Bivariate Regression - Part I

I. Background. We have previously studied relationships between (a) Continuous dependent variable and a categorical independent variable (T-Test, ANOVA); and (b) Categorical Dependent variable and a categorical independent variable (Categorical data analysis, or Nonparametric tests).

We will now examine relationships between continuous dependent variables and continuous independent variables.

The following typology may be helpful:

Indep var \ Dep var	**Continuous**	**Discrete**
Continuous	OLS Regression	Logistic Regression
Discrete	T-Test, ANOVA	Categorical Data Analysis

II. Regression.

A. OLS. With OLS (Ordinary Least Squares) Regression, we are interested in how changes in one set of variables are related to changes in another set. That is, we want to describe or estimate the value of one variable, called the dependent variable, on the basis of one or more other variables, called independent variables.

Examples:

✓ What is the relationship between education and income? For each year of education, how much does income increase (on average)?

✓ What will be the rate of return on investment? For each dollar invested, how much will sales increase?

✓ For a political candidate, how many votes will she get for each dollar she spends on advertising?

It is usually not the case that the independent variables will perfectly predict the values of the dependent variable. For the most part, we are interested in determining the average relationship between the dependent and independent variable. That is, we want to know E(Y | X), i.e. for a particular value of X, what value, on average, do people have on Y?

For most regression problems, the average relationship between the dependent variable (y) and the independent variable (x) is assumed to be linear. That is, the *Population regression line* is

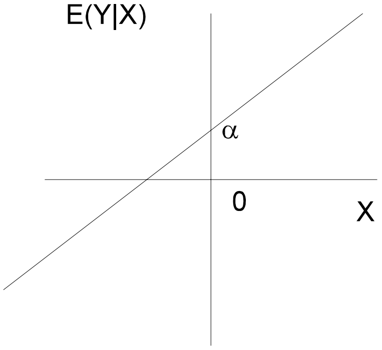
$$E(Y \mid X) = \alpha + \beta X$$

E(Y | X) = the average value of Y for a given value of X

ß = slope coefficient. This tells you how much a 1-unit increase in X affects the value of Y.

α = intercept. This is the point where the regression line crosses the Y axis, i.e. when X = 0, E (Y | X) = α. (NOTE: Do not confuse this α with the α we use when specifying significance levels!)

Graphically, we can show this as



Example. Suppose E(Y | X) = $5,000 + $1,500 X where Y = income and X = years of education. This means that, on average, a person with no education makes $5,000 a year. People with 12 years of education average $23,000, while those with 16 years of education average $29,000.

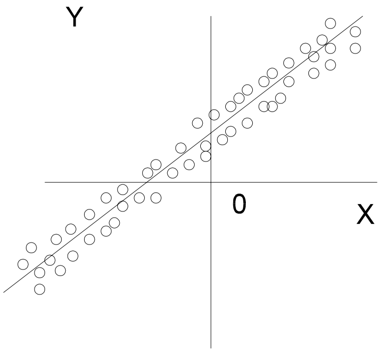
Of course, not all people with 16 years of education will make $29,000. Some will make more, some will make less. The score a particular individual has on Y can be written as:

$$y_i = \alpha + \beta\, x_i + \varepsilon_i\,; or,$$

$$y_i = E(y \mid x_i) + \varepsilon_i$$

Here, ε_i is a *random error term*, or *disturbance*.

A scatter diagram reflects this:



Note that, for any particular x_i, some values of y lie above the regression line, some below it.

B. *Sample estimation.* Of course, we don't know the values of the population parameters. They must be estimated from sample data. The *Sample regression line* is:

$$\hat{Y} = a + bX; or$$

$$\hat{Y} = \hat{\alpha} + \hat{\beta}X$$

and the *Sample regression model* is:

$$y_i = a + bx_i + e_i = \hat{y}_i + e_i\,; or$$

$$y_i = \hat{\alpha} + \hat{\beta}\,x_i + \hat{\varepsilon}_i = \hat{y}_i + \hat{\varepsilon}_i$$

Question: How do we determine values for a and b?
Answer: We could just plot the values, and draw a line by hand. But, two people might draw two different lines - and we would not have any means for determining sampling errors.

A better approach is to try to find values for a and b so as to minimize the values of e_i, where e_i = y_i - $\hat{y}_i$. The approach used to do this is called *Ordinary Least Squares*. With OLS, we choose a and b so as to minimize

$$\sum e_i^2 = \sum(y_i - \hat{y}_i)^2$$

Through calculus, we can show that the best values are

$$b = \frac{\frac{1}{N-1}\sum(x_i - \overline{x})(y_i - \overline{y})}{\frac{1}{N-1}\sum(x_i - \overline{x})^2} = \frac{s_{xy}}{s_x^2}\,,$$

$$a = \overline{y} - b\overline{x}$$

(NOTE: Hayes offers an informal proof on pp. 548-549).

Question: suppose x = $\overline{x}$. What does $\hat{y}$ equal?

$$\hat{y} = a + b\overline{x} = \overline{y} - b\overline{x} + b\overline{x} = \overline{y}$$

Hence, the regression line includes $(\overline{x}, \overline{y})$.

Question: Suppose b = 0. What does $\hat{y}$ equal?

$$\hat{y} = a + bx = \overline{y} - b\overline{x} + bx = \overline{y}$$

Hence, if the slope is zero, the best estimate of $\hat{y}$ is $\bar{y}$ - put another way, knowing x is of no value to you when predicting y.

C. Hypothesis testing. We are interested in whether the population parameter ß differs from zero. If ß = 0, then knowing X is of no use to us (since no matter what the value of X is, E(Y | X) = μ_y.) Therefore, we want to test

H_0: ß = 0
H_A: ß <> 0.

To do this, we have to make certain assumptions:

0. relationship between x and y is linear. This may or may not be a reasonable assumption.

1. cov(ε, x) = 0. The error terms are independent of the values of X. That is, the size of X has no relation to the size of the error term. An example of where this might not be true: Errors in predicted income may get bigger as education increases. At low education levels, most errors may be within a few thousand dollars; at higher education levels the errors may tend to be in the tens of thousands.

2. ε ~ Normal. Most of the actual values of Y will be close to the regression line.

3. E(ε) = 0. The average error will be zero; positive errors will be offset by negative errors.

4. COV(ε_k, ε_j) = 0 for j <> k. Knowing one error term tells you nothing about the value of another error term. A violation of this assumption might occur if samples are not independent of each other (e.g. husbands and their wives are treated as separate cases in one sample). Serial correlation is another common violation (Errors are correlated across time, as when you collect data on industries at multiple points in time).

5. V(ε_i | x_i) = σ_e^2 for all x. This is referred to as the assumption of homoskedasticity. Populations that do not have a constant variance are heteroskedastic.

D. Gauss-Markov Theorem: If assumptions 1, 3, 4, and 5 hold true, then estimators a and b determined by the least squares method are BLUE (Best Linear Unbiased Estimate) i.e. they are unbiased and have the smallest possible variance.