Tutorial 04, Hilary Term

Research Methods for Political Science (PO3600)

Stefan Müller

13 February 2018

Trinity College Dublin
http://muellerstefan.net/research-methods

Session Outline

1. Discussion Homework #2, HT

Definitions

How to get good marks in defining concepts:

1. Explain what the concept is
2. Highlight why it is important
3. Might provide an example of an instance where it can arise
4. Mention how concept can be measured

Exercise I

Define and explain: **Inter-Coder Reliability**

Exercise I

Define and explain: **Inter-Coder Reliability**

Intercoder reliability refers to the extent to which at least two independent coders agree on the coding of contents applying the same coding scheme. Intercoder reliability is a critical component in the content analysis of open-ended survey responses or textual data. With low inter-coder reliability the interpretation of the content cannot be considered objective and valid. Yet, even high intercoder reliability is not the only criteria necessary to argue that a coding is valid.

Exercise II

- What is the difference between correlation and regression analysis?
- What are advantages of regression?

Exercise II

a. We want to know whether two variables, `GDP02` and `Co2_2003` (CO_2 emissions per capita in metric tonnes) are correlated. Compute and interpret the appropriate correlation coefficient. [10 points]

Exercise II

b. Run a simple linear regression, where `Co2_2003` is the dependent variable and `GDP02` is the independent variable. [5 points]

c. Identify and interpret the intercept in the model. Is it meaningful in this case? [10 points]

d. Interpret the coefficient in the model. What does it mean in substantive terms? (hint: the SPSS dataset indicates the units that each of the variables is measured in) [10 points]

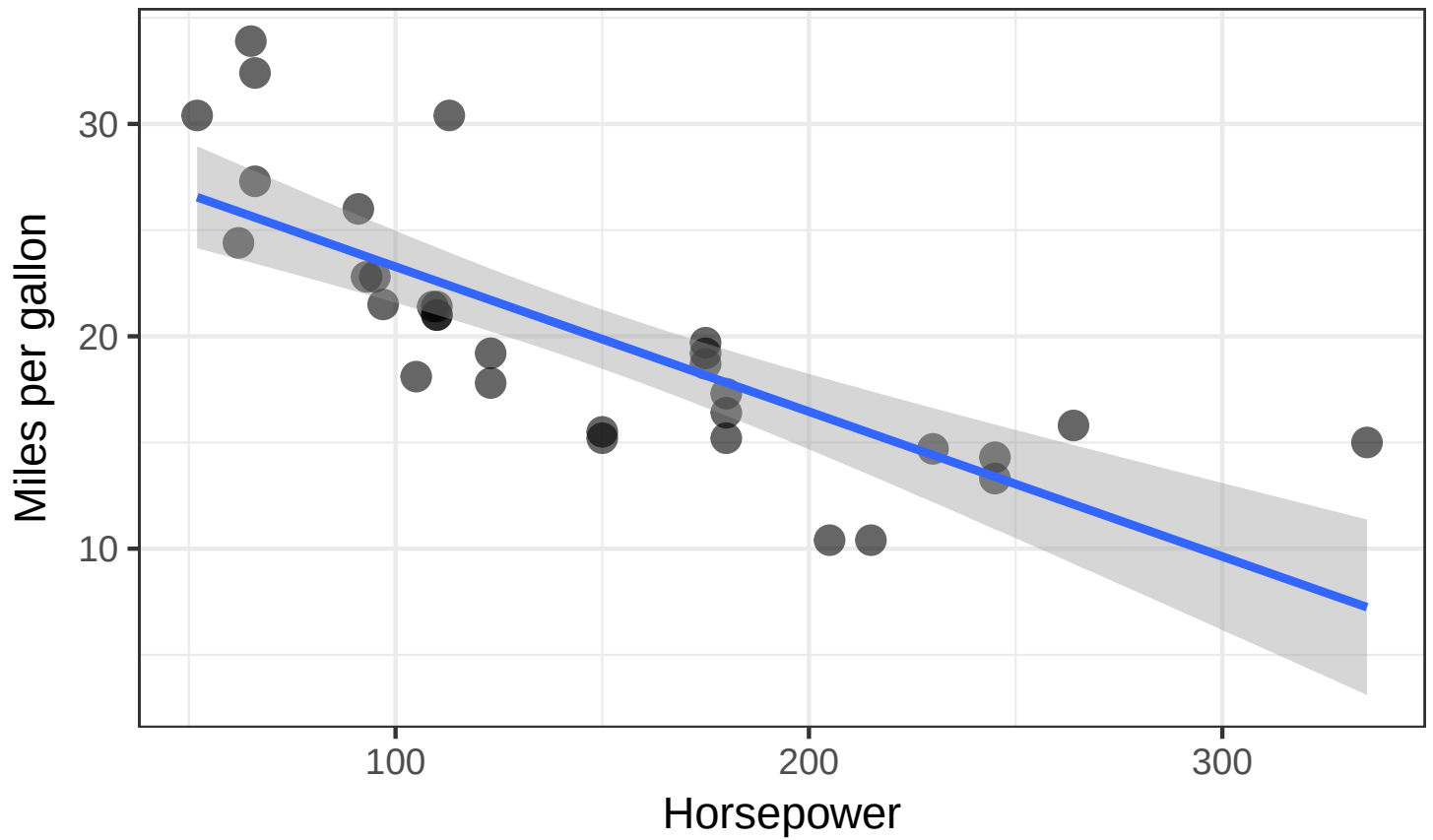
Exercise II

e. Interpret the R Square in the model output. [10 points]

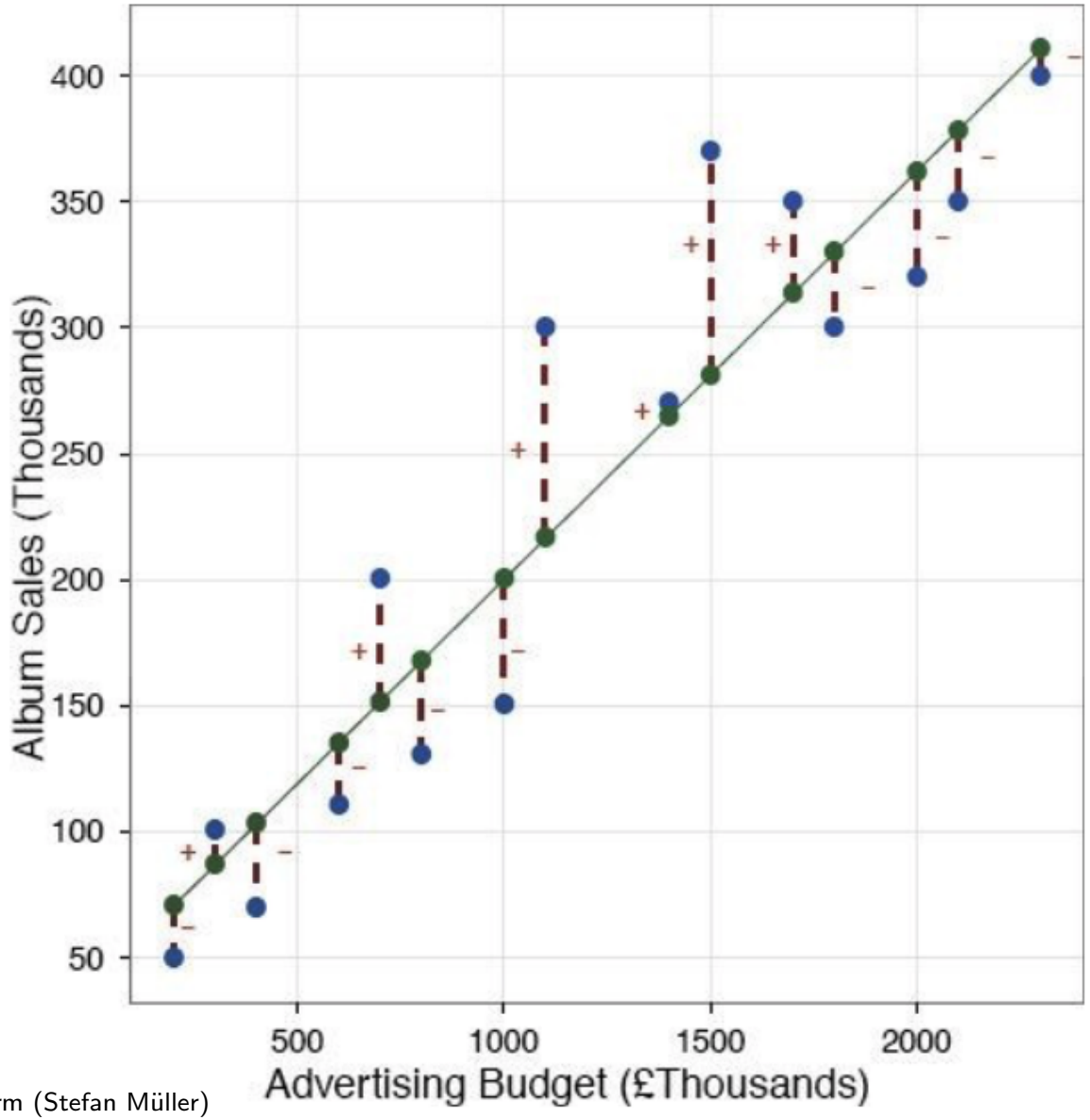
f. Using the ANOVA section from the regression output, explain how this R square is computed. (Hint: what do the numerator and the denominator of the formula measure?) [10 points]

g. Explain the relationship between the R square and the correlation coefficient computed in part a) of this question. [10 points]

Example: Regression Line



Example: Residuals



Regression

$$Y = \alpha + \beta \times X + \epsilon$$

- Y is the outcome/response/dependent variable
- X is the predictor/independent variable
- Intercept α represents the average value of Y when X is zero
- Slope β 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

$$Y = \alpha + \beta \times X + \epsilon$$

- Once we obtain the estimates of α and β 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 "line of best fit" because it minimises the magnitude of prediction error

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

Regression

$$Y = \alpha + \beta \times X + \epsilon$$

- General idea: choose $\hat{\alpha}$ and $\hat{\beta}$ such that together they minimise the sum of squared residuals (SSR). $SSR = \sum(Y_i - \hat{\alpha} - \hat{\beta} \times X_i)^2$
- R square is 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

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

Definitions from Field (2013)

- **R2** "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)."

The 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

Useful Resources

- Chapter 8 of A. Field (2013). *Discovering Statistics Using IBM SPSS Statistics*. 4th ed. Los Angeles: SAGE
- Links on tutorial website:
  http://muellerstefan.net/research-methods