Principle Component Analysis of Sugar Bowl Weather Data

John Lomas

11/2/2018

Principle Component Analysis

Principle component analysis is often performed during exploratory data analysis when there are many varaibles and/or multicollinearty is an issue. It is an orthogonal transformation which simply computes a new set of axes from which to observe the data. The first axis (or principle component) is always in the direction of greatest variability. The seond principle component is in the second greatest direction of variability, and so on. What PCoA gives you in terms of dimensionality reduction and independence, you lose in interpretability (since every principle component is a linear combination of every original varaible).

In the previous modeling section we saw clear evidence of multicollinearty and decided to remove 4 variables because of it. Here, we'll use PCoA to eliminate the multicollinearity issue while retaining as much of the information as possible.

Generating Principal Components

We'll begin by loading the data and cleaning it up for the analysis:

```r
# Load libraries
library(ggbiplot)
library(ggplot2)
library(plotly)
library(rsm)
library(spatial)
library(car)

# Load data and format the months factor for nice plotting
weather <- read.csv("weather_df.csv")
weather$Month = factor(weather$Month, levels(weather$Month)[c(5, 4, 8, 1, 9,
                                                              7, 6, 2, 12, 11,
                                                              10, 3)])
head(weather)
```

```
##      Date Year Day_of_Year Day_of_Run Solar_Rad Wind_Speed Wind_Direction
## 1 12/1/05 2005         335          1     0.097       2.99              2
## 2 12/2/05 2005         336          2     0.841       2.02            295
## 3 12/3/05 2005         337          3     2.794       2.22            126
## 4 12/4/05 2005         338          4     2.295       0.82             90
## 5 12/5/05 2005         339          5     2.322       0.46            148
## 6 12/6/05 2005         340          6     2.230       0.55            198
##   Wind_Gust Tavg Tmax  Tmin Havg Hmax Hmin Pressure Snow_Depth
## 1     19.89  1.9  3.0  -0.1   98   99   95      796      244.8
## 2     13.88 -4.7 -0.2  -7.3   90   98   79      799      250.8
## 3     18.19 -6.5 -3.5 -10.0   72   94   51      806      453.6
## 4     13.00 -5.9  0.6 -10.7   68   88   49      811      253.2
## 5      6.47 -3.1  3.7  -7.4   51   78   25      810      253.6
## 6      6.92 -2.1  4.9  -7.0   45   63   21      808      251.2
##   Total_Precip    Month
## 1        99.06 December
## 2         3.05 December
## 3         1.52 December
## 4         0.00 December
## 5         0.25 December
## 6         0.00 December
```

```
#Remove samples where data is missing from one or more variables
weather.complete <- weather[complete.cases(weather),]

#Generate principle components
weather.pca <- prcomp(weather.complete[,c(5, 9:15)],
                      center = TRUE,
                      scale. = TRUE)

#View the principle components
summary(weather.pca)
```

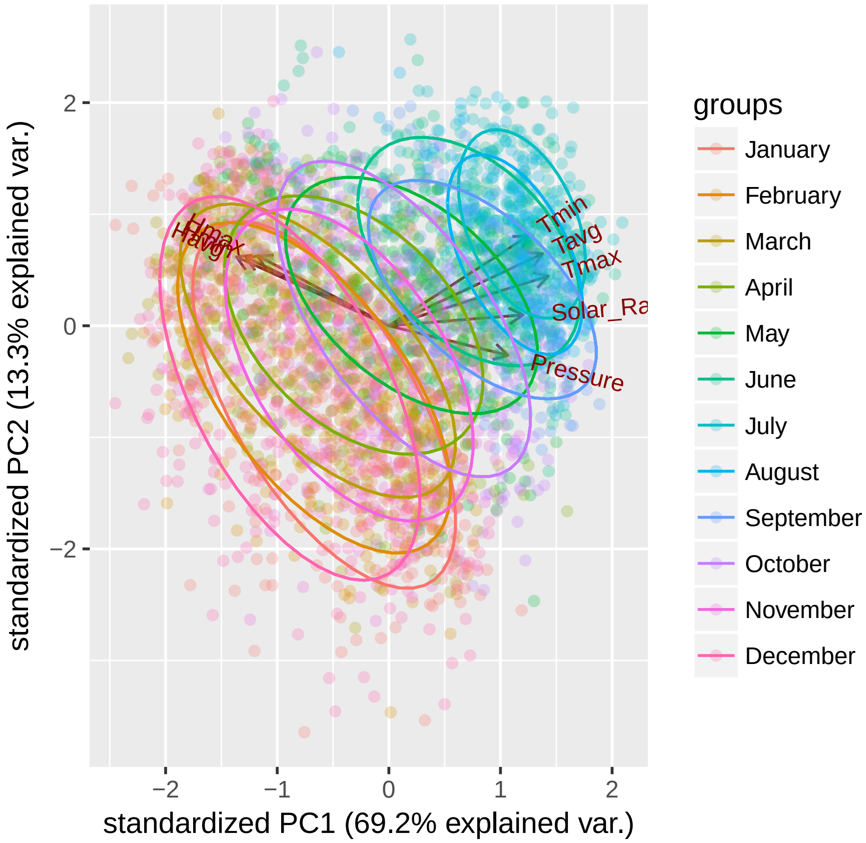
```
## Importance of components:
##                           PC1    PC2     PC3     PC4     PC5     PC6
## Standard deviation     2.3525 1.0328 0.77055 0.70533 0.46621 0.23583
## Proportion of Variance 0.6918 0.1333 0.07422 0.06219 0.02717 0.00695
## Cumulative Proportion  0.6918 0.8251 0.89932 0.96151 0.98868 0.99563
##                            PC7     PC8
## Standard deviation     0.16444 0.08904
## Proportion of Variance 0.00338 0.00099
## Cumulative Proportion  0.99901 1.00000
```

After building the principle components for the data set, the sumary tells us that PC1 contains 69% of the total variability, PC2 contains 13% of the variability, and almost all of the variability is accounted for by PC1 - PC4. Interestingly, 82.5% of the total variability is accounted for in the first two principle componets. Because most of the varaition is in the first two principle components, we can move forward with PC1 and PC2 for further analysis.

Plotting the Principle Components

One useful chart for visualizing principle components is a Biplot. In this type of chart, the original axes are projected onto a scaatter plot using principle components as the primary axes. This allows you to see how the principle components interact with the original variables.

```
ggbiplot(weather.pca, alpha = .25, ellipse = TRUE, groups = weather.complete$Month)
```

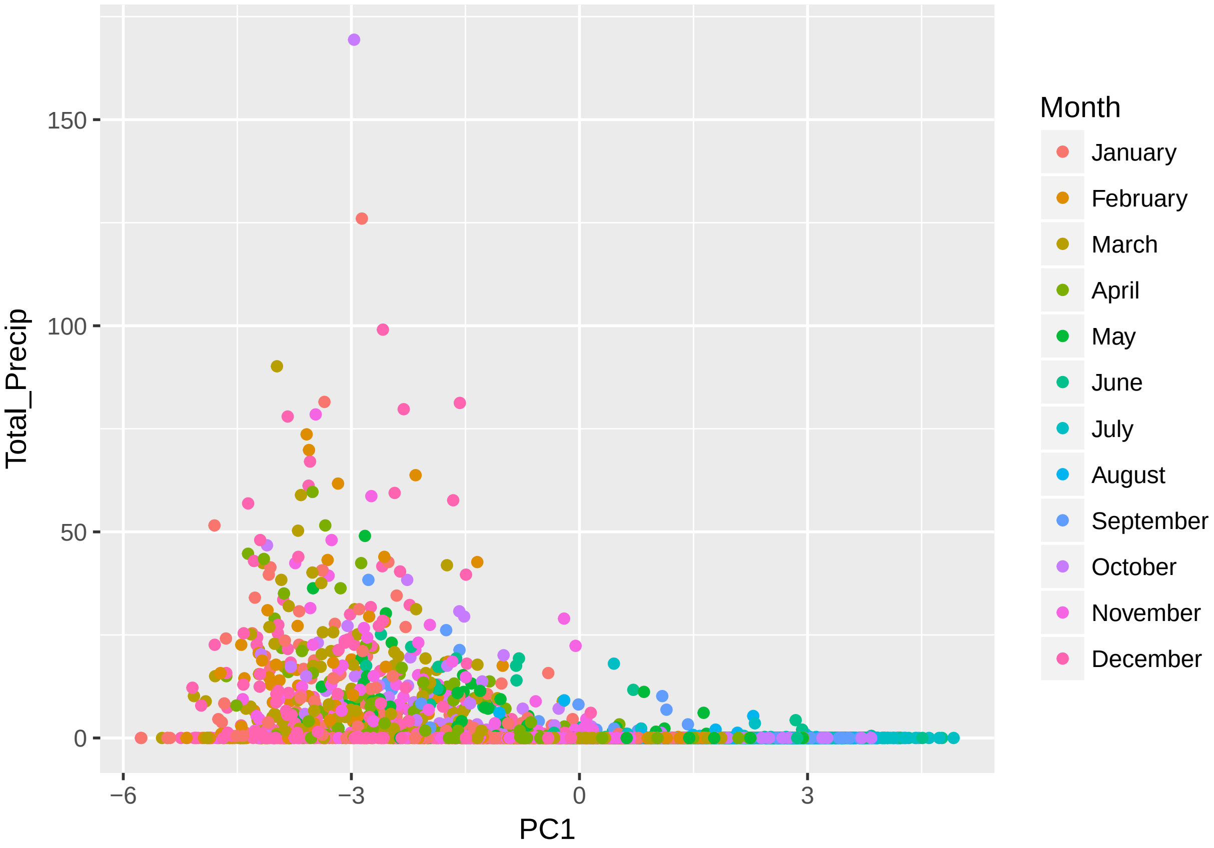


In this biplot it is evident that Solar Radiation plays a major role in determining PC1 and, therefore, the whole dataset. Furthermore, we can see how the variance increases in the winter months (pink) and decreases in the summer months (green/blue).

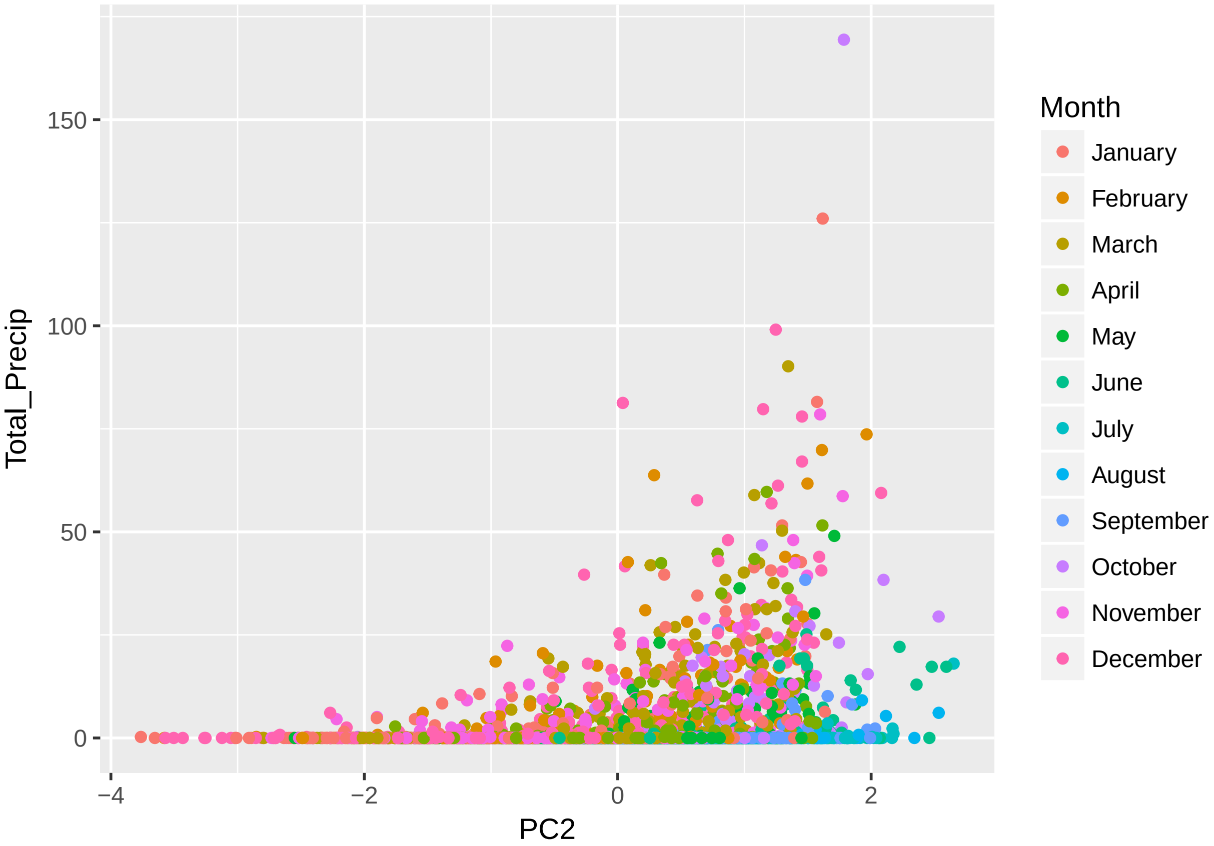
How do the Principle Components Relate to Total Precipitation?

Now, we can look at the relationship between the principle components and the total precipitation by making scatter plots.

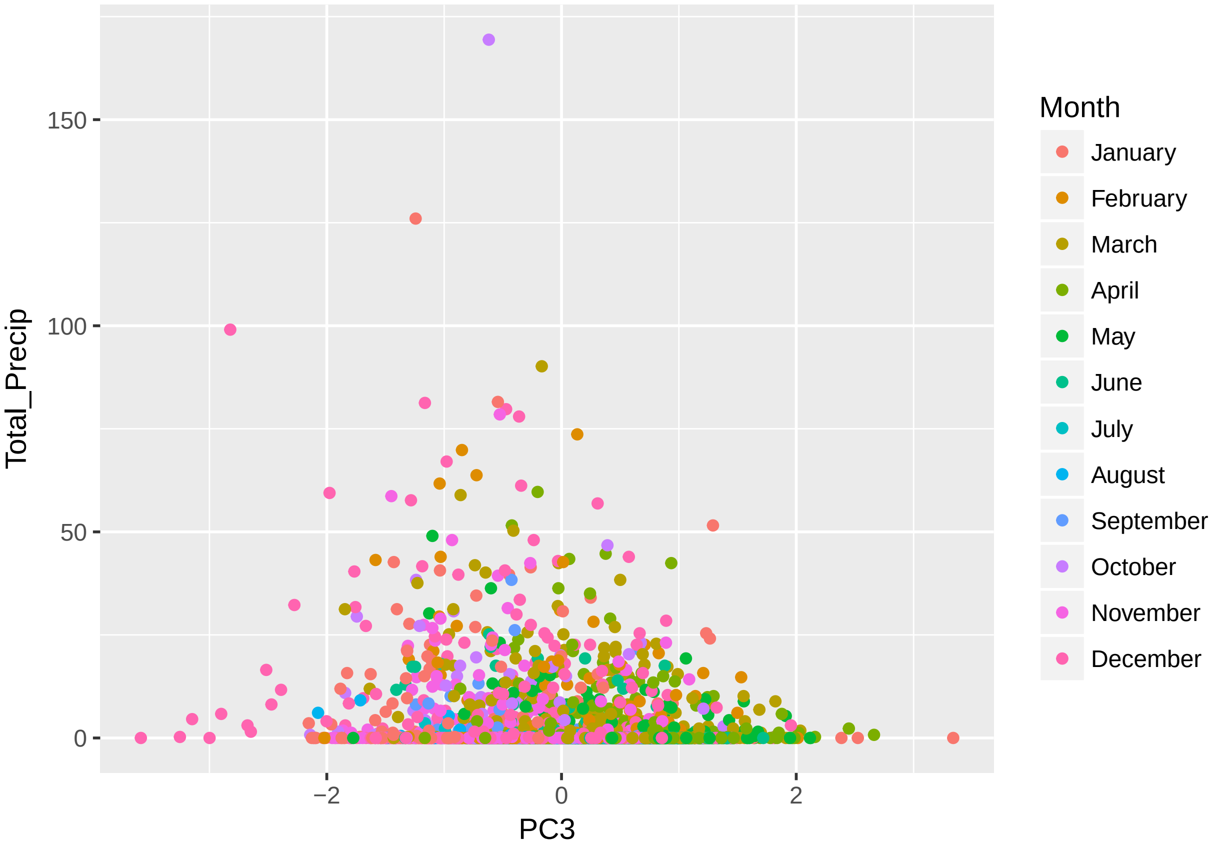
```
PCs <- weather.pca$x
PCs <- as.data.frame(PCs)
PC.complete <- cbind(PCs,weather.complete$Total_Precip, weather.complete$Month)
colnames(PC.complete)[9:10] <- c("Total_Precip", "Month")
ggplot(data = PC.complete, mapping = aes(x = PC1, y = Total_Precip)) +
  geom_point(mapping = aes(color = Month))
```



```r
ggplot(data = PC.complete, mapping = aes(x = PC2, y = Total_Precip)) +
  geom_point(mapping = aes(color = Month))
```



```
ggplot(data = PC.complete, mapping = aes(x = PC3, y = Total_Precip)) +
  geom_point(mapping = aes(color = Month))
```



```
plot_ly(PC.complete, x = ~PC1,
                     y = ~PC2,
                     z = ~Total_Precip,
                     type = "scatter3d",
                     color = ~Month,
                     mode = "markers",
                     marker = list(line = list(color = "black", width = .75)))
```

```
## Warning in RColorBrewer::brewer.pal(N, "Set2"): n too large, allowed maximum for palette Set2 is 8
## Returning the palette you asked for with that many colors
```

Evidently, precipitation events are favored by low values of PC1 (the component related to solar radiation and temperature) and high values of PC2. I say favored because there are clearly many low PC1, high PC2 days on which precipitation did not occur.

Principle Component Regression

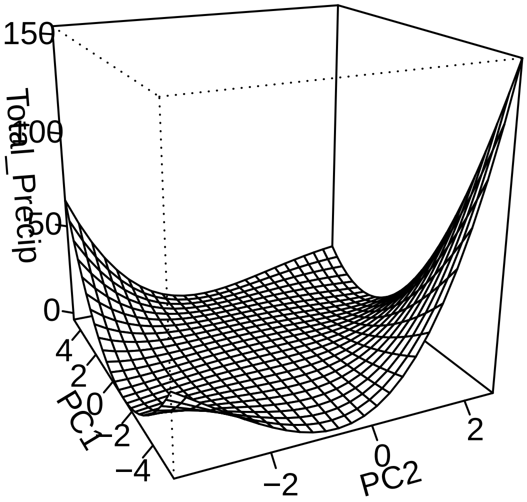
Having analyzed the principle components and seen a clear response between the first two principle components and total precipitation, we can build a new regression model based on the principle components. I'm moving on with a fourth degree polynomial type model based on trail and error using the fit diagnostics in the previous section to guide me.

```
fit <- lm(Total_Precip ~ PC1 + PC1^2 + PC1^3 + PC1^4 +
               PC1:PC2 + I(PC1^2):PC2 + #I(PC1^3):PC2 +
               PC1:I(PC2^2) + I(PC1^2):I(PC2^2) +
               I(PC2^3) + PC1:I(PC2^3) +
               I(PC2^4), data = PC.complete)
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = Total_Precip ~ PC1 + PC1^2 + PC1^3 + PC1^4 + PC1:PC2 +
##     I(PC1^2):PC2 + PC1:I(PC2^2) + I(PC1^2):I(PC2^2) + I(PC2^3) +
##     PC1:I(PC2^3) + I(PC2^4), data = PC.complete)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -40.701  -1.207  -0.364   0.288 139.264
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         1.19168    0.14530   8.202 3.34e-16 ***
## PC1                -0.50659    0.06778  -7.474 9.86e-14 ***
## I(PC2^3)            0.34575    0.05381   6.426 1.50e-10 ***
## I(PC2^4)            0.07823    0.01983   3.945 8.15e-05 ***
## PC1:PC2            -0.67140    0.10348  -6.488 9.96e-11 ***
## PC2:I(PC1^2)        0.30590    0.02315  13.215  < 2e-16 ***
## PC1:I(PC2^2)       -0.69515    0.05404 -12.863  < 2e-16 ***
## I(PC1^2):I(PC2^2)  0.21304    0.01782  11.957  < 2e-16 ***
## PC1:I(PC2^3)       -0.22581    0.03425  -6.594 4.97e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.162 on 3346 degrees of freedom
## Multiple R-squared:  0.3768, Adjusted R-squared:  0.3753
## F-statistic: 252.9 on 8 and 3346 DF,  p-value: < 2.2e-16
```

```r
persp(fit, PC1 ~ PC2, zlab = "Total_Precip")
```



Though the model has all significant regression coefficients, the Adjusted R-squared value is still relatively low, indicating poor predictive ability. Now we can run the model diagnostics to see if principle component regression improved our model at all.

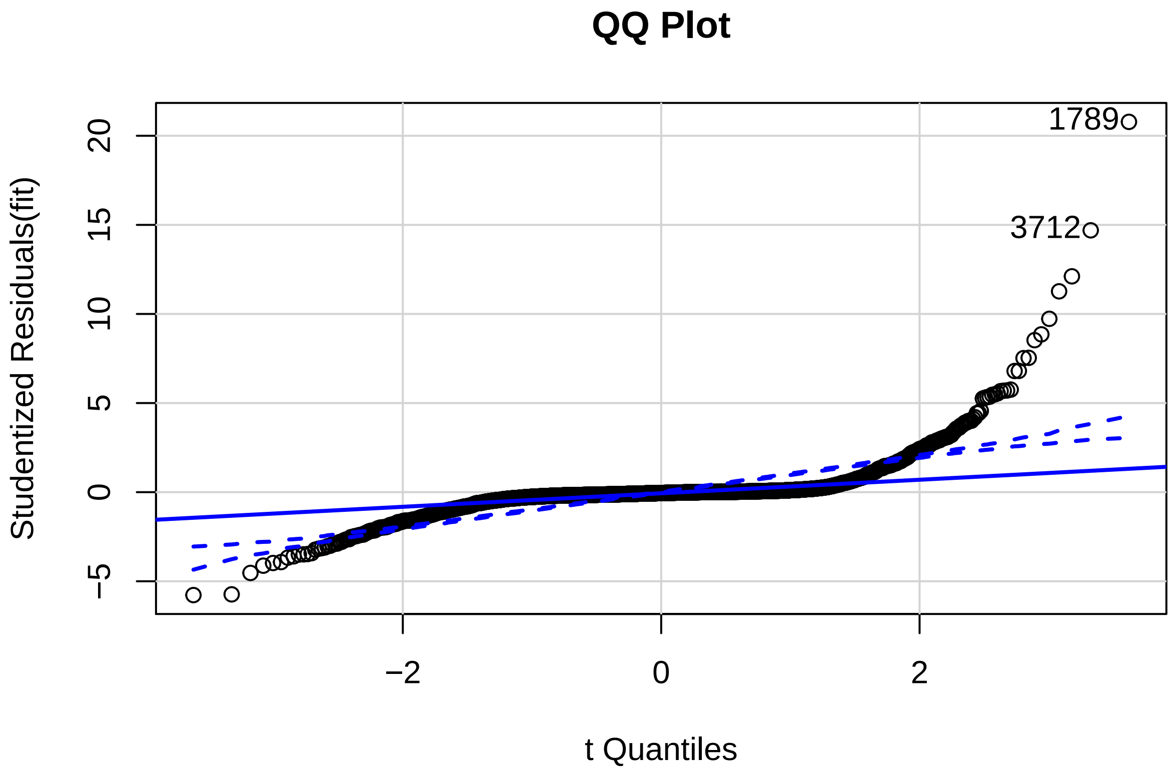
```r
all_vifs <- car::vif(fit)
print(all_vifs)
```

```
##              PC1         I(PC2^3)         I(PC2^4)          PC1:PC2
##         1.662429         3.129844         2.893834         3.088225
##      PC2:I(PC1^2)    PC1:I(PC2^2) I(PC1^2):I(PC2^2)    PC1:I(PC2^3)
##         1.445386         1.912445         1.294419         3.058461
```

Notice that we have no multicollinearity issues as indicated by the above variance inflation factors.

```
# qq plot for studentized residuals

qqPlot(fit, main="QQ Plot")
```



```
## 1789 3712
## 1717 3299
```

The Q-Q Plot indicates that we are still not satisfying the normality assumption.

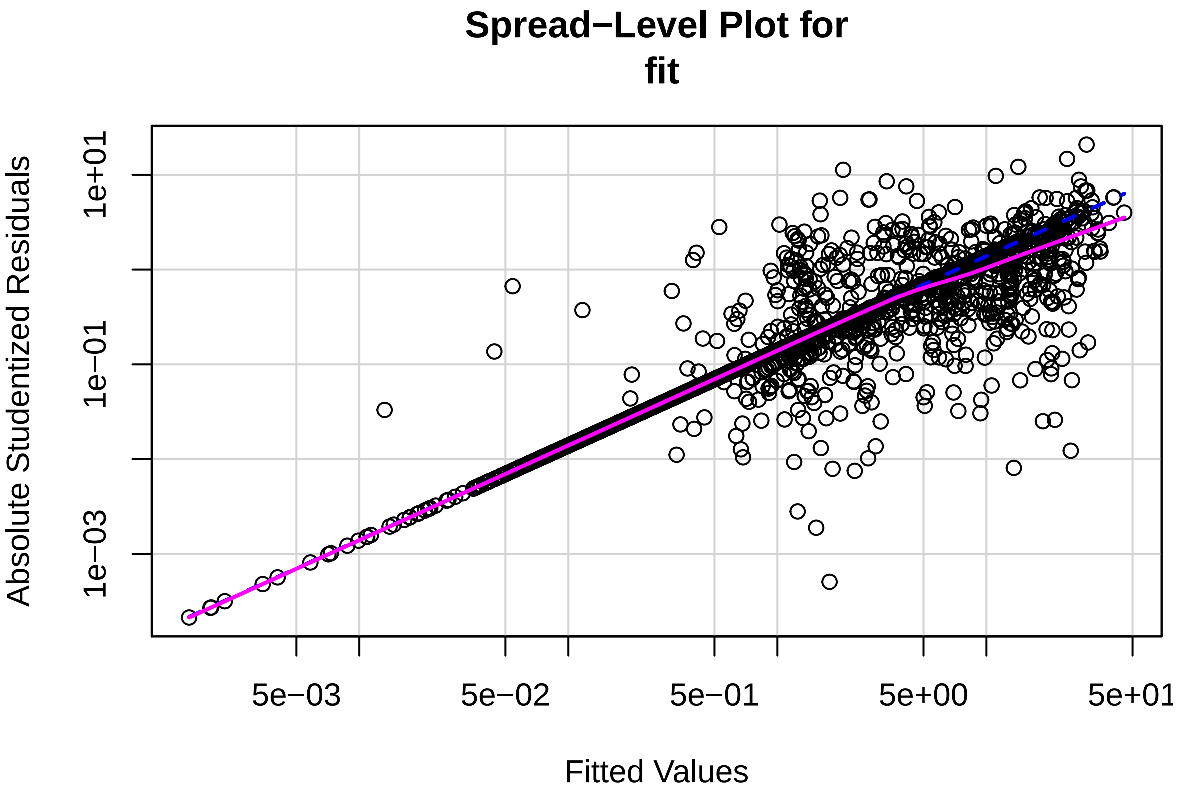
```
# Breusch-Pagan test
ncvTest(fit)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 11732.51, Df = 1, p = < 2.22e-16
```

```
# plot studentized residuals vs. fitted values
spreadLevelPlot(fit)
```

```
## Warning in spreadLevelPlot.lm(fit):
## 882 negative fitted values removed
```

```
## Warning in rlm.default(x, y, weights, method = method, wt.method =
## wt.method, : 'rlm' failed to converge in 20 steps
```



```
##
## Suggested power transformation:  0.002351599
```

Again the varainces are not constant.

Conclusions

The principle component analysis allowed us to observe some interesting trends in the weather data: including the importance of solar radiation and the increased varaince in the winter. Additinoally, the dimensionality reduction alowed us to build a visually appealing model, although principle component regression did not improve the performance of our model.