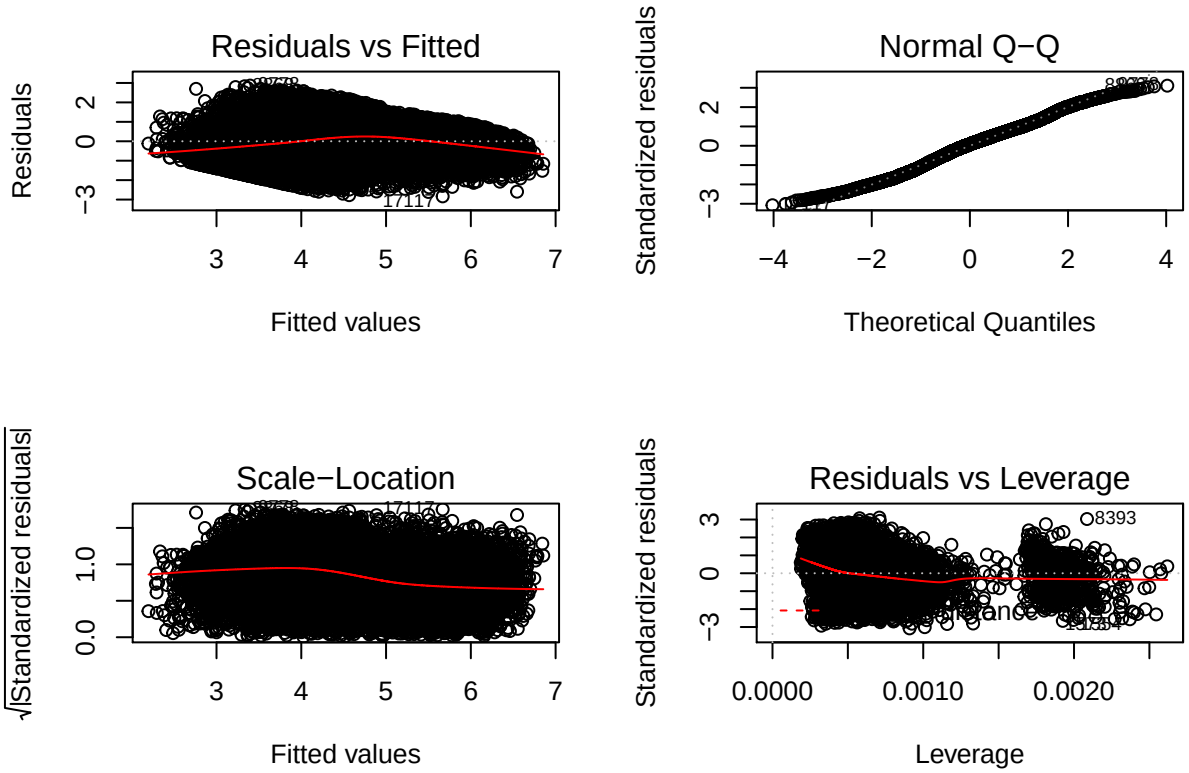
Code

Yakub Akhmerov

November 17, 2016

```
library(leaps)
library(ggplot2)
library(boot)
setwd("/Users/yakubakhmerov/Downloads")
BikeSharing <- read.csv("BikeSharingDataset.csv")
```

```
par(mfrow = c(2, 2))
plot(lm(log(cnt + 4)~ season + yr + mnth + hr + weekday + temp + weathersit + hum + windspeed, data=Bik
```




```
env_season_casual.lm = lm(log(casual + 4)~ season + yr + mnth + hr + weekday + temp + weathersit + hum +

env_season_casual = regsubsets(log(casual + 4)~ season + yr + mnth + hr + weekday + temp + weathersit +

n = nrow(BikeSharing)
```

Computing Adjusted R squared

```r
summary(env_season_casual)$adjr2
```

```
## [1] 0.3248451 0.4767720 0.5419084 0.5482062 0.5529379 0.5529832 0.5530194
## [8] 0.5530276
```

```r
which.max(summary(env_season_casual)$adjr2)
```

```
## [1] 8
```

```r
summary(env_season_casual)$which #throwout windspeed
```

```
##   (Intercept) season    yr  mnth    hr weekday temp weathersit   hum
## 1        TRUE  FALSE FALSE FALSE FALSE   FALSE TRUE      FALSE FALSE
## 2        TRUE  FALSE FALSE FALSE  TRUE   FALSE TRUE      FALSE FALSE
## 3        TRUE  FALSE FALSE FALSE  TRUE   FALSE TRUE      FALSE  TRUE
## 4        TRUE  FALSE  TRUE FALSE  TRUE   FALSE TRUE      FALSE  TRUE
## 5        TRUE   TRUE  TRUE FALSE  TRUE   FALSE TRUE      FALSE  TRUE
## 6        TRUE   TRUE  TRUE  TRUE  TRUE   FALSE TRUE      FALSE  TRUE
## 7        TRUE   TRUE  TRUE  TRUE  TRUE    TRUE TRUE      FALSE  TRUE
## 8        TRUE   TRUE  TRUE  TRUE  TRUE    TRUE TRUE       TRUE  TRUE
##   windspeed
## 1     FALSE
## 2     FALSE
## 3     FALSE
## 4     FALSE
## 5     FALSE
## 6     FALSE
## 7     FALSE
## 8     FALSE
```

Per the analysis, it showed that it was appropriate to throw out the variable "windspeed".

Computing AIC

```r
step(env_season_casual.lm) #throw out windspeed and weathersit
#hid results because it makes appendix difficult to read
```

Per the analysis of AIC, it seemed best to throwout the "windspeed" and "weathersit" variable.

Computing BIC

```r
step(env_season_casual.lm, direction="both", k = log(n)) #windspeed, weathersit, weekday, mnth out
#hid results because it makes appendix difficult to read
```

BIC showed that it'd be best to throw out "windspeed", weathersit, "weekday" and "mnth".

Computing Mallow's CP

```
summary(env_season_casual)$cp
```

```
## [1] 8873.840895 2968.019039  436.748098  192.918052    9.991672    9.231224
## [7]    8.822592    9.503468
```

```
which.min(summary(env_season_casual)$cp) #throw out weathersit and windspeed
```

```
## [1] 7
```

Per the analysis of Mallow's CP, it seemed best to throwout the "windspeed" and "weathersit" variable.

Leave One Out Cross-Validation

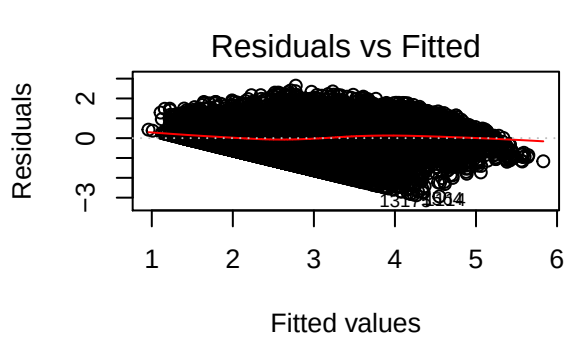
```
m1 = lm(log(casual + 4)~ season + yr + mnth + hr + weekday + temp + weathersit + hum, data=BikeSharing)
m2 = lm(log(casual + 4)~ season + yr + mnth + hr + weekday + temp + hum, data=BikeSharing)
m3 = lm(log(casual + 4)~ season + yr + hr + temp + hum, data=BikeSharing)
#m4 = lm(log(cnt + 4)~ season + yr + mnth + hr + weekday + temp + weathersit + hum + windspeed, data=Bi

cv.scores = rep(-999, 3)
cv.scores[1] = sum((m1$residuals^2)/((1 - influence(m1)$hat)^2))
cv.scores[2] = sum((m2$residuals^2)/((1 - influence(m2)$hat)^2))
cv.scores[3] = sum((m3$residuals^2)/((1 - influence(m3)$hat)^2))
#cv.scores[4] = sum((m4$residuals^2)/((1 - influence(m4)$hat)^2))
cv.scores
```

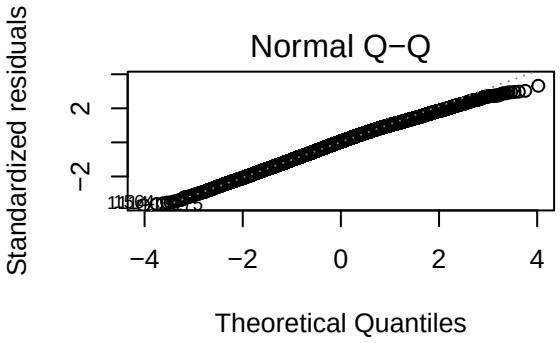
```
## [1] 9948.264 9947.836 9948.306
```

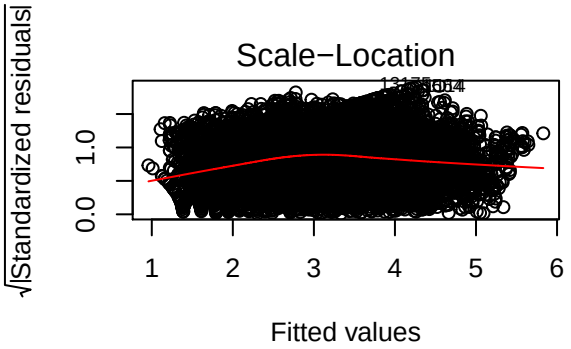
The lowest CV svore is 9947.836, which is the AIC and Mallow's CP models.

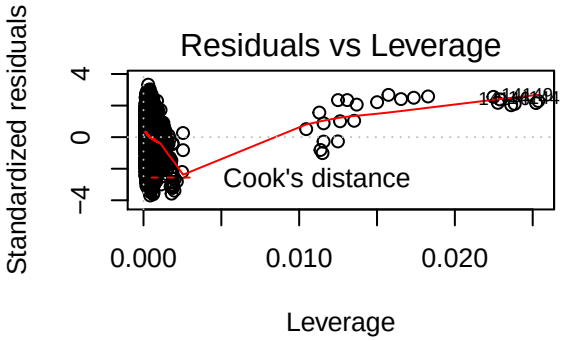
what is the effect of temperature on the number of bike rentals

```
#Hypothesis test: Testing the null hypothesis that the working day variable has no impact on the number
M = lm(log(casual + 4) ~ workingday + weathersit + temp + atemp + hum + windspeed, data = BikeSharing)

par(mfrow = c(2, 2))

plot(M) #Testing Normality
```










```
m = lm(log(casual + 4) ~ weathersit + temp + atemp + hum + windspeed, data = BikeSharing)

fst = (deviance(m) - deviance(M))/(deviance(M)/M$df.residual)
sqrt(fst)
```

```
## [1] 42.66776
```

```
1 - pf(fst, 1, M$df.residual)
```

```
## [1] 0
```

```
summary(M)
```

```
##
## Call:
## lm(formula = log(casual + 4) ~ workingday + weathersit + temp +
##     atemp + hum + windspeed, data = BikeSharing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.93870 -0.54279  0.03637  0.57761  2.64451
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.86789    0.03422  83.807  < 2e-16 ***
## workingday  -0.55494    0.01301 -42.668  < 2e-16 ***
## weathersit   0.08014    0.01060   7.562 4.15e-14 ***
```

```
## temp         0.70115    0.20697   3.388 0.000706 ***
## atemp        2.95256    0.23227  12.712  < 2e-16 ***
## hum         -2.14784    0.03659 -58.707  < 2e-16 ***
## windspeed    0.23901    0.05369   4.452 8.56e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7957 on 17372 degrees of freedom
## Multiple R-squared:  0.5056, Adjusted R-squared:  0.5054
## F-statistic:  2961 on 6 and 17372 DF,  p-value: < 2.2e-16
```

```
#The p-value is simply the probability that a t-variable with 17379 degrees of freedom exceeds the t-va
1-pt(1.96, n)
```

```
## [1] 0.02500587
```

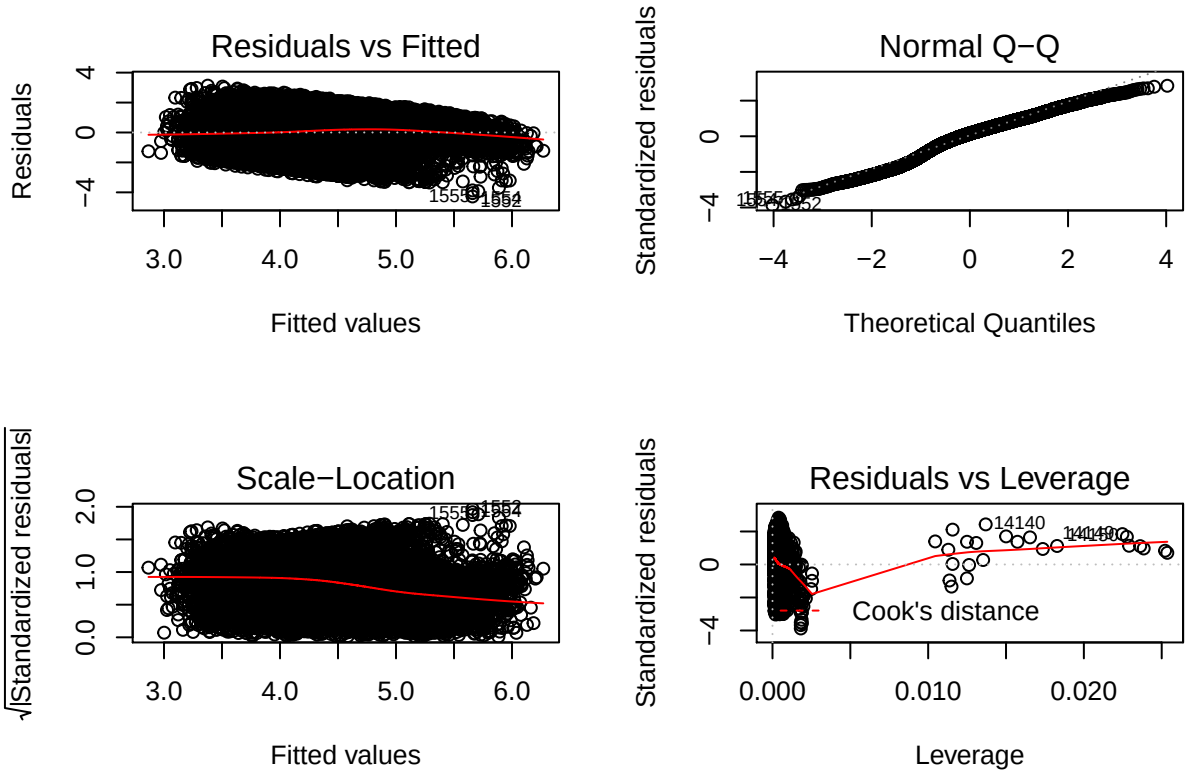
```
anova(M, m)
```

```
## Analysis of Variance Table
##
## Model 1: log(casual + 4) ~ workingday + weathersit + temp + atemp + hum +
##     windspeed
## Model 2: log(casual + 4) ~ weathersit + temp + atemp + hum + windspeed
##   Res.Df   RSS Df Sum of Sq      F    Pr(>F)
## 1  17372 10998
## 2  17373 12151 -1   -1152.6 1820.5 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value computed is 2.2e-16, which is smaller than our critical value of 0.025. Thus, we reject the null.

Does the effect of temperature on the number of bike rentals effect vary from weekday to weekend?

```
#Hypothesis Test
M = lm(log(registered + 4) ~ workingday + weathersit + temp + atemp + hum + windspeed, data = BikeShari
m = lm(log(registered + 4) ~ weathersit + temp + atemp + hum + windspeed, data = BikeSharing)

par(mfrow = c(2, 2))

plot(M) #Testing Normality
```




```
fst = (deviance(m) - deviance(M))/(deviance(M)/M$df.residual)
sqrt(fst)
```

```
## [1] 6.779401
```

```
1 - pf(fst, 1, M$df.residual)
```

```
## [1] 1.245559e-11
```

```
#The p-value is simply the probability that a t-variable with 17379 degrees of freedom exceeds the t-va
1-pt(1.96, n)
```

```
## [1] 0.02500587
```

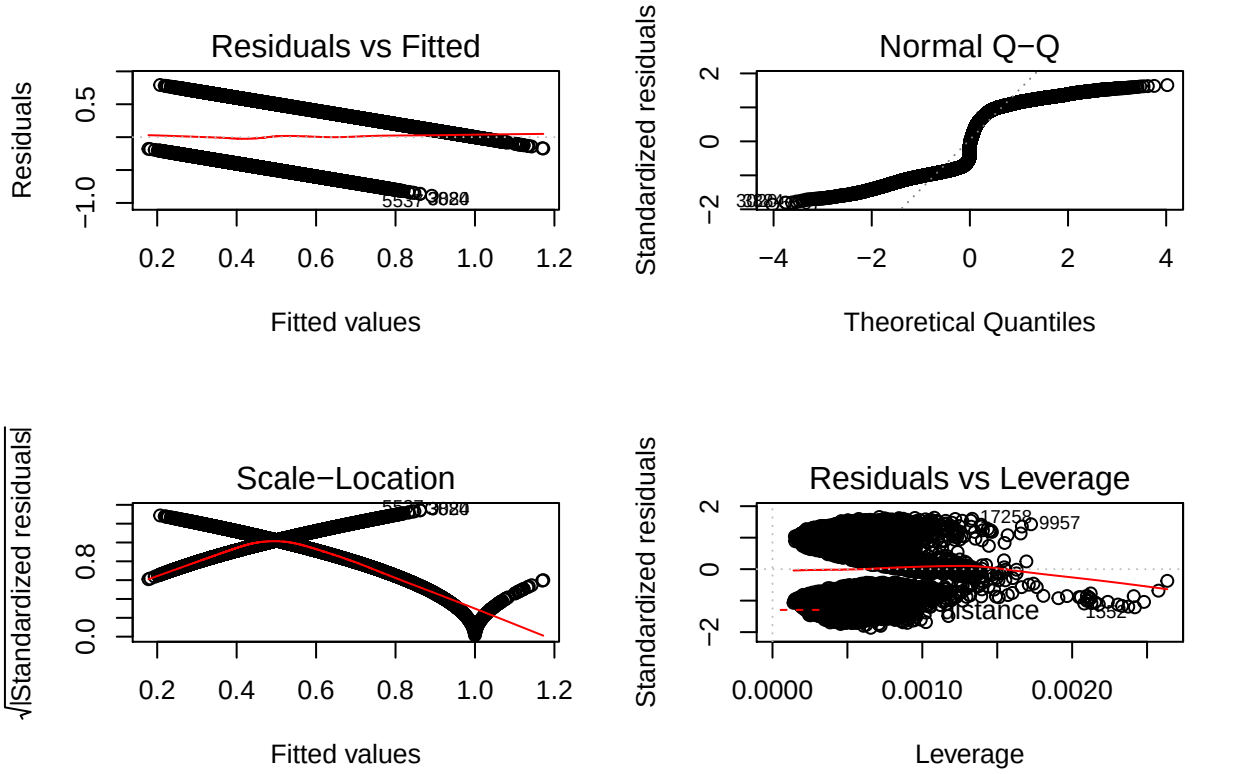
```
anova(M, m)
```

```
## Analysis of Variance Table
##
## Model 1: log(registered + 4) ~ workingday + weathersit + temp + atemp +
##     hum + windspeed
## Model 2: log(registered + 4) ~ weathersit + temp + atemp + hum + windspeed
##   Res.Df   RSS Df Sum of Sq     F    Pr(>F)
## 1  17372 21153
## 2  17373 21209 -1   -55.965 45.96 1.246e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value computed is 1.245559e-11, which is smaller than our critical value of 0.025. Thus, we reject the null.

Between 2011 and 2012, was there a significant difference in the seasonal conditions for these years or is the only difference between these years is the number of bikes rented?

```
#Hypothesis test:
M = lm( yr ~ mnth + hr + workingday + weathersit + atemp + hum + windspeed + cnt, data = BikeSharing)
m = lm( yr ~ cnt, data = BikeSharing)

par(mfrow = c(2, 2))

plot(M) #testing normality
```




```
#nonparametric bootstrap
rsq <- function(formula, data, indices) {
  d <- BikeSharing[indices,] # allows boot to select sample
  fit <- lm(formula, data=d)
  return(summary(M)$coef[6,4])
}

result <- boot(data=BikeSharing, statistic=rsq,
    R=100, formula=yr ~ mnth + hr + workingday + weathersit + atemp + hum + windspeed + cnt)
result
```

```
##
## ORDINARY NONPARAMETRIC BOOTSTRAP
##
```

```
##
## Call:
## boot(data = BikeSharing, statistic = rsq, R = 100, formula = yr ~
##     mnth + hr + workingday + weathersit + atemp + hum + windspeed +
##         cnt)
##
##
## Bootstrap Statistics :
##         original  bias    std. error
## t1* 2.572306e-17       0           0
```

The test statistic is computed to 2.57230562912941e-17. Which is very small and can be safely assumed that it is in the rejection region.