Applied Econometrics

Multiple Regression Analysis

Part 1: Estimation

Text: Wooldridge, Chapter 3

May 24, 2011

The Multiple Regression Model

- We have seen how to use the simple linear model to analyze the variations of a dependent variable y in response to changes of an independent variable x.

- A major drawback of the simple linear model is that it is difficult to identify ceteris paribus effects of x on y, given that the assumption of uncorrelatedness between x and the other factors affecting y is often unrealistic.

- The **multiple regression model** is much better suited for such analysis, since it allows to explicitly control for other factors that simultaneously affect the dependent variable y.

The Multiple Regression Model

- Consider the *wage equation* example, which may be expanded to include *labor market experience* (*exper*),

$$wage = \beta_0 + \beta_1 edu + \beta_2 exper + u, \tag{1}$$

  where edu is years of education.

- Assume we are still primarily interested in the effect of *edu* on *wage*, holding fixed all the other factors affecting *wage*.

- Just as in the simple linear model, we will need several assumptions about the relationship (or lack thereof) between u and the variables *edu* and *exper*.

- However, and in contrast to the simple regression framework, we do know that we will be able to measure the impact of education on wage while holding experience fixed.

The Multiple Regression Model

- The multiple regression model with k explanatory variables $x_1, x_2, \ldots, x_k$ is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + u. \tag{2}$$

- The terminology is rather similar to the simple linear model:

- y is the **dependent** or **explained** or **response** variable, or the **regressand**, and

- $x_1, x_2, \ldots, x_k$ are the **independent** or **explanatory** or **control** variables, or the **regressors**.

- u is the **error term** or **disturbance**, capturing factors other than $x_1, \ldots, x_k$ that affect y.

- As before, β_0 is the **intercept parameter**, and the $\beta_1, \beta_2, \ldots, \beta_k$ will be referred to as the **slope parameters**.

The Ordinary Least Squares (OLS) Estimator

- Assume that we observe a sample of size n,

$$\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, \ldots, n\}, \tag{3}$$

  from the population generated by our multiple linear regression model.

- That is, we have n observations for each of our k explanatory variables and the explained variable.

- x_{ij} denotes the ith observation of variable j, i.e.,

  – i is the observation number, and
  – j is the variable number.

- We can write

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + u_i, \quad i = 1, \ldots, n. \tag{4}$$

- Our estimators of $\beta_0, \beta_1, \ldots, \beta_k$ will be denoted by $\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k$, and we may define the fitted value for observation i as

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \cdots + \widehat{\beta}_k x_{ik}, \tag{5}$$

and the **regression residuals**

$$\widehat{u}_i = y_i - \widehat{y}_i = y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik}, \quad i = 1, \ldots, n. \tag{6}$$

- As for the simple linear model, where $k = 1$, the OLS estimator chooses the estimates so that $\widehat{\beta}_0$ the sum of squared regression residuals is minimized, i.e., we want to minimize

$$\begin{aligned} S(\widehat{\beta}_0, \widehat{\beta}_1, \ldots, \widehat{\beta}_k) &= \sum_{i=1}^{n} \widehat{u}_i^2 \\ &= \sum_{i=1}^{n} (y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik})^2. \end{aligned}$$

- The first–order conditions are

$$\begin{aligned}
\frac{\partial S}{\partial \widehat{\beta}_0} &= -2\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik}) = 0 \\
\frac{\partial S}{\partial \widehat{\beta}_1} &= -2\sum_{i=1}^{n} x_{i1}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik}) = 0 \\
\frac{\partial S}{\partial \widehat{\beta}_2} &= -2\sum_{i=1}^{n} x_{i2}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik}) = 0 \qquad (7) \\
&\vdots \\
\frac{\partial S}{\partial \widehat{\beta}_k} &= -2\sum_{i=1}^{n} x_{ik}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_k x_{ik}) = 0.
\end{aligned}$$

- The first–order conditions (7) imply that

$$\sum_i \widehat{u}_i = 0, \quad \sum_i x_{i1}\widehat{u}_i = 0, \quad \ldots, \quad \sum_i x_{ik}\widehat{u}_i = 0, \qquad (8)$$

which resembles our result from the simple regression model.

- Several aspects of the analysis can be considerably simplified by writing the multiple regression model in matrix form.

- Let us first consider observation i.

- By defining the vectors (both of dimension $k+1$)

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \quad \text{and} \quad \boldsymbol{x}_i = \begin{bmatrix} 1 & x_{i1} & \cdots & x_{ik} \end{bmatrix}, \tag{9}$$

  we can write

$$y_i = \boldsymbol{x}_i\boldsymbol{\beta} + u_i, \quad i = 1, \ldots, n. \tag{10}$$

- Next, we define the data matrix of dimension $n \times (k+1)$,

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}. \tag{11}$$

- The ith row of this matrix contains the explanatory variables (including the constant) for observation i.

- The jth column of this matrix contains the i observations of the jth explanatory variable, $x_{1j}, x_{2j}, \ldots, x_{nj}$.

- Then we can write

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}}_{=\boldsymbol{y}} = \underbrace{\begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{bmatrix}}_{=\boldsymbol{X}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}}_{=\boldsymbol{\beta}} + \underbrace{\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix}}_{=\boldsymbol{u}}, \tag{12}$$

i.e.,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}. \tag{13}$$

- Likewise, we can define

$$\widehat{\boldsymbol{y}} = \begin{bmatrix} \widehat{y}_1 \\ \widehat{y}_2 \\ \vdots \\ \widehat{y}_n \end{bmatrix}, \quad \widehat{\boldsymbol{u}} = \begin{bmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ \vdots \\ \widehat{u}_n \end{bmatrix}, \quad \widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix}. \tag{14}$$

- Inspection of the conditions (8),

$$\sum_i \widehat{u}_i = 0, \quad \sum_i x_{i1}\widehat{u}_i = 0, \quad \ldots, \quad \sum_i x_{ik}\widehat{u}_i = 0,$$

shows that these may be written

$$\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & \cdots & x_{n1} \\ \vdots & \vdots & \vdots & & \vdots \\ x_{1k} & x_{2k} & x_{3k} & \cdots & x_{nk} \end{bmatrix} \begin{bmatrix} \widehat{u}_1 \\ \widehat{u}_2 \\ \vdots \\ \widehat{u}_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \tag{15}$$

that is,

$$\boldsymbol{X}'\widehat{\boldsymbol{u}} = \boldsymbol{0}, \tag{16}$$

where the the prime means "transposed".

- We now use condition (16),

$$\boldsymbol{X}'\widehat{\boldsymbol{u}} = \boldsymbol{0}, \tag{17}$$

to find the ordinary least squares estimator.

- We have

$$\boldsymbol{y} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} + \widehat{\boldsymbol{u}}. \tag{18}$$

- Now multiply both sides of the equation by $\boldsymbol{X}'$ as given in (15),

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} + \underbrace{\boldsymbol{X}'\widehat{\boldsymbol{u}}}_{=\boldsymbol{0}}, \tag{19}$$

so (the *normal equations*)

$$\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}, \tag{20}$$

or

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}, \tag{21}$$

provided $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists, as we shall assume.

- Note that $\boldsymbol{X}'\boldsymbol{X}$ is the second moment matrix of the independent variables, and the normal equations (20), written in extensive form, are

$$\underbrace{\begin{bmatrix} n & \sum_i x_{i1} & \sum_i x_{i2} & \cdots & \sum_i x_{ik} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} & \cdots & \sum_i x_{i1}x_{ik} \\ \sum_i x_{i2} & \sum_i x_{i1}x_{i2} & \sum_i x_{i2}^2 & \cdots & \sum_i x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \sum_i x_{ik} & \sum_i x_{ik}x_{i1} & \sum_i x_{ik}x_{i2} & \cdots & \sum_i x_{ik}^2 \end{bmatrix}}_{=\boldsymbol{X}'\boldsymbol{X}} \begin{bmatrix} \widehat{\beta}_0 \\ \widehat{\beta}_1 \\ \widehat{\beta}_2 \\ \vdots \\ \widehat{\beta}_k \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} \sum_i y_i \\ \sum_i x_{i1}y_i \\ \sum_i x_{i2}y_i \\ \vdots \\ \sum_i x_{ik}y_i \end{bmatrix}}_{=\boldsymbol{X}'\boldsymbol{y}}.$$

- The condition that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists requires that the explanatory variables are linearly independent, and is a generalization of our earlier condition $s_x^2 > 0$ in the simple linear model (since if there is no variation in one of the explanatory variables, then there is linear dependence between this "variable" and the constant in the regression equation).

- An obvious condition for $\boldsymbol{X}'\boldsymbol{X}$ to be nonsingular is that $n \geq k+1$ (usually n should be considerably larger than k).

- Also note from the first–order condition for $\widehat{\beta}_0$ that the OLS estimator satisfies

$$\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{k} \widehat{\beta}_j \overline{x}_j, \tag{22}$$

  where

$$\overline{x}_j = \frac{1}{n} \sum_{i=1}^{n} x_{ij}. \tag{23}$$

- This implies, for example, that the point $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k, \overline{y})$ is always on the regression line.

Interpretation of the OLS Regression Equation

- The OLS regression line is

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \cdots + \widehat{\beta}_k x_k. \tag{24}$$

- Just as in the simple linear model, the intercept parameter $\widehat{\beta}_0$ may or may not have an interesting interpretation, depending on whether setting all the explanatory variables equal to zero is a relevant scenario.

- The estimates $\widehat{\beta}_j$, $j = 1, \ldots, k$, have a **partial effect**, or **ceteris paribus**, interpretation.

- From (24), with Δ denoting "change",

$$\Delta\widehat{y} = \widehat{\beta}_1 \Delta x_1 + \widehat{\beta}_2 \Delta x_2 + \cdots + \widehat{\beta}_k \Delta x_k, \tag{25}$$

  we can obtain the (predicted) change in y in response to changes in x_j, $j = 1, \ldots, k$.

- When all independent variables except x_1 are held fixed, so that $\Delta x_2 = \Delta x_3 = \cdots = \Delta x_k = 0$, then

$$\Delta \widehat{y} = \widehat{\beta}_1 \Delta x_1. \tag{26}$$

  In this sense, when using multiple regression, we can control for the variables $x_2, \ldots, x_k$ when measuring the effect of x_1 on y.

- Clearly the other coefficients of the equation have an analogous interpretation.

More flexible functional forms

- For example, with multiple regression, we may model y as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + u. \tag{27}$$

- Clearly the coefficients β_1 and β_2, separately, do not have the interpretation of slope coefficients.

- Rather, to find the effect of x_1 we have to calculate

$$\frac{\partial y}{\partial x_1} = \beta_1 + 2\beta_2 x_1. \tag{28}$$

Coefficient of Determination (R^2)

- We have the same relation as in the simple linear model, i.e.,

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{=\text{SST}}=\underbrace{\sum_{i=1}^{n}\widehat{u}_i^2}_{=\text{SSR}}+\underbrace{\sum_{i=1}^{n}(\widehat{y}_i-\bar{y})^2}_{=\text{SSE}}$$

  where

- **SST** means **total sum of squares**, measuring the total sample variation in the y_is,

- **SSE** means **explained sum of squares**, i.e., the part of variation in the y_is that is explained by the fitted regression line,

- **SSR** means **residual sum of squares**, i.e., the part of the variation that is not explained by the fitted line.

- The coefficient of determination, R^2,

$$R^2 = \frac{\text{SSE}}{\text{SST}} = 1 - \frac{\text{SSR}}{\text{SST}}, \tag{29}$$

  can be interpreted as the *fraction of the sample variation in y that is explained by x* (via the fitted linear regression line).

- Clearly

$$0 \leq R^2 \leq 1.$$

- The fraction–of–variance interpretation will hold only if there is a constant in the regression, since otherwise $\sum_i \widehat{u}_i \neq 0$ generally.

- R^2 cannot be used for selecting the regressors to be included into the model since it never decreases (and usually increases) with the inclusion of an additional variable.

- For example, if we have n linearly independent regressors (including the constant), then $R^2 = 1$.

- Economically, the relevant question is whether an explanatory variable has an economically and statistically significant partial effect on y.

- Thus, R^2 will not play an important role in building econometric models if such questions are of primary interest.

- Measures of fit can be more important when forecasting is the main goal of modeling, however (e.g., the *adjusted* R^2).

Statistical Properties of OLS: Assumptions (Gauß–Markov Assumptions)

- The assumptions required to derive the properties of the ordinary least squares estimator are straightforward extensions of those of the simple linear model.

- In particular, the following assumptions will be made:

  1) The linear model is correctly specified, i.e., y is related to the independent variables and u as $y = \beta_0 + \beta_1 x + \cdots + \beta_k x_k + u$. (**Linearity in Parameters**)
  2) We observe a **random sample** of size n, $\{(x_{i1}, x_{i2}, \ldots, x_{ik}, y_i) : i = 1, \ldots, n\}$, generated from the linear model.
  3) The independent variables are not linearly dependent, i.e., $\det(\boldsymbol{X}'\boldsymbol{X}) > 0$, so that $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ exists. (**No Perfect Collinearity**)
  4) $\mathrm{E}(u|x_1, x_2, \ldots, x_k) = 0$. (**Zero Conditional mean**)
  5) $\mathrm{E}(u^2|x) = \mathrm{Var}(y|x_1, x_2, \ldots, x_k) = \sigma^2$, i.e., the variance of u (and hence y) does not depend on x. (**Homoskedasticity**)

- Note that the assumption of random sampling implies

$$\mathsf{E}(u_i u_j) = \mathsf{Cov}(u_i, u_j) = 0, \quad i \neq j. \tag{30}$$

- We can also combine the assumption of random sampling with that of homoskedasticity to obtain a simple and useful expression for the covariance matrix of the vector $\boldsymbol{u}$. Namely. as all the variances are equal to σ^2, and all covariances are zero, we have

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{u}) &= \mathsf{E}(\boldsymbol{u}\boldsymbol{u}') = \mathsf{E}\left(\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \begin{bmatrix} u_1 & u_2 & \cdots & u_n \end{bmatrix} \right) && (31) \\
&= \begin{bmatrix} \mathsf{E}(u_1^2) & \mathsf{E}(u_1 u_2) & \cdots & \mathsf{E}(u_1 u_n) \\ \mathsf{E}(u_1, u_2) & \mathsf{E}(u_2^2) & \cdots & \mathsf{E}(u_2 u_n) \\ \vdots & \vdots & \cdots & \vdots \\ \mathsf{E}(u_1 u_n) & \mathsf{E}(u_2 u_n) & \cdots & \mathsf{E}(u_n^2) \end{bmatrix} && (32) \\
&= \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} = \sigma^2 \boldsymbol{I}_n, && (33)
\end{aligned}$$

where $\boldsymbol{I}_n$ is the identity matrix of dimension n, i.e., an $n \times n$ matrix with ones on the diagonal and zeros elsewhere.

Statistical Properties: Unbiasedness

- We have seen that the OLS estimator is unbiased in the simple linear model.

- The same can be shown in the multiple regression model, as follows:

- We write our expression for $\widehat{\boldsymbol{\beta}}$ and, assuming our model is correctly specified (Assumption 1), substitute for $\boldsymbol{y}$,

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\underbrace{\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{u}}_{=\boldsymbol{y}}) \\
&= \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\boldsymbol{I}_k \text{ (identity matrix)}}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u} \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}. \qquad (34)
\end{aligned}$$

Taking expectations, using Assumption 4,

$$\begin{aligned}
\mathsf{E}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \boldsymbol{\beta} + \mathsf{E}[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}|\boldsymbol{X}] \\
&= \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\underbrace{\mathsf{E}(\boldsymbol{u}|\boldsymbol{X})}_{=\mathbf{0}} \\
&= \boldsymbol{\beta}.
\end{aligned}$$

- Thus, the OLS estimator is unbiased.
- Note that the assumption of homoskedasticity of the regression errors is not required to derive the unbiasedness of the OLS estimator.
- It is also apparent that $\widehat{\boldsymbol{\beta}}$ is a **linear estimator**.

The Omitted Variable Bias

- Suppose we omit a variable that actually affect y in the population, which is referred to as the probelm of **excluding a relevant variable**.

- Consider the simplest possible case, where the true model is

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u, \tag{35}$$

whereas we estimate the equation

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \tilde{u}. \tag{36}$$

Our OLS estimator of the misspecified model is

$$\widehat{\tilde{\beta}}_1 = \frac{\sum_{i=1}^{n}(x_{i1} - \overline{x}_1)y_i}{ns_{x_1}^2}. \tag{37}$$

- Substituting (35) into (37) gives

$$\begin{aligned}
\widehat{\widetilde{\beta}}_1 &= \frac{\sum_{i=1}^n (x_{i1}-\overline{x}_1)(\beta_0+\beta_1 x_{i1}+\beta_2 x_{i2}+u_i)}{n s_{x_1}^2} \\
&= \frac{\sum_{i=1}^n (x_{i1}-\overline{x}_1)\beta_0 + \beta_1 \sum_{i=1}^n (x_{i1}-\overline{x}_1)x_{i1}}{n s_{x_1}^2} \\
&\quad + \frac{\beta_2 \sum_{i=1}^n (x_{i1}-\overline{x}_1)x_{i2} + \sum_{i=1}^n (x_{i1}-\overline{x}_1)u_i}{n s_{x_1}^2} \\
&= 0 + \beta_1 + \beta_2 \frac{s_{x_1,x_2}}{s_{x_1}^2} + \frac{\sum_i (x_{i1}-\overline{x}_1)u_i}{n s_{x_1}^2}, \qquad (38)
\end{aligned}$$

where

$$s_{x_1,x_2} = \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)x_{i2} = \frac{1}{n}\sum_{i=1}^n (x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)$$

is the sample covariance between x_1 and x_2.

- Taking expectations in (38) gives

$$\mathsf{E}(\widehat{\tilde{\beta}}_1|\boldsymbol{X}) = \beta_1 + \beta_2 \frac{s_{x_1,x_2}}{s_{x_1}^2} \neq \beta_1, \tag{39}$$

provided that $\beta_2 \neq 0$ and $s_{x_1,x_2} \neq 0$. Note that

$$\frac{s_{x_1,x_2}}{s_{x_1}^2} \tag{40}$$

is the OLS slope coefficient $\widehat{\delta}_1$ of the regression

$$\widehat{x}_{i2} = \widehat{\delta}_0 + \widehat{\delta}_1 x_{i1}, \tag{41}$$

i.e., the slope coefficient of a regression of x_2 on x_1.

- It follows that the OLS estimator of the misspecified equation (37) is unbiased only if either

  – variable x_2 does not affect y, i.e., $\beta_2 = 0$, or
  – x_1 and x_2 are uncorrelated in the sample, i.e., $s_{x_1,x_2} = 0$.

- The bias is due to the fact that x_1 affects y *directly* via β_1 but also via $x_1 \overset{s_{x_1,x_2}}{\longrightarrow} x_2 \overset{\beta_2}{\longrightarrow} y$.

- Thus to obtain an unbiased estimator of β_1 the impact of x_2 must be controlled for.

- The bias in this case is

$$\mathsf{E}(\widehat{\tilde{\beta}}_1|\boldsymbol{X}) - \beta_1 = \widehat{\delta}_1\beta_2, \tag{42}$$

  and the following table provides information about the sign of the bias.

Table 1: Bias of $\widehat{\widetilde{\beta}}_1$

	$s_{x_1,x_2} > 0$	$s_{x_1,x_2} < 0$
$\beta_2 > 0$	positive bias	negative bias
$\beta_2 < 0$	negative bias	positive bias

- In more general settings, it is much more difficult to make precise statements about the sign and the magnitude of the bias, since the entire correlation structure of the variables in $\boldsymbol{X}$ determine the bias.

- For example, if in a true model with three variables,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u, \tag{43}$$

we omit x_3, and x_1 is correlated with x_3, whereas x_2 is not correlated with x_3, then the estimator of β_2 based on the wrong model (with x_3 left out) will still be biased unless x_1 and x_2 are likewise uncorrelated.

- To illustrate, assume that all variables have zero mean, so a model with zero intercept is appropriate, i.e.,

$$y = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + u. \tag{44}$$

- Write the true model as

$$\boldsymbol{y} = \boldsymbol{X}_{1:2}\boldsymbol{\beta}_{1:2} + \boldsymbol{x}_3\beta_3 + \boldsymbol{u}, \tag{45}$$

where

$$\boldsymbol{X}_{1:2} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ \vdots & \vdots \\ x_{n1} & x_{n2} \end{bmatrix}, \quad \boldsymbol{\beta}_{1:2} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}, \quad \boldsymbol{x}_3 = \begin{bmatrix} x_{13} \\ x_{23} \\ \vdots \\ x_{n3} \end{bmatrix}. \tag{46}$$

- The estimator of $\boldsymbol{\beta}_{1:2}$ which results from omitting x_3 is

$$\widehat{\widetilde{\boldsymbol{\beta}}}_{1:2} = (\boldsymbol{X}'_{1:2}\boldsymbol{X}_{1:2})^{-1}\boldsymbol{X}'_{1:2}\boldsymbol{y},$$

and, upon substituting (45), its expectation is

$$\begin{aligned}
\mathsf{E}(\widehat{\tilde{\boldsymbol{\beta}}}_{1:2}) &= \mathsf{E}[(\boldsymbol{X}'_{1:2}\boldsymbol{X}_{1:2})^{-1}\boldsymbol{X}'_{1:2}\boldsymbol{y}] \\
&= \mathsf{E}[(\boldsymbol{X}'_{1:2}\boldsymbol{X}_{1:2})^{-1}\boldsymbol{X}'_{1:2}(\boldsymbol{X}_{1:2}\boldsymbol{\beta}_{1:2} + \boldsymbol{x}_3\beta_3 + \boldsymbol{u})] \\
&= \boldsymbol{\beta}_{1:2} + (\boldsymbol{X}'_{1:2}\boldsymbol{X}_{1:2})^{-1}\boldsymbol{X}'_{1:2}\boldsymbol{x}_3\beta_3.
\end{aligned}$$

- The bias term $\mathsf{E}(\widehat{\tilde{\boldsymbol{\beta}}}_{1:2}) - \boldsymbol{\beta}_{1:2}$ is

$$\begin{aligned}
(\boldsymbol{X}'_{1:2}\boldsymbol{X}_{1:2})^{-1}\boldsymbol{X}'_{1:2}\boldsymbol{x}_3\beta_3 &= \begin{bmatrix} \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} \\ \sum_i x_{i1}x_{i2} & \sum_i x_{i2}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i x_{i1}x_{i3} \\ \sum_i x_{i2}x_{i3} \end{bmatrix} \beta_3 \\
&= \frac{\begin{bmatrix} \sum_i x_{i2}^2 & -\sum_i x_{i1}x_{i2} \\ -\sum_i x_{i1}x_{i2} & \sum_i x_{i1}^2 \end{bmatrix}}{\sum_i x_{i1}^2 \sum_i x_{i2}^2 - (\sum_i x_{i1}x_{i2})^2} \begin{bmatrix} \sum_i x_{i1}x_{i3} \\ \sum_i x_{i2}x_{i3} \end{bmatrix} \beta_3.
\end{aligned}$$

- Hence, the expected value of x_2's slope coefficient from the incomplete model, $\widehat{\tilde{\beta}}_2$, is

$$\mathsf{E}(\widehat{\tilde{\beta}}_2) = \beta_2 + \frac{\sum_i x_{i1}^2 \sum_i x_{i2}x_{i3} - \sum_i x_{i1}x_{i2} \sum_i x_{i1}x_{i3}}{\sum_i x_{i1}^2 \sum_i x_{i2}^2 - (\sum_i x_{i1}x_{i2})^2}\beta_3, \tag{47}$$

  where, due to the zero mean assumption, terms of the form

$$\sum_i x_{i1}x_{i2} \tag{48}$$

  can be interpreted as the covariance between x_1 and x_2 in the sample.

- Therefore, even if x_2 is uncorrelated with x_3, $\widehat{\tilde{\beta}}_2$ will be biased as long as either x_1 is also uncorrelated with x_3 or x_2 is also uncorrelated with x_1.

- However, the reasoning leading to Table 1 is often followed as a rough guide for obtaining the probable bias in estimators in more complex settings.

Including Irrelevant Variables

- The opposite of the problem considered above is the inclusion of irrelevant variables in the model.

- Actually, this means considering variables with a population slope coefficient of zero.

- As OLS is unbiased, the expected value of the slope estimators of these variables will be zero, and the unbiasedness of OLS will not be affected.

- However, including irrelevant variables may increase the variance of the OLS estimators, as we shall see below.

The Covariance Matrix of the OLS Estimator

- Among other things, the covariance matrix is required for testing statistical hypotheses about the parameter vector $\boldsymbol{\beta}$.

- The covariance matrix of $\widehat{\boldsymbol{\beta}}$ is

$$
\begin{aligned}
\mathsf{Cov}(\widehat{\boldsymbol{\beta}}|\boldsymbol{X}) &= \begin{bmatrix} \mathsf{Var}(\widehat{\beta}_0|\boldsymbol{X}) & \mathsf{Cov}(\widehat{\beta}_0,\widehat{\beta}_1|\boldsymbol{X}) & \cdots & \mathsf{Cov}(\widehat{\beta}_0,\widehat{\beta}_k|\boldsymbol{X}) \\ \mathsf{Cov}(\widehat{\beta}_0,\widehat{\beta}_1|\boldsymbol{X}) & \mathsf{Var}(\widehat{\beta}_1|\boldsymbol{X}) & \cdots & \mathsf{Cov}(\widehat{\beta}_1,\widehat{\beta}_k|\boldsymbol{X}) \\ \vdots & \vdots & \ddots & \vdots \\ \mathsf{Cov}(\widehat{\beta}_0,\widehat{\beta}_k|\boldsymbol{X}) & \mathsf{Cov}(\widehat{\beta}_k,\widehat{\beta}_1|\boldsymbol{X}) & \cdots & \mathsf{Var}(\widehat{\beta}_k|\boldsymbol{X}) \end{bmatrix} \\
&= \mathsf{E}[(\widehat{\boldsymbol{\beta}} - \mathsf{E}(\widehat{\boldsymbol{\beta}}))(\widehat{\boldsymbol{\beta}} - \mathsf{E}(\widehat{\boldsymbol{\beta}}))'|\boldsymbol{X}] \\
&\overset{(34)}{=} \mathsf{E}\left[(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{u}\boldsymbol{u}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}|\boldsymbol{X}\right] \\
&= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathsf{E}(\boldsymbol{u}\boldsymbol{u}'|\boldsymbol{X})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&\overset{(33)}{=} (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\sigma^2\boldsymbol{I}_T)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} \\
&= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1},
\end{aligned}
$$

where $(\boldsymbol{A}\boldsymbol{B})' = \boldsymbol{B}'\boldsymbol{A}'$ was used.

- To illustrate, consider the familiar simple regression model, where we have

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix}, \quad \boldsymbol{X'X} = \begin{bmatrix} n & \sum_i x_{i1} \\ \sum_i x_{i1} & \sum_i x_{i1}^2 \end{bmatrix}, \quad \boldsymbol{X'y} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_{i1} y_i \end{bmatrix}.$$

Thus,

$$\begin{aligned} \mathsf{Cov}([\widehat{\beta}_0, \widehat{\beta}_1]') &= \sigma^2 \frac{\begin{bmatrix} \sum_i x_{i1}^2 & -\sum_i x_{i1} \\ -\sum_i x_{i1} & n \end{bmatrix}}{n \sum_i x_{i1}^2 - \left(\sum_i x_{i1}\right)^2} \\ &= \frac{\sigma^2}{n s_{x_1}^2} \begin{bmatrix} \overline{x_1^2} & -\overline{x}_1 \\ -\overline{x}_1 & 1 \end{bmatrix}. \end{aligned}$$

- The variance of the OLS estimator is important for statistical inference, and it also helps figuring out several further properties of OLS, as we shall now see.

Components of OLS Variances

- Consider a simplified setting where all variables have zero mean, so that there is no intercept in the model

$$y = \beta_1 x_1 + \beta_2 x_2 + u. \tag{49}$$

We can calculate the covariance matrix of the OLS estimator $\widehat{\boldsymbol{\beta}} = [\widehat{\beta}_1, \widehat{\beta}_2]'$ by using the formula

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

$$\begin{aligned} \sigma^2 (\boldsymbol{X}'\boldsymbol{X})^{-1} &= \sigma^2 \begin{bmatrix} \sum_i x_{i1}^2 & \sum_i x_{i1} x_{i2} \\ \sum_i x_{i1} x_{i2} & \sum_i x_{i2}^2 \end{bmatrix}^{-1} \\ &= \frac{\sigma^2}{\sum_i x_{i1}^2 \sum_i x_{i2}^2 - \left(\sum_i x_{i1} x_{i2}\right)^2} \begin{bmatrix} \sum_i x_{i2}^2 & -\sum_i x_{i1} x_{i2} \\ -\sum_i x_{i1} x_{i2} & \sum_i x_{i1}^2 \end{bmatrix}, \end{aligned}$$

so

$$\begin{aligned}
\mathsf{Var}(\widehat{\beta}_1) &= \frac{\sigma^2 \sum_i x_{i2}^2}{\sum_i x_{i1}^2 \sum_i x_{i2}^2 - \left(\sum_i x_{i1} x_{i2}\right)^2} \\
&= \frac{\sigma^2}{\sum_i x_{i1}^2 \left(1 - \frac{\left(\sum_i x_{i1} x_{i2}\right)^2}{\sum_i x_{i1}^2 \sum_i x_{i2}^2}\right)} \\
&= \frac{\sigma^2}{\sum_i x_{i1}^2 \left(1 - r_{x_1,x_2}^2\right)} \\
&= \frac{\sigma^2}{n s_{x_1}^2 \left(1 - r_{x_1,x_2}^2\right)},
\end{aligned}$$

where r_{x_1,x_2}^2 is the squared correlation coefficient between x_1 and x_2, and therefore the coefficient of determination of a regression of x_1 on x_2 (or vice versa).

- This result holds generally, i.e., for regressions with an arbitrary number of regressors and intercept:

- Under the Gauß–Markov Assumptions,

$$\mathsf{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{n s_{x_j}^2 (1 - R_j^2)}, \tag{50}$$

where σ^2 is the variance of the error term,

$$s_{x_j}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_{ij} - \overline{x}_j)^2, \tag{51}$$

and R_j^2 is the coefficient of determination from a regression of x_j on all the other variables in $\boldsymbol{X}$ (and a constant, but not y).

- So the components of the sampling variance of $\widehat{\beta}_j$ are as follows.

1) The sample size, n.

2) The error term variance, σ^2. This can only be reduced by adding more independent variables to the regression, that is, taking some factors out of the error term. Clearly this requires that additional factors can be identified that substantially affect y.

3) The variation in x_j, $s^2_{x_j}$. As the slope coefficient β_j measures how y changes with x_j, we may expect that learning about β_j is easier when x_j changes a lot.

4) The correlation between x_j and the other independent variables in the sample.
   Perfect correlation, i.e., linear dependence, where $R^2_j = 1$, is the case of perfect collinearity, whereas high (but not perfect) correlation is referred to as **multicollinearity**.
   When the regressors are highly correlated, it is statistically difficult to disentangle the impact of x_1 from that of x_2. As a consequence, the precision of individual estimates is reduced.

- Factor $(1 - R_j^2)^{-1}$ is also referred to as *variance inflation factor* (VIF).

- If correlation is very high, this may indicate that the questions we want the data to answer are just a bit too complex.

- An example in the textbook refers to the situation where student performance is regressed on various school expenditure categories (in the US) like teacher salaries, instructional materials, sports, etc.

- All of these tend to be highly correlated, and finding out their partial impacts may just be too ambitious.

- It may be more reasonable to just consider the variable "expenditure per student" as a summary of all of them.

"Partialling Out" Interpretation of Multiple Regression

- This helps illustrating several of the properties of OLS discussed so far.

- We consider again the model with $k = 2$ with all variables having zero mean, i.e.,

$$y = \beta_1 x_1 + \beta_2 x_2 + u.$$

- The OLS estimator is

$$\begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \sum_i x_{i1}^2 & \sum_i x_{i1}x_{i2} \\ \sum_i x_{i1}x_{i2} & \sum_i x_{i2}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum_i x_{i1}y_i \\ \sum_i x_{i2}y_i \end{bmatrix},$$

which gives, with $s_1^2 = \sum_i x_{1i}^2$, $s_2^2 = \sum_i x_{2i}^2$, $s_{12} = \sum_i x_{i1}x_{i2}$,

$$\begin{aligned} \widehat{\beta}_1 &= \frac{\sum_i x_{i2}^2 \sum_i x_{i1}y_i - \sum_i x_{i1}x_{i2} \sum_i x_{i2}y_i}{\sum_i x_{i1}^2 \sum_i x_{i1}^2 - (\sum_i x_{i1}x_{i2})^2} \\ &= \frac{\sum_i y_i \left(x_{i1} - \frac{s_{12}}{s_2^2} x_{i2} \right)}{\sum_i x_{i1} \left(x_{i1} - \frac{s_{12}}{s_2^2} x_{i2} \right)} = \cdots \end{aligned}$$

$$\cdots = \widehat{\beta}_1 = \frac{\sum_i y_i \widehat{v}_i}{\sum_i \widehat{v}_i^2}, \tag{52}$$

where the $\widehat{v}_i$ are the residuals from the simple regression of x_1 on x_2,

$$\begin{aligned} x_{i1} &= \widehat{\delta}_1 x_{i2} + \widehat{v}_i, \quad \widehat{\delta}_1 = \frac{s_{12}}{s_2^2}, \\ \sum_i \widehat{v}_i^2 &= \sum_i \left(x_{i1} - \frac{s_{12}}{s_2^2} x_{i2} \right)^2 = \sum_i x_{i1} \left(x_{i1} - \frac{s_{12}}{s_2^2} x_{i2} \right) = \sum_i x_{i1} \widehat{v}_i, \end{aligned}$$

since

$$\sum_i x_{2i} \widehat{v}_i = 0$$

by the OLS first–order condition.

- Equation (52) states that, in order to calculate $\widehat{\beta}_1$, we may first regress x_1 on x_2 with residuals $\widehat{v}$, and then calculate the regression of y on $\widehat{v}$ to obtain $\widehat{\beta}_1$.

- Thus, $\widehat{\beta}_1$ measures the relationship between y and x_1 after the effects of x_2 have been partialled out.

- What is left in the residuals $\widehat{v}$ (and measured by $\widehat{\beta}_1$) is the variation in y *that matches up uniquely with variation in* x_1.

- Clearly the same reasoning applies to $\widehat{\beta}_2$.

- If more than two variables are in the model, the same interpretation is true for each of the slope coefficients $\widehat{\beta}_j$, but the residuals in (52) are then obtained from a regression of all independent variables under study except variable x_j (when $\widehat{\beta}_j$ is computed).

- This can be related to the multicollinearity issue: If the variables are highly correlated, there is not enough *independent* variation in each variable for OLS to provide precise estimates.

Gauß–Markov Theorem

- The Gauß–Markov Theorem can also be derived for the multiple regression model.

- It states that, provided the Gauß–Markov Assumptions hold, the OLS estimators have the smallest variance in the class of linear unbiased estimators.

- That is, for any other linear unbiased estimator, $\tilde{\beta}_j$,

$$\text{Var}(\tilde{\beta}_j) \geq \text{Var}(\widehat{\beta}_j), \quad j = 0, \ldots, k. \tag{53}$$

- To see this, consider any other linear unbiased estimator of $\boldsymbol{\beta}$, say

$$\tilde{\boldsymbol{\beta}} = \boldsymbol{A}\boldsymbol{y} = \boldsymbol{A}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}). \tag{54}$$

- For (54) to be unbiased, we need to have $\boldsymbol{A}\boldsymbol{X} = \boldsymbol{I}$ (the identity of dimension $k + 1$), and then, proceeding in the same way as for the OLS estimator,

$$\text{Cov}(\tilde{\boldsymbol{\beta}}) = \sigma^2 \boldsymbol{A}\boldsymbol{A}'. \tag{55}$$

- Now define

$$\boldsymbol{B} = \boldsymbol{A} - (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}',$$

so that (55) may be written

$$\begin{aligned}\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) &= \sigma^2(\boldsymbol{B} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')(\boldsymbol{B}' + \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}) \\ &= \sigma^2\left(\boldsymbol{B}\boldsymbol{B}' + (\boldsymbol{X}'\boldsymbol{X})^{-1} + \boldsymbol{B}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{B}'\right).\end{aligned}$$

- Since, by the definition of $\boldsymbol{B}$,

$$\boldsymbol{B}\boldsymbol{X} = \underbrace{\boldsymbol{A}\boldsymbol{X}}_{=\boldsymbol{I}} - \underbrace{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}}_{=\boldsymbol{I}} = \boldsymbol{0},$$

we get

$$\mathrm{Cov}(\tilde{\boldsymbol{\beta}}) = \sigma^2\left(\boldsymbol{B}\boldsymbol{B}' + (\boldsymbol{X}'\boldsymbol{X})^{-1}\right).$$

- The variances are on the diagonal of the covariance matrix.

- Thus, with the $(k+1)$–dimensional column vector $\boldsymbol{e}_j$ defined as[1]

$$\boldsymbol{e}_j = [\underbrace{0 \ldots 0}_{j \text{ times}} \ 1 \ \underbrace{0 \ldots 0}_{k-j \text{ times}} \ ]', \quad j = 0, \ldots, k,$$

$$\begin{aligned} \text{Var}(\tilde{\beta}_j) &= \sigma^2 \boldsymbol{e}_j' (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{e}_j + \sigma^2 \boldsymbol{e}_j' \boldsymbol{B}\boldsymbol{B}' \boldsymbol{e}_j \\ &= \text{Var}(\widehat{\beta}_j) + \sigma^2 \boldsymbol{e}_j' \boldsymbol{B}\boldsymbol{B}' \boldsymbol{e}_j. \end{aligned} \tag{56}$$

- The second term on the right hand side of (56) is nonnegative, since, with vector $\boldsymbol{a}_j := \boldsymbol{B}'\boldsymbol{e}_j$,

$$\boldsymbol{e}_j' \boldsymbol{B}\boldsymbol{B}' \boldsymbol{e}_j = (\boldsymbol{B}\boldsymbol{e}_j)'(\boldsymbol{B}'\boldsymbol{e}_j) = \boldsymbol{a}_j' \boldsymbol{a}_j = \sum_{i=1}^{k+1} a_{ij}^2 \geq 0.$$

[1] If $j = 0$, corresponding to β_0, then the 1 is in the first position of $\boldsymbol{e}_0$.

- The same reasoning can be applied to any linear combination of the β_js.

- I.e., if we are interested in estimating a linear combination of the parameters such as $\beta_1 + \beta_2$, or $\beta_1 + 2\beta_2 - 5\beta_3$, the linear combination of the OLS estimators provides best linear unbiased estimator of this linear combination.

- Summarizing, the Gauß–Markov Theorem implies that, under the Gauß–Markov Assumptions, we need not look for for alternative unbiased estimators of the simple form $\tilde{\boldsymbol{\beta}} = \boldsymbol{Ay}$, as any estimator of this form will be inferior to OLS.

- The result is also useful as it helps to construct efficient estimators in cases where one of the Gauß–Markov Assumptions turns out to be violated.

Estimation of the Error Variance

- For conducting statistical inference about the elements of $\boldsymbol{\beta}$, we need to estimate their variances, and this requires an estimator of the error term variance σ^2.

- An unbiased estimator of σ^2 is given by

$$\widehat{\sigma}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\widehat{u}_i^2 = \frac{\widehat{\boldsymbol{u}}'\widehat{\boldsymbol{u}}}{n-k-1}, \tag{57}$$

  which has been shown for $k = 1$ (simple regression) in the exercises.

- Note that, without a constant term, we have to divide by $n - k$.