Bike Sharing

Bike Sharing Demand Prediction

Contents

Abstract

The purpose of this study is to understand what influences bike rental usage and also predict it in order to satisfy demand. This study is based on **regression analysis** methods with aim to create a prediction model for bike usage. We will use a dataset from Capital Bikeshare system, Washington D.C., USA which is publicly available in http://capitalbikeshare.com/system-data. This database was aggregated with another database with weather and seasonal information from http://www.freemeteo.com. For our analysis purpose we will use a data sample with 1500 observations for model training and another smaller sample with 500 observations for model evaluation.

Chapter 1

Introduction - Description of the problem

Background information

Bike sharing systems are new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return it back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousands bicycles. Today, there exists great interest in these systems due to their important role in traffic, environmental and health issues.

Apart from interesting real world applications of bike sharing systems, the characteristics of data being generated by these systems make them attractive for the research. Opposed to other transport services such as bus or subway, the duration of travel, departure and arrival position is explicitly recorded in these systems. This feature turns bike sharing system into a virtual sensor network that can be used for sensing mobility in the city. Hence, it is expected that most of the important events in the city could be detected via monitoring these data.

The data

Bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc. can affect the rental behaviors. The core data set is related to the 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. We aggregated the data on hourly basis and then extracted and added the corresponding weather and seasonal information. Weather information are extracted from http://www.freemeteo.com.

Dataset characteristics

All datasets are random subsamples of 1500 hour occasions and have the following fields:

- **X:** record index
- **instant:** record index
- **dteday:** date
- **season:** season (1: Spring, 2: Summer, 3: Fall, 4: Winter)
- **yr:** year (0: 2011, 1:2012)
- **mnth:** month (1 to 12)
- **holiday:** weather day is holiday or not (1=Yes, 0=No)
- **weekday:** day of the week
- **workingday:** if day is neither weekend nor holiday is 1, otherwise is 0.
- **weathersit :** Possible outcomes
    - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
    - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
    - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
    - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- **temp :** Normalized temperature in Celsius. The values are divided to 41 (max)
- **atemp:** Normalized feeling temperature in Celsius. The values are divided to 50 (max)
- **hum:** Normalized humidity. The values are divided to 100 (max) (% percentage)
- **windspeed:** Normalized wind speed. The values are divided to 67 (max)
- **casual:** count of casual users
- **registered:** count of registered users
- **cnt:** count of total rental bikes including both casual and registered (we will use it as response)

Chapter 2

Descriptive Analysis[1]

Firstly we imported our data in **R Studio**. We observed that the two columns **X** and **instant** are ids of the records, so we excluded them, because are useless in our analysis. Also, we excluded from the beginning the **dteday**, because we have **hr**, **mnth**, **yr**, **workingday** variables so we don't need the date. In our analysis we have to divide our attributes to numerical (continues & discrete) and categorical (factors) variables because numerical and categorical variables need different visualization and also different model interpretation. As numerical[2], we have:

- **mnth** (discrete 1-12),
- **hr** (discrete 1-24)
- **weekday** (discrete 0-6)
- **temp** (continues)
- **atemp** (continues)
- **hum** (continues)
- **windspeed** (continues)
- **casual** (discrete)
- **registered** (discrete)
- **cnt** (discrete)

As categorical/factors we have:

- **yr** (2 levels)
- **season** (4 levels)
- **holiday** (2 levels)
- **workingday** (2 levels)
- **weathersit** (4 levels)

For **numerical variables** visualization we used histograms (hist() in R). With a first look (Figure 1-Numerical Variables Distribution) it's clear that no one variable has high symmetry or follows a normal distribution. **Mnth** distribution seems to have an outlier, in January and **hr** have outlier at 12:00 o'clock.

1 See Command 1: Insert Libraries - Command 8: Visual Analysis for Factors

[2] We preferred to consider **weekdays** and **hr** as discrete numeric because in the other case, we will have a separate, independent coefficient (or more precisely, degree of freedom) for each hour of the day. This could be too many variables to fit.

Temp and **atemp** distributions look quite similar to Normal distribution (but are not Normal distributed, KS p<<0.05); as we will discuss later these two variables are high correlated. Also we can observe that **casual**, **registered** and cnt variables have the same trend, are very high right skewed, something logical (because, casual users + registered users = cnt users). In **hum** distribution we have left skewness and the majority of records are in conditions about ≈ 50% humidity.

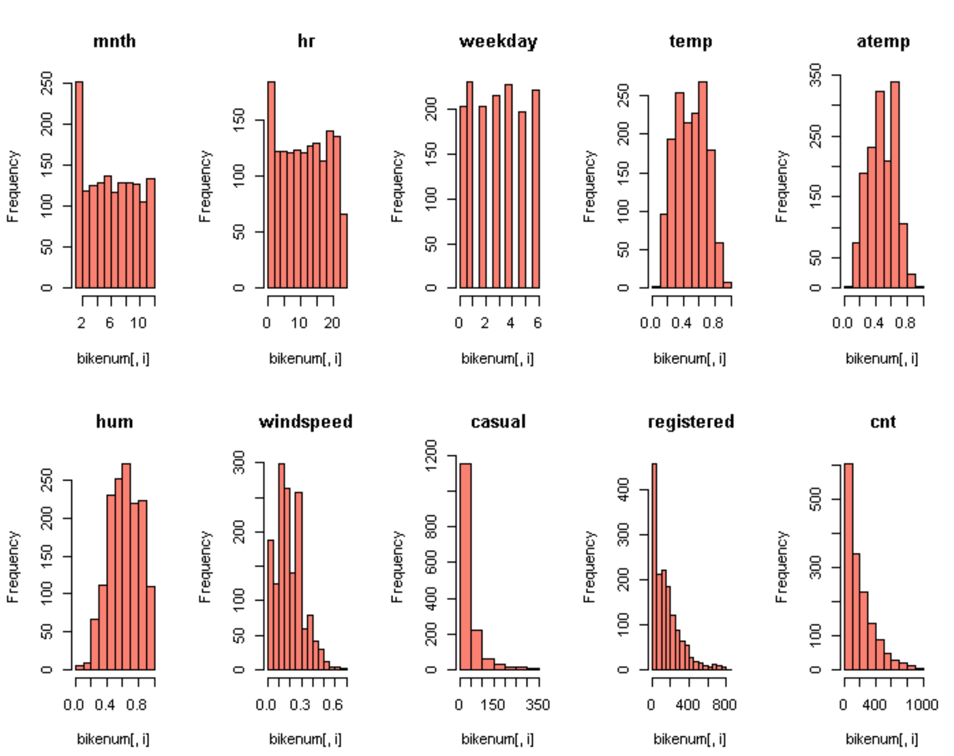

Figure 1-Numerical Variables Distribution

For **categorical variables** we used barplots (barplot() in R).We see that **Seasons** (see Figure 2-Categorical Variables Distribution**)** distribution is approximately in the same level with small fluctuations. Specifically, in winter we have a little drop, but not significant. In **Year** plot we see a small drop in records in 2012. On the other hand **holiday** and **workingday** have huge difference between distributions. In **weathersit** is clear that people don't like to use bikes in very bad weather conditions.

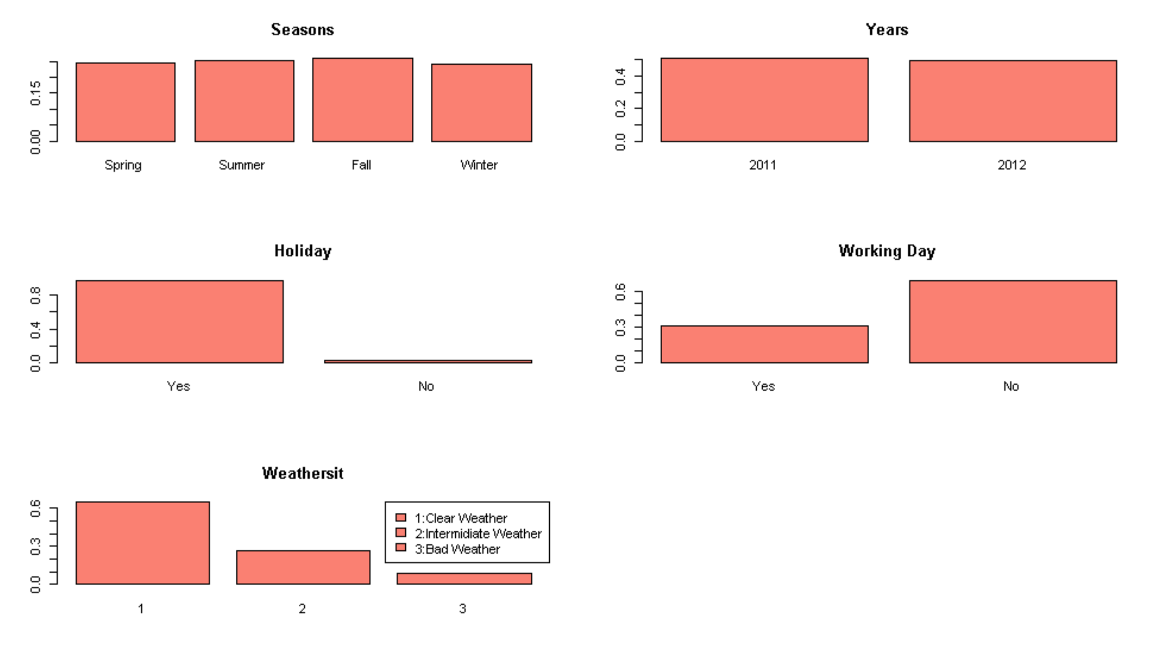

Figure 2-Categorical Variables Distribution

Chapter 3

Pairwise Comparisons[3]

In that stage we have to find variable correlations (corrplot() in R). So, we created a correlation figure between all numerical variables (see Figure 3- Numeric Variables Correlation). Firstly, we observed a high correlation (value around 1) between temp and atemp. This correlation was expected because the actual temperature is high related with the “real feel” temperature (“real feel” usually depends on actual temperature, humidity and wind speed). Also, we have intermediate/low correlations between cnt and weather conditions temp, atemp and humidity (hum); hum and cnt looks inversely proportional variables. On the other hand windspeed doesn’t look to affect bike sharing. Another variable, that is correlated with cnt, is the hour (hr) of the day, something that we expected because users can’t produce the same traffic all the day (at night hours we expect a usage drop).

NOTE: We removed the columns with casual and registered users because in our analysis we have to make a model only for the total rentals (cnt) and not for each type of users.

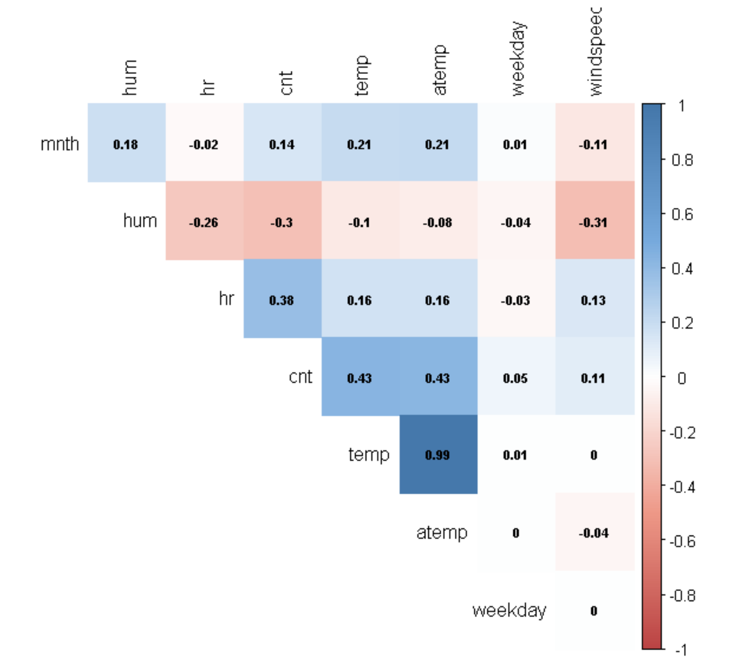

Figure 3- Numeric Variables Correlation

[3] See Command 9: Pairs of Numerical Variables - Command 11: Cnt on Factor Variables

In the following graphs we have analyzed the distribution of total bike demand according to the variables, that we mentioned above. In the following figure (see Figure 4- Scatterplots For Weather Variables) we observe that there is an increasing trend between cnt and actual, as cnt and real feel temperature.

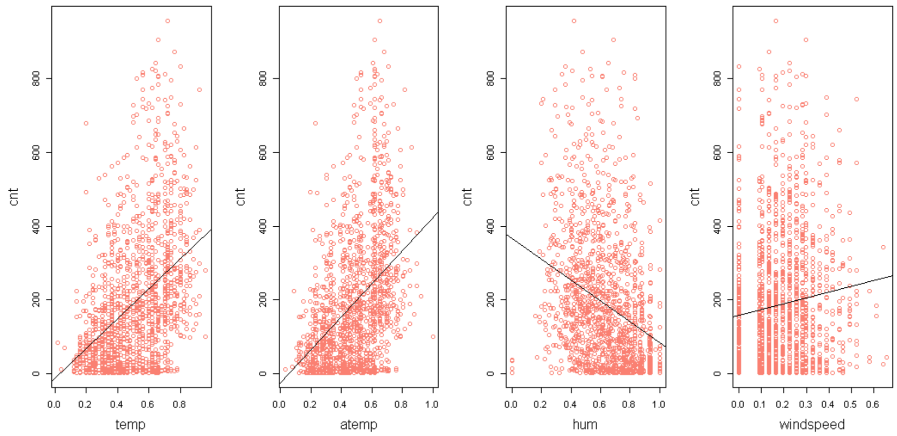

Figure 4- Scatterplots For Weather Variables

Heteroscedasticity also increases, and maybe it's a sign that we will have problem with model assumptions, if we will include temp or atemp as predictors. Humidity (hum) preserves also a lot of variance. Wind speed (windspeed) doesn't affect the bike sharing usage as we said in the previous paragraph; but also in the following graph we can observe this independence between cnt and windspeed because the fitting of the line doesn't follow the distribution of scatters. Also, it's important to mention that we have a lot of outliers and maybe influential points that will affect our models. Hence, we have to keep in mind these figures when we create our models.

Afterwards, we have to check the distribution of bike share demand in every hour of a day. So, we created a boxplot graph. As we can see in the following figure (see Figure 5 – Total Usage in a Daily Bases) the distribution has high variance during a day. Although, we can divide the day in the following periods:

- **24:00 - 07:00 :** Low bike usage
- **07:00 -16:00 :** Middle bike usage
- **16:00 - 19:00 :** High bike usage
- **19:00 - 24:00 :** Middle-Low bike usage

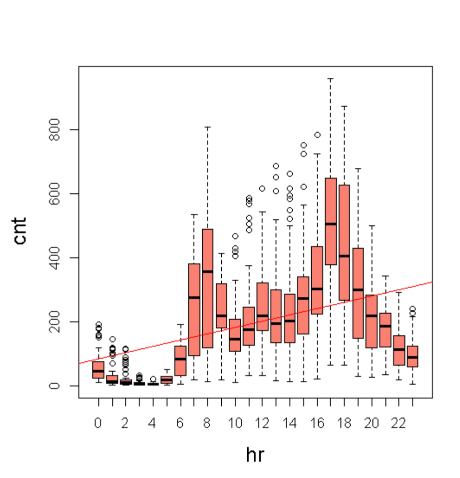

Figure 5 – Total Usage in a Daily Bases

At that stage we have to check the relation between cnt and **categorical variables**. We created the following boxplots (with function boxplot() in R) (Figure 6- Cnt on Factor Variables). With these boxplots we can understand the relation between cnt and factors and how will affect our models. Specifically, bike usage tends to be increased at 2012, and also when people don't work (holiday = 1 **or/and** workingday=0). Seasons look to have a positive relation with cnt, as we go from Spring to Winter, it also looks that we have influential points in Winter because the line doesn't fit very well in the boxes. On the other hand, in weathersit and cnt we have negative relation; in bad weather conditions (weathersit =3) the bike usage is very low.

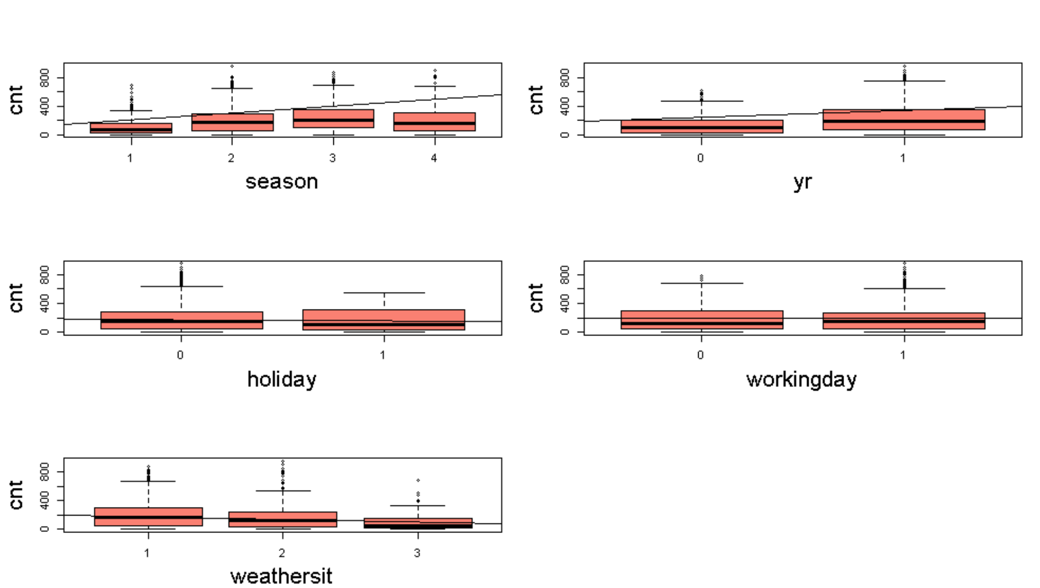

Figure 6- Cnt on Factor Variables

Chapter 4

Model Selection[4]

We have already observed the correlations and relations between variances and now we can construct our regression models and try to find the most appropriate of them for prediction of the total bike demand. As a response we have the cnt and all the other variables as predictors. We start our searching for a linear model ($Y = b_0 + b_1 x_1 + \ldots b_n x_n$, with $\varepsilon \approx N[0,\sigma^2]$).

Firstly, we constructed our first model with only numerical variables, **initial model** (lm() in R)(see Table 3- Initial Model). In this model we have temp, atemp and windspeed as insignificant variables (because $P_r > 0.05$ and we can't reject the null hypothesis, $H_0 =$ Coefficient equals to zero). Also, if we check Variance Inflection Factor – VIF based on **Akaike** criterion (vif() in R) we observe that temp and atemp have very high value (around 42.4), something that we expected (see Figure 3- Numeric Variables Correlation).

In order to take a better decision we used Stepwise Regression method (step() function in R) and Backward Elimination in initial model and we created a new model, **model 1**, without atemp, windspeed, (see Table 4 - Model 1). This model has similar R^2 & $R^2_{Adj} \approx 0.33$; now we have low collinearity between variables, VIF (value ≈ 1).

In this stage we added Factors to our model and we created the **full model** (see Table 6- Full Model). This model has higher R^2 & $R^2_{Adj} \approx 0.39$, but the R^2 increasing was expected, because we increased the number of predictors. In that model we have intercept, atemp, mth, windspeed, workingday and holiday as insignificant variables ($P_r > 0.05$, we reject the null Hypothesis, $H_0 =$ Coefficient equals to zero) and also it's appeared again the high VIF between temp and atemp. So, we apply stepwise function (we prefer AIC criterion than BIC, because BIC has higher penalize and simplify a lot the model) in order to decrease the number of predictors.

After Stepwise Regression method and Backward Elimination we had the same model in both cases, **model 3** (see Table 7- Model 3). This model has less predictors, method removed atemp, mth, windspeed,

[4] See Command 12: Initial Regression Model - Command 21: Anova Test, Full with Null Model - Command 20: Model 3, Stepwise Method

workingday and holiday; low VIF (because the atemp was excluded), R^2 & $R^2_{Adj} \approx 0.39$ and standard error ≈ 140 bikes. So, we can assume that model 3 has the best fitting for now.

Assumptions[5]

There are four principal assumptions which justify the use of linear regression models because for purposes of inference or prediction, the R-squared doesn't tell us the entire story. We should evaluate R-squared values in conjunction with residual plots in order to round out the picture. At that stage of our analysis, we have to check the assumptions of linear regression. So, we test:

- ❖ Check for Normality
- ❖ Variance in Quantiles
- ❖ Residuals Variance (we prefer for R^2)
- ❖ Residuals Linearity
- ❖ Residuals Independence
- ❖ Check For Outliers

In the following figure (see Figure 7- Normality, Variance, Linearity & Independence Plots) we clearly see that the residuals don't follow a Normal distribution, because in Q-Q plot only in the center area they are is a straight line. Also, we observe a lot of heteroscedasticity of the residuals and variance increasing (isn't constant in all quantiles). In addition, we confirm the previous results with **Shapiro-Wilk** test, (shapiro.test(), $P_r<0.05$ and we reject the null hypothesis, H_o = Normal Distribution), **Levene** test (leveneTest(), $P_r<0.05$ and we reject the null hypothesis, H_o = Homogeneity of variance) and **Non-constant Variance** test (ncvTest(), $P_r<0.05$ and we reject the null hypothesis, H_o = Constant Variance). Finally, about residuals independence we can not observe a pattern, so we have independence of errors. We check it further with independence test (runs.test(), $P_r>0.05$ and we can't reject the null hypothesis, H_o = Randomness).

[5] See Command 22: Check Assumption - Check Normality of the Residuals & Constant Variance – Command 26: Check Assumption- Check for Outliers

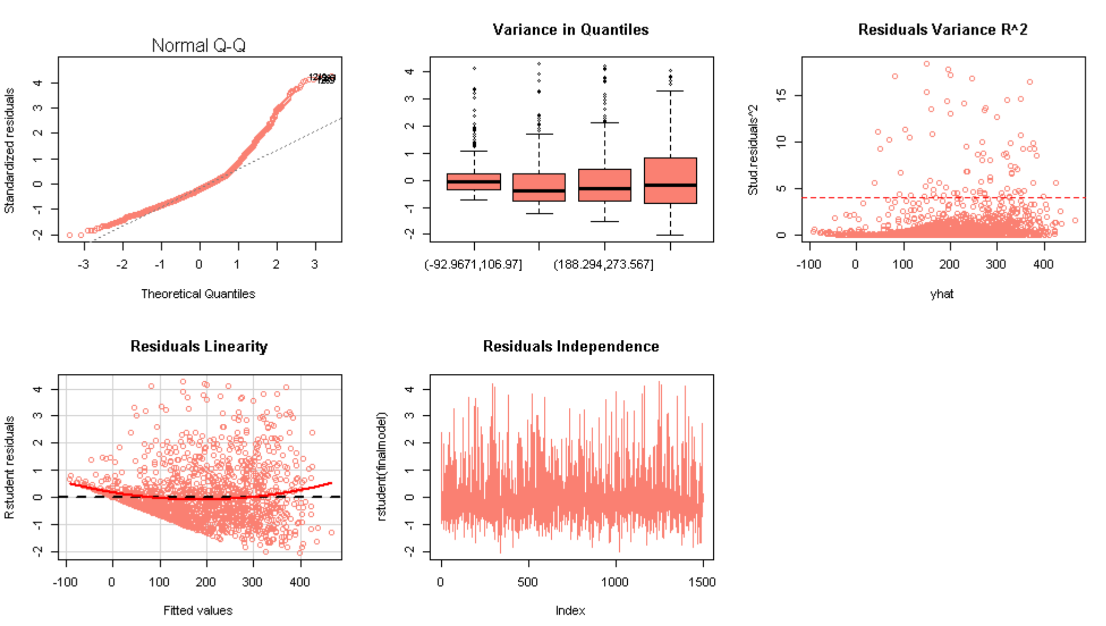

Figure 7- Normality, Variance, Linearity & Independence Plots

Also Leverage test (LeveragePlots() in R) shows that we have extreme values on predictor variables, maybe outliers or influential points.

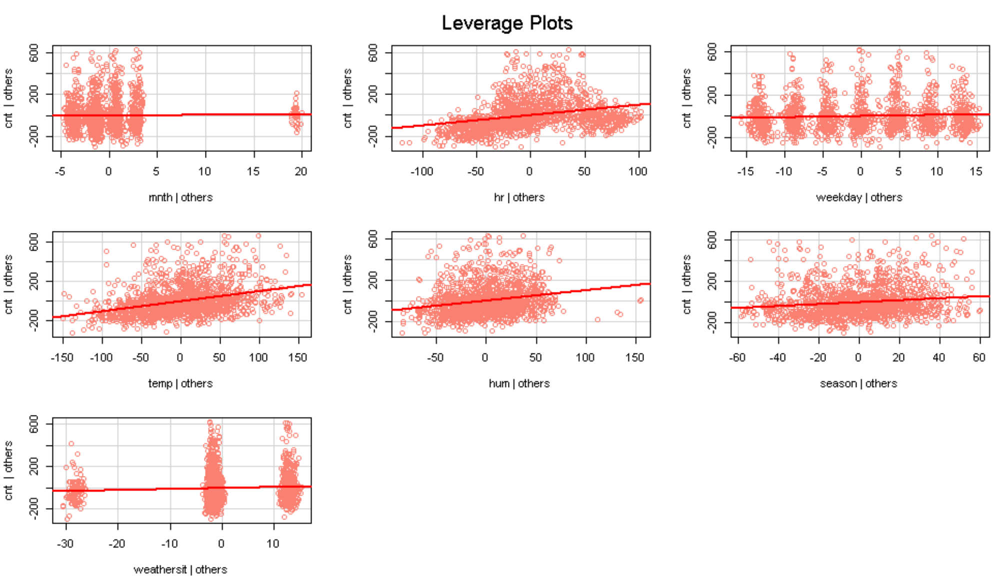

Figure 8- Leverage Test

Now, our goal is to fix our assumptions by adding no-linear terms in our model. We test all predictors with response in order to understand the relation between them and to correct them one by one. After a lot of trials we concluded that we have to logarithm the response (cnt) in order to meliorate the normality of the residuals. We observed that hr has a periodic form. So we added a linear combination of trigonometric functions:

$$cnt \sim Asin(2\pi * hr * \omega/24) + Bcos(2\pi * hr * \varphi/24), with\ A, B, \omega, \varphi\ are\ constants$$

Also, we exclude all the predictors that aren't significant (see Table 9- Model with Corrected Assumptions (Trigonometrical Terms)). Moreover, we try, instead of adding trigonometric terms, to add polynomial terms. The result is a model with better fitting (see Table 8- Model with Corrected Assumptions (Polynomial Terms)). But, the interpretation of the model becomes extremely difficult and also we don't know if our new model is over fitted only to this sample. Thus, we prefer to continue with the model with trigonometrical terms.

$$\log(cnt) = 3.627 + -0.002\, Summer - 0.065\, Fall + 0.202\, Winter + 0.207\, Year(2012) + 0.011\, Weekday + 0.053\, Weathersit2 - 0.164\, Weathersit3 + 1.237\, Temp - 0.548\, hum - 0.194 \sin\left(6.28\frac{hr}{24}\right) - 2.244 \cos(0.628\frac{hr}{24})$$

With the above model we achieve to fix assumptions in a certain extent, not completely. In that model the residuals are more close to Normal distribution, but in the edges we still have problem. The variance isn't constant (we achieved constant variance only when we excluded the hr). Also, the residuals have linearity and independence.

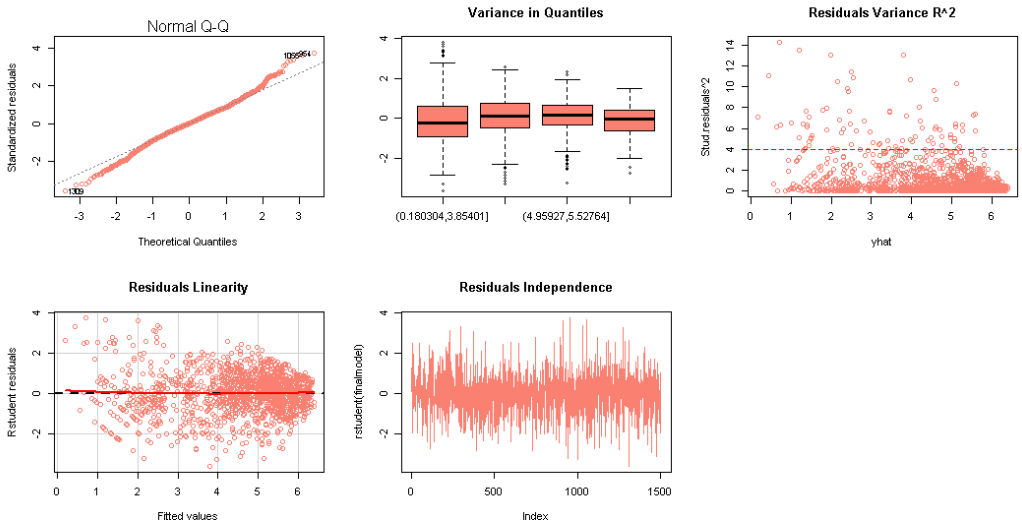

Figure 9- Corrected Assumptions

In conclusion, I would like to mention that it was impossible to fix all the assumptions simultaneously. We used Cox-Box transformations (for polynomial terms) to our model in order to find the appropriate terms but, we were not satisfied with the results. In next paragraphs we will evaluate the above model.

Lasso Model[6]

At that stage we create another model by using **Lasso Regression** (least absolute shrinkage and selection operator). Lasso performs model selection. In Lasso we have to tune the parameter lambda (λ) that controls the amount of regularization. As the λ increases, Lasso sets more coefficients to zero. We choose the largest value of λ, in order to limit the error within 1 standard error of the minimum. In R we create a matrix (without the intercept) and by using the **glmnet** library we create the relation between λ and Lasso coefficients (see Figure 10- Lasso Coefficients Shrinkage).

[6] See Command 27: Lasso

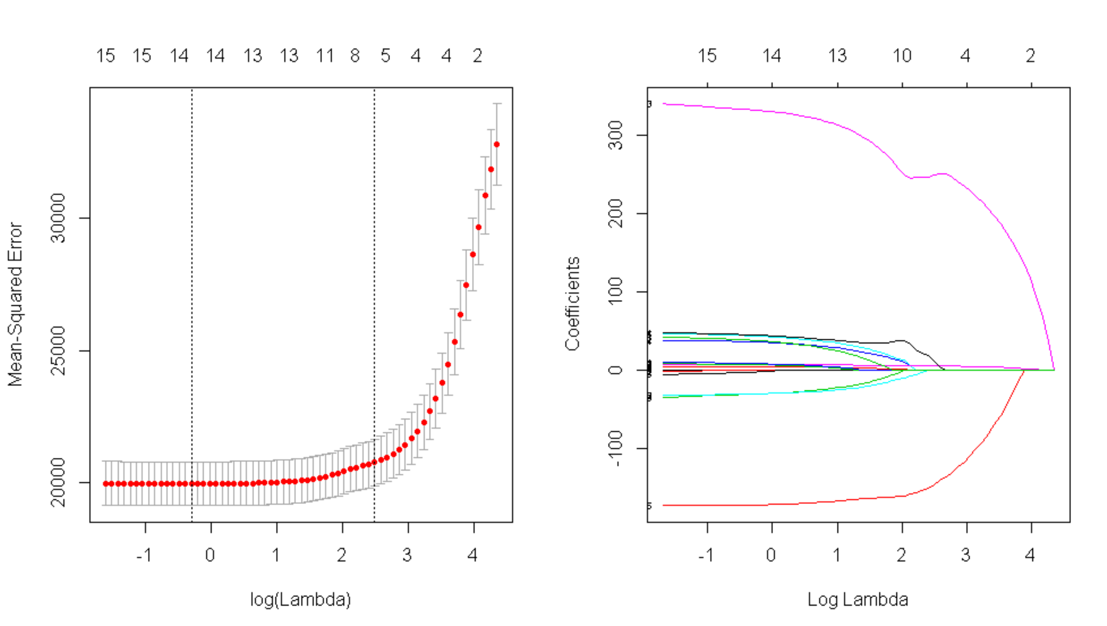

Figure 10- Lasso Coefficients Shrinkage

After we find the value of λ (by running lasso1$lambda.1se, in R Studio). Finally, we create a table with the shrinkage of coefficients by putting the s= lasso1$lambda.1se. Lasso shrinkage (see Table 10 - Lasso Coefficients Shrinkage Table) weathersit, mnth, holiday, weekday, workingday and windspeed. After Lasso we calculate again the coefficients (see Table 11 - Lasso Model Summary) and we have season and atemp as insignificant predictors (P_r>0.05 in regression model). So, Lasso model has the following format (beta parameter values are given in the first column in the Table 9):

$$cnt = -23.57 - 4.77 * Summer - 38.7 * Fall - 38.7 * Winter$$
$$+81.8 * Yr(2012) + 3 * Month + 6 * Hr + 380.7 * temp + 12.6 * atemp - 189.1 * hum, with\ \varepsilon \approx N[0, 141]$$

Models Evaluation[7]

At that point, we have to test/evaluate our models and assess the out-of-sample predictive ability with another sample, an evaluation/test dataset of 500 observations. We will test models from Stepwise Regression method, Lasso method, compare full and null model and also we have to check the model with corrected assumptions (with no linear terms).

[7] See Command 28: Import Evaluation Data - Command 30: Create Factor Variables

Firstly, we created distributions for each variable[8], for our new sample, in order to understand how close our samples are and in addition to understand better the behavior of our models. We found that they are similar because we implemented a no parametric test (distributions aren't normal or symmetric, Wilcox.test, P_r=0.96>>0.05, so we can't reject the null Hypothesis, H_o= Similar distributions).

Then we tested **model 3** (model from Stepwise method) with the evaluation data. The prediction/fitting ability is lower ($R^2_{Adj} \approx 0.36$) and few predictors are insignificant (P_r>0.05 in regression model), also the standard error had increased a little (SE= 151 bikes). With **Lasso model** we have the same behavior like we had with training dataset. The goodness of fit indicators are only approximately 1% less in evaluation sample, with temp and atemp now insignificant variables in training dataset (see Table 13 - Lasso Model (Evaluation Dataset)).

The **full_model** has approximately the same fitting ability in both datasets (training and evaluation dataset) because the $R^2_{Adj} \approx 0.39$ (see Table 14 - Full Model (Evaluation Dataset)). But, full model is more complicated and has a lot of insignificant predictors in both cases. Then, we have to check with Anova test (anova() in R) to find out if the additional parameters of the full model are zero. So, we compare the full model with the null model (null_model = $b_o + \theta$) and we reject the null hypothesis (Annova P_r<<0.05, H_o = Additional parameters are equal to zero)

Finally, we evaluate the **model with no linear terms** (see Table 15 - No linear Model (Evaluation Dataset)). This model has $R^2_{Adj} \approx 0.47$, a bit slight higher than we had in training data. Moreover, the assumptions of this model are quite similar with the assumptions we had in training data.

Final Model & Interpretation

At the end, we have to choose our final model for total bike sharing demand. At this stage we will take into our consideration all previous data from descriptive analysis, pairwise methods, Lasso, and also the prediction abilities in evaluation sample.

Firstly, we checked with Anova that the extra parameters from our models weren't zero (we keep in mind that from pairwise analysis we found that temp and atemp have high correlation). **Full model** is very complicated with a lot of insignificant variables (because P_r>0.05 and we can't reject the null hypothesis, H_o = Coefficient equals to zero). **Model 3**, model after pairwise method, has exclude all the insignificant variables from the full model and also has very low VIF in all variables. The prediction abilities of this model are quite similar in the evaluation dataset.

No-linear model, with corrected assumptions, has the highest goodness of fit. But we think it's over parametrized and has the most complicated interpretation. Also, No-linear model has extremely low standard error, maybe it's a signal of overfitting in the errors of the training sample. Although, we had also higher goodness of fit in the evaluation sample; but we are not sure about their prediction abilities in other samples, but a further check is out of this study.

Lasso model is a model without workingday, windspeed and weathersit variables. Lasso model has R^2 and $R^2_{Adj} \approx 0.39$ and SE=141 bikes as we had in the full model but with less predictors, so it's a more simple model with better goodness of fit than full model. But that model don't have shrink of temp or atemp, so VIF between these two variables is high. Also, I would like to mention that atemp is insignificant variable.

In our view **model 3 (Stepwise model)** is appropriate for our analysis purpose. We don't prefer to choose models with only the higher goodness of fit either if they have better assumptions. We choose a simple model with significant covariates, low VIF between variables and simple interpretation in order to describe a typical day for each season. So, we will continue with the following model:

$$cnt = -36.6 - 0.6\, Summer - 25.6\, Fall + 59.1\, Winter + 81.6\, Yr(2012) + 6.7\, Hr + 4.8\, Weekday$$
$$+11.07\, Weathersit2\ 29.2\, Weathersit3 + 391.3\, Temp - 178.4\, Hum + e,\ with\ e = N[0,140]$$

"Everything should be made simple as possible, but not simpler – Albert Einstein"

We have negative value in the intercept, something meaningful. So we can centralize[9] the model for better interpretation (see Table 16 - Cetralized model 3 (Coefficients))

Finally we have (we round the coefficients):

[9] See Command 31: Centralize model 3

$$cnt = 140 + 82\,Yr(2012) - Summer - 26\,Fall + 59\,Winter + 7\,Hr + 5\,Weekday$$
$$+11 Weathersit2 - 29\,Weathersit3 + 391\,Temp - 178\,Hum\,, \qquad with\ \ e = N[0{,}140]$$

So we can understand that if we are in 2011, its Spring, not a week day with good weather, temperature around 20.5 Celsius degrees (Actual_temp= temp*median, we don't have a symmetric distribution in temp so we use median as a representative statistic) and also hum 89% (Actual_hum= hum*median, for the same reason with temperature)the total bike users will be 140. If we compare the same situations with 2012, the users will increase +82. Moreover, if we be in Summer we will have 1 user less, or in Fall we will have 26 users less, but in Winter we will have 59 users more. The bike share survey was implemented in Washington, in Washington the winter is mild[10] (We can say the climate is similar to Greece).

The coefficients of Weekday and Hr have different interpretation in comparison with the other numerical variables (Temp and Hum). These coefficients mean that if we change one day (e.g from Monday to Tuesday), or one hour (from 11:00 to 12:00) we will have 59 bikes more in a different day and 7 bikes more in a different hour. **These two variables don't have good interpretation if they are numerical instead of factors.**

Temperature and humidity[11] are very important variables. If we are in a specific date (year, season, day, hour) and increase for 1 (+41 in Celsius degrees) that means the increasing of +391 bikes. On the other hand if we are in the same date and increase for 1 (100% humidity) then we will have less 178 users (negative association between users and humidity). All in all, the variability of predictions is ±140 bikes, not so low variance.

[10] Reference: https://www.worldtravelguide.net/guides/north-america/united-states-ofamerica/washington-state/weather-climate-geography/

[11] Temperature and Humidity have lower and upper limits, 0≤hum≤100, -10≤Temp≤41

Chapter 5

Conclusion & Discussion

Finally, after a lot of work, we have a prediction model for bike demand. As it was our initial thought, weather conditions play a certain role in bike usage. The most major factor is weather temperature and humidity; people don't like to use bicycle for commuting in extreme weather temperatures so this behavior is also significant connected with seasons. In an annual bases, winter (a mild winter) looks the most convenient period for bike usage. Also, in extreme weather conditions people avoid to use bikes. I would like to mention that our analysis didn't take into consideration the two different type of bike users, casual and registered, because in our study we didn't have this goal. So, if in the future we want better prediction model we could split our model in different kind of users, because maybe they will have different bike culture.

Another topic that I would like to mention is that, we could have done a different analysis if we had used hr and weekday as categorical variables (factor). We could have implemented it with different ways, e.g use 24 levels for hour or create classes. But in my point of view, that analysis is very interesting and it would be very helpful in a model that tries to describe the fluctuation of bike usage during a day or during a week. In our analysis we didn't have this purpose and we used them as numeric.

Also, in this analysis we didn't achieve to correct all assumptions, something that is very important if we want to have an accurate prediction model. Finally, we didn't preferred the no-linear model although it had the higher goodness of fit and better assumptions. We preferred a more simple model that will help us to describe a typical day for each season. In real conditions, we could have a no linear model as supplementary model for prediction, in specific occasions. For instance, the bike demand in a week day of September at noon. Also we don't forget that the training and evaluation sample are a very little piece from a huge database. So, if we had the opportunity to train and evaluate our model with more data it would help a lot the prediction behavior.

Table of Figures

Appendix I

Tables

Table 1 - Stucturre of Training Dataset

```
data.frame':                 1500 obs. of  13 variables:
 $ season    : Factor w/ 4 levels "1","2","3","4": 2 4 4 4 3 3 1 2 1 4 ...
 $ yr        : Factor w/ 2 levels "0","1": 2 1 2 1 1 1 2 1 2 1 ...
 $ mnth      : num  5 12 9 10 7 8 1 4 3 10 ...
 $ hr        : num  3 12 1 17 21 6 1 20 4 10 ...
 $ holiday   : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 2 1 1 1 ...
 $ weekday   : num  4 1 6 1 4 6 1 0 0 6 ...
 $ workingday: Factor w/ 2 levels "0","1": 2 2 1 2 2 1 1 1 1 1 ...
 $ weathersit: Factor w/ 3 levels "1","2","3": 2 1 1 1 1 1 1 3 1 2 ...
 $ temp      : num  0.46 0.28 0.54 0.52 0.74 0.64 0.14 0.42 0.24 0.44 ...
 $ atemp     : num  0.455 0.288 0.515 0.5 0.682 ...
 $ hum       : num  0.88 0.52 0.6 0.68 0.62 0.73 0.43 0.41 0.6 0.62 ...
 $ windspeed : num  0 0.104 0.224 0.134 0.164 ...
 $ cnt       : num  6 145 101 614 209 22 20 79 22 236 ...
```

Table 2 – Summary of Training Dataset

```
season  yr        mnth           hr        holiday    weekday      workingday weathersit
 1:366   0:759   Min.   : 1.000   Min.   : 0.00   0:1466   Min.   :0.000   0: 460     1:971
 2:382   1:741   1st Qu.: 4.000   1st Qu.: 6.00   1:  34   1st Qu.:1.000   1:1040     2:403
 3:388           Median : 6.000   Median :12.00            Median :3.000              3:126
 4:364           Mean   : 6.467   Mean   :11.71            Mean   :3.009
                 3rd Qu.: 9.000   3rd Qu.:18.00            3rd Qu.:5.000
                 Max.   :12.000   Max.   :23.00            Max.   :6.000
      temp            atemp             hum           windspeed          cnt
 Min.   :0.0200   Min.   :0.0303   Min.   :0.0000   Min.   :0.0000   Min.   :  1.0
 1st Qu.:0.3400   1st Qu.:0.3333   1st Qu.:0.4800   1st Qu.:0.1045   1st Qu.: 41.0
 Median :0.5000   Median :0.4848   Median :0.6400   Median :0.1642   Median :143.5
 Mean   :0.4991   Mean   :0.4782   Mean   :0.6293   Mean   :0.1920   Mean   :188.9
 3rd Qu.:0.6600   3rd Qu.:0.6212   3rd Qu.:0.7800   3rd Qu.:0.2836   3rd Qu.:281.2
 Max.   :0.9600   Max.   :1.0000   Max.   :1.0000   Max.   :0.6567   Max.   : 957.0
```

Table 3- Initial Model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.469     23.768   0.314   0.7534
```

```
mnth          6.232      1.157   5.389 8.24e-08 ***
hr            6.856      0.576  11.902  < 2e-16 ***
weekday       4.334      1.923   2.254   0.0244 *
temp        175.652    129.160   1.360   0.1740
atemp       164.763    144.715   1.139   0.2551
hum        -199.878     21.929  -9.115  < 2e-16 ***
windspeed    36.976     33.668   1.098   0.2723
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 148.2 on 1492 degrees of freedom
Multiple R-squared:  0.3339,	Adjusted R-squared:  0.3308
F-statistic: 106.8 on 7 and 1492 DF,  p-value: < 2.2e-16
```

Table 4 - Model 1

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   23.1660    20.4851   1.131   0.2583
mnth           6.1969     1.1551   5.365 9.37e-08 ***
hr             6.9184     0.5743  12.047  < 2e-16 ***
weekday        4.2576     1.9221   2.215   0.0269 *
temp         320.3072    20.7006  15.473  < 2e-16 ***
hum         -203.4884    21.0376  -9.673  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 148.2 on 1494 degrees of freedom
Multiple R-squared:  0.333,	Adjusted R-squared:  0.3308
F-statistic: 149.2 on 5 and 1494 DF,  p-value: < 2.2e-16
```

Table 5- Model 2 (No Intercept)

```
Coefficients:
         Estimate Std. Error t value Pr(>|t|)
mnth       6.3426     1.1480   5.525 3.88e-08 ***
hr         7.2000     0.5176  13.911  < 2e-16 ***
weekday    4.9643     1.8178   2.731  0.00639 **
temp     330.6769    18.5606  17.816  < 2e-16 ***
hum     -186.2994    14.5460 -12.808  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 148.2 on 1495 degrees of freedom
Multiple R-squared~~: 0.6802,~~	Adjusted R-squared:  0.6792
F-statistic: 636.1 on 5 and 1495 DF,  p-value: < 2.2e-16
```

Calculated Independetly:
n <- nrow(bikenum)
true.r2 <- 1-sum(model2$res^2)/((n-1)*var(bikenum$cnt))
Multiple R-squared : 0.33

Table 6- Full Model

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6475.9546  6153.1789  -1.052  0.29276
dteday         0.4288     0.4114   1.042  0.29746
season2       -4.0065    13.5984  -0.295  0.76832
season3      -37.1546    19.2522  -1.930  0.05381 .
season4       37.9517    19.3645   1.960  0.05020 .
yr1          -74.4480   150.6390  -0.494  0.62123
mnth         -10.2758    12.6834  -0.810  0.41797
hr             6.7446     0.5525  12.208  < 2e-16 ***
holiday1      -5.4059    25.2660  -0.214  0.83061
weekday        4.8610     1.8393   2.643  0.00831 **
workingday1    8.5894     8.1412   1.055  0.29157
weathersit2   10.6109     8.9463   1.186  0.23579
weathersit3  -32.6926    14.6412  -2.233  0.02570 *
temp         342.5302   129.4994   2.645  0.00825 **
atemp         51.2414   139.4601   0.367  0.71335
hum         -173.3892    23.6002  -7.347 3.33e-13 ***
windspeed     42.9621    32.6925   1.314  0.18901
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 140.7 on 1483 degrees of freedom
Multiple R-squared:  0.4031,	Adjusted R-squared:  0.3966
F-statistic: 62.58 on 16 and 1483 DF,  p-value: < 2.2e-16
```

Table 7- Model 3

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -3.398e+03  3.041e+02 -11.175  < 2e-16 ***
dteday       2.243e-01  2.005e-02  11.186  < 2e-16 ***
season2     -3.908e+00  1.358e+01  -0.288  0.77350
season3     -3.826e+01  1.913e+01  -2.000  0.04573 *
season4      3.823e+01  1.935e+01   1.976  0.04837 *
mnth        -4.127e+00  2.092e+00  -1.973  0.04873 *
hr           6.757e+00  5.506e-01  12.273  < 2e-16 ***
weekday      4.808e+00  1.828e+00   2.630  0.00861 **
weathersit2  1.140e+01  8.875e+00   1.284  0.19929
weathersit3 -2.880e+01  1.432e+01  -2.012  0.04442 *
temp         3.907e+02  3.195e+01  12.229  < 2e-16 ***
hum         -1.812e+02  2.222e+01  -8.155 7.33e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 140.6 on 1488 degrees of freedom
Multiple R-squared:  0.4018,	Adjusted R-squared:  0.3974
F-statistic: 90.85 on 11 and 1488 DF,  p-value: < 2.2e-16
```

Table 8- Model with Corrected Assumptions (Polynomial Terms)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.40568    0.04941  89.167  < 2e-16 ***
weekday         0.04086    0.00865   4.724 2.53e-06 ***
season2         0.19318    0.06491   2.976  0.00296 **
season3         0.22968    0.08193   2.803  0.00512 **
season4         0.40657    0.05572   7.297 4.77e-13 ***
weathersit2    -0.08344    0.04046  -2.062  0.03937 *
weathersit3    -0.69297    0.06421 -10.792  < 2e-16 ***
poly(hr, 8)1   30.08278    0.68857  43.689  < 2e-16 ***
poly(hr, 8)2  -22.06754    0.68554 -32.190  < 2e-16 ***
poly(hr, 8)3  -12.49909    0.70511 -17.727  < 2e-16 ***
poly(hr, 8)4   11.56932    0.67347  17.179  < 2e-16 ***
poly(hr, 8)5  -15.39761    0.67180 -22.920  < 2e-16 ***
poly(hr, 8)6    6.67845    0.66872   9.987  < 2e-16 ***
poly(hr, 8)7    8.67208    0.66849  12.973  < 2e-16 ***
poly(hr, 8)8   -9.22313    0.66797 -13.808  < 2e-16 ***
poly(temp, 2)1  12.65635    1.20392  10.513  < 2e-16 ***
poly(temp, 2)2  -4.11826    0.75684  -5.441 6.18e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.6671 on 1483 degrees of freedom
Multiple R-squared:  0.8014,
F-statistic: 374.1 on 16 and 1483 DF,  p-value: < 2.2e-16
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
```

Table 9- Model with Corrected Assumptions (Trigonometrical Terms)

```
Coefficients:

              Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.627999   0.357872  10.138  < 2e-16 ***
season2           -0.001627   0.046178  -0.035  0.97190
season3           -0.064733   0.059081  -1.096  0.27341
season4            0.202802   0.039658   5.114 3.57e-07 ***
yr1               0.207256   0.025387   8.164 6.84e-16 ***
weekday            0.010970   0.006312   1.738  0.08241 .
weathersit2        0.053087   0.030688   1.730  0.08386 .
weathersit3       -0.163694   0.049860  -3.283  0.00105 **
temp              1.236955   0.114141  10.837  < 2e-16 ***
hum              -0.547595   0.080182  -6.829 1.24e-11 ***
sin(6.28/24 * hr)  -0.194688   0.032148  -6.056 1.76e-09 ***
cos(6.28/240 * hr) -2.243780   0.371773  -6.035 2.00e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 0.4863 on 1488 degrees of freedom
Multiple R-squared:  0.4388,	Adjusted R-squared:  0.4347 
F-statistic: 105.8 on 11 and 1488 DF,  p-value: < 2.2e-16
```

Table 10 - Lasso Coefficients Shrinkage Table

```
(Intercept)   26.5601492
season2        .        
season3        .        
season4       20.0424587
yr1           60.0816884
mnth           0.3465592
hr             5.5524157
holiday1       .        
weekday        .        
workingday1    .        
weathersit2    .        
weathersit3    .        
temp         282.4038797
atemp          2.3611210
hum         -129.6109561
windspeed      .        
```

Table 11 - Lasso Model Summary

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    9.2105    19.0365   0.484    0.629    
yr1           83.7185     7.4809  11.191   <2e-16 ***
hr             6.9267     0.5571  12.433   <2e-16 ***
temp         327.0966    19.6413  16.653   <2e-16 ***
hum         -168.5345    20.0090  -8.423   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 143.9 on 1495 degrees of freedom
Multiple R-squared:  0.3705,	Adjusted R-squared:  0.3688 
F-statistic:   220 on 4 and 1495 DF,  p-value: < 2.2e-16
```

Table 12 - Model 3 (Evaluation Dataset)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -41.064     38.804  -1.058  0.29047    
yr1           74.573     13.614   5.478 6.89e-08 ***
hr             8.159      1.061   7.689 8.14e-14 ***
weekday        4.325      3.428   1.262  0.20762    
weathersit2    3.504     16.295   0.215  0.82982    
weathersit3  -76.030     26.232  -2.898  0.00392 ** 
temp         358.572     36.684   9.775  < 2e-16 ***
hum         -134.805     41.348  -3.260  0.00119 ** 
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 151.6 on 492 degrees of freedom
Multiple R-squared:  0.3696,	Adjusted R-squared:  0.3606
F-statistic: 41.21 on 7 and 492 DF,  p-value: < 2.2e-16
```

Table 13 - Lasso Model (Evaluation Dataset)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -18.9542    38.1135  -0.497   0.6192
season2       -0.0859    23.5797  -0.004   0.9971
season3       34.5160    31.4568   1.097   0.2731
season4       75.2354    20.0763   3.747   0.0002 ***
yr1           75.4480    13.5118   5.584 3.90e-08 ***
hr             7.4130     1.0490   7.067 5.47e-12 ***
temp         136.1402   196.0123   0.695   0.4877
atemp        237.0292   208.8966   1.135   0.2571
hum         -192.3772    38.4744  -5.000 7.99e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 149.8 on 491 degrees of freedom
Multiple R-squared:  0.3854,	Adjusted R-squared:  0.3754
F-statistic: 38.48 on 8 and 491 DF,  p-value: < 2.2e-16
```

Table 14 - Full Model (Evaluation Dataset)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -85.194     45.802  -1.860 0.063487 .
season2       -9.082     24.141  -0.376 0.706936
season3       17.253     35.518   0.486 0.627360
season4       46.608     34.142   1.365 0.172851
yr1           75.379     13.475   5.594 3.71e-08 ***
mnth           3.484      3.653   0.954 0.340637
hr             7.720      1.078   7.164 2.93e-12 ***
holiday1     -42.363     39.945  -1.061 0.289434
weekday        4.212      3.437   1.226 0.220946
workingday1    4.336     14.990   0.289 0.772514
weathersit2    3.921     16.167   0.243 0.808451
weathersit3  -71.626     26.233  -2.730 0.006557 **
temp          86.613    204.679   0.423 0.672363
atemp        295.511    217.297   1.360 0.174481
hum         -150.602     42.492  -3.544 0.000432 ***
windspeed     81.705     62.044   1.317 0.188501
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
Residual standard error: 148.9 on 484 degrees of freedom
Multiple R-squared:  0.4019,	Adjusted R-squared:  0.3834 
F-statistic: 21.68 on 15 and 484 DF,  p-value: < 2.2e-16
```

Table 15 - No linear Model (Evaluation Dataset)

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)    
(Intercept)         3.68580    0.59574   6.187 1.30e-09 ***
season2            -0.06419    0.07380  -0.870 0.384782    
season3            -0.04951    0.09836  -0.503 0.614941    
season4             0.18676    0.06302   2.964 0.003190 ** 
yr1                 0.06754    0.04225   1.599 0.110525    
weekday             0.01146    0.01060   1.081 0.280273    
weathersit2         0.14641    0.05042   2.904 0.003856 ** 
weathersit3        -0.13805    0.08153  -1.693 0.091030 .  
temp                1.45019    0.18687   7.761 4.99e-14 ***
hum                -0.54373    0.13240  -4.107 4.71e-05 ***
sin(6.28/24 * hr)  -0.20430    0.05078  -4.023 6.65e-05 ***
cos(6.28/240 * hr) -2.35482    0.61988  -3.799 0.000164 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4678 on 488 degrees of freedom
Multiple R-squared:  0.4774,	Adjusted R-squared:  0.4656 
F-statistic: 40.53 on 11 and 488 DF,  p-value: < 2.2e-16
```

Table 16 - Cetralized model 3 (Coefficients)

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  140.1589    11.2359  12.474  < 2e-16 ***
hr             6.7498     0.5508  12.255  < 2e-16 ***
weekday        4.8924     1.8276   2.677  0.00751 ** 
temp         391.2920    31.9625  12.242  < 2e-16 ***
hum         -178.4329    22.1732  -8.047 1.71e-15 ***
season2        0.6450    13.1409   0.049  0.96086    
season3      -25.6131    16.7042  -1.533  0.12541    
season4       59.1319    11.4666   5.157 2.85e-07 ***
yr1           81.6102     7.3449  11.111  < 2e-16 ***
weathersit2   11.0756     8.8795   1.247  0.21248    
weathersit3  -29.2185    14.3155  -2.041  0.04142 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 140.7 on 1489 degrees of freedom
Multiple R-squared:  0.4007,	Adjusted R-squared:  0.3967 
F-statistic: 99.56 on 10 and 1489 DF,  p-value: < 2.2e-16
```

Appendix II

R-code

Command 1: Insert Libraries

```
require(psych)
require(corrplot)
library(car)
library(randtests)
library(lmtest)
require(glmnet)
```

Command 2: Import Data

```
bike_11 <- read.csv2("C:/bike_11.csv")
#Delete columns that we don't need to our analysis
bike_11$X<-NULL
bike_11$instant<-NULL
bike_11$casual<-NULL
bike_11$registered<-NULL
bike_11$dteday<-NULL
```

Command 3: Create Numeric Variables

```
bike_11$dteday <- as.Date(bike_11$dteday)
bike_11<-transform(bike_11, temp = as.numeric(temp))
bike_11<-transform(bike_11, atemp = as.numeric(atemp))
bike_11<-transform(bike_11, hum = as.numeric(hum))
bike_11<-transform(bike_11, windspeed = as.numeric(windspeed))
bike_11<-transform(bike_11, cnt = as.numeric(cnt))
bike_11<-transform(bike_11, mnth = as.numeric(mnth))
bike_11<-transform(bike_11, hr = as.numeric(hr))
```

Command 3: Create Factor Variables

```
bike_11<-transform(bike_11, yr = factor(yr))
bike_11<-transform(bike_11, season = factor(season))
```

```
bike_11<-transform(bike_11, holiday = factor(holiday))
bike_11<-transform(bike_11, workingday = factor(workingday))
bike_11<-transform(bike_11, weathersit = factor(weathersit))
```

Command 4: Structure of Data and Basic Statistic

```
str(bike_11)
summary(bike_11)
```

Command 5: Divide Numeric Variables

```
index <- sapply(bike_11, class) == "numeric"
bikenum <- bike_11[,index]
```

Command 6: Divide Factor Variables

```
index <- sapply(bike_11, class) == "factor"
bikefact <- bike_11[,index]
```

Command 7: Visual Analysis for numerical variables

```
par(mfrow=c(2,5))
for (i in 1:nrow(bikenum))
{
  hist(bikenum[,i],col="salmon", main=names(bikenum)[i])
}
```

Command 8: Visual Analysis for Factors

```
par(mfrow=c(3,2))
n <- nrow(bikefact)
barplot(table(bikefact[,1])/n, main="Seasons", col=c("salmon"), names.arg = c("Spring","Summer","Fall","Winter"))
barplot(table(bikefact[,2])/n, main="Years", col=c("salmon"), names.arg = c("2011","2012"))
barplot(table(bikefact[,3])/n, main="Holiday", col=c("salmon"), names.arg = c("Yes","No"))
barplot(table(bikefact[,4])/n, main="Working Day", col=c("salmon"), names.arg = c("Yes","No"))
barplot(table(bikefact[,5])/n, main="Weathersit", col=c("salmon"),args.legend = list(x = "topright"), legend = c("1:Clear Weather","2:Intermidiate Weather","3:Bad Weather"))
```

Command 9: Pairs of Numerical Variables

```
cor(bikefact)
col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))
corrplot(cor(bikenum), method = "color", col = col(200),
         type = "upper", order = "hclust", number.cex = .7,
         addCoef.col = "black", # Add coefficient of correlation
         tl.col = "black", tl.srt = 90, # Text label color and rotation
         # Combine with significance
         sig.level = 0.05, insig = "blank",
         # hide correlation coefficient on the principal diagonal
         diag = FALSE)
```

Command 10: Cnt on Each Numerical Variable

```
par(mfrow=c(2,5))
for(j in 1:(nrow(bikenum)-1)){
plot(bikenum[,j],bikenum[,ncol(bikenum)],xlab=names(bikenum)[j],        ylab='cnt',cex.lab=1.5,col="salmon")
abline(lm(bikenum[,ncol(bikenum)]~bikenum[,j]))
}

par(mfrow=c(2,5))
for(j in 1:(nrow(bikenum)-1)){
boxplot(bikenum[,ncol(bikenum)]~bikenum[,j],   xlab=names(bikenum)[j],   ylab='cnt',cex.lab=1.5,col="salmon")
abline(lm(bikenum[,ncol(bikenum)]~bikenum[,j]),col=2)
}

par(mfrow=c(1,2))
plot(bikenum[,4],bikenum[,ncol(bikenum)],xlab=names(bikenum)[4],ylab='cnt',cex.lab=1.5,col="salmon")
abline(lm(bikenum[,ncol(bikenum)]~bikenum[,4]))

plot(bikenum[,5],bikenum[,ncol(bikenum)],xlab=names(bikenum)[5],ylab='cnt',cex.lab=1.5,col="salmon")
 abline(lm(bikenum[,ncol(bikenum)]~bikenum[,5]))
```

Command 11: Cnt on Factor Variables

```
par(mfrow=c(3,2))
for(j in 1:(nrow(bikenum)-1)){boxplot(bikenum[,ncol(bikenum)]~bikefact[,j], xlab=names(bikefact)[j],
ylab='cnt',cex.lab=2.0,col="salmon") abline(lm(bikenum[,ncol(bikenum)]~bikefact[,j]))
```

```
}
```

Command 12: Initial Regression Model

```
model <- lm(cnt ~., data = bikenum)
summary(model)
```

Command 13: Collinearity Check

```
round(vif(model),1)#Using VIF
vif(step(model, direction = "both"))
```

Command 14: Mode 1

```
model1<- lm(cnt ~.-atemp-windspeed, data = bikenum)
summary(model1)
```

Command 15: No intercept Model

```
model2<- lm(cnt ~.-1-atemp-windspeed, data = bikenum)
summary(model2)
```

Command 16: R^2 Calculation

```
n <- nrow(bikenum)
true.r2 <- 1-sum(model2$res^2)/((n-1)*var(bikenum$cnt))
```

Command 17: Constant Model

```
model0 <- lm(cnt ~ 1, data = bikenum)
summary(model0)
```

Command 18: Check With Anova, If The Extra Parameters Are Insignificant

```
anova(model2,model0)
```

Command 19: Adding Factors, AIC & VIF Calculation

```
fullmodel<-lm(cnt~., data=bike_11) #FULL MODEL
summary(fullmodel)
```

```
vif(fullmodel)
AIC(fullmodel)
```

Command 20: Model 3, Stepwise Method

```
model3<-step(fullmodel, direction = "both")
summary(model3)
vif(model3)
AIC(model3)
```

Command 21: Anova Test, Full with Null Model

```
#Now we test whether the additional parameters in two nested models are zero or not
anova(model2,model3)
```

Command 22: Check Assumption- Check Normality of the Residuals & Constant Variance

```
finalmodel<-model3
Stud.residuals <- rstudent(finalmodel)
yhat <- fitted(finalmodel)
par(mfrow=c(1,3))

plot(finalmodel,col="salmon", which = 2)
{plot(yhat, Stud.residuals,main="Residuals Variance",col="salmon")
  abline(h=c(-2,2), col=2, lty=2)}

{plot(yhat, Stud.residuals^2,main="Residuals Variance R^2",col="salmon")
  abline(h=4, col=2, lty=2)}

shapiro.test(finalmodel$residuals)
# ------------------
ncvTest(finalmodel)
```

Command 23: Check Assumption- Check for the Variance in Quantiles

```
par(mfrow=c(1,3))
yhat.quantiles<-cut(yhat, breaks=quantile(yhat, probs=seq(0,1,0.25)), dig.lab=6)
table(yhat.quantiles)
leveneTest(rstudent(finalmodel)~yhat.quantiles)
```

```
boxplot(rstudent(finalmodel)~yhat.quantiles,col="salmon",main="Variance in Quantiles")
```

Command 24: Check Assumption- Check for residuals linearity

```
residualPlot(finalmodel, type='rstudent',col="salmon",main="Residuals Linearity")
residualPlots(finalmodel, plot=F)
```

Command 25: Check Assumption- Check for Residuals Independence

```
plot(rstudent(finalmodel), type='l',col="salmon",main="Residuals Dependence")
runs.test(finalmodel$res)
dwtest(finalmodel)
durbinWatsonTest(finalmodel)
```

Command 26: Check Assumption- Check for Outliers

```
leveragePlots(finalmodel,col="salmon")
```

Command 27: Lasso

```
X <- model.matrix(fullmodel)[,-1]
lasso <- glmnet(X, bike_11$cnt)
plot(lasso, xvar = "lambda", label = T)

lasso1 <- cv.glmnet(X, bike_11$cnt, alpha = 1)
#Comments:Now we want to find the minimum lamda value.

par(mfrow=c(1,2))

# Results
plot(lasso1)
plot(lasso1$glmnet.fit, xvar="lambda", label=TRUE)

log(lasso1$lambda.1se)
coef(lasso1, s=lasso1$lambda.1se)

Lasso_model<-lm(cnt~.-mnth-holiday-workingday-windspeed-weathersit-weekday, data=bike_11)
summary(Lasso_model)
```

Command 28: Import Evaluation Data

```
bike_test <- read.csv2("C:/bike_test.csv")
#Delete X and instance columns
bike_test$X<-NULL
bike_test$instant<-NULL
bike_test$casual<-NULL
bike_test$registered<-NULL
bike_test$dteday<-NULL
```

Command 29: Create Numeric Variables

```
bike_test<-transform(bike_test, temp = as.numeric(temp))
bike_test<-transform(bike_test, atemp = as.numeric(atemp))
bike_test<-transform(bike_test, hum = as.numeric(hum))
bike_test<-transform(bike_test, windspeed = as.numeric(windspeed))
bike_test<-transform(bike_test, cnt = as.numeric(cnt))
bike_test<-transform(bike_test, mnth = as.numeric(mnth))
bike_test<-transform(bike_test, hr = as.numeric(hr))
bike_test<-transform(bike_test, weekday = as.numeric(weekday))
```

Command 30: Create Factor Variables

```
bike_test<-transform(bike_test, yr = factor(yr))
bike_test<-transform(bike_test, season = factor(season))
bike_test<-transform(bike_test, holiday = factor(holiday))
bike_test<-transform(bike_test, workingday = factor(workingday))
bike_test<-transform(bike_test, weathersit = factor(weathersit))
```

Command 31: Centralize model 3

```
bike_Central<- as.data.frame(scale(bikenum, center = TRUE, scale = F))
bike_Central$cnt<-bikenum$cnt
bike_Central<-as.data.frame(c(bike_Central,bikefact))
sapply(bike_Central,mean)
sapply(bike_Central,sd)
central_model<-lm(cnt~.-mnth-atemp-windspeed-holiday-workingday, data=bike_Central)
summary(central_model)
```

Appendix III

Figures

Figure 11 - Evaluation Data Analysis (Numerical Variables)

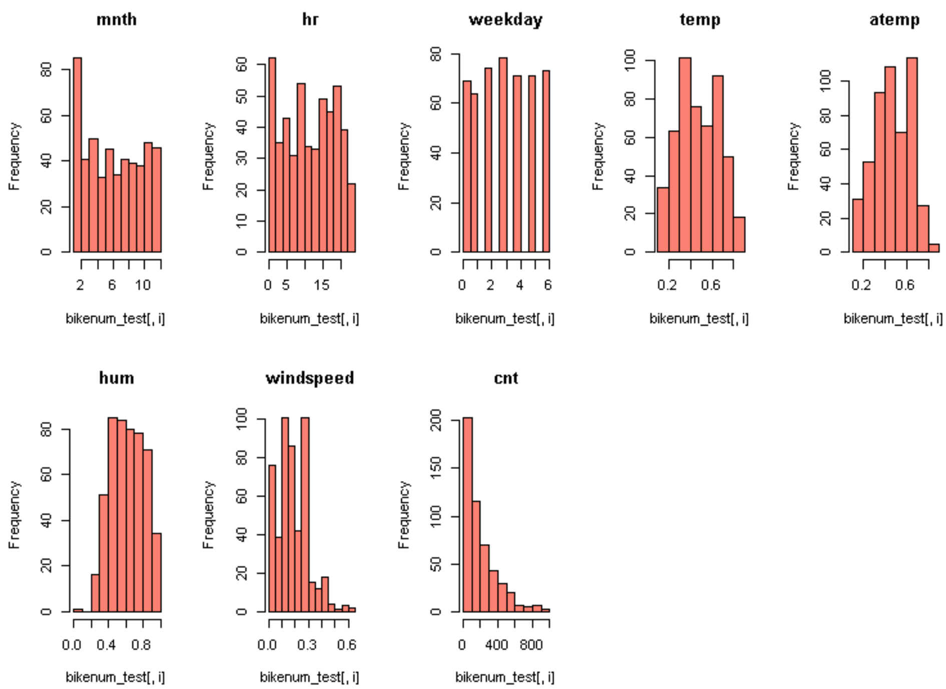

Figure 12 - Evaluation Data Analysis (Categorical Variables)