$ since $s_i > 0$. Hence in the simulation where $\beta_i = 0$, the coverage of the CI for $\beta_i(\boldsymbol{x}, Y)$ and the coverage of the "CI" for $\beta_i(\boldsymbol{w}, Y)$ will be exactly the same. When $\beta_i \neq 0$, we expect that the coverages will differ, and that the "CI" for $\beta_i(\boldsymbol{w}, Y)$ will often have undercoverage. Here the coverage is the observed proportion of intervals that contained the population parameter. Hence if 5000 CIs for β_i were made, and 4750 of the CIs contained β_i, then the (observed) coverage is $4750/5000 = 0.95$.

The simulations used $\boldsymbol{L} = \boldsymbol{L}_O$ where $\boldsymbol{L}_O \boldsymbol{\beta} = \boldsymbol{c} = \boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_k})^T$ and $O = \{i_1, ..., i_k\}$.

3 Example and Simulations

Example. 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 x_1 and predictors $x_2 =$ the *population of the district in 1843*, $x_3 =$ the *number of married civilian men* in the district, $x_4 =$ the *number of married men in the military* in the district, and $x_5 =$ 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_3 are highly correlated but not equal. Similarly, x_4 and x_5 are highly correlated but not equal. Then $\hat{\boldsymbol{\beta}}_{OLS} = (242.3910, 0.00035, 0.9995, -0.2328, 0.1531)^T$, and forward selection with OLS and the C_p criterion used $\hat{\boldsymbol{\beta}}_{I,0} = (\hat{\beta}_1, 0, 1.0010, 0, 0)^T$. With the scaled data, $\hat{\boldsymbol{\beta}}_{OLS}(\boldsymbol{w}, Y) = (242.3910, 81.0283, 40877.4086, -104.8576, 66.2739)^T$.

Next, we describe a small OLS simulation study. The simulation used $\psi = 0$ and 0.5; and $k = 1$ and $p - 1$ where k and ψ are defined in the following paragraph.

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 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 $\boldsymbol{u}_i = \boldsymbol{A}\boldsymbol{w}_i$ so that $Cov(\boldsymbol{u}_i) = \boldsymbol{\Sigma}_{\boldsymbol{u}} = \boldsymbol{A}\boldsymbol{A}^T = (\sigma_{ij})$ where the diagonal entries $\sigma_{ii} = [1 + (m-1)\psi^2]$ and the off diagonal entries $\sigma_{ij} = [2\psi + (m-2)\psi^2]$. Hence the correlations are $cor(x_i, x_j) = \rho = (2\psi + (m-2)\psi^2)/(1 + (m-1)\psi^2)$ for $i \neq j$ where x_i and x_j are nontrivial predictors. If $\psi = 1/\sqrt{cp}$, then $\rho \to 1/(c+1)$ as $p \to \infty$ where $c > 0$. As ψ gets close to 1, the predictor vectors cluster about the line in the direction of $(1, ..., 1)^T$. Let $Y_i = 1 + 1x_{i,1} + \cdots + 1x_{i,k} + e_i$ 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.

When $\psi = 0$, the OLS confidence intervals for β_i should have length near $2t_{96,0.975}\sigma/\sqrt{n}$. Hence the scaled CI length $= \sqrt{n}$ CI length $\approx 2(1.96)\sigma = 3.92\sigma$ when the iid zero mean errors have variance σ^2. The simulation gave the average scaled CI lengths.

For the unscaled predictors, the simulation computed the large sample 95% CIs $[L_{in}, U_{in}]$ for β_i and $i = 1, ..., p$. The test for $H_0 : (\beta_{i_1}, \beta_{i_2})^T = (\beta_{i_1,0}, \beta_{i_2,0})^T$ was also performed using equation (11) with $\{i_1, i_2\} = \{p-1, p\}$. 5000 CIs were generated for each β_i, and the coverage was the proportion of times β_i was in its CI. Hence if β_1 was in its interval $4750/5000 = 0.95$, then the observed coverage was 0.95.

For the scaled predictors, the simulation computed the "95% CIs" $[s_i L_{in}, s_i U_{in}]$ for $\sigma_i \beta_i$ and $i = 1, ...p$ with $\{i_1, i_2\} = \{p-1, p\}$. The coverage was the proportion of times $\sigma_i \beta_i$ was in its "CI." The "test" for $H_0 : (\beta_{i_1}(\boldsymbol{w}, Y), \beta_{i_2}(\boldsymbol{w}, Y))^T = (\sigma_{i_1}\beta_{i_1,0}, \sigma_{i_2}\beta_{i_2,0})^T$ was also performed using equation (11) on the scaled data $\boldsymbol{W}$. The "test" is a valid large sample test if $(\beta_{i_1}, \beta_{i_2})^T = (0, 0)^T$. When $k = 1$, the test is valid and the "95% CI" can be used as a large sample test for $H_0 : \sigma_i \beta_i = 0$ except for β_2 since $\beta_3 = \cdots = \beta_p = 0$. When $k = p - 1$ the "test" and "95% CIs" are not valid large sample tests and CIs (except for β_1). The undercoverage can be rather large when the test is not valid.

Table 1: n=100,p=5,indices=(4,5), k=1

psi	etype	β_1	β_2	β_3	β_4	β_5	testcov
0, cov	1	0.9488	0.9452	0.9536	0.9482	0.9540	0.9526
u, len		4.0386	4.0676	4.0651	4.0634	4.0705	
0, cov	1	0.9488	0.8874	0.9536	0.9482	0.9540	0.9526
s, len		4.0386	4.0382	4.0396	4.0393	4.0380	
0.5, cov	1	0.9484	0.9530	0.9484	0.9510	0.9514	0.9504
u, len		4.0413	7.1015	7.1073	7.0910	7.1041	
0.5, cov	1	0.9484	0.9332	0.9484	0.9510	0.9514	0.9504
s, len		4.0413	9.3737	9.3807	9.3653	9.3790	

Each table has 4 lines for each type. The first line gives the coverages for the β_i while the second line gives the scaled CI lengths. There is not a length for testcov since the test corresponds to a confidence region instead of a confidence interval. The third and fourth lines are for the scaled data where cov is the proportion of times $\sigma_i \beta_i$ was in its interval. With 5000 runs, coverage between 0.94 and 0.96 suggests that the actual coverage is near the nominal large sample coverage of 0.95.

For Table 1, H_0 is true except for the scaled data with $\sigma_2 \beta_2$. With error type 1 and psi $= \psi = 0$, the average scaled CI lengths were near 4.07 which is not too far from 3.92 considering that $n = 100$ and $p = 5$. In the third line under β_2, the coverage is 0.8874. With $\psi = 0.5$, the sixth line under β_2 has coverage 0.9333. Increasing ψ often decreased the undercovaerage.

Table 2: n=100,p=5,indices=(4,5), k=5

psi	etype	β_1	β_2	β_3	β_4	β_5	testcov
0, cov	1	0.9448	0.9448	0.9536	0.9492	0.9500	0.9528
u, len		4.0419	4.0730	4.0797	4.0690	4.0689	
0, cov	1	0.9448	0.8976	0.8984	0.8882	0.8902	0.8654
s, len		4.0419	4.0421	4.0417	4.0422	4.0424	
0.5, cov	1	0.9548	0.9582	0.9486	0.9530	0.9472	0.9506
u, len		4.0431	7.0952	7.1066	7.1151	7.1100	
0.5, cov	1	0.9548	0.9360	0.9354	0.9338	0.9310	0.9130
s, len		4.0431	9.3555	9.3679	9.3722	9.3643	

For Table 2 with the scaled data, H_0 is only true for β_1. For the scaled data, the "CI" undercoverage was more severe for $\psi = 0$ than for $\psi = 0.5$, and the testcov was worse than that for the CIs. With the unscaled data, H_0 was always true.

4 Conclusions

For multiple linear regression with standardized data, OLS software tests of the form $H_0 : \boldsymbol{\beta}_O = \mathbf{0}$ are valid large sample tests where $\boldsymbol{\beta}_O = (\beta_{i_1}, ..., \beta_{i_k})^T$. However, OLS software does not give correct confidence intervals for $\beta_i(\boldsymbol{w}, Y) = \sigma_i \beta_i$ for $i = 2, ..., p$ unless $\beta_i = 0$.

Software

The *R* software was used in the simulations. See R Core Team (2024). Programs are in the Olive (2025) collections of *R* functions *slpack.txt*, available from (http://parker.ad.siu.edu/Olive/slpack.txt). The function `mlrsim` was used to make the tables.

References

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

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.

Olive, D.J. (2017), *Robust Multivariate Analysis*, Springer, New York, NY.

Olive, D.J. (2025), *Prediction and Statistical Learning*, online course notes, see (http://parker.ad.siu.edu/Olive/slearnbk.htm).

R Core Team (2024), "R: a Language and Environment for Statistical Computing," R Foundation for Statistical Computing, Vienna, Austria, (www.R-project.org).

Seber, G.A.F., and Lee, A.J. (2003), *Linear Regression Analysis,* 2nd ed., Wiley, New York, NY.

Sen, P.K., and Singer, J.M. (1993), *Large Sample Methods in Statistics: an Introduction with Applications,* Chapman & Hall, New York, NY.