ΟΙΚΟΝΟΜΙΚΟ
ΠΑΝΕΠΙΣΤΗΜΙΟ
ΑΘΗΝΩΝ

ATHENS UNIVERSITY
OF ECONOMICS
AND BUSINESS

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS

DEPARTMENT OF MANAGEMENT, SCIENCE & TECHNOLOGY

MSc BUSINESS ANALYTICS

“Main Assignment in course: Statistics for Business Analytics I”

Full Name: STAMATIOS SIDERIS

Register Number: f2822113

ATHENS, 2022

Table of Contents

1. Introduction

The main purpose of the assignment is to understand what influences bike rentals hourly and predict them to satisfy demand. To do so, we will create a statistical model based on data from the dataset “bike_13.csv” and we will test our model using the dataset “bike_test.csv”. For the purposes of the assignment to highlight the use of multiple linear regression on a dependent variable that follows the Normal Distribution, the multiple linear regression is used, although the most appropriate distribution would be the Poisson Distribution as bike rentals is a discrete non-negative number. The datasets “bike_13.csv” and “bike_test.csv” are random sub-samples extracted from a combination of datasets:

- Core dataset is related to the, aggregated on hourly basis, two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA which is publicly available in http://capitalbikeshare.com/system-data
- Weather information are extracted from http://www.freemeteo.com

2. Descriptive analysis and exploratory data analysis

The dataset “bike_13.csv” consists of 1500 observations of 17 variables of which 13 are of class integer, 4 are of class numeric and 1 of class character. No NAs and blank values exist. The variable names and descriptions are:

- instant (int): record index
- dteday (chr): date
- season (int): season (1: winter, 2: spring, 3: summer, 4: autumn)
- yr (int): year (0: 2011, 1:2012)
- mnth (int): month (1 to 12)
- hr (int): hour (0 to 23)
- holiday (int): day (0: not holiday, 1: holiday)
- weekday (int): day of the week (0: Sunday to 6: Saturday)
- workingday (int): day (0: weekend or holiday, 1: working day)
- Weathersit (int): Possible outcomes

  1: Good: Clear, Partly cloudy

  2: Moderate: Cloudy/Broken clouds/Few clouds/Misty

  3: Bad: Light Rain + Thunderstorm + Scattered clouds/Light Rain + Scattered clouds

  4: Worse: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp (num): Normalized temperature in Celsius. The values are divided to 41 (max)
- atemp (num): Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- hum (num): Normalized humidity. The values are divided to 100 (max)
- windspeed (num): Normalized wind speed. The values are divided to 67 (max)
- casual (int): count of casual users

- registered (int): count of registered users
- cnt (int): count of total rental bikes including both casual and registered (response)

First, we exclude variable instant as it adds no material information to the dataset and variable dteday as it can be described by variables yr, mnth and weekday. Also, we exclude variables registered and casual as they include the count of bike rentals which we want to predict affecting the fitting of our final model. Afterwards, we change the class of variables season, yr, mnth, hr, holiday, weekday, weathersit to factor and the rest of integer variables to numeric. By calculating the descriptive measures of numeric variables in **Table 1**, it is observed that:

Variables temp, atemp, hum, windspeed have equal mean and median indicating their data are following the Normal Distribution. This means that 99.7% of their data would be spread |3σ| around the mean. Specifically, temp data would be spread |3 x 0.19|, atemp |3 x 0.17|, hum |3 x 0.19| and windspeed |3 x 0.12| around the mean.

The skewness of variables windspeed and cnt is positive indicating right skewed distributions and data concentrated to lower values while the skewness of variables temp, atemp and hum is almost zero which is another indicator of Normal Distribution.

Variables temp, atemp and hum have higher negative kurtosis and variable cnt higher positive kurtosis exhibiting tail data that exceed the tails of the normal distribution.

Table 1 *The descriptive measures of numeric variables*

	temp	atemp	hum	windspeed	cnt
vars	1.00	2.00	3.00	4.00	5.00
n	1500.00	1500.00	1500.00	1500.00	1500.00
mean	0.50	0.48	0.62	0.19	191.54
sd	0.19	0.17	0.19	0.12	181.61
median	0.50	0.48	0.62	0.19	151.00
trimmed	0.50	0.48	0.62	0.19	164.68
mad	0.24	0.20	0.24	0.13	177.91
min	0.02	0.03	0.00	0.00	1.00
max	1.00	0.91	1.00	0.66	968.00
range	0.98	0.88	1.00	0.66	967.00
skew	0.00	-0.07	-0.02	0.50	1.22
kurtosis	-0.96	-0.91	-0.88	0.31	1.26
se	0.01	0.00	0.00	0.00	4.69

To emphasize more in variables practical interpretation, we create, in **Figure 1**, the histogram of each variable. It is observed that temperature is balanced in Washington D.C. throughout the years 2011 and 2012, with medium to high levels of humidity and medium to low speeds of wind. Approximately, 200 rentals of bikes happen every hour from which about 25% are from casual users.

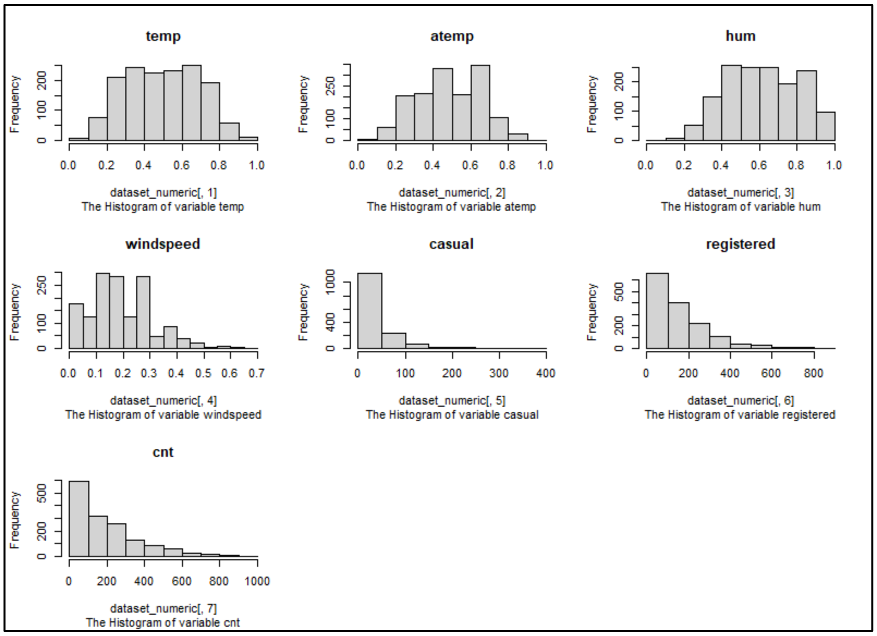


Figure 1 The Histograms of numeric Variables

Moreover, we create the barplots of binary factor variables in **Figure 2**. We observe that only a few holidays are included in our sample while days are balanced between 2011 and 2012. 70% of the days are working days.

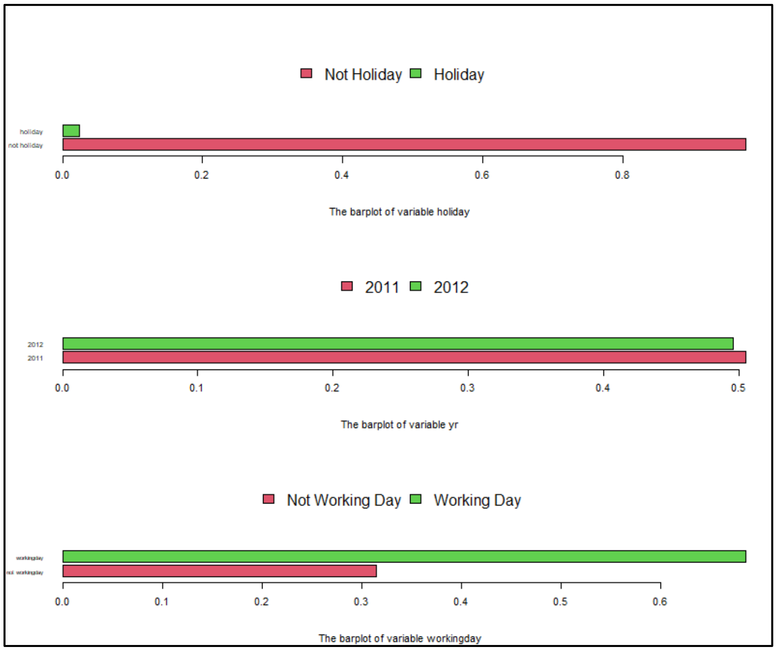


Figure 2 The barplots of binary factor variables holiday, yr and workingday

Finally, we create the Plots for the rest of the factor variables in **Figure 3**. It is observed that, most rentals happen in spring and summer, during months April, May and July, on days Monday, Wednesday and Saturday and especially on day times 3-5 am, 10-12 pm, 17-19 pm. The weather is mostly good and clear on those days.

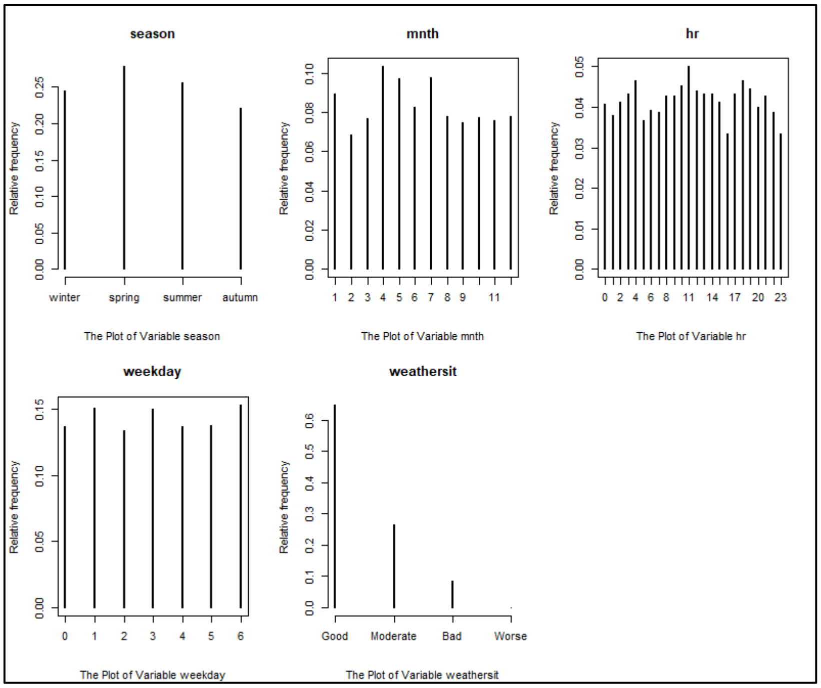


Figure 3 Plots of factor variables season, mnth, hr, weekday and weathersit

3. Pairwise comparisons

We create the correlation Plot and check the correlations between the numeric variables in **Figure 4**. It is observed that temperature and feeling temperature are highly correlated to each other as their correlation is almost 1. We must take it into consideration to avoid multi-collinearity in our final model. It seems that number of rentals is low correlated to the rest of independent variables as their correlations are much lower than 0.7 but they might play a critical role on our final fitted model.

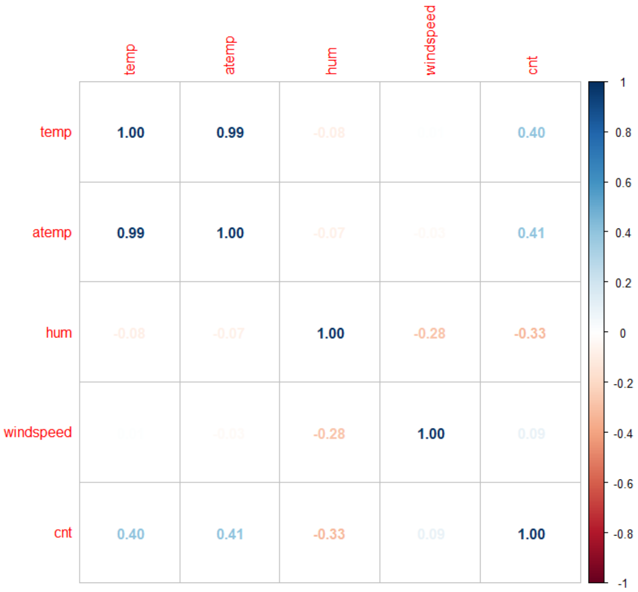


Figure 4 The correlation plot for numeric variables

We create the plots of bike rentals against the numeric variables and draw the linear regression line on them in **Figure 5**. It is observed that bike rentals per hour increase on moderate temperatures around 25°C and 50-60% humidity while decrease as windspeeds get higher.

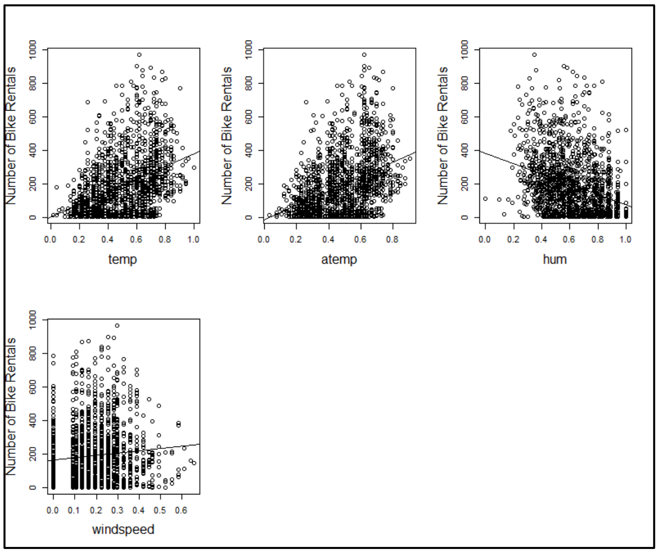


Figure 5 The Plots of numeric variables

We create the boxplots of factor variables in **Figure 6**. It is observed that, bike rentals increased in 2012. Weather plays a critical role for bike rentals as users prefer to rent a bike on days with good or slightly bad weather. This happens mostly in seasons spring, summer, autumn and their included months. Users do not seem to have a specific preference between days of the week or holidays, but they clearly prefer renting a bike early in the morning and evening around 7-9 am and 17-19 pm.

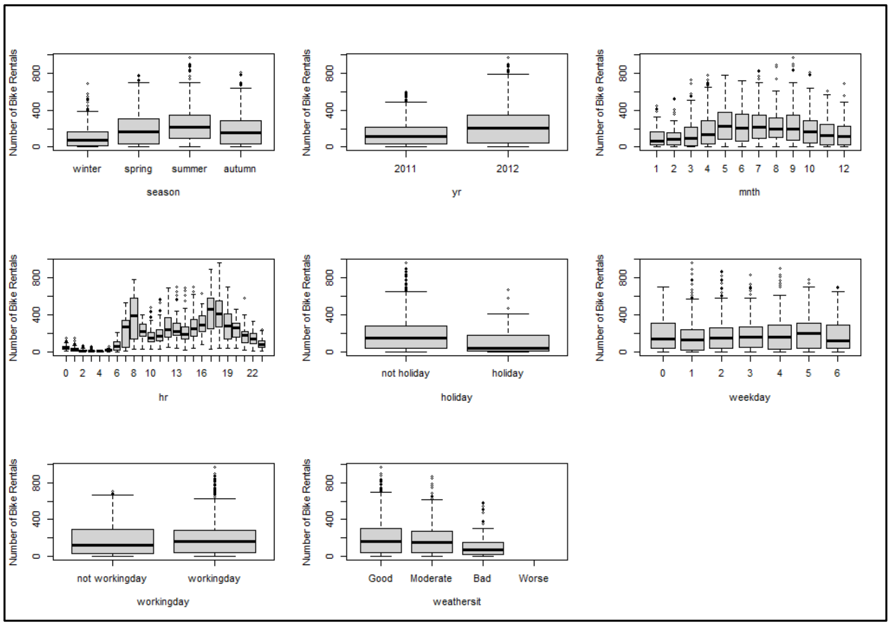


Figure 6 The boxplots of factor variables

4. Predictive or Descriptive Models

We compare the constant model (includes only the intercept as independent variable) to the full model (includes the intercept and all the independent variables of our dataset) via an ANOVA test. It results that p-value is lower than 0.05 so we reject hypothesis H0 that the constant model is statistically more significant than the full model. We will base our final model on the full model. (Anova's p-value = < 2.2e-16 < 0.05); see **Table 2** in Appendix for details.

We implement Lasso to minimize the number of covariates for our model. To find a reasonable value for λ, we use cross validation. We choose a grid of λ values and compute the cross-validation error rate for each value of λ. We then select the tuning parameter value for which the cross-validation error is smallest and choose the λ that is 1 standard deviation from the minimum value of λ to be more parsimonious and avoid overfitting. In **Figure 7**, are presented the variables' coefficients as log λ increases. The intermittent lines show the minimum log λ (on the left) and the log λ 1se from minimum (on the right). For a λ 1se from minimum equal to 2.13, variables holiday, weekday, temp and windspeed are excluded from our model as they are penalized the most.

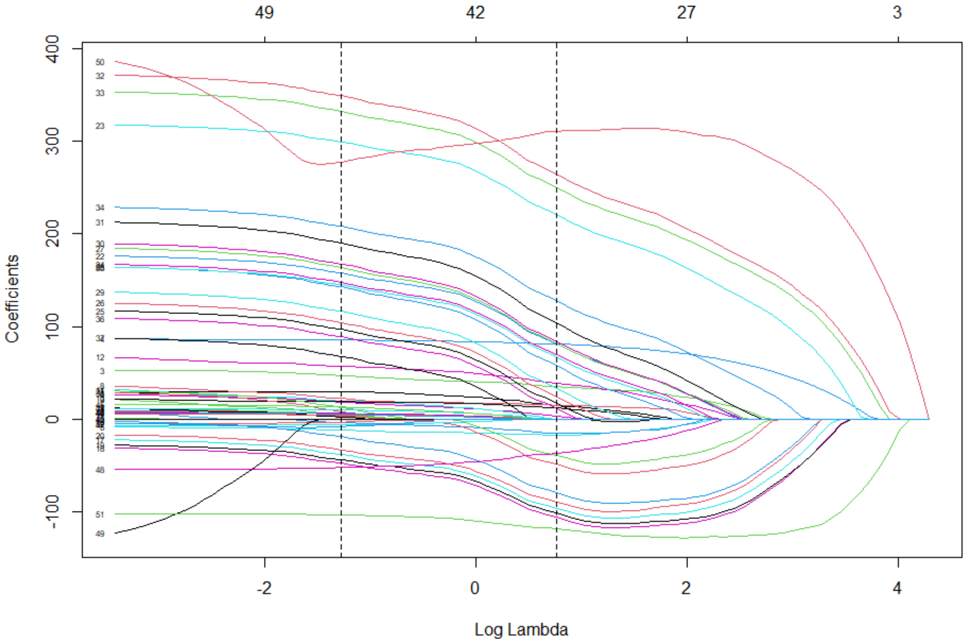


Figure 7 The variables' coefficients as Log Lambda increases

We proceed by implementing Stepwise Procedures to find the best covariates for our model. We choose both ways and AIC criterion as it is preferred for predictive models instead of BIC which is preferred for interpretation models. The results, as in **Table 3**, show that any covariates are excluded from our model as <none> is the one with the lowest AIC value. Lasso implementation has excluded all needed to be excluded covariates.

Table 3 Step Wise Procedure

Start: AIC=13959.73
cnt ~ season + yr + mnth + hr + weathersit + atemp + hum

	Df	Sum of Sq	RSS	AIC
<none>			15593673	13960
-season	3	143907	15737581	13968
-mnth	11	381437	15975110	13974
-weathersit	2	267967	15861640	13981
-hum	1	286049	15879722	13985
-atemp	1	584864	16178538	14013
-yr	1	2681802	18275475	14196
-hr	23	17233369	32827042	15030

The model, now, includes the intercept and variables season, yr, mnth, hr, weathersit, temp and hum against the dependent variable cnt with an adjusted R-squared of 0.67 meaning that 67% of the variance of the dependent variable is explained by our model, as it appears in **Table 4** (see Appendix for details). The residual standard error is high and equal to 103.5 meaning that residuals have a high variance of 103.5^2 around the mean. We want to improve our model so that the variance of dependent variable is decreased and explained by at least 70% by our model.

We perform audit on multi-Collinearity and residuals Assumptions based on linear regression to check for problems and improve our model:

We check the variance inflation factors to detect multi-Collinearity between our independent variables in **Table 5**. As our model includes factors with over 2 levels of data, the generalized variance inflation factors are used. All variables are below 3.16 and as a result no multi-Collinearity is detected.

Table 5 Generalized Variance Inflation Factors

	GVIF	Df	GVIF^(1/(2*Df))
season	229.8	3	2.5
yr	1	1	1
mnth	477.4	11	1.3
hr	2.1	23	1
weathersit	1.4	2	1.1
atemp	5.2	1	2.3
hum	1.9	1	1.4

We perform normality test of residuals by using the Normal QQ Plot of standardized residuals against the theoretical quantiles in **Figure 8, Diagram 1**. It is observed that the points fall along a line in the middle of the graph but curve off in the extremities. This probably means that residuals have more extreme values than would be expected if they truly came from a normal distribution.

We perform non-constant variance test to check for linear effect on the variance of errors. We reject hypothesis Ho that residuals have constant variance (ncv's p-value=<2.2e-16 < 0.05; see **Table 6** in Appendix for details). We assure our results by observing the QQplot of Studentized residuals against the fitted values in **Figure 8, Diagram 2**. It is observed that as the predicted value increases, the residual errors variance increases which indicates a problem of heteroscedasticity.

We perform the Tukey test to check our model for non-linearity. We reject hypothesis Ho that the quadratic term is equal to zero (Tukey's test p-value=<2e-16 < 0.05; see **Table 7** in Appendix for details). We assure our results by creating the QQplot of Rstudent residuals against the fitted values in **Figure 8, Diagram 3**. It is observed that Rstudent residuals are following a bell-shaped curve against the fitted values.

We perform the Durbin Watson test to check for autocorrelation of residuals. We reject hypothesis Ho that residuals are not autocorrelated and we keen on the possibility that the sequence was not produced in a random manner (Durbin Watson's test p-value=0.036<0.05; see **Table 8** in Appendix for details). To assure our results, we create a simple time-sequence plot in **Figure 8, Diagram 4** where patterns and high spikes are observed.

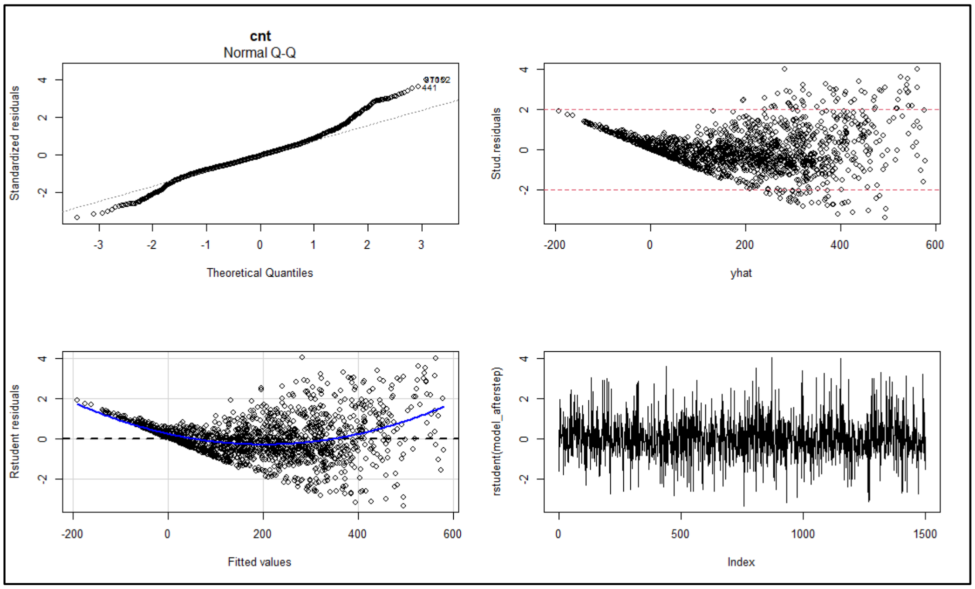


Figure 8 The Plots of residuals for the model after stepwise procedure. (1) Residuals against Fitted values, (2) Studentized residuals against predicted value, (3) The Rstudent residuals against Fitted values, (4) Rstudent residuals against Index

As all Assumptions are violated, we proceed on fixing the above problems by transforming our model:

1. First, we check our modified dataset for influential points (large residuals and/or high leverage) that affect our model predictions. In **Figure 9**, it is observed that the top 2 highest in Cooks Distance observations, those with the highest effect on our model for deleting them, are those with index 1267 and 563. We keep them into consideration in case we could not fit our model later.

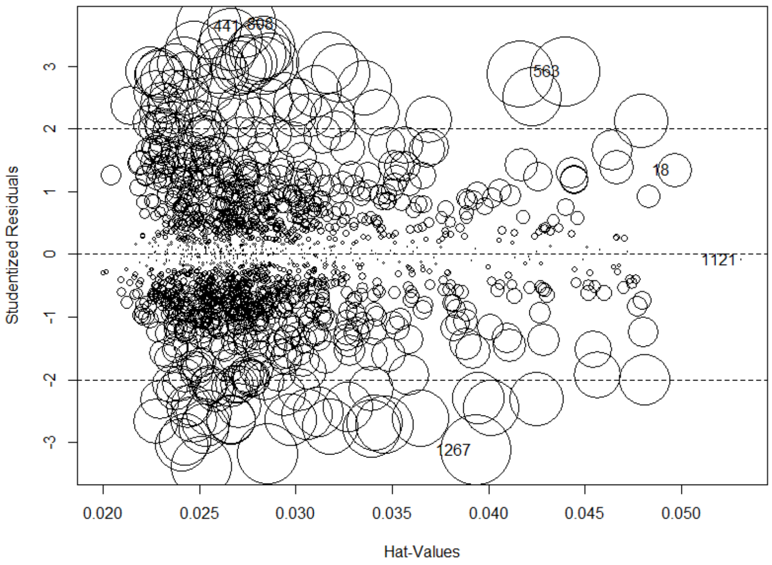


Figure 9 Studentized Residuals against Hat-Values for first model

2. We perform Box-Cox transformation to find if a transformation of the dependent variable is needed. In **Table 9** (see Appendix), it is observed that a transformation of the dependent variable is needed as p-value = 2.22e-16 <0.05, rejecting Ho that no transformation of dependent variable is needed. Also, the Ho that the right transformation for the dependent variable is the logarithmic transformation is rejected. The optimal value of parameter λ is that of 0.1961 ≃ 0.20. We assure our result by performing a transformation test for that value, in **Table 10** (see Appendix), where p-value = 0.77 >0.05 accepting our hypothesis. As a result, the dependent variable must be converted to y = (cnt^λ-1)/λ. Trying to avoid this transformation for interpretation reasons, we perform the Box-Cox transformation for our dataset abstracting the 2 influential points we targeted on the step before hoping that the value of λ will change to a more easily interpreted. The results differ slightly but not enough. Nevertheless, for interpretation reasons, we avoid the Box-Cox suggested transformation and we proceed with the logarithmic transformation as λ = 0.2 ≃ 0 affecting our Residual Normality Assumption.
3. We add polynomial effects on our model to fix the non-linearity of residuals. We first add polynomials of 5th grade for both numeric variables atemp and hum and proceed backwards abstracting polynomial terms that are insignificant for our model. The final added polynomials are that of atemp^2 and hum^2.
4. It is observed that non-linearity is not fixed and we proceed to abstract variables being careful to not under-fit our model. The chosen variables are yr and mnth as it is observed that yr affects linearity on a high level and mnth is insignificant for our model's predictability. As a result, non-linearity problem is fixed.
5. We perform the weighted least squares method to solve the problem of Heteroskedasticity. By adding weights Heteroskedasticity is fixed.

6. As a last attempt to try and fix non-Normality of residuals we detect and abstract the highest observations by Cooks Distance on our logarithmic model, in **Figure 10**. These are observations 1269 and 1278. Normality is not fixed.

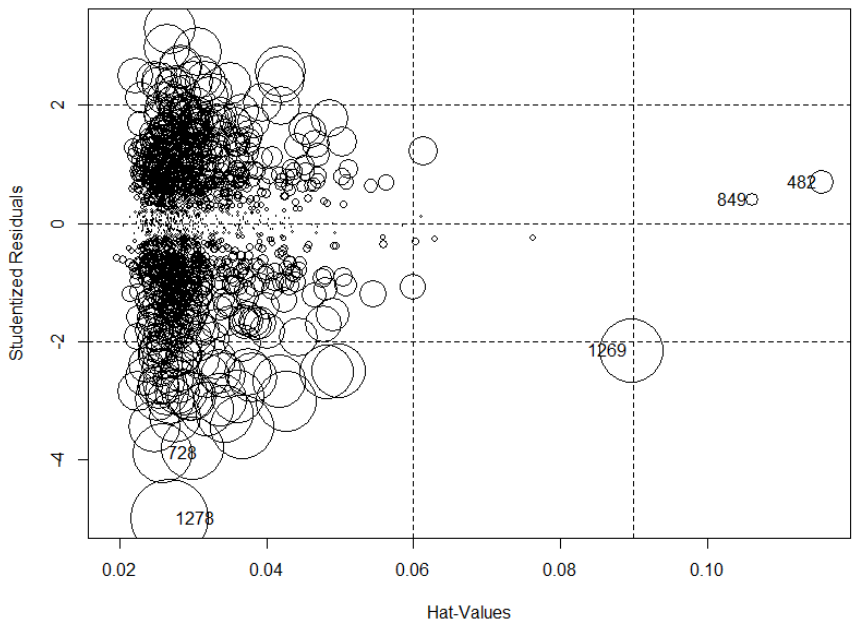


Figure 10 Studentized Residuals against Hat-Values for logarithmic model

As a conclusion, the model fails to pass the normality of residuals test (Lilliefors' test p-value= 2.452e-07<0.05; see **Table 11** for details) but passes the test of non-constant variance (nvc's p-value=0.77>0.05; see **Table 12** for details), passes test of non-linearity of residuals (Tukey's test p-value=0.28>0.05; see **Table 13** for details) and test of autocorrelation of residuals (Durbin Watson's test p-value=0.71>0.05; see **Table 14** for details). This means that the predictive ability of our model is not the same across the full range of the dependent variable but the differences between predicted and actual values present constant variance, linearity and no patterns on a chronological axis for all the predicted values. The problem of non-Normality of Residuals maybe could be avoided by selecting a high number of samples (n>15) to train our model but as we have only one subset to train it we will proceed by assuming that residuals are following the Normal Distribution for assignment interpretation reasons, even though they are not.

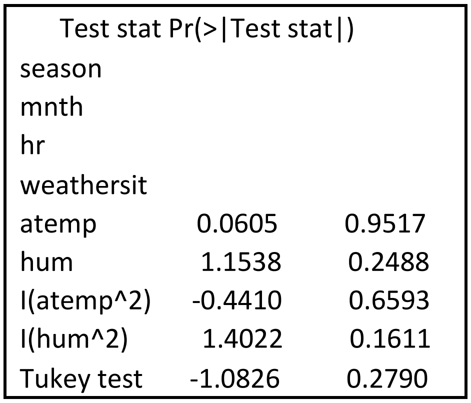
Table 11 Lilliefors' Test on Normality of Residuals

```
Lilliefors (Kolmogorov-Smirnov) normality test

data:  final_model$residuals
D = 0.044535, p-value = 2.452e-07
```

Table 12 Non-constant Variance test Score Test on Residuals

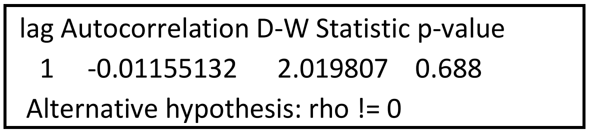
```
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.08635535, Df = 1, p = 0.76886
```

Table 13 Tukey's test on Residuals



```
     Test stat Pr(>|Test stat|)
season
mnth
hr
weathersit
atemp          0.0605       0.9517
hum             1.1538       0.2488
I(atemp^2)   -0.4410       0.6593
I(hum^2)        1.4022       0.1611
Tukey test    -1.0826       0.2790
```

Table 14 Durbin Watson's test on Residuals



```
lag Autocorrelation D-W Statistic p-value
  1    -0.01155132      2.019807    0.688
 Alternative hypothesis: rho != 0
```

The equation of our final model, as observed in **Table 15** and assuming the Normal Distribution of residuals although it is not corrected, is:

$$\log(cnt) = \mathbf{1.55} + \mathbf{0.17}*seasonspring + \mathbf{0.17}*seasonsummer + \mathbf{0.28}*seasonautumn - \mathbf{0.61}*hr1 - \mathbf{1.21}*hr2 - \mathbf{1.75}*hr3 - \mathbf{1.93}*hr4 - \mathbf{0.89}*hr5 + \mathbf{0.31}*hr6 + \mathbf{1.33}*hr7 + \mathbf{1.99}*hr8 + \mathbf{1.66}*hr9 + \mathbf{1.31}*hr10 + \mathbf{1.39}*hr11 + \mathbf{1.57}*hr12 + \mathbf{1.53}*hr13 + \mathbf{1.38}*hr14 + \mathbf{1.58}*hr15 + \mathbf{1.81}*hr16 + \mathbf{2.19}*hr17 + \mathbf{2.12}*hr18 + \mathbf{1.77}*hr19 + \mathbf{1.59}*hr20 + \mathbf{1.31}*hr21 + \mathbf{1.10}*hr22 + \mathbf{0.64}*hr23 + \mathbf{0.03}*weathersitModerate - \mathbf{0.42}*weathersitBad + \mathbf{7.17}*atemp + \mathbf{0.84}*hum - \mathbf{5.54}*atemp\text{^}2 - \mathbf{1.24}*hum\text{^}2 + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon \sim N(0,1.30\text{^}2)}$$

In **Table 15** (see Appendix)**,** it is observed that R squared is equal to 0.77, improved by 10% from the model with the unfixed assumptions, meaning that 77% of the difference of our predicted value from the actual value of the dependent variable is explained by our model. The Residual Standard error has decreased critically meaning that, if the residuals were to follow the Normal Distribution, the error in the estimated number of rentals would be (±2 x 1.30^2) $ around the expected value. P-value = < 2.2e-16 < 0.05 meaning that our model is much better than the constant model.

For a more practical interpretation of our model, we will exponentiate it and as a result we express the effect of a one-unit change in x on cnt as a percent.:

The intercept coefficient is equal to ($\mathbf{e}^{1.55}$ -1)*100 ≃ 371 meaning that on a winter day, at 00:00 – 01:00 am, having a good weather, 0 °C and 0% humidity the predicted number of bike rentals is 371.

The bike rentals will be ($\mathbf{e}^{0.17}$ -1)*100 ≃ 19% higher in season Spring and Summer than in Winter and ($\mathbf{e}^{0.28}$ -1)*100 ≃ 32% higher in Autumn than in Winter having all other variables constant.

The coefficients of variable hr are 23 in number and based on hr0 equal to 00:00-01:00 am. Hours hr1 to hr5 have a negative coefficient meaning that for one of those hours predicted bike rentals are getting decreased. The rest of hour coefficients are positive meaning that for one of those hours predicted bike rentals are getting increased. To implement an example, the bike rentals will have ($\mathbf{e}^{-0.61}$ -1)*100 $\simeq$ -47% decrease at 01:00-02:00 (hr1) in comparison to 00:00-01:00 (hr0), having all other variables constant.

The bike rentals will be ($\mathbf{e}^{0.03}$ -1)*100 $\simeq$ 3% higher when weather is Moderate than when it is Good and ($\mathbf{e}^{-0.42}$ -1)*100 $\simeq$ -34% lower when weather is Bad than when it is Good, having all other variables stable.

The bike rentals will increase by ($\mathbf{e}^{7.17}$ -1)*100 $\simeq$ 1849 for 1 unit of higher feeling temperature and by ($\mathbf{e}^{0.84}$ -1)*100 $\simeq$ 132 for 1 unit of higher humidity, having all other variables stable.

We use dataset “bike_test.csv” to assess the out of sample predictive ability of our model. By fitting our model to the new data we calculate the new R squared which is equal to 0.5063. Our model is able to explain the difference between predicted and actual values by almost 51% which is good but could become better. Also, our model is based on non-normal residuals, assuming they are normal for the assignment needs, which affects its accuracy and credibility.

We compare model after lasso, model after stepwise, constant only model and the full model using the mean absolute error. Models after lasso and after stepwise are the same as no extra variable was abstracted between them and the models are also the best models in comparison to the other 2 models as their value is equal to 76.05, the lowest, while full model is next with 76.48 and constant model last with 146.27.

Table 16 *R2 metric for out-of-sample predictions using the final model and MAE metrics for constant, full, after stepwise and after lasso models*

R2	MAE_2	MAE_3	MAE_4	MAE_5
0.5063733	146.2671	76.47961	76.04623	76.04623

5. Further Analysis

We create the descriptive diagrams for each season and proceed to describe a typical day for each season:

A winter day is characterized by low temperatures of around 10°C, with humidity of around 40% and low windspeeds around 25%. The weather is mostly good and if not, it is misty. Bike rentals become higher at 07:00-09:00 am and 15:00-18:00 pm reaching a total, at the end of the day, of 90 (see **Figure 11**).

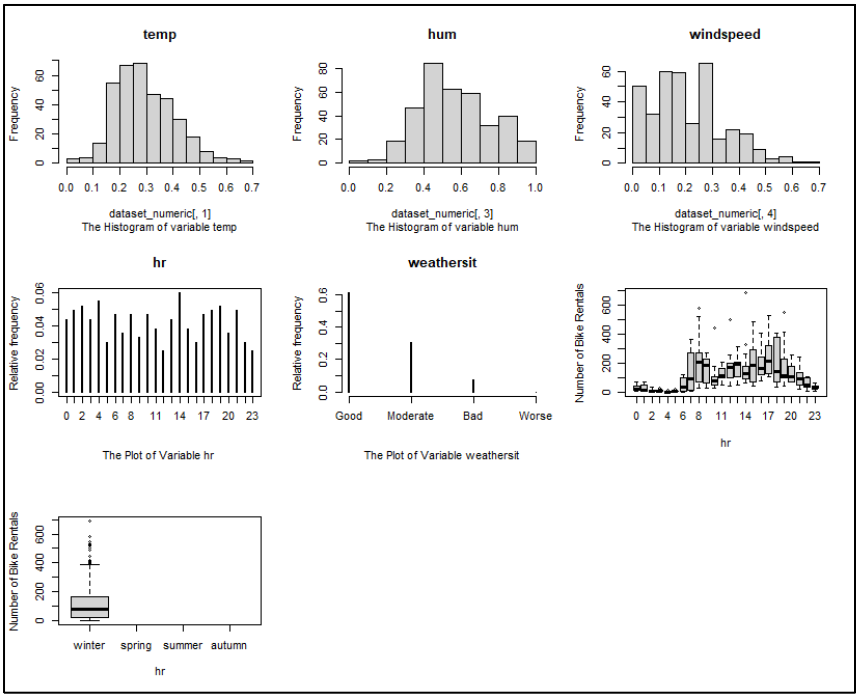


Figure 11 Descriptive Diagrams of Winter days

A spring day is characterized by medium temperatures of around 20.5°C, with humidity of around 65% and low windspeeds around 20%. The weather is mostly good and if not, it is misty. Bike rentals become higher at 07:00-09:00 am and 15:00-18:00 pm reaching a total, at the end of the day, of 180 (see **Figure 12**).

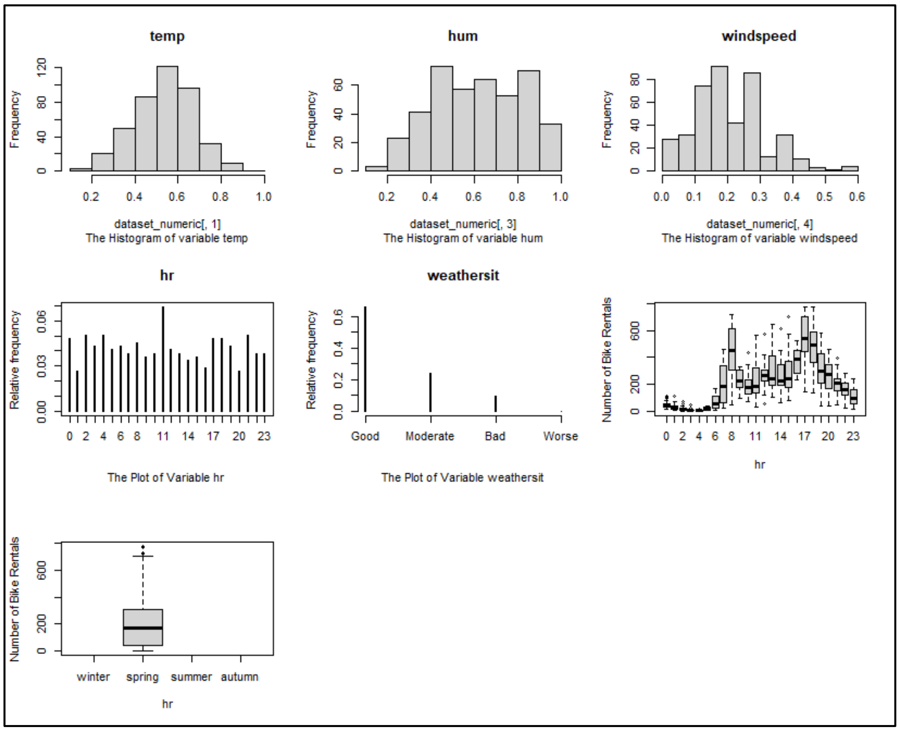


Figure 12 Descriptive Diagrams of Spring days

A summer day is characterized by high temperatures of around 31°C, with humidity of around 65% and low windspeeds around 20%. The weather is mostly good and if not, it is misty. Bike rentals become higher at 07:00-09:00 am and 17:00-19:00 pm reaching a total, at the end of the day, of 200 (see **Figure 13**).

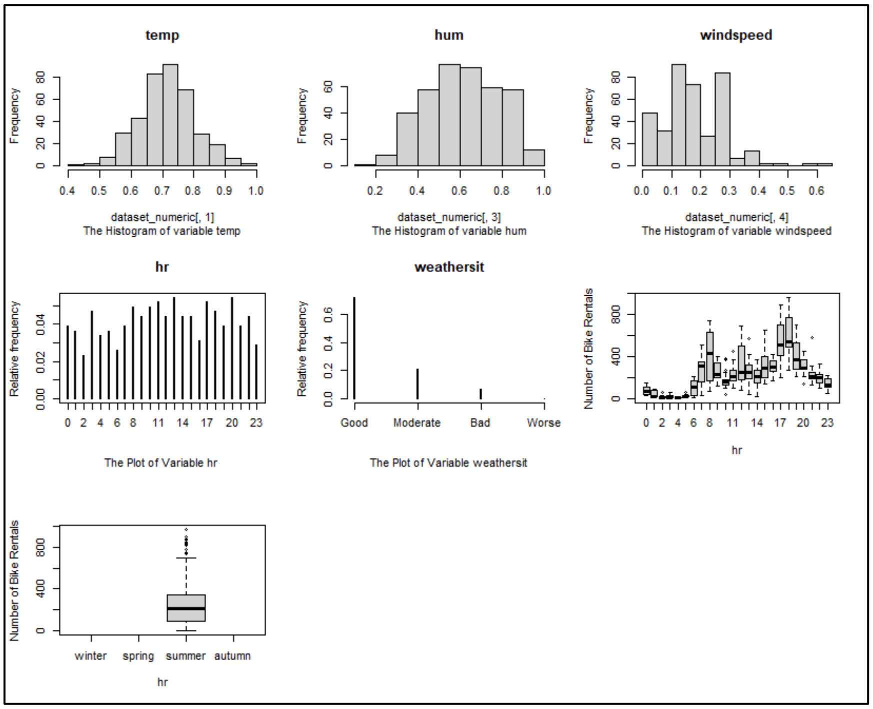


Figure 13 Descriptive Diagrams of Summer days

An autumn day is characterized by medium temperatures of around 16°C, with humidity of around 70% and low windspeeds around 15%. The weather is mostly good and if not, it is misty. Bike rentals become higher at 07:00-09:00 am and 16:00-18:00 pm reaching a total, at the end of the day, of 180 (see **Figure 14**).

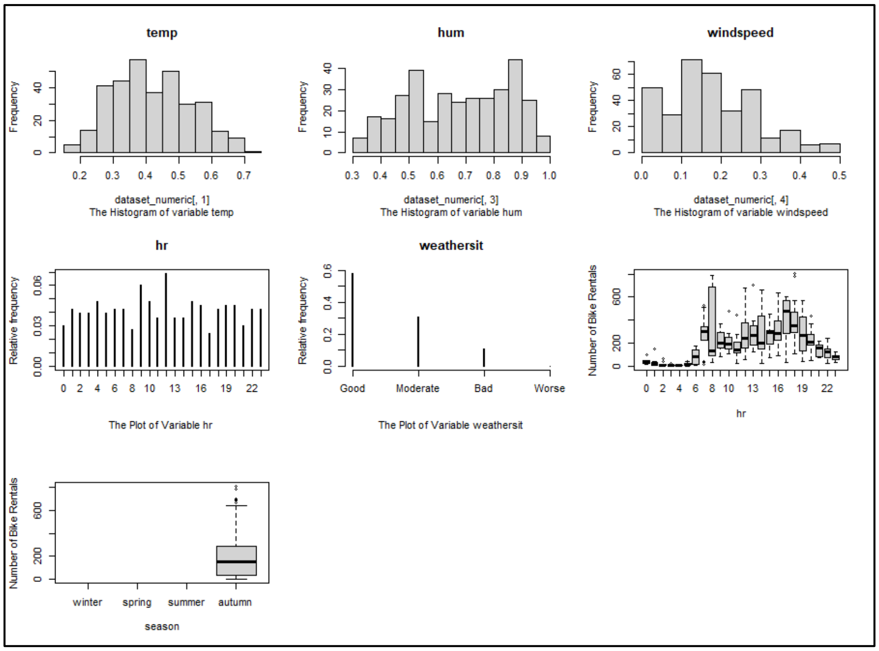


Figure 14 Descriptive Diagrams of Autumn days

6. Conclusion and Discussion

Bike rentals are highly affected by weather conditions. Seasons with mostly good weather and controlled temperature, humidity and windspeed like spring, summer and autumn present higher numbers of bike rentals than in winter. People mostly prefer to rent a bike early in the morning at 06:00-09:00 am and early in the evening 16:00-19:00 pm.

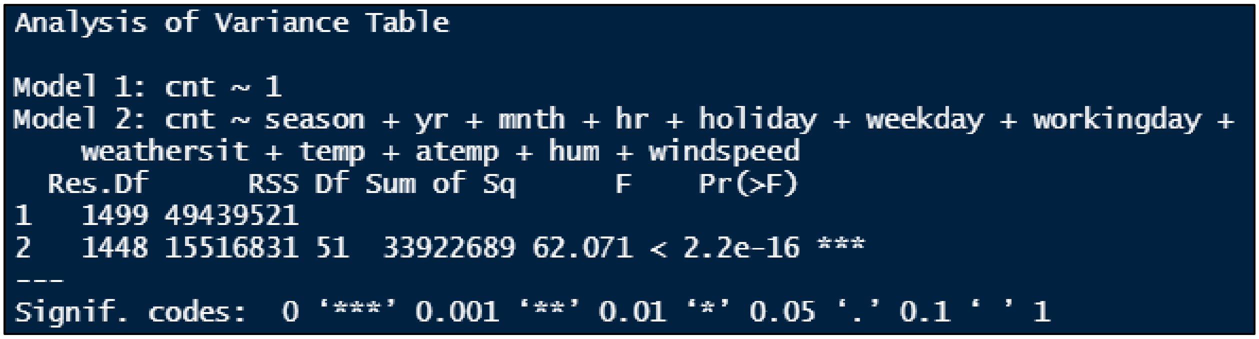
We created a multiple linear regression model with a logarithmic transformation to predict the bike rentals per hour given specific characteristics of the day. The transformation was not the best choice but helps us to interpretate the model more easily. The model rejects the assumption that residuals are following the Normal distribution leading to problematic credibility for model's predictions. For all the other assumptions not rejected, we proceeded with the existing model assuming the Normality of residuals for assignment's interpretation reasons. Our model explains the difference between predicted and actual values by 77%. The model was tested on an out-of-sample dataset and according to the predictions it scored an R squared of 51%.

If we were to use a different approach creating our model but out of the scope of the assignment, we could follow the Poisson Distribution as bike rentals is a discrete non-negative number, use the log link function and avoid in this way to normalize the residuals as it is not mandatory.

s

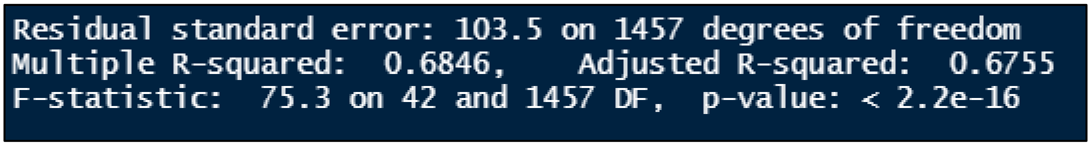
7. Appendix

Table 2 The ANOVA test between Full Model and Model including only the Constant



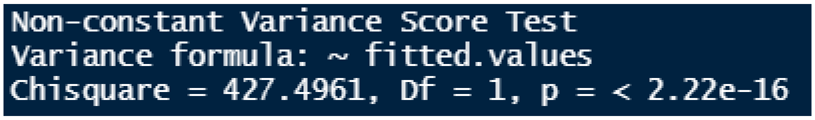
```
Analysis of Variance Table

Model 1: cnt ~ 1
Model 2: cnt ~ season + yr + mnth + hr + holiday + weekday + workingday + 
    weathersit + temp + atemp + hum + windspeed
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)    
1   1499 49439521                                   
2   1448 15516831 51  33922689 62.071 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 4 Interpretation of the model after step wise procedure



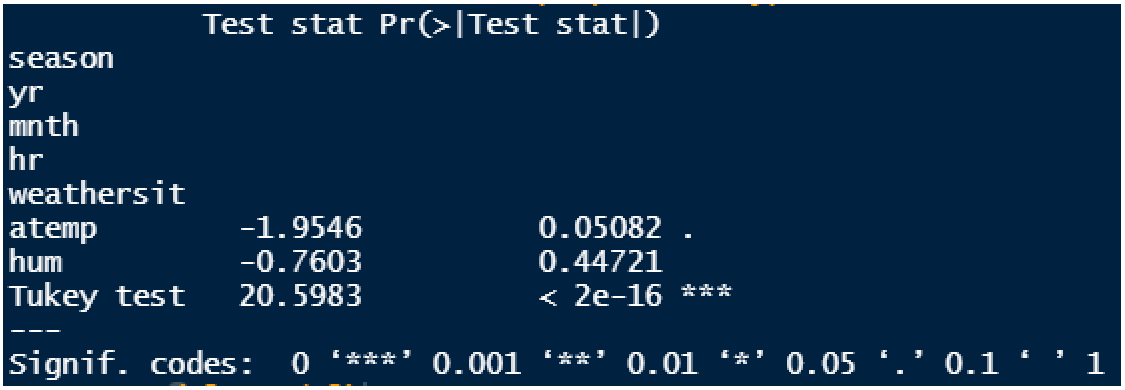
```
Residual standard error: 103.5 on 1457 degrees of freedom
Multiple R-squared:  0.6846,	Adjusted R-squared:  0.6755 
F-statistic:  75.3 on 42 and 1457 DF,  p-value: < 2.2e-16
```

Table 6 Non-constant Variance Score Test



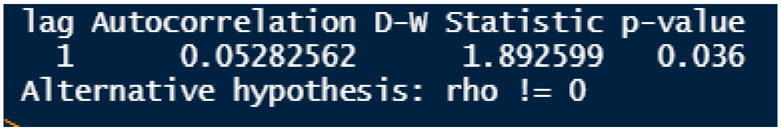
```
Non-constant Variance Score Test 
Variance formula: ~ fitted.values 
Chisquare = 427.4961, Df = 1, p = < 2.22e-16
```

Table 7 *Tukey's Test for Non-linearity of Residuals*



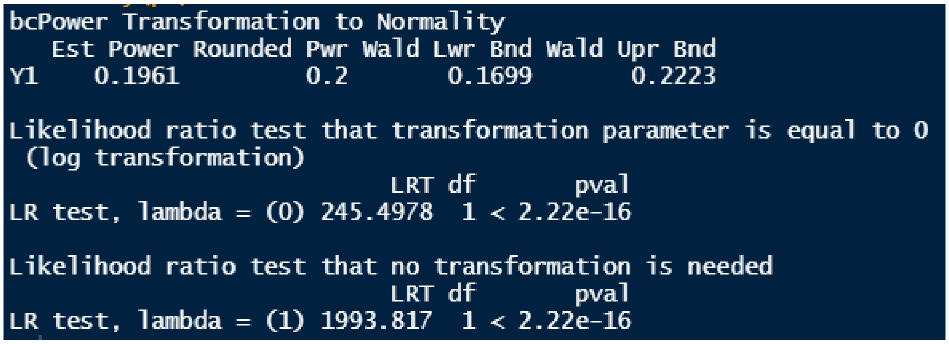
```
           Test stat Pr(>|Test stat|)    
season                                   
yr                                       
mnth                                     
hr                                       
weathersit                               
atemp         -1.9546          0.05082 .  
hum           -0.7603          0.44721    
Tukey test    20.5983          < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Table 8 Darwin-Watson Autocorrelation Test



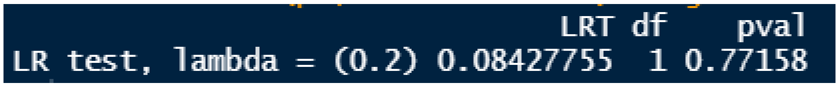
```
 lag Autocorrelation D-W Statistic p-value
   1      0.05282562      1.892599   0.036
 Alternative hypothesis: rho != 0
```

Table 9 Box Cox power transformation to Normality



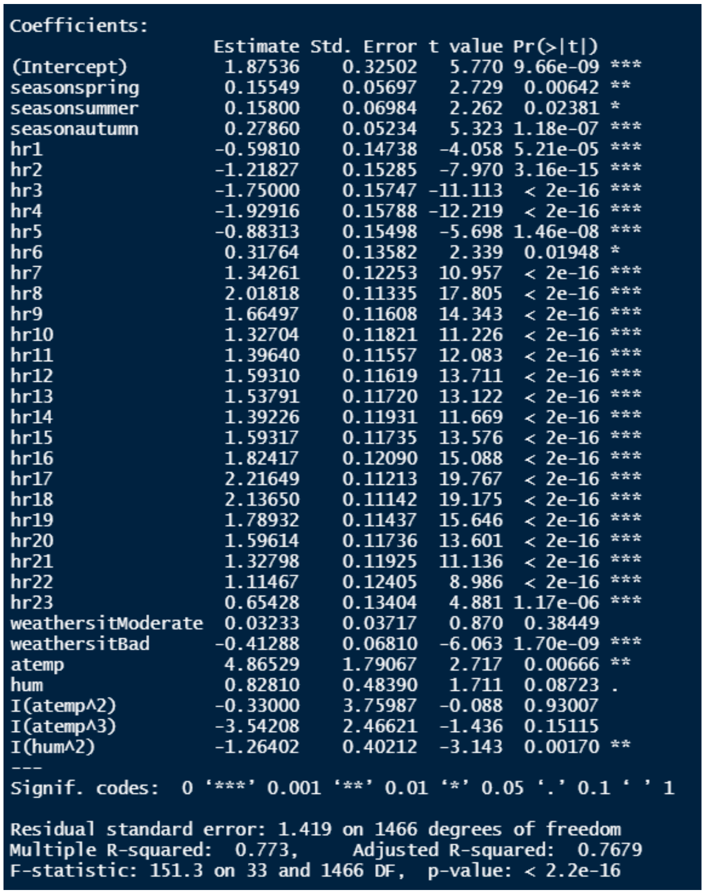
```
bcPower Transformation to Normality
   Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Y1    0.1961         0.2       0.1699       0.2223

Likelihood ratio test that transformation parameter is equal to 0
 (log transformation)
                           LRT df       pval
LR test, lambda = (0) 245.4978  1 < 2.22e-16

Likelihood ratio test that no transformation is needed
                           LRT df       pval
LR test, lambda = (1) 1993.817  1 < 2.22e-16
```

Table 10 Box Cox transformation of y Test for goodness of fit



```
                               LRT df    pval
LR test, lambda = (0.2) 0.08427755  1 0.77158
```

Table 15 The interpretation of final model



```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         1.87536    0.32502   5.770 9.66e-09 ***
seasonspring        0.15549    0.05697   2.729  0.00642 ** 
seasonsummer        0.15800    0.06984   2.262  0.02381 *  
seasonautumn        0.27860    0.05234   5.323 1.18e-07 ***
hr1                -0.59810    0.14738  -4.058 5.21e-05 ***
hr2                -1.21827    0.15285  -7.970 3.16e-15 ***
hr3                -1.75000    0.15747 -11.113  < 2e-16 ***
hr4                -1.92916    0.15788 -12.219  < 2e-16 ***
hr5                -0.88313    0.15498  -5.698 1.46e-08 ***
hr6                 0.31764    0.13582   2.339  0.01948 *  
hr7                 1.34261    0.12253  10.957  < 2e-16 ***
hr8                 2.01818    0.11335  17.805  < 2e-16 ***
hr9                 1.66497    0.11608  14.343  < 2e-16 ***
hr10                1.32704    0.11821  11.226  < 2e-16 ***
hr11                1.39640    0.11557  12.083  < 2e-16 ***
hr12                1.59310    0.11619  13.711  < 2e-16 ***
hr13                1.53791    0.11720  13.122  < 2e-16 ***
hr14                1.39226    0.11931  11.669  < 2e-16 ***
hr15                1.59317    0.11735  13.576  < 2e-16 ***
hr16                1.82417    0.12090  15.088  < 2e-16 ***
hr17                2.21649    0.11213  19.767  < 2e-16 ***
hr18                2.13650    0.11142  19.175  < 2e-16 ***
hr19                1.78932    0.11437  15.646  < 2e-16 ***
hr20                1.59614    0.11736  13.601  < 2e-16 ***
hr21                1.32798    0.11925  11.136  < 2e-16 ***
hr22                1.11467    0.12405   8.986  < 2e-16 ***
hr23                0.65428    0.13404   4.881 1.17e-06 ***
weathersitModerate  0.03233    0.03717   0.870  0.38449    
weathersitBad      -0.41288    0.06810  -6.063 1.70e-09 ***
atemp               4.86529    1.79067   2.717  0.00666 ** 
hum                 0.82810    0.48390   1.711  0.08723 .  
I(atemp^2)         -0.33000    3.75987  -0.088  0.93007    
I(atemp^3)         -3.54208    2.46621  -1.436  0.15115    
I(hum^2)           -1.26402    0.40212  -3.143  0.00170 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.419 on 1466 degrees of freedom
Multiple R-squared:  0.773,	Adjusted R-squared:  0.7679 
F-statistic: 151.3 on 33 and 1466 DF,  p-value: < 2.2e-16
```