Tutorial 2 HT

Research Methods for Political Science - PO3110

Andrea Salvi

5 February 2019

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

Table of contents

Notes on Homework Exercises

Notes on Homework Exercises

- Need to be submitted on Blackboard;
- The best course of action is that of generating a syntax file. It has several advantages:
  1. Reproducible code (you can re-run it whenever you want and share it!)
  2. Clear and legible (avoid copy-pasting from the output, it's quite "noisy!")
- Export graphs
- Add comments!

Example: Calculate the descriptive statistics of a a given distribution

25 : example_var

	example_var
1	10
2	20
3	30
4	50
5	100
6	100
7	10
8	50
9	30
10	40
11	55
12	34
13	21
14	124
15	18
16	30
17	293
18	2
19	3
20	4
21	5
22	65
23	65
24	20

The wrong way

Example: Calculate the descriptive statistics of a a given distribution

”Here it is.”

```
DATASET ACTIVATE DataSet0.
DESCRIPTIVES VARIABLES=example_var
  /SAVE
  /STATISTICS=MEAN STDDEV VARIANCE MIN MAX SEMEAN.
```

Descriptives

[DataSet0]

Descriptive Statistics

	N	Minimum	Maximum	Mean		Std. Deviation	Variance
	Statistic	Statistic	Statistic	Statistic	Std. Error	Statistic	Statistic
example_var	24	2	293	49.13	12.558	61.523	3785.071
Valid N (listwise)	24						

The nice way

- Firstly, make sure you have working a syntax file.

The nice way

- Firstly, make sure you have working a syntax file.
- Secondly, run it!

The nice way

- Firstly, make sure you have working a syntax file.
- Secondly, run it!
- Export the output

The nice way

- Firstly, make sure you have working a syntax file.
- Secondly, run it!
- Export the output
- Write a comment if required.

Other minor things:

- Use mathematical notation: x bar is $\bar{x}$, mu is μ, var(x) is σ^2 etc.

The nice way

- Firstly, make sure you have working a syntax file.
- Secondly, run it!
- Export the output
- Write a comment if required.

Other minor things:

- Use mathematical notation: x bar is $\bar{x}$, mu is μ, var(x) is σ^2 etc.
- Make sure the document you are working on is clear and legible.

The nice way

- Firstly, make sure you have working a syntax file.
- Secondly, run it!
- Export the output
- Write a comment if required.

Other minor things:

- Use mathematical notation: x bar is $\bar{x}$, mu is μ, var(x) is σ^2 etc.
- Make sure the document you are working on is clear and legible.
- The homework exercises are based on stuff we did, don't panic if you don't remember how to do something and check your notes/previous slides.

Correlation and Regression

Correlation

Definition

A standardized statistic that provides a measure of the strength and direction of a relationship between two variables. It can take any value from -1 to +1. A result of +1 shows that there is a perfect positive relationship between the two variables: as one variable increases, the other increases. A correlation of -1 indicates that if one variable increases, the other one decreases. A result of 0 shows that there is no discernible pattern to the relation between the two variables.

Correlation

Definition

A standardized statistic that provides a measure of the strength and direction of a relationship between two variables. It can take any value from -1 to +1. A result of +1 shows that there is a perfect positive relationship between the two variables: as one variable increases, the other increases. A correlation of -1 indicates that if one variable increases, the other one decreases. A result of 0 shows that there is no discernible pattern to the relation between the two variables.

That is:

- $\text{correlation} = \frac{\text{covariance}}{\text{sd x} \times \text{sd y}}$

Correlation

Definition

A standardized statistic that provides a measure of the strength and direction of a relationship between two variables. It can take any value from -1 to +1. A result of +1 shows that there is a perfect positive relationship between the two variables: as one variable increases, the other increases. A correlation of -1 indicates that if one variable increases, the other one decreases. A result of 0 shows that there is no discernible pattern to the relation between the two variables.

That is:

- $\text{correlation} = \frac{\text{covariance}}{\text{sd x} \times \text{sd y}}$
- $r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y}$

Correlation

Definition

A standardized statistic that provides a measure of the strength and direction of a relationship between two variables. It can take any value from -1 to +1. A result of +1 shows that there is a perfect positive relationship between the two variables: as one variable increases, the other increases. A correlation of -1 indicates that if one variable increases, the other one decreases. A result of 0 shows that there is no discernible pattern to the relation between the two variables.

That is:

- $\text{correlation} = \frac{\text{covariance}}{\text{sd x} \times \text{sd y}}$
- $r = \frac{\sigma_{xy}}{\sigma_x \times \sigma_y}$
- $-1 \leq r \leq 1$

Similarities and Differences between Correlation and Regression

BASIS FOR COMPARISON	CORRELATION	REGRESSION
Meaning	Correlation is a statistical measure which determines co-relationship or association of two variables.	Regression describes how an independent variable is numerically related to the dependent variable.
Usage	To represent linear relationship between two variables.	To fit a best line and estimate one variable on the basis of another variable.
Dependent and Independent variables	No difference	Both variables are different.
Indicates	Correlation coefficient indicates the extent to which two variables move together.	Regression indicates the impact of a unit change in the known variable (x) on the estimated variable (y).
Objective	To find a numerical value expressing the relationship between variables.	To estimate values of random variable on the basis of the values of fixed variable.

Regression

Regression

- $Y = \beta_0 + \beta_1 X + \epsilon$

Regression

- $Y = \beta_0 + \beta_1 X + \epsilon$
- Y is the outcome/response/dependent variable

Regression

- $Y = \beta_0 + \beta_1 X + \epsilon$
- Y is the outcome/response/dependent variable
- X is the predictor/independent variable

Regression

- $Y = \beta_0 + \beta_1 X + \epsilon$
- Y is the outcome/response/dependent variable
- X is the predictor/independent variable
- β_0 is the Intercept, it represents the average value of Y when X is zero

Regression

- $Y = \beta_0 + \beta_1 X + \epsilon$
- Y is the outcome/response/dependent variable
- X is the predictor/independent variable
- β_0 is the Intercept, it represents the average value of Y when X is zero
- β_0 is the Slope and measures the average increase in Y when X increases by one unit.

Regression

- $Y = \beta_0 + \beta_1 X + \epsilon$
- Y is the outcome/response/dependent variable
- X is the predictor/independent variable
- β_0 is the Intercept, it represents the average value of Y when X is zero
- β_0 is the Slope and measures the average increase in Y when X increases by one unit.

See extensively K. Imai (2017). Quantitative Social Science. An Introduction. Princeton: Princeton University Press

Regression

- $Y = \beta_0 + \beta_1 X + \epsilon$
- Y is the outcome/response/dependent variable
- X is the predictor/independent variable
- β_0 is the Intercept, it represents the average value of Y when X is zero
- β_0 is the Slope and measures the average increase in Y when X increases by one unit.

See extensively K. Imai (2017). Quantitative Social Science. An Introduction. Princeton: Princeton University Press

Regression

- Once we obtain the estimates of β_0 and β_1 we have the so-called regression line

Regression

- Once we obtain the estimates of β_0 and β_1 we have the so-called regression line
- We can use the regression line to predict the value of the outcome variable given that of a predictor

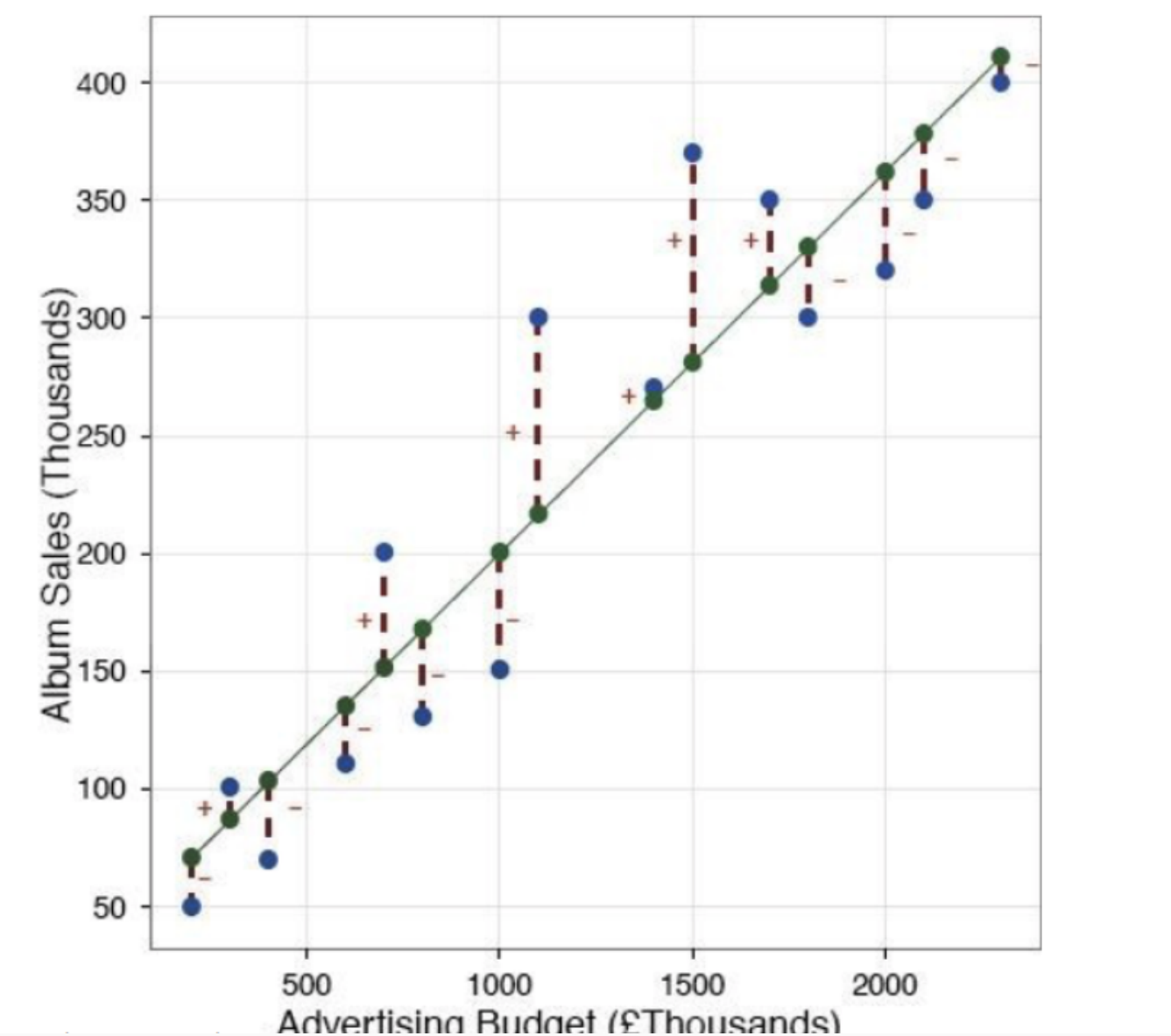
Regression

- Once we obtain the estimates of β_0 and β_1 we have the so-called regression line
- We can use the regression line to predict the value of the outcome variable given that of a predictor
- Regression line is the f best fit because it minimizes the magnitude of prediction error (OLS).

Regression

- Once we obtain the estimates of β_0 and β_1 we have the so-called regression line
- We can use the regression line to predict the value of the outcome variable given that of a predictor
- Regression line is the f best fit because it minimizes the magnitude of prediction error (OLS).

Regression


Album Sales (Thousands)
400
350
300
250
200
150
100
50
500
1000
1500
2000
Advertising Budget (£Thousands)

Regression

- General idea: choose $\hat{\beta}_0$ and $\hat{\beta}_1$ such that together they minimize the sum of squared residuals (SSR).
- $SSR = \sum(Y_i + \hat{\beta}_0 + \hat{\beta}_1 \times X_i)^2$
- R^2 the root-mean squared error, calculated as $\sqrt{\frac{1}{n} \times SSR}$
- R square indicates how much (ratio) of the variance in the dependent variable can be explained by our regression model

Regression: Definitions from Field

- ""R^2 tells us how much variance is explained by the model compared to how much variance there is to explain in the first place. It is the proportion of variance in the outcome variable that is shared by the predictor variable."
- "F tells us how much variability the model can explain relative to how much it can't explain (i.e., it's the ratio of how good the model is compared to how bad it is)."

Anova box

- includes tests whether the model is significantly better at predicting the outcome than using the mean as a ‘best guess’

Anova box

- includes tests whether the model is significantly better at predicting the outcome than using the mean as a ‘best guess’
- F-ratio represents the ratio of the improvement in prediction that results from fitting the model, relative to the inaccuracy that still exists in the model

Anova box

- includes tests whether the model is significantly better at predicting the outcome than using the mean as a 'best guess'
- F-ratio represents the ratio of the improvement in prediction that results from fitting the model, relative to the inaccuracy that still exists in the model
- F-ratio is calculated by dividing the average improvement in prediction by the model (MS_M) by the average difference between the model and the observed data (MS_R).

Anova box

- includes tests whether the model is significantly better at predicting the outcome than using the mean as a 'best guess'
- F-ratio represents the ratio of the improvement in prediction that results from fitting the model, relative to the inaccuracy that still exists in the model
- F-ratio is calculated by dividing the average improvement in prediction by the model (MS_M) by the average difference between the model and the observed data (MS_R).
- If the improvement due to fitting the regression model is much greater than the inaccuracy within the model, then the value of F will be greater than 1, and SPSS calculates the exact probability of obtaining the value of F by chance

Anova box

- includes tests whether the model is significantly better at predicting the outcome than using the mean as a 'best guess'
- F-ratio represents the ratio of the improvement in prediction that results from fitting the model, relative to the inaccuracy that still exists in the model
- F-ratio is calculated by dividing the average improvement in prediction by the model (MS_M) by the average difference between the model and the observed data (MS_R).
- If the improvement due to fitting the regression model is much greater than the inaccuracy within the model, then the value of F will be greater than 1, and SPSS calculates the exact probability of obtaining the value of F by chance

Regression output from SPSS

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	CO2 emissions, per capita (metric tons), 2003 (world bank 2007)[b]	.	Enter

a. Dependent Variable: GDP 2002 (UNDP 2004)

b. All requested variables entered.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.204[a]	.042	.036	859.53477

a. Predictors: (Constant), CO2 emissions, per capita (metric tons), 2003 (world bank 2007)

Regression output from SPSS

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	5530758.59	1	5530758.59	7.486	.007[b]
	Residual	127073605	172	738800.026		
	Total	132604363	173			

a. Dependent Variable: GDP 2002 (UNDP 2004)

b. Predictors: (Constant), CO2 emissions, per capita (metric tons), 2003 (world bank 2007)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	65.115	77.873		.836	.404
	CO2 emissions, per capita (metric tons), 2003 (world bank 2007)	24.319	8.888	.204	2.736	.007

a. Dependent Variable: GDP 2002 (UNDP 2004)