# Lecture 2
# Multiple Linear Regression: Estimation

*CHUNG-MING KUAN*

*Department of Finance & CRETA*

*National Taiwan University*

February 21, 2020

# Lecture Outline

1. Mulple Linear Regression: Estimation
   - Least-Squares 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 behavior of the dependent variable $y$ may be better characterized by a linear function of $k$ ($k > 1$) explanatory variables (regressors) such that

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

where $\beta_0$, $\beta_1, \ldots, \beta_k$ are unknown parameters, and the error term summarizes the non-systematic part of $y$ and varies with the parameter values. 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.

# Least-Squares Minimization

To find the hyperplane that "best" fits the sample data $(x_{i1}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots, n$, we minimize the LS criterion function:

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

and solve for $k+1$ unknown parameters $\beta_0, \beta_1, \ldots, \beta_k$ from the FOCs:

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

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

$$\vdots$$

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

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

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

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

then imply that the FOC: $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik}) x_{i3} = 0$ must also hold and hence is redundant. As such, the number of effective FOCs is less than $k+1$, and the OLS estimators cannot be uniquely solved from the FOCs.

Given the OLS estimators $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$, 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

For the OLS estimator $\hat{\beta}_1$ we shall use the following analytic formula (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 formula is also the OLS estimator of regressing $y$ on $\hat{r}_1$ (without the constant term) and hence the marginal effect of $\hat{r}_1$ on $y$.
- By definition, $\hat{r}_1$ is part of $x_1$ that is uncorrelated with $x_2, \ldots, x_k$. Hence, $\hat{\beta}_1$ can be understood as the "pure" effect of $x_1$ on $y$, after the effects of $x_2, \ldots, x_k$ on $x_1$ have been "partialled out" or "purged away".

From the formula of $\hat{\beta}_1$ we can see 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 $x_{i,1} - \bar{x}_1 = \hat{r}_{i,1}$. When $x_2, \ldots, x_k$ are not linearly related to $x_1$, regressing $x_1$ on the constant one and $x_2, \ldots, x_k$ yield:

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

so that $\hat{\beta}_1 = \hat{b}_1$. On the other hand, when $x_2, \ldots, x_k$ are linearly related to $x_1$, $\hat{\beta}_1 \neq \hat{b}_1$. In this case, $\hat{b}_1$ is the marginal effect of $x_1$ on $y$ without controlling other regressors and hence must involve both the "pure" effect ($\hat{\beta}_1$) of $x_1$ on $y$ as well as 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, $\hat{\beta}_j \neq \hat{b}_j$, the OLS estimator of regressing $y$ on $x_j$ without controlling other regressors. 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 seen from simple linear gression that SST = SSR + SSE, i.e.,

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

For multiple regressions, a measure of goodness of fit is the coefficient of determination:

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

which measures the proportion of the total variation (SST) of $y_i$ due to the variation of $\hat{y}_i$ (SSE). As in simple linear regression, $0 \leq R^2 \leq 1$. A specification has a better (worse) fit of the 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:

$$y_i = \hat{b}_0 + \hat{b}_1 x_{i1} + \hat{b}_2 x_{i2} + \hat{v}_i,$$
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{3i} + \hat{u}_i.$$

Note that the former can be written as:

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

Clearly, the estimates $\hat{b}_0, \hat{b}_1, \hat{b}_2, 0$ would not minimize the error sum of squares in the 3-regressor regression, because the last coefficient is restricted to zero. This suggests 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 additional regressor(s).

# 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 $\bar{R}^2$, which is $R^2$ adjusted for the degrees of freedom:

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

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

## Classical Assumption II

The random variables $y_i$, $i = 1, \ldots, n$, follow the population model:

$$y_i = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik} + u_i,$$

for some numbers $b_0, b_1, \ldots, b_k$ (parameters of interest), where (i) $x_{i1}, \ldots, x_{ik}$ are non-random, (ii) $\mathbb{E}(y_i) = b_0 + b_1 x_{i1} + \cdots + b_k x_{ik}$, and (iii) $\text{var}(y_i) = \sigma_o^2$, $\text{cov}(y_i, y_j) = 0$ for $i \neq j$.

## Unbiasedness of the OLS Estimators

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

# Some Algebra

Recall that $\hat{r}_{i,1}$ are the OLS residuals of regressing $x_1$ on the constant one and $x_2, \ldots, x_k$ and hence are non-random by Classical Assumption II(i). Using the formula for "partialling out" argument, we have

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

where the second equality follows from Classical Assumption II(ii). By the FOCs of OLS estimation, we have $\sum_{i=1}^n \hat{r}_{i,1} = 0$, $\sum_{i=1}^n \hat{r}_{i,1} x_{i2} = 0, \ldots,$ $\sum_{i=1}^n \hat{r}_{i,1} x_{ik} = 0$. Consequently,

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

because $\sum_{i=1}^n \hat{r}_{i,1} x_{i1} = \sum_{i=1}^n \hat{r}_{i,1}^2$ (Verify!). 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), (ii) and (iii),

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

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 estimates 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.

# More Algebra

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

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

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

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

This proves the formula for $\text{var}(\hat{\beta}_1)$.

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

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

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

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

Replacing $\sigma_o^2$ 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

Again consider the OLS formula:

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

Thus, $\hat{\beta}_j$ is in effect a weighted sum of $y_i$ and hence an estimator linear in $y_i$: $\sum_{i=1}^n a_{i,j} y_i$, with $a_{i,j} = \hat{r}_{i,j} / \sum_{i=1}^n \hat{r}_{i,j}^2$. 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 the most efficient. A proof will be given later using matrix notations.

## Gauss-Markov Theorem

Under Classical Assumption II(i), (ii) and (iii), the OLS estimators $\hat{\beta}_j$ are best linear unbiased 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:

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

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 $\text{exper}^2$, the estimated parameters are:

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

- The new regressor $\text{exper}^2$ 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 $\text{exper} = 36.5$. Thus, log(wage) increases with a decreasing rate $(-0.001)$ before experience reaches 36.5 years.

# LS Estimation in Matrix Notations

The specification is: $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{u}(\beta)$, 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 $\boldsymbol{u}(\beta) = (u_1(\beta)\ u_2(\beta) \ldots u_n(\beta))'$. The LS criterion function is

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

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

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

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 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, we obtain the OLS estimator of $\beta$:

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

**Remarks**:

- The inverse $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists provided that $\boldsymbol{X}$ 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:

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

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

These are exactly the algebraic properties we observed earlier.

# Some Matrix Results

- Two $n \times 1$ vectors, $\boldsymbol{x}$ and $\boldsymbol{z}$, are said to be orthogonal if $\boldsymbol{x}'\boldsymbol{z} = 0$. We also call the product $\boldsymbol{x}'\boldsymbol{z}$ an inner product.
- Let $\boldsymbol{x}$ be an $n \times 1$ vector. Then, $\boldsymbol{x}'\boldsymbol{x} = \sum_{i=1}^{n} x_i^2$, so that the Euclidean norm of $\boldsymbol{x}$ is $\|\boldsymbol{x}\| = (\boldsymbol{x}'\boldsymbol{x})^{1/2}$.
- A matrix $\boldsymbol{A}$ is said to be a projection matrix if it is idempotent ($\boldsymbol{AA} = \boldsymbol{A}$). Writing $\boldsymbol{x} = \boldsymbol{Ax} + (\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x}$, the projection $\boldsymbol{Ax}$ would be orthogonal if $\boldsymbol{x}'\boldsymbol{A}'(\boldsymbol{I} - \boldsymbol{A})\boldsymbol{x} = 0$. When $\boldsymbol{A}$ is symmetric ($\boldsymbol{A} = \boldsymbol{A}'$), we have $\boldsymbol{x}'(\boldsymbol{A} - \boldsymbol{AA})\boldsymbol{x} = \boldsymbol{x}'(\boldsymbol{A} - \boldsymbol{A})\boldsymbol{x}$ which is zero. Thus, $\boldsymbol{A}$ is an orthogonal projection matrix if it is idempotent and symmetric. 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 the orthogonal projection matrix that projects vectors onto the space spanned by the column vectors of $\boldsymbol{X}$, $\text{span}(\boldsymbol{X})$. Similarly, $\boldsymbol{I} - \boldsymbol{P}$ is the orthogonal projection matrix that projects vectors onto the orthogonal complement of $\text{span}(\boldsymbol{X})$, $\text{span}(\boldsymbol{X})^{\perp}$. Thus, $\boldsymbol{PX} = \boldsymbol{X}$, and $(\boldsymbol{I} - \boldsymbol{P})\boldsymbol{X} = \boldsymbol{0}$.

- The vector of fitted values is

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

  which is the orthogonal projection of $\boldsymbol{y}$ onto $\text{span}(\boldsymbol{X})$.

- The residual vector is the orthogonal projection of $\boldsymbol{y}$ onto $\text{span}(\boldsymbol{X})^{\perp}$:

$$\hat{\boldsymbol{u}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = (\boldsymbol{I} - \boldsymbol{P})\boldsymbol{y}.$$

- The orthogonal projection $\boldsymbol{P}\boldsymbol{y}$ provides the "best approximation" to $\boldsymbol{y}$, in the sense that the Euclidean norm of $\boldsymbol{y} - \boldsymbol{P}\boldsymbol{y} = \hat{\boldsymbol{u}}$, $\|\hat{\boldsymbol{u}}\|$, is the smallest possible, compared with $\|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{y}\|$, where $\boldsymbol{A}\boldsymbol{y}$ is any other projection of $\boldsymbol{y}$. This is precisely what the LS optimization 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 $\boldsymbol{X}$.
- The algebraic property $\hat{\boldsymbol{y}}'\hat{\boldsymbol{u}} = 0$ holds because $\hat{\boldsymbol{u}}$ must be orthogonal to $\hat{\boldsymbol{y}}$ which is the orthogonal projection of $\boldsymbol{y}$ onto $\text{span}(\boldsymbol{X})$.

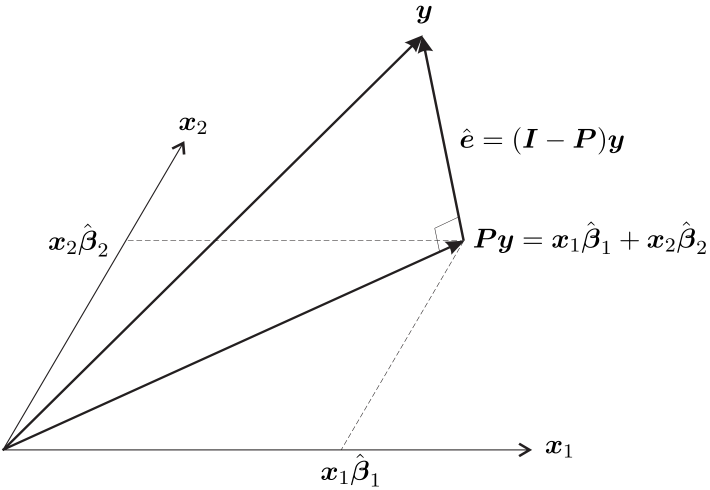


Figure: The orthogonal projection of $\boldsymbol{y}$ onto span$(\boldsymbol{x}_1, \boldsymbol{x}_2)$.

# An Example

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

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

and

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

Using these results it is readily verified that $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$ leads to the OLS estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ obtained in the simple linear regression.

## Statistical Properties

### Classical Assumption II in Matrix Notations

The random vector $\boldsymbol{y}$ follows the population model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{b}_o + \boldsymbol{u}$ for some parameter vector $\boldsymbol{b}_o$, where (i) $\boldsymbol{X}$ is non-random, (ii) $\mathbb{E}(\boldsymbol{y}) = \boldsymbol{X}\boldsymbol{b}_o$, and (iii) $\operatorname{var}(\boldsymbol{y}) = \sigma_o^2 \boldsymbol{I}$.

- Unbiasedness: $\mathbb{E}(\hat{\beta}) = \mathbb{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}]$. By Classical Assumption II(i) and II(ii),

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

- Variance: $\operatorname{var}(\hat{\beta}) = \operatorname{var}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}]$, and by Classical Assumption II(i) and II(iii),

$$\operatorname{var}(\hat{\beta}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'[\operatorname{var}(\boldsymbol{y})]\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma_o^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

## Gauss-Markov Theorem

Under Classical Assumption II(i), (ii) and (iii), the OLS estimators $\hat{\beta}$ is best linear unbiased 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

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

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 $\boldsymbol{C}\boldsymbol{y} = \boldsymbol{C}(\boldsymbol{X}\boldsymbol{b}_o + \boldsymbol{u}) = \boldsymbol{C}\boldsymbol{u}$ and hence

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

**Proof** (Cont'd): It follows that

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

that is, $\text{var}(\check{\boldsymbol{\beta}}) - \text{var}(\hat{\boldsymbol{\beta}}) = \sigma_o^2 \boldsymbol{C}\boldsymbol{C}'$, 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:

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

# Inclusion of Irrelevant Variables

For a specification that includes irrelevant variables, we will show the OLS estimators remain unbiased but are less efficient. Suppose that Classical Assumption II holds with $\mathbb{E}(y) = b_0 + b_1 x_1 + b_2 x_2$. We estimate the specification A below to obtain the OLS estimators $\hat{\beta}_j$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u,$$

which includes the irrelevant variable $x_3$. As $\mathbb{E}(y)$ can also be expressed as $b_0 + b_1 x_1 + b_2 x_2 + 0 \cdot x_3$, Classical Assumption II(ii) holds for $b_0, b_1, b_2, 0$. It follows that $\mathbb{E}(\hat{\beta}_j) = b_j$, $j = 0, 1, 2$, and $\mathbb{E}(\hat{\beta}_3) = 0$, proving their unbiasedness.

To see $\hat{\beta}_j$ are less efficient, note that

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

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:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u,$$

and obtain the OLS estimators $\tilde{\beta}_j$. Then, $\mathbb{E}(\tilde{\beta}_j) = b_j$, $j = 0, 1, 2$, and

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

where $R_1^2(B)$ is $R^2$ of regressing $x_1$ on 1 and $x_2$. Noting $R_1^2(A) \geq R_1^2(B)$ (Why?), we have $\mathrm{var}(\tilde{\beta}_1) \leq \mathrm{var}(\hat{\beta}_1)$. Similarly, $\mathrm{var}(\tilde{\beta}_2) \leq \mathrm{var}(\hat{\beta}_2)$.

# Exclusion of Important Variables

For a specification that excludes important variables, the OLS estimators become biased. Suppose that Classical Assumption II holds with $\mathbb{E}(y) = b_0 + b_1 x_1 + b_2 x_2$, but we excludes the variable $x_2$ and estimate the simple specification: $y = \beta_0 + \beta_1 x_1 + u$, and obtain the OLS estimators $\hat{\beta}_j$. It is then easy to see that

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

Hence, $\hat{\beta}_1$ is biased for $b_1$, unless $\sum_{i=1}^n (x_{i1} - \bar{x}_1) x_{i2} = 0$.

**Remark**: When the sample covariance of $x_{i1}$ and $x_{i2}$ is zero, the second term on the right-hand side is zero, so that $\hat{\beta}_1$ remains unbiased for $b_1$.