Tutorial 4 HT

Research Methods for Political Science - PO3110

Andrea Salvi

19 February 2019

Trinity College Dublin,
https://andrsalvi.github.io/research-methods/

Table of contents

Multiple Regression

Basics of Regression (once again)

- $y = \beta_0 + \beta_1 x + \epsilon$

Basics of Regression (once again)

- $y = \beta_0 + \beta_1 x + \epsilon$
- We want to estimate the line of best fit using OLS (ordinary least squares).

Basics of Regression (once again)

- $y = \beta_0 + \beta_1 x + \epsilon$
- We want to estimate the line of best fit using OLS (ordinary least squares).
- That is: we want to minimise SS_R (sum of square residuals)

Basics of Regression (once again)

- $y = \beta_0 + \beta_1 x + \epsilon$
- We want to estimate the line of best fit using OLS (ordinary least squares).
- That is: we want to minimise SS_R (sum of square residuals)
- $SS_R = \sum (y_i - \hat{y}_i)^2$ (or RSS)

Basics of Regression (once again)

- $y = \beta_0 + \beta_1 x + \epsilon$
- We want to estimate the line of best fit using OLS (ordinary least squares).
- That is: we want to minimise SS_R (sum of square residuals)
- $SS_R = \sum(y_i - \hat{y}_i)^2$ (or RSS)
- Given that we know how to calculate $\hat{y}_i$, we can rewrite it as: $SS_R = \sum(y_i - \beta_0 - \beta_1 \times x_i)^2$

Basics of Regression (once again)

- $y = \beta_0 + \beta_1 x + \epsilon$
- We want to estimate the line of best fit using OLS (ordinary least squares).
- That is: we want to minimise SS_R (sum of square residuals)
- $SS_R = \sum(y_i - \hat{y}_i)^2$ (or RSS)
- Given that we know how to calculate $\hat{y}_i$, we can rewrite it as: $SS_R = \sum(y_i - \beta_0 - \beta_1 \times x_i)^2$
- Accordingly: $\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x}$

Basics of Regression (once again)

- $y = \beta_0 + \beta_1 x + \epsilon$
- We want to estimate the line of best fit using OLS (ordinary least squares).
- That is: we want to minimise SS_R (sum of square residuals)
- $SS_R = \sum(y_i - \hat{y}_i)^2$ (or RSS)
- Given that we know how to calculate $\hat{y}_i$, we can rewrite it as: $SS_R = \sum(y_i - \beta_0 - \beta_1 \times x_i)^2$
- Accordingly: $\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x}$
- $\hat{\beta}_0 = \bar{y} - \beta_1 \times \bar{x}$

Basics of Regression (once again)

- $y = \beta_0 + \beta_1 x + \epsilon$
- We want to estimate the line of best fit using OLS (ordinary least squares).
- That is: we want to minimise SS_R (sum of square residuals)
- $SS_R = \sum(y_i - \hat{y}_i)^2$ (or RSS)
- Given that we know how to calculate $\hat{y}_i$, we can rewrite it as: $SS_R = \sum(y_i - \beta_0 - \beta_1 \times x_i)^2$
- Accordingly: $\hat{\beta}_1 = \frac{\sigma_{xy}}{\sigma_x}$
- $\hat{\beta}_0 = \bar{y} - \beta_1 \times \bar{x}$
- This formulas apply when you have ONE independent variable.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.940[a]	.883	.863	64.13553

a. Predictors: (Constant), seats

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	185831.801	1	185831.801	45.178	.001[b]
	Residual	24680.199	6	4113.367		
	Total	210512.000	7			

a. Dependent Variable: proposals

b. Predictors: (Constant), seats

$$R^2 = 1 - \frac{SS_R}{SS_T} = 1 - \frac{24680.199}{210512} = 0.883$$

R^2

But also:

- $SS_T = SS_R + SS_M$

R^2

But also:

- $SS_T = SS_R + SS_M$
- All info we can get from our ANOVA table!

R^2

But also:

- $SS_T = SS_R + SS_M$
- All info we can get from our ANOVA table!
- $R^2 = \frac{SS_M}{SS_T}$

R^2

But also:

- $SS_T = SS_R + SS_M$
- All info we can get from our ANOVA table!
- $R^2 = \frac{SS_M}{SS_T}$
- $\frac{185831.8}{210512} = 0.883$

Some more stuff we care about

F-test

- Helps us to assess the overall fit of the model

Some more stuff we care about

F-test

- Helps us to assess the overall fit of the model
- Tests H_0 that all slopes in the model = 0
- $F = \frac{ModelMeanSquares}{ResidualMeanSquares} = \frac{MS_M}{MS_R}$

Some more stuff we care about

F-test

- Helps us to assess the overall fit of the model
- Tests H_0 that all slopes in the model = 0
- $F = \frac{ModelMeanSquares}{ResidualMeanSquares} = \frac{MS_M}{MS_R}$
- SPSS provide us with the exact p value. If significant we reject H_0.

Some more stuff we care about

F-test

- Helps us to assess the overall fit of the model
- Tests H_0 that all slopes in the model = 0
- $F = \frac{ModelMeanSquares}{ResidualMeanSquares} = \frac{MS_M}{MS_R}$
- SPSS provide us with the exact p value. If significant we reject H_0.
- If we have a model with only one independent variable, the F test and the t-test give the same result, because both test the null hypothesis that the one slope in the model is equal to zero (see slides from Stat HT3 lecture to review t-tests in regression analysis).

Multiple Regression

The equation would look like this:

Multiple Regression

The equation would look like this:

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i$

Multiple Regression

The equation would look like this:

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i$
- This allows you to "control" for things!

Multiple Regression

The equation would look like this:

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i$
- This allows you to ”control” for things!
- Remember the example from the lecture with Happiness as a DV and ”Work Hours” as an IV. When introducing ”Education” in the model, the effect of ”Work Hours” was no longer significant.

Multiple Regression

The equation would look like this:

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i$
- This allows you to ”control” for things!
- Remember the example from the lecture with Happiness as a DV and ”Work Hours” as an IV. When introducing ”Education” in the model, the effect of ”Work Hours” was no longer significant.
- Slightly different interpretation. We look at the change in Y when X changes by 1 unit CETERIS PARIBUS.

Multiple Regression

The equation would look like this:

- $y_i = \beta_0 + \beta_1 x_i + \beta_2 x_2 + ... + \beta_k x_k + \epsilon_i$
- This allows you to ”control” for things!
- Remember the example from the lecture with Happiness as a DV and ”Work Hours” as an IV. When introducing ”Education” in the model, the effect of ”Work Hours” was no longer significant.
- Slightly different interpretation. We look at the change in Y when X changes by 1 unit **CETERIS PARIBUS**.
- Let’s look at an example using ”Norris.sav” (`http://tinyurl.com/norris-ht4`).

Multiple Regression: an example

What predicts wealth (measured as GDP per capita)?

- Dependent variable: GDP per capita (US$) 2002 (UNDP 2004)
- Independent variables:
  - FM_Lit2002: Adult illiteracy rate (% ages 15 and above) 2002 (UNDP 2004)
  - F_Work2002: Female economic activity rate (% ages 15 and above) 2002 (UNDP 2004)
  - SDI: Social Diversity Index, primary data source 2001 (Okediji 2005)

Multiple Regression: an example

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-1052.329	3854.899		-.273	.786
	Adult literacy rate (female rate as % of male rate) 2002 (UNDP 2004)	72.153	28.127	.276	2.565	.012
	Female economic activity rate (% ages 15 and above) 2002 (UNDP 2004)	-89.083	34.071	-.302	-2.615	.011
	Social Diversity Index, primary data source 2001 (Okediji 2005)	3266.519	2976.317	.126	1.098	.276

a. Dependent Variable: GDP per capita (US$) 2002 (UNDP 2004)

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	-1615.23	6343.40	2927.33	2056.694	84
Residual	-5941.399	23353.250	.000	4386.110	84
Std. Predicted Value	-2.209	1.661	.000	1.000	84
Std. Residual	-1.330	5.227	.000	.982	84

a. Dependent Variable: GDP per capita (US$) 2002 (UNDP 2004)

Diagnostics

1. Influential data points/outliers
2. Independence/autocorrelation (errors associated with one observation not correlated with errors in any other observation)
3. Linearity (relationship should be linear)
4. Homoscedasticity (constant error variance)
5. Normality (errors should be normally distributed)
6. Model specification
7. Multicollinearity (predictors are highly correlated)
8. Leverage (extent to which predictor differs from mean of predictor)

Let's have a look at some of them!