Regression Exercises

Ananthan Ambikairajah

27 February, 2020

Contents

1 Exercise 1

1.1 Visualising your data/models

Load the dataset **Album Sales 1.dat** and save the output to a variable (i.e. album_data). The dataset consists of two variables including *adverts* – the amount of money spent on advertisements in the week before the release of the album (thousands of dollars) and *sales* – the number of record sales in the week after the release of the album (thousands).

1. Inspect the data using head(), tail(), summary() and str().
2. Construct a baseline linear model with no predictors to examine *sales.* Save the output to a variable. Inspect the output of the model using summary().
3. Construct a simple linear regression of *sales* as a function of *adverts.* Save the output to a variable. Inspect the output of the model using summary().
4. Create a scatterplot of the data (i.e. the relationship between *adverts* and *sales*) with the mean of the response variable as the predictive line.
5. Create a second scatterplot of the data with the predictor (i.e. *adverts* coefficient) as the predictive line.

2 Exercise 2

2.1 Deriving the output – Baseline linear model

Using the theory taught so far, use R to work out the following:

1. The estimate (i.e. the mean).
2. The standard error of the estimate.
3. The t value for the estimate.
4. An approximate p-value for the intercept and for *adverts.* Note – you will need to use the table of critical values (for the t-distribution) for this.

5. The residuals for the estimate.
6. The minimum, 1st quartile, median, 3rd quartile and maximum of the residuals.
7. The standard deviation of the residuals (confusingly named as the residual standard error by R for your regression output).
8. The upper and lower confidence intervals for the estimate.

2.2 Deriving the output – Simple linear regression

Using the theory taught so far, use R to work out the following:

1. The standard deviation of the residuals (hint – extract the residuals from the model)
2. The R2 value – (i.e the coefficient of determination – called multiple R-squared by R).
3. The adjusted R2 value.
4. The F-statistic.
5. The t-value for the intercept and *adverts*.
6. An approximate p-value for the intercept and for *adverts*.
7. The minimum, 1st quartile, median, 3rd quartile and maximum of the residuals.

2.3 Extension activity - Functions

Create your own function, which produces the output for all calculations in Exercise 2.2, except for the p-values.

Hint: You will need to store the output in a list.

3 Exercise 3

3.1 Anscombe's dataset

Load the dataset **Anscombe.csv**. It consists of four datasets, each consisting of 11 observations. For each of the four datasets:

1. Calculate:

i) the mean of x,

ii) the variance of x,

iii) the mean of y,

iv) the variance of y.

2. Fit a linear regression model of y as a function of x for each of the four distributions within the dataframe. Does x have a significant effect on y? Are the models for each data set similar or different to one another?

Hint: You will need to use subsetting techniques to select individual distribution groups within the dataset.

3.2 Plotting data

Visualise the Anscombe dataset (i.e. construct a scatterplot each of the four datasets). Then fit your linear regression output on top of your graph. Are the models for each data set similar or different to one another?

3.3 Extension activity - For loops

Use the power of for loops to complete exercises 3.1 and 3.2. Store the output of 3.1 (Question 1) in a dataframe and 3.1 (Question 2) in a list