Lecture 2:
Simple Regression I

Jan Zouhar

Introductory Econometrics

Introducing Simple Regression

Introducing Simple Regression

- simple regression = regression with 2 variables

y	x
dependent variable	independent variable
explained variable	explanatory variable
response variable	control variable
predicted variable	predictor variable
regressand	regressor

- we are actually going to derive the linear regression model in three very different ways
- these three ways reflect three types of econometric questions we discussed in the first lecture (*descriptive*, *causal* and *forecasting*)
- while the math for doing it is identical, conceptually they are very different ideas

Descriptive Approach

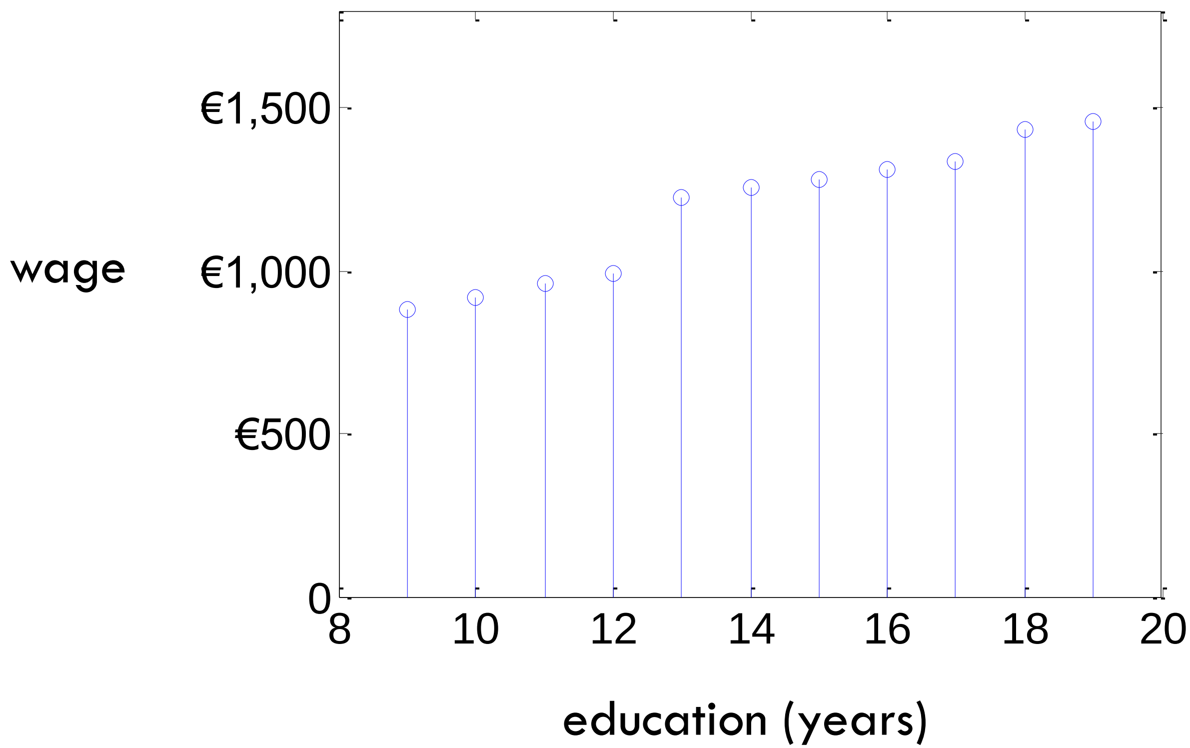
Why do we need a regression model?

Estimation & interpretation.

Correlation vs. causation.

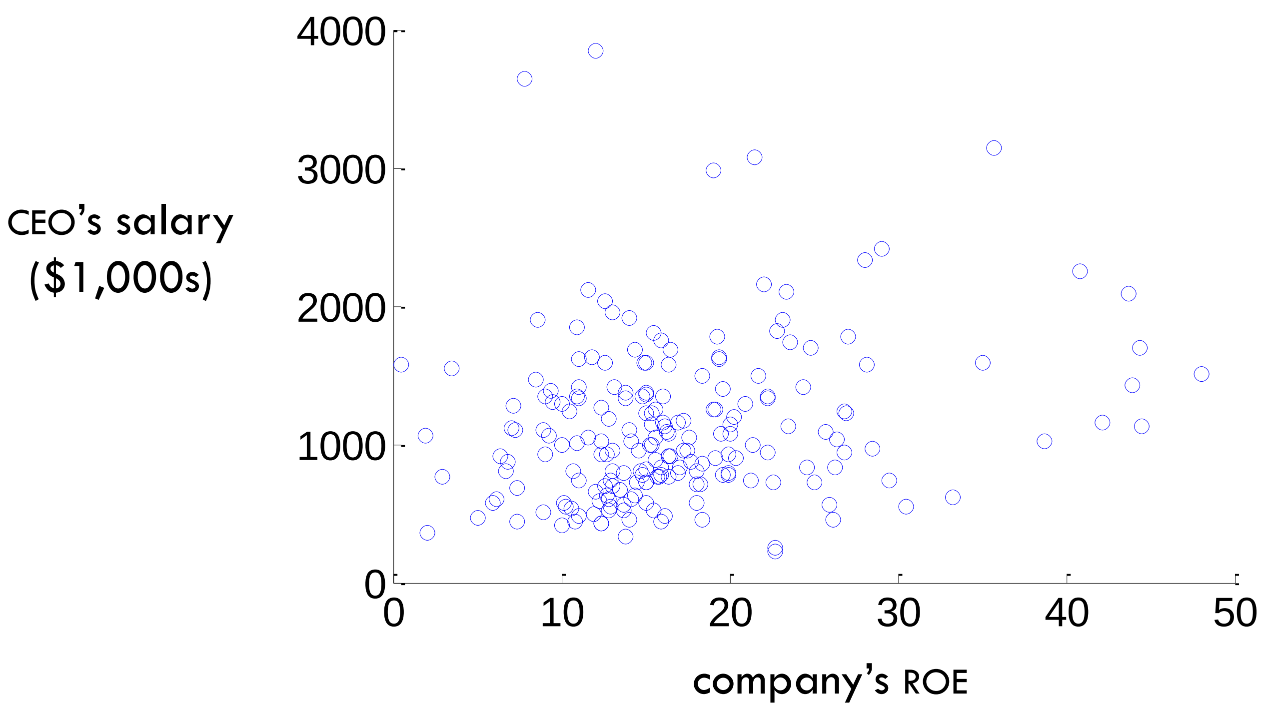
Descriptive Analysis

- **goal**: estimate $\mathsf{E}[y \mid x]$ (called the **population regression function**)
  - x is discrete (i.e., categories) – *wages* vs. *education* example
    - one can collect data for the individual categories (the more categories, the more difficult)



Descriptive Analysis (cont'd)

- x is continuous – problem: no categories
  - **example**: CEO's salary vs. company's ROE
    - ROE = *return on equity* = net income as a percentage of common equity (ROE = 10 means if I invest $100 in equity in a firm I earn $10 a year)
    - does a CEO's salary (typically huge) reflect her performance?



Descriptive Analysis (cont'd)

- therefore, we need a model for $\mathsf{E}[y \mid x]$
  - in other words, we need to find a "good" mathematical expression for f in $\mathsf{E}[y \mid x] = f(x)$)
- the simplest model I can think of is $\mathsf{E}[y \mid x] = \beta_0 + \beta_1 x$

Why use simple models:

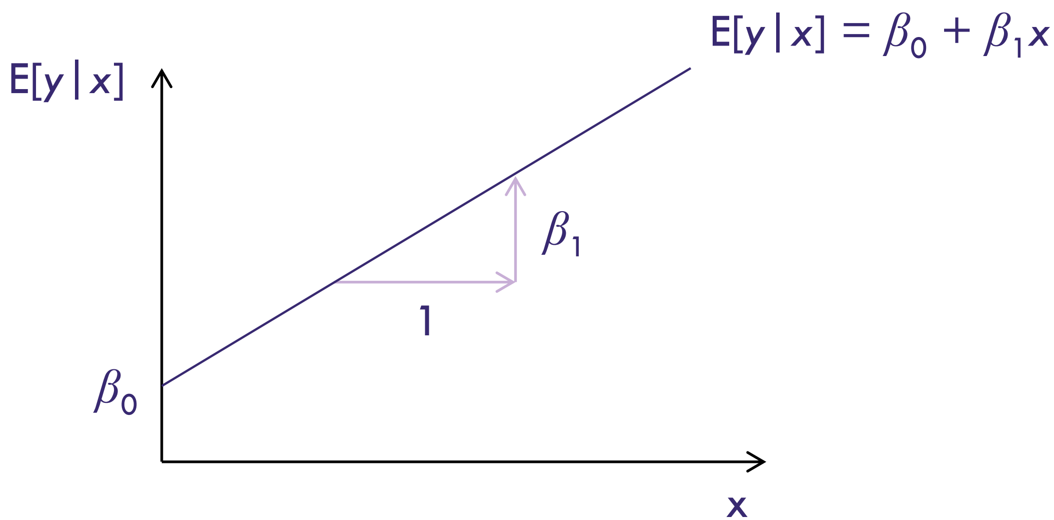
Simple models are:

- easier to estimate.
- easier to interpret (e.g., $\beta_1 = \Delta wage / \Delta educ$ etc.).
- easier to analyze from the statistical standpoint.
- safe: they serve as a good approximation to the real relationship, the functional nature of which might be unknown and/or complicated. Things can't go too wrong when using a simple model.

Further reading: Angrist and Pischke (2008): *Mostly Harmless Econometrics: An Empiricist's Companion.*

Linear Model for E[y | x]

- interepretation:
  - *intercept*: $\beta_0 = \mathsf{E}[y \mid x = 0]$
  - *slope*: $\beta_1 = \dfrac{\partial \mathsf{E}[y \mid x]}{\partial x} = \dfrac{\Delta \mathsf{E}[y \mid x]}{\Delta x}$



Estimation

- once we've decided about the functional form of the model (in our case, it's $\mathsf{E}[y \mid x] = \beta_0 + \beta_1 x$), we need to develop techniques to obtain estimates of the parameters β_0 and β_1
- we base our estimates on a sample of the population → we need to make inferences about the whole population
  - sample:
    - n observations (people, countries, etc.) indexed with i
    - x and y values for ith person denoted as y_i, x_i (for $i = 1,\ldots,n$)

Preliminaries

- it will be useful to *define* $u = y - \beta_0 - \beta_1 x$
- what do we know about u?
  - first, because $\mathsf{E}[y \mid x] = \beta_0 + \beta_1 x$, we have

$$\begin{aligned}\mathsf{E}[u \mid x] &= \mathsf{E}[y - \beta_0 - \beta_1 x \mid x] = \\ &= \mathsf{E}[y \mid x] - \beta_0 - \beta_1 x = \\ &= \beta_0 + \beta_1 x - \beta_0 - \beta_1 x = 0\end{aligned}$$

- this means two things:
  1. $\mathsf{E}\,u = 0$. (This should be intuitive: $\mathsf{E}[u\,|\,x] = 0$ for all $x \;\rightarrow\; \mathsf{E}\,u = 0$; alternatively, you can plug in CE.4 from Wooldridge, page 687.)
  2. the expected value of u does not change when we change x
- the second property has numerous implications, the most important being:
  - $\mathsf{cov}(x,u) = 0$ (this is property CE.5 from Wooldridge, page 687)
  - $\mathsf{E}[xu] = 0$ (because $\mathsf{cov}(x,u) = \mathsf{E}[xu] - \mathsf{E}x\,\mathsf{E}u$)

- how is this useful in estimation?
  - we arrived at two important facts about expectations of u:

$$\mathsf{E}\,u = 0$$
$$\mathsf{E}[xu] = 0$$

  - typically, expectations are estimated using a *sample mean* (e.g., how would you estimate *mean wage* with a sample of 10 people?)
  - we'll use the idea of *sample analogue*, forcing the sample means of u and xu to equal zero (see below)

Estimation (cont'd)

- before we move on, we'll revise some more statistical concepts connected with random sample
- remember we need to make inference about the population based on our sample and its characteristics

Population vs. sample characteristics:

Random variable	Population (size *N*)	Sample (size *n*)
$\mu_x = \mathrm{E}\,x$	$\mu_x = \frac{1}{N}\sum_{i=1}^{N} x_i$	$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$
$\mathrm{var}\,x = \mathrm{E}(x-\mu_x)^2$	$\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_x)^2$	$\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$
$\mathrm{cov}(x,y) = \mathrm{E}(x-\mu_x)(y-\mu_y)$	$\frac{1}{N}\sum_{i=1}^{N}(y_i-\mu_y)(x_i-\mu_x)$	$\frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})$

- the expressions on the right are called *sample mean*, *sample variance* and *sample covariance*

Estimation (cont'd)

- for ith person (i.e., observation) in our sample, we have the **population regression model**

$$y_i = \beta_0 + \beta_1 x_i + u_i \,,$$

where β_0 and β_1 are the unknown parameters to be estimated

- to think about estimation let's define the **sample regression model**

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i,$$

where
  - $\hat{\beta}_0$ and $\hat{\beta}_1$ are our estimates of β_0 and β_1 from the sample
  - $\hat{u}_i$ is defined as $\hat{u}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$
- *note*: if $\hat{\beta}_0$ and $\hat{\beta}_1$ are like β_0 and β_1, then $\hat{u}_i$ should be like u_i
- **sample analogue**:
  - in order to make inferences about the population, we have to believe that our sample looks like the population
  - if it is so, then let's force things to be true in the sample which we know would be true in the population

Estimation (cont'd)

- from the discussion about (the population's) u, we know that

$$\mathsf{E}\,u = 0$$

$$\mathsf{E}[xu] = 0$$

- the sample analogue to this is:

$$0 = \tfrac{1}{n}\sum_{i=1}^{n}\hat{u}_i \quad = \tfrac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

$$0 = \tfrac{1}{n}\sum_{i=1}^{n}x_i\hat{u}_i = \tfrac{1}{n}\sum_{i=1}^{n}x_i(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

- as y_i and x_i are the data, the above equations are in fact two linear equations in variables $\hat{\beta}_0$ and $\hat{\beta}_1$; they can be solved (fairly) easily (see Wooldridge, pages 28 and 29)
- note that the first equation tells us that $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$, where $\bar{x}$ and $\bar{y}$ are the sample means of x_i's and y_i's
- solving the equations to get $\hat{\beta}_1$ yields: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Estimation (cont'd)

- we can rewrite the formula for the slope as

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

  which can be viewed as the sample analogue to

$$\frac{\mathrm{cov}(x, y)}{\mathrm{var}\, x}$$

- both var x and its sample counterpart are always positive, therefore:
  - the regression coefficient (β_1), the covariance, and the correlation coefficient must all have the same sign
  - one of them is zero only if all of them are zero

Correlation Vs. Causation

- the difference between causation and correlation is a key concept in econometrics
- you saw that in the model with conditional expectations, the estimates were based on the *correlation* between x and y (remember the formulas)
- there are many ways a *causal* interpretation can be given that is consistent with the (correlation-based) results (see next slide)
  - as we'll see, no econometric tool can ever "prove" or "find" a causal relationship on itself
  - having an economic model is essential in establishing the causal interpretation (we'll talk about this in the causal-analytic part)
- **conclusions**:
  - correlation $\neq$ causation
  - statistical significance $\neq$ the *effect* of x on y is significant (only that they are "tightly associated")
- these issues are confused all the time by politicians and the popular press

Three Causation Schemes

- looking at the relationship between x and y from the *causal* standpoint, the association (or correlation) between x and y can represent three basic situations:

$$y \longleftarrow x$$

x causes y: if a CEO's performance is good (high ROE), he gets paid a lot.

$$y \longrightarrow x$$

y causes x: high salaries motivate CEOs, thus making them perform well (resulting in high ROE)

$$\begin{tikzcd} y \arrow[rr, leftrightarrow, dashed] & & x \\ & z \arrow[ul] \arrow[ur] & \end{tikzcd}$$

there is another factor z that causes both x and y, which makes x and y be associated: if a CEO is good (clever, motivated, etc.), he both performs well and gets paid a lot

- the descriptive approach makes no claims about which one is the case

Causal Approach

The need for causal analysis.

Structural model and its assumptions.

Estimation.

Examples.

Causal Analysis

- in the descriptive framework, we couldn't really say anything about causality
- however, in decision-making situations, causal questions are typically those we need to answer:
  - if I face a decision, it means I can influence something (I have control over an economic variable)
  - in order to decide effectively, I need to know the impact of my decision on things I'm interested in
  - examples:
    - central bank sets interest rates in order to keep inflation within specified bounds (inflation targeting)
    - companies charge prices so as to maximize profit / revenues / market share
    - ministry of education introduces new school fees scheme; the aim is to collect money without discouraging students from education. Therefore, one has to find the effect of education on future income

Causal Analysis (cont'd)

- to say something about causality we need to make some more assumptions
- in practice, these assumptions will have to be checked using an economic theory / common sense + knowledge of the real problem
- mathematically, the formulation of these assumptions consists in writing down a *structural data-generating model*
- this will look similar to what we have been doing, but conceptually it is very different
  - for the description-type analysis, we started with the data and then asked which model could help summarize the conditional expectation
  - for the structural (causal) case we start with the model and then use it to say what the data will look like (even before we actually collect them)

Structural model

- a simple structural model may look like this

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

- what has changed from the descriptive analysis?
  - *descriptive model*: conditional expectation of y = a function of x

$$\mathsf{E}[y \mid x] = f(x)$$

  - *structural model*: y = a function of x and u

$$y = f(x,u)$$

- it should be clear that modeling y is much more ambitious than modeling the expectation of y
- we have already encountered u, but this time it has a real content (see later)
- in choosing a particular structural model, we're actually saying that we believe that, in reality, the value of y is "created" from x and u using function f
  - this is a daring claim, so we need to choose the model very carefully

How About u?

- u is probably the most important part of the structural model
- we'll spend a lot of time talking about u and its relation to x (note that the relation of u and y is obvious from the equation)
- what does u contain? Everything that the rest of the right-hand side of the equation failed to capture
  - in the previous example, this means everything that affects your wages besides education:
    - intelligence
    - work effort
    - …other suggestions?
  - in some textbooks, you can read that u contains a couple more things:
    - measurement errors (or poor *proxy variables*)
    - the intrinsic randomness (in human behaviour etc.)
    - model specification errors

  → these will not be all that relevant for our causal discussion

Crucial Assumption: $E[u|x] = 0$

- once we have specified the structural model, we need to estimate its parameters, i.e. β_0 and β_1
- in order to be able to do so, we need to assume something about the relation between u and x
- mathematically, the crucial assumption takes on the form $E[u|x] = 0$
- this is the same as what we did before (in the descriptive analysis), but conceptually very different
  - before u was defined simply $u = y - \beta_0 - \beta_1 x$. It didn't actually mean anything
  - now we think of u as this real thing that is actually out there and means something – it is just that we don't / can't observe it
    - even though we don't observe it, we can still argue whether the assumption is fulfilled or not
    - in order to do so, we need to know the assumption tells us in the first place

Crucial Assumption: $E[u | x] = 0$ (cont'd)

- as before, the assumption that $\mathsf{E}[u\,|\,x] = 0$ really means two separate things, one of which is a big deal, the other is not:

  1. $\mathsf{E}[u\,|\,x]$ doesn't vary with x (i.e., it's a constant).
     - this holds true if…
       - …u is assigned at random
       - …u and x are independent
       - …perhaps something else
     - and is not true if u and x are correlated
       - **example**: intelligence (a part of u) is correlated with education (x) $\rightarrow$ we're in trouble
         (we'll discuss the possible solutions as we go)

  2. $\mathsf{E}u = 0$ (i.e., the "unconditioned" expectation is zero).

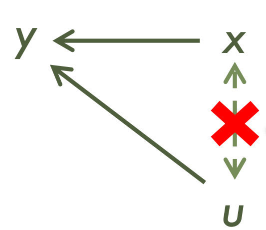
- the important part is 1; we assume 2 just for convenience

$E[u \mid x] = 0$ Vs. Causation

- what does the assumption tell us about causation?
  - remember the three causation schemes between x and y:

$$y \leftarrow x \quad , \quad y \rightarrow x \quad , \quad y \leftarrow z \rightarrow x$$

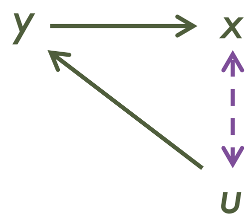
  - in the causal analysis, we want to rule out all but the first one
  - this is exactly what the assumption about $E[u \mid x]$ effectively does
- the "arrow scheme" now contains three letters: y, x and u
  - we know that u affects y (by definition), $\quad y \leftarrow u$
  - note that $E[u \mid x] = constant$ implies $\text{cov}(x,u) = 0$
  - therefore, we'd like the arrow scheme to look like this:



this suggests there should be no association (correlation) between x and u

$E[u\,|\,x] = 0$ Vs. Causation (cont'd)

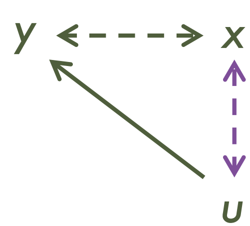
- imagine we have the reversed causality $y \rightarrow x$:



here u affects y which in turn affects x, so u affects x, and u and x are correlated

  - therefore $\mathrm{cov}(x,u) \neq 0$ and the assumption is necessarily violated

- now consider the case $y \leftarrow z \rightarrow x$:
  - if there's any z that affects y, it is a part of u (by definition), and therefore z (and u) affect x



u contains z, so u affects x, meaning the two are correlated

Estimation

- now, let's take the assumption $\mathsf{E}[u \mid x] = 0$ for granted
- then, nothing is really any different than in the descriptive case
- we can write down the sample regression function as

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i,$$

- we know that

$$\mathsf{E} u_i = 0$$
$$\mathsf{E}[x_i u_i] = 0$$

- the sample analogue is $0 = \frac{1}{n}\sum_{i=1}^{n} \hat{u}_i \quad = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

$$0 = \frac{1}{n}\sum_{i=1}^{n} x_i \hat{u}_i = \frac{1}{n}\sum_{i=1}^{n} x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$$

which gives us $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ exactly as before

- it is only the interpretation that has changed

Forecasting Approach

Forecasting framework.

Ordinary least squares estimation (OLS).

Forecasting Analysis

- let's shift gears completely
- imagine the following
  - you have a bunch of data on x and y now (i.e., the x_i's and y_i's)
  - you know the value of x tomorrow (we'll denote the value x^*) and want to predict what the value of y will be tomorrow (denoted y^*)
  - actually, when talking about dynamic models, there are often lags in economic responses, so that today's cause (x) is the yesterday's value of an economic variable
  - **examples**:
    - inflation rate this year, unemployment rate next year
    - corporate profits today, stock price tomorrow
    - SAT scores, college GPA
  - this can make x^* available in advance for prediction

Forecasting Analysis (cont'd)

- in order to do prediction we need a model
- let's once again use the linear sample regression model

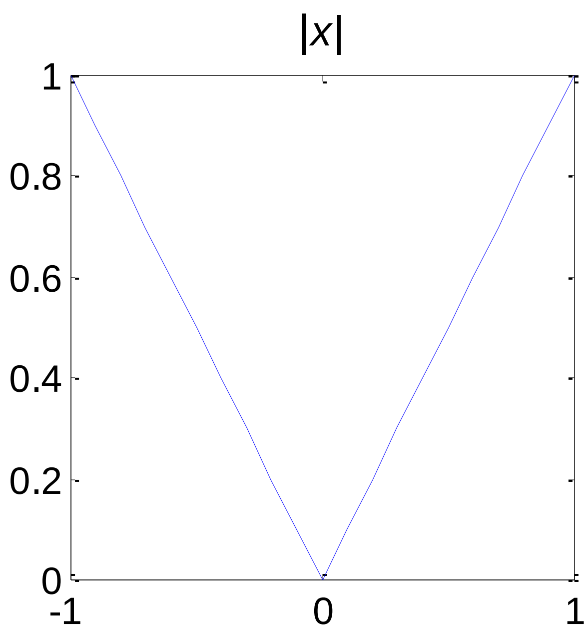
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \hat{u}_i$$

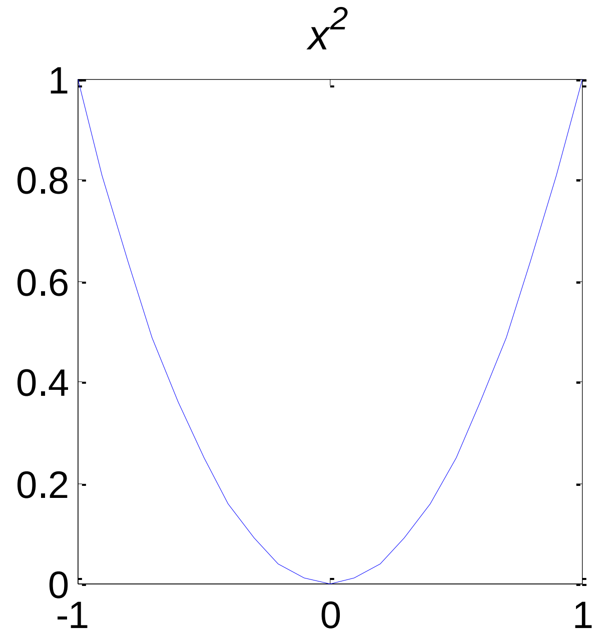
- once we estimate the parameters, we can predict the future value of y as

$$\hat{\beta}_0 + \hat{\beta}_1 x^*$$

- note that unlike in the causal analysis, *we do not choose x^** here!
- from the forecasting point of view, the fitted values $\hat{y}_i$ can be regarded as the would-be predictions we'd use if we didn't have the appropriate y_i's
- therefore, it seems sensible to find $\hat{\beta}_0$ and $\hat{\beta}_1$ such that the overall distance between $\hat{y}_i$'s and y_i's is minimized
- how to measure this distance?
  - obvious idea: absolute difference between the two: $|y_i - \hat{y}_i|$
  - however, absolute value is a really ugly function

Forecasting Analysis (cont'd)





- a much smoother function is $(y_i - \hat{y}_i)^2$
- we want to aggregate this distance across all data points:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- this says how close our model is to the data

Ordinary Least Squares (OLS) Estimation

- note that the overall distance distance between $\hat{y}_i$'s and y_i's is a function of variables $\hat{\beta}_0$ and $\hat{\beta}_1$ (because y_i's and x_i's are known – the data)
- we can call this function *sum of squares* (SS) and write

$$SS(\hat{\beta}_0, \hat{\beta}_1) = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

- we want to keep the model as close to the data as possible, which means we minimize SS, so we take the derivative of this function with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$ and set to zero:

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

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

- note that we're back to the same system of 2 linear equations as before (after dividing both equations by –2)
- the estimates are the same as before, but with a different kind of reasoning again

Three Approaches: A Comparison

- descriptive approach
  - *goal*: find the association between x and y expressed in terms of conditional expectation $\mathsf{E}[y\,|\,x]$
  - *assumptions*: approximate functional form of $\mathsf{E}[y\,|\,x]$
- causal approach
  - *goal*: find the causal effect of x on y
  - *assumptions*:
    - structural model for y ("how y is created")
    - assumptions about u: $\mathsf{E}[u\,|\,x] = 0$ (and thus, $\mathsf{E}[xu] = 0$)
- forecasting approach
  - *goal*: predict future values of y based on the knowledge of future x
  - *assumptions*: approximate functional form of the relation "y vs. x"

→ different goals, different assumptions, same formulas for estimates
→ the econometric software has only one procedure for all three cases, you have to know what you're doing, check the assumptions etc.

Lecture 2:
Simple Regression I

Jan Zouhar

Introductory Econometrics