The Technical Report – British Household Panel Survey (BHPS)

Introduction

This report investigates the dataset Group5.csv which contains data on the BHPS for 13 different waves of respondents from 1991 to 2004. The main research question is to determine how the annual income (*afiyr*) depends on other factors. We aim to do this with a simple linear regression model with adequate interpretability and sufficient predictive power.

This study first provides an explanatory analysis of the data, covariates and the outcome variable of interest. The methodology is to fit a model with all explanatory variables. Next, we are going to simplify the model with variable selection. Then, we will evaluate the goodness of fit and the residual assumptions of this model. Potential transformations will be explored to improve the model. Lastly, we will investigate outliers and possibly refit our model without them.

The BHPS contains a large number of variables but in this report we focus on the following variables.

Variable	Type	Definition	Baseline
aage	Continuous	Age of the individual at the date of interview	-
sex	Binary	Gender of individual	Male
AJBHRS	Continuous	Number of hours normally worked per week	-
ancars	Categorical	The availability of a car or van for private use	0
ahhsize	Continuous	Number of individuals in the household	-
aqfedhi	Categorical	Highest educational qualification achieved	No QF
arace	Categorical	Ethnicity of individual	White
aregion	Categorical	Region of household	Inner London
atenure	Categorical	Housing tenure	Owned outright
avote	Categorical	Political party supported	Conservative

Our results show that all explanatory variables contribute to explaining the annual income except the ethnicity of an individual. The final model we present is given by the below equation:

√afiyr= 39.7505 + 0.7233**aage** - 15.5256**sex_female** + 2.5179**ancars_1** + 6.5851**ancars_2** – 4.4688**ancars_3** - 1.8738**ahhsize** + a + e_i , where a denotes the remaining categorical variables (equation would be too cumbersome with them) and e_i is the error tem.

We choose the first baseline for all categorical variables except aqfedhi which is changed to No QF. This regression indicates that the higher aage, the higher is afiyr. Additionally, a man is expected to earn on average 241.04 more annual income than a woman. Also, the higher is ahhsize, the lower is afiyr.

Main Results

1 Data

Our dataset consists of 11 variables accounting for a total of 2000 entries. The dependent is *afiyr* which is simply the annual individual income. Of the 10 explanatory variables, seven of them were categorical, two were continuous and one was binary as shown in the above table.

2 Explanatory analysis

The defining variable *afiyr* has a mean of 10980, with minimum value of 0 and maximum of 85830. We identified six individuals whose annual income was 0, as they did not appear to be outliers we left them in the dataset for the time being. However, we also found missing values for *aqfedhi* and *atenure* which were removed from the data set. The scatterplots of *afiyr* against the continuous values of *aage* and *AJBHRS* revealed a discernible positive relationship. However, the relationship was less obvious with *ahhsize*. Also, we ran individual regressions for afiyr against all the explanatory variables. This showed that *aqfedhi* and *AJBHRS* provided the highest adjusted R2 of 0.2087 and 0.2065 respectively. The variable *ancars* had the lowest adjusted R2 of 0.00494.

3 Regression Model and Variable Selection

A linear regression model of all explanatory variables on *afiyr* was ran, denoted by '*lm.full'* .Next, we computed the backward elimination. All covariates that were not significant at the 5% level were removed and this yielded the model *'lm.full3'* (treating categorical variables as single variables and using an anova with the F test). Then, we computed the forward selection method which produced the model *'lm.forward'* and also the stepwise selection to get the model *'lm.step'.* It was found that these 3 models provided the same results in terms of regression statistics and also produced the same model. The variable *'arace'* was removed from the model because it was not statistically significant at the 5% level. The best subset selection was not run as it would be too time consuming.

4 Regression diagnostics

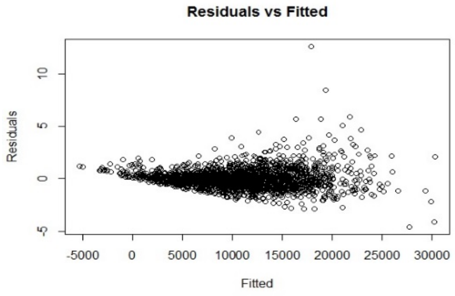
# 4.1 Goodness of Fit

This model has a p-value of the F-statistic of < 2.2e-16 which has a very high statistical significance. The adjusted R2 is equal to 0.4859 and is within 4% of the multiple R2. Also, the residual standard error of 5438 is considerably small compared to the range of afiyr values.

# 4.2 Residual Assumptions

Constant variance of residuals

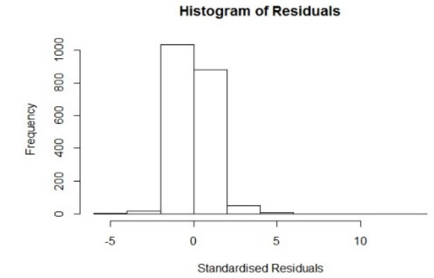
The 'Residuals vs Fitted' plot shows heteroscedasticity, synonymous of non-constant variance. This is because the variance of the residuals increases along with the fitted values.



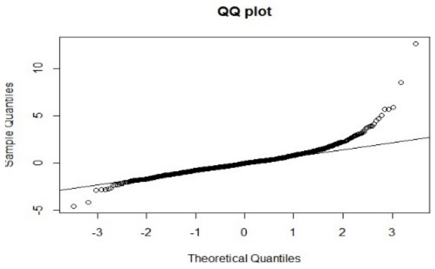
Graph 1: 'Residuals vs Fitted' of lm.backward

Normality of residuals

The residuals of the model roughly follow a normal distribution but the data is too peaked in the middle according to the histogram of residuals. This is reflected on the QQ plot which has heavy tails and which shows positive kurtosis.



Graph 2: Histogram of the standard residuals of lm.backward



Graph 3: QQ plot of lm.backward

# 4.3 Multicollinearity and the Variance Inflation Factor (VIF)

Variables	sex	aage	AJBHRS	ancars	ahhsize	aqfedhi	aregion	atenure	avote
VIF	1.1503	1.2152	1.1385	1.0659	1.0855	1.0388	1.0206	1.0393	1.0416

The VIF was significantly low for all variables. All of them were below than 5 as shown in the above table, indicating no multicollinearity. It can be concluded that explanatory variables do not depend on each other, possibly because most of them are categorical.

5 Transformations

Potential transformations were considered because of the following reasons

I. The adjusted R-squared is 0.4994 which is smaller than 0.6, showing that the goodness of fit statistics can be improved
II. The ‘Residuals vs Fitted’ plot showed heteroscedasticity.
III. The Histogram of standard residuals showed that the distributions of the residuals model is skewed.
IV. The QQ plot has heavier tails than a regular normal distribution and showed positive kurtosis.

# 5.1 Transformations on outcome variable of interest

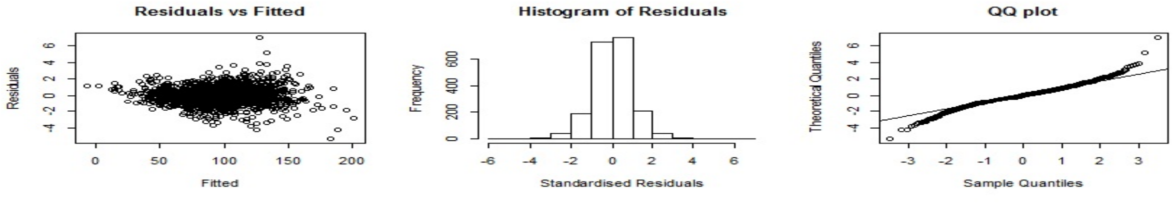
Before performing the transformations, the zero entries were removed .The **Box-Cox transform** showed λ=14/33 minimises SS(reg). However, raising the power of the dependent variable to λ would be too cumbersome and would reduce the interpretability of the model.

Some potential transformations on the dependent variable *afiyr* were considered, notably a logarithmic, square root, reciprocal and quadratic transform.

Transformation	Logarithmic	Square Root	Reciprocal	Quadratic
Adjusted R2	0.5056	0.5514	-0.01219	0.2346

As predicted by the Box-Cox, the logarithmic and square root provide the best response variable transformation since they are closest to λ=14/33, with the adjusted R2 of the square root transformation being better and within 4% of the multiple R2. The reciprocal and quadratic transforms have very poor adjusted R2 and are discarded.

The residual plots for the logarithmic transform are worse than with no transform. The residuals vs fitted plot demonstrate a high degree of heteroscedasticity. The histogram of residuals and QQ plot are skewed to the left. However, the residual plots for the square root transform have improved. The residuals vs fitted plot show roughly a reduced degree of heteroscedasticity with points being more randomly scattered. The transformation has serve its purpose to stabilize variance The histogram of residuals show more evidence of a normal distribution as we have the traditional ‘bell-curve’ and it is more symmetric. However, the data is still peaked in the middle and the QQ plot still shows positive kurtosis.

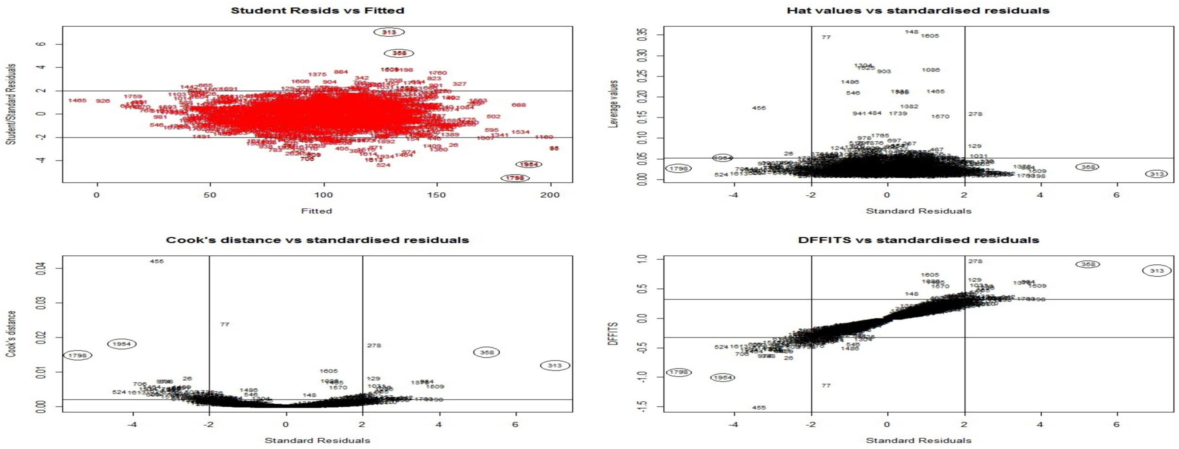


Graph 4: Residual plots of *lm.sqrt*

Therefore, a square root transformation on afiyr is needed to improve the fit of the model.

6. Outliers

We have also investigated the possibility of outliers and high leverage points. We have considered the following 4 criteria: hat/leverage values, residuals, cook’s distance and DFFITS. If data points are above the hat value 0.010096, they are high leverage points (given by the formula 2p/n where p is the number of explanatory variables and n the sample size). Values exceeding a deletion residual of 2 are considered potential outliers. Cook's values exceeding 1 are considered influential points. Our criteria for DFFITS is 0.14210 given by the formula 2(p/n)^1/2



Graph 5: Outlier plots

We identified 1833 high leverage points (hat value > 0.010096), 48 potential outliers (rstudent>2)), 1981 influential entries based on cook's distances and 285 influential points according to the DFFITS. From these high number of outliers/influential points 4 had exceptionally large values for all four criteria considered. They were: 313,358,1798,1954 as illustrated above. We found that apart from the fact that they were all male, they did not have any common properties. Also, points 313 and 358 were the individuals earning the highest annual incomes respectively.

7. Refit of model

Finally, we reran our regression without the outliers. This resulted in a model which had the same explanatory variables as *lm.full3.* The square root transformation was applied to obtain our final model *lmfinal.sqrt* with coefficient and p values given below for √afiyr.

(Intercept)		**aqfedhi_higher degree**		**aqfedhi _first degree**		**aqfedhi _teaching QF**	
39.7505	7.56e-14	46.7321	< 2e-16	34.1968	<2e-16	29.9830	4e-15
aqfedhi _Other higher QF		**aqfedhi _Nursing QF**		**aqfedhi _GCE A Levels**		**aqfedhi _GCE O Levels or equivalent**	
21.3596	2e-16	25.4108	<2e-16	12.2558	1.97e-09	7.7806	5.64e-06
aqfedhi _Commercial QF/No O Levels		**aqfedhi _CSE Grade 2-5/ Scot Grade 4-5**		**aqfedhi _Apprenticeship**		**aqfedhi _Other QF**	
5.5094	0.0734	7.2840	0.0082	2.6891	0.51324	1.8967	0.7580
aqfedhi _still at school		**atenure _owned with mortgage**		**atenure _local authority rented**		**atenure _housing assoc. rented**	
4.8679	0.6612	8.8454	2.00e-07	2.0546	0.3751	5.5107	0.2100
atenure _rented from employer		**atenure _rented private unfurnished**		**atenure _rented private furnished**		**atenure _other rented**	
-9.2742	0.0312	-3.7328	0.2660	-0.1578	0.9635	-3.6696	0.7570
aregion _Outer London		**aregion _R. of SE**		**aregion _South West**		**aregion _East Anglia**	
-5.7443	0.0998	-10.3700	0.0008	-14.7023	1.78e-05	-14.3180	0.0003

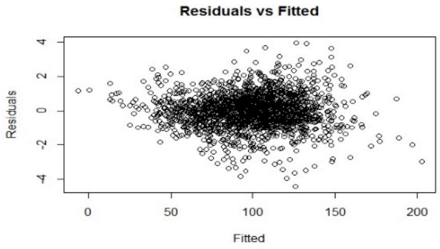
aregion _East Midlands		aregion _West Midlands Conurbation		aregion _R. of West Midlands		aregion _Greater Manchester	
-17.1993	5.90e-07	19.2623	8.37e-07	16.2264	7.49e-06	-14.7094	0.0001
aregion _Merseyside		**aregion _Region. of North West**		**aregion _South Yorkshire**		**aregion _West Yorkshire**	
-8.1200	0.0694	18.5529	6.43e-07	-14.6542	0.0005	-16.4916	2.46e-05
aregion _R. of Yorks and Humberside		**aregion _Tyne and Wear**		**aregion _ R. of North**		**aregion _Wales**	
16.9827	0.0001	19.5082	1.13e-05	14.1814	0.0002	14.0107	0.0001
aregion _Scotland		**avote _Labour**		**avote _Lib Dem/Lib/SDP**		**avote _Scot Nat**	
13.1510	8.59e-05	-3.9498	0.0026	-3.2943	0.0600	-9.8956	0.0361
avote _Plaid Cymru		**avote _Green Party**		**avote _Other Party**		**avote _Other answer**	
-27.8435	0.0409	1.0491	0.7841	-21.2719	0.0169	-10.6442	0.0141
avote _none		**avote _can't vote**		**ancars _1**		**ancars _2**	
-5.6504	0.0099	-28.3071	3.65e-10	2.5179	0.1712	6.5851	0.0012
ancars _3		**aage**		**AJBHRS**		**ahhsize**	
-4.4688	0.0939	0.7233	< 2e-16	1.2057	< 2e-16	-1.8738	8.28e-05
sex_female							
-15.5256				< 2e-16			

#7.1 Regression Diagnostics of final model

The goodness of fit statistics for our final model have improved. The adjusted R2 has increased to 0.5685 within 4% of the multiple R2. Also, the residual standard errors have decreased significantly to 23.13. The residual plots for our final model are also better.

Constant variance of residuals

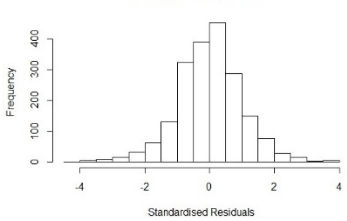
It can be observed that the points are scattered more randomly. There is not any definite pattern and this suggest constant variance residuals. The Residuals vs Fitted plot is thus satisfactory.



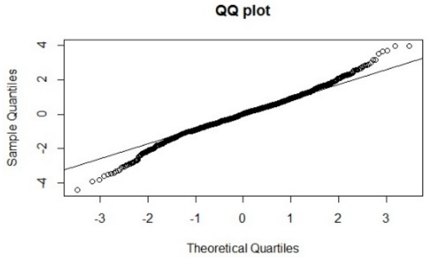
Graph 6: 'Residuals vs Fitted' of lm.full

Normality of residuals

The histogram of residuals is symmetric and centred around 0. It has a traditional 'bell curve' and strongly follows a normal distribution. The QQ plot still have a few points at the ends that deviate on the QQ line. However, most points now fit on the line .This is a satisfactory indicator that the residuals follow a normal distribution.



Graph 7: Histogram of the standard residuals of lm.final



Graph 8: QQ plot of lm.final

8 Conclusion

Our final model reasonably fits the data. Regarding GoF statistics, the adjusted R2 is slightly below 0.6 but all explanatory variables are significant at the 5% level. Also, the residual plots suggest a good fit. Our results can be used as a first evaluation of which variables affect the annual income which is discussed in the lay report.

Our model could possibly be improved by considering interaction terms but this would create more variables and reduce the interpretability of our model. Also, the explanatory variables already have low dependence on each other.

An interesting characteristic of the dataset is that some people have 0 annual income. This can be further investigated. Also, some missing values were found. The accuracy of this report can be increased if the missing data could be requested from original investigators so that the data used for our analysis comprises only of respondents who completed all questions. It is possible that the analysis is incorrect due to biasness of respondents while answering questions. For a more complete understanding behind the factors driving the annual income, a greater number of explanatory variables can be taken into account to increase predictability of the model. A larger and newer dataset can be considered as this survey was carried out more than 10 years ago.

The Lay Report – British Household Panel Survey

Introduction

The study of annual individual income is an important indicator in determining socio-economic wealth. By knowing the factors that contribute annual individual income, the government can prioritise areas of funding in order to successfully raise the socio-economic wealth of the population.

We analysed a subset of 2000 units taken from the British Panel Household Survey (BHPS). The BHPS began in 1991 and is a study that is used for a variety of purposes. It follows the same representative sample of individuals over a number of years and it is household based, meaning that every adult member of the sampled households is interviewed. The BHPS has transformed our understanding of complex trends affecting the UK society and the survey is used in helping to inform the government when making long-term policies.

In this report we analysed how the annual income of individuals can be related to the following 10 variables: gender, age, number of hours normally worked per week, the availability of a car or van for private use, the number of individuals in the household in which they lived in, highest educational qualification achieved, ethnicity, region they live in, housing tenure and the political party that the individual supports. Each of the eight categorical variables were measured according to their categories, and variables that were continuous were measured as normal. We ran a regression analysis to determine whether certain variables had influential relationships with annual individual income.

Firstly, we found that annual individual income increases with the age of the individual. In fact, for every increase of one year in age, annual individual income increases by £0.49 on average. This result was to be anticipated as naturally younger workers have to ‘work their way up’ and as they progress through their career.

Moreover, we were able to see the expected result that when the individual changed from having no qualifications to having a higher degree, the annual income increased by almost £2184 on average. In fact, for each of the other 11 categories in variable regarding highest educational qualification achieved, the individual’s annual income earned more than if they had no qualification.

Additionally, the gender of the individual proved to be an important factor in annual individual income. Our analysis showed that females were on average expected to earn around £241 less per year than males. We anticipated this result as similar findings have been found in other surveys and the gender pay gap is a well-known issue in the UK with regular news coverage. According to the Annual Survey of Hours and Earnings conducted by the Office for National Statistics, men working full time earn £96 more per week than women. These findings confirmed the validity of the conclusion that two individuals of different gender are not expected to have the same annual individual income. [1]

Although the majority of results that our analysis produced were expected, some results were unexpected. For instance, when the region of the household changes from inner London to Tyne and Wear, the annual income of the individual increases on average by around £380. This result disagreed with an article about Poverty in England published by the Guardian in 2012. The article showed that the many of the local authorities in the Tyne and Wear region were among the poorest in the England.[2] A possible reason for this unexpected result could be attributed to the areas of inner London where the household data was gathered, thus not giving a true representative sample of the population.

Furthermore, we found several individuals with an annual income of 0. It would be beneficial to the accuracy of our results if we could find out whether these individuals were unemployed or retired. This would give us a better idea of the reasons for having no annual individual income which we could then incorporate into our conclusions for our analysis.

The ethnicity of the individual was included in our initial analysis, but we observed that it was not related to annual individual income so we omitted it from our final analysis. Thus we concluded that with all other variables constant, individuals of different ethnicity are expected to attain the same annual income. This result was to be expected as recognised early on that the majority of individuals were of white ethnicity, so only a small sample of other ethnicities was taken thus leading to variable results.

From our data set we can see that there are several measures in which the government take to raise the annual individual income. Firstly, the difference in annual individual income between genders could be addressed by actively encouraging more women to further their careers or by investigating whether there is a bias towards towards men when hiring for roles of high responsibility. Additionally, the government could encourage and inform school students about the importance of gaining qualifications for future annual income. The government could also increase the awareness of qualification opportunities to individuals in order for them to increase their annual income earning potential.

The Technical Report- Moon Madness

- **Introduction**

The dataset shown to us, represents the hospital admission rates of a UK psychiatric clinic in the week before, the week during and the week after the 36 full moons from August 1971 to July 1974.

- **Results and Method:**

Prior to designing any models, we established which variables were dependent and independent. The dependent variable is the hospital admission rate, and the two independent variables are the month the admission rate was recorded and whether the admission rate was recorded before, during or after the full moon.

Variable	Type	Levels	Baseline

[1] BBC website: "Gender pay gap almost unchanged, says ONS", URL: http://www.bbc.co.uk/news/business-34855056
[2] The Guardian website: "Poverty in England", URL: http://www.theguardian.com/news/datablog/2012/mar/06/poverty-england-experian-rankings

Month	Categorical	Apr, Aug, Dec, Feb, Jan, Jul, Jun, Mar, May, Nov, Oct, Sep	Apr
Moon	Categorical	Before, during, after	During
Admission rate	Continuous	-	-

Table 1. The definition of the variables

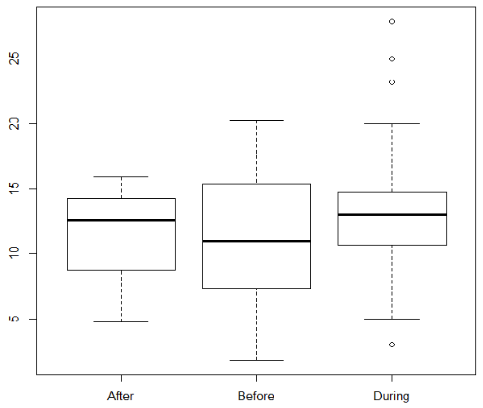


Figure 1: Boxplot (Moon vs. Admission)

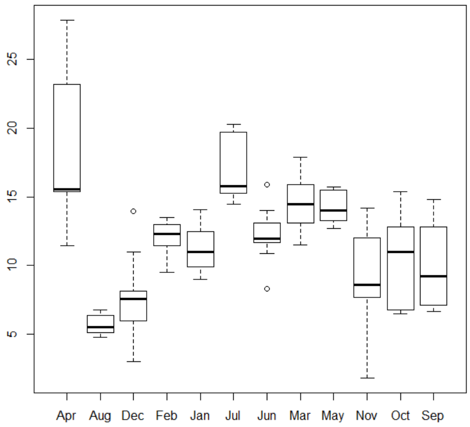


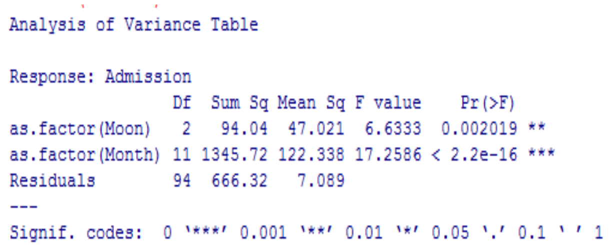
Figure 2: Boxplot (Month vs. Admission)

From *Figure 1: Boxplot (Moon vs. Admission)*, we can deduce that the mean admission rate during the full moon is higher than that of before or after the full moon. We can also observe that the mean after the full moon is higher than before the full moon. From *Figure 2: Boxplot (Month vs. Admission),* we can deduce that the highest mean is in the month of July and the lowest being August, and also we can see that all the means are different in every month.

Now, we produced a linear model called *moonmadness.lm*. The regression produced a p-value of <2.2e^-16 for our F-test. That means the p-value is significant at the 5% level. Secondly, we produced a linear model called *moonmadness.lm.2* we carried out a pairwise t-test on this specific model to compare the mean values between the moon. From the pairwise t-test, the *during-before* and *after-during* comparison is statistically significant. This suggests that the *during* is superior to the *before* treatment, and the *after* is superior to *during* treatment but that there is insufficient statistical support to distinguish between the after and before treatments. This means there is sufficient evidence to reject the hypothesis. That is to reject the fact that the group means for *during-before* and *after-during* are identical.

Our next step was to create a full model and we called this *full.lm*. By using variance inflation factor we made sure that there was no multi-collinearity. Moreover, we ran a regression for the full model and found that nearly all variables were significant with the exception of (Month)July and (Moon)after. The F-statistic was also significant and the Adjusted R-squared is 0.6399 which is within 4% of the Multiple R-squared of 0.6836. Therefore, the goodness of fit assumptions holds with our full model.

We ran an ANOVA on *full.lm* and found that the p-value of the F-statistics is significant at the 5% level giving us a p-value of <2.2e^-16 for Month and 0.002019 for Moon. This is shown in figure 3 below.



```
Analysis of Variance Table

Response: Admission
                  Df  Sum Sq Mean Sq F value    Pr(>F)
as.factor(Moon)    2   94.04  47.021  6.6333  0.002019 **
as.factor(Month)  11 1345.72 122.338 17.2586 < 2.2e-16 ***
Residuals         94  666.32   7.089
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 3. ANOVA of full.lm

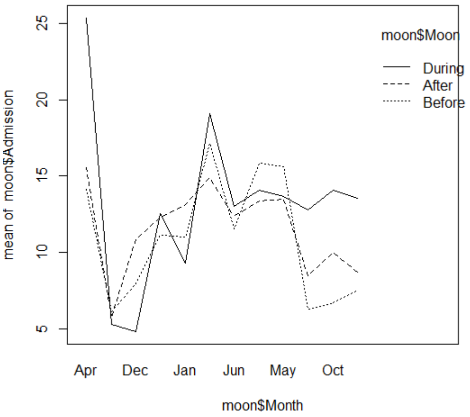


Figure 4: interaction.lm

There may be relationships between month and moon so we decided to run an interaction test on these two variables. Figure 4 conveys some evidence of interaction so we decided to investigate further by putting in the interaction term 'Moon*Month' to the full model and carry out an ANOVA test on it. By looking at the ANOVA results, we saw that the p-value of the F-statistic is < 2.2e-16. The summary of our models shows that the adjusted R-Squared is now 0.8609.

Additionally, we needed to determine whether the residual assumptions still hold so we carried out a residual analysis on the model. As shown in Figure 5, there were some negative aspects. The Histogram of the Residuals shows some negative skew and the QQ plot is slightly S shaped indicating that there is some positive kurtosis. In order to try to mitigate these possible violations of our residual assumptions, we tried some transformations.

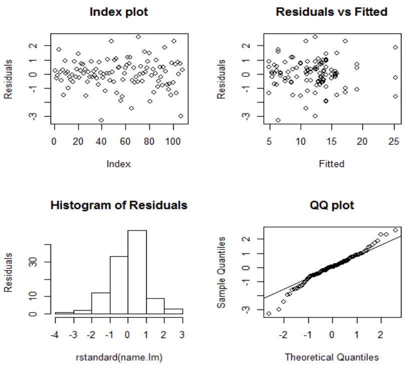
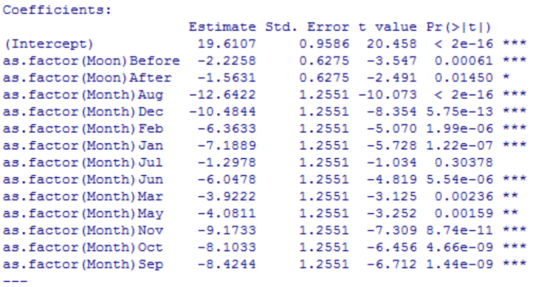


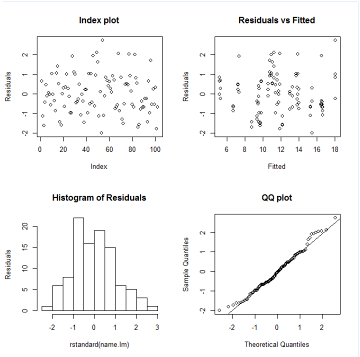
Figure 5: Residual plots for interaction.lm

On the other hand, having tried out the transformations, we found that these produced worse results, so we concluded that our full model was the best model as all the Goodness of fit assumptions hold and the adjusted R-Squared is within 4% of our Multiple R-Squared. After changing the baseline of our Moon variable to "During", all the categories within Moon are significant and only "July" within month is non-significant as shown in figure 6. However, we decided to keep it in our final model as our ANOVA test shows that the full model is significant which was discussed earlier.



```
Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            19.6107     0.9586  20.458  < 2e-16 ***
as.factor(Moon)Before  -2.2258     0.6275  -3.547  0.00061 ***
as.factor(Moon)After   -1.5631     0.6275  -2.491  0.01450 *
as.factor(Month)Aug   -12.6422     1.2551 -10.073  < 2e-16 ***
as.factor(Month)Dec   -10.4844     1.2551  -8.354 5.75e-13 ***
as.factor(Month)Feb    -6.3633     1.2551  -5.070 1.99e-06 ***
as.factor(Month)Jan    -7.1889     1.2551  -5.728 1.22e-07 ***
as.factor(Month)Jul    -1.2978     1.2551  -1.034  0.30378
as.factor(Month)Jun    -6.0478     1.2551  -4.819 5.54e-06 ***
as.factor(Month)Mar    -3.9222     1.2551  -3.125  0.00236 **
as.factor(Month)May    -4.0811     1.2551  -3.252  0.00159 **
as.factor(Month)Nov    -9.1733     1.2551  -7.309 8.74e-11 ***
as.factor(Month)Oct    -8.1033     1.2551  -6.456 4.66e-09 ***
as.factor(Month)Sep    -8.4244     1.2551  -6.712 1.44e-09 ***
---
```

Figure 6; summary for full model



Furthermore, our residual assumptions all generally hold, although it can be argued that our Histogram of Residuals is a little bit skewed and our QQ plot is slightly S shaped. Therefore we decided to identify possible outliers using the plot.outliers.1 function and found that data 101,93,21,40,53 and 81 can be possible outliers. We therefore removed them and when we ran our linear model again, this produced better residual plots while all the assumptions still hold.

Figure 7; Residual plots for full.lm.1 (with outliers removed)

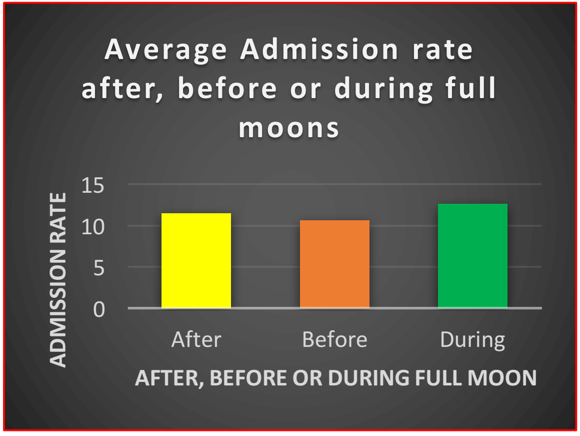
Conclusion

Although the categories of our variables are not all significant, we managed to find that most of the Goodness of Fit and Residual assumptions hold which would give us some evidence that these results are valid. However, we should be cautious when drawing conclusions from our analysis as our dataset is quite small which can result in issues with precision. The data is also collected from one clinic which may lead to bias results. For example, it may be due to confounding where there may be external factors at that location that cause changes in the hospital admission rates which affected our results. In order to better understand whether the moon makes people behave strangely, we need to have more explanatory variables as well to make sure that we take into account other relationships which may affect hospital admission rates.

Lay Report

The aim of our report is to assess whether there is a relationship between the full moon and people's behaviours. We did this by looking at data of emergency hospital admission rates of a UK psychiatric clinic at different times of the year. We ran a regression on this by looking at the admission rates the week before, during and after 36 full moons ranging from January to December so we can also see if the time of the year makes any difference.

Below is a bar chart of our data showing the average admission rates the week before, the week during and the week after the full moons. We can see that on the week during the full moon, the admission rate is at its highest and the week before is the lowest. This tells us that there is evidence that shows that during the full moon, more people go to the hospital for emergency treatment.



Our regression shows that for every one increase in the admission rate during the full moon on average, there is a one to two decrease in the admission rate before and after the full moon. This means that on average more people are going to the hospital during the week of the full moon which agrees with the fact that the full moon affects people's behaviours negatively.

The time of the year did contribute to our understanding of the effects of the full moon although there are some odd results. The results for July doesn't seem to be consistent with other months. This can however be explained by our small sample size; there is only 9 pieces of data per month. Therefore, it is very difficult to see whether the time of the year affects the admissions rate. Our regression does show that there is evidence that the month affects the admission rates of the hospital which implies that it may be not due to the moon. For example, there may be special events happening in a few of the months which cause more hospital admissions.

Figure 8: Average Admissions Rate

Based on our analysis it is hard to say whether there is a relationship between the full moon and people's behaviours. Although our data shows that the admissions rate is higher during the week of the full moon, the difference is only small and our dataset is not large enough to make comprehensive conclusions. Furthermore, other factors can play a role in the admission rates too. For example, in the winter more people may get ill which may be a reason behind the increase in admissions rate in specific months. However, our dataset is not helpful in helping us identify these factors that may affect our results due to the small dataset and the large number of months. Our data also only comes from one UK clinic and location may also play a role in the admission rate which is not reflected in our dataset. If we were to do this again, we would gather more data from many different locations over a few years so we have quality data for each month. This would allow us to draw more reliable conclusions from our analysis.