

Ch03. Multiple Regression Analysis: Estimation

Ping Yu

HKU Business School
The University of Hong Kong

Motivation for Multiple Regression

Motivation for Multiple Regression

- The multiple linear regression (MLR) model is defined as

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u,$$

which tries to explain variable y in terms of variables x_1, \dots, x_k .

- The terminology for $y, (x_1, \dots, x_k), u, \beta_0, (\beta_1, \dots, \beta_k)$ is the same as in the SLR model.
- **Motivation:**
 - 1 Incorporate more explanatory factors into the model;
 - 2 Explicitly hold fixed other factors that otherwise would be in u ;
 - 3 Allow for more flexible functional forms;
- Motivation 1 is easy to understand, so we will provide four examples to illustrate the other two motivations: Examples 1 and 2 for motivation 2 and Examples 3 and 4 for motivation 3.

Example: Wage Equation

- Suppose

$$wage = \beta_0 + \beta_1 educ + \beta_2 exper + u,$$

where

$wage$ = hourly wage

$educ$ = years of education

$exper$ = years of labor market experience

u = all other factors affecting wage

- Now, β_1 measures effect of education explicitly holding experience fixed.
- If omitting $exper$, then $E[u|educ] \neq 0$ given that $educ$ and $exper$ are correlated
 $\implies \hat{\beta}_1$ is biased.

Example: Average Test Scores and Per Student Spending

- Suppose

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u,$$

where

avgscore = average standardized test score of school

expend = per student spending at this school

avginc = average family income of students at this school

u = all other factors affecting *avgscore*

- Per student spending is likely to be correlated with average family income at a given high school because of school financing.
- Omitting average family income in regression would lead to biased estimate of the effect of spending on average test scores.
- In a simple regression model, effect of per student spending would partly include the effect of family income on test scores. [[intuition here](#), direct and indirect effects]

Example: Family Income and Family Consumption

- Suppose

$$cons = \beta_0 + \beta_1 inc + \beta_2 inc^2 + u,$$

where

$cons$ = family consumption

inc = family income

inc^2 = family income squared

u = all other factors affecting $cons$

- Model has two explanatory variables: income and income squared.
- Consumption is explained as a quadratic function of income.
- One has to be very careful when interpreting the coefficients:

$$\frac{\partial cons}{\partial inc} = \beta_1 + 2\beta_2 inc,$$

which depends on how much income is already there. [[intuition here](#)]

Example: CEO Salary, Sales and CEO Tenure

- Suppose

$$\log(\text{salary}) = \beta_0 + \beta_1 \log(\text{sales}) + \beta_2 \text{ceoten} + \beta_3 \text{ceoten}^2 + u,$$

where

$\log(\text{salary})$ = log of CEO salary

$\log(\text{sales})$ = log sales

ceoten = CEO tenure with the firm

- Model assumes a constant elasticity relationship between CEO salary and the sales of his or her firm.
- Model assumes a quadratic relationship between CEO salary and his or her tenure with the firm.
- Recall that the "linear" in linear regression means linear in parameter, not "linear in the variables".

Mechanics and Interpretation of Ordinary Least Squares

a: Obtaining the OLS Estimates

- Suppose we have a random sample $\{(x_{j1}, \dots, x_{jk}, y_i) : i = 1, \dots, n\}$, where the first subscript of x_{ij} , i , refers to the observation number, and the second subscript j , $j = 1, \dots, k$, refers to different independent variables.
- Define the residuals at arbitrary $\beta \equiv (\beta_0, \beta_1, \dots, \beta_k)$ as

$$\hat{u}_i(\beta) = y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik}.$$

- Minimize the sum of squared residuals:

$$\begin{aligned} \min_{\beta} SSR(\beta) &= \min_{\beta} \sum_{i=1}^n \hat{u}_i(\beta)^2 = \min_{\beta_0, \beta_1, \dots, \beta_k} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik})^2 \\ &\implies \hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_k), \end{aligned}$$

where $\hat{\beta}$ is the solution to the FOCs,

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) &= 0, \\ \sum_{i=1}^n x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) &= 0, \\ &\vdots \\ \sum_{i=1}^n x_{ik} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \dots - \hat{\beta}_k x_{ik}) &= 0, \end{aligned} \tag{1}$$

which can be carried out through standard econometric softwares.

b: Interpreting the OLS Regression Equation

- In the MLR model, $y = \beta_0 + \beta_1 x_1 + \dots + \beta_k x_k + u$, so

$$\beta_j = \frac{\partial y}{\partial x_j}$$

means "by how much does the dependent variable change if the j -th independent variable is increased by one unit, holding all other independent variables and the error term constant".

- The multiple linear regression model manages to hold the values of other explanatory variables fixed even if, in reality, they are correlated with the explanatory variable under consideration.
- This is the usual "ceteris paribus"-interpretation.
- It has still to be assumed that unobserved factors do not change if the explanatory variables are changed.

Example: Determinants of College GPA

- The fitted regression is

$$\widehat{colGPA} = 1.29 + .453hsGPA + .0094ACT,$$

where

colGPA = grade point average at college

hsGPA = high school grade point average

ACT = achievement test score¹

- Holding ACT fixed, another point on high school GPA is associated with another .453 points college GPA.
- Or: If we compare two students with the same ACT, but the *hsGPA* of student A is one point higher, we predict student A to have a *colGPA* that is .453 higher than that of student B.
- Holding high school GPA fixed, another 10 points on ACT are associated with less than one-tenth point on college GPA.

¹An achievement test is a test of developed skill or knowledge. The most common type of achievement test is a standardized test developed to measure skills and knowledge learned in a given grade level such as SAT, usually through **planned** instruction, such as training or classroom instruction. Achievement tests are often contrasted with tests that measure aptitude, a more general and stable cognitive trait.

f: A "Partialling Out" Interpretation of Multiple Regression

- One can show that the estimated coefficient of an explanatory variable in a multiple regression can be obtained in two steps:
 - 1 Regress the explanatory variable on all other explanatory variables.
 - 2 Regress y on the residuals from this regression
- Mathematically, suppose we regress y on the constant 1, x_1 and x_2 (denoted as $y \sim 1, x_1, x_2$), and want to get $\hat{\beta}_1$.
 - 1 $x_{i1} \sim 1, x_{i2} \implies \hat{r}_{i1}$
 - 2 $y_i \sim 1, x_{i2} \implies \hat{r}_{iy}$ (no need) [[Assignment II](#), Problem 1(i)]
 - 3 y_i (or \hat{r}_{iy}) $\sim \hat{r}_{i1} \implies$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2}$$

- In Step 3, the constant regressor is not required since the mean of \hat{r}_{i1} is equal to zero from Step 1. If the constant regressor is added in, the formula of $\hat{\beta}_1$ is the same:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (\hat{r}_{i1} - \bar{\hat{r}}_1) y_i}{\sum_{i=1}^n (\hat{r}_{i1} - \bar{\hat{r}}_1)^2} = \frac{\sum_{i=1}^n \hat{r}_{i1} y_i}{\sum_{i=1}^n \hat{r}_{i1}^2}$$

since $\bar{\hat{r}}_1 = \frac{1}{n} \sum_{i=1}^n \hat{r}_{i1} = 0$.

(*) A Formal Derivation of the $\hat{\beta}_1$ Formula (Appendix 3A.2)

- From Step 1, $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$ with $\hat{x}_{i1} = \hat{\gamma}_0 + \hat{\gamma}_1 x_{i2}$, $\sum_{i=1}^n \hat{r}_{i1} = 0$, $\sum_{i=1}^n x_{i2} \hat{r}_{i1} = 0$ and $\sum_{i=1}^n \hat{x}_{i1} \hat{r}_{i1} = 0$ (recall from the SLR).
- Recall that the FOCs of OLS when $k = 2$ are

$$\begin{aligned} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) &= 0, \\ \sum_{i=1}^n x_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) &= 0, \\ \sum_{i=1}^n x_{i2} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) &= 0. \end{aligned}$$

- From the second FOC,

$$\begin{aligned} & \sum_{i=1}^n (\hat{\gamma}_0 + \hat{\gamma}_1 x_{i2} + \hat{r}_{i1}) (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) \\ &= \hat{\gamma}_0 \sum_{i=1}^n \hat{u}_i + \hat{\gamma}_1 \sum_{i=1}^n x_{i2} \hat{u}_i + \sum_{i=1}^n \hat{r}_{i1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}) \\ &= -\hat{\beta}_0 \sum_{i=1}^n \hat{r}_{i1} - \hat{\beta}_2 \sum_{i=1}^n x_{i2} \hat{r}_{i1} + \sum_{i=1}^n \hat{r}_{i1} [y_i - \hat{\beta}_1 (\hat{x}_{i1} + \hat{r}_{i1})] \\ &= \sum_{i=1}^n \hat{r}_{i1} (y_i - \hat{\beta}_1 \hat{r}_{i1}) = 0, \end{aligned}$$

where $\hat{u}_i \equiv y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2}$, the second equality is from the first and third FOCs, and the third equality is from the properties of \hat{r}_{i1} above.


- Solving the last equality, $\sum_{i=1}^n \hat{r}_{i1} y_i = \hat{\beta}_1 \sum_{i=1}^n \hat{r}_{i1}^2$, we get the $\hat{\beta}_1$ formula.

Why does This Procedure Work?

- This procedure is usually called the **FWL theorem** [[photo here](#)] and was proposed in the following two papers:
 - Frisch, R. And F. Waugh, 1933, Partial Time Regressions as Compared with Individual Trends, *Econometrica*, 1, 387-401.
 - Lovell, M.C., 1963, Seasonal Adjustment of Economic Time Series, *Journal of the American Statistical Association*, 58, 993-1010.
- The residuals from the first regression is the part of the explanatory variable that is **uncorrelated** with the other explanatory variables.
- The slope coefficient of the second regression therefore represents the **isolated** (or **pure**) effect of the explanatory variable on the dependent variable.
- Recall that in the SLR,

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

so in the MLR, we replace $x_i - \bar{x}$ by \hat{r}_{i1} . Actually, $x_i - \bar{x}$ is the residual in the regression of x_i on all other explanatory variables,² where \bar{x} is the coefficient of the only regressor 1.

²In the SLR, all other explanatory variables include only the constant 1. 

History of FWL



R. Frisch (1895-1973), Oslo, 1969NP



F.V. Waugh (1898-1974), USDA



M.C. Lovell (1930-), Wesleyan

(**) The Moment Conditions for OLS

- The least squares estimator (LSE) can be interpreted as a MoM estimator.
- (1) are the sample counterparts of the population moment conditions

$$\begin{aligned} E[u] &= 0, \\ E[x_j u] &= 0, \end{aligned}$$

where

$$u = y - \beta_0 - \beta_1 x_1 - \dots - \beta_k x_k$$

and $\beta_j, j = 0, 1, \dots, k$, is the true value of the coefficient of x_j .

- From the FWL theorem, if we define \tilde{u} as $\tilde{y} - \beta_1 \tilde{x}_1 - \dots - \beta_k \tilde{x}_k$,³ then these moment conditions can be rewritten as

$$\begin{aligned} E[\tilde{u}] &= 0, \\ E[\tilde{x}_j \tilde{u}] &= 0, \end{aligned}$$

where $\tilde{y} = y - E[y]$ and $\tilde{x}_j = x_j - E[x_j], j = 1, 2, 3$, are the demeaned y and x_j 's.
 - Note that $E[\tilde{x}_j \tilde{u}] = \text{Cov}(x_j, \tilde{u}) = \text{Cov}(x_j, u)$.

³Actually, $\tilde{u} = u$. Why? $u - \tilde{u} = E[y] - \beta_0 - \beta_1 E[x_1] - \dots - \beta_k E[x_k] = 0$ because $E[u] \equiv 0$.

e: Properties of OLS on Any Sample of Data

- Fitted values and residuals:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_k x_{ik} \text{ and } \hat{u}_i = \hat{u}_i(\hat{\beta}) = y_i - \hat{y}_i$$

- Algebraic properties of OLS regression:

- (i) $\sum_{i=1}^n \hat{u}_i = 0$: deviations from the fitted regression "plane" sum up to zero.

$-\bar{y} = \hat{\beta}_0 + \bar{x}_1 \hat{\beta}_1 + \cdots + \bar{x}_k \hat{\beta}_k$: sample averages of y and of the regressors lie on the fitted regression plane.

- (ii) $\sum_{i=1}^n x_{ij} \hat{u}_i = 0, j = 1, \dots, k$: correlations between deviations and regressors are zero.

- These properties are corollaries of the FOCs for the OLS estimates.

h: Goodness-of-Fit

- Decomposition of total variation:

$$SST = SSE + SSR.$$

- R-squared:

$$R^2 = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}.$$

- Alternative expression for R-squared [proof not required]:

$$\begin{aligned} R^2 &= \frac{\left(\sum_{i=1}^n (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})\right)^2}{\left(\sum_{i=1}^n (y_i - \bar{y})^2\right) \left(\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2\right)} \\ &= \frac{\widehat{Cov}(y, \hat{y})^2}{\widehat{Var}(y) \widehat{Var}(\hat{y})} = \widehat{Corr}(y, \hat{y})^2, \end{aligned}$$

i.e., R -squared is equal to the squared correlation coefficient between the actual and the predicted value of the dependent variable.

- Because $\widehat{Corr}(y, \hat{y}) \in [-1, 1]$, $R^2 \in [0, 1]$. (when will $R^2 = 0$ and when will $R^2 = 1$?)

[Review] Correlation

- One useful property of covariance: For any constants a_1, b_1, a_2 and b_2 ,

$$\text{Cov}(a_1 + b_1 X, a_2 + b_2 Y) = b_1 b_2 \text{Cov}(X, Y).$$

- From this property, the covariance depends on units of measurement.⁴ For example, the covariance between education and earnings depends on whether earnings are measured in dollars or thousands of dollars, or whether education is measured in months or years.
- The population **correlation coefficient** [photo here] between X and Y , sometimes denoted as ρ_{XY} , is free of units:

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\text{sd}(X) \text{sd}(Y)} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.^5$$

- Because σ_X and σ_Y are positive, $\text{Cov}(X, Y)$ and $\text{Corr}(X, Y)$ always have the same sign, and $\text{Corr}(X, Y) = 0$ if, and only if, $\text{Cov}(X, Y) = 0$.
- Like $\text{Cov}(X, Y)$, $\text{Corr}(X, Y)$ is also a measure of linear dependence.

⁴The unit of $\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])]$ is the product of the units of X and Y .

⁵ $\text{Corr}(a_1 + b_1 X, a_2 + b_2 Y) = \frac{\text{Cov}(a_1 + b_1 X, a_2 + b_2 Y)}{\sqrt{\text{Var}(a_1 + b_1 X)} \sqrt{\text{Var}(a_2 + b_2 Y)}} = \frac{b_1 b_2 \text{Cov}(X, Y)}{\sqrt{b_1^2 \text{Var}(X)} \sqrt{b_2^2 \text{Var}(Y)}} = \frac{b_1 b_2 \text{Cov}(X, Y)}{|b_1 b_2| \text{sd}(X) \text{sd}(Y)} =$

$\text{sign}(b_1 b_2) \text{Corr}(X, Y)$, where $\text{sign}(x) = 1$ if $x > 0$, $\text{sign}(x) = -1$ if $x < 0$ and $\text{sign}(x) = 0$ if $x = 0$.



Karl Pearson (1857-1936), UCL

- Karl Pearson (1857-1936) is the inventor of the correlation coefficient, so the correlation coefficient is also called the **Pearson correlation coefficient**. He is also the inventor of the method of moments as discussed in Chapter 2.
- The **sample correlation**, $\widehat{Corr}(X, Y) = \frac{\widehat{Cov}(X, Y)}{\sqrt{\widehat{Var}(X)}\sqrt{\widehat{Var}(Y)}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$, is the MoM estimator of $Corr(X, Y)$.

(*) R -Squared Cannot Decrease When One More Regressor Is Added

- why? Recall that when there are k regressors,

$$SSR_k = \min_{\beta_0, \beta_1, \dots, \beta_k} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik})^2.$$

- When one more regressor is added to the regression,

$$SSR_{k+1} = \min_{\beta_0, \beta_1, \dots, \beta_k, \beta_{k+1}} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \dots - \beta_k x_{ik} - \beta_{k+1} x_{i,k+1})^2.$$

- Treat SSR_{k+1} as a function of β_{k+1} , i.e., for each value of β_{k+1} , we minimize the objective function of SSR_{k+1} with respect to $(\beta_0, \beta_1, \dots, \beta_k)$. Denote the resulting function as $SSR_{k+1}(\beta_{k+1})$.
- Obviously, when $\beta_{k+1} = 0$, the two objective functions of SSR_k and SSR_{k+1} are the same. So $SSR_{k+1}(0) = SSR_k$.
- However, in SSR_{k+1} , we search for the optimal β_{k+1} , so $SSR_{k+1} < SSR_k$ unless the optimal β_{k+1} (i.e., $\hat{\beta}_{k+1}$) is zero. [\[figure here\]](#)
- Because SST is the same in the two regressions, $SSR_{k+1} < SSR_k$ implies $R_{k+1}^2 > R_k^2$.

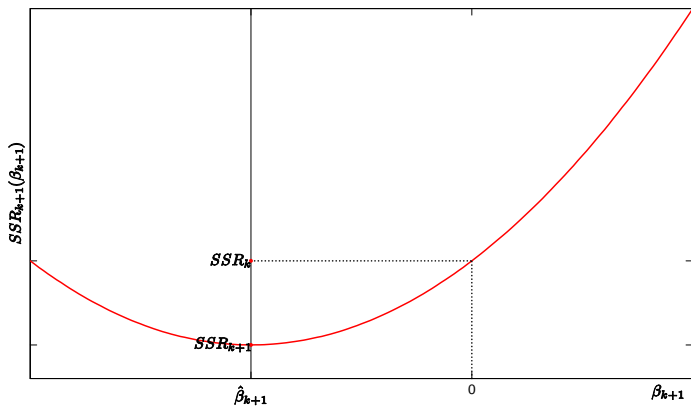


Figure: SSR_{k+1} As a Function of β_{k+1}

Example: Explaining Arrest Records

- The fitted regression line is

$$\widehat{narr86} = .712 - .150pcnv - .034ptime86 - .104qemp86$$

$$n = 2,725, R^2 = .0413 \text{ (quite small, not unusual)}$$

where

$narr86$ = number of times arrested during 1986

$pcnv$ = proportion (not percentage) of prior arrests that led to conviction

$ptime86$ = months spent in prison during 1986

$qemp86$ = the number of quarters employed in 1986

- $pcnv : +0.5 \implies -0.075$, i.e., -7.5 arrests per 100 men.
- $ptime86 : +12 \implies -.034 \times 12 = -0.408$ arrests for a given man.
- $qemp86 : +1 \implies -.104$, i.e., -10.4 arrests per 100 men - economic policies are effective.

continue

- An additional explanatory variable is added:

$$\widehat{narr86} = .707 - .151pcnv + .0074avg\textit{sen} - .037ptime86 - .103qemp86$$

$$n = 2,725, R^2 = .0422 \text{ (increases only slightly)}$$

where

*avg*sen = average sentence length in prior convictions (zero for most people)

- Average prior sentence increases number of arrests (anti-intuitive but $\hat{\beta}_2 \approx 0$).
- Limited additional explanatory power as R -squared increases by little (why? $\hat{\beta}_2 \approx 0$).
- General remark** on R -squared: even if R -squared is small (as in the given example), regression may still provide good estimates (i.e., s.e.'s are small) of ceteris paribus effects.
 - why? Recall in the SLR model, $\widehat{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{SST_x}$. $R^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}_y^2}$ is small means that $\hat{\sigma}^2$ is large (or $\hat{\sigma}_y^2$ is large), but SST_x can still be large to make $\widehat{Var}(\hat{\beta}_1)$ (or $se(\hat{\beta}_1)$) small.

The Expected Value of the OLS Estimators

Standard Assumptions for the MLR Model

- **Assumption MLR.1** (Linear in Parameters):

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u.$$

- In the population, the relationship between y and x is linear.
- The "linear" in linear regression means "linear in parameter".

- **Assumption MLR.2** (Random Sampling): The data $\{(x_{i1}, \dots, x_{ik}, y_i) : i = 1, \dots, n\}$ is a random sample drawn from the population, i.e., each data point follows the population equation,

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + u_i.$$

continue

- **Assumption MLR.3** (No Perfect Collinearity): In the sample (and therefore in the population), none of the independent variables is constant and there are no **exact** relationships among the independent variables.
- **Remarks:**
 - The assumption only rules out perfect collinearity/correlation between explanatory variables; imperfect correlation⁶ is allowed.
 - If an explanatory variable is a perfect linear combination of other explanatory variables it is redundant and may be eliminated.
 - Constant variables are also ruled out (collinear with the regressor 1).
- This is an extension of $\sum_{i=1}^n (x_i - \bar{x})^2 > 0$ in the SLR model. (why?)

⁶This is referred to as multicollinearity in the later discussion.

Example for Perfect Collinearity

- **Small Sample:**

$$avgscore = \beta_0 + \beta_1 expend + \beta_2 avginc + u,$$

- In a small sample, *avginc* may accidentally be an exact multiple of *expend*; it will not be possible to disentangle their separate effects because there is exact covariation (**why?** we cannot move one while fixing another).

- **Relationships Between Regressors:**

$$voteA = \beta_0 + \beta_1 shareA + \beta_2 shareB + u,$$

- Either *shareA* or *shareB* will have to be dropped from the regression because there is an exact linear relationship between them: $shareA + shareB = 1$.

- Another related example is the so-called **dummy variable trap**, which will be discussed in Chapter 7.

Standard Assumptions for the MLR Model (continue)

- Assumption MLR.4 (Zero Conditional Mean):

$$E[u_j | x_{i1}, \dots, x_{ik}] = 0.$$

- The value of the explanatory variables must contain no information about the mean of the unobserved factors.
- In a multiple regression model, the zero conditional mean assumption is much more likely to hold because fewer things end up in the error.
- Example: Reconsider

$$\text{avgscore} = \beta_0 + \beta_1 \text{expend} + \beta_2 \text{avginc} + u.$$

If *avginc* was not included in the regression, it would end up in the error term; it would then be hard to defend that *expend* is uncorrelated with the error.

Discussion of the Zero Conditional Mean Assumption

- Explanatory variables that are correlated with the error term are called **endogenous**; endogeneity is a violation of assumption MLR.4.
- Explanatory variables that are uncorrelated with the error term are called **exogenous**; MLR.4 holds if all explanatory variables are exogenous.
- Exogeneity is the **key** assumption for a causal interpretation of the regression, and for unbiasedness of the OLS estimators.
- **Theorem 3.1** (Unbiasedness of OLS): Under assumptions MLR.1-MLR.4,

$$E[\hat{\beta}_j] = \beta_j, j = 0, 1, \dots, k,$$

for any values of β_j .

- Unbiasedness is an average property in repeated samples; in a given sample, the estimates may still be far away from the true values.

a: Including Irrelevant Variables in a Regression Model

- Suppose

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

where $\beta_3 = 0$, i.e., x_3 is irrelevant to y .

- No problem because $E[\hat{\beta}_3] = \beta_3 = 0$.
- However, including irrelevant variables may increase sampling variance $\hat{\beta}_1$ and $\hat{\beta}_2$.

b: Omitted Variable Bias: the Simple Case

- Suppose the true model is

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

i.e., the true model contains both x_1 and x_2 ($\beta_1 \neq 0, \beta_2 \neq 0$), while the estimated model is

$$y = \alpha_0 + \alpha_1 x_1 + w,$$

i.e., x_2 is omitted.

- If x_1 and x_2 are correlated, assume a linear regression relationship between them:

$$x_2 = \delta_0 + \delta_1 x_1 + v.$$

- Then

$$\begin{aligned} y &= \beta_0 + \beta_1 x_1 + \beta_2 (\delta_0 + \delta_1 x_1 + v) + u \\ &= \beta_0 + \beta_2 \delta_0 + (\beta_1 + \beta_2 \delta_1) x_1 + (u + \beta_2 v). \end{aligned}$$

- If y is only regressed on x_1 , the estimated intercept is $\beta_0 + \beta_2 \delta_0 = \alpha_0$ and the estimated slope is $\beta_1 + \beta_2 \delta_1 = \alpha_1$.
- **why?** The new error term $w = u + \beta_2 v$ satisfies the zero conditional mean assumption: $E[u + \beta_2 v | x_1] = E[u | x_1] + \beta_2 E[v | x_1] = 0$.
- That is, all estimated coefficients will be biased.

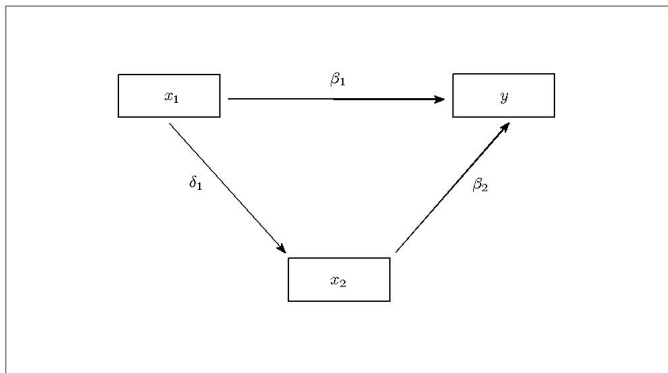


Figure: Chain of Omitted Variable Bias: Direct and Indirect Effects of x_1 on y

Example: Omitting Ability in a Wage Equation

- Suppose the true wage equation is

$$wage = \beta_0 + \beta_1 educ + \beta_2 abil + u, \text{ where } \beta_2 > 0,$$

and the estimated equation is

$$wage = \alpha_0 + \alpha_1 educ + w.$$

- Suppose

$$abil = \delta_0 + \delta_1 educ + v, \text{ where } \delta_1 > 0,$$

when

$$wage = \beta_0 + \beta_2 \delta_0 + (\beta_1 + \beta_2 \delta_1) educ + (u + \beta_2 v).$$

- The return to education β_1 will be overestimated because $\beta_2 \delta_1 > 0$. It will look as if people with many years of education earn very high wages, but this is partly due to the fact that people with more education are also more able on average.
- When is there no omitted variable bias? If the omitted variable is irrelevant ($\beta_2 = 0$) or uncorrelated ($\delta_1 = 0$).

c: Omitted Variable Bias: More General Cases

- Suppose the true model is

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

i.e., the true model contains x_1 , x_2 and x_3 ($\beta_j \neq 0, j = 1, 2, 3$), while the estimated model is

$$y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + w,$$

i.e., x_3 is omitted.

- No general statements possible about direction of bias.
- Analysis of β_1 is as in simple case if x_2 is uncorrelated with other regressors (x_1 and x_3). [see the next two slides for details]
- **Example:** Omitting ability in the following wage equation,

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{abil} + u,$$

if *exper* is approximately uncorrelated with *educ* and *abil*, then the direction of the omitted variable bias can be as analyzed in the simple two variable case.

(**) Appendix 3A.4

- Long regression: $y \sim 1, x_1, \dots, x_k \implies \hat{\beta}_j, j = 0, 1, \dots, k.$
- Short regression: $y \sim 1, x_1, \dots, x_{k-1} \implies \tilde{\beta}_j, j = 0, 1, \dots, k-1.$
- x_k regression: $x_k \sim 1, x_1, \dots, x_{k-1} \implies \tilde{\delta}_j, j = 0, 1, \dots, k-1.$
- Plugging the x_k regression to the long regression and collecting terms to have $\tilde{\beta}_j = \hat{\beta}_j + \hat{\beta}_k \tilde{\delta}_j, j = 0, 1, \dots, k-1.$
- $E[\tilde{\beta}_j | \mathbf{X}_{k-1}] = E[\hat{\beta}_j | \mathbf{X}_{k-1}] + E[\hat{\beta}_k \tilde{\delta}_j | \mathbf{X}_{k-1}] = \beta_j + \beta_k \delta_j,$ where \mathbf{X}_{k-1} collects $(x_{i1}, \dots, x_{i,k-1}), i = 1, \dots, n.$
- $E[\hat{\beta}_j | \mathbf{X}_{k-1}] \stackrel{(*)}{=} E[E[\hat{\beta}_j | \mathbf{X}_k] | \mathbf{X}_{k-1}] \stackrel{(1)}{=} \beta_j,$ and $E[\hat{\beta}_k \tilde{\delta}_j | \mathbf{X}_{k-1}] \stackrel{(*)}{=} E[E[\hat{\beta}_k \tilde{\delta}_j | \mathbf{X}_k] | \mathbf{X}_{k-1}] \stackrel{(2)}{=} E[\tilde{\delta}_j E[\hat{\beta}_k | \mathbf{X}_k] | \mathbf{X}_{k-1}] \stackrel{(1)}{=} \beta_k E[\tilde{\delta}_j | \mathbf{X}_{k-1}] \stackrel{(1)}{=} \beta_k \delta_j.$
 - (*) is due to the **law of iterated expectations (LIE)**: For random variables $Y, X, Z,$ $E[Y|X] = E[E[Y|X,Z]|X].$
 - (1) is due to the unbiasedness of $\hat{\beta}_j, \hat{\beta}_k$ and $\tilde{\delta}_j.$
 - (2) is because $\tilde{\delta}_j$ is a function of $\mathbf{X}_k.$
- $\tilde{\beta}_j$ is unbiased only if $\beta_k = 0$ or $\delta_j = 0.$

(**) The $k = 3$ Case

- When $k = 3$, $E[\tilde{\beta}_1 | \mathbf{X}_2] = \beta_1 + \beta_3 \delta_1$, so if δ_1 , the coefficient of x_1 in the regression $x_3 \sim 1, x_1, x_2$, is the same as the coefficient of x_1 in the regression $x_3 \sim 1, x_1$, then we are done.
- Does

$$\begin{pmatrix} E[x_3 - \delta_0 - \delta_1 x_1 - \delta_2 x_2] \\ E[\tilde{x}_1 (\tilde{x}_3 - \delta_1 \tilde{x}_1 - \delta_2 \tilde{x}_2)] \\ E[\tilde{x}_2 (\tilde{x}_3 - \delta_1 \tilde{x}_1 - \delta_2 \tilde{x}_2)] \end{pmatrix} = 0 \quad (2)$$

imply

$$\begin{pmatrix} E[x_3 - \delta_0 - \delta_1 x_1] \\ E[\tilde{x}_1 (\tilde{x}_3 - \delta_1 \tilde{x}_1)] \end{pmatrix} = 0 \quad (3)$$

for the same δ_0 and δ_1 if $\text{Cov}(x_2, x_1) = E[\tilde{x}_2 \tilde{x}_1] = E[\tilde{x}_2 \tilde{x}_3] = \text{Cov}(x_2, x_3) = 0$? where $\tilde{x}_j = x_j - E[x_j]$, $j = 1, 2, 3$, is the demeaned x_j , and the second and third equations in (2) and the second equation in (3) are from the FWL theorem.

- $E[\tilde{x}_2 \tilde{x}_1] = 0$ is enough to guarantee the same δ_1 from the second equation of (2), but for the same δ_0 , we further need $E[\tilde{x}_2 \tilde{x}_3] = 0$ to make sure $\delta_2 = 0$ from the third equation of (2).

The Variance of the OLS Estimators

Standard Assumptions for the MLR Model (continue)

- **Assumption MLR.5** (Homoskedasticity): $\text{Var}(u_i | \mathbf{x}_{i1}, \dots, \mathbf{x}_{ik}) = \sigma^2$.
- The value of the explanatory variables must contain no information about the variance of the unobserved factors.
- **Short Hand Notation:** $\text{Var}(u_i | \mathbf{x}_i) = \sigma^2$, where $\mathbf{x}_i = (x_{i1}, \dots, x_{ik})$, i.e., all explanatory variables are collected in a random vector.
- **Example:** In the wage equation

$$\text{wage} = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{exper} + \beta_3 \text{tenure} + u,$$

the homoskedasticity assumption

$$\text{Var}(u_i | \text{educ}_i, \text{exper}_i, \text{tenure}_i) = \sigma^2$$

may also be hard to justify in many cases.

Sampling Variances of the OLS Slope Estimators

- **Theorem 3.2** (Sampling Variances of the OLS Slope Estimators): Under assumptions MLR.1-MLR.5,

$$\text{Var}(\hat{\beta}_j) = \frac{\sigma^2}{\text{SST}_j(1 - R_j^2)}, j = 1, \dots, k,$$

where σ^2 is the variance of error term, $\text{SST}_j = \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2$ is the total sample variation in explanatory variable x_j , and R_j^2 is the R -squared from a regression of explanatory variable x_j on all other independent variables (including a constant), i.e., the R^2 in the regression

$$x_j \sim 1, x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_k. \quad (4)$$

- Note that

$$\text{SST}_j(1 - R_j^2) = \text{SSR}_j = \sum_{i=1}^n \hat{r}_{ij}^2,$$

where \hat{r}_{ij} is the residual in the regression (4).

- Compared with the SLR case where $\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\text{SST}_x} = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$, the MLR case replaces $x_i - \bar{x}$ by \hat{r}_{ij} .

a: The Components of OLS Variances

(i) The error variance, σ^2 :

- A high error variance increases the sampling variance because there is more "noise" in the equation.
- A large error variance necessarily makes estimates imprecise.
- The error variance does not decrease with sample size.

(ii) The total sample variation in the explanatory variable x_j , SST_j :

- More sample variation leads to more precise estimates.
- Total sample variation automatically increases with the sample size.
 - For any $n+1$ values $\{x_i : i = 1, \dots, n+1\}$,

$$\sum_{i=1}^n (x_i - \bar{x}_n)^2 \leq \sum_{i=1}^n (x_i - \bar{x}_{n+1})^2 \leq \sum_{i=1}^{n+1} (x_i - \bar{x}_{n+1})^2$$

unless $x_{n+1} = \bar{x}_n$, where the first inequality is from Assignment 1.4(i).

- $SST_j = n \left[\frac{1}{n} \sum_{i=1}^n (x_{ij} - \bar{x}_j)^2 \right] = n \cdot \widehat{\text{Var}}(x_j)$, where $\widehat{\text{Var}}(x_j)$ tends to be stable.

- Increasing the sample size n is thus a way to get more precise estimates.
- These two components are similar as in the SLR model.

continue

(iii) The linear relationships among the independent variables, R_j^2 :

- In the regression of x_j on all other independent variables (including a constant), the R-squared will be the higher the better x_j can be linearly explained by the other independent variables.
- Sampling variance of $\hat{\beta}_j$ will be the higher the better explanatory variable x_j can be linearly explained by other independent variables.
- The problem of almost linearly dependent explanatory variables is called **multicollinearity** (i.e., $R_j^2 \rightarrow 1$ for some j).
- If $R_j^2 = 1$, i.e., there is perfect collinearity between x_j and other regressors, then β_j cannot be identified. This is why $\text{Var}(\hat{\beta}_j) = \infty$ now.
- Multicollinearity plays a similar role as the sample size n in $\text{Var}(\hat{\beta}_j)$: both work to increase $\text{Var}(\hat{\beta}_j)$. Arthur Goldberger [[photo here](#)] coined the term **micronumerosity**, which he defines as the “problem of small sample size.”
- Like perfect collinearity, multicollinearity is a small-sample problem. As larger and larger data sets are available nowadays, i.e., n is much larger than k , it is seldom a problem in current econometric practice.

History of Micronumerosity



Arthur S. Goldberger (1930-2009),
Wisconsin, 1958MichiganPhD

An Example for Multicollinearity

- Consider the following MLR model,

$$avgscore = \beta_0 + \beta_1 teacherexp + \beta_2 matexp + \beta_3 othexp + \dots,$$

where

avgscore = average standardized test score of school

teacherexp = expenditures for teachers

matexp = expenditures for instructional materials

othexp = other expenditures

- The different expenditure categories will be strongly correlated because if a school has a lot of resources it will spend a lot on everything.
- It will be hard to estimate the differential effects of different expenditure categories because all expenditures are either high or low. For precise estimates of the differential effects, one would need information about situations where expenditure categories change differentially.
- As a consequence, sampling variance of the estimated effects will be large.

Discussion of the Multicollinearity Problem

- In the above example, it would probably be better to lump all expenditure categories together because effects cannot be disentangled.
- In other cases, dropping some independent variables may reduce multicollinearity (but this may lead to omitted variable bias).
- Only the sampling variance of the variables involved in multicollinearity will be inflated; the estimates of other effects may be very precise.
- Note that multicollinearity is not a violation of MLR.3 in the strict sense.
- Multicollinearity may be detected through "**variance inflation factors**":

$$VIF_j = \frac{1}{1 - R_j^2}.$$

- As an (arbitrary) rule of thumb, the variance inflation factor should not be larger than 10 (or R_j^2 should not be larger than 0.9). [[figure here](#)]

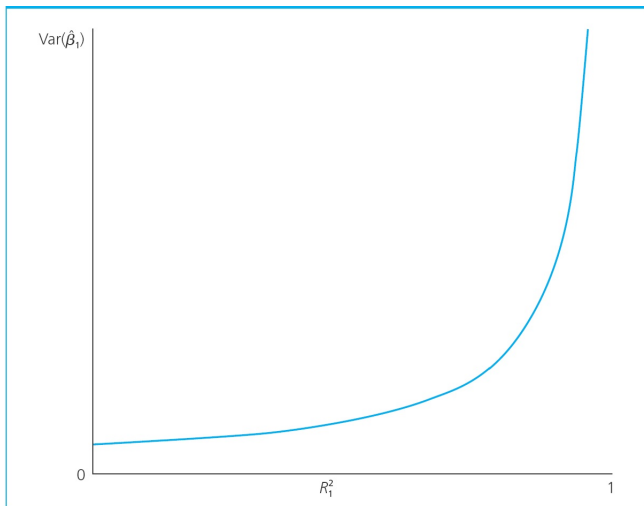


Figure: $\text{Var}(\hat{\beta}_1)$ as a Function of R_1^2

b: Variances in Misspecified Models

- The choice of whether to include a particular variable in a regression can be made by analyzing the trade-off between bias and variance.
- Suppose the true model is

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

the fitted regression line in model 1 is

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2,$$

and in model 2 is

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

- It might be the case that the likely omitted variable bias of $\tilde{\beta}_1$ in the misspecified model 2 is overcompensated by a smaller variance.
- (*) **Mean Squared Error (MSE)**: For a general estimator, say, $\hat{\beta}$,

$$\begin{aligned} \text{MSE}(\hat{\beta}) &\equiv E\left[(\hat{\beta} - \beta)^2\right] = E\left[(\hat{\beta} - E[\hat{\beta}] + E[\hat{\beta}] - \beta)^2\right] \\ &= E\left[(\hat{\beta} - E[\hat{\beta}])^2\right] + (E[\hat{\beta}] - \beta)^2 + 2(E[\hat{\beta}] - \beta)E[\hat{\beta} - E[\hat{\beta}]] \\ &= E\left[(\hat{\beta} - E[\hat{\beta}])^2\right] + (E[\hat{\beta}] - \beta)^2 = \text{Var}(\hat{\beta}) + \text{Bias}(\hat{\beta})^2. \end{aligned}$$

(*) Wrong Results in the Textbook

- Since

$$\text{Var}(\tilde{\beta}_1) = \frac{\sigma^2}{SST_1},$$

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{SST_1(1 - R_1^2)}$$

with $0 \leq R_1^2 \leq 1$, conditional on x_1 and x_2 , the variance in model 2 is always smaller than that in model 1.

- So there are two cases:

- 1 $\beta_2 = 0 \implies E[\hat{\beta}_1] = \beta_1, E[\tilde{\beta}_1] = \beta_1, \text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1).$
- 2 $\beta_2 \neq 0 \implies E[\hat{\beta}_1] = \beta_1, E[\tilde{\beta}_1] \neq \beta_1, \text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1).$

- Actually, there are four cases, depending on the relationship between x_1 and x_2 ($\delta_1 = 0?$) and whether x_2 is relevant ($\beta_2 = 0?$).

Correct Results


- The key is that the two σ^2 's in $\text{Var}(\tilde{\beta}_1)$ and $\text{Var}(\hat{\beta}_1)$ are not the same unless $\beta_2 = 0$.

	$\beta_2 = 0$ (two σ^2 are the same)	$\beta_2 \neq 0$ (σ^2 in $\text{Var}(\hat{\beta}_1)$ is smaller) ⁷
$R_1^2 = 0$ ($\delta_1 = 0$)	$E[\hat{\beta}_1] = \beta_1 = E[\tilde{\beta}_1]$ $\text{Var}(\tilde{\beta}_1) = \text{Var}(\hat{\beta}_1)$	$E[\hat{\beta}_1] = \beta_1 = E[\tilde{\beta}_1]$ $\text{Var}(\hat{\beta}_1) < \text{Var}(\tilde{\beta}_1)$
$R_1^2 \neq 0$ ($\delta_1 \neq 0$)	$E[\hat{\beta}_1] = \beta_1 = E[\tilde{\beta}_1]$ $\text{Var}(\tilde{\beta}_1) < \text{Var}(\hat{\beta}_1)$	$E[\hat{\beta}_1] = \beta_1 \neq E[\tilde{\beta}_1]$ undetermined

Table: Biases and Variances in Misspecified Models

- $R_1^2 = 0$ ($\beta_2 = 0$) plays two roles: (i) in mean: implies the unbiasedness of $\tilde{\beta}_1$;⁸ (ii) in variance: guarantees the denominators (numerators) of $\text{Var}(\tilde{\beta}_1)$ and $\text{Var}(\hat{\beta}_1)$ to be the same.

⁷why? If $\beta_2 \neq 0$, then x_2 can explain some "extra" variation in y beyond x_1 , so the remaining variation in y , i.e., the variance of the error term, is smaller. (Compare $\text{Var}(u)$ and $\text{Var}(w)$ in slide 31.)

⁸ $R_1^2 = 0 \xrightarrow{\text{Slide 44 of Chapter 2}} \hat{\gamma}_1 = 0 \implies \widehat{\text{Cov}}(x_1, x_2) = 0 \implies \hat{\delta}_1 = 0$ (or $\delta_1 = 0$ if n is large enough) 

c: Estimating the Error Variance

- The unbiased estimator of σ^2 is

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

where

$$\begin{aligned} n-k-1 &= n-(k+1) \\ &= (\text{number of observations}) - (\text{number of estimated parameters}) \end{aligned}$$

is called the **degree of freedom (df)**.

- The n estimated squared residuals $\{\hat{u}_i : i = 1, \dots, n\}$ in the sum are not completely independent but related through the $k+1$ equations that define the first order conditions of the minimization problem.
- In the SLR, $k = 1$. That is why $\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2$.
- Theorem 3.3** (Unbiased Estimation of σ^2): Under assumptions MLR.1-MLR.5,

$$E[\hat{\sigma}^2] = \sigma^2.$$

Estimation of the Sampling Variances of the OLS Estimators

- The true sampling variation of the estimated β_j is

$$sd(\hat{\beta}_j) = \sqrt{\text{Var}(\hat{\beta}_j)} = \sqrt{\frac{\sigma^2}{SST_j(1-R_j^2)}}.$$

- The estimated sampling variation of the estimated β_j , or the standard error of $\hat{\beta}_j$, is

$$se(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}(\hat{\beta}_j)} = \sqrt{\frac{\hat{\sigma}^2}{SST_j(1-R_j^2)}},$$

i.e., we plug in $\hat{\sigma}^2$ for the unknown σ^2 .

- Note that these formulas are only valid under assumptions MLR.1-MLR.5 (in particular, there has to be homoskedasticity).

Efficiency of OLS: The Gauss-Markov Theorem

Efficiency of OLS

- Under assumptions MLR.1-MLR.5, OLS is unbiased.
- However, under these assumptions there may be many other estimators that are unbiased.
- Which one is the unbiased estimator with the smallest variance?
- In order to answer this question one usually limits oneself to **linear** estimators, i.e., estimators linear in the dependent variable:

$$\tilde{\beta}_j = \sum_{i=1}^n w_{ij} y_i,$$

where w_{ij} may be an arbitrary function of the sample values of all the explanatory variables.

- The OLS estimator can be shown to be of this form. For example, in the SLR,

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

i.e.,

$$w_{j1} = \frac{x_i - \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{x_i - \bar{x}}{SST_x}$$

is a complicated function of $\{x_i : i = 1, \dots, n\}$. (How about $\hat{\beta}_j$ in MLR?)

The Gauss-Markov Theorem

- **Theorem 3.4** (The Gauss-Markov Theorem): [[photo here](#)] Under assumptions MLR.1-MLR.5, the OLS estimators are the best linear unbiased estimators (**BLUEs**) of the regression coefficients, i.e.,

$$\text{Var}(\widehat{\beta}_j) \leq \text{Var}(\widetilde{\beta}_j)$$

for all $\widetilde{\beta}_j = \sum_{i=1}^n w_{ij} y_i$ for which $E[\widetilde{\beta}_j] = \beta_j, j = 0, 1, \dots, k$.

- OLS is only the best estimator if MLR.1-MLR.5 hold; if there is heteroskedasticity for example, there are better estimators (see Chapter 8).
- The key assumption for the Gauss-Markov theorem is Assumption MLR.5.
- Due to the Gauss-Markov Theorem, assumptions MLR.1-MLR.5 are collectively known as the **Gauss-Markov assumption**.

History of the Gauss-Markov Theorem



Carl Friedrich Gauss (1777-1855), Göttingen



Andrey Markov (1856-1922), St. Petersburg

The Gauss-Markov Theorem

- **Theorem 3.4** (The Gauss-Markov Theorem): [[photo here](#)] Under assumptions MLR.1-MLR.5, the OLS estimators are the best linear unbiased estimators (**BLUEs**) of the regression coefficients, i.e.,

$$\text{Var}(\widehat{\beta}_j) \leq \text{Var}(\widetilde{\beta}_j)$$

for all $\widetilde{\beta}_j = \sum_{i=1}^n w_{ij}y_i$ for which $E[\widetilde{\beta}_j] = \beta_j, j = 0, 1, \dots, k$.

- OLS is only the best estimator if MLR.1-MLR.5 hold; if there is heteroskedasticity for example, there are better estimators (see Chapter 8).
- The key assumption for the Gauss-Markov theorem is Assumption MLR.5.
- Due to the Gauss-Markov Theorem, assumptions MLR.1-MLR.5 are collectively known as the **Gauss-Markov assumption**.

(**) Proof of The Gauss-Markov Theorem

Proof.

WLOG, let $j = 1$. Plugging y_i in $\tilde{\beta}_1$ to have

$$\tilde{\beta}_1 = \beta_0 \sum_{i=1}^n w_{i1} + \beta_1 \sum_{i=1}^n w_{i1} x_{i1} + \cdots + \beta_k \sum_{i=1}^n w_{i1} x_{ik} + \sum_{i=1}^n w_{i1} u_i.$$

Since the w_{i1} are functions of \mathbf{x}_{ij} ,

$$\begin{aligned} E[\tilde{\beta}_1 | \mathbf{X}] &= \beta_0 \sum_{i=1}^n w_{i1} + \beta_1 \sum_{i=1}^n w_{i1} x_{i1} + \cdots + \beta_k \sum_{i=1}^n w_{i1} x_{ik} + \sum_{i=1}^n w_{i1} E[u_i | \mathbf{X}] \\ &= \beta_0 \sum_{i=1}^n w_{i1} + \beta_1 \sum_{i=1}^n w_{i1} x_{i1} + \cdots + \beta_k \sum_{i=1}^n w_{i1} x_{ik}, \end{aligned}$$

where \mathbf{X} collects \mathbf{x}_i , $i = 1, \dots, n$, and $E[u_i | \mathbf{X}] = E[u_i | \mathbf{x}_i] = 0$. Therefore, for $E[\tilde{\beta}_1 | \mathbf{X}] = \beta_1$ for any values of the parameters, we must have

$$\sum_{i=1}^n w_{i1} = 0, \sum_{i=1}^n w_{i1} x_{i1} = 1, \sum_{i=1}^n w_{i1} x_{ij} = 0, j = 2, \dots, k. \quad (5)$$

Proof continue

Now, let \hat{r}_{i1} be the residuals from the regression of x_{i1} on $1, x_{i2}, \dots, x_{ik}$. Then, from (5), it follows that

$$\sum_{i=1}^n w_{i1} \hat{r}_{i1} = 1 \quad (6)$$

because $x_{i1} = \hat{x}_{i1} + \hat{r}_{i1}$ and $\sum_{i=1}^n w_{i1} \hat{x}_{i1} = 0$. Now, consider the difference between $\text{Var}(\tilde{\beta}_1 | \mathbf{X})$ and $\text{Var}(\hat{\beta}_1 | \mathbf{X})$:

$$\begin{aligned} & \sigma^2 \sum_{i=1}^n w_{i1}^2 - \frac{\sigma^2}{\sum_{i=1}^n \hat{r}_{i1}^2} \stackrel{(6)}{=} \sigma^2 \left[\sum_{i=1}^n w_{i1}^2 - \frac{(\sum_{i=1}^n w_{i1} \hat{r}_{i1})^2}{\sum_{i=1}^n \hat{r}_{i1}^2} \right] \\ &= \sigma^2 \sum_{i=1}^n \left(w_{i1} - \frac{\sum_{i=1}^n w_{i1} \hat{r}_{i1}}{\sum_{i=1}^n \hat{r}_{i1}^2} \hat{r}_{i1} \right)^2 \geq 0, \end{aligned}$$

where the last equality can be verified by direct calculation, and the term in the last parenthesis is the residual of w_{i1} regressed on \hat{r}_{i1} without intercept. \square