Financial Econometrics

Simple regression

Maximilian Rohrer

Overview

- Introduction
- Simple regression model
- Derive OLS regressors
- Estimating and interpreting the effect of class size on teaching outcome
- Binary regressors
- Hypothesis testing
- Goodness of fit

Introduction

Steps in empirical economic analysis

1. Careful formulation of research question (RQ)
2. Construct economic model
3. Turn into econometric model
   - Specify econometric model
   - Variable choices
   - Hypothesis development
   - Data gathering
   - Estimation of model parameters

Common data structures

- Cross-sectional data: sample of firms, households, firms, cities, states, countries, or other units of interest observed at a given point in time or in a given period
- Time-series data: observations on a given variable or several variables over time such as stock prices, money supply, consumer prices, GDP, etc.
- Panel data: the same cross sectional units are followed over time

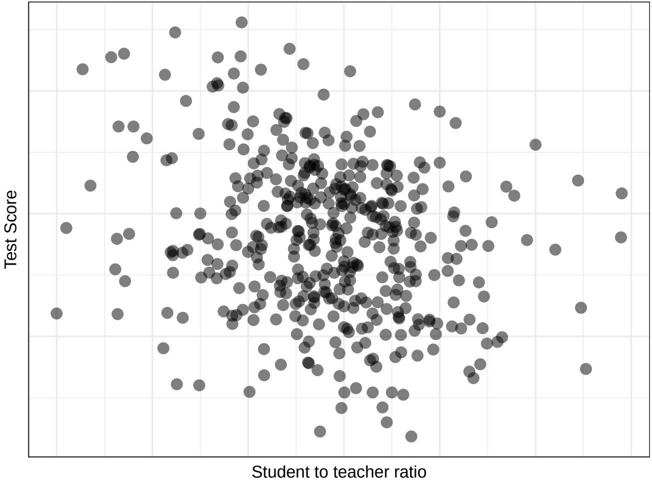
Simple regression model

Simple regression model

Does the class size matter for teaching outcome?

- Data: `School.Rdata`
- Teaching outcome - test score (testscr)
- Student-to-teacher ratio (str)

Let's look at the data



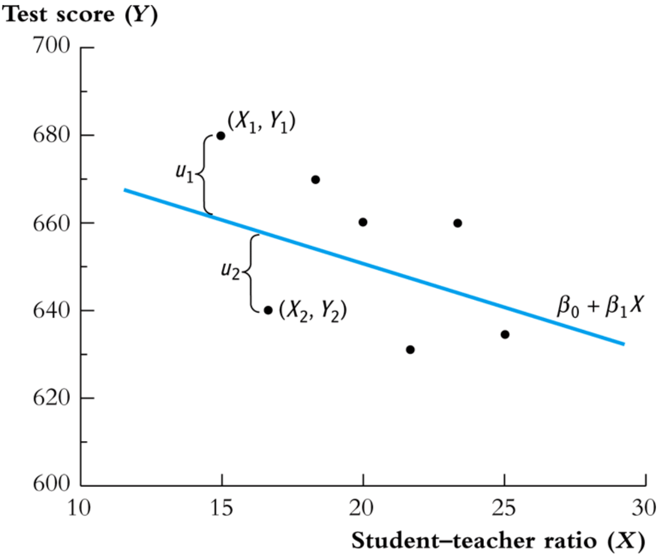
Simple regression model

$$y = \beta_0 + \beta_1 x + u \tag{1}$$

- y = dependent variable, explained variable, response variable, predicted variable, left-hand side variable (LHS)
- x = independent variable, explanatory variable, control variable, predictor variable, regressor, right-hand side variable (RHS)
- u = error term, disturbance, residual
- β_0 = intercept, constant
- $\beta_1 = \frac{\Delta y}{\Delta x}$ = slope coefficient; if x increases by one unit y increases by β units
- **Please note that in the recorded lecture I falsely say α instead of β_0 and β instead of β_1**

Simple regression model

- We assume a linear relationship in the population, but how to set the slope and intercept of the line?



Derive OLS estimator

Derive OLS estimators

- Our data (y_i, x_i) for $i = 1, \ldots, n$
- Assume linear population regression model

$$y_i = \beta_0 + \beta x_i + u_i$$

- Given estimates $\hat{\beta}_0$, $\hat{\beta}_1$ of β_0, β_1, we can predict y with $\hat{y}$

$$\hat{y}_i = \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- Given the prediction, we can cumpute the prediction error

$$\hat{u}_i = y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$$

- How to find best linear estimators $\hat{\beta}_0$, $\hat{\beta}_1$ of β_0, β_1?
- OLS minimizes the sum of squared residuals

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0, \beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \hat{u}_i^2$$

Derive OLS estimators

- Function to be minimized

$$\hat{\beta}_0, \hat{\beta}_1 = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \hat{u}_i^2 = \underset{\beta_0,\beta_1}{\operatorname{argmin}} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2$$

- Set partial derivatives to zero

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 = -2 * \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0$$

$$\frac{\partial}{\partial \beta_1} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right)^2 = -2 * \sum_{i=1}^{n} \left[\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) * x_i \right] = 0$$

- Apply following manipulations
  - $\sum_{i=1}^{n} x_i = \bar{x}$
  - $\sum_{i=1}^{n} x_i = \bar{y}$
  - multiply with $1/n$
  - multiply by 0.5

Derive OLS estimators

- After manipulations

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \hat{\beta}_0 \bar{x} - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0$$

Derive OLS estimator

- Substitude one equation within the other and express $\hat{\beta}_1$

$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i - (\bar{y} - \hat{\beta}_1 \bar{x})\bar{x} - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0$$

$$\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{y}\bar{x} + \hat{\beta}_1 \bar{x}^2 - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 \left(\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2 \right) = \frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{y}\bar{x}$$

$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^{n} y_i x_i - \bar{y}\bar{x}}{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \bar{x}^2} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Insert for $\hat{\beta}_0$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Effect of class size on teaching outcome

Effect of class size on teaching outcome

$$testscr = \beta_0 + \beta_1 str$$

```r
# load data
load("School.Rdata")

# inspect data
# summary(School) # use this on your machine
str(School)
```

```
## 'data.frame':    420 obs. of  18 variables:
##  $ observat: num  1 2 3 4 5 6 7 8 9 10 ...
##  $ dist_cod: num  75119 61499 61549 61457 61523 ...
##  $ county  : chr  "Alameda" "Butte" "Butte" "Butte" ...
##  $ district: chr  "Sunol Glen Unified" "Manzanita Elementary" "Therm
##  $ gr_span : chr  "KK-08" "KK-08" "KK-08" "KK-08" ...
##  $ enrl_tot: num  195 240 1550 243 1335 ...
##  $ teachers: num  10.9 11.1 82.9 14 71.5 ...
##  $ calw_pct: num  0.51 15.42 55.03 36.48 33.11 ...
##  $ meal_pct: num  2.04 47.92 76.32 77.05 78.43 ...
##  $ computer: num  67 101 169 85 171 25 28 66 35 0 ...
##  $ testscr : num  691 661 644 648 641 ...
##  $ comp_stu: num  0.344 0.421 0.109 0.35 0.128
```

Effect of class size on teaching outcome

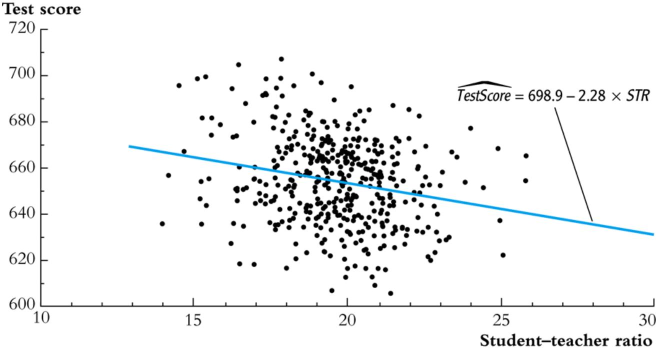
```r
# estimate beta.1
y_bar <- mean(School$testscr)
x_bar <- mean(School$str)

beta_hat_1 <- sum((School$str - x_bar)*(School$testscr - y_bar)) /
  sum((School$str - x_bar)^2)

# estimate beta.2
beta_hat_0 <- mean(School$testscr) - beta_hat_1 * x_bar

# display
beta_hat_0
```

```
## [1] 698.933
```

```r
beta_hat_1
```

```
## [1] -2.279808
```

Fitted regression line



Estimation with built-in regression function

```r
# estimate regression
fit <- lm(testscr ~ str, data = School)

# display result
summary(fit)
```

Estimation with built-in regression function

```
##
## Call:
## lm(formula = testscr ~ str, data = School)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.727 -14.251   0.483  12.822  48.540
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 698.9330     9.4675  73.825  < 2e-16 ***
## str          -2.2798     0.4798  -4.751 2.78e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
##
## Residual standard error: 18.58 on 418 degrees of freedom
## Multiple R-squared:  0.05124,    Adjusted R-squared:  0.04897
## F-statistic: 22.58 on 1 and 418 DF,  p-value: 2.783e-06
```

Let's answer some questions

- How much variation is there in *str* and *testscr*?
- How much does *testscr* change if *str* increases from 16 to 19 $(+3)$?
- How much does *testscr* change if *str* increases from 20 to 23 $(+3)$?
- Is this a big effect?
- How much does *testscr* change if there is a small class $str < 20$ rather than a big one?
- Are the effects significant?
- How well does the regression model describe the data?
- Is this linear model with only one explanatory variable a good model?

Let's answer some questions

How much variation is there in *str* and *testscr*?

```
summary(School$str)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   14.00   18.58   19.72   19.64   20.87   25.80
```

```
sd(School$str)
```

```
## [1] 1.891812
```

```
summary(School$testscr)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   605.5   640.0   654.5   654.2   666.7   706.8
```

```
sd(School$testscr)
```

```
## [1] 19.05335
```

Let's answer some questions

- How much does *testscr* change if *str* improves from 16 to 19 (+3)?

```
# predicted value for str = 16
y_x16 <- fit$coefficients[1] + fit$coefficients[2]*16
# predicted value for str = 19
y_x19 <- fit$coefficients[1] + fit$coefficients[2]*19
# difference
as.numeric(y_x19 - y_x16)
```

```
## [1] -6.839425
```

- How much does *testscr* change if *str* improves from 20 to 23 (+3)?

```
# predicted value for str = 20
y_x20 <- fit$coefficients[1] + fit$coefficients[2]*20
# predicted value for str = 23
y_x23 <- fit$coefficients[1] + fit$coefficients[2]*23
# difference
as.numeric(y_x23 - y_x20)
```

```
## [1] -6.839425
```

Let's answer some questions

- Linear model means that the change in y is independent from the level of x
- β_1 measures marginal effect on y for a unit change in x
- Simple way to get marginal effects on previous slide

```
fit$coefficients[2]*3
```

```
##       str
## -6.839425
```

Let's answer some questions

Is this a big effect?

- So-called economic significant
- Sometimes magnitudes are enough
- Express results in standard deviations

Is this a big effect?

```r
# std x variables
sd(School$str)
```

```
## [1] 1.891812
```

```r
# effect on y if x changes by one std
sd(School$str)*fit$coefficients[2]
```

```
##       str
## -4.312968
```

```r
# std y var
sd(School$testscr)
```

```
## [1] 19.05335
```

```r
# effect on y if x changes by one std expressed
# in std of the dependent var
sd(School$str)*fit$coefficients[2] / sd(School$testscr)
```

```
##        str
## -0.2263628
```

Let's answer some more questions

- How much does *testscr* change if there is a small class $str < 20$ rather than a big one? Binary regressors
- Are the effects significant? Hypothesis testing
- How well does the regression model describe the data? Measures of fit, R2
- Is this linear model with only one explanatory variable a good model? Multiple regression (next lecture)

Binary regressor

Binary regressor

$$y = \beta_0 + \beta_1 x + u$$

- Sometimes a regressor is binary
  - x = 1 if small class size, x = 0 if not
  - x = 1 if female, x = 0 if male
- Binary regressors are sometimes called dummy variables
- So far, β_1 has been called a "slope", but that does not make sense if x is binary

Binary regressors

$$y = \beta_0 + \beta_1 x + u$$

- Interpretation as conditional mean
- when $x_i = 0$, $y_i = \beta_0 + u_i$
  - the expected value for all i with $x_i = 0$ is β_0
  - $E(y_i|x_i = 0) = \beta_0$
- when $x_i = 1$, $y_i = \beta_0 + \beta_1 + u_i$
  - the expected value for all i with $x_i = 1$ is $\beta_0 + \beta_1$
  - $E(y_i|x_i = 1) = \beta_0 + \beta_1$
- Hence, β_1 is the difference in population group means,

$$\beta_1 = E(y_i|x_i = 1) - E(y_i|x_i = 0)$$

Example: test scores

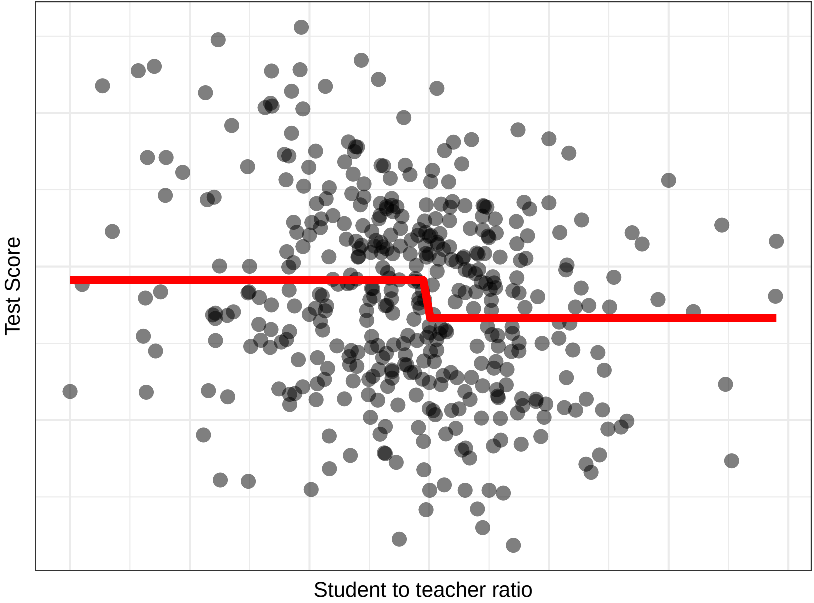
- Define small class size as < 20 students

```
# define small class size
School <- mutate(School, small = str < 20)
# School$small <- School$str < 20 # alternative code

# estimate regression
fit.bin <- lm(testscr ~ small, data = School)

# display result
fit.bin$coefficients
```

```
## (Intercept)   smallTRUE
##   649.97885     7.37241
```

$$\widehat{testscr} = 649.98 + 7.73 * small$$

Predicted regression



Hypothesis testing

Hypothesis testing

- Let's assume we know that in the population, the relationship between x and y can be described as:

$$y = 0 + 3 * x + u$$

- In other words, in the population $\beta_1 = 3$
- The econometrician does not know, he collects 100 observations on x and y, runs a regression, and gets following estimates $\hat{\beta}_0 = 0.2$ and $\hat{\beta}_1 = 2.87$

$$\hat{y} = 0.2 + 2.87 * x$$

Hypothesis testing

- Can he prove that $\beta_1 = 3$ is true?
- But maybe he can disprove that $\beta_1 = 0$? (H0)
- Note that we use sample parameters $\hat{\beta}_1$ to test hypotheses on the population parameter β_1
- If the population β_1 equals 0, what is the probability of observing an estimate $\hat{\beta}_1$ that is further away from the hypothesised value than our estimate $\hat{\beta}_1 = 2.87$? (p-value; probability of Type 1 error)
- If that probability is sufficiently small, shall we reject the hypothesis that the population parameter equals 0? (critical value)

Hypothesis testing - more formally

- Construct a hypothesis: remember, statistical inference relies on disproving and not proving!
- Null hypothessis and two-sided alternative
  - H0: $\beta = \beta_{H0}$ vs. H1: $\beta \neq \beta_{H0}$
- Null hypothessis and one-sided alternative
  - H0: $\beta = \beta_{H0}$ vs. H1: $\beta < \beta_{H0}$ or $\beta > \beta_{H0}$
- β_{H0} is the hypothesized value under the null. Often the null equals "no effect" which often impleis $\beta_{H0} = 0$

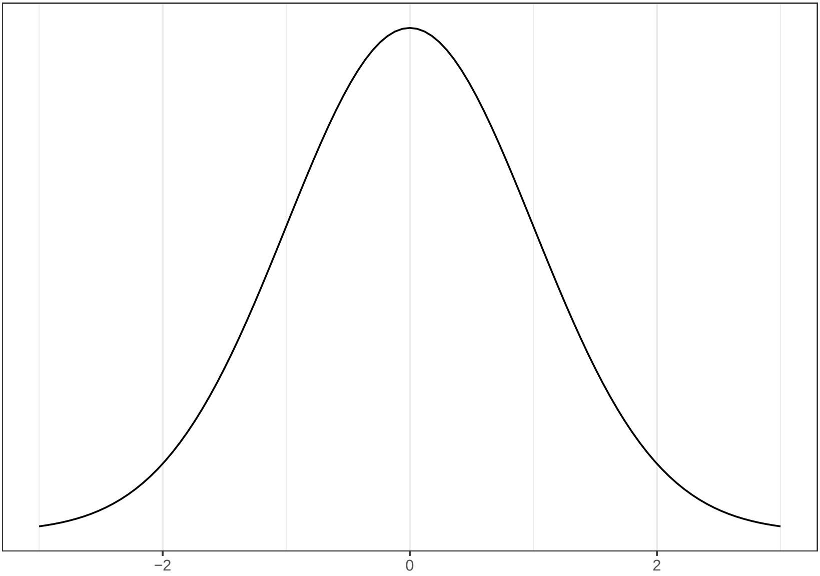
Hypothesis testing

- How far is $\hat{\beta}$ from β_{H0}? Distance and sampling error matters
- Distribution of the estimator

$$\frac{(\hat{\beta} - \beta_{H0})}{se(\hat{\beta})} \equiv t_{\hat{\beta}} \sim t_{df}$$

- Degrees of freedom df of t distribution (t_{df}) depends on the unknown parameters (intercept + slope) and the sample size
- If you have enough data, you can use normal distribution (Textbook p. 127)

Hypothesis testing



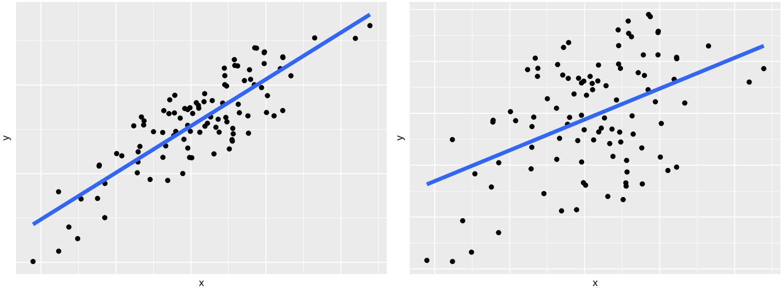
Summary of hypothesis testing on a single parameter two-sided

$$\frac{(\hat{\beta} - \beta_{H0})}{se(\hat{\beta})} \equiv t_{\hat{\beta}}$$

- Reject H0: $\hat{\beta} = \beta_{H0}$ at 5% significance if $|t| > 1.96$
- The p-value is $p = P(|t| > |t_{\hat{\beta}}|)$
- Reject H0: $\hat{\beta} = \beta_{H0}$ at 5% significance if p-value $< 5\%$
- 95% confidence interval for $\beta = \hat{\beta} \pm 1.96 \times se(\hat{\beta})$

Goodness of fit

Goodness of fit



- Total Sum of Squares (SST): $\sum_{i=1}^{N}(y_i - \bar{y})^2$
- Explained Sum of Squares (SSE): $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$
- Residual Sum of Squares (SSR): $\sum_{i=1}^{N} \bar{u}_i{}^2$
- SST = SSE + SSR
- R^2 = SSE/SST = 1 - SSR/SST

Goodness of fit, pitfalls

An increasing R2 does not mean that

- the effect becomes causal
- that your models suffers from less biases

```r
# extract from lm output
summary(fit)$r.squared
```

```
## [1] 0.0512401
```

```r
# compute yourself
var(fit$fitted.values) / var(fit$residuals +
                                fit$fitted.values)
```

```
## [1] 0.0512401
```

What are we missing?

- Is this linear model with only one explanatory variable a good model? Multiple regression (next lecture)
- We have not proved that using OLS to estimate the population parameters is valid
- Necessary assumptions will be covered in two weeks

Summary

- Introduction
- Simple regression model
- Derive OLS regressors
- Estimating and interpreting the effect of class size on teaching outcome
- Binary regressors
- Hypothesis testing
- Goodness of fit

Textbook

- Chapter 1-3, 4, 5, Stock and Watson, Introduction to Econometrics, Global Edition, 4th edition