Lab1

Sahil Jain
September 9, 2017

Reading the csv file into the R console.

```
BikeShare <- read.csv("/Users/sahiljain/Downloads/bike_share.csv")
```

Initializing all the variable equal to y = count (response variable), x1,x2,x3,x4,x5 = explanatory variables

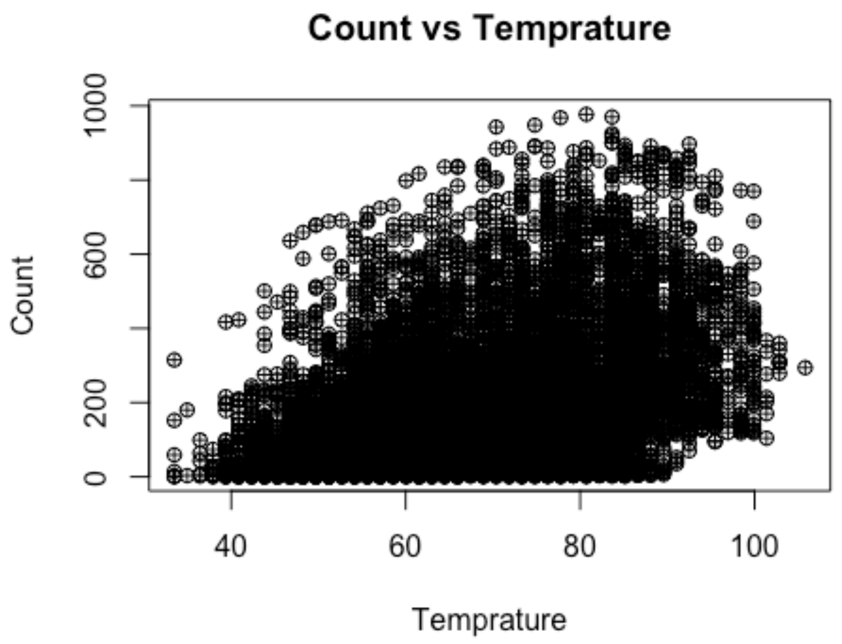
```
y <- BikeShare$count
x1 <- BikeShare$temp
x2 <- BikeShare$humidity
x3 <- BikeShare$windspeed
x4 <- BikeShare$season
x5 <- BikeShare$weather
```

Q1 Construct scatter plots of count versus temp, humidity and windspeed, being sure to appropriately label your axes. In each case describe the linear relationship you observe in terms of 'direction' and 'strength'. Use the correlation coefficient in each case to formalize this interpretation.

There will be 3 different scatter plots and interpretations are made with respect to the plot.

Count vs temp

```
plot(x1, y, ylab = "Count", xlab = "Temprature", main = "Count vs Temprature", pch = 10)
```

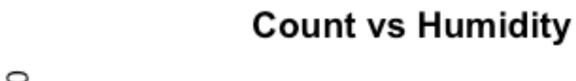


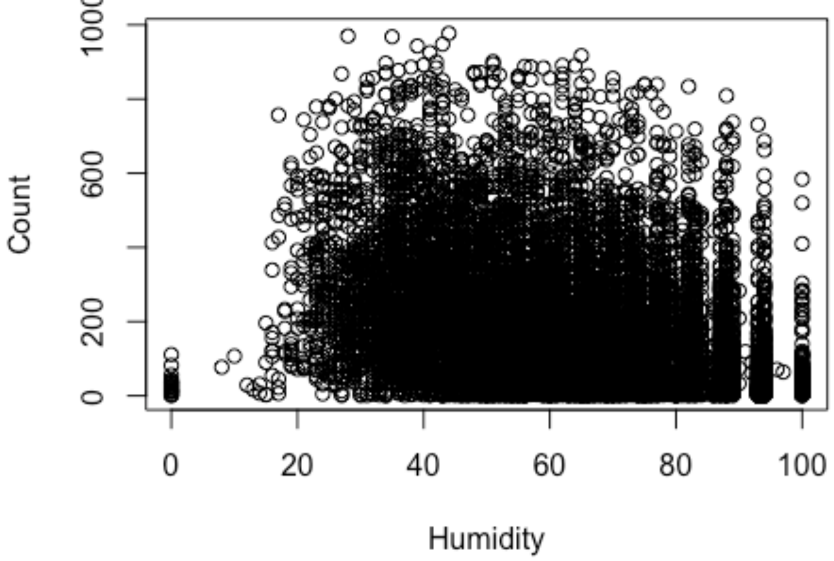
```
cor(x1,y)
## [1] 0.3944536
```

Interpretation : So from the scatter plot and correlation (.3944) of count vs temp we can see that they have a very strong linear relation as value of rho (correlation) is close to 1. From the scatter plot we can see that the count of bike renting goes up as the temprature rises above 40 degrees fahrenheit.

Count vs humidity

```
plot(x2,y, ylab = "Count", xlab = "Humidity", main = "Count vs Humidity")
```



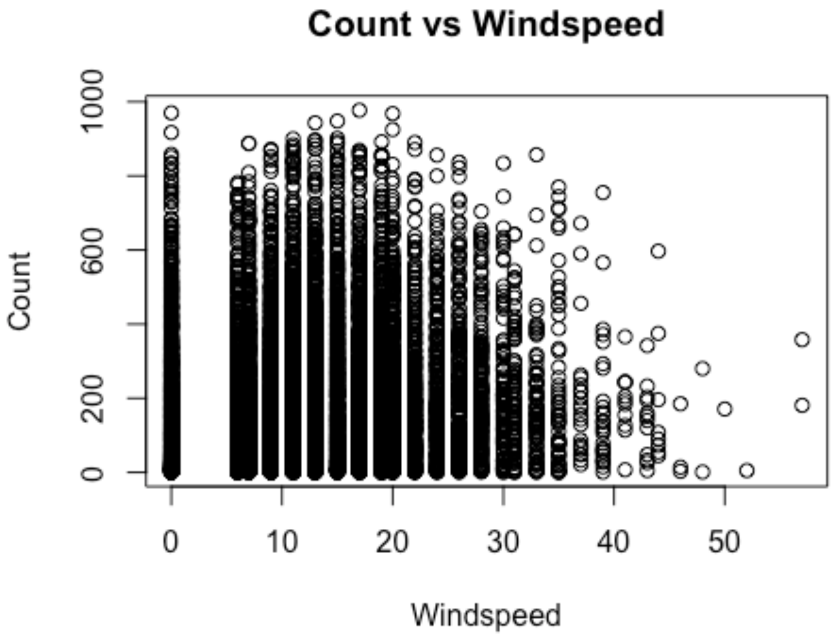


```
cor(x2,y)
## [1] -0.3173715
```

Interpretation : In this case we see that as humidity increseas bike share count decreases. Correlation coefficient between two variables is -0.317 which is closer to -1 hence a stronger linear relationship. Count starts to decrease as humidity starts to go above 40%.

Count vs Windspeed

```
plot(x3,y, ylab = "Count", xlab = "Windspeed", main = "Count vs Windspeed")
```



```
cor(x3,y)
## [1] 0.1013695
```

Interpretation : Bike share count increases as windspeed decreases, but count remains somewhat constant upto windspeed ~22 mph, and then starts to decrease as windspeed increases above 22 mph. Correlation coeffiecient for count vs windspeed in .1013 which is a weak linear relation as it is much closer to zero.

Q2 For each relationship in part (a) calculate the equation of the line-of-best-fit, treating count as the response variable and temp, humidity and windspeed as explanatory variables. Note that you must use the equations derived in class to perform these calculations. You may, however, use automated functions (such as lm() in R and OLS() in Python) to check your answers.

Calculating the parameters and the line of best fit in each of the previous 3 cases.

Count vs temp

```
beta1a_hat <- cor(x1,y) * sd(y) / sd(x1)
```

```
beta1a_hat
## [1] 5.094745
beta0a_hat <- mean(y) - beta1a_hat * mean(x1)
beta0a_hat
## [1] -156.9856
```

beta1_hat = 5.094745 beta0_hat = -156.9856 Line of best fit (Count vs Temp) : Yi = -156.9856 + 5.094745Xi

Count vs Humidity

```
beta1b_hat <- cor(x2,y) * sd(y) / sd(x2)
beta1b_hat
## [1] -2.987269
beta0b_hat <- mean(y) - beta1b_hat * mean(x2)
beta0b_hat
## [1] 376.4456
```

beta1_hat = -2.987269 beta0_hat = 376.4456 Line of best fit (Count vs humidity) : Yi = 376.4456 - 2.987Xi

Count vs Windspeed

```
beta1c_hat <- cor(x3,y) * sd(y) / sd(x3)
beta1c_hat
## [1] 2.249058
beta0c_hat <- mean(y) - beta1c_hat * mean(x3)
beta0c_hat
## [1] 162.7876
```

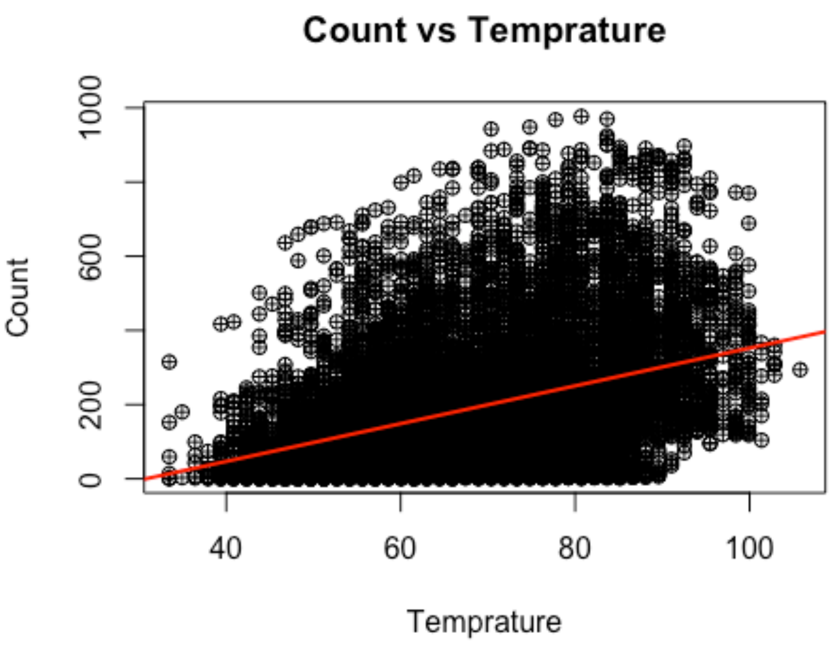
beta1_hat = 2.249 beta0_hat = 162.7876 Line of best fit (Count vs Windspeed) : Yi = 162.7876 + 2.24Xi

Q3 Add the fitted regression lines from part (b) to the appropriate scatter plots constructed in part (a).

Adding fitting line to the scatter plot

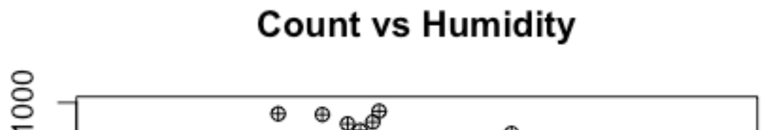
Count vs Temprature

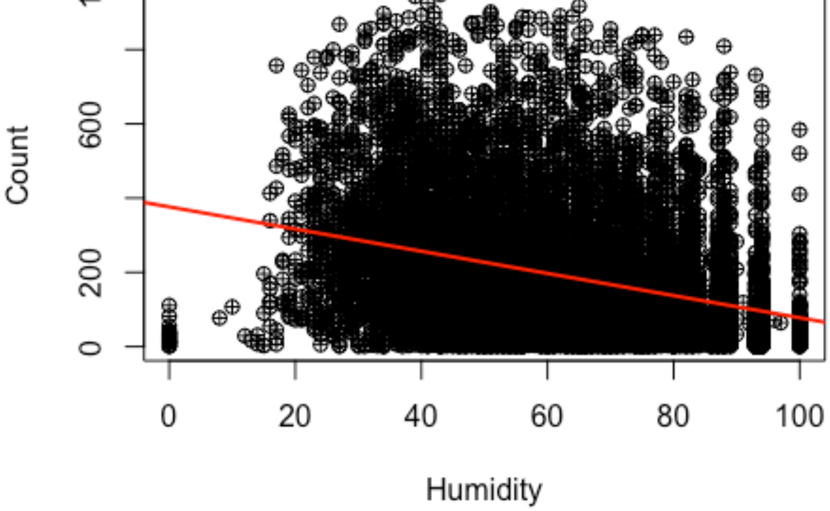
```
plot(x1, y, ylab = "Count", xlab = "Temprature", main = "Count vs Temprature", pch = 10)
abline(beta0a_hat, beta1a_hat, col = "red", lwd = 2)
```



Count vs Humidity

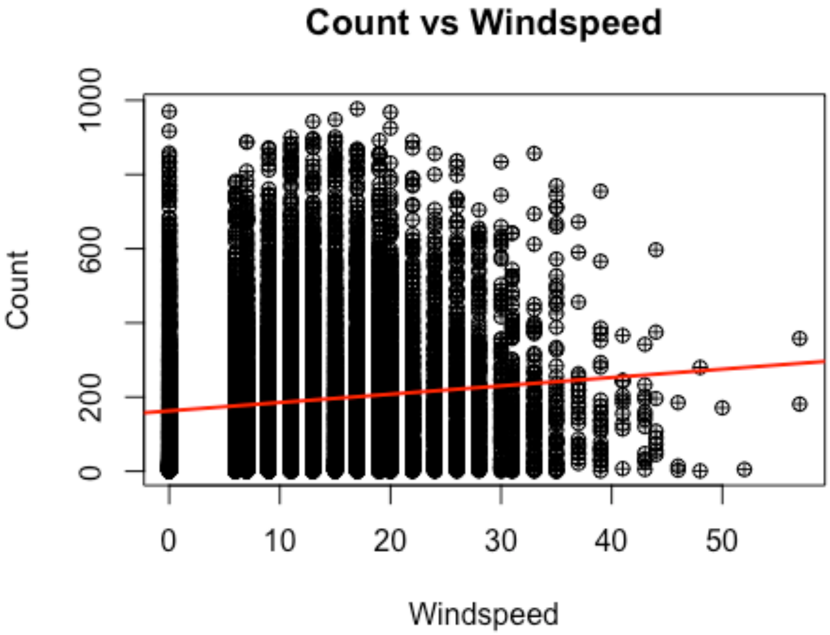
```
plot(x2,y, ylab = "Count", xlab = "Humidity", main = "Count vs Humidity", pch = 10)
abline(beta0b_hat, beta1b_hat, col = "red", lwd = 2)
```





Count vs Windspeed

```
plot(x3,y, ylab = "Count", xlab = "Windspeed", main = "Count vs Windspeed", pch = 10)
abline(beta0c_hat, beta1c_hat, col = "red", lwd = 2)
```



Q4 Based on your findings thus far, rank the variables temp, humidity and windspeed in terms of the strength of their relationship with bike rentals, from most weakly associated to most strongly associated.

Ranking : (1) Count vs Windspeed (2) Count vs Humidity (3) Count vs Temprature

Q5 Using your lines-of-best-fit calculated in (b), calculate the expected number of bike rentals in hours for which (i) the outside temperature is 80 degrees Fahrenheit (ii) the wind speed is 15 miles per hour (iii) the relative humidity is 100%

At temp = 80

```
x <- 80
mu_hat <- beta0a_hat + beta1a_hat * x
mu_hat
## [1] 250.594
```

At windspeed = 15 mph

```
x <- 15
mu_hat <- beta0c_hat + beta1c_hat * x
mu_hat
## [1] 196.5234
```

At humidity 100%

```
x <- 100
mu_hat <- beta0b_hat + beta1b_hat * x
mu_hat
## [1] 77.71875
## [1] 77.71875
```

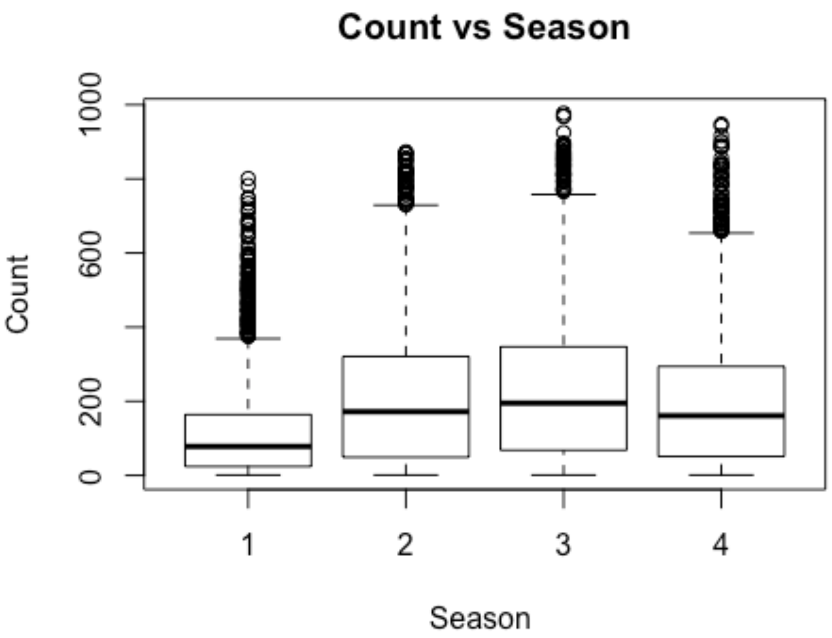
Q6 In each of the cases in part (e) the value of the explanatory variable lies within the range of values actually observed. What risk does one face when predicting outside the range of observed explanatory variable values

The risk one person face while predicting outside the range of observed explanatory varibale values is that the model wont be as consistent as compared to predictic within the range.

Q7 Construct boxplots of count vs. season and count vs. weather, being sure to appropriately label your axes. Comment on the relationship between bike rentals and these two variables.

Box plot for Count vs Season

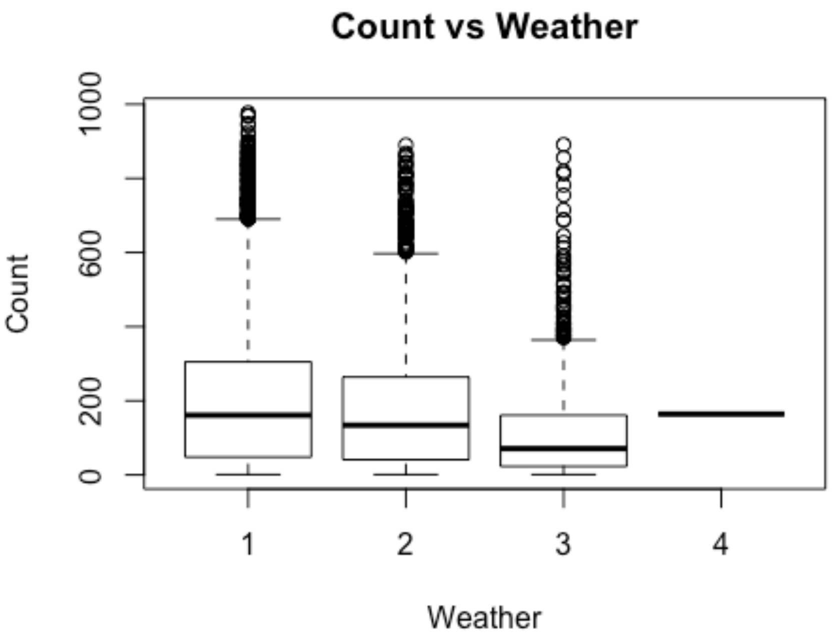
```
boxplot(y~x4, xlab = "Season ", ylab = "Count", main = "Count vs Season")
```



Interpretation : From the box plot we can interpret this thing that there were more bike rentals in season 1 but the median was less than that of season 2,3 and 4.

Box plot for Count vs Weather

```
boxplot(y~x5, xlab = "Weather", ylab = "Count", main = "Count vs Weather")
```



Interpretation : Weather 3 had the most of number of rentals as compared to weather 1 and 2, however weather 1 and 3 have high median value as compared to weather 3. Whereas 4 has least number of bike rentals because it was stormy weather.

Q8 Using automated functions (such as lm() in R and OLS() in Python) fit a simple linear regression model between count, and each of season and weather. Interpret the regressioncoefficients in each case. Do these interpretations seem practically useful?

```
model2 <- lm(y~x4)
summary(model2)
##
## Call:
## lm(formula = y ~ x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -230.19 -138.61  -44.90   88.86  772.34
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  125.087      4.211   29.70   <2e-16 ***
## x4            26.525      1.535   17.28   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 178.7 on 10884 degrees of freedom
## Multiple R-squared:  0.02671,    Adjusted R-squared:  0.02662
## F-statistic: 298.7 on 1 and 10884 DF,  p-value: < 2.2e-16
```

From the model it looks like depending upon the season bike rentals vary. For instance in season 1 or 2 the number would be greater as compared to season 3 or 4.

```
model2 <- lm(y~x5)
summary(model2)
##
## Call:
## lm(formula = y ~ x5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -205.96 -139.96  -44.96   91.04  770.04
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  243.727      4.221   57.75   <2e-16 ***
## x5           -36.768      2.717  -13.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 179.6 on 10884 degrees of freedom
## Multiple R-squared:  0.01655,    Adjusted R-squared:  0.01646
## F-statistic: 183.2 on 1 and 10884 DF,  p-value: < 2.2e-16
```

In the regression result of count vs weather, it seems like number of bike rentals depends upon the weather. Depending on what number 1,2,3 or 4 depicts the bike rentals vary.

In both of the above cases the interpretations are not practically useful.

Q9 Explain why the linear regressions in part (h) are inappropriate. Suggest an alternative approach that would be more appropriate

The linear model in part (h) is inappropriate because the these variables make the model inconsisten and hard to interpret the model in a proper way.