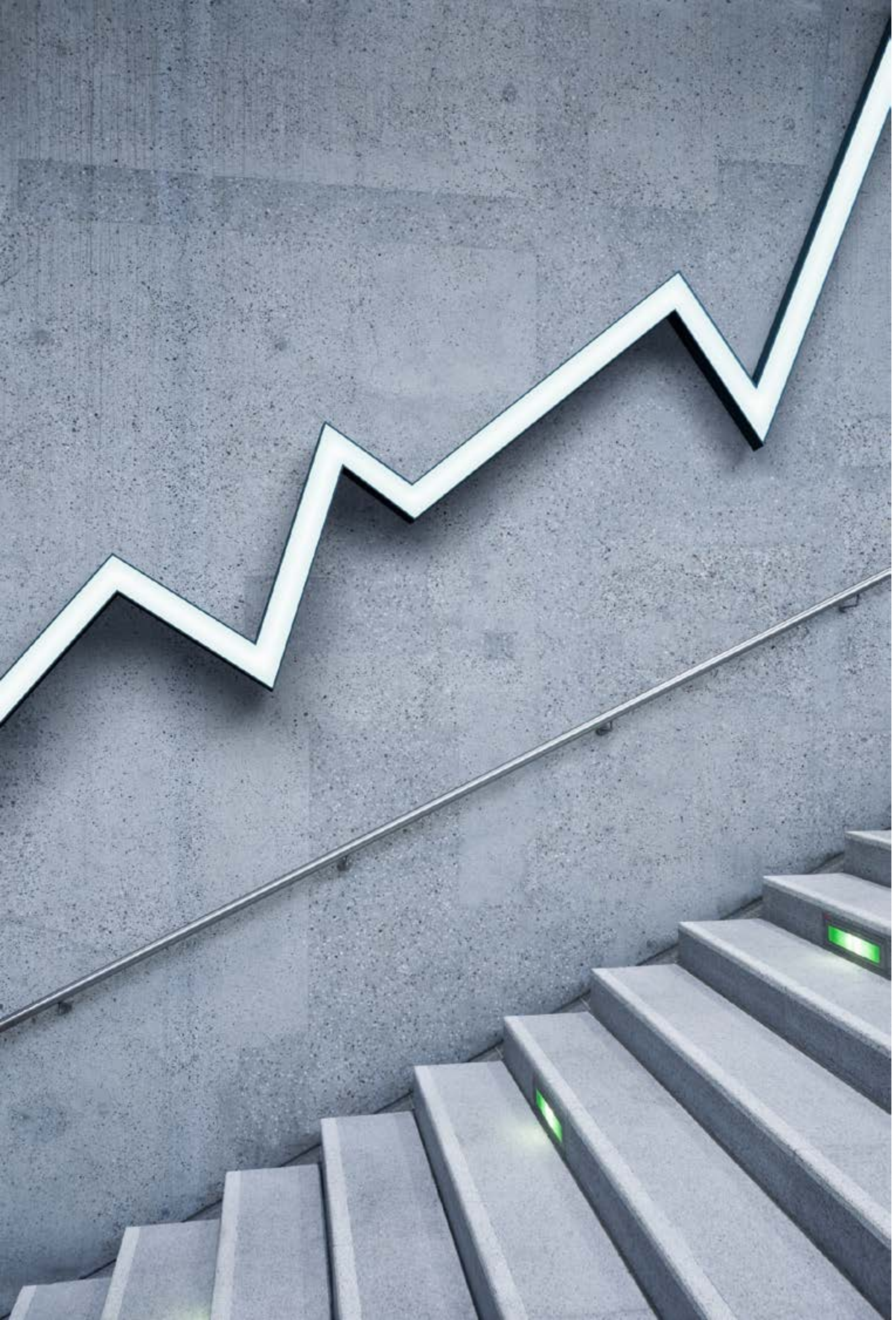


Lecture 1 The Simple Regression Model

LECTURED BY WENYA WANG

Outline

1. Definition of the simple regression model

2. Deriving the Ordinary-Least-Square (OLS) estimates

3. Properties of OLS estimates

4. Expected values and variances of OLS estimators

5. Issues on units of measurement and functional form

Introduction

Goal: study the relationship between two variables

1. Can a company's profit rate explain its stock return?

2. Can one's education level explain his/her earnings? ...

Explained variable (y):

stock return, earnings

Explanatory variable (x):

profit rate, education level

Error term μ
Unknowns

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

Example of error terms: 1. the competitiveness of the company in the industry
2. one's job experience

Introduction

Simple linear regression model

1. a.k.a., two-variable linear regression model or bivariate linear regression model
2. unknown factors in the big residual bin of error term: limit the analytical power of the model
3. nevertheless a good starting point for statistical analysis

TABLE 2.1 Terminology for Simple Regression

y	x
Dependent variable	Independent variable
Explained variable	Explanatory variable
Response variable	Control variable
Predicted variable	Predictor variable
Regressand	Regressor

Ceteris Paribus

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

Interpretation:

- Holding μ constant, i.e., $\Delta\mu = 0$ (ceteris paribus), effect of Δx units increase of x on y:

$$\Delta y = \beta_1 \Delta x$$

*Linear effect

* β_1 slope parameter, β_0 intercept parameter

Example: 1. if the profit rate increases by 0.01, the increase of stock return, ceteris paribus?

2. if the education year increases by 1, the increase of one’s earning, ceteris paribus?

Zero conditional mean

Economic data: observable y and x

- Unobservable β_0, β_1, μ
- How are we going to uncover them, particularly β_1?

*Important for economic decisions, e.g., portfolio choice, government policy on education

Crucial assumption to proceed (zero conditional mean):

$$\boxed{E(\mu|x) = E(\mu) = 0}$$

i.e., the expected value of the error term, conditional on the observed x, equals to its unconditional mean

*the observed x values do not tell us more information on the error term, i.e., not a function of x!

Zero conditional mean

It is ok that

$$E(\mu|x) = E(\mu) = constant(e.g., 20)$$

*If this is the case, our following ordinary least square estimator would load 20 onto β_0

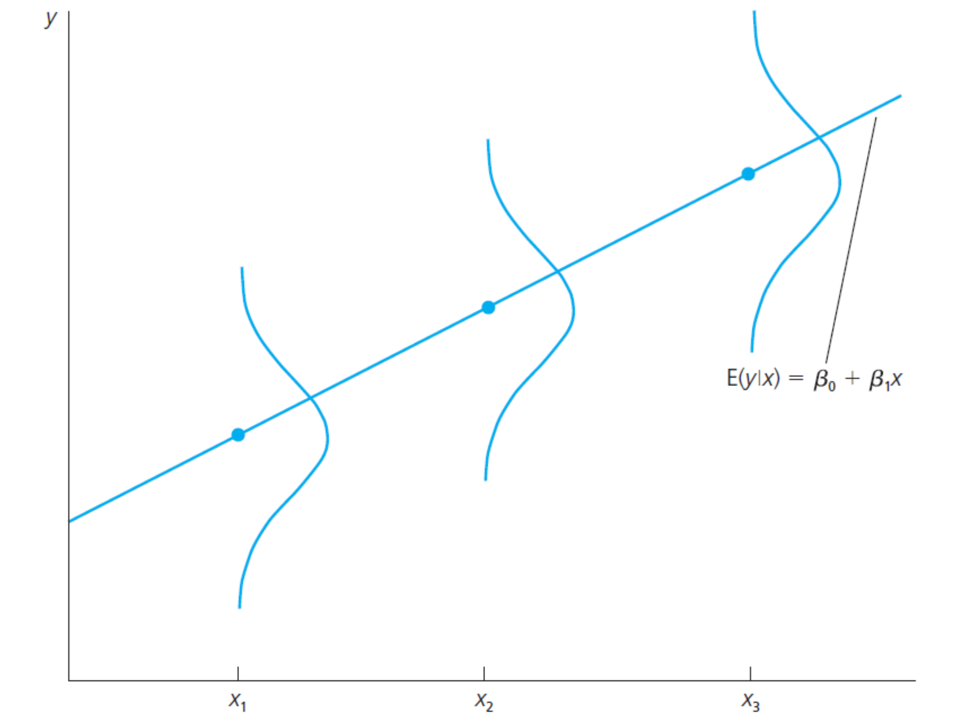
It is not ok that

$$E(\mu|x) = ax$$

Let's see this graphically

Zero condition mean

FIGURE 2.1 $E(y|x)$ as a linear function of x.



The vertical bell shaped lines are the probability density functions of μ.

*Note its relative position compared to the line of $\beta_0 + \beta_1 x$ is not changing when x changes

* $E(\mu|x)$ is a constant

In other words,

- μ and x are uncorrelated
- μ is mean independent with x

  *weaker than the statement that μ and x are independent

Let's see through an example of Mincer equation

Example: Mincer equation

Famous Mincer equation in labor economics

$$Hourly\ Wage = \beta_0 + \beta_2 Education + \beta_1 Experience + \beta_3 Experience^2 + \epsilon$$

Suppose you have a data only on hourly wage and experience, and mis-specify the equation as

$$Hourly\ Wage = \beta_0 + \beta_1 Experience + \mu$$

Suppose $E(\epsilon|x)$=0, and Education is uncorrelated with experience, we have

$$\mathrm{E}(\mu|Exeprience) = E(\beta_3 Experience^2|Experience) = Constant$$

* Is μ independent with Experience? Much stricter assumption

* Is μ mean independent with Experience? Easier to satisfy in most data

Population regression function

Given that

$$E(\mu|x) = E(\mu) = 0$$

We take expectations on both sides of

$$y = \beta_0 + \beta_1 x + \mu$$
$$\Rightarrow E(y|x) = \beta_0 + \beta_1 x$$

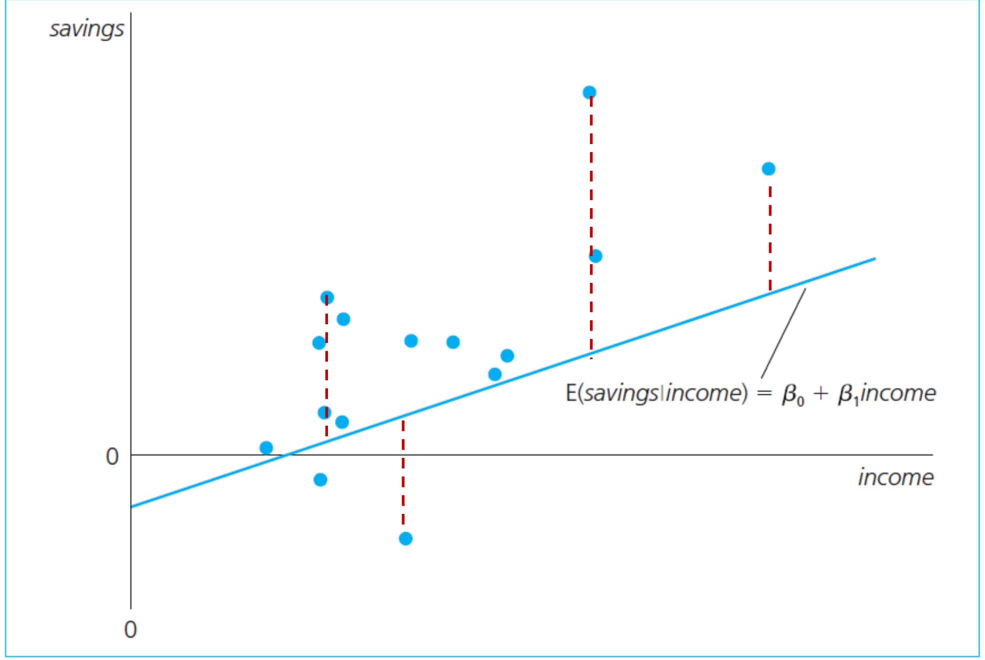
Population regression function (PRF)

Deriving the Ordinary-Least-Squares estimates

- Now we're ready use *one* method to uncover β_0 *and* β_1…
- … given a sample of the population, $\{(x_i, y_i): i = 1, \dots, n\}$
- The method is called Ordinary-Least-Squares, and the idea is pretty simple (and powerful)

Find $\hat{\beta}_0$, $\hat{\beta}_1$ to minimize the sum of distance between the predicted level and the actual level of y

FIGURE 2.2 Scatterplot of savings and income for 15 families, and the population regression $E(savings|income) = \beta_0 + \beta_1 income$.



Deriving the Ordinary-Least-Squares estimates

Actual value: y_i

Predicted (Fitted) value: $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}$ (Ask yourself: why there is a hat on both betas?)

Distance (Residual): $\widehat{\mu_i} = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x})$

Minimizing the sum of distance:

$$\min \mathrm{D} = \sum_{i=1}^{n} \hat{\mu}^2 = \sum_{i=1}^{n} [\, y_i - (\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x})]^2$$

*Why squared? Punishing the over- and under-predicted levels symmetrically

Deriving the Ordinary-Least-Squares estimates

General approach to solve the minimization problem

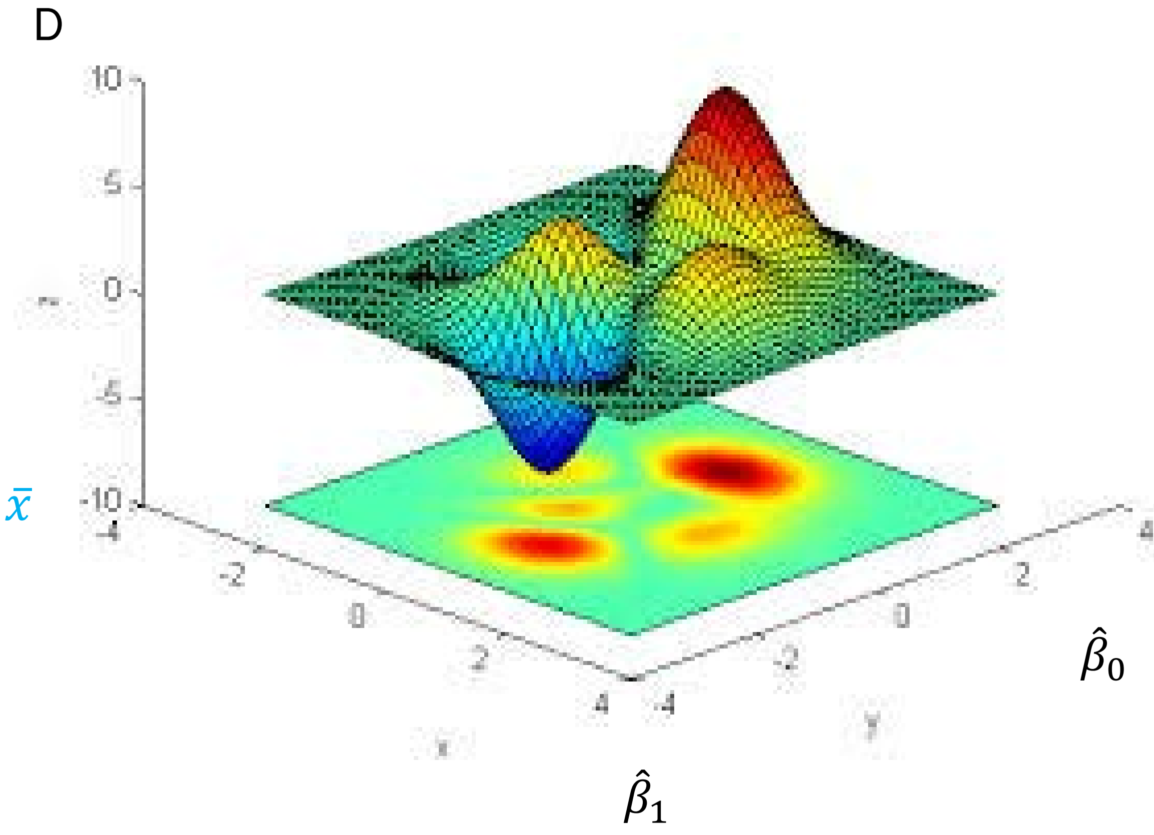
* take derivative with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$

* at minimum, we have $\frac{\partial D}{\partial \hat{\beta}_0}=0$, $\frac{\partial D}{\partial \hat{\beta}_1}=0$

Hence, we have

(i) $2\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0$ $\Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$

(ii) $2\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]x_i = 0$



Deriving the Ordinary-Least-Squares estimates

Plug the blue equation into (ii), we'll arrive

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

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

Deriving the Ordinary-Least-Squares estimates (alternative approach)

Method of moments: Note for the true β_0, β_1, the population has

(1) $E(y - \beta_0 - \beta_1 x) = 0$

(2) $E(xu) = E(x)E(u|x) = 0 \Rightarrow E\big(x(y - \beta_0 - \beta_1 x)\big) = 0$

Two unknowns, two moments.

Our sample analog?

(1) $n^{-1} \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] = 0$

(2) $n^{-1} \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)] x_i = 0$

$\Rightarrow$

Exactly the same equations on Page 13!

Yet method of moments: a more general approach

A note on terminology

For the sake of brevity, running a regression

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

is equivalent to say

1. run the regression of y on x
2. or, even more simply speaking, regress y on x

Example: household credit card limit

Data source: the U.S. Survey of Consumer Finance (2003), sponsored by Federal Reserve Bank

Task: quantify how does the change of household income affect household credit card limit

Follow these steps:

1. Explained variable y? Explanatory variable x?
2. Go to the sample that is assigned to your group (I randomly sampled 6 samples from the data).
3. Calculate $\bar{\mathrm{x}}, \bar{y}$.
4. Calculate $\hat{\beta}_1$ first and $\hat{\beta}_0$ second.

Do you have the same $\hat{\beta}_1$ and $\hat{\beta}_0$?

Expected values and variances of the OLS estimators

Estimators vs estimates:

- estimates, values of $\hat{\beta}_0$ and $\hat{\beta}_1$
- estimators, random variables of $\hat{\beta}_0$ and $\hat{\beta}_1$

You would appreciate the idea of randomness in $\hat{\beta}_0$ and $\hat{\beta}_1$ by

- By finding your estimates differ between any two groups in the above example
- Why? Your sample differ from each other.

Since the population is too expensive to be exhausted

- i.e., surveying 1.4 billion in China sounds incredibly expensive
- we could scientifically sample, e.g., 1% of the population (The idea behind Census)

    *Who gets sampled? Random

Expected values and variances of the OLS estimators

How could we be assured that our sampling gives "good" estimates of β_0 and β_1?

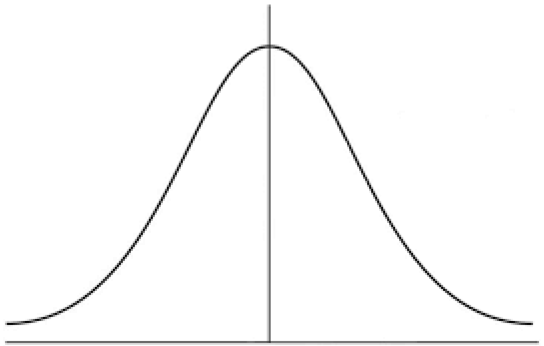
Two measures describing whether our estimators are "good"

1. Unbiased

   After sampling many times, I'm getting the right β_0 and β_1 on average.

2. Most efficient

   After sampling many times, my estimates of β_0 and β_1 are bouncing around with the possibly smallest magnitudes.



$E(\hat{\beta}_0) = \beta_0, and\ Var(\hat{\beta}_0)$ smallest

Expected values and variances of the OLS estimators

These ideas are associated with expected values and variances of the OLS estimators.

Suppose you sample many many many times, i.e., T=1e23 times

1. averagely right (mean): $T^{-1}\sum_{t=1}^{T}\hat{\beta}_o^t = \beta_0$? ; $T^{-1}\sum_{t=1}^{T}\hat{\beta}_1^t = \beta_1$?
2. magnitude of bouncing (variance): $T^{-1}\sum_{t=1}^{T}(\hat{\beta}_o^t - \beta_0)^2$; $T^{-1}\sum_{t=1}^{T}(\hat{\beta}_1^t - \beta_1)^2$

Their population analog

1. E($\hat{\beta}_o$) $= \beta_0$; E($\hat{\beta}_1$) $= \beta_1$
2. Var($\hat{\beta}_o$) ; Var($\hat{\beta}_1$)

Let's check the expected values and variances of the OLS estimators

Expected values of the OLS estimators

E($\hat{\beta}_1$) = β_1? Yes!

$$
\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})y_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(\beta_0 + \beta_1 x_i + \mu_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mu_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}
\end{aligned}
$$

$$
\Rightarrow E(\hat{\beta}_1) = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})E(\mu_i)}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \beta_1
$$

Expected values of the OLS estimators

E($\hat{\beta}_0$) = β_0? Yes!

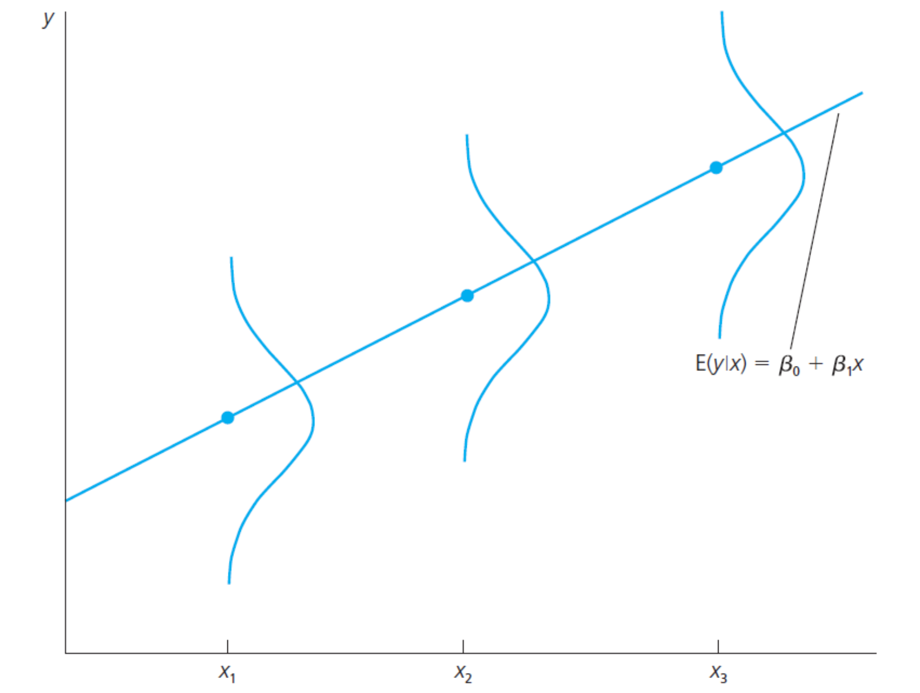
$$E(\hat{\beta}_0) = E(\bar{y} - \hat{\beta}_1\bar{x})$$
$$= E(\beta_0 + \beta_1\bar{x} + \bar{\mu} - \hat{\beta}_1\bar{x})$$
$$= \beta_0$$

Theorem 2.1 the OLS estimators of $\hat{\beta}_0$ and $\hat{\beta}_1$ are unbiased estimators for β_0 *and* β_1 under the assumptions:

1, y and x are linearly related
2, the sample is a random sample from the population
3, there are variations of y and x within the sample
4, zero mean condition

Variances of the OLS estimators

FIGURE 2.1 E(y|x) as a linear function of x.



Additional assumption 5： Homoskedasticity, i.e.,

$$var(\mu|x) = \sigma^2$$

- variance of the error term does not vary
- e.g., the dispersion of other factors that influence wage is the same for college vs high school educated

*Note zero conditional mean concerns about the mean, not the variance/dispersion

Graph a picture when the homoscedasticity assumption is violated.

Variances of the OLS estimators

Recall

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_{i=1}^{n}(x_i - \bar{x})\mu_i}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

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

$$= \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

and

$$\hat{\beta}_0 = \beta_0 + \bar{\mu} + (\beta_1 - \hat{\beta}_1)\bar{x}$$

$$\Rightarrow var(\hat{\beta}_0) = n^{-1}\sigma^2 + var(\hat{\beta}_1)\bar{x}^2$$

$$= n^{-1}\sigma^2 + \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\bar{x}^2$$

$$= \frac{\sigma^2 n^{-1} \sum_{i=1}^{n} x_i^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

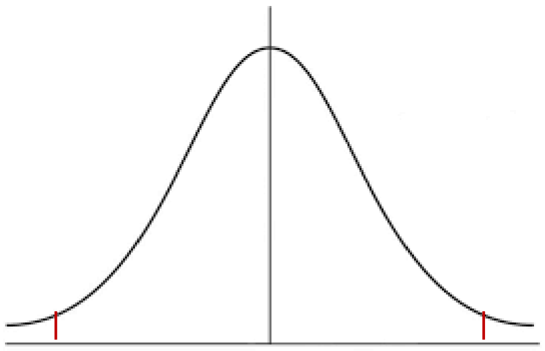
Variances of the OLS estimators

Theorem 2.2 the OLS estimators of $\hat{\beta}_0$ and $\hat{\beta}_1$ have the variances shown above under the assumptions:

1, y and x are linearly related
2, the sample is a random sample from the population
3, there are variations of y and x within the sample
4, zero mean condition
5. the error term is homoskedasticity

Inference of the OLS estimators

Recall from our overview class:



95% confidence interval:

[-1.96*sd($\hat{\beta}_1$)+mean($\hat{\beta}_1$), 1.96*sd($\hat{\beta}_1$)+mean($\hat{\beta}_1$)]

*For the model to have explanatory power, we hope that 0 does not fall into this range, i.e., the increase of x does "cause" an increase of y

*caution on causality, more rigorously speaking:

the increase of x is statistically significantly associated with an increase in y

Estimating the error variance

Now suppose you want to calculate the variances of the OLS estimators in your SCF sample

*to form the confidence interval

What is unknown though:σ^2

- variance of the error term μ, which is unobserved
- estimate the unobserved error term by the observed residual

Note

$$y_i = \beta_0 + \beta_1 x_i + \underbrace{\mu_i}_{\text{error term}} = \hat{\beta}_0 + \hat{\beta}_1 x_i + \underbrace{\hat{\mu}_i}_{\text{residual}}$$

error term vs residual

Estimator of σ^2, variance of the residual: $\hat{\sigma}^2 = n^{-1} \sum_{i=1}^{n} \hat{\mu}_i^2$

Estimating the error variance

This is however, a biased estimator of σ^2

To see this, note

$$\hat{\mu}_i^2 = [\mu_i - (\hat{\beta}_0 - \beta_0) - (\hat{\beta}_1 - \beta_1)x_i]^2$$

$$= \mu_i^2 + (\hat{\beta}_0 - \beta_0)^2 + (\hat{\beta}_1 - \beta_1)^2 x_i^2$$

$$+ 2(\hat{\beta}_0 - \beta_0)(\hat{\beta}_1 - \beta_1)x_i$$

$$\Rightarrow E(\hat{\sigma}_i^2) = n^{-1}\sum_{i=1}^{n} E(\hat{\mu}_i^2) = \frac{n-2}{n}\sigma^2 \neq \sigma^2$$

To adjust for that, we usually use another estimator $\hat{\hat{\sigma}}^2 = (n-2)^{-1}\sum_{i=1}^{n}\hat{\mu}_i{}^2$

*averagely right ⇑

Now you should be able to implement statistical inference in a simple regression model!

Estimating the error variance

Why $\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}\hat{\mu}_i{}^2$ is biased?

Recall how we solve the OLS estimators:

(1) $n^{-1}\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 \mathrm{x}_i)] = n^{-1}\sum_{i=1}^{n}[\hat{\mu}_i] = 0$

(2) $n^{-1}\sum_{i=1}^{n}[y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]x_i = n^{-1}\sum_{i=1}^{n}[\hat{\mu}_i x_i] = 0$

Here $\{\hat{\mu}_i\}, i = 1,2,\ldots,n$ must satisfy the above two equations

- when you know n-2 $\hat{\mu}_i$, you know the rest 2
- degree of freedom is n-2

Goodness-of-fit

Not only we care whether there is a statistically significant impact of x on y

But also the idea: how much of the variation in y could be explained by the variation of x

- Exactly the idea behind goodness-of-fit

Define R-squared as $R^2 \equiv \mathrm{SSE}/\mathrm{SST} = 1 - \mathrm{SSR}/\mathrm{SST}.$

where total sum of squares (SST) $\mathrm{SST} \equiv \sum_{i=1}^{n} (y_i - \bar{y})^2.$

explained sum of squares (SSE) $\mathrm{SSE} \equiv \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2.$

residual sum of squares (SSR) $\mathrm{SSR} \equiv \sum_{i=1}^{n} \hat{u}_i^2.$

Goodness-of-fit

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$
$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

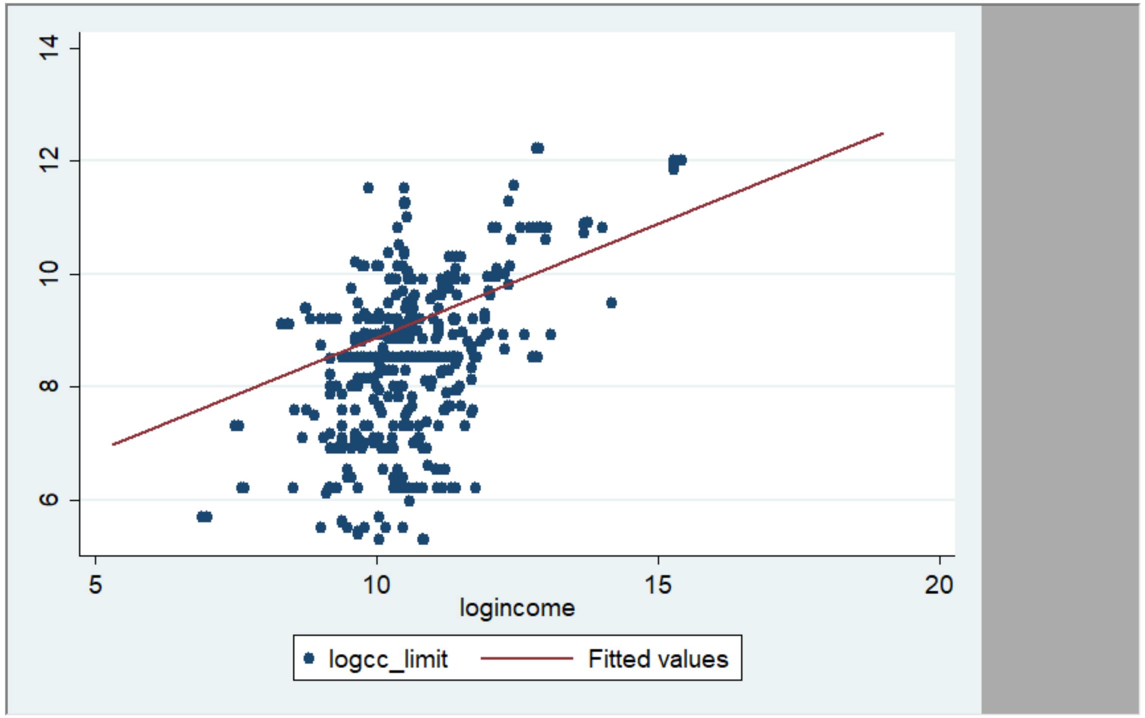
By merging the first and third items,

$$\Rightarrow = \sum_{i=1}^{n}(y_i^2 - \hat{y}_i^2 + 2\hat{y}_i\bar{y} - 2\hat{y}_i\bar{y}) + SSE$$
$$= \sum_{i=1}^{n}\hat{\mu}_i^2 + 2\sum_{i=1}^{n}\hat{y}_i - 2n\bar{y} + SSE$$
$$= SST + SSE$$

Goodness-of-fit (graphically)

y=Grumpiness, x=Sleep hours. High R-squared

y=Credit card limit, x=Income. Low R-squared



Goodness-of-fit

Calculate the goodness-of-fit in your SCF sample.

Other issues: 1. units and functional forms

1. Units

 (i) rescaling

 Example: salary in thousand dollars or in dollars, corresponding rescaling in β_0 *and* β_1

 (ii) logs

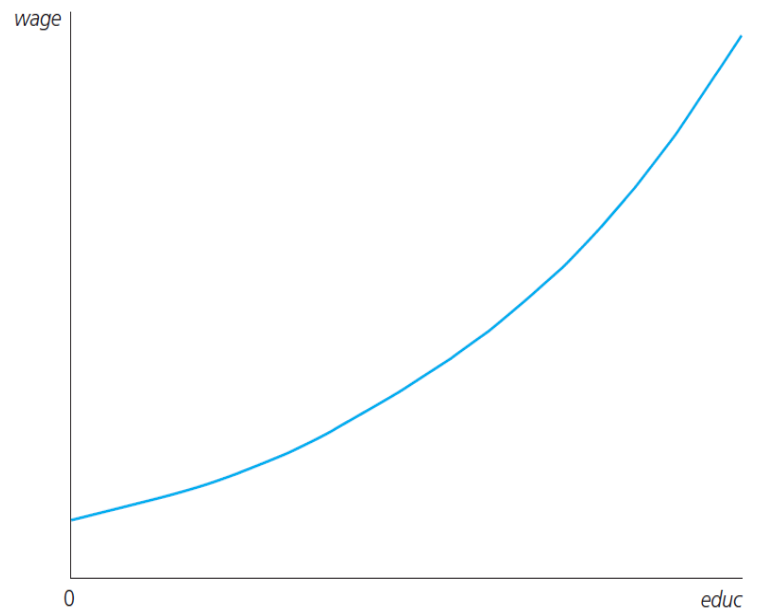
 (a) Non-linear model

$$\log(wage) = \beta_0 + \beta_1 educ + u,$$

i.e., the impact of education on wage is not a constant

Not a sense of the right model, but a sense of a model that better captures the data

FIGURE 2.6 $wage = \exp(\beta_0 + \beta_1 educ)$, with $\beta_1 > 0$.



Other issues: 1. units and functional forms

(b) Constant elasticity model

Economists care about the elasticity in many applications, i.e., percentage changes in y caused by 1 percent increase in x

Example of production function

$$log(output) = \beta_0 + \beta_1 log(capital) + \mu$$

Elasticity is β_1

Other issues: 2. regression without a constant

In some cases, we wish to impose the assumption that y=0 if x=0

$$\tilde{y} = \tilde{\beta}_1 x,$$

i.e., regression through the origin

The OLS estimator is then

$$\tilde{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2},$$

Summary

1. Definition of the simple regression model

2. Deriving the Ordinary-Least-Square (OLS) estimates

3. Properties of OLS estimates

4. Expected values and variances of OLS estimators

5. Issues on units of measurement and functional form