Turning your R scripts into reports

David JPO'Sullivan

2/25/2020

What is R Markdown?

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown, see the following websites here and here. R Markdown is a powerful tool for automating your report creation process.

When you click the **Knit** button, a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this.

```
# readin the data
glimpse(credit_slr_df)
```

```
## Observations: 226
## Variables: 2
## $ bill_amt1 <dbl> 119287, 4670, 12547, 277822, 59143, 874, 21854, 41906, 57...
## $ bill_amt2 <dbl> 116995, 4670, 14699, 255167, 58612, -256, 17376, 17969, 2...
```

The above is referred to as a code chunk. These chunks run segments of code from the analysis. They can be printed in the result document or not, but the information they produce is still available to us for analysis.

In the following section we are going to turn the R script that we used to do the hypothesis testing and to fit a linear model to the credit data into an automatically generated report.

Credit modelling

Hypothesis testing on credit data

Difference in population means

We are interested in the differences in the population means of `total_bill_amt` between those that `default` and those that do not. To answer this question, we will help us hypothesis testing with a 5% level of significance ($\alpha = 5$).

The null and alternative hypothesis for the t-test are

- H_0 : There is no difference between the population means ($\mu_1 = \mu_2$).
- H_A : There is a difference between the population means ($\mu_1 \neq \mu_2$).

The resulting p-value from the t-test is 0, which is less than $\alpha = 0.05$ than we reject the null hypothesis. The population means are different between the default and non-default group.

Additionally, we can examine the confidence interval for the mean difference between the two groups. The 95% CI for the mean difference between the groups is $[754, 1.4853 \times 10^4]$. There, we are 95% certain the non-default group has, on average, between 754 and 1.4853×10^4 more than the default group.

Paired sample t-test

Insert markdown text and/or code here to:

- Do a paired sample t-test between `bill_amt1` and `bill_amt2`.
- Comment on the set up of the test and result using the steps outlined in the lecture notes.
- Calculate the confidence interval for the difference and comment

Modelling customer spending behaviour between months

Here we want to investigate if by fitting a linear regression model between `bill_amt1` and `bill_amt2`. The Pearson's correlation coefficient between the two variables is 0.97, which indicate that there is a very strong linear relationship. To confirm this, we visually inspect a scatter plot of the two variables. From the following graph, we note that there is a linear trend between `bill_amt1` and `bill_amt2`.

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

The line of best fit is estimated as:

$$y = -1838.74 + 1.07 \times x$$

Examining the confidence intervals for the slope. We are 95% certain that the true slope (b_1) for `bill_amt2` is in the range (1.03, 1.1).

Insert markdown text and/or code here to:

1. Comment on the confidenc interval.
2. Interpret the slope.
3. State the R^2 and commonent on the goodness of fit

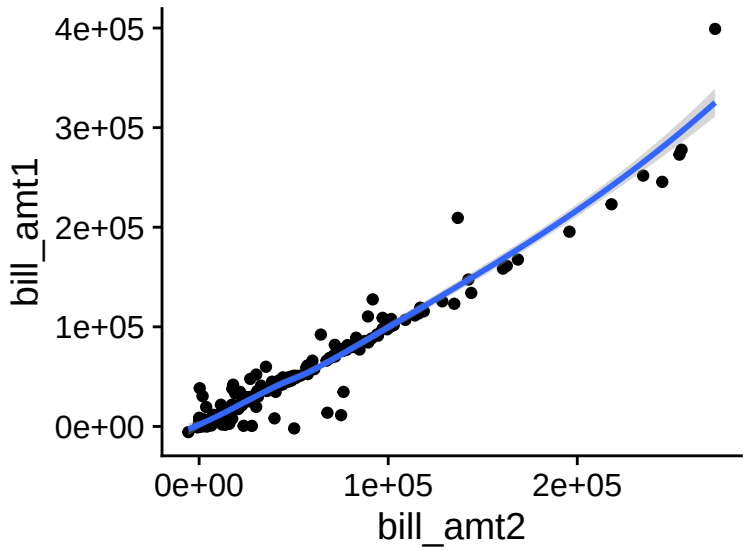

Figure 1: Scatter plot of `bill_amt1` and `bill_amt2`

Model diagnostics

We can examine the quality of the fits and check that the model statistics the underlying assumptions of the model using the following diagnostic plot of the residuals.

There is some evidence of departures from normality in the tails of the Q-Q plot (but the values between -2 to 2 fit reasonably well). We also have evidence of a point of high leverage (sample number 119), might be work to see what is so special about that person.

It seems that it is maybe reasonable to use this model to predict the `bill_amt1` using `bill_amt2`. But we should expect the model to be inaccurate for very large or very small values.

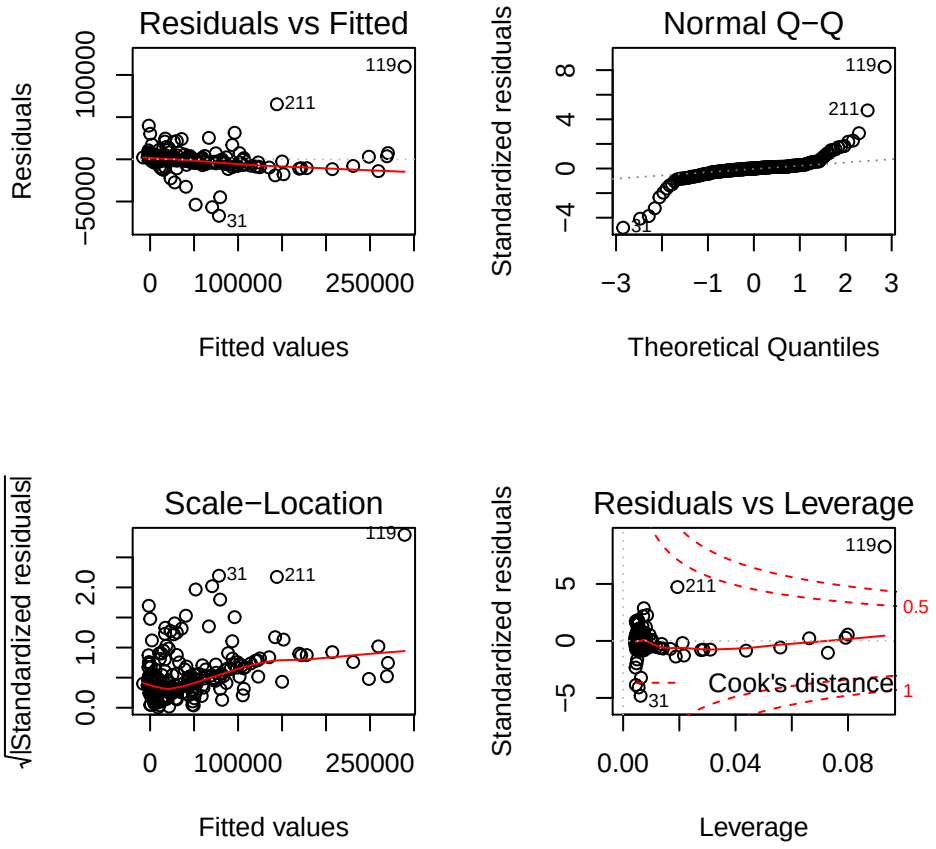

Figure 2: Diagnostic plot for regression.‘

Accuracy of predictions

Insert markdown text and/or code here to:

1. Visually inspect the a plot of the fitted linear regression model, the fitted data and the test data.

2. Comment on why the model may not predict as well on the test data than the training data.