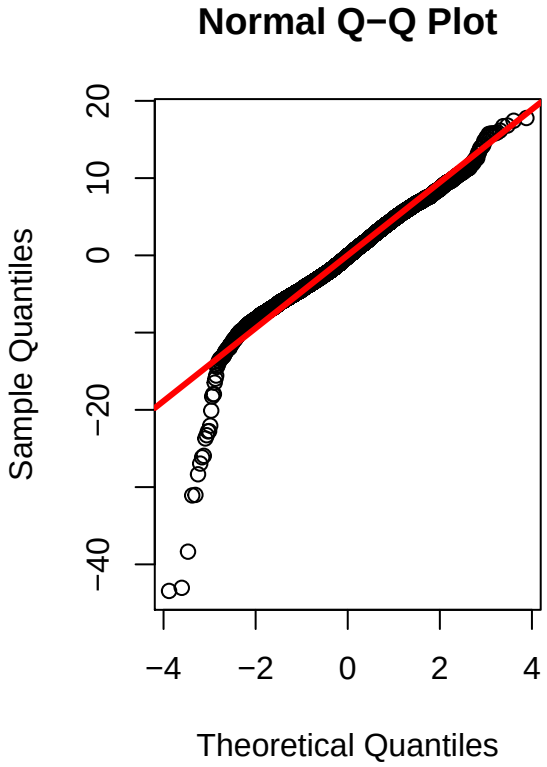
Final Problem 2

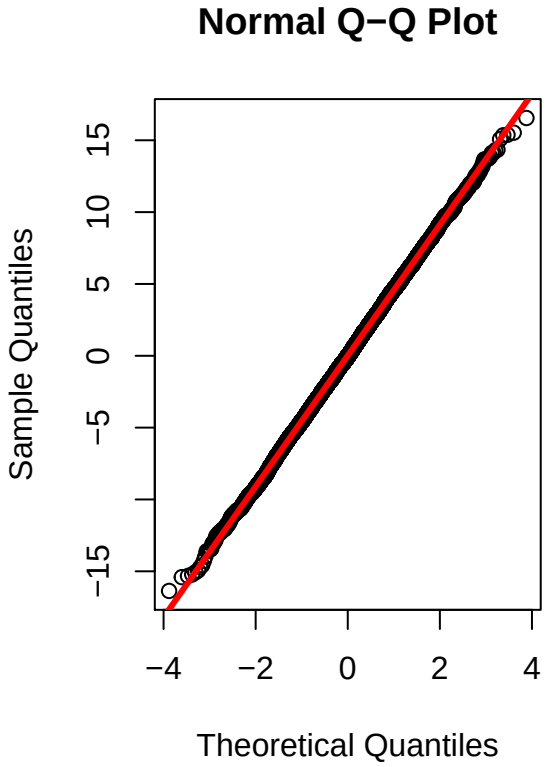
Michael Frasco

December 4, 2014

Fit a linear model of PE on all other variables. Do the proper diagnostics indivate whether it seems that i) the normality assumption is holding

Since the shapiro-wilk test tends to favor rejecting the normality assumption for very large samples, we create a qqplot of the residuals of the fit to examine the normality assumption visually. Below on the right we create a qqplot of the same number of observations of synthetic normals with the same standard deviation. We can uses this graph for comparison.

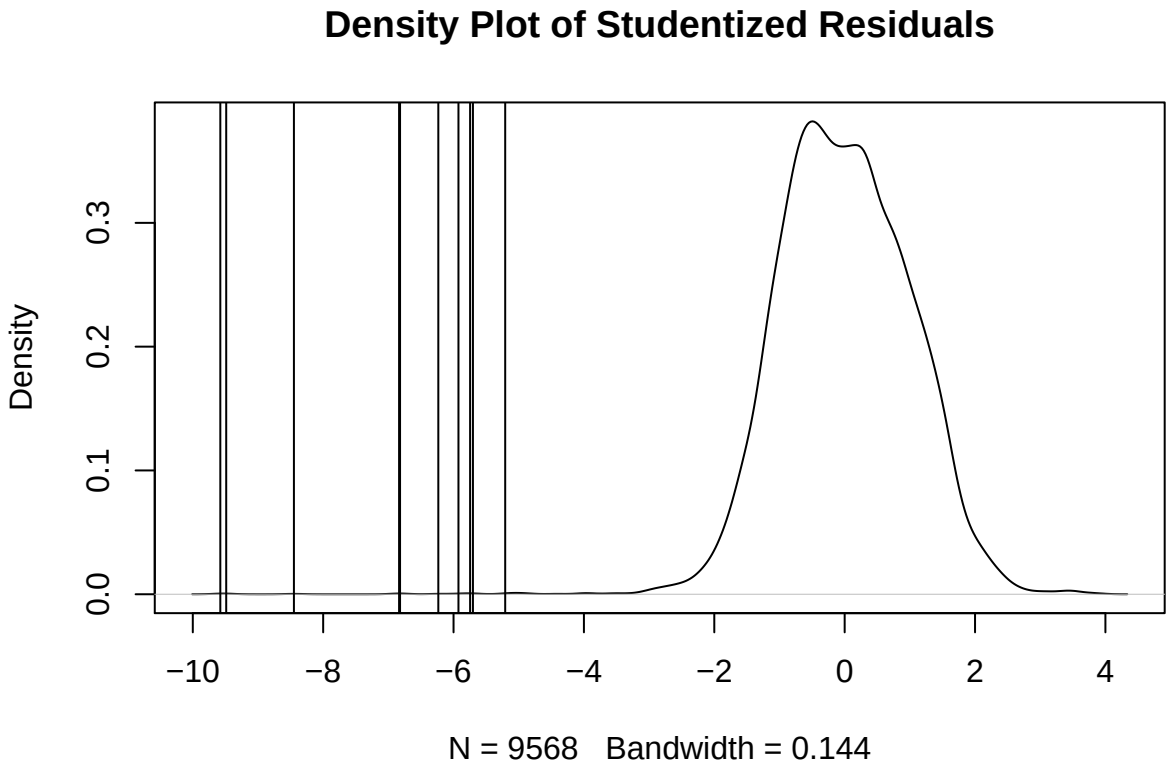




Compared to the plot of synthetic normals, the normal assumption seems to be severely violated. The lower tail experiences residuals with much greater absolute value than would be expected under a normal distribution. Although the rest of the data appear to follow the normal line, the existence of the large tail is enough to reject the normality assumption. Even though, in a sample size of almost 10,000 points, we would expect a handful of observations to have large residuals, the number and magnitude of this lower tail is indicative a a heavy tailed, non-normal distribution.

Are there any outliers?

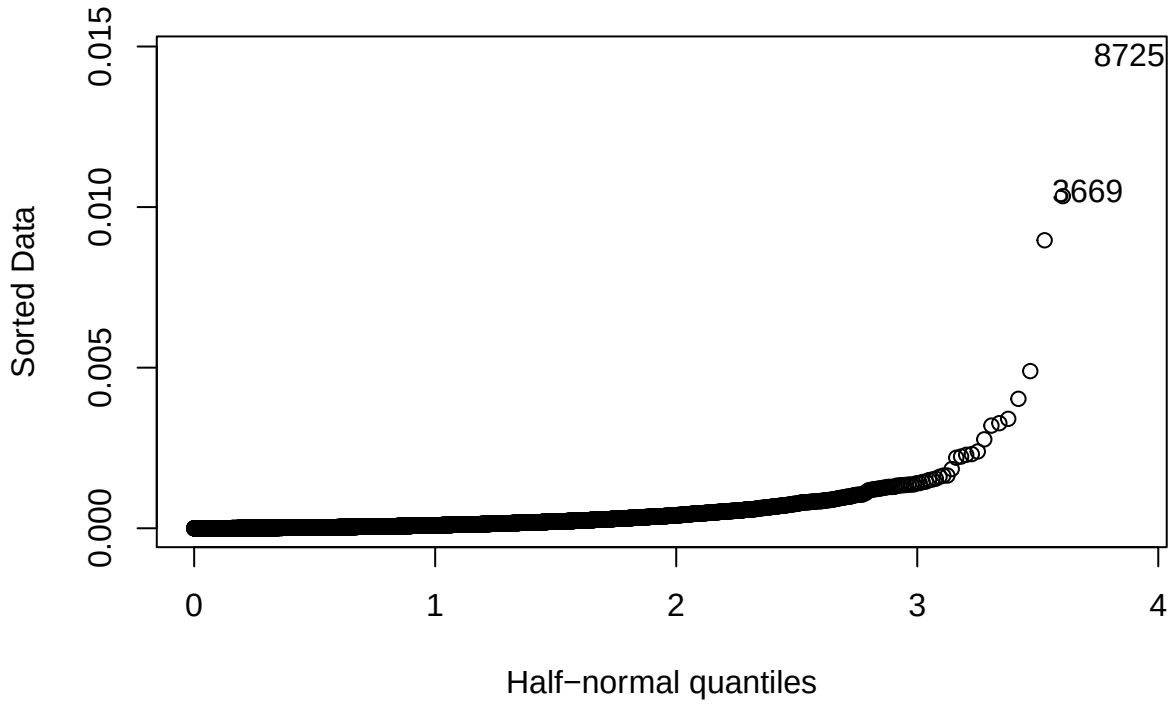
Observations with large residuals are candidate outliers. We can examine the externally studentized residuals (a.k.a. the jackknifed residuals) to see if any residual values are so much larger than the rest of the observations that we can call them outliers. Below we provide a density plot of the studentized residuals, with the value of the largest ten observations represented by vertial lines. Also note that if ignore the normality assumption, the 95% bonferroni cutoff from the t-distribution is -4.558. And this is a very conservative estimate.



Clearly, we have some evidence to claim that these points are outliers. However, since there are so many observations with larger than expected studentized residual values, it may also be that the distribution of studentized residuals has very heavy tails.

We can also check a half-normal plot of the cook statistic values for these observations to visually see if some points break away from the trend.

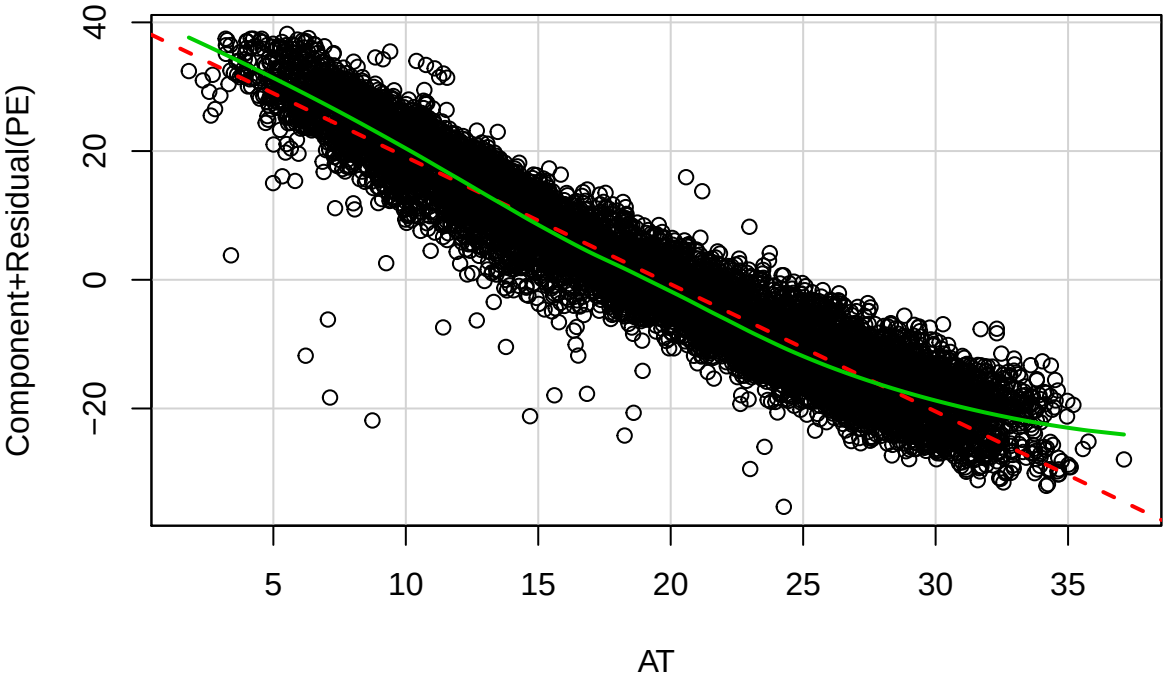
```
cook = cooks.distance(fitCch)
halfnorm(cook)
```



Here, it is quite clear that the largest two points are outliers. However, it is difficult to say whether the other largest points break away from the graph enough. We will make a more formal test in part b).

Do you see any nonlinearity

After looking at partial residual plots for all four predictors, there is no significant evidence if nonlinearity in the relationship with the response. Below I have shown the partial residual plot for ambient temperature, which exhibits slight evidence of nonlinearity



There might be some slight curvature in the partial residual plot for the AT variable. However, it is nothing too extreme. The other three variables all seem to exhibit a linear relationship.

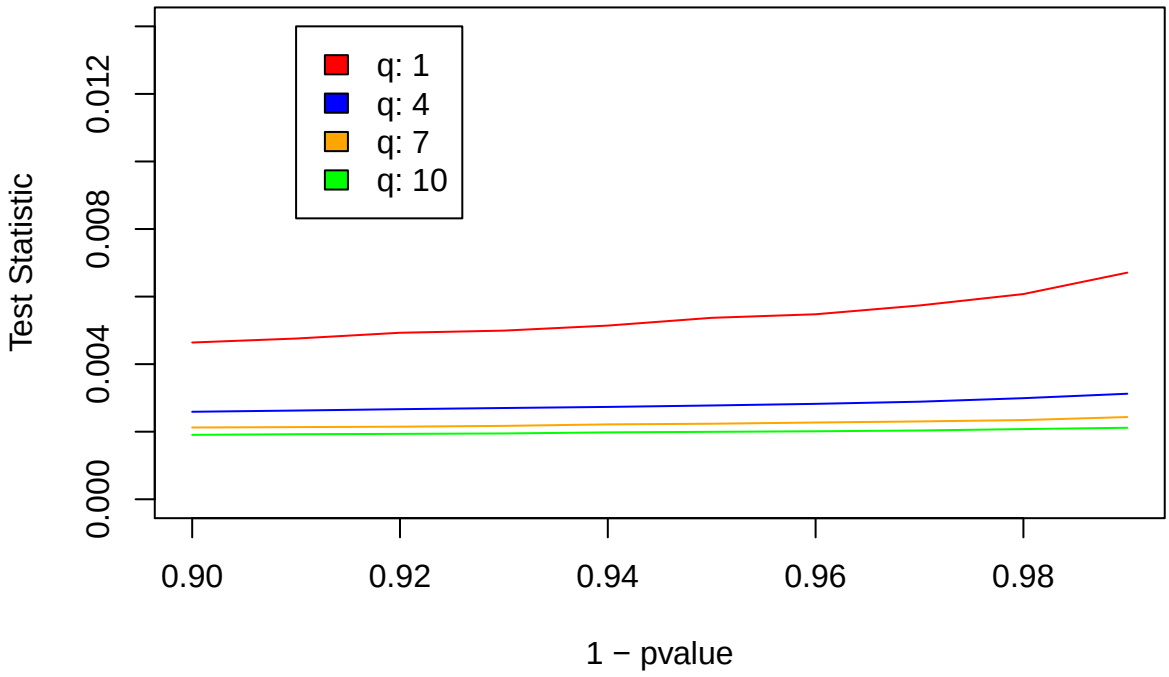
Focus now on deciding whether there are some outliers under the normality assumption. Design a sharp compute-intensive test to identify the worst q outliers by means of i) the Cook Distances and ii) the externally studentized residuals

I will start by examining the cook distances. To accomplish this task, I used resampling techniques to simulate the distribution of the qth worst outlier. I operated under the assumption that the residuals are normally distributed (even though this is questionable). As a result, when constructing a new data set using the coefficients from the original, I added an error term from a normal distribution with mean equal to 0 and standard deviation equal to the standard deviation from the original model. I then calculated the cook statistics for my new model and stored the value of the qth largest in a vector. Repeating the process 1,000 times gave me a distribution for the qth largest cook statistic. In order to test if my observed qth largest cook statistic is significant at a 5% level, I compared it with the value of the 95th percentile of simulated distribution. If it is greater than this value, I claim that it is significant.

```
maxQ = 10
n = nrow(cch)
beta = fitCch$coefficients
sigma = summary(fitCch)$sigma
qCooks = sort(cooks.distance(fitCch), decreasing = TRUE)[1:maxQ]
MM = model.matrix(fitCch)
predictors = MM[, 2:5]
ten_cook_distributions = list()

for(q in 1:maxQ) {
      q_cook_list = numeric()
      for(i in 1:1000) {
            new_Y = MM %*% beta + rnorm(n, mean = 0, sd = sigma)
            new_data = data.frame(new_Y, predictors)
            new_Fit = lm(new_Y ~ ., data = new_data)
            new_cook = sort(cooks.distance(new_Fit), decreasing = TRUE)[q]
            q_cook_list[i] = new_cook
      }
      ten_cook_distributions[[q]] = q_cook_list
}
```

As a result of this function, the statistics for the qth worst outlier are stored as a numeric vector in a giant list. In order to show how the test value depends on the p-value, I find the (1 - p)th percentile in each sorted vector. If I do this for values of p between 0.1 and 0.01, I can plot the results and show how the test value depends on p.



Lastly, I can create a table to indicate whether the qth outlier is significant at various pvalues. Below is such a table. Each row represents an outlier. So the first row represents the largest outlier. Each column represents a significance level of either 0.1, 0.05, or 0.01. A value of 1 in the table indicates that the outlier is significant at this level. A value of 0 indicates that the outlier is not significant.

```
##       [,1] [,2] [,3]
##  [1,]    1    1    1
##  [2,]    1    1    1
##  [3,]    1    1    1
##  [4,]    1    1    1
##  [5,]    1    1    1
##  [6,]    1    1    1
##  [7,]    1    1    1
##  [8,]    1    1    1
##  [9,]    1    1    1
## [10,]    1    1    1
```

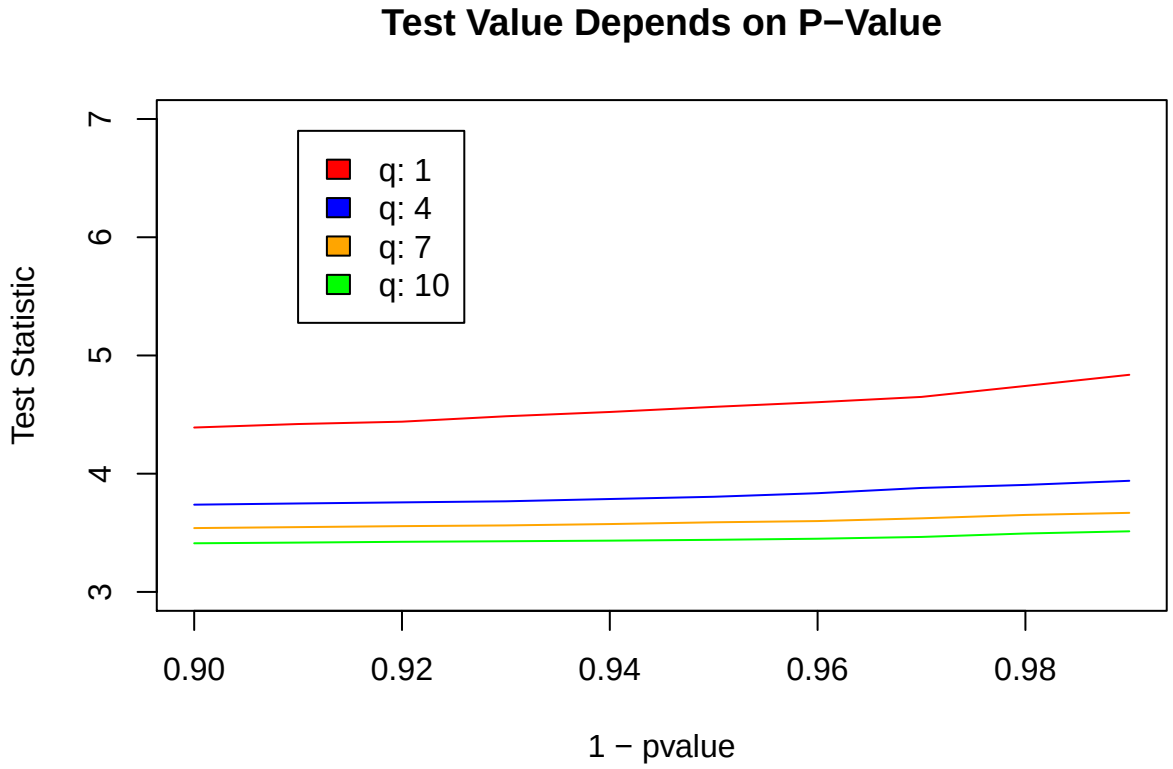
Now, I can repeat the process using the externally studentized residuals.

```
maxQ = 10
n = nrow(cch)
beta = fitCch$coefficients
sigma = summary(fitCch)$sigma
qStudRes = sort(abs(studres(fitCch)), decreasing = TRUE)[1:maxQ]
MM = model.matrix(fitCch)
predictors = MM[, 2:5]
ten_studres_distributions = list()

for(q in 1:maxQ) {
```

```
      q_studres_list = numeric()
      for(i in 1:1000) {
            new_Y = MM %*% beta + rnorm(n, mean = 0, sd = sigma)
            new_data = data.frame(new_Y, predictors)
            new_Fit = lm(new_Y ~ ., data = new_data)
            new_studres = sort(abs(studres(new_Fit)), decreasing = TRUE)[q]
            q_studres_list[i] = new_studres
      }
      ten_studres_distributions[[q]] = q_studres_list
}
```

I create the same plot that I created before. I use the p value to find the value of the (1 - p)th percentile in each sorted distribution. I plot the results below.



Here is the significance table. The organization is the same as before.

```
##       [,1] [,2] [,3]
##  [1,]    1    1    1
##  [2,]    1    1    1
##  [3,]    1    1    1
##  [4,]    1    1    1
##  [5,]    1    1    1
##  [6,]    1    1    1
##  [7,]    1    1    1
##  [8,]    1    1    1
##  [9,]    1    1    1
## [10,]    1    1    1
```

Would you conclude that the departure from the standard assumptions can best be represented by declaring a number of points as outliers, or can you think of a more satisfactory answer?

The results of my simulations in part b indicated that the points I observed were not merely outliers: they were enormous outliers. In fact, the size of the observed cook statistics and studentized residuals were so large that it might be better to consider those data points as extreme anomalies or mistakes, instead of declaring them as outliers. I considered of a handful of outside-the-problem hypotheses to explain the extreme nature of these points.

I do not think it was a data entry error. I inspected the distribution of the variables in the data set and each does not seem to contain absurd values.

Since the data contains points collected each hour over 6 years, it might be that the outliers could be explained temporally. Each observation represents the output from a power plant for a given hour. I notice that all of the biggest residuals have negative values, meaning that the observed value is much less than the predicted value. Perhaps the powerplant was not operating at full capacity during these hours. Perhaps all of the outliers occured at the same time of day. Perhaps the outliers occured in the same week every year when the power plant was being tested. I would like to have information about the specific times these outliers occured and what other events (natural disasters, national holidays, etc.) occured during these times. It might be that the ambient temperature happened to drop suddenly for a single hour or maybe there was a data measurement error in the thermometer.