QS101: Introduction to Quantitative Methods in Social Science

Week 19: Multivariate Regression

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

March 5, 2015

Recap

Queries

- What is regression?

Queries

- What is regression?
- What does the intercept tell us?

Queries

- What is regression?
- What does the intercept tell us?
- What does the slop indicate?

Queries

- What is regression?
- What does the intercept tell us?
- What does the slop indicate?
- What is OLS?

Multivariate Regression

Reminder

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

$$Y_i = \beta_0 + \beta_1 X_i \tag{1}$$

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

- In such a setup we only have ONE independent variable
- This setup is not terribly realistic
- For example, your time spent on Facebook might also depend on revision time, time spent in societies, etc.
- Therefore, we need to extend the model, as follows:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \quad (2)$$

Example

Suppose we receive the following results of our estimation:

$$\widehat{\text{Facebook}}_i = 5 + 2.2\text{FRIEND}_i - 1.9\text{SOCIETY}_i \quad (3)$$

where

- FRIEND is the number of friends on Facebook
- SOCIETY is the amount of hours spent in societies

Interpretation of Coefficients in Multiple Regression

- How do we interpret these coefficients?

Interpretation of Coefficients in Multiple Regression

- How do we interpret these coefficients?
- 2.2 means that for each additional friend on Facebook, we spend 2.2 hours more online, holding constant SOCIETIES

Interpretation of Coefficients in Multiple Regression

- How do we interpret these coefficients?
- 2.2 means that for each additional friend on Facebook, we spend 2.2 hours more online, holding constant SOCIETIES
- This is referred to as *ceteris paribus*, Latin for "all other things being equal"

Goodness of Fit

Total, Explained and Residual Sums of Squares

- Goodness of fit: How much of the variation in the dependent variable is explained by the estimated regression equation?

Total, Explained and Residual Sums of Squares

- Goodness of fit: How much of the variation in the dependent variable is explained by the estimated regression equation?
- For this, we can use the Total Sum of Squares

Total, Explained and Residual Sums of Squares

- Goodness of fit: How much of the variation in the dependent variable is explained by the estimated regression equation?
- For this, we can use the Total Sum of Squares
- This is the squared variation of Y around its mean and is written as:

Total, Explained and Residual Sums of Squares

- Goodness of fit: How much of the variation in the dependent variable is explained by the estimated regression equation?
- For this, we can use the Total Sum of Squares
- This is the squared variation of Y around its mean and is written as:

$$TSS = \Sigma_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{4}$$

Decomposition of TSS

- The TSS can be decomposed into two parts

Decomposition of TSS

- The TSS can be decomposed into two parts
- First: Variation that can be explained by the regression

Decomposition of TSS

- The TSS can be decomposed into two parts
- First: Variation that can be explained by the regression
- Second: Variation that cannot be explained by the regression

Mathematically . . .

$$\Sigma_{i=1}(Y_i - \bar{Y})^2 = \Sigma_{i=1}(\hat{Y}_i - \bar{Y})^2 + \Sigma_{i=1}e_i^2$$

TOTAL Sum of Squares (TSS)	=	EXPLAINED Sum of Squares (ESS)	+	RESIDUAL Sum of Squares (RSS)

Graphically . . .

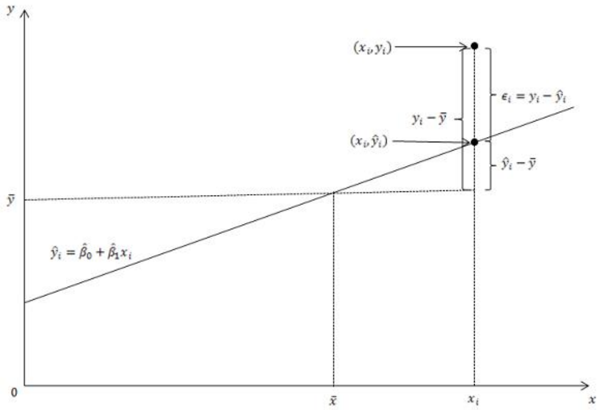

Figure: Decomposition of the Variance in Y, source: Studenmund, 2014, p. 49

Describing the Overall Fit

- We want the regression function to explain as much variation as possible, of course

Describing the Overall Fit

- We want the regression function to explain as much variation as possible, of course
- We can use this to compare different regression models, for example

Describing the Overall Fit

- We want the regression function to explain as much variation as possible, of course
- We can use this to compare different regression models, for example
- We need to apply this criterion with caution, however

R^2

Simplest, commonly used measure is R^2 (AKA coefficient of determination), and is the ratio of the explained sum of squares over the total sum of squares:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS} = 1 - \frac{\Sigma e_i^2}{\Sigma(Y_i - \bar{Y})^2} \tag{5}$$

Interpretation of R^2

- $0 \leq R^2 \leq 1$

Interpretation of R^2

- $0 \leq R^2 \leq 1$
- The closer the regression function fits the data, the closer to 1 R^2 is going to be

Interpretation of R^2

- $0 \leq R^2 \leq 1$
- The closer the regression function fits the data, the closer to 1 R^2 is going to be
- A value close to 0 would indicate, that the function fails to describe the data better than the sample mean $\bar{Y}$

Adjusted R^2

- Problem: Adding another variable to the regression equation, can *never* decrease R^2

Adjusted R^2

- Problem: Adding another variable to the regression equation, can *never* decrease R^2
- Hence, the equation with more variables will always have a better (or at least equal) fit

Adjusted R^2

- Problem: Adding another variable to the regression equation, can *never* decrease R^2
- Hence, the equation with more variables will always have a better (or at least equal) fit
- This is due to the added variable usually reducing RSS (it never increases RSS)

Adjusted R^2 (contd.)

- Imagine we include something non-sensical, such as the colour of your bedsheet in the Facebook equation

Adjusted R^2 (contd.)

- Imagine we include something non-sensical, such as the colour of your bedsheet in the Facebook equation
- Makes no sense theoretically, and it requires the estimation of another coefficient

Adjusted R^2 (contd.)

- Imagine we include something non-sensical, such as the colour of your bedsheet in the Facebook equation
- Makes no sense theoretically, and it requires the estimation of another coefficient
- This lessens the degrees of freedom in the estimation

Degrees of Freedom

- Here: The excess number of observations (N) over the number of coefficient (including the intercept) estimated $(K + 1)$

Degrees of Freedom

- Here: The excess number of observations (N) over the number of coefficient (including the intercept) estimated ($K + 1$)
- Lower degrees of freedom mean less reliable estimates

Degrees of Freedom

- Here: The excess number of observations (N) over the number of coefficient (including the intercept) estimated ($K + 1$)
- Lower degrees of freedom mean less reliable estimates
- This leads to an R^2 that is adjusted for degrees of freedom

So, what's the Equation?

$$\bar{R}^2 = 1 - \frac{\Sigma e_i^2/(N-K-1)}{\Sigma(Y_i - \bar{Y})^2/(N-1)} \tag{6}$$

R^2 Conclusion

- Use $\bar{R}^2$ instead of R^2

R^2 Conclusion

- Use $\bar{R}^2$ instead of R^2
- Useful to compare the fit of different models

R^2 Conclusion

- Use $\bar{R}^2$ instead of R^2
- Useful to compare the fit of different models
- BUT: $\bar{R}^2$ is only *one* measure to compare models

Significance Testing

Back to the t-test

- Regression uses the t-test for the test of significance

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{7}$$

Back to the t-test

- Regression uses the t-test for the test of significance
- Our regression equation is:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i \tag{7}$$

t-test for Regression

- H_0: $\beta_k = 0$ (k=1,2, . . . , K)

t-test for Regression

- H_0: $\beta_k = 0$ (k=1,2, . . . , K)
- We therefore write:

t-test for Regression

- H_0: $\beta_k = 0$ (k=1,2, . . . , K)
- We therefore write:

$$t_k = \frac{(\hat{\beta}_k - \beta_{H_0})}{SE(\hat{\beta}_k)} \tag{8}$$

where $SE(\hat{\beta}_k)$ is the estimated standard error of $\hat{\beta}_k$

t-test for Regression

- H_0: $\beta_k = 0$ (k=1,2, ..., K)
- We therefore write:

$$t_k = \frac{(\hat{\beta}_k - \beta_{H_0})}{SE(\hat{\beta}_k)} = \frac{(\hat{\beta}_k - 0)}{SE(\hat{\beta}_k)} \tag{9}$$

where $SE(\hat{\beta}_k)$ is the estimated standard error of $\hat{\beta}_k$

t-test for Regression

- H_0: $\beta_k = 0$ (k=1,2, . . . , K)
- We therefore write:

$$t_k = \frac{(\hat{\beta}_k - \beta_{H_0})}{SE(\hat{\beta}_k)} = \frac{(\hat{\beta}_k - 0)}{SE(\hat{\beta}_k)} = \frac{\hat{\beta}_k}{SE(\hat{\beta}_k)} \tag{10}$$

where $SE(\hat{\beta}_k)$ is the estimated standard error of $\hat{\beta}_k$

Interpretation

- The larger the t-value, the greater the evidence against H_0

Interpretation

- The larger the t-value, the greater the evidence against H_0
- This would be a two-sided test

Interpretation

- The larger the t-value, the greater the evidence against H_0
- This would be a two-sided test
- Stata reports the SE, as well as the p-value for you