Intermediate Econometrics

Lecture 8

Vincenzo Gioia

11 November 2026

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 respectively) to the unit price and to some covariates specific to the demand and to the supply side of the market

- The equilibrium is 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) and p the price (in logarithm)
2. d and s are variables that shift the demand and the supply, respectively
3. $\beta_2^d < 0$ (demand curve decreasing) and $\beta_2^s > 0$ (supply curve increasing)

- What we observe for each observation is **only the equilibrium** price-quantity combination (p_i, q_i)

Linear Model

3. Simultaneity bias: what happens when a shock hits

- A positive shock ε^d moves the demand curve upward and leads to a new equilibrium with a higher price p' and a higher quantity q'

- Hence p is **correlated with the demand error** ε^d, and OLS applied to the demand equation is biased

- Two special cases:

1. vertical supply (zero price elasticity): the price increases, the quantity does not change
2. horizontal supply (infinite price elasticity): the quantity increases, the price does not change (and then p and ε^d are uncorrelated)

- The same reasoning applies to the supply equation: a positive shock ε^s moves the supply curve and **lowers** the equilibrium price, so p is correlated also with ε^s

- The observed points are intersections of curves that keep shifting: a cloud of points that does not trace either curve

Linear Model

3. Simultaneity bias: the equilibrium price

- Setting $q^d = q^s$ and solving for p:

$$\beta_1^d + \beta_2^d p + u^d = \beta_1^s + \beta_2^s p + u^s$$
$$(\beta_2^s - \beta_2^d)\, p = (\beta_1^d - \beta_1^s) + u^d - u^s$$

  where $u^d = \beta_3^d d + \varepsilon^d$ and $u^s = \beta_3^s s + \varepsilon^s$ are the "demand" and "supply" disturbances, assumed independent

- Therefore, with $\Delta = \beta_2^s - \beta_2^d > 0$,

$$p = \frac{(\beta_1^d - \beta_1^s) + u^d - u^s}{\Delta}$$

- The price depends on **both** errors:

$$\text{cov}(p, \varepsilon^d) = \frac{\sigma^2_{\varepsilon^d}}{\Delta} > 0, \qquad \text{cov}(p, \varepsilon^s) = -\frac{\sigma^2_{\varepsilon^s}}{\Delta} < 0$$

Linear Model

3. Simultaneity bias: what does OLS estimate?

- Regress q on p. Since $q = \beta_1^s + \beta_2^s p + u^s$,

$$\hat{\beta}_2^{OLS} \xrightarrow{p} \frac{\text{cov}(p,q)}{\text{var}(p)} = \beta_2^s + \frac{\text{cov}(p,u^s)}{\text{var}(p)} = \beta_2^s - \frac{\Delta\, V_S}{V_D + V_S}$$

with $V_D = \text{var}(u^d)$ and $V_S = \text{var}(u^s)$ (using $\text{var}(p) = (V_D + V_S)/\Delta^2$ and $\text{cov}(p,u^s) = -V_S/\Delta$)

- Replacing $\Delta = \beta_2^s - \beta_2^d$ we get

$$\hat{\beta}_2^{OLS} \xrightarrow{p} \frac{\beta_2^s\, V_D + \beta_2^d\, V_S}{V_D + V_S}$$

- OLS estimates a **weighted average of the supply and of the demand slope**, not the demand slope β_2^d

- If the demand shocks dominate (V_D large) the demand curve moves a lot and traces the supply curve: OLS is close to β_2^s;

- If the supply shocks dominate, OLS is close to β_2^d

Linear Model

3. Simultaneity bias: a small simulation

- Demand: $q^d = 10 - 1.0\,p + 1.0\,d + \varepsilon^d$
- Supply: $q^s = 2 + 1.5\,p + 1.0\,s + \varepsilon^s$
- With $d, s, \varepsilon^d, \varepsilon^s \sim \mathcal{N}(0, 1)$
- For each observation we solve $q^d = q^s$ and **observe only the equilibrium** (p_i, q_i):

$$p_i = \frac{(10 - 2) + d_i - s_i + \varepsilon_i^d - \varepsilon_i^s}{1.5 - (-1.0)}$$

Linear Model

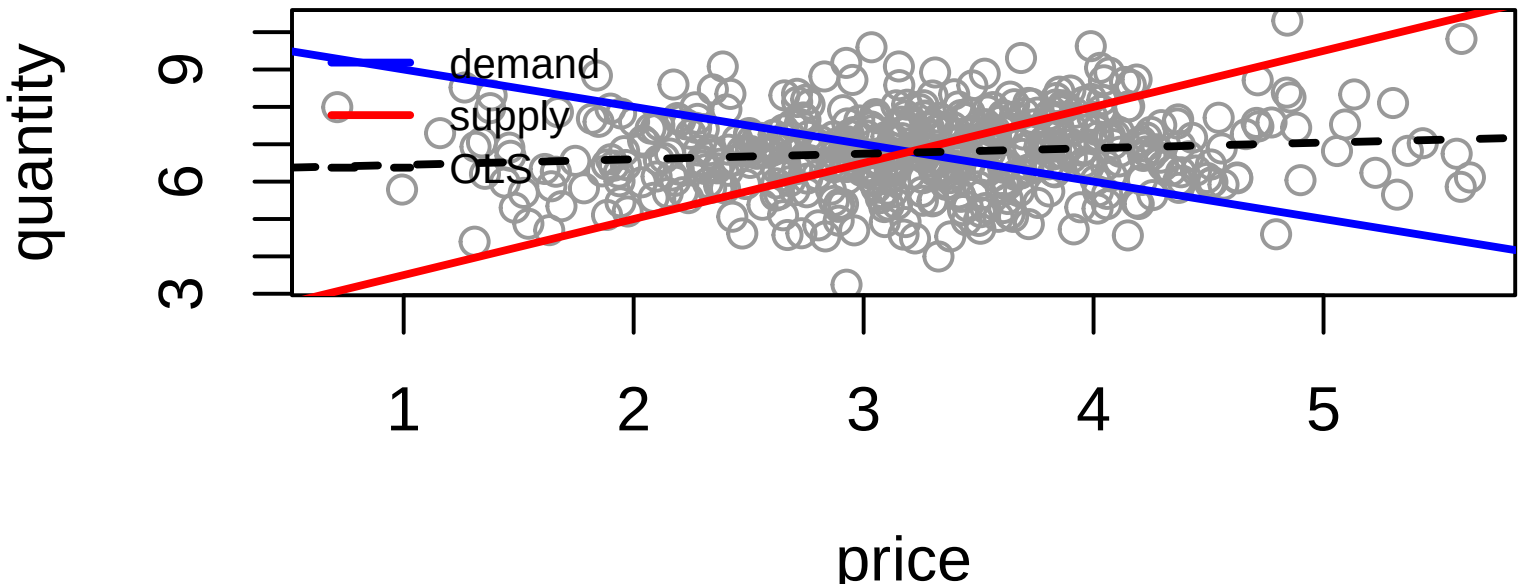
3. Simultaneity bias: a small simulation

```
set.seed(1)
n <- 500
bd <- c(10, -1.0, 1.0)
bs <- c(2, 1.5, 1.0)
d <- rnorm(n)
s <- rnorm(n)
ed <- rnorm(n)
es <- rnorm(n)
p <- ((bd[1] - bs[1]) + bd[3]*d - bs[3]*s + ed - es) / (bs[2] - bd[2])
q <- bs[1] + bs[2]*p + bs[3]*s + es                  # equilibrium quantity
```

Linear Model

3. Simultaneity bias: a small simulation

```
plot(p, q, col = "grey60", xlab = "price", ylab = "quantity")
abline(bd[1], bd[2], col = "blue", lwd = 2)       # demand (d = 0, no shock)
abline(bs[1], bs[2], col = "red",  lwd = 2)       # supply (s = 0, no shock)
abline(lm(q ~ p), lwd = 2, lty = 2)               # OLS
legend("topleft", c("demand", "supply", "OLS"), lwd = 2, lty = c(1, 1, 2),
       col = c("blue", "red", "black"), bty = "n", cex = 0.7)
```



Linear Model

3. Simultaneity bias: what OLS estimates

- The cloud of points is made of **intersections of curves that keep shifting**: the OLS line is neither the demand nor the supply curve, but a mixture of the two

- The price is correlated with both errors: $\text{cov}(p, \varepsilon^d) > 0$ and $\text{cov}(p, \varepsilon^s) < 0$ (a positive demand shock raises p and q)

```
coef(lm(q ~ p))["p"]        # true demand slope: -1.0; true supply slope: +1.5
```

```
##         p
## 0.1488941
```

```
cor(p, ed); cor(p, es)
```

```
## [1] 0.5053838
```

```
## [1] -0.5004643
```

Linear Model

3. Simultaneity bias: the weight of the shocks

- The OLS slope is a weighted average of the demand and supply slopes:
  - it is closer to the supply slope when the **demand** shocks are large (the demand curve moves a lot and traces the supply curve)
  - it is closer to the demand slope when the **supply** shocks are large

```
ols_slope <- function(sd_d, sd_s, n = 1e5) {
  d <- rnorm(n)
  s <- rnorm(n)
  ed <- rnorm(n, 0, sd_d)
  es <- rnorm(n, 0, sd_s)
  p <- ((bd[1] - bs[1]) + bd[3]*d - bs[3]*s + ed - es) / (bs[2] - bd[2])
  q <- bs[1] + bs[2]*p + bs[3]*s + es
  coef(lm(q ~ p))["p"]
}
```

Linear Model

3. Simultaneity bias: the weight of the shocks

- The OLS slope is a weighted average of the demand and supply slopes:
  - it is closer to the supply slope when the **demand** shocks are large (the demand curve moves a lot and traces the supply curve)
  - it is closer to the demand slope when the **supply** shocks are large

```
# varying sd of eps^d
sapply(c(0.3, 1, 3, 10), function(sd_d) ols_slope(sd_d, sd_s = 1))
```

```
##          p          p          p          p
## -0.1144756  0.2485118  1.0849837  1.4517865
```

```
# varying sd of eps^s
sapply(c(0.3, 1, 3, 10), function(sd_s) ols_slope(sd_d = 1, sd_s))
```

```
##          p          p          p          p
##  0.6110347  0.2468782 -0.5812709 -0.9518287
```

Linear Model

Summary and examples of the three sources of endogeneity

Source	Example (y on x)	Why $\text{cov}(x, \varepsilon) \neq 0$
Errors in variables	log(wage) on IQ (measuring ability)	we observe IQ $= a + \nu$ instead of a: the measurement error ν ends up in ε and is correlated with IQ $\Rightarrow$ attenuation
Omitted variable	log(wage) on educ (kpt.dat)	ability is not observed, so it is in ε; it raises wages and is correlated with educ
Omitted variable	log(wage) on educ (Card, 1995)	ability and family background are in ε; they raise wages and are correlated with educ
Simultaneity	log(q) on log(p) (Fulton fish market)	p and q are determined jointly: a demand shock ε^d moves q and, through the equilibrium, also p

- In all cases $\text{cov}(x, \varepsilon) \neq 0$ and OLS is biased and inconsistent, **whatever the sample size**
- Solutions: (i) more or better controls, when they are available; (ii) **instrumental variables**, that find a source of variation of x unrelated to the error

Econometric Example

1. Errors in the variables: the data

- Data: **wage2** (package `wooldridge`): 935 men in the US in 1980 (Blackburn and Neumark, 1992)

- Main variables:

1. **wage**: monthly earnings; **lwage**: log(wage)
2. **educ**: years of education; **exper**: years of work experience; **tenure**: years with the current employer
3. **IQ**: IQ score; **KWW**: score on the "Knowledge of the World of Work" test
4. **age**, **married** (=1 if married), **black** (=1 if black), **south** (=1 if living in the South), **urban** (=1 if living in an SMSA)
5. **hours**: weekly hours of work; **sibs**: number of siblings; **brthord**: birth order; **meduc**, **feduc**: mother's and father's education

- Innate **ability is not observed**: IQ and KWW are only two noisy measurements of it

Econometric Example

1. Errors in the variables: ability measured by IQ

- Model of interest, where ability a_i is **not observed**:

$$\log(w_i) = \beta_1 + \beta_2 \, \text{educ}_i + \beta_3 \, \text{exper}_i + \cdots + \beta_k \, a_i + \varepsilon_i$$

- We only observe a noisy measurement of ability, the **IQ score**:

$$\text{IQ}_i = a_i + \nu_i$$

that is the classical errors-in-variables setting, with the error ν_i uncorrelated with a_i

- Consequences of using IQ in place of a:

1. the coefficient of IQ is **attenuated** towards zero
2. ability is only partly controlled for: since IQ and education are positively correlated, the coefficient of educ is still **too large** (part of the ability effect is left in it)

Econometric Example

1. Errors in the variables: ability measured by IQ

- OLS with IQ used as a proxy for ability (controls: experience, tenure, married, south, urban, black):

```
library(wooldridge)
```

```
## Warning: il pacchetto 'wooldridge' è stato creato con R versione 4.5.3
```

```
data("wage2")
fit_iq <- lm(lwage ~ educ + exper + tenure +
               married + south + urban + black + IQ,
             data = wage2)
round(summary(fit_iq)$coefficients[c("educ", "IQ"), ], 4)
```

```
##      Estimate Std. Error t value Pr(>|t|)
## educ   0.0544     0.0069  7.8532    0e+00
## IQ     0.0036     0.0010  3.5885    4e-04
```

- One more year of education: about $+5.4\%$ of the wage; one more IQ point: about $+0.36\%$
- But the IQ coefficient is attenuated and, as a consequence, the education one is probably overestimated

Econometric Example

1. Errors in the variables: a second measurement

- The dataset contains a second, different indicator of ability: **KWW** (Knowledge of the World of Work score)

- IQ and KWW are both noisy measures of the same unobserved ability: they are positively correlated...

```
with(wage2, cor(IQ, KWW))
```

```
## [1] 0.4135155
```

- ...but their measurement errors are (plausibly) unrelated

- **Idea**: KWW is correlated with IQ (through ability) and its error is unrelated to the error of IQ; it can be used as an **instrumental variable** for IQ (we will see later)

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
```

Econometric Example

2. Returns from education (Card, 1995)

- Data: **SchoolingReturns** (package `ivreg`): 3010 men from the National Longitudinal Survey of Young Men, 1976

- Variables:

1. **wage**: hourly wage (in cents); we use log(wage)
2. **education**: years of schooling
3. **experience**: years of labor market experience
4. **ethnicity**: African American (`afam`) or other
5. **smsa**, **south**: lives in a metropolitan area / in the South
6. **nearcollege**: grew up near a four-year college (`yes`/`no`)

- Ability is **not** in the dataset

Econometric Example

2. Returns from education (Card, 1995)

- OLS of log(wage) on education and the observed controls:

```
data("SchoolingReturns", package = "ivreg")

ols_ed <- lm(log(wage) ~ education + experience + I(experience^2) +
               ethnicity + smsa + south, data = SchoolingReturns)
round(summary(ols_ed)$coefficients["education", ], 4)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##     0.0740     0.0035    21.1126     0.0000
```

- One more year of schooling is associated with an increase of about 7.4% in the wage
- Is this the **causal** effect of education?

Econometric Example

2. Returns from education (Card, 1995)

- Ability (motivation, family background...) is unobserved and goes into the error:

$$\log(w_i) = \beta_1 + \beta_2 e_i + \cdots + \underbrace{\beta_a a_i + \varepsilon_i}_{u_i}$$

- More able individuals earn more ($\beta_a > 0$) and, on average, study more ($\beta_2^* > 0$ in $a_i = \beta_1^* + \beta_2^* e_i + \varepsilon_i^*$)

- Thus

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

  and OLS would overestimate the return to education

- **But** education is also measured with error, which pushes the OLS estimate in the opposite direction (attenuation): the net sign of the bias is not known a priori

Econometric Example

2. Returns from education: what can we do?

1. **A proxy for ability as a control** (e.g. a test score such as AFQT): if good enough, the omitted variable problem is (almost) solved by OLS with more controls.

2. **No good proxy available**: we need an exogenous source of variation in education, that is an **instrumental variable**

- Card's idea: the distance from a college changes the cost of studying, but (plausibly) it does not affect the wage in other ways

Econometric Example

3. Fulton Fish Market (Graddy, 1995)

- Data: **fish**, 97 days of trading of whiting at the Fulton Fish Market (New York), December 1991 – May 1992

- Variables:

1. **ltotqty**: log of the total quantity sold in the day
2. **lavgprc**: log of the average price in the day
3. **mon, tues, wed, thurs**: day-of-the-week dummies
4. **wave2, wave3**: height of the waves at sea (recent days), weather conditions that affect fishing

- Demand for fish (log-log, so β_2^d is the price elasticity of demand):

$$\text{ltotqty}_t = \beta_1^d + \beta_2^d \, \text{lavgprc}_t + \beta_3^d \, \text{day}_t + \varepsilon_t^d$$

where the day dummies shift the demand (e.g. restaurants buy more on some days)

Econometric Example

3. Fulton Fish Market: OLS

- Price and quantity are set **together** by the market: the price of a day reflects both the demand and the supply (fishing) shocks of that day

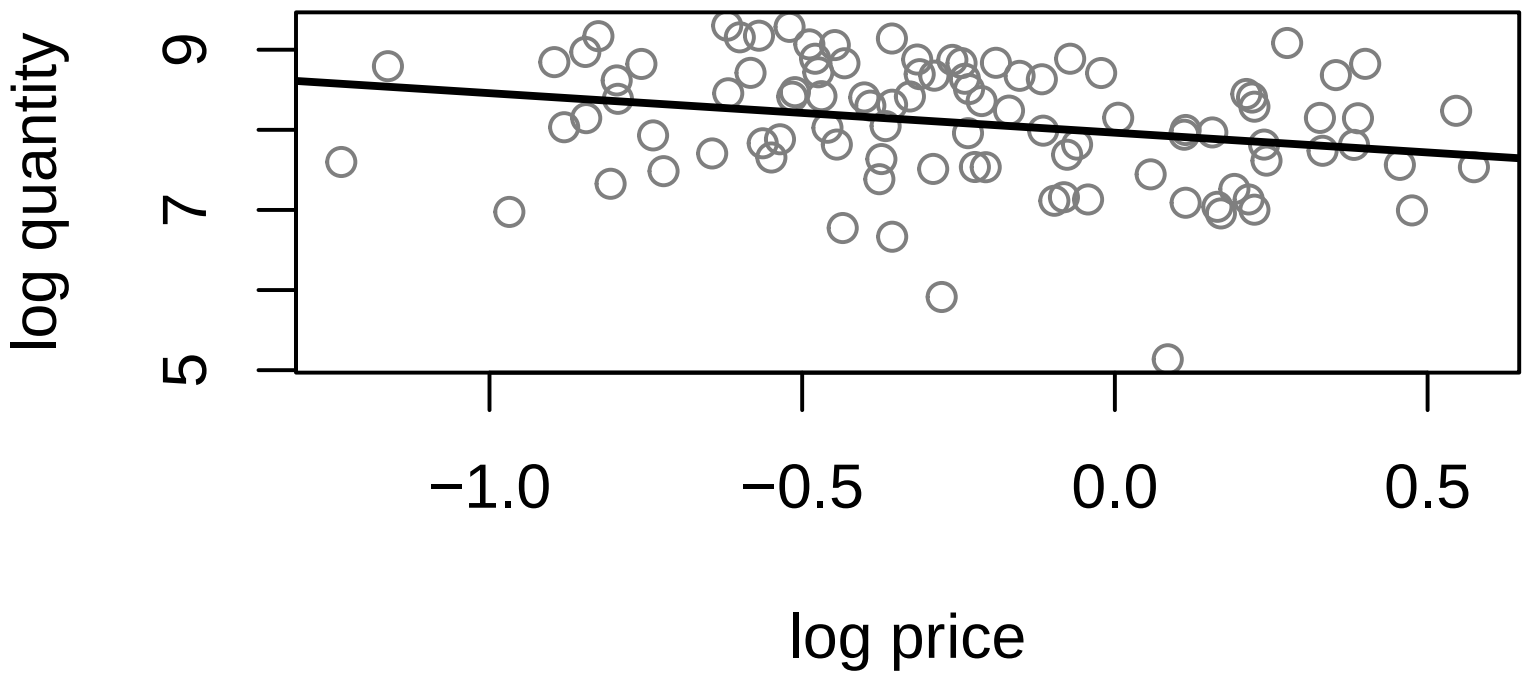
```
data("fish")
ols_fish <- lm(ltotqty ~ lavgprc + mon + tues + wed + thurs, data = fish)
round(summary(ols_fish)$coefficients["lavgprc", ], 4)
```

```
##   Estimate Std. Error    t value   Pr(>|t|)
##    -0.5247     0.1761    -2.9791     0.0037
```

Econometric Example

3. Fulton Fish Market: the data

```
plot(ltotqty ~ lavgprc, data = fish, col = "grey50",
     xlab = "log price", ylab = "log quantity")
abline(lm(ltotqty ~ lavgprc, data = fish), lwd = 2)
```



Econometric Example

3. Fulton Fish Market: reading the OLS result

- The OLS estimate of the price elasticity is -0.53 (standard error 0.18): the sign is the expected one for a demand curve
- However, the cloud of points is made of intersections of demand and supply curves, so the OLS slope is a mixture of the two (weighted average, see before): **we do not know which curve we are estimating**
- To estimate the demand curve we need something that **moves only the supply**, so that the price changes while the demand error does not: a *supply shifter*
- **The weather at sea is such a variable**: bad weather makes fishing harder, reduces the supply and raises the price, but it does not change how much New York buyers want to purchase