QS101: Introduction to Quantitative Methods in Social Science

Week 18: Linear Regression

Dr. Florian Reiche
Teaching Fellow in Quantitative Methods
Course Director BA Politics and Sociology
Deputy Director of Student Experience and Progression, PAIS

February 27, 2015

Linear Relationships

Our Enquiry

- Suppose we have data on how much time students spend on Facebook every day.

Our Enquiry

- Suppose we have data on how much time students spend on Facebook every day.
- We can see that the time spent online is higher for students who have many friends on Facebook

Our Enquiry

- Suppose we have data on how much time students spend on Facebook every day.
- We can see that the time spent online is higher for students who have many friends on Facebook
- So we we hypothesise that online times can be explained by the number of friends.

Our Enquiry

- Suppose we have data on how much time students spend on Facebook every day.
- We can see that the time spent online is higher for students who have many friends on Facebook
- So we we hypothesise that online times can be explained by the number of friends.
- If we want to put this hypothesis to a test, we can use regression analysis to establish whether this relationship exists:

Definition

Regression analysis is a statistical technique that attempts to "explain" movements in one variable, the dependent variable, as a function of movements in a set of other variables, called the independent (or explanatory) variables, through the quantification of a single equation. (Studenmund, 2006, p. 6, emphasis removed)

In its simplest setup, such an equation takes the following form:

$$y = \beta_0 + \beta_1 X \tag{1}$$

where y is the dependent variable, x is an independent variable and β_0 and β_1 are coefficients to be estimated.

- What is our dependent variable?

In its simplest setup, such an equation takes the following form:

$$y = \beta_0 + \beta_1 X \tag{1}$$

where y is the dependent variable, x is an independent variable and β_0 and β_1 are coefficients to be estimated.

- What is our dependent variable?
- What is our independent variable?

Graphical Depiction

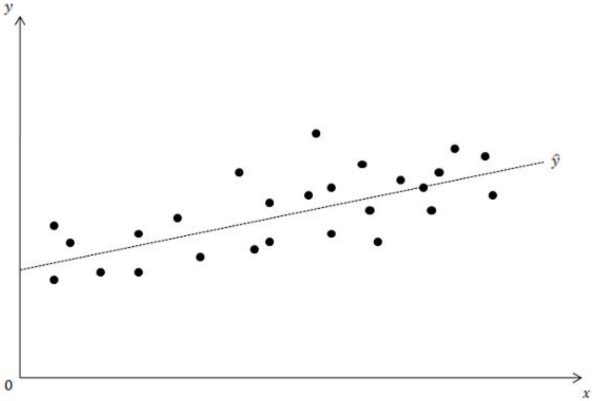

Figure: The Intuition of Regression

Interpretation

- We can see that there is indeed a positive relationship between X and Y and taking pen and ruler we can draw a "regression-line" $\hat{Y}$ through the plot which fits the data reasonably well.

Interpretation

- We can see that there is indeed a positive relationship between X and Y and taking pen and ruler we can draw a "regression-line" $\hat{Y}$ through the plot which fits the data reasonably well.
- The notation $\hat{Y}$ is chosen to denote the estimated regression line.

The Equation

- The βs are the coefficients that determine the coordinates of the straight line

The Equation

- The βs are the coefficients that determine the coordinates of the straight line
- β_0 is the intercept, or constant, it indicates where the line intercepts the y-axis

The Equation

- The βs are the coefficients that determine the coordinates of the straight line
- β_0 is the intercept, or constant, it indicates where the line intercepts the y-axis
- Another way of expressing this is: the value of Y when X equals zero

The Equation

- The βs are the coefficients that determine the coordinates of the straight line
- β_0 is the intercept, or constant, it indicates where the line intercepts the y-axis
- Another way of expressing this is: the value of Y when X equals zero
- β_1 is the slope coefficient

The Slope

- The slope indicates the amount that Y will change, if X increases by one unit

The Slope

- The slope indicates the amount that Y will change, if X increases by one unit
- Therefore:

The Slope

- The slope indicates the amount that Y will change, if X increases by one unit
- Therefore:

$$\frac{Y_2 - Y_1}{X_2 - X_1} = \frac{\Delta Y}{\Delta X} = \beta_1 \tag{2}$$

The Stochastic Error Term

Our Scatter Plot Again

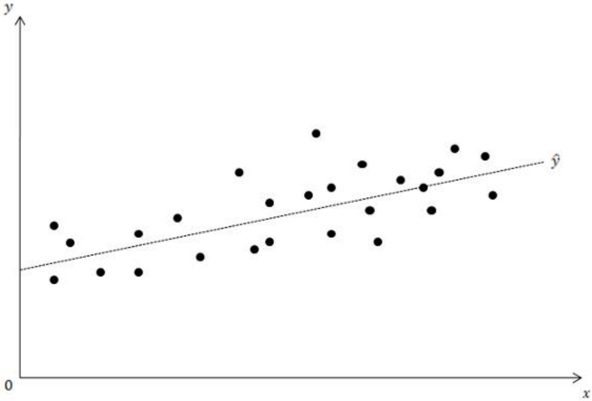

Figure: The Intuition of Regression

Remember we have fitted the following equation to the plot:

$$Y = \beta_0 + \beta_1 X \tag{3}$$

- However well this function is placed in the plot, there obviously remain differences between the observations and the regression line.
- These differences are called error terms, denoted as ϵ
- This is due to omitted influences, measurement error, purely random, . . .
- The inclusion of this term leads to the regression equation in its usual form

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{4}$$

The Full Equation

- This equation has two parts:
  - The deterministic part $\beta_0 + \beta_1 X$

The Full Equation

- This equation has two parts:
  - The deterministic part $\beta_0 + \beta_1 X$
  - The stochastic part ϵ

The Expected Value

- The deterministic component can be thought of as the expected value of Y, given X

The Expected Value

- The deterministic component can be thought of as the expected value of Y, given X
- Formally: $E(Y|X) = \beta_0 + \beta_1 X$

The Expected Value

- The deterministic component can be thought of as the expected value of Y, given X
- Formally: $E(Y|X) = \beta_0 + \beta_1 X$
- For example: The average amount of time spent on Facebook for a person with 100 friends is 3h per month

- The introduction of this error term is necessary, because "there are at least four sources of variation in [Y] other than the variation in the included [X]s:"

1. Many minor influences on [Y] are *omitted* from the equation (for example, because data are unavailable).
2. It is virtually impossible to avoid some sort of *measurement error* in the dependent variable.
3. The underlying theoretical equation might have a *different functional form* (or shape) than the one chosen for the regression. For example the underlying equation might be nonlinear.
4. All attempts to generalize human behavior must contain at least some amount of unpredictable or *purely random* variation.

(Studenmund, 2006, p. 11, see also Greene, 2008, p.9)

The Estimated Regression Equation

- The theoretical equation is abstract in nature:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{5}$$

- The theoretical equation is abstract in nature:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{6}$$

- The actual, estimated equation has numbers in it:

$$\hat{Y}_i = 50 + 12.5 X_i \tag{7}$$

where the subscript i denotes the i^{th} observation.

More Formally Again . . .

- We can re-write the estimated equation more generally again as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{8}$$

More Formally Again . . .

- We can re-write the estimated equation more generally again as:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i \tag{8}$$

- These "beta-hats" are empirical best guesses of the true regression coefficients from our sample data

Summary

- $\hat{Y}_i$ is the estimated value of Y_i

Summary

- $\hat{Y}_i$ is the estimated value of Y_i
- It represents the the value of Y calculated from the estimated regression equation for the i^{th} observation

Summary

- $\hat{Y}_i$ is the estimated value of Y_i
- It represents the the value of Y calculated from the estimated regression equation for the i^{th} observation
- The closer these $\hat{Y}$s are to the Ys, the better the fit of the equation

Residuals

- The difference between the estimated value of the dependent variable $\hat{Y}_i$ and the actual value of the dependent variable Y_i is defined as residual e_i

$$e_i = Y_i - \hat{Y}_i \tag{9}$$

Residuals versus Error Terms

- The error term is the difference between the observed Y and the true regression equation (the expected value of Y)

Residuals versus Error Terms

- The error term is the difference between the observed Y and the true regression equation (the expected value of Y)
- It is a purely theoretical concept and can NEVER be observed

Residuals versus Error Terms

- The error term is the difference between the observed Y and the true regression equation (the expected value of Y)
- It is a purely theoretical concept and can NEVER be observed
- The residual e_i meanwhile is the difference between the observed value Y and the estimated value $\hat{Y}$

Residuals versus Error Terms

- The error term is the difference between the observed Y and the true regression equation (the expected value of Y)
- It is a purely theoretical concept and can NEVER be observed
- The residual e_i meanwhile is the difference between the observed value Y and the estimated value $\hat{Y}$
- The residual can therefore be thought of as an estimate of the error term (e could be denoted as $\hat{\epsilon}$)

True and Estimated Regression Lines

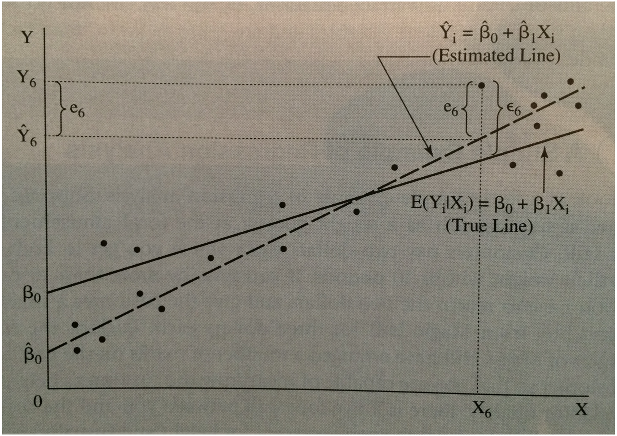

Figure: True and Estimated Regression Lines (source: Studenmund, 2014, p. 17)

Where do we go from here?

- The residual e_i are proving useful in estimating the regression line

Where do we go from here?

- The residual e_i are proving useful in estimating the regression line
- The associated method is called Ordinary Least Squares (OLS)

Where do we go from here?

- The residual e_i are proving useful in estimating the regression line
- The associated method is called Ordinary Least Squares (OLS)
- As the most frequently used estimation technique, we are going to look at it in more detail

Ordinary Least Squares (OLS)

- OLS follows the intuition that a regression line $\hat{Y}$ should fit the plot of data as well as possible (see Greene, 2008, p. 20)

- OLS follows the intuition that a regression line $\hat{Y}$ should fit the plot of data as well as possible (see Greene, 2008, p. 20)
- In order to achieve this, it minimises the sum of the squared residuals e_i

- OLS follows the intuition that a regression line $\hat{Y}$ should fit the plot of data as well as possible (see Greene, 2008, p. 20)
- In order to achieve this, it minimises the sum of the squared residuals e_i

$$\sum_i e_i^2 = \sum_i (Y_i - \hat{Y})^2 = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \qquad (10)$$

Graphical Depiction

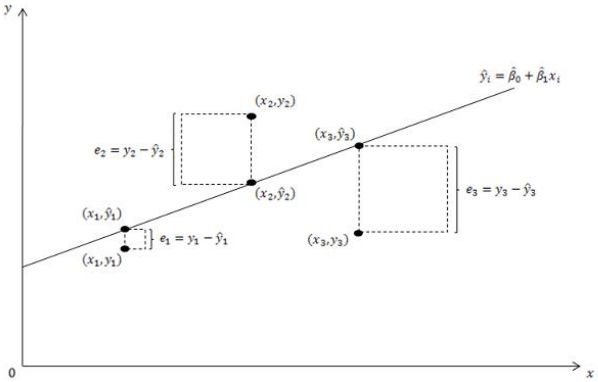

Figure: Ordinary Least Squares (OLS)

- Each residual e_i is equal to the distance between a data point Y_i and the corresponding estimated point $\hat{Y}_i$ on the regression line.

- Each residual e_i is equal to the distance between a data point Y_i and the corresponding estimated point $\hat{Y}_i$ on the regression line.
- OLS does not use the mere distance in its process, however, but squares it so as to prevent negative distances levelling out positive ones when taking the sum.

- Each residual e_i is equal to the distance between a data point Y_i and the corresponding estimated point $\hat{Y}_i$ on the regression line.
- OLS does not use the mere distance in its process, however, but squares it so as to prevent negative distances levelling out positive ones when taking the sum.
- Rather than fiddling with pen and ruler (and very probably rubber) which becomes impossible with more than two variables anyway, OLS allows the researcher to estimate the coefficients minimising the residuals.

Outlook

- Transforming equation 10 into matrix terms, it can be re-written as . . .

Outlook

- Transforming equation 10 into matrix terms, it can be re-written as . . .
- . . . we will see this in week 10

Next Week

- How do we extend this linear model to incorporate more independent variables?

Next Week

- How do we extend this linear model to incorporate more independent variables?
- How does this relate to correlation and to ANOVA?