Quantitative analysis methods for public policies

Regression

Co-funded by the
Erasmus+ Programme
of the European Union

Keywords

- Simple linear regression
- Multiple linear regression
- Dependent variable
- Independent variables
- R^2 value
- Regression coefficients

Co-funded by the
Erasmus+ Programme
of the European Union

When do we apply it?

- It is a step forward from the correlation analysis in the sense that prediction is now the new element in the analysis
- More specifically, there is a set of variables (called independent variables-IV) that presumably affect one other variable (called dependent variable-DV).
- Examples include the relationship between income and a set of independent variables such as years of education, previous job experience, etc.
- The researcher wishes to investigate two things:

(a) Do the IV as a whole affect the DV?

(b) If yes, which is the contribution of each IV?

(c) How well can we predict the DV for specific values of the IV?

Comments

- Initially, only continuous variables will be considered.
- Extensions include discrete DV (logistic regression) and a mix of discrete and continuous IV (Analysis of variance/covariance)
- There may other IV which fit our regression model better but are not available to the researcher.
- In case of one DV (IV is always one), we have the case of simple linear regression denoted by the equation

$$Y = \alpha + \beta X + \varepsilon$$

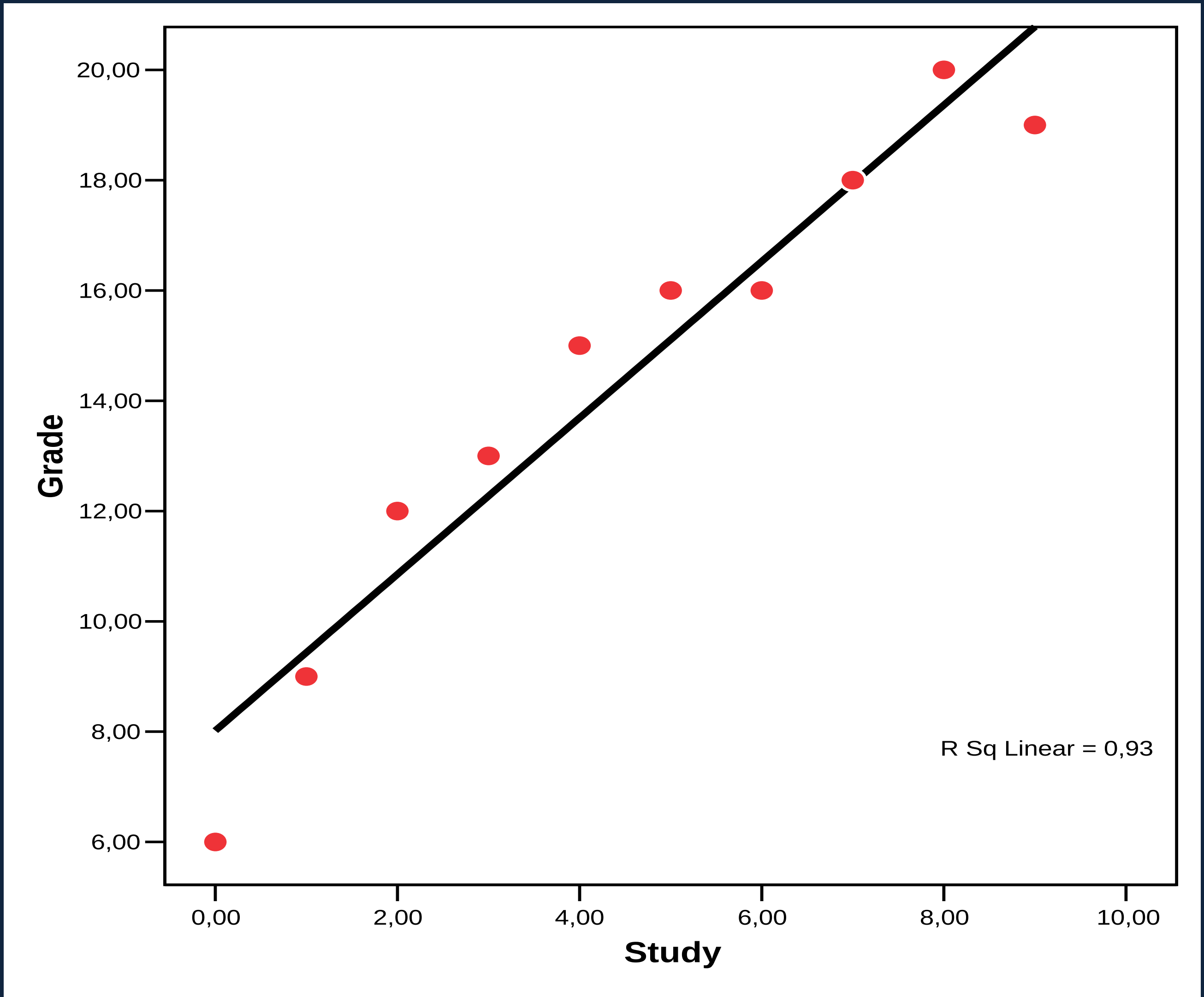
Where Y stands for the DV, X for the IV and ε for the omnipresent statistical error.

Co-funded by the Erasmus+ Programme of the European Union

Comments (2)

- In the previous simple linear regression equation a and b are the regression coefficients.
- Since they are not known (the error is unknown), they are estimated in a way that the error is made as small as possible
- Thus, a and b are called the regression coefficients and help us predict the value of Y for specific value of X.
- In the simple linear regression model their relationship may be described with a line among the data. The direction of the line is based on the estimation of the regression coefficients a and b.
- Example: Let Y=student grade in a 20-point scale and X=hours of daily study. The regression line may look like the one in the next diagram.

Regression line



Co-funded by the
Erasmus+ Programme
of the European Union

Interpretation of the results

- The software provides an estimation of the regression coefficients as well as a hypothesis test to check if they are statistically significant from zero (i.e., if the regression model is meaningful or not).
- Suppose that a=8.02 and b=1.42. Thus we have the equation that:

Predicted Grade=8.02+1.42*Hours

- Assuming that these tests are significant, the researcher can gain valuable information about the predicted grade. Thus for 4 hours of study the predicted grade is expected to be 13.7 (8.02+1.42*4)

Interpretation of the results (2)

- The value of b denotes the expected gain in grade if the student studies one hour more.
- Note the use of the term “expected” since the actual change may vary (although not by much if the error term is small)
- Thus, 5 *more* hours of study are expected to increase the grade by 1.42*5=7.1 points.
- The value of a means that if a student does not study at all (X=0), she he is expected to take 8.02 in the test.
- Note that the interpretation of a may not always be meaningful. For example if X is years of age then X=0 does not apply to students.
- Another important value is R^2 . The closer to 100%, the better the estimation is. In social sciences most R^2 values are around 30%-40%

Co-funded by the Erasmus+ Programme of the European Union

Multiple linear regression

- It includes more than one IV.
- It becomes exceedingly difficult or simply impossible in most cases to produce a regression line or a plot
- A difference from the case of the simple linear regression is the fact that the researcher initially examines if the IV as a set significantly affect the DV
- If the overall effect is significant then each separate DV is examined to gauge the biggest impact
- The interpretation of the regression coefficients remain mostly the same with the addition that a change in one of the X's (the IV) should not result to the simultaneous change of some of the other X's.

Co-funded by the Erasmus+ Programme of the European Union

Analysis of Variance (ANOVA)

- The main difference from regression analysis is that the IV are discrete.
- It investigates the effect of one or more discrete IV to a continuous DV.
- Examples include the effect of Gender and educational level in people’s income
- In the case of one DV, the model is called one-way Analysis of Variance or one-way ANOVA.

Co-funded by the Erasmus+ Programme of the European Union

One-way ANOVA

- Let's assume that a researcher wishes to examine the effect of educational level (recorded as primary/secondary/tertiary) to employees income.
- This is a straightforward setup of one-way ANOVA examining the effect of the 3-level education IV on the continuous income DV
- The idea is to examine if the average income across the three educational level varies significantly.
- If a statistical difference is established among the three level of education, then the researcher wishes to examine in which pair(s) this difference lies (primary-second / primary-tertiary /secondary-tertiary)

Co-funded by the Erasmus+ Programme of the European Union

Logistic regression

- The main difference from regression analysis is that the DV is discrete.
- It investigates the effect of one or more discrete or continuous IV to a discrete DV.
- Examples include the effect of Gender and educational level in being employed
- The case of a DV having only two values is the most common one and is referred to as binary logistic regression.

Co-funded by the Erasmus+ Programme of the European Union

Binary logistic regression

- Let’s assume that a researcher wishes to examine the effect of being a minority (yes/no) to the probability of someone being employed or not.
- This is a straightforward setup of a binary logistic regression on a discrete IV. It may be handled also through chi-square testing, although in the case of more than one IV a logistic regression models is necessary.
- The idea is to examine if the probability of having a job (or not) differs across the two minority levels significantly. Furthermore, one of the two minority levels is set arbitrarily as a reference level to examine the difference in the probability of the other one compared to that level.

Co-funded by the Erasmus+ Programme of the European Union