Turning your R scripts into reports

David JPO'Sullivan

25/5/2020

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.

```
# Peek at 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...
```

These chunks run segments of code from the analysis. They can be printed in the resulting document or not, but the information they produce is still available for analysis. In the following section we are going to turn the R script that we used to perform the hypothesis tests and estimation of the linear regression model into an automatically generated report.

Credit modelling

Hypothesis testing on credit data

Difference in population means

We are interested in the difference between population means of `total_bill_amt` for those who *defaulted* and those who do not. To answer this question, we use a 5% level of significance ($\alpha = 0.05$).

The null and alternative hypotheses 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 0.03, which is less than $\alpha = 0.05$. Therefore, we reject the null hypothesis. It appears that the population means for the default and non-default groups differ.

Additionally, we can examine the confidence interval for the difference between the two population means, which is given by $[754, 1.4853 \times 10^4]$. Therefore, we are 95% confident that the non-default group owes, on average, between 754 and 1.4853×10^4 more than the default group.

Paired sample t-test

We are interested in whether or not there is a difference in the mean amount owed in `bill_amt1` and `bill_amt2`.These two amounts come as a pair for each individual in the dataset which is why a paired t-test is required. As before we will use a 5% level of significance ($\alpha = 0.05$).

The null and alternative hypotheses for the paired sample t-test are

- H_0 : The mean difference is zero ($\mu_d = 0$).
- H_A : The mean difference is not zero ($\mu_d \neq 0$).

The resulting p-value from the t-test is 0, which is less than $\alpha = 0.05$. Thus, we reject the null hypothesis, and conclude that there appears to be a difference in the amount owed for the two months (`bill_amt1` and `bill_amt2`).

Additionally, we can examine the confidence interval for the mean difference, and this is given by $[1592, 2234]$. Therefore, we are 95% confident that a person, on average, owes between 1592 and 2234 more in `bill_amt1` than `bill_amt2`.

Modelling customer spending behaviour between months

We have just discovered that there is a difference between `bill_amt1` and `bill_amt2`. We will now investigate how well we can predict `bill_amt1` using `bill_amt2`. Pearson's correlation coefficient for these two variables is 0.97, which indicates that there is a very strong linear relationship between the variables. To confirm this, we visually inspect a scatter plot of the two variables. From the following graph, we can see that there is indeed a linear trend.

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

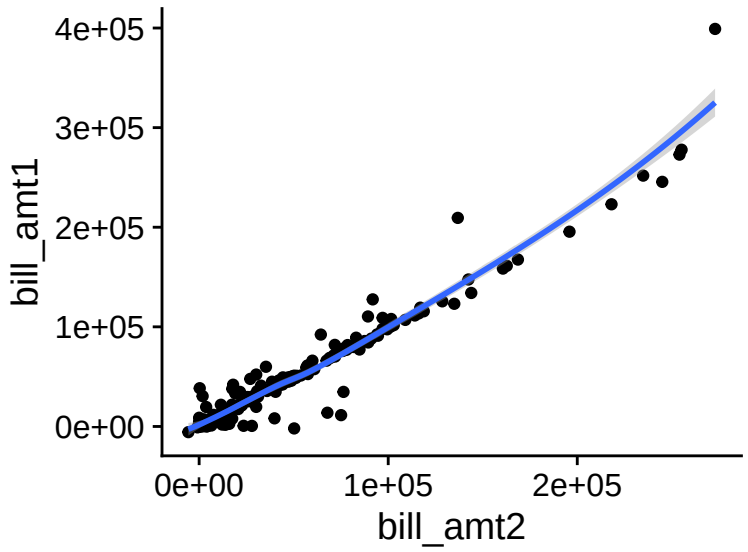

Figure 1: Scatter plot of `bill_amt1` and `bill_amt2`

The line of best fit is estimated as:

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

We are 95% confident that the true slope (b_1) is in the range (1.03, 1.1). The confidence interval does not contain zero, so there is a statistically significant **linear** relationship between this and last month's bill. In particular, examining the value of the slope, for a one-unit increase in `bill_amt2` we expect `bill_amt1` to increase by 1.07 units, on average.

In terms of model fit, we have that the coefficient of determination is $R^2 = 0.94$. In other words, 94% of the variability in `bill_amt1` is explained by `bill_amt2` which is an excellent fit.

Model diagnostics

Here we will assess the model adequacy using the following diagnostic plots of residuals which highlight departures from normality, any particularly large residuals, and cases with high leverage.

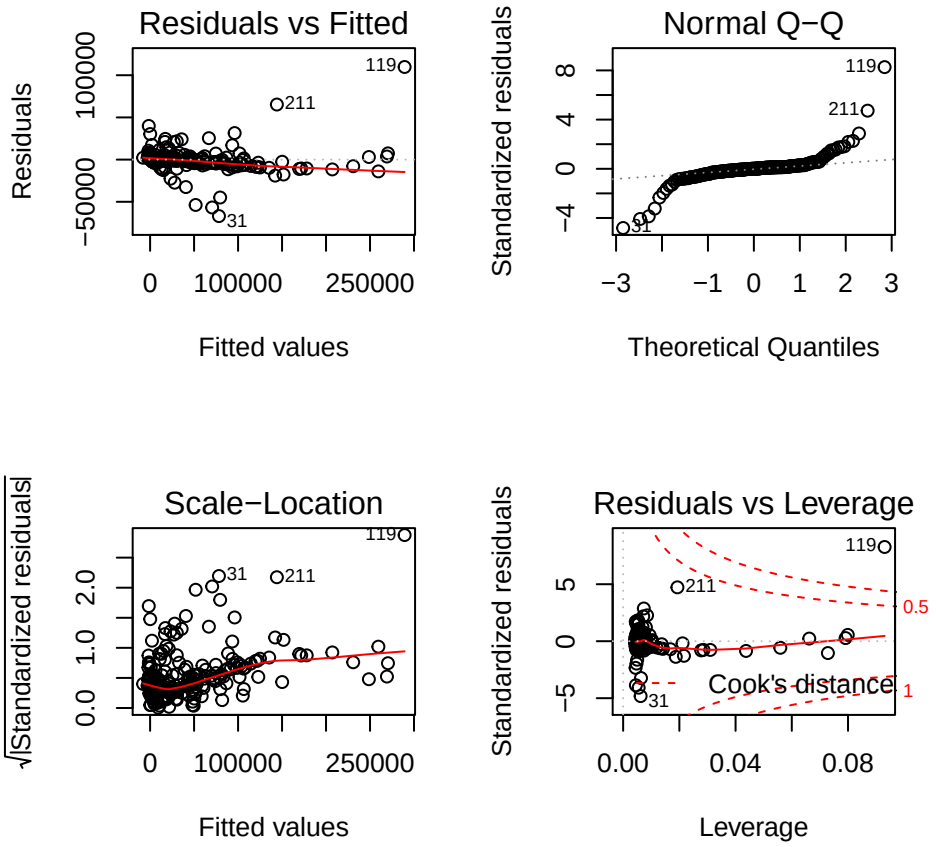

Figure 2: Diagnostic plot for regression.‘

There is some evidence of departures from normality in the tails of the Q-Q plot. We also have evidence of a point of high leverage (sample number 119), and it might be worth investigating what is different about that particular person.

It seems that this is a reasonable model for predicting `bill_amt1` using `bill_amt2`.

Accuracy of predictions

We will visually assess the accuracy of the linear regression model when using it to predict spending behaviour on unseen test data.

It looks like there still is a linear trend in the data but shifted. What could be the cause of this? Model drift: there maybe some unforseen factor at play (like seasonal effects) resulting in different amounts spent at different times of year. For example, you expect people, on average, to spend more during Christmas than at other times of the year and our linear regression model does not account for such temporal features.

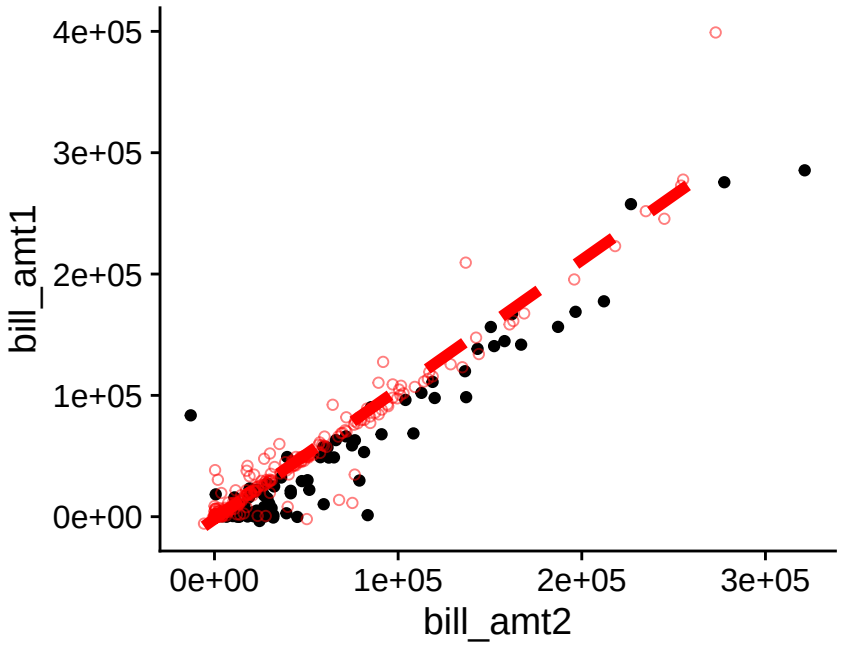

Figure 3: Scatter plot to assess accuracy of predictions.