Regression Analysis of Shared Bike Rentals

PSTAT 126 Regression Analysis

Randy Zhu, 8554123

Brian Fernandez, 4398335

Introduction

Bike sharing systems are new generation of traditional bike rentals where renting and returning back become automatic. The data set we analyze has various information of bike sharing such as season, holiday, weather, temperature, humidity, etc. We will regress the number of rented bikes, cnt(number of rented bikes), on humidity, workday, temperature to tell what factors are influential to bike sharing and predict in what condition would the bike sharing program be most successful. There are other two predictors, atemp(feeling temperature) and windspeed that will be used in stepwise regression.

Questions of Interest

Does a simple linear regression model apply to cnt and temperature?

Is humidity related to cnt?

Is cnt related to workday/non-workday, after taking into account of temperature and humidity?

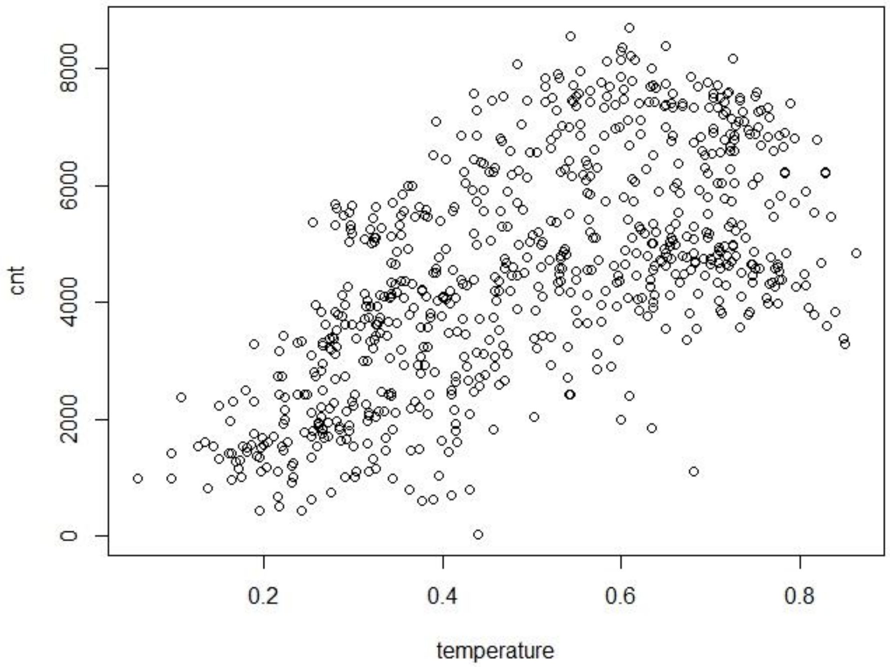
What is the best model to predict the number of rental bikes?

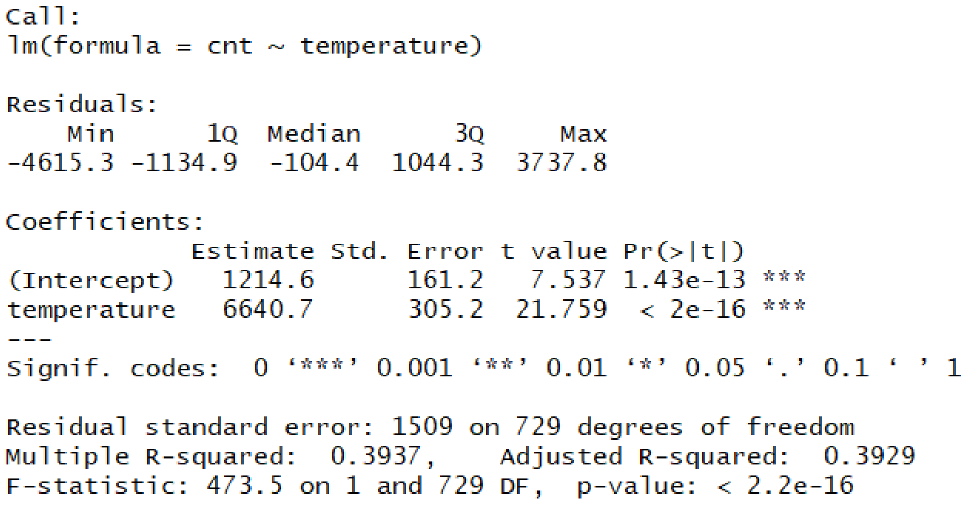
Regression Method

We regress cnt, count of total rental bikes including both casual and registered, on all predictors and do residual analysis and hypothesis tests for different questions. Then, use step wise regression with AIC to build the optimal model to predict the number of total rental bikes.

Cnt vs temperature (normalized)

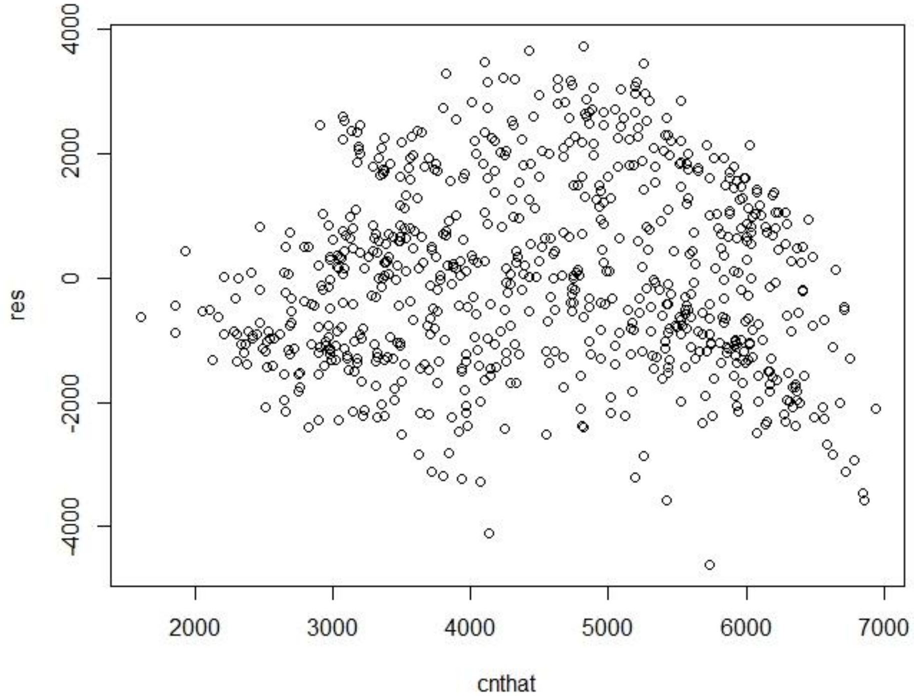
The scatter plot below shows the result of regressing cnt on normalized temperature. Normalized temperature is derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39. According to the scatter plot, there is a rough linear relationship between cnt and temperature. The summary report of this model indicates that the p-value of temperature is smaller than 0.01, so we are positive that temperature is a significant predictor.



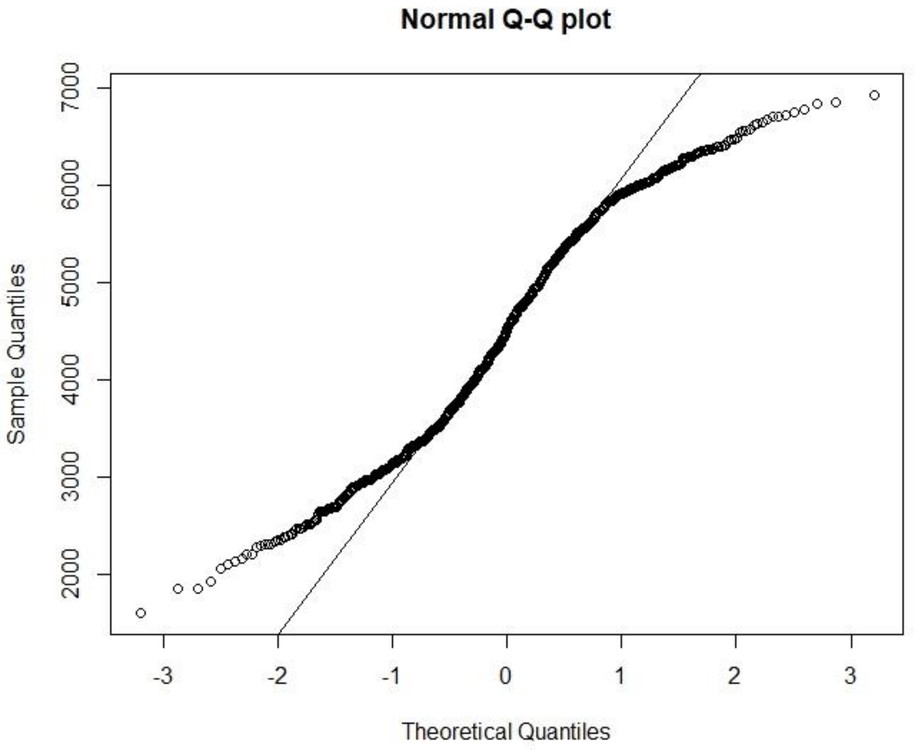


```
Call:
lm(formula = cnt ~ temperature)

Residuals:
    Min      1Q  Median      3Q     Max 
-4615.3 -1134.9  -104.4  1044.3  3737.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   1214.6      161.2   7.537 1.43e-13 ***
temperature   6640.7      305.2  21.759  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1509 on 729 degrees of freedom
Multiple R-squared:  0.3937,	Adjusted R-squared:  0.3929 
F-statistic: 473.5 on 1 and 729 DF,  p-value: < 2.2e-16
```

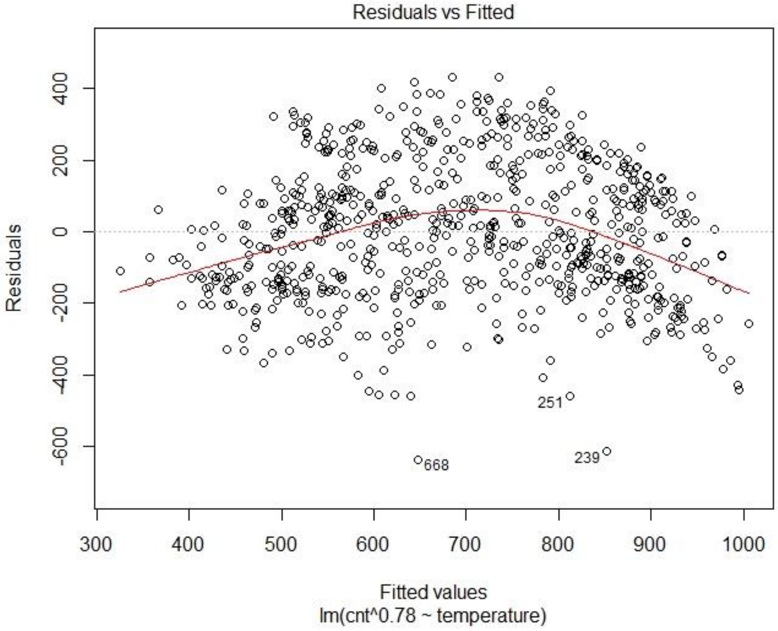
Then we do residuals vs fitted plot to test linearity and equal variance. According to the plot below, the residuals are bouncing randomly around the 0 line and roughly form a horizontal band around the 0 line. So, linearity and equal variances are fulfilled.

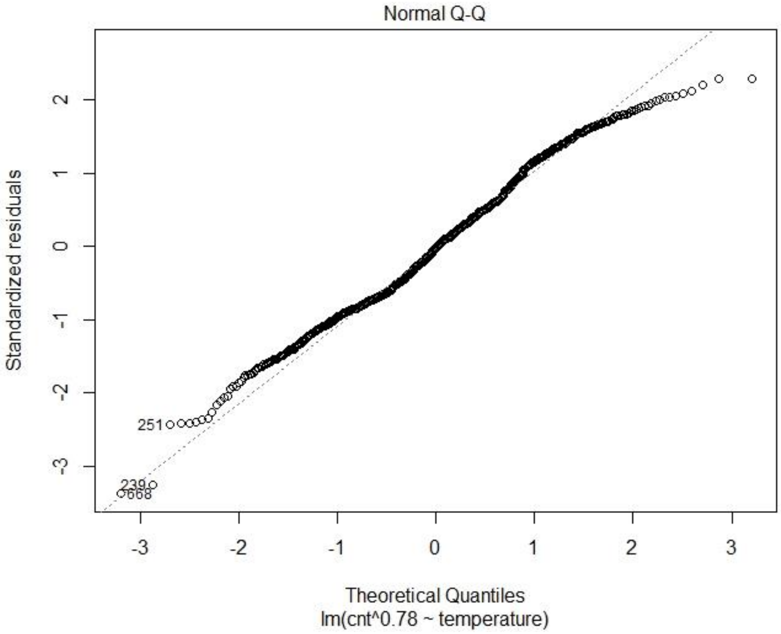


Then we do the Normal Q-Q plot. Obviously, residuals are not normally distributed.



Therefore, we use Box-Cox transformation to get a better model. 0.78 is the best lambda value we get. So we regress cnt^0.78 on temperature again and get the following two plots.





Research question: Does a simple linear regression model apply to cnt and temperature?

No, because residuals are not normally distributed, that is “N” requirement of LINE is not met. However, if we do Box-Cox transformation and take cnt^0.78, LINE requirements are all met.

Cnt vs humidity (normalized)

To answer the question “Is humidity (linearly) related to cnt?”, we test the hypothesis:

H_0 = slope of humidity equals 0 and H_1 = slope of humidity doesn’t equal 0

The full model:

$$Y_i = \beta0 + \beta1xi1 + \beta2xi2 + \beta3xi3 + \varepsilon i$$

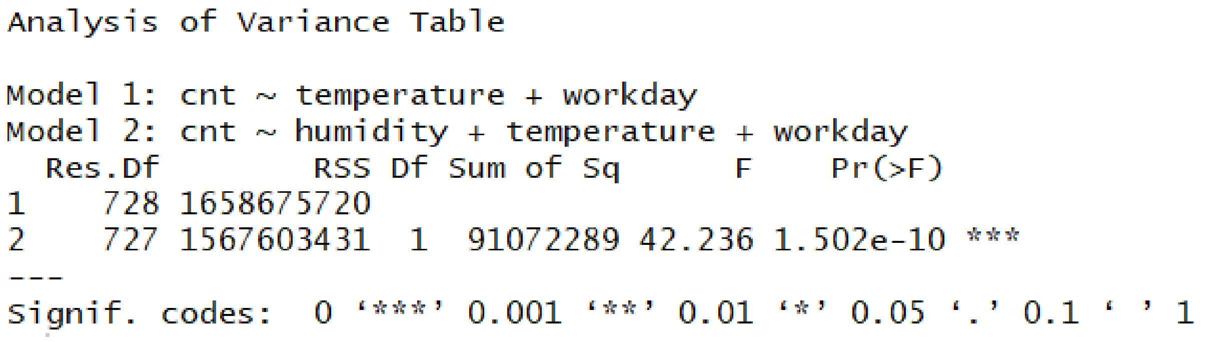
The reduced model:

$$Y_i = \beta0 + \beta2xi2 + \beta3xi3 + \varepsilon i$$

Where $\beta1$ is the slope of humidity, $\beta2$ is the slope of temperature, $\beta3$ is the coefficient of workday/non-workday.

Research question: Is humidity related to cnt?

Yes. Conduct the overall F test in R by anova():



```
Analysis of Variance Table

Model 1: cnt ~ temperature + workday
Model 2: cnt ~ humidity + temperature + workday
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)
1    728 1658675720
2    727 1567603431  1  91072289 42.236 1.502e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

We reject H_0 because there is sufficient evidence (F=42.236, p-value<0.001) to conclude that humidity is related to cnt.

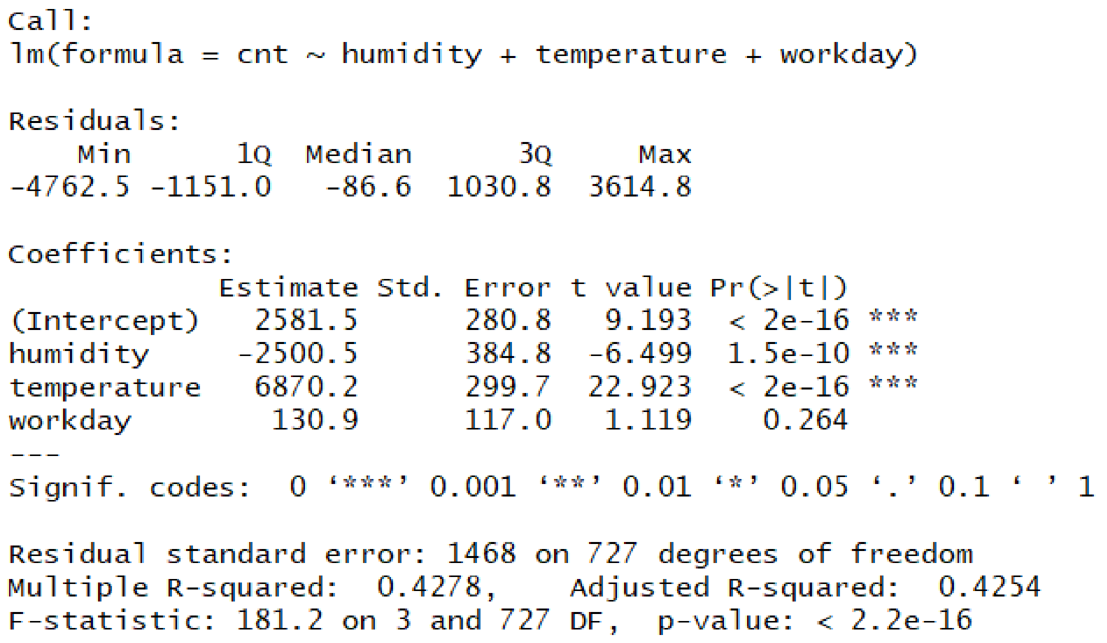
Categorical Predictors: workday

A first-order model with one binary predictor and two quantitative predictors that helps us answer the question is:

$$Y_i = \beta0 + \beta1xi1 + \beta2xi2 + \beta3xi3 + \varepsilon i$$

Where Yi is cnt, xi1 is humidity value, xi2 is temperature value, xi3 is a binary predictor coded as a 1 if it’s a workday, 0 if it’s not.

We get the following estimated regression equation in R:



```
Call:
lm(formula = cnt ~ humidity + temperature + workday)

Residuals:
    Min      1Q  Median      3Q     Max 
-4762.5 -1151.0   -86.6  1030.8  3614.8 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2581.5      280.8   9.193  < 2e-16 ***
humidity     -2500.5      384.8  -6.499  1.5e-10 ***
temperature   6870.2      299.7  22.923  < 2e-16 ***
workday        130.9      117.0   1.119    0.264    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1468 on 727 degrees of freedom
Multiple R-squared:  0.4278,	Adjusted R-squared:  0.4254 
F-statistic: 181.2 on 3 and 727 DF,  p-value: < 2.2e-16
```

Cnt = 2581.5 – 2500.5*humidity +6870.2*temperature +130.9*workday

Therefore, the estimated model for workday = 1 is:

Cnt = 2712.4 – 2500.5*humidity +6870.2*temperature

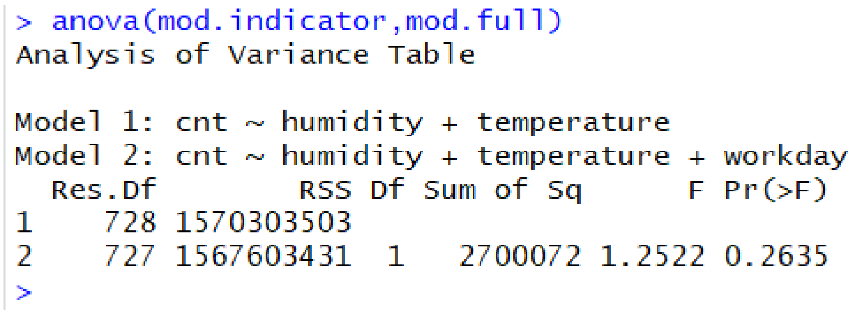
And the estimated model for workday = 0 is:

Cnt = 2581.5 – 2500.5*humidity +6870.2*temperature

That is, we obtain two different parallel estimated lines. The difference between the two lines, 130.9, represents the difference in the average number of rented bikes for a fixed temperature and humidity for workday and non-workday in the sample.

Research question: Is cnt related to workday/non-workday, after taking into account of temperature and humidity?

No. We can answer our research question by testing the null hypothesis H0 : β3 = 0 vs H1 : β3 != 0. Using anova() in R we get the following report:

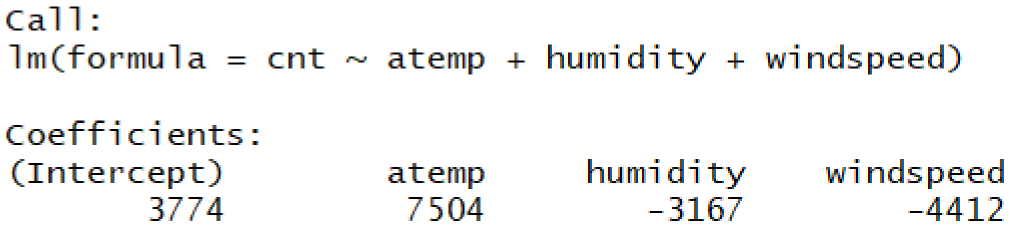


```
> anova(mod.indicator,mod.full)
Analysis of Variance Table

Model 1: cnt ~ humidity + temperature
Model 2: cnt ~ humidity + temperature + workday
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    728 1570303503                           
2    727 1567603431  1   2700072 1.2522 0.2635
> 
```

Since p-value = 0.2635, we fail to reject H_0. Thus, we cannot conclude that there is a statistically significant difference in the mean cnt of workday and non-workday.

Best model with Stepwise Regression with AIC

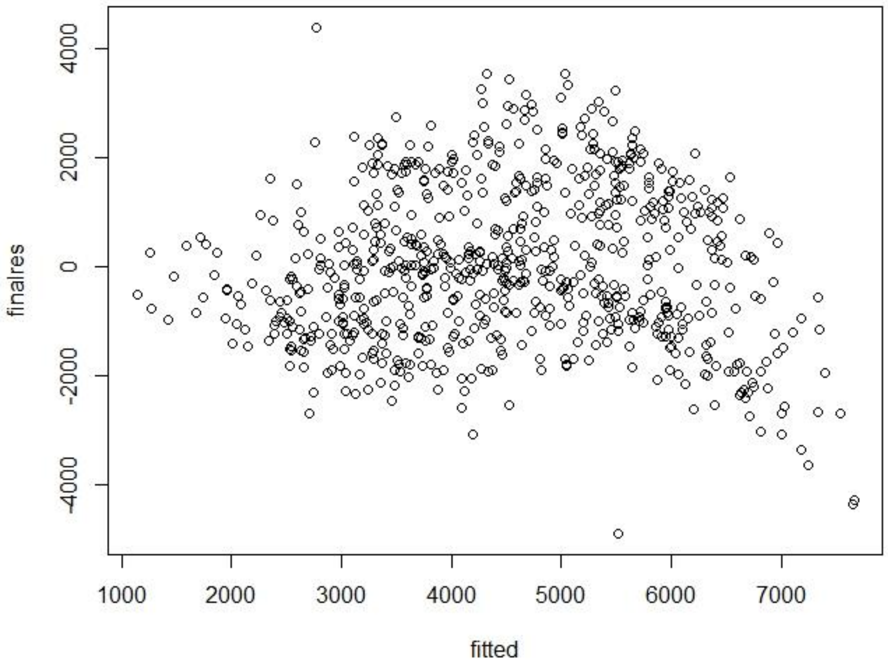
We use step() function in R to build the optimal model with AIC. The final model we obtained is:



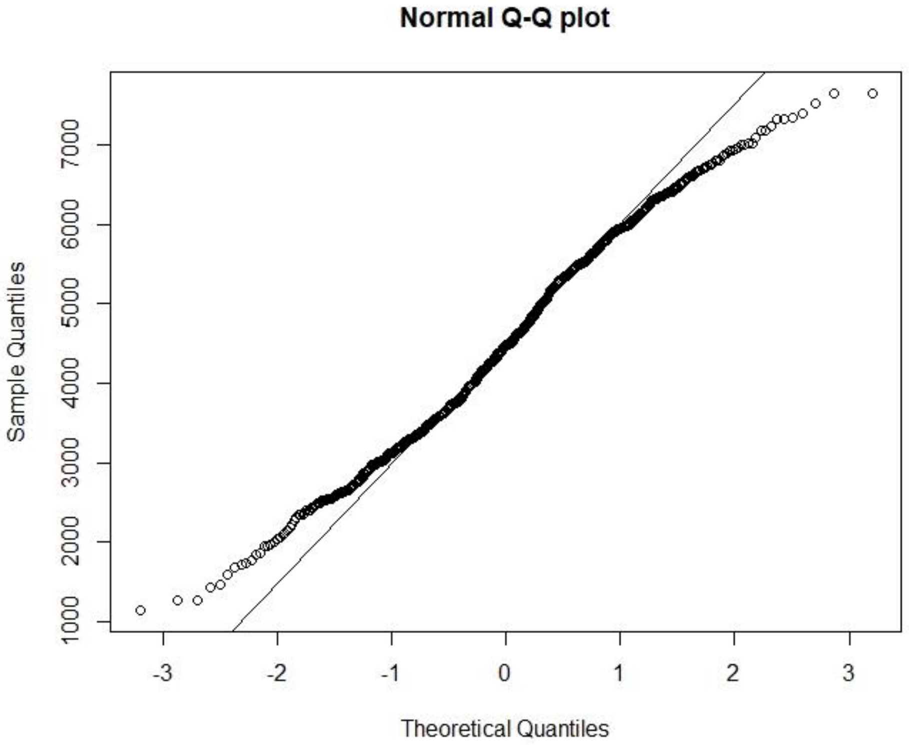
```
Call:
lm(formula = cnt ~ atemp + humidity + windspeed)

Coefficients:
(Intercept)        atemp     humidity    windspeed
       3774         7504        -3167        -4412
```

Cnt = 3774 + 7504*atemp -3167*humidity -4412*windspeed

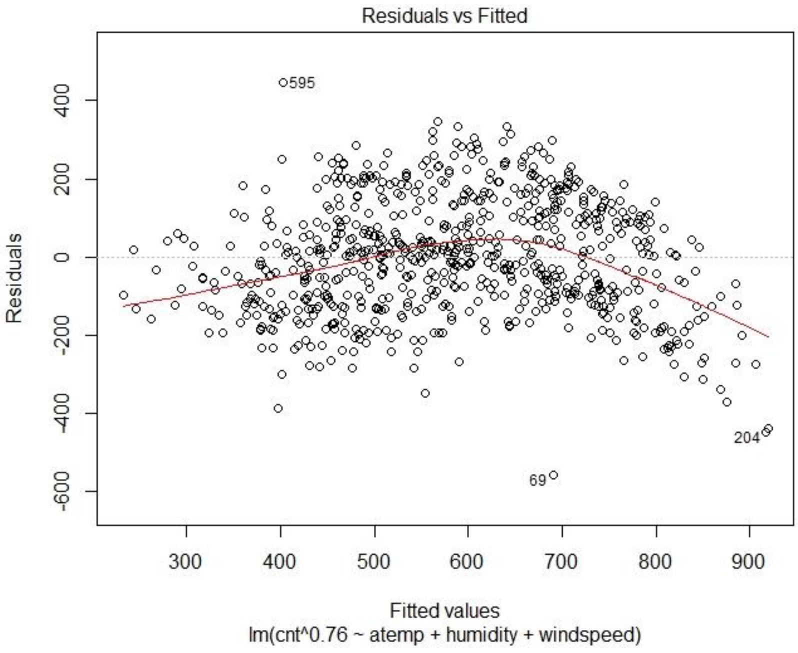
We do the “residuals vs fitted” plot and Normal Q-Q plot to diagnose the model.

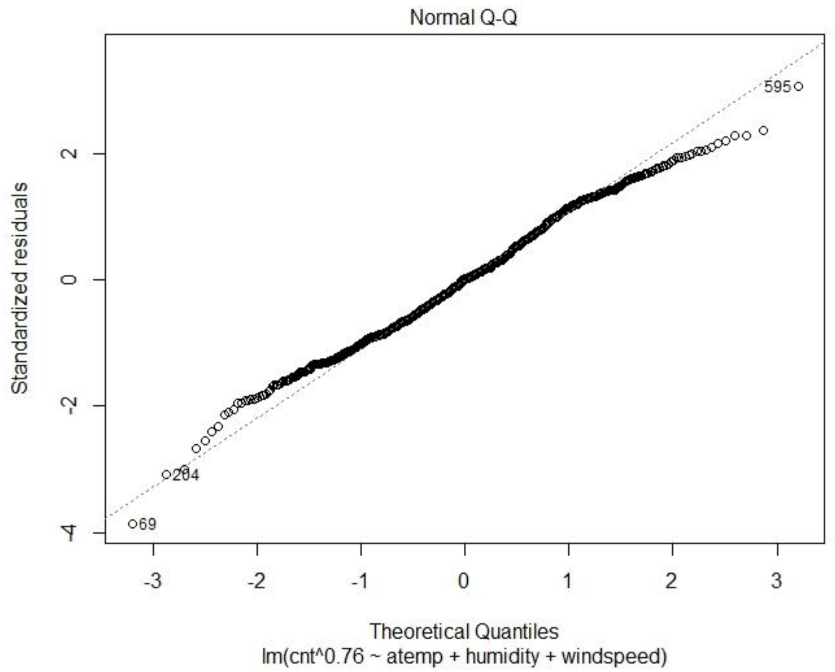


Based on this plot, linearity and equal variances are met.



However, according to the Normal Q-Q plot, normality is violated. Hence, we use Box-Cox transformation again to obtain the optimal model. We get lambda=0.76 from the Box-Cox method. So we regress cnt^0.76 again and we have the revised plots below:



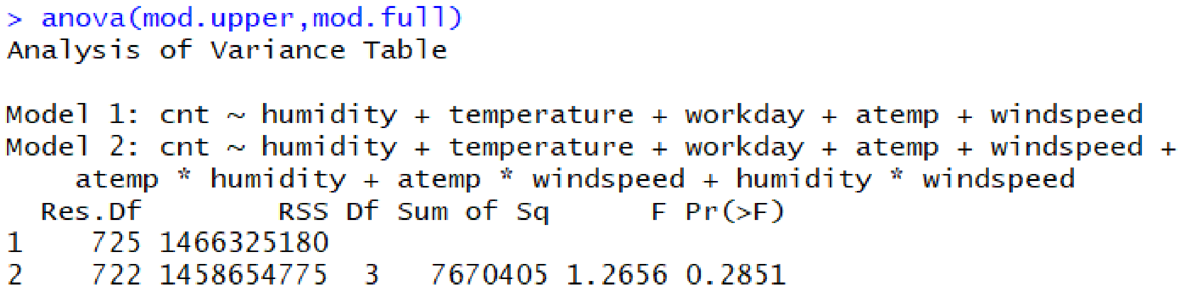


Both plots look better. So we have our final model:

Cnt^0.76 = 3774 + 7504*atemp -3167*humidity -4412*windspeed

This model answers our last question: "What is the best model to predict the number of rental bikes?"

In case there are interaction terms, we regress the model again. This time, regress cnt on all variables and atemp*humidity, atemp*windspeed, humidity*windspeed. Do the anova test in R we get:



```
> anova(mod.upper,mod.full)
Analysis of Variance Table

Model 1: cnt ~ humidity + temperature + workday + atemp + windspeed
Model 2: cnt ~ humidity + temperature + workday + atemp + windspeed +
    atemp * humidity + atemp * windspeed + humidity * windspeed
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    725 1466325180
2    722 1458654775  3   7670405 1.2656 0.2851
```

There is insufficient evidence at the 0.05 level to conclude that at least one of the interaction parameters is not 0.

Conclusion:

In summary, we managed to filter out the best variables to predict the number of rented bikes. There is a simpler linear regression model between cnt and temperature if we transform cnt to cnt^0.78. Humidity is proven to be statistically related to cnt. We can’t conclude that there is a statistically significant difference in the mean cnt of workday and non-workday. Finally, according to AIC method, an optimal model to predict cnt is calculated where feeling temperature, humidity and windspeed are the three best variables to put in the model. According to the anova test for interaction terms, we fail to reject H0, that is all interaction terms’ parameters are 0.

Appendix

```
#PSTAT 126 Project
#Randy Zhu, 8554123  Brian Fernandez, 4398335
library(MASS)

cnt <- day$cnt
humidity <- day$hum #values are divided to 100
workday <- day$workingday
temperature <- day$temp #temperature = (temperature-tmin)/(tmax-tmin)
atemp <- day$atemp
windspeed <- day$windspeed

#1
plot(cnt~temperature)
fit_t <- lm(cnt~temperature)
summary(fit_t)
#According to the linear model we get cnt=1214.6+6640.7*temperature
#Since p-value for temperature is way smaller than 0.01 we are positive that temperature is a significant predictor
cnthat <- 1214.6+6640.7*temperature #predicted cnt values
res <- cnt-cnthat #residuals
plot(res~cnthat)
qqnorm(cnthat,main="Normal Q-Q plot")
qqline(cnthat)
boxcox(cnt~temperature,lambda=seq(0.5,1,length=10)) #lambda=0.78
fit_boxcox <- lm(cnt^0.78~temperature)
summary(fit_boxcox)
plot(fit_boxcox,which=c(1,2))

#2
mod.full <- lm(cnt~humidity+temperature+workday)
mod.reduced <- lm(cnt~temperature+workday)
anova(mod.reduced,mod.full)


#3
fit.all <- lm(cnt~humidity+temperature+workday)
summary(fit.all)
mod.indicator <- lm(cnt~humidity+temperature)
anova(mod.indicator,mod.full)


#4
mod0 <- lm(cnt~1)
mod.upper <- lm(cnt~humidity+temperature+workday+atemp+windspeed)
step(mod0,scope=list(lower=mod0,upper=mod.upper))
fitted <- 3774+7504*atemp-3167*humidity-4412*windspeed
finalres <- cnt-fitted
plot(finalres~fitted)
qqnorm(fitted,main="Normal Q-Q plot")
qqline(fitted)
boxcox(cnt~atemp+humidity+windspeed,lambda=seq(0.7,0.8,length=10)) #lambda=0.76
optimal <- lm(cnt^0.76~atemp+humidity+windspeed)
summary(optimal)
plot(optimal,which=c(1,2))
mod.full <- lm(cnt~humidity+temperature+workday+atemp+windspeed+atemp*humidity+atemp*windspeed+humidity*windspeed)
anova(mod.upper,mod.full)
```