The Bike-sharing rental process

1. **Problem description**

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return has become automatic. Through these systems, user can easily rent a bike from a position and return at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed of over 500 thousand 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 important events in the city could be detected via monitoring these data.

1. **Analysis process**

2.1 **Data collection and description**

The source for the dataset is the Capital Bikeshare system, USA. 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 two hourly and daily basis and then extracted and added the corresponding weather and seasonal information. Weather information are extracted from <http://www.freemeteo.com>. The prediction of bike rental count hourly or daily based on the environmental and seasonal settings.

2.2 **Understanding DataSet**:

-- holiday: This is a categorical variable with levels (extracted from <http://dchr.dc.gov/page/holiday-schedule>).

--weathersit: This is a categorical variable with levels [1,2,3,4]

--temp : Normalized temperature in Celsius. The values are divided to 41 (max)

-- hum: Normalized humidity. The values are divided to 100 (max)

-- windspeed: Normalized wind speed. The values are divided to 67 (max)

--cnt: count of total rental bikes including both casual and registered. This is the response variable

--season: seasons (1: spring, 2: summer, 3: fall, 4: winter)

--yr: year (0: 2011, 1: 2012)

--Month: month (1 to 12)

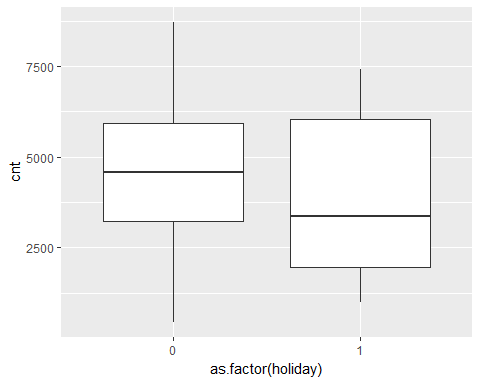
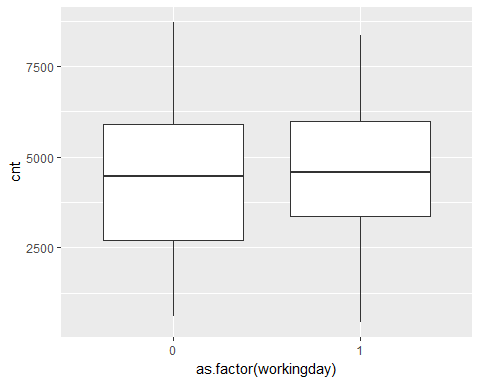
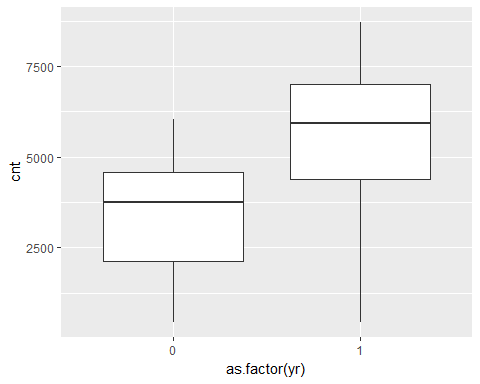
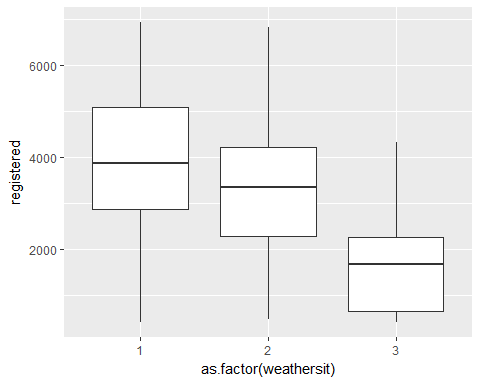
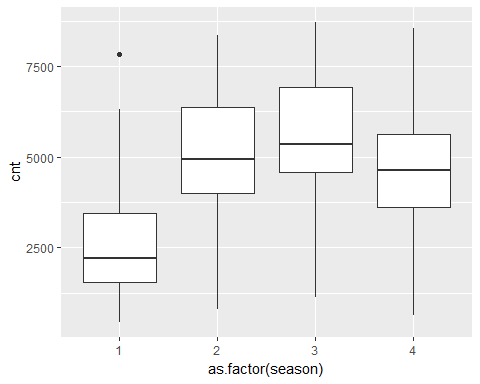
--Casual: count of casual users

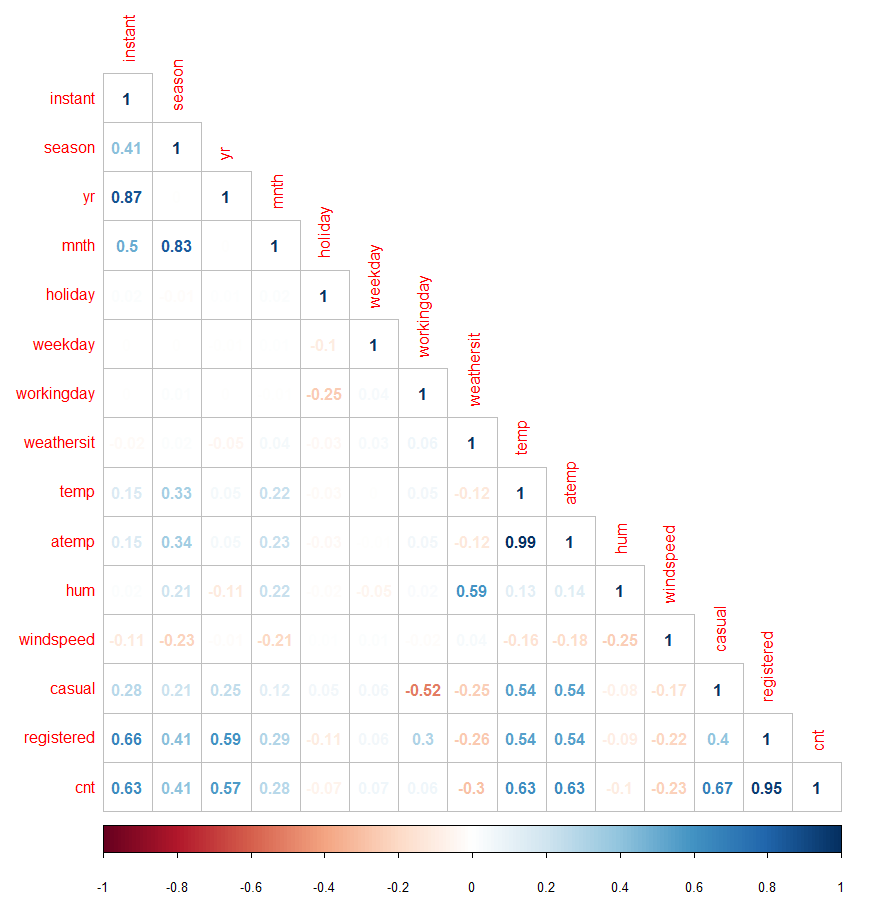
--registered: count of registered users

--workingday: factor variable with levels [0,1]

2.3 **Data Analysis with Plots**

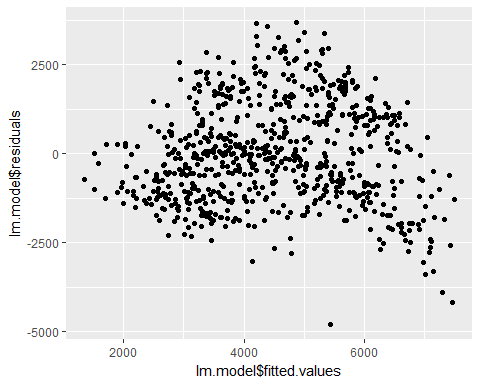
Here we have three response variables, cnt, registered and casual.





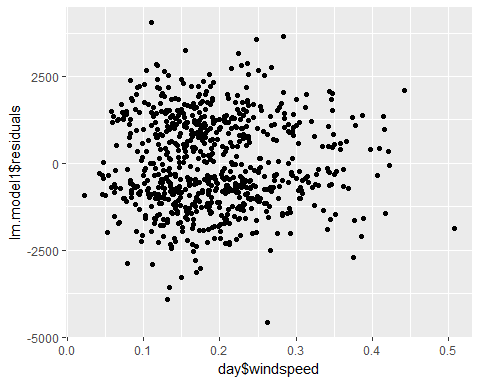
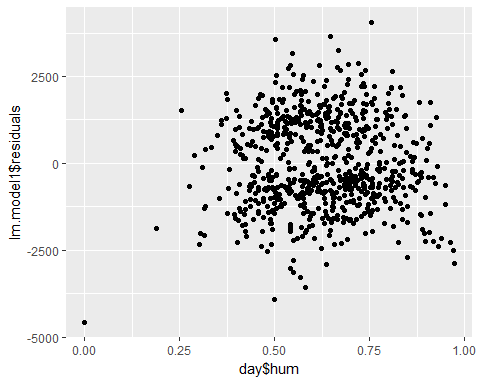
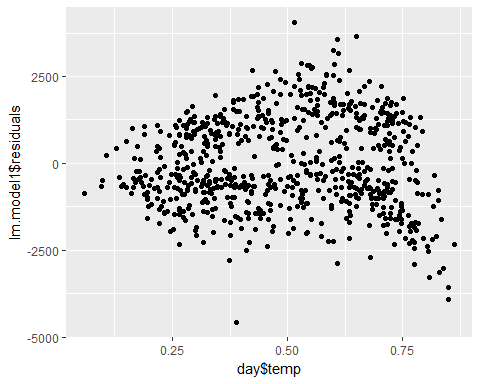
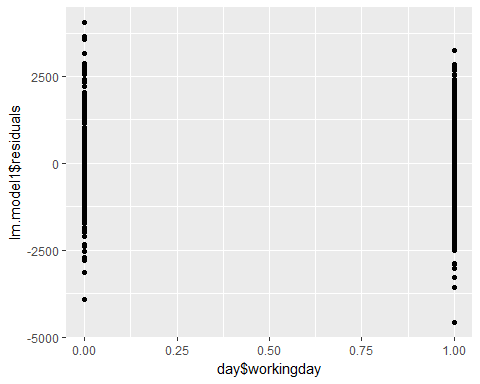
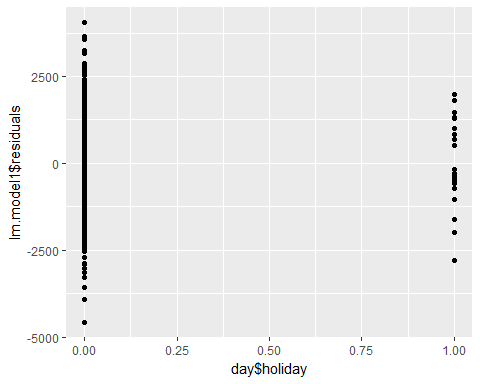
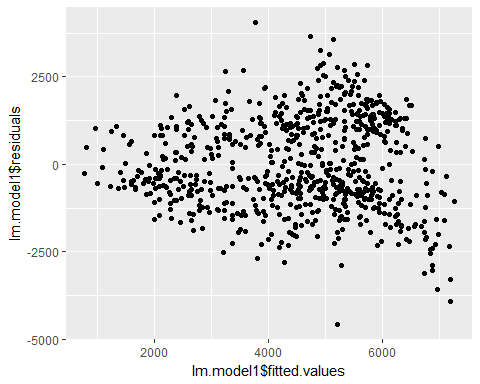
From the above corelation plot we see that the 'temp','hum' and 'windspeed' are good predictors for fitting the model.

##   
## Call:  
## lm(formula = cnt ~ as.factor(holiday) + as.factor(workingday) +   
## temp + hum + windspeed, data = day)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4806.6 -1085.0 -67.9 1037.0 3689.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4007.22 346.82 11.554 < 2e-16 \*\*\*  
## as.factor(holiday)1 -560.71 325.25 -1.724 0.0851 .   
## as.factor(workingday)1 80.26 117.02 0.686 0.4930   
## temp 6597.99 292.57 22.552 < 2e-16 \*\*\*  
## hum -3050.68 384.37 -7.937 7.89e-15 \*\*\*  
## windspeed -4678.19 710.16 -6.587 8.60e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1421 on 724 degrees of freedom  
## Multiple R-squared: 0.4625, Adjusted R-squared: 0.4588   
## F-statistic: 124.6 on 5 and 724 DF, p-value: < 2.2e-16



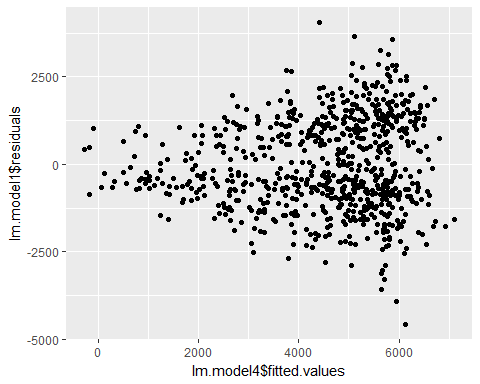
From the above summary statistics the p-value is not significant for workingday. And the plot shows that there is no constant variance. So, we will fit another model by considering the 'season' column in the dataset.

##   
## Call:  
## lm(formula = cnt ~ as.factor(holiday) + as.factor(workingday) +   
## temp + hum + windspeed + as.factor(season), data = day)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4582.2 -893.7 -223.6 1077.0 4057.6   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3631.05 330.60 10.983 < 2e-16 \*\*\*  
## as.factor(holiday)1 -557.76 302.12 -1.846 0.0653 .   
## as.factor(workingday)1 78.93 108.67 0.726 0.4679   
## temp 6566.29 480.91 13.654 < 2e-16 \*\*\*  
## hum -3679.23 361.89 -10.167 < 2e-16 \*\*\*  
## windspeed -4046.53 669.70 -6.042 2.43e-09 \*\*\*  
## as.factor(season)2 873.97 181.53 4.815 1.80e-06 \*\*\*  
## as.factor(season)3 359.41 239.06 1.503 0.1332   
## as.factor(season)4 1460.52 154.67 9.443 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1319 on 721 degrees of freedom  
## Multiple R-squared: 0.5391, Adjusted R-squared: 0.534   
## F-statistic: 105.4 on 8 and 721 DF, p-value: < 2.2e-16



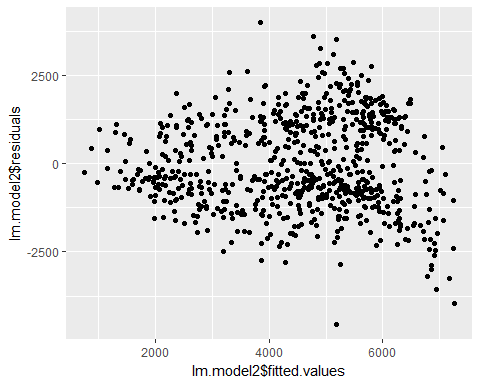
We still see some curvature in the residual vs fitted values plot and the p-value for working day is not significant. From the residual vs temp plot we see there is some curvature. So, adding the higher order term for temp and fitting the model again.

##   
## Call:  
## lm(formula = cnt ~ as.factor(holiday) + as.factor(workingday) +   
## temp + X + hum + windspeed + as.factor(season), data = day)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5506.7 -945.8 -76.6 1057.8 3421.5   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1128.79 409.99 2.753 0.00605 \*\*   
## as.factor(holiday)1 -469.58 285.43 -1.645 0.10037   
## as.factor(workingday)1 61.87 102.63 0.603 0.54683   
## temp 22498.18 1752.11 12.841 < 2e-16 \*\*\*  
## X -16847.12 1789.45 -9.415 < 2e-16 \*\*\*  
## hum -4357.85 349.24 -12.478 < 2e-16 \*\*\*  
## windspeed -4827.61 637.79 -7.569 1.15e-13 \*\*\*  
## as.factor(season)2 542.34 174.99 3.099 0.00202 \*\*   
## as.factor(season)3 671.83 228.16 2.945 0.00334 \*\*   
## as.factor(season)4 1029.31 153.06 6.725 3.59e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1245 on 720 degrees of freedom  
## Multiple R-squared: 0.5896, Adjusted R-squared: 0.5845   
## F-statistic: 114.9 on 9 and 720 DF, p-value: < 2.2e-16



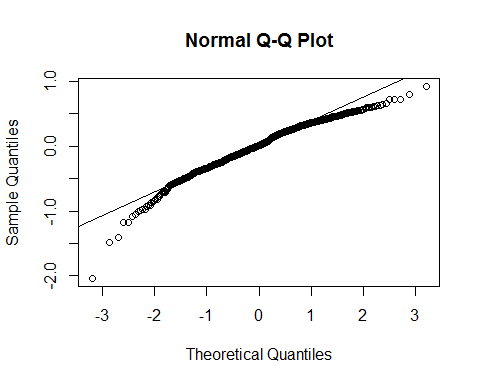
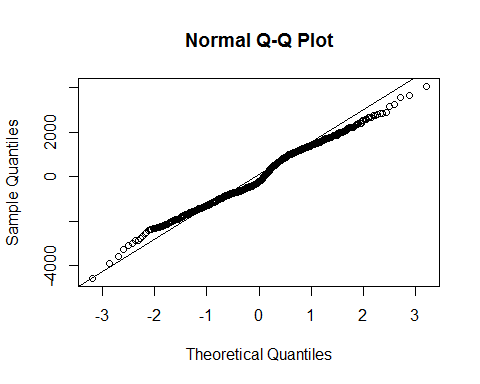
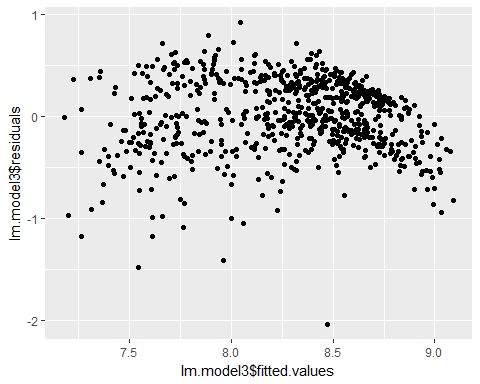
The residuals Vs fitted values we see there is increasing constant variance. The standard error value for temp increased alot eventhough p-value is sognificant. So, we will remove the quadratic term from the model as we need standard error 's' to be small. As p-value is not significant for 'workingday' we fitted the model by removing the 'working day'.

##   
## Call:  
## lm(formula = cnt ~ as.factor(holiday) + temp + hum + windspeed +   
## as.factor(season), data = day)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -4557.9 -922.9 -222.6 1095.6 3998.0   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3679.6 323.7 11.368 < 2e-16 \*\*\*  
## as.factor(holiday)1 -613.2 292.2 -2.099 0.0362 \*   
## temp 6585.8 480.0 13.720 < 2e-16 \*\*\*  
## hum -3677.0 361.8 -10.164 < 2e-16 \*\*\*  
## windspeed -4052.1 669.4 -6.053 2.28e-09 \*\*\*  
## as.factor(season)2 870.6 181.4 4.799 1.94e-06 \*\*\*  
## as.factor(season)3 352.8 238.8 1.477 0.1400   
## as.factor(season)4 1458.6 154.6 9.435 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1318 on 722 degrees of freedom  
## Multiple R-squared: 0.5387, Adjusted R-squared: 0.5343   
## F-statistic: 120.5 on 7 and 722 DF, p-value: < 2.2e-16



As we see from the above plot, still we have some curvature in the residual vs fitted values.So, applied the box-cox transformation for the response variable and fitting the model by adding the 'workingday' column.

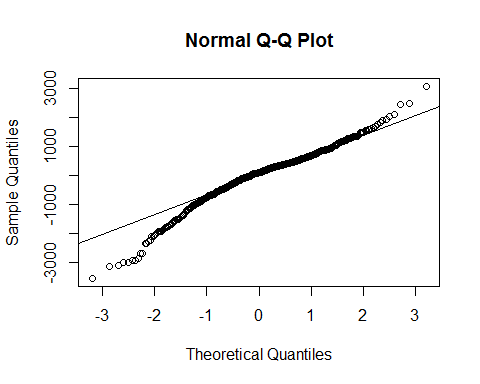
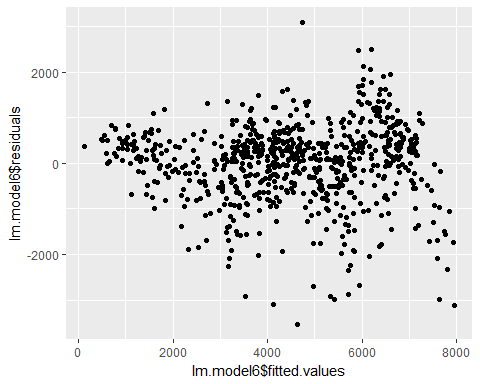
##   
## Call:  
## lm(formula = y ~ as.factor(holiday) + as.factor(workingday) +   
## temp + hum + windspeed + as.factor(season), data = day)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.03752 -0.22422 0.01253 0.27035 0.92219   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 7.97374 0.09037 88.236 < 2e-16 \*\*\*  
## as.factor(holiday)1 -0.17485 0.08258 -2.117 0.0346 \*   
## as.factor(workingday)1 0.04664 0.02970 1.570 0.1168   
## temp 1.92652 0.13145 14.656 < 2e-16 \*\*\*  
## hum -1.05058 0.09892 -10.620 < 2e-16 \*\*\*  
## windspeed -1.13760 0.18306 -6.214 8.71e-10 \*\*\*  
## as.factor(season)2 0.28006 0.04962 5.644 2.39e-08 \*\*\*  
## as.factor(season)3 0.09816 0.06535 1.502 0.1335   
## as.factor(season)4 0.47515 0.04228 11.238 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3604 on 721 degrees of freedom  
## Multiple R-squared: 0.577, Adjusted R-squared: 0.5724   
## F-statistic: 123 on 8 and 721 DF, p-value: < 2.2e-16



Eventhorugh transformation reduced the standard error values, the plot between fitted values and residuals have little curvature. Comparing the QQplots for both model1 by adding the season column and model3 for which we have transformed the response variable, we can see that model1 is fitting better than model3.

By considering the 'yr' and fitted the model.

##   
## Call:  
## lm(formula = cnt ~ as.factor(holiday) + temp + hum + windspeed +   
## as.factor(season) + as.factor(yr), data = day)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3536.0 -430.2 111.2 494.5 3088.8   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2273.53 214.61 10.594 < 2e-16 \*\*\*  
## as.factor(holiday)1 -646.14 189.53 -3.409 0.000688 \*\*\*  
## temp 5606.34 312.87 17.919 < 2e-16 \*\*\*  
## hum -2699.23 236.67 -11.405 < 2e-16 \*\*\*  
## windspeed -3527.27 434.51 -8.118 2.05e-15 \*\*\*  
## as.factor(season)2 1079.98 117.85 9.164 < 2e-16 \*\*\*  
## as.factor(season)3 730.80 155.36 4.704 3.06e-06 \*\*\*  
## as.factor(season)4 1532.34 100.30 15.278 < 2e-16 \*\*\*  
## as.factor(yr)1 2021.71 64.08 31.548 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 854.9 on 721 degrees of freedom  
## Multiple R-squared: 0.8062, Adjusted R-squared: 0.8041   
## F-statistic: 375 on 8 and 721 DF, p-value: < 2.2e-16



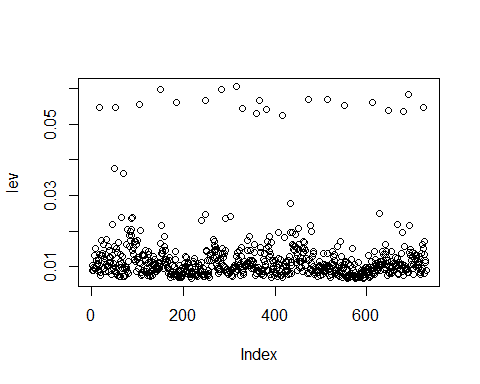
From the summary of above model, p-values are significant and 'S' is small and p-values are significant.

Finally, the model6 is best fit for predicting the values.

## model model1 model2 model3 model4 model5 model6  
## [1,] 1.068766 2.084264 2.216261 2.084264 2.084264 2.052362 2.078666

The mean values for VIF for all models is less than 10. So, all models are significant. The R-squared is high for model7 compared to other models and Standard Error is less compared to other models.

**Leverages**:



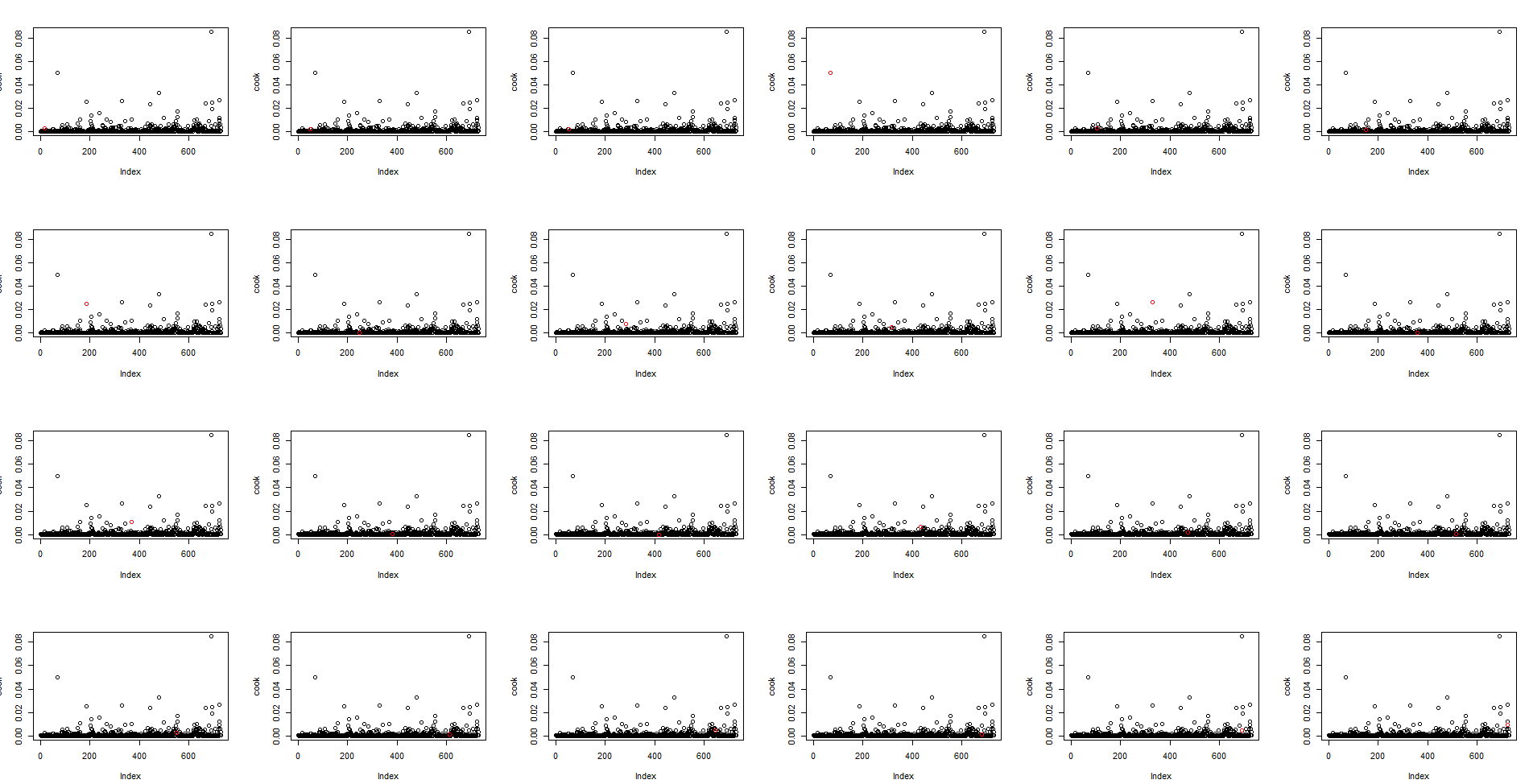
## [1] 17 50 52 69 105 150 185 248 283 315 328 360 367 381 416 433 472  
## [18] 514 551 612 647 682 692 725

## [1] 24

The maximum leverage value should be (2\*(k+1)/n) which is approximately 0.025. There are around 24 observations having higher leverage than all other points.

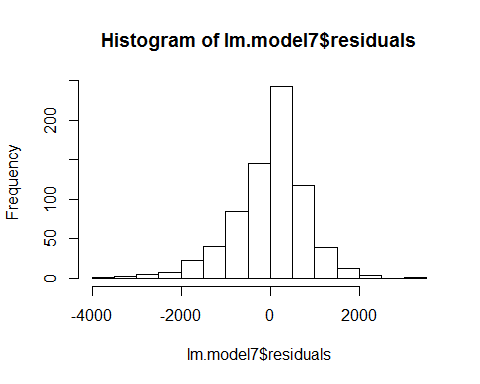
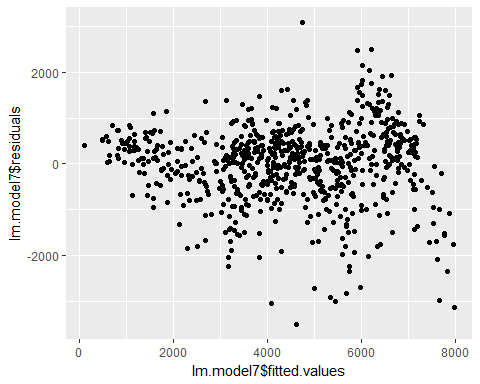
An influential point is one if removed from the data would significantly change the fit. An influential point may either be an outlier or have large leverage, or both, but it will tend to have at least one of those properties. Cook's distance is a commonly used influence measure that combines these two properties.

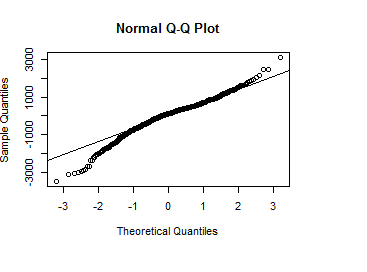
Plotting all the 24 observations for cook's distance.



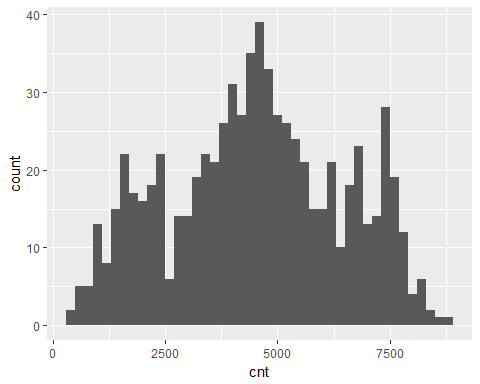
From the above plots, we can say that, the observation 69 and 692 are influential. So, removing that data and fitting the model again.

##   
## Call:  
## lm(formula = cnt ~ as.factor(holiday) + temp + hum + windspeed +   
## as.factor(season) + as.factor(yr), data = day2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3513.0 -446.9 118.0 484.0 3088.4   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2376.01 213.95 11.106 < 2e-16 \*\*\*  
## as.factor(holiday)1 -656.78 187.54 -3.502 0.00049 \*\*\*  
## temp 5652.53 310.12 18.227 < 2e-16 \*\*\*  
## hum -2849.26 237.50 -11.997 < 2e-16 \*\*\*  
## windspeed -3561.49 430.01 -8.282 5.89e-16 \*\*\*  
## as.factor(season)2 1057.93 116.85 9.053 < 2e-16 \*\*\*  
## as.factor(season)3 701.17 154.07 4.551 6.27e-06 \*\*\*  
## as.factor(season)4 1531.97 99.48 15.400 < 2e-16 \*\*\*  
## as.factor(yr)1 2013.07 63.57 31.667 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 845.9 on 719 degrees of freedom  
## Multiple R-squared: 0.8097, Adjusted R-squared: 0.8076   
## F-statistic: 382.5 on 8 and 719 DF, p-value: < 2.2e-16





From the residuals vs fitted values we see there is little curvature in the model. But the qqplot and histogram for residuals we see that the model is good fit for the data.



**Other Distributions tested on Data Set**:

From the histogram above, it seems that the number of total rented bikes follow a nearly normal distribution. Many studies have revealed that Poisson distribution can be used to analyze count data. The mean and variance of Poisson distribution are the same, and when the mean is getting larger, Poisson distribution approximates a normal distribution.Poisson regression is good for modeling count data. It assumes the response variable follows Poisson distribution and can be explained by linear combination of the explanatory variables.

## mean variance ratio  
## [1,] 4510.489 3730338 827.0363

We can see there is huge difference between the mean and variance, So, clearly it doesnot follow the standard poisson distribution. For the count data, if the variance is larger than mean, it means the data has over dispersion.

We first fit the standard poisson distribution.

##   
## Call:  
## glm(formula = cnt ~ as.factor(holiday) + as.factor(workingday) +   
## temp + hum + windspeed + as.factor(season), family = "poisson",   
## data = day)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -74.024 -16.819 -3.544 16.203 66.568   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 8.079250 0.003976 2032.19 <2e-16 \*\*\*  
## as.factor(holiday)1 -0.143246 0.003723 -38.47 <2e-16 \*\*\*  
## as.factor(workingday)1 0.020501 0.001235 16.60 <2e-16 \*\*\*  
## temp 1.466066 0.005504 266.34 <2e-16 \*\*\*  
## hum -0.829311 0.004240 -195.58 <2e-16 \*\*\*  
## windspeed -0.894635 0.007971 -112.23 <2e-16 \*\*\*  
## as.factor(season)2 0.305348 0.002264 134.87 <2e-16 \*\*\*  
## as.factor(season)3 0.175895 0.002821 62.36 <2e-16 \*\*\*  
## as.factor(season)4 0.451465 0.001982 227.84 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for poisson family taken to be 1)  
##   
## Null deviance: 660064 on 729 degrees of freedom  
## Residual deviance: 312590 on 721 degrees of freedom  
## AIC: 320002  
##   
## Number of Fisher Scoring iterations: 4

Even though p values are significant the Residual deviance and degree of freedom suggest that model doesn't fit the data well. The reason for that is probably because of over dispersion in the dataset.

To resolve the over-dispersion issue, we use 2 models, namely

1. Quassi-Poisson Regression
2. Negative-binomial

**Quasi-Poisson Regression**:

Quasi-Poisson is remedy for over dispersion in Poisson model. It estimates the scale parameter as well, and fixes the estimated standard error. It gives us more conservation results.

##   
## Call:  
## glm(formula = cnt ~ as.factor(holiday) + as.factor(workingday) +   
## temp + hum + windspeed + as.factor(season), family = quasipoisson(link = "log"),   
## data = day)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -74.024 -16.819 -3.544 16.203 66.568   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.07925 0.08170 98.886 < 2e-16 \*\*\*  
## as.factor(holiday)1 -0.14325 0.07652 -1.872 0.0616 .   
## as.factor(workingday)1 0.02050 0.02538 0.808 0.4196   
## temp 1.46607 0.11312 12.960 < 2e-16 \*\*\*  
## hum -0.82931 0.08714 -9.517 < 2e-16 \*\*\*  
## windspeed -0.89464 0.16382 -5.461 6.52e-08 \*\*\*  
## as.factor(season)2 0.30535 0.04653 6.563 1.01e-10 \*\*\*  
## as.factor(season)3 0.17590 0.05797 3.034 0.0025 \*\*   
## as.factor(season)4 0.45147 0.04072 11.087 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for quasipoisson family taken to be 422.3375)  
##   
## Null deviance: 660064 on 729 degrees of freedom  
## Residual deviance: 312590 on 721 degrees of freedom  
## AIC: NA  
##   
## Number of Fisher Scoring iterations: 4

Comparing the two tables above, we can see that the estimated standard errors in Quasi-Poisson model are larger than that in Poisson model. And the dispersion parameter given by Quasi-Poisson model is 422.3375, which means there is indeed over dispersion in the data set and Poisson model underestimated the standard errors.

**Negative-Binomial model**:

Negative Binomial model is also very useful for modeling count variable, especially for the data set with over dispersion. Compared to Poisson model, it contains an extra parameter theta, which is the parameter of multiplicative random effect.

##   
## Call:  
## glm.nb(formula = cnt ~ as.factor(holiday) + as.factor(workingday) +   
## temp + hum + windspeed + as.factor(season), data = day, init.theta = 8.700179902,   
## link = log)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -4.5879 -0.7985 -0.1391 0.6864 2.8207   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 8.01540 0.08511 94.172 < 2e-16 \*\*\*  
## as.factor(holiday)1 -0.16988 0.07779 -2.184 0.029 \*   
## as.factor(workingday)1 0.03562 0.02797 1.273 0.203   
## temp 1.90802 0.12380 15.413 < 2e-16 \*\*\*  
## hum -0.98294 0.09317 -10.550 < 2e-16 \*\*\*  
## windspeed -1.02101 0.17242 -5.922 3.18e-09 \*\*\*  
## as.factor(season)2 0.23518 0.04674 5.032 4.85e-07 \*\*\*  
## as.factor(season)3 0.06220 0.06154 1.011 0.312   
## as.factor(season)4 0.43248 0.03982 10.860 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for Negative Binomial(8.7002) family taken to be 1)  
##   
## Null deviance: 1566.80 on 729 degrees of freedom  
## Residual deviance: 744.42 on 721 degrees of freedom  
## AIC: 12647  
##   
## Number of Fisher Scoring iterations: 1  
##   
##   
## Theta: 8.700   
## Std. Err.: 0.448   
##   
## 2 x log-likelihood: -12626.796

## [1] 2.693456e-195

The Chi-Square test p-value for the negative binomial is less than 0.05. So, negative binomial model is also fitting the model correctly.

**Conclusion**:

The count of total rental bikes depends on factors like the temperature, the windspeed, the type of weather, holiday, working day and humidity. We haven't considered the weather type for linear regression models as the weather type is already covered using the parameters temp, Humidity and Windspeed. For linear regression models, box-cox transformation has been applied for the best fit of the model. Eventhough, the R-squared and p-values are significant the plot for residuals and fitted values has little curvature.

Since the response is a count variable with high values we have chosen the Negative Binomial method. Based on the dispersion parameter, the maximum likelihood, the residual deviance, p values, standard error, the AIC values, the theta values, and the chi square test are significant but the log likehood value is not significant. So, we conclude that this model suits the data better than the Poisson model but not better than linear regression model.

**Contribution**:

Contribution calculated the total hours spent and work done by each member in the team.

Divya: 65%

Harsha: 35%