TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA

MSc European Economy and Business Law

Statistics Pre-course 2022

Topic 5: Basic Regression Models

Francesca Centofanti

francesca.centofanti@students.uniroma2.eu

Outline

- Introduction
- Regression Line
- Least squares Method
- Regression Analysis
- Appendix: OLS inference

Introduction: Recap of Association

- We start recalling the concept of **Association**:

  An *association* exists between two variables if a particular value for one variable is more likely to occur with certain values of the other variable
- When there is an association, the **likelihood** of a particular value for one variable depends on the value of the other variable
- When we analyse data on two variables, our first step is to distinguish between the **response variable** and the **explanatory variable**:
  1. The *response variable* is the **outcome variable** on which comparisons are made
  2. When the *explanatory variable* is **categorical**, it defines the groups to be compared with respect to values for the response variable; when the *explanatory variable* is **quantitative**, it defines the change in different numerical values to be compared with respect to values for the response variable

Introduction: Response and Explanatory Variables

The *data analysis* examines how the outcome on the response variable depends on or is explained by the value of the explanatory variable, describing the **nature of the association** (if there is any)

- With two quantitative variables, it is common to denote the **response variable** y and the **explanatory variable** x
- We use this notation because graphical plots for examining the association use the $y - axis$ for values of the response variable and the $x - axis$ for values of the explanatory variable
- This graphical plot is called **scatterplot**

Graphical Representation: Scatterplot

A *scatterplot* is a graphical display for two quantitative variables using the horizontal $x - axis$ for the explanatory variable x and the vertical $y - axis$ for the response variable y. The values of x and y for a subject are represented by a **point** relative to the two axes. The observations for the n subjects are n points on the scatterplot.

Example: How to examine a scatterplot

- We examine a scatterplot to study association:
  1. How do values on the response variable change as values of the explanatory variable change? As Internet use gets higher, for instance, we see that Facebook use gets higher
  2. When there is a trend in a scatterplot, what is the direction? Is the association positive or negative? The figure displays a positive association, because high (low) values of Internet use tend to occur with high (low) values of Facebook use

REMEMBER! *Positive association*: As x goes up, y tends to go up; *Negative association*: As x goes up, y tends to go down

Introduction: Recap of Correlation

- When the data points follow a roughly *straight-line trend*, the variables are said to have an approximately **linear relationship**
- In some cases, the data points fall close to a straight line, but more often there is quite a bit of variability of the points around the straight-line trend
- A *summary measure* called **correlation** describes the strength of the linear association

The correlation summarizes the *direction* of the association between two quantitative variables and the *strength* of its linear (straight-line) trend.

Denoted by r, it takes values between -1 and $+1$

1. A positive value for r indicates a positive association and a negative value for r indicates a negative association
2. The closer r is to ± 1 the closer the data points fall to a straight line, and the stronger the linear association is. The closer r is to 0, the weaker the linear association is

Regression Line

- We have seen that when the relationship has a straight-line pattern, the correlation describes it *numerically*
- We can analyse the data further by finding an *equation* for the straight line that best describes that pattern

The *regression line* predicts the value for the response variable y as a straight-line function of the value x of the explanatory variable

- Let $\widehat{y}$ denote the predicted value of y. The equation for the regression line has the form

$$\widehat{y} = a + bx$$

- The **Regression Line** is an equation for *predicting* the response outcome (it is often called a *prediction equation*)
  1. a denotes the $y - intercept$
  2. b denotes the slope

Regression Line: y – *intercept* and Slope interpretation

- The **y-intercept** is the predicted value of y when $x = 0$. This fact helps us plot the line, but it may not have any interpretative value if no observations had x values near 0
- The **slope** b equals the amount that $\widehat{y}$ changes when x increases by one unit. For two x values that differ by 1.0, the $\widehat{y}$ values differ by b
  1. When the slope is **positive**, the predicted value $\widehat{y}$ increases as x increases. The straight line then goes upward, and the *association is positive*
  2. When the slope is **negative**, the predicted value $\widehat{y}$ decreases as x increases. The straight line then goes downward, and the *association is negative*
  3. When $b = 0$, the regression line is horizontal (parallel to the x – *axis*). The predicted value $\widehat{y}$ of y stays constant at the y – *intercept* for any value of x. Then the predicted value does not change as x changes, and the *variables do not exhibit an association*

Regression Line: Find the Regression Equation

How can we use the **data** to find the *regression equation*?

- We should first construct a *scatterplot* to make sure that the relationship has a roughly straight line trend
- If so, then software or calculators can easily find the straight line that best fits the data
- Once we have used the regression equation, we can compare the predicted values to the actual values to check the accuracy of those *predictions*

Regression Equation: Residuals

The *prediction error* is the difference between the actual y value and the predicted value, which is $y - \widehat{y}$

- These prediction errors are called **residuals**
- Each observation has a residual (positive or negative)
  1. A *positive residual* occurs when the actual y is larger than the predicted value $\widehat{y}$, so that $y - \widehat{y} > 0$
  2. A *negative residual* results when the actual y is smaller than the predicted value $\widehat{y}$, so that $y - \widehat{y} < 0$
- The smaller the absolute value of a residual, the closer the predicted value is to the actual value, so the better is the prediction
- In a *scatterplot*, the vertical distance between the point and the regression line is the absolute value of the residual

Least Squares Method: How to get the Regression Line

- We have seen that software find the regression line choosing the optimal line to fit through the data points by making the residuals as small as possible
- This process involves compromise because a line can perfectly predict one point (resulting in a residual of 0), but poorly predict many other points (resulting in larger residuals)
- To *evaluate* a regression line, we can construct the summary measure called **Residual Sum of Squares** (RSS):

$$RSS = \sum (residuals)^2 = \sum (y - \hat{y})^2$$

- This formula squares each vertical distance between a point and the line and then adds them up
- Each potential line has a set of predicted values, a set of residuals, and a residual sum of squares
- The line that software report is the one having the *minimum* residual sum of squares and this way of selecting a line is called the **least squares method**

Least Squares Method: Properties

Among the many possible lines that could be drawn through data points in a scatterplot, the *least squares method* gives what we call the regression line. This method produces the line that has the **smallest value** for the *residual sum of squares* using $\widehat{y} = a + bx$ to predict y

- **Properties** of the Regression Line:
  1. It has some positive residuals and some negative residuals, and the sum (and mean) of the residuals equals 0 $\Rightarrow$ It tells us that the too-low predictions are balanced by the too-high predictions
  2. It passes through the point $(\overline{x}, \overline{y})$ $\Rightarrow$ It tells us that the line passes through the center of the data.

BE CAREFUL! The fist property ($\sum(residuals) = 0$) is the reason why we use the Residual Sum of Squares ($\sum(residuals)^2$) to evaluate the regression line

Least Squares Method: Theoretical Formulas

- Even though we usually rely on technology to compute the regression line, the method of least squares does provide formulas for the $y-$ *intercept* and slope, based on *summary statistics* for the sample data
- Let $\overline{x}$ denote the *mean* of x, $\overline{y}$ the *mean* of y, s_x the *standard deviation* of the x values and s_y the *standard deviation* of the y values:
  1. The **slope** equals $b = r(\frac{s_y}{s_x})$, where r is the correlation
  2. The **y-intercept** equals $a = \overline{y} - b(\overline{x})$

Regression Analysis: Regression Model

- At a given value of x, the equation $\widehat{y} = a + bx$ predicts a single value $\widehat{y}$ of the response variable. However, we should not expect all subjects at that value of x to have the same value of y. **Variability** occurs in their y values

Example: let x =number of years of education and y =annual income in dollars for the adult residents in the workforce of your hometown. For a random sample, suppose you find $\widehat{y} = -20,000 + 4000x$. Those workers with $x = 12$ years of education have predicted annual income

$$\widehat{y} = -20,000 + 4000 \times 12 = 28,000$$

It is not the case that every worker with 12 years of education would have annual income \$28,000, since income is not completely dependent upon education.

- We can think of $\widehat{y} = 28,000$ as estimating the *mean annual income* for all workers with $x = 12$
- Likewise, there is a mean of the annual income values at each separate education value

Regression Model: Population Regression Equation

- A similar equation describes the relationship in the *population* between x and the means of y

$$\mu_y = \alpha + \beta x$$

- We estimate the **population regression equation** using the *prediction equation* for the sample data
- A straight line is the simplest way to describe the relationship between two quantitative variables. In practice, most relationships are not exactly linear. The equation $\mu_y = \alpha + \beta x$ **merely approximates** the actual relationship between x and the population means of y

A *model* is a simple approximation for how variables relate in a population

Regression Model: Definitions

A *regression model* describes how the population mean μ_y of each conditional distribution for the response variable depends on the value x of the explanatory variable

- A straight-line regression model uses the line $\mu_y = \alpha + \beta x$ to connect the means
- The model also has a parameter σ (standard deviation) that describes the *variability* of observations around the mean of y at each x value

Regression Model: Predictive Power

- Another way to describe the *strength* of association refers to how close predictions for y tend to be to observed y values
- The variables are **strongly associated** if you can predict y much better by substituting x values into the prediction equation $\widehat{y}$ than by merely using the sample mean $\overline{y}$ and ignoring x
- Recalling that, for a given subject, the *prediction error* is the difference between the observed and predicted values of y:
  1. The error using the regression line to make a prediction is $y - \widehat{y}$
  2. The error using $\overline{y}$ to make a prediction is $y - \overline{y}$
- For each potential predictor ($\widehat{y}$ and $\overline{y}$) we can summarize the sizes of the errors by the sum of their squared values
  1. **Residual Sum of Squares**: when we predict y x (with the regression equation) $\Rightarrow \sum(y - \widehat{y})^2$
  2. **Total Sum of Squares**: when we predict y using $\overline{y}$ (that is, ignoring x) $\Rightarrow \sum(y - \overline{y})^2$

Regression Model: Predictive Power (cont'd)

- The difference between the two *error summaries* depends on the units of measure
- We can eliminate this dependence on units by converting the difference to a proportion, obtaining the *summary measure of association*:

$$r^2 = \frac{\sum(y - \overline{y})^2 - \sum(y - \widehat{y})^2}{\sum(y - \overline{y})^2}$$

- This measure is interpreted as the *proportional reduction in error*
- We use the notation r^2 for this measure because, in fact, it can be shown that this measure equals the square of the correlation r
- In practise, if you know the correlation r, it is simple to calculate r^2 by squaring the correlation (the formula shown previously for r^2 is useful for interpretation but it is not needed for calculation)

Regression Model: Final Remarks on Association

- **Properties** of r^2
  1. Since $-1 \leq r \leq 1$, r^2 falls between 0 and 1
  2. $r^2 = 1$ when $\sum(y - \widehat{y})^2 = 0$, which happens only when all the data points fall exactly on the regression line. There is then no prediction error using x to predict y (that is, $y = \widehat{y}$ for each observation). This corresponds to $r = \pm 1$
  3. $r^2 = 0$ when $\sum(y - \widehat{y})^2 = \sum(y - \overline{y})^2$. This happens when the slope $b = 0$, in which case each $\widehat{y} = \overline{y}$. The regression line and then give the same predictions
  4. The closer r^2 is to 1, the stronger the linear association; the more effective the regression equation $\widehat{y} = a + bx$ then is compared to $\overline{y}$ in predicting y
- Both the correlation r and its square r^2 describe the strength of association, but they have different interpretations:
  1. The correlation represents the slope of the regression line when x and y have equal standard deviations (it governs the extent of "regression toward the mean")
  2. r^2 summarizes the reduction in sum of squared errors in predicting y using the regression line instead of using the mean of y

Appendix: (Ordinary) Least Squares inference

- Recalling that any Regression Model is just a simple *approximation* for how variables relate in a population, the linear regression equation $\mu_y = \alpha + \beta x$ can be expressed in an alternative way:

$$Y_i = \alpha + \beta x_i + \epsilon_i \qquad i = 1, ..., n$$

  where $\epsilon_1, ..., \epsilon_n$ are uncorrelated random variables, with $\mathbb{E}[\epsilon_i] = 0$ and $\mathbb{V}[\epsilon_i] = \sigma^2$, called **random errors**
- Intuitively, the *error term* ϵ_i includes all the "elements" (typically unobserved) that concur to predict the dependent variable y_i but cannot be included in the explanatory variable x_i

(Ordinary) Least Squares inference

- As we know, the **OLS method** aims to estimate the parameters β_0 (intercept) and β_1 (slope) by *minimizing*:

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_1))^2$$

- To obtain the formulas for the estimated parameters, we apply the *maximum likelihood* to the OLS equation:

$$\begin{cases} \frac{\delta L(\beta_0, \beta_1)}{\delta \beta_0} = 0 \\ \frac{\delta L(\beta_0, \beta_1)}{\delta \beta_1} = 0 \end{cases}$$

- Solving the system, one yields the following (alternative) expressions for the estimated parameters:
  1. $b_1 = \frac{SS_{xy}}{SS_{xx}}$
  2. $b_0 = \overline{y} - b_1$

  where $SS_{xy} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$ and $SS_{xx} = \sum_{i=1}^{n} (x_i - \overline{x})^2$

OLS inference

- To establish the *statistical significance* of the model, the researcher use the following **hypothesis system** (the intercept should be kept in the model anyway):

$$\begin{cases} H_0 : \beta_1 = 0 \\ H_1 : \beta_1 \neq 0 \end{cases}$$

- As we know from the steps of the *Testing Procedure*, we need a Test Statistics for the Hypothesis Testing
- In our previous *Example* we used a **z-statistics**, distributed as a Standard Normal ($Z \sim N(0,1)$)
- For this test is usually used a **t-statistics**, distributed ad a t-student:

$$t = \frac{B_1}{s(B_1)} \sim t_{n-2}$$

where B_1 is the estimated value of the coefficient β_1 and $s(B_1)$ is the standard error, obtained by the square root of $\mathbb{V}[B_1] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$

OLS inference: t-student distribution

- The *t-student* distribution is a symmetric and unimodal distribution with $mean = 0$.
- This statistics is constructing using the *Standard Normal* and the *Chi-square* (χ^2) distribution, using **degrees of freedom** (numbers of observations minus the parameters to be estimated)

$$t = \frac{Z}{\sqrt{\frac{V}{v}}}$$

where Z is a Standard Normal distribution and V has a Chi-square distribution with degree of freedom v (is also the sum of Z^2)

OLS inference

- Further, for ascertain the *goodness of fit* of the model, we use an alternative expression of the summary measure of association, called **coefficient of determination**:

$$r^2 = \frac{SSR}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

where

1. SSR(sum of squares of regression)$= \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2$
2. SSE(sum of squares of error or *residual sum of squares*)$= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum(\widehat{e}_i)^2$
3. SS_{yy}(sum of squares of y or *total sum of squares*)$= \sum_{i=1}^{n}(y_i - \overline{y})^2 = SSR + SSE$

OLS inference: OLS Assumptions

- Methods of estimation are based on *theoretical assumptions*
- If these assumptions do not hold, the estimator proposed by the method could be not the best choice for estimating the parameters of the regression model
- For using OLS, the **assumptions** are:
  1. Linearity (in the parameters and in x_i)
  2. Random sampling: the sample taken for the linear regression model must be drawn randomly from the population; the number of observations should be greater than the number of parameters; xs should be fixed (dependent variable should be affected by independent ones)
  3. ε_i are i.i.d (independently and identically distributed) and $\mathbb{E}[\varepsilon_i|x]$
  4. No multicollinearity (for multiple regressions): one explanatory variable can be linearly predicted from the others used in the same regression
  5. $\mathbb{V}[\varepsilon_i|x] = \sigma^2$ (homoschedasticity) and $Cov(\varepsilon_i\varepsilon_j) = 0$ (no autocorrelation: error terms of different observations should not be correlated with each other)

OLS inference: OLS Assumptions

- If **all** the assumptions hold, then OLS method is **BLUE** (Best Linear Unbiased Estimator)
- If $\mathbb{E}[\varepsilon_i|x]$ does not hold, than the OLS estimates are biased
- Looking in particular at **Assumption 5**:
  1. If the *homoschedasticity* assumption ($\mathbb{V}[\varepsilon_i|x] = \sigma^2$) does not hold, then the standard errors $s(B_1)$ are not reliable and the hypothesis test for the statistical significance of the estimated parameter can be falsified
  2. If the *no-autocorrelation* assumption ($Cov(\varepsilon_i \varepsilon_j) = 0$) does not hold, OLS is no longer BLUE