ISM 6137 – Statistical Data Mining
Data Analysis Report
on
Capital Bikeshare System

Team:

Sanketh Bhagavanthi

Sai Kumar Nanjala

Raju Mohan Reddy Bakaram

Divya Rekha Surnilla

Table of Contents

Problem Statement

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 particular position and return back at another position. 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.

Unlike car sharing, the bike sharing will be majorly impacted by the weather and seasonal conditions. So, we considered taking all the weather factors on a particular day into account and using these variables we try to build a model that can forecast, based on the past data, the count of bikes that will be rented on a particular day.

Dataset Description

Data Source: Kaggle

There is a total of 731 observations. Each observation pertains to one day. The weather and seasonal conditions of each day have been recorded for a span of two years (2011 and 2012). The final model developed can be used by the capital bike share company to predict the number of bikes that might be rented on a particular day based on the environmental and weather conditions in Washington DC.

The dataset consists of the following variables:

S.No	Variable	Description
1	instant	Number of the observation
2	dteday	Date on that day in mm/dd/yyyy format.
3	season	1 = spring; 2 = summer; 3 = fall; 4 = winter
4	yr	0 indicates 2011 1 indicates 2012

5	mnth	This field has values from 1 to 12 indicating the month.
6	holiday	If holiday = 1 Not a holiday = 0
7	weekday	What day of the week it is. 0=Sunday; 1 = Monday; 2= Tuesday; 3= Wednesday; 4= Thursday; 5= Friday; 6= Saturday.
8	workingday	If working day, 1 Not a working day = 0
9	weathersit	Has values 1 to 3. 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
10	temp	Normalized values of temperature on that day.
11	atemp	Normalized values of feels like temperature on that day.
12	hum	Normalized value Humidity on that day.
13	windspeed	Wind Speed on that day.
14	casual	Number of unregistered user rentals initiated.
15	registered	Number of registered user rentals initiated.
16	cnt	Total number of rentals.

Data Preprocessing

We have done a correlation test between all the variables. As the correlation co-efficient between “temp” and “atemp” is almost 1, we chose to eliminate the variable temp from our analysis. We chose “atemp” because the users would consider what the temperature feels like rather than the actual temperature.

The values of temp, hum(humidity), windspeed have been normalized to have a common scale. This field had about less than 4% of missing values. We took the mean of the windspeed for that week and imputed the values.

The variable count is the sum of casual and registered users. As the correlation between these variables is high and because we are not considering the attributes casual and registered for our forecasting model, we have ignored the attributes ‘casual and ‘registered’ from our analysis.

The initial dataset did not have the column describing the month. We added the column month and populated the value by referring to the ‘mm’ information in the ‘*dteday’* value.

Hypothesis

The core dependent variable in our case would be: ‘count’, as we are predicting the count of bikes rented based on other dependent variables such as month, holiday or not, which day of the week is it, weather situation, temperature, humidity and wind speed.

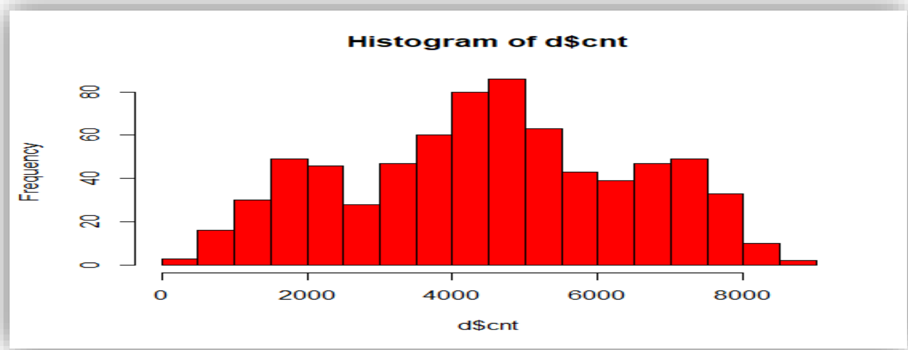
1. Temperature has a positive correlation with demand for bikes.
2. The number of bike rentals would be high in the month of June as it is a summer month and the usage of bikes would increase with temperature.
3. The total number of bike rentals on a holiday will be more than on a working day.

Descriptive Analysis

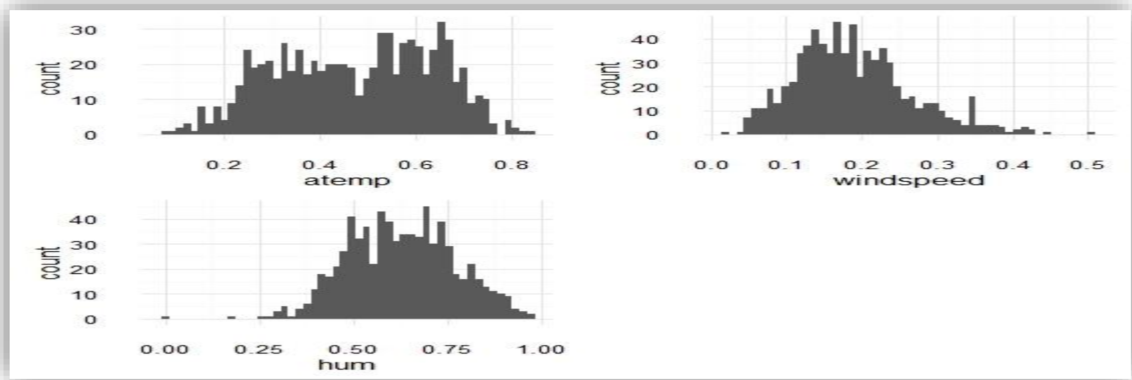
Univariate Analysis

Histogram for Target Variable:

The dependent variable, count has a normal spread.



Histogram for Predictor Variables:

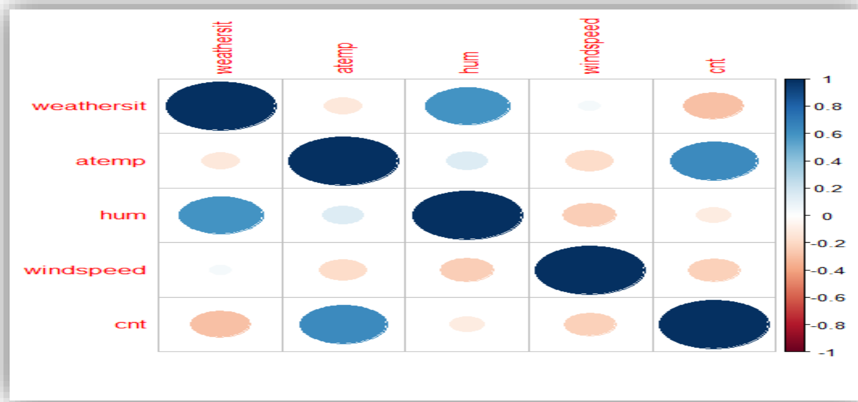


Bivariate Analysis:

Correlation Test:

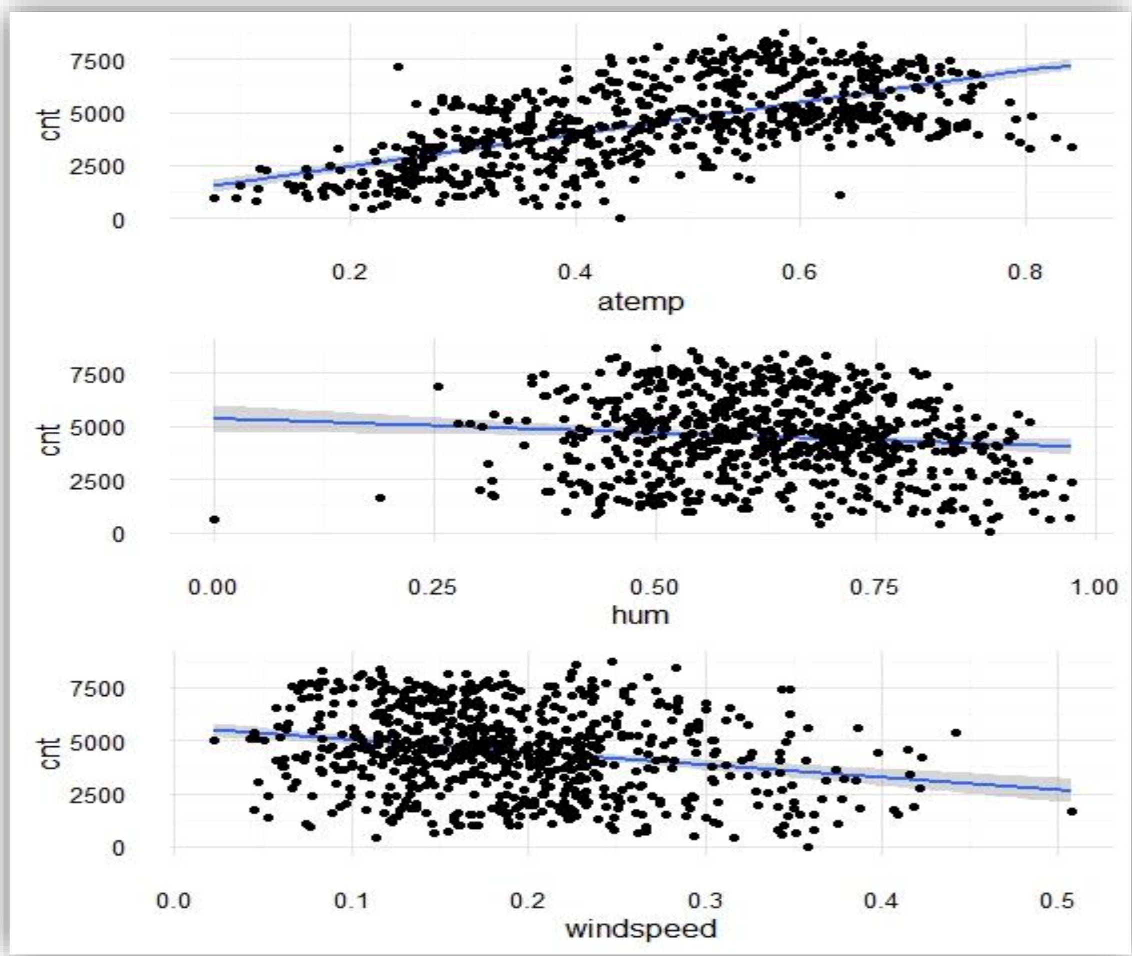
Correlation test is used to investigate the dependencies between multiple variables and Correlation matrix is a table that gives the correlation coefficients between each variable.

	weathersit	atemp	hum	windspeed	cnt
weathersit	1.000	-0.122	0.591	0.040	-0.297
atemp	-0.122	1.000	0.140	-0.184	0.631
hum	0.591	0.140	1.000	-0.248	-0.101
windspeed	0.040	-0.184	-0.248	1.000	-0.235
cnt	-0.297	0.631	-0.101	-0.235	1.000



Ggplot:

From the ggplot between the dependent and independent variables, we see that the relationship is almost linear.



Models

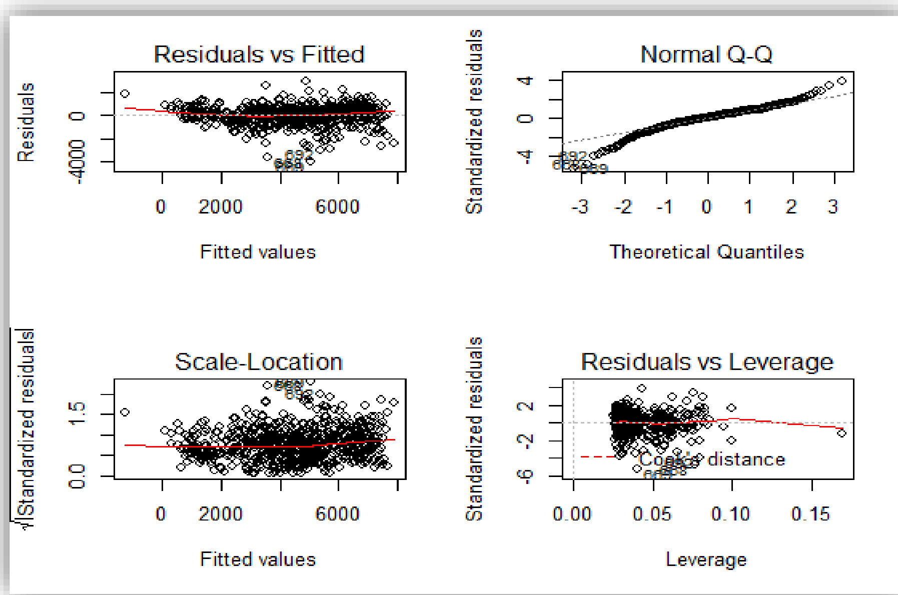
Kitchen Sink Model

```
lm(formula = d$cnt ~ d$season + d$yr + d$mnth + d$holiday + d$workingday +
    d$weekday + d$weathersit + d$atemp + d$hum + d$windspeed)

Residuals:
    Min      1Q  Median      3Q     Max
-3967.2  -347.0    69.3   449.1  2929.7

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1419.97     238.10   5.964 3.90e-09 ***
d$season2       876.41     179.85   4.873 1.36e-06 ***
d$season3       848.80     213.46   3.976 7.72e-05 ***
d$season4      1578.83     181.40   8.704  < 2e-16 ***
d$yr1          2027.87      58.21  34.834  < 2e-16 ***
d$mnth2         137.39     144.07   0.954  0.34058
d$mnth3         577.61     164.92   3.502  0.00049 ***
d$mnth4         500.97     246.92   2.029  0.04284 *
d$mnth5         839.63     263.32   3.189  0.00149 **
d$mnth6         660.67     273.92   2.412  0.01612 *
d$mnth7         177.75     306.15   0.581  0.56168
d$mnth8         607.53     293.18   2.072  0.03861 *
d$mnth9        1112.07     260.39   4.271 2.22e-05 ***
d$mnth10        566.62     241.00   2.351  0.01899 *
d$mnth11       -104.09     231.21  -0.450  0.65272
d$mnth12        -84.48     182.61  -0.463  0.64377
d$holiday1     -107.30     206.77  -0.519  0.60397
d$workingday1   453.87     107.36   4.227 2.68e-05 ***
d$weekday1     -244.15     109.37  -2.232  0.02590 *
d$weekday2     -138.74     107.51  -1.290  0.19731
d$weekday3      -61.33     107.96  -0.568  0.57014
d$weekday4      -59.73     107.33  -0.557  0.57804
d$weekday5          NA         NA      NA       NA
d$weekday6      444.32     106.79   4.161 3.56e-05 ***
d$weathersit2  -467.10      77.23  -6.048 2.38e-09 ***
d$weathersit3 -1955.77     197.44  -9.906  < 2e-16 ***
d$atemp        4639.41     431.50  10.752  < 2e-16 ***
d$hum         -1516.84     292.96  -5.178 2.94e-07 ***
d$windspeed   -2647.09     406.38  -6.514 1.40e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 771 on 703 degrees of freedom
Multiple R-squared:  0.8475,	Adjusted R-squared:  0.8416
F-statistic: 144.7 on 27 and 703 DF,  p-value: < 2.2e-16
```



Kitchen sink model seems like a good candidate for linear regression from the above plots.

```
> alias(m1)
Model :
d$cnt ~ d$season + d$yr + d$mnth + d$holiday + d$workingday +
    d$weekday + d$weathersit + d$atemp + d$hum + d$windspeed

Complete :
           (Intercept) d$season2 d$season3 d$season4 d$yr1 d$mnth2 d$mnth3 d$mnth4 d$mnth5 d$mnth6 d$mnth7 d$mnth8
d$weekday5  0           0         0         0         0     0       0       0       0       0       0       0
           d$mnth9 d$mnth10 d$mnth11 d$mnth12 d$holiday1 d$workingday1 d$weekday1 d$weekday2 d$weekday3 d$weekday4
d$weekday5  0       0        0        0        1          1            -1         -1         -1         -1
           d$weekday6 d$weathersit2 d$weathersit3 d$atemp d$hum d$windspeed
d$weekday5  0          0             0             0       0     0
```

But, as seen from the alias values, weekday is perfectly collinear with working day and holiday. Hence remove working day and weekday variables in the next model and keep holiday because it gives us the information about working day too. Weekday is not very useful in our analysis because we only need to analyze the trend on holidays and working days.

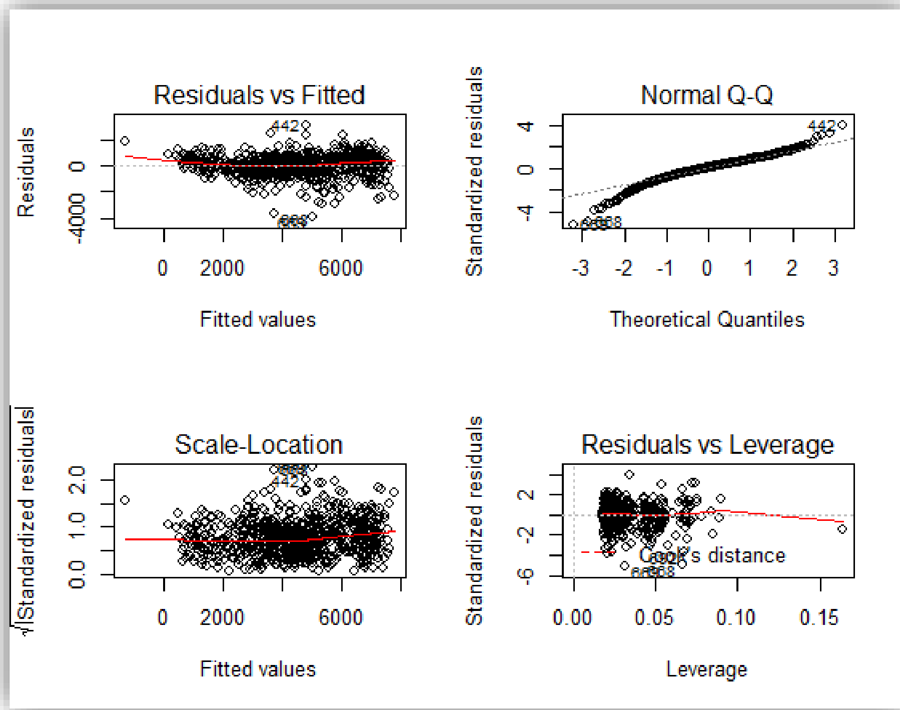
MODEL 1

```
Call:
lm(formula = d$cnt ~ d$season + d$yr + d$mnth + d$holiday + d$weathersit +
    d$atemp + d$hum + d$windspeed)

Residuals:
    Min      1Q  Median      3Q     Max
-3939.3  -369.7    77.8   468.8  3041.9

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    1777.93     229.90   7.734 3.59e-14 ***
d$season2       885.39     182.47   4.852 1.50e-06 ***
d$season3       866.26     216.53   4.001 6.98e-05 ***
d$season4      1566.42     184.02   8.512  < 2e-16 ***
d$yr1          2022.56      59.07  34.240  < 2e-16 ***
d$mnth2         142.85     146.20   0.977 0.328880
d$mnth3         575.92     167.29   3.443 0.000610 ***
d$mnth4         478.17     250.19   1.911 0.056374 .
d$mnth5         816.06     266.80   3.059 0.002307 **
d$mnth6         626.69     277.21   2.261 0.024080 *
d$mnth7         121.62     309.48   0.393 0.694450
d$mnth8         569.12     296.75   1.918 0.055537 .
d$mnth9        1090.26     263.47   4.138 3.92e-05 ***
d$mnth10        567.35     244.21   2.323 0.020447 *
d$mnth11        -81.35     234.41  -0.347 0.728674
d$mnth12        -68.71     185.14  -0.371 0.710645
d$holiday1     -612.95     175.35  -3.496 0.000503 ***
d$weathersit2  -430.72      77.91  -5.528 4.55e-08 ***
d$weathersit3 -1873.44     199.03  -9.413  < 2e-16 ***
d$atemp        4747.14     435.31  10.905  < 2e-16 ***
d$hum         -1655.88     295.20  -5.609 2.91e-08 ***
d$windspeed   -2688.05     412.07  -6.523 1.31e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 782.6 on 709 degrees of freedom
Multiple R-squared:  0.8415,    Adjusted R-squared:  0.8368
F-statistic: 179.3 on 21 and 709 DF,  p-value: < 2.2e-16
```



From the above plots it looks like the model m2 is linear and passes homoscedasticity, normality assumption tests.

```
> vif(m2)
                    GVIF Df GVIF^(1/(2*Df))
d$season      168.658132  3        2.350541
d$yr            1.041278  1        1.020430
d$mnth        353.228482 11        1.305632
d$holiday       1.024143  1        1.012000
d$weathersit    1.842316  2        1.165041
d$atemp         5.998721  1        2.449229
d$hum           2.107330  1        1.451665
d$windspeed     1.215657  1        1.102568
```

Then we ran a test to check multi collinearity. We can see that there is high multi collinearity between season and month. After careful analysis we excluded the variable season from our next model because month is a part of season.

MODEL 2

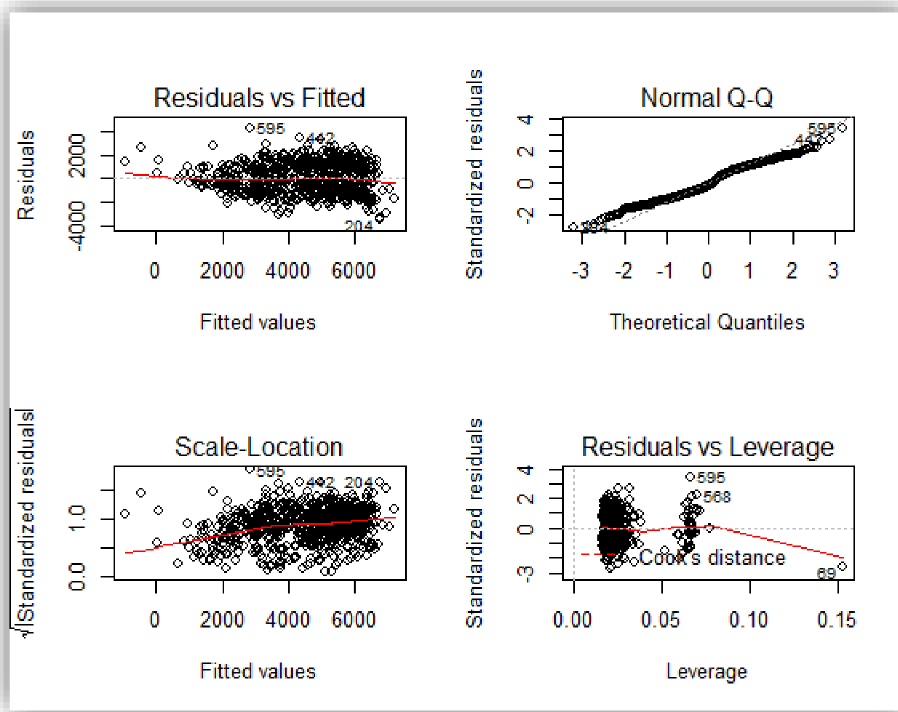
```
> m4 <- lm(d$cnt ~   d$mnth +d$holiday +d$weathersit + d$atemp + d$hum + d$windspeed )
> summary(m4)

Call:
lm(formula = d$cnt ~ d$mnth + d$holiday + d$weathersit + d$atemp +
    d$hum + d$windspeed)

Residuals:
    Min      1Q  Median      3Q     Max
-3495.9 -1014.7  -177.1  1084.0  4282.1

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    3201.31     374.84   8.540  < 2e-16 ***
d$mnth2          16.52     242.39   0.068 0.945695
d$mnth3         580.61     255.05   2.276 0.023113 *
d$mnth4         905.92     281.77   3.215 0.001363 **
d$mnth5        1093.68     322.59   3.390 0.000737 ***
d$mnth6         603.68     365.75   1.651 0.099277 .
d$mnth7         -26.53     398.94  -0.067 0.946992
d$mnth8         568.03     368.23   1.543 0.123371
d$mnth9        1528.23     331.11   4.616 4.65e-06 ***
d$mnth10       1736.16     283.51   6.124 1.51e-09 ***
d$mnth11       1268.46     251.38   5.046 5.73e-07 ***
d$mnth12        827.14     241.71   3.422 0.000657 ***
d$holiday1     -730.57     289.89  -2.520 0.011947 *
d$weathersit2  -204.68     128.81  -1.589 0.112494
d$weathersit3 -1849.19     329.82  -5.607 2.95e-08 ***
d$atemp        7004.35     707.35   9.902  < 2e-16 ***
d$hum         -3132.01     483.36  -6.480 1.71e-10 ***
d$windspeed   -3516.00     680.18  -5.169 3.06e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1298 on 713 degrees of freedom
Multiple R-squared:  0.5613,    Adjusted R-squared:  0.5509
F-statistic: 53.67 on 17 and 713 DF,  p-value: < 2.2e-16
```



From the above plots, we can say that the model is linear, homoscedastic and follows normality.

```
> vif(m4)
                  GVIF Df GVIF^(1/(2*Df))
d$mnth        6.781114 11        1.090904
d$holiday     1.016926  1        1.008427
d$weathersit  1.822924  2        1.161963
d$atemp       5.754740  1        2.398904
d$hum         2.052713  1        1.432729
d$windspeed   1.203433  1        1.097011
```

When tested for multi collinearity, the VIF values indicated that there is no multi collinearity.

```
> AIC(m1,m2,m4)
   df      AIC
m1 29 11822.82
m2 23 11838.83
m4 19 12575.03
> BIC(m1,m2,m4)
   df      BIC
m1 29 11956.06
m2 23 11944.50
m4 19 12662.33
```

After comparison of the Adjusted R square, AIC and BIC values, though the values are better for the first two models, we reject those because they did not pass multi collinearity test.

Therefore, our final model would be m4.

```
> shapiro.test(m4$residuals[sample(731,100)])

        Shapiro-Wilk normality test

data:  m4$residuals[sample(731, 100)]
W = 0.98478, p-value = 0.3059
```

From the Shapiro test for normality, the p value is greater then 0.05 satisfying the cut off (> 0.05) value and hence this model satisfies the normality test.

Forecasting:

We divided the data into 75% train data and 25% test data to run our forecasting model. The observed RMSE is 1295.147

```
> RMSE(preds,test$cnt)
[1] 1295.147
```

Insights

From our analysis we found that *month*, *holiday*, *weather situation*, *temperature*, *humidity* and *windspeed* are statistically significant for forecasting the bike rentals.

Some of the insights we got from the analysis are:

1. Temperature has a positive correlation with Count. With the increase in temperature, the demand for bikes also increases.
2. On days when the humidity is high, the number of bikes rented reduces. Same is the case with windspeed.
3. To our surprise, the demand for bikes is less on holidays than on working days.
4. The demand for bikes is more in September and October months.