Intermediate Econometrics

Lecture 7

Vincenzo Gioia

06 November 2026

Linear Model

Endogeneity

- The unbiasedness and the consistency of the OLS estimator rest on the hypothesis that the conditional expectation of the error is constant (and can be set to zero if the model contains an intercept)

- Consider the simple linear model:

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$$

with

$$\begin{aligned}\mathbb{E}(Y_i|x_i) &= \mathbb{E}[\beta_1 + \beta_2 x_i + \varepsilon_i|x_i] \\ &= \beta_1 + \beta_2 x_i + \mathbb{E}[\varepsilon_i|x_i] = \beta_1 + \beta_2 x_i\end{aligned}$$

if $\mathbb{E}(\varepsilon_i|x_i) = 0$

- When x is correlated with ε, we face **endogeneity**

- **Endogeneity** leads to **biased** and **inconsistent** OLS estimates

- It is common in observational data

Linear Model

Why endogeneity matters

- The same property can also be described using the covariance between the covariate and the error that can be written, using the rule of repeated expectation (tower property):

$$\begin{aligned}\text{cov}(x, \varepsilon) &= \mathbb{E}[(x - \mu_x)\varepsilon] = \mathbb{E}_x[\mathbb{E}_\varepsilon[(x - \mu_x)\varepsilon|x]] \\ &= \mathbb{E}_x[(x - \mu_x)\mathbb{E}_\varepsilon[\varepsilon|x]]\end{aligned}$$

- **Key consequence:**

$$\text{cov}(x, \varepsilon) \neq 0 \Rightarrow \hat{\beta}_{2,\text{OLS}} \xrightarrow{p} \beta_2 + \frac{\text{cov}(x, \varepsilon)}{\text{var}(x)}$$

where $\xrightarrow{p}$ stands for convergence in probability

- Indeed, considering a simple linear regression model

$$\begin{aligned}\hat{\beta}_{2,\text{OLS}} &= \frac{cov(x, y)}{var(x)} = \frac{cov(x, \beta_1 + \beta_2 x + \varepsilon)}{var(x)} \\ &= \beta_2 + \frac{cov(x, \varepsilon)}{var(x)}\end{aligned}$$

Linear Model

Why endogeneity matters

- Cases that we could encounter in applications:

1. if $cov(x, \varepsilon) > 0 \implies$ OLS overestimates β_2
2. if $cov(x, \varepsilon) < 0 \implies$ OLS underestimates β_2
3. if $cov(x, \varepsilon) = 0 \implies$ OLS is consistent (converges to β_2)

- OLS incorrectly attributes variation in y to x
- Standard errors are no longer meaningful
- Inference becomes unreliable

Linear Model

Endogeneity

- If the conditional expectation of ε is a constant, $\mathbb{E}_\varepsilon[\varepsilon|x] = \mu_\varepsilon$ (not necessarily 0), the covariance is

$$cov(x, \varepsilon) = \mu_\varepsilon \mathbb{E}_x[x - \mu_x] = 0$$

- Stated in a different way, x is supposed to be exogenous, or x is assumed to be uncorrelated with ε

- Endogeneity arises when $cov(x, \varepsilon) \neq 0$

- Sources of endogeneity:

1. **There are errors in the variables**
2. **There are omitted variables**
3. **Simultaneity bias**

Linear Model

1. Errors in the variables

- Data used in economics, especially micro-data, are prone to errors of measurement (in both outcome and predictors)

- Suppose that the model that we seek to estimate is

$$y_i^* = \beta_1 + \beta_2 x_i^* + \varepsilon_i^*$$

where the covariate is exogenous ($cov(x^*, \varepsilon^*) = 0$)

- We will distinguish two cases:

1. The outcome is observed with error (the predictor is observed correctly, i.e., $x_i^* = x_i$), that is
$$y_i^* = y_i - \nu_i$$
2. The predictor is observed with error (the outcome is observed correctly, $y_i^* = y_i$)

$$x_i^* = x_i - \nu_i$$

where ν_i is the measurement error.

Linear Model

1. Errors in the variables (outcome)

- Then, if $y_i^* = y_i - \nu_i$

$$y_i = \beta_1 + \beta_2 x_i + (\varepsilon_i^* + \nu_i)$$

- The error of the model is $\varepsilon_i = \varepsilon_i^* + \nu_i$, which is still uncorrelated with x if ν is uncorrelated with x, which means that the error of measurement of the response is uncorrelated with the covariate

- The measurement error only increases the size of the error, which implies:

1. the coefficients are estimated less precisely
2. the R^2 is lower

compared to a model with a correctly measured response

Linear Model

1. Errors in the variables (predictor)

- Then, if $x_i^* = x_i - \nu_i$

$$y_i = \beta_1 + \beta_2(x_i - \nu_i) + \varepsilon_i^* = \beta_1 + \beta_2 x_i + (\varepsilon_i^* - \beta_2 \nu_i)$$

- If the measurement error is uncorrelated with the value of the covariate, the variance of the observed covariate is

$$\sigma_x^2 = \sigma_x^{*2} + \sigma_\nu^2$$

and the covariance between the observed covariate and the measurement error is equal to the variance of the measurement error, that is $\text{cov}(x, \nu) = \mathbb{E}((x^* + \nu - \mu_x)\nu) = \sigma_\nu^2$ because the measurement error is uncorrelated with the covariate x^*

- The error of the model is correlated with x, as

$$cov(x, \varepsilon) = cov(x^* + \nu, \varepsilon^* - \beta_2 \nu) = -\beta_2 \sigma_\nu^2$$

Linear Model

1. Errors in the variables (predictor): the OLS estimator

- Start from the model $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ and from its average $\bar{y} = \beta_1 + \beta_2 \bar{x} + \bar{\varepsilon}$. Subtracting the second from the first (the intercept cancels):

$$y_i - \bar{y} = \beta_2 (x_i - \bar{x}) + (\varepsilon_i - \bar{\varepsilon})$$

- Substitute it in the numerator of the OLS estimator:

$$\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) = \beta_2 \sum_{i=1}^{n} (x_i - \bar{x})^2 + \sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})$$

- The deviations of x from its mean sum to zero, $\sum_{i=1}^{n} (x_i - \bar{x}) = 0$, so

$$\sum_{i=1}^{n} (x_i - \bar{x})\bar{\varepsilon} = \bar{\varepsilon} \sum_{i=1}^{n} (x_i - \bar{x}) = 0 \quad \Rightarrow \quad \sum_{i=1}^{n} (x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) = \sum_{i=1}^{n} (x_i - \bar{x})\varepsilon_i$$

Linear Model

1. Errors in the variables (predictor)

- Dividing by $\sum_{i=1}^{n}(x_i - \bar{x})^2$, the OLS estimator can be written as

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
$$= \beta_2 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where x_i is the observed covariate and $\varepsilon_i = \varepsilon_i^* - \beta_2\nu_i$ is the error of the model

- Taking the expectation of the numerator, we have $\mathbb{E}[(x - \bar{x})\varepsilon] = -\beta_2\sigma_\nu^2$ and, approximating the expectation of the ratio with the ratio of the expectations (exact in the limit as $n \to \infty$), the expected value of the estimator is

$$\mathbb{E}(\hat{\beta}_2) \approx \beta_2 \left(1 - \frac{\sigma_\nu^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}\right) = \beta_2 \left(1 - \frac{\sigma_\nu^2}{\hat{\sigma}_x^2}\right)$$

- The OLS estimator is biased and the term in brackets is 1 minus the share of the variance of x that is due to measurement errors

Linear Model

1. Errors in the variables (predictor)

$$\mathbb{E}(\hat{\beta}_2) \approx \beta_2 \left(1 - \frac{\sigma_\nu^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2/n}\right) = \beta_2 \left(1 - \frac{\sigma_\nu^2}{\hat{\sigma}_x^2}\right)$$

- Then, $|\mathbb{E}(\hat{\beta}_2)| < |\beta_2|$
- This is called **attenuation bias**: it can be either a downward or an upward bias depending on the sign of β_2
- This bias clearly does not vanish in large samples. As n grows, the empirical variances/covariances converge to the population ones, and the estimator therefore converges to $\beta_2(1 - \sigma_\nu^2/\sigma_x^2)$

Linear Model

Errors in the variables: a small simulation

- True model:

$$y_i^* = 1 + 2x_i^* + \varepsilon_i^*$$

- $x_i^* \sim \mathcal{N}(0,1)$, $\varepsilon_i^* \sim \mathcal{N}(0,1)$ and measurement error $\nu_i \sim \mathcal{N}(0, 0.5)$ (the variance is 0.5)
- We compare three scenarios:

1. no measurement error
2. error on the outcome
3. error on the predictor

```
set.seed(1)
beta_1 <- 1
beta_2 <- 2
s2x <- 1
s2nu <- 0.5
nobs <- 200
B <- 1000
est <- matrix(NA, B, 3,
              dimnames = list(NULL, c("no err", "err y", "err x")))
r2  <- matrix(NA, B, 2,
              dimnames = list(NULL, c("no err", "err y")))
```

Linear Model

Errors in the variables: a small simulation

```
for (b in 1:B) {
  xs    <- rnorm(nobs, 0, sqrt(s2x))
  ys    <- beta_1 + beta_2 * xs + rnorm(nobs)
  y_obs <- ys + rnorm(nobs, 0, sqrt(s2nu))
  x_obs <- xs + rnorm(nobs, 0, sqrt(s2nu))
  est[b, ] <- c(coef(lm(ys ~ xs))[2],
                coef(lm(y_obs ~ xs))[2],
                coef(lm(ys ~ x_obs))[2])
  r2[b, ]  <- c(summary(lm(ys ~ xs))$r.squared,
                summary(lm(y_obs ~ xs))$r.squared)
}
```

```
colMeans(est)
```

```
##   no err    err y    err x
## 1.999818 2.001234 1.330538
```

```
beta_2 * (1 - s2nu / (s2x + s2nu))
```

```
## [1] 1.333333
```

```
colMeans(r2)
```

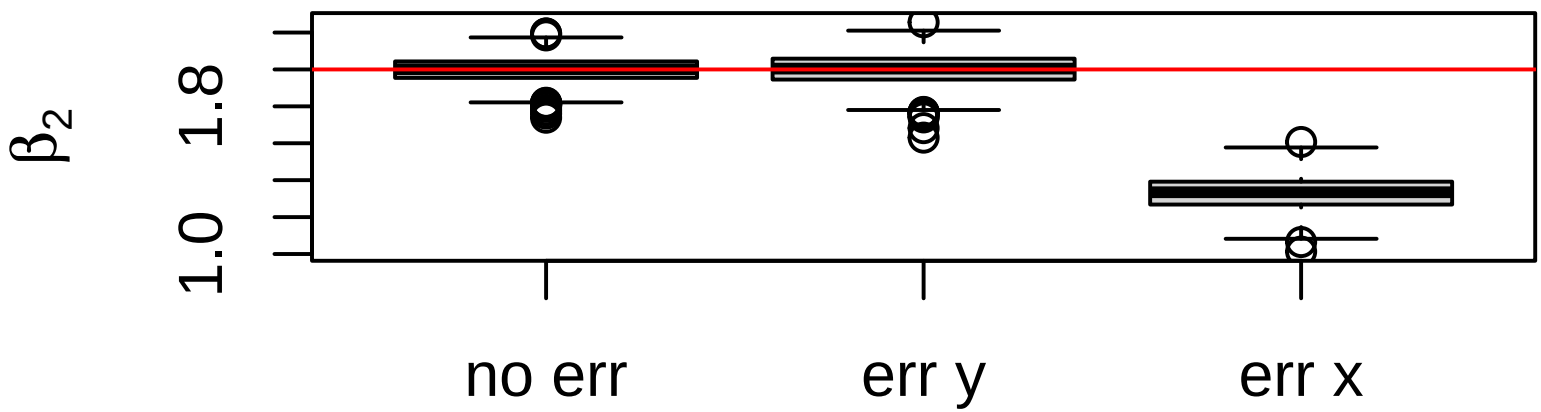
```
##    no err     err y
## 0.7999061 0.7275084
```

Linear Model
Errors in the variables: a small simulation

- Error on the response: unbiased, but larger variance and lower R^2

- $E[\hat{\beta}_2] \approx \beta_2(1 - \sigma_\nu^2/\sigma_x^2) = 2 \times (1 - 0.5/1.5) \approx 1.33$

```
boxplot(est, ylab = expression(hat(beta)[2]))
abline(h = beta_2, col = "red")
```



Linear Model

Errors in the variables: the attenuation bias does not vanish

- Error on the predictor: $\hat{\beta}_2$ converges to $\beta_2(1 - \sigma_\nu^2/\sigma_x^2)$ (here $2 \times (1 - 0.5/1.5) \approx 1.33$), whatever n is

```
est <- rep(NA, B)
size <- c(50, 500, 5000)
for (nn in 1:length(size)) {
  for(b in 1:B){
    xs <- rnorm(size[nn])
    x_obs <- xs + rnorm(size[nn], 0, sqrt(s2nu))
    ys <- beta_1 + beta_2 * xs + rnorm(size[nn])
    est[b] <- coef(lm(ys ~ x_obs))[2]
    }
  print(paste0("n = ", size[nn], " mean of hat_b2 = ", round(mean(est), 3)))
}
```

```
## [1] "n = 50 mean of hat_b2 = 1.334"
## [1] "n = 500 mean of hat_b2 = 1.334"
## [1] "n = 5000 mean of hat_b2 = 1.333"
```

Linear Model

2. Omitted variables bias

- Suppose that the true model is:

$$y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$$

where $E[\varepsilon|x, z] = 0$ and the model can be estimated consistently using OLS

- **Consider that z is unobserved**

- The model we can estimate is

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i^*$$

where $\varepsilon_i^* = \varepsilon_i + \beta_3 z_i$

Linear Model

2. Omitted variables bias

- The omission of a relevant covariate ($\beta_3 \neq 0$) has two consequences

1. The variance of the error is

$$\sigma^2_{\varepsilon^*} = \beta_3^2 \sigma_z^2 + \sigma_\varepsilon^2$$

and it is greater than the one of the initial model for which z is observed and used as a covariate

2. the covariance between the error and x is

$$\text{cov}(x, \varepsilon^*) = \text{cov}(x, \varepsilon + \beta_3 z) = \beta_3 \text{cov}(x, z)$$

if the covariate is correlated with the omitted variable, the covariate and the error of the model are correlated.

Linear Model

2. Omitted variables bias: the OLS estimator

- The true model is $y_i = \beta_1 + \beta_2 x_i + \beta_3 z_i + \varepsilon_i$, but we regress y on x only. The estimator is the usual OLS slope

$$\hat{\beta}_2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Taking the average of the true model and subtracting it from the model itself (the intercept cancels):

$$y_i - \bar{y} = \beta_2(x_i - \bar{x}) + \beta_3(z_i - \bar{z}) + (\varepsilon_i - \bar{\varepsilon})$$

- Substituting in the numerator:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \beta_2 \sum_{i=1}^{n}(x_i - \bar{x})^2 + \beta_3 \sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z}) + \sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon})$$

- Dividing by $\sum_{i=1}^{n}(x_i - \bar{x})^2$ we get the decomposition in the next slide. Since $\sum_{i=1}^{n}(x_i - \bar{x}) = 0$, the means can be dropped when they multiply the deviations of x, e.g. $\sum_{i=1}^{n}(x_i - \bar{x})\bar{\varepsilon} = 0$, so that $\sum_{i=1}^{n}(x_i - \bar{x})(\varepsilon_i - \bar{\varepsilon}) = \sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i$ (and similarly $\sum_{i=1}^{n}(x_i - \bar{x})(\beta_3 z_i + \varepsilon_i) = \sum_{i=1}^{n}(x_i - \bar{x})(\beta_3(z_i - \bar{z}) + \varepsilon_i)$)

Linear Model

2. Omitted variables bias

- As the variance of the OLS estimator is proportional to the variance of the errors, omission of a relevant covariate will (in general) induce a less precise estimation of the slopes and a lower R^2

- Moreover, if the omitted covariate is correlated with the covariate used in the regression, the estimation will be biased and inconsistent

- This omitted variable bias can be computed as follows:

$$\begin{aligned}\hat{\beta}_2 &= \beta_2 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_3 z_i + \varepsilon_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ &= \beta_2 + \beta_3 \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\varepsilon_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\end{aligned}$$

- Taking the conditional expectation given x and z, the last term disappears

Linear Model

2. Omitted variables bias: why the last term disappears

- Conditioning on x and z, the quantities x_i, z_i, $\bar{x}$, $\bar{z}$ are treated as constants and come out of the expectation: the only random terms are the errors ε_i

- The assumption of the true model is $\mathbb{E}(\varepsilon_i|x,z) = 0$. Then, for the last term,

$$\mathbb{E}\left[\left.\frac{\sum_{i=1}^n (x_i-\bar{x})\varepsilon_i}{\sum_{i=1}^n (x_i-\bar{x})^2}\right| x,z\right] = \frac{\sum_{i=1}^n (x_i-\bar{x})\,\mathbb{E}(\varepsilon_i|x,z)}{\sum_{i=1}^n (x_i-\bar{x})^2} = \frac{\sum_{i=1}^n (x_i-\bar{x})\cdot 0}{\sum_{i=1}^n (x_i-\bar{x})^2} = 0$$

- The other terms do not change, since they depend only on β_2, β_3, x and z. Dividing numerator and denominator of the middle term by n we get the sample covariance and variance, so that

$$\mathbb{E}(\hat{\beta}_2|x,z) = \beta_2 + \beta_3 \widehat{\text{cov}}(x,z)/\hat{\sigma}_x^2$$

- The bias comes from the middle term: it is zero only if $\beta_3 = 0$ or $\widehat{\text{cov}}(x,z) = 0$. Therefore, there is

1. an upward bias if the signs of the covariance between x and z and β_3 are the same
2. a downward bias if they have opposite signs

Linear Model

2. Omitted variables bias: a small simulation

- True model:

$$y_i = 1 + 2x_i + 3z_i + \varepsilon_i$$

- $z_i, \varepsilon_i, u_i \sim \mathcal{N}(0, 1)$, independent of each other
- $x_i = \rho z_i + \sqrt{1 - \rho^2}\, u_i$, with $\rho = 0.5$ or $\rho = 0$. Then $\text{var}(x) = \rho^2 + (1 - \rho^2) = 1$ and $\text{cov}(x, z) = \rho\, \text{var}(z) = \rho$
- Three fits:

1. full model with x and z;
2. z omitted and $\text{cov}(x, z) = 0.5$;
3. z omitted and $\text{cov}(x, z) = 0$

- Expected limit in 2.: $\beta_2 + \beta_3\, \text{cov}(x, z)/\text{var}(x) = 2 + 3 \times 0.5 = 3.5$

```
set.seed(1)
beta_1 <- 1; beta_2 <- 2; beta_3 <- 3; nobs <- 200; B <- 1000
gen <- function(n, rho) {                          # rho = cov(x, z), var(x) = 1
  z <- rnorm(n)
  x <- rho * z + sqrt(1 - rho^2) * rnorm(n)
  y <- beta_1 + beta_2 * x + beta_3 * z + rnorm(n)
  data.frame(x, z, y)
}
```

Linear Model

2. Omitted variables bias: a small simulation

```
est <- r2 <- matrix(NA, B, 3,
                    dimnames = list(NULL, c("full", "omit_corr", "omit_uncorr")))
for (b in 1:B) {
  d  <- gen(nobs, 0.5)
  d0 <- gen(nobs, 0)
  f1 <- lm(y ~ x + z, d)
  f2 <- lm(y ~ x, d)
  f3 <- lm(y ~ x, d0)
  est[b, ] <- c(coef(f1)["x"], coef(f2)["x"], coef(f3)["x"])
  r2[b, ]  <- c(summary(f1)$r.squared, summary(f2)$r.squared, summary(f3)$r.squa
}
colMeans(est)           # mean of the estimates of beta_2
```

```
##        full   omit_corr omit_uncorr
##    2.002181    3.494401    1.982486
```

```
apply(est, 2, sd)       # precision
```

```
##        full   omit_corr omit_uncorr
##   0.0821090   0.1950371   0.2282719
```

```
colMeans(r2)            # R^2
```

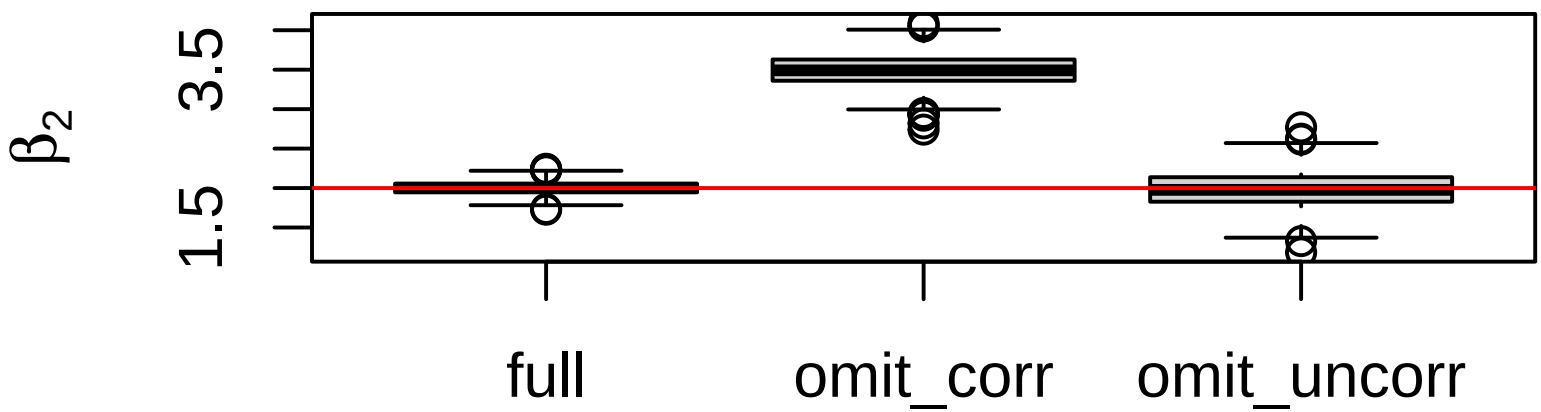
```
##        full   omit_corr omit_uncorr
##   0.9500063   0.6119085   0.2851264
```

Linear Model

2. Omitted variables bias: a small simulation

- Omitting z uncorrelated with x: no bias, but larger variance and lower R^2
- Omitting z correlated with x: $\hat{\beta}_2$ converges to $\beta_2 + \beta_3 \, \text{cov}(x, z)/\text{var}(x)$

```
boxplot(est, ylab = expression(hat(beta)[2]))
abline(h = beta_2, col = "red")
```



Linear Model

2. Omitted variables bias: the bias does not vanish

- Omitting z correlated with x: $\hat{\beta}_2$ converges to $\beta_2 + \beta_3 \, \text{cov}(x, z)/\text{var}(x)$, whatever n is

```
est <- rep(NA, B)
for (nn in c(50, 500, 5000)) {
  for(b in 1:B){
  est[b] <- coef(lm(y ~ x, gen(nn, 0.5)))["x"]
  }
  print(paste0("n = ", nn, " mean of hat_b2 = ", round(mean(est), 3)))
}
```

```
## [1] "n = 50 mean of hat_b2 = 3.484"
## [1] "n = 500 mean of hat_b2 = 3.496"
## [1] "n = 5000 mean of hat_b2 = 3.501"
```

Econometric Example

2. Returns from education: how to read β_2

- Mincer earnings function:

$$\log(w_i) = \beta_1 + \beta_2 e_i + \beta_3 s_i + \beta_4 s_i^2 + \varepsilon_i$$

  where w is the wage, e the years of education and s the experience

- Holding experience fixed, only e changes and

$$\frac{\partial \log(w)}{\partial e} = \beta_2$$

- Since $d\log(w) = \frac{dw}{w}$ (derivative of the logarithm), we get

$$\beta_2 = \frac{dw/w}{de}$$

  that is, β_2 is the **relative change of the wage** for one more year of education

- Multiplying by 100 gives the percentage change (an approximation, good for small β_2)
- Exactly, for one more year: $\frac{w_1}{w_0} = e^{\beta_2}$, that is a change of $100\,(e^{\beta_2} - 1)\%$
- **Example**: $\beta_2 = 0.10 \implies$ about $+10\%$ (exactly $e^{0.10} - 1 \approx 10.5\%$)

Econometric Example

2. Returns from education

- Sample of 303 white males taken from the National Longitudinal Survey of Youth in 1992

- Variables:

1. **wage**: Log hourly wage
2. **educ**: Education (years of schooling)
3. **AFQT**: Ability based on standardized AFQT test score
4. **educSibl**: Education of oldest sibling (years of schooling)
5. **exper**: Experience (total weeks of labor market experience)
6. **tenure**: Tenure (weeks on current job)
7. **educM**: Mother's education (years of schooling)
8. **educF**: Father's education (years of schooling)
9. **urban**: Dummy variable for urban residence
10. **home**: Broken home dummy (dummy for living with both parents at age 14)

```
data <- read.table("kpt.dat")
colnames(data) <- c("wage", "educ", "AFQT", "educSibl", "exper",
                    "tenure", "educM", "educF", "urban", "home")
```

Econometric Example

2. Returns from education

- Let's fit an OLS regression: one more year of education increases the average wage by 10%, holding fixed the experience

```
data$exper <- data$exper/52

fit1 <- lm(wage ~ educ + poly(exper, 2, raw = T), data = data)
summary(fit1)$coefficients
```

```
##                                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)                0.698683243 0.445792129  1.5672848 1.181058e-01
## educ                       0.100037599 0.011796664  8.4801601 1.049101e-15
## poly(exper, 2, raw = T)1 -0.014636178 0.078373261 -0.1867496 8.519836e-01
## poly(exper, 2, raw = T)2  0.003813165 0.003591661  1.0616717 2.892415e-01
```

Econometric Example

2. Returns from education

- Concern: individuals have different abilities (a); higher ability has a positive effect on wage, but may also have a positive effect on education

- If so, adding ability in the model

$$\log(w_i) = \beta_1 + \beta_2 e_i + \beta_3 s_i + \beta_4 s_i^2 + \beta_5 a_i + \varepsilon_i$$

will provide $\beta_5 > 0$ and regressing ability on the education, that is

$$a_i = \beta_1^* + \beta_2^* e_i + \varepsilon_i^*$$

will provide $\beta_2^* > 0$

- Thus

$$\hat{\beta}_2 \xrightarrow{p} \beta_2 + \beta_5 \times \beta_2^* > \beta_2$$

  and the OLS estimator is upward biased

Econometric Example

2. Returns from education

- This is the case, because more education

1. increases, for a given level of ability, the expected wage is β_2
2. on average, the level of ability is higher, this effect being β_2^*, so the wage will also be higher

- In our data set we have the (proxy of) ability (**AFQT**), which is the standardized AFQT test score.

- If we introduce ability in the regression, education is no more endogenous and least squares will give a consistent estimation of the effect of education on wage.

- We first check that education and ability are positively correlated:

```
with(data, cor(educ, AFQT))
```

```
## [1] 0.6055756
```

Econometric Example

2. Returns from education

- Adding ability as a covariate should decrease the coefficient on education
- The effect of one more year of education is now an increase of 8.7% of the wage (compared to the 10%)
- The relatively small decrease suggests that the correlation between education and ability is limited, so the omitted variable bias in this example is not very large

```
fit2 <- lm(wage ~ educ + poly(exper, 2, raw = T) + AFQT, data = data)
summary(fit2)$coefficients
```

```
##                                Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)                0.826938825 0.457647134  1.8069354 7.178086e-02
## educ                       0.088748184 0.014981282  5.9239381 8.660934e-09
## poly(exper, 2, raw = T)1 -0.011399438 0.078353913 -0.1454865 8.844251e-01
## poly(exper, 2, raw = T)2  0.003555693 0.003594909  0.9890910 3.234210e-01
## AFQT                       0.046939742 0.038447331  1.2208843 2.230949e-01
```

Linear Model

3. Simultaneity bias

- Often in economics, the phenomenon of interest is not described by a single equation, but by a system of equations

- Consider for example a market equilibrium. The two equations relate the quantity demanded / supplied (q^d and q^s) to the unit price and to some covariates specific to the demand and to the supply side of the market

- The equilibrium on the loan market is then defined by a system of three equations:

$$\begin{cases} q^d = \beta_1^d + \beta_2^d p + \beta_3^d d + \varepsilon^d \\ q^s = \beta_1^s + \beta_2^s p + \beta_3^s s + \varepsilon^s \\ q^d = q^s \end{cases}$$

where

1. q is the quantity (in logarithm)
2. p is the price
3. d and s two rates

To be continued...