

Chapter 12

Ordinary Least Squares

In chapter 11 we imposed only mild regularity conditions on the data. To further investigate linear least squares estimation and provide additional interpretation of the estimated coefficients, we need to make more assumptions. The purpose of this chapter is to describe the properties of linear least squares estimation under the classical OLS assumptions, where OLS stands for **ordinary least squares**.

12.1 Estimation under OLS

In this section we introduce the OLS assumptions and study the finite sample properties of the OLS estimators under these assumptions.

12.1.1 Assumptions

Let $\mathbf{X}$ and $\mathbf{y}$ be as defined in §11.1. The OLS assumptions are as follows:

Assumption 12.1.1 (Full rank) $\mathbf{X}$ has full column rank with probability one.

Assumption 12.1.2 (Linearity) The observations satisfy $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ for some unknown K -vector of parameters $\boldsymbol{\beta}$ and unobservable vector of shocks $\mathbf{u}$.

Assumption 12.1.3 (Orthogonal regressors) $\mathbb{E}[\mathbf{u} | \mathbf{X}] = \mathbf{0}$.

Assumption 12.1.4 (Spherical errors) $\mathbb{E}[\mathbf{u}\mathbf{u}^\top | \mathbf{X}] = \sigma^2\mathbf{I}$ for some unknown $\sigma > 0$.

Let's break these assumptions down and see what they imply.

Assumption 12.1.1 is just a repetition of assumption 11.1.1 from chapter 11. Assumption 12.1.2 can be decomposed into the separate equations

$$y_n = \mathbf{x}_n^\top \boldsymbol{\beta} + u_n, \quad n = 1, \dots, N \quad (12.1)$$

Here are some examples of models that produce relationships in the form of (12.1).

Example 12.1.1 The **Cobb–Douglas production function** relates capital and labor inputs with output via $y = Ak^a \ell^\delta$, where A is a random, firm-specific productivity term and a and δ are parameters. Taking logs yields the linear regression model

$$\ln y = \gamma + a \ln k + \delta \ln \ell + u$$

where the random term $\ln A$ is represented by $\gamma + u$.

Example 12.1.2 The **gravity model** relates international trade flows between country ℓ and country n via the equation $T_{\ell n} = \lambda \xi_{\ell n} G_\ell^\alpha G_n^\beta / D_{\ell n}^\gamma$. Here $T_{\ell n}$ is exports from country ℓ to country n , λ is a constant term, $\xi_{\ell n}$ is a shock, G_ℓ and G_n are GDP in country ℓ and n respectively, and $D_{\ell n}$ is distance between them. Taking logs gives

$$\ln T_{\ell n} = \ln \lambda + \alpha \ln G_\ell + \beta \ln G_n - \gamma \ln D_{\ell n} + \ln \xi_{\ell n} \quad (12.2)$$

For motivation and discussion, see, for example, Bergeijk and Brakman (2010).

Fact 12.1.1 If the linearity assumption 12.1.2 holds, then

- (i) $\mathbf{My} = \mathbf{Mu}$.
- (ii) $\mathbf{Py} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Pu}$.
- (iii) $\text{RSS} = \mathbf{u}^\top \mathbf{Mu}$.

The notation in fact 12.1.1 follows §11.1.2.2. Exercise 12.4.2 requests a proof.

Fact 12.1.2 If assumption 12.1.3 holds, then

- (i) $\mathbb{E}[\mathbf{u}] = \mathbf{0}$,
- (ii) $\mathbb{E}[u_m | x_{nk}] = 0$ for any m, n, k ,
- (iii) $\mathbb{E}[u_m x_{nk}] = 0$ for any m, n, k (orthogonality), and
- (iv) $\text{cov}[u_m, x_{nk}] = 0$ for any m, n, k .

Exercise 12.4.3 asks you to give proofs.

Fact 12.1.3 If assumption 12.1.4 holds, then

- (i) $\text{var}[\mathbf{u}] = \mathbb{E}[\mathbf{uu}^\top] = \sigma^2 \mathbf{I}$,

- (ii) $\mathbb{E}[u_i^2 | \mathbf{X}] = \mathbb{E}[u_j^2 | \mathbf{X}] = \sigma^2$ for any i, j in $1, \dots, N$, and
- (iii) $\mathbb{E}[u_i u_j | \mathbf{X}] = 0$ whenever $i \neq j$.

Parts (ii) and (iii) are called **homoskedasticity** and zero correlation respectively. Combining assumptions 12.1.3 and 12.1.4 gives

$$\text{var}[\mathbf{u} | \mathbf{X}] := \mathbb{E}[\mathbf{u}\mathbf{u}^\top | \mathbf{X}] - \mathbb{E}[\mathbf{u} | \mathbf{X}]\mathbb{E}[\mathbf{u}^\top | \mathbf{X}] = \mathbb{E}[\mathbf{u}\mathbf{u}^\top | \mathbf{X}]$$

Thus the conditional variance–covariance matrix is the diagonal matrix $\sigma^2 \mathbf{I}$.

12.1.2 The OLS Estimators

The standard estimator of β in (12.1) is the least squares estimator $\hat{\beta}$ defined in (11.9). That is,

$$\hat{\beta} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} \quad (12.3)$$

Unlike chapter 11, it's now understood as an estimator of the unknown parameter vector β . In this context, $\hat{\beta}$ is also called the **OLS estimator** of β . Substituting $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ into (12.3) and cancel terms gives the useful comparison

$$\hat{\beta} = \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \quad (12.4)$$

The usual OLS estimator of the parameter σ^2 introduced in assumption 12.1.4 is

$$\hat{\sigma}^2 := \frac{\text{RSS}}{N - K} \quad (12.5)$$

While we can compute $\hat{\beta}$ and $\hat{\sigma}^2$ directly using the expressions in (12.3) and (12.5), in practice it's more common to pass them to a routine in a regression package. Here's an example that runs the gravity model regression (12.2) on world trade data using Python. Like other worked examples, the full set of code and data for replicating these results can be found at johnstachurski.net/emet.html, along with references, details on the data source and the definitions of the variables.

First we import some Python libraries often used for statistics:

```
import pandas as pd
import statsmodels.formula.api as smf
```

Then we read in the data to a pandas DataFrame from a local CSV file:

```
data = pd.read_csv("../data/gravity_dataset_2013.csv")
```

Next we build a model using a formula to indicate the regression we want to run symbolically (similar to R).

```
formula = "log(value) ~ log(egdp) + log(igdp) + log(dist)"
model = smf.ols(formula, data)
```

The formula used above is analogous to (12.2). We omit the shock (the last term in (12.2)) because it is not part of the data set, as well as the constant term, which is included automatically. The names `egdp` and `igdp` represent exporter and importer GDP respectively.

Now we can perform the estimation and print a table summarizing results:

```
result = model.fit(cov_type='HC1')
print(result.summary())
```

The output is as follows:

```

=====
                        OLS Regression Results
=====
Dep. Variable:          log(value)      R-squared:                0.652
Model:                  OLS             Adj. R-squared:           0.652
Method:                 Least Squares    F-statistic:             1.203e+04
Date:                  Wed, 04 Nov 2015  Prob (F-statistic):      0.00
Time:                  16:02:42          Log-Likelihood:          -47185.
No. Observations:      19655            AIC:                    9.438e+04
Df Residuals:          19651            BIC:                    9.441e+04
Df Model:               3
Covariance Type:       HC1
=====

```

	coef	std err	z	P> z	[95.0% Conf. Int.]
Intercept	-30.2350	0.394	-76.773	0.000	-31.007 -29.463
log(egdp)	1.2783	0.008	153.772	0.000	1.262 1.295
log(igdp)	1.0287	0.009	118.885	0.000	1.012 1.046
log(dist)	-1.3483	0.023	-58.113	0.000	-1.394 -1.303

```

=====
Omnibus:                1819.031    Durbin-Watson:           1.798
Prob(Omnibus):           0.000      Jarque-Bera (JB):        3361.159
Skew:                   -0.639      Prob(JB):                0.00
Kurtosis:                4.571      Cond. No.:               708.
=====

```

The value of the OLS estimate $\hat{\beta}$ is the column under `coef`. Objects such as standard errors are discussed below. The argument `cov_type='HC1'` to the `model.fit` call

that produced this table requests heteroskedasticity-consistent or “robust” standard errors, which will be discussed in §12.2.1.

Let’s check that the values for $\hat{\beta}$ reported under `coef` coincide with the expression for $\hat{\beta}$ given in (12.3). First we build the design matrix $\mathbf{X}$ (see the URL given above for details) and then compute as follows:

```
betahat = inv(X.T @ X) @ X.T @ y
```

When we print the result the output agrees with the output in the table

```
[-30.23498073   1.27825004   1.02865139  -1.34830012]
```

The expression for the R^2 in the table above is for R_c^2 , as given in (11.31). A verification is provided in the code available at johnstachurski.net/emet.html.

Figure 12.1 shows a **partial regression plot** for the linear model just described. It is produced from the `result` object generated above via the commands

```
import matplotlib.pyplot as plt
import statsmodels.api as sm
fig = plt.figure()
fig = sm.graphics.plot_partregress_grid(result, fig=fig)
```

To understand the figure, consider, for example, the bottom left subplot, which shows the relationship between $\log(\text{igdp})$ and the dependent variable. The data points in the plot correspond to the residuals in (11.26) and (11.27). In particular, the values on the horizontal axis are the residuals from regressing $\log(\text{igdp})$ on all other columns in $\mathbf{X}$, while those on the vertical axis are the residuals from regressing the dependent variable $\log(\text{value})$ on all other columns in $\mathbf{X}$.

The relationship between these residuals measures the change in $\log(\text{value})$ associated with changing $\log(\text{igdp})$ while holding all other variables fixed. The slope of the regression line running through the plot is exactly the coefficient estimate on $\log(\text{igdp})$. For clarification see the discussion surrounding (11.26) and (11.27).

The partial regression plot can be useful both for interpretation and diagnostics. For now, however, let’s look at some general theory concerning properties of the OLS estimators.

12.1.3 Finite Sample Properties

Let’s start with classical results on finite sample properties under the OLS assumptions. These are elegant results, although you should also read §14.2 if you want to view them with proper perspective. Our first result shows that, under the OLS assumptions, the OLS estimators are unbiased.

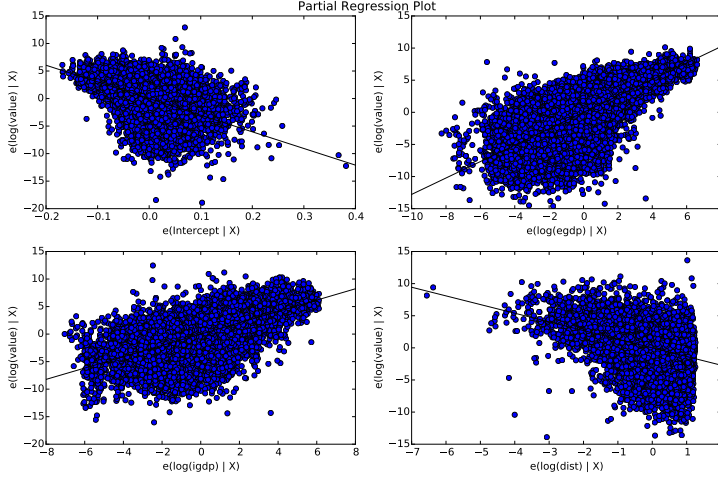


Figure 12.1 Partial regression plot for the gravity model

Theorem 12.1.1 *If assumptions 12.1.1–12.1.3 hold, then*

$$\mathbb{E} \hat{\beta} = \mathbb{E} [\hat{\beta} | \mathbf{X}] = \beta \quad (12.6)$$

If assumption 12.1.4 also holds, then

$$\mathbb{E} \hat{\sigma}^2 = \mathbb{E} [\hat{\sigma}^2 | \mathbf{X}] = \sigma^2 \quad (12.7)$$

Proof. By (12.4) and assumption 12.1.3, we have $\mathbb{E} [\hat{\beta} | \mathbf{X}] = \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E} [\mathbf{u} | \mathbf{X}] = \beta$. Taking the unconditional expectation gives (12.6).

Regarding (12.7), if we use the expression for RSS in fact 12.1.1, we get

$$\mathbb{E} [\text{RSS} | \mathbf{X}] = \mathbb{E} [\mathbf{u}^\top \mathbf{M} \mathbf{u} | \mathbf{X}] = \sigma^2 \mathbb{E} [\boldsymbol{\xi}^\top \mathbf{M} \boldsymbol{\xi} | \mathbf{X}] \quad \text{where} \quad \boldsymbol{\xi} := \sigma^{-1} \mathbf{u}$$

It follows now from fact 5.1.3 on page 127 that $\mathbb{E} [\text{RSS} | \mathbf{X}] = \sigma^2 \text{trace} \mathbf{M}$. Hence

$$\mathbb{E} [\hat{\sigma}^2 | \mathbf{X}] = \frac{\sigma^2 \text{trace}(\mathbf{M})}{N - K}$$

By fact 3.3.4 on page 61 we have $\text{trace} \mathbf{M} = N - K$. Inserting this into the preceding equation and taking unconditional expectations gives (12.7). $\square$

Now that we know $\hat{\beta}$ is unbiased, let's consider the variance.

Theorem 12.1.2 *If assumptions 12.1.1–12.1.4 hold, then*

$$\text{var}[\hat{\beta} | \mathbf{X}] = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1} \quad (12.8)$$

For a proof, see exercise 12.4.5 and its solution.

Is the value on the right-hand side of (12.8) small or large? A famous result in OLS theory is that under the stated conditions, the OLS estimator $\hat{\beta}$ is best linear unbiased. That is, in the class of linear unbiased estimators of β , the OLS estimator has the lowest variance.

Theorem 12.1.3 (Gauss–Markov) *Let assumptions 12.1.1–12.1.4 hold and let $\mathbf{b}$ be an estimator of β . If $\mathbf{b}$ is linear and unbiased, then*

$$\text{var}[\hat{\beta} | \mathbf{X}] \leq \text{var}[\mathbf{b} | \mathbf{X}]$$

Here $\text{var}[\hat{\beta} | \mathbf{X}] \leq \text{var}[\mathbf{b} | \mathbf{X}]$ means that $\text{var}[\mathbf{b} | \mathbf{X}] - \text{var}[\hat{\beta} | \mathbf{X}]$ is nonnegative definite. Since matrices have no standard ordering, this is one way to assert that $\text{var}[\mathbf{b} | \mathbf{X}]$ is “larger.” One implication is that $\text{var}[b_k | \mathbf{X}] \geq \text{var}[\hat{\beta}_k | \mathbf{X}]$ for all k . (Why?)

Also, linearity of $\mathbf{b}$ means that $\mathbf{b}$ is linear as a function of $\mathbf{y}$, taking $\mathbf{X}$ as fixed. In view of theorem 3.1.1 on page 48, this is equivalent to requiring that $\mathbf{b} = \mathbf{C}\mathbf{y}$ for some matrix $\mathbf{C}$. The matrix $\mathbf{C}$ is allowed to depend on $\mathbf{X}$ but not $\mathbf{y}$.

The statement that $\mathbf{b}$ is unbiased means $\mathbb{E}[\mathbf{b} | \mathbf{X}] = \mathbb{E}[\mathbf{C}\mathbf{y} | \mathbf{X}] = \beta$ for all $\beta \in \mathbb{R}^K$.

Proof of theorem 12.1.3. Let $\mathbf{b} = \mathbf{C}\mathbf{y}$, as described above, and let $\mathbf{D} := \mathbf{C} - \mathbf{A}$, where $\mathbf{A} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$. Then

$$\mathbf{b} = \mathbf{C}\mathbf{y} = \mathbf{D}\mathbf{y} + \mathbf{A}\mathbf{y} = \mathbf{D}(\mathbf{X}\beta + \mathbf{u}) + \hat{\beta} = \mathbf{D}\mathbf{X}\beta + \mathbf{D}\mathbf{u} + \hat{\beta} \quad (12.9)$$

Taking conditional expectations and using the fact that $\mathbf{D}$ is a function of $\mathbf{X}$, we obtain

$$\begin{aligned} \mathbb{E}[\mathbf{b} | \mathbf{X}] &= \mathbb{E}[\mathbf{D}\mathbf{X}\beta | \mathbf{X}] + \mathbb{E}[\mathbf{D}\mathbf{u} | \mathbf{X}] + \mathbb{E}[\hat{\beta} | \mathbf{X}] \\ &= \mathbf{D}\mathbf{X} \mathbb{E}[\beta | \mathbf{X}] + \mathbf{D} \mathbb{E}[\mathbf{u} | \mathbf{X}] + \mathbb{E}[\hat{\beta} | \mathbf{X}] \\ &= \mathbf{D}\mathbf{X}\beta + \mathbf{0} + \beta \end{aligned}$$

Since $\mathbf{b}$ is unbiased and, in particular, $\mathbb{E}[\mathbf{b} | \mathbf{X}] = \beta$ for any given β , we have

$$\begin{aligned} \beta &= \mathbf{D}\mathbf{X}\beta + \beta \quad \text{for all } \beta \in \mathbb{R}^K \\ \therefore \quad \mathbf{0} &= \mathbf{D}\mathbf{X}\beta \quad \text{for all } \beta \in \mathbb{R}^K \end{aligned}$$

In light of exercise 3.5.13 on page 71, we conclude that $\mathbf{D}\mathbf{X} = \mathbf{0}$. Combining this result with (12.9), we obtain $\mathbf{b} = \mathbf{D}\mathbf{u} + \hat{\beta}$. Hence $\mathbf{b}$ is equal to the OLS estimator plus

zero-mean noise.

To complete the proof, observe that

$$\text{var}[\mathbf{b} | \mathbf{X}] = \text{var}[\mathbf{D}\mathbf{u} + \hat{\boldsymbol{\beta}} | \mathbf{X}] = \text{var}[(\mathbf{D} + \mathbf{A})\mathbf{u} | \mathbf{X}] = (\mathbf{D} + \mathbf{A}) \text{var}[\mathbf{u} | \mathbf{X}] (\mathbf{D} + \mathbf{A})^\top$$

Using assumption 12.1.4 and fact 3.2.4, the right-hand side of this expression becomes

$$\sigma^2(\mathbf{D} + \mathbf{A})(\mathbf{D}^\top + \mathbf{A}^\top) = \sigma^2(\mathbf{D}\mathbf{D}^\top + \mathbf{D}\mathbf{A}^\top + \mathbf{A}\mathbf{D}^\top + \mathbf{A}\mathbf{A}^\top)$$

Since

$$\mathbf{D}\mathbf{A}^\top = \mathbf{D}\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{0}(\mathbf{X}^\top\mathbf{X})^{-1} = \mathbf{0}$$

and

$$\mathbf{A}\mathbf{A}^\top = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1} = (\mathbf{X}^\top\mathbf{X})^{-1}$$

we conclude that

$$\text{var}[\mathbf{b} | \mathbf{X}] = \sigma^2[\mathbf{D}\mathbf{D}^\top + (\mathbf{X}^\top\mathbf{X})^{-1}] = \sigma^2\mathbf{D}\mathbf{D}^\top + \text{var}[\hat{\boldsymbol{\beta}} | \mathbf{X}]$$

The matrix $\sigma^2\mathbf{D}\mathbf{D}^\top$ is nonnegative definite, so the proof is now done. $\square$

12.1.3.1 Precision of the Estimates

Theorem 12.1.2 shows that the variance–covariance matrix of $\hat{\boldsymbol{\beta}}$ given $\mathbf{X}$ is $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$. The scalar variances of the individual OLS coefficient estimates $\hat{\beta}_1, \dots, \hat{\beta}_K$ are given by the principal diagonal of this matrix. Since each $\hat{\beta}_k$ is unbiased (theorem 12.1.1), small variance means probability mass concentrated around the true parameter β_k . In this case, we say that the estimator has high **precision**.

The Gauss–Markov theorem states that, as far as unbiased linear estimators go, the OLS estimates will have highest precision. But what if we hold the estimation technique fixed, as well as the sample size, and vary the application? Which problems will have high precision estimates, and which will have low precision estimates?

To answer this question, let's focus on the variance of a fixed coefficient β_k . We can write $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \text{col}_k(\mathbf{X})\beta_k + \mathbf{u} \quad (12.10)$$

where $\text{col}_k(\mathbf{X})$ is the vector of observations of the k th regressor, $\mathbf{X}_1$ contains as its columns the observations of the other regressors, and $\hat{\boldsymbol{\beta}}_1$ is the OLS estimates of the corresponding coefficients. By the FWL theorem,

$$\hat{\beta}_k = (\text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \text{col}_k(\mathbf{X}))^{-1} \text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \mathbf{y} \quad (12.11)$$

where $\mathbf{M}_1$ is the residual projection $\mathbf{M}_1 := \mathbf{I} - \mathbf{P}_1$ and $\mathbf{P}_1 := \mathbf{X}_1(\mathbf{X}_1^\top\mathbf{X}_1)^{-1}\mathbf{X}_1^\top$ projects

onto colspace $\mathbf{X}_1$. Applying $\mathbf{M}_1$ to both sides of (12.10) yields

$$\mathbf{M}_1 \mathbf{y} = \mathbf{M}_1 \text{col}_k(\mathbf{X}) \beta_k + \mathbf{M}_1 \mathbf{u}$$

Substituting into (12.11) gives

$$\hat{\beta}_k = \beta_k + (\text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \text{col}_k(\mathbf{X}))^{-1} \text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \mathbf{u} \quad (12.12)$$

Some calculations then show (ex. 12.4.11) that

$$\text{var}[\hat{\beta}_k | \mathbf{X}] = \sigma^2 (\text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \text{col}_k(\mathbf{X}))^{-1} = \sigma^2 \|\mathbf{M}_1 \text{col}_k(\mathbf{X})\|^{-2} \quad (12.13)$$

Thus the variance of $\hat{\beta}_k$ depends on two components, the variance σ^2 of the shock u and the norm of the vector $\mathbf{M}_1 \text{col}_k \mathbf{X}$.

The variance in σ^2 is unavoidable: some data are noisier than others. The larger the variance in the unobservable shock, the harder it will be to estimate β_k with good precision.

The other term is more interesting. $\mathbf{M}_1 \text{col}_k \mathbf{X}$ is the residuals from regressing $\text{col}_k \mathbf{X}$ on $\mathbf{X}_1$, and $\|\mathbf{M}_1 \text{col}_k \mathbf{X}\|$ is the norm of this vector. If this norm is small, then the variance of $\hat{\beta}_k$ will be large.

When will this norm be small? It will be so when $\text{col}_k \mathbf{X}$ is “almost” a linear combination of the other regressors and hence close to colspace $\mathbf{X}_1$. Then

$$\|\mathbf{M}_1 \text{col}_k \mathbf{X}\| = \|\text{col}_k \mathbf{X} - \mathbf{P}_1 \text{col}_k \mathbf{X}\| \approx 0$$

This situation is sometimes referred to as **multicollinearity**.

Figure 12.2 gives an illustration of the effect of multicollinearity on the variance of OLS estimate $\hat{\beta}_2$. In this example $\mathbf{X}$ has two columns. They are related by $\text{col}_2 \mathbf{X} = \delta \text{col}_1 \mathbf{X} + (1 - \delta) \mathbf{z}$, where $\mathbf{z}$ is a vector of N independent draws from the standard normal distribution. Larger δ means more dependence between $\text{col}_1 \mathbf{X}$ and $\text{col}_2 \mathbf{X}$. The true parameter is $\beta_2 = 1$. The figure shows the distribution of $\hat{\beta}_2$ for different values of δ . The estimator $\hat{\beta}_2$ is unbiased but its variance increases with δ . Further details on the simulation can be found at johnstachurski.net/emet.html. (The normal shape of the distribution is explained in the next section.)

12.1.4 Inference with Normal Errors

If we want to compute confidence intervals or test hypotheses about the coefficients in finite samples (i.e., without appealing to asymptotics), we need to strengthen our OLS assumptions even further, by specifying the parametric class of the error vector $\mathbf{u}$. Because of its many attractive properties, the normal distribution is the default.

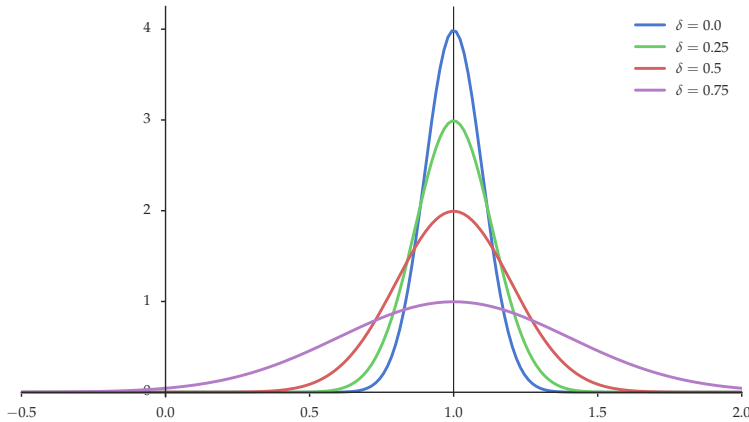


Figure 12.2 Distribution of $\hat{\beta}_2$ when $\text{col}_2 \mathbf{X} = \delta \text{col}_1 \mathbf{X} + (1 - \delta)\mathbf{z}$

Assumption 12.1.5 $\mathbf{X}$ and $\mathbf{u}$ are independent and $\mathcal{L}(\mathbf{u}) = \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

Assumption 12.1.5 implies both assumption 12.1.3 and assumption 12.1.4. It also implies that the conditional distribution of $\hat{\boldsymbol{\beta}}$ given $\mathbf{X}$ is normal, as can be seen from (12.4) and fact 5.1.5 on page 133. The next theorem records this result.

Theorem 12.1.4 *If assumptions 12.1.1–12.1.2 and 12.1.5 hold, then*

$$\mathcal{L}[\hat{\boldsymbol{\beta}} | \mathbf{X}] = \mathcal{N}(\boldsymbol{\beta}, \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1})$$

It follows from theorem 12.1.4 and (5.12) on page 133 that the distribution of individual coefficient $\hat{\beta}_k$ given $\mathbf{X}$ is also normal, with

$$\mathcal{L}[\hat{\beta}_k | \mathbf{X}] = \mathcal{L}[\mathbf{e}_k^\top \hat{\boldsymbol{\beta}} | \mathbf{X}] = \mathcal{N}(\beta_k, \sigma^2 v_k(\mathbf{X})) \quad (12.14)$$

Here $v_k(\mathbf{X}) :=$ the (k, k) th element of $(\mathbf{X}^\top \mathbf{X})^{-1}$.

Theorem 12.1.4 and the normality of the marginal distributions along the individual coefficients is illustrated in figure 12.3. The central plot shows the contours of a density estimate that can be understood as a smoothed histogram based on simulated observations of $\hat{\boldsymbol{\beta}}$ (see §14.1 for details on these density estimates). Each observation $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \hat{\beta}_2)$ is obtained by fitting the regression model $y_n = \beta_1 + \beta_2 x_n + u_n$, where $\beta_1 = 0.5$, $\beta_2 = 1$ and u_n is standard normal. The sample size for each fit is $N = 250$. The univariate distributions on the top and right are univariate density

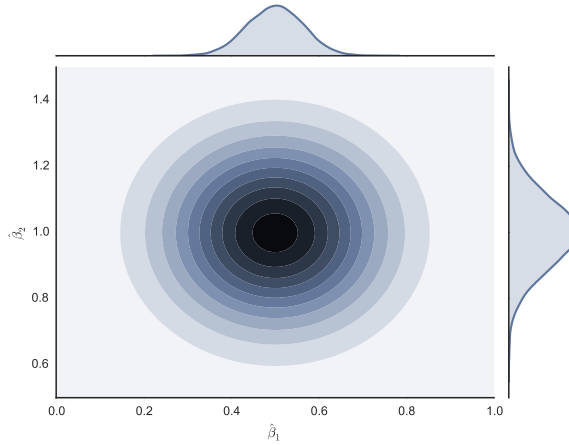


Figure 12.3 Simulated observations of $\hat{\beta}$ and its marginals

estimates based on the observations of $\hat{\beta}_1$ and $\hat{\beta}_2$ respectively.¹

We also want to know the distribution of $\hat{\sigma}^2$, as defined in (12.5). It turns out to be more convenient to work with a transformation of $\hat{\sigma}^2$ defined as

$$Q := \frac{\text{RSS}}{\sigma^2} = (N - K) \frac{\hat{\sigma}^2}{\sigma^2} \quad (12.15)$$

Theorem 12.1.5 *If assumptions 12.1.1–12.1.2 and 12.1.5 hold, then*

$$\mathcal{L}[Q | \mathbf{X}] = \chi^2(N - K)$$

Proof. Recalling fact 12.1.1 on page 324, we have

$$Q = \frac{\mathbf{u}^\top \mathbf{M} \mathbf{u}}{\sigma^2} = (\sigma^{-1} \mathbf{u})^\top \mathbf{M} (\sigma^{-1} \mathbf{u})$$

The claim now follows from fact 5.1.19 on page 141 and the fact that $\text{trace } \mathbf{M} = N - K$. (The last equality was discussed in the proof of theorem 12.1.1.) $\square$

Fact 12.1.4 *If assumptions 12.1.1–12.1.2 and 12.1.5 hold, then the random elements $\hat{\sigma}^2$ and $\hat{\beta}$ are independent given $\mathbf{X}$ (see ex. 12.4.6).*

1. The code for the simulation can be found at johnstachurski.net/emet.html.

12.1.4.1 The t -Test

Let's consider the problem of testing a hypothesis about an individual coefficient β_k . Specifically, we consider the null hypothesis

$$H_0: \beta_k = \beta_k^0$$

where β_k^0 is any number. If we knew σ^2 , we could construct a test of H_0 based around the following observation:

$$z_k := \frac{\hat{\beta}_k - \beta_k}{\sigma \sqrt{v_k(\mathbf{X})}} \implies \mathcal{L}[z_k | \mathbf{X}] = \mathcal{N}(0, 1) \quad (12.16)$$

(The implication follows from (12.14).) Since we don't, the standard methodology is to replace σ^2 with its estimator $\hat{\sigma}^2$, and determine the distribution of the resulting test statistic. Our next result implements this idea. In doing so, we will make use of the following notation:

$$\text{se}(\hat{\beta}_k) := \sqrt{\hat{\sigma}^2 v_k(\mathbf{X})}$$

The term $\text{se}(\hat{\beta}_k)$ is called the **standard error** of $\hat{\beta}_k$. It can be regarded as the sample estimate of the standard deviation of $\hat{\beta}_k$. Replacing this standard deviation with its sample estimate $\text{se}(\hat{\beta}_k)$ and β_k with β_k^0 in (12.16), we obtain the **t -statistic**

$$t_k := \frac{\hat{\beta}_k - \beta_k^0}{\text{se}(\hat{\beta}_k)} \quad (12.17)$$

associated with the hypothesis H_0 . The distribution of this statistic under the null is described in the next theorem.

Theorem 12.1.6 *Let assumptions 12.1.1–12.1.2 and 12.1.5 hold. If H_0 is true, then*

$$\mathcal{L}[t_k | \mathbf{X}] = \text{Student's } t \text{ with } N - K \text{ degrees of freedom}$$

Proof. Taking $\mathbf{X}$ as fixed, we need to show that t_k can be written as

$$t_k = z_k \sqrt{\frac{N - K}{Q}} \quad (12.18)$$

where z_k is standard normal, Q is $\chi^2(N - K)$ and z_k and Q are independent. The desired result then follows from fact 5.1.16 on page 140.

To establish these properties, observe that, under H_0 ,

$$t_k = \frac{\hat{\beta}_k - \beta_k^0}{\sqrt{\hat{\sigma}^2 v_k(\mathbf{X})}} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 v_k(\mathbf{X})}} \sqrt{\frac{\sigma^2}{\hat{\sigma}^2}}$$

In view of the definition of z_k in (12.16), the definition of Q in (12.15) and theorem 12.1.5, it remains only to show that z_k and Q are independent given $\mathbf{X}$. Since we can write z_k as a function of $\hat{\beta}$ and Q as a function of $\hat{\sigma}^2$, this follows from fact 12.1.4. $\square$

Let $T := |t_k|$ and let a desired size α for our test of H_0 be given. In view of (10.18) on page 285, to generate a test of size α , we choose $c = c_\alpha$ to solve $\alpha = \mathbb{P}_\theta\{T > c\}$, or

$$1 - \alpha = \mathbb{P}_\theta\{|t_k| \leq c\}$$

From (4.36) on page 115, the solution is $c_\alpha = F^{-1}(1 - \alpha/2)$, where F is the Student's t CDF with $N - K$ degrees of freedom. In view of example 10.2.7, the corresponding p -value is $2F(-|t_k|)$.

Example 12.1.3 A common implementation of the t -test is the test that a given coefficient is equal to zero. For the k th coefficient β_k , this leads to the statistic

$$t_k := \frac{\hat{\beta}_k}{\text{se}(\hat{\beta}_k)}$$

This statistic is sometimes called the **Z-score**. An application with simulated data is given in listing 8. Running the code shows that the Z-scores calculated by the `zscore` function agree with the “t value” column of the summary table produced by R's `summary` function (the last line of listing 8, output not shown). It is left as an exercise to check that the p -values in the same table agree with the formula $2F(-|t_k|)$ given above.

12.1.4.2 The F -Test

The t -test is used to test hypotheses about individual regressors. For hypotheses concerning multiple regressors, the most common test is the F -test. The F -test can test quite general hypotheses. For simplicity we will focus on null hypotheses that restrict a subset of the coefficients to be zero.

Let $\mathbf{X}_1, \mathbf{X}_2, \hat{\beta}_1, \hat{\beta}_2, \mathbf{P}_1$, and $\mathbf{M}_1$ be as defined in §11.2.2. As our null hypothesis we take

$$H_0: \beta_2 = \mathbf{0}$$

Listing 8 Calculating Z-scores (R)

```

set.seed(1234)
N <- 50; K <- 3
beta <- rep(1, K)
X <- cbind(runif(N), runif(N), runif(N))
u <- rnorm(N)
y <- X %%% beta + u

betahat <- solve(t(X) %%% X) %%% t(X) %%% y
residuals <- y - X %%% betahat
sigmahat <- sqrt(sum(residuals^2) / (N - K))

# Compute t-stat (Z-score) for k-th regressor
zscore <- function(k) {
  se <- sigmahat * sqrt(solve(t(X) %%% X)[k, k])
  return(betahat[k] / se)
}
# Print t-stats
for (k in 1:3) {
  cat("t-stat, k =", k, ":", zscore(k), "\n")
}
# For comparison:
print(summary(lm(y ~ X - 1)))

```

Since

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \quad (12.19)$$

it follows that under the null hypothesis we have

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u} \quad (12.20)$$

Letting

$$\text{USSR} := \|\mathbf{M}\mathbf{y}\|^2 \quad \text{and} \quad \text{RSSR} := \|\mathbf{M}_1\mathbf{y}\|^2$$

be the residual sums of squares for the unrestricted regression (12.19) and restricted regression (12.20) respectively, the standard test statistic for our null hypothesis is

$$F := \frac{(\text{RSSR} - \text{USSR})/K_2}{\text{USSR}/(N - K)} \quad (12.21)$$

Large residuals in the restricted regression (12.20) relative to those in (12.19) result in large values for F , which translates to evidence against the null hypothesis.

Theorem 12.1.7 *Let assumptions 12.1.1–12.1.2 and 12.1.5 hold. If H_0 is true, then, given $\mathbf{X}$, the statistic F defined in (12.21) has the F distribution with parameters $(K_2, N - K)$.*

Proof. Let $Q_1 := (\text{RSSR} - \text{USSR})/\sigma^2$ and let $Q_2 := \text{USSR}/\sigma^2$, so that

$$F = \frac{Q_1/K_2}{Q_2/(N - K)}$$

In view of fact 5.1.17 on page 140, it now suffices to show that, under the null hypothesis,

- (a) Q_1 is chi-squared with K_2 degrees of freedom,
- (b) Q_2 is chi-squared with $N - K$ degrees of freedom, and
- (c) Q_1 and Q_2 are independent.

Part (b) was established in theorem 12.1.5. Regarding part (a), observe that, under the null hypothesis,

- $\text{USSR} = \|\mathbf{M}\mathbf{y}\|^2 = \|\mathbf{M}(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u})\|^2 = \|\mathbf{M}\mathbf{u}\|^2 = \mathbf{u}^\top \mathbf{M}\mathbf{u}$ and
- $\text{RSSR} = \|\mathbf{M}_1\mathbf{y}\|^2 = \|\mathbf{M}_1(\mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u})\|^2 = \|\mathbf{M}_1\mathbf{u}\|^2 = \mathbf{u}^\top \mathbf{M}_1\mathbf{u}$.

It follows that

$$\text{RSSR} - \text{USSR} = \mathbf{u}^\top \mathbf{M}_1\mathbf{u} - \mathbf{u}^\top \mathbf{M}\mathbf{u} = \mathbf{u}^\top (\mathbf{M}_1 - \mathbf{M})\mathbf{u}$$

Using the definitions of $\mathbf{M}$ and $\mathbf{M}_1$, we obtain

$$Q_1 = \frac{\text{RSSR} - \text{USSR}}{\sigma^2} = \frac{\mathbf{u}^\top (\mathbf{I} - \mathbf{P}_1 - \mathbf{I} + \mathbf{P})\mathbf{u}}{\sigma^2} = (\sigma^{-1}\mathbf{u})^\top (\mathbf{P} - \mathbf{P}_1)(\sigma^{-1}\mathbf{u})$$

It is an exercise to show that $(\mathbf{P} - \mathbf{P}_1)$ is symmetric and idempotent.² Hence

$$\text{rank}(\mathbf{P} - \mathbf{P}_1) = \text{trace}(\mathbf{P} - \mathbf{P}_1) = \text{trace } \mathbf{P} - \text{trace } \mathbf{P}_1 = K - K_1 = K_2$$

Via fact 5.1.19, we conclude that $\mathcal{L}(Q_1) = \chi^2(K_2)$, as was to be shown.

To complete the proof, it remains to show that, under the null hypothesis and taking $\mathbf{X}$ as given, Q_1 and Q_2 are independent. To see this, observe that Q_1 is a function of $(\mathbf{P} - \mathbf{P}_1)\mathbf{u}$, while Q_2 is a function of $\mathbf{M}\mathbf{u}$. Hence it suffices to show that $(\mathbf{P} - \mathbf{P}_1)\mathbf{u}$ and $\mathbf{M}\mathbf{u}$ are independent given $\mathbf{X}$. By fact 5.1.12 on page 137, it suffices to show that

2. Hint: See fact 2.2.7 on page 32.

their covariance is zero. To verify this, observe that

$$\text{cov}[(\mathbf{P} - \mathbf{P}_1)\mathbf{u}, \mathbf{M}\mathbf{u} \mid \mathbf{X}] = \mathbb{E}[(\mathbf{P} - \mathbf{P}_1)\mathbf{u}(\mathbf{M}\mathbf{u})^\top \mid \mathbf{X}] = \mathbb{E}[(\mathbf{P} - \mathbf{P}_1)\mathbf{u}\mathbf{u}^\top \mathbf{M} \mid \mathbf{X}]$$

Since $\mathbf{P}$, $\mathbf{P}_1$, and $\mathbf{M}$ are just functions of $\mathbf{X}$, this becomes

$$(\mathbf{P} - \mathbf{P}_1)\mathbb{E}[\mathbf{u}\mathbf{u}^\top \mid \mathbf{X}]\mathbf{M} = \sigma^2(\mathbf{P} - \mathbf{P}_1)\mathbf{M} = \sigma^2(\mathbf{P} - \mathbf{P}_1)(\mathbf{I} - \mathbf{P})$$

Using idempotence and fact 2.2.7, the matrix product on the right is

$$(\mathbf{P} - \mathbf{P}_1)(\mathbf{I} - \mathbf{P}) = \mathbf{P} - \mathbf{P}^2 - \mathbf{P}_1 + \mathbf{P}_1\mathbf{P} = \mathbf{P} - \mathbf{P} - \mathbf{P}_1 + \mathbf{P}_1 = \mathbf{0}$$

This completes the proof of independence, and hence of theorem 12.1.7. $\square$

The most common implementation of the F -test is the test that all coefficients of nonconstant regressors are zero. In this case (12.19) becomes

$$\mathbf{y} = \mathbf{1}\beta_1 + \mathbf{X}_2\beta_2 + \mathbf{u} \quad (12.22)$$

where β_2 is the vector of coefficients corresponding to the nonconstant regressors. Since $\mathbf{X}_1 = \mathbf{1}$, we then have $\mathbf{M}_1 = \mathbf{M}_c$, where the latter is defined in (3.10) on page 61, and hence RSSR is the squared norm of $\mathbf{y} - \bar{y}\mathbf{1}$. An application with simulated data is given in listing 9. If you run the program, you will find that the F statistic calculated by the theoretical formula agrees with the F statistic produced by R's `summary` function.

It is an exercise to show that in the case of (12.22), the F statistic in (12.21) can be rewritten as

$$F = \frac{R_c^2}{1 - R_c^2} \frac{N - K}{K_2} \quad (12.23)$$

where R_c^2 is the centered R squared defined in §11.2.3. Can you provide some intuition as to why large F is evidence against the null?

12.2 Problems and Extensions

The standard OLS assumptions are strict and the results we have obtained are sensitive to their failure. Let's discuss some situations where the OLS assumptions fail and investigate the implications.

12.2.1 Nonspherical Errors

Heteroskedasticity occurs when the variance of the error term is not constant across observations, in contradiction to assumption 12.1.4. In particular, the errors are called

Listing 9 Calculating the F statistic (R)

```

set.seed(1234)
N <- 50; K <- 3
beta <- rep(1, K)
x2 <- runif(N); x3 <- runif(N)
X <- cbind(rep(1, N), x2, x3)
u <- rnorm(N)
y <- X %*% beta + u

betahat <- solve(t(X) %*% X) %*% t(X) %*% y
residuals <- y - X %*% betahat
ussr <- sum(residuals^2)
rssr <- sum((y - mean(y))^2)

Fa <- (rssr - ussr) / 2
Fb <- ussr / (N - K)

cat("F =", Fa / Fb, "\n")          # Print result
print(summary(lm(y ~ x2 + x3)))    # For comparison

```

heteroskedastic if the diagonal terms of $\mathbb{E}[\mathbf{u}\mathbf{u}^\top | \mathbf{X}]$ are not constant. If the off-diagonal terms are nonzero, then the errors are said to have **serial correlation**.

Failure of assumption 12.1.4 does not alone cause bias in the OLS estimator $\hat{\beta}$. This is easily verified by reading the proof of the first part of theorem 12.1.1. Indeed, the proof that $\hat{\beta}$ is unbiased for β does not require assumption 12.1.4. At the same time, the claim that $\hat{\sigma}^2$ is unbiased for σ in the second part of theorem 12.1.1 does not make sense without assumption 12.1.4, and the expression $\text{var}[\hat{\beta} | \mathbf{X}] = \sigma^2(\mathbf{X}^\top \mathbf{X})^{-1}$ from (12.8) is no longer valid. The Gauss–Markov theorem also breaks down. The most important practical consequence is that the results on inference in §12.1.4, which depend on the expression for the variance given above, are no longer valid.

What happens if we replace the assumption $\mathbb{E}[\mathbf{u}\mathbf{u}^\top | \mathbf{X}] = \sigma^2 \mathbf{I}$ with the more general assumption

$$\mathbb{E}[\mathbf{u}\mathbf{u}^\top | \mathbf{X}] = \mathbf{\Omega} \quad (12.24)$$

for some positive definite matrix $\mathbf{\Omega}$? (Assumptions 12.1.1–12.1.3 are maintained.) Using (12.4) on page 325, we can write the variance of the OLS estimator $\hat{\beta}$ under these conditions as

$$\text{var}[\hat{\beta} | \mathbf{X}] = \mathbb{E}[(\hat{\beta} - \beta)(\hat{\beta} - \beta)^\top | \mathbf{X}] = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{\Omega} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \quad (12.25)$$

If Ω is known, then we can estimate under assumptions 12.1.1–12.1.3 and (12.24) using **generalized least squares**. The idea is as follows: In view of fact 3.3.10 on page 67, there exists a nonsingular matrix $\mathbf{C}$ such that $\mathbf{C}^\top \mathbf{C} = \Omega^{-1}$. Using $\mathbf{C}$, we transform the regression model by multiplying both sides of $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$ by $\mathbf{C}$. This gives

$$\mathbf{y}_c = \mathbf{X}_c \beta + \mathbf{u}_c \quad \text{where} \quad \mathbf{y}_c := \mathbf{C}\mathbf{y}, \mathbf{X}_c := \mathbf{C}\mathbf{X} \text{ and } \mathbf{u}_c := \mathbf{C}\mathbf{u}$$

Exercise 12.4.17 asks you to show that

$$\mathbb{E}[\mathbf{u}_c | \mathbf{X}_c] = \mathbf{0} \quad \text{and} \quad \mathbb{E}[\mathbf{u}_c \mathbf{u}_c^\top | \mathbf{X}_c] = \mathbf{I} \quad (12.26)$$

Hence the full set of OLS assumptions 12.1.1–12.1.4 is satisfied for the transformed data. If we use the transformed data to estimate β via $(\mathbf{X}_c^\top \mathbf{X}_c)^{-1} \mathbf{X}_c^\top \mathbf{y}_c$, we recover the best linear unbiased property. Expanding out the definitions of $\mathbf{X}_c$ and $\mathbf{y}_c$ and using $\mathbf{C}^\top \mathbf{C} = \Omega^{-1}$, we can show the new estimator to be

$$\hat{\beta}_{\text{GLS}} := (\mathbf{X}^\top \Omega^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \Omega^{-1} \mathbf{y} \quad (12.27)$$

This estimator is called the **generalized least squares (GLS) estimator** of β .

In practice, Ω is rarely known. While we can estimate it, estimation is difficult unless additional structure is imposed. This is because Ω is $N \times N$, implying that, absent further restrictions, the number of parameters contained in Ω grows *faster* than the sample. Our estimation problem gets harder the more data that we have.

A common way of adding structure is to assume that Ω is diagonal. That is, $\Omega = \text{diag}(\sigma_1^2, \dots, \sigma_N^2)$, where each σ_n is strictly positive and unknown. The diagonal assumption rules out correlation but maintains the possibility of heteroskedasticity. One estimator of Ω is

$$\hat{\Omega} := \text{diag}(\hat{u}_1^2, \dots, \hat{u}_N^2) \quad \text{where} \quad \hat{\mathbf{u}} := \mathbf{M}\mathbf{y}$$

Replacing Ω with $\hat{\Omega}$ in (12.25) leads to the estimate of $\text{var}[\hat{\beta} | \mathbf{X}]$ given by

$$\hat{\mathbf{V}} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \hat{\Omega} \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} \quad (12.28)$$

If we use this expression in place of the usual variance estimate $\hat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}$ when we form our standard errors, we get the **heteroskedasticity-consistent standard errors** proposed by White (1980).

12.2.2 Bias

In §12.2.1 we considered what happens when assumption 12.1.4 (spherical errors) fails. What about the other OLS assumptions? Assumption 12.1.1 (full rank) is fairly

benign. Assumptions 12.1.2 and 12.1.3 (linearity and exogeneity) are more problematic. Let's have a look at how they might fail and the consequences.

The linearity assumption $y_n = \mathbf{x}_n^\top \boldsymbol{\beta} + u_n$ is fundamental. If we replace it with a more general assumption such as $y_n = f(\mathbf{x}_n) + u_n$ for some arbitrary function f , then OLS theory falls apart. For example, we cannot even *ask* if the OLS estimator is unbiased because $\boldsymbol{\beta}$ has vanished. There is no parameter vector for $\hat{\boldsymbol{\beta}}$ to estimate.³

The remaining assumption we need to consider is exogenous regressors. What happens if all other OLS assumptions hold but assumption 12.1.3 fails?

Observe that, by (12.4) on page 325,

$$\mathbb{E}[\hat{\boldsymbol{\beta}} | \mathbf{X}] - \boldsymbol{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbb{E}[\mathbf{u} | \mathbf{X}] \quad (12.29)$$

Without $\mathbb{E}[\mathbf{u} | \mathbf{X}] = \mathbf{0}$, we cannot assert that the right-hand side is zero. Hence $\hat{\boldsymbol{\beta}}$ is biased. Bias due to failure of assumption 12.1.3 is called **endogeneity bias**.

As discussed in several parts of this text, striving for unbiasedness is somewhat misguided. Better estimators can usually be constructed by admitting some degree of bias in order to penalize complexity (thereby acknowledging that the empirical distribution is not the true distribution). However, there is an important difference with the bias in (12.29). Good estimators reduce bias as the sample size increases, so that the object they aim to compute can be recovered asymptotically. This is not true for the OLS estimator when exogeneity fails.

To see this, recall from theorem 11.1.2 on page 304 that $\hat{\boldsymbol{\beta}} \xrightarrow{p} \mathbb{E}[\mathbf{xx}^\top]^{-1} \mathbb{E}[\mathbf{x}y]$ as $N \rightarrow \infty$. The right-hand side is the vector of coefficients in the best linear predictor (see page 147). If $y = \mathbf{x}^\top \boldsymbol{\beta} + u$, then this becomes

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \mathbb{E}[\mathbf{xx}^\top]^{-1} \mathbb{E}[\mathbf{x}(\mathbf{x}^\top \boldsymbol{\beta} + u)] = \boldsymbol{\beta} + \mathbb{E}[\mathbf{xx}^\top]^{-1} \mathbb{E}[\mathbf{x}u]$$

With exogeneity we get $\mathbb{E}[\mathbf{x}u] = \mathbb{E}[\mathbf{x} \mathbb{E}[u | \mathbf{x}]] = \mathbf{0}$ and the OLS estimator is consistent. Without exogeneity we cannot assert the same.

12.2.2.1 Examples of Endogeneity Bias

Let's look at some applications where assumption 12.1.3 fails. First up, consider again the Cobb–Douglas example on page 324, which yields the regression model

$$\ln y_n = \gamma + a \ln k_n + \delta \ln \ell_n + u_n$$

3. It is an interesting habit of economists to quibble endlessly about whether someone's estimators are biased, even when their models are wildly speculative in terms of specific functional forms. (For example, do we really believe that log GDP per capita is a linear function of various measures of institutional quality and an Africa dummy, say, or even that this is a close depiction of reality? If not, what exactly does it mean to say that $\hat{\boldsymbol{\beta}}$ is biased?)

Here y is output, k is capital, ℓ is labor, and subscript n indicates observation on the n th firm. The term u_n is a firm specific productivity shock. A likely problem here is that the firm will choose higher levels of both capital and labor when it anticipates high productivity in the current period. This will lead to endogeneity bias.

To illustrate, suppose that $u_{n,-1}$ is the productivity shock received by firm n last period and this value is observable to the firm. Suppose further that productivity follows a random walk, with $u_n = u_{n,-1} + \eta_n$, where η_n is zero-mean white noise. As a result the firm forecasts period n productivity as $\mathbb{E}[u_n | u_{n,-1}] = u_{n,-1}$. Finally, suppose that the firm increases labor input when productivity is anticipated to be high, with the specific relationship $\ell_n = a + b\mathbb{E}[u_n | u_{n,-1}]$ for $b > 0$. When all shocks are zero-mean, we then have

$$\mathbb{E}[\ell_n u_n] = \mathbb{E}[(a + bu_{n,-1})(u_{n,-1} + \eta_n)] = \mathbb{E}[bu_{n,-1}^2]$$

This term will be strictly positive whenever $u_{n,-1}$ has positive variance. Thus the conditions of fact 12.1.2 (page 324) fail, and therefore assumption 12.1.3 does not hold (because assumption 12.1.3 implies fact 12.1.2).

This source of endogeneity bias in estimating production functions has been discussed many times in the literature. The best solution is better modeling. For one illustration of careful modeling, see Olley and Pakes (1996).

As a second example of endogeneity bias, suppose next that we have in hand data generated according to the simple AR(1) model

$$x_0 = 0 \quad \text{and} \quad x_n = \beta x_{n-1} + u_n \quad \text{for } n = 1, \dots, N \quad (12.30)$$

Let $\{u_n\}_{n=1}^N$ be IID with distribution $N(0, \sigma^2)$. The unknown parameters are β and σ^2 . Letting $\mathbf{y} := (x_1, \dots, x_N)$, $\mathbf{x} := (x_0, \dots, x_{N-1})$, and $\mathbf{u} := (u_1, \dots, u_N)$, we can write the N equations in (12.30) as $\mathbf{y} = \beta \mathbf{x} + \mathbf{u}$. The OLS estimate of β is $\hat{\beta} := (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}$.

By fact 12.1.2, if assumption 12.1.3 holds, then $\mathbb{E}[u_m x_{n+1}] = 0$ for any m and n . In the current setup this equality fails. For example, if $n \geq m$, then we can (ex. 12.4.15) write x_n as

$$x_n = \sum_{j=0}^{n-1} \beta^j u_{n-j} \quad (12.31)$$

Hence $\mathbb{E}[x_n u_m] = \sum_{j=0}^{n-1} \beta^j \mathbb{E}[u_{n-j} u_m] = \beta^{n-m} \sigma^2$. In particular, assumption 12.1.3 fails whenever $\beta \neq 0$.

To illustrate the bias in $\hat{\beta}$, listing 10 generates the data 10,000 times, computes $\hat{\beta}$ on each occasion and histograms the result. The parameters are $N = 25$ and $\beta = 0.9$. The output is figure 12.4. If you modify the code to print the sample mean, it gives around 0.82. Since the number of replications is large, this is close to $\mathbb{E}[\hat{\beta}]$. An asymptotic

95% confidence interval for our estimate is around (0.818, 0.824).

Listing 10 Generates and histograms observations of $\hat{\beta}$ (Python)

```
import numpy as np
import matplotlib.pyplot as plt

N = 25
x = np.zeros(N)
beta = 0.9
num_reps = 10000
betahat_obs = np.empty(num_reps)

for j in range(num_reps):
    u = np.random.randn(N)
    for t in range(N-1):
        x[t+1] = beta * x[t] + u[t+1]
    y = x[1:]          # x_1, ..., x_N
    x_vec = x[:-1]     # x_0, ..., x_{N-1}
    betahat_obs[j] = np.sum(x_vec * y) / np.sum(x_vec**2)

plt.hist(betahat_obs, bins=50, alpha=0.6, normed=True)
plt.show()
```

12.2.3 Instrumental Variables

The discussion in §12.2.2 shows that the OLS estimator is biased and inconsistent when regressors are endogenous. Here we show that a consistent estimator of β can be found when extra information is available in the form of “exogenous” variables called **instruments**. These instruments will be collected into a $N \times J$ matrix $\mathbf{Z}$, where each column is observations on one exogenous variable. We connect $\mathbf{Z}$ and the other variables via the following assumptions:

- (i) $\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$.
- (ii) $\mathbb{E}[\mathbf{u} | \mathbf{Z}] = \mathbf{0}$.
- (iii) $\mathbb{E}[\mathbf{u}\mathbf{u}^T | \mathbf{Z}] = \sigma^2\mathbf{I}$ for some positive constant σ .
- (iv) $\mathbf{Z}^T\mathbf{X}$ has full column rank.

As exercise 12.4.18 asks you to show, assumption (iv) implies that $J \geq K$, which is to say that we have at least as many instruments as regressors. If $J > K$, the model

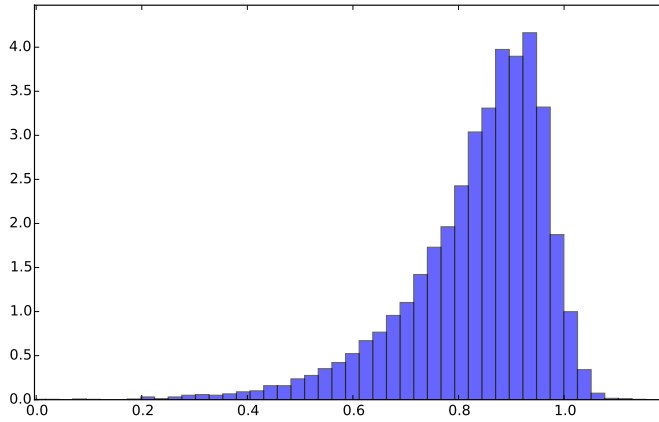


Figure 12.4 Observations of $\hat{\beta}$ when $\beta = 0.9$

is said to be **overidentified**. If $J = K$, the model is called **exactly identified**. (When $J < K$, the model is called **underidentified**.)

The idea of instrumental variables least squares is to multiply $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ by $\mathbf{Z}^\top$ to produce

$$\mathbf{Z}^\top \mathbf{y} = \mathbf{Z}^\top \mathbf{X} \boldsymbol{\beta} + \mathbf{w} \quad \text{where} \quad \mathbf{w} := \mathbf{Z}^\top \mathbf{u} \quad (12.32)$$

Applying (ii) and (iii) from the IV assumptions gives $\mathbb{E}[\mathbf{w} | \mathbf{Z}] = \mathbf{0}$ and $\mathbb{E}[\mathbf{w}\mathbf{w}^\top | \mathbf{Z}] = \sigma^2 \mathbf{Z}^\top \mathbf{Z}$. This suggests that we might apply generalized least squares estimation to the linear equation in (12.32). We can use the expression for the GLS estimator from (12.27) after substituting $\mathbf{X}$ for $\mathbf{Z}^\top \mathbf{X}$, $\mathbf{y}$ for $\mathbf{Z}^\top \mathbf{y}$, and $\boldsymbol{\Omega}$ for $\sigma^2 \mathbf{Z}^\top \mathbf{Z}$. This leads to the **instrumental variable least squares (IVLS) estimator**

$$\hat{\boldsymbol{\beta}}_{\text{IVLS}} := (\mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z} (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{y} \quad (12.33)$$

In the case where $J = K$, this reduces to the simpler expression

$$\hat{\boldsymbol{\beta}}_{\text{IVLS}} := (\mathbf{Z}^\top \mathbf{X})^{-1} \mathbf{Z}^\top \mathbf{y} \quad (12.34)$$

In the preceding discussion, the analogy with GLS was not exact. (Direct application of GLS to (12.32) would require that $\mathbb{E}[\mathbf{w} | \mathbf{Z}^\top \mathbf{X}] = \mathbf{0}$ rather than $\mathbb{E}[\mathbf{w} | \mathbf{Z}] = \mathbf{0}$.) As a consequence we cannot claim unbiasedness and other properties by appealing to the GLS theory. Indeed $\hat{\boldsymbol{\beta}}_{\text{IVLS}}$ is, in general, biased (see ex. 12.4.20).

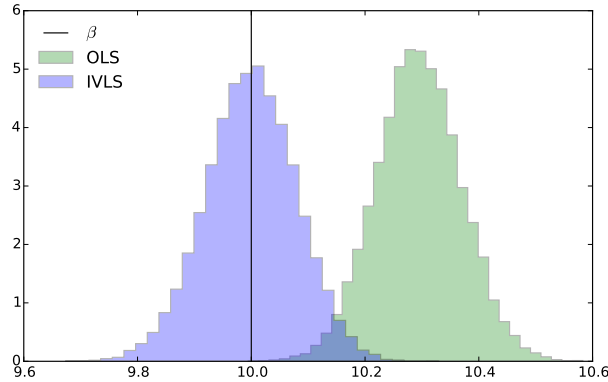


Figure 12.5 Simulated draws of OLS and IVLS estimators of β

At the same time, the bias is typically smaller than OLS and vanishes asymptotically (see, e.g., §8.3 of Davidson and MacKinnon 2004). Figure 12.5 gives an illustration. This figure corresponds to OLS and IVLS estimates of β in $y = x\beta + u$ when $\beta = 10$, z is drawn independently of u , and x is set equal to $\alpha z + (1 - \alpha)u$. Thus z is exogenous and x is a mixture of z and u . The value α is a constant. In the simulation, α is set to 0.8, so z is a relatively strong instrument. The histograms are observations of the two estimators over 25,000 repetitions, each one having a sample size of 250. IV provides a much better estimator of β in this case.⁴

There is another way to derive the expression given for $\hat{\beta}_{\text{IVLS}}$ on the right-hand side of (12.33). It involves a method called **two-stage least squares**. The key idea is to project the endogenous regressors onto the column space of the exogenous instruments before running the regression. Exercise 12.4.21 walks you through the details.

12.2.4 Causality

Causality is a slippery concept. Let's touch on some basic points, particularly concerning the connection between causality and regression. (References to more detailed treatments are given in §12.3.)

You might have heard R^2 interpreted as measuring the “explanatory power” of the regressors in a particular econometric problem. The idea is that regression amounts to decomposing $\mathbf{y}$ into the sum of two orthogonal parts, the fitted values $\mathbf{P}\mathbf{y}$ and the residuals $\mathbf{M}\mathbf{y}$. As in (11.12), the squared norm $\|\mathbf{y}\|^2 =: \text{TSS}$ can then be decomposed

4. For code and full details of the simulation, see johnstachurski.net/emet.html.

as $\|\mathbf{Py}\|^2 =: \text{ESS}$ plus $\|\mathbf{My}\|^2 =: \text{RSS}$. This is sometimes paraphrased as “the total variation in $\mathbf{y}$ is the sum of explained and unexplained variation.” The value of R^2 is then claimed to be the fraction of the variation in $\mathbf{y}$ “explained” by the regressors.

This terminology is problematic because the notion that the regressors “explain” variation in $\mathbf{y}$ suggests causation, and R^2 says nothing about causation *per se*. Instead, R^2 is better thought of as a measure of correlation (see §11.2.3). As has been observed on many occasions, correlation and causation are not the same thing. Two informal examples are as follows:

- We often see car crashes and ambulances together (correlation). This does not imply that ambulances cause crashes.
- Say we observe that people sleeping with their shoes on often wake up with sore heads. One possible explanation is that wearing shoes to bed causes headaches. Another is that both phenomena are caused by too many pints at the pub the night before.

Identifying causality in observational data can be very challenging. Here’s another example, which, although it falls in the realm of traffic safety, is nonetheless interesting because it involves choices and human behavior: It has been observed that motorcycles fitted with ABS brakes have much lower crash rates than those without anti-lock brakes (see, e.g., Teoh 2011). Does that mean fitting ABS to a given motorcycle will reduce the probability that the bike is involved in an accident by a similar amount? If so, should traffic authorities mandate the use of ABS brakes?

There’s an obvious problem in jumping to this conclusion: selection bias in the sample. Cautious motorcyclists are more likely to choose bikes with safety features such as ABS. Lower crash rates for ABS bikes might therefore be caused primarily by more careful riding.

We can eliminate the selection bias by randomly assigning bikes of both types to a large group of riders and observing the results. One might expect that such randomization would reduce the observed effect, although perhaps not eliminate it entirely. In fact such a study has already been conducted for the case of passenger vehicles. In Munich, half of a large fleet of taxis were fitted with anti-lock brakes and half were fitted with conventional brakes (Aschenbrenner and Biehl 1994). The study lasted three years and drivers were randomly assigned across vehicles. A total of 747 accidents were observed. The authors found that accidents rates were slightly *higher* for taxis fitted with ABS.

The conclusion of the authors is that, while anti-lock brakes do improve braking performance, drivers compensate by driving faster, making sharper turns, poorer lane changes, and so forth. (This altered behavior is called “risk compensation.”) The net effect is roughly zero: mandatory ABS brakes will cause no appreciable drop in accident rates.

In general, understanding causality requires either good experiments or good theory. There is no magic econometric procedure that can reliably extract causality from observational data without a good model.

12.3 Further Reading

General econometrics texts treating OLS and its extensions include Hayashi (2000), Kennedy (2008), Greene (2011), Davidson and MacKinnon (2004), Wooldridge (2010), and Cameron and Trivedi (2005). A comprehensive treatment of theory can be found in Ruud (2000).

Regarding robust standard errors, asymptotics of the estimator in (12.28) are provided in White (1980). Closely related is the covariance matrix estimator proposed in Newey and West (1987). See Hayashi (2000), Hansen (2015), and Cameron and Trivedi (2005) for textbook treatments.

Measurement errors form another source of bias in OLS estimators. There is a standard approach to so-called errors-in-variables models based around instrumental variables. See, for example, chapter 8 of Davidson and MacKinnon (2004). For a treatment of nonlinear errors-in-variables, see Chen et al. (2011).

For more discussion of causality, see Freedman (2009), Angrist and Pischke (2009), Pearl (2009), or Imbens and Rubin (2015). An important notion of causality in time series models is the ability of one time series to predict another. A seminal contribution is Granger (1969). See Dufour and Taamouti (2010) for a more recent discussion. Chernozhukov et al. (2013) consider how to perform inference on counterfactual distributions. Belloni et al. (2014) discuss treatment effects in high dimensional settings.

12.4 Exercises

Ex. 12.4.1 Argue that the sample mean of a random sample $y_1, \dots, y_N$ from a given distribution F can be viewed as a least squares estimator of the mean of F .

Ex. 12.4.2 Prove the claims in fact 12.1.1 on page 324.

Ex. 12.4.3 Prove the claims in fact 12.1.2.

Ex. 12.4.4 Confirm part (i) of fact 12.1.3.

Ex. 12.4.5 Prove theorem 12.1.2 on page 328.

Ex. 12.4.6 Prove fact 12.1.4 on page 333.⁵

5. Hint: Use fact 5.1.13 on page 137.

Ex. 12.4.7 Show by direct calculation that $\text{cov}[\mathbf{Py}, \mathbf{My} \mid \mathbf{X}] = \mathbf{0}$.

Ex. 12.4.8 Show that under assumptions 12.1.1–12.1.4, we have $\text{cov}[\hat{\beta}, \mathbf{Mu} \mid \mathbf{X}] = \mathbf{0}$.

Ex. 12.4.9 Show that $\mathbb{E}[\mathbf{Py} \mid \mathbf{X}] = \mathbf{X}\beta$ and $\text{var}[\mathbf{Py} \mid \mathbf{X}] = \sigma^2 \mathbf{P}$.

Ex. 12.4.10 Show that $\mathbb{E}[\mathbf{My} \mid \mathbf{X}] = \mathbf{0}$ and $\text{var}[\mathbf{My} \mid \mathbf{X}] = \sigma^2 \mathbf{M}$.

Ex. 12.4.11 Confirm (12.13) on page 331.

Ex. 12.4.12 Using (12.13), show that for the simple OLS model $y = \beta_1 \mathbf{1} + \beta_2 \mathbf{x} + \mathbf{u}$, the variance of $\hat{\beta}_2$ given $\mathbf{x}$ is $\sigma^2 / \sum_{n=1}^N (x_n - \bar{x})^2$.

Ex. 12.4.13 Show that in the case of (12.22), the F statistic in (12.21) can be rewritten as (12.23).

Ex. 12.4.14 Suppose that assumption 12.1.5 holds, so that $\mathbf{X}$ and $\mathbf{u}$ are independent and $\mathcal{L}(\mathbf{u}) = N(\mathbf{0}, \sigma^2 \mathbf{I})$. Show that, conditional on $\mathbf{X}$,

(i) $\mathbf{Py}$ and $\mathbf{My}$ are normally distributed, and

(ii) $\mathbf{Py}$ and $\mathbf{My}$ are independent.

Ex. 12.4.15 Verify the expression for x_n in (12.31).

Ex. 12.4.16 Obtain fact 10.1.1 on page 277 as a special case of theorem 12.1.6.⁶

Ex. 12.4.17 Prove the claims $\mathbb{E}[\mathbf{u}_c \mid \mathbf{X}_c] = \mathbf{0}$ and $\mathbb{E}[\mathbf{u}_c \mathbf{u}_c^\top \mid \mathbf{X}_c] = \mathbf{I}$ from (12.26).

Ex. 12.4.18 In connection with the theory of instrumental variables in §12.2.3, show that if $\mathbf{Z}$ is $N \times J$, $\mathbf{X}$ is $N \times K$ and $\mathbf{Z}^\top \mathbf{X}$ has full column rank, then $J \geq K$.

Ex. 12.4.19 Let all OLS assumptions hold except 12.1.2 (page 323). Suppose that the data-generating process is

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{Z}\theta + \mathbf{u} \quad (12.35)$$

with $\theta \neq \mathbf{0}$ and $\mathbb{E}[\mathbf{u} \mid \mathbf{X}, \mathbf{Z}] = \mathbf{0}$. Suppose further that we mistakenly ignore $\mathbf{Z}$, and simply regress $\mathbf{y}$ on $\mathbf{X}$. Show that the OLS estimator is biased whenever $\mathbf{X}^\top \mathbf{Z} \neq \mathbf{0}$.

Ex. 12.4.20 Show that the assumptions of §12.2.3 are not sufficient for unbiasedness of the IVLS estimator of β .

Ex. 12.4.21 This exercise asks you to construct $\hat{\beta}_{\text{IVLS}}$ in (12.33) via an alternative approach called two-stage least squares. Your task is to show that the estimator produced by the following two steps is identical to the expression given in the right-hand side of (12.33).

(i) Construct $\hat{\mathbf{X}}$ by projecting each column of $\mathbf{X}$ onto colspace $\mathbf{Z}$.

(ii) Regress $\mathbf{y}$ on $\hat{\mathbf{X}}$ using OLS.

6. Hint: Consider the regression model $y = \beta + u$.

12.4.1 Solutions to Selected Exercises

Solution to Ex. 12.4.1. Letting $\mu := \mathbb{E}[y_n]$ we can write $y_n = \mu + u_n$ when $u_n := y_n - \mu$. In other words, $\mathbf{y} = \mathbf{1}\mu + \mathbf{u}$. The OLS estimate of μ is

$$\hat{\mu} := (\mathbf{1}^\top \mathbf{1})^{-1} \mathbf{1}^\top \mathbf{y} = \frac{1}{N} \mathbf{1}^\top \mathbf{y} = \frac{1}{N} \sum_{n=1}^N y_n = \bar{y}_N$$

Reading right to left, the sample mean of $\mathbf{y}$ is the OLS estimate of the mean. □

Solution to Ex. 12.4.2. From fact 3.3.1 on page 61 we have $\mathbf{M}\mathbf{X} = \mathbf{0}$ and $\mathbf{P}\mathbf{X} = \mathbf{X}$. From this (i) and (ii) follow easily. For example,

$$\mathbf{M}\mathbf{y} = \mathbf{M}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) = \mathbf{M}\mathbf{u}$$

Regarding (iii), since $\mathbf{M}$ is symmetric and idempotent (page 61), we have

$$\text{RSS} := \|\mathbf{M}\mathbf{y}\|^2 = \|\mathbf{M}\mathbf{u}\|^2 = (\mathbf{M}\mathbf{u})^\top (\mathbf{M}\mathbf{u}) = \mathbf{u}^\top \mathbf{M}^\top \mathbf{M}\mathbf{u} = \mathbf{u}^\top \mathbf{M}\mathbf{u} \quad \square$$

Solution to Ex. 12.4.3. The respective proofs for claims 1–4 are as follows:

- (i) $\mathbb{E}[\mathbf{u}] = \mathbb{E}[\mathbb{E}[\mathbf{u} | \mathbf{X}]] = \mathbb{E}[\mathbf{0}] = \mathbf{0}$
- (ii) $\mathbb{E}[u_m | x_{nk}] = \mathbb{E}[\mathbb{E}[u_m | \mathbf{X}] | x_{nk}] = \mathbb{E}[\mathbf{0} | x_{nk}] = 0$
- (iii) $\mathbb{E}[u_m x_{nk}] = \mathbb{E}[\mathbb{E}[u_m x_{nk} | x_{nk}]] = \mathbb{E}[x_{nk} \mathbb{E}[u_m | x_{nk}]] = 0$
- (iv) $\text{cov}[u_m, x_{nk}] = \mathbb{E}[u_m x_{nk}] - \mathbb{E}[u_m] \mathbb{E}[x_{nk}] = 0$

□

Solution to Ex. 12.4.4. By definition, $\text{var } \mathbf{u} = \mathbb{E}[\mathbf{u}\mathbf{u}^\top] - \mathbb{E}[\mathbf{u}]\mathbb{E}[\mathbf{u}^\top]$. Since $\mathbb{E}[\mathbf{u}] = \mathbb{E}[\mathbb{E}[\mathbf{u} | \mathbf{X}]] = \mathbf{0}$, this reduces to $\text{var } \mathbf{u} = \mathbb{E}[\mathbf{u}\mathbf{u}^\top]$. Moreover

$$\mathbb{E}[\mathbf{u}\mathbf{u}^\top] = \mathbb{E}[\mathbb{E}[\mathbf{u}\mathbf{u}^\top | \mathbf{X}]] = \mathbb{E}[\sigma^2 \mathbf{I}] = \sigma^2 \mathbf{I} \quad \square$$

Solution to Ex. 12.4.5. If $\mathbf{A} := (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$, then by (12.4) we have $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{A}\mathbf{u}$, and

$$\text{var}[\hat{\boldsymbol{\beta}} | \mathbf{X}] = \text{var}[\boldsymbol{\beta} + \mathbf{A}\mathbf{u} | \mathbf{X}] = \text{var}[\mathbf{A}\mathbf{u} | \mathbf{X}]$$

Treating $\mathbf{A}$ as nonrandom given $\mathbf{X}$ and applying fact 5.1.2 on page 127, we have

$$\text{var}[\mathbf{A}\mathbf{u} | \mathbf{X}] = \mathbf{A} \text{var}[\mathbf{u} | \mathbf{X}] \mathbf{A}^\top = \mathbf{A}(\sigma^2 \mathbf{I}) \mathbf{A}^\top$$

Moreover, $\mathbf{A}(\sigma^2 \mathbf{I}) \mathbf{A}^\top = \sigma^2 \mathbf{A} \mathbf{A}^\top = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{X} (\mathbf{X}^\top \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1}$. Hence

$$\text{var}[\hat{\boldsymbol{\beta}} | \mathbf{X}] = \text{var}[\mathbf{A}\mathbf{u} | \mathbf{X}] = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \quad \square$$

Solution to Ex. 12.4.6. Note first that $\mathbf{Pu}$ and $\mathbf{Mu}$ are independent by fact 5.1.13 on page 137. By fact 5.1.10 on page 137, functions of independent random vectors are likewise independent. Hence it suffices to show that $\hat{\beta}$ and $\hat{\sigma}^2$ are functions of $\mathbf{Pu}$ and $\mathbf{Mu}$ respectively. The case of $\hat{\sigma}^2$ is straightforward. For $\hat{\beta}$ we have

$$\hat{\beta} = \beta + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Pu}$$

as can be seen by writing out $\mathbf{P}$ in terms of $\mathbf{X}$ and canceling to get (12.4). $\square$

Solution to Ex. 12.4.7. Note that

$$\mathbb{E}[\mathbf{My} \mid \mathbf{X}] = \mathbb{E}[\mathbf{Mu} \mid \mathbf{X}] = \mathbf{M}\mathbb{E}[\mathbf{u} \mid \mathbf{X}] = \mathbf{0}$$

Using these facts, we obtain

$$\text{cov}[\mathbf{Py}, \mathbf{My} \mid \mathbf{X}] = \text{cov}[\mathbf{X}\beta + \mathbf{Pu}, \mathbf{Mu} \mid \mathbf{X}] = \mathbb{E}[(\mathbf{X}\beta + \mathbf{Pu})(\mathbf{Mu})^\top \mid \mathbf{X}]$$

By linearity of expectations and symmetry of $\mathbf{M}$, this becomes

$$\text{cov}[\mathbf{Py}, \mathbf{My} \mid \mathbf{X}] = \mathbb{E}[\mathbf{X}\beta \mathbf{u}^\top \mathbf{M} \mid \mathbf{X}] + \mathbb{E}[\mathbf{Pu} \mathbf{u}^\top \mathbf{M} \mid \mathbf{X}]$$

Regarding the first term on the right-hand side, we have

$$\mathbb{E}[\mathbf{X}\beta \mathbf{u}^\top \mathbf{M} \mid \mathbf{X}] = \mathbf{X}\beta \mathbb{E}[\mathbf{u}^\top \mid \mathbf{X}] \mathbf{M} = \mathbf{0}$$

Regarding the second term on the right-hand side, we have

$$\mathbb{E}[\mathbf{Pu} \mathbf{u}^\top \mathbf{M} \mid \mathbf{X}] = \mathbf{P} \mathbb{E}[\mathbf{u} \mathbf{u}^\top \mid \mathbf{X}] \mathbf{M} = \mathbf{P} \sigma^2 \mathbf{I} \mathbf{M} = \sigma^2 \mathbf{P} \mathbf{M} = \mathbf{0}$$

$$\therefore \text{cov}[\mathbf{Py}, \mathbf{My} \mid \mathbf{X}] = \mathbf{0}$$

$\square$

Solution to Ex. 12.4.8. Since $\mathbf{M}$ is a function of $\mathbf{X}$, we have $\mathbb{E}[\mathbf{Mu} \mid \mathbf{X}] = \mathbf{M}\mathbb{E}[\mathbf{u} \mid \mathbf{X}] = \mathbf{0}$. As a result

$$\begin{aligned} \text{cov}[\hat{\beta}, \mathbf{Mu} \mid \mathbf{X}] &= \mathbb{E}[\hat{\beta} (\mathbf{Mu})^\top \mid \mathbf{X}] \\ &= \mathbb{E}[\beta (\mathbf{Mu})^\top + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} (\mathbf{Mu})^\top \mid \mathbf{X}] \\ &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} (\mathbf{Mu})^\top \mid \mathbf{X}] \\ &= \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u} \mathbf{u}^\top \mathbf{M} \mid \mathbf{X}] = \sigma^2 (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{M} \end{aligned}$$

The claim now follows from $\mathbf{X}^\top \mathbf{M} = (\mathbf{MX})^\top = \mathbf{0}$. $\square$

Solution to Ex. 12.4.9. Regarding the claim that $\mathbb{E}[\mathbf{Py} \mid \mathbf{X}] = \mathbf{X}\beta$, our previous results

and linearity of expectations gives

$$\mathbb{E}[\mathbf{Py} | \mathbf{X}] = \mathbb{E}[\mathbf{X}\boldsymbol{\beta} + \mathbf{Pu} | \mathbf{X}] = \mathbf{X}\boldsymbol{\beta} + \mathbf{P}\mathbb{E}[\mathbf{u} | \mathbf{X}] = \mathbf{X}\boldsymbol{\beta}$$

Regarding the claim that $\text{var}[\mathbf{Py} | \mathbf{X}] = \sigma^2 \mathbf{P}$, our rules for manipulating variances yield

$$\text{var}[\mathbf{Py} | \mathbf{X}] = \text{var}[\mathbf{X}\boldsymbol{\beta} + \mathbf{Pu} | \mathbf{X}] = \text{var}[\mathbf{Pu} | \mathbf{X}] = \mathbf{P} \text{var}[\mathbf{u} | \mathbf{X}] \mathbf{P}^\top = \mathbf{P} \sigma^2 \mathbf{I} \mathbf{P}^\top$$

Using symmetry and idempotence of $\mathbf{P}$, we obtain $\text{var}[\mathbf{Py} | \mathbf{X}] = \sigma^2 \mathbf{P}$. $\square$

Solution to Ex. 12.4.10. Similar to the solution of exercise 12.4.9. $\square$

Solution to Ex. 12.4.11. Repeating (12.12) we have

$$\hat{\beta}_k = \beta_k + (\text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \text{col}_k(\mathbf{X}))^{-1} \text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \mathbf{u} \quad (12.36)$$

Since β_k is constant, taking the variance of (12.36) conditional on $\mathbf{X}$, we obtain

$$\text{var}[\hat{\beta}_k | \mathbf{X}] = \text{var}[\mathbf{A}\mathbf{u} | \mathbf{X}] \quad \text{where } \mathbf{A} := (\text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \text{col}_k(\mathbf{X}))^{-1} \text{col}_k(\mathbf{X})^\top \mathbf{M}_1$$

Since $\mathbf{A}$ is a function of $\mathbf{X}$, we can treat it as constant given $\mathbf{X}$, and we obtain

$$\text{var}[\hat{\beta}_k | \mathbf{X}] = \mathbf{A} \text{var}[\mathbf{u} | \mathbf{X}] \mathbf{A}^\top = \mathbf{A} \sigma^2 \mathbf{I} \mathbf{A}^\top = \sigma^2 \mathbf{A} \mathbf{A}^\top$$

To complete the proof, we just observe that

$$\begin{aligned} \mathbf{A} \mathbf{A}^\top &= (\text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \text{col}_k(\mathbf{X}))^{-1} \text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \mathbf{M}_1^\top \text{col}_k(\mathbf{X}) (\text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \text{col}_k(\mathbf{X}))^{-1} \\ &= (\text{col}_k(\mathbf{X})^\top \mathbf{M}_1 \text{col}_k(\mathbf{X}))^{-1} \end{aligned}$$

where the last equality is due to symmetry and idempotence of $\mathbf{M}_1$. We conclude that (12.13) is valid. $\square$

Solution to Ex. 12.4.12. We will apply (12.13). For the model $\mathbf{y} = \beta_1 \mathbf{1} + \beta_2 \mathbf{x} + \mathbf{u}$, the matrix $\mathbf{M}_1$ is the residual projection $\mathbf{M}_c$ associated with $\mathbf{1}$, and $\text{col}_k \mathbf{X}$ is just $\mathbf{x}$. Hence, from (12.13), we have

$$\text{var}[\hat{\beta}_2 | \mathbf{X}] = \sigma^2 (\mathbf{x}^\top \mathbf{M}_c \mathbf{x})^{-1}$$

Using symmetry and idempotence of $\mathbf{M}_c$, this becomes

$$\text{var}[\hat{\beta}_2 | \mathbf{X}] = \sigma^2 [(\mathbf{M}_c \mathbf{x})^\top \mathbf{M}_c \mathbf{x}]^{-1} = \sigma^2 / \sum_{n=1}^N (x_n - \bar{x})^2$$

Finally, since the only random variables in $\mathbf{X}$ are the random variables in $\mathbf{x}$, we obtain

the solution

$$\text{var}[\hat{\beta}_2 | \mathbf{x}] = \sigma^2 / \sum_{n=1}^N (x_n - \bar{x})^2 \quad \square$$

Solution to Ex. 12.4.13. We need to show that in the special case (12.22) we have

$$\frac{(\text{RSSR} - \text{USSR})/K_2}{\text{USSR}/(N - K)} = \frac{R_c^2}{1 - R_c^2} \frac{N - K}{K_2}$$

or, equivalently,

$$\frac{\text{RSSR} - \text{USSR}}{\text{USSR}} = \frac{R_c^2}{1 - R_c^2} \quad (12.37)$$

Consider first the left-hand side of (12.37). In the case of (12.22), this becomes

$$\frac{\text{RSSR} - \text{USSR}}{\text{USSR}} = \frac{\|\mathbf{M}_c \mathbf{y}\|^2 - \|\mathbf{M} \mathbf{y}\|^2}{\|\mathbf{M} \mathbf{y}\|^2}$$

On the other hand, regarding the right-hand side of (12.37), the definition of R_c^2 and some minor manipulation gives

$$\frac{R_c^2}{1 - R_c^2} = \frac{\|\mathbf{P} \mathbf{M}_c \mathbf{y}\|^2}{\|\mathbf{M}_c \mathbf{y}\|^2 - \|\mathbf{P} \mathbf{M}_c \mathbf{y}\|^2}$$

Hence, to establish (12.37), we need to show that

$$\frac{\|\mathbf{M}_c \mathbf{y}\|^2 - \|\mathbf{M} \mathbf{y}\|^2}{\|\mathbf{M} \mathbf{y}\|^2} = \frac{\|\mathbf{P} \mathbf{M}_c \mathbf{y}\|^2}{\|\mathbf{M}_c \mathbf{y}\|^2 - \|\mathbf{P} \mathbf{M}_c \mathbf{y}\|^2}$$

This is can be established using (11.32). $\square$

Solution to Ex. 12.4.14. We have shown previously that $\mathbf{M} \mathbf{y} = \mathbf{M} \mathbf{u}$ and $\mathbf{P} \mathbf{y} = \mathbf{X} \boldsymbol{\beta} + \mathbf{P} \mathbf{u}$. Since we are conditioning on $\mathbf{X}$, we can treat it as constant. When $\mathbf{X}$ is constant, $\mathbf{P}$, $\mathbf{M}$ and $\mathbf{X} \boldsymbol{\beta}$ are all constant. Since linear (or affine) transformations of normal random vectors are normal, $\mathbf{M} \mathbf{u}$ and $\mathbf{X} \boldsymbol{\beta} + \mathbf{P} \mathbf{u}$ are both normally distributed.

It remains to show that $\mathbf{P} \mathbf{y}$ and $\mathbf{M} \mathbf{y}$ are independent given $\mathbf{X}$. Since they are normally distributed given $\mathbf{X}$, we need only show that they are uncorrelated given $\mathbf{X}$. This was already proved in exercise 12.4.7. $\square$

Solution to Ex. 12.4.17. The proof of $\mathbb{E}[\mathbf{u}_c | \mathbf{X}_c] = \mathbf{0}$ is easy. Regarding the second claim, recall that $\mathbf{C}$ is a nonsingular matrix chosen to satisfy $\mathbf{C}^\top \mathbf{C} = \boldsymbol{\Omega}^{-1}$. (Existence

is from the Cholesky decomposition; see page 67.) From $\mathbf{u}_c = \mathbf{C}\mathbf{u}$ we have

$$\mathbb{E}[\mathbf{u}_c \mathbf{u}_c^\top | \mathbf{X}_c] = \mathbf{C} \mathbb{E}[\mathbf{u} \mathbf{u}^\top | \mathbf{X}_c] \mathbf{C}^\top = \mathbf{C} \boldsymbol{\Omega} \mathbf{C}^\top$$

Since $\mathbf{C}$ is nonsingular, we have

$$\mathbf{C} \boldsymbol{\Omega} \mathbf{C}^\top = \mathbf{C}(\mathbf{C}^\top \mathbf{C})^{-1} \mathbf{C}^\top = \mathbf{C} \mathbf{C}^{-1} (\mathbf{C}^\top)^{-1} \mathbf{C}^\top = \mathbf{I} \quad \square$$

Solution to Ex. 12.4.18. If not then $\mathbf{Z}^\top \mathbf{X}$ has fewer rows than columns, and hence its columns cannot be linearly independent. See theorem 2.1.3 on page 20. $\square$

Solution to Ex. 12.4.19. The OLS estimator is $\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$. Substituting into (12.35) gives

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top (\mathbf{X} \boldsymbol{\beta} + \mathbf{Z} \boldsymbol{\theta} + \mathbf{u}) = \boldsymbol{\beta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z} \boldsymbol{\theta} + (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{u}$$

Therefore

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \mathbb{E}[\mathbb{E}[\hat{\boldsymbol{\beta}} | \mathbf{X}, \mathbf{Z}]] = \boldsymbol{\beta} + \mathbb{E}[(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{Z}] \boldsymbol{\theta}$$

If $\mathbf{X}^\top \mathbf{Z} = \mathbf{0}$ then the last term on the right-hand side drops out, and $\hat{\boldsymbol{\beta}}$ is unbiased. Otherwise $\hat{\boldsymbol{\beta}}$ is biased. $\square$

Solution to Ex. 12.4.20. To solve this exercise it suffices to give an example of a model that satisfies the assumptions of §12.2.3 and yet produces a biased IV estimator. To this end, suppose that we have only one observation, and let all variables be scalar. Let z and u be independent random variables with $\mathcal{L}(u) = N(0, \sigma^2)$ and let $x = u$. Let $y = x\beta + u$. While the IV assumptions of §12.2.3 are satisfied, the IVLS estimator is biased. Indeed, the expression (12.34) combined with $y = x\beta + u$ gives us

$$\hat{\beta}_{\text{IVLS}} = (zx)^{-1}zy = \beta + (zx)^{-1}zu = \beta + \frac{u}{x}$$

Using $x = u$ gives $\mathbb{E}[\hat{\beta}_{\text{IVLS}}] = \hat{\beta}_{\text{IVLS}} = \beta + 1$. $\square$

Solution to Ex. 12.4.21. By theorem 3.3.1 on page 60, the matrix that implements orthogonal projection onto $\text{colspace } \mathbf{Z}$ is $\mathbf{P}_Z := \mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top$. Applying this operation to each column of $\mathbf{X}$ is equivalent to premultiplying $\mathbf{X}$ by $\mathbf{P}_Z$. Regressing $\mathbf{y}$ on $\mathbf{P}_Z \mathbf{X}$ using OLS produces the estimator

$$\hat{\boldsymbol{\beta}}_{\text{2SLS}} = [(\mathbf{P}_Z \mathbf{X})^\top (\mathbf{P}_Z \mathbf{X})]^{-1} (\mathbf{P}_Z \mathbf{X})^\top \mathbf{y} = [\mathbf{X}^\top \mathbf{P}_Z \mathbf{X}]^{-1} \mathbf{X}^\top \mathbf{P}_Z \mathbf{y}$$

(The second equality uses symmetry and idempotence of $\mathbf{P}_Z$. See fact 3.3.2 on page 61.) Using the definition of $\mathbf{P}_Z$ gives (12.33). $\square$