Linear Regression with Multiple Regressors

Osman DOGAN

Outline: Linear Regression with Multiple Regressors

- Linear Regression with Multiple Regressors
  1. Omitted variable bias
  2. Multiple regression and OLS
  3. Measures of fit
  4. Sampling distribution of the OLS estimator
  5. Multicollinearity
- Readings
  1. Stock and Watson (2020, Chapter 6),
  2. Hanck et al. (2021, Chapters 6).

Omitted variable bias

- Recall the OLS model:

$$Y = \beta_0 + \beta_1 X + u. \tag{1}$$

- The error u arises because of factors, or variables, that influence Y but are not included in the regression function.

Definition 1 (Omitted variable bias)

The bias in the OLS estimator that occurs as a result of an omitted factor, or variable, is called omitted variable bias.

- For omitted variable bias to occur, the omitted variable Z must satisfy two conditions:
  1. Z is a determinant of Y (i.e., Z is part of u); and
  2. Z is correlated with the regressor X (i.e., corr$(Z, X) \neq 0$).
- Both conditions must hold for the omission of Z to result in omitted variable bias.

Omitted variable bias

- To see the omitted variable bias, recall the following formula that we derived in Chapter 5:

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\frac{1}{n}\sum_{i=1}^{n} v_i}{\left(\frac{n-1}{n}\right) s_X^2}, \tag{2}$$

where $v_i = (X_i - \bar{X})u_i \approx (X_i - \mu_X)u_i$.

- If $\mathrm{E}\left[(X_i - \mu_X)u_i\right] = \text{cov}(X_i, u_i) = \sigma_{Xu} \neq 0$, then we have

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})u_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \xrightarrow{p} \frac{\sigma_{Xu}}{\sigma_X^2} = \frac{\sigma_u}{\sigma_X} \times \frac{\sigma_{Xu}}{\sigma_X \sigma_u} = \frac{\sigma_u}{\sigma_X} \times \rho_{Xu}, \tag{3}$$

where $\rho_{Xu} = \frac{\sigma_{Xu}}{\sigma_X \sigma_u}$.

- If Assumption 1 is correct, then $\rho_{Xu} = 0$, but if **not** we have the **omitted variable bias formula**:

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \frac{\sigma_u}{\sigma_X} \times \rho_{Xu}, \tag{4}$$

- **Conclusion:** This formula shows that the OLS estimator $\hat{\beta}_1$ is biased and inconsistent.

How to avoid the omitted variable bias?

- Three possible ways to overcome omitted variable bias:
    1. Use a regression in which the omitted variable is no longer omitted: run a multiple regression.
    2. Run a randomized controlled experiment in which treatment is randomly assigned.
    3. Include some variables to the regression that can control for the omitted variable bias, i.e., includes control variables to the regression.
- To introduce the first and the third approach, we need to introduce the multiple linear regression model.

The population multiple regression model

- Consider the case of two regressors:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \quad i = 1, 2, \ldots, n, \tag{5}$$

where

- Y is the dependent variable,
- X_1 and X_2 are the two independent variables (regressors),
- (Y_i, X_{1i}, X_{2i}) denote the ith observation on Y, X_1, and X_2,
- β_0 is the unknown population intercept,
- β_1 is the effect on Y of a change in X_1, holding X_2 constant,
- β_2 is the effect on Y of a change in X_2, holding X_1 constant,
- u is the regression error (omitted factors).

- Population regression function is:

$$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2. \tag{6}$$

The population multiple regression model

- Consider changing X_1 by ΔX_1 while holding X_2 constant:
  - **Before the change**: $\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.
  - **After the change**: $\mathrm{E}(Y^*) = \beta_0 + \beta_1 (X_1 + \Delta X_1) + \beta_2 X_2$.
- Then, the expected change in Y is $\Delta Y = \mathrm{E}(Y^*) - \mathrm{E}(Y) = \beta_1 \Delta X_1$. This result suggests that

$$\beta_1 = \frac{\Delta Y}{\Delta X_1}, \quad \text{holding } X_2 \text{ constant,}$$
$$\beta_2 = \frac{\Delta Y}{\Delta X_2}, \quad \text{holding } X_1 \text{ constant.}$$

- Thus, β_1 gives the effect of a change in X_1 on the expected change in Y, holding X_2 constant. Similarly, β_2 gives the effect of a change in X_2 on the expected change in Y, holding X_1 constant.
- Alternatively, we can call β_1 the **partial effect** on Y of X_1, holding X_2 constant. Similarly, β_2 is the **partial effect** on Y of X_2, holding X_1 constant.
- Finally, the intercept term β_0 is the expected value of Y when $X_1 = X_2 = 0$.

The OLS estimator in multiple regression

- With two regressors, the OLS estimator solves:

$$\min_{b_0,b_1,b_2} \sum_{i=1}^{n} \left[Y_i - (b_0 + b_1 X_{1i} + b_2 X_{2i})\right]^2 .$$

- The OLS estimator minimizes the sum of the squared residuals.
- This minimization problem is solved using calculus.
- This yields the OLS estimators of β_0 ,β_1 and β_2.

Application to test score and the student-teacher ratio

- In the test score example, an omitted student characteristic is the percentage of students who are still learning English in the districts. These students are largely located in the immigrant communities.
- These communities are relatively poor and thus have smaller school budgets and higher student-teacher ratio.
- Therefore, the percentage of English learners (PctEL) is correlated with the student–teacher ratio, the first condition for omitted variable bias holds.
- It is plausible that students who are still learning English will do worse on standardized tests than native English speakers, in which case the percentage of English learners is a determinant of test scores and the second condition for omitted variable bias holds.

Application to test score and the student-teacher ratio

- Using caschools.xlsx, we will consider the following linear model.

$$\text{TestScore} = \beta_0 + \beta_1 \times \text{STR} + \beta_2 \times \text{PctEL} + u \tag{7}$$

Listing 1: Using R

```
# Regression Models----
install.packages("stargazer")
library(stargazer)
install.packages("readxl")
library(readxl)

# Import data
mydata<-read.table("caschool.csv", sep=",", header=TRUE)
# mydata=read_excel("caschool.xlsx",col_names=T,skip=0)
head(mydata)

# OLS estimates----
results1<-lm(testscr~str,data=mydata)
results2<-lm(testscr~str+el_pct,data=mydata)
# Print results
stargazer(results1,results2, type="text", title="OLS Estimation Results")
```

Application to test score and the student-teacher ratio

Listing 2: Using Python

```
# Import data
mydata=pd.read_excel("caschool.xlsx")
# Column names
mydata.columns

# Estimation
model1=smf.ols(formula='testscr~str',data=mydata)
result1=model1.fit()
model2=smf.ols(formula='testscr~str+el_pct',data=mydata)
result2=model2.fit()

# Print results in table form
models=['Model 1', 'Model 2']
results_table=summary_col(results=[result1,result2],
                          float_format='%0.3f',
                          stars=True,
                          model_names=models)
results_table
```

Application to test score and the student-teacher ratio

- According to results in Table 1, the estimated OLS regression equations are

$$\widehat{TestScore} = 698.9 - 2.280 \times STR,$$
$$\widehat{TestScore} = 686 - 1.10 \times STR - 0.65 \times PctEL.$$

Table 1: OLS Estimation Results

	Dependent variable:	
	testscr	
	(1)	(2)
str	-2.280^{***}	-1.101^{***}
	(0.480)	(0.380)
el_pct		-0.650^{***}
		(0.039)
Constant	698.933^{***}	686.032^{***}
	(9.467)	(7.411)
Observations	420	420
R^2	0.051	0.426
Adjusted R^2	0.049	0.424
Residual Std. Error	18.581 (df = 418)	14.464 (df = 417)
F Statistic	22.575^{***} (df = 1; 418)	155.014^{***} (df = 2; 417)
Note:		*p<0.1; **p<0.05; ***p<0.01

Application to test score and the student-teacher ratio

- According to results in Table 1, the estimated OLS regression equations are

$$\text{SLRM}: \widehat{TestScore} = 698.9 - 2.280 \times STR,$$

$$\text{MLRM}: \widehat{TestScore} = 686 - 1.10 \times STR - 0.65 \times PctEL.$$

- Notice here that when PctEL is a part of the error term in the SLRM, it is negatively related to Test Score.
- Hence, ρ_{Xu} is negative here and there is downward bias on the OLSE in the SLRM.
- The estimate $\hat{\beta}_1 = -1.10$ in MLRM indicates that when STR increases by one, the test score on average declines by 1.10 points, holding PctEL constant.
- The estimate $\hat{\beta}_2 = -0.65$ indicates that a one percent increase in PctEL reduces the test score on average by 0.65, holding STR constant.

Measure of fit in multiple regression

- The predicted values are $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i}$ for $i = 1, 2, \ldots, n$.
- The residuals are $\hat{u}_i = Y_i - \hat{Y}_i$. Thus

$$\text{Actual} = \text{Predicted} + \text{Residual}$$
$$\implies Y_i = \hat{Y}_i + \hat{u}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \hat{u}_i.$$

- The standard error of the regression (SER) is the standard deviation of residuals:

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

  where SSR is the sum of squared residuals $SSR = \sum_{i=1}^{n} \hat{u}_i^2$ and k is the number of regressors.
- The root mean square errors: $RMSE = \sqrt{\frac{1}{n} SSR}$.
- Both SER and RMSE are measures of the spread of the distribution of Y around the regression line.

Measure of fit in multiple regression

- The R^2 is the fraction of the variance explained - same definition as in regression with a single regressor:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS},$$

  where
  1. Explained sum of squares: $ESS = \sum_{i=1}^{n} \left(\hat{Y}_i - \bar{\hat{Y}}\right)^2$,
  2. Sum of squared residuals: $SSR = \sum_{i=1}^{n} \hat{u}_i^2$,
  3. Total sum of squares: $TSS = \sum_{i=1}^{n} \left(Y_i - \bar{Y}\right)^2$.

- In general, the SSR will decrease whenever a new regressor is added to the regression.

- But this means that R^2 generally increases (and never decreases) when a new regressor is added - a bit of a problem for a measure of "fit".

- Because the R^2 increases when a new variable is added, an increase in the R^2 does not mean that adding a variable actually improves the fit of the model.

- In this sense, the R^2 gives an inflated estimate of how well the regression fits the data.

Measure of fit in multiple regression

- The $\bar{R}^2$ (the adjusted R^2) corrects this problem by "penalizing" you for including another regressor. This measure does not necessarily increase when you add another regressor.

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \times \frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2}.$$

- Note that $\bar{R}^2 < R^2$, however if n is large the two will be very close.
- Application to test score:
  1. Table 1 shows that including PctEL in the regression increases the R^2 from 0.051 to 0.426.
  2. Thus, including PctEL substantially improves the fit of the regression.
  3. Because n is large and we have only two regressors, the difference between R^2 and $\bar{R}^2$ is very small: $R^2 = 0.426$ vs. $\bar{R}^2 = 0.424$.
  4. The SER value 18.6 falls to 14.5 when PctEL is included as a second regressor.

The sampling distribution of OLS estimators

- Consider the following multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i, \quad i = 1, 2, \ldots, n. \tag{8}$$

- We will consider this model under the following assumptions:

Assumption 1

The conditional distribution of u given $X_{1i}, \ldots, X_{ik}$ has mean zero, that is, $\mathrm{E}\left(u_i | X_{1i}, X_{i2}, \ldots, X_{ik}\right) = 0$ for $i = 1, 2, \ldots, n$.

Assumption 2

$(X_{1i}, X_{i2}, \ldots, X_{ik}, Y_i)$, $i = 1, 2, \ldots, n$, are i.i.d. draws from their joint distribution.

Assumption 3

Large outliers are unlikely: $X_1, \ldots, X_k$, and Y have finite fourth moments.

Assumption 4

There is no perfect multicollinearity.

The sampling distribution of OLS estimators

- Assumption 1 has the same interpretation as in regression with a single regressor.
- Failure of this condition leads to omitted variable bias, specifically, if an omitted variable
  1. belongs in the equation (so is in u) and
  2. is correlated with an included X then this condition fails and there is omitted variable bias.
- Assumption 2 is satisfied automatically if the data are collected by simple random sampling.
- Assumption 3 is the same assumption as we had before for a single regressor. As in the case of a single regressor, OLS can be sensitive to large outliers, so you need to check your data (scatterplots!) to make sure there are no crazy values (typos or coding errors).
- Perfect multicollinearity in Assumption 4 is when one of the regressors is an exact linear function of the other regressors.

The sampling distribution of OLS estimators

- Under Assumptions 1–4, we have the following properties:
  1. The sampling distribution of $\hat{\beta}_1$ has mean β_1, i.e., $\mathrm{E}(\hat{\beta}_1) = \beta_1$.
  2. var$(\hat{\beta}_1)$ is inversely proportional to n.
  3. When n is large, we have
     1. $\hat{\beta}_1$ is consistent: $\hat{\beta}_1 \xrightarrow{p} \beta_1$ (law of large numbers), and
     2. $\frac{\hat{\beta}_1 - \beta_1}{\sqrt{\text{var}(\hat{\beta}_1)}}$ is approximately distributed $N(0, 1)$ (CLT)
  4. All these results hold for $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$.
- Conceptually, there is nothing new here!

Multicollinearity

- Perfect multicollinearity is when one of the regressors is an exact linear function of the other regressors.
- Suppose there are multiple categories and every observation falls in one and only one category (Freshmen, Sophomores, Juniors, Seniors, Other).
- If you include all these dummy variables and a constant, you will have perfect multicollinearity. This is sometimes called the dummy variable trap.
- Solutions to the dummy variable trap:
  1. Omit one of the groups (e.g. Senior), or
  2. Omit the intercept
- **Imperfect multicollinearity** occurs when two or more regressors are very highly correlated.
  1. Imperfect multicollinearity (correctly) results in large standard errors for one or more of the OLS coefficients.

Dummy variable trap

- To illustrate the dummy variable trap, we will use a data set from the textbook.
- This data set comes from a survey known as "Current Population Survey" (CPS), done by the Bureau of Labor Statistics in the U.S. Department of Labor.
- The data set provides data on labor force characteristics of the population, including the level of employment, unemployment, and earnings. Approximately 60,000 randomly selected U.S. households are surveyed each month.
- These data are from the March 2016 survey. The file `ch8_cps.xlsx` contains the following variables.

Variable	Description
`Female`	1 if female; 0 if male
`Age`	Age (in Years)
`Ahe`	Average Hourly Earnings in 2015
`Yrseduc`	Years of Education
`Northeast`	1 if from the Northeast, 0 otherwise
`Midwest`	1 if from the Midwest, 0 otherwise
`South`	1 if from the South, 0 otherwise
`West`	1 if from the West, 0 otherwise

- There are four regional dummies: northeast, midwest, south and west. Observations are collected from these regions.

Dummy variable trap

- Consider the following interaction with R.

```
# Dummy Variable Trap----
install.packages("stargazer")
library(stargazer)
install.packages("readxl")
library(readxl)

# Import data
mydata=read_excel("ch8_cps.xlsx",col_names=T,skip=0)

# Summary Statistics
stargazer(as.data.frame(mydata),type="text",title="Summary Statistics")

# Regional dummay variables sum to one: dummy variable trap
mydata$northeast+mydata$midwest+mydata$south+mydata$west+mydata$midwest

# Run Regression----
result=lm(ahe~yrseduc+female+northeast+midwest+south+west,data=mydata)
stargazer(result,type = "latex", title = "Estimation Results")
```

Dummy variable trap

- The following table includes the summary statistics on the variables.

Table 2: Summary Statistics

Statistic	N	Mean	St. Dev.	Min	Max
ahe	59,668	25.214	17.289	2.000	228.365
yrseduc	59,668	14.094	2.598	6	20
female	59,668	0.439	0.496	0	1
age	59,668	42.885	12.584	15	85
northeast	59,668	0.157	0.364	0	1
midwest	59,668	0.195	0.396	0	1
south	59,668	0.370	0.483	0	1
west	59,668	0.278	0.448	0	1

- We will consider the following regression model:

$$\begin{aligned} \text{Ahe} &= \beta_0 + \beta_1 \text{Yrseduc} + \beta_2 \text{female} + \beta_3 \text{Northeast} \\ &+ \beta_4 \text{Midwest} + \beta_5 \text{South} + \beta_6 \text{West} + u \end{aligned}$$

- This regression suffers from the dummy variable trap since

$$\text{Northeast} + \text{Midwest} + \text{South} + \text{West} = 1.$$

- If you run this regression, R will drop the last dummy variable. That is, the last dummy variable will be the base category.

Table 3: Estimation Results

	Dependent variable:
	ahe
yrseduc	2.981***
	(0.024)
female	−6.367***
	(0.126)
northeast	1.105***
	(0.197)
midwest	−1.382***
	(0.184)
south	−1.108***
	(0.156)
west	
Constant	−13.501***
	(0.356)
Observations	59,668
R^2	0.226
Adjusted R^2	0.226
Residual Std. Error	15.214 (df = 59662)
F Statistic	3,478.146*** (df = 5; 59662)
Note:	*p<0.1; **p<0.05; ***p<0.01

Dummy variable trap

- In Table 3, R dropped the `west` variable to avoid the dummy variable trap.
- How should we interpret the coefficients on the regional dummies?
- The coefficient on the dummy variable shows the average difference in the dependent variable relative to the base category.
- For example, the estimated coefficient on northeast is 1.105: Holding the other variables constant, a person living in the northeast on average earns 1.105 more dollars hourly than a person living in the west.
- How should we interpret the coefficients on `female` and `yrseduc`?
- The coefficient on female is -6.367: Holding other variables constant, women on average earn 6.367 fewer dollars hourly than men.
- The coefficient on yrseduc is 2.981: When yrseduc increases by one year, the average hourly earnings increase by 2.981 dollars, holding other variables constant.

OLS in matrix form

- Consider the following multiple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i,\ i = 1, 2, \ldots, n. \tag{9}$$

- Define the following vectors and matrices:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix},$$

where

1. $\boldsymbol{Y}$ is the $n \times 1$ dimensional vector of n observations on the dependent variable,
2. $\boldsymbol{X}$ is the $n \times (k+1)$ dimensional matrix of n observations on the $k+1$ regressors (including the “constant” regressor for the intercept),
3. $\boldsymbol{U}$ is the $n \times 1$ dimensional vector of the n error terms, and
4. $\boldsymbol{\beta}$ is the $(k+1) \times 1$ dimensional vector of the $k+1$ unknown regression coefficients.

OLS in matrix form

- We can also express $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1^{'} \\ \boldsymbol{X}_2^{'} \\ \vdots \\ \boldsymbol{X}_n^{'} \end{pmatrix},$$

where $\boldsymbol{X}_i = (1, X_{1i}, \cdots, X_{ki})^{'}$ is the $(k+1) \times 1$ dimensional column vector for $i = 1, 2, \ldots, n$.

- Alternatively, we can express $\boldsymbol{X}$ as

$$\boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix} = \left(\boldsymbol{l}, \boldsymbol{X}_1, \boldsymbol{X}_2, \ldots, \boldsymbol{X}_k\right),$$

where $\boldsymbol{l}$ is the $n \times 1$ vector of ones and $\boldsymbol{X}_j$ is the jth column of $\boldsymbol{X}$ for $j = 1, 2, \ldots, k$.

OLS in matrix form

- Using $\boldsymbol{X}_i$ and $\boldsymbol{\beta}$, we can express the model for the ith observation as

$$Y_i = \boldsymbol{X}_i^{'} \boldsymbol{\beta} + u_i,\ i = 1, \ldots, n. \tag{10}$$

- Stacking all n observations in (10) yields the multiple regression model in matrix form:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{U}. \tag{11}$$

OLS in matrix form

- The OLS estimator is defined by

$$\widehat{\boldsymbol{\beta}} = \operatorname{argmin}_{\boldsymbol{b}} \sum_{i=1}^{n} (Y_i - \boldsymbol{X}_i^{'}\boldsymbol{b})^2 = \operatorname{argmin}_{\boldsymbol{b}} (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^{'}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}). \quad (12)$$

- Let $\mathcal{S}(\boldsymbol{b})$ denote $(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})^{'}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b})$.

- The first-order condition for a minimum is

$$\frac{\partial \mathcal{S}(\boldsymbol{b})}{\partial \boldsymbol{b}} = -2\boldsymbol{X}^{'}\boldsymbol{Y} + 2\boldsymbol{X}^{'}\boldsymbol{X}\boldsymbol{b} = 0. \quad (13)$$

- Then, the solution from (13) follows given $\boldsymbol{X}$ is full column rank, i.e., $\boldsymbol{X}^{'}\boldsymbol{X}$ is invertible,

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{Y}. \quad (14)$$

Two algebraic properties of the OLS estimator

(1) $\boldsymbol{X}$ is orthogonal to the residuals by construction, i.e., the inner product of every column of $\boldsymbol{X}$ and $\widehat{\boldsymbol{U}}$ is zero (they form a right angle):

$$\begin{aligned} \boldsymbol{X}^{'}\widehat{\boldsymbol{U}} &= \boldsymbol{X}^{'}(\boldsymbol{Y}-\boldsymbol{X}\widehat{\boldsymbol{\beta}}) = \boldsymbol{X}^{'}\boldsymbol{Y}-\boldsymbol{X}^{'}\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}^{'}\boldsymbol{Y}-\boldsymbol{X}^{'}\boldsymbol{X}(\mathbf{X}^{'}\boldsymbol{X})^{-1}\boldsymbol{X}^{'}\boldsymbol{Y} \\ &= \boldsymbol{X}^{'}\boldsymbol{Y}-\boldsymbol{X}^{'}\boldsymbol{Y} = \mathbf{0}. \end{aligned} \tag{15}$$

(2) The least-squares residuals sum to zero. From (15),

$$\boldsymbol{X}^{'}\widehat{\boldsymbol{U}} = (\boldsymbol{l}, \boldsymbol{X}_1, \ldots, \boldsymbol{X}_k)^{'}\widehat{\boldsymbol{U}} = \begin{pmatrix} \boldsymbol{l}^{'}\widehat{\boldsymbol{U}} \\ \boldsymbol{X}_1^{'}\widehat{\boldsymbol{U}} \\ \vdots \\ \boldsymbol{X}_k^{'}\widehat{\boldsymbol{U}} \end{pmatrix} = \mathbf{0}.$$

- Notice that $\boldsymbol{l}^{'}\widehat{\boldsymbol{U}} = \sum_{i=1}^{n}\hat{u}_i$ is the sum of least squares residuals and equals zero.

Bibliography I

Hanck, Christoph et al. (2021). *Introduction to Econometrics with R*. URL: https://www.econometrics-with-r.org/index.html.

Stock, James H. and Mark W. Watson (2020). *Introduction to Econometrics*. Fourth. Pearson.