FORK 1002 Preparatory Course in Statistics:

2 Regression Analysis

Genaro Sucarrat (BI)
http://www.sucarrat.net/

Contents:

1. Bivariate Correlation Analysis
2. Simple Regression
3. Estimation and Fit
4. T-Test: Single Hypothesis Testing
5. Multiple Regression
6. F-Test: Multiple Hypothesis Testing

Bivariate Correlation Analysis

Bivariate Correlation: A statistical association between two variables

Pearson's r ("the" correlation): A linear measure that varies between -1 and 1

- $r = 1$: Perfect positive linear correlation
- $r = 0$: No correlation
- $r = -1$: Perfect negative linear correlation

Formula of Pearson's r:

$$r = \frac{s_{xy}}{s_x \cdot s_y}$$

where s_{xy} is the sample covariance between x and y

Simple Regression:

The simple regression model:

$$Y = B_1 + B_2 X + u \tag{1}$$

where

- Y: Dependent, endogenous or left-hand side variable
- X: Independent, exogenous or right-hand side variable
- B_1: Intercept, i.e. the average value of Y when $X = 0$
- B_2: Slope coefficient, the effect on Y when X increases by 1 unit
- $B_1 + B_2 X$: The regression line, i.e. the "explanation" or prediction of Y
- u: The error, i.e. the unexplained part or prediction error

Estimation and Fit:

- How do we (usually) estimate B_1 and B_2? OLS!
- OLS = Ordinary Least Squares
- OLS consists of choosing the values of B_1 and B_2 such that the sum of squared errors (i.e. $\sum u^2$) is minimised
- Note (for simple regression): $\sum u^2 = \sum (Y - B_1 - B_2 X)^2$
- The OLS formulas for simple regression:

  Estimate of B_2: $\frac{s_{XY}}{s_X^2}$

  Estimate of B_1: $\overline{Y} - \hat{B}_2 \overline{X}$, where $\hat{B}_2$ is the estimate of B_2
- *R-squared* (R^2): A measure of a model's explanatory or predictive power
  - Varies between 0 and 1
  - $R^2 = 1$: The model perfectly predicts or explains the variation in Y
  - $R^2 = 0$: The model predicts or explains nothing of the variation in Y
- In simple regression we have that $R^2 = r^2$ (Pearson's r squared)
- The formula of R^2: $R^2 = 1 - RSS/TSS$, where RSS = Residual Sum of Squares (i.e. $\sum u^2$) and TSS = Total Sum of Squares

The t-Test: Simple Hypothesis Testing

The t-test in 4 steps:

1. Choose α, formulate H_0 and H_A. Examples:
   - $H_0 : B = 0$ $H_A : B \neq 0$ (two-sided)
   - $H_0 : B = 3$ $H_A : B > 3$ (right-sided)
   - $H_0 : B = 1$ $H_A : B < 1$ (left-sided)
2. Find the critical value(s) and identify the rejection area:
   - Use a t-distribution with degrees of freedom $df = n - k$, where k is the number of Bs (in simple regression: $k = 2$)
3. Compute the value of the test expression:

$$\frac{\hat{B} - H_0 \text{ value}}{\text{Standard error of } \hat{B}}$$

   where $\hat{B}$ is an estimate of B
4. Conclude: Reject H_0 if the value of the test expression is in the rejection area

Multiple Regression

The multiple regression model:

$$Y = B_1 + B_2 X_2 + \cdots + B_k + u \tag{2}$$

where

- Y: Dependent, endogenous or left-hand side variable
- Xs: Independent, exogenous or right-hand side variables
- B_1: Intercept, i.e. the average value of Y when all the Xs are 0
- $B_2, \ldots, B_k$: Slope coefficients. Each B is typically interpreted as the effect on Y when the associated X increases by 1 unit, *given* that the other Xs do not change
- $B_1 + B_2 X_2 + \cdots + B_k X_k$: The regression line, i.e. the "explanation" or prediction of Y
- u: The error, i.e. the unexplained part or prediction error
- RSS and R^2 defined as in simple regression, that is, $RSS = \sum u^2$ and $R^2 = 1 - RSS/TSS$

The F-Test: Multiple Hypothesis Testing

The F-test in 4 steps:

1. Choose α, formulate H_0 and H_A. Example:
   - $H_0 : B_1 = 39$ and $B_2 = 5$
   - H_A : One or more of the equalities in H_0 are false
2. Find the critical value(s) and identify the rejection area:
   - Use an F-distribution
   - Numerator degrees of freedom: $df_1 =$ number of equalities in H_0
   - Denominator degrees of freedom: $n - k$, where k is the number of Bs in the unrestricted model
3. Compute the value of one of the following test expressions:
   - $\frac{(R^2_{UR} - R^2_R)/df_1}{(1 - R^2_{UR})/df_2}$
   - $\frac{(RSS_R - RSS_{UR})/df_1}{RSS_{UR}/df_2}$
4. Conclude: Reject H_0 if the value of the test expression is in the rejection area