Lecture 2

Multiple Linear Regression: Estimation

CHUNG-MING KUAN

Department of Finance & CRETA

National Taiwan University

February 24, 2022

Lecture Outline

1. Multiple Linear Regression: Estimation
   - Algebraic Properties of LS Estimation
   - Statistical Properties of LS Estimation
   - LS Estimation in Matrix Notations
   - Consequence of Over- and Under-Specification

Linear Specification

In practice, the systematic part of the dependent variable y may be better characterized by a collection of k ($k > 1$) explanatory variables, such that

$$y = f(x_1, \ldots, x_k) + u,$$

where u is the error term (non-systematic part of y). As in simple linear regression, it is convenient to postulate f as the linear function, and

$$y = \underbrace{\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k}_{\text{systematic part}} + \underbrace{u(\beta_0, \beta_1, \ldots, \beta_k)}_{\text{error}},$$

with $k+1$ unknown parameters β_0, $\beta_1, \ldots, \beta_k$.

Least-Squares Minimization

Given the sample data $(x_{i1}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots, n$, we have

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i, \quad i = 1, \ldots, n,$$

where $u_i = u_i(\beta_0, \beta_1, \ldots, \beta_k)$ is the i th error. Our goal now is to find a hyperplane that "best" fits the sample data. The best fit of data can be obtained by minimizing the sum of squared errors with respect to $\beta_0, \beta_1, \ldots, \beta_k$:

$$Q_n(\beta_0, \beta_1, \ldots, \beta_k) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2.$$

The FOCs of the LS problem now contain $k+1$ equations with $k+1$ unknowns:

$$\frac{\partial Q_n(\beta_0, \beta_1, \ldots, \beta_k)}{\partial \beta_0} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}) = 0,$$

$$\frac{\partial Q_n(\beta_0, \beta_1, \ldots, \beta_k)}{\partial \beta_1} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})x_{i1} = 0,$$

$$\vdots$$

$$\frac{\partial Q_n(\beta_0, \beta_1, \ldots, \beta_k)}{\partial \beta_k} = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})x_{ik} = 0.$$

The solutions to the FOCs are the OLS estimators: $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$. We shall present the analytic forms of these estimators using matrix notations later.

Remark: The OLS method does not require any assumptions on sample data, except that there should be no exact linear relations among regressors and the constant term. To see this, suppose $x_{i3} = x_{i1} + x_{i2}$ for all i. Then, the following two FOCs:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})x_{i1} = 0,$$
$$\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})x_{i2} = 0,$$

imply that the FOC: $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})x_{i3} = 0$ also holds. That is, when there is an exact linear relation among regressors, some FOC must be redundant, and the number of effective FOCs would be less than $k+1$. As such, he OLS estimators cannot be uniquely solved from the FOCs.

Given $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, the estimated regression hyperplane is:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_k x_k,$$

with the i th fitted value $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}$; the i th residual is $\hat{u}_i = y_i - \hat{y}_i$.

- $\hat{\beta}_j = \mathrm{d}\hat{y}/\mathrm{d}x_j$, still known as a "slope" parameter, predicts how much y would change when the j th regressor changes by one unit, while holding other regressors fixed. We usually say $\hat{\beta}_j$ is the marginal effect of x_j after the effects of other regressors are "controlled."
- $\hat{\beta}_j$ is not the same as the OLS estimate of regressing y on x_j only, because the latter is obtained without controlling other regressors; see the following slides.
- $\hat{\beta}_0$ is the intercept and predicts the level of y when $x_1 = \cdots = x_k = 0$.

A "Partialling Out" Interpretation

We shall use the following analytic formula to illustrate the marginal effect of the OLS estimator (we omit the proof):

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} \hat{r}_{i,1} y_i}{\sum_{i=1}^{n} \hat{r}_{i,1}^2},$$

where $\hat{r}_{i,1}$ are the i th OLS residuals of regressing x_1 on the constant one and $x_2, \ldots, x_k$.

- This is also the OLS estimator of regressing y on $\hat{r}_1$ (without the constant term) and represents the marginal effect of $\hat{r}_1$ on y.
- By definition, $\hat{r}_1$ is part of x_1 that is not linearly related with $x_2, \ldots, x_k$. Hence, $\hat{\beta}_1$ can be understood as the "pure" effect of x_1 on y, because the effects of $x_2, \ldots, x_k$ on x_1 have been "partialled out" or "purged away".

Note that $\hat{\beta}_1$ is, in general, not the same as the OLS estimator of regressing y on the constant one and x_1:

$$\hat{b}_1 = \frac{\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1) y_i}{\sum_{i=1}^{n}(x_{i,1} - \bar{x}_1)^2},$$

unless $x_{i,1} - \bar{x}_1 = \hat{r}_{i,1}$. These two estimators would coincide when $x_2, \ldots, x_k$ are not linearly related to x_1, so that regressing x_1 on the constant one and $x_2, \ldots, x_k$ yields:

$$x_{i,1} = \bar{x}_1 + \hat{r}_{i,1}.$$

On the other hand, this equality above fails when $x_2, \ldots, x_k$ are linearly related to x_1, so that $\hat{\beta}_1 \neq \hat{b}_1$. As $\hat{b}_1$ is the marginal effect of x_1 on y without controlling other regressors, it involves both the "pure" effect ($\hat{\beta}_1$) of x_1 on y and the "indirect" effects of $x_2, \ldots, x_k$ on y via x_1.

Similarly, let $\hat{r}_{i,j}$ denote the i th OLS residuals of regressing x_j on 1 and x_h, $h \neq j$. Then,

$$\hat{\beta}_j = \frac{\sum_{i=1}^{n} \hat{r}_{i,j} y_i}{\sum_{i=1}^{n} \hat{r}_{i,j}^2}, \quad j = 2, \ldots, k,$$

which represent the "pure" effect of x_j on y when other regressors (x_h, $h \neq j$) are controlled. In general, $\hat{\beta}_j$ is not the same as $\hat{b}_j$, the OLS estimator of regressing y on the constant one and x_j only. These results show that including all relevant variables in a multiple linear regression is important because it allows us to identify the "pure" effect of each regressor.

Algebraic Properties

- Plugging $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ into the FOCs we obtain:

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) = \sum_{i=1}^{n} \hat{u}_i = 0,$$

so that the positive and negative residuals cancel out, and

$$\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_k x_{ik}) x_{ij} = \sum_{i=1}^{n} \hat{u}_i x_{ij} = 0, \quad j = 1, \ldots, k,$$

so that the sample covariance between x_{ij} and $\hat{u}_i$ is zero.

- As $\sum_{i=1}^{n} \hat{u}_i = 0$, we can see:

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}_1 + \cdots + \hat{\beta}_k \bar{x}_k,$$

which shows the estimated regression hyperplane must pass through $(\bar{x}_1, \ldots, \bar{x}_k, \bar{y})$.

- Knowing that $\sum_{i=1}^{n} \hat{u}_i = 0$ and $\sum_{i=1}^{n} \hat{u}_i x_{ij} = 0$, we have

$$\begin{aligned}\sum_{i=1}^{n} \hat{u}_i \hat{y}_i &= \sum_{i=1}^{n} \hat{u}_i(\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik}) \\ &= \hat{\beta}_0 \sum_{i=1}^{n} \hat{u}_i + \hat{\beta}_1 \sum_{i=1}^{n} \hat{u}_i x_{i1} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} \hat{u}_i x_{ik} \\ &= 0,\end{aligned}$$

  so that the sample covariance between the fitted values and the residuals is also zero.

- It follows that

$$\sum_{i=1}^{n} \hat{u}_i y_i = \sum_{i=1}^{n} \hat{u}_i(\hat{y}_i + \hat{u}_i) = \sum_{i=1}^{n} \hat{u}_i^2.$$

Goodness of Fit: R^2

We have learned that the total sum of squares (SST) is the sum of the residual sum of squares (SSR) and the explained sum of squares (SSE):

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n} \hat{u}_i^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2.$$

For multiple regressions, we also use the coefficient of determination as a measure of goodness of fit:

$$R^2 = \text{SSE}/\text{SST} = 1 - \text{SSR}/\text{SST},$$

which measures the proportion of the total variation (SST) of y_i due to the variation of $\hat{y}_i$ (SSE). Again, $0 \leq R^2 \leq 1$, and a specification has a better (worse) fit of data if its R^2 is closer to one (zero).

Drawback: R^2 is non-decreasing in the number of regressors. That is, adding regressors to a regression will result in higher R^2. As such, one would tend to choose a more complex model if R^2 is the criterion for determining a model. To see this, consider two estimated regressions:

$$\begin{aligned} y_i &= \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \tilde{\beta}_2 x_{i2} + \tilde{v}_i, \\ y_i &= \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{3i} + \hat{u}_i. \end{aligned}$$

Note that the former can be written as:

$$y_i = \tilde{\beta}_0 + \tilde{\beta}_1 x_{i1} + \tilde{\beta}_2 x_{i2} + 0 \cdot x_{3i} + \tilde{v}_i.$$

Clearly, the estimates $\tilde{\beta}_0, \tilde{\beta}_1, \tilde{\beta}_2, 0$ do not minimize the sum of squared errors in the 3-regressor regression, because the last coefficient is restricted to zero. This shows that R^2 of a 2-regressor regression must be smaller (or no greater) than R^2 of the regression with these two regressors and an additional regressor.

Goodness of Fit: Adjusted R^2

To avoid the problem of non-decreasing R^2, a modified measure of goodness of fit is usually adopted. This is known as $\bar{R}^2$, defined as R^2 adjusted for the degrees of freedom:

$$\bar{R}^2 = 1 - \frac{\mathrm{SSR}/(n-k-1)}{\mathrm{SST}/(n-1)}.$$

It can also be written as the difference between R^2 and a penalty term:

$$\bar{R}^2 = R^2 - \frac{k}{n-k-1}(1-R^2),$$

where the penalty term depends on the trade-off between model complexity (k) and model explanatory ability (R^2). Thus, $\bar{R}^2$ may be decreasing when the contribution of additional regressors to model fitness does not outweigh the penalty on model complexity. In practice, we compare models based on $\bar{R}^2$, rather than R^2.

Statistical Properties

The assumption below is analogous to Classical Assumption I.

Classical Assumption II

The random variables y_i, $i = 1, \ldots, n$, are such that:

(i) $\mathbb{E}(y_i) = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}$ for some $b_0, b_1, \ldots, b_k$, where $x_{i1}, \ldots, x_{ik}$ are non-random;

(ii) $\text{var}(y_i) = \sigma_o^2$, $\text{cov}(y_i, y_j) = 0$ for $i \neq j$.

Letting ε_i denote the errors evaluated at $b_0, b_1, \ldots, b_k$, this assumption is equivalent to: $y_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + \varepsilon_i$, with $x_{i1}, \ldots, x_{ik}$ non-random, $\mathbb{E}(\varepsilon_i) = 0$, $\text{var}(\varepsilon_i) = \sigma_o^2$, and $\text{cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$.

Unbiasedness

.

Unbiasedness of the OLS Estimators

Under Classical Assumption II(i), $\hat{\beta}_j$ are unbiased for b_j, $j = 0, 1, \ldots, k$.

Proof: Note that $\hat{r}_{i,1}$ are non-random because they are the OLS residuals of regressing x_1 on the constant one and $x_2, \ldots, x_k$. By Classical Assumption II(i) and the formula for the "partialling out" argument,

$$\begin{aligned}
\mathbb{E}(\hat{\beta}_1) &= \frac{\sum_{i=1}^n \hat{r}_{i,1} \mathbb{E}(y_i)}{\sum_{i=1}^n \hat{r}_{i,1}^2} \\
&= \frac{\sum_{i=1}^n \hat{r}_{i,1}(b_0 + b_1 x_{i1} + \cdots + b_k x_{ik})}{\sum_{i=1}^n \hat{r}_{i,1}^2} \\
&= \frac{b_0 \sum_{i=1}^n \hat{r}_{i,1} + b_1 \sum_{i=1}^n \hat{r}_{i,1} x_{i1} + \cdots + b_k \sum_{i=1}^n \hat{r}_{i,1} x_{ik}}{\sum_{i=1}^n \hat{r}_{i,1}^2}.
\end{aligned}$$

Recall that the FOCs of the LS problem imply: $\sum_{i=1}^n \hat{r}_{i,1} = 0$, $\sum_{i=1}^n \hat{r}_{i,1} x_{i2} = 0, \ldots$, $\sum_{i=1}^n \hat{r}_{i,1} x_{ik} = 0$. Consequently,

$$\mathbb{E}(\hat{\beta}_1) = \frac{b_1 \sum_{i=1}^n \hat{r}_{i,1} x_{i1}}{\sum_{i=1}^n \hat{r}_{i,1}^2}.$$

By the algebraic properties of OLS regression (Verify!),

$$\sum_{i=1}^n \hat{r}_{i,1} x_{i1} = \sum_{i=1}^n \hat{r}_{i,1}^2,$$

so that $\mathbb{E}(\hat{\beta}_1) = b_1$. This proves unbiasedness of $\hat{\beta}_1$. Similarly, we can show $\mathbb{E}(\hat{\beta}_j) = b_j$, $j = 2, \ldots, k$.

Variance of the OLS Estimators

Under Classical Assumption II(i) and (ii),

$$\mathrm{var}(\hat{\beta}_j) = \sigma_o^2 \frac{1}{\sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}, \quad j = 1, \ldots, k,$$

where R_j^2 is R^2 of regressing x_j on 1 and other regressors x_h, $h \neq j$; $\mathrm{var}(\hat{\beta}_0)$ has a different form and is omitted.

Remarks

1. When x_j is highly linearly related to other regressors, R_j^2 would be high, so that $\mathrm{var}(\hat{\beta}_j)$ is large; otherwise, the OLS estimators have a smaller variance and hence are more stable.
2. When the regressors satisfy an exact linear relation so that $R_j^2 = 1$, the variance would be infinitely large, and the OLS method breaks down, as discussed earlier.

Proof of var($\hat{\beta}_j$)

To derive var($\hat{\beta}_1$), note that under Classical Assumptions,

$$\begin{aligned}\operatorname{var}(\hat{\beta}_1) &= \operatorname{var}\left(\frac{\sum_{i=1}^n \hat{r}_{i,1} y_i}{\sum_{i=1}^n \hat{r}_{i,1}^2}\right) = \frac{\sum_{i=1}^n \hat{r}_{i,1}^2 \operatorname{var}(y_i)}{\left(\sum_{i=1}^n \hat{r}_{i,1}^2\right)^2} \\ &= \sigma_o^2 \frac{\sum_{i=1}^n \hat{r}_{i,1}^2}{\left(\sum_{i=1}^n \hat{r}_{i,1}^2\right)^2} = \sigma_o^2 \frac{1}{\sum_{i=1}^n \hat{r}_{i,1}^2}.\end{aligned}$$

For the regression of x_1 on the constant one and $x_2, \ldots, x_k$,

$$\begin{aligned}\sum_{i=1}^n \hat{r}_{i,1}^2 &= \mathrm{SSR}_1 = \mathrm{SST}_1 - \mathrm{SSE}_1 = \mathrm{SST}_1(1 - \mathrm{SSE}_1/\mathrm{SST}_1) \\ &= \sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 (1 - R_1^2),\end{aligned}$$

proving the formula for var($\hat{\beta}_1$). Other var($\hat{\beta}_j$) can be derived similarly.

As $\hat{u}_i$ in multiple linear regression must satisfy $k+1$ FOCs and hence lose $k+1$ degrees of freedom, the OLS estimator of σ_o^2 is computed as:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2.$$

This estimator is unbiased for σ_o^2 (proof omitted here).

Unbiasedness of $\hat{\sigma}^2$

Under Classical Assumption II(i) and (ii), $\mathbb{E}(\hat{\sigma}^2) = \sigma_o^2$.

Replacing σ_o^2 with $\hat{\sigma}^2$, we obtain the following variance estimators:

$$\widehat{\mathrm{var}(\hat{\beta}_j)} = \hat{\sigma}^2 \frac{1}{\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2(1 - R_j^2)}, \quad j = 1, \ldots, k,$$

which are also unbiased for $\mathrm{var}(\hat{\beta}_j)$. The square root of $\widehat{\mathrm{var}(\hat{\beta}_j)}$ is referred to as the standard error of $\hat{\beta}_j$.

Efficiency of the OLS Estimators

From the OLS formula:

$$\hat{\beta}_j = \frac{\sum_{i=1}^n \hat{r}_{i,j} y_i}{\sum_{i=1}^n \hat{r}_{i,j}^2}, \quad j = 1, \ldots, k,$$

we can see that $\hat{\beta}_j$ is a linear combination of y_i: $\sum_{i=1}^n a_{i,j} y_i$, with $a_{i,j} = \hat{r}_{i,j} / \sum_{i=1}^n \hat{r}_{i,j}^2$, i.e., an estimator linear in y_i. The result below asserts that, compared with all linear unbiased estimators for b_j, $\hat{\beta}_j$ is the best in the sense that it has the smallest variance or is the most efficient. A proof will be given later using matrix notations.

Gauss-Markov Theorem

Under Classical Assumption II(i) and (ii), $\hat{\beta}_j$ are the best linear unbiased estimators (BLUEs) for b_j, $j = 0, 1, \ldots, k$.

Example: Wage Regression with 2 Regressors

The estimated wage model based on Taiwan's 2010 male data (11561 obs): The dependent variable is log(wage), and the estimated parameters are:

$$
\begin{array}{cccl}
3.8939 & +\ 0.0800\ \text{educ} & +\ 0.0166\ \text{exper}, & \bar{R}^2 = 0.2893, \\
(0.0198) & (0.0012) & (0.0003) & \hat{\sigma} = 0.3595; \\
4.5929 & +\ 0.0494\ \text{educ}, & & \bar{R}^2 = 0.1329, \\
(0.0156) & (0.0012) & & \hat{\sigma} = 0.3971; \\
5.1208 & & +\ 0.0059\ \text{exper}, & \bar{R}^2 = 0.0263, \\
(0.0073) & & (0.0003) & \hat{\sigma} = 0.4208;
\end{array}
$$

where the numbers in the parentheses are the standard errors. Note that for the regression with two regressors, $\bar{R}^2$ is much larger than those with only one regressor, and the marginal effect of educ is also larger (8%) when exper is controlled (Why?).

Example: Wage Regression with 3 Regressors

Adding a new regressor exper^2, the estimated parameters are:

$$\begin{array}{cccc} 3.790 & +\,0.0779\,\text{educ} & +\,0.0365\,\text{exper} & -\,0.0005\,\text{exper}^2, \\ (0.0199) & (0.0012) & (0.0009) & (0.00002) \end{array}$$
$$\bar{R}^2 = 0.319, \quad \hat{\sigma} = 0.3519.$$

- The new regressor exper^2 is a nonlinear function of exper, so that there is no linear relation among regressors. Note that $\bar{R}^2$ increases.
- The marginal effect of exper is $(0.0365 - 0.001\,\text{exper})$. Setting this effect to zero, we find that the effect of the years of working experience on log(wage) reaches the maximum when exper $= 36.5$. Thus, log(wage) increases with a decreasing rate (-0.001) before experience reaches 36.5 years.

LS Estimation in Matrix Notations

The specification is: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}(\boldsymbol{\beta})$, where

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix},$$

and $\boldsymbol{u}(\boldsymbol{\beta}) = (u_1(\boldsymbol{\beta})\ u_2(\boldsymbol{\beta}) \ldots u_n(\boldsymbol{\beta}))'$. The LS problem is to minimize

$$Q_n(\boldsymbol{\beta}) := (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

The FOCs are $-2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) = \boldsymbol{0}$, leading to the normal equations:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y},$$

where $\boldsymbol{X}'\boldsymbol{X}$ is $(k+1) \times (k+1)$ and $\boldsymbol{X}'\boldsymbol{y}$ is $(k+1) \times 1$.

The OLS Estimator

Pre-multiplying both sides of the normal equations by $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ (provided that the inverse exists), we obtain the OLS estimator of β:

$$\hat{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Remarks:

- The inverse $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists if $\boldsymbol{X}$ is of full column rank $k+1$, i.e., any column of $\boldsymbol{X}$ is not a linear combination of other columns. As the inverse matrix $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is unique, $\hat{\beta}$ is also unique.
- When $\boldsymbol{X}$ is not of full column rank, we say there exists exact multicollinearity among regressors. In this case, the matrix $\boldsymbol{X}'\boldsymbol{X}$ is not invertible, and the OLS method breaks down.

Given the OLS estimator $\hat{\beta}$, the vector of the OLS fitted values is $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\beta}$, and the vector of the OLS residuals is $\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}}$. The FOCs yield the following algebraic properties:

$$\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\hat{\beta}) = \boldsymbol{X}'\hat{\boldsymbol{u}} = \begin{bmatrix} \sum_{i=1}^{n} \hat{u}_i \\ \sum_{i=1}^{n} x_{i1}\hat{u}_i \\ \vdots \\ \sum_{i=1}^{n} x_{ik}\hat{u}_i \end{bmatrix} = \mathbf{0},$$

$$\hat{\boldsymbol{y}}'\hat{\boldsymbol{u}} = \sum_{i=1}^{n} \hat{y}_i \hat{u}_i = \hat{\beta}'\boldsymbol{X}'\hat{\boldsymbol{u}} = 0.$$

These are exactly the algebraic properties we observed earlier.

Some Matrix Results

- Given two $n \times 1$ vectors, $\boldsymbol{x}$ and $\boldsymbol{z}$, their inner product is defined as $\boldsymbol{x}'\boldsymbol{z} = \sum_{i=1}^{n} x_i z_i$.
- The Euclidean norm of $\boldsymbol{x}$ is

$$\|\boldsymbol{x}\| = (\boldsymbol{x}'\boldsymbol{x})^{1/2} = \left(\sum_{i=1}^{n} x_i^2 \right)^{1/2}.$$

- The inner product $\boldsymbol{x}'\boldsymbol{z} = \|\boldsymbol{x}\| \, \|\boldsymbol{z}\| \cos\theta$, where θ is the angle between $\boldsymbol{x}$ and $\boldsymbol{z}$. Thus, $\boldsymbol{x}$ and $\boldsymbol{z}$ are said to be orthogonal if $\boldsymbol{x}'\boldsymbol{z} = 0$.
- The matrix $\boldsymbol{A}$ is said to be a projection matrix if it is idempotent ($\boldsymbol{A}\boldsymbol{A} = \boldsymbol{A}$). That is, given the projection of $\boldsymbol{x}$, $\boldsymbol{A}\boldsymbol{x}$, projecting it again will not alter the projection: $\boldsymbol{A}\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x}$.

- The projection matrix $\boldsymbol{A}$ is the orthogonal projection matrix if it is also symmetric ($\boldsymbol{A} = \boldsymbol{A}'$). To see this, write $\boldsymbol{x} = \boldsymbol{A}\boldsymbol{x} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}$. Then, provided that $\boldsymbol{A} = \boldsymbol{A}'$, we have

$$(\boldsymbol{A}\boldsymbol{x})'(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{x}'\boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = \boldsymbol{x}'(\boldsymbol{A} - \boldsymbol{A}\boldsymbol{A})\boldsymbol{x} = 0.$$

  This shows that $\boldsymbol{A}\boldsymbol{x}$ is orthogonal to $(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}$, so that $\boldsymbol{A}\boldsymbol{x}$ is the orthogonal projection of $\boldsymbol{x}$. Note that when $\boldsymbol{A}$ is an orthogonal projection matrix, so is $\boldsymbol{I} - \boldsymbol{A}$.

- For two $n \times n$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, $\boldsymbol{A} - \boldsymbol{B}$ is positive semi-definite (p.s.d.) if $\boldsymbol{x}'(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{x} \geq 0$ for all $\boldsymbol{x}$ such that $\|\boldsymbol{x}\| = 1$; $\boldsymbol{A} - \boldsymbol{B}$ is positive definite (p.d.) if the inequality above holds strictly.

Geometric Illustration

Let $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. It can be seen that $\boldsymbol{P}$ is symmetric and idempotent, and hence an orthogonal projection matrix. Note that $\boldsymbol{P}$ projects vectors onto the space spanned by the column vectors of $\boldsymbol{X}$, $\mathrm{span}(\boldsymbol{X})$. Similarly, $\boldsymbol{I} - \boldsymbol{P}$ is the orthogonal projection matrix that projects vectors onto the orthogonal complement of $\mathrm{span}(\boldsymbol{X})$, $\mathrm{span}(\boldsymbol{X})^{\perp}$. Thus, $\boldsymbol{P}\boldsymbol{X} = \boldsymbol{X}$, and $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$.

- The vector of the OLS fitted values is

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{P}\boldsymbol{y},$$

  the orthogonal projection of $\boldsymbol{y}$ onto $\mathrm{span}(\boldsymbol{X})$.
- The residual vector is $\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}$, the orthogonal projection of $\boldsymbol{y}$ onto $\mathrm{span}(\boldsymbol{X})^{\perp}$, and must be orthogonal to $\hat{\boldsymbol{y}}$. That is, $\hat{\boldsymbol{y}}'\hat{\boldsymbol{u}} = 0$.

- Compared with any other projection of $\boldsymbol{y}$ (say, $\boldsymbol{A}\boldsymbol{y}$), the orthogonal projection $\boldsymbol{P}\boldsymbol{y}$ provides the "best approximation" to $\boldsymbol{y}$. Indeed, it is easily verified that the Euclidean norm of $\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = \hat{\boldsymbol{u}}$ is the smallest possible:

$$\|\hat{\boldsymbol{u}}\| = \|\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y}\| \leq \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}\|,$$

  for any other projection matrix $\boldsymbol{A}$. This is precisely what the LS minimization problem does.
- The algebraic property $\boldsymbol{X}'\hat{\boldsymbol{u}} = \mathbf{0}$ holds because $\hat{\boldsymbol{u}}$ is in $\text{span}(\boldsymbol{X})^{\perp}$ and hence must be orthogonal to every column vector of $\boldsymbol{X}$.

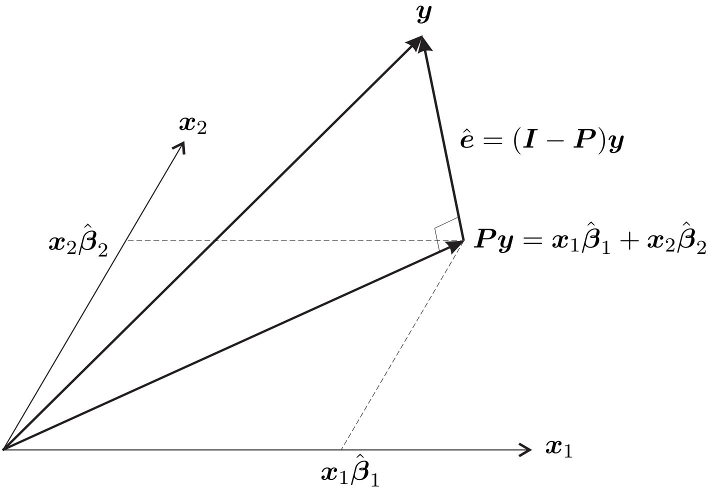

Figure: The orthogonal projection of $\boldsymbol{y}$ onto span($\boldsymbol{x}_1$,$\boldsymbol{x}_2$).

Example: Simple Linear Regression

The simple linear regression in matrix notations: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$, with $\boldsymbol{\beta} = (\beta_0 \ \beta_1)'$,

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix}, \quad \boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} n & \sum_{i=1}^n x_i \\ \sum_{i=1}^n x_i & \sum_{i=1}^n x_i^2 \end{bmatrix},$$

$$\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_i y_i \end{bmatrix}.$$

In this case, $\boldsymbol{X}$ has rank 2 (full column rank) provided that x_i are not a constant. For if x_i are a constant, the second column of $\boldsymbol{X}$ would be a multiple of the first column, so that the rank of $\boldsymbol{X}$ is 1.

When $\boldsymbol{X}$ has full column rank, $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists and reads:

$$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2} \begin{bmatrix} \sum_{i=1}^{n} x_i^2 & -\sum_{i=1}^{n} x_i \\ -\sum_{i=1}^{n} x_i & n \end{bmatrix}.$$

Noting that $n\sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2 = n\sum_{i=1}^{n}(x_i - \bar{x})^2$, it is readily verified that

$$(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \begin{bmatrix} \bar{y} - \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x} \\ \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \end{bmatrix},$$

which are exactly the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained earlier in the simple linear regression.

"Partialling Out" Interpretation in Matrix Notations

Frisch-Waugh-Lovell Theorem

For $\boldsymbol{y} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u}$, the OLS estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are:

$$\hat{\boldsymbol{\beta}}_1 = [\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y},$$
$$\hat{\boldsymbol{\beta}}_2 = [\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{X}_2]^{-1}\boldsymbol{X}_2'(\boldsymbol{I} - \boldsymbol{P}_1)\boldsymbol{y},$$

where $\boldsymbol{P}_1 = \boldsymbol{X}_1(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'$ and $\boldsymbol{P}_2 = \boldsymbol{X}_2(\boldsymbol{X}_2'\boldsymbol{X}_2)^{-1}\boldsymbol{X}_2'$.

Remark: Let $\tilde{\boldsymbol{X}}_1 = (\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1$, the matrix of residuals from regressing $\boldsymbol{X}_1$ on $\boldsymbol{X}_2$. As $\boldsymbol{I} - \boldsymbol{P}_2$ is idempotent, we have

$$\hat{\boldsymbol{\beta}}_1 = [\tilde{\boldsymbol{X}}_1'\tilde{\boldsymbol{X}}_1]^{-1}\tilde{\boldsymbol{X}}_1'\boldsymbol{y},$$

the "pure" marginal effect of $\boldsymbol{X}_1$ on $\boldsymbol{y}$, where the effect of $\boldsymbol{X}_2$ on $\boldsymbol{y}$ (via $\boldsymbol{X}_1$) has been "partialled out". The interpretation of $\hat{\boldsymbol{\beta}}_2$ is similar.

Proof: Letting $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]$ and $\boldsymbol{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$, we can write

$$\boldsymbol{y} = \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.$$

Pre-multiplying both sides by $\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)$, we have

$$\begin{aligned}&\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y}\\ &\quad = \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1 + \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2 + \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.\end{aligned}$$

Clearly, $(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_2 = \boldsymbol{0}$, so that the second term vanishes. As $\operatorname{span}(\boldsymbol{X}_2) \subseteq \operatorname{span}(\boldsymbol{X})$, we have $\operatorname{span}(\boldsymbol{X})^{\perp} \subseteq \operatorname{span}(\boldsymbol{X}_2)^{\perp}$, and hence $(\boldsymbol{I} - \boldsymbol{P}_2)(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{I} - \boldsymbol{P}$. It follows that the third term also vanishes because $\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}) = \boldsymbol{0}$. Consequently,

$$\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y} = \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1\hat{\boldsymbol{\beta}}_1,$$

from which we obtain the expression for $\hat{\boldsymbol{\beta}}_1$.

Statistical Properties

Classical Assumption II in Matrix Notations

The random vector $\boldsymbol{y}$ is such that:

(i) $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}_o$ for some $\boldsymbol{b}_o$, where $\boldsymbol{X}$ is non-random;

(ii) $\text{var}(\boldsymbol{y}) = \sigma_o^2 \boldsymbol{I}$.

This assumption is equivalent to: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}_o + \boldsymbol{\varepsilon}$, with $\boldsymbol{X}$ non-random, $\mathbb{E}(\boldsymbol{\varepsilon}) = \boldsymbol{0}$, $\text{var}(\boldsymbol{\varepsilon}) = \sigma_o^2 \boldsymbol{I}$.

- Unbiasedness: By Classical Assumption II(i),

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_o = \boldsymbol{b}_o.$$

- Variance: By Classical Assumption II(i) and (ii),

$$\begin{aligned}\text{var}(\hat{\boldsymbol{\beta}}) &= \text{var}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \\ &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\text{var}(\boldsymbol{y})]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.\end{aligned}$$

Gauss-Markov Theorem

Under Classical Assumption II(i) and (ii), $\hat{\beta}$ is the BLUE for $\boldsymbol{b}_o$.

Proof: Consider an arbitrary linear estimator $\check{\beta} = \boldsymbol{A}\boldsymbol{y}$, where $\boldsymbol{A}$ is a non-random matrix, say, $\boldsymbol{A} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' + \boldsymbol{C}$. Then, $\check{\beta} = \hat{\beta} + \boldsymbol{C}\boldsymbol{y}$, and

$$\operatorname{var}(\check{\beta}) = \operatorname{var}(\hat{\beta}) + \operatorname{var}(\boldsymbol{C}\boldsymbol{y}) + 2\operatorname{cov}(\hat{\beta}, \boldsymbol{C}\boldsymbol{y}).$$

By Classical Assumption II(i) and (ii), $\mathbb{E}(\check{\beta}) = \boldsymbol{b}_o + \boldsymbol{C}\boldsymbol{X}\boldsymbol{b}_o$, which is unbiased if, and only if, $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{0}$. The condition $\boldsymbol{C}\boldsymbol{X} = \boldsymbol{0}$ implies

$$\begin{aligned}
\operatorname{cov}(\hat{\beta}, \boldsymbol{C}\boldsymbol{y}) &= \mathbb{E}[(\hat{\beta} - \boldsymbol{b}_o)\boldsymbol{y}'\boldsymbol{C}'] = \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o)\boldsymbol{y}'\boldsymbol{C}'] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}[(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{b}_o)\boldsymbol{y}']\boldsymbol{C}' \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\operatorname{var}(\boldsymbol{y})\boldsymbol{C}' \\
&= \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{C}' = \boldsymbol{0}.
\end{aligned}$$

Proof (Cont'd): It follows that

$$\operatorname{var}(\check{\boldsymbol{\beta}}) = \operatorname{var}(\hat{\boldsymbol{\beta}}) + \operatorname{var}(\boldsymbol{C}\boldsymbol{y}) = \operatorname{var}(\hat{\boldsymbol{\beta}}) + \sigma_o^2 \boldsymbol{C}\boldsymbol{C}';$$

that is, $\operatorname{var}(\check{\boldsymbol{\beta}}) - \operatorname{var}(\hat{\boldsymbol{\beta}}) = \sigma_o^2 \boldsymbol{C}\boldsymbol{C}'$, a p.s.d. matrix (Verify!). This shows that $\hat{\boldsymbol{\beta}}$ must be more efficient than any linear unbiased estimator $\check{\boldsymbol{\beta}}$. □

Note that the estimator $\hat{\sigma}^2$ can be expressed as:

$$\hat{\sigma}^2 = \frac{1}{n-k-1} \sum_{i=1}^{n} \hat{u}_i^2 = \frac{\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}}{n-k-1}.$$

It can be shown that, under Classical Assumption II,

$$\mathbb{E}(\hat{\boldsymbol{u}}'\hat{\boldsymbol{u}}) = \sigma_o^2(n-k-1).$$

Hence, $\hat{\sigma}^2$ is unbiased for σ_o^2.

Inclusion of Irrelevant Variables

For a specification that includes irrelevant variables, we will show the OLS estimators remain unbiased but are not the most efficient.

Suppose we estimate the specification A below:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u, \quad (A)$$

and obtain the OLS estimators $\tilde{\beta}_j$, $j = 0, 1, 2, 3$. Assume that Classical Assumption II(i) holds with

$$\mathbb{E}(y) = b_0 + b_1 x_1 + b_2 x_2 = b_0 + b_1 x_1 + b_2 x_2 + 0 \cdot x_3;$$

this suggests that the specification A includes an irrelevant regressor x_3. As Classical Assumption II(i) still holds for b_0, b_1, b_2, and 0, we have $\mathbb{E}(\tilde{\beta}_j) = b_j$, $j = 0, 1, 2$, and $\mathbb{E}(\tilde{\beta}_3) = 0$.

To examine efficiency of $\tilde{\beta}_j$, note that

$$\text{var}(\tilde{\beta}_1) = \sigma_o^2 \frac{1}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 [1 - R_1^2(A)]},$$

where $R_1^2(A)$ is R^2 of regressing x_1 on 1, x_2 and x_3. Suppose we estimate the specification B instead:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \quad (B)$$

the resulting OLS estimators $\hat{\beta}_j$, $j = 0, 1, 2$, are such that $\mathbb{E}(\hat{\beta}_j) = b_j$, and

$$\text{var}(\hat{\beta}_1) = \sigma_o^2 \frac{1}{\sum_{i=1}^n (x_{i1} - \bar{x}_1)^2 [1 - R_1^2(B)]},$$

where $R_1^2(B)$ is R^2 of regressing x_1 on 1 and x_2. Clearly, $R_1^2(B) \leq R_1^2(A)$ (Why?), and hence $\text{var}(\hat{\beta}_1) \leq \text{var}(\tilde{\beta}_1)$. That is, $\tilde{\beta}_1$ is less efficient than $\hat{\beta}_1$. Similarly, $\tilde{\beta}_2$ is less efficient than $\hat{\beta}_2$.

General Case

Suppose we estimate the specification

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u} = \boldsymbol{X}_1\boldsymbol{\beta}_1 + \boldsymbol{X}_2\boldsymbol{\beta}_2 + \boldsymbol{u},$$

where $\boldsymbol{X} = [\boldsymbol{X}_1 \ \boldsymbol{X}_2]$ with $\boldsymbol{X}_1$ an $n \times k_1$ matrix and $\boldsymbol{X}_2$ an $n \times k_2$ matrix, and the OLS estimator is $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1' \ \tilde{\boldsymbol{\beta}}_2')'$. When the mean function of $\boldsymbol{y}$ is

$$\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}_1\boldsymbol{b}_1 + \boldsymbol{X}_2\mathbf{0} = \boldsymbol{X}\boldsymbol{b}_0, \quad \boldsymbol{b}_o = (\boldsymbol{b}_1' \ \mathbf{0}')',$$

our specification in fact includes k_2 irrelevant regressors $\boldsymbol{X}_2$. Then,

$$\mathbb{E}(\tilde{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathbb{E}(\boldsymbol{y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{b}_o = (\boldsymbol{b}_1' \ \mathbf{0}')',$$

showing that $\tilde{\boldsymbol{\beta}}_1$ is unbiased for $\boldsymbol{b}_1$ and $\tilde{\boldsymbol{\beta}}_2$ is unbiased for $\mathbf{0}$.

Recall that $\tilde{\beta}_1 = [\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}[\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{y}]$ by the Frisch-Waugh-Lovell Theorem. It is easy to see that, given $\mathrm{var}(\boldsymbol{y}) = \sigma_o^2\boldsymbol{I}$,

$$\begin{aligned}
&\mathrm{var}(\tilde{\beta}_1) \\
&= [\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\,\mathrm{var}(\boldsymbol{y})(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1[\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1} \\
&= \sigma_o^2[\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1[\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1} \\
&= \sigma_o^2[\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1}.
\end{aligned}$$

If we estimate $\boldsymbol{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{u}$ without irrelevant regressors $\boldsymbol{X}_2$, the OLS estimator $\hat{\beta}_1 = (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1'\boldsymbol{y}$ has variance: $\sigma_o^2(\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$. As

$$\boldsymbol{X}_1'\boldsymbol{X}_1 - \boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1 = \boldsymbol{X}_1'\boldsymbol{P}_2\boldsymbol{X}_1,$$

which is a p.s.d. matrix (Why?),

$$[\boldsymbol{X}_1'(\boldsymbol{I} - \boldsymbol{P}_2)\boldsymbol{X}_1]^{-1} - (\boldsymbol{X}_1'\boldsymbol{X}_1)^{-1}$$

is a p.s.d. matrix. This shows that $\tilde{\beta}_1$ is less efficient than $\hat{\beta}_1$.

Exclusion of Important Variables

For a specification that excludes important variables, the OLS estimators become biased in general. Suppose that we estimate the specification: $y = \beta_0 + \beta_1 x_1 + u$, and obtain the OLS estimators $\tilde{\beta}_0$ and $\tilde{\beta}_1$. Assume $\mathbb{E}(y) = b_0 + b_1 x_1 + b_2 x_2$; this suggests that our specification excludes an important variable x_2. Then,

$$
\begin{aligned}
\mathbb{E}(\tilde{\beta}_1) &= \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)\mathbb{E}(y_i)}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} \\
&= \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(b_0 + b_1 x_{i1} + b_2 x_{i2})}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} \\
&= b_1 + b_2 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2}}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}.
\end{aligned}
$$

Thus, $\tilde{\beta}_1$ is no longer unbiased for b_1, unless $\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2} = 0$, i.e., the sample covariance of x_{i1} and x_{i2} is zero.

General Case

Suppose we estimate the specification $\boldsymbol{y} = \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{u}$ and obtain the OLS estimator $\tilde{\boldsymbol{\beta}}_1$. If

$$\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}_1 \boldsymbol{b}_1 + \boldsymbol{X}_2 \boldsymbol{b}_2, \quad \boldsymbol{b}_2 \neq \boldsymbol{0},$$

so that our specification excludes relevant regressors $\boldsymbol{X}_2$, we then have

$$\begin{aligned} \mathbb{E}(\tilde{\boldsymbol{\beta}}_1) &= (\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1' \mathbb{E}(\boldsymbol{y}) \\ &= (\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1' (\boldsymbol{X}_1 \boldsymbol{b}_1 + \boldsymbol{X}_2 \boldsymbol{b}_2) \\ &= \boldsymbol{b}_1 + (\boldsymbol{X}_1' \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1' \boldsymbol{X}_2 \boldsymbol{b}_2. \end{aligned}$$

That is, the OLS estimator $\tilde{\boldsymbol{\beta}}_1$ is biased for $\boldsymbol{b}_1$, unless $\boldsymbol{X}_1' \boldsymbol{X}_2 = \boldsymbol{0}$, i.e., every column of $\boldsymbol{X}_1$ is orthogonal to the columns of $\boldsymbol{X}_2$.