University of Glasgow
School of Mathematics and Statistics

Investigating the Bike Sharing Dataset

By David Quinlan

Contents

1 Introduction

An investigation was conducted to investigate the effect of a number of different factors on the total number of hourly bike rentals of a bike sharing scheme during the years 2011 and 2012. It was of interest to examine which of the explanatory variables had a significant effect on the hourly number of bike rentals. The dataset contains 17379 observations in total with 13 variables associated with each observation. More details about each of the variables contained within this dataset are given as follows:

- **Cnt:** The total number of hourly bike rentals.
- **Dteday:** The date on which the bikes were rented.
- **Season:** A categorical variable ranging from 1 to 4 indicating the season in which the bikes were rented, where 1 is winter, 2 spring, 3 summer and 4 is autumn.
- **Yr:** A binary variable, where 1 indicates the year 2012 and 0 the year 2011.
- **Mnth:** The month of the bike rentals.
- **Hr:** The hour of the day in which the bikes were rented.
- **Holiday:** A binary variable where 1 indicates that day is a public holiday, 0 otherwise.
- **Weekday:** A categorical variable indicating the day of the week the bike rentals took place.
- **Weathersit:** A categorical variable were 1 represents a clear, few clouds or partly cloudly day, 2 a misty or cloudy day, 3 a day where there was light snow or rainfall and 4 a day where heavy rainfall or a thunderstorm occurred.
- **Temp:** The normalised temperature given in Celsius.
- **Atemp:** The normalised felt temperature given in Celsius.
- **Hum:** The normalised humidity.
- **Windspeed:** The windspeed at the time of rental.

The main aims of this investigation are:

- To conduct initial exploratory analysis on the bike sharing dataset.
- To fit a normal linear regression model to the dataset between the response variable Cnt and the remaining explanatory variables.
- To fit a Poisson regression model between the explanatory variables and the response variable Cnt.
- To fully evaluate and explain the final linear and Poisson models.

1.1 Initial Exploratory Analysis

The bike sharing dataset was loaded into R and cleaned prior to analysis. This process included the examination of the dataset's dimensions and calculation of each variables summary statistics. The index variable Dteday was removed from the dataset as it contained a date value for each observation. This variable was removed as it would add little to no additional information to any future models fitted to the dataset. Next, a series of pairs plots were created to examine the correlations between the continuous variables contained in this dataset, these plots can be seen below in Figure 1.

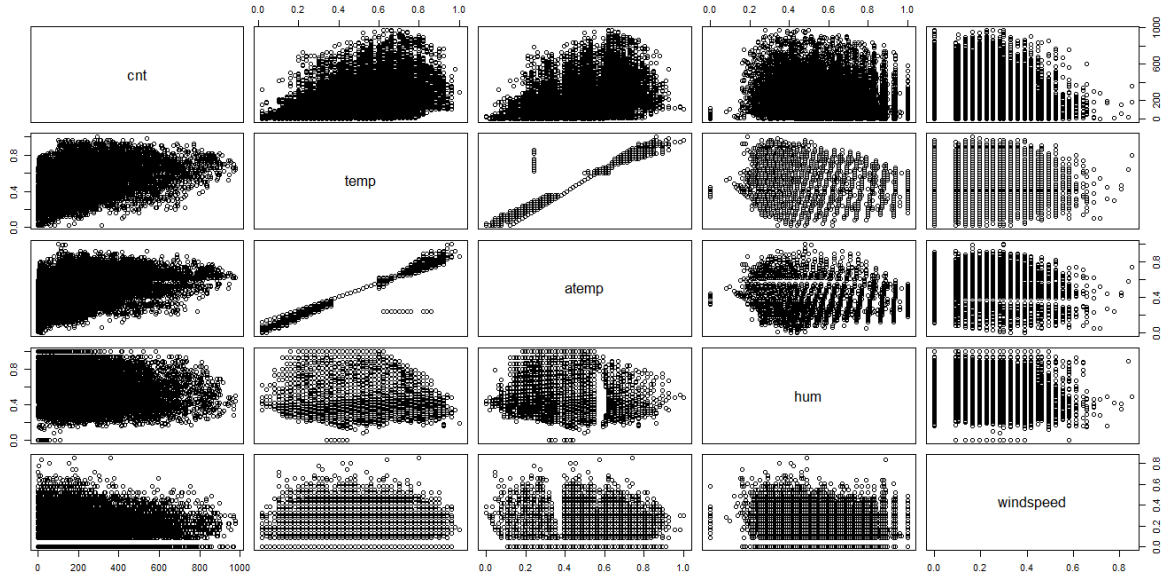

Figure 1: Pairs plot of the continuous variables contained within the bike rental dataset.

From the pairs plot it is clear that there is a strong positive relationship between the Temp and Atemp variables. It is difficult to distinguish any other significant relationships between variables from the pairs plot. Correlations between the continuous covariates were also assessed. Examining these results it was felt that the felt-temperature, or Atemp variable contained very similar information to the actual temperature variable, therefore, it was decided to remove the Atemp variable from the dataset. The following variables were treated as factor variables: Yr, Season, Month, Hour, Holiday, Weekday, and Weathersit.

As part of the initial exploratory analysis boxplots were created of the various continuous variables contained within the bike dataset. Boxplots provide a nice graphical summary of a variable and can also be used to identify possible outliers. These boxplots are shown below in Figure 2.

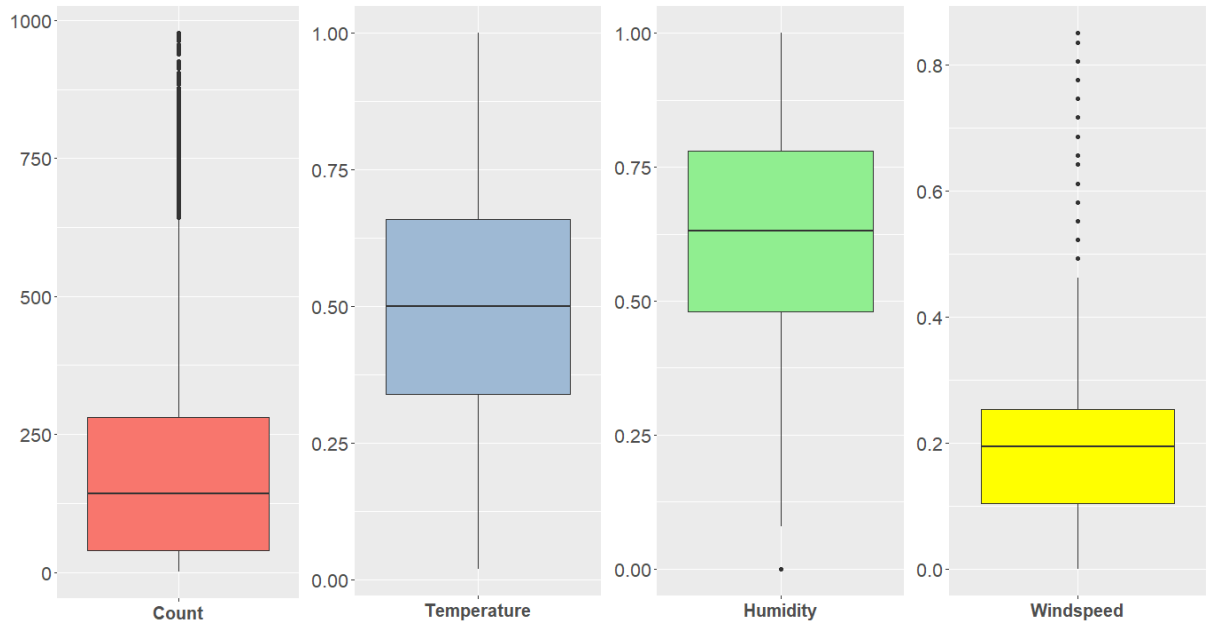

Figure 2: Boxplots of the remaining continuous variables contained within the bike rental dataset.

From the boxplots it is clear that Windspeed, Humidity and Counts all contain some possible outliers. However, given the size of the dataset and the fact that the dataset has been obtained from a real world source, outliers are to be expected. These outliers will not be removed for the moment, however, further analysis may be conducted on any possible outliers that appear to be influential in the linear or Poisson models fitted later in this investigation.

Following the creation of boxplots, histograms were created for each of the continuous variables. The mean and standard deviation of these continuous variables were also calculated with these values displayed above their corresponding histogram. These histograms are shown below in Figure 3.

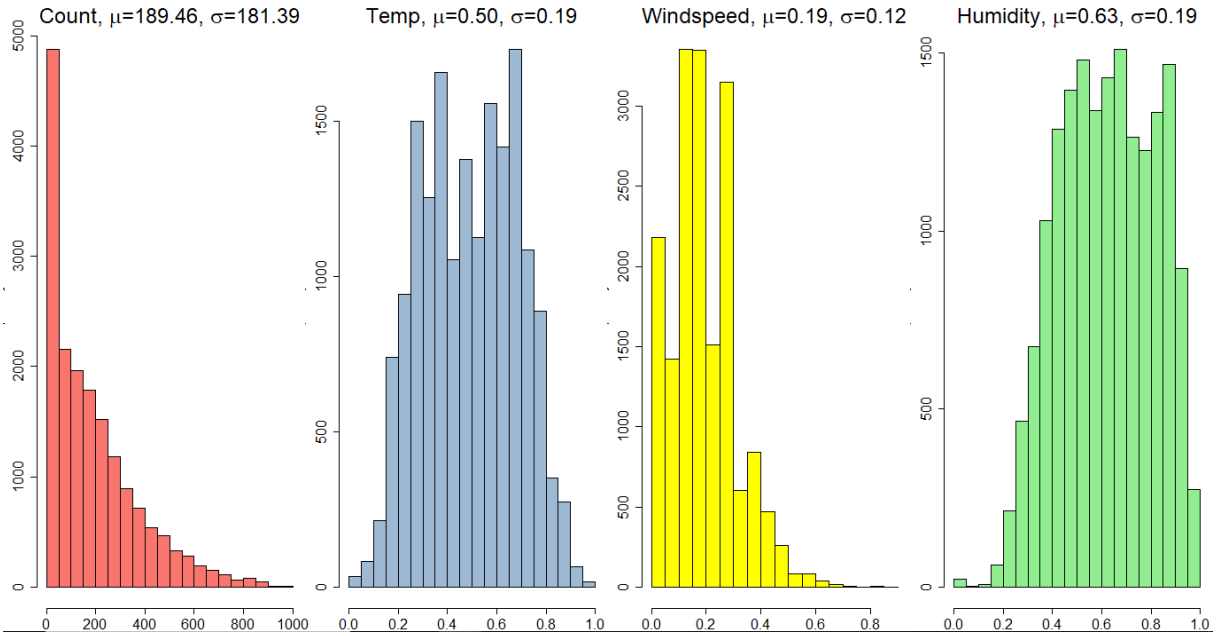

Figure 3: Histograms of the continuous variables contained in the bike rental dataset.

From the histograms it is clear to see that the variables Windspeed, Temp and Hum are approximately normally distributed, however, the Cnt variable appears to be skewed to the right. It may be necessary to apply some sort of transformation to the Cnt response variable to normalise the data when fitting a linear model to the dataset. This

is due to the fact that for a linear model, the response variable Y is assumed to be normally distributed. The fitting of a linear model between the non-transformed response variable and the corresponding explanatory variables may also violate the linear models assumption of constant variance.

Finally, boxplots of the categorical variables against the response variable, Cnt, were produced in order to explore the relationship between the different categorical explanatory variables and the response variable. One of the more interesting relationships was identified was between the categorical variable Hr and the response variable Cnt. The boxplots of the Hr variable against the response variable Cnt are shown below in Figure 4.

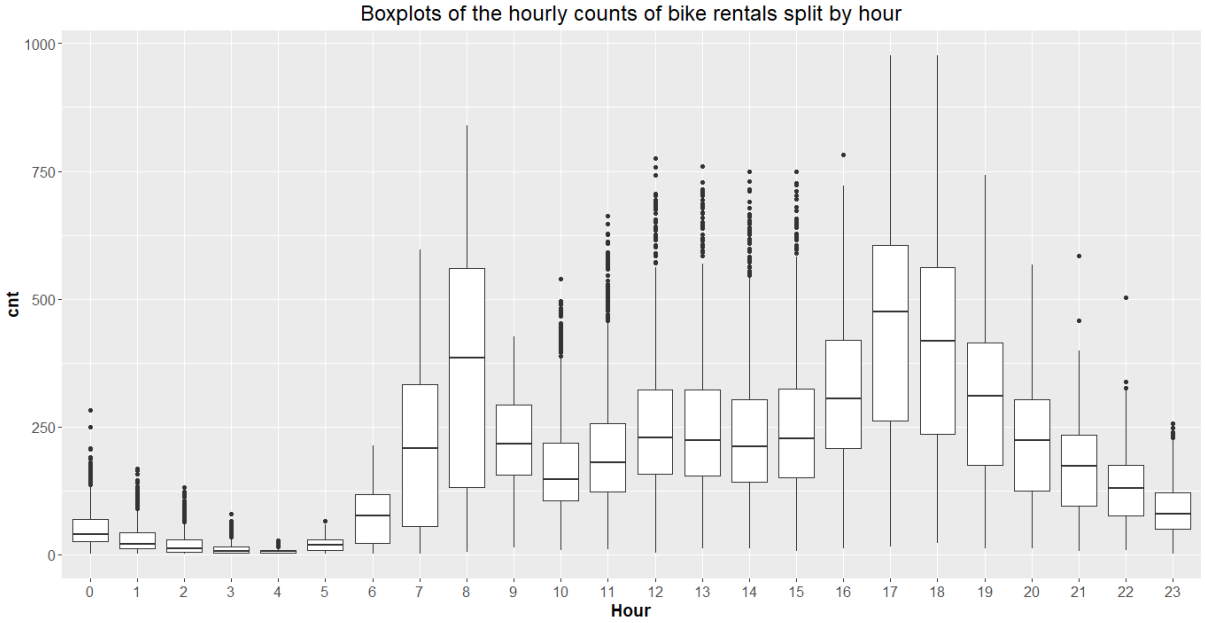

Figure 4: Boxplots of the Cnt variable split by Hour.

Examining the boxplots of the Cnt variable split by Hr, it is possible to identify different levels of the Hr variable which obtain similar numbers of bike rentals. Therefore, it was decided to reduce the number of levels contained within the Hr variable. When fitting models to data it can be very useful to reduce the number of levels in a factor variable, especially, if the reduced number of parameters explains the same level of information as the original variable. This process reduces the number of parameters in the model and can thus help to create a more parsimonious model.

To reduce the number of levels contained within the Hr variable it was decided to split the 24 hours into 5 different groups. These groups represented, early mornings (00:00-06:59), morning rush-hour (07:00-08:59), mid-day (09:00-16:59), evening rush-hour (17:00-19:59) and late evenings (20:00-23:59), with this new variable called Work. It is believe that if similar factors are grouped together, that the predictive power of the overall Hr variable would only be minimally reduced. Boxplots of the new explanatory variable Work against the response variable Cnt are shown below in Figure 5.

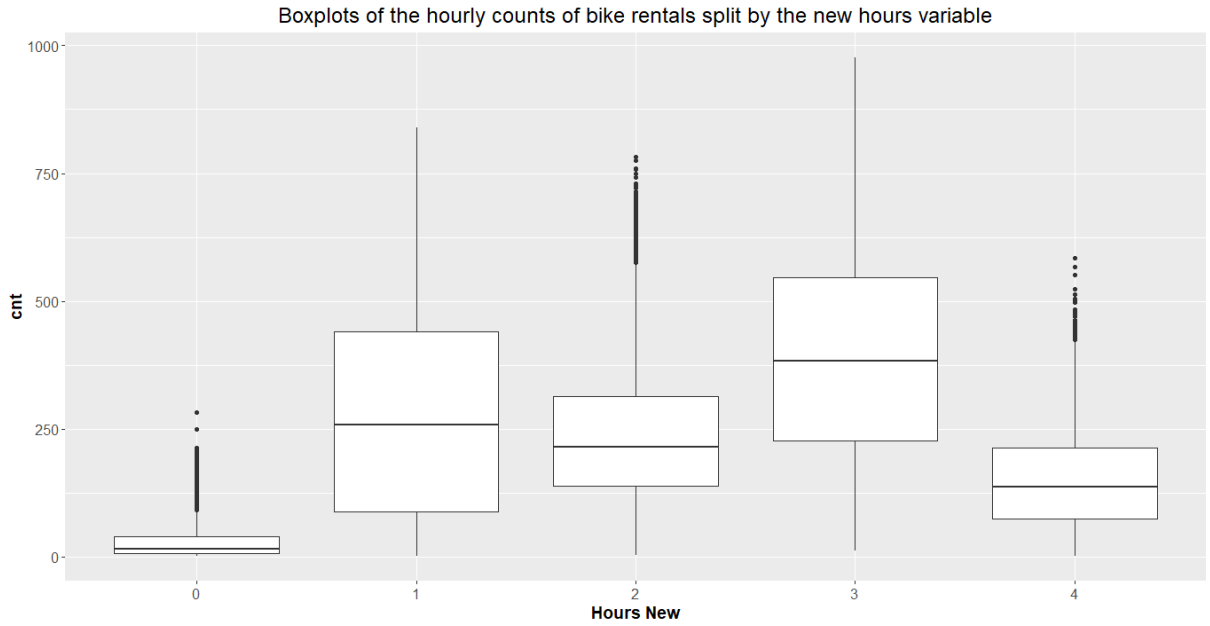

Figure 5: Boxplots of the Cnt variable split by the new Work variable.

Finally, it was felt that the information contained in the Month variable could be similarly explained by the Season variable. For this reason it was decided to remove the Month variable from the dataset. The variable Weekday was transformed into a binary variable Weekend where 1 represents the weekend and 0 the weekdays. It is believed that the feature engineering conducted above was justified as it should help to produced a simpler model both in structure and interpretation.

2 Methods

2.1 Multiple Linear Regression

A multiple linear regression model is a modelling approach whereby a linear relationship is fitted between a continuous response variable and a number of explanatory variables.

The linear regression model is given by the following formula:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip} + \epsilon_i, \quad \text{for } i = 1, ..., n \tag{1}$$

Where β_0 is the intercept term and $\beta_1, ..., \beta_p$ are the coefficients of the explanatory variables.

The linear regression assumptions are given as follows:

- There is a linear relationship between the response variable and the explanatory variables.
- The values $x_{i1}, ..., x_{ip}$ are fixed values.
- The explanatory variables are independent (i.e. no presence of multicollinearity).
- Error terms are independent and $N(0, \sigma^2)$ distributed with zero mean and constant variance σ^2.

2.2 Poisson Regression

A Poisson regression model is a specific form of a generalized linear model (GLM), whereby the conditional distribution of the response variable given the data is Poisson, i.e. $Y|x_1 + ... + x_p \sim Poisson(\lambda)$.

A Poisson regression model is given as follows:

$$\lambda = E(Y|x_1 + ... + x_p) = e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p} \tag{2}$$

Where β_0 is the intercept term and $\beta_1, ..., \beta_p$ are the coefficients of the explanatory variables. Taking logs of both sides, it is found that:

$$log(\lambda) = log(E(Y|x_1 + ... + x_p)) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{3}$$

This indicates how the natural logarithm represents the link function for a Poisson regression model. The assumptions of a Poisson regression model are given as follows:

- The response variable consists of count data.
- The distribution of the response variable conditional on the data follows a Poisson distribution, i.e. $Y|x_1 + ... + x_p \sim Poisson(\lambda)$.
- The mean and variance of the model are equal, i.e. $\lambda = E(Y|x_1 + ... + x_p) = Var(Y|x_1 + ... + x_p)$.
- Error terms are independent.
- Observations and variables are independent (i.e. there is no presence of multicollinearity).

It is possible to extend the Poisson regression model to allow for the modelling of the rate of occurrences of some variable of interest. Thus, it is possible to extend the model found above by including an offset parameter k, therefore:

$$log(E(Y|x_1 + ... + x_p)) = log(k) + \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{4}$$

Which equals:

$$log\Big(\frac{E(Y|x_1 + ... + x_p)}{k}\Big) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p \tag{5}$$

Where k is the offset parameter whose parameter estimate for $log(k)$ is constrained to 1.

2.3 Quasi-Poisson Regression

Over-dispersion is the phenomenon whereby the observed data has a much greater variance than expected given a certain statistical model. In the case of a Poisson regression model, over-dispersion occurs when the variance is much larger than the mean. The presence of over-dispersion violates the Poisson regression assumption that the mean and variance of a Poisson model are equal i.e. $\lambda = E(Y|x_1 + ... + x_p) = Var(Y|x_1 + ... + x_p)$.

It is possible to account for over-dispersion using a quasi-Poisson regression model. A quasi-Poisson model suggests an alternative mean-variance relationship structure for a Poisson regression model, this relationship is defined below:

$$\begin{aligned} E(Y|x_1 + ... + x_p) &= \lambda \\ Var(Y|x_1 + ... + x_p) &= \phi\lambda \end{aligned} \tag{6}$$

Where ϕ is some constant denoting the dispersion parameter.

Therefore, instead of having equal mean and variance, it is suggested that the variance for the Poisson distribution is represented by a product of λ and ϕ; where ϕ is a scaling parameter which is estimated from the observed data. A quasi-likelihood approach is used to estimate the models parameters without fully knowing the relationship between the mean and variance of the model. This approach is equivalent to an iteratively reweighted least squares procedure whereby the weights depend solely on the current parameter estimates. In the case of a quasi-Poisson model the weights are inversely proportional to the models variance and directly proportional to its mean [1] [2]. Apart from the different mean-variance relationship, all of the other Poisson regression models assumptions remain in the quasi-Poisson case.

3 Results and Discussion

3.1 Linear Regression

First a linear model was fitted to the bike rental counts data. This model, which will be defined as Model 1, fitted a linear relationship between the total rental counts response variable and all remaining explanatory variables, this includes the newly created variables Work and Weekend. It was found that all terms were statistically significant at a 5% level of significance. However, it was clear when examining the models diagnostic plots, that the residuals were not normally distributed nor did the linear model assumption of constant variance appear to hold. Therefore, it was felt that applying a transformation to the response variable Cnt might help solve the issue of non-constant variance.

A Box-Cox transformation was applied to the linear model and identified that the optimal transformation to apply to the response variable Cnt was either a log transform or a square root transform as the global maximum was found to be near 0. Therefore, it was decided to fit two linear models to the dataset. The first, defined as Model 2, fits a linear relationship between the explanatory variables and the log of the response variable. The second model fits a linear relationship between the explanatory variables and the square-root of the response variable and will be defined as Model 3.

Examining Model 2 or the linear model with a log transformed response variable, it was found that all the explanatory variables were statistically significant at a 5% level of significance bar the Weekend variable. However, when examining the diagnostic plots it was found that the residuals were not normally distributed with the assumption of constant variance also broken.

Examining the linear model with a square root transformed response variable or Model 3, it was found that all the explanatory variables were statistically significant at a 5% level of significance. The diagnostic plots of this model were also examined with these plots shown below in Figure 6.

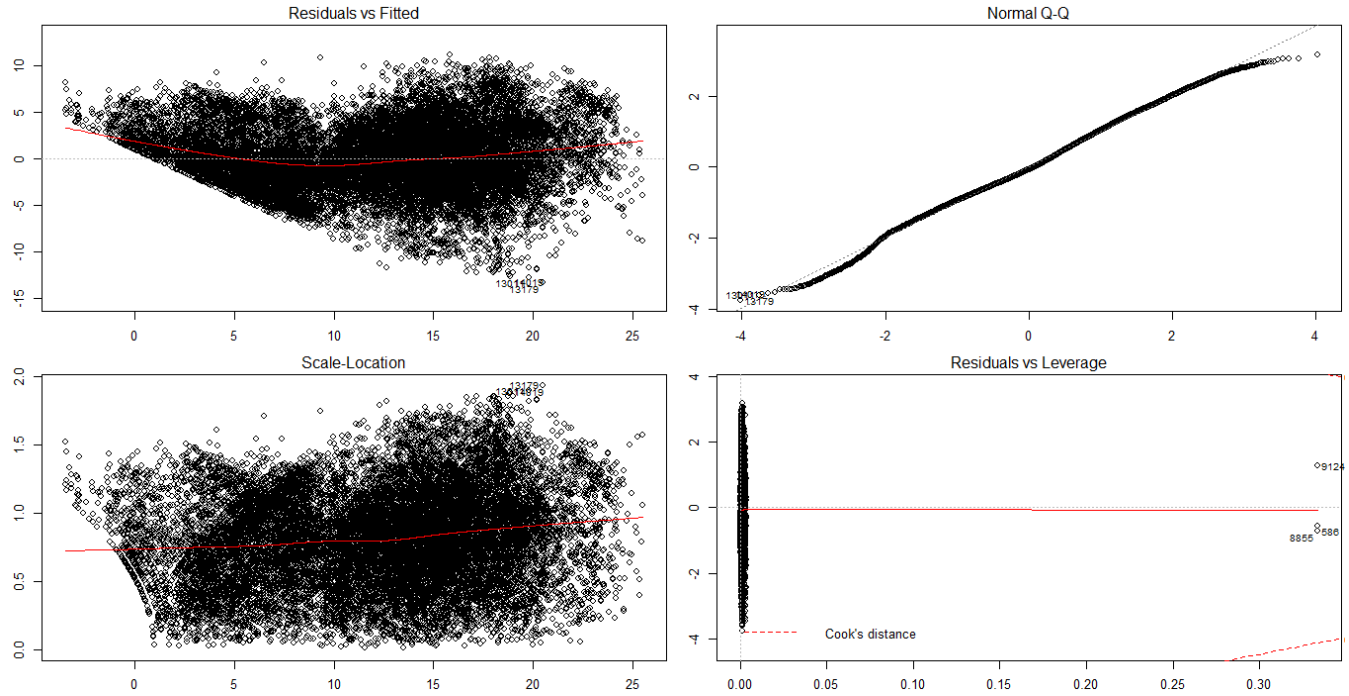

Figure 6: Diagnostic plots of the model $Cnt \sim Season + Yr + Holiday + Weekend + Weathersit + Temp + Windspeed + Work$.

Examining the QQ-plot it appears that the residuals are approximately normally distributed as the residuals follow the diagonal line. A similar conclusion can be made when examining the residuals vs fitted values plot where the residuals appear to be approximately scattered evenly above and below 0. However, the residuals form a strange diagonal pattern on the left hand side of the plot which may be a cause for concern. Examining the scale-location plot one can tell that the constant variance assumption holds with the residuals evenly distributed across the range

of residual values. When examining the residuals vs leverage plot the majority of values appear to not approach the cooks distance line. However, there does appear to be a small cluster of residuals which are much further away from the main group of residuals. Upon further investigation it was found that these residuals were associated with the counts of bike rentals which were identified as being rented during the three occasions when the weather was classified as stormy with heavy rain, therefore, since these three observations represent a specific group of individual cyclists they were not removed from the model. After examining the residuals of the square root transformed response linear model it appears that all of the linear models assumptions hold. Therefore, the final model is defined as:

$$\begin{aligned}\sqrt{Cnt} = 0.57 &+ 1.47x_{\text{Season2}} + 0.83x_{\text{Season3}} + 2.50x_{\text{Season4}} + 2.75x_{\text{Yr1}} - 1.12x_{\text{Holiday1}} - 0.07x_{\text{Weekend1}} \\ &- 0.22x_{\text{Weathersit2}} - 2.46x_{\text{Weathersit3}} - 0.56x_{\text{Weathersit4}} + 9.75x_{\text{Temp}} - 3.17x_{\text{Hum}} - 0.90x_{\text{Windspeed}} \\ &+ 10.68x_{\text{Work1}} + 8.66x_{\text{Work2}} + 12.96x_{\text{Work3}} + 6.31x_{\text{Work4}}\end{aligned} \tag{7}$$

It is now possible to fully interpret and explain the above linear regression model in more detail:

- **Intercept:** The intercept in this model is defined as the null model where all parameter values are equal to 0. In this case the intercept is given as the number of bike rentals during the winter of 2011 on a non-public holiday during the week between the hours of 00:00 and 06:00 when the weather is classed as being clear and partly cloudy. The intercept coefficient value is given as 0.57, thus, the square root of the number of bikes rented during the early morning of a partly cloudy working weekday during the winter of 2011 is on average between 0 and 1.
- **Season:** While all other variables remain constant the square root of the number of rented bikes increases on average by 1.47 during Spring, increases on average by 0.83 during the Summer and increases on average by 2.50 during the Autumn season.
- **Yr:** It was found that there was an average increase of 2.75 in the square root of the number of rented bikes during the year 2012, provided all other variables remain constant.
- **Holiday:** On public holiday days the square root of the number of bikes rented decreases by 1.12 on average, provided all other variables remain constant.
- **Weekend:** Provided all other variables remain constant, on average, the square root of the number of rented bikes during the weekend was found to decrease by 0.07.
- **Weathersit:** While all other variables remain constant the square root of the number of rented bikes decreases on average by 0.22 when the weather is misty or cloudy, decreases on average by 2.46 during light snow or rain showers and decreases on average by 0.56 during thunderstorms, stormy weather and when there is the occurrence of heavy rainfall.
- **Temp:** For a one unit increase in the normalised temperature, the square root of the number of rented bikes increases on average by 9.75, provided all other variables remain constant.
- **Hum:** For a one unit increase in the normalised humidity, the square root of the number of bike rentals decreases by 3.17 on average, provided all other variables remain constant.
- **Windspeed:** A one unit increase in windspeed decreases, on average, the square root number of rented bikes by 0.90, provided all other variable remain constant.
- **Work:** When all other variables are held constant the square root of the number of rented bikes increases on average by 10.68, 8.66, 12.96, 6.31 for the hour ranges of 07:00-08:59, 09:00-16:59, 17:00-20:59 and 21:00-23:59, respectively.

A table of the above linear models coefficients, standard errors and p-values is shown below in Table 1. From the above analysis one can see that the Work and Temp variables have the most influence on the square root of the number of hourly rented bikes. A negative aspect of this model is that it is not very easy to interpret. For example, a unit increase in one of the explanatory variables causes an increase/decrease in the square root of the number of rented bikes which is not always easy to interpret. However, this interpretation is necessary as $E(\sqrt{y}) \neq \sqrt{E(y)}$, therefore, it is incorrect to simply square the coefficients of the explanatory variables when interpreting the model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.5690	0.1667	3.41	0.0006
season2	1.4744	0.0965	15.27	0.0000
season3	0.8305	0.1244	6.68	0.0000
season4	2.5061	0.0838	29.89	0.0000
yr1	2.7503	0.0539	51.04	0.0000
holiday1	-1.1227	0.1610	-6.97	0.0000
weekend1	-0.0656	0.0595	-1.10	0.2699
weathersit2	-0.2172	0.0661	-3.28	0.0010
weathersit3	-2.4586	0.1108	-22.20	0.0000
weathersit4	-0.5604	2.0355	-0.28	0.7831
temp	9.7541	0.2400	40.64	0.0000
hum	-3.1663	0.1815	-17.44	0.0000
windspeed	-0.8967	0.2352	-3.81	0.0001
work1	10.6775	0.1050	101.66	0.0000
work2	8.6593	0.0780	110.96	0.0000
work3	12.9591	0.0985	131.53	0.0000
work4	6.3075	0.0840	75.10	0.0000

Table 1: Table of the coefficients, standard errors and p-values associated with each of the parameters from the square root transformed linear model fitted to the bike rental dataset.

3.2 Poisson Regression

After fitting a linear model it was important to examine and fit other modelling approaches to the dataset. The response variable Cnt which represents the total number of bikes rented during a specific hour contains strictly positive count values. Examining the assumptions of a Poisson regression model it appears that this model is an appropriate choice to fit the relationship between the response and the explanatory variables. A Poisson regression model was fitted between response variable Cnt and the explanatory variables Season, Year, Holiday, Weekend, Weathersit, Temp, Hum, Windspeed and Work.

After fitting the model it was found that all of the parameter estimates were statistically significant at a 5% level of significance. Before examining this model in more detail it is important to ensure that the Poisson regression assumption of equal model variance and mean holds. Therefore, it was decided to test for over-dispersion using the dispersion test. The null hypothesis of the dispersion test states that there is no presence of dispersion in the model, while, the alternative hypothesis states that there is evidence of over-dispersion in the model. A p-value of 2.2e-16 was obtained indicating that we reject the null hypothesis and conclude that there is evidence to suggest the presence of over-dispersion in the model. The presence of over-dispersion suggests that a Poisson regression is not an appropriate model and therefore another type of model should be fitted to the dataset. More suitable models include using a quasi-Poisson model and a negative binomial model both of which have an additional parameter which account for any dispersion present in the dataset.

After discovering that the data was over-dispersed a quasi-Poisson model was fitted to the dataset between the response variable and the explanatory variables mentioned above. It was found that all explanatory variables bar the Windspeed variable were statistically significant at a 5% level of significance. Therefore, the Windspeed variable was removed and the new model was fitted to the dataset. This time all parameters were statistically significant at a 5% level of significant. The dispersion parameter was estimated to be 45.83, this value is much greater than 1, the dispersion parameter for a Poisson regression model. After fitting the quasi-Poisson model to the dataset it is important to ensure that all of the quasi-Poisson model assumptions hold. Diagnostic plots of the quasi-Poisson model fitted above were created and are shown below.

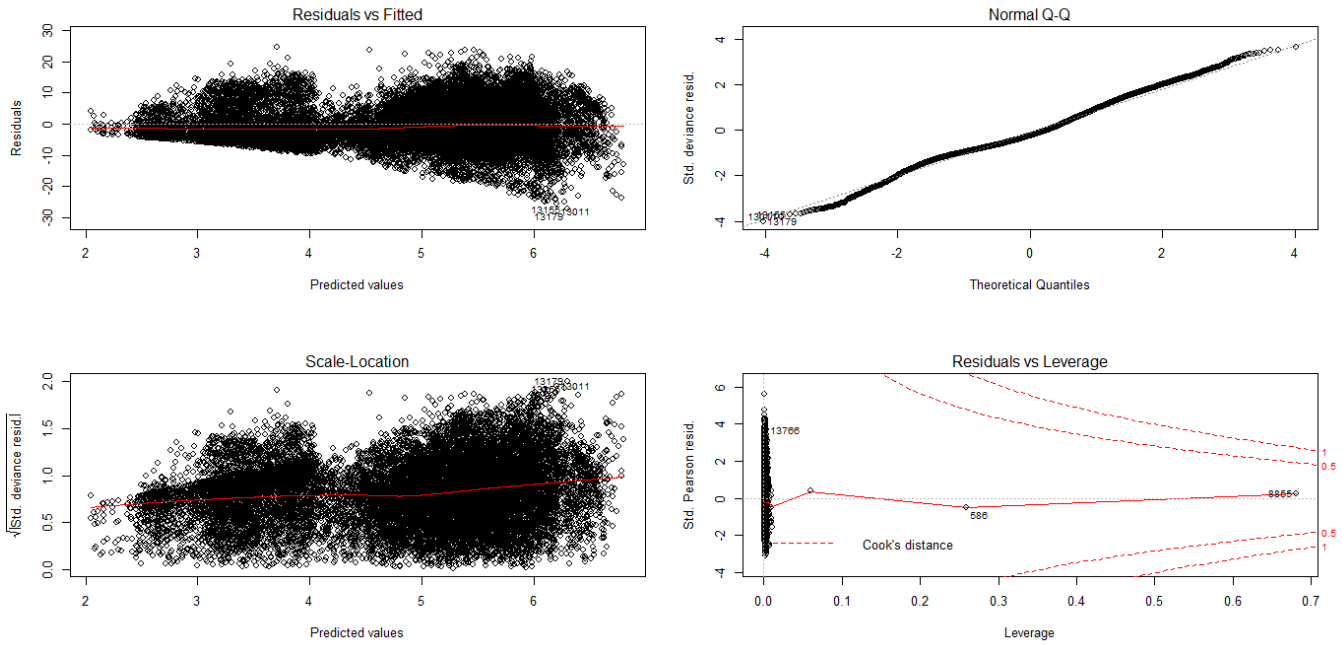

Figure 7: Diagnostic plots of the quasi-Poisson model: $Cnt \sim Season + Yr + Holiday + Weekend + Weathersit + Temp + Hum + Work$.

Examining the diagnostic plots above it is clear that the residuals are approximately normally distributed when examining the QQ-plot as the residuals follow the 45 degree line. Although, the normality of residuals for a Poisson (or quasi-Poisson) regression model is not one of the models assumptions, it is expected, given the large number of observations in the dataset, that the residuals would tend to a normal distribution, for a correctly specified model. When examining the residuals vs fitted value plot one can see that the residuals are randomly scattered above and below 0, with the fitted LOWESS smoother approximately following the 0 mean line. Both of these plots would suggest that the residuals are approximately normal distributed. However, a cause for concern is that the residuals do not appear to be evenly scattered above and below 0 across all of the predicted values, this may suggest that the constant variance assumption may have been broken. When examining the scale-location plot it appears that the constant variance assumption holds as the variance of the residuals seems to be even across all of the predicted values. The red smooth line is flat which suggests that the variance remains constant across all of the residual values.

Finally, possible outliers can be identified when examining the residual vs leverage plot. Observations 8865 and 596 seem to deviate much further away from the majority of the other observations. They also appear to have a larger leverage than the other observations, however, neither observation obtained a large Cook's distance. As previously mentioned above the values 8865 and 596 form part of a group of only 3 observations where bike rental counts were recorded during stormy weather. For this reason, it was felt that this group of observations should not be removed from the dataset. Therefore, the above model is identified as our final quasi-Poisson model and is defined as follows:

$$\begin{aligned} log(Cnt) = {} & 2.60 + 0.31x_{\text{Season2}} + 0.21x_{\text{Season3}} + 0.47x_{\text{Season4}} + 0.46x_{\text{Yr1}} - 0.19x_{\text{Holiday1}} - 0.02x_{\text{Weekend1}} \\ & - 0.05x_{\text{Weathersit2}} - 0.44x_{\text{Weathersit3}} - 0.36x_{\text{Weathersit4}} + 1.33x_{\text{Temp}} - 0.31x_{\text{Hum}} - 2.19x_{\text{Windspeed}} \\ & + 2.19x_{\text{Work1}} + 1.82x_{\text{Work2}} + 2.33x_{\text{Work3}} + 1.48x_{\text{Work4}} \end{aligned} \tag{8}$$

In the following section the final quasi-Poisson models coefficients will be fully explained and interpreted:

- **Intercept:** The intercept in this model is defined as the null model where all parameter values are equal to 0. In this case the intercept is given as the number of bike rentals during the winter of 2011 on a non-public holiday during the week between the hours of 00:00 and 06:00 when the weather is classes as being clear and partly cloudy. The intercept coefficient value is given as 13.4, thus, the number of bikes rented during the early morning of a partly cloudy working weekday during the winter of 2011 is between 13 and 14 rented bikes.

- **Season:** While all other variables remain constant the number of rented bikes increases by 35.9% during Spring, increases by 23.7% during the Summer and increases by 59.2% during the Autumn season.
- **Yr:** It was found that there was a 58.8% increase in the number of rented bikes during the year 2012, when all other variables remain constant. This would suggest that the bike rental scheme is getting more popular as time progresses.
- **Holiday:** On public holiday days the number of rented bikes decreased by 17.4%, provided all other variables remain constant.
- **Weekend:** Provided all other variables remain constant, the number of rented bikes during the weekend was found to decrease by 2.3%.
- **Weathersit:** While all other variables remain constant the number of rented bikes decreases by 4.5% when the weather is misty or cloudy, decreases by 35.7% during light snow or rain showers and decreases by 30.4% during thunderstorms, stormy weather and when there is the occurrence of heavy rainfall.
- **Temp:** For a one unit increase in the normalised temperature the number of rented bikes increases by a multiplicative factor of 3.80, provided all other variables remain constant.
- **Hum:** For a one unit increase in the normalised humidity the number of bike rentals decreases by 26.8%, provided all other variables remain constant.
- **Work:** When all other variables are held constant the number of rented bikes increases by a multiplicative factor of 8.90, 6.15, 10.25, 4.40 for the hour ranges of 07:00-08:59, 09:00-16:59, 17:00-20:59 and 21:00-23:59, respectively. Therefore, this model seems to indicate that the number of bike rentals increases during the hours when individuals are commuting to and from work, with less demand for bike rentals found during the hours of the working day or during the late evening time.

It is interesting to note how the number of bike rentals changes with respect to the different explanatory variables. For example, increases in bike rentals are seen during the hours just before and immediately after work. Also the number of rented bikes decreases on weekends and public holidays, this would suggest that the majority of bike rental users are using the bikes to commute to and from work. It is also interesting to note that the model has identified some factors which may reduce the number of bike rentals. These include poor weather conditions and the time of the year, both of these reductions would make sense as less people are likely to cycle for example when it is raining or during the winter when it is cold.

A table of each parameters associated coefficients, standard errors and p-values is shown below in Table 2.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.5981	0.0280	92.88	0.0000
season2	0.3070	0.0149	20.60	0.0000
season3	0.2129	0.0181	11.76	0.0000
season4	0.4652	0.0131	35.45	0.0000
yr1	0.4629	0.0077	59.73	0.0000
holiday1	-0.1916	0.0246	-7.78	0.0000
weekend1	-0.0237	0.0084	-2.83	0.0047
weathersit2	-0.0459	0.0095	-4.81	0.0000
weathersit3	-0.4416	0.0190	-23.27	0.0000
weathersit4	-0.3628	0.4537	-0.80	0.4238
temp	1.3350	0.0337	39.66	0.0000
hum	-0.3129	0.0252	-12.39	0.0000
work1	2.1862	0.0199	110.09	0.0000
work2	1.8167	0.0187	96.96	0.0000
work3	2.3276	0.0193	120.83	0.0000
work4	1.4807	0.0198	74.71	0.0000

Table 2: Table of the coefficients, standard errors and p-values associated with each of the parameters from the quasi-Poisson model fitted to the bike rental dataset.

3.3 Other Modelling Strategies

A negative binomial model was also fitted to the dataset, modelling the relationship between the response variable Cnt and the explanatory variables examined above using the Poisson and quasi-Poisson models. It was felt that this model was not an appropriate model for the dataset as the constant variance assumption was violated. Similarly, a Poisson generalised additive model (GAM) was also fitted to the dataset fitting smooth curves to the continuous variables Hum, Temp and Windspeed while also including the remaining categorical variables. This model was found to have an R^2 value of 0.77 with a deviance explained of 81.4%. However, examining this models diagnostic plots it was found that the residuals were not normally distributed which as mentioned above should probably be the case given the number of observations contained within the dataset.

4 Conclusion

After fitting multiple types of models to the bike rental dataset, two models were identified as providing a good fit to the data. The first of these models was a linear regression model with a square root transformed response variable, the second model was quasi-Poisson model. Both of these models appear to provide a good fit to the dataset. However, the quasi-Poisson model was identified as being the optimal model as it is far easier to interpret than the square root transformed linear model. Also from a theoretical standpoint the quasi-Poisson model would appear to be better suited to model the Cnt response variable given that it contains positive count bike rental data.

The quasi-Poisson regression model provides some very interesting insights into the dataset. For example, it was identified that the majority of bike rentals occur during weekdays directly before and after work, this would suggest that the majority of the bikes are being rented by individuals undertaking their daily commutes. It was also found that the time of year and the weather played a major role in the numbers of bike rentals, with poorer weather more likely to lead to lower levels of bike rentals. All of these results make sense which is a sign that the model is a good fit to the dataset.

It would be interesting to see if we could further improve our analysis by collecting more data. I believe that this model could be improved by examining the spatial aspect of bike rentals. For example, the level of public transport of an area or the number of employees working in a certain district could influence the number of bike rentals in those locations. Therefore, it is felt that if such spatial information was available that this might help further improve the accuracy of the model.

References

[1] P McCullagh et al. Nelder. ja (1989), generalized linear models. *CRC Monographs on Statistics & Applied Probability, Springer Verlag, New York*, 1973.

[2] Jay M Ver Hoef and Peter L Boveng. Quasi-poisson vs. negative binomial regression: How should we model overdispersed count data? *Ecology*, 88(11):2766–2772, 2007.