day_8_nov_12_2016_occupancy_vif

Going to look at the multi-colinearity problem for the dataset. As we discovered previously, the correlation between variables are high.

```
##   Temperature       Humidity         Light             CO2        
##  Min.   :20.20   Min.   :22.10   Min.   :   0.0   Min.   : 427.5  
##  1st Qu.:20.65   1st Qu.:23.26   1st Qu.:   0.0   1st Qu.: 466.0  
##  Median :20.89   Median :25.00   Median :   0.0   Median : 580.5  
##  Mean   :21.43   Mean   :25.35   Mean   : 193.2   Mean   : 717.9  
##  3rd Qu.:22.36   3rd Qu.:26.86   3rd Qu.: 442.5   3rd Qu.: 956.3  
##  Max.   :24.41   Max.   :31.47   Max.   :1697.2   Max.   :1402.2  
##  HumidityRatio        Occupancy     
##  Min.   :0.003303   Min.   :0.0000  
##  1st Qu.:0.003529   1st Qu.:0.0000  
##  Median :0.003815   Median :0.0000  
##  Mean   :0.004027   Mean   :0.3647  
##  3rd Qu.:0.004532   3rd Qu.:1.0000  
##  Max.   :0.005378   Max.   :1.0000
```

```r
library(MASS)
df <- as.data.frame(scale(df))
fit <- lm(Occupancy~., data=df)
summary(fit)
```

```
## 
## Call:
## lm(formula = Occupancy ~ ., data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -4.0893 -0.0821 -0.0307  0.1150  1.1763 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)    3.127e-15  6.524e-03   0.000    1.000    
## Temperature   -1.186e+00  8.078e-02 -14.683   <2e-16 ***
## Humidity      -1.304e+00  1.158e-01 -11.261   <2e-16 ***
## Light          9.513e-01  1.243e-02  76.503   <2e-16 ***
## CO2           -3.681e-02  3.026e-02  -1.217    0.224    
## HumidityRatio  2.379e+00  1.868e-01  12.737   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.3368 on 2659 degrees of freedom
## Multiple R-squared:  0.8868, Adjusted R-squared:  0.8866 
## F-statistic:  4165 on 5 and 2659 DF,  p-value: < 2.2e-16
```

So first things first, we notice that the adjusted R2 is 0.905 right now with all the variables.

Then we can also look at the p-values for the coefficients. The raw p-value is against the H0 that the coefficients can be equal to 0. If we take a naive alpha of 0.05, then we can reject all the nulls except for CO2, CO2 seems to be not as important to the model.

```r
fit$coefficients
```

```
##   (Intercept)   Temperature      Humidity         Light           CO2
##  3.126976e-15 -1.186048e+00 -1.304426e+00  9.513004e-01 -3.681144e-02
## HumidityRatio
##  2.378987e+00
```

Looking at the coefficients, we see that CO2, Temperature and Humidity have negative coefficients.

This is pretty weird because you'd think that more CO2 or higher temperature in office means there are people.

```r
t.test(df[df$Occupancy<0,]$CO2, df[df$Occupancy>0,]$CO2)
```

```
##
##  Welch Two Sample t-test
##
## data:  df[df$Occupancy < 0, ]$CO2 and df[df$Occupancy > 0, ]$CO2
## t = -55.432, df = 1450.9, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.651714 -1.538810
## sample estimates:
##  mean of x  mean of y
## -0.5818367  1.0134255
```

This simple t-test has a large negative value, meaning that the first sample (no one in office)'s CO2 level is lower than the second sample.

This suggests that there are colinearity in our dataset, because CO2 should have a positive coefficient, but instead have a negative one.

```r
cor(df)
```

```
##               Temperature  Humidity     Light       CO2 HumidityRatio
## Temperature     1.0000000 0.7169882 0.7684279 0.8702265     0.8945481
## Humidity        0.7169882 1.0000000 0.5619048 0.9116106     0.9519135
## Light           0.7684279 0.5619048 1.0000000 0.7691675     0.6932860
## CO2             0.8702265 0.9116106 0.7691675 1.0000000     0.9644402
## HumidityRatio   0.8945481 0.9519135 0.6932860 0.9644402     1.0000000
## Occupancy       0.7057838 0.6107640 0.9279491 0.7680296     0.7003300
##               Occupancy
## Temperature   0.7057838
## Humidity      0.6107640
## Light         0.9279491
## CO2           0.7680296
## HumidityRatio 0.7003300
## Occupancy     1.0000000
```

The pearson r suggests that we have some highly correlated variables.

```
library(car)
vif(fit)
```

```
##   Temperature      Humidity         Light           CO2 HumidityRatio 
##    153.233964    315.084070      3.631164     21.501371    819.260074
```

Wow, uhhh that's some really high values for VIF. Typically greater than 4 means need investigation, greather than 10 means something is seriously not right.

Let's remove some variables, which should be a surrogate for HumidityRatio.

```
df2 <- df[,c("Temperature", "Humidity", "Light", "CO2", "Occupancy")]
cor(df2)
```

```
##             Temperature  Humidity     Light       CO2 Occupancy
## Temperature   1.0000000 0.7169882 0.7684279 0.8702265 0.7057838
## Humidity      0.7169882 1.0000000 0.5619048 0.9116106 0.6107640
## Light         0.7684279 0.5619048 1.0000000 0.7691675 0.9279491
## CO2           0.8702265 0.9116106 0.7691675 1.0000000 0.7680296
## Occupancy     0.7057838 0.6107640 0.9279491 0.7680296 1.0000000
```

```
df3 <- df[,c("Temperature", "Light", "CO2", "Occupancy")]
cor(df3)
```

```
##             Temperature     Light       CO2 Occupancy
## Temperature   1.0000000 0.7684279 0.8702265 0.7057838
## Light         0.7684279 1.0000000 0.7691675 0.9279491
## CO2           0.8702265 0.7691675 1.0000000 0.7680296
## Occupancy     0.7057838 0.9279491 0.7680296 1.0000000
```

```
df4 <- df[,c("Light", "CO2", "Occupancy")]
cor(df4)
```

```
##               Light       CO2 Occupancy
## Light     1.0000000 0.7691675 0.9279491
## CO2       0.7691675 1.0000000 0.7680296
## Occupancy 0.9279491 0.7680296 1.0000000
```

```
summary(lm(Occupancy~., data=df4))
```

```
## 
## Call:
## lm(formula = Occupancy ~ ., data = df4)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3.7501 -0.0480 -0.0035  0.1238  1.3659 
## 
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 1.957e-15  7.032e-03    0.00        1
## Light       8.257e-01  1.101e-02   75.02   <2e-16 ***
## CO2         1.329e-01  1.101e-02   12.08   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.363 on 2662 degrees of freedom
## Multiple R-squared:  0.8683, Adjusted R-squared:  0.8682
## F-statistic:  8776 on 2 and 2662 DF,  p-value: < 2.2e-16
```

We can see that after getting rid of more than half our covariates, the adjusted R2 is still pretty high, only .03 lower than before.