A Primer in Econometric Theory

Lecture 11: Ordinary Least Squares

John Stachurski
Lectures by Akshay Shanker

March 26, 2017

Ordinary Least Squares

In chapter 11 we imposed only mild regularity conditions on the data

To provide additional interpretation of the estimated coefficients, we now make more assumptions — the classical OLS assumptions

The first assumption is a repetition of assumption 11.1.1 from chapter 11

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

Assumption.(12.1.2) 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) $\mathbb{E}[\mathbf{u} \,|\, \mathbf{X}] = \mathbf{0}$

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

Assumption 12.1.2 can be decomposed into the separate equations

$$y_n = \mathbf{x}_n^\mathsf{T}\boldsymbol{\beta} + u_n, \qquad n = 1, \ldots, N \tag{1}$$

Examples of models that produce relationships in the form of (1):

Example. 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. The **gravity model** relates international trade flows between country ℓ and country n via the equation

$$T_{\ell n} = \lambda \xi_{\ell n} G_{\ell}^{\alpha} G_{n}^{\beta} / D_{\ell n}^{\gamma}$$

- $T_{\ell n}$ 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

$$\ln T_{\ell n} = \ln \lambda + \alpha \ln G_{\ell} + \beta \ln G_n - \gamma \ln D_{\ell n} + \ln \xi_{\ell n} \qquad (2)$$

Fact. (12.1.1) If the linearity assumption 12.1.2 holds, then

1. $\mathbf{My} = \mathbf{Mu}$
2. $\mathbf{Py} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Pu}$
3. $\text{RSS} = \mathbf{u}^\mathsf{T}\mathbf{Mu}$

For a proof, see exercise 12.4.2

Fact. (12.1.2) If assumption 12.1.3 holds, then

1. $\mathbb{E}[\mathbf{u}] = \mathbf{0}$
2. $\mathbb{E}[u_m \mid x_{nk}] = 0$ for any m, n, k
3. $\mathbb{E}[u_m x_{nk}] = 0$ for any m, n, k (orthogonality)
4. $\text{cov}[u_m, x_{nk}] = 0$ for any m, n, k

For a proof, see exercise 12.4.3

Fact. (12.1.3) If assumption 12.1.4 holds, then

1. $\text{var}[\mathbf{u}] = \mathbb{E}[\mathbf{u}\mathbf{u}^\mathsf{T}] = \sigma^2\mathbf{I}$,
2. $\mathbb{E}[u_i^2 \,|\, \mathbf{X}] = \mathbb{E}[u_j^2 \,|\, \mathbf{X}] = \sigma^2$ for any i, j in $1, \ldots, N$, and
3. $\mathbb{E}[u_i u_j \,|\, \mathbf{X}] = 0$ whenever $i \neq j$.

Parts 1 and 2 are called **homoskedasticity** and zero correlation respectively

Combining assumptions 12.1.3 and 12.1.4 gives

$$\text{var}[\mathbf{u} \mid \mathbf{X}] := \mathbb{E}[\mathbf{u}\mathbf{u}^\mathsf{T} \mid \mathbf{X}] - \mathbb{E}[\mathbf{u} \mid \mathbf{X}]\mathbb{E}[\mathbf{u}^\mathsf{T} \mid \mathbf{X}] = \mathbb{E}[\mathbf{u}\mathbf{u}^\mathsf{T} \mid \mathbf{X}]$$

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

The standard estimator of $\boldsymbol{\beta}$ in (1) is the least squares estimator $\hat{\boldsymbol{\beta}}$ defined by Equation (11.9) in ET:

$$\hat{\boldsymbol{\beta}} := (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y} \tag{3}$$

Unlike chapter 11, $\hat{\boldsymbol{\beta}}$ now understood as an estimator of the unknown parameter vector $\boldsymbol{\beta}$

Substituting $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ into (3) and cancelling terms gives the useful comparison

$$\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} + (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{u} \tag{4}$$

$\hat{\boldsymbol{\beta}}$ is also called the **OLS estimator** of $\boldsymbol{\beta}$

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

$$\hat{\sigma}^2 := \frac{\mathrm{RSS}}{N-K} \tag{5}$$

Example that runs the gravity model regression (2) on world trade data using Python — full set of code and data can be found at johnstachurski.net/emet

First import some Python libraries often used for statistics:

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

Rad in the data to a `pandas DataFrame` from local CSV file:

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

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)
```

Perform the estimation and print a table summarizing results:

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

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
No. Observations:	19655	Prob (F-statistic):	0.00
Df Residuals:	19651	Log-Likelihood:	-47185
Df Model:	3	AIC:	9.438e+04
Covariance Type:	HC1	BIC:	9.441e+04\docume

	coef	std err	z	P>|z|
Intercept	-30.2350	0.394	-76.773	0.000
log(egdp)	1.2783	0.008	153.772	0.000
log(igdp)	1.0287	0.009	118.885	0.000
log(dist)	-1.3483	0.023	-58.113	0.000

Let's check the values for $\hat{\beta}$ reported under `coef` coincide with the expression for $\hat{\beta}$ given in (3)

First build the design matrix $\mathbf{X}$ (see the URL above for details) and then compute as follows:

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

The output agrees with the output in the table

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

A **partial regression plot** for the linear model just described:

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

- horizontal axis — residuals from regressing `log(igdp)` on all other columns in $\mathbf{X}$
- vertical axis — residuals from regressing the dependent variable `log(value)` on all other columns in $\mathbf{X}$

(Refer to discussion surrounding equations (11.26) and (11.27) in ET)

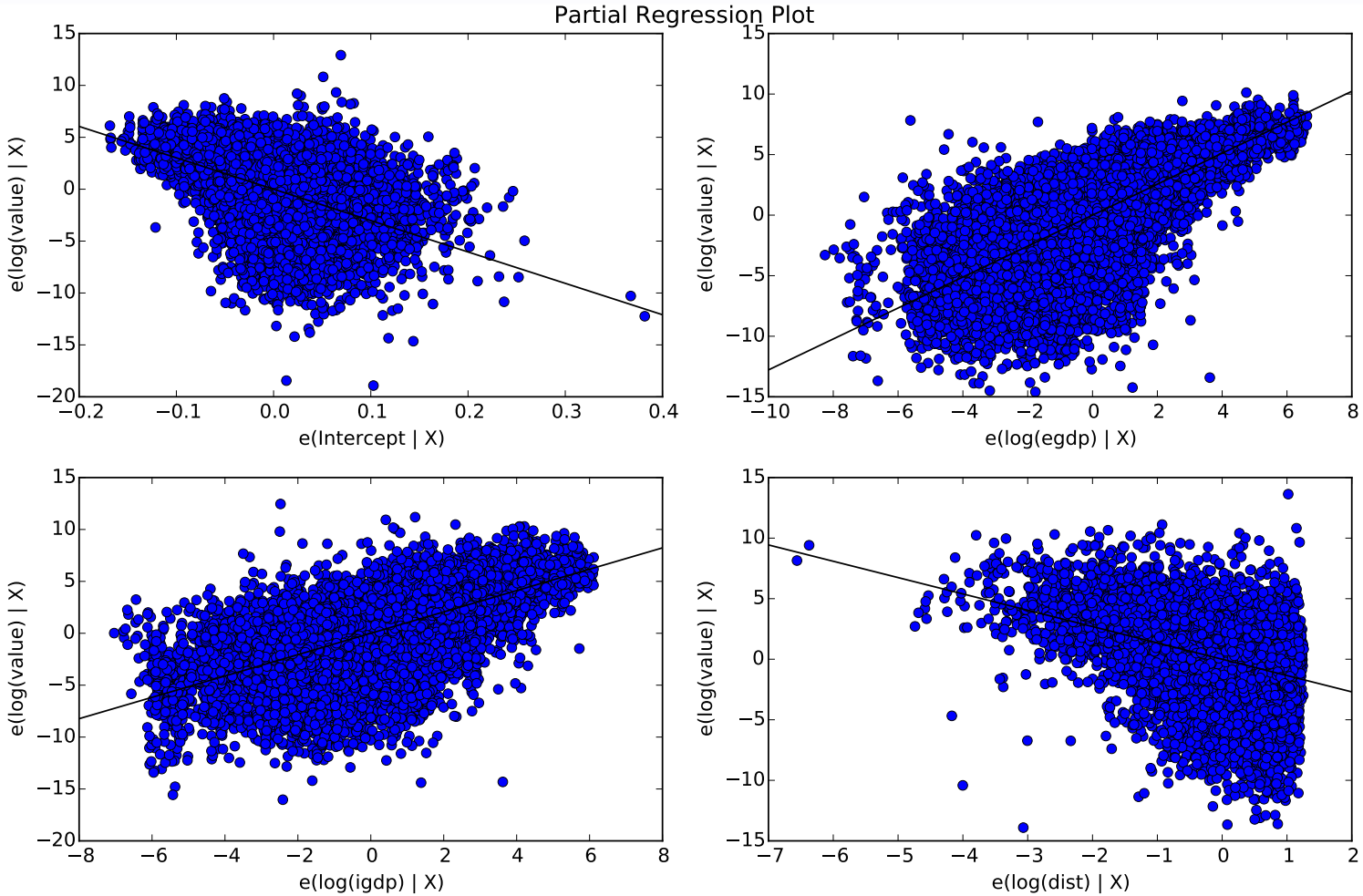

Figure: Partial regression plot for the gravity model

Finite Sample Properties

Theorem. (12.1.1) If assumptions 12.1.1–12.1.3 hold, then

$$\mathbb{E}\,\hat{\boldsymbol{\beta}} = \mathbb{E}\left[\hat{\boldsymbol{\beta}} \,|\, \mathbf{X}\right] = \boldsymbol{\beta} \tag{6}$$

If assumption 12.1.4 also holds, then

$$\mathbb{E}\,\hat{\sigma}^2 = \mathbb{E}\left[\hat{\sigma}^2 \,|\, \mathbf{X}\right] = \sigma^2 \tag{7}$$

Proof. By (12.4) and assumption 12.1.3, we have

$$\mathbb{E}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] = \boldsymbol{\beta} + (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbb{E}[\mathbf{u} \mid \mathbf{X}] = \boldsymbol{\beta}$$

Taking the unconditional expectation gives (6)

Regarding (7), we use the expression for RSS in fact 12.1.1:

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

Proof.(cont.) From fact 5.1.3, $\mathbb{E}\left[\text{RSS} \mid \mathbf{X}\right] = \sigma^2 \operatorname{trace} \mathbf{M}$

Hence

$$\mathbb{E}\left[\hat{\sigma}^2 \mid \mathbf{X}\right] = \frac{\sigma^2 \operatorname{trace}(\mathbf{M})}{N - K}$$

By fact 3.3.4, we have $\operatorname{trace} \mathbf{M} = N - K$

Inserting this into the preceding equation and taking unconditional expectations gives (7) □

Theorem. (12.1.2) If assumptions 12.1.1–12.1.4 hold, then

$$\text{var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1} \tag{8}$$

For a proof, see exercise 12.4.5

Under the stated conditions, the OLS estimator $\hat{\boldsymbol{\beta}}$ is best linear unbiased:

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

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

Here $\text{var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] \leq \text{var}[\mathbf{b} \mid \mathbf{X}]$ means $\text{var}[\mathbf{b} \mid \mathbf{X}] - \text{var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}]$ is nonnegative definite — one way to assert $\text{var}[\mathbf{b} \mid \mathbf{X}]$ is "larger"

- an implication: $\text{var}[b_k \mid \mathbf{X}] \geq \text{var}[\hat{\beta}_k \mid \mathbf{X}]$ for all k

Linearity of $\mathbf{b}$ means $\mathbf{b}$ is linear as a function of $\mathbf{y}$, taking $\mathbf{X}$ as fixed

- equivalent to requiring $\mathbf{b} = \mathbf{C}\mathbf{y}$ for some matrix $\mathbf{C}$; $\mathbf{C}$ can depend on $\mathbf{X}$ but not $\mathbf{y}$

Saying $\mathbf{b}$ is unbiased means $\mathbb{E}[\mathbf{b} \mid \mathbf{X}] = \mathbb{E}[\mathbf{C}\mathbf{y} \mid \mathbf{X}] = \boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^K$

Proof. [proof for theorem (12.1.3)]

Let $\mathbf{b} = \mathbf{C}\mathbf{y}$, as described above, and let $\mathbf{D} := \mathbf{C} - \mathbf{A}$, where $\mathbf{A} := (\mathbf{X}^\intercal \mathbf{X})^{-1}\mathbf{X}^\intercal$

Then

$$\mathbf{b} = \mathbf{C}\mathbf{y} = \mathbf{D}\mathbf{y} + \mathbf{A}\mathbf{y} = \mathbf{D}(\mathbf{X}\boldsymbol{\beta} + \mathbf{u}) + \hat{\boldsymbol{\beta}} = \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \mathbf{D}\mathbf{u} + \hat{\boldsymbol{\beta}} \quad (9)$$

Take conditional expectations and use the fact that $\mathbf{D}$ is a function of $\mathbf{X}$:

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

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

$$\boldsymbol{\beta} = \mathbf{D}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^K$$

$$\therefore \quad \mathbf{0} = \mathbf{D}\mathbf{X}\boldsymbol{\beta} \quad \text{for all } \boldsymbol{\beta} \in \mathbb{R}^K$$

In light of exercise 3.5.13, we conclude $\mathbf{D}\mathbf{X} = \mathbf{0}$

Combining this result with (9), we obtain $\mathbf{b} = \mathbf{D}\mathbf{u} + \hat{\boldsymbol{\beta}}$

Hence $\mathbf{b}$ is equal to the OLS estimator plus zero-mean noise

To complete the proof, observe

$$\begin{aligned}\operatorname{var}[\mathbf{b} \mid \mathbf{X}] &= \operatorname{var}[\mathbf{D}\mathbf{u} + \hat{\boldsymbol{\beta}} \mid \mathbf{X}] \\ &= \operatorname{var}[(\mathbf{D} + \mathbf{A})\mathbf{u} \mid \mathbf{X}] = (\mathbf{D} + \mathbf{A}) \operatorname{var}[\mathbf{u} \mid \mathbf{X}](\mathbf{D} + \mathbf{A})^{\mathsf{T}}\end{aligned}$$

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

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

Since

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

and

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

we conclude

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

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

Precision of Estimators

By theorem 12.1.2, the variance — covariance matrix of $\hat{\boldsymbol{\beta}}$ given $\mathbf{X}$ is $\sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$

- scalar variances of the individual OLS coefficient estimate $\hat{\beta}_1, \ldots, \hat{\beta}_K$ 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 — the estimator has high **precision**

Gauss–Markov Theorem: the OLS estimates will have highest precision

Hold the estimation technique fixed, as well as the sample size, and vary the application

- which problems will have high precision estimates?
- which will have low precision estimates?

Consider variance of a fixed coefficient β_k

Write $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ as

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathrm{col}_k(\mathbf{X})\beta_k + \mathbf{u} \tag{10}$$

- $\mathrm{col}_k(\mathbf{X})$ is the vector of observations of the kth regressor
- $\mathbf{X}_1$ contains as its columns the observations of the other regressors
- $\hat{\boldsymbol{\beta}}_1$ is the OLS estimates of the corresponding coefficients

By the FWL theorem:

$$\hat{\beta}_k = (\text{col}_k(\mathbf{X})^\intercal \mathbf{M}_1 \, \text{col}_k(\mathbf{X}))^{-1} \, \text{col}_k(\mathbf{X})^\intercal \mathbf{M}_1 \mathbf{y} \tag{11}$$

- $\mathbf{M}_1$ is the residual projection $\mathbf{M}_1 := \mathbf{I} - \mathbf{P}_1$
- $\mathbf{P}_1 := \mathbf{X}_1 (\mathbf{X}_1^\intercal \mathbf{X}_1)^{-1} \mathbf{X}_1^\intercal$ projects onto colspace $\mathbf{X}_1$

Apply $\mathbf{M}_1$ to both sides of (10)

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

Substituting into (11) gives

$$\hat{\beta}_k = \beta_k + (\text{col}_k(\mathbf{X})^\intercal \mathbf{M}_1 \, \text{col}_k(\mathbf{X}))^{-1} \, \text{col}_k(\mathbf{X})^\intercal \mathbf{M}_1 \mathbf{u} \tag{12}$$

Some calculations show (ex. 12.4.11)

$$\operatorname{var}[\hat{\beta}_k \mid \mathbf{X}] = \sigma^2 (\operatorname{col}_k(\mathbf{X})^\intercal \mathbf{M}_1 \operatorname{col}_k(\mathbf{X}))^{-1} = \sigma^2 \|\mathbf{M}_1 \operatorname{col}_k(\mathbf{X})\|^{-2}$$

Variance of $\hat{\beta}_k$ depends on:

1. variance σ^2 of the shock u — unavoidable
2. norm of the vector $\mathbf{M}_1 \operatorname{col}_k \mathbf{X}$

$\mathbf{M}_1 \operatorname{col}_k \mathbf{X}$ is the residuals from regressing $\operatorname{col}_k \mathbf{X}$ on $\mathbf{X}_1$

If $\|\mathbf{M}_1 \operatorname{col}_k \mathbf{X}\|$ is small, then variance of $\hat{\beta}_k$ large

- small when $\operatorname{col}_k \mathbf{X}$ is "almost" a linear combination of the other regressors and hence close to $\operatorname{colspace} \mathbf{X}_1$

Then

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

Situation referred to as **multicollinearity**

Figure to show effect of multicollinearity on the variance of OLS estimate $\hat{\beta}_2$

$\mathbf{X}$ has two columns related by $\text{col}_2 \mathbf{X} = \delta \, \text{col}_1 \mathbf{X} + (1 - \delta)\mathbf{z}$

- $\mathbf{z}$ is a vector of N independent draws from the standard normal distribution

Larger δ means more dependence between $\text{col}_1 \mathbf{X}$ and $\text{col}_2 \mathbf{X}$

Figure shows the distribution of $\hat{\beta}_2$ for different values of δ

- true parameter is $\beta_2 = 1$
- $\hat{\beta}_2$ is unbiased but its variance increases with δ

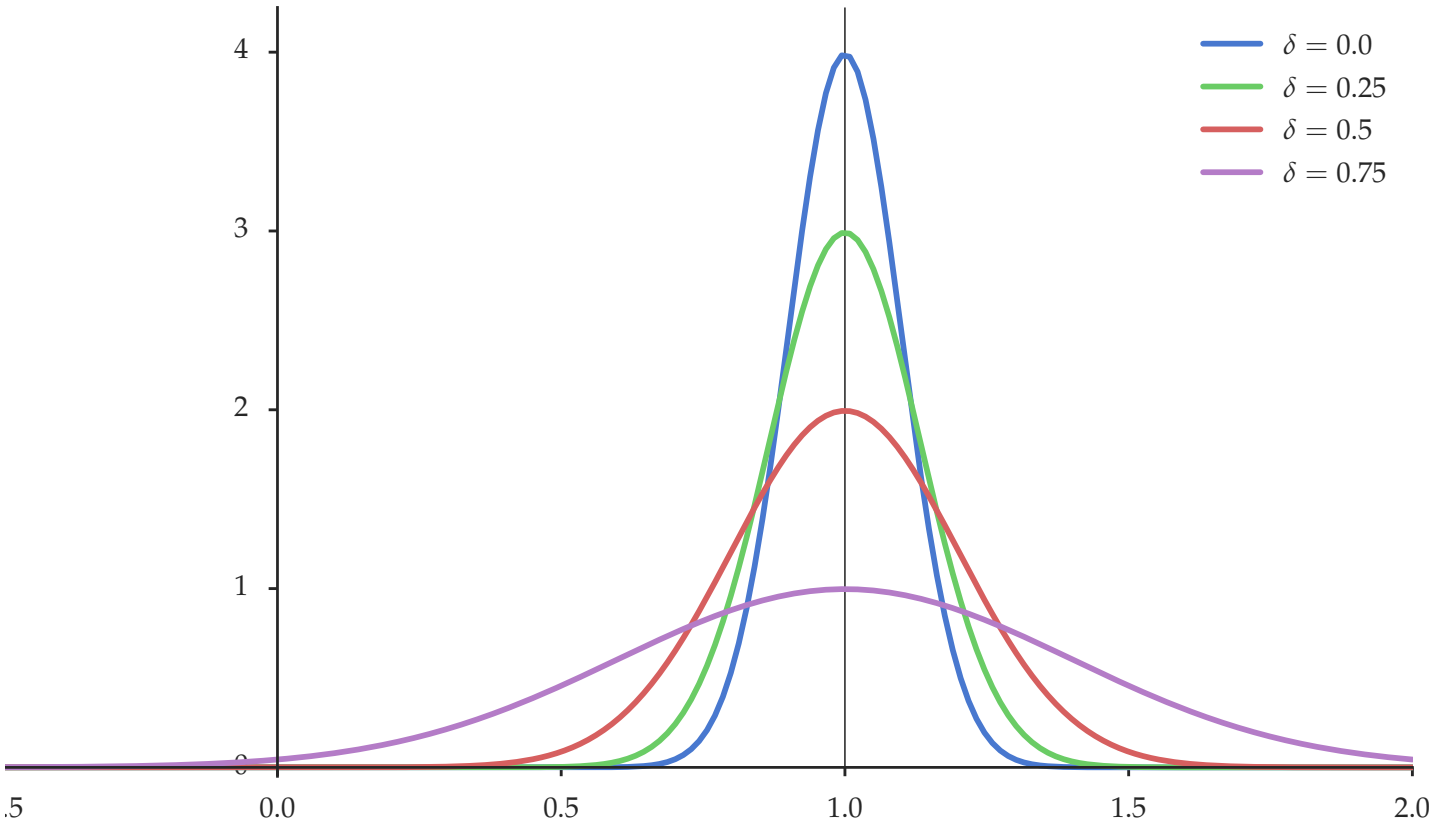

Figure: Distribution of $\hat{\beta}_2$ when $\text{col}_2 \mathbf{X} = \delta \, \text{col}_1 \mathbf{X} + (1 - \delta)\mathbf{z}$

Inference with Normal Errors

We want to compute confidence intervals or test hypotheses about the coefficients in finite samples (i.e., without appealing to asymptotics)

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

Assumption 12.1.5 implies both assumption 12.1.3 and assumption 12.1.4

Theorem. (12.1.4) If assumptions 12.1.1–12.1.2 and 12.1.5 hold, then

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

The theorem implies the distribution of individual coefficient $\hat{\beta}_k$ given $\mathbf{X}$ is also normal, with

$$\mathcal{L}\left[\hat{\beta}_k \,|\, \mathbf{X}\right] = \mathcal{L}\left[\mathbf{e}_k^\intercal\,\hat{\boldsymbol{\beta}} \,|\, \mathbf{X}\right] = \mathrm{N}(\beta_k, \sigma^2 v_k(\mathbf{X})) \qquad (13)$$

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

(Recall (5.12) in ET: if $\mathbf{x} = (x_1, \ldots, x_N)$ is multivariate normal, then

Regarding the distribution of $\hat{\sigma}^2$, convenient to work with a transformation:

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

Theorem. (12.1.5) If assumptions 12.1.1–12.1.2 and 12.1.5 hold, then

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

See page 333 in ET for a proof

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 $\hat{\boldsymbol{\beta}}$ are independent given $\mathbf{X}$

See ex. 12.4.6

The t-Test

Consider problem of testing hypothesis about an individual coefficient β_k

The null hypothesis

$$H_0 \colon \beta_k = \beta_k^0$$

where β_k^0 is any number

If σ^2 is known, construct a test of H_0 based on:

$$z_k := \frac{\hat{\beta}_k - \beta_k}{\sigma \sqrt{v_k(\mathbf{X})}} \implies \mathcal{L}[\, z_k \,|\, \mathbf{X} \,] = \mathrm{N}(0,1) \tag{14}$$

We don't know σ^2 — replace σ^2 with its estimator $\hat{\sigma}^2$

Some 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 $\hat{\beta}_k$

Replace standard deviation with its sample estimate $\text{se}(\hat{\beta}_k)$ and β_k with β_k^0 in (14) and obtain the ***t*-statistic**

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

associated with the hypothesis H_0

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

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

See page 334 for a proof.

Let $T := |t_k|$ and let a desired size α for our test of H_0 be given

To generate a test of size α, we choose $c = c_\alpha$ to solve $\alpha = \mathbb{P}_\theta\{T > c\}$, or

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

The solution is $c_\alpha = F^{-1}(1 - \alpha/2)$

- F is the Student's t CDF with $N - K$ degrees of freedom (Equation (4.36) in ET)

The corresponding p-value is $2F(-|t_k|)$

Example. A common implementation of the t-test is the test that a given coefficient is equal to zero

For the kth 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**

The F-Test

Return to the setting of when we discussed the FWL theorem (§11.2.2):

Let $\mathbf{y}$ and $\mathbf{X}$ be given and let $\hat{\boldsymbol{\beta}}$ be the least squares estimator. Let K_1 be an integer with $1 \leq K_1 < K$, and let

- $\mathbf{X}_1$ be a matrix consisting of the first K_1 columns of $\mathbf{X}$,
- $\mathbf{X}_2$ be a matrix consisting of the remaining $K_2 := K - K_1$ columns,
- $\hat{\boldsymbol{\beta}}_1$ be the $K_1 \times 1$ vector consisting of the first K_1 elements of $\hat{\boldsymbol{\beta}}$.
- $\hat{\boldsymbol{\beta}}_2$ be the $K_2 \times 1$ vector consisting of the remaining K_2 elements of $\hat{\boldsymbol{\beta}}$,
- $\mathbf{P}_1 := \text{proj}(\text{colspace}\, \mathbf{X}_1)$, and
- $\mathbf{M}_1 := \mathbf{I} - \mathbf{P}_1 =$ the corresponding residual projection

Hypotheses concerning multiple regressors

Null hypothesis:

$$H_0 : \boldsymbol{\beta}_2 = \mathbf{0}$$

Since

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u} \tag{15}$$

Under the null hypothesis,

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{u} \tag{16}$$

Let

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

be the residual sums of squares for the unrestricted regression (15) and restricted regression (16)

Test statistic for our null hypothesis

$$F := \frac{(\text{RSSR} - \text{USSR})/K_2}{\text{USSR}/(N-K)} \tag{17}$$

Large residuals in the restricted regression (16) relative to those in (15) result in large values for F — 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 (17) 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

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

Recall that if x_1 and x_2 are independent with $\mathcal{L}(x_i) = \chi^2(k_i)$ for $i = 1, 2$, then

$$\frac{x_1/k_1}{x_2/k_2} \quad \text{is } F \text{ distributed with parameters } k_1, k_2$$

Proof.(cont.)

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), under the null hypothesis,

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

It follows that

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

Proof.(cont.)

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

$$Q_1 = \frac{\text{RSSR} - \text{USSR}}{\sigma^2}$$

$$= \frac{\mathbf{u}^\mathsf{T}(\mathbf{I} - \mathbf{P}_1 - \mathbf{I} + \mathbf{P})\mathbf{u}}{\sigma^2} = (\sigma^{-1}\mathbf{u})^\mathsf{T}(\mathbf{P} - \mathbf{P}_1)(\sigma^{-1}\mathbf{u})$$

Exercise: show $(\mathbf{P} - \mathbf{P}_1)$ is symmetric and idempotent. Hence

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

$$= \text{trace}\,\mathbf{P} - \text{trace}\,\mathbf{P}_1 = K - K_1 = K_2$$

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

Now we show that, under the null hypothesis and taking $\mathbf{X}$ as given, Q_1 and Q_2 are independent

Q_1 is a function of $(\mathbf{P} - \mathbf{P}_1)\mathbf{u}$, while Q_2 is a function of $\mathbf{Mu}$

- we show $(\mathbf{P} - \mathbf{P}_1)\mathbf{u}$ and $\mathbf{Mu}$ are independent given $\mathbf{X}$ by showing their covariance is zero

Observe

$$\operatorname{cov}[(\mathbf{P} - \mathbf{P}_1)\mathbf{u}, \mathbf{Mu} \,|\, \mathbf{X}]$$

$$= \mathbb{E}[(\mathbf{P} - \mathbf{P}_1)\mathbf{u}(\mathbf{Mu})^\mathsf{T} \,|\, \mathbf{X}] = \mathbb{E}[(\mathbf{P} - \mathbf{P}_1)\mathbf{u}\mathbf{u}^\mathsf{T}\mathbf{M} \,|\, \mathbf{X}]$$

Proof.(cont.) Since $\mathbf{P}$, $\mathbf{P}_1$, and $\mathbf{M}$ are just functions of $\mathbf{X}$, the above becomes

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

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 the theorem □

Common F-Test: test that all coefficients of nonconstant regressors are zero

Consider:

$$\mathbf{y} = \mathbf{1}\beta_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \mathbf{u}$$

where $\boldsymbol{\beta}_2$ is the vector of coefficients corresponding to the nonconstant regressors. Null hypothesis:

$$\mathbf{y} = \mathbf{1}\beta_1 + \mathbf{u}$$

F statistic in (17) becomes

$$F = \frac{R_c^2}{1 - R_c^2} \frac{N - K}{K_2}$$

(proof is exercise — see page 338)

Why is large F is evidence against the null?

Nonspherical Errors

Heteroskedasticity occurs when the variance of the error term is not constant across observations

- violate assumption 12.1.4

Errors **heteroskedastic** if diagonal terms of $\mathbb{E}\left[\mathbf{u}\mathbf{u}^{\mathsf{T}} \mid \mathbf{X}\right]$ 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 $\hat{\boldsymbol{\beta}}$

However, without assumption 12.1.4, no longer true that $\hat{\sigma}^2$ is unbiased for σ

- expression $\operatorname{var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] = \sigma^2(\mathbf{X}^\intercal\mathbf{X})^{-1}$ from (8) is no longer valid
- Gauss–Markov theorem also breaks down
- results on inference we discussed above from §12.1.4 no longer valid

Replace the assumption $\mathbb{E}[\mathbf{u}\mathbf{u}^\mathsf{T} \mid \mathbf{X}] = \sigma^2\mathbf{I}$ with more general assumption

$$\mathbb{E}[\mathbf{u}\mathbf{u}^\mathsf{T} \mid \mathbf{X}] = \mathbf{\Omega} \tag{18}$$

for some positive definite matrix $\mathbf{\Omega}$

Write the variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ under these conditions as

$$\operatorname{var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] = \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^\mathsf{T} \mid \mathbf{X}] = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{\Omega}\mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$$

Maintain assumptions 12.1.1–12.1.3

Write the variance of the OLS estimator $\hat{\boldsymbol{\beta}}$ under these conditions as

$$\begin{aligned}\operatorname{var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}] &= \mathbb{E}\left[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathsf{T}} \mid \mathbf{X}\right] \\ &= (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathsf{T}}\boldsymbol{\Omega}\mathbf{X}(\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1}\end{aligned}$$

If $\boldsymbol{\Omega}$ is known, then we can estimate using **generalized least squares**

- Cholesky decomposition: there exists a nonsingular matrix $\mathbf{C}$ such that $\mathbf{C}^{\mathsf{T}}\mathbf{C} = \boldsymbol{\Omega}^{-1}$
- Use $\mathbf{C}$ to transform the regression model by multiplying both sides of $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ by $\mathbf{C}$

This gives

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

Exercise 12.4.17 asks you to show

$$\mathbb{E}\left[\mathbf{u}_c \mid \mathbf{X}_c\right] = \mathbf{0} \quad \text{and} \quad \mathbb{E}\left[\mathbf{u}_c\mathbf{u}_c^\mathsf{T} \mid \mathbf{X}_c\right] = \mathbf{I}$$

All OLS assumptions satisfied for the transformed data

Use the transformed data to estimate $\boldsymbol{\beta}$ via $(\mathbf{X}_c^\mathsf{T}\mathbf{X}_c)^{-1}\mathbf{X}_c^\mathsf{T}\mathbf{y}_c$ — recover the best linear unbiased property

The new estimator:

$$\hat{\boldsymbol{\beta}}_{\text{GLS}} := (\mathbf{X}^\intercal \boldsymbol{\Omega}^{-1} \mathbf{X})^{-1} \mathbf{X}^\intercal \boldsymbol{\Omega}^{-1} \mathbf{y} \tag{19}$$

- estimator is the **generalized least squares (GLS) estimator** of $\boldsymbol{\beta}$

In practice, $\boldsymbol{\Omega}$ is rarely known

Estimation difficult unless additional structure is imposed

- $\boldsymbol{\Omega}$ is $N \times N$, implying that, absent further restrictions, the number of parameters contained in $\boldsymbol{\Omega}$ grows *faster* than the sample

Assume $\boldsymbol{\Omega}$ is diagonal

- rules out correlation but maintains the possibility of heteroskedasticity

One estimator of $\boldsymbol{\Omega}$:

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

Replace $\boldsymbol{\Omega}$ with $\hat{\boldsymbol{\Omega}}$ in expression for $\operatorname{var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}]$, get estimate of $\operatorname{var}[\hat{\boldsymbol{\beta}} \mid \mathbf{X}]$:

$$\hat{\mathbf{V}} := (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \hat{\boldsymbol{\Omega}} \mathbf{X} (\mathbf{X}^\intercal \mathbf{X})^{-1}$$

Using this in place of the variance estimate $\hat{\sigma}^2 (\mathbf{X}^\intercal \mathbf{X})^{-1}$ to form standard errors, we get the **heteroskedasticity-consistent standard errors** proposed by White (1980)

Bias

We will look at what happens when Assumptions 12.1.2 and 12.1.3 (linearity and exogeneity) fail

Linearity fundamental to OLS. Without lineariry:

- suppose $y_n = f(\mathbf{x}_n) + u_n$ for some arbitrary function f
- cannot even *ask* if the OLS estimator is unbiased because no parameter vector for $\hat{\boldsymbol{\beta}}$ to estimate

Endogeniety Bias

Bias due to failure of exogeniety assumption 12.1.3 is called **endogeneity bias**

Recalling Equation (4) at the start of the lecture,

$$\mathbb{E}\left[\hat{\boldsymbol{\beta}} \,|\, \mathbf{X}\right] - \boldsymbol{\beta} = (\mathbf{X}^\intercal \mathbf{X})^{-1} \mathbf{X}^\intercal \mathbb{E}\left[\mathbf{u} \,|\, \mathbf{X}\right] \tag{20}$$

Without $\mathbb{E}\left[\mathbf{u} \,|\, \mathbf{X}\right] = \mathbf{0}$, we cannot assert right-hand side is zero, hence $\hat{\boldsymbol{\beta}}$ is biased

Striving for unbiasedness somewhat misguided. Some degree of bias to

- penalize complexity
- recognize empirical distribution not the true distribution

However, good estimators reduce bias as the sample size increases, so that the object they aim to compute can be recovered asymptotically

- not true for the OLS estimator when exogeneity fails

Recall from theorem 11.1.2 that:

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathsf{T}}\right]^{-1} \mathbb{E}\left[\mathbf{x}\, y\right]$$

as $N \to \infty$

The right-hand side is the vector of coefficients in the best linear predictor

If $y = \mathbf{x}^{\mathsf{T}}\boldsymbol{\beta} + u$, then

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathsf{T}}\right]^{-1} \mathbb{E}\left[\mathbf{x}\left(\mathbf{x}^{\mathsf{T}}\boldsymbol{\beta} + u\right)\right] = \boldsymbol{\beta} + \mathbb{E}\left[\mathbf{x}\mathbf{x}^{\mathsf{T}}\right]^{-1} \mathbb{E}\left[\mathbf{x}\, u\right]$$

With exogeneity we get $\mathbb{E}\left[\mathbf{x}\, u\right] = \mathbb{E}\left[\mathbf{x}\, \mathbb{E}\left[u \mid \mathbf{x}\right]\right] = 0$ and the OLS estimator is consistent — cannot assert without exogeniety

Examples of Exogeniety

Consider again the Cobb–Douglas production function which yields the regression model

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

- y is output, k is capital and ℓ is labor
- subscript n indicates observation on the nth firmand
- term u_n is a firm specific productivity shock

Source of endogeniety bias: firm chooses higher levels of capital and labor when it anticipates high productivity in the current period

To illustrate, suppose firm n recieved and observed $u_{n,-1}$ last period

Productivity follows random walk, with $u_n = u_{n,-1} + \eta_n$, where η_n is zero-mean white noise

Firms

- forecast period n productivity as $\mathbb{E}\left[u_n \mid u_{n,-1}\right] = u_{n,-1}$
- increase labor input when productivity is anticipated to be high, with the specific relationship $\ell_n = a + b\mathbb{E}\left[u_n \mid u_{n,-1}\right]$ for $b > 0$

With zero-mean shocks:

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

Strictly positive whenever $u_{n,-1}$ has positive variance

- exogeniety fails as conditions of fact 12.1.2 fail

Now consider data generated according to AR(1) model

$$x_0 = 0 \quad \text{and} \quad x_n = \beta x_{n-1} + u_n \quad \text{for } n = 1, \ldots, N \tag{21}$$

Let $\{u_n\}_{n=1}^N$ be IID with distribution $\mathrm{N}(0, \sigma^2)$

The unknown parameters are β and σ^2

Let $\mathbf{y} := (x_1, \ldots, x_N)$, $\mathbf{x} := (x_0, \ldots, x_{N-1})$, and $\mathbf{u} := (u_1, \ldots, u_N)$,

Write N equations in (21) as $\mathbf{y} = \beta \mathbf{x} + \mathbf{u}$

The OLS estimate of β is $\hat{\beta} := (\mathbf{x}^\intercal \mathbf{x})^{-1} \mathbf{x}^\intercal \mathbf{y}$

If $n \geq m$, then we can write x_n as

$$x_n = \sum_{j=0}^{n-1} \beta^j u_{n-j}$$

Hence $\mathbb{E}[x_n u_m] = \sum_{j=0}^{n-1} \beta^j \mathbb{E}[u_{n-j} u_m] = \beta^{n-m} \sigma^2$

Once again exogeneity fails since conditions of fact 12.1.2 fail

```
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()
```

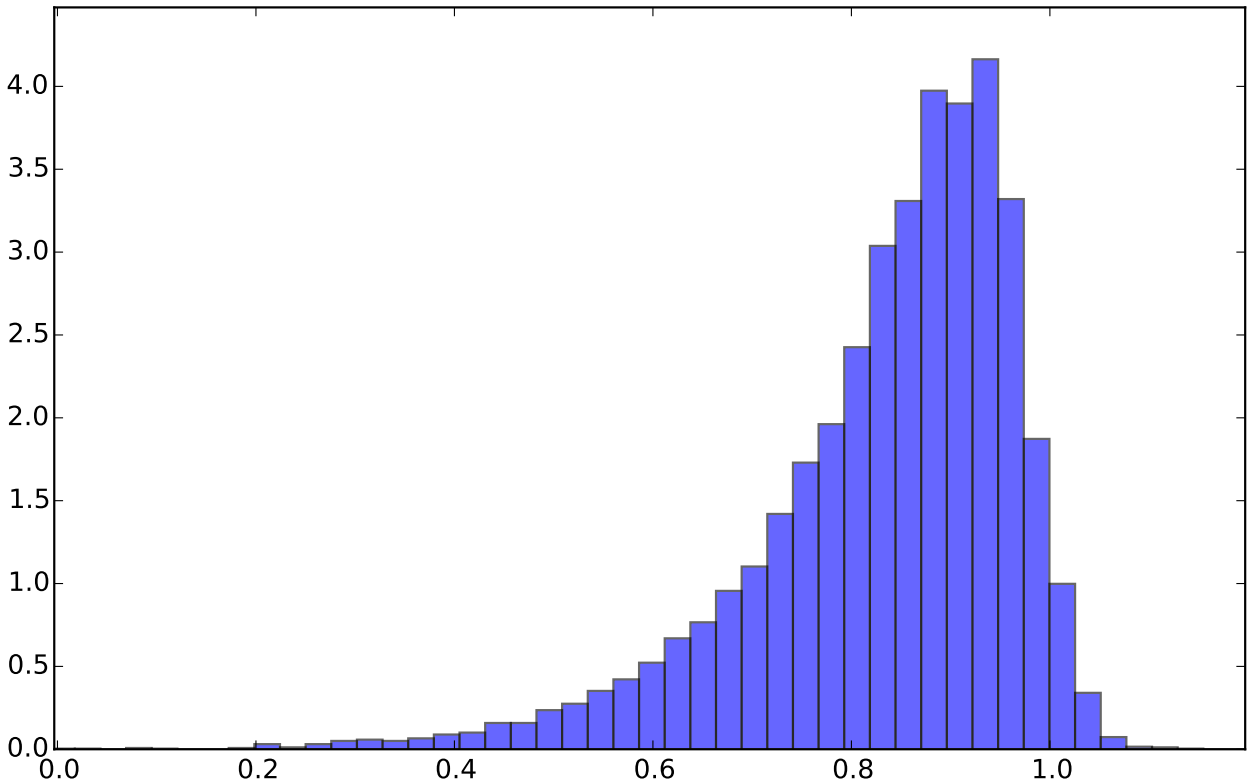

Figure: Observations of $\hat{\beta}$ when $\beta = 0.9$

Instrumental Variables

When regressors are endogenous, we can find consistent estimators of $\boldsymbol{\beta}$ when extra information is available in the form of "exogenous" variables called **instruments**

These collected into a $N \times J$ matrix $\mathbf{Z}$, where each column is observations on one exogenous variable

Assumptions:

1. $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$
2. $\mathbb{E}[\mathbf{u} \mid \mathbf{Z}] = \mathbf{0}$
3. $\mathbb{E}[\mathbf{u}\mathbf{u}^\mathsf{T} \mid \mathbf{Z}] = \sigma^2\mathbf{I}$ for some positive constant σ
4. $\mathbf{Z}^\mathsf{T}\mathbf{X}$ has full column rank

Assumption 4. implies $J \geq K$

- If $J > K$, the model is said to be **overidentified**
- If $J = K$, the model is called **exactly identified**
- If $J < K$, the model is called **underidentified**

Multiply $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}$ by $\mathbf{Z}^\mathsf{T}$ to produce

$$\mathbf{Z}^\mathsf{T}\mathbf{y} = \mathbf{Z}^\mathsf{T}\mathbf{X}\boldsymbol{\beta} + \mathbf{w} \quad \text{where} \quad \mathbf{w} := \mathbf{Z}^\mathsf{T}\mathbf{u} \tag{22}$$

Recall the GLS estimator

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

Substitute $\mathbf{X}$ for $\mathbf{Z}^\intercal \mathbf{X}$, $\mathbf{y}$ for $\mathbf{Z}^\intercal \mathbf{y}$, and $\boldsymbol{\Omega}$ for $\sigma^2 \mathbf{Z}^\intercal \mathbf{Z}$

The **instrumental variable least squares (IVLS) estimator**

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

If $J = K$, simplify to

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

Cannot claim unbiasedness and other properties by appealing to the GLS theory — $\hat{\boldsymbol{\beta}}_{\text{IVLS}}$ biased in general (see ex. 12.4.20)

- because direct application of GLS to (22) requires $\mathbb{E}[\mathbf{w} \mid \mathbf{Z}^\mathsf{T}\mathbf{X}] = \mathbf{0}$ rather than $\mathbb{E}[\mathbf{w} \mid \mathbf{Z}] = \mathbf{0}$

However, bias typically smaller than OLS and vanishes asymptotically

Figure: OLS and IVLS estimates of β in $y = x\beta + u$ when $\beta = 10$:

- z drawn independently of u
- $x = \alpha z + (1 - \alpha)u$
- α is constant

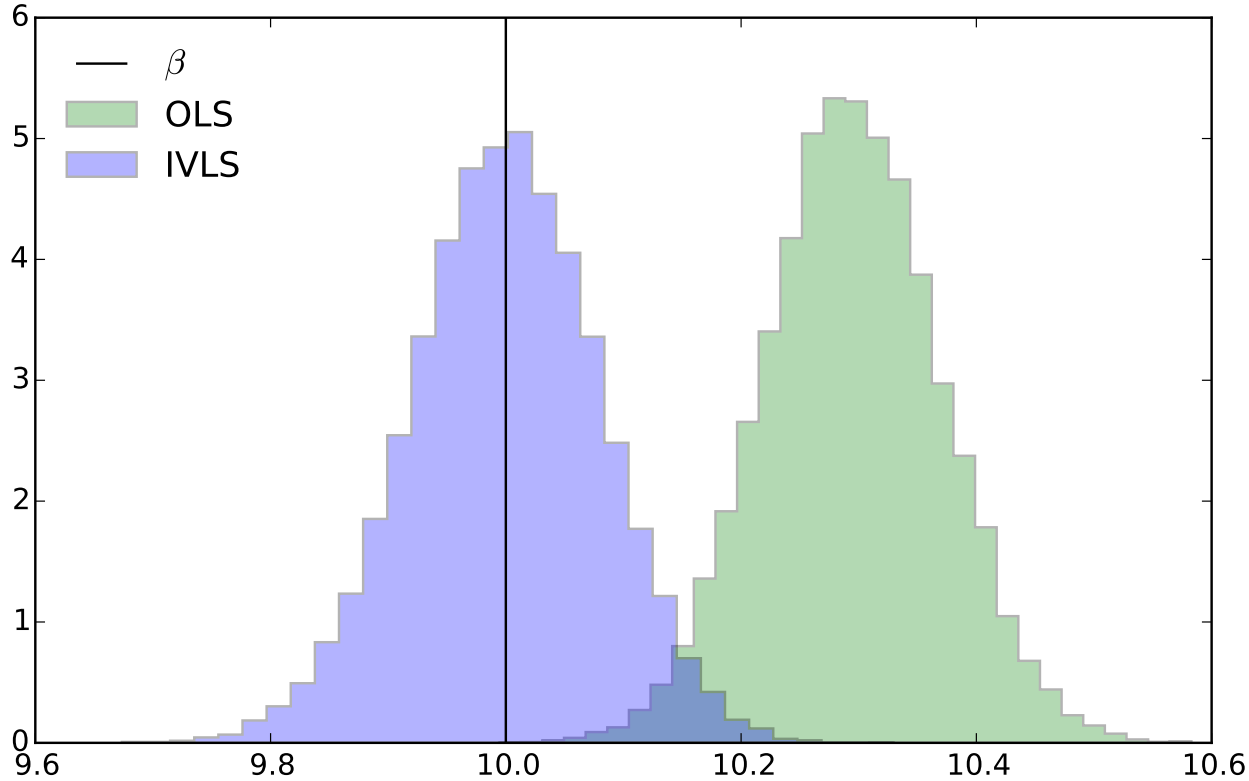

Figure: Simulated draws of OLS and IVLS estimators of β

Causality

Interpret R^2 as measuring the "explanatory power"?

- common terminology: "the total variation in $\mathbf{y}$ is the sum of explained and unexplained variation"
- value of R^2 claimed to be the fraction of the variation in $\mathbf{y}$ "explained" by the regressors

Above terminology misleading: R^2 says nothing about causation *per se*

- R^2 better thought of as a measure of correlation

Understanding causality requires either good experiments or good theory — no magic econometric procedure to reliably extract causality from observational