ST1131 Assignment 2

By Toh Yi Fan (A0272396U)

INTRODUCTION:

The purpose of this report is to propose a linear regression model for the response variable (count of bike rental) and investigate if the proposed model is adequate.

We are interested in finding an adequate linear regression model for the count of bike rental using the following explanatory variables: season, working day, weather situation, temperature, humidity, and windspeed.

PART I: EXPLORING VARIABLES

1) Suitability of Response Variable to fit linear model:

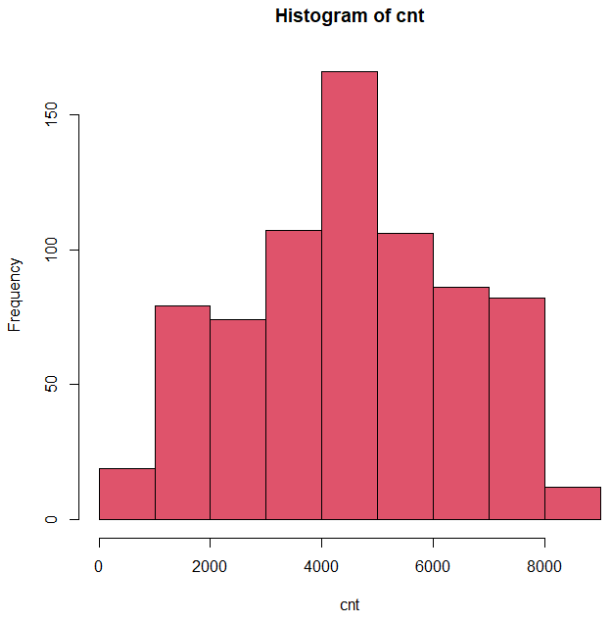

Figure 1: Histogram

- Based on this histogram of count of rental bikes, the graph is unimodal, symmetric and there are no obvious outliers. Thus, it is suitable to fit a linear regression model for the count of rental bikes as it seems to be normally distributed.

2) Association between the response and quantitative explanatory variable

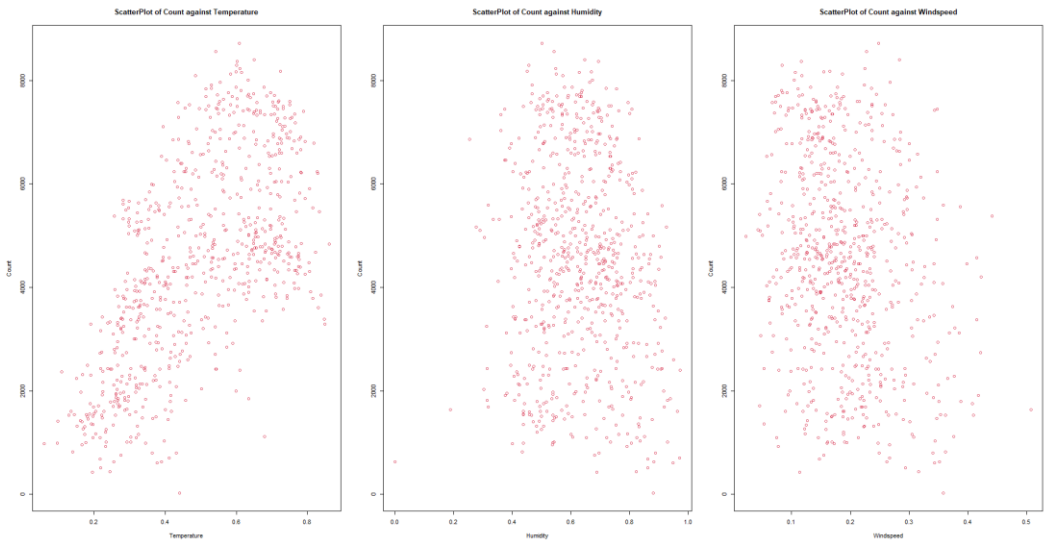

Figure 2: Scatterplot graphs

- Based on the above scatterplots between the response and the quantitative explanatory variable, we conclude the following associations:
  1) There is a somewhat strong positive linear association between bike rental count and temperature (cor = 0.627494).
  2) There is a weak negative association between bike rental count and humidity (cor = -0.1006586).
  3) There is a somewhat weak negative association between bike rental and humidity (cor = -0.234545).

3) Association between the response and categorical explanatory variable

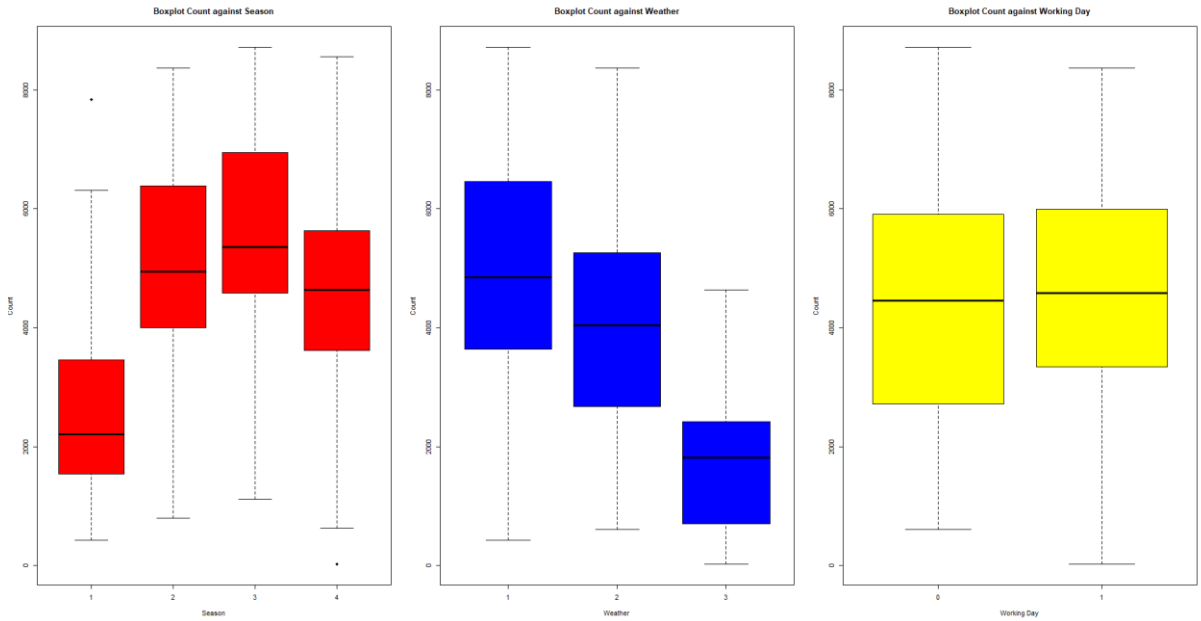

Figure 3: Boxplot graphs

- Based on the above bar plot graphs, we conclude the following associations:

1) The median bike rental count increases from Spring to Autumn and decrease from Autumn to Winter. There are 2 outliers, 1 outlier above median during spring and 1 outlier below median during winter.
2) The median bike rental count decreases as weather worsens.
3) The median bike rental count on working days seems to be slightly higher and the interquartile range is smaller than on non-working days.

In conclusion, the least significant variable (based on the association of the variables with the response variable) are humidity, windspeed, and working days.

PART II: BUILDING MODEL

Proposed regressors: Season, temperature, and weather situation.

Initial Model, M1:

Fitted regression line:

Point estimate of mean of Count = 18.23 + 3136.50(I(season = 2)) + 11102.19(I(season = 3)) + 1347.47(I(season = 4)) - 193.59(I(weathersit = 2)) + 1705.69(I(weathersit = 3)) + 9271.97(temp) - 5023.11(I(season = 2) * temp) - 16544.05(I(season = 3) * temp) - 188.98(I(season = 4) * temp) - 2390.34(I(season = 2) * I(weathersit = 2)) - 9656.00(I(season = 3) * I(weathersit = 2)) + 1318.57(I(season = 4) * I(weathersit = 2)) - 1725.86(I(season = 2) * I(weathersit = 3)) - 5517.14(I(season = 3) * I(weathersit = 3)) - 5851.17 (I(season = 3) * I(weathersit = 3)) - 648.47(I(weathersit = 2) * temp) - 12210.88(I(weathersit = 3) * temp)) + 3439.43(I(season = 2) * I(weathersit = 2) * temp) + 13778.46(I(season = 3) * I(weathersit = 2) * temp) - 3586.93(I(season = 4) * I(weathersit = 2) * temp) + 2581.04(I(season = 2) * I(weathersit = 3) * temp) + 11787.92(I(season = 3) * I(weathersit = 3) * temp) + 13517.89 I(season = 4) * I(weathersit = 3) * temp)

Residual Plot:

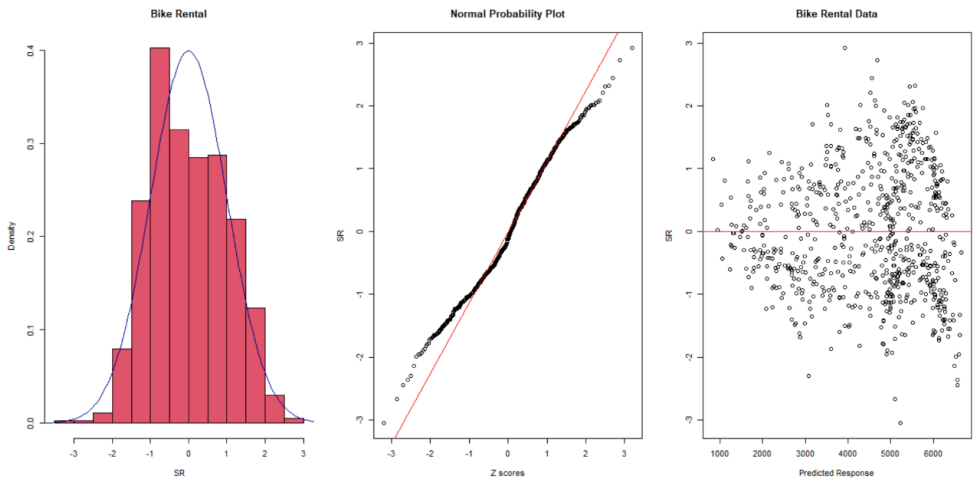

Figure 4: Residual Plot for M1

Comments:

- The histogram is unimodal, symmetric, has no obvious outliers, and ranges from -3 to 3, thus it is normally distributed.
- From the QQ plot, both right and left tails are **NOT** normal (both shorter than normal), hence normality assumption **is violated**.
- From the Scatter plot graph, there is 1 outlier. In addition, the outwards funnel shape suggests that the variance is **NOT** constant, thus **violating** the homoscedasticity assumption.

Conclusion: M1 is not adequate as it violates both normality and homoscedasticity assumption.

Outliers & Influential Points:

```
> which(SR>3 | SR<(-3) )
239 442
239 442
```

```
> Cook = cooks.distance(M1)
> which(Cook >1)
named integer(0)
```

- There are two outliers at index 239 and index 442.
- There are no influential points in M1.

Significance of Regressors:

```
Response: cnt
                        Df     Sum Sq   Mean Sq  F value    Pr(>F)
season                   3  950595868 316865289 199.0040 < 2.2e-16 ***
weathersit               2  243512285 121756142  76.4677 < 2.2e-16 ***
temp                     1  255086767 255086767 160.2047 < 2.2e-16 ***
season:temp              3   98643137  32881046  20.6506 7.788e-13 ***
season:weathersit        6   20075457   3345909   2.1014 0.0510934 .
weathersit:temp          2    1695196    847598   0.5323 0.5874731
season:weathersit:temp   6   44202032   7367005   4.6268 0.0001269 ***
Residuals              707 1125724650   1592256
```

Figure 5: Anova table summary for M1

- Using a significance level of 0.05, we can observe that the p-value for the interaction terms weathersit * season and weathersit * temp is larger than 0.05. Thus, this suggests that the 2 interaction terms are not significant, therefore we can safely drop these 2 interaction terms.

Goodness-of-fit:

```
Multiple R-squared:  0.5891,    Adjusted R-squared:  0.5757
F-statistic: 44.07 on 23 and 707 DF,  p-value: < 2.2e-16
```

- The p-value of F test is very small, thus suggesting that the overall model is significant.
- But R^2 is somewhat small as 0.5891 suggest that only 58.91% of the variation in bike rental count is explained by the fitted regression. Adjusted R^2 = 0.5757.

Next step: We can fix the homoscedasticity assumption by transforming the response variable to log(count) and drop non-significant terms to make the model less complex.

Intermediate Model, M2:

Fitted regression line:

Point estimate of mean of log(Count) = 6.86883 + 0.85656(I(Season = 2)) + 2.41614(I(Season = 3)) + 0.93404(I(Season = 4)) - 0.18382(I(weathersit = 2)) - 1.26867(I(weathersit = 3)) + 3.16145(Temp) - 1.68402(I(Season = 2) * Temp) - 4.02643(I(Season = 3) * Temp) - 1.50831(I(Season = 4) * Temp)

Next step: We can remove the outlier at index 668 as the extremely low count is likely due to the Hurricane Sandy that affected 24 states in the US on 29 October 2012. Since this is a rare occurrence of extreme weathers, we can remove this outlier.

Final Model, M3:

Proposed regressors: Season, temperature, weather situation and humidity.

Fitted regression line:

Point estimate of mean of log(Count) = 6.26504 + 0.89432(I(Season = 2)) + 2.61267(I(Season = 3)) + 0.93867(I(Season = 4)) - 0.04433(I(Weathersit = 2)) - 0.65931(I(Weathersit = 3)) + 3.39087(Temp) - 1.79477(I(Season = 2) * Temp) - 4.39606(I(Season = 3) * Temp) - 1.45054(I(Season = 4) * Temp) + 2.39676(Hum) - 2.55277(I(Hum^2)

Residual plot:

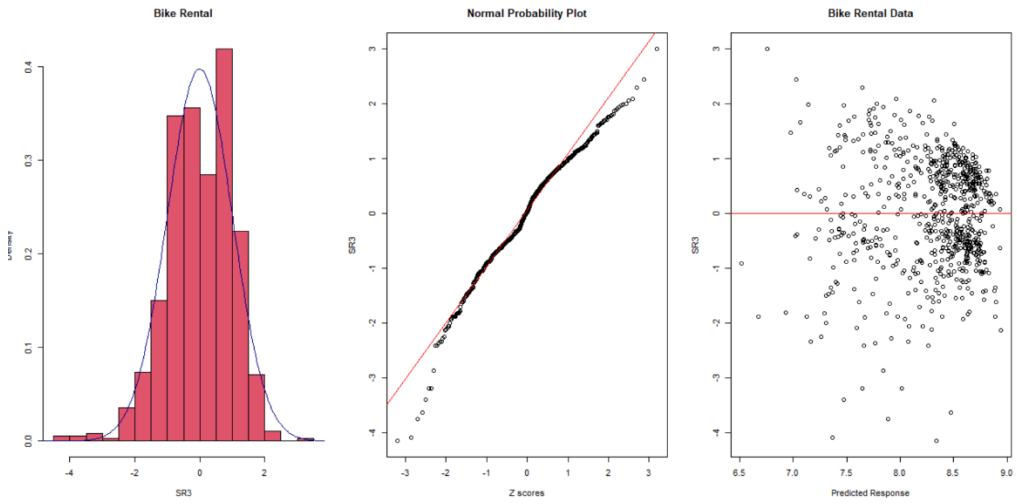

Figure 6: Residual plot for final model

Comments:

- The histogram is unimodal, left skewed, with no obvious outliers. The range of the histogram is not from -3 to 3. Thus, this is **NOT** a normal distribution.
- From the QQ plot, both right and left tails are **NOT** normal (left longer than normal, right shorter than normal), hence normality assumption **is violated**.

- From the Scatter plot graph, there are multiple outliers, but there is **NO** funnel shape observed, thus homoscedasticity assumption is **not violated**.

Conclusion: M3 is inadequate due to violation of normality assumption.

Outliers & Influential Points:

```
> which(SR3 >3 | SR3<(-3) )
  2  27  65 239 328 359 407 669
  2  27  65 239 328 359 407 668
```

```
> Cook = cooks.distance(M3)
> which(Cook >1)
named integer(0)
```

- There are outliers at index 2, 27, 65, 239, 328, 359, 407, 669.
- There are no influential points in M1.

Significance of Regressors:

```
Response: log(cnt)
             Df Sum Sq Mean Sq F value    Pr(>F)
season        3 83.997 27.9992 270.353 < 2.2e-16 ***
temp          1 25.937 25.9371 250.442 < 2.2e-16 ***
weathersit    2 20.942 10.4708 101.103 < 2.2e-16 ***
hum           1  2.356  2.3560  22.749 2.236e-06 ***
I(hum^2)      1  1.968  1.9680  19.003 1.496e-05 ***
season:temp   3 11.866  3.9553  38.191 < 2.2e-16 ***
Residuals   718 74.360  0.1036
```

Figure 7: Anova table summary for M3

- Using a significance level of 0.05, we can observe that the p-value for all the terms is less than 0.05, therefore all the terms in M3 are significant.

Goodness-of-fit:

```
Multiple R-squared:  0.6642,    Adjusted R-squared:  0.659
F-statistic: 129.1 on 11 and 718 DF,  p-value: < 2.2e-16
```

- The p-value of F test is very small, thus suggesting that the overall model is significant.
- R^2 is moderate as 0.6642 suggest that 66.42% of the variation in bike rental count is explained by the fitted regression. Adjusted R^2 = 0.659.
-

REPORT CONCLUSION

While my final model M3 is inadequate due to the violation of normality assumption, M3 is still the best model out of the 3 as it is less complex than the initial model, M1. Furthermore, M3 has the highest R^2 value out of all the 3 models proposed, which shows that higher variation of bike rental count can be explained by model M3. Additionally, model M3 also fixes the violation of homoscedasticity assumption that is prevalent in model M1.

Thus, after considering factors such as goodness-of-fit and complexity of the models, I feel that M3 is the best model in comparison to my other models.