

Quant II - Problem Set II

The Linear Regression Model

Kenneth Benoit

Assigned: Wednesday, January 14, 2009

Due: Wednesday, January 21, 2009

1. You have a dataset with five observations as follows:

x	y
3	8.9
5	15.2
6	15.7
10	23.7
12	28

You want to regress y on x .

- (a) Calculate the OLS estimators of slope and intercept for this example. You may do this either manually or by showing the individual calculation steps in R.
 - (b) Verify your results in R with the regression command.
 - (c) Create a scatterplot for x and y . (Make sure y is on the vertical axis!)
 - (d) Extra credit: add the regression line to the plot.
2. Use the familiar dataset by Benoit and Marsh (2008) and regress first preference votes on total spending and incumbency.
 - (a) What percentage of variation in the response variable is explained by the two predictors?
 - (b) Which observation has the largest positive, which the largest negative residual? Give the case numbers. What does it mean substantively in this regression to have a large residual?
 - (c) Compute the mean and median of the residuals.
 - (d) Compute the correlation of the residuals with the fitted values.
 - (e) Compute the correlation of the residuals with total spending.

- (f) For total spending held constant, what would be the difference in predicted first preference votes for an incumbent compared to a challenger?
3. Load the SPSS dataset from the Irish National Election Study (ignore the warnings R issues when doing so). For more information on the data, please refer to the questionnaire.

We are interested in the popularity of Gerry Adams, the Sinn Fein leader. The variable C1A9 is a so-called thermometer score, where respondents indicated how they feel about Gerry Adams. We would like to examine, how gender, age, and the attitude towards Irish unification affect Gerry Adams' ratings.

The raw variables are the Adams thermometer score (C1A9), gender (E4), year of birth (E5YR) and the attitude to Irish Unification (C6_1).

- (a) First of all, you have to prepare the variables for analysis. Some may have to be changed from factor to numeric or vice versa, many will contain missing values that are stored as numbers, some have to be recoded. Please provide the full R code, so that others can replicate what you have done.

Hints: You may use `class()` and `attributes()` to learn more about the variables, `table()` them etc.

A command useful for recoding numeric missing values is e.g.

```
data$variable[data$variable == missing_code_value] <- NA
```

Note that C6_1 contains two different numeric codes for missing values.

- (b) On basis of year of birth, generate a new variable that contains the age at the time of the survey (2002).
- (c) Regress the thermometer score on gender, the new age variable, and the attitude towards Irish Unification. Briefly discuss the results.
- (d) Calculate SSR, SSE, and SST “manually” in R (don't use the `anova()` function).
- (e) Which r is R-squared the square of? Verify this result with R for the above regression.
- (f) R-squared is sometimes referred to as a Proportional-Reduction-of-Error (PRE) measure. Does this make sense? Why (not)?

- (g) Create a new dependent variable by dividing the old dependent variable by 10 and then subtract 5 from the result. What is the theoretical and empirical range of the new variable compared to the range of the old? Re-run the model from above with the new dependent variable. How does the coefficient of the gender variable change? How do the fit statistics from a regression using the new variable compare to the fit statistics from the previous model?