1. Regressions and Regression Models

Simply put, economists use regression models to study the relationship between two variables. If Y and X are two variables, representing some population, we are interested in "explaining Y in terms of X", or in determining "how Y varies with changes in X".

The classic example, common in labor economics, is to try and understand the relationship between income (Y) and education (X). When we talk about adding other X's (covariates or regressors) into our estimation, this means that we believe that other variables aside from education are also important in explaining variation in income such as work experience or parents' education.

We will go into this in much more detail as the course progresses, but for now, we can think of regression models as an estimated relationship between X and Y variables found in actual data.

The linear regression model assumes that the relationship between Y and X is linear - and as economists we then try to find the line that most closely approximates the true relationship. The appropriate picture to have in mind is the following:[1]

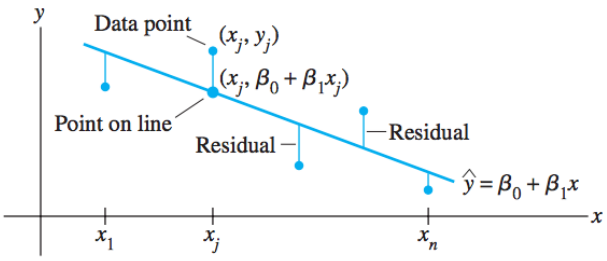

FIGURE 1 Fitting a line to experimental data.

2. Model Example

Watching TV one evening you come across a news program talking about Berkeley residents' health. The video shows an image of an emergency room in a hospital packed with people. The conditions of the hospital look to be very poor.

The news anchor says: "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." The following table is then shown on the screen:

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

What is the implied research question in this news story?

Do you agree with the news anchors conclusion? What other factors might contribute to whether or not someone recovers from illness? How could additional data or information improve your confidence in the anchors conclusion?

What are the following components of the regression model that would analyze this question if you had the data?

1. Dependent Variable (Y):
2. Explanatory/ independent of primary interest (X_1):

[1]Figure taken from: Lay, David C. Linear Algebra and It's Applications. 4th ed. Boston: Addison-Wesley, 2012. Print.

3. Additional Explanatory/ independent variables or covariates ($X_2, X_3, ...$):

3. Data Types

Knowing and understanding your data is *critical* to being a good economist. The form of your data will dictate which methods of data analysis we can choose from, which will, in turn, determine which different types of questions we can answer with it.

1. **Cross-Sectional Data:** Contains observations of different people, countries, firms, farmer etc. *at a single point in time*.
   - Example: a 2014 survey of Berkeley seniors on their academic record and extra-curricular participation.
2. **Time Series Data:** Contains a single person, country, firm, farmer etc. *over multiple points in time*.
   - Example: Data containing the weekly number of violent crimes committed in Los Angeles from 2010-2014.
3. **Pooled/Repeated Cross Section:** Contains multiple cross sections of people, countries, firms, farmers, etc *over multiple points in time where the observations are not necessarily repeated across rounds*.
   - Example: Current Population Survey in the United States. Each month a different set of households is surveyed about employment, unemployment etc.
   - This is also often referred to as a *repeated* cross section.
4. **Panel or Longitudinal Data:** You observe *the same* set of people, countries, firms, farmers, etc. *over multiple points in time*.
   - Example: The Indonesian Family Life Survey has been conducted across five rounds of data collection between 1993 and 2007 returning to the same households throughout the course of the study.
   - Panel data can either be shaped in long or wide format: below is an example of each:

Table 1: Data in Wide Format

	ID	Income_97	Income_98
1.	1	1000	2000
2.	2	4320	5000

Table 2: Data in Long Format

	ID	Year	Income
1.	1	97	1000
2.	1	98	2000
3.	2	97	4320
4.	2	98	5000

4. Random Variables and Distributions

a. Definitions

Let's briefly review the definitions of random variables and distributions

Random variables and their probability distributions: In essence, a random variable is a number that is taken from some distribution of possible outcomes. It can be **discrete** where there are a finite number of possible values (number of completed years of school) or **continuous** where there are infinite possible

values (a person's height). Once a random variable is drawn from the distribution, it becomes the **realization** of a random number.

Any discrete random variable can be completely described by detailing the possible values it takes, as well as the associated probability that it takes each value. The **probability density function** (**pdf**) of X summarizes the information concerning the possible outcomes of X and the associated probabilities. We can define a probability density function for continuous variables as well. However it doesn't make sense to talk about the probability that a continuous random variables takes on a particular value, rather we use the pdf to compute the probability of events involving a certain range.

- A bit of notation: We write f_X to denote the pdf of X. Then the probability that X takes on a certain discrete value j is $f(x_j) = P(X = x_j)$. The probability that X takes on a value within the interval $[a, b]$ is given by $Pr(a < X < b) = \int_a^b xf(x)dx$

Joint Distributions, Conditional Distributions, and Independence: Let X, and Y be discrete random variables. Then (X,Y) have a joint distribution, which can be described by 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. We might also be interested in establishing how X varies with different values of Y: this is the conditional distribution of Y given X, which is described by the **conditional probability density function** : $f_{(Y|X)}(y|x) = P(Y = y|X = x)$

b. Example

Take the following example of a survey of 652 women applying for a job at a factory. Two pieces of information that were collected include whether a woman was the head of her household and how much education she had completed. Look below at the following charts:

	Head of Household	
	Yes	No
No or incomplete primary	50	124
Primary only	84	192
Secondary	63	139

	Head of Household	
	Yes	No
No or incomplete primary	0.08	0.19
Primary only	0.13	0.29
Secondary	0.10	0.21

Note that the chart on the left gives the total number of women who fit in each *cell* of the chart. The sum of these cells is 652. From this chart, we could then calculate the chart on the right which tells us what proportion of women fall into each category. Each cell of the chart on the right provides us with the joint probability of two events happening.

- What is the joint probability that a randomly drawn person from the sample is a secondary school graduate and not a head of household?
- What is the conditional probability that a randomly drawn head of household has not completed primary school? Is this the same as the (unconditional) probability of someone randomly drawn from the full sample not having completed primary school?
- Are head of household status and education independent variables?
- If we only had the chart on the right, would we be able to recreate the one on the left? What other piece of information would we need?

5. Features of Probability Distributions

a. The Expected Value

If X is a random variable, the **expected value** (or expectation) of X, denoted E(X), is the weighted average of all possible values of X. The weights are determined by the probability density function. The expected

value is also called the population mean. Formally

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

Note if X is continuous

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

Now for a quick example:

x_j	$p(X = x_j)$
-1	1/8
0	1/2
2	3/8

$$E(X) = (-1)(1/8) + 0(1/2) + 2(3/8) = 5/8$$

b. The Variance and Standard Deviation

The **variance** tells us the exxpected distance from X to its mean:

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

Note the squaring eliminates the sign from the distance measure; the resulting positive value corresponds to the notion of distance, and treats values above and below symmetrically.

The **standard deviation** of a random variable, denoted sd(X) is the positive square root of the variance: $sd(X) = +\sqrt{Var(X)}$

Note

$$Var(aX + b) = a^2 Var(X)$$
$$sd(aX + b) = a \cdot sd(X)$$

This last property makes the standard deviation more natural to work with than the variance. As an example, take a random variable X measured in dollars. Next define Y=1000X. Suppose E(X)=20 and sd(X)=6. Then:

$$E(Y) = 1000E(X) = 20,000$$
$$sd(Y) = 1000sd(X) = 6,000$$
$$Var(Y) = (1000)^2 sd(X) = 6,000,000$$

The expected value and the standard deviation both increase by the same factor (1,000), whereas the variance of Y scales by 1,000,000.

c. Sample Mean and Law of large numbers

Let X denote a random variable, with Expected Value E(X) -population mean- . The sample mean can be expressed as:

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

The law of large numbers says that if we draw a sample consisting of n realizations of our random variable, and take the average (sample mean), this sample mean will approach the population mean as n approaches infinity. As an example, let X be the roll of a die, which can take on values 1,2,3,4,5,6. First, we can calculate the population mean (expected value)

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

Next, lets calculate the sample mean:

n	x_j	$\bar{X}_n$
2	6,6	12/2=6
3	1,2,2	5/3= 1.67
5	1,1,6,3	11/4=2.75
⋮		

As n$\rightarrow \infty$ any irregularities that occur due to the small sample size are muted, and the sample mean will converge to the population mean.

D. Sample Variance

Let X denote a random variable, with Variance Var(X). The sample variance is given by

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

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 (lots of terminology there, which we will explore later)