Regression Analysis of Car Showroom Profits

STAT0006 Coursework

Yu Kate Chan

Part 1: Normal linear model

Introduction to the data

This report aims to investigate the potential factors affecting the daily profits of a car showroom. The provided dataset comprises 314 profit entries over a five-year period (2019-2023), featuring four categorical and three numerical variables. Notably, the dataset is complete with no missing values.

The mean profit is approximately 2798.8 GBP, ranging from -118.3 GBP to 13632.8 GBP (with negative profits indicating a loss). From analysing the density plot of profits below, the data shows a positive skewness, indicating potential overestimation when using the mean for analysis as the mean is greater than the median. Thus, we prefer to use the median values in this case [1].

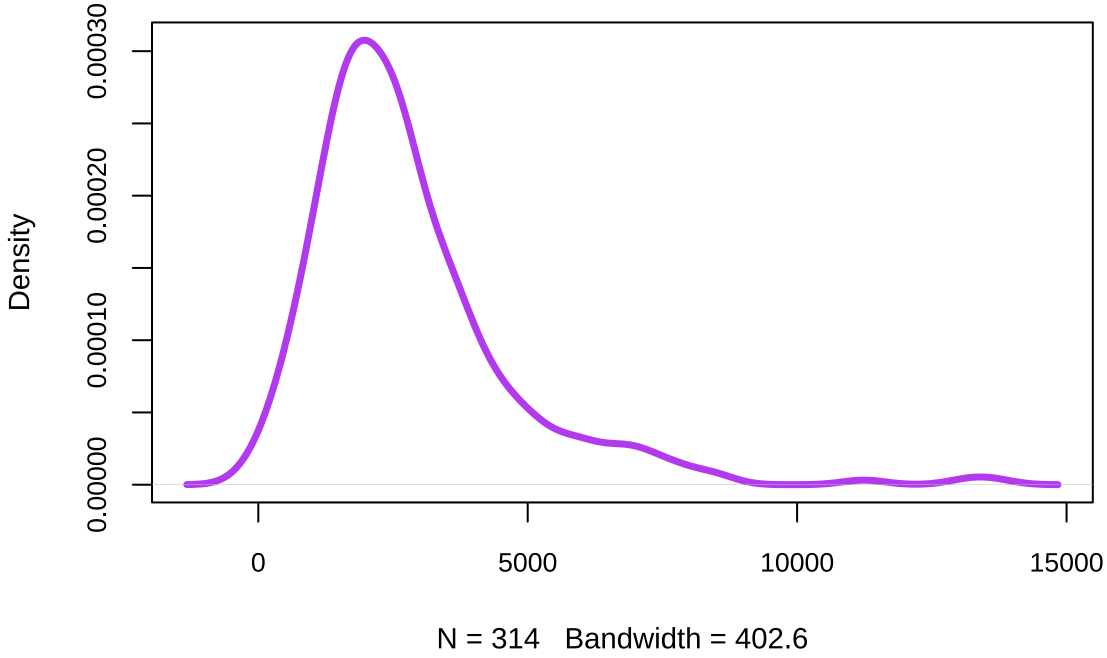


Figure 1: Density plot of profits

[1] Frost, J. (2023) Skewed distribution: Definition & examples, Statistics By Jim. Available at: https://statisticsbyjim.com/basics/skewed-distribution/ (Accessed: 07 January 2024).

We analyse the categorical variables which are year, day type, rainy condition, and new car release using boxplots as follows. The black line represents the median. All the boxplots indicate some observations far beyond the upper whisker, while there is no observations locating beyond the lower whisker. Considering that the mean is more sensitive to extreme values than the median, the median might therefore provide a more realistic representation of the data. All categorical variables seem to exert little influence on the daily profit.

No significant trend in median profit is observed across the five years of collected data. However, the increased occurrences of outliers in 2020, 2021, and 2022 may be attributed to higher data volume in these years. Notably, the relatively small box for the year 2019 indicates significantly less dispersion in the data compared to other years. Considering data on whether a new car model has been released in the past seven days, it is observed that when there is no new release, the median profit is higher at 2425.15 GBP, whereas the median profit is at 2064.10 GBP when there is a new car release. Notably, there are 26 entries for new releases and 288 entries for days with no release.

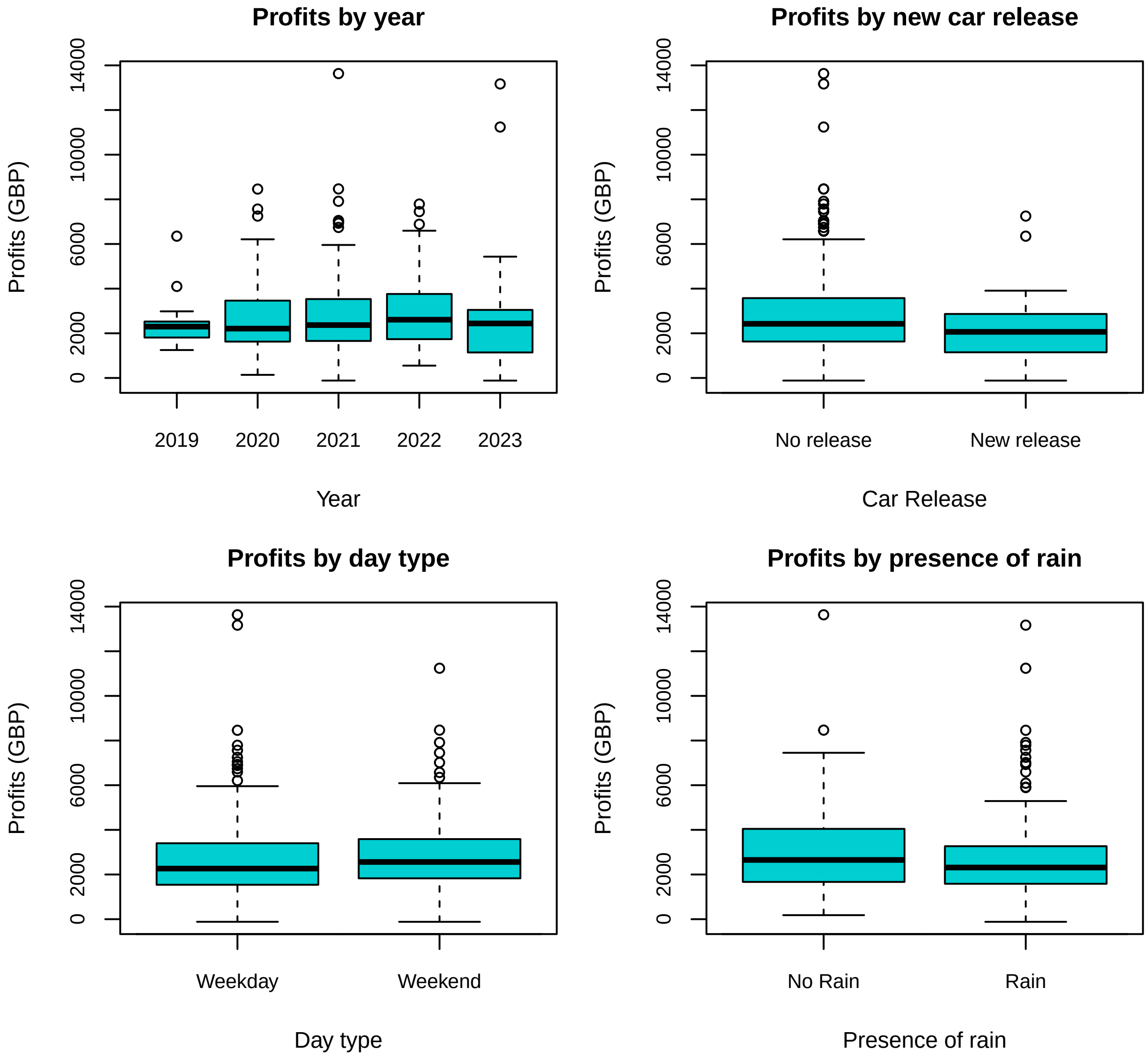


Figure 2: Profits by categorical variables

Examining profits by day type, weekends show a higher median at 2561.92 GBP compared to weekdays at 2264.93 GBP, with both day types exhibiting positive skewness in profit distribution. Additionally, considering data on whether there was any rain at the showroom on a given day, lower profits at 2313.57 GBP are noted when rain is present compared to 2652.45 GBP when there is no rain, as depicted in the boxplot.

The temperature data is analyzed using the scatter plot below. The temperature ranges from -2.42 degrees Celsius to 25 degrees Celsius, with a mean of 2.72 degrees Celsius. A positive relationship is observed between temperature and daily profits, with higher temperatures generally associated with greater profits. Notably, the variability of profits seems to increase as temperature rises. Additionally, the data reveals a weak positive association between advertising expenditure and daily profits, and a right-skewed distribution is evident. Moreover, the variability of profits appears to increase with higher advertising expenditure. Data on the number of staff working on a given day ranges from 1 to 5 staff members. The median profits appear to increase with the number of staff members, except for 5 staff members, for which only 1 data point is available.

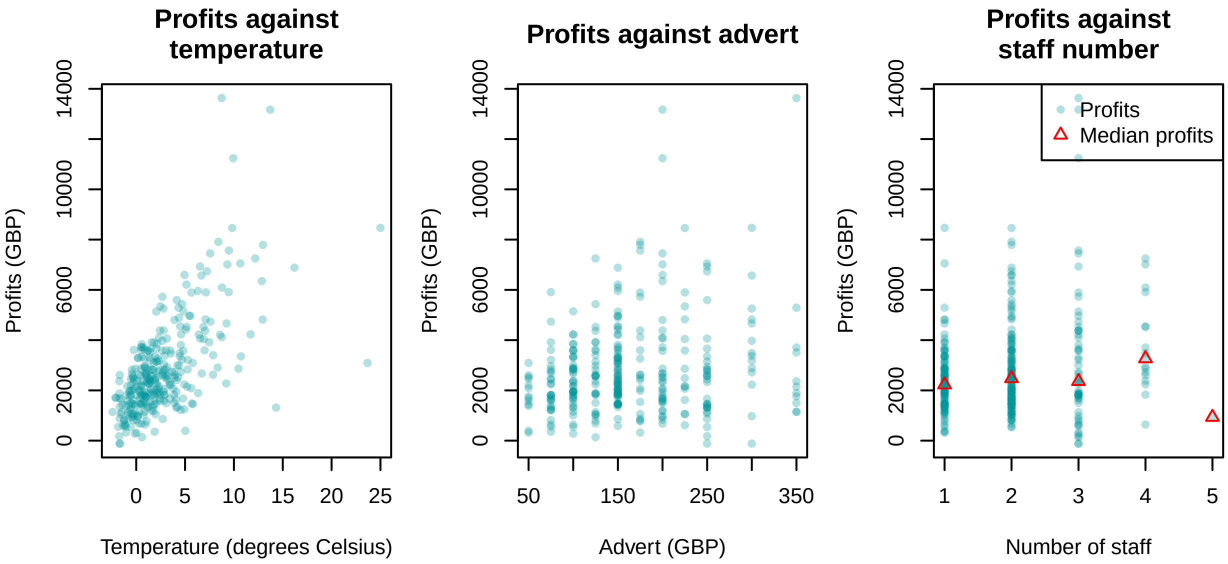


Figure 3: Profits by numerical covariates

In conclusion, the analysis reveals that the covariates that might be potentially useful in modeling the factors that affect profits are the number of staff, new car release, day of the week, temperature, and presence of rain. Though the categorical covariates do not appear to have much influence on the profits, the interactions between them and other numerical covariates might be insightful.

Model building

When building the model, we used backward elimination with α_{remove} set to 0.05. Through each iteration of the process, we remove the covariate that has the highest p-value. This allows us to determine the covariates that are most significant in predicting daily profits.

Firstly, we start with a full model including all the covariates, which are `staff`, `advert`, `new_release`, `temperature`, `weekend`, `rain` and `year`. We obtain Model 1 summarised below:

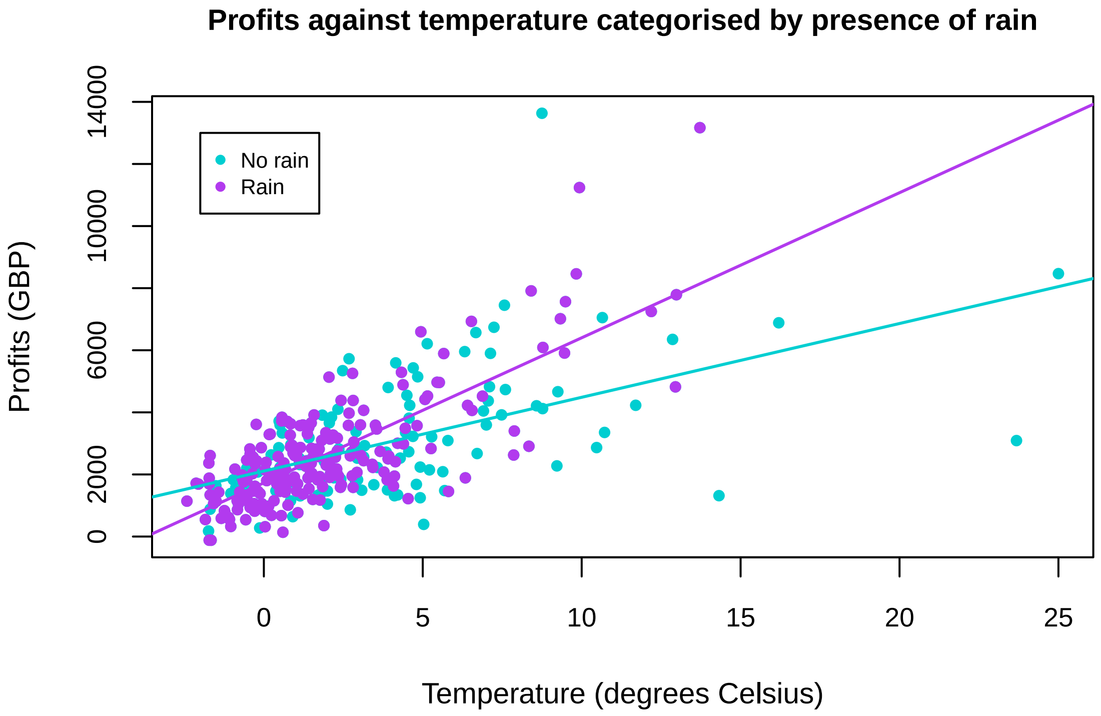
```
##
## Call:
## lm(formula = profits ~ staff + advert + temperature + new_release +
##     weekend + rain + year, data = prof)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5628.0  -806.5   -92.3   642.6  7544.8
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   211.298    406.211   0.520   0.6033
## staff         448.873     90.841   4.941 1.29e-06 ***
## advert          4.391      1.069   4.107 5.16e-05 ***
## temperature   338.804     21.589  15.694  < 2e-16 ***
## new_releaseY -557.909    283.299  -1.969   0.0498 *
## weekend1       35.791    173.099   0.207   0.8363
## rainY         204.815    169.853   1.206   0.2288
## year2020      -92.570    356.746  -0.259   0.7954
## year2021       28.715    351.911   0.082   0.9350
## year2022        7.017    355.452   0.020   0.9843
## year2023     -184.069    385.508  -0.477   0.6334
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1355 on 303 degrees of freedom
## Multiple R-squared:  0.5041, Adjusted R-squared:  0.4878
## F-statistic: 30.81 on 10 and 303 DF,  p-value: < 2.2e-16
```

From the initial model, the dummy variable `weekend1` has the highest p-value overall since some of the dummy variables for `year` are lower. From this, we remove `weekend` from our initial model and refit the model without this covariate.

By repeating the process above, we removed the covariates in the order of `weekend`, `year`, and finally, `rain`. The remaining covariates are `staff`, `advert`, `temperature` and `new_release`. This is summarised by the following Model 2:

```
##
## Call:
## lm(formula = profits ~ staff + advert + temperature + new_release,
##     data = prof)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5784.9  -800.2   -28.7   649.7  7499.3
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   329.123    259.882   1.266   0.2063
## staff         453.338     88.385   5.129 5.15e-07 ***
## advert          4.385      1.062   4.131 4.66e-05 ***
## temperature   332.506     20.516  16.207  < 2e-16 ***
## new_releaseY -598.069    277.165  -2.158   0.0317 *
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1347 on 309 degrees of freedom
## Multiple R-squared:  0.5001, Adjusted R-squared:  0.4936
## F-statistic: 77.28 on 4 and 309 DF,  p-value: < 2.2e-16
```

In the backward elimination method, interactions between the covariates were not considered. We created a scatterplot of `profits` against `temperature`, with different colours corresponding to the presence of rain and their individual line of best fits. Since the purple line is steeper than the blue line, there is evidence that the effect of temperature on daily profits changes with the presence of rain. Hence, the interaction between `temperature` and `rain` would be useful to consider in the final model.



By including the interaction between temperature and rain, we add two additional coefficients to our model the dummy variable (`rainY`) and the interaction between temperature and rainY (`temperature:rainY`). This is our final model, called Model 3:

```
##
## Call:
## lm(formula = profits ~ staff + advert + temperature * rain +
##     new_release, data = prof)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4035.5  -797.8   -73.2   620.2  8095.3
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        675.923    289.182   2.337   0.0201 *
## staff              393.731     86.246   4.565 7.23e-06 ***
## advert               4.351      1.023   4.252 2.82e-05 ***
## temperature        246.574     28.025   8.798  < 2e-16 ***
## rainY             -389.101    201.631  -1.930   0.0546 .
## new_releaseY      -537.830    267.736  -2.009   0.0454 *
## temperature:rainY  204.387     41.611   4.912 1.47e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1298 on 307 degrees of freedom
## Multiple R-squared:  0.5388, Adjusted R-squared:  0.5298
## F-statistic: 59.77 on 6 and 307 DF,  p-value: < 2.2e-16
```

`temperature:rainY` has a very small p-value, therefore it is useful to include in the final model. Although the p-value of `rainY` is statistically insignificant, it is included due to it being a main effect, which influences the interpretation of the interaction terms.

An ANOVA test is performed comparing Model 2 and Model 3 to confirm that a more complex model with interactions better predicts the daily profits. The p-value is very small so the additional terms are worth including in the final model.

```
## Analysis of Variance Table
##
## Model 1: profits ~ staff + advert + temperature + new_release
## Model 2: profits ~ staff + advert + temperature * rain + new_release
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1    309 560989574
## 2    307 517612860  2  43376714 12.864 4.317e-06 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

Model checking for final chosen model

To assess for linearity we plot the standardised residuals with each numerical covariate (`staff`, `advert`, `temperature`) included in the final model. Under the assumption of linearity, the plots should have a random scatter of points around zero. For all plots, we can observe that the data points are scattered randomly around the horizontal line. There are a few extreme values when there are 1 or 3 members of staff working. However, a few extreme values are normal due to randomness, suggesting no evidence that the linear assumption is violated for our model. Also, the variability of the standardised residuals increases at higher temperatures.

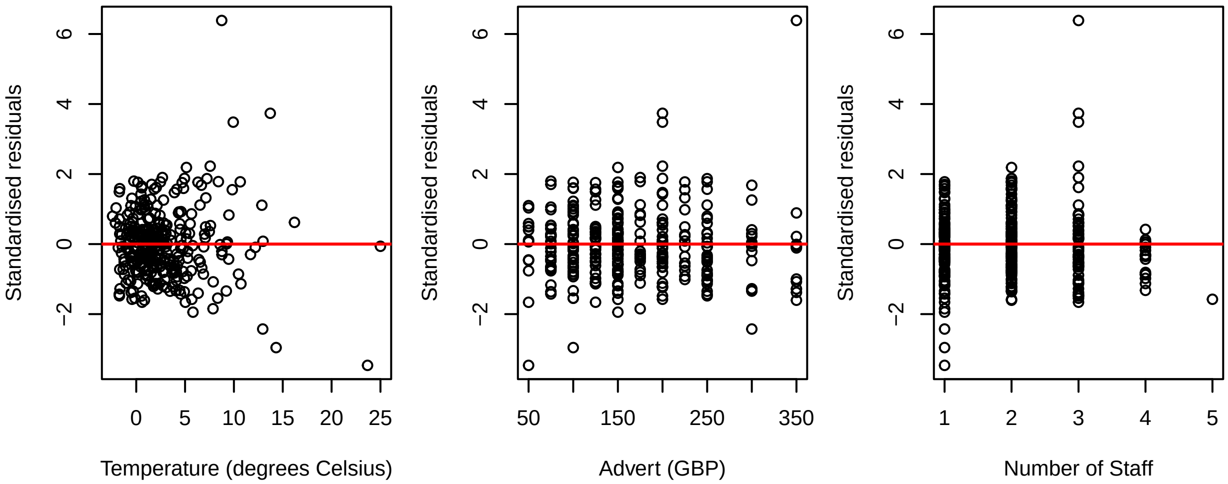


Figure 4: Plots to assess the assumption of linearity by covariates

To assess for homoscedasticity of the error terms, we plot the standardised residuals against their corresponding fitted values. There is a random scatter of points along the horizontal line. However, there is increased variability for higher fitted values, so there may be some evidence against homoscedasticity. There was some consideration into taking the log of temperature to possibly reduce heteroscedasticity, but since it has an interaction with rain, it would cause interpretation issues.

To assess for normality of the error terms, we use a Q-Q plot, where the points should fall along the 45-degree angle line. The majority of the points in the middle lie close to this line. At either end, the points curve away from the line, especially in the positive direction. There is no significant violation of the normality assumption since minor deviations due to extreme values are expected.

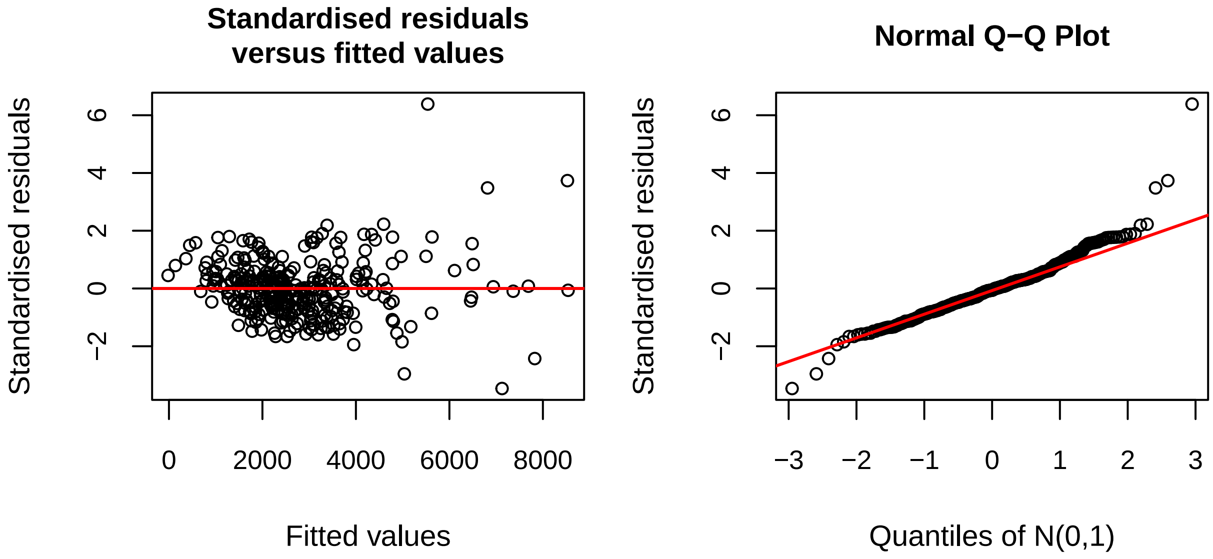
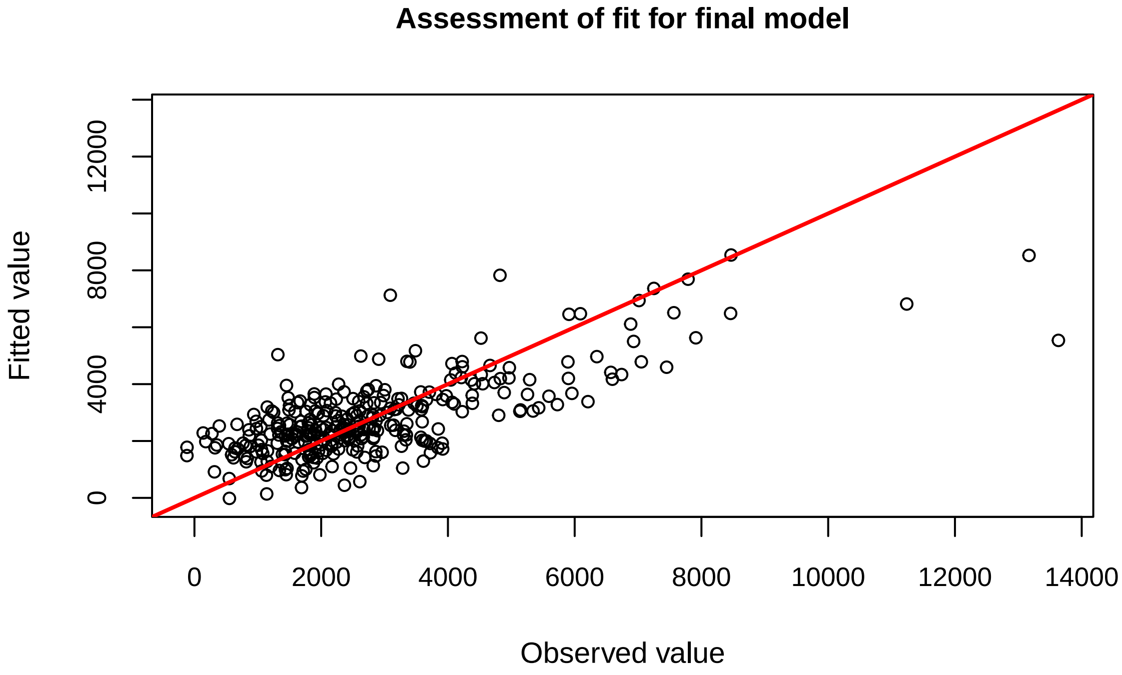


Figure 5: Plots to assess the assumptions of homoscedasticity (left), and normality (right)

Since the observations are sampled from random days, they are randomly ordered, so we are unable to check for serial correlation.

Overall, none of the plots suggest any major violation of the modelling assumptions.

Finally, to check the fit of the final model, we plot all the fitted values against the corresponding observed values of profits. For the model to be a good fit, the points on the plot should lie closely along the 45-degree line. In general, the majority of the points are scattered along this line, with the exception of some larger values. Hence, we can conclude that our final model would be suitable for predicting profits based on the covariates and interactions in the model.



Conclusion

According to Model 3, the main factors that influence the daily profits of the showroom are the number of staff, the amount spent on advertising in the previous seven days, the release of new car models, the temperature and the rainy conditions. In general, the release of new car models in the previous seven days is expected to reduce the daily profits by 537.83 GBP. Likewise, the daily profit is anticipated to shrink by 389.10 GBP if there is rain at the showroom. On the contrary, given that all the other factors are unchanged, a relatively high temperature is rather promising for greater profits. While profits are estimated to be worse on rainy days, the influence of the temperature on profits is greater on such days. Specifically, for every one degree Celsius increase in temperature, the profit is expected to rise by 246.57 GBP, and an additional 204.39 GBP when it rains. Having more staff on site and increasing expenditure on advertisement also appear to have a positive impact on profits. For each additional staff, the daily profit is predicted to grow by 393.73 GBP; and for each GBP increase in the amount spent on advertising in the previous seven days, the daily profit is expected to climb an extra 4.35 GBP. Among the controllable factors, the number of staff has the most positive influence on the profit of each day. But it might not be practical to increase the staff number to a great extent. Alternatively, it may be more reasonable and feasible to raise the budget for advertisement.