Regression analysis of car showroom profit and sales

Justin Lo

2024-01-24

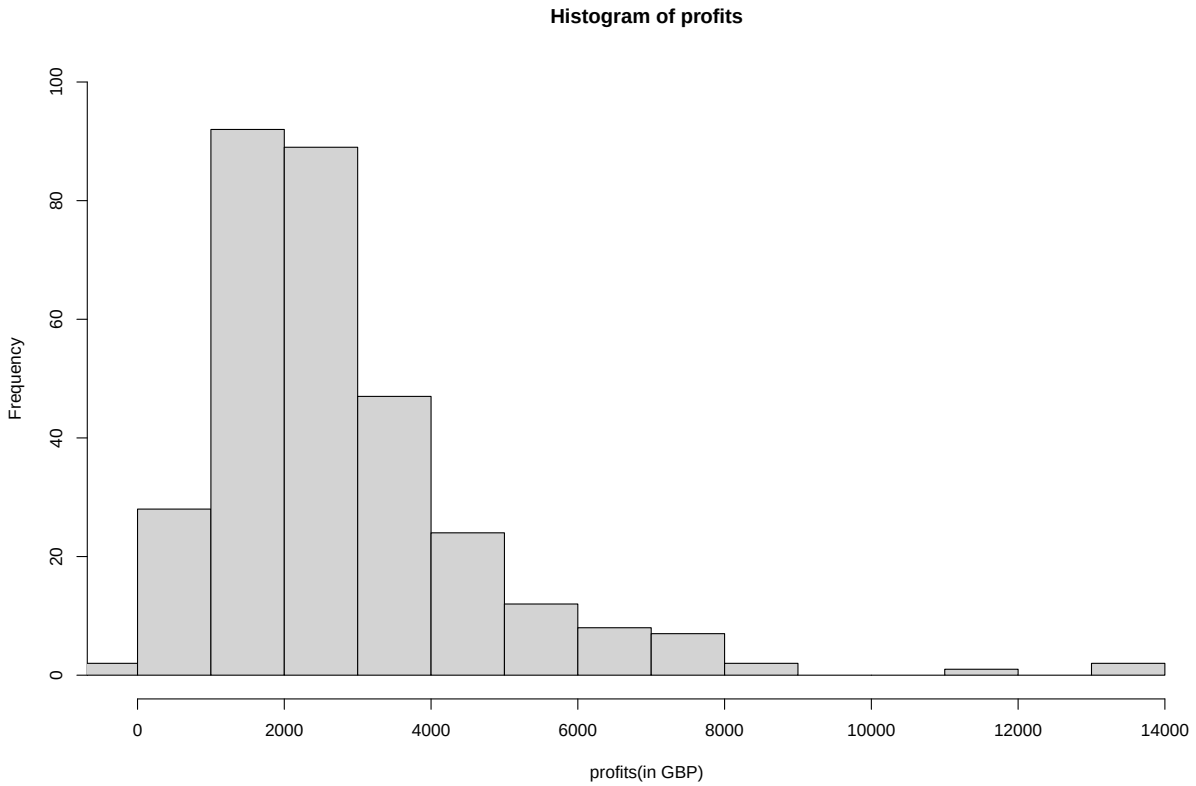
Part 1: Normal linear model

```
profits <- read.csv("profits.csv")
head(profits)
```

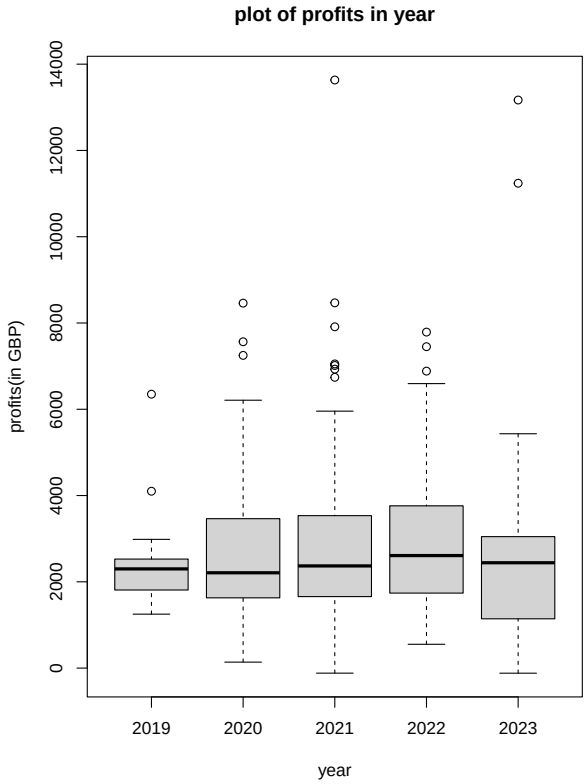
```
##    profits staff advert new_release weekend temperature rain year
## 1  2562.57     1    250           N       0       -0.34    Y 2023
## 2  1800.83     3    100           Y       0        2.88    N 2020
## 3  2487.57     2    225           N       0        1.98    Y 2023
## 4 13168.57     3    200           N       0       13.72    Y 2023
## 5  4224.64     3    100           N       1        6.41    Y 2023
## 6  1140.39     1    250           Y       0       -2.42    Y 2020
```

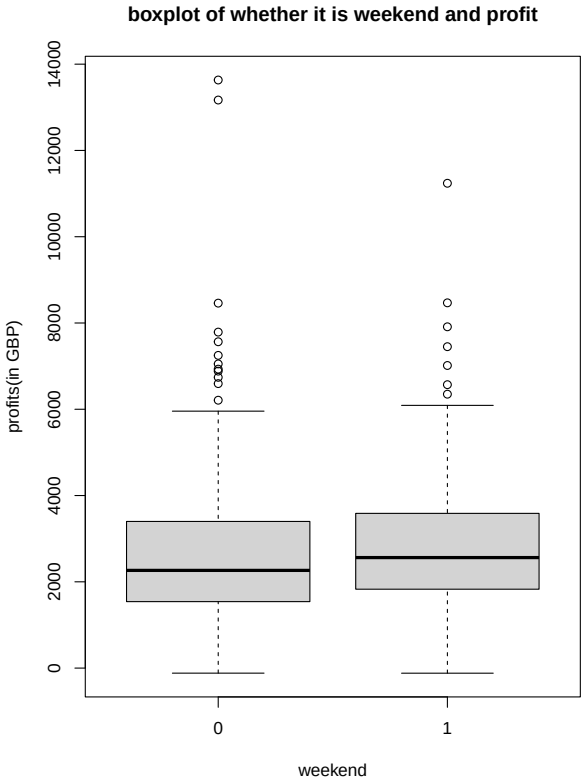
```
profits$year <- as.factor(profits$year)
profits$rain <- as.factor(profits$rain)
```

```
hist(profits$profits, main = 'Histogram of profits', xlab = 'profits(in GBP)', xlim = c(-120,14000), yl
```

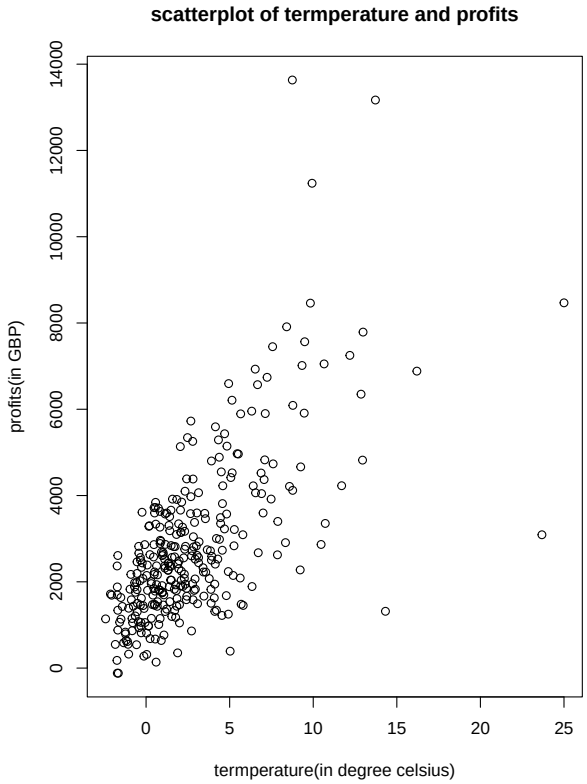


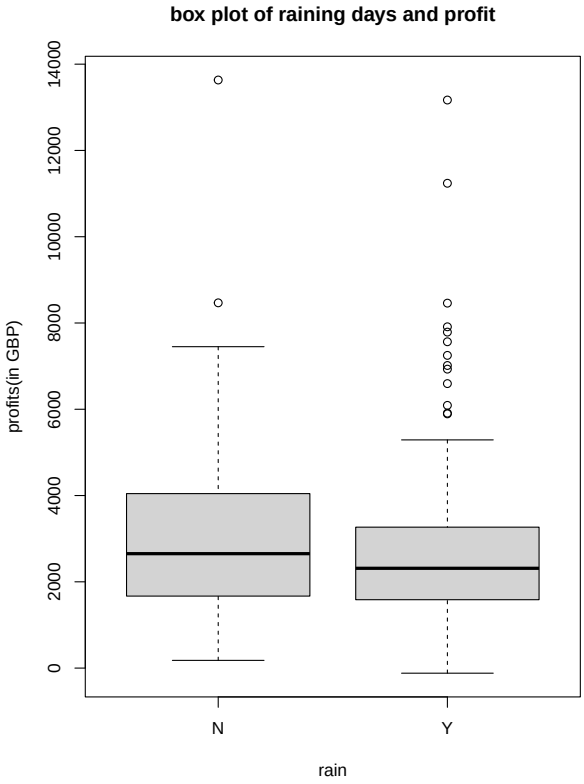

```
par(mfrow=c(1,2))
plot(profits$year, profits$profits, main='plot of profits in year', xlab='year', ylab='profits(in GBP)'
plot(as.factor(profits$weekend), profits$profits, main='boxplot of whether it is weekend and profit', xl
```



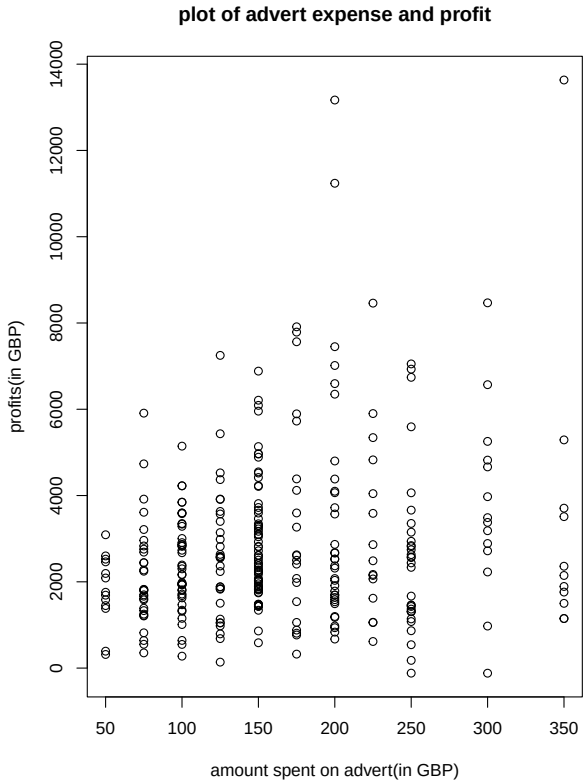


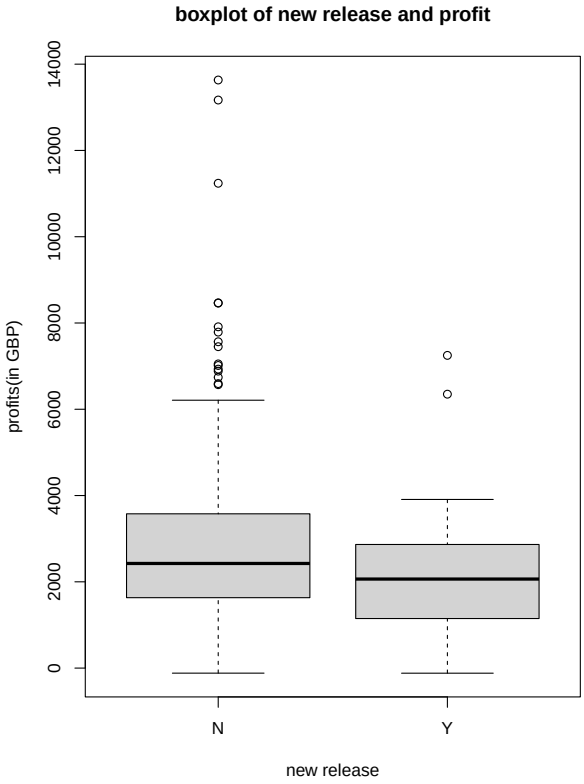

```
par(mfrow=c(1,2))
plot(profits$temperature, profits$profits, main = "scatterplot of termperature and profits", xlab = "te
plot(as.factor(profits$rain), profits$profits, main = 'box plot of raining days and profit', xlab = 'ra
```






```
par(mfrow=c(1,2))
plot(profits$advert, profits$profits, main = 'plot of advert expense and profit', xlab = 'amount spent
plot(as.factor(profits$new_release), profits$profits, main='boxplot of new release and profit', xlab='n
```





Introduction to the data

Starting off this project with an exploratory analysis, I will introduce the overview of the dataset and the main aim of this analysis.

There is a total of 8 columns and 314 rows(observations). As seen from the first five rows of the dataset, this dataset contains information about the profits made from a car showroom and relevant information on that day. The relevant information includes the number of staff, the amount spent on the advert in the previous seven days if the car is a new release model, if it is a weekend, the temperature of the day if it is raining and also what year that day is in. In which there are no NA values. Of 314 observations of profits, two were negative; the lowest value was -118.33, the highest value was 13632.8, and the mean was 2798.796. The histogram shows that the profit value mainly lies in the range of 1500 - 4000. The main aim of this analysis is to explore the factors that affect the amount of profit.

Observations are in the 2019 - 2023 year range. Eighteen of them are in 2019, 78 of them are in 2020, 92 of them are in 2021, 83 of them are in 2022, and 43 of them are in 2023. 2022 sees the highest mean profit value of 2982.84, and 2019 sees the lowest mean profit of 2429.203. Profits remain relatively stable throughout the years, as seen from the boxplot, but the management team should review the year 2022 and see what they have done well.

Furthermore, weekend observations see higher mean profit than weekdays for weather conditions. The lowest temperature in the observations is -2.4200, the highest is 25.0, and the mean is 2.72. The 4x4 scatterplot shows a relationship between temperature and profits; higher temperature leads to higher profits and vice versa. Eighty-nine observations see rain, and 225 see no rain. However, the boxplot shows that the mean values of non-rainy and rainy days are similar; further regression analysis is needed to see whether rain affects profit. 1 to 5 staff members worked at the showroom on the day of observation; the mean number of staff members was 1.97. Intuitively thinking, there should be a positive relationship, as more staff can get more work done, so there should be a higher profit. However, from the 4x4 scatterplot, no distinct pattern can be seen.

Advert expenses range from 50-350. As seen from the plot, there is a vague positive relationship where the

profits increase when the advert increases. Further analysis should be done to see its impact on profits.

Model building

I would like to approach the model building phase by starting off with a model that includes all the given variables. Then, determine which to keep. After that, I would consider adding interactions to better capture the relationship.

I would employ backward elimination to strike a balance between complexity and performance. Removing insignificant variables reduces model complexity, which can help avoid overfitting and improve model interpretability. Each time, one variable would be removed to make the elimination process more systematic. Variables that do not contribute significantly to the model's explanatory power can be identified and removed step by step.

```
profits$weekend <- ifelse(profits$weekend == "0", "no", profits$weekend)
profits$weekend <- ifelse(profits$weekend == "1", "yes", profits$weekend)
model_1<-lm(profits~staff + advert + new_release + weekend + temperature + rain + as.factor(year), data=
summary(model_1)
```

```
##
## Call:
## lm(formula = profits ~ staff + advert + new_release + weekend +
##     temperature + 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 ***
## new_releaseY          -557.909    283.299  -1.969   0.0498 *
## weekendyes              35.791    173.099   0.207   0.8363
## temperature            338.804     21.589  15.694  < 2e-16 ***
## 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
```

model_1 has profits as the response variable. Staff, advert, temperature as numerical covariates and new_release, weekend, rain, year as categorical covariates. With not new release, not weekend, no rain and year 2019 as the reference category.

For interpretation of model_1. The coefficient of intercept is 211.298, meaning the expected value of profit is 211.298 when the coefficient of numerical covariates is 0 and that it is the case of reference category for

the categorical covariates. For the interpretation of the numerical covariates, taking staff as an example. The coefficient of staff is 448.873, meaning there is an expected increase of 448.873 in profit when there is a one unit increase in staff. For the interpretation of the categorical covariates, taking rain as an example. The reference category is not raining and the coefficient of rain is 204.815, meaning that there is an expected 204.815 difference in the value of profit between no raining and raining considering other numerical covariates coefficient is 0 and the other categorical category takes the reference category.

model_1 has an r-squared of 0.5041. I will now start determining which covariates to keep. From the summary of model_1, staff, advert, and temperature have low p-values thus, they are statistically significant in explaining the profit variance. I would keep the three covariates. I went through the step-by-step process of eliminating variables that do not contribute much to the model's explanatory power(reflected by the high p-value); the two removed variables are weekend and year.

To support the decision, the p-values of weekend and year are very large, indicating that they are highly statistically insignificant and that there is no strong evidence against removing the two variables. Also, there is a lack of trend looking from the plots of year and weekend against profits that support a relationship. Therefore, the two variables from the model are eliminated.

I kept rain in the model despite its high p-value due to the interaction I would like to consider later in the development. It is a better practice to keep the covariate if the interaction between that covariate and another is in the model.

```
model_2<-lm(profits~ staff + advert + new_release + rain + temperature , data=profits)
summary(model_2)
```

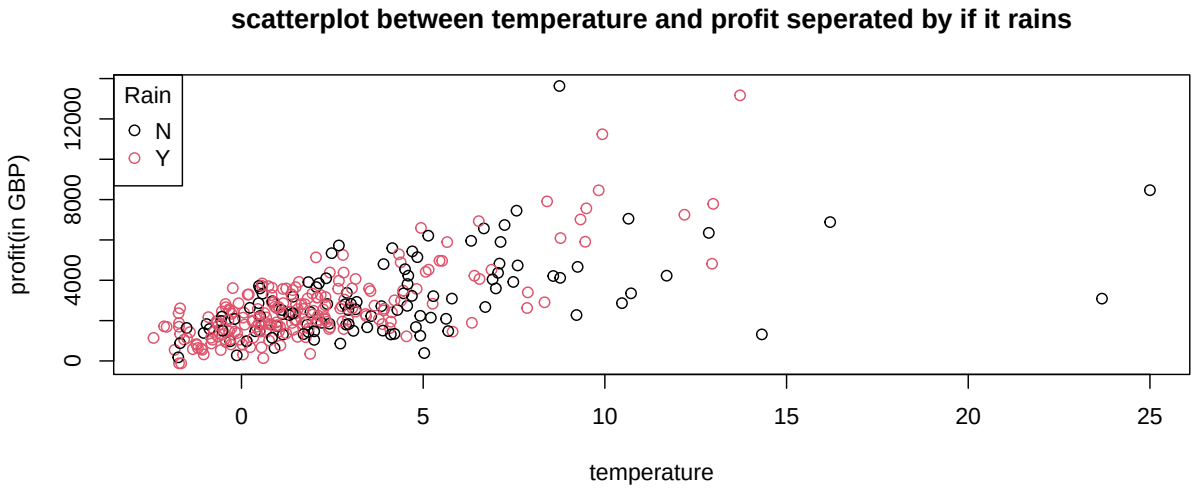
```
##
## Call:
## lm(formula = profits ~ staff + advert + new_release + rain +
##     temperature, data = profits)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5813.8  -773.3   -83.0   600.7  7597.4
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   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 ***
## new_releaseY -577.030    277.482  -2.080   0.0384 *
## rainY         204.273    167.385   1.220   0.2233
## temperature   339.863     21.368  15.905  < 2e-16 ***
## ---
## 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
```

From the summary of model_2, we can see an increase in adjusted r-squared, which is r-squared adjusted for the number of predictors. That indicates a higher proportion of variance in profit is explained by the model, an increase in fit of the model

Now, I would explore the possible dependencies between variables. I have two intuitive thoughts which I would like to show by plotting them. The first is whether the effect of advert expenses is more significant

on newly released models than on older models. Second is whether the impact of temperature on profits is different on rainy days compared to non-rainy days.

```
par(mfrow=c(2,1))
plot(profits$temperature, profits$profits, col=as.factor(profits$rain), main = "scatterplot between temp
legend("topleft", legend = levels(as.factor(profits$rain)), col = 1:2, pch = 1, title = "Rain" )

plot(profits$advert, profits$profits, col=as.factor(profits$new_release), main = "scatterplot between ad
legend("topleft", legend = levels(as.factor(profits$new_release)), col = 1:2, pch = 1, title = "New rele
```



From the plot of temperature and profit separated by if it rains, it can be seen that rainy days generally have a lower temperature. Secondly, profits are low if it rains and has a low temperature. Hence, the effect of temperature on profit on rainy and non rainy days should be assessed and considered in the model. This could add insights into the effect of the weather conditions on consumer behavior.

Whereas, for the plot between advert and profit separated by if it is a new release, it does not show any significant pattern. Also, the model that includes the interaction between advert and new release shows poor explanatory power and the p-value of the coefficient of the interaction is very large. Hence, The interaction will not be considered.

```r
model_3<-lm(profits~ staff + advert + new_release + rain + temperature + temperature*rain, data=profits)
summary(model_3)
```

```
##
## Call:
## lm(formula = profits ~ staff + advert + new_release + rain +
##     temperature + 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 ***
## new_releaseY      -537.830    267.736  -2.009   0.0454 *
## rainY             -389.101    201.631  -1.930   0.0546 .
## temperature        246.574     28.025   8.798  < 2e-16 ***
## rainY:temperature  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
```

model_3 sees a further increase in r-squared, indicating the increase in the model's ability to explain the profit's variance.

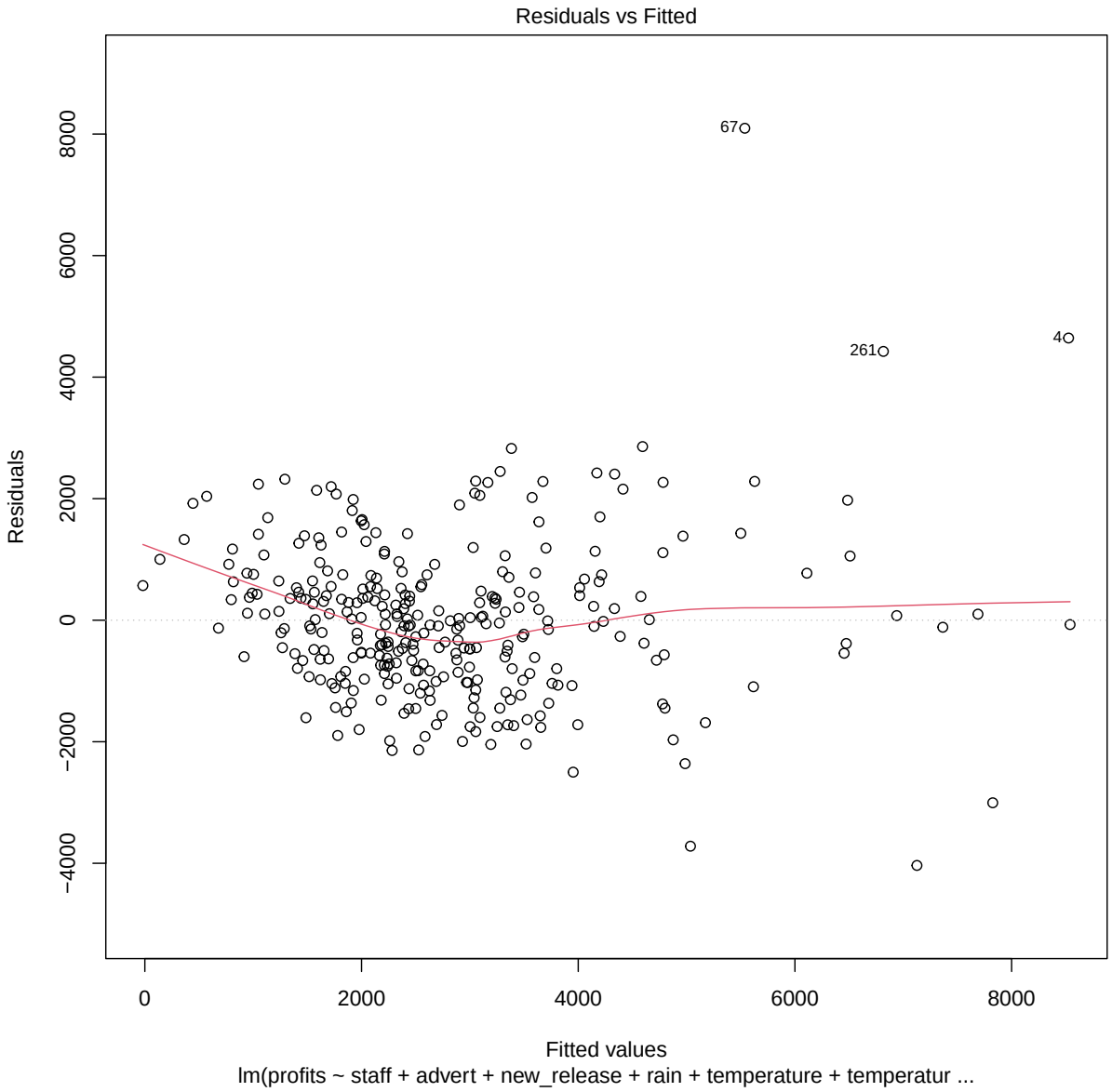
The coefficient of the interaction is 204.387, indicating that the effect on profit of a one unit change in temperature for rainy days is 204.387, holding all else constant.

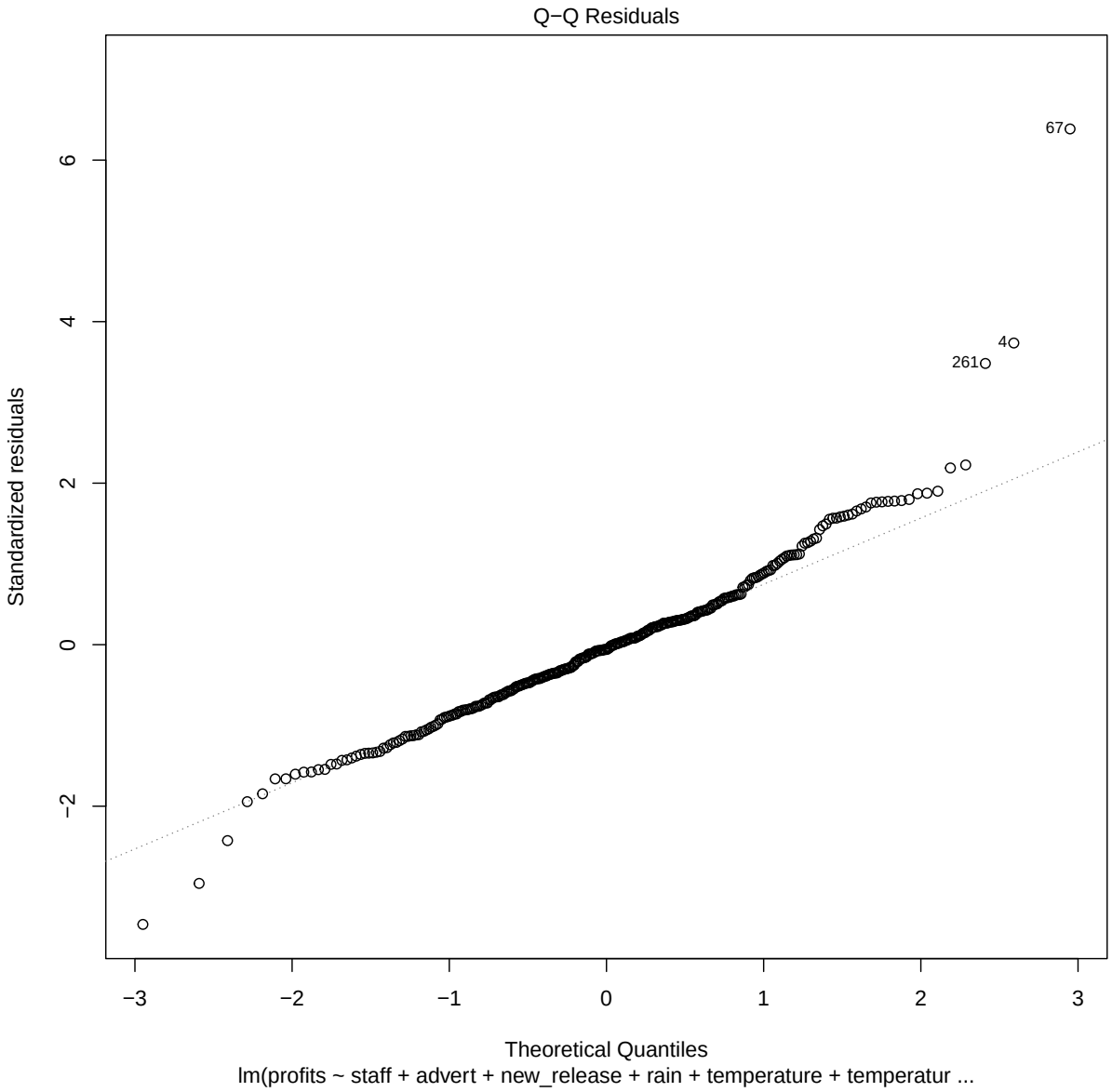
Model checking for final chosen model

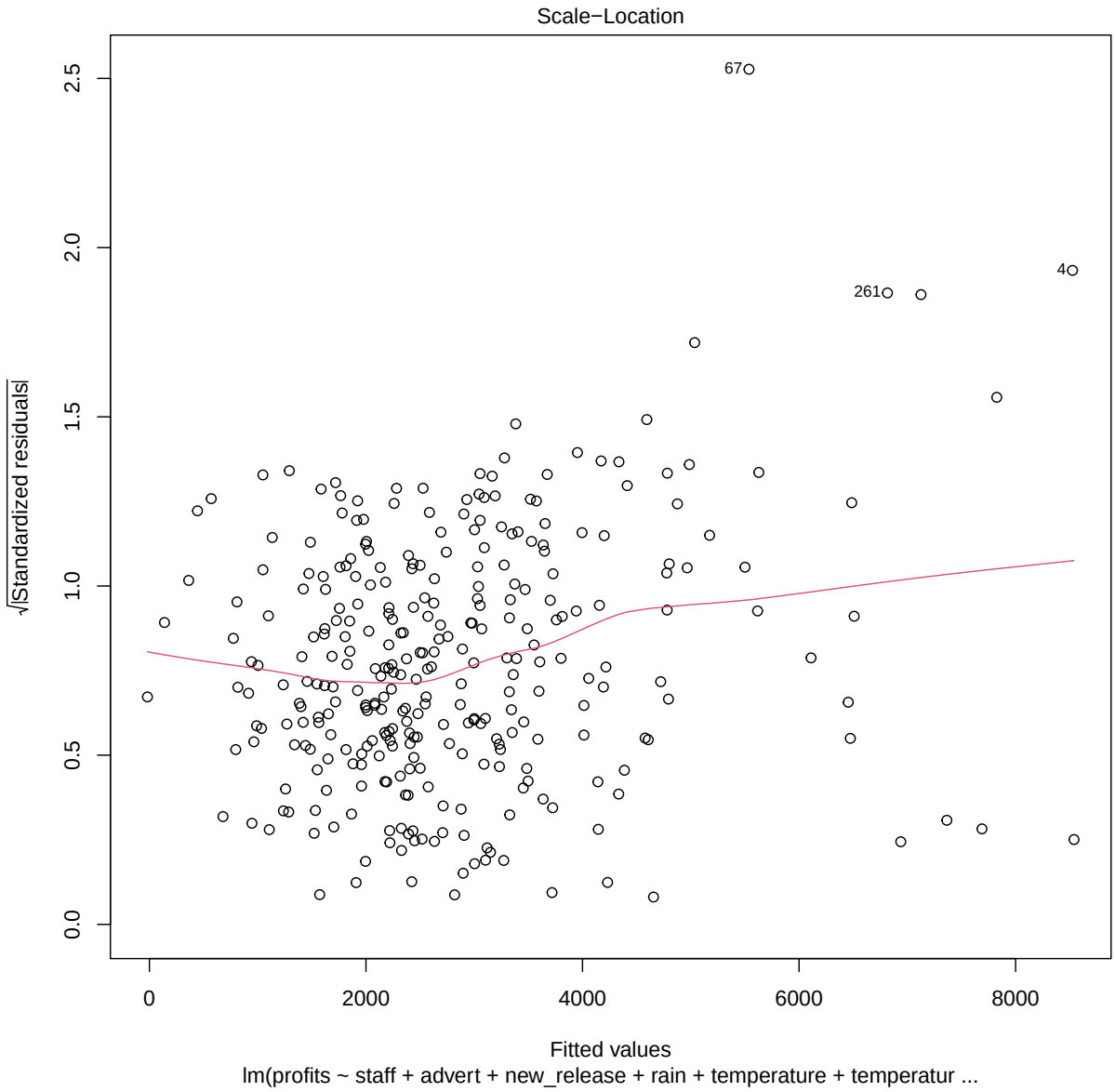
The assumptions are linearity between dependent and independent variables, independence of error, homoscedasticity, and normality of residuals.

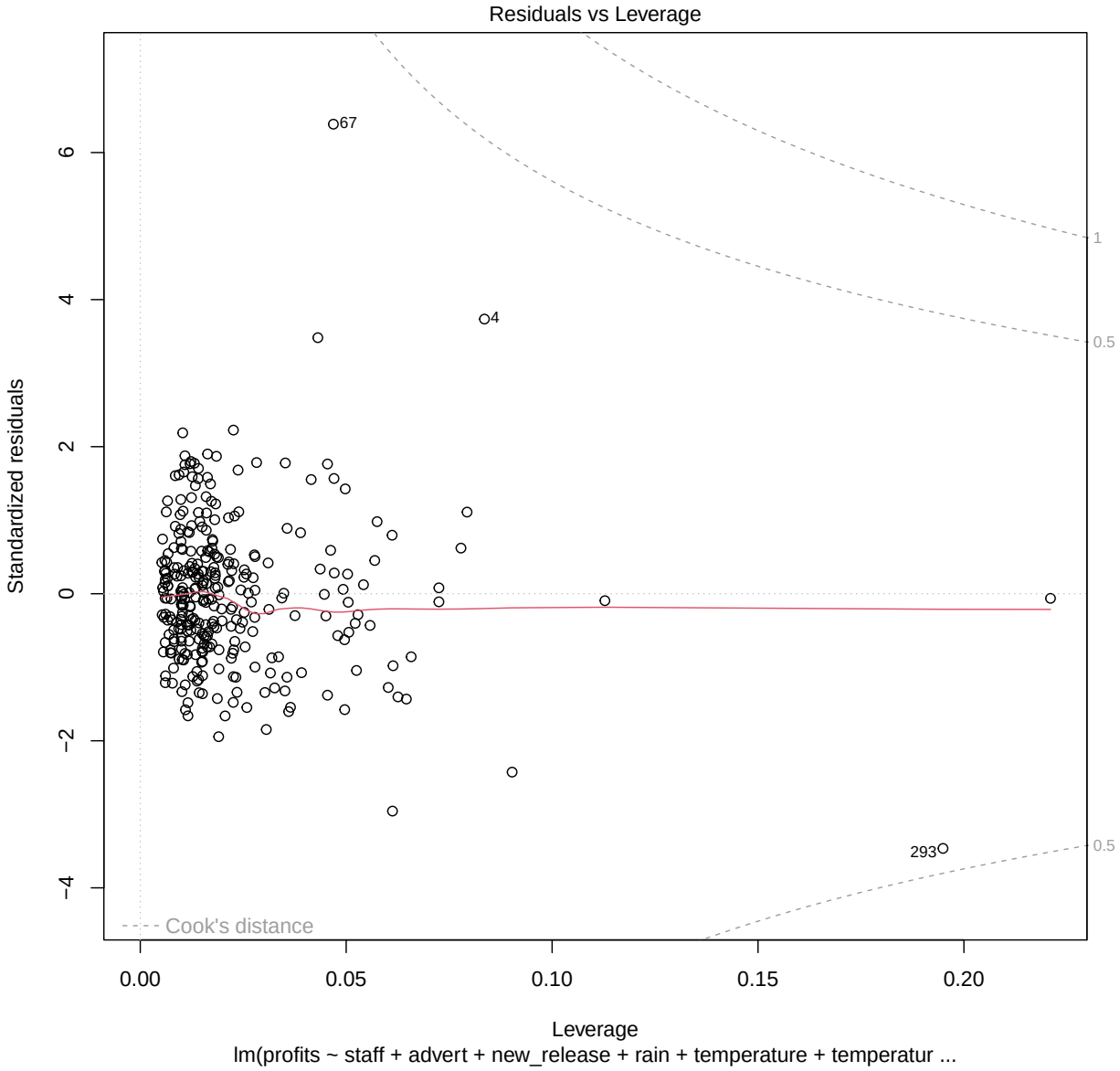
Here are some plots for assessing the assumptions.

```r
plot(model_3)
```


Residuals vs Fitted
Residuals
Fitted values
lm(profits ~ staff + advert + new_release + rain + temperature + temperatur ...
67
261
4
8000
6000
4000
2000
0
−2000
−4000
0
2000
4000
6000
8000


Q−Q Residuals
67
4
261
Standardized residuals
6
4
2
0
−2
−3
−2
−1
0
1
2
3
Theoretical Quantiles
lm(profits ~ staff + advert + new_release + rain + temperature + temperatur ...


Scale−Location
67
261
4
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
0
2000
4000
6000
8000
Fitted values
lm(profits ~ staff + advert + new_release + rain + temperature + temperatur ...



Based on the qq-plot, there is a big deviation on both tails, this means that the error have distribution with a heavier tail than the normal distribution. Indicating a violation of the assumption of normality of errors. Since we are trying to make inferences based on this model, this assumption is crucial for capturing the underlying relationship.

To alleviate the problem, I believe a transformation to the response variable should be made to the response variable, so that the assumption can be met. I have tried different transformation and have decided that a square root transformation fits the data the best.

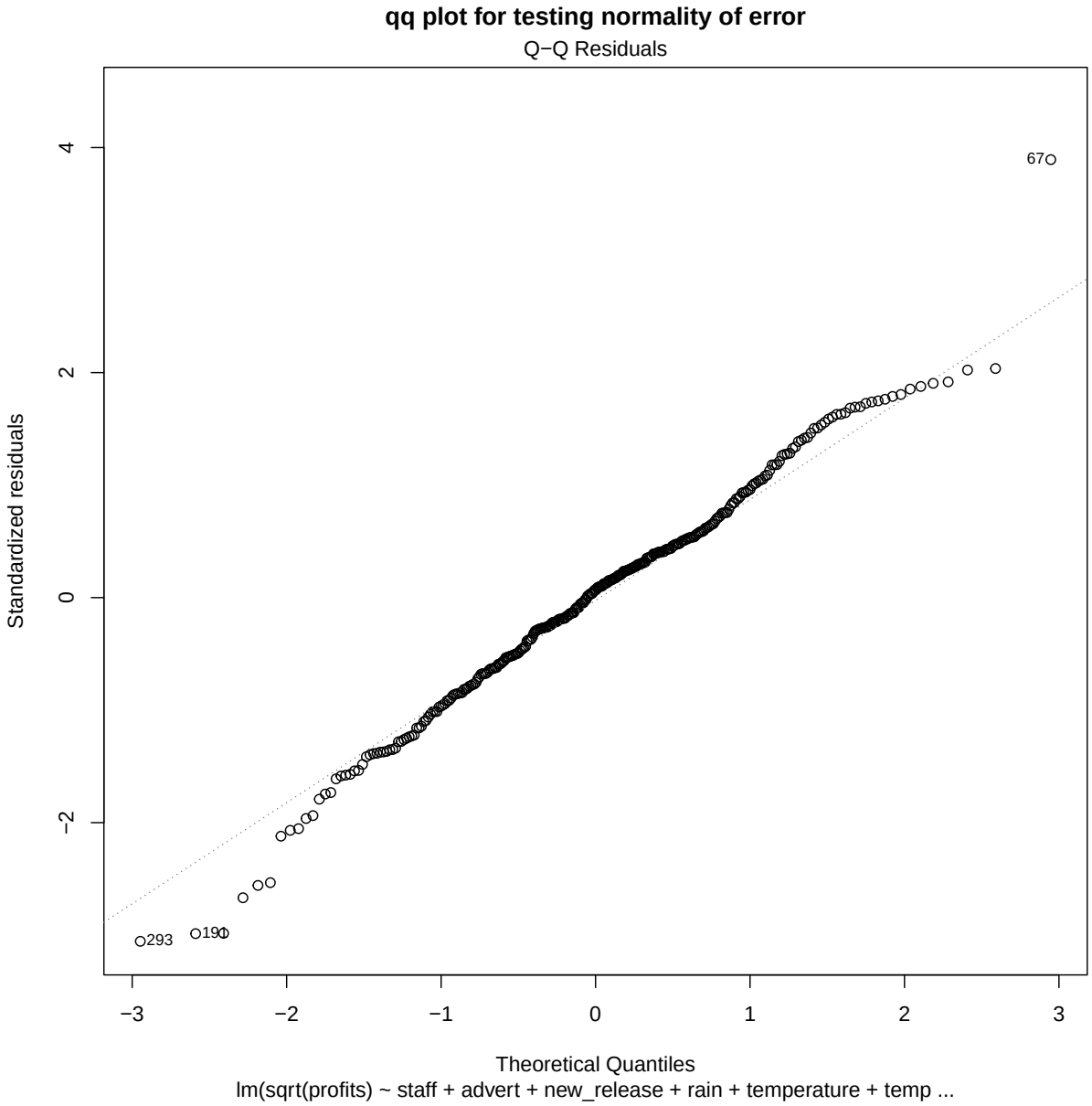
```
model_4<-lm(sqrt(profits)~ staff + advert + new_release + rain + temperature + temperature*rain, data=p
```

```
## Warning in sqrt(profits): NaNs produced
```

```r
summary(model_4)
```

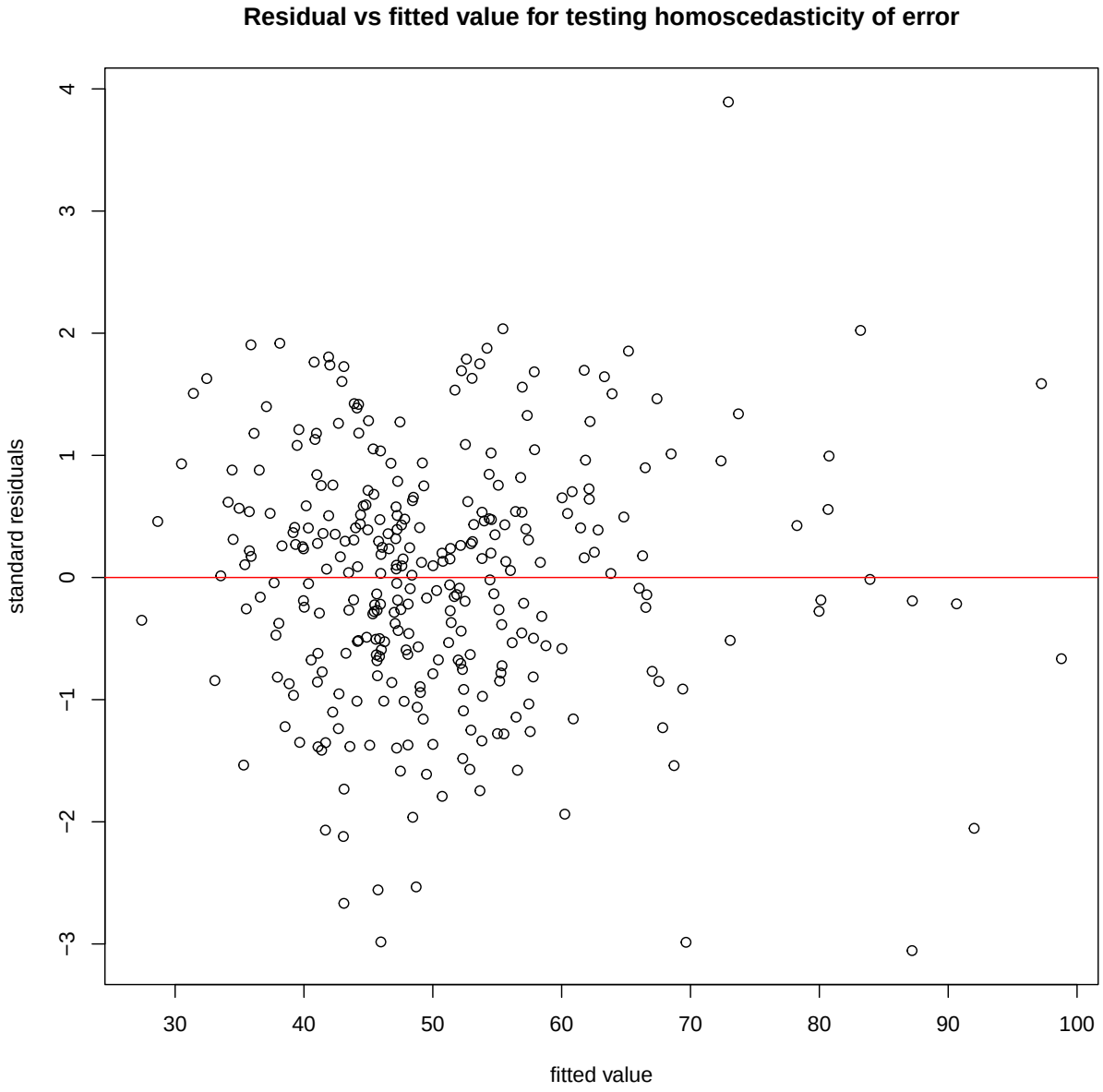
```
##
## Call:
## lm(formula = sqrt(profits) ~ staff + advert + new_release + rain +
##     temperature + temperature * rain, data = profits)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.207  -7.212   0.797   6.611  43.820
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       33.595838   2.592078  12.961  < 2e-16 ***
## staff              2.958042   0.768894   3.847 0.000146 ***
## advert             0.035463   0.009164   3.870 0.000133 ***
## new_releaseY      -4.646662   2.418380  -1.921 0.055614 .
## rainY             -3.156548   1.797733  -1.756 0.080118 .
## temperature        2.063726   0.248927   8.290 3.64e-15 ***
## rainY:temperature  1.641181   0.371450   4.418 1.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.53 on 305 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.5002, Adjusted R-squared:  0.4904
## F-statistic: 50.88 on 6 and 305 DF,  p-value: < 2.2e-16
```

```r
plot(model_4, which = 2, main = 'qq plot for testing normality of error')
```



From looking at the qq-plot, the deviation from the ends of the two tails is not severe. Indicating there is weaker evidence against the assumption of normality of error as compared to model_3.

```
plot(model_4$fitted.values,rstandard(model_4), main = "Residual vs fitted value for testing homoscedasti
abline(h=0, col='red')
```



Now, the other assumptions are to be checked. From looking at the updated plot of residual vs fitted value, the data scatter around 0 randomly, with no visible pattern. That indicates there is no strong evidence against the assumption of homoscedasticity and linearity is violated.

There is a NaNs produced error on model_4, this is due to the fact that square root of a negative value return undefined and there is two negative value in profit. I have updated the dataset by taking the negative profit away. Below is the same model trained on only positive profit.

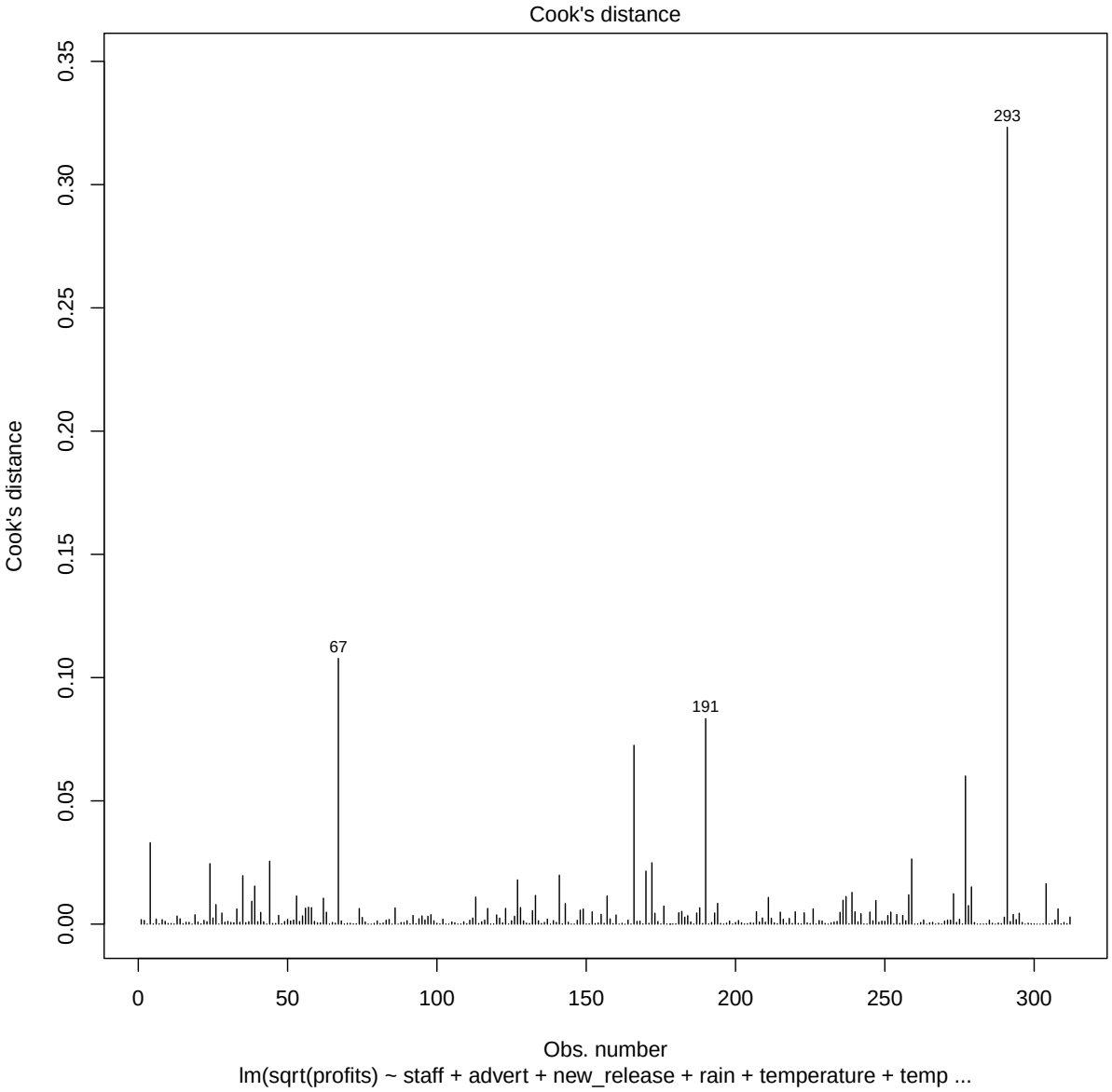
```
profits_2<-profits[-c(137, 219),]
model_5<-lm(sqrt(profits)~ staff + advert + new_release + rain + temperature + temperature*rain, data=p
summary(model_5)
```

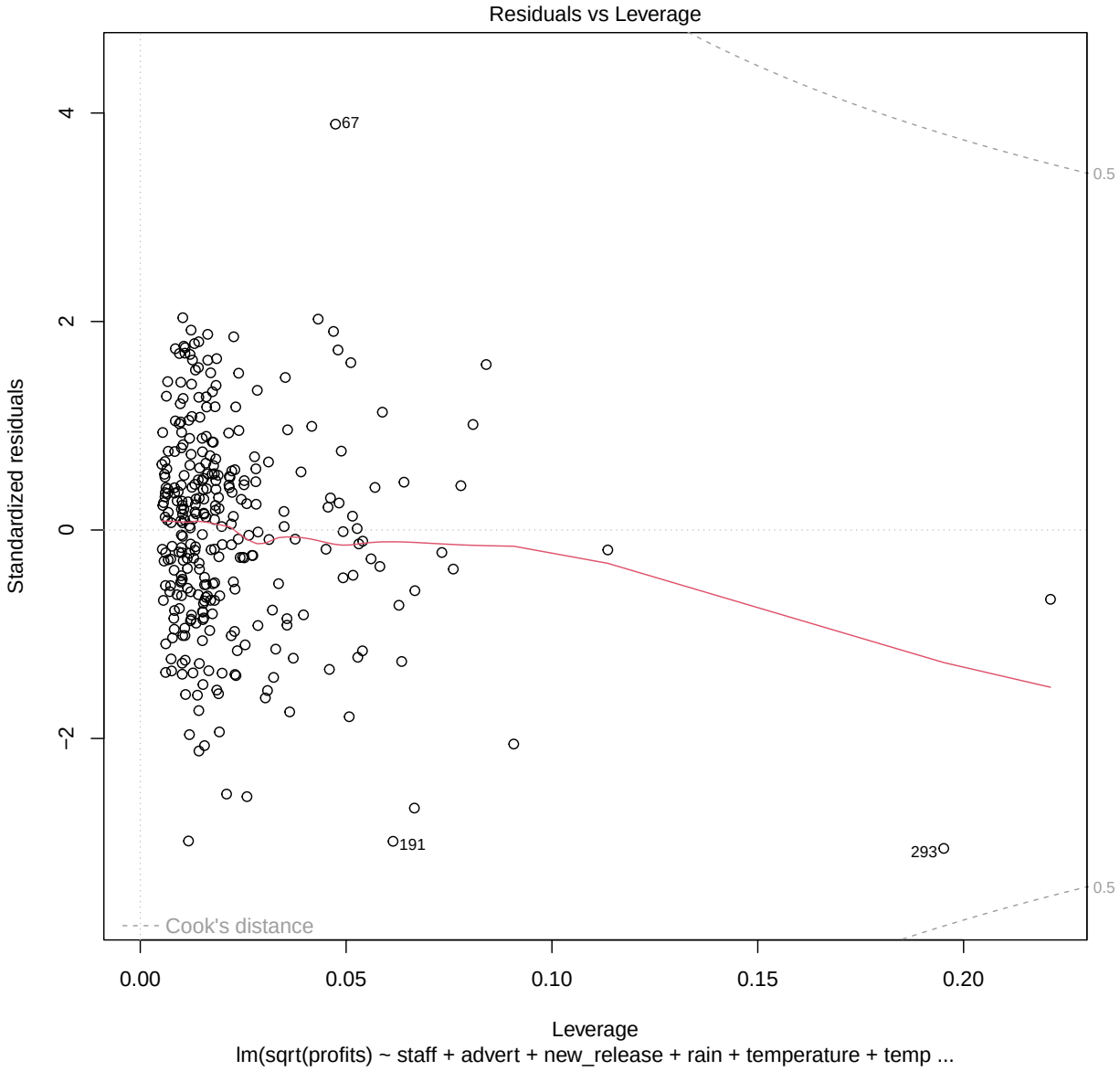
```
##
## Call:
## lm(formula = sqrt(profits) ~ staff + advert + new_release + rain +
```

```
##     temperature + temperature * rain, data = profits_2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.207  -7.212   0.797   6.611  43.820
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       33.595838   2.592078  12.961  < 2e-16 ***
## staff              2.958042   0.768894   3.847 0.000146 ***
## advert             0.035463   0.009164   3.870 0.000133 ***
## new_releaseY      -4.646662   2.418380  -1.921 0.055614 .
## rainY             -3.156548   1.797733  -1.756 0.080118 .
## temperature        2.063726   0.248927   8.290 3.64e-15 ***
## rainY:temperature  1.641181   0.371450   4.418 1.38e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.53 on 305 degrees of freedom
## Multiple R-squared:  0.5002, Adjusted R-squared:  0.4904
## F-statistic: 50.88 on 6 and 305 DF,  p-value: < 2.2e-16
```

I would like to explore, if there are highly influential data points in the dataset and see how it impacts the model.

```
plot(model_5, which = c(4,5))
```


Cook's distance
Cook's distance
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
293
67
191
0
50
100
150
200
250
300
Obs. number
lm(sqrt(profits) ~ staff + advert + new_release + rain + temperature + temp ...



Looking at the residuals vs leverage plot, most points lie far from the cook's distance contours, indicating not many points are highly influential. However, the datapoint 293 carries a high magnitude of residual and at the same time a high leverage.

Let's test how the model is after taking datapoint 293 out.

```
profits_3<-profits[-c(137,219,293),]
model_6<-lm(sqrt(profits)~ staff + advert + new_release + rain + temperature + temperature*rain, data=p
summary(model_6)
```

```
##
## Call:
## lm(formula = sqrt(profits) ~ staff + advert + new_release + rain +
##     temperature + temperature * rain, data = profits_3)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.570  -7.370   0.577   6.568  42.428
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       33.072798   2.561890  12.910  < 2e-16 ***
## staff              2.912104   0.758431   3.840 0.000150 ***
## advert             0.032478   0.009089   3.574 0.000409 ***
## new_releaseY      -4.755239   2.385274  -1.994 0.047091 *
## rainY             -2.060585   1.807899  -1.140 0.255280
## temperature        2.417728   0.270795   8.928  < 2e-16 ***
## rainY:temperature  1.292267   0.383256   3.372 0.000843 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.37 on 304 degrees of freedom
## Multiple R-squared:  0.5154, Adjusted R-squared:  0.5058
## F-statistic: 53.88 on 6 and 304 DF,  p-value: < 2.2e-16
```

The model has not changed much, with r-squared and also the p-value of the coefficient of covaraites roughly the same.

Taking a deeper look into datapoint 291. It is a day with very high temperature, yet with rather average profit. It could have been reflected as highly influential simply because not many observations see high temperature and that on that particular day of high temperature, it does not see particular high profit. I believe this datapoint still reflect and says a lot on the profit situation, hence not removing the datapoint.

Conclusion

The model indicates that staff numbers and advertising expenditure are critical drivers of daily profits, with both showing a positive association. The positive effect of staff could be due to better customer service and a bigger sales capacity. Recruiting more staff and training staff could benefit the firm's profit. Advertisements lead to increased sales or attract more customers, thereby boosting profits. Note that the coefficient of staff is 388, whereas advert is 3.9, indicating a one-unit increase in staff contributes to more profit than a one-unit increase in advert; the firm should allocate more expenses to staff than in advert.

Interestingly, the presence of a new release impacts profits, which warrants further investigation negatively. This could imply several things, such as high costs associated with new releases that are not immediately offset by sales or a consumer shift focus away from other profitable items. The firm should be wary of the cost associated with newly released models.

Although rain alone does not show a significant impact, its interaction with temperature suggests a more complex relationship, where the positive effect of temperature on profits is amplified during rainy days. Intuitively, people tends to go out less during cold and rainy days, hence explained by the model a decrease in profit.

This model gives the firm information regarding the influential factors affecting profit.

Discussion of limitations

A big limitation is that the model identifies correlation but cannot establish causality. The relationship revealed by the model does not indicate a causation effect, i.e, higher staff count lead to higher profit, it could be the case that a high sales day is anticipated, therefore more staff is called upon work that day. The coefficient indicates that the presence of more staff and higher profit has a positive relationship.

Second, the dataset might not have covered up all factors, i.e, location, season. Missing variables weaken the model's explanatory power in explaining profit. On top of that, 314 observations do not seem to be big enough for a robust model. Also, the management team should be aware that two observation of losses are removed from the data, this model is trained on. The management team should look back on those days and check out what happens and taking reference when making inferences with this model. Square root transformation can make it harder for the reader to interpret the model's inferences. The management team should be careful when making inferences.

Third, there is a likely a natural upper limit on profit each day, but using a linear model does not reflect that. A normal linear regression model does not inherently account for such upper limits and other nature of the relationships.

Last, this model is intended to understand the influential factors that account for profit, less for predicting profits. Because this is a explanatory model, interpretability is very important, hence there are covariates or interactions that could have been useful in predicting. This could lead to under fitting and therefore making predictions from this model might not be desirable. (2087 words)

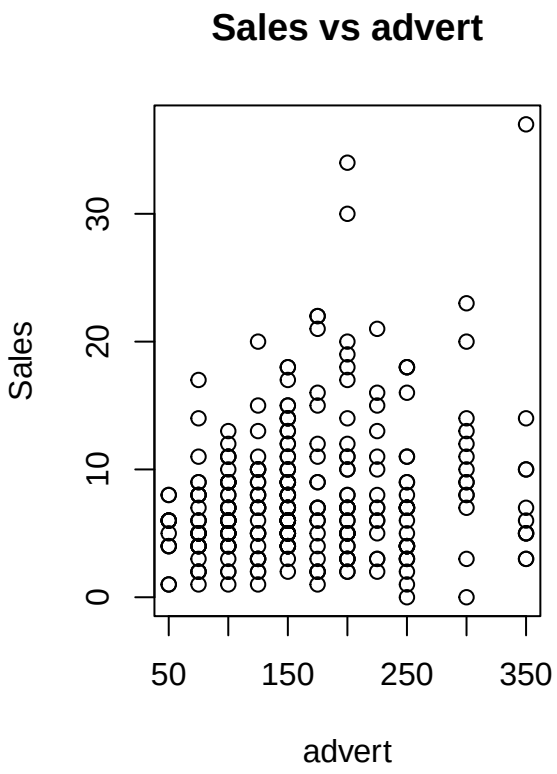
Part 2: Generalised linear model OR Generalised additive model

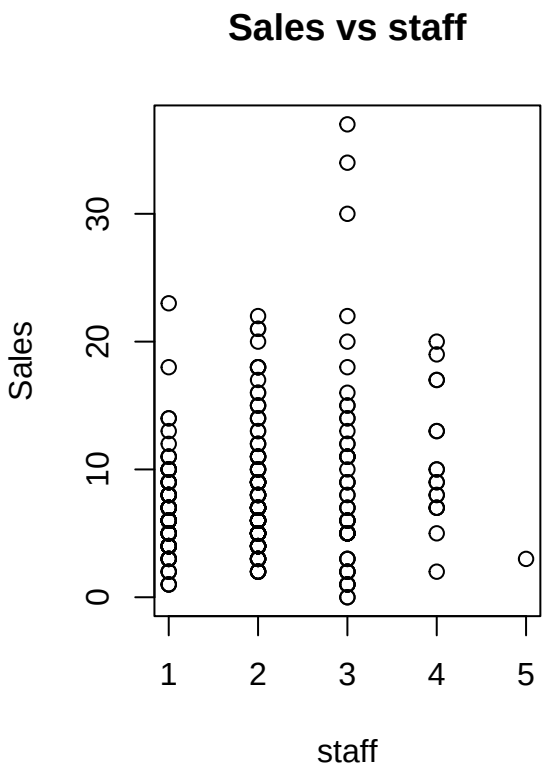
An explanatory analysis would be done to study the underlying relationship between sales and the covariates. The most suitable covariate would be picked to develop the model. Then, we will justify our decision on whether to use glm or gam. The model's fit will then be evaluated and conclusion of the model will be discussed.

```
sales <- read.csv("~/Desktop/STAT0006/ICA 3/sales.csv")
head(sales)
```

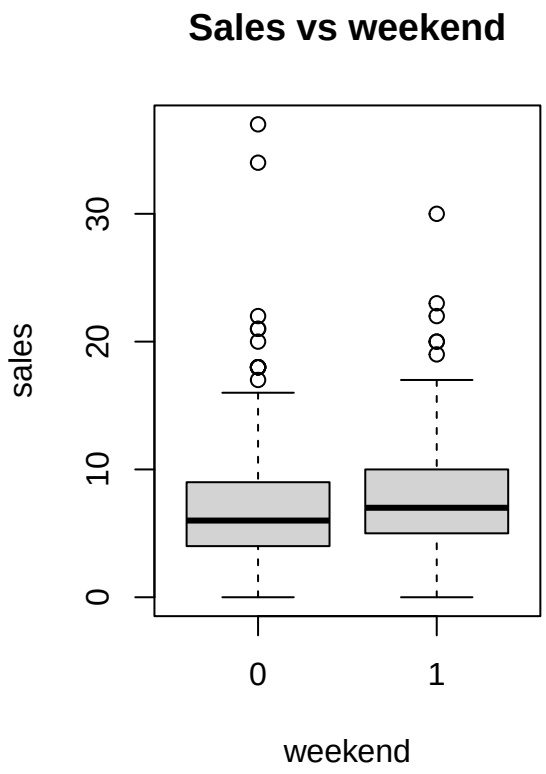
```
##   sales staff advert new_release weekend temperature rain year
## 1     7     1    250           N       0       -0.34    Y 2023
## 2     5     3    100           Y       0        2.88    N 2020
## 3     7     2    225           N       0        1.98    Y 2023
## 4    34     3    200           N       0       13.72    Y 2023
## 5    12     3    100           N       1        6.41    Y 2023
## 6     3     1    250           Y       0       -2.42    Y 2020
```

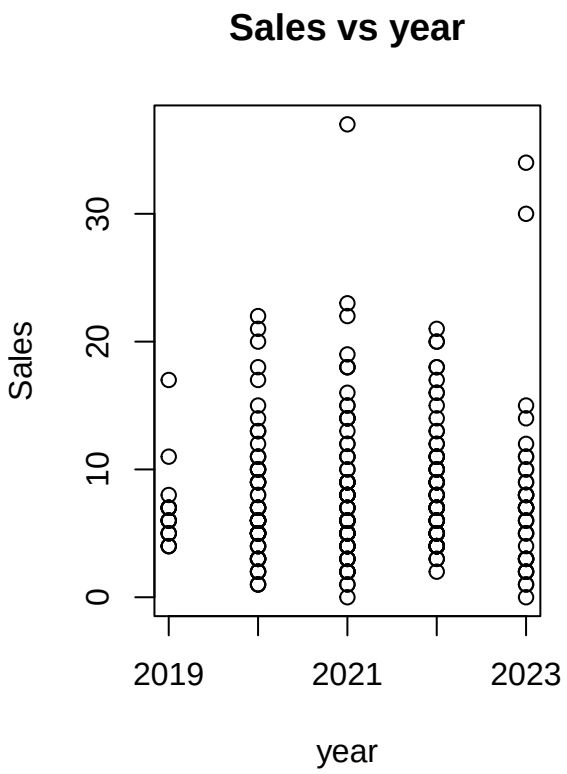
```
sales$weekend<- as.factor(sales$weekend)
sales$rain<- as.factor(sales$rain)

par(mfrow=c(1,2))
plot(sales$advert, sales$sales, main = "Sales vs advert", xlab = "advert", ylab = "Sales")
plot(sales$staff, sales$sales, main = "Sales vs staff", xlab = "staff", ylab = "Sales")
```



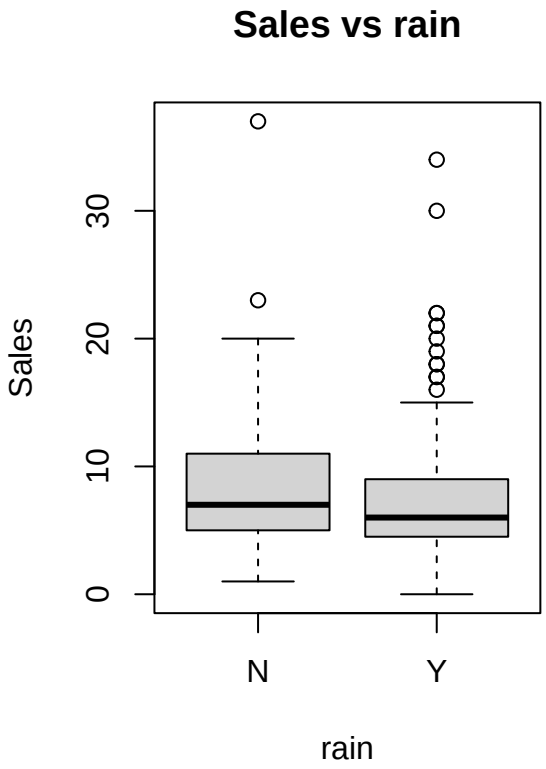


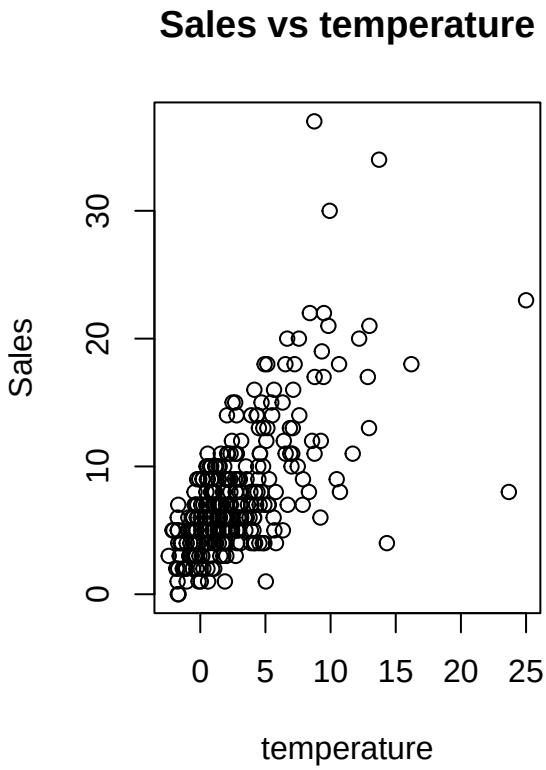

```
par(mfrow=c(1,2))
plot(sales$weekend, sales$sales, main = "Sales vs weekend", xlab = "weekend", ylab = "sales")
plot(sales$year, sales$sales, main = "Sales vs year", xlab = "year", ylab = "Sales")
```






```
par(mfrow=c(1,2))
plot(sales$rain, sales$sales, main = "Sales vs rain", xlab = "rain", ylab = "Sales")
plot(sales$temperature, sales$sales, main = "Sales vs temperature", xlab = "temperature", ylab = "Sales
```





Starting off with the exploratory analysis, we explore the overview of the dataset. Shown, by the first 5 rows, the dataset consists pretty much identical information as the profit dataset. The difference is that this records sales, not profit. Sales is recorded as the number of car sales made each day, it is a count, non-negative and discrete data. The minimum count of sales is 0, maximum count of sales is 37, mean is 7.72.

Information regarding the covariates by itselves are the same as the one in part(1). I will explore the relationship between sales and the covariates. As shown from the plot displayed, there is a slight indication that a higher advert expenses lead to higher sales count, but the visual evidence is not strong. Same for staff, the plot shows sales see its highest when there are 3 staffs and surprisingly low with 5 staffs. There is no clear positive relationship that one would intuitively think of for sales and staff, like more staff would make more sales. The median sales on and off weekend, with and without rain are very close. Sales are relatively similar throughout the years as well.

From the plot of sales vs temperature, there is a clear pattern showing a positive relationship between temperature and sales where a higher temperature leads to higher sales. This suggests temperature exhibits a relationship with sales worthy of investigation. Also, in retail sales, it is well established that weather conditions, especially temperature, can influence customer shopping behavior and purchases. Hence, temperature is selected to be the covariate of the model.

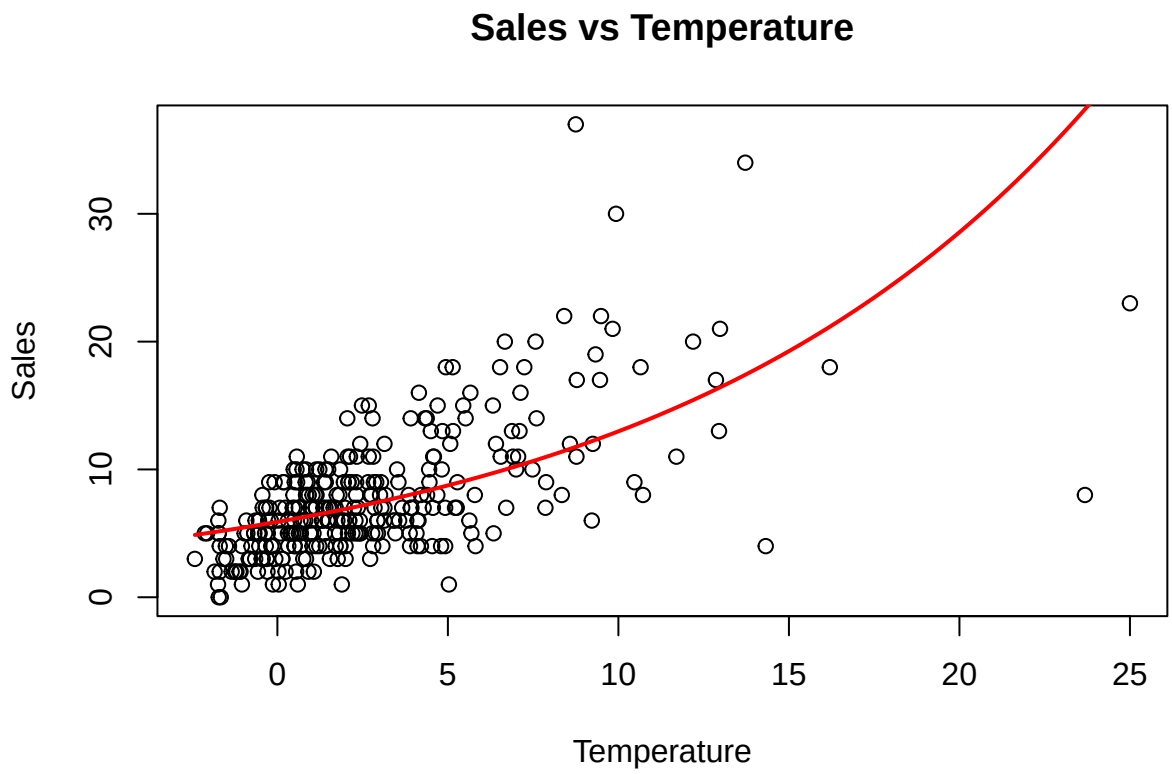
A generalised linear model will be developed with sales as the response variable and temperature as the covariate. One of the reason why glm is picked over gam is because glm has a clearer interpretability and that the model estimates and coefficients are more straightforward to explain to business stakeholders. Also, visual inspection of the sales versus temperature relationship did not conclusively establish strong evidence of nonlinearity. A linear specification may thus be reasonable. On top of that, the response variable sales, is a count, discrete and non-negative data, this can be assumed to be a poisson distribution which is a member of the exponential family which can then be modeled by a glm. A log link function is also included in the model, in attempt to improve model fitting as sales data exhibits right skew.

```r
model1 <- glm(sales ~ temperature, family = "poisson"(link = 'log'), data = sales)
summary(model1)
```

```
##
## Call:
## glm(formula = sales ~ temperature, family = poisson(link = "log"),
##     data = sales)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 1.774977   0.026049   68.14   <2e-16 ***
## temperature 0.078852   0.003803   20.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 895.76  on 313  degrees of freedom
## Residual deviance: 559.58  on 312  degrees of freedom
## AIC: 1728.7
##
## Number of Fisher Scoring iterations: 5
```

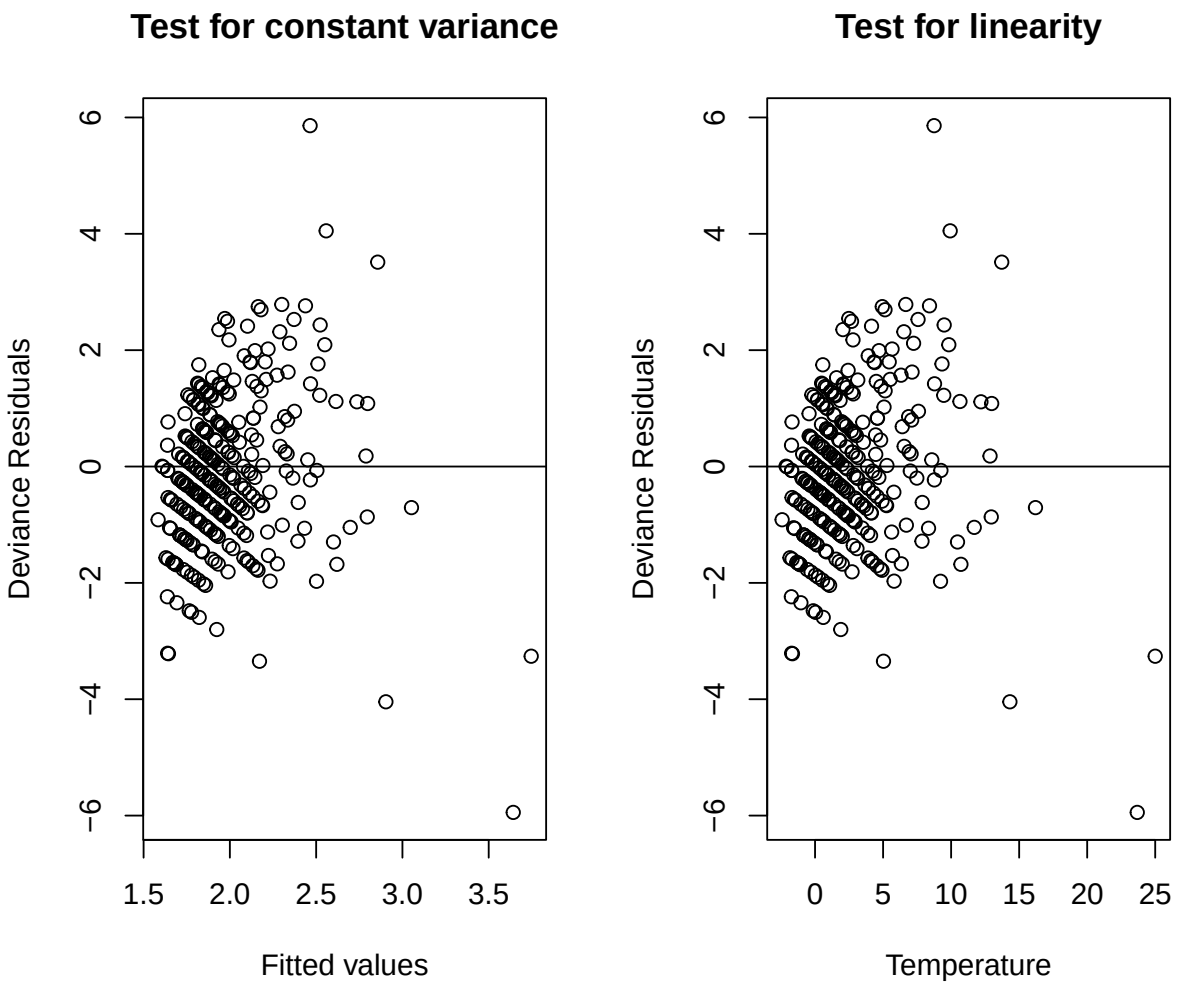
Report on modelling number of car sales

```r
temperature_range <- seq(from = min(sales$temperature), to = max(sales$temperature), length.out = 100)
predict_data <- data.frame(temperature = temperature_range)
predicted_sales <- predict(model1, newdata = predict_data, type = "response")
plot(sales$temperature, sales$sales, main = "Sales vs Temperature", xlab = "Temperature", ylab = "Sales

lines(temperature_range, predicted_sales, col = "red", lwd = 2)
```

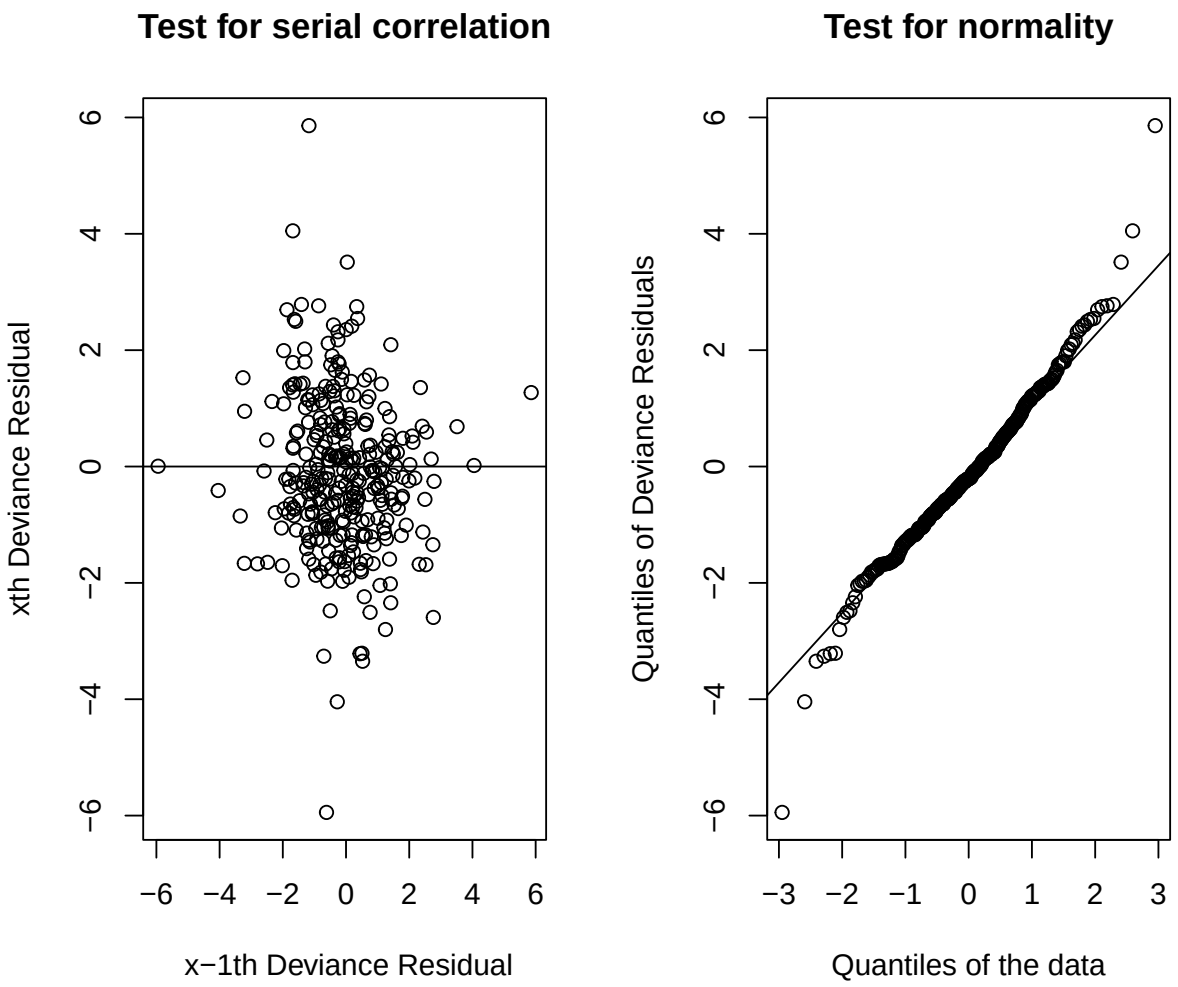


Our model(red line) seems to fit the central trend of the data points reasonably well, especially at higher temperatures where the sales numbers are more spread out. It captures the upward trend in sales as temperature increases, but it's worth noting that the spread of sales also increases with temperature, suggesting potential heteroscedasticity.

```
deviance.resid <- residuals(model1, type="deviance")
fitted.glm <- fitted(model1)

par(mfrow=c(1,2))


plot(log(fitted.glm), deviance.resid,
     xlab="Fitted values", ylab="Deviance Residuals")
abline(h=0)
title("Test for constant variance")

plot(sales$temperature, deviance.resid,
     xlab="Temperature", ylab="Deviance Residuals")
abline(h=0)
title("Test for linearity")
```




```
par(mfrow=c(1,2))
plot(deviance.resid[1:(nrow(sales)-1)], deviance.resid[2:nrow(sales)],
     xlab=" x-1th Deviance Residual", ylab="xth Deviance Residual")
abline(h=0)
title("Test for serial correlation")

qqnorm (deviance.resid, main="",
        xlab = "Quantiles of the data", ylab = "Quantiles of Deviance Residuals")
qqline (deviance.resid)
title("Test for normality")
```



From these plots we can see that there are no strong departures from the assumptions of: constant variance, linearity and serial correlation. However, there seems to be a slight departure from the normality assumption meaning that the deviance residual may have a distribution with heavier tails than a standard normal distribution. Our dataset is relatively small, which could influence the results on this plot meaning that we should still consider this model to be well fitted as it upholds a large majority of the assumptions.

Conclusion

The analysis indicates a statistically significant relationship between temperature and car sales in the showroom. Specifically, for each one-unit increase in temperature, the expected log count of sales increases by approximately 0.0788.

In practical terms, this positive coefficient for temperature suggests that warmer days are associated with an increase in the number of cars sold. The model predicts that on average, as the temperature increases, so does the number of car sales, holding all else constant. The significance of the temperature coefficient (p-value $<$ 2e-16) reinforces the robustness of this finding.

However, there are some concerns. The residual deviance of 559.58 on 312 degrees of freedom suggests that there is still unexplained variation in the data that the current model does not capture. This could be due to the fact that only one covariate is included. Also, since we are using a Poisson model, we assume that the mean and variance of the count of sales are equal. If the actual data exhibit overdispersion (variance greater

than the mean), which is common in count data, this model may not be the most appropriate. This could lead to underestimating the standard errors and thus overestimating the significance of predictors. Another model using a negative binomial distribution can be considered as a next step.