Overview: The data contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

The data set has 3 files inside, day.csv, hour.csv and Readme.txt.

Aim: Based on the file day.csv propose a linear regression model for the response variable (count of total rental bike daily).

EXPLORATORY DATA ANALYSIS:

1) Response variable: Count of total daily rental bikes

a) The response variable cnt is quantitative

b) Summary of cnt:
Min = 22
1st Quartile = 3152
2nd Quartile (median) = 4548
Mean = 4504
3rd Quartile = 5956
Max = 8714

```
> summary(cnt)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
     22    3152    4548    4504    5956    8714
```

c) Nomarlity:
- From the histogram of cnt in Fig 1.1, we can conclude that the distribution is bell-shaped and symmetric.
- From the boxplot of cnt in Fig 1.2, we can conclude that the distribution has no outliers and is not skewed.
Thus, from the observations we can conclude that the distribution is normal.

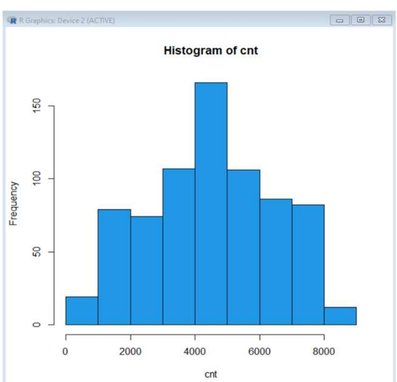


FIG 1.1

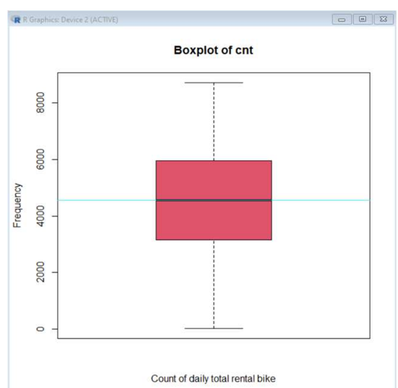


FIG 1.2

Hence, since the response is quantitative and its distribution is symmetric, the response is suitable to fit a linear regression model.

2) - Regressor: temp
Correlation between cnt and temp = 0.627494
Association between cnt and temp is **positive** and **weak**
From scatter plot: temp might be linear and have a constant variance
From histogram: multimodal

- Regressor: hum

Correlation between cnt and hum = -0.1006586
Association between cnt and hum is **negative** and **weak**
From scatter plot: hum might be linear but its variance might not be constant
From histogram: left skewed

- Regressor: windspeed
  Correlation between cnt and windspeed = -0.234545
  Association between cnt and windspeed is **negative** and **weak**
  From scatter plot: windspeed might be linear and have a constant variance
  From histogram: right skewed

- Regressor: season
  From the boxplot in Fig 2.1 we can observe that the IQR for summer and winter season is almost same thus, the spread of data for both categories is almost same but, the median of winter vs cnt is higher than the median of summer vs cnt.

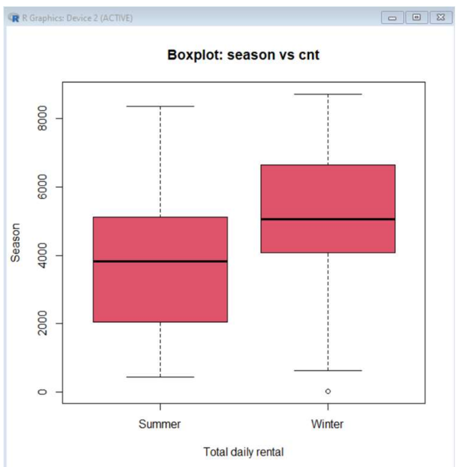


FIG 2.1

- Regressor: workingday
  From the boxplot in Fig 2.2 we can observe that the IQR for a non-working day is greater than the IQR for a working day thus, the spread of data for both categories is different. Moreover, the median of a working day is slightly more, almost same, than the median of a non-working day.

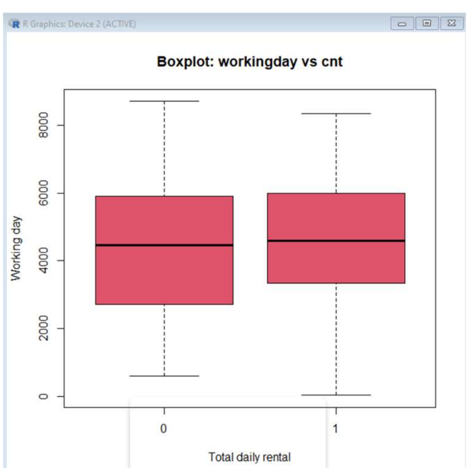


FIG 2.2

- Regressor: weathersit
  From the boxplot in Fig 2.3 we can observe that the IQR for good and bad weather is almost same

thus, the spread of data is almost same but, the median of good weather is higher than the median of bad weather.

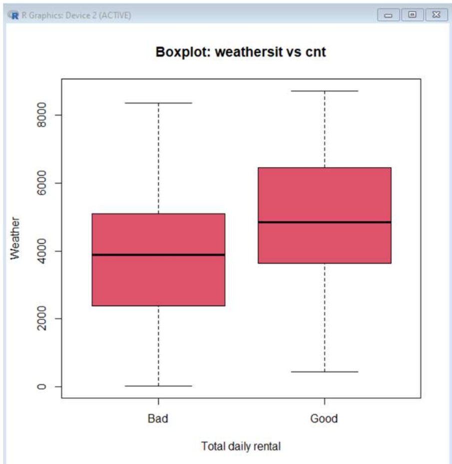


FIG 2.3

MODEL:

3) Proposed regressors for model M1 are:
season, workingday, weathersit, temp, hum and windspeed

In R created a model M1 with the above mentioned regressors and got the following model summary:

```
> M1= lm(cnt~seasons+workday+weather+temp+hum+windspeed, data=data)
> summary(M1) # adj r sqr = 0.5157, workday is insignificant

Call:
lm(formula = cnt ~ seasons + workday + weather + temp + hum +
    windspeed, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4208.2 -1031.9  -142.7  1100.1  3750.7

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     3409.6      432.3   7.886 1.15e-14 ***
seasonsWinter    466.7      116.2   4.015 6.55e-05 ***
workdayYes       150.1      112.3   1.337    0.182
weatherGood      296.6      138.0   2.149    0.032 *
temp            6004.0      317.3  18.924  < 2e-16 ***
hum            -2594.9      481.7  -5.387 9.68e-08 ***
windspeed      -4065.3      718.0  -5.662 2.16e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1406 on 724 degrees of freedom
Multiple R-squared:  0.4775,	Adjusted R-squared:  0.4732
F-statistic: 110.3 on 6 and 724 DF,  p-value: < 2.2e-16
```

For the fitted model:
Let Y^ denoted the predicted value of cnt
B0 denote the intercept
B1 denote the coefficient and X1 denote the value for the regressor seasons
B2 denote the coefficient and X2 denote the value for the regressor workday
B3 denote the coefficient and X3 denote the value for the regressor weather
B4 denote the coefficient and X4 denote the value for the regressor temp
B5 denote the coefficient and X5 denote the value for the regressor hum
B6 denote the coefficient and X6 denote the value for the regressor windspeed
I denote an indicator variable

The fitted model equation:

Y^ = B0 + B1.I(X1=Winter) + B2.I(X2=Yes) + B3.I(X3=Good) + B4.X4 + B5.X5 + B6.X6
Y^ = 3409.6 + 466.7(1) + 150.1(1) + 296.6(1) + 6004.0(X4) - 2594.9(X5) - 4065.3(X6)

4) Checking if M1 is adequate:

a) Calculated the standard residuals of the model and stored it in SR
b) Using R we can infer that:
Outliers: at 69
Influential points: none
c) Checking whether the assumptions (normality, constant variance and linearity) hold:

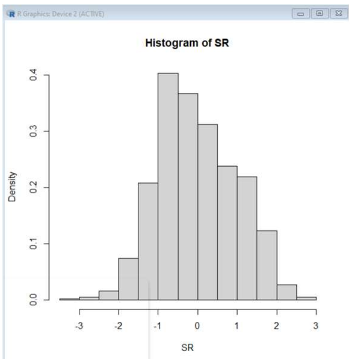


FIG 4.1

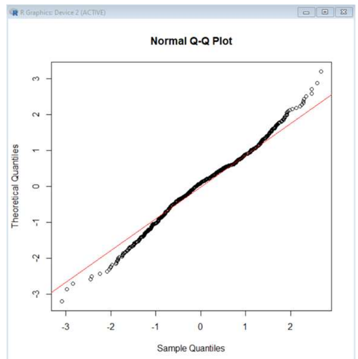


FIG 4.2

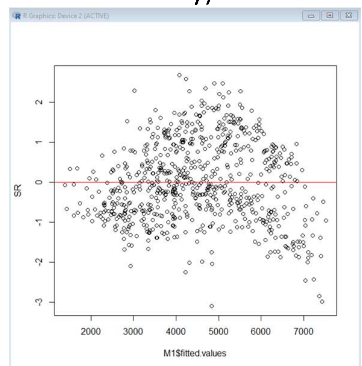


FIG 4.3

From histogram of SR (Fig 4.1): Slightly left skewed, normality violated
From QQ plot of SR (Fig 4.2): Both tails are slightly shorter than normal, normality violated
From scatterplot of predicted Y^ vs SR: Slightly funnel shaped, constant variance violated

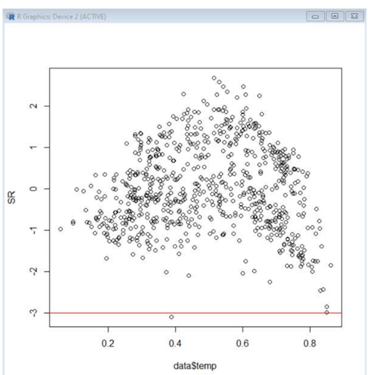


FIG 4.4

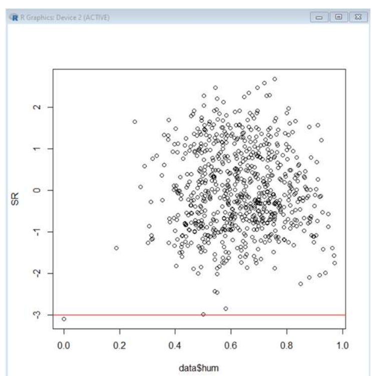


FIG 4.5

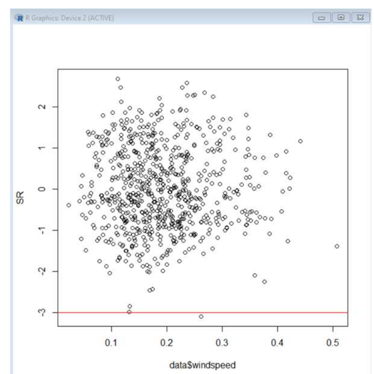


FIG 4.6

From scatterplots of regressors against SR (FIG 4.4, 4.5, 4.6):
For temp linearity might be violated (FIG 4.4)
hum might be linear (FIG 4.5)
windspeed might be linear (FIG 4.6)

d) Adjusted R^2 = 0.4732
e) F-statistic = 110.3 on a null distribution F(6,724)
f) p-value < 0.05; f-test is significant
g) The model **M1 is not adequate** because the **assumptions are violated** and the **adjusted R^2 suggests that the model is not good enough.**

5) From model M1 we can see that:
seasons- significant as p-value < 0.05
workday- insignificant as p-value> 0.05
weather- significant as p-value<0.05
temp- significant as p-value<0.05

hum- significant as p-value<0.05
windspeed- significant as p-value<0.05

Proposal: Although the regressor workday is insignificant I did not choose to remove it from the model as in the following model M2 I decided to include the interaction regressors in the model and, since there exists significant interaction variables involving workday, I cannot remove workday from the model.

6) Series of linear models developed to reach the final linear model is as follows:

a) Model: M2
Difference from M1: Added interaction terms

```
> summary(M2)

Call:
lm(formula = cnt ~ seasons + weather + hum + temp + windspeed +
    workday + seasons * hum + seasons * temp + seasons * windspeed +
    weather * windspeed + weather * temp + weather * hum + hum *
    workday + hum * temp + hum * windspeed + temp * windspeed +
    temp * workday + windspeed * workday + seasons * weather +
    seasons * workday + weather * workday, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4469.2  -927.6  -124.8   975.6  3437.4

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                3369.35    1525.18   2.209  0.02748 *
seasonsWinter              4258.66     929.30   4.583 5.42e-06 ***
weatherGood               -2257.23     825.79  -2.733  0.00642 **
hum                       -4096.69    1925.13  -2.128  0.03368 *
temp                      10922.21    2551.10   4.281 2.11e-05 ***
windspeed                 -8436.87    4298.09  -1.963  0.05004 .
workdayYes                  274.75     871.55   0.315  0.75267
seasonsWinter:hum         -1277.98    1029.52  -1.241  0.21489
seasonsWinter:temp        -5956.04     622.80  -9.563  < 2e-16 ***
seasonsWinter:windspeed    -571.36    1505.09  -0.380  0.70434
weatherGood:windspeed      5731.18    1782.25   3.216  0.00136 **
weatherGood:temp           -651.93     797.44  -0.818  0.41390
weatherGood:hum            2111.30     985.24   2.143  0.03246 *
hum:workdayYes               67.11     991.05   0.068  0.94603
hum:temp                   -623.50    2986.69  -0.209  0.83470
hum:windspeed              3304.15    5446.49   0.607  0.54427
temp:windspeed            -1403.27    4424.17  -0.317  0.75120
temp:workdayYes           -1344.44     634.38  -2.119  0.03441 *
windspeed:workdayYes        282.04    1448.24   0.195  0.84564
seasonsWinter:weatherGood   -35.89     281.70  -0.127  0.89866
seasonsWinter:workdayYes    227.77     232.17   0.981  0.32690
weatherGood:workdayYes      492.13     288.31   1.707  0.08827 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1306 on 709 degrees of freedom
Multiple R-squared:  0.5585,    Adjusted R-squared:  0.5454
F-statistic:  42.7 on 21 and 709 DF,  p-value: < 2.2e-16
```

Remarks:
- The F-test is significant and Adjusted R^2 increased.
- Interation terms that are insignificant: seasonsWinter*windspeed, weatherGood*temp, seasonsWinter*hum, hum*workdayYes,hum*temp,hum*windspeed, temp*windspeed, workdayYes*windspeed, seasonsWinter*weatherGood and seasonsWinter*workdayYes

b) Model: M3
Difference from M2: Removed insignificant interaction terms

```
> summary(M3)

Call:
lm(formula = cnt ~ seasons + weather + hum + temp + windspeed +
    workday + seasons * temp + weather * windspeed + weather *
    hum + temp * workday + weather * workday, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4945.4  -953.3  -162.6   980.4  3334.1

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)              3596.2      604.0   5.954 4.10e-09 ***
seasonsWinter            3489.4      311.2  11.214  < 2e-16 ***
weatherGood             -2275.2      721.7  -3.153 0.001685 **
hum                     -4116.0      655.3  -6.281 5.81e-10 ***
temp                     9763.9      565.4  17.270  < 2e-16 ***
windspeed               -6741.1     1034.7  -6.515 1.37e-10 ***
workdayYes                357.2      315.9   1.131 0.258429
seasonsWinter:temp      -6005.6      584.2 -10.281  < 2e-16 ***
weatherGood:windspeed    4930.2     1354.8   3.639 0.000293 ***
weatherGood:hum          1875.9      889.9   2.108 0.035381 *
temp:workdayYes         -1076.2      564.3  -1.907 0.056917 .
weatherGood:workdayYes    481.3      220.4   2.184 0.029290 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1302 on 719 degrees of freedom
Multiple R-squared:  0.5553,	Adjusted R-squared:  0.5485
F-statistic: 81.64 on 11 and 719 DF,  p-value: < 2.2e-16
```

Remarks:
- F-test is significant and Adjusted R^2 increased
- hum and windspeed are insignificant but since they are involved in interaction terms, did not remove them
- temp*workingday is insignificant
- Modifying M3 itself and removing the insignificant term
- Modified M3 contains another insignificant term seasonsWinter*hum
- Thus, again modified M3 and removed the insignificant term

Checking whether M3 is adequate:
- Calculated standard residuals and stored in SR3
- Outliers: at 69
- Influential point: none
- From
hist(SR3, prob=TRUE): the plot is slighlt left skewed

qqnorm(SR3, datax = TRUE): normality violated, tails are slightly shorter than normal
qqline(SR3, datax = TRUE, col="red")

plot(M3$fitted.values, SR3) : slightly funnel shaped, equal variance might be violated
abline(h=0, col="red")

Remarks:
- F-test is significant and Adjusted R^2 = 0.5469
- No insignificant terms except workday which is involved in significant interaction terms
- One outlier at 69 which is non-influential
- Model M3 is still not adequate as the assumptions of normality and constant variance are slightly violated plus the adjusted R^2 is also not good enough

c) Model: M4
Changes: Transformations made to the regressors and response to rectify the violations
Transformations:
n_hum = hum^3 #for higher powers the scatterplot might be funnel shaped
n_wind = windspeed^2
n_cnt = log(cnt)

n_temp = log(temp)

```
> summary(M4)

Call:
lm(formula = n_cnt ~ seasons + weather + workday + n_wind + n_hum +
    n_temp + weather * n_wind + seasons * n_temp + weather *
    n_hum + weather * workday, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-4.2725 -0.2063  0.0267  0.2439  1.0431

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            9.63402    0.08795 109.538  < 2e-16 ***
seasonsWinter         -0.37163    0.06148  -6.045 2.40e-09 ***
weatherGood           -0.36049    0.08801  -4.096 4.68e-05 ***
workdayYes            -0.01582    0.05051  -0.313  0.75427
n_wind               -16.91005    1.81065  -9.339  < 2e-16 ***
n_hum                 -1.16052    0.12254  -9.470  < 2e-16 ***
n_temp                 1.06689    0.04440  24.032  < 2e-16 ***
weatherGood:n_wind    14.24731    2.21882   6.421 2.46e-10 ***
seasonsWinter:n_temp  -0.68321    0.07314  -9.341  < 2e-16 ***
weatherGood:n_hum      0.56641    0.20458   2.769  0.00577 **
weatherGood:workdayYes 0.10338    0.06227   1.660  0.09734 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.37 on 720 degrees of freedom
Multiple R-squared:  0.6031,	Adjusted R-squared:  0.5976
F-statistic: 109.4 on 10 and 720 DF,  p-value: < 2.2e-16
```

Checking whether the model is adequate:
- Calculated standard residuals and stored in SR4
- Outliers: at 2, 27, 69, 668
- Influential point: none
- From
  hist(SR4, prob=TRUE): highly left skewed, might be due to outliers

qqnorm(SR4, datax = TRUE): left tail is longer & right tail is almost normal, normality violated,
qqline(SR4, datax = TRUE, col="red")

plot(M4$fitted.values, SR4): normality violated
abline(h=0, col="red")

plot(n_temp, SR4): linear, constant variance, normality violated
abline(h=3, col="red")
abline(h=(-3), col="red")

plot(n_hum, SR4): linear, constant variance, normality violated
abline(h=3, col="red")
abline(h=(-3), col="red")

plot(n_wind, SR4): linear, constant variance, normality violated
abline(h=3, col="red")
abline(h=(-3), col="red")

Remarks:
- F-test is significant and Adjusted R^2 increased to 0.5976
- The transformations elimininated the moost of the violations
- Still the model is not adequate
- Suspected reason is due to outliers

d) Model: M5
   Changed: Removed Outliers

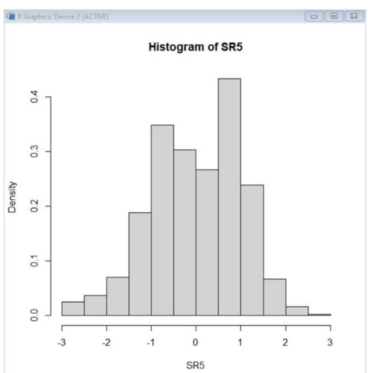
```
> summary(M5)

Call:
lm(formula = n_cnt ~ n_season + n_weathersit + n_workingday +
    n_wind + n_hum + n_temp + n_weathersit * n_wind + n_season *
    n_temp + n_season * n_hum + n_weathersit * n_hum + n_weathersit *
    n_workingday, data = new_data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09090 -0.21015  0.01707  0.24189  0.79580

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      9.869180   0.094244 104.719  < 2e-16 ***
n_seasonWinter                  -0.382474   0.067676  -5.652 2.30e-08 ***
n_weathersitGood                -0.561435   0.102140  -5.497 5.39e-08 ***
n_workingdayYes                  0.001238   0.043681   0.028 0.977391
n_wind                          -2.143430   0.256177  -8.367 3.10e-16 ***
n_hum                           -1.177816   0.132184  -8.910  < 2e-16 ***
n_temp                           1.052943   0.038497  27.352  < 2e-16 ***
n_weathersitGood:n_wind          1.739532   0.333714   5.213 2.44e-07 ***
n_seasonWinter:n_temp           -0.675396   0.062916 -10.735  < 2e-16 ***
n_seasonWinter:n_hum             0.063278   0.135459   0.467 0.640545
n_weathersitGood:n_hum           0.601752   0.181207   3.321 0.000943 ***
n_weathersitGood:n_workingdayYes 0.091722   0.053722   1.707 0.088193 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.3175 on 715 degrees of freedom
Multiple R-squared:  0.6575,    Adjusted R-squared:  0.6523
F-statistic: 124.8 on 11 and 715 DF,  p-value: < 2.2e-16
```

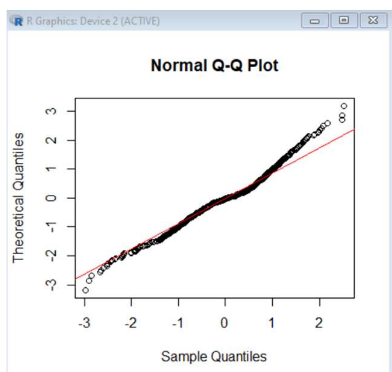
Remarks:

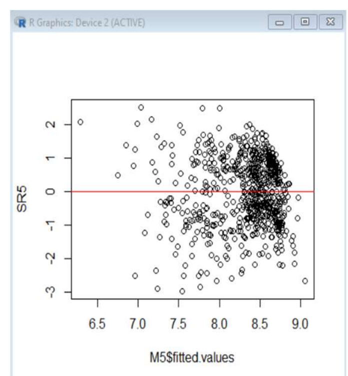
- While checking for the adequacy of M5, got more outliers that might be affecting the model and could eliminate the violations if these outliers were removed.
- Thus, on modifying M5 repeatedly removed the outliers at the following positions: 1,239,2,27,69,668, 302,328, 341, 358, 359, 669, 726, 299, 325, 338, 355, 356, 665, 722

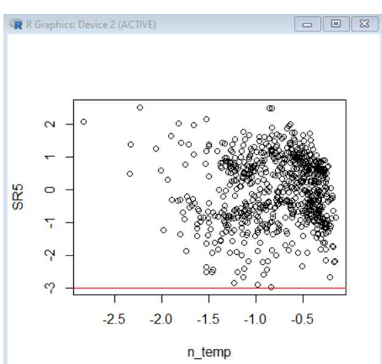
Checking the adequacy of the final model:

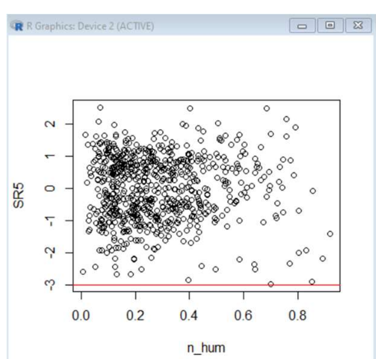
- No outliers
- No influential points
-

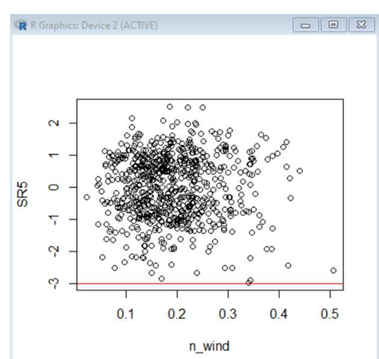












From these graphs we can conclude that:

Almost all violations have been removed and the regressors are linear with constant variance and are normally distributed whereas the response is also linear, has constant variance and is almost normal. QQ plot suggests that since the right tail is slightly shorter than normal thus, the response is slightly not normal.

Remarks:
- At last, the model has no outliers and almost all violations are eliminated from the model
-So, M5 is the final model that has:
-F-statistic = 127 and a null distribution of $F(11,699)$
-Residual standard error = 0.2981
-P-value of F-test < 0.05
-Adjusted R^2 = 0.6613, suggests that the model is good
-Thus, the model is almost adequate and significant

7) Had re-categorized season, weathersit and workingday and stored it in new variables seasons, weather and workday as follows:
weather = ifelse(weathersit == 1, "Good", "Bad")
seasons = ifelse(season == 1 | season == 2, "Summer", "Winter")
workday = ifelse(workingday == 1, "Yes", "No")

8) **Final model: M5**
Assumptions: - The data is random
- The response and regressors have a linear relationship
- The residuals of the build model are normal
- The regressors and response have constant variance

R output:

```
> M5 = lm(n_cnt~n_season+n_weathersit+n_workingday+n_wind+n_hum+n_temp+n_weat$
> summary(M5)

Call:
lm(formula = n_cnt ~ n_season + n_weathersit + n_workingday +
    n_wind + n_hum + n_temp + n_weathersit * n_wind + n_season *
    n_temp + n_season * n_hum + n_weathersit * n_hum + n_weathersit *
    n_workingday, data = new_data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.86469 -0.21960  0.01278  0.23739  0.73653

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                       9.75557    0.09056 107.723  < 2e-16 ***
n_seasonWinter                   -0.41681    0.06449  -6.463 1.92e-10 ***
n_weathersitGood                 -0.43631    0.09793  -4.456 9.74e-06 ***
n_workingdayYes                  -0.01878    0.04191  -0.448  0.65425
n_wind                           -1.74862    0.25153  -6.952 8.29e-12 ***
n_hum                            -1.06483    0.12613  -8.442  < 2e-16 ***
n_temp                            1.03336    0.03634  28.437  < 2e-16 ***
n_weathersitGood:n_wind           1.33020    0.32278   4.121 4.22e-05 ***
n_seasonWinter:n_temp            -0.70099    0.05982 -11.718  < 2e-16 ***
n_seasonWinter:n_hum              0.16018    0.13121   1.221  0.22257
n_weathersitGood:n_hum            0.47214    0.17261   2.735  0.00639 **
n_weathersitGood:n_workingdayYes  0.09092    0.05135   1.770  0.07708 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2981 on 699 degrees of freedom
Multiple R-squared:  0.6665,	Adjusted R-squared:  0.6613
F-statistic:   127 on 11 and 699 DF,  p-value: < 2.2e-16
```

Final fitted equation:
X1: season (Winter=1, Summer=0)
X2: weathersit (Good=1, Bad=0)
X3: workingdays (Yes=1, No=0)
X4: windspeed
X5: hum
X6: temp

Y^ = 9.75557 - 0.41681*I(X1=Winter) - 0.43631*I(X2=Good) - 0.01878*(X3=Yes) - 1.74862*X4

- 1.06483*X5 + 1.03336*X6 + 1.33020*X4*I(X2=Good) - 0.70099*X5*I(X1=Winter) + 0.16018*X5*I(X1=Winter) + 0.47214*X5*I(X2=Good) + 0.09092*I(X3=Yes)*I(X2=Good) +

0.2981

Overall Significance of regression:

- The first thing we should test for is whether overall, any terms in the model are significant or not.
- This is analogous F-test.
- The null hypothesis is
  H0 : _1 = _2 = _3 = 0
- The alternative is
  H1 : at least one of _1; _2 or _3 is not zero
- This F-test has test statistic F = 127 on a null distribution F(11,699) and has p-value < 0:00001.
- Thus, at _ = 0:05, we would reject H0.

Conclusion: model adequacy:

- As explained in the point number 6 d.
- The model M5 is adequate after removing the outliers.

R CODE:

```
setwd("C:\\Users\\Anjali\\Desktop\\NUS\\YEAR 1\\SEM 1\\ST1131 Introduction to Statistics and Statistical Computing\\Bike-Sharing-Dataset")

data = read.csv("day.csv")

attach(data)

#PART 1:

# RESPONSE VARIABLE: cnt

# REGRESSORS: season, workingday, weathersit, temp, hum and windspeed

weather = ifelse(weathersit == 1, "Good", "Bad")

seasons = ifelse(season == 1 | season == 2, "Summer", "Winter")

workday = ifelse(workingday == 1, "Yes", "No")


#Q1

summary(cnt)

hist(cnt, col=4) #symmetric distribution

boxplot(cnt, ylab= "Frequency", xlab= "Count of daily total rental bike", col= 2, main = "Boxplot of cnt")

abline(h=median(cnt), col= 5)


#Q2

cor(cnt,temp) #positive and weak association
```

```
plot(cnt~temp) #linear, constant variance
hist(temp) #multimodal

cor(cnt,hum) #negative and weak association
plot(cnt~hum) #linear, constant variance violated
hist(hum) #left skewed

cor(cnt, windspeed) #negative and weak association
plot(cnt~windspeed) #linear, variance might be constant
hist(windspeed) #right skewed

#IQR is almost same, median of winter> summer season
boxplot(cnt~seasons,ylab="Season", xlab="Total daily rental",, main="Boxplot: season vs cnt", col=2)

#IQR range of 0>1, spread of data 0>1, median alomost same, 1 had median slightly higher than 0
boxplot(cnt~workday, col=2, ylab="Working day", xlab="Total daily rental",, main="Boxplot: workingday vs cnt")

#IQR is almost same, median of good weather>bad weather
boxplot(cnt~weather, ylab="Weather", xlab="Total daily rental",, main="Boxplot: weathersit vs cnt", col=2)

#PART 2:
#Q3,4,5,6,7
#MODEL 1
M1= lm(cnt~seasons+workday+weather+temp+hum+windspeed, data=data)
summary(M1) # adj r sqr = 0.4732, workday is insignificant

raw.res = M1$res
SR = rstandard(M1)

which(SR>3 | SR<(-3)) #one outlier at 69
c = cooks.distance(M1)
which(c>1) #no influential points
```

```
hist(SR, prob=TRUE)#slightly left skewed, normality violated

qqnorm(SR, datax = TRUE) #normality violated, tails are slightly shorter than normal
qqline(SR, datax = TRUE, col="red")

#PLOTTING PREDICTED Y^ VS SR TO CHECK FOR CONSTANT VARIANCE AND NORMALITY
plot(M1$fitted.values, SR) #slightly funnel shaped, equal variance violated
abline(h=0, col="red")

#PLOTTING EACH X AGAINST SR TO CHECK FOR X'S LINEARITY ASSUMPTION
plot(data$temp, SR) #linearity might be violated
abline(h=3, col="red")
abline(h=(-3), col="red")

plot(data$hum, SR) #linearity check of hum
abline(h=3, col="red")#might be linear
abline(h=(-3), col="red")

plot(data$windspeed, SR) #linearity check of windspeed
abline(h=3, col="red") #might be linear
abline(h=(-3), col="red")

#Workingday is insignificant
#There is one outlier which is not influential
#The model is not adequate due to the violations of the assumptions
#There might interaction terms among the regressors



#MODEL 2
#Since we added interaction terms of workday with other regressors, workday is still a regressor
#Added interaction terms
M2 = lm(cnt~seasons+weather+hum+temp+windspeed+workday+
        seasons*hum+seasons*temp+seasons*windspeed+
        weather*windspeed+weather*temp+weather*hum+
```

```
        hum*workday+hum*temp+hum*windspeed+
        temp*windspeed+temp*workday+
        windspeed*workday+
        seasons*weather+seasons*workday+weather*workday, data=data)
summary(M2)
# Adjusted R^2 = 0.5454
# Interation terms that are insignificant:
# seasonsWinter*windspeed, weatherGood*temp, seasonsWinter*hum, hum*workdayYes,hum*temp,hum*windspeed,
# temp*windspeed, workdayYes*windspeed, seasonsWinter*weatherGood and seasonsWinter*workdayYes


#MODEL 3
#Removed all insignificant regressors
M3 = lm(cnt~seasons+weather+hum+temp+windspeed+workday+
        seasons*temp+
        weather*windspeed+weather*hum+
        temp*workday+
        weather*workday, data=data)
summary(M3)
#Adjusted R^2 = 0.5485
#hum and windspeed are insignificant but since they are involved in interaction terms, did not remove them
#temp*workingday is insignificant
#Modifying model 3 itself and removing the insignificant term


M3 = lm(cnt~seasons+weather+hum+temp+windspeed+workday+
        seasons*temp+
        weather*windspeed+weather*hum+
        weather*workday, data=data)
summary(M3)


raw.res = M3$res
SR3 = rstandard(M3)
```

```
which(SR3>3 | SR3<(-3)) #no outliers
c3 = cooks.distance(M3)
which(c3>1) #no influential points

hist(SR3, prob=TRUE) #slighlt left skewed

qqnorm(SR3, datax = TRUE) #normality violated, tails are slightly shorter than normal
qqline(SR3, datax = TRUE, col="red")

plot(M3$fitted.values, SR3) #slightly funnel shaped, equal variance might be violated
abline(h=0, col="red")

#PLOTTING EACH X AGAINST SR
plot(data$temp, SR3) #constant variance might be violated
abline(h=3, col="red")
abline(h=(-3), col="red")

plot(data$hum, SR3)
abline(h=3, col="red")
abline(h=(-3), col="red")

plot(data$windspeed, SR3)
abline(h=3, col="red")
abline(h=(-3), col="red")

#Adjusted R^2 = 0.5524
#Model is not adequate due to the violations
#Doing transformations to the response and regressors might help in eliminating the violations


#MODEL 4
n_hum = hum^3 #for higher powers the scatterplot might be funnel shaped
n_wind = windspeed^2
n_cnt = log(cnt)
```

```
n_temp = log(temp)

M4 =
lm(n_cnt~seasons+weather+workday+n_wind+n_hum+n_temp+weather*n_wind+seasons*n_temp+weat
her*n_hum+weather*workday, data=data)

summary(M4)


raw.res = M4$res

SR4 = rstandard(M4)


which(SR4>3 | SR4<(-3)) #outliers at 2, 27, 69, 668

c4 = cooks.distance(M4)

which(c4>1) #no influential point


hist(SR4, prob=TRUE) #highly left skewed, might be due to outliers


qqnorm(SR4, datax = TRUE) #left tail is longer & right tail is almost normal, normality violated,

qqline(SR4, datax = TRUE, col="red")


#PLOTTING PREDICTED Y^ VS SR TO CHECK FOR CONSTANT VARIANCE AND NORMALITY

plot(M4$fitted.values, SR4) #normality violated

abline(h=0, col="red")


#PLOTTING EACH X AGAINST SR TO CHECK FOR X'S LINEARITY ASSUMPTION

plot(n_temp, SR4) #linear

abline(h=3, col="red")

abline(h=(-3), col="red")


plot(n_hum, SR4) #linearity check of hum

abline(h=3, col="red")#linear

abline(h=(-3), col="red")


plot(n_wind, SR4) #linearity check of windspeed

abline(h=3, col="red") #linear

abline(h=(-3), col="red")
```

```
#Adjusted R^2 = 0.5976

#After the transformations the adjusted R^2 increased and it also elimininated the linearity violations of the regressors

#Still the model is not adequate

#Suspected reason is due to outliers



#MODEL 5

new_data = data[-c(1,239,2,27,69,668, 302,328, 341, 358, 359, 669, 726, 299, 325, 338, 355, 356, 665, 722 ),]

n_hum = (new_data$hum)^3 #for higher powers the scatterplot might be funnel shaped

n_wind = new_data$windspeed

n_cnt = log(new_data$cnt)

n_temp = log(new_data$temp)

n_season = seasons[-c(1,239,2,27,69,668, 302,328, 341, 358, 359, 669, 726, 299, 325, 338, 355, 356, 665, 722 )]

n_weathersit = weather[-c(1,239,2,27,69,668, 302,328, 341, 358, 359, 669, 726, 299, 325, 338, 355, 356, 665, 722 )]

n_workingday = workday[-c(1,239,2,27,69,668, 302,328, 341, 358, 359, 669, 726, 299, 325, 338, 355, 356, 665, 722 )]

M5 = lm(n_cnt~n_season+n_weathersit+n_workingday+n_wind+n_hum+n_temp+n_weathersit*n_wind+n_season*n_temp+n_season*n_hum+n_weathersit*n_hum+n_weathersit*n_workingday, data=new_data)

summary(M5)


raw.res = M5$res

SR5 = rstandard(M5)


which(SR5>3 | SR5<(-3)) #outliers at 2, 239, 358, 669, 668, 667, 662, 353 (removed this from new_data and modified M5)

c5 = cooks.distance(M5)

which(c5>1) #no influential point


hist(SR5, prob=TRUE) #slighlty left skewed, might be due to outliers


qqnorm(SR5, datax = TRUE) #both tails are slighlty shorter than normal, normality slighlty violated

qqline(SR5, datax = TRUE, col="red")
```

```
#PLOTTING PREDICTED Y^ VS SR TO CHECK FOR CONSTANT VARIANCE AND NORMALITY
plot(M5$fitted.values, SR5) #constant variance but normality is minimally violated
abline(h=0, col="red")

#PLOTTING EACH X AGAINST SR TO CHECK FOR X'S LINEARITY ASSUMPTION
plot(n_temp, SR5) #linear, constant variance
abline(h=3, col="red")
abline(h=(-3), col="red")

plot(n_hum, SR5) #linearity check of hum
abline(h=3, col="red")#linear, constant variance
abline(h=(-3), col="red")

plot(n_wind, SR5) #linearity check of windspeed
abline(h=3, col="red") #linear, constant variance
abline(h=(-3), col="red")

#After removing outliers calculated from M4
#This model had an Adjusted R^2 = 0.6613 and almost all major violations eliminated
#Thus, removed more outliers, found in M5, from new_data and modified M5 itself
#There are some insignificant regressors in the model but did not remove them for better estimation of Y^

#At last, the model has no outliers and almost all violations are eliminated from the model
#So, M5 is the final model that has:
#F-statistic = 127 and a null distribution of F(11,699)
#Residual standard error = 0.2981
#P-value of F-test < 0.05
#Adjusted R^2 = 0.6613, suggests that the model is good
#Thus, the model is almost adequate and significant
```

---- THE END ----