Modeling Rider Count

Amy Tan

With daily information on the season, day type and temperature, this project works to determine the relationship between these attributes on Rider Counts. Methods used includes Principle Component Analysis and Hypothesis Testing.

```
# Loading the Dataset
data<-read.csv("hourlybikes.csv")

  # Data Variates

    # - instant: record index
    # - dteday : date
    # - season : season (1:springer, 2:summer, 3:fall, 4:winter)
    # - yr : year (0: 2011, 1:2012)
    # - mnth : month ( 1 to 12)
    # - hr : hthe (0 to 23)
    # - holiday : weather day is holiday or not (extracted from http://dchr.d
c.gov/page/holiday-schedule)
    # - weekday : day of the week
    # - workingday : if day is neither weekend nor holiday is 1, otherwise is
0.
    # + weathersit :
    #   - 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
    #   - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
    # - temp : Normalized temperature in Celsius. The values are divided to 4
1 (max)
    # - atemp: Normalized feeling temperature in Celsius. The values are divi
ded to 50 (max)
    # - hum: Normalized humidity. The values are divided to 100 (max)
    # - windspeed: Normalized wind speed. The values are divided to 67 (max)
    # - casual: count of casual users
    # - registered: count of registered users
    # - cnt: count of total rental bikes including both casual and registered

# Initializing the explanatory variables
season <-data$season
yr <-data$yr
month <-data$mnth
hr <-data$hr
```

```
holiday <-data$holiday
weekday <-data$weekday
atemp <-data$atemp
temp <-data$temp
workingday <- data$workingday
weathersit <- data$weathersit
windspeed <- data$windspeed
humidity <- data$hum

# Changing seasons, month and time into categorical variables
spring <-ifelse(season==1,1,0)   # Seasons
summer <-ifelse(season==2,1,0)
fall <-ifelse(season==3,1,0)

feb <-ifelse(month==2,1,0)      # Months
mar <-ifelse(month==3,1,0)
apr <-ifelse(month==4,1,0)
may <-ifelse(month==5,1,0)
jun <-ifelse(month==6,1,0)
jul <-ifelse(month==7,1,0)
aug <-ifelse(month==8,1,0)
sep <-ifelse(month==9,1,0)
oct <-ifelse(month==10,1,0)
nov <-ifelse(month==11,1,0)
dec <-ifelse(month==12,1,0)


oneam <-ifelse(hr==1,1,0)      # Time of Day
twoam <-ifelse(hr==2,1,0)
threeam <-ifelse(hr==3,1,0)
fouram <-ifelse(hr==4,1,0)
fiveam <-ifelse(hr==5,1,0)
sixam <-ifelse(hr==6,1,0)
sevenam <-ifelse(hr==7,1,0)
eightam <-ifelse(hr==8,1,0)
nineam <-ifelse(hr==9,1,0)
tenam <-ifelse(hr==10,1,0)
elevenam <-ifelse(hr==11,1,0)
noon <-ifelse(hr==12,1,0)
onepm <-ifelse(hr==13,1,0)
twopm <-ifelse(hr==14,1,0)
threepm <-ifelse(hr==15,1,0)
fourpm <-ifelse(hr==16,1,0)
fivepm <-ifelse(hr==17,1,0)
sixpm <-ifelse(hr==18,1,0)
sevenpm <-ifelse(hr==19,1,0)
eightpm <-ifelse(hr==20,1,0)
ninepm <-ifelse(hr==21,1,0)
tenpm <-ifelse(hr==22,1,0)
```

```r
elevenpm <-ifelse(hr==23,1,0)

rider_count <- data$cnt     # Response Variable

revised_data <- data.frame(rider_count, yr, holiday, weekday,
                           atemp, temp, workingday,
                           weathersit, windspeed, humidity,
                           spring, summer, fall,
                           # categorical variables for seasons
                           feb, mar, apr, may, jun, jul, aug,
                           sep, oct, nov, dec,
                           # categorical variables for Month
                           oneam, twoam, threeam, fouram, fiveam, sixam, seve
nam,
                           eightam, nineam, tenam, elevenam, noon,
                           onepm, twopm, threepm, fourpm, fivepm, sixpm, seve
npm,
                           eightpm, ninepm, tenpm, elevenpm
                           # categorical variables for Hours of the Day
                           )

# Accessing the explanatory variables for better analysis
revised_reg <- lm(rider_count~., data= revised_data)
summary(revised_reg)

##
## Call:
## lm(formula = rider_count ~ ., data = revised_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -400.30  -60.86   -6.98   51.68  438.45
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  12.6808    16.7740   0.756  0.44967
## yr           85.6690     1.5676  54.648  < 2e-16 ***
## holiday     -20.0380     4.8663  -4.118 3.84e-05 ***
## weekday       2.4045     0.3894   6.175 6.75e-10 ***
## atemp       131.8651    30.6750   4.299 1.73e-05 ***
## temp         22.8037     5.6933   4.005 6.22e-05 ***
## workingday    4.8060     1.7295   2.779  0.00546 **
## weathersit  -24.1453     1.4092 -17.135  < 2e-16 ***
## windspeed    -4.2500     0.8624  -4.928 8.37e-07 ***
## humidity    -16.3527     1.0729 -15.241  < 2e-16 ***
## spring      -68.4387     4.8957 -13.979  < 2e-16 ***
## summer      -29.8970     5.7454  -5.204 1.98e-07 ***
## fall        -36.1285     5.1737  -6.983 2.99e-12 ***
## feb           1.9855     3.9282   0.505  0.61325
## mar          13.1557     4.4176   2.978  0.00291 **
```

```
## apr           3.7224     6.5601   0.567  0.57044
## may          17.5594     7.0188   2.502  0.01237 *
## jun           2.2390     7.2104   0.311  0.75617
## jul         -17.8233     8.0839  -2.205  0.02748 *
## aug           3.9727     7.8851   0.504  0.61439
## sep          29.2350     7.0095   4.171 3.05e-05 ***
## oct          13.1402     6.4956   2.023  0.04309 *
## nov         -11.6206     6.2523  -1.859  0.06310 .
## dec          -6.5713     4.9669  -1.323  0.18585
## oneam       -17.3829     5.3610  -3.242  0.00119 **
## twoam       -26.4177     5.3804  -4.910 9.19e-07 ***
## threeam     -36.8645     5.4192  -6.803 1.06e-11 ***
## fouram      -40.1095     5.4246  -7.394 1.49e-13 ***
## fiveam      -22.7914     5.3891  -4.229 2.36e-05 ***
## sixam        36.2103     5.3747   6.737 1.67e-11 ***
## sevenam     170.7910     5.3646  31.837  < 2e-16 ***
## eightam     311.6619     5.3577  58.171  < 2e-16 ***
## nineam      164.2234     5.3628  30.623  < 2e-16 ***
## tenam       109.1723     5.3857  20.271  < 2e-16 ***
## elevenam    134.5890     5.4256  24.806  < 2e-16 ***
## noon        173.5228     5.4724  31.709  < 2e-16 ***
## onepm       167.8991     5.5102  30.470  < 2e-16 ***
## twopm       151.6896     5.5410  27.376  < 2e-16 ***
## threepm     160.7194     5.5512  28.952  < 2e-16 ***
## fourpm      222.2980     5.5376  40.144  < 2e-16 ***
## fivepm      375.8370     5.5055  68.265  < 2e-16 ***
## sixpm       344.3369     5.4697  62.954  < 2e-16 ***
## sevenpm     236.0025     5.4199  43.544  < 2e-16 ***
## eightpm     155.8430     5.3893  28.917  < 2e-16 ***
## ninepm      107.0328     5.3687  19.937  < 2e-16 ***
## tenpm        70.3277     5.3589  13.124  < 2e-16 ***
## elevenpm     31.3859     5.3543   5.862 4.66e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.1 on 17332 degrees of freedom
## Multiple R-squared:  0.6843, Adjusted R-squared:  0.6835
## F-statistic: 816.7 on 46 and 17332 DF,  p-value: < 2.2e-16
```

```
# Checking for Multicollinearity
summary(lm(data$atemp~ data$temp))
```

```
##
## Call:
## lm(formula = data$atemp ~ data$temp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.55336 -0.01181  0.00173  0.01386  0.13340
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.4757751  0.0002041  2331.5   <2e-16 ***
## data$temp   0.1697317  0.0002041   831.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0269 on 17377 degrees of freedom
## Multiple R-squared:  0.9755, Adjusted R-squared:  0.9755
## F-statistic: 6.918e+05 on 1 and 17377 DF,  p-value: < 2.2e-16
```

```
summary(lm(data$weekday~ data$workingday))
```

```
##
## Call:
## lm(formula = data$weekday ~ data$workingday)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.89790 -2.05284 -0.05284  1.94716  3.10210
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.89790    0.02699 107.350  < 2e-16 ***
## data$workingday  0.15495    0.03267   4.743 2.13e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.005 on 17377 degrees of freedom
## Multiple R-squared:  0.001293,   Adjusted R-squared:  0.001235
## F-statistic: 22.49 on 1 and 17377 DF,  p-value: 2.125e-06
```

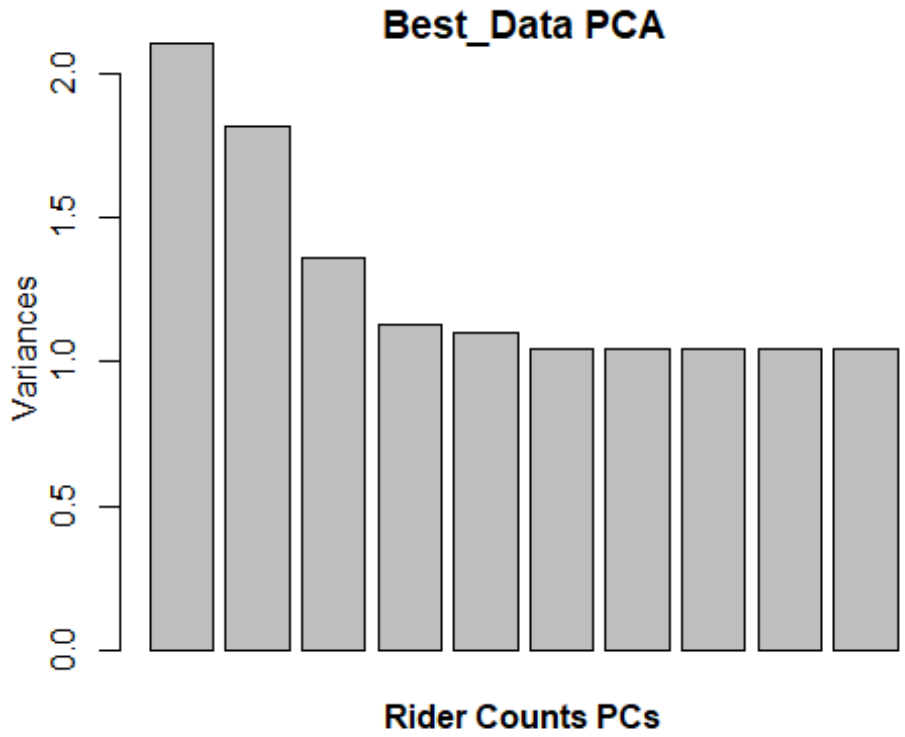
When loading the data I wanted to change the data into something I can use to obtain a better analysis. I took the initiative to take out the variables “instant”, “dteday”, “casual” and “registered”. Leaving in instant and dteday would result in a time series and the casual and registered counts are subsets of rider counts “cnt”. I also switched in categorical variables for season, month and hr. I decided to leave weathersit as a numeric value because it can be viewed as a spectrum of weather harshness. After regressing the variables together, I find that month is not significant to the rider count.The reason could be the relationship it has with season. I also find that temp and atemp are linearly correlated. Along with weekday and workingday. After taking everything into consideration I created a better data set to work with.

```
best_data <- data.frame(rider_count, yr, holiday, weekday, temp,
                        weathersit, windspeed, humidity,
                        spring, summer, fall,
                        # categorical variables for seasons
                        oneam, twoam, threeam, fouram, fiveam, sixam, seve
nam,
                        eightam, nineam, tenam, elevenam, noon,
```

```
                          onepm, twopm, threepm, fourpm, fivepm, sixpm, seve
npm,
                          eightpm, ninepm, tenpm, elevenpm
                          # categorical variables for Hours of the Day
                          )

better_reg <- lm(rider_count~., data=best_data)
summary(better_reg)

##
## Call:
## lm(formula = rider_count ~ ., data = best_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -407.59  -60.45   -6.71   50.12  471.39
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  81.6239     4.7280  17.264  < 2e-16 ***
## yr           85.7856     1.5734  54.523  < 2e-16 ***
## holiday     -25.8780     4.6938  -5.513 3.57e-08 ***
## weekday       2.4055     0.3913   6.148 8.02e-10 ***
## temp         47.6945     1.3703  34.807  < 2e-16 ***
## weathersit  -24.5477     1.4129 -17.373  < 2e-16 ***
## windspeed    -4.8714     0.8377  -5.815 6.17e-09 ***
## humidity    -13.6423     1.0513 -12.977  < 2e-16 ***
## spring      -65.5322     2.4469 -26.782  < 2e-16 ***
## summer      -23.5809     2.3915  -9.860  < 2e-16 ***
## fall        -39.8393     2.9850 -13.347  < 2e-16 ***
## oneam       -17.6334     5.3982  -3.267  0.00109 **
## twoam       -26.7857     5.4172  -4.945 7.70e-07 ***
## threeam     -37.2150     5.4550  -6.822 9.26e-12 ***
## fouram      -40.5656     5.4592  -7.431 1.13e-13 ***
## fiveam      -23.1671     5.4223  -4.273 1.94e-05 ***
## sixam        35.7903     5.4075   6.619 3.73e-11 ***
## sevenam     170.5520     5.3990  31.590  < 2e-16 ***
## eightam     311.7637     5.3943  57.795  < 2e-16 ***
## nineam      164.6022     5.3998  30.483  < 2e-16 ***
## tenam       109.7161     5.4190  20.246  < 2e-16 ***
## elevenam    135.4834     5.4516  24.852  < 2e-16 ***
## noon        174.7015     5.4897  31.824  < 2e-16 ***
## onepm       169.2582     5.5192  30.667  < 2e-16 ***
## twopm       153.1132     5.5435  27.621  < 2e-16 ***
## threepm     162.0819     5.5512  29.197  < 2e-16 ***
## fourpm      223.5303     5.5397  40.350  < 2e-16 ***
## fivepm      376.9303     5.5145  68.352  < 2e-16 ***
## sixpm       345.4216     5.4867  62.956  < 2e-16 ***
## sevenpm     236.9370     5.4451  43.514  < 2e-16 ***
## eightpm     156.6928     5.4200  28.910  < 2e-16 ***
```

```
## ninepm      107.6905     5.4028  19.932  < 2e-16 ***
## tenpm        70.7326     5.3950  13.111  < 2e-16 ***
## elevenpm     31.5852     5.3915   5.858 4.76e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.8 on 17345 degrees of freedom
## Multiple R-squared:  0.6796, Adjusted R-squared:  0.679
## F-statistic:  1115 on 33 and 17345 DF,  p-value: < 2.2e-16
```

```
# PC Analysis of best_data
best.hourly.pca <-prcomp(best_data[,-1],center = TRUE,scale. = TRUE)

# Plotting the variance
par(mar = c(3.1, 3.1, 1.1, 1.1), mgp = 2:0)
plot(best.hourly.pca, main = "Best_Data PCA")
mtext(side = 1, "Rider Counts PCs", line = 1, font = 2)
```



From this first screeplot I find that in PC1-6 captures the most variation and information as the PCs beyond those have variances that levels off and gives you around the same information.

```
# Using the principle components for each data set, I can look at the rotatio
ns
round(best.hourly.pca$rotation[,1:33],2)
```

```
##                PC1   PC2   PC3   PC4   PC5   PC6   PC7   PC8   PC9  PC10  PC
11
## yr          -0.03  0.07  0.04 -0.20  0.00 -0.05  0.00  0.03  0.02  0.00  0.
00
## holiday      0.04  0.03  0.06 -0.22 -0.66 -0.01  0.00  0.00  0.01  0.00  0.
00
## weekday     -0.01  0.03  0.00  0.19  0.69  0.00  0.00  0.00  0.00  0.00  0.
00
## temp        -0.64  0.00 -0.04  0.07 -0.05 -0.01  0.00  0.00  0.00  0.00  0.
00
## weathersit  0.15 -0.30 -0.18  0.59 -0.16 -0.05  0.00  0.00  0.00  0.00  0.
00
## windspeed    0.05  0.41 -0.18  0.34 -0.09  0.06  0.00  0.00  0.00  0.00  0.
00
## humidity     0.08 -0.64 -0.04  0.15 -0.06  0.03  0.00  0.00  0.00  0.00  0.
00
## spring       0.51  0.20  0.33  0.06  0.01 -0.01  0.00  0.00  0.00  0.00  0.
00
## summer      -0.07 -0.02 -0.78 -0.20  0.02  0.00  0.00  0.00  0.00  0.00  0.
00
## fall        -0.52 -0.10  0.44  0.15 -0.03  0.02  0.00  0.00  0.00  0.00  0.
00
## oneam        0.03 -0.10  0.03 -0.16  0.06 -0.05  0.00  0.00 -0.01 -0.01 -0.
02
## twoam        0.03 -0.11  0.02 -0.15  0.07 -0.09  0.01  0.06  0.05  0.01  0.
01
## threeam      0.03 -0.13  0.02 -0.11  0.10 -0.08  0.00  0.02  0.02  0.00  0.
01
## fouram       0.03 -0.14  0.01 -0.08  0.02  0.08  0.01  0.08  0.07  0.01  0.
02
## fiveam       0.04 -0.14  0.03 -0.12  0.05  0.00  0.01  0.09  0.08  0.01  0.
01
## sixam        0.05 -0.14  0.02  0.00  0.00  0.13 -0.01  0.03  0.03  0.01 -0.
03
## sevenam      0.05 -0.13  0.01  0.10 -0.05 -0.02 -0.03 -0.09 -0.05  0.00 -0.
06
## eightam      0.04 -0.09  0.00  0.12 -0.06  0.17 -0.03 -0.02 -0.03  0.00 -0.
08
## nineam       0.02 -0.04 -0.01  0.18 -0.08  0.13 -0.03  0.01 -0.01  0.01 -0.
05
## tenam        0.00  0.01 -0.01  0.14 -0.06  0.08 -0.02  0.04  0.01  0.01  0.
01
## elevenam    -0.02  0.06 -0.01  0.08 -0.03 -0.21  0.00  0.00  0.03  0.01  0.
14
## noon        -0.03  0.10 -0.02  0.13 -0.05 -0.45  0.00 -0.12 -0.07 -0.01  0.
69
## onepm       -0.04  0.13 -0.02  0.08 -0.03 -0.45  0.03 -0.05  0.60  0.32 -0.
45
## twopm       -0.05  0.15 -0.03  0.11 -0.04 -0.15  0.01  0.29 -0.07 -0.81 -0.
28
```

```
## threepm    -0.05  0.16 -0.03  0.11 -0.04 -0.06  0.01  0.37 -0.63  0.48 -0.
22
## fourpm     -0.05  0.16 -0.03  0.07 -0.02  0.27 -0.70 -0.47  0.00 -0.01 -0.
05
## fivepm     -0.04  0.14 -0.03  0.09 -0.03  0.28  0.71 -0.49 -0.04 -0.02 -0.
04
## sixpm      -0.03  0.11 -0.02  0.06 -0.03  0.21  0.01  0.33  0.27  0.06  0.
32
## sevenpm    -0.02  0.07 -0.01 -0.05  0.01  0.32  0.01  0.24  0.16  0.04  0.
08
## eightpm    -0.01  0.03  0.01 -0.11  0.04  0.15  0.01  0.11  0.09  0.02  0.
15
## ninepm      0.00 -0.01  0.02 -0.17  0.06  0.02  0.01 -0.06 -0.08 -0.02  0.
02
## tenpm       0.01 -0.03  0.02 -0.18  0.06  0.05  0.00 -0.09 -0.12 -0.03 -0.
11
## elevenpm    0.02 -0.06  0.02 -0.13  0.04 -0.32 -0.01 -0.29 -0.29 -0.07 -0.
07
##             PC12  PC13  PC14  PC15  PC16  PC17  PC18  PC19  PC20  PC21  PC
22
## yr          0.02  0.00  0.00  0.00  0.04  0.01  0.01  0.00  0.05 -0.02 -0.
07
## holiday     0.00  0.00  0.00  0.00  0.02  0.00  0.00  0.00  0.02  0.00 -0.
01
## weekday     0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.
00
## temp        0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.
00
## weathersit  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.
00
## windspeed   0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.
00
## humidity    0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.
00
## spring      0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.
00
## summer      0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.
00
## fall        0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.00  0.
00
## oneam       0.04  0.00  0.00  0.00  0.15  0.05  0.01  0.00  0.22 -0.53  0.
71
## twoam       0.06  0.01  0.00  0.00  0.14  0.03  0.01  0.00  0.17 -0.10 -0.
31
## threeam     0.02  0.00  0.00  0.00  0.02  0.01  0.00  0.00  0.02 -0.01 -0.
06
## fouram      0.05  0.01  0.00  0.00  0.12  0.02  0.01  0.00  0.13 -0.04 -0.
20
## fiveam      0.07  0.01  0.00  0.00  0.19  0.04  0.02  0.00  0.22 -0.11 -0.
46
```

```
## sixam       0.03  0.02  0.00 -0.01  0.18  0.07  0.01  0.00  0.28  0.77  0.
34
## sevenam    -0.06  0.04  0.00 -0.03  0.34  0.19  0.13 -0.35 -0.66 -0.09  0.
01
## eightam    -0.09  0.02  0.00 -0.03 -0.23  0.17 -0.66  0.45 -0.08 -0.14 -0.
03
## nineam     -0.06  0.01  0.00 -0.02 -0.36 -0.12  0.69  0.30  0.12 -0.15 -0.
02
## tenam      -0.05  0.00  0.00 -0.02 -0.33 -0.36 -0.24 -0.70  0.27 -0.10 -0.
02
## elevenam    0.00  0.01  0.00  0.00  0.36 -0.79 -0.10  0.29 -0.15  0.03  0.
04
## noon       -0.27 -0.05  0.00  0.04 -0.03  0.31  0.00  0.00  0.14  0.03 -0.
01
## onepm      -0.07  0.01  0.00 -0.01 -0.08  0.12 -0.01  0.00  0.00  0.06  0.
03
## twopm       0.00  0.04  0.00 -0.01  0.08  0.11  0.00  0.00  0.05  0.01 -0.
02
## threepm     0.02  0.06  0.00 -0.01  0.13  0.11  0.00  0.00  0.07  0.01 -0.
04
## fourpm      0.08  0.03  0.00  0.00  0.18  0.06  0.01  0.00  0.11 -0.03 -0.
07
## fivepm      0.08  0.02  0.00  0.00  0.16  0.06  0.01  0.00  0.10 -0.04 -0.
06
## sixpm       0.56 -0.07  0.44 -0.02 -0.09  0.09  0.00  0.00 -0.19  0.00  0.
07
## sevenpm    -0.27 -0.60 -0.50  0.13  0.01  0.01  0.02  0.00 -0.14  0.02  0.
02
## eightpm    -0.07  0.75 -0.37  0.18 -0.18 -0.01  0.02  0.00 -0.19  0.07  0.
04
## ninepm     -0.22 -0.02  0.13 -0.83 -0.20 -0.05  0.02  0.00 -0.16  0.11  0.
01
## tenpm      -0.40 -0.08  0.58  0.49 -0.17 -0.08  0.02  0.00 -0.15  0.10  0.
00
## elevenpm    0.52 -0.23 -0.27  0.15 -0.37 -0.05 -0.01  0.00 -0.18  0.13  0.
01
##             PC23  PC24  PC25  PC26  PC27  PC28  PC29  PC30  PC31  PC32  PC
33
## yr         -0.01 -0.09  0.02 -0.51  0.81 -0.01  0.03 -0.07  0.05 -0.04  0.
00
## holiday    -0.01  0.01  0.03  0.03 -0.05 -0.71 -0.01  0.01  0.00 -0.01  0.
00
## weekday     0.00  0.00 -0.04 -0.08 -0.02 -0.69 -0.04 -0.02  0.04  0.00  0.
00
## temp        0.00  0.01  0.00  0.03  0.06 -0.02  0.00  0.28  0.29  0.65 -0.
03
## weathersit  0.00  0.00  0.00  0.25  0.39 -0.07 -0.14  0.30 -0.38  0.06 -0.
01
## windspeed   0.00  0.00 -0.01  0.12  0.09 -0.04  0.78 -0.17  0.10  0.01 -0.
01
```

```
## humidity    0.00  0.00  0.00  0.09  0.08 -0.02  0.07 -0.36  0.63 -0.08  0.
03
## spring      0.00 -0.01  0.01  0.02  0.04  0.04 -0.04  0.58  0.49  0.05  0.
00
## summer      0.00  0.01 -0.01 -0.07 -0.05 -0.02  0.00  0.41  0.19 -0.36  0.
01
## fall        0.00  0.01  0.00  0.03  0.01 -0.01  0.18  0.28  0.01 -0.62  0.
02
## oneam       0.07  0.07 -0.08  0.10  0.07 -0.01  0.13  0.06 -0.06  0.05  0.
20
## twoam      -0.76  0.34 -0.08  0.12  0.02  0.00  0.15  0.07 -0.07  0.06  0.
20
## threeam     0.00 -0.28  0.88  0.08 -0.03 -0.01  0.16  0.07 -0.07  0.06  0.
20
## fouram     -0.02 -0.80 -0.40  0.03 -0.10  0.00  0.13  0.09 -0.08  0.07  0.
20
## fiveam      0.64  0.37 -0.07  0.02 -0.02  0.00  0.16  0.09 -0.08  0.07  0.
20
## sixam      -0.01  0.10 -0.03 -0.20 -0.07  0.00  0.13  0.09 -0.08  0.07  0.
20
## sevenam    -0.02  0.05 -0.03 -0.35 -0.14  0.00  0.11  0.07 -0.06  0.07  0.
20
## eightam    -0.04  0.05  0.00 -0.32 -0.12  0.00  0.06  0.06 -0.05  0.05  0.
21
## nineam     -0.04  0.04  0.02 -0.34 -0.15  0.01  0.00  0.03 -0.02  0.03  0.
21
## tenam      -0.03  0.02  0.02 -0.20 -0.09  0.01 -0.04  0.00  0.01  0.01  0.
21
## elevenam    0.01  0.00 -0.02 -0.05 -0.03  0.01 -0.06 -0.04  0.04 -0.02  0.
21
## noon        0.02 -0.03 -0.03 -0.07 -0.06  0.01 -0.09 -0.06  0.06 -0.04  0.
21
## onepm       0.03 -0.02 -0.03  0.05  0.00  0.01 -0.11 -0.08  0.08 -0.05  0.
21
## twopm       0.01 -0.03  0.01  0.03  0.00  0.01 -0.16 -0.09  0.08 -0.06  0.
22
## threepm     0.01 -0.03  0.02  0.05  0.02  0.01 -0.17 -0.09  0.08 -0.06  0.
22
## fourpm     -0.01 -0.02  0.05  0.11  0.07  0.01 -0.19 -0.08  0.07 -0.06  0.
22
## fivepm     -0.02 -0.01  0.05  0.05  0.04  0.01 -0.17 -0.07  0.06 -0.05  0.
21
## sixpm      -0.01  0.04  0.02  0.06  0.04  0.00 -0.14 -0.05  0.05 -0.04  0.
21
## sevenpm     0.00  0.04  0.00  0.15  0.10  0.00 -0.08 -0.03  0.02 -0.02  0.
21
## eightpm     0.02  0.03 -0.04  0.19  0.12 -0.01 -0.02 -0.01  0.00  0.00  0.
21
## ninepm      0.04  0.02 -0.07  0.23  0.14 -0.02  0.04  0.01 -0.02  0.01  0.
21
```

```
## tenpm       0.04  0.02 -0.08  0.20  0.13 -0.02  0.07  0.03 -0.03  0.02  0.
21
## elevenpm    0.05  0.01 -0.11  0.08  0.05 -0.01  0.11  0.03 -0.04  0.03  0.
21
```

Looking at the rotations vectors, by picking out the higher magnitude values, we can speculate the meaning of the following principle components:

PC1: -temp PC1 is for cooler temperature weathers. PC1 may be for riders who ride in cooler temperatures. The cooler temperatures might attract riders who would otherwise walk in a cooler day but chooses to ride a bike instead.

PC2: -humidity PC2 is for less humid weathers. PC2 may be for riders who ride bikes during times with more windspeed and less humidity.

PC3: spring, -summer, fall PC3 counts for the seasonal effect on riders. PC3 may be for seasonal riders who ride less during the summer. Maybe for students who use the bikes to ride to school.

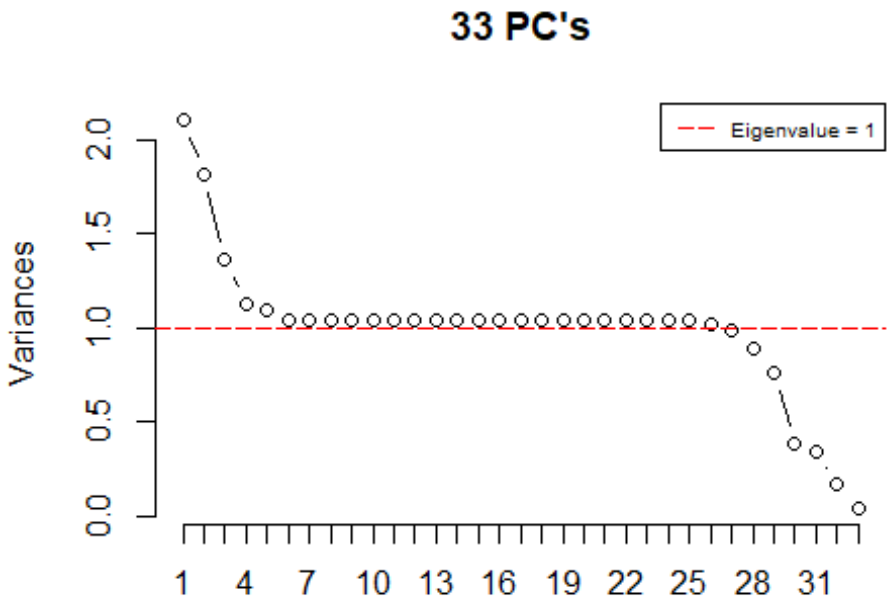
PC4: weathersit PC4 is for harsher and less ideal weathers. PC4 may be for riders who ride bikes during times where the weather is harder to walk in. (ie. harsher weathers and windy days)

PC5: -holiday, weekday PC5 is for working days vs holidays. PC5 may be for employees and workers who rides the bikes to work during the week. Since I took out workday, weekday suggests work day and with the negative holiday rotation value this set of riders only use the bikes to ride to work.

PC6: -noon, -onepm, sevenpm, -elevenpm PC6 is time component. PC6 may be a group of regular riders that uses the bikes at similar and specific times.

These rotations suggests the different factors that may affect the rider count. However PC6 seems suspicious because it only uses certain parts of the categorical variable hr.

```r
# screeplot
screeplot(best.hourly.pca, type = "l", npcs = 33, main = "33 PC's")
abline(h = 1, col="red", lty=5)
legend("topright", legend=c("Eigenvalue = 1"),
col=c("red"), lty=5, cex=0.6)
```



This screeplot illustrates how PC1-6 make up most of the variation before it levels off.

```r
#PCA components
best.hourly.components <- predict(best.hourly.pca)[, 1:5]
best.regression <- lm(rider_count ~ ., data = as.data.frame(best.hourly.compo
nents))


# Using AICc and BIC to choose the number of factors
full.pca <- predict(best.hourly.pca)
pca <- as.data.frame(full.pca) # new predict

kfits <- lapply(1:33, function(k){lm(rider_count~ .,data= pca[, 1:k, drop = F
ALSE])})

n <- nrow(pca)
```

```r
m2ll <- function(reg2) { n * (1 + log(2 * pi) + log(mean(reg2$resid^2))) }
aicc <- function(reg) { m2ll(reg) / n + (2 * reg$rank) / (n - reg$rank - 1) }
bic <- function(reg) { (m2ll(reg) + (log(n) * reg$rank)) / n }


# Plots looks suspiciously weird.
aic_vec <- sapply(kfits, aicc)
#which.min(aic_vec)

aic_vec <- sapply(kfits, aicc)
#which.min(aic_vec)

bic_vec <- sapply(kfits, bic)
#which.min(bic_vec)

# Creating a Lasso Regression
cvlassoPC <- cv.glmnet(x=full.pca, y= rider_count, nfold = 20)
drop(coef(cvlassoPC))
```

```
## (Intercept)         PC1         PC2         PC3         PC4         PC5
## 189.4630876 -49.5013307  49.8044843 -12.2167625  19.7687765  -8.1855777
##         PC6         PC7         PC8         PC9        PC10        PC11
##  37.0832890  19.2260000 -10.5494411  11.2025270   3.1374183  10.2154044
##        PC12        PC13        PC14        PC15        PC16        PC17
##   2.9775552  -2.3488286   2.2102143  -1.3376079  -8.0785987  15.1854679
##        PC18        PC19        PC20        PC21        PC22        PC23
## -16.3572778  16.9198587 -30.6522249  -0.9322585   0.0000000   0.0000000
##        PC24        PC25        PC26        PC27        PC28        PC29
##   7.7276474   0.0000000 -41.9853959  28.5733864   0.0000000 -48.7603356
##        PC30        PC31        PC32        PC33
## -30.4559689  11.8871574  26.9318177 115.6708482
```

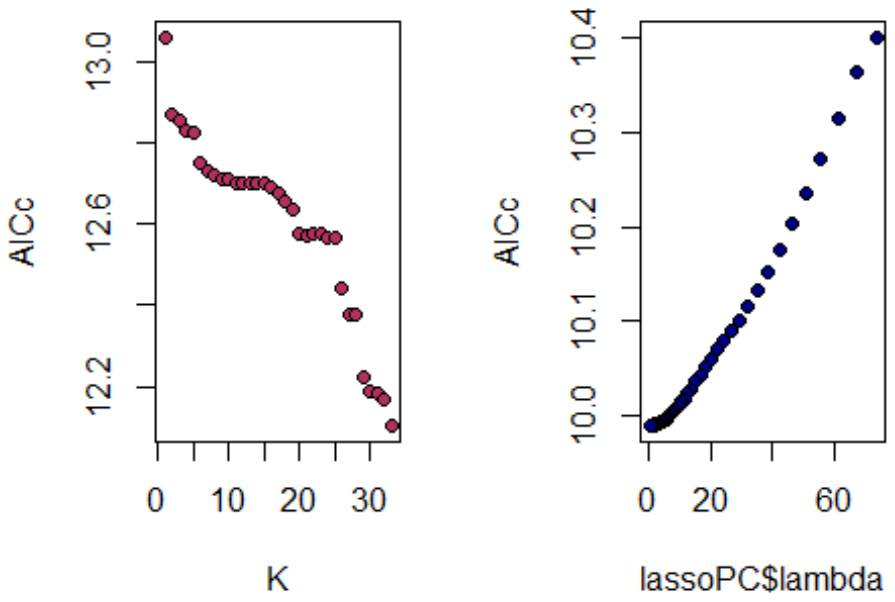
```r
# PC number 33 is very correlated

all_coef <- cbind(coef(cvlassoPC, s = "lambda.min"), coef(cvlassoPC),
                  c(coef(kfits[[33]]), rep(0, 13)),
                  c(coef(kfits[[33]]), rep(0, 17)))
```

```
## Warning in cbind2(argl[[i]], r): number of rows of result is not a multipl
e of
## vector length (arg 1)
```

```r
colnames(all_coef) <- c("lambda.min", "1se", "AICc", "BIC")

all_coef # Matrix
```

```
## 34 x 4 sparse Matrix of class "dgCMatrix"
##              lambda.min         1se        AICc         BIC
## (Intercept) 189.4630876 189.4630876 189.4630876 189.4630876
## PC1         -50.7317656 -49.5013307 -50.8520046 -50.8520046
```

```
## PC2          51.1285544  49.8044843  51.2579435  51.2579435
## PC3         -13.7476007 -12.2167625 -13.8971953 -13.8971953
## PC4          21.4473870  19.7687765  21.6114221  21.6114221
## PC5          -9.8892670  -8.1855777 -10.0557528 -10.0557528
## PC6          38.8297796  37.0832890  39.0004480  39.0004480
## PC7          20.9744631  19.2260000  21.1453242  21.1453242
## PC8         -12.2979330 -10.5494411 -12.4687969 -12.4687969
## PC9          12.9510412  11.2025270  13.1219073  13.1219073
## PC10          4.8859336   3.1374183   5.0567999   5.0567999
## PC11         11.9639528  10.2154044  12.1348223  12.1348223
## PC12          4.7261190   2.9775552   4.8969900   4.8969900
## PC13         -4.0973956  -2.3488286  -4.2682669  -4.2682669
## PC14          3.9587820   2.2102143   4.1296534   4.1296534
## PC15         -3.0861757  -1.3376079  -3.2570471  -3.2570471
## PC16         -9.8272038  -8.0785987  -9.9980788  -9.9980788
## PC17         16.9340782  15.1854679  17.1049537  17.1049537
## PC18        -18.1058976 -16.3572778 -18.2767740 -18.2767740
## PC19         18.6684789  16.9198587  18.8393554  18.8393554
## PC20        -32.4008516 -30.6522249 -32.5717288 -32.5717288
## PC21         -2.6809878  -0.9322585  -2.8518749  -2.8518749
## PC22          0.1661512   .           0.3370471   0.3370471
## PC23         -1.0811175   .          -1.2520502  -1.2520502
## PC24          9.4773925   7.7276474   9.6483789   9.6483789
## PC25         -0.1573946   .          -0.3284329  -0.3284329
## PC26        -43.7542040 -41.9853959 -43.9270533 -43.9270533
## PC27         30.3739001  28.5733864  30.5498477  30.5498477
## PC28          0.9909049   .           1.1754294   1.1754294
## PC29        -50.8069930 -48.7603356 -51.0069939 -51.0069939
## PC30        -33.3435532 -30.4559689 -33.6257302 -33.6257302
## PC31         14.9219650  11.8871574  15.2185287  15.2185287
## PC32         31.2305961  26.9318177  31.6506760  31.6506760
## PC33        124.2944501 115.6708482 125.1371550 125.1371550

lassoPC <- glmnet(best.hourly.components, data$cnt)
aicc_la <- log(deviance(lassoPC) / n) + (2 * lassoPC$df) / (n - lassoPC$df -
1)
par(mfrow = c(1,2))
plot(aic_vec, pch = 21, bg = "maroon", xlab = "K", ylab = "AICc")
plot(lassoPC$lambda, aicc_la, pch = 21, bg = "navy", ylab = "AICc")
```



From the other methods of finding the number of factors, the results were suspicious because they all chose PC33 as the most significant. This may look weird because I cleared out the attributes before running the PCA so there isn't much more I can clear but upon closer investigation PC33 basically takes in the hourly effects which is the same as PC6. There is also something weird going on with PC33 as there is a positive relationship between every hthe and rider counts. After considering all the methods and the spreeplot I made from earlier I decided to stick with PC 1-5 to make the final estimate.

```r
# Reduced

summary(best.regression)

##
## Call:
## lm(formula = rider_count ~ ., data = as.data.frame(best.hourly.components)
)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -387.57  -97.12  -28.84   56.94  740.59
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 189.4631     1.1193  169.28   <2e-16 ***
## PC1         -50.8520     0.7710  -65.96   <2e-16 ***
## PC2          51.2579     0.8296   61.78   <2e-16 ***
## PC3         -13.8972     0.9592  -14.49   <2e-16 ***
## PC4          21.6114     1.0518   20.55   <2e-16 ***
## PC5         -10.0558     1.0675   -9.42   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 147.6 on 17373 degrees of freedom
## Multiple R-squared:  0.3385, Adjusted R-squared:  0.3383
## F-statistic:  1778 on 5 and 17373 DF,  p-value: < 2.2e-16
```

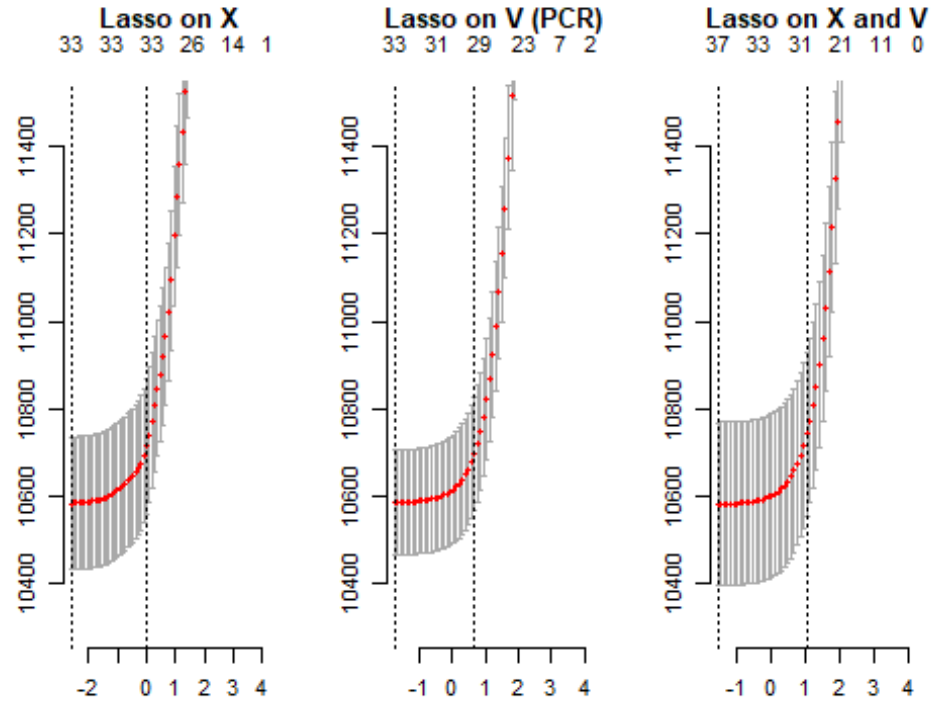
From the regression I find how the PCs correlate to the count of riders and it appears they are all significant.Using the summary of the regression I can make some observation on how mean rider counts depend on the factors found in part 1. A day with one unit of “cooler temperature” (PC1) there will be around -50 mean riders. A day with more “less humidity” there will be around +51 mean riders (PC2). Since PC3 suggests a time that matches the seasonal component of less in the summer, I expect it to reduce the mean riders by 13. For PC 4 I find a unit harshness and windiness of the weather will increase mean rider count by around 21 riders. PC5 suggests a day that is a holiday will result in 10 less mean rider counts.

```r
# estimating the model with both factors

maybe_better_reg <- lm(rider_count ~ yr + holiday + weekday + temp +
                         weathersit + windspeed + humidity +
                         spring + summer + fall +
                           # categorical variables for seasons
                         oneam + twoam + threeam + fouram + fiveam + sixam +
sevenam +
                         eightam + nineam + tenam + elevenam + noon +
                         onepm + twopm + threepm + fourpm + fivepm + sixpm +
sevenpm +
                         eightpm + ninepm + tenpm + elevenpm +
                           # categorical variables for Hours of the Day
                         pca$PC1 + pca$PC2 +pca$PC3 + pca$PC4 + pca$PC5)


# testing if best.regression is the best with an anova test
# Ho: best.regression is the best and the untransformed variables are insigni
ficant
anova(maybe_better_reg, best.regression)
```

```
## Analysis of Variance Table
##
## Model 1: rider_count ~ yr + holiday + weekday + temp + weathersit + windsp
eed +
##     humidity + spring + summer + fall + oneam + twoam + threeam +
##     fouram + fiveam + sixam + sevenam + eightam + nineam + tenam +
##     elevenam + noon + onepm + twopm + threepm + fourpm + fivepm +
##     sixpm + sevenpm + eightpm + ninepm + tenpm + elevenpm + pca$PC1 +
##     pca$PC2 + pca$PC3 + pca$PC4 + pca$PC5
## Model 2: rider_count ~ PC1 + PC2 + PC3 + PC4 + PC5
##   Res.Df       RSS  Df  Sum of Sq      F    Pr(>F)
## 1  17345 183172753
## 2  17373 378236522 -28 -195063768 659.68 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

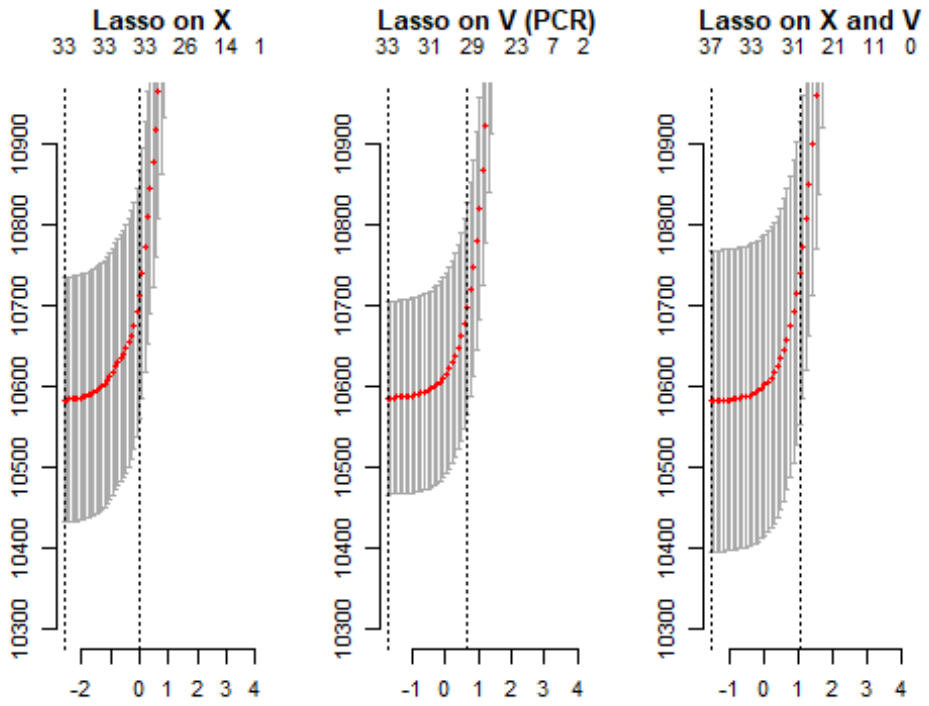
Pvalue < 0.05 suggests I reject the null hypothesis and only using the transforemd regressors is not preferred over the model that includes all transformed and untransformed regressors. Thus I may want to include the untransformed regressors as they are significant to rider counts.

```r
# comparing regressors

xlasso <- cv.glmnet(x = as.matrix(best_data[,-1]), y = rider_count, nfold = 20)
# creating the transformed lasso regression to compare with the factorized and transformed regression

xvlasso <- cv.glmnet(x = as.matrix(cbind(best_data[-1], full.pca)), y = rider_count, nfold = 20)
# creating a lasso that combines the transformed regressors and the transformed


# plotting the relationships
par(mfrow = c(1, 3), mar = c(3.1, 3.1, 5.1, 1.1), mgp = 2:0)
plot(xlasso, main = "Lasso on X", ylim = c(10300, 11500),     # setting a boundary on the y axis for a closer look
     ylab = "", xlab = "", bty = "n")
plot(cvlassoPC, main = "Lasso on V (PCR)", ylim = c(10300, 11500),
     ylab = "", xlab = "", bty = "n")
plot(xvlasso, main = "Lasso on X and V", ylim = c(10300, 11500),
     ylab = "", xlab = "", bty = "n")
mtext(side = 2, "mean squared error", outer = TRUE, line = 2)
mtext(side = 1, "log lamba", outer = TRUE, line = 2)
```



A closer look at the plot

```
# taking a closer look
par(mfrow = c(1, 3), mar = c(3.1, 3.1, 5.1, 1.1), mgp = 2:0)
plot(xlasso, main = "Lasso on X", ylim = c(10300, 10950),     # setting a bou
ndary on the y axis for a closer look
     ylab = "", xlab = "", bty = "n")
plot(cvlassoPC, main = "Lasso on V (PCR)", ylim = c(10300, 10950),
     ylab = "", xlab = "", bty = "n")
plot(xvlasso, main = "Lasso on X and V", ylim = c(10300, 10950),
     ylab = "", xlab = "", bty = "n")
mtext(side = 2, "mean squared error", outer = TRUE, line = 2)
mtext(side = 1, "log lamba", outer = TRUE, line = 2)
```



The graphs above shows the relationship between a regression with only untransformed regressors, X, a regression with only transfomed regressors/ PCs, V and a regression with both regressors combined. I can see that the combined model have less variation and error than both the untransformed an transformed model. This suggests that the better model should include both transformed and untransformed regressors.

I know that from the f-test I did that the untransformed regressors are significant to rider counts.

```r
# incorporating the seasonal components and taking out insignificant variable
s
maybe_better_reg1 <- lm (rider_count ~ yr +
                           spring + summer + fall +
                             # categorical variables for seasons
                           oneam + twoam + threeam + fouram + fiveam + sixam +
sevenam +
                           eightam + nineam + tenam + elevenam + noon +
                           onepm + twopm + threepm + fourpm + fivepm + sixpm +
sevenpm +
                           eightpm + ninepm + tenpm + elevenpm +
                             # categorical variables for Hours of the Day
                           pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 )

# Ho: maybe_better_reg is the best and the variables removed are significant
anova(maybe_better_reg1, maybe_better_reg)
```

```
## Analysis of Variance Table
##
## Model 1: rider_count ~ yr + spring + summer + fall + oneam + twoam + three
am +
##     fouram + fiveam + sixam + sevenam + eightam + nineam + tenam +
##     elevenam + noon + onepm + twopm + threepm + fourpm + fivepm +
##     sixpm + sevenpm + eightpm + ninepm + tenpm + elevenpm + pca$PC1 +
##     pca$PC2 + pca$PC3 + pca$PC4
## Model 2: rider_count ~ yr + holiday + weekday + temp + weathersit + windsp
eed +
##     humidity + spring + summer + fall + oneam + twoam + threeam +
##     fouram + fiveam + sixam + sevenam + eightam + nineam + tenam +
##     elevenam + noon + onepm + twopm + threepm + fourpm + fivepm +
##     sixpm + sevenpm + eightpm + ninepm + tenpm + elevenpm + pca$PC1 +
##     pca$PC2 + pca$PC3 + pca$PC4 + pca$PC5
##   Res.Df       RSS Df Sum of Sq      F Pr(>F)
## 1  17347 183198938
## 2  17345 183172753  2     26185 1.2397 0.2895
```

```r
summary(maybe_better_reg1) # summary of the better model
```

```
##
## Call:
## lm(formula = rider_count ~ yr + spring + summer + fall + oneam +
##     twoam + threeam + fouram + fiveam + sixam + sevenam + eightam +
##     nineam + tenam + elevenam + noon + onepm + twopm + threepm +
##     fourpm + fivepm + sixpm + sevenpm + eightpm + ninepm + tenpm +
##     elevenpm + pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -413.98  -60.33   -6.69   50.11  471.98
```

```
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   59.2448     4.2719  13.868  < 2e-16 ***
## yr            83.6894     1.7156  48.782  < 2e-16 ***
## spring       -37.4193     9.3256  -4.013 6.03e-05 ***
## summer       107.2337    19.3884   5.531 3.23e-08 ***
## fall        -212.4079    11.6565 -18.222  < 2e-16 ***
## oneam          6.2301     5.7351   1.086  0.27736
## twoam         -0.5598     5.7993  -0.097  0.92310
## threeam      -11.8503     5.7747  -2.052  0.04018 *
## fouram       -15.9994     5.6659  -2.824  0.00475 **
## fiveam         5.3644     5.6774   0.945  0.34474
## sixam         57.5482     5.5181  10.429  < 2e-16 ***
## sevenam      182.5964     5.7233  31.904  < 2e-16 ***
## eightam      317.6485     5.6992  55.736  < 2e-16 ***
## nineam       159.1859     5.9269  26.858  < 2e-16 ***
## tenam         98.8947     5.7282  17.264  < 2e-16 ***
## elevenam     120.3881     5.6019  21.490  < 2e-16 ***
## noon         150.1021     5.8446  25.682  < 2e-16 ***
## onepm        143.0740     5.7562  24.856  < 2e-16 ***
## twopm        122.1609     5.8958  20.720  < 2e-16 ***
## threepm      130.5605     5.8980  22.137  < 2e-16 ***
## fourpm       195.7791     5.7987  33.762  < 2e-16 ***
## fivepm       351.1521     5.7733  60.823  < 2e-16 ***
## sixpm        326.5561     5.6346  57.956  < 2e-16 ***
## sevenpm      231.0659     5.5336  41.757  < 2e-16 ***
## eightpm      160.0026     5.5982  28.581  < 2e-16 ***
## ninepm       119.6491     5.7818  20.694  < 2e-16 ***
## tenpm         86.6937     5.7864  14.982  < 2e-16 ***
## elevenpm      47.0280     5.5555   8.465  < 2e-16 ***
## pca$PC1      -78.4506     2.0349 -38.552  < 2e-16 ***
## pca$PC2       12.4127     1.1171  11.111  < 2e-16 ***
## pca$PC3       73.9462     9.7240   7.605 3.01e-14 ***
## pca$PC4       22.2111     3.4926   6.359 2.08e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.8 on 17347 degrees of freedom
## Multiple R-squared:  0.6796, Adjusted R-squared:  0.679
## F-statistic:  1187 on 31 and 17347 DF,  p-value: < 2.2e-16
```

With the anova test, maybe_better_reg1 is preferred, I found that the seasonal and hourly component I included are very significant to the data. PC5 are not significant. Although the summary of the model shows that there are certain hours that are insignificant to the mean rider count, they have to be included because they are categorical variables. The times that appear to be less significant are between 1am- 5am. This issue may be a result of inconclusive data during those times or not enough data. There could also be other interactions between time and season that could be causing this.

The comparison between the new fit and one in part 2 suggests that the new model performs better than the previous estimator. The new model has lower residual standard error and higher adjusted r-squared. When compared to the regression I ran on all the untransformed data, I found the coefficients made more sense and were interpretable.

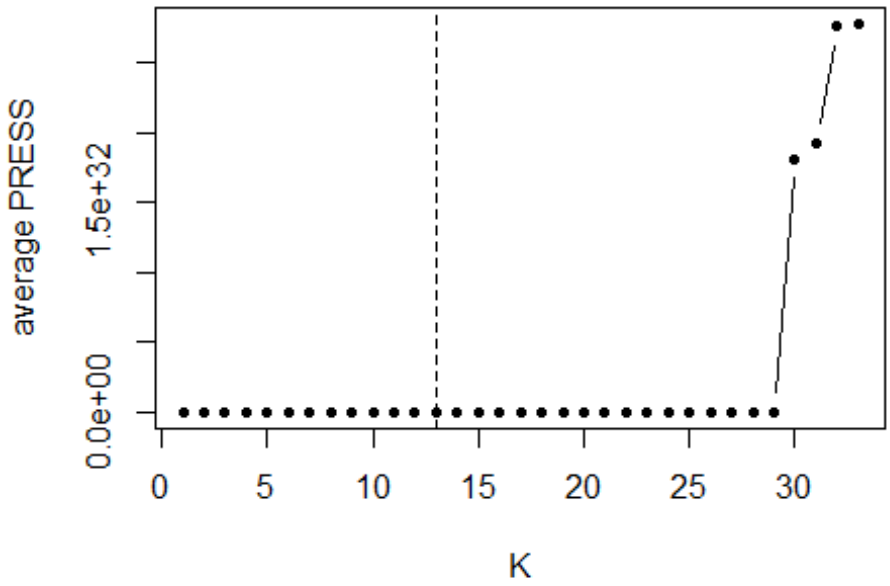
```
# partial least squares [PLS]
plsreg<- plsr(rider_count ~., data=best_data, method = "oscorespls", validati
on = "CV", segments = 6)
summary(plsreg)

## Data:    X dimension: 17379 33
##  Y dimension: 17379 1
## Fit method: oscorespls
## Number of components considered: 33
##
## VALIDATION: RMSEP
## Cross-validated using 6 random segments.

## Warning in sqrt(z$val): NaNs produced

##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           181.4    153.6    150.9    136.9    128.7      113    107.7
## adjCV        181.4    153.6    150.9    136.9    128.7      113    107.7
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       106.9    105.6    105.3       105     104.3     102.9     102.9
## adjCV    106.9    105.5    105.2       105     104.0     102.9     102.9
##        14 comps  15 comps  16 comps  17 comps  18 comps  19 comps  20 comp
s
## CV        102.9     102.9     102.9     102.9     102.9     102.9     102.
9
## adjCV     102.9     102.9     102.9     102.9     102.9     102.9     102.
9
##        21 comps  22 comps  23 comps  24 comps  25 comps  26 comps  27 comp
s
## CV        102.9     102.9     102.9     102.9     102.9     102.9      177
8
## adjCV     102.9     102.9     102.9     102.9     102.9     102.9       Na
N
##        28 comps   29 comps   30 comps   31 comps   32 comps   33 comps
## CV     28246547  2.657e+11  1.022e+14  1.054e+14  1.260e+14  1.265e+14
## adjCV       NaN        NaN  5.131e+12  3.638e+12  1.054e+13  1.038e+13
##
## TRAINING: % variance explained
##               1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X               14.81    55.06    62.73    72.88    76.73    83.10    85.82
## rider_count     28.40    30.89    43.14    49.77    61.34    64.89    65.40
##               8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14
comps
## X               86.42    87.76     89.58     90.18     90.64     91.00
```

```
91.38
## rider_count    66.28    66.47     66.61     67.27     67.96     67.96
67.96
##               15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
## X                91.83     92.28     92.73     93.17     93.62     94.08
## rider_count      67.96     67.96     67.96     67.96     67.96     67.96
##               21 comps  22 comps  23 comps  24 comps  25 comps  26 comps
## X                94.53     94.99     95.45     95.90     96.36     96.81
## rider_count      67.96     67.96     67.96     67.96     67.96     67.96
##               27 comps  28 comps  29 comps  30 comps  31 comps  32 comps
## X                97.27     97.72     98.18     98.19     98.63     99.09
## rider_count      67.96     67.96     67.96     67.96     67.96     67.96
##               33 comps
## X                99.54
## rider_count      67.96
```

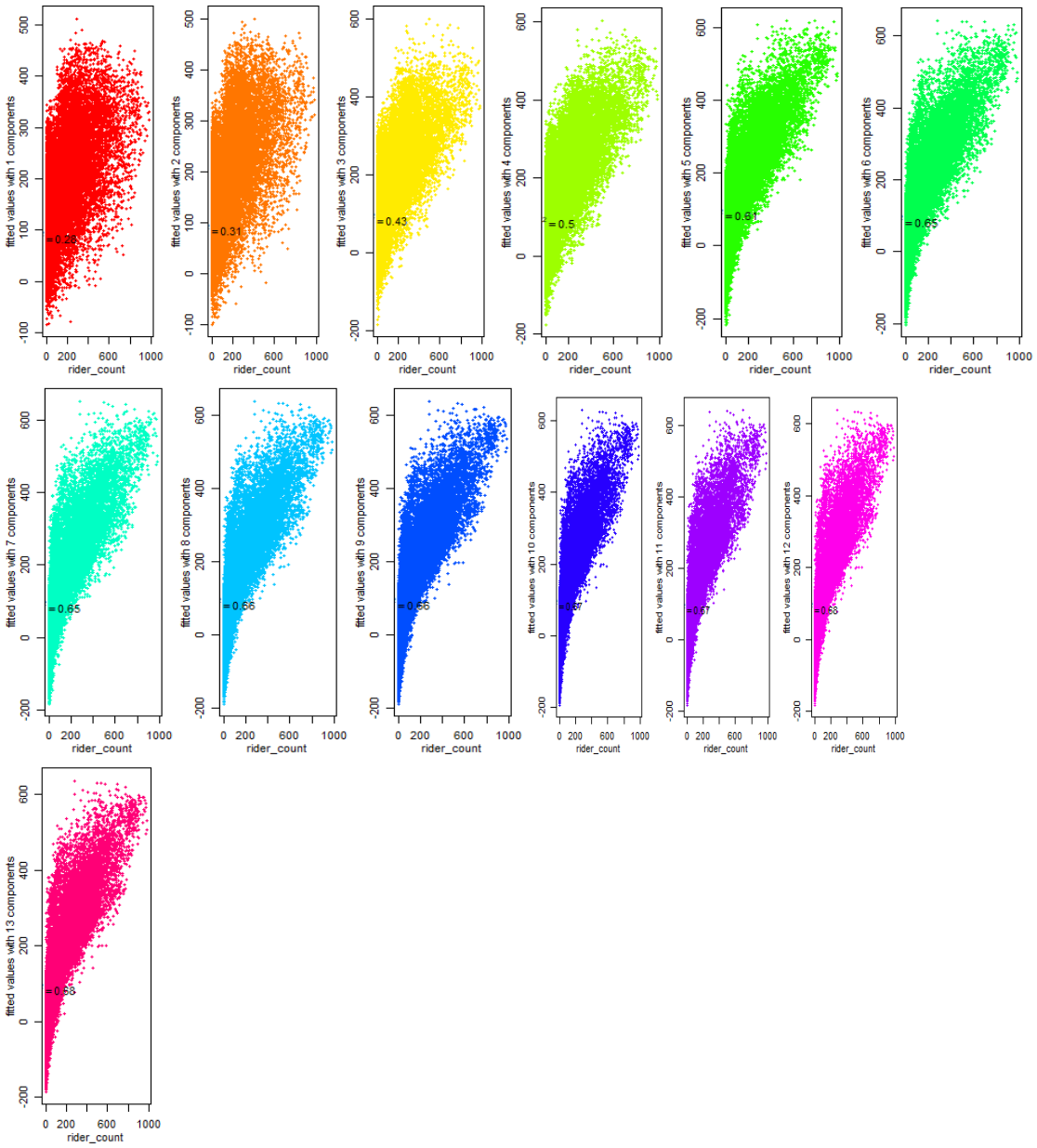
```
# Finding the Min predicted residual error sum of squares [PRESS]
min_AVGPRESS = which.min(plsreg$validation$PRESS)
min_AVGPRESS
```

```
## [1] 13
```

```
# PLS CV plot average PRESS
plot(1:plsreg$ncomp, plsreg$validation$PRESS, type = "b", pch = 20,
     xlab = "K", ylab = "average PRESS",)
abline(v = which.min(plsreg$validation$PRESS), lty = 2)
```



```r
# graphing the fitted values to 13
pal <- rainbow(13)

par(mfrow = c(1, 3), mar = c(3.1, 3.1, 1.1, 1.1), mgp = 2:0)
for (k in 1:13) {
  r2 <- cor(rider_count, plsreg$fitted[, , k])^2

  plot(rider_count, plsreg$fitted[, , k], pch = 20, col = pal[k],
       ylab = paste("fitted values with", k, "components"))
  text(84, 88.5, bquote(R^2 == .(round(r2, 2))))}
```



After plotting the partial least square I find at the minimum average PRESS, R^2 is 0.68 which is very close to the revised model in part 3.

```r
# using partial least squares to model the data
pls.fit = plsr(rider_count~., data=best_data, scale=TRUE, ncomp=33)
# The lowest cross-validation error occurs when M=33 PLS directions are used
summary(pls.fit)
```

```
## Data:    X dimension: 17379 33
##  Y dimension: 17379 1
## Fit method: kernelpls
## Number of components considered: 33
## TRAINING: % variance explained
##              1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps
## X              5.411    9.318    13.84    16.49    18.54    21.10    22.86
## rider_count   57.912   65.186    65.90    66.51    67.01    67.39    67.76
##              8 comps  9 comps  10 comps  11 comps  12 comps  13 comps  14
comps
## X              25.17    26.63     29.43     30.81     33.92     36.93
39.92
## rider_count    67.87    67.96     67.96     67.96     67.96     67.96
67.96
##              15 comps  16 comps  17 comps  18 comps  19 comps  20 comps
## X               43.08     46.24     49.41     52.56     55.72     58.88
## rider_count     67.96     67.96     67.96     67.96     67.96     67.96
##              21 comps  22 comps  23 comps  24 comps  25 comps  26 comps
## X               62.04     65.20     68.37     71.53     74.69     77.85
## rider_count     67.96     67.96     67.96     67.96     67.96     67.96
##              27 comps  28 comps  29 comps  30 comps  31 comps  32 comps
## X               81.01     84.17     87.33     90.49     93.65     96.81
## rider_count     67.96     67.96     67.96     67.96     67.96     67.96
##              33 comps
## X               99.97
## rider_count     67.96
```

```r
# looking at the rotations from Comp 1 to 9
pls.fit$projection[,1:9]
```

```
##                Comp 1      Comp 2      Comp 3      Comp 4       Comp 5
## yr          0.25736880  0.32858540  0.08296356 -0.02936531 -0.060098210
## holiday    -0.03177599 -0.01065242 -0.02615709  0.01057138 -0.024247996
## weekday     0.02763803  0.01814281  0.02672875  0.04816401 -0.002639878
## temp        0.41587975 -0.19650959  0.29701039  0.14020778  0.344366424
## weathersit -0.14633450  0.17430310 -0.08359622 -0.37722682 -0.085444720
## windspeed   0.09579224 -0.22462921 -0.53917690 -0.04149940  0.426240735
## humidity   -0.33177183  0.18100614  0.54161955  0.22126385  0.304620813
## spring     -0.25219121  0.11554792 -0.35905451 -0.07263779 -0.167571414
## summer      0.06235719 -0.08558971 -0.32061361 -0.43441068 -0.146220172
## fall        0.15578141 -0.41342566 -0.13648216 -0.62108647 -0.410064561
## oneam      -0.18434303 -0.16354712  0.03907854  0.10604195  0.226066671
## twoam      -0.19547106 -0.16934587  0.03180629  0.08912542  0.238065382
## threeam    -0.20579219 -0.17907354  0.02694294  0.07382885  0.247867498
## fouram     -0.21201504 -0.17709447  0.02736953  0.07093423  0.244707515
```

```
## fiveam     -0.19925780 -0.13645484  0.02642368  0.05426262  0.252520511
## sixam      -0.13404705 -0.02949102  0.03409357  0.02914286  0.265068328
## sevenam     0.02675065  0.19799069  0.04119884 -0.03950226  0.303995188
## eightam     0.20067274  0.41829345  0.05833907 -0.08431741  0.314311861
## nineam      0.03532547  0.11698306  0.06844738  0.08610478  0.255105173
## tenam      -0.01869408 -0.02453232  0.08069486  0.19064025  0.212712093
## elevenam    0.02210915 -0.01881675  0.08913298  0.22720242  0.204191019
## noon        0.07562889  0.01929238  0.09784152  0.25063502  0.206008943
## onepm       0.07609230 -0.01574527  0.10543553  0.29012351  0.185416643
## twopm       0.06102518 -0.06303253  0.11631249  0.33707849  0.158405242
## threepm     0.07321447 -0.05321175  0.11990592  0.34105154  0.152683609
## fourpm      0.14532420  0.05788951  0.12610344  0.30140654  0.154335777
## fivepm      0.32261203  0.34822150  0.12445613  0.16977404  0.206494395
## sixpm       0.27958096  0.31521118  0.11396475  0.15190221  0.214432135
## sevenpm     0.14457119  0.15689951  0.09963941  0.16333331  0.198032706
## eightpm     0.04331101  0.04511611  0.08345799  0.16104020  0.200677808
## ninepm     -0.02031114 -0.01203977  0.06871525  0.13658710  0.208853059
## tenpm      -0.06884815 -0.05690552  0.06108476  0.12857355  0.210978451
## elevenpm   -0.12037550 -0.10375434  0.04657888  0.11433416  0.233397729
##                 Comp 6       Comp 7        Comp 8       Comp 9
## yr           0.12327541 -0.069828840 -0.2982867082  0.299637244
## holiday      0.01814837 -0.018863948  0.0335985311  0.004463817
## weekday     -0.01285101 -0.004483397 -0.0008867234  0.005114580
## temp         0.28413043  0.203014868 -0.4663004452 -0.826325423
## weathersit  -0.03861221  0.313653348  0.0962118334 -0.342550794
## windspeed   -0.03257530 -0.326888350  0.0233559227  0.047461728
## humidity    -0.16715490 -0.311840988  0.0892194259  0.202956594
## spring       0.25353274  0.121527246 -0.0696701207 -0.283373726
## summer      -0.33094603  0.334652657  0.0136225876  0.379758054
## fall        -0.02299635 -0.155741188  0.4695432334  0.539934179
## oneam        0.35973385  0.408862202  0.1269199176  0.260390288
## twoam        0.35942208  0.406029880  0.1335155115  0.242549755
## threeam      0.35148918  0.404213837  0.1419303223  0.220463046
## fouram       0.34226018  0.420549692  0.1702145333  0.181155663
## fiveam       0.37690202  0.410024579  0.1298508592  0.212335691
## sixam        0.37500853  0.407555568  0.1573112972  0.187979611
## sevenam      0.38257116  0.363241124  0.2061761618  0.179780758
## eightam      0.38796902  0.329512470  0.2420131453  0.196149469
## nineam       0.30445395  0.387600152  0.2642530109  0.190954137
## tenam        0.26213605  0.407031040  0.2665883455  0.224658024
## elevenam     0.24965874  0.385116738  0.2702976461  0.272746108
## noon         0.22679225  0.368259848  0.3024422896  0.286871970
## onepm        0.21305846  0.371075668  0.3018822873  0.313833428
## twopm        0.18433456  0.394665560  0.3194248595  0.307649873
## threepm      0.18182249  0.396100161  0.3219112683  0.311522981
## fourpm       0.20488431  0.387608165  0.3137463706  0.318061497
## fivepm       0.25447833  0.335193804  0.3291501496  0.310976806
## sixpm        0.27033466  0.338842708  0.3076585274  0.302416698
## sevenpm      0.29152175  0.368015130  0.2475791010  0.306087544
## eightpm      0.31134184  0.378261098  0.2040010495  0.307559576
```

```
## ninepm      0.33741702  0.379658639  0.1662032279  0.311595610
## tenpm       0.34587610  0.390333740  0.1511941333  0.298561075
## elevenpm    0.34790587  0.385965379  0.1524020548  0.279589773

# coefficients
# pls.fit$coefficients
```

I notice that after the 9-component PLS fit, the percentage of variance in rider counts are all 67.96. I get that the determinants of rider counts includes up to 9comps. From looking at the rotations up to 9comps, I see that the values are closer in values compared to the rotations I got from doing the PC analysis, suggesting that each component in this pls is using alot more attributes from the original data. This may result in over fitting. Although the PLS gives a clearer picture of which component to choose to create a model as competitive as the modified "better model", this information doesn't help us explain and interpret the model. I found the rotations created by this method is not as clear. When looking at the coefficients I find that pls does not help us understand the underlying relationships between variables.

```
# observing the humidity effect on rider counts

humidregpc<-lm(rider_count~ pca$PC2)
# PC2 addresses the humidity and windy effect on rider_count
coef(humidregpc)

## (Intercept)     pca$PC2
##   189.46309    51.25794

humidreg <- lm(rider_count~ humidity)
coef(humidreg)

## (Intercept)    humidity
##    189.4631    -58.5720

# testing interactions between humidity and other regressors

humidregtemp <- lm(rider_count~ temp + humidity + temp:humidity) # interactio
ns with temp
summary(humidregtemp)

##
## Call:
## lm(formula = rider_count ~ temp + humidity + temp:humidity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -398.34  -99.29  -36.21   64.38  708.54
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)    188.031      1.185  158.66   <2e-16 ***
## temp            64.170      1.232   52.09   <2e-16 ***
## humidity       -54.220      1.185  -45.75   <2e-16 ***
## temp:humidity  -20.492      1.257  -16.30   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 155.8 on 17375 degrees of freedom
## Multiple R-squared:  0.2623, Adjusted R-squared:  0.2622
## F-statistic:  2060 on 3 and 17375 DF,  p-value: < 2.2e-16
```

When looking at the interaction between humidity and temperature I find that given an increase in temperature, an increase in humidity will decrease mean rider count by 20 riders.

Using the best model estimator from combining transformed and untransformed regressors, I included interaction terms to see the causual effect humidity has on riders given other factors:

```
# testing causual effect of humidity on rider counts given other factors from
the estimator

# maybe_better_reg1 with interactions on hourly componenet
humidreg_hrly<-lm (rider_count ~ yr +
                                 spring + summer + fall +
                                 # categorical variables for seasons
                                 oneam + twoam + threeam + fouram + fiveam + s
ixam + sevenam +
                                 eightam + nineam + tenam + elevenam + noon +
                                 onepm + twopm + threepm + fourpm + fivepm + s
ixpm + sevenpm +
                                 eightpm + ninepm + tenpm + elevenpm +
                                 # time component
                                 pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 +
                                 # transformed component

                                 # interaction variables
                                 + humidity +

                                 oneam:humidity + twoam:humidity + threeam:hum
idity +
                                 fouram:humidity + fiveam:humidity + sixam:hum
idity +
                                 sevenam: humidity + eightam:humidity + nineam
:humidity +
                                 tenam:humidity + elevenam:humidity + noon:hum
idity +
                                 onepm:humidity + twopm:humidity + threepm:hum
```

```
idity + fourpm:humidity +
                                fivepm:humidity + sixpm:humidity + sevenpm:hu
midity +
                                eightpm:humidity + ninepm:humidity + tenpm:hu
midity + elevenpm:humidity )

summary(humidreg_hrly)

##
## Call:
## lm(formula = rider_count ~ yr + spring + summer + fall + oneam +
##     twoam + threeam + fouram + fiveam + sixam + sevenam + eightam +
##     nineam + tenam + elevenam + noon + onepm + twopm + threepm +
##     fourpm + fivepm + sixpm + sevenpm + eightpm + ninepm + tenpm +
##     elevenpm + pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 + +humidity +
##     oneam:humidity + twoam:humidity + threeam:humidity + fouram:humidity +
##     fiveam:humidity + sixam:humidity + sevenam:humidity + eightam:humidity
+
##     nineam:humidity + tenam:humidity + elevenam:humidity + noon:humidity +
##     onepm:humidity + twopm:humidity + threepm:humidity + fourpm:humidity +
##     fivepm:humidity + sixpm:humidity + sevenpm:humidity + eightpm:humidity
+
##     ninepm:humidity + tenpm:humidity + elevenpm:humidity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -436.79  -59.93   -6.59   49.92  460.89
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         56.1335     4.5161  12.430  < 2e-16 ***
## yr                  83.8480     1.7159  48.866  < 2e-16 ***
## spring             -42.6267    10.6203  -4.014 6.00e-05 ***
## summer             111.4879    22.0491   5.056 4.32e-07 ***
## fall              -209.0459    12.8566 -16.260  < 2e-16 ***
## oneam                6.3847     6.3297   1.009  0.31314
## twoam                0.1153     6.5311   0.018  0.98591
## threeam            -10.8400     6.6949  -1.619  0.10543
## fouram             -15.6449     6.7022  -2.334  0.01959 *
## fiveam               7.1817     6.6895   1.074  0.28303
## sixam               55.7883     6.4195   8.691  < 2e-16 ***
## sevenam            179.7452     6.4101  28.041  < 2e-16 ***
## eightam            317.4581     6.1411  51.694  < 2e-16 ***
## nineam             159.1461     6.1586  25.841  < 2e-16 ***
## tenam              101.3167     5.9172  17.122  < 2e-16 ***
## elevenam           124.1835     5.8967  21.060  < 2e-16 ***
## noon               151.3447     6.3329  23.898  < 2e-16 ***
## onepm              143.4004     6.4026  22.397  < 2e-16 ***
## twopm              122.5293     6.6417  18.448  < 2e-16 ***
## threepm            132.4092     6.6511  19.908  < 2e-16 ***
```

```
## fourpm              192.5949     6.4981  29.639  < 2e-16 ***
## fivepm              337.4119     6.3525  53.115  < 2e-16 ***
## sixpm               320.0307     6.0691  52.731  < 2e-16 ***
## sevenpm             231.1165     5.7986  39.857  < 2e-16 ***
## eightpm             162.9338     5.8187  28.002  < 2e-16 ***
## ninepm              122.3072     6.0408  20.247  < 2e-16 ***
## tenpm                88.1370     6.0938  14.463  < 2e-16 ***
## elevenpm             47.1047     5.9187   7.959 1.85e-15 ***
## pca$PC1             -76.0689     2.0565 -36.990  < 2e-16 ***
## pca$PC2              13.6071     2.3140   5.880 4.17e-09 ***
## pca$PC3              75.5392    11.0964   6.808 1.03e-11 ***
## pca$PC4              22.8400     3.8197   5.980 2.28e-09 ***
## humidity              7.3668     4.8259   1.527  0.12690
## oneam:humidity       -0.4117     6.3356  -0.065  0.94819
## twoam:humidity       -2.3029     6.3725  -0.361  0.71782
## threeam:humidity     -3.3569     6.5686  -0.511  0.60932
## fouram:humidity      -2.9490     6.6392  -0.444  0.65691
## fiveam:humidity      -5.6686     6.5701  -0.863  0.38826
## sixam:humidity       -0.4972     6.4727  -0.077  0.93877
## sevenam:humidity      1.8618     6.4695   0.288  0.77352
## eightam:humidity     -1.7081     6.4511  -0.265  0.79118
## nineam:humidity       3.0386     6.3081   0.482  0.63003
## tenam:humidity        1.7904     6.2196   0.288  0.77345
## elevenam:humidity    -1.2291     6.1451  -0.200  0.84147
## noon:humidity        -8.6954     6.0982  -1.426  0.15392
## onepm:humidity      -10.1446     6.0678  -1.672  0.09457 .
## twopm:humidity       -9.6363     6.0107  -1.603  0.10891
## threepm:humidity     -7.4654     5.9491  -1.255  0.20954
## fourpm:humidity     -14.9073     5.9002  -2.527  0.01153 *
## fivepm:humidity     -33.8669     5.8623  -5.777 7.73e-09 ***
## sixpm:humidity      -25.3680     5.9039  -4.297 1.74e-05 ***
## sevenpm:humidity    -15.4442     5.9557  -2.593  0.00952 **
## eightpm:humidity     -4.6660     6.0173  -0.775  0.43810
## ninepm:humidity       0.4830     6.1086   0.079  0.93698
## tenpm:humidity        3.2767     6.1406   0.534  0.59361
## elevenpm:humidity     4.1086     6.2180   0.661  0.50878
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.5 on 17323 degrees of freedom
## Multiple R-squared:  0.6818, Adjusted R-squared:  0.6808
## F-statistic: 674.9 on 55 and 17323 DF,  p-value: < 2.2e-16
```

By looking at the coefficients of the interaction variables, I can see a causual effect of humidity on rider counts given the hour: The coefficients tell us that at 12am, a unit of humidity would increase the number of mean rider count by 7 riders. During the times 1am - 6am, 8am, 11am - 8pm, the effects humidity has on mean rider count would decrease relative to 7 riders. During the times 7am, 9am, 10am, 9pm-11pm, the effects of humidity on rider count will increase relative to 7 riders.

```r
# maybe_better_reg1 with interactions on seasons
humidregseason<-lm (rider_count ~ yr +
                                spring + summer + fall +
                                # categorical variables for seasons
                                oneam + twoam + threeam + fouram + fiveam + s
ixam + sevenam +
                                eightam + nineam + tenam + elevenam + noon +
                                onepm + twopm + threepm + fourpm + fivepm + s
ixpm + sevenpm +
                                eightpm + ninepm + tenpm + elevenpm +
                                # time component
                                pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 +
                                # transformed component

                                # interaction variables
                                + humidity +

                                spring:humidity + summer:humidity + fall: hum
idity )

summary(humidregseason)
```

```
##
## Call:
## lm(formula = rider_count ~ yr + spring + summer + fall + oneam +
##     twoam + threeam + fouram + fiveam + sixam + sevenam + eightam +
##     nineam + tenam + elevenam + noon + onepm + twopm + threepm +
##     fourpm + fivepm + sixpm + sevenpm + eightpm + ninepm + tenpm +
##     elevenpm + pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 + +humidity +
##     spring:humidity + summer:humidity + fall:humidity)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -427.94  -59.83   -6.00   50.48  464.17
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        63.6145     4.2850  14.846  < 2e-16 ***
## yr                 81.3705     1.7154  47.435  < 2e-16 ***
## spring            -31.3801    10.6300  -2.952 0.003161 **
## summer             90.9014    22.0590   4.121 3.79e-05 ***
## fall             -203.1717    12.8050 -15.867  < 2e-16 ***
## oneam               5.6749     5.8629   0.968 0.333095
## twoam              -0.6488     5.9727  -0.109 0.913500
## threeam           -11.4940     5.9768  -1.923 0.054485 .
## fouram            -15.1887     5.8145  -2.612 0.009003 **
## fiveam              6.0233     5.8123   1.036 0.300079
## sixam              58.9724     5.5310  10.662  < 2e-16 ***
## sevenam           184.3764     5.6888  32.410  < 2e-16 ***
## eightam           318.1175     5.6629  56.176  < 2e-16 ***
```

```
## nineam            159.0360     5.9144  26.890  < 2e-16 ***
## tenam              97.4475     5.7184  17.041  < 2e-16 ***
## elevenam          118.0401     5.5849  21.136  < 2e-16 ***
## noon              147.7659     5.8654  25.193  < 2e-16 ***
## onepm             139.9333     5.7641  24.277  < 2e-16 ***
## twopm             119.0427     5.9342  20.061  < 2e-16 ***
## threepm           127.5275     5.9358  21.484  < 2e-16 ***
## fourpm            192.3368     5.8183  33.057  < 2e-16 ***
## fivepm            348.4141     5.7963  60.110  < 2e-16 ***
## sixpm             324.0448     5.6239  57.619  < 2e-16 ***
## sevenpm           228.4608     5.5008  41.532  < 2e-16 ***
## eightpm           157.4903     5.5961  28.143  < 2e-16 ***
## ninepm            117.6242     5.8426  20.132  < 2e-16 ***
## tenpm              85.0070     5.8689  14.484  < 2e-16 ***
## elevenpm           46.6104     5.6202   8.293  < 2e-16 ***
## pca$PC1           -77.2113     2.0587 -37.504  < 2e-16 ***
## pca$PC2            13.6879     2.3178   5.906 3.58e-09 ***
## pca$PC3            66.4777    11.0831   5.998 2.04e-09 ***
## pca$PC4            19.5695     3.8120   5.134 2.87e-07 ***
## humidity           -9.0531     2.7204  -3.328 0.000877 ***
## spring:humidity    29.9870     2.3004  13.035  < 2e-16 ***
## summer:humidity     3.8721     2.2486   1.722 0.085085 .
## fall:humidity       2.6888     2.4244   1.109 0.267429
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.1 on 17343 degrees of freedom
## Multiple R-squared:  0.684,  Adjusted R-squared:  0.6833
## F-statistic:  1072 on 35 and 17343 DF,  p-value: < 2.2e-16
```

Humidity on rider counts given the season looks more intuiative. From looking at the coefficients of the interaction variables, I see that at winter, the effects of humidity on mean rider counts is -9 riders per unit humidity. During the spring time, the effects of humidity on mean rider count increases by 29 so instead of -9, the mean rider count is expected to rise by 20 riders for every increase of unit humidiy. The interaction variables suggests that a unit rise in humidity will result in less mean rider counts for all seasons but spring.

```
# maybe_better_reg1 with interactions on PC1
humidregPC1 <-lm (rider_count ~ yr +  # maybe_better_reg1 interaction model
                                spring + summer + fall +
                                # categorical variables for seasons
                                oneam + twoam + threeam + fouram + fiveam + s
ixam + sevenam +
                                eightam + nineam + tenam + elevenam + noon +
                                onepm + twopm + threepm + fourpm + fivepm + s
ixpm + sevenpm +
                                eightpm + ninepm + tenpm + elevenpm +
                                # time component
                                pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 +
                                # transformed component
```

```
                              # interaction variables
                              humidity:pca$PC1 )

summary(humidregPC1)
```

```
##
## Call:
## lm(formula = rider_count ~ yr + spring + summer + fall + oneam +
##     twoam + threeam + fouram + fiveam + sixam + sevenam + eightam +
##     nineam + tenam + elevenam + noon + onepm + twopm + threepm +
##     fourpm + fivepm + sixpm + sevenpm + eightpm + ninepm + tenpm +
##     elevenpm + pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 + humidity:pca$PC1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -413.91  -59.04   -4.95   49.80  457.88
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         59.3213     4.2524  13.950  < 2e-16 ***
## yr                  82.3915     1.7108  48.160  < 2e-16 ***
## spring             -40.3397     9.2859  -4.344 1.41e-05 ***
## summer             102.2245    19.3039   5.296 1.20e-07 ***
## fall              -198.8898    11.6522 -17.069  < 2e-16 ***
## oneam                4.5565     5.7104   0.798  0.42493
## twoam               -1.9948     5.7739  -0.345  0.72973
## threeam            -12.6898     5.7487  -2.207  0.02730 *
## fouram             -16.9444     5.6405  -3.004  0.00267 **
## fiveam               3.6179     5.6531   0.640  0.52220
## sixam               56.2694     5.4938  10.242  < 2e-16 ***
## sevenam            181.8756     5.6974  31.923  < 2e-16 ***
## eightam            316.7910     5.6736  55.836  < 2e-16 ***
## nineam             158.8893     5.8999  26.931  < 2e-16 ***
## tenam               97.7450     5.7027  17.140  < 2e-16 ***
## elevenam           118.2532     5.5789  21.197  < 2e-16 ***
## noon               147.5604     5.8214  25.348  < 2e-16 ***
## onepm              139.1989     5.7380  24.259  < 2e-16 ***
## twopm              117.8235     5.8789  20.042  < 2e-16 ***
## threepm            125.9196     5.8824  21.406  < 2e-16 ***
## fourpm             190.7957     5.7856  32.978  < 2e-16 ***
## fivepm             347.0672     5.7560  60.297  < 2e-16 ***
## sixpm              323.3567     5.6145  57.593  < 2e-16 ***
## sevenpm            228.3384     5.5125  41.422  < 2e-16 ***
## eightpm            157.5695     5.5760  28.259  < 2e-16 ***
## ninepm             117.5697     5.7577  20.420  < 2e-16 ***
## tenpm               84.7551     5.7620  14.709  < 2e-16 ***
## elevenpm            45.8953     5.5309   8.298  < 2e-16 ***
## pca$PC1            -73.7985     2.0586 -35.849  < 2e-16 ***
## pca$PC2             14.2731     1.1217  12.725  < 2e-16 ***
```

```
## pca$PC3            70.7657     9.6828   7.308 2.82e-13 ***
## pca$PC4            19.9007     3.4815   5.716 1.11e-08 ***
## pca$PC1:humidity    7.1916     0.5674  12.675  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.3 on 17346 degrees of freedom
## Multiple R-squared:  0.6825, Adjusted R-squared:  0.6819
## F-statistic:  1165 on 32 and 17346 DF,  p-value: < 2.2e-16
```

Here I see for every unit increase in "cooler weather", the effects of humidity on mean rider counts is 7 more mean riders for each unit increase of humidity.

```
# maybe_better_reg1 with interactions on PC2
humidregPC2 <-lm (rider_count ~ yr +  # maybe_better_reg1 interaction model
                                spring + summer + fall +
                                # categorical variables for seasons
                                oneam + twoam + threeam + fouram + fiveam + s
ixam + sevenam +
                                eightam + nineam + tenam + elevenam + noon +
                                onepm + twopm + threepm + fourpm + fivepm + s
ixpm + sevenpm +
                                eightpm + ninepm + tenpm + elevenpm +
                                # time component
                                pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 +
                                # transformed component

                                # interaction variables
                                humidity:pca$PC2 )

summary(humidregPC2)
```

```
##
## Call:
## lm(formula = rider_count ~ yr + spring + summer + fall + oneam +
##     twoam + threeam + fouram + fiveam + sixam + sevenam + eightam +
##     nineam + tenam + elevenam + noon + onepm + twopm + threepm +
##     fourpm + fivepm + sixpm + sevenpm + eightpm + ninepm + tenpm +
##     elevenpm + pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 + humidity:pca$PC2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -411.41  -60.02   -6.75   50.36  475.23
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        61.3216     4.3039  14.248  < 2e-16 ***
## yr                 83.7111     1.7149  48.814  < 2e-16 ***
## spring            -36.3311     9.3261  -3.896 9.83e-05 ***
## summer            106.4794    19.3816   5.494 3.99e-08 ***
```

```
## fall              -212.8089    11.6523 -18.263  < 2e-16 ***
## oneam                7.4094     5.7409   1.291 0.196850
## twoam                0.9278     5.8098   0.160 0.873119
## threeam            -10.1883     5.7884  -1.760 0.078406 .
## fouram             -14.1731     5.6833  -2.494 0.012647 *
## fiveam               7.4216     5.7000   1.302 0.192924
## sixam               59.3268     5.5350  10.718  < 2e-16 ***
## sevenam            183.6773     5.7278  32.068  < 2e-16 ***
## eightam            317.8590     5.6972  55.792  < 2e-16 ***
## nineam             158.5217     5.9270  26.746  < 2e-16 ***
## tenam               98.0224     5.7304  17.106  < 2e-16 ***
## elevenam           119.8569     5.6014  21.398  < 2e-16 ***
## noon               149.9106     5.8425  25.659  < 2e-16 ***
## onepm              143.5260     5.7550  24.939  < 2e-16 ***
## twopm              123.0243     5.8977  20.860  < 2e-16 ***
## threepm            131.7045     5.9030  22.311  < 2e-16 ***
## fourpm             197.0635     5.8059  33.942  < 2e-16 ***
## fivepm             351.9370     5.7746  60.946  < 2e-16 ***
## sixpm              326.9153     5.6331  58.035  < 2e-16 ***
## sevenpm            231.1748     5.5314  41.793  < 2e-16 ***
## eightpm            160.0310     5.5960  28.597  < 2e-16 ***
## ninepm             119.8848     5.7798  20.742  < 2e-16 ***
## tenpm               87.1202     5.7852  15.059  < 2e-16 ***
## elevenpm            47.6095     5.5553   8.570  < 2e-16 ***
## pca$PC1            -78.7372     2.0354 -38.683  < 2e-16 ***
## pca$PC2             12.7685     1.1205  11.396  < 2e-16 ***
## pca$PC3             73.3893     9.7212   7.549 4.59e-14 ***
## pca$PC4             22.6844     3.4934   6.494 8.61e-11 ***
## pca$PC2:humidity     2.3066     0.5970   3.864 0.000112 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.7 on 17346 degrees of freedom
## Multiple R-squared:  0.6799, Adjusted R-squared:  0.6793
## F-statistic:  1151 on 32 and 17346 DF,  p-value: < 2.2e-16
```

Here I see for every unit increase in "humid weather", the effects of humidity on mean rider counts is 2 more mean riders for each unit increase of humidity.

```
# maybe_better_reg1 with interactions on PC3
humidregpc3 <-lm (rider_count ~ yr +  # maybe_better_reg1 interaction model
                                spring + summer + fall +
                                # categorical variables for seasons
                                oneam + twoam + threeam + fouram + fiveam + s
ixam + sevenam +
                                eightam + nineam + tenam + elevenam + noon +
                                onepm + twopm + threepm + fourpm + fivepm + s
ixpm + sevenpm +
                                eightpm + ninepm + tenpm + elevenpm +
                                # time component
```

```
                              pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 +
                              # transformed component

                              # interaction variables
                              humidity:pca$PC3 )

summary(humidregpc3)
```

```
##
## Call:
## lm(formula = rider_count ~ yr + spring + summer + fall + oneam +
##     twoam + threeam + fouram + fiveam + sixam + sevenam + eightam +
##     nineam + tenam + elevenam + noon + onepm + twopm + threepm +
##     fourpm + fivepm + sixpm + sevenpm + eightpm + ninepm + tenpm +
##     elevenpm + pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 + humidity:pca$PC3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -406.17  -59.92   -6.66   50.95  480.19
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         59.2751     4.2650  13.898  < 2e-16 ***
## yr                  83.1841     1.7141  48.530  < 2e-16 ***
## spring             -32.4950     9.3332  -3.482  0.00050 ***
## summer             100.2368    19.3790   5.172 2.34e-07 ***
## fall              -208.8886    11.6469 -17.935  < 2e-16 ***
## oneam                5.5752     5.7265   0.974  0.33027
## twoam               -1.2452     5.7906  -0.215  0.82974
## threeam            -12.6742     5.7664  -2.198  0.02797 *
## fouram             -16.6828     5.6575  -2.949  0.00319 **
## fiveam               4.4449     5.6695   0.784  0.43305
## sixam               57.0434     5.5095  10.354  < 2e-16 ***
## sevenam            182.5749     5.7140  31.952  < 2e-16 ***
## eightam            317.7824     5.6900  55.849  < 2e-16 ***
## nineam             159.6521     5.9176  26.979  < 2e-16 ***
## tenam               99.6075     5.7197  17.415  < 2e-16 ***
## elevenam           121.3057     5.5942  21.684  < 2e-16 ***
## noon               151.1979     5.8370  25.904  < 2e-16 ***
## onepm              144.1375     5.7485  25.074  < 2e-16 ***
## twopm              123.0801     5.8875  20.905  < 2e-16 ***
## threepm            131.3879     5.8894  22.309  < 2e-16 ***
## fourpm             196.5202     5.7902  33.940  < 2e-16 ***
## fivepm             351.9671     5.7650  61.053  < 2e-16 ***
## sixpm              327.0764     5.6258  58.138  < 2e-16 ***
## sevenpm            231.4354     5.5248  41.890  < 2e-16 ***
## eightpm            160.2821     5.5893  28.677  < 2e-16 ***
## ninepm             119.7011     5.7724  20.737  < 2e-16 ***
## tenpm               86.6815     5.7771  15.004  < 2e-16 ***
## elevenpm            46.8751     5.5466   8.451  < 2e-16 ***
```

```
## pca$PC1           -78.5143     2.0316 -38.646  < 2e-16 ***
## pca$PC2            11.5049     1.1218  10.256  < 2e-16 ***
## pca$PC3            70.1596     9.7211   7.217 5.52e-13 ***
## pca$PC4            21.5540     3.4881   6.179 6.58e-10 ***
## pca$PC3:humidity    5.0245     0.6634   7.574 3.80e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.6 on 17346 degrees of freedom
## Multiple R-squared:  0.6806, Adjusted R-squared:  0.6801
## F-statistic:  1155 on 32 and 17346 DF,  p-value: < 2.2e-16
```

Here I see for every unit increase in "not summer", the effects of humidity on mean rider counts is 5 more mean riders for each unit increase of humidity.

```
# maybe_better_reg1 with interactions on PC4
humidregPC4 <-lm (rider_count ~ yr +  # maybe_better_reg1 interaction model
                                spring + summer + fall +
                                # categorical variables for seasons
                                oneam + twoam + threeam + fouram + fiveam + s
ixam + sevenam +
                                eightam + nineam + tenam + elevenam + noon +
                                onepm + twopm + threepm + fourpm + fivepm + s
ixpm + sevenpm +
                                eightpm + ninepm + tenpm + elevenpm +
                                # time component
                                pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 +
                                # transformed component

                                # interaction variables
                                humidity:pca$PC4 )

summary(humidregPC4)

##
## Call:
## lm(formula = rider_count ~ yr + spring + summer + fall + oneam +
##     twoam + threeam + fouram + fiveam + sixam + sevenam + eightam +
##     nineam + tenam + elevenam + noon + onepm + twopm + threepm +
##     fourpm + fivepm + sixpm + sevenpm + eightpm + ninepm + tenpm +
##     elevenpm + pca$PC1 + pca$PC2 + pca$PC3 + pca$PC4 + humidity:pca$PC4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -407.90  -60.35   -6.59   50.33  471.94
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        58.9914     4.2730  13.806  < 2e-16 ***
## yr                 84.0825     1.7247  48.752  < 2e-16 ***
```

```
## spring             -37.9524     9.3277  -4.069 4.75e-05 ***
## summer             107.9160    19.3888   5.566 2.65e-08 ***
## fall              -211.5229    11.6622 -18.137  < 2e-16 ***
## oneam                5.8833     5.7366   1.026  0.30511
## twoam               -1.0702     5.8033  -0.184  0.85369
## threeam            -12.4245     5.7800  -2.150  0.03160 *
## fouram             -16.6258     5.6725  -2.931  0.00338 **
## fiveam               4.5564     5.6887   0.801  0.42316
## sixam               57.0560     5.5220  10.332  < 2e-16 ***
## sevenam            182.4542     5.7230  31.881  < 2e-16 ***
## eightam            317.5347     5.6988  55.719  < 2e-16 ***
## nineam             159.0503     5.9266  26.837  < 2e-16 ***
## tenam               98.8010     5.7277  17.250  < 2e-16 ***
## elevenam           120.5221     5.6016  21.515  < 2e-16 ***
## noon               149.9817     5.8443  25.663  < 2e-16 ***
## onepm              143.3716     5.7571  24.903  < 2e-16 ***
## twopm              122.2571     5.8953  20.738  < 2e-16 ***
## threepm            130.7704     5.8981  22.172  < 2e-16 ***
## fourpm             196.3872     5.8047  33.833  < 2e-16 ***
## fivepm             351.5775     5.7760  60.869  < 2e-16 ***
## sixpm              327.0343     5.6381  58.004  < 2e-16 ***
## sevenpm            232.0365     5.5506  41.804  < 2e-16 ***
## eightpm            161.0066     5.6162  28.668  < 2e-16 ***
## ninepm             120.3702     5.7904  20.788  < 2e-16 ***
## tenpm               87.1528     5.7896  15.053  < 2e-16 ***
## elevenpm            47.1122     5.5550   8.481  < 2e-16 ***
## pca$PC1            -77.9107     2.0495 -38.015  < 2e-16 ***
## pca$PC2             12.1595     1.1229  10.828  < 2e-16 ***
## pca$PC3             73.9185     9.7229   7.602 3.05e-14 ***
## pca$PC4             22.7377     3.5005   6.496 8.49e-11 ***
## pca$PC4:humidity    -1.7019     0.7745  -2.197  0.02802 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 102.8 on 17346 degrees of freedom
## Multiple R-squared:  0.6797, Adjusted R-squared:  0.6791
## F-statistic:  1150 on 32 and 17346 DF,  p-value: < 2.2e-16
```

Here I see for every unit increase in “harsher weather”, the effects of humidity on mean rider counts is 1 less mean riders for each unit increase of humidity.