EEP/IAS 118 - Introductory Applied Econometrics, Section 1

Pierre Biscaye and Jed Silver

September 2021

Intro

- Brief Intro
- Section
    - In-person; lecture notes posted online beforehand (usually), slides and videos posted afterward
    - Not required
    - Mainly review material from lectures; sometimes discuss assignments
    - Try to come to your scheduled section - fine to swap here and there

Talk to us

- Office hours: in-person, Giannini 203
  - Pierre: Tuesdays from 2-4PM
  - Jed: Wednesdays from 10AM-12PM
  - Email us if these times don't work for you
- Email policy: we'll respond with 48 hours during the week; don't expect responses after 6PM or on weekends
- Piazza: similar response time; encourage you to assist each other first
  - You will often be each other's best resources

Don't be afraid to reach out - we can't help you if we don't know!

- Econometrics material can be hard to grasp initially - but that's okay!
- Tell us if you are having technology problems (e.g., bad Internet connection, rolling power outages, issues with DataHub)

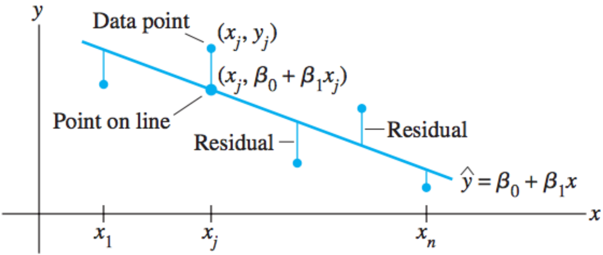
Class websites

- Lecture: Zoom for now, in-person next week; slides posted online beforehand (usually), videos posted afterward
- bCourses: course announcements, files, videos, assignments
- Piazza: forum for questions and discussions; GSIs monitor but student interactions encouraged!
- DataHub server: host for assignment Jupyter notebooks; do your work here and then save as PDF to submit
- If you haven't checked out videos on Jupyter notebook and R yet, please do so!
- Gradescope: website for submitting completed assignments

Topics for today

- Linear regression models
- Functional forms
- Random variable review

Linear Regression Models



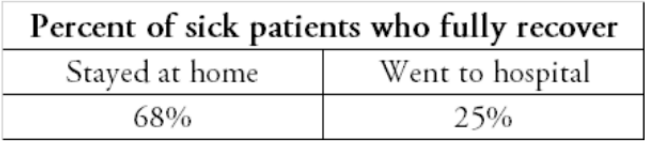
Key components of figure:

- Actual data: x_j, y_j - observations of the two variables
- Line equation for $\hat{y} = \beta_0 + \beta_1 x$ - this is the predicted value of the outcome variable y
- Residuals ($\hat{u}_i$) - the difference between the predicted $\hat{y}_i$ and the actual observed y_i

Relationships between variables

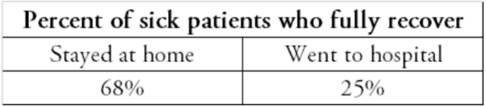
A model is simply trying to describe a relationship between variables. However, we need to be careful.

A news host shows a hospital in poor condition, and states “As you can see, health services here are so bad that going to a hospital is actually worse than staying at home. The following statistics demonstrate that you are better off staying away from hospitals” (see page 2 of notes)



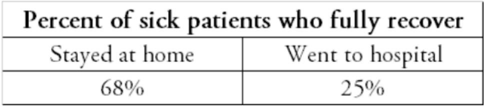
Percent of sick patients who fully recover	
Stayed at home	Went to hospital
68%	25%

A news host states “ The following statistics demonstrate that you are better off staying away from hospitals” (see page 1 of notes)



Percent of sick patients who fully recover	
Stayed at home	Went to hospital
68%	25%

- What is the implied research question from this story?
- Do you agree with the news anchor’s conclusion? Why or why not?
- What are the components of the regression model you would use to analyze this question (if you had the data)?



Percent of sick patients who fully recover	
Stayed at home	Went to hospital
68%	25%

- What is the implied research question from this story? *What is the effect of going to the hospital on full recovery from an illness?*
- Do you agree with the news anchor's conclusion? *No, because the sample of people who go to the hospital is different from the sample that does not.*
- What are the components of the regression model you would use to analyze this question (if you had the data)?
  - *Dependent variable* (Y) = *Fully Recover*
  - *Explanatory variable of interest* (X_1) = *Went to hospital*
  - *Other explanatory variables* ($X_1, X_2, ...$) = *Age, Medical History, Severity of illness*

Correlation vs. Causation

- In this example, we do see a negative correlation between recovery and visiting the hospital
- So, what does newspaper article get wrong? There *is* a correlation!
    - The article falsely assigns *causality* to the relationship - this is the classic correlation $\neq$ causation
- The statistic is misleading (if improperly understood) because it omits other important variables associated with recovery from the model (age, medical history, severity of illness, etc.)

The core of this class is to understand how and when we can assign *causality* to observed relationships.

Review on functional forms

Preliminary concepts:

- *Proportional change*: $\frac{x_1 - x_0}{x_0} = \frac{\Delta x}{x_0}$
- *Percentage change*: $\frac{x_1 - x_0}{x_0} \times 100 = \frac{\Delta x}{x_0} \times 100$
- *Elasticity*: $\frac{\Delta z / z}{\Delta x / x} = \frac{\partial z}{\partial x} \frac{x}{z}$

Note, percent change is just proportional change times 100.

Elasticity is the "percent change in one variable in response to a given (one) percent change in another variable"

Functional forms and Marginal Effects

This Table (Table 2.3 in Wooldridge) is meant to practice and continue familiarizing ourselves with these functional forms (found on page 5 of notes).

Model	DepVar	Ind. Var	Δy relates to Δx?	Interpretation
Linear	y	x	$\Delta y = \beta_1 \Delta x$	$\Delta y = \beta_1 \Delta x$
Logarithmic	y	$\log(x)$	$\Delta y = \beta_1 \frac{\Delta x}{x}$	$\Delta y = (\beta_1/100)\%\Delta x$
Exponential	$\log(y)$	x	$\frac{\Delta y}{y} = \beta_1 \Delta x$	$\%\Delta y = (100\beta_1)\Delta x$
Log-Log	$\log(y)$	$\log(x)$	$\frac{\Delta y}{y} = \beta_1 \frac{\Delta x}{x}$	$\%\Delta y = \beta_1 \%\Delta x$

Ex: $\%\Delta y = (100\beta_1)\Delta x$ - Read this as "$\hat{y}$ increases by $100 * \beta_1$% for a one unit increase in x." Derivations of these are in the notes! **Tip:** When

you see logs, think percent changes!

Examples: Interpreting Marginal Effects

Suppose you've collected data on household gasoline consumption (gallons) in the Bay Area and gas prices ($ per gallon), and you estimate the following model:

$$\log(gasoline) = 12 - 0.21 price$$

According to the model, how does gas consumption change when *price* increases by $1?

Examples: Interpreting Marginal Effects

Suppose you've collected data on household gasoline consumption (gallons) in the Bay Area and gas prices ($ per gallon), and you estimate the following model:

$$\log(gasoline) = 12 - 0.21 price$$

According to the model, how does gas consumption change when $price$ increases by $1?

If $price$ increases by $1, then predicted $gasoline$ consumption will decrease by 21%

Examples: Interpreting Marginal Effects

Professor Magruder uses firm data from Kenya to investigate how basket sales were affected by straw prices. In this example, he looks at the share of basket purchases that were made while baskets were on sale. The following model can be estimated:

$$\log(\mathit{basketshare}) = 0.83 + 0.491 \log(\mathit{strawprice})$$

How does *basketshare* change if straw prices rise by 2%?

Examples: Interpreting Marginal Effects

Professor Magruder uses firm data from Kenya to investigate how basket sales were affected by straw prices. In this example, he looks at the share of basket purchases that were made while baskets were on sale. The following model was estimated:

$$\log(basketshare) = 0.83 + 0.491 \log(strawprice)$$

How does *basketshare* change if straw prices rise by 2%?

This is a log-log model, so if the price of straw increases by 2%, then the predicted share of baskets sold on sale increases by 0.98%.

$$\%\Delta y = 0.491 * 2\% = 0.98\%$$

Examples: Interpreting Marginal Effects

Suppose you've collected data on CEO salaries (hundred thousand \$) and annual firm sales (million \$), and you estimate the following model:

$$salary = 2.23 + 1.1 \log(sales)$$

According to the model, how does *salary* change if annual firm sales increase by 10%?

Examples: Interpreting Marginal Effects

$$salary = 2.23 + 1.1 \log(sales)$$

According to the model, how does $salary$ change if annual firm sales increase by 10%?

Sol. If annual firm sales increase by 10%, the model predicts that CEO salary increases by $11,000.

If annual firm sales increase by 10%, then we know $\%\Delta x = 10$.

$$\Delta y = (\beta_1/100)\%\Delta x$$

We can plug this and our estimate of β_1 into the formula from the table to see that $\Delta y = \frac{1.1}{100} * 10 = 0.11$. Since the units of CEO salaries is $100,000, an increase of 0.11 units is an increase of $11,000.

Stat Review: Random Variables

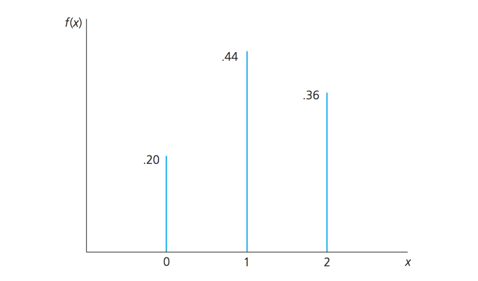
Random variables are numbers that are taken from a distribution of possible outcomes. A fundamental way to describe a random variable is through its probability distribution function.

Discrete random variable pdf:

$$f(x_j) = P(X = x_j), \;\; j = \{1, 2, 3, 4, 5, ...k\}$$

$$f(0) = 0.20\,; f(1) = 0.44\,; f(2) = 0.36$$

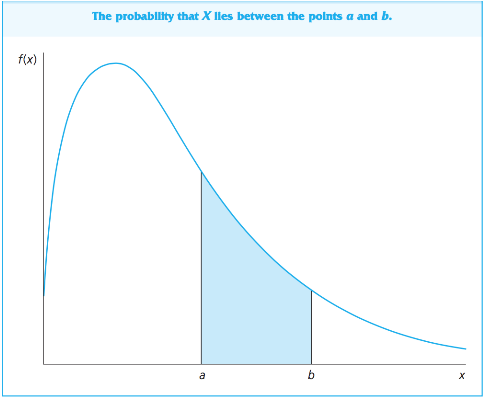
The pdf of the number of free throws made out of two attempts.



Stat Review: Random Variables

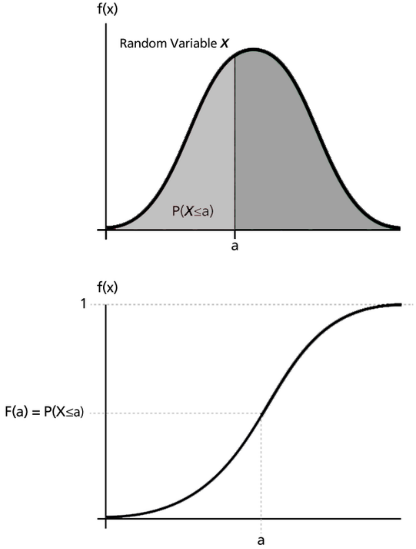
Continuous variable (pdf):

$$Pr(a < X < b) = \int_a^b f(x)dx$$



Stat Review: Random Variables

The cumulative distribution function is another useful way to visualize a random variable:



Stat Review: Random Variables

- If we have two discrete random variables X and Y, we can define the **joint probability density function** of (X,Y):

$$f_{X,Y} = P(X = x, Y = y)$$

- Two variables are **independent** if the joint PDF is equal to the product of the individual variables' pdf.

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

- The conditional distribution of Y given X, which is described by the **conditional probability density function** :

$$f_{(Y|X)}(y|x) = P(Y = y|X = x)$$

Stat Review: Random Variables

Let's do an example using survey data from 652 women:

	Head of household	
	Yes	No
Incomplete primary	30	124
Primary only	44	192
Secondary	123	139

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

- What is joint probability that a random person is a secondary school graduate and not a head of household?

Stat Review: Random Variables

Let's do an example using survey data:

	Head of household	
	Yes	No
Incomplete primary	30	124
Primary only	44	192
Secondary	123	139

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

- What is joint probability that a random person is a secondary school graduate and not a head of household?

$$f(Secondary, no) = 0.21$$

Stat Review: Random Variables

Let's do an example using survey data:

	Head of household	
	Yes	No
Incomplete primary	30	124
Primary only	44	192
Secondary	123	139

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

- What is the conditional probability that a randomly drawn head of household did NOT complete primary school?

Stat Review: Random Variables

Let's do an example using survey data:

	Head of household	
	Yes	No
Incomplete primary	30	124
Primary only	44	192
Secondary	123	139

	Head of household	
	Yes	No
Incomplete primary	0.05	0.19
Primary only	0.07	0.29
Secondary	0.19	0.21

- What is the conditional probability that a randomly drawn head of household did NOT complete primary school?

$$f(Incomplete|yes) = 30/197 = 0.15$$

Features of Probability Distributions

- **The expected value of X:**

$$E(X) = x_1 f(x_1) + x_2 f(x_2) + \cdots + x_k f(x_k) = \sum_{j=1}^{k} x_j f(x_j)$$

If X is continuous

$$E(X) = \int_{-\infty}^{+\infty} x f(x) d(x)$$

- **The variance of X:**

$$Var(X) = E[(X - E(X))^2]$$

- **The standard deviation of X**

$$sd(X) = \sqrt{Var(X)}$$

Sample Properties

An important distinction in this class is the difference between *populations* and *samples*. We deal with samples, so we can never know the real pdf or cdf of the population at large.

However, we can calculate the statistical properties of these samples:

- **Sample Mean:**

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- **Sample Variance:**

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

  Note: it seems like we should divide by n, but instead we divide by $n - 1$. We do this to ensure that the sample variance estimator is an unbiased estimator of population variance

Sample Properties: Law of Large Numbers

- In small samples, the sample mean can be quite different from the true population mean
  - For example, if I roll a five and a six on a die the sample mean will be $\frac{1}{2}(6+5) = 5.5$, even when we know the true *population* expected value of a die roll is 3.5:

$$E(X) = 1\left(\tfrac{1}{6}\right) + 2\left(\tfrac{1}{6}\right) + 3\left(\tfrac{1}{6}\right) + 4\left(\tfrac{1}{6}\right) + 5\left(\tfrac{1}{6}\right) + 6\left(\tfrac{1}{6}\right) = 3.5$$

- Usefully, the **law of large numbers** says that if we draw a sample consisting of *n* realizations of our random variable, and take the average, this sample mean will approach the population mean as *n* approaches infinity.
  - This means that if I roll a die more and more, my sample mean will approach the true population mean of 3.5