Regression Modelling Final Assignment

Ambika Huluse Kapanaiah (u3227622)

Overview:

The aim of this article is to illustrate the following two main topics.

1. **Textbook Segment:** Which comprises of the description and summary of the regression models such as, multiple linear regression, elastic net (penalized model), K-Nearest-Neighbors (KNN) regression and Poisson model.
2. **Practical Segment:** Which includes the analysis, modeling and evaluation of Seoul Bike sharing dataset using the 5 models mentioned in textbook segment.

Textbook Segment:

There are 4 regression models which are discussed under this section.

1. Multiple linear regression,
2. Elastic net (penalized model)
3. K-Nearest-Neighbors (KNN) regression
4. Poisson model (GLM)

Multiple Linear Regression:

Background: In simple linear regression we have one dependent and one independent variable.

In case of multiple linear regression, we have multiple independent (predicting) variables. In multiple linear regression, the value of response y is discrete or numerical.

Y = b0 + b1x1 + b2x2 +b3x3 +…… bnxn { where B0 is the intercept, and B1,b2,b3…bn are the coefficients of predictor variables x1, x2, x3…..xn

For example, the grades obtained by student depends on the number of hours they studied, Number of hours they slept. So here the response variable Y is the grades obtained by student and the predictors x1 and x2 are number of hours they studied and Number of hours they slept.

Assumption of Multiple linear regression:

We can categorize the assumption under 2 sections here.

1. Based on the Relationship among variables
   a. Linear relationship between variables and response: the predictors must have some linear relationship (either positive or negative) with the response variable.

We can use correlation coefficients or can plot scatterplots to check. If it is nonlinear, we could consider dropping those features or can take lor or some statistical transformation.

b. no multicollinearity: If the predictor variables are highly correlated among themselves, it is better to drop one of the features among two. It makes it hard to interpret the model and creates problem of overfitting.
c. No endogeneity: If one or more predictors are correlated with error term (actual-predicted y value) it is called endogeneity.
   So, the multiple linear regression equation will be:
   Y = b0 + b1x1 + b2x2 +b3x3 +…… bnxn + Error
   If, Error = f(x)?. It is also called as omitted variable bias. If we have omitted some predictor due to various reasons, but that predictor plays an important role in predicting response or if that predictor has some relationship with another predictor. The Multiple linear regression equation will be having the error term replacing the omitted predictor. This is called endogeneity. So, it is better to start modelling with all the predictors involved. Removal of columns should be done using feature selection or adjusted R-Squared value.

2. Based on the Behavior of the data:
   a. Large sample size: If sample size is large enough, we would be able to get higher adjusted R-squared value at the same time the standard error of adjusted r-squared will get smaller approaching to zero.
   b. Normality of residuals: when the scatter plot of the residuals is plotted, it should follow normal distribution. Majority of the observation to be centered around the mean. If this is not satisfied, we need to go for statistical transformation, or chose other than multiple linear regression such as polynomial and so on.
   c. Homoscedasticity: Spread of the data place an important role in Multiple linear regression. If the spread of the data is not good it is more likely not passing the assumption of linearity, we need to check for outliers and remove them or we need to choose another model such as polynomial.

Along with the above assumptions we have one more issue to deal with, which is called as dummy variable trap:

The multiple linear regression can handle only numbers. So, we convert categorical variables to ordered data. For example, if Season has Autumn Spring Winter and Summer, we will have 4 columns respectively for each saying Autumn as 0 Spring as 1 Winter as 2 and Summer as 3.

This is called Label encoding. But there will be a redundant variable here, if we know the First 3 values from 0,1,2 we will know what is the value of the column 3(Summer). If we add the redundant variable, it overfit the model. Better to drop one of the variables among the 4.

When to use Multiple linear regression:

We can use Multiple linear regression,

1. If the business application is to know the relationship between the features and to know them at certain value of an independent feature. For example, how rainfall, temperature, and amount of fertilizer added affect crop growth or may be expected yield at certain levels of rainfall, temperature, and amount of fertilizer.
2. It passes all the assumptions mentioned above.

Advantages:

1. Finding relationship between one or more predictive variables with response variables is easy. For example, number of bedrooms and size of the home has strong relation to prize of the home at the same time proximity to schools will not have if the retired group is considered.
2. The ability to find outliers. For example, the salary paid in an organization is dependent on the number of hours worked, department size and budget while seniority doesn't. we will have exceptions like some person is being paid high (overpaid) compared to others at same level.

Disadvantages:

1. If the data is incomplete, it would end up in bad prediction. For example, if the 7 houses are reviewed out of 10 which are bought by only young parents, we would end up deciding proximity to school is also a strong predictor to Sale price.
2. Based on the outliers we can predict that one person is highly paid and can remove that outlier completely. But what if the person is involved in some other work and highly skilled compared to others which are not considered which comes under falsely concluding that a correlation is a causation

Elastic net (penalized model):

Background:

The penalty models (regularized models) are used to improve generalization and to prevent overfitting. We have Lasso, Ridge and Elastic net penalty models here. Basically, they add penalty to the error term.

Ridge makes the coefficients to shrink, and Lasso makes the coefficients to zero. Elastic uses both the methods from Lasso and Ridge. They achieve this by modifying the coefficient values.

Penalized models are also called as shrinkage or regularization methods. Groupings and variables selection are the key roles of the elastic net technique. Elastic net produces a regression model

which is penalized with both the L1-norm (Lasso) and L2-norm (Ridge). Some of the coefficients will be shrinking as in Ridge and some of the coefficients made equal to zero as in Lasso. We will select the best parameters α and λ which minimize the cross-validation error.

The mathematical formula for Elastic net regression is as follows.

SSResid + λ * (1 − α Σβi 2 /2 + αΣ|βi|), lambda is the amount of penalty can be fine-tuned by a constant.

When to use Elastic net Regression:

Whenever the number of features is huge in the dataset, we can choose elastic net regression. Elastic net decides on which coefficients to make zero, which one to shrink and to do both, when the variables are correlated and need to drop one among them. This technique is most appropriate when the number of features is greater than the number of samples used.

Advantages:

1. Produces sparse model with good prediction accuracy by encouraging grouping effect of both Lasso, Ridge regression and both.
2. Overcomes with the problem of overfitting due to huge number of features.

Disadvantages:

1. Computational cost is high.

K-Nearest Neighbors Regression:

Background:

KNN is a machine learning algorithm used both for regression and classification. It approximates the association between independent variables and the continuous response variable by taking the average of the observations in the same neighborhood. The algorithm uses feature similarity as the key.

KNN makes use of mean of the k nearest values for the response feature. If we need to predict the continuous response, KNN calculates the average of the k most similar training observations.

The similarity measure uses a distance function most commonly Euclidean distance.

The value of K could be chosen from a range of 1 to the size of training data set based on the accuracy measure ex: RMSE

For example, if we want to predict the loan size to be approved for a new individual, based on Age income credit history we need to place the data from the new individual to check where he

stands. If the number of neighbors around him are more from the one with high income and good credit history, the loan will be approved in certain range size.

As KNN used distance calculation, normalization plays an important role here. So, it is always better to take all the features on the same scale before proceeding to fitting the model.

Choosing K value is another important fact. Smaller the value of K model will be less accurate and unstable. If it is larger, more stable, and accurate model fitting and prediction and subjected to less error. But larger K make sit computationally expensive. For regression, the best value of k is chosen to minimize an error metric such as RMSE.

When to use KNN:

Whenever the business application has incomplete data and needed a quick learning method, we can use this KNN as it has no learning period as well as it doesn't effect on the accuracy or R-squared value even after adding more data into the dataset.

Advantages:

1. KNN algorithm much faster than other algorithms. It never learns from data. It just keeps the data and predicts when the data is fed. So, it is also called as lazy learner.
2. New data can be added anytime, which will not impact the accuracy of the algorithm.
3. Easy to implement. As we need only K value and distance calculation
4. KNN works better than linear regression when the data have high Signal to Noise ratio.

Disadvantages:

1. Doesn't work well with large dataset.
2. Doesn't work with larger number of dimensions.
3. Needs feature scaling.
4. Sensitive to Noisy data such as missing values and outliers.

Poisson Model:

Background:

Poisson model is often used to model the data which includes counts as response variables. For example, Number of patients entering the clinic each day. Poisson model comes under generalized linear models. GLM involves a set of independent response variables $Y1, \dots, Yn$ with $E\ Yi = \mu$. And a set of vector predictors $x1, \dots, x$k. A link function relating to the mean of the responses.

$g\ \mu i = \beta 0 + xi1\beta 1 + \dots + xik\beta$

So, for Poisson g,

$g(\mu)$ = log(μ), the log link is used.

If we chose $g(\mu)$ = μ and error distribution of residuals as normal, we would end up in linear regression again.

So Linear regression is also a special case of GLM. In Poisson, the response is a count which can vary from 0 to infinity. We should map this range to real number line that is -Infinity to +Infinity. This can be achieved using log link.

$g(\mu)$ = log(μ)

Interpretation: In the following equation

log(μ) = $\beta 0 + \beta 1 x 1 + \beta 2 x 2 + \cdots + \beta k x$, consider holding the other factors constant.

And if x1 increases by 1 unit log(μ) = $\beta 0 + \beta 1(1)$. … The response will change by the multiplicative factor of e^ $\beta 1$ or increase by e^ $\beta 1$ -1

When to use Poisson:

Whenever we have the response variable as Counting rare events, for example the number of volcanoes in a year or counting events in a time interval for example, number of visits to a clinic each day. In such situations we could chose to model using Poisson. And if the response variable follows rightly skewed curve, the Poisson modeling can be applied.

Advantages:

It overcomes some of the problems of the normal model.

1. Poisson model has minimum value of 0 and will not predict negative values. If the mean and most typical values are close to 0, this model is the ideal one.
2. It is fundamentally rightly skewed model. With the data characterized with long tail, this modelling works the best.
3. Very slight round off errors could be found as Poisson estimates by maximum likelihood method by making estimates identical to actual data.
4. When count response is to be predicted Poisson is better compared to Normal models. Poisson is very less prone to over or underestimating of number of incidents for most records as in normal model.

Disadvantages:

On the other hand, the Poisson model is not the perfect one.

1. The problem is that count data are usually over dispersed. Leading to the residual errors that is the difference between the actual and predicted values for each segment of the prediction, will be greater than what is expected. So based on the probability significant test we will end up in selecting many predictors that really should not be selected.

2. There is a condition with under-dispersion in the residual errors. With under distribution, one cannot assume a normal distribution because, it will still underestimate the high values of the Response variable.

Practical Segment:

The chosen dataset is **Seoul Bike Sharing Demand Data Set**. The response variable is the **Rented Bike count** depends on various predictors from dataset such as Date : year-month-day, Hour - Hour of the day, Temperature-Temperature in Celsius, Humidity - %, Windspeed - m/s, Visibility - 10m, Dew point temperature – Celsius, Solar radiation - MJ/m2, Rainfall – mm, Snowfall – cm, Seasons - Winter, Spring, Summer, Autumn, Holiday - Holiday/No holiday, Functional Day – NoFunc (Non Functional Hours), Fun(Functional hours).

To begin with I have loaded the below packages in the R studio.

```
library(data.table)
library(ggplot2)
library(purrr)
library(tidyr)
library(dplyr)
library(GGally)
library(rmarkdown)
library(caret)
library(reshape2)
library(car)
library(modelr)
```

Read and prepare the data:

Reading the SeoulBikeData.csv file which is downloaded from the weblink: *https://archive.ics.uci.edu/ml/machine-learning-databases/00560/*

```
#####**************************************************************************
##### Read and prepare the data
#####**************************************************************************
Seoul_BikeSharing_df <- read.csv("SeoulBikeData.csv")
head(Seoul_BikeSharing_df)
View(Seoul_BikeSharing_df)
```

There are no missing values found in the data.

```
> colSums(is.na(Seoul_BikeSharing_df))
                     Date         Rented.Bike.Count                      Hour
                        0                         0                         0
          Temperature..C.                Humidity...          Wind.speed..m.s.
                        0                         0                         0
        Visibility..10m. Dew.point.temperature..C.    Solar.Radiation..MJ.m2.
                        0                         0                         0
            Rainfall.mm.             Snowfall..cm.                   Seasons
                        0                         0                         0
                 Holiday           Functioning.Day
                        0                         0
```

After careful observation I found that Date feature is not important to be considered for modelling. So decided to drop “Date”. The Hour column has the values with levels 0 to 23, and cannot be considered as Numerical so, it is converted to factor. I have also renamed the column names so that makes easy to understand and access features. Mainly, the name of the response variable “**Rented.Bike.Count**” is renamed to “**Demand**” which makes it easy to understand.

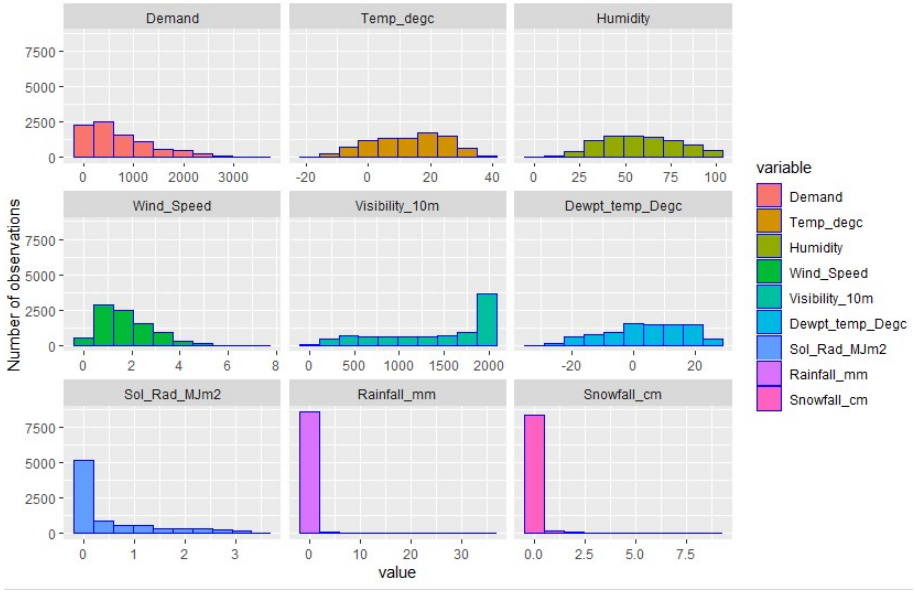
```
> colnames(Seoul_BikeSharing_df)
 [1] "Demand"           "Hour"             "Temp_degc"        "Humidity"
 [5] "Wind_Speed"       "Visibility_10m"   "Dewpt_temp_Degc"  "Sol_Rad_MJm2"
 [9] "Rainfall_mm"      "Snowfall_cm"      "Seasons"          "Holiday"
[13] "Functioning.Day"
```

Exploratory Data analysis:

Exploratory data analysis For Numerical features:

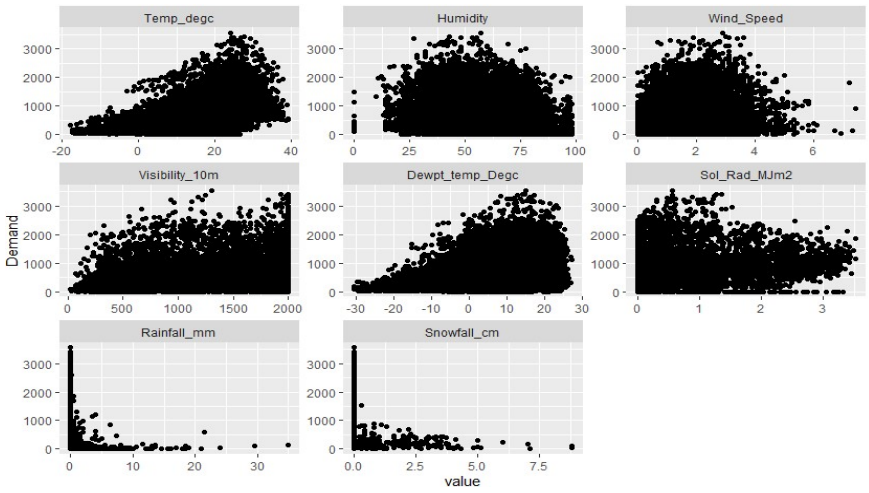
For data analysis I have chosen first to visualize the numerical columns from the dataset.

1. Below histogram shows the distribution of all the numerical features including the response variable Demand.

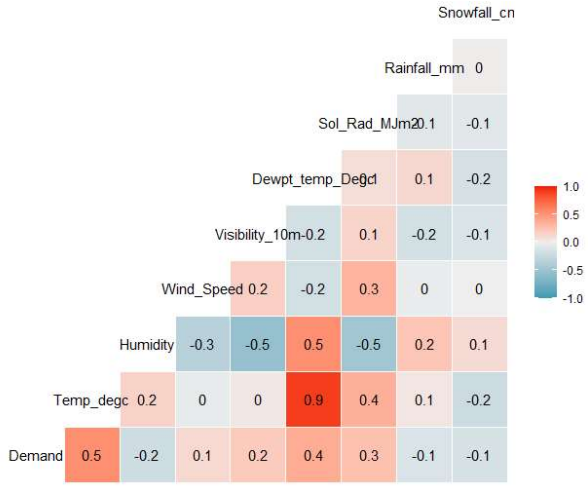


Distribution of response **Demand** looks rightly skewed, same with Windspeed except few counts on the left side. Temp_degc , Humidity and Dewpt_temp are mostly normal distributed.

2. The scatterplot below shows the relationship of all numerical features versus Response variable Demand (Rental Bike count). This shows strong relation of Demand Resose with most of the columns. Rainfall, Snowfall, Solar_rad, windspeed are in negative relation. with Demand. Temp, Visibility. Dewpt_temp are in positive relationship with Demand.



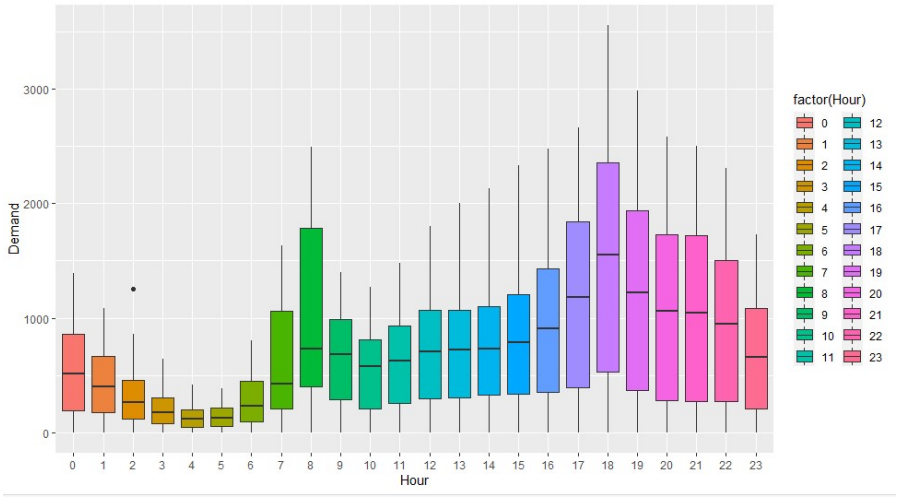
3. This below correlation plots show the correlation coefficients of all numerical features.



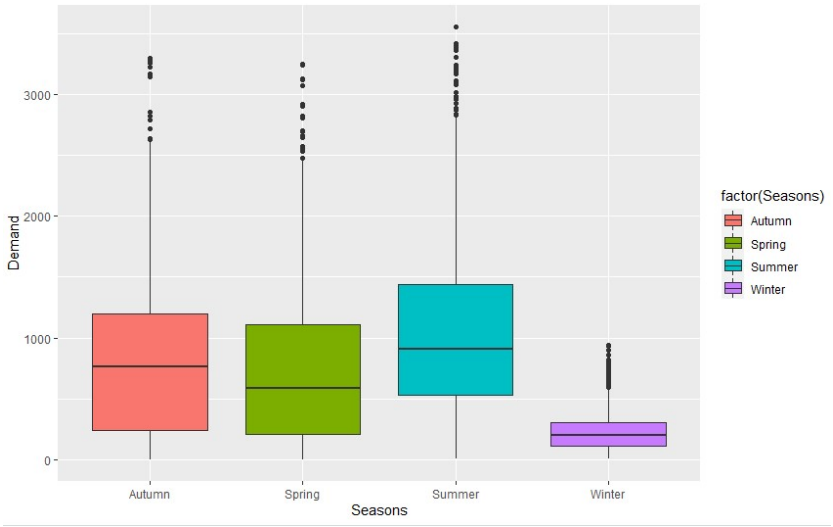
Looks like Temp_degc and Dewpt_Degc has multicollinearity issue (with 0.9 correlation value). So, for Multiple linear regression we could use only one feature among them to fit the model. The correlation coefficients varying from -0.1 to +0.4 with Demand variable. Temo_degc has has highest correlation of 0.5 with the Demand and next comes Dewpt_temp with 0.4 and Sol_rad with 0.3. Visibility and humidity share same correlation of 0.2 but Humid is negatively related to Demand.

Exploratory data analysis For Categorical variables:

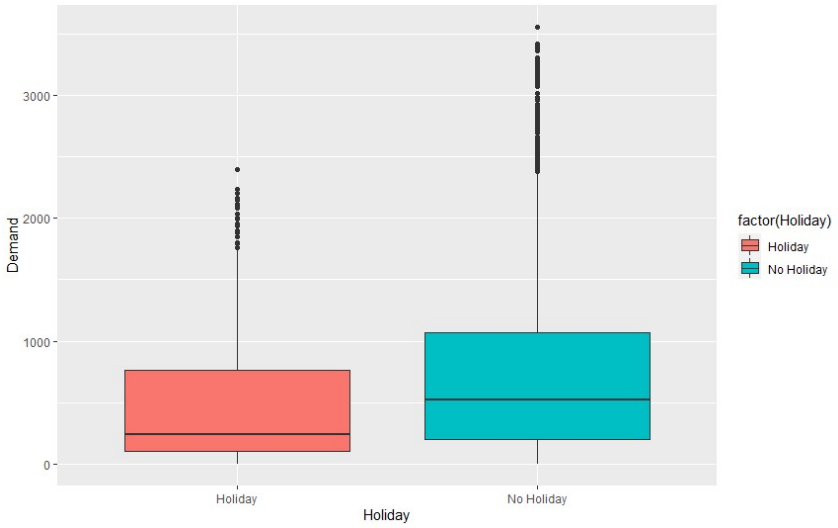
1. This boxplot is for Demand versus Hour. The number of rented bikes count varies throughout the day.



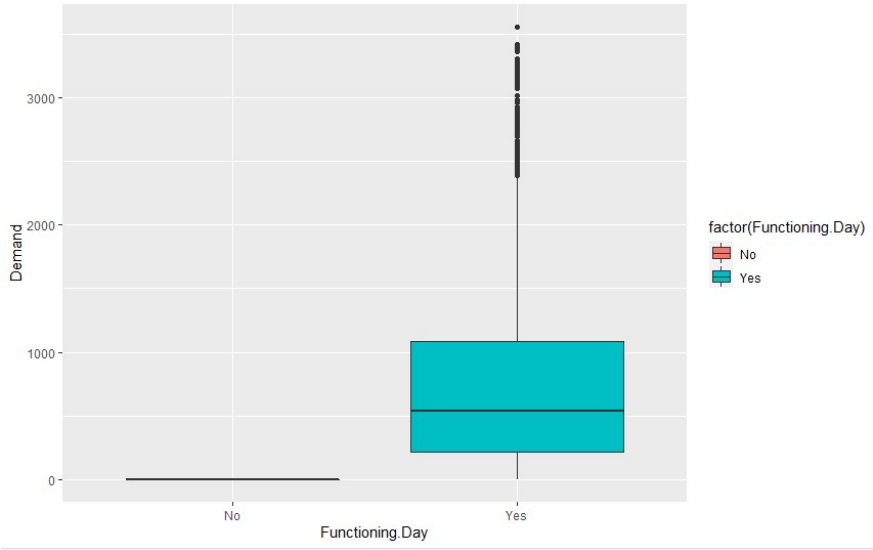
2. This boxplot is for Demand versus Seasons: During Winter the Demand for rented bikes is lesser compared to other seasons, During Summer it is more and next comes Autumn followed by Spring. We can also notice few outliers here on each season which comes under exceptional observations.



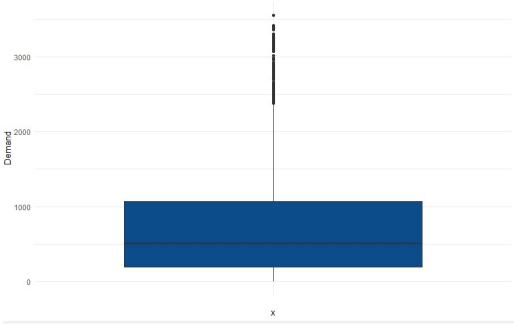
3. The following boxplot is for demand versus Holidays
During working days use of rented bikes is in demand compared to Holidays with some exceptional cases in each segment.



4. This below boxplot is for Demand versus Functioning. Day: Almost all the days the rented bikes are available for rent. Only countable number of days it is not available. And the usage is more when it is available.



Outlier checks: Few outliers seen on Demand, the response variable.



The problem statements for the Bike-sharing dataset can be considered as follows.

1. Did the humidity effect the demand of rental bikes usage?
2. Did really on raining and snowfall days the demand shoot-up? how could we improve if no demand on these days.
3. Did each hour showed any difference in the demand of the bikes which were rented?
4. What is the relation between temperature, Solar radiation, Dewpt_temperature with Demand and so on…

I have chosen the following 4 regression models to fit the model to solve Bike sharing Dataset problem.

1. Multiple linear regression,
2. Elastic net (penalized model)
3. K-Nearest-Neighbors (KNN) regression
4. Poisson model (GLM)

Reasons behind choosing them to solve this problem is as follows.

1. Multiple Linear regression:
   If we need to find out a solution on how to improve rented bike demand on a particular day, we need to understand the relationship between the various features such as temperature, humidity, rainfall, working day/not and so on. And if we assume that the response variable depends on various predictors it is always better to start with Multiple linear regression as a baseline. Simple linear regression doesn't hold good here as the number of predictors are more.
2. Elastic net (penalized model):
   Considering all the predictors to fit model may result in overfitting. After the exploratory data analysis, I was unable to select especially the numerical features, which showed strong relationship with Demand in scatterplot but low correlation value from the correlation plot. As the Elastic net helps in shrinking coefficients to 0 or removing coefficients thereby minimizes the overfitting issues by not considering the few predictors (by giving least lambda weightage). So, this model helps is understanding whether the model fitted using multiple linear regression is overfitted?

3. K-Nearest-Neighbors (KNN) regression:
   For example, if we need to predict the Rented bike demands for a future day based on the historical pattern of data, it is easy for KNN as it works on the method of feature similarity by checking how closely the new point resembles the points in training set. While checking for outliers in Bike sharing dataset, there are prominent outliers found in both predictors and response variable. When this Signal to noise ratio is high, it is better to fit the model using KNN regression.

4. Poisson model (GLM):
   The response variable Rented Bike count (Demand), fundamentally count of rented bikes on hour basis which comes under time interval counts. And also the histogram of the response variable Demand has rightly skewed cure which makes it an obvious choice to select Poisson model for fitting the model using Bike-sharing dataset.

Note: testing and training ratio taken here is 20:80

Fitting the model using Multiple linear Regression:

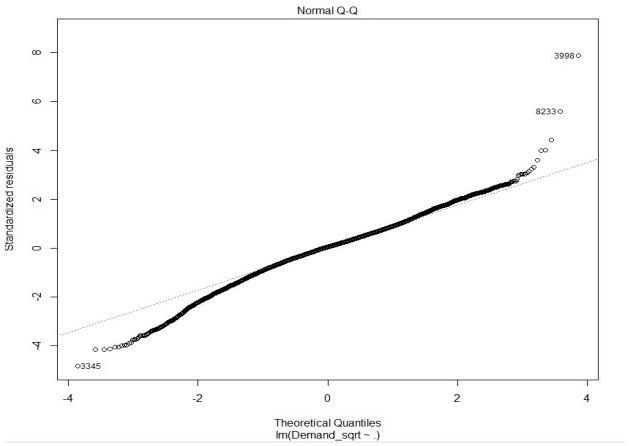
As there are few assumptions, we need to make sure to check on them before modelling with Multiple linear regression.

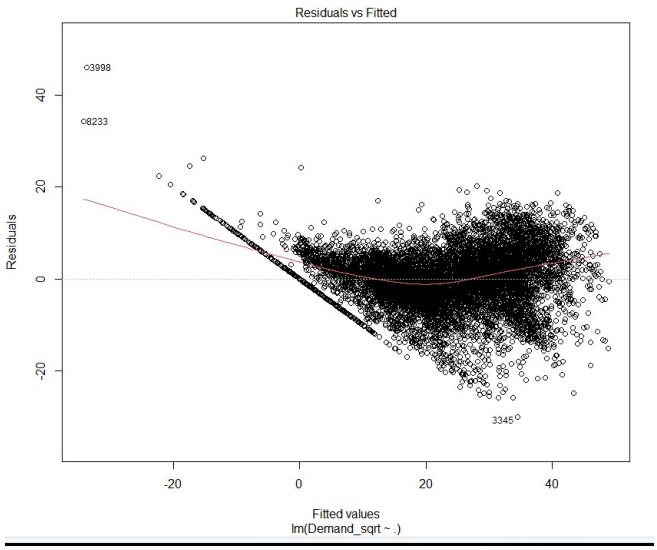
1. Linear relationship: As we checked under EDA section on the scatterplot and correlation coefficients, there is no non-linear relationship between features. Even though the correlation coefficients were lesser than or equal to 0.5, we didn't find very low coefficients such as lesser than 0.05. We can assume here that Linear relationship assumption is passed.
2. Normality: not to violate normality assumption, I have considered the square root transformation of the response variable. As the histogram of response variable shown non normality.
3. no multicollinearity: As we saw on the correlation matrix, there is a multicollinearity issue found between 2 features that is Temp_degc and Dewpt_degc, We can remove Dewpt_degc before fitting the model which is having less coefficient than Temp_degc. so that, this issue will be solved.
4. No endogeneity: As discussed in the textbook section, we should be considering all the features to fit the model so that the error term is very less.
5. Normality of residuals: this should be checked once the model is fitted.
6. Homoscedasticity:
   As discussed before, the variance of the data should be checked, if the spread is more then we will fail in the linear assumptions which we made. So, we need to check for outliers and remove them before fitting the model.
7. Dummy variable trap:
   As we have 4 categorical features such as Seasons, Holiday, functioning day and Hour, we need to create dummy variables and consider only n-1 levels of columns from each categorical feature. For example, if the holiday has holiday Yes and no, we can consider only Holiday No and omit holiday yet, which is obviously the negate of the holiday no value (that means if holiday no is 0 holiday yes is 1). This removes redundancy and solves overfitting.

Fitting 2 models here for Multiple linear regression:

1. First dataset **Seoul_BikeSharingMLM1**: where in no outliers are removed/treated. And considering all the predictors for fitting the model.

```
Residual standard error: 6.229 on 8723 degrees of freedom
Multiple R-squared:  0.7507,    Adjusted R-squared:  0.7497
F-statistic: 729.7 on 36 and 8723 DF,  p-value: < 2.2e-16
```

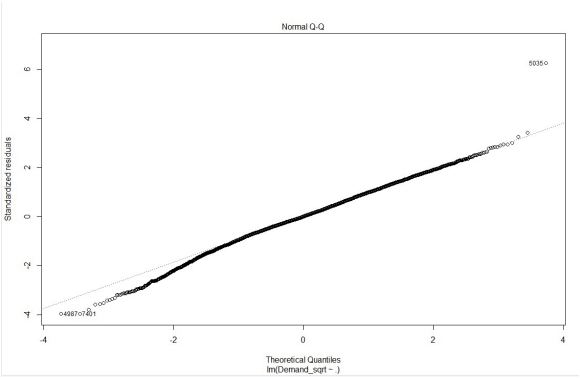


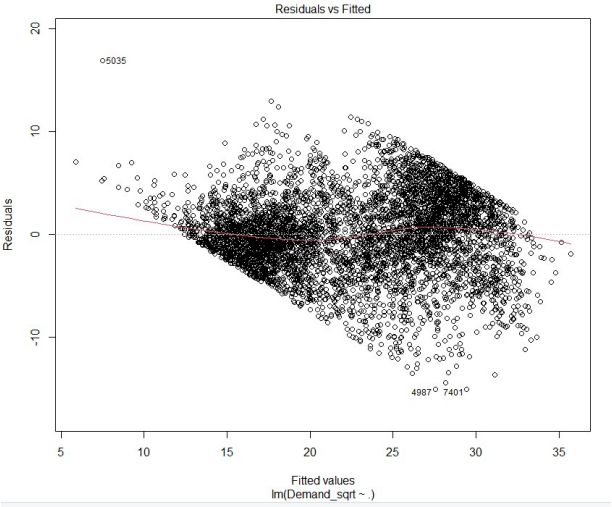


Able to get Adjusted R-squared value of 0.7497 and RSE 6.229 on 8723 degrees of freedom. Few features have more than alpha (0.05), so they are insignificant. The residual plot is somewhat linear and not normal. P value of the model is p-value: < 2.2e-16 which is less than 0.05 alpha model is significant.

2. Second dataset **Seoul_BikeSharingMLM2**: Dewpt_Dedc feature is removed to remove multicollinearity, removed outliers in the Inter quantile rage of 0.2 to 0.8.

```
Residual standard error: 3.817 on 5224 degrees of freedom
Multiple R-squared:  0.6721,    Adjusted R-squared:  0.6699
F-statistic: 314.9 on 34 and 5224 DF,  p-value: < 2.2e-16
```





Adjusted R-2 is 0.6699 and RSE is 3.817. Residual plot is somewhat linear and somewhat normal. p-value: < 2.2e-16 which is less than 0.05 alpha model is significant.

Multiple linear regression Conclusion:

As the error is reduced in 2nd model even though adjusted R-squared value is low in 2nd model. 2nd model can be considered. Even Residual plot shows more linear and more normal than from model 1. We could write the formula as

Y = intercept+/-coefficient1*x1+/-coefficient2*x2……..and so on

Fitting the model using Elastic Net Regression:

Two models are used to fit the elastic net model

1. First dataset **Seoul_BikeSharingEL1:** considering all features and Without normalizing the output feature Demand (no square root).

```
+ )
  RMSE.net Rsquare.net
1  953.109   0.6349481
```

2. Second dataset **Seoul_BikeSharingEL2:** with squareroot normalization of response feature Demand

```
  RMSE.net Rsquare.net
1 6.276002   0.7539977
```

Elastic net regression Conclusion:

R-squared value is more and RMSE is less for the model with taking normalization of response variable.

So the 2nd model **Seoul_BikeSharingEL2** better than the **Seoul_BikeSharingEL1**. When we look into RMSE of first model, 953.109 is too high. We conclude that normalization is very important factor to be considered.

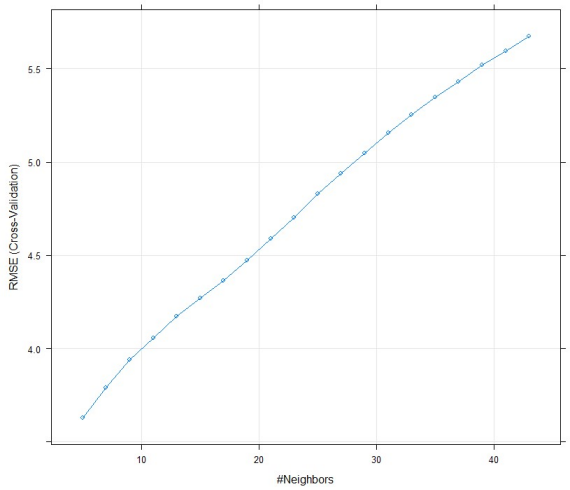
Fitting the model using KNN Regression:

Considered the dataset for fitting with (on normalized sqrt (Demand) response)

Dataset name: **Seoul_BikeSharingKNN1**

model Knn_1: with Tune Length = 20 (number of K neighbors) and 10- fold cross validation

```
> knnsummary1
      RMSE       RSQ      MAE
1 2.657094 0.9563218 1.767294
```



RMSE is very less for the K value = 5. R-Squared = 0.9563, MAE = 1.7672

Conclusion from KNN regression:

Got the highest R-Squared value-> 0.9563218 with RMSE-> 2.657094 and MAE->0. 1.767294. Plot is shown for less value of RMSE, we have got for K = 5.

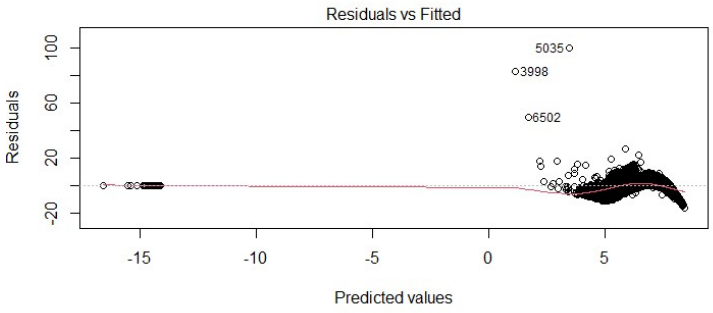
Fitting the model using Poisson Regression:

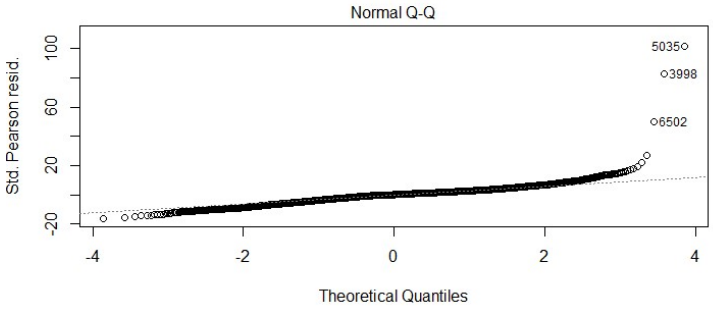
Using all the features here except Dewpt_temp:

AIC: 200663,

Interpretation: one unit change in the predictor variable, the difference in the logs of expected counts is expected to change by the respective regression coefficient, given the other predictor variables in the model are held constant.

```
> PoissonSummary1
      RMSE       RSQ       MAE
1 21.18306 0.326077 18.21061
```





Final-Conclusion: Out of all the model fitted above:

Out of all KNN regression best fits the model and predicts. The best R-Squared value achieved here-0.9563.

Multiple Linear Regression:	
MLM1:	Residual standard error: 6.229 on 8723 degrees of freedom Multiple R-squared: 0.7507, Adjusted R-squared: 0.7497 F-statistic: 729.7 on 36 and 8723 DF, p-value: < 2.2e-16
MLM2:	Residual standard error: 3.817 on 5224 degrees of freedom Multiple R-squared: 0.6721, Adjusted R-squared: 0.6699 F-statistic: 314.9 on 34 and 5224 DF, p-value: < 2.2e-16
Elastic net Regression:	
Elastic1	+) RMSE.net Rsquare.net 1 953.109 0.6349481
Elastic2	RMSE.net Rsquare.net 1 6.276002 0.7539977
KNN:	
	> knnsummary1 RMSE RSQ MAE 1 2.657094 0.9563218 1.767294
POISSON	
	> PoissonSummary1 RMSE RSQ MAE 1 21.18306 0.326077 18.21061