Does lower windspeed lead to a larger number of hourly users?
An Analysis of Bike Share Usage in Arlington, VA Area

Introduction:

Bike sharing systems, a new generation of traditional bike rentals, is an effective solution to traffic, environment and health issues. Members of this system simply rent a bike from a particular position and return it at another position. To better exploit and develop the many perks of this brand new bike rental system, it is important to understand and characterize how the bike share system is used. There are two types of users renting the bikes, regular users and casual users. Although their use patterns are different, their use patterns are both affect by the humidity and temperature, based on our observation. What we don't know is how much windspeed is going to affect their use pattern. Here we introduce two variables. One is the total number of rental bikes being used that particular hour and the other is windspeed, in km/h. It has been hypothesized that a lower windspeed corresponds to more hourly users.

Exploratory Data Analysis:

We have a dataset of 165 samples and two variables, abbreviated as the independent predictor variable "windspeed" and the dependent response variable "count". The summary of these two variables are down below.

Windspeed: Mean = 0.1906903; Median = 0.1940; Minimum: 0.0000; 1st Qu.: 0.1045; 3rd Qu.: 0.2836; Maximum: 0.4627; Variance: 0.01273391; Standard Deviation: 0.1128446

Count: Mean = 102.8; Median = 69.0; Minimum: 1.0; 1st Qu.: 21.0; 3rd Qu.: 139.0; Maximum: 599.0; Variance: 12249.3; Standard Deviation: 110.6766

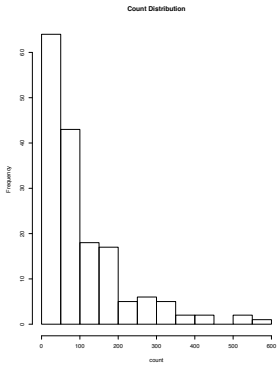


Figure 1.A

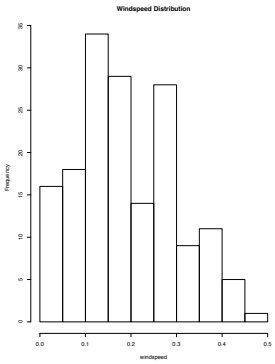


Figure 1.B

Figure 1.A shows the distribution of count. It is skewed to the right with no significant outliers.

Figure 1.B shows the distribution of windspeed. Windspeed distribution is bi-modal, one located at around 0.15 and the other located at around 0.25. There is no significant skewness.

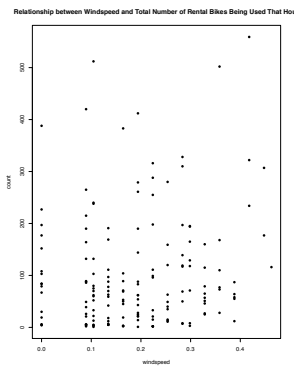


Figure 1.C

Figure 1.C on the right shows the relationship between windspeed and count. We can tell from the graph that the samples seems random enough and have a fairly wide spread. It is hard to tell whether or not there is a linear relationship between these two variables. There are also a couple of outliers on the top of the graph which may potentially affect the linear model we build later on.

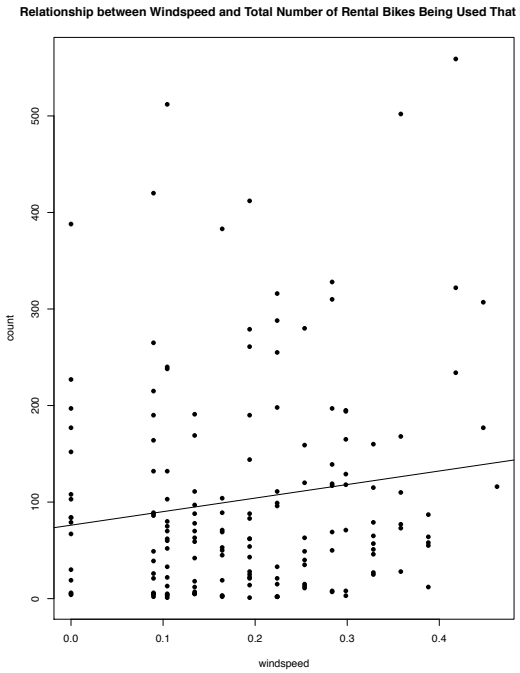


Figure 2.A

Simple Linear Regression Estimation (SLR):

Using the lm() function in R, we get an SLR line "lineDA": Yi = 140.35Xi + 76.07. (Figure 2.A). Every unit increase in windspeed will increment count by 140.35. And if windspeed is zero, the expect count to be 76.07. The line shows a slightly positive linear relationship. We have several concerns for this model. The variance alone the line is not constant. A relatively large group of low count values, although balanced out the outliers on the top, drags the line towards the negative side.

Diagnostic:

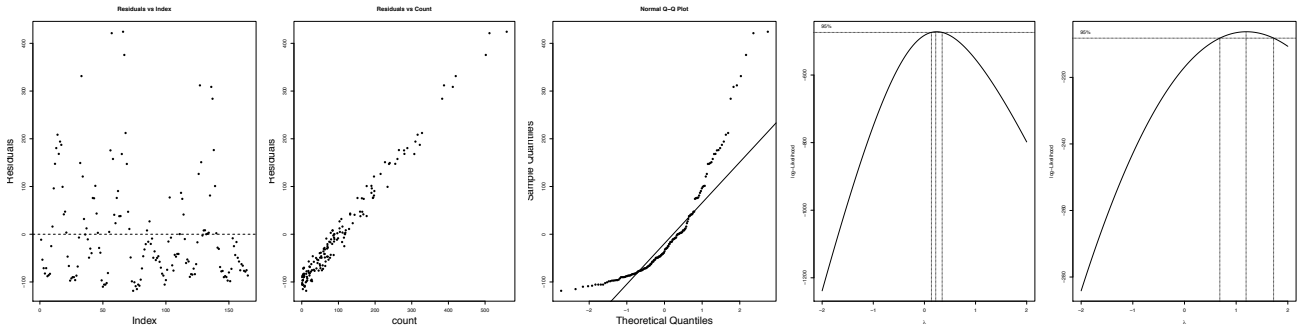


Figure 3.A Figure 3.B Figure 3.C Figure 3.D Figure 3.E

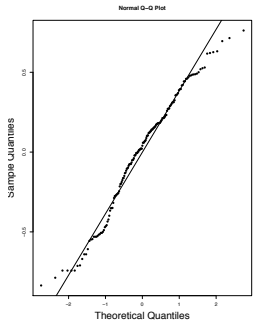
Figure 3.A is the distribution of residuals. The distribution is random with no visible patterns. However, the expectation of the variance is very likely not zero. We have too many values on the top of the graph, pulling the variance to be slightly greater than zero.

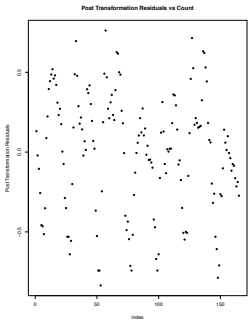
Figure 3.B shows the relationship between count and residuals. We see an extremely strong linear relationship, which indicated dependence. In other words, residuals are strongly associated with "count" the variable, which is not desirable.

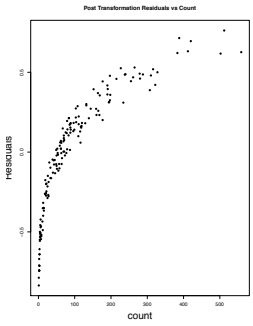
Figure 3.C is the normal probability plot (qq-plot) of "lineDA". It indicated tail problems, which makes sense since our response variable "count" is severely right skewed.

Figure 3.D on the right is the Box-Cox plot, an indication of whether or not this model needs transformation. According to the Box-Cox, the 95% range is between 0 and 1, more closer to 0. This is a clear sign that the model needs transformation. We choose lambda to be 0.2.

Figure 3.E shows the Box-Cox after the transformation on y. We can see that 1 is within the 95% range of lambda now, which is good. Next we will test the new model after transformation.







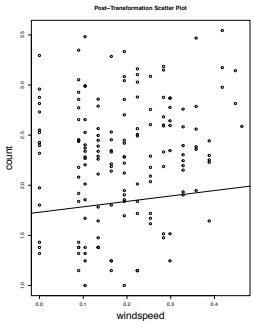


Figure 3.F Figure 3.G Figure 3.H Figure 3.I

Figure 3.F is the qq-plot of the new model after transformation. The previous tail issue is fixed now.
Figure 3.G shows the relationship between post-transformation residuals and their index. Decently spread out, decently random, no pattern at all.
Figure 3.H shows the relationship between post-transformation residuals and counts. It looks like there is a quadratic relationship between these two factors. This is not good because it is very likely that these two factors are correlated.
Finally, take a look at the summary of this new transformed line “lineDAA”. “F-statistic: 4.438 on 1 and 163 DF, p-value: 0.03668” shows that this new line is statistically significant since the p value is less than 0.05. See Figure 3.I.

Conclusion & Discussions

The new linear regression model after transformation fits most of our diagnostic criteria and is statistically significant. However, the strong quadratic pattern on the qq-plot is worth concerning. So basically there does exist an acceptable linear relationship between the windspeed and the number of bikes rented out at that hour. Yet, it does not accord with our hypothesis, which is the lower the windspeed the greater the number of biked rented out.

R Code Appendix

Univariate EDA:

```
summary(windspeed); var(windspeed); sd(windspeed);
summary(count)l var(count); sd(count);
hist(windspeed, main = "Windspeed Distribution", breaks = 10)
hist(count, main = "Count Distribution", breaks = 10)
plot(windspeed, count, pch=16, cex=1, main ="Relationship between Windspeed and
Total Number of Rental Bikes Being Used That Hour")
```

Simple Linear Regression Estimation:

```
lineDA<-lm(count~windspeed)
summary(lineDA)
plot(windspeed, count, pch=16, cex=1, main ="Relationship between Windspeed and
Total Number of Rental Bikes Being Used That Hour")
abline(lineDA)
confint(lineDA, level=0.95)
cor(windspeed, count)
```

Diagnostic:

```
res <- lineDA$res
fit <- lineDA$fit
plot(lineDA$res,pch=16,ylab="Residuals",cex.lab=2,main="Residuals vs Index")
abline(h=0,lwd=2,lty=2)

> count<-DA[,3]
> countNew<-sign(count)*abs(count)^0.15
> lineDAA<-lm(countNew~windspeed)
> boxcox(lineDAA)
> countNew<-sign(count)*abs(count)^0.2
> boxcox(lineDAA)
> ress<-lineDAA$res
> qqnorm(ress)
> qqnorm(ress,pch=16,cex.lab=2);qqline(ress,lwd=2)
> summary(lineDAA)
```

```
Call:
lm(formula = countNew ~ windspeed)

Residuals:
    Min      1Q  Median      3Q     Max
-0.83574 -0.26073  0.03997  0.25868  0.76176

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.73087    0.05681  30.465   <2e-16 ***
windspeed    0.54058    0.25661   2.107   0.0367 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3708 on 163 degrees of freedom
Multiple R-squared:  0.0265,         Adjusted R-squared:  0.02053
F-statistic: 4.438 on 1 and 163 DF,  p-value: 0.03668

> plot(lineDAA$res,pch=16)
> plot(lineDAA$res,pch=16)
> plot(count,ress,pch=16,ylab="Residuals",cex.lab=2,main="Post Transformation
Residuals vs Count")
> plot(lineDAA$res,pch=16,ylab = "Post Transformation Residuals",main = "Post
Transformation Residuals vs Count")
> lineDAA

Call:
lm(formula = countNew ~ windspeed)

Coefficients:
(Intercept)    windspeed
     1.7309       0.5406

> summary(lineDAA)

Call:
lm(formula = countNew ~ windspeed)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.83574 -0.26073  0.03997  0.25868  0.76176

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.73087    0.05681  30.465   <2e-16 ***
windspeed    0.54058    0.25661   2.107   0.0367 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3708 on 163 degrees of freedom
Multiple R-squared:  0.0265,        Adjusted R-squared:  0.02053
F-statistic: 4.438 on 1 and 163 DF,  p-value: 0.03668

> plot(windspeed,count)
> abline(lineDAA)
> lineDA

Call:
lm(formula = count ~ windspeed)

Coefficients:
(Intercept)    windspeed
      76.07       140.35


> plot(windspeed,countNew)
> abline(lineDAA)
> plot(windspeed,countNew,ylab="count",main="Post-Transformation Scatter Plot")
> abline(lineDAA)
> plot(windspeed,countNew,ylab="count",main="Post-Transformation Scatter
Plot",cex.lab=2)
> abline(lineDAA)
> countt<-count^0.4
> lm(countt~windspeed)
```

```
> plot(res)
> plot(lineDA$res,pch=16)
> plot(lineDA$res,pch=16,ylab="Residuals")
> plot(lineDA$res,pch=16,ylab="Residuals",cex.lab=1.5)
> plot(lineDA$res,pch=16,ylab="Residuals",cex.lab=2)
> lineAA<-lm(index~res)
Error in eval(expr, envir, enclos) : object 'index' not found
> lineAA<-lm(res)
Error in formula.default(object, env = baseenv()) : invalid formula
> plot(lineDA$res,pch=16,ylab="Residuals",cex.lab=2,main="Residuals vs Index")
> abline(h=0,lwd=2,lty=2)
> plot(count,res,pch=16,ylab="Residuals",cex.lab=2,main="Residuals vs Count")
> plot(windspeed,res,pch=16,ylab="Residuals",cex.lab=2,main="Residuals vs Count")
> plot(windspeed,lineDA$res,pch=16,ylab="Residuals",cex.lab=2,main="Residuals vs
Count")
> plot(count,res,pch=16,ylab="Residuals",cex.lab=2,main="Residuals vs Count")
> boxcox(lineDA)
> qqnorm(res,pch=16,cex.lab=2);qqline(res,lwd=2)
> count<-sign(count)*abs(count)^0.20
> lineDAA<-lm(count~windspeed)
> lineDAA
```