STAT0006 ICA 3

Group 97

Student numbers: 22087103, 22112264, 21132230, 22026212

Part 1: Normal linear model

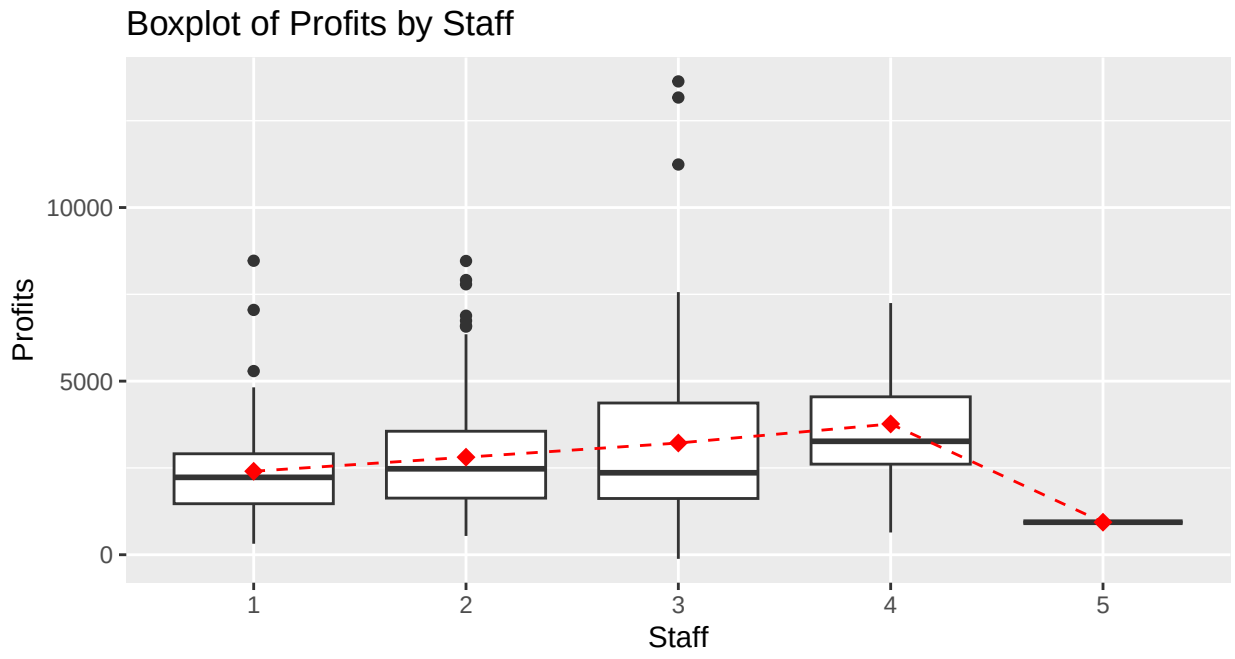
Introduction to the data

Data containing information on profits of car sales at a showroom were randomly collected to analyze how various factors may affect profits. There are 314 observations in total with no missing values.

The dataset includes information on showroom operations, featuring the following variables. The response `Profits` represents the financial outcomes of the showroom on a given day, denominated in GBP, with negative values indicating losses. `Staff` signifies the number of employees present at the showroom during the recorded day. `Advert` reflects the expenditures on advertising by the showroom in the preceding seven days. The binary variable `New Release` indicates whether any new car models were introduced in the last week (Y/N). `Weekend` is a binary indicator distinguishing between weekdays (0) and weekends (1). `Temperature` quantifies the average temperature at the showroom in degrees Celsius. The binary variable `Rain` signifies the occurrence of rain at the showroom on the recorded day (Y/N). Lastly, `Year` denotes the calendar year in which the observation was documented, ranging from 2019 to 2023.

Then, the relationships between profits and each covariate are investigated by plotting profits against each covariate respectively.

Profits vs Staff

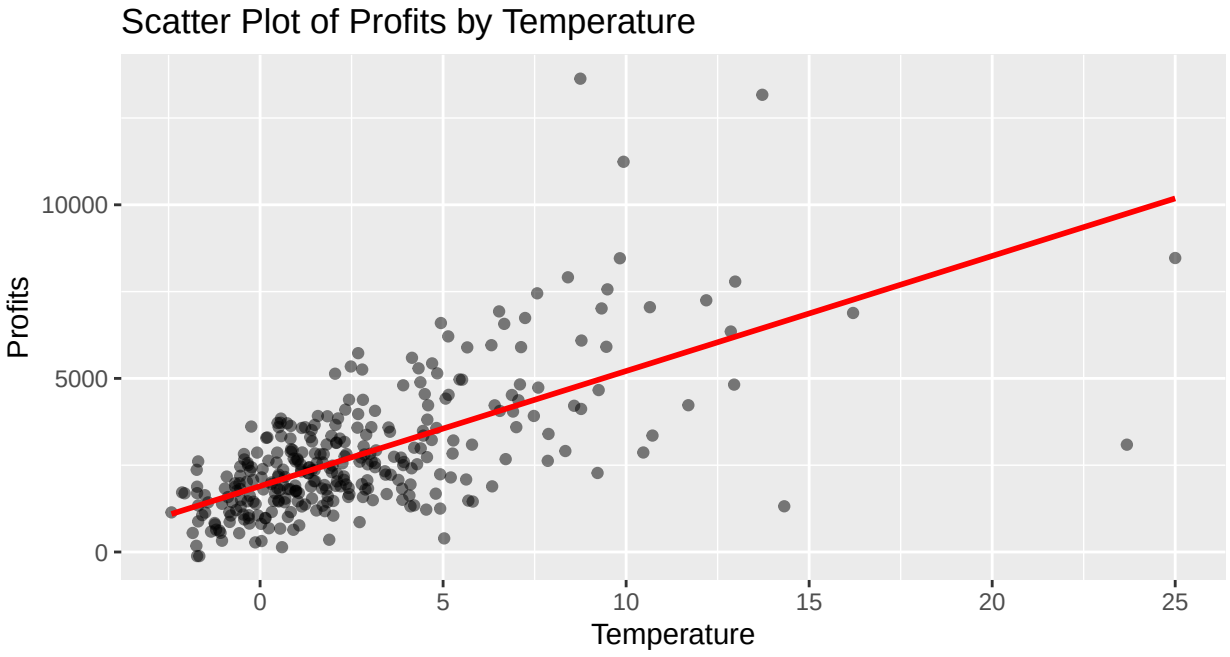


Although staff is a numeric covariate, we use boxplots to gain insight into the relationship since staff only take integers from 1 to 5 in the dataset. It is notable that when the number of staff is 5, there is only one data point in the dataset, so the plot is merely a single line. The red points represent the values of the mean in each boxplot. From the trend of the means, it seems that in the range from 1 to 4 staff, there is a positive linear relationship between profits and staff.

Profits vs Advert

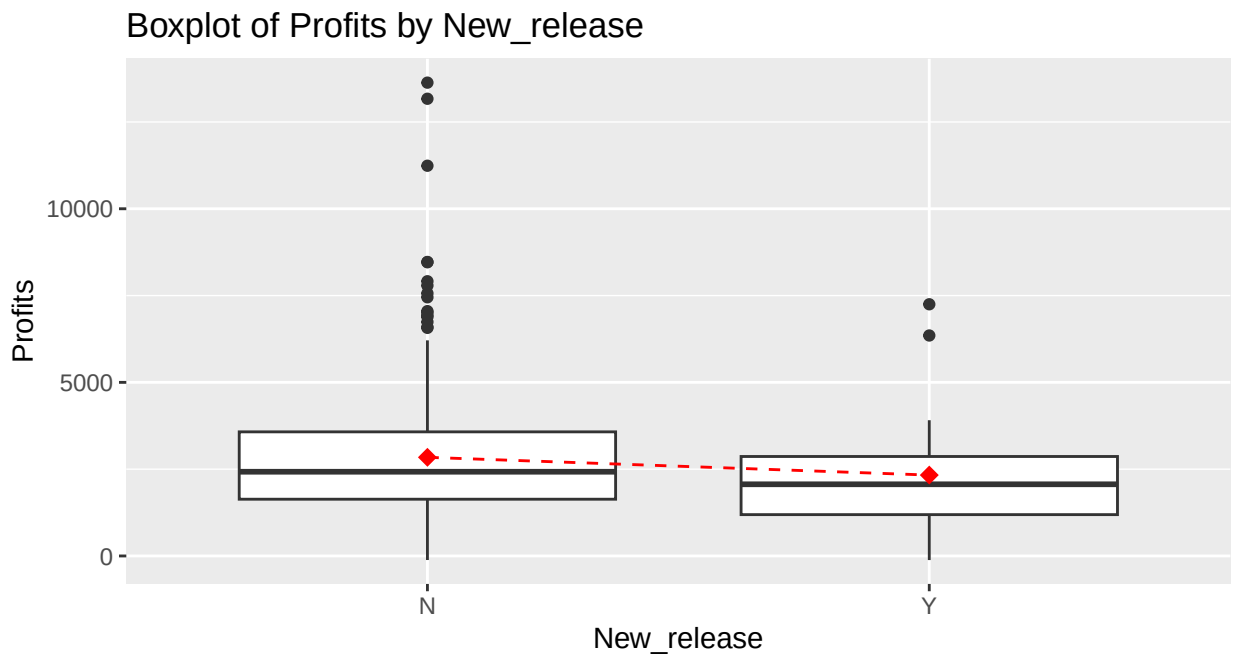
The scatter plot above shows a slightly positive correlation between profit and the advert.

Profits vs Temperature



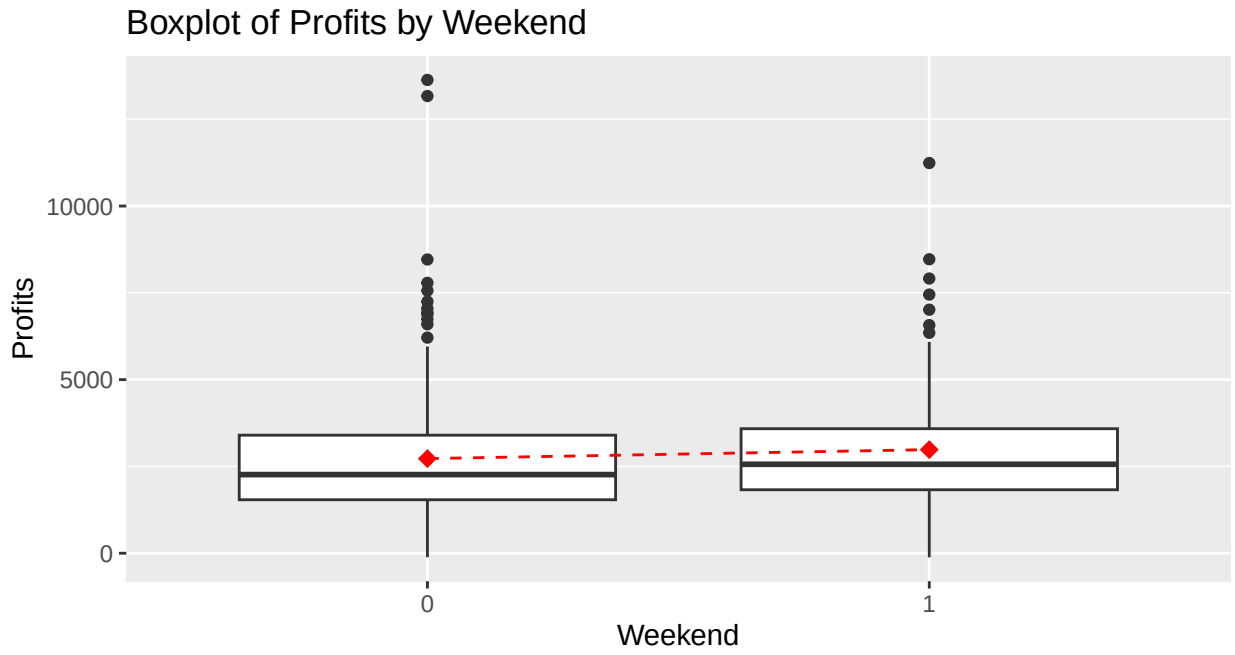
The graph showcases an apparent positive correlation between profits and temperature since profits increase as temperature rises and the trend potentially demonstrates a linear relationship.

Profits vs New_release



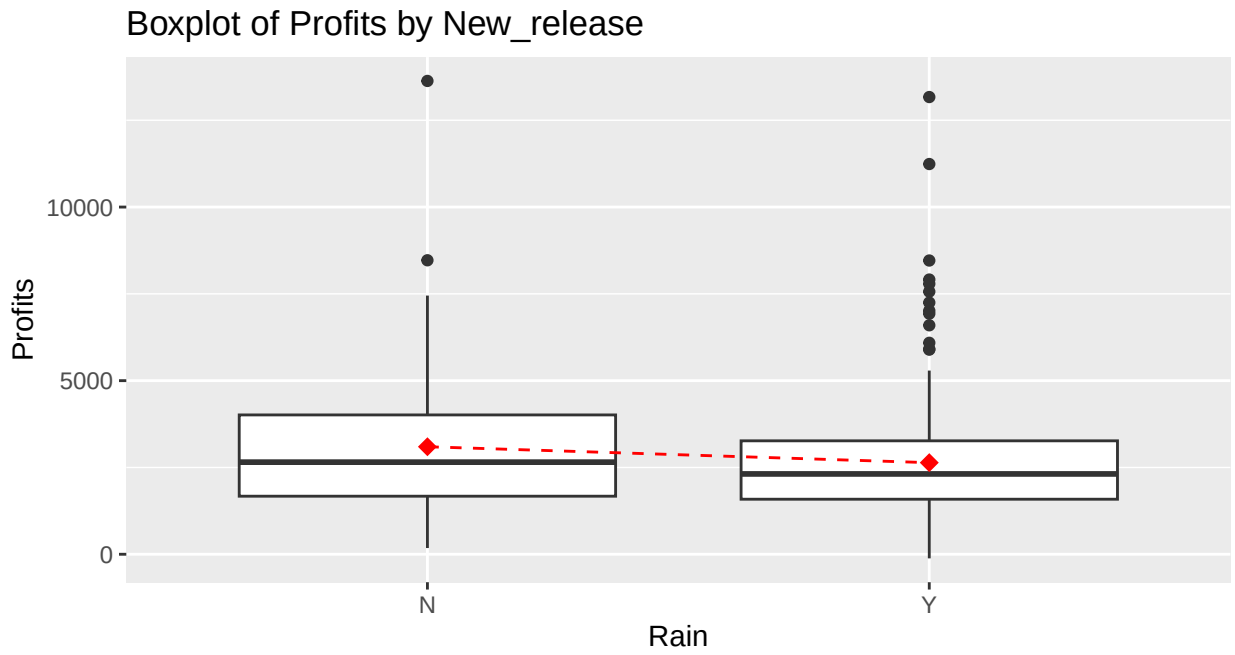
The medians, means and interquartile range across states of new release are very similar, while with new release there are several more extreme values outside the interquartile range.

Profits vs Weekend



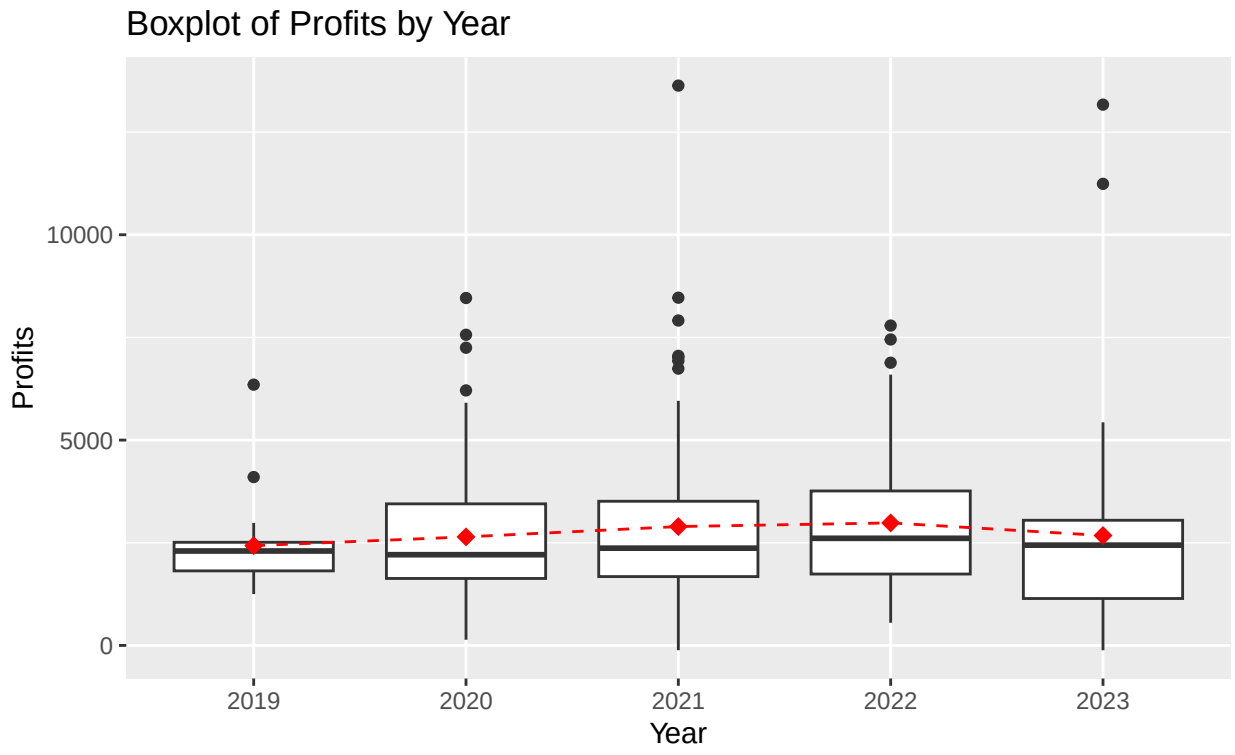
For the categorical covariate, weekend, the means, medians, interquartile ranges and ranges of profits across states of weekend are again very similar. And in both states, a few extreme values exist.

Profits vs Rain



For the categorical covariate, rain, the medians and means are close. When there is rain, the interquartile range is relatively smaller, and several more profits data points fall outside the interquartile range.

Profits vs Year



Year is a numerical covariate, but here we use boxplots of profits against year since it only takes 5 distinct values. On the boxplot, there is no obvious trend in the median across the 5 years, while there is a slightly

increasing trend in the mean.

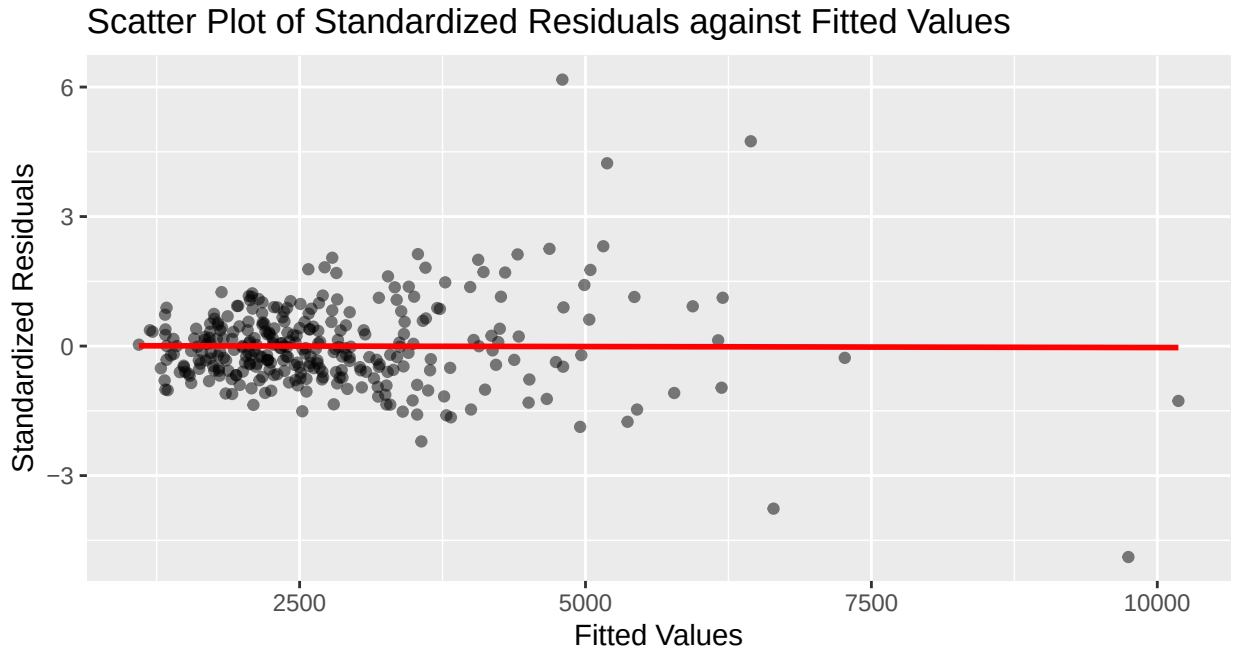
Model building

From the exploratory data analysis above, it seems that there is a positive linear relationship between profits and temperature. So we start with a simple linear regression model using temperature as the covariate (model 0).

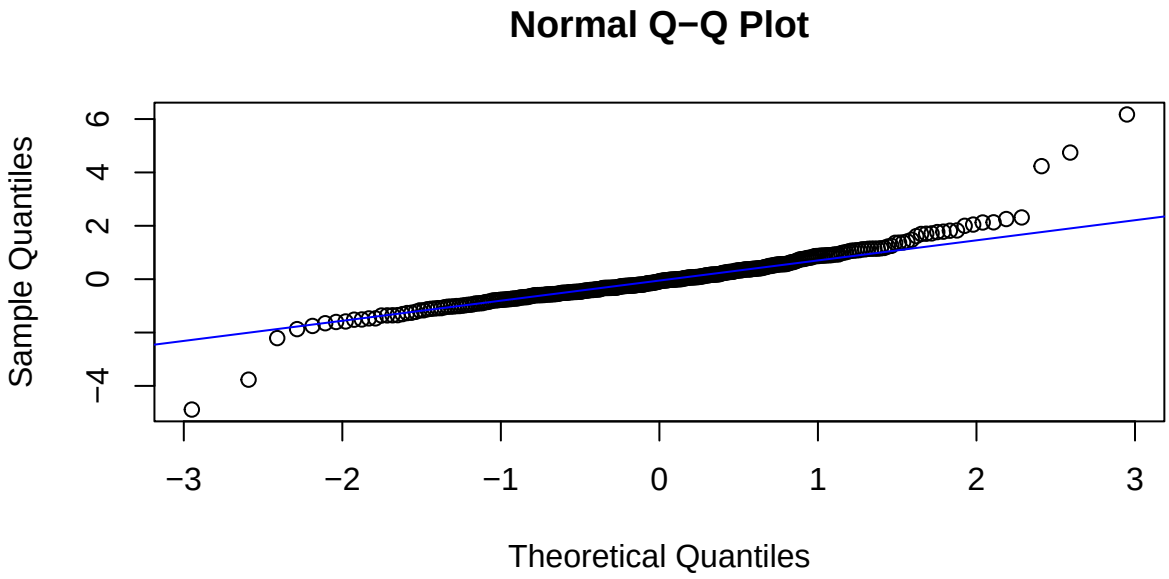
```
##
## Call:
## lm(formula = profits ~ temperature, data = profits)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6656.5  -806.6  -116.4   654.7  8835.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1897.13     100.72   18.84   <2e-16 ***
## temperature   331.50      21.89   15.15   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1440 on 312 degrees of freedom
## Multiple R-squared:  0.4237, Adjusted R-squared:  0.4219
## F-statistic: 229.4 on 1 and 312 DF,  p-value: < 2.2e-16
```

The p-value for temperature is very tiny, so there is evidence to reject the hypothesis that there is no relationship between profits and temperature. The coefficient of determination is about 0.424, which indicates that the covariate temperature captures a proportion of 42.4% of the variability in the response profits.

We also plotted some plots to check the model assumptions. Linearity was already checked by the scatter plot of profits against temperature in the previous EDA.



Since most values of standardized residuals scatter around 0 within the range from -2 to 2, the assumption of homoscedasticity is not violated.



On the normal Q-Q plot above, most data points follow the Q-Q line, so the assumption of normality is also satisfied.

Thus we should include temperature in our final model and there is no need to carry out any transformation.

Likewise, we also tried building simple linear models using advert and staff respectively. The p-value for advert is about 0.0006 and for staff is about 0.0016. Therefore there is evidence to reject the hypothesis that there is no relationship between profits and advert or profits and staff. Model assumptions were also checked by plotting plots.

Then, we build a general linear regression model using the three numeric covariates (model1).

```
##
## Call:
## lm(formula = profits ~ temperature + staff + advert, data = profits)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5773.8  -782.1   -44.6   647.8  7575.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  322.539    261.392   1.234    0.218
## temperature  333.168     20.634  16.147  < 2e-16 ***
## staff        436.909     88.575   4.933 1.33e-06 ***
## advert         4.311      1.067   4.039 6.78e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1355 on 310 degrees of freedom
## Multiple R-squared:  0.4926, Adjusted R-squared:  0.4877
```

```
## F-statistic: 100.3 on 3 and 310 DF,  p-value: < 2.2e-16
```

We can see that all p-values are very tiny. So we can include all the three covariates in our final model. Now the coefficient of determination is about 0.493, suggesting that the model accounts for 49.3% of variability in the response.

For categorical covariates, it seems that there is no obvious pattern from the exploratory analysis. To spot any problem, we first try to build a model with all the 7 covariates given in the dataset (model 2). In the model, year is treated as a categorical covariate, as it only has 5 values and there is no obvious trend of profits as year increases from the previous analysis. We also tried including year as a numerical covariate in the model, but the result was not satisfying.

```
##
## Call:
## lm(formula = profits ~ staff + advert + temperature + new_release +
##     as.factor(weekend) + rain + as.factor(year), data = profits)
##
## 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 *
## as.factor(weekend)1    35.791    173.099   0.207   0.8363
## rainY                 204.815    169.853   1.206   0.2288
## as.factor(year)2020   -92.570    356.746  -0.259   0.7954
## as.factor(year)2021    28.715    351.911   0.082   0.9350
## as.factor(year)2022     7.017    355.452   0.020   0.9843
## as.factor(year)2023 -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
```

The p-values for weekend, rain and year are larger than 0.05. Especially p-values for weekend and year are very large. Also, the coefficient of determination only makes a little improvement from 0.493 to 0.504 after adding weekend, rain and year as covariates. From the exploratory analysis, there is no obvious pattern of profits in the boxplots of profits against weekend and against year, so there may be no relationship between profits and weekend or year.

To test whether we can remove them, we built a model without weekend and year and compared the nested model (model 3) with the full model with a F-test.

```
##
## Call:
## lm(formula = profits ~ staff + advert + temperature + new_release +
##     rain, data = profits)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5813.8  -773.3   -83.0   600.7  7597.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   194.605    282.102   0.690   0.4908
## staff         441.377     88.858   4.967 1.13e-06 ***
## advert          4.408      1.061   4.155 4.22e-05 ***
## temperature   339.863     21.368  15.905  < 2e-16 ***
## new_releaseY -577.030    277.482  -2.080   0.0384 *
## rainY         204.273    167.385   1.220   0.2233
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1346 on 308 degrees of freedom
## Multiple R-squared:  0.5025, Adjusted R-squared:  0.4944
## F-statistic: 62.22 on 5 and 308 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Model 1: profits ~ staff + advert + temperature + new_release + rain
## Model 2: profits ~ staff + advert + temperature + new_release + as.factor(weekend) +
##     rain + as.factor(year)
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1    308 558290001
## 2    303 556470747  5   1819254 0.1981 0.9631
```

The p-value of the F-test is very large, which indicates that there is no evidence that the nested model is not as good as the full model. Therefore, we decided to choose the nested model.

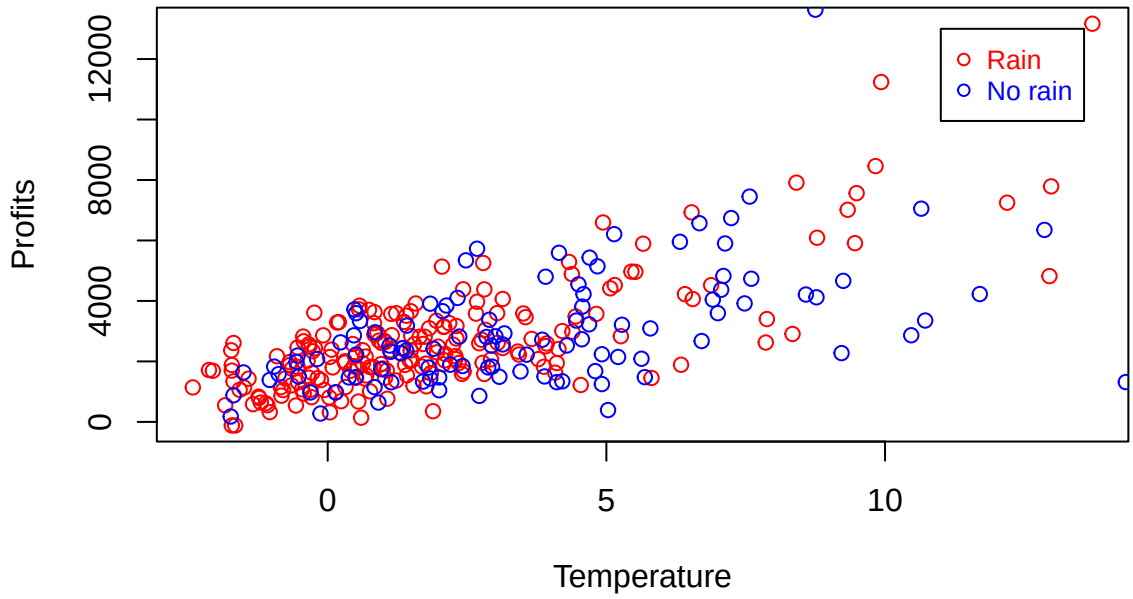
In the previous full model, the p-value for rain is 0.223. Should we remove it as well? Since distinctively, the relatively large p-value can result from an interaction between rain and temperature, we built a general linear model using advert, staff, new_release and rain (model 4) to see how the p-value for rain would change.

```
##
## Call:
## lm(formula = profits ~ as.factor(rain) + advert + staff + new_release,
##     data = profits)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3642.0 -1076.4  -238.4   838.4  9887.4
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1567.983    361.861   4.333 1.99e-05 ***
## as.factor(rain)Y -546.914    216.371  -2.528 0.011981 *
## advert              4.874      1.429   3.411 0.000734 ***
## staff             428.440    119.721   3.579 0.000401 ***
## new_releaseY     -717.655    373.689  -1.920 0.055720 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 1814 on 309 degrees of freedom
## Multiple R-squared:  0.0939, Adjusted R-squared:  0.08217
## F-statistic: 8.005 on 4 and 309 DF,  p-value: 3.753e-06
```

It can be seen that the p-value for rain becomes smaller than 0.05 in the model without temperature. This suggests that adding temperature to covariates increases the p-value for rain.

To further investigate whether there is interaction between rain and temperature, the following plot is plotted.



It appears that on rainy days, profits tend to increase slightly more as temperature increases. Now we try to add the interaction between rain and temperature to the model (model 5).

```
##
## Call:
## lm(formula = profits ~ staff + advert + temperature + new_release +
##     rain + temperature * rain, data = profits)
##
## 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 ***
## new_releaseY       -537.830    267.736  -2.009   0.0454 *
## rainY              -389.101    201.631  -1.930   0.0546 .
## 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
```

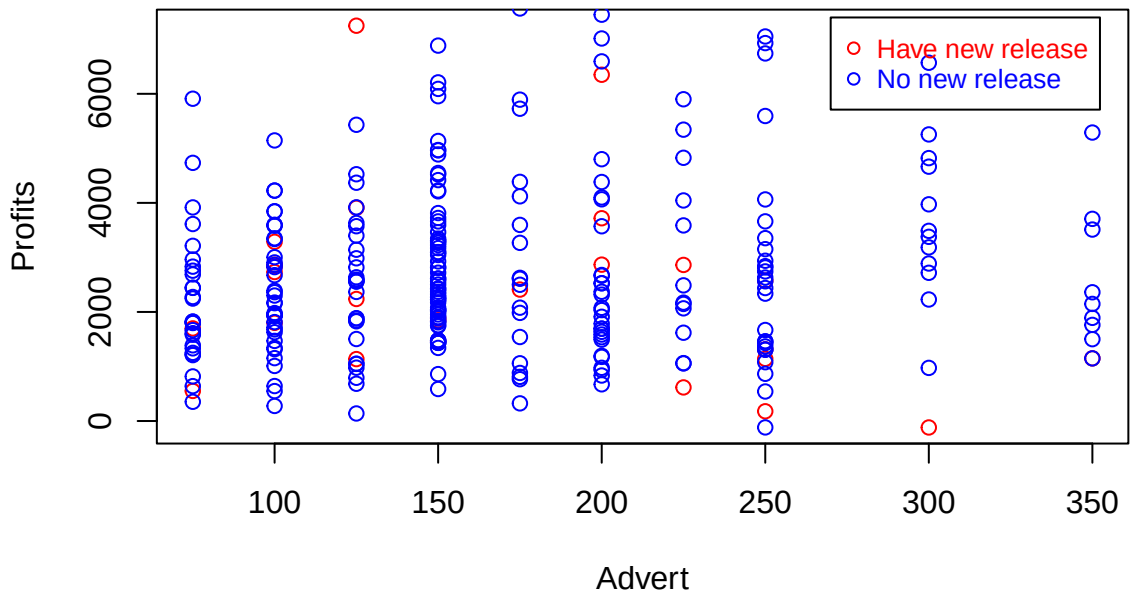
We can see that the coefficient of determination has increased to 0.5388, showing that the new model can explain more variability in the response.

Again, a F-test was carried out to compare the model with the previous one.

```
## Analysis of Variance Table
##
## Model 1: profits ~ as.factor(rain) + advert + staff + new_release
## Model 2: profits ~ staff + advert + temperature + new_release + rain +
##     temperature * rain
##   Res.Df        RSS Df Sum of Sq      F    Pr(>F)
## 1    309 1016855797
## 2    307  517612860  2 499242936 148.05 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value of the F-test is very tiny, suggesting there is evidence that the nested model is not as good as the full model, so we can keep the interaction in our final model.

We also investigated whether there is any interaction between advert and new_release with the following plot.



As in the dataset, the number of days with new release is very small compared to the other days, no pattern can be seen from the plot. We tried to add the interaction to model 3, the model using the original covariates in the dataset except weekend and year. The p-value for new_release becomes 0.993 and the p-value for the interaction dummy variable is 0.391. Thus it is not considered a good idea to add the interaction to our final model.

Regarding model 5, we have observed an issue where some covariates have relatively high standard errors, for instance, `rain_Y` and `new_release`. Therefore, it is necessary to check for collinearity by VIF:

```
##            staff           advert      temperature      new_release
##         1.036068         1.003730         2.015716         1.013871
##             rain temperature:rain
##         1.723219         2.141992
```

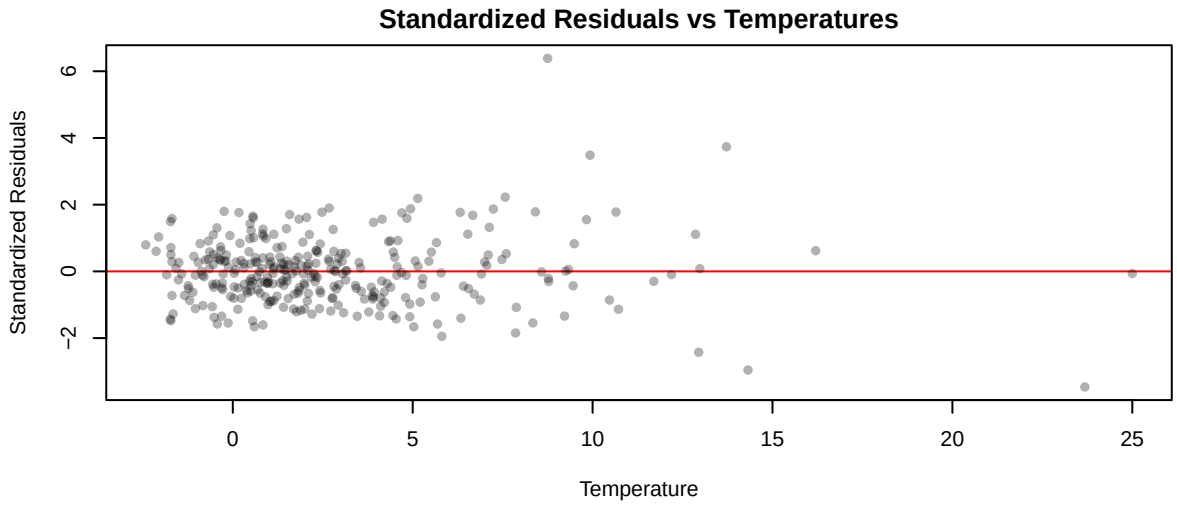
The results indicate that our choice of variables satisfies the conditions for collinearity. The high standard errors could be due to model selection or the nature of the data itself, and these points will be further elaborated in the **Discussion of Limitations**.

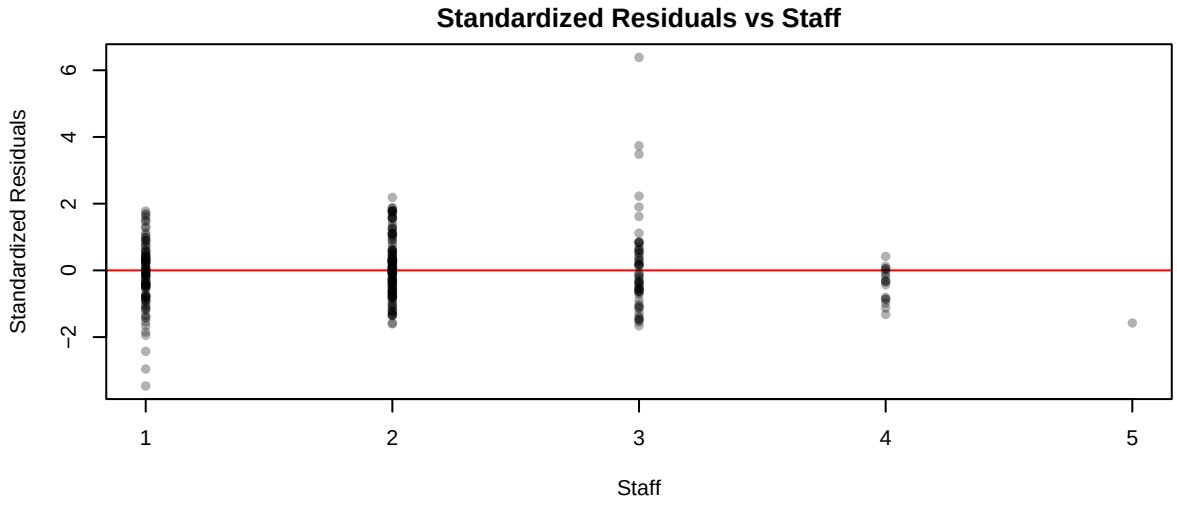
As a result, we chose model 5 as our final model.

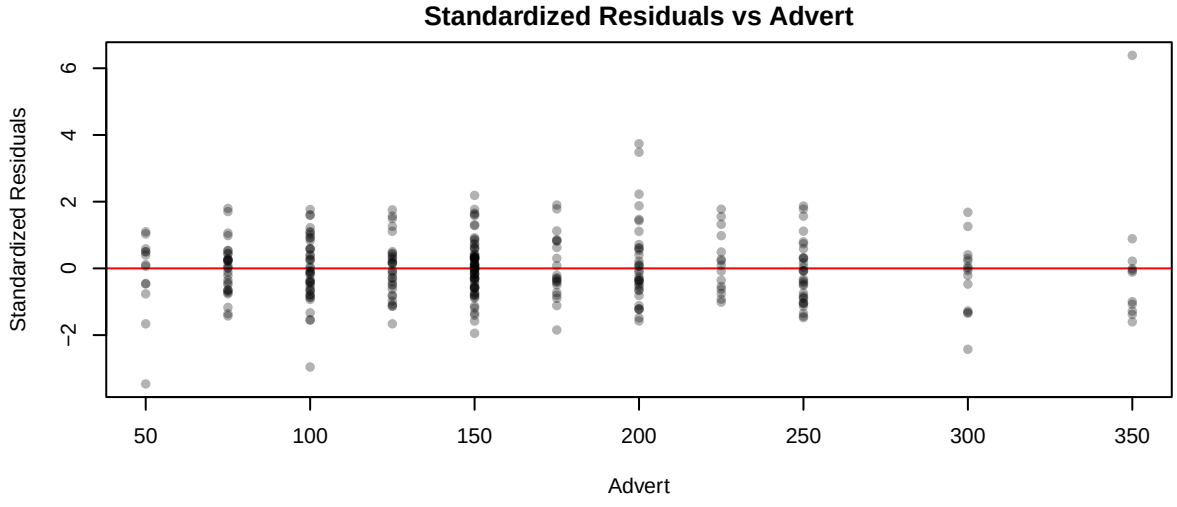
Check assumptions of our final model

Linearity

Linearity is checked by plotting the standardised residuals against each numeric covariate.

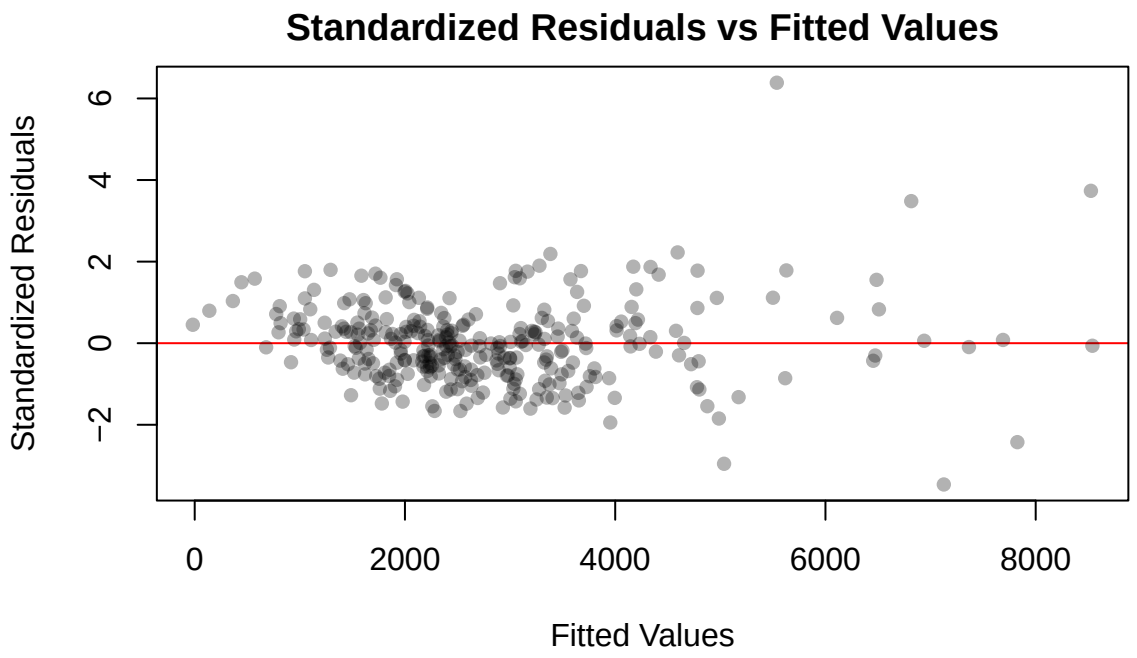






From the exploratory analysis and the plots above, since in all three plots, the standardized residuals scatter around zero with no systematic pattern, the linearity assumption is valid.

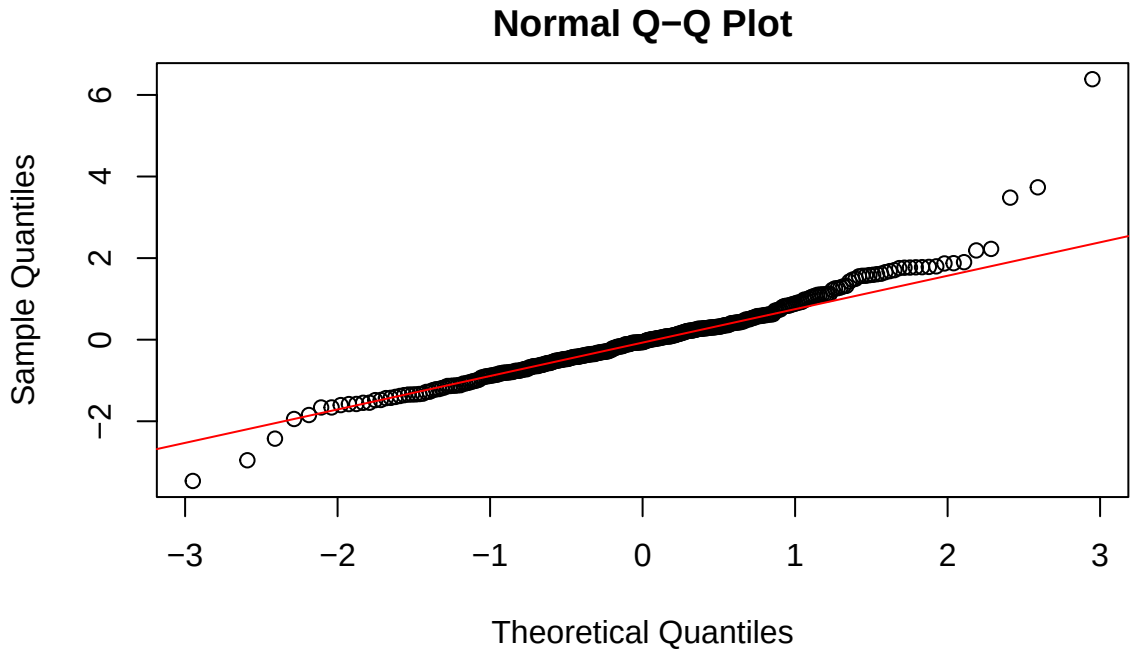
Homoscedasticity



By inspecting the plot of standardized residuals against fitted values, we can see that the standardized residuals roughly cluster around 0 across different fitted values and fall in the range from -2 to 2. However, there is a slightly potential non-linear pattern that can be observed. Therefore, the homoscedasticity assumption is generally satisfied but further close investigation into the pattern could be taken.

Normality

By checking the model Q-Q plot, we can see the majority of standardized residuals fall on the Q-Q line that represents the theoretical normal distribution quantile, which indicates the normality assumption is valid.



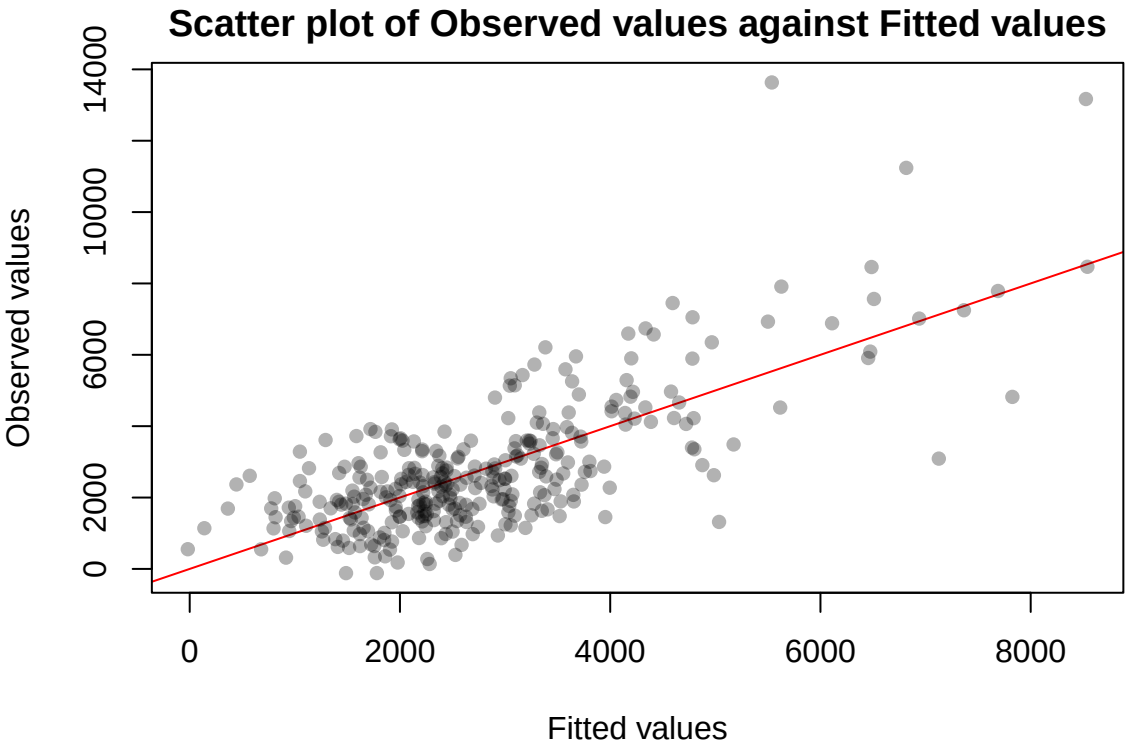
Independence

For independence, the data provided are randomly selected with no time order or details about its collection

process, so we cannot check the serial correlation here. Since we have not learned tests to check independence other than serial correlation, we just assume that independence is satisfied.

Check model fit

The fitness of the final model is checked by plotting the observed values against fitted values.



As we can see by the plot, most points scatter around the line y=x, this implies a generally good fitting of our model to the dataset.

Conclusion

In conclusion, the regression analysis provides us an insight into how each variable influences the daily profits. By looking at the estimated coefficients, it seems that advertising expenditure (advert), staff number and temperature have a positive impact on profits, while the introduction of new releases and rain negatively influence the profits.

Notably, increased advertising expenditure (advert) and a higher staff count contribute positively to profits, with coefficients of 4.351 and 393.731, respectively. That is, when other covariates are fixed, increasing one GBP spend on advertisement leads to a 4.351 GBP increase in profits and increasing one staff leads to a 393.731 GBP increase in profits.

Conversely, the introduction of new releases has a negative impact on daily profits, as indicated by the coefficient of -537.830, which means when other covariates are fixed, new release leads to a 537.830 GBP decrease in profits. This suggests that while new releases may bring potential revenue, other factors may offset the anticipated gains.

Weather variables also play a significant role. Higher temperatures positively influence profits with a coefficient of 246.574, while rain has a negative impact with a coefficient of -389.101. The interaction between temperature and rain further complicates the picture, with a positive coefficient of 204.387. This suggests that on rainy days, when other covariates are fixed, profits may potentially increase more as temperature increases.

Discussion of limitations

1) The model might not encompass all critical factors influencing daily profits. Variables like competitor strategies, macroeconomic policies, or specific market trends, not included in the model, could substantially contribute to profit variations.

2) The model assumes a linear relationship between independent variables and daily profits. If the actual relationships are non-linear, the model's predictions may deviate from the true profit values, potentially leading to inaccurate insights.

3) With only 314 data points, the dataset size may be insufficient for training the model. The model's ability to generalize to new data could be compromised, especially in capturing the complexity of real-world profit dynamics.

4) Linear models may not adequately capture complex nonlinear relationships inherent in real-world scenarios. Exploring more advanced machine learning or deep learning approaches is a good way to enhance the predicting ability.

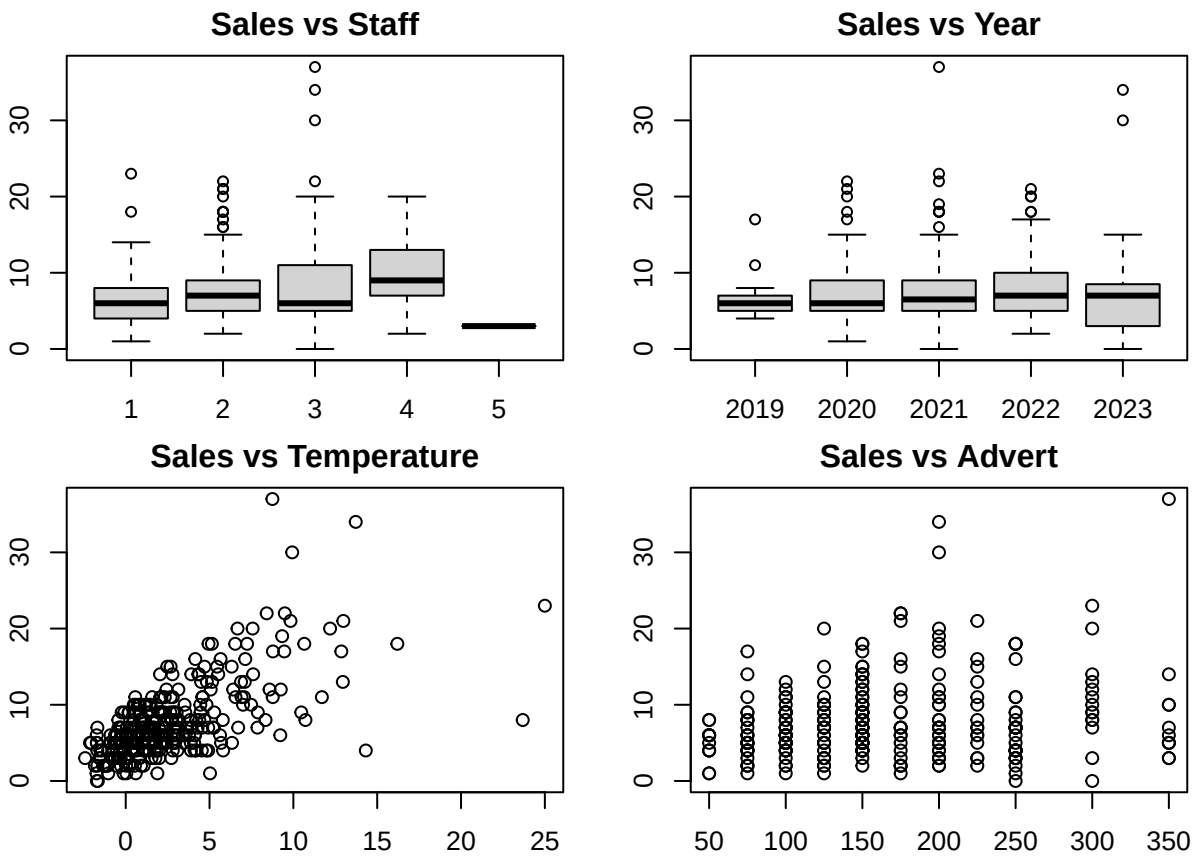
5) The absence of a split between a training set and a validation set poses a risk of overfitting. Without a separate validation set, the model might memorize the training data rather than learning underlying patterns, potentially compromising its performance on new, unseen data.

6) The data may contain outliers or leverage points. Particularly in smaller datasets, even a few influential points can significantly distort estimates of coefficients and standard errors. Therefore, it is crucial to conduct detection and apply appropriate remedies, such as removing outliers or utilizing robust regression.

Part 2: Generalised linear model OR Generalised additive model

Report on modelling number of car sales

To consider the number of car sales made in the showroom each day, we start with a simple EDA. To select the most suitable covariate for a one-covariate regression model, we initially excluded three yes-or-no categorical variables: `new_release`, `weekend`, and `rain`, as splitting the data into just two categories usually does not yield a good prediction. For the four numerical covariates, we assess their relationship with sales using scatterplots and boxplots:

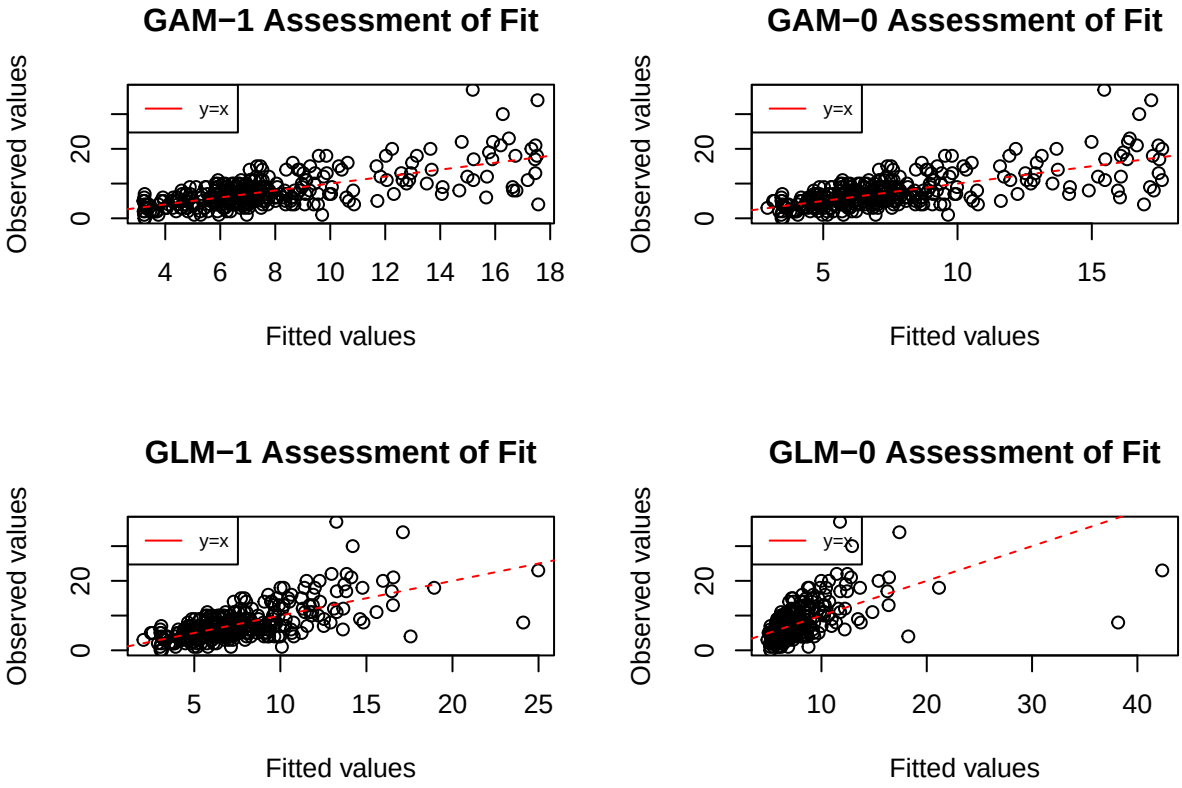


In the two boxplots, `Sales vs. Staff` and `Sales vs. Year` respectively, and the scatterplot on the bottom-right `Sales vs. Advert`, we can assert that the relationship between sales and these three covariates does not exhibit a systematic pattern. The data points are rather evenly distributed across most levels, except for a few levels where the number of observations is too small to form a definitive trend. Despite some fluctuations, it appears there is no clear trend between the number of sales and either the number of staff or advert spending, and there is no evident time trend or seasonal effect concerning the year. This suggests that these covariates may not be strong predictors for the number of car sales on their own.

Regarding the relationship between sales and temperature in the bottom-left plot, it's apparent that in the lower temperature range (from negative to 10 degrees Celsius), sales increase significantly with rising temperatures. As temperatures continue to rise, the increment in sales with each additional degree Celsius begins to diminish. This trend resembles the curve of a logarithmic function, leading to the suspicion that sales may have a logarithmic relationship with temperature. Moreover, considering the positively skewed distribution of temperature, it seems worthwhile to attempt a log-transformation of temperature to make data points more horizontally even.

Therefore, we decide to use temperature as the final covariate.

Based on the nature of the count data, we decided to use a Poisson model with a log-link for our analysis. We evaluated four different models: GAM without transformation (GAM-0), GAM with transformed temperature (GAM-1), GLM without transformation (GLM-0), and GLM with transformation (GLM-1). The transformation applied was $log(temperature + 3.42)$ to accommodate negative temperature values.



Initially, we assessed model fit by examining Observed sales vs. Fitted sales plots above for each model. The assessment plot for GLM-0 showed that the data points were furthest from the y=x line, indicating the poorest fit among the models. Then when comparing deviance across the four models, we concluded:

$$D_{GLM-0} >> D_{GLM-1} > D_{GAM-0} \approx D_{GAM-1}$$

We then conducted ANOVA tests for the two GAM models to confirm whether incorporating nonlinear effects considerably improved the model.

```
## Analysis of Deviance Table
##
## Model 1: sales ~ temperature
## Model 2: sales ~ s(temperature)
##   Resid. Df Resid. Dev     Df Deviance  Pr(>Chi)
## 1    312.00     559.58
## 2    306.41     456.94 5.5911   102.64 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Analysis of Deviance Table
##
## Model 1: sales ~ temperature
## Model 2: sales ~ s(temperature)
##   Resid. Df Resid. Dev     Df Deviance Pr(>Chi)
## 1    312.00     479.68
## 2    305.91     455.07 6.0872   24.613 0.000433 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

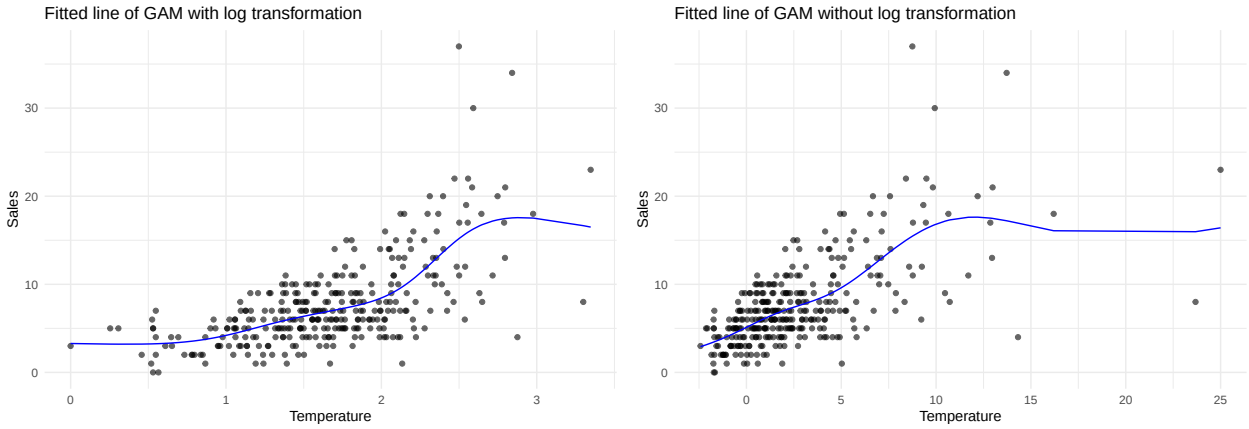
The very small p-value from the chi-square test demonstrates that adding smooth functions and nonlinear effects significantly enhanced our model. Thus, we decided to use GAM approach.

Upon comparing GAM-0 and GAM-1 in terms of R^2, deviance explained, and GCV score, we found that these metrics were very similar, making it difficult to distinguish between the two.

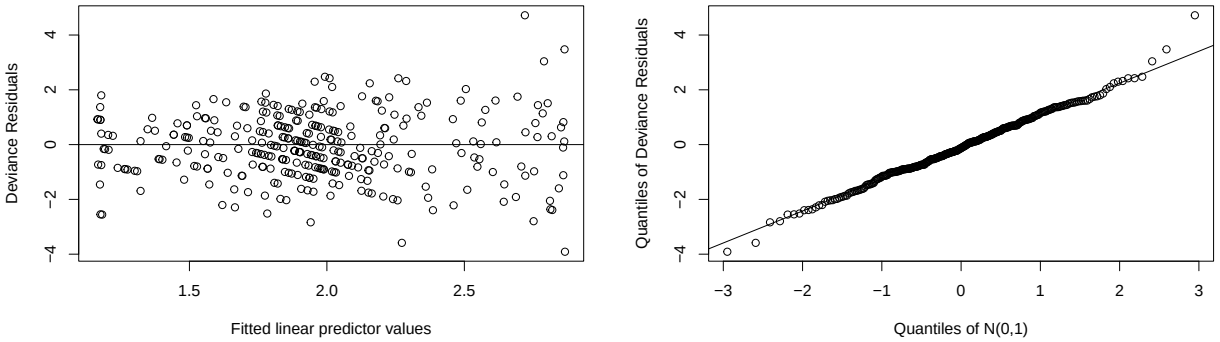
Table 1: Comparison of GAM Models Metrics

Model	Rsquared	DevianceExplained	GCVscore
GAM-0	0.4609610	0.4898826	0.4966518
GAM-1	0.4599649	0.4919725	0.4932601

The deciding factor came from examining the fitted lines on the Observed Sales vs. temperature plots. The plot for GAM-1 displayed a more horizontally even distribution of data points, whereas the GAM-0 plot showed a dense clustering of points at lower temperatures and a sparse distribution at higher temperatures. This resulted in a fitted straight horizontal line between 15 to 25 degrees, leaving this range almost devoid of data points. Due to this drawback, we ultimately preferred the GAM with the log-transformed temperature.



Considering the assumptions for our model, we can check normality and constant variance of deviance residuals from the plots below:



The assumption of homoscedasticity is satisfied because the Deviance Residual vs. Fitted Values plot above shows a random scatter evenly distributed around zero with no systematic pattern.

For normality, aside from a few outliers in the tails, the majority of points in the central portion of the plot adhere closely to the theoretical quantiles of $N(0, 1)$, which satisfies the assumption of normality.

For independence, the data provided are randomly selected with no told order or detail about its collection process, so we do not need to check for serial correlation here and can just assume independence is satisfied.

Thus, our GAM model satisfied all the assumptions required. As mentioned in model selection before, our model has a good fit as well, which can be concluded from Observed sales vs. Fitted sales plot, R^2, deviance explained, GCV score, ANOVA and fitted line on observed data points.

In general, car sales appear to be positively correlated with the temperature of the showroom in our model. Although we cannot claim that an increase in temperature causes an increase in sales, the management team might experiment with enhancing the showroom temperature, such as turning on the heating, to improve the customer experience during selection, and observe if this leads to an increase in sales. At the same time, although our GAM model fits quite well, a model using only temperature as a covariate is unlikely to perform well in sales prediction. This can lead to predictions that are overly sensitive to minor fluctuations in temperature and may increase the residuals due to the neglect of effects from other potential covariates, reducing accuracy. Therefore, the management team should consider using a multivariate model for more accurate prediction purposes.

Total word count: 2970