STAT0006 ICA 3

Group 86

Student numbers: 21000790, 20058715, 20017151, 20181056

Introduction to the data

Original Dataset and Adjustment

The given dataset `icecream.csv` includes data on 314 weekly sales of various ice cream brands in a supermarket chain over the past five years, each linked with data on 10 corresponding variables. It contained 3 sales records with missing values, which were removed, resulting in a modified dataset with 311 records. The number of ice creams sold per week ranges from 0 to 2444, with a mean of 530.4 ice creams.

Variables Interpretation

The variables `brand`, `brand_competitors`, `distance`, `holiday`, `milk`, `promotion`, `store_type`, `temperature`, `wind`, and `year` represent, respectively, the brand of the ice cream being sold; the number of other ice cream brands available in the store during that week; the distance (in miles) to the nearest another supermarket; whether there was a national bank holiday during the week; the national average wholesale price of milk during the week; whether there was a promotion campaign for this brand of ice cream during that week; the size of the store (Small, Medium, or Large); the average weekly store temperature (in °C); the average weekly wind speed at the store (in knots); and the year in which the sales were recorded.

Approach

The aim of this analysis is to determine the extent to which the 10 factors influence the sales of a particular brand of ice cream.

Figure 1.1 illustrates the relationship between the number of ice creams sold and the variables `brand` and `store_type`. The first plot indicates that ice cream from Brand A appears to be more popular than the other brands, while Brand B has moderate popularity, and Brand C seems to have the lowest popularity. The second plot shows that as the size of store decreases, the number of ice creams sold also decreases.

Figure 1.1: The boxplots show the number of ice cream sold each week in a store plotted against the categorical variables brand and storesize, with orange dots representing the mean number of sales in each category.

Figure 1.2 illustrates the relationship between the number of ice creams sold and the variables `promotion` and `holiday`. The first plot suggests that the weekly sales of ice cream tend to be higher when there is a promotion campaign for the particular brand, compared to weeks without such campaigns. Additionally, the second plot shows that there is a slight increase in the number of ice creams sold during weeks with national bank holidays, compared to weeks without such holidays.

Figure 1.2: Boxplots of the weekly sales amount against the binary variables promotionand holiday.

Figures 1.3 and 1.4 do not appear to demonstrate a clear linear relationship between the variables `sales` and `year`, or `sales` and `brand_competitors`, respectively. Therefore, it may be necessary to exclude the variables `year` and `brand_competitors` when constructing a normal linear regression model for ice cream sales.

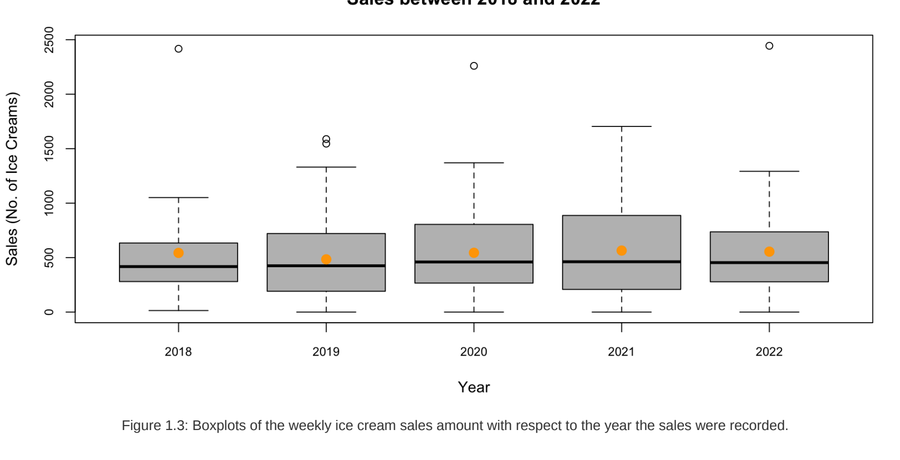

Figure 1.3: Boxplots of the weekly ice cream sales amount with respect to the year the sales were recorded.

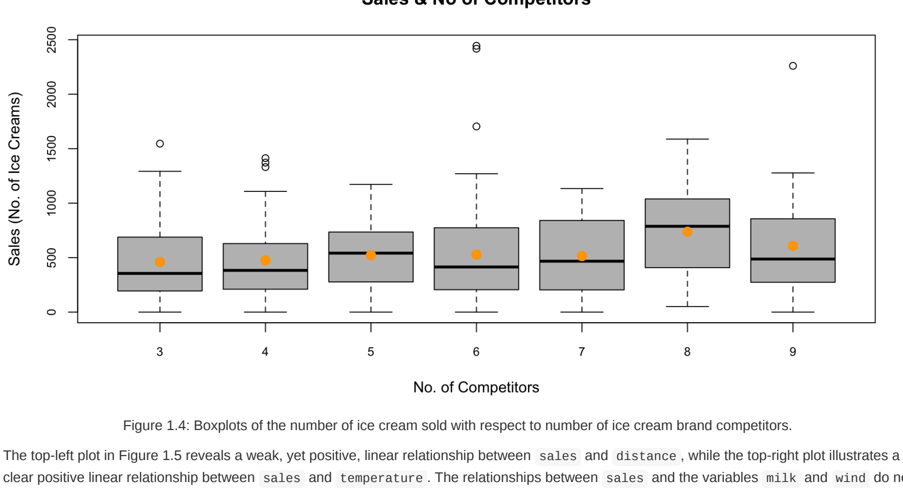

Figure 1.4: Boxplots of the number of ice cream sold with respect to number of ice cream brand competitors.

The top-left plot in Figure 1.5 reveals a weak, yet positive, linear relationship between `sales` and `distance`, while the top-right plot illustrates a clear positive linear relationship between `sales` and `temperature`. The relationships between `sales` and the variables `milk` and `wind` do not exhibit linearity, so it may be advisable to exclude both variables when constructing a linear model.

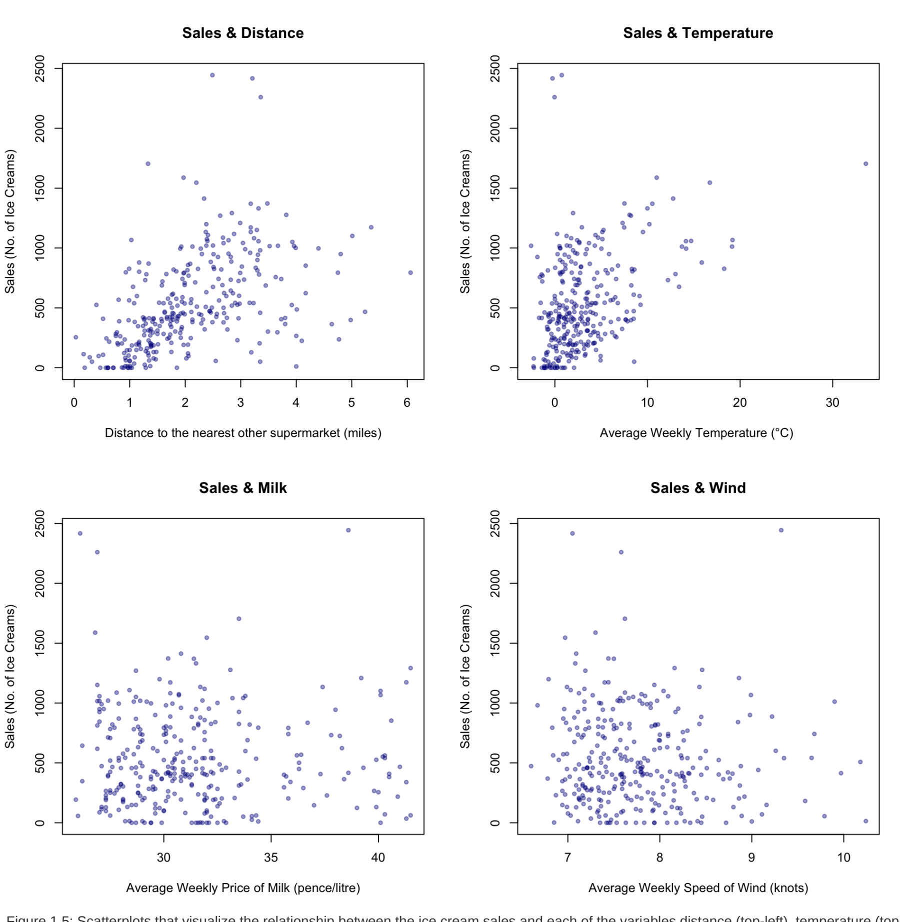

Figure 1.5: Scatterplots that visualize the relationship between the ice cream sales and each of the variables distance (top-left), temperature (top-right), milk (bottom-left) and wind (bottom-right).

Model Building and Checking

Step 1: Select important covariates for normal linear model

Due to unclear linear relationship with `sales`, we decided to omit the covariates `year`, `wind`, `milk` and `brand_competitors` in Model 1. Model 1 is as below:

```
##
## Call:
## lm(formula = sales ~ brand + holiday + promotion + store_type +
##     temperature + distance, data = ic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -501.49 -140.36  -15.78  117.53 1652.33
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       240.070     64.182   3.740  0.00022 ***
## brandBrandB      -237.980     35.061  -6.788 6.04e-11 ***
## brandBrandC      -382.910     36.095 -10.608  < 2e-16 ***
## holidayY           91.528     40.712   2.248  0.02529 *
## promotionY        223.796     48.075   4.655 4.86e-06 ***
## store_typeMedium  -65.097     40.458  -1.609  0.10866
## store_typeSmall   -34.515     44.981  -0.767  0.44349
## temperature        39.385      3.577  11.012  < 2e-16 ***
## distance          173.065     17.394   9.950  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 255 on 302 degrees of freedom
## Multiple R-squared:  0.6088, Adjusted R-squared:  0.5984
## F-statistic: 58.74 on 8 and 302 DF,  p-value: < 2.2e-16
```

The p-values of `store_typeMedium` and `store_typeSmall` in Model 1 are large, thus we might want to remove `store_type` from our model, keeping all the other covariates. However, Figure 1.1 suggests that `store_type` might influence `sales`, since ice cream sales increase as store size increases. Therefore, we kept `store_type` in the model and looked for further interactions between `store_type` and other covariates.

Step 2: Add suitable interaction terms

The relationship between `store_type` and `sales` might be more complex and might be dependent on other factors, such as the distance to the nearest store, the brand of ice cream being sold or whether there was a holiday or not. We therefore consider including interaction terms for `store_type*distance`, `store_type*brand` and `store_type*holiday`. The small p-value for `store_typeMedium` and `store_typeSmall` in the resulting Model 2 is a sufficient evidence for keeping `store_type` in the model. Model 2 is as below:

```
##
## Call:
## lm(formula = sales ~ +holiday + temperature + promotion + distance *
##     store_type + store_type * brand + store_type * holiday, data = ic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -616.81  -56.72  -24.81   46.91  873.43
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    737.771     43.060  17.133  < 2e-16 ***
## holidayY                       631.301     47.523  13.284  < 2e-16 ***
## temperature                     49.197      1.942  25.337  < 2e-16 ***
## promotionY                     209.209     25.648   8.157 1.01e-14 ***
## distance                        66.301     12.426   5.336 1.90e-07 ***
## store_typeMedium              -976.890     60.826 -16.060  < 2e-16 ***
## store_typeSmall              -1123.666     61.617 -18.236  < 2e-16 ***
## brandBrandB                   -529.394     33.830 -15.649  < 2e-16 ***
## brandBrandC                   -907.698     33.410 -27.169  < 2e-16 ***
## distance:store_typeMedium      242.830     21.028  11.548  < 2e-16 ***
## distance:store_typeSmall       390.339     27.108  14.400  < 2e-16 ***
## store_typeMedium:brandBrandB   476.251     48.064   9.909  < 2e-16 ***
## store_typeSmall:brandBrandB    524.242     45.227  11.591  < 2e-16 ***
## store_typeMedium:brandBrandC   808.795     46.409  17.428  < 2e-16 ***
## store_typeSmall:brandBrandC    890.517     47.925  18.582  < 2e-16 ***
## holidayY:store_typeMedium     -769.397     59.970 -12.830  < 2e-16 ***
## holidayY:store_typeSmall      -704.748     58.115 -12.127  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 134.6 on 294 degrees of freedom
## Multiple R-squared:  0.8939, Adjusted R-squared:  0.8881
## F-statistic: 154.8 on 16 and 294 DF,  p-value: < 2.2e-16
```

- Holiday and Store Type

In Figure 2.1, there is an unclear distinction among the different store types when considering the effects of `holiday` on `sales`. Therefore, there is no strong evidence to keep `store_type * holiday` in the model.

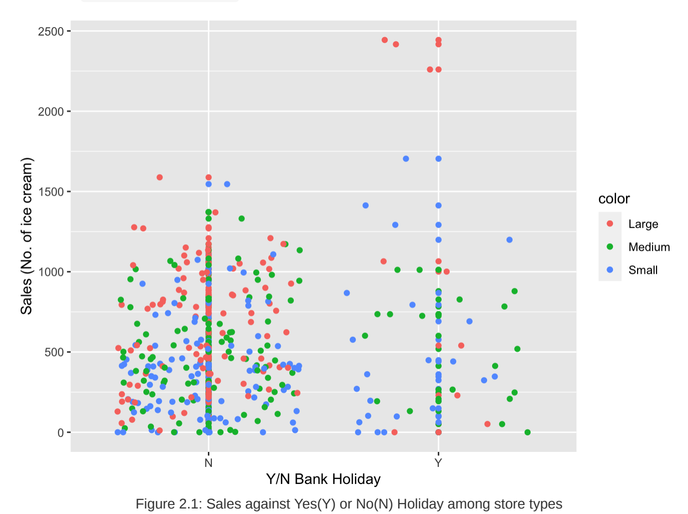

Figure 2.1: Sales against Yes(Y) or No(N) Holiday among store types

- Distance and Store Type

Based on the plot below, the data points dependent on each store type seem to behave in a different slope when plotting `distance` against `sales`. For this reason, there is evidence supporting the claim that the effect of `distance` on `sales` depends on `store_type`.

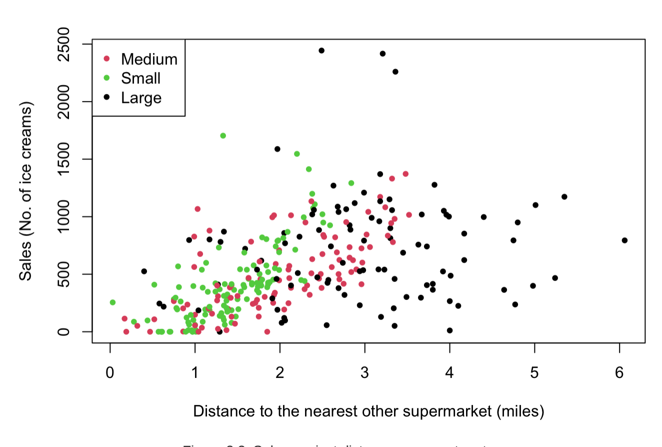

Figure 2.2: Sales against distance, among store types

- Brand and Store Type

From the figure below, the data points dependent on `Brand` appear to be randomly distributed in medium and small store categories. However, a clear ordering in sales appears in the large store category (Brand A > Brand B > Brand C), which is an evidence for retaining `brand*store_type` in the model. Therefore, only `store_type * distance` and `store_type * brand` are added as interactions.

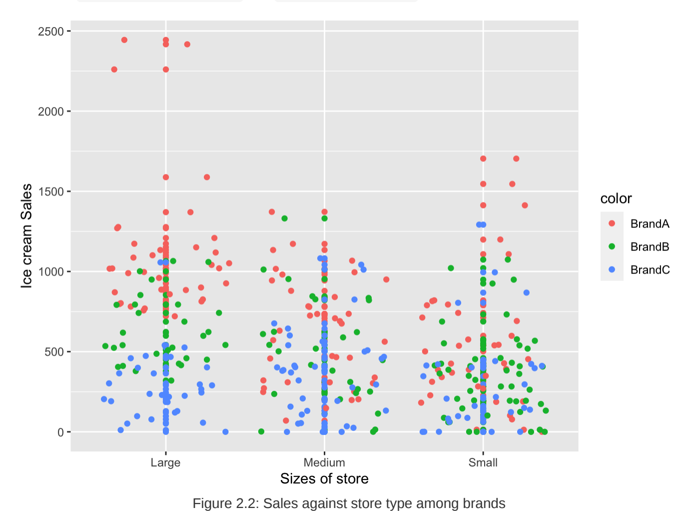

Figure 2.2: Sales against store type among brands

Step 3: Finalise interaction term selection

Interaction term `store_type * brand` is removed. The resulting Model 3 is below:

```
##
## Call:
## lm(formula = sales ~ temperature + promotion + holiday + distance *
##     store_type + store_type * brand, data = ic)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -317.51  -70.36  -14.12   30.33 1393.90
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    799.743     54.607  14.646  < 2e-16 ***
## temperature                     44.346      2.433  18.230  < 2e-16 ***
## promotionY                     210.679     32.546   6.473 3.97e-10 ***
## holidayY                        54.413     27.548   1.975   0.0492 *
## distance                        65.514     15.847   4.134 4.64e-05 ***
## store_typeMedium             -1036.789     77.279 -13.416  < 2e-16 ***
## store_typeSmall              -1166.165     78.471 -14.861  < 2e-16 ***
## brandBrandB                   -536.347     43.137 -12.434  < 2e-16 ***
## brandBrandC                   -876.147     42.504 -20.613  < 2e-16 ***
## distance:store_typeMedium      236.929     26.808   8.838  < 2e-16 ***
## distance:store_typeSmall       376.893     34.409  10.953  < 2e-16 ***
## store_typeMedium:brandBrandB   476.528     61.292   7.775 1.26e-13 ***
## store_typeSmall:brandBrandB    524.082     57.676   9.087  < 2e-16 ***
## store_typeMedium:brandBrandC   765.171     59.034  12.962  < 2e-16 ***
## store_typeSmall:brandBrandC    834.197     60.785  13.724  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 171.7 on 296 degrees of freedom
## Multiple R-squared:  0.8263, Adjusted R-squared:  0.8181
## F-statistic: 100.6 on 14 and 296 DF,  p-value: < 2.2e-16
```

Assumption Evaluation

It is crucial to verify the assumptions of an ordinary least squares model using diagnostic plots to ensure its validity. These assumptions include: the normality, the constant variance and independence of the error terms. It is also advisable to examine multicollinearity afterwards.

- Normality: Regarding the assumption of normality, the QQ-plot in Figure 3.2 indicates that the error term is approximately normally distributed, although the tails are slightly heavier than expected under the normality assumption. Consequently, there are no major normality concerns.

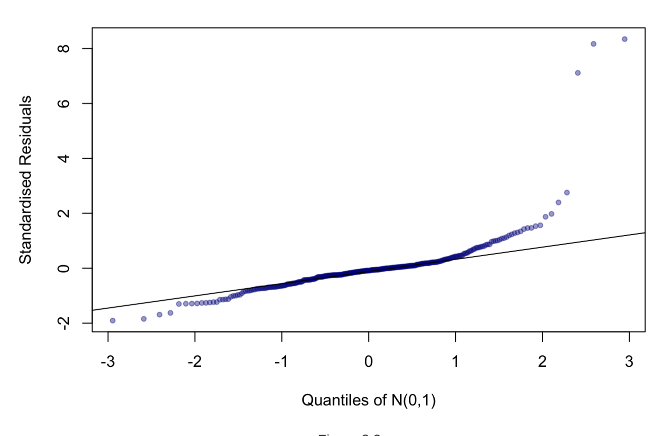

Figure 3.2

- *Homoscedasticity and Independence:* These assumptions are not violated if there is no systematic pattern in Standardised Residuals-Fitted Values plot. The left side of Figure 3.3 reveals a linear pattern in a range of fitted values, particularly where the fitted values of `sales` are negative. Transformations can be applied to check if homoscedasticity can be resolved.

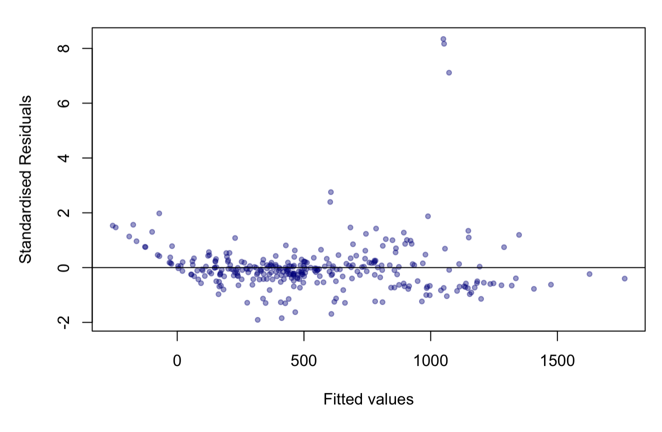

Figure 3.3

- *Multicollinearity:* Multicollinearity can be assessed by evaluating the variance inflation factors (VIFs) for Model 3. As all VIF < 5, indicating that multicollinearity is not a concern. This is further supported by the fact that the estimated coefficients are relatively stable, meaning that they do not change much when the model is retrained.
- *Potentially omitted important covariates:* No systematic relationship is observed in the standardised residual values against the omitted `wind` and `milk` (Figure 3.4), suggesting there is no need to include these variables in the model.

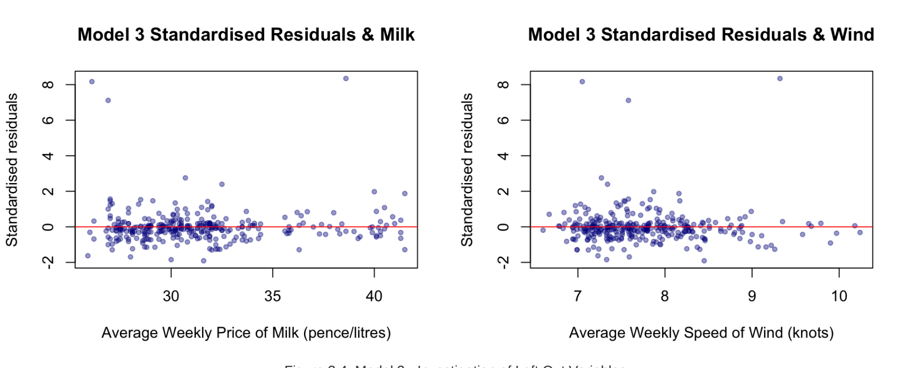

Figure 3.4: Model 3 - Investigation of Left Out Variables

In summary, Model 3 exhibits a strong fit for the observed data, but some predicted values are negative, and there is a linear pattern in the plot of standardised residuals against fitted values, suggesting potential concerns about homoscedasticity and independence.

Step 4: Transformation of Response Variable

To address the homoscedasticity issue in Model 3, we applied Box-Cox transformation to `sales` to the variance. The optimal transformation function is determined by identifying the value of lambda that maximises the log-likelihood, as depicted in Figure 3.5 below.

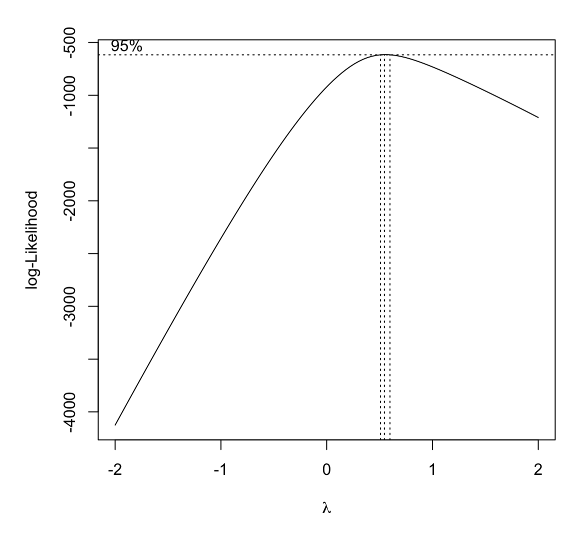

Figure 3.5: Box-Cox Transformation after Model 3

The optimal value of lambda is approximately 0.5, indicating that the suitable function for transforming `sales` is the square root function. Note that Box-Cox transformation only works on strictly positive values, yet there are several zero sales observations in the dataset. Adding a constant can make them positive, without changing the original data distribution. Model 4 is as below:

```
##
## Call:
## lm(formula = sqrt(sales) ~ +temperature + promotion + holiday +
##     distance * store_type + store_type * brand, data = ic)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.0045  -1.4775   0.1824   1.3466  20.2065
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   25.73322    1.10206  23.350  < 2e-16 ***
## temperature                    1.09694    0.04909  22.343  < 2e-16 ***
## promotionY                     5.39047    0.65684   8.207 7.06e-15 ***
## holidayY                      -0.95008    0.55598  -1.709   0.0885 .
## distance                       1.46003    0.31981   4.565 7.32e-06 ***
## store_typeMedium             -23.90709    1.55964 -15.329  < 2e-16 ***
## store_typeSmall              -28.98044    1.58368 -18.299  < 2e-16 ***
## brandBrandB                   -9.39458    0.87059 -10.791  < 2e-16 ***
## brandBrandC                  -18.32343    0.85780 -21.361  < 2e-16 ***
## distance:store_typeMedium      6.57211    0.54103  12.147  < 2e-16 ***
## distance:store_typeSmall      11.07364    0.69443  15.946  < 2e-16 ***
## store_typeMedium:brandBrandB   7.65257    1.23699   6.186 2.04e-09 ***
## store_typeSmall:brandBrandB    9.66340    1.16400   8.302 3.69e-15 ***
## store_typeMedium:brandBrandC  15.34787    1.19141  12.882  < 2e-16 ***
## store_typeSmall:brandBrandC   17.14927    1.22676  13.979  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.464 on 296 degrees of freedom
## Multiple R-squared:  0.8718, Adjusted R-squared:  0.8658
## F-statistic: 143.8 on 14 and 296 DF,  p-value: < 2.2e-16
```

- *Assumptions Evaluation*

Model 4 exhibits normality to a satisfactory extent, despite the fatter tails compared to Model 3. The transformation of `sales` with square root function did ensure the positive range of sales values. However, the linear pattern was not removed in Figure 3.6, indicating that the homoscedasticity issue in Model 3 was not fully resolved.

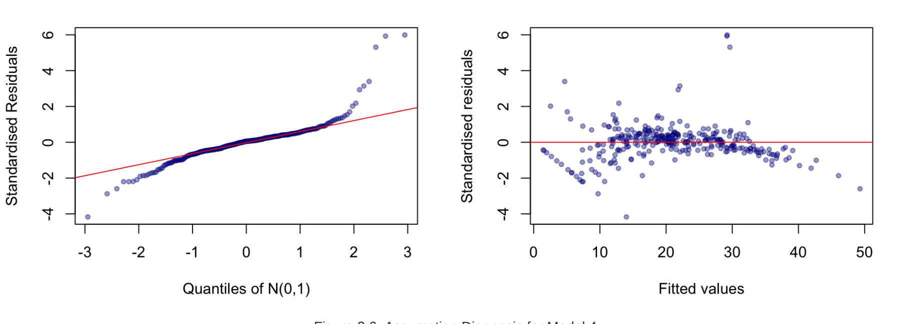

Figure 3.6: Assumption Diagnosis for Model 4

Comparing fit of all models

Model 3's performance appears to be satisfactory when compared to its observed ice cream sales data plot to all other models, depicted in Figure 3.7.

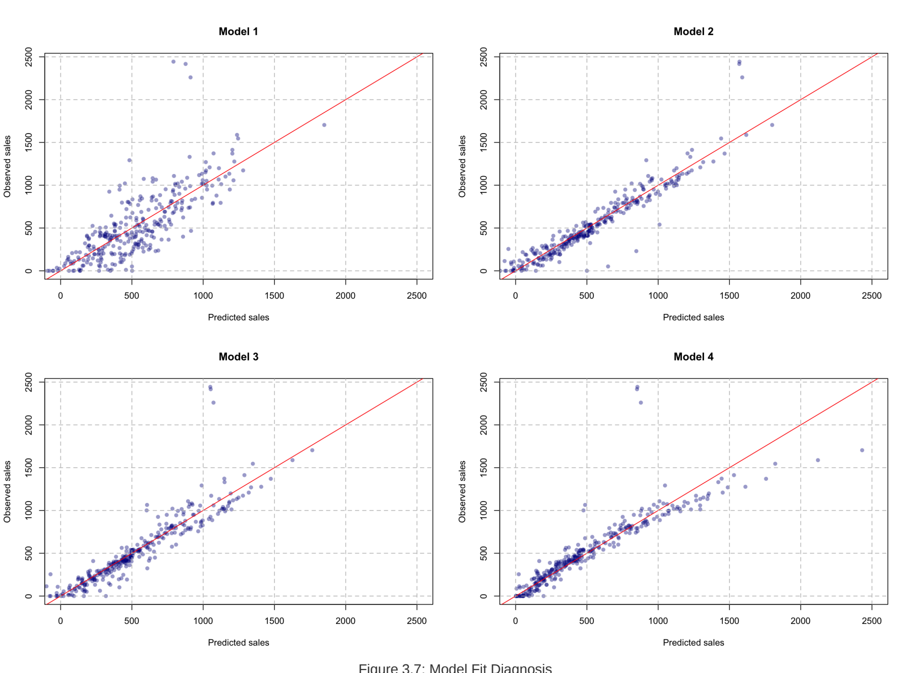

Figure 3.7: Model Fit Diagnosis

The fit was highly improved in Model 2 from Model 1, supported by the increase in the R-Squared value (0.6088 < 0.8939). When comparing Model 2 to Model 3, Model 3 has fewer overvalued points in the [500,1000] range and appears to fit the lower sales points more strongly.

Evaluating the fit of Model 4 raises concerns in higher sales values, when the fit starts to form a curved relationship, although its R-Squared value improved from Model 3 (0.8718 > 0.8263). Thus, Model 3 remains the most suitable predictions for the Ice Cream Sales dataset.

Conclusion

Model 3 suggests the expected ice cream sale is around 800 units for the reference group. The chosen reference group in our final model (Model 3) is `BrandA`, `N` for no promotion, and the store_type `Large`.

`Promotion`: Running a promotion can have a positive effect on weekly sales, which is expected to be 210.7-unit higher than when there is no promotion, holding remaining covariates constant.

`Temperature`: People buy ice cream when the weather is hotter, which is at 44 units increase of weekly sales per 1°C increase, holding all other covariates constant.

`Distance` * `store_type`: Looking at the coefficients of `distance:store_typeMedium` and `distance:store_typeSmall`, the final model suggests that when `distance` increases by one mile, the increase in ice cream sales in small stores > medium stores > large stores.

Figure 4.1 shows that small and medium stores tend to have very low sales when the distance to the nearest stores is small. This could be explained by reasons such as people would want to go to larger stores at such a distance so that they can shop for more other goods/ have more ice cream options. When this distance increases, people might adhere to the current store for convenience.

Also, the increase in ice cream sales is more sticky in large stores because shoppers have more options within that shop. They will be less willing to travel to nearby stores, thus ice cream sales in large stores are less influenced by the distance to the nearest store.

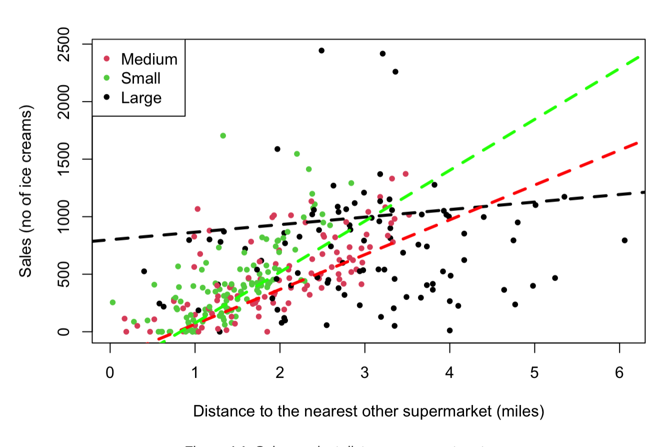

Figure 4.1: Sales against distance among store types

`store_type*brand`: Figure 2.2 showed consumers seem to care less about the brand of ice cream when they are shopping in small stores, the points don't display a clear separation. However, in larger stores, there is a clearer distinction between the sales of ice cream: Brand A > Brand B > Brand C. This could be explained by shopping tastes or availability problems in smaller stores.

Discussion of limitations

Data

Even though the dataset already included the most important factors, some important variables might have been left out. Some possibilities can be the population density in the store area, marketing spending or demographic factors.

Model

From the bottom-left plot in Figure 3.7, most of the negative predicted values are aligned with the observed sales values at 0. From Figure 3.3, these values also form the linear pattern, causing concerns about the homoscedasticity and independence assumptions.

We investigated the observations in the dataset where sales are zero, which seem reasonable: mostly among small and medium stores, the nearest stores are within 1-mile, there is no promotion etc. There appears to be no systematic problems with these data, we call these the 'empty season', where sales happen to be zero.

We also investigated the extreme points in the dataset. The model also does not fit as strongly with higher values of sales, particularly the three extreme values at 'peak seasons' where sales are over 2000. These values have high values of sales, and after investigation, we realised that they seem to be valid because they are observed in large stores, and during holiday seasons. Although they might pull the fitted hyperplane towards the higher values, these values only make up less than 1% of the observations, they are therefore not a major concern.

Total word count: 1989