Adam Gincel

Professor Li

MA 331

PCB Study Report

Polychlorinated biphenyls, or PCBs, are collections of synthetic compounds, also known as congeners. They are very toxic to young children and the environment. Despite no longer being produced in the United States, they are still present in many places, particularly marine wildlife. As such, the EPA measures levels of PCBs in fish as a means of determining PCB presence. There are many different PCBs present in fish, so to compare them all more simply the Toxic Equivalent Score, or TEQ, was invented. Estimating TEQ efficiently based on a few measurements greatly speeds along the process of measuring how safe, or unsafe, fish may be to consume. Doing this successfully would also greatly reduce costs.

Graphs and numerical summaries described the variables and compared them to one another. After that, a multiple regression was used to see how many measurements were needed to accurately predict the total PCB content of a given fish. Outliers were later accounted for, making the measurements even more accurate. Later in the assignment the log of each measurement was used to further improve predictive accuracy using different datasets, which decreased standard error even further.

Using PCBs 52, 118, 138, and 180 and removing two outliers, I got an R-squared value of 0.9941, denoting a very highly accurate multiple regression prediction. Later on I was asked to use the TEQPCB, TEQDIOXIN, and TEQFURAN rows to predict total TEQ content, which resulted in an R-squared value of 1, meaning these three rows were able to perfectly accurately predict the Toxic Equivalent Scores of marine wildlife. Through this and other values, we were able to successfully predict information with less information than given, which could potentially allow for accurate prediction with less comprehensive measurement, saving both time and money in the short- and long-run.

Adam Gincel

MA331

17 December 2016

*I pledge my honor that I have abided by the Stevens Honor System*

Statistics Final Project 11.42 – 11.52

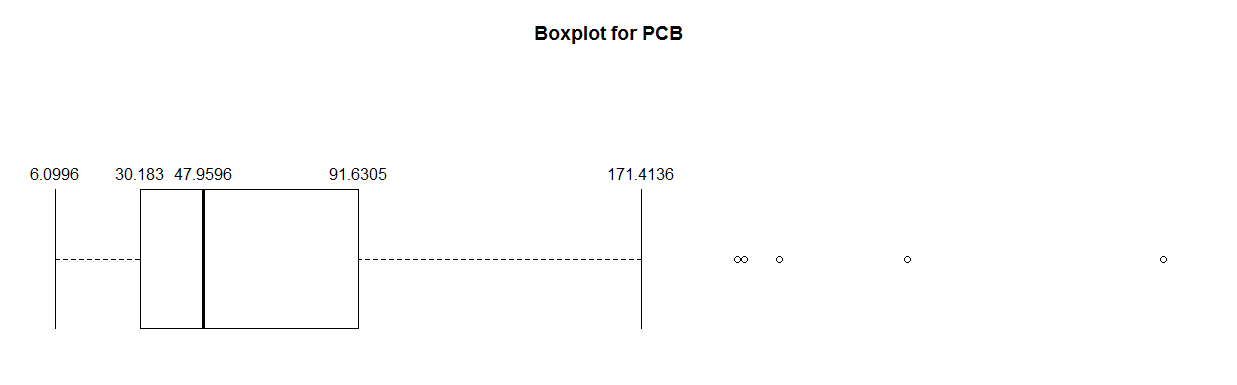
Chapter 11:

42. Consider the following variables: PCB (the total amount of PCB) and four congeners: PCB52, PCB118, PCB138, and PCB180.

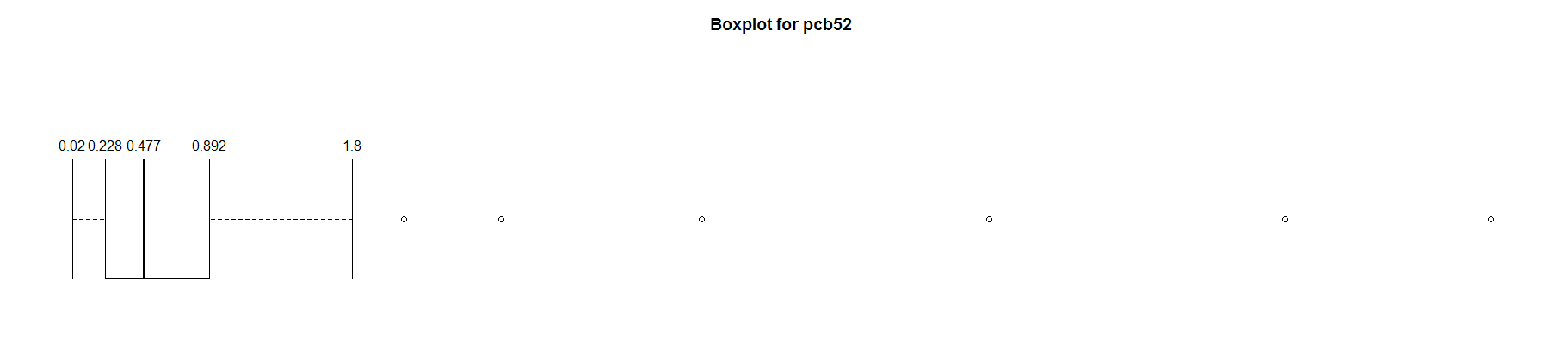
1. Using numerical and graphical summaries, describe the distribution of each of these variables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **PCB** | **PCB52** | **PCB118** | **PCB138** | **PCB180** |
| **Min** | 6.0996 | 0.02 | 0.236 | .64 | 0.395 |
| **Q1** | 30.138 | 0.228 | 1.49 | 3.18 | 1.24 |
| **Median** | 47.9596 | 0.477 | 2.42 | 4.92 | 2.69 |
| **Q3** | 91.6305 | 0.892 | 3.89 | 8.65 | 4.49 |
| **Maximum (not outlier)** | 171.4136 | 1.8 | 6.9 | 13.1 | 9.33 |
| **Mean** | 68.46737 | 0.9580435 | 3.256319 | 6.826812 | 4.15842 |
| **Standard Deviation** | 59.3906 | 1.59828 | 3.019118 | 5.862651 | 4.986439 |
| **Variance** | 3527.243 | 2.5545 | 9.115071 | 34.37068 | 24.86458 |

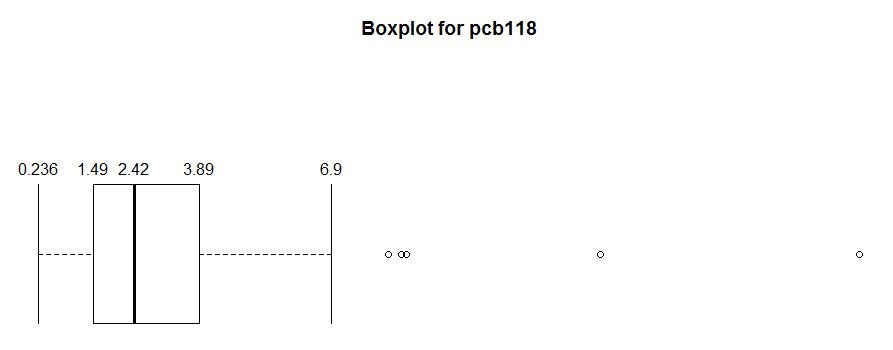
**Boxplots:**



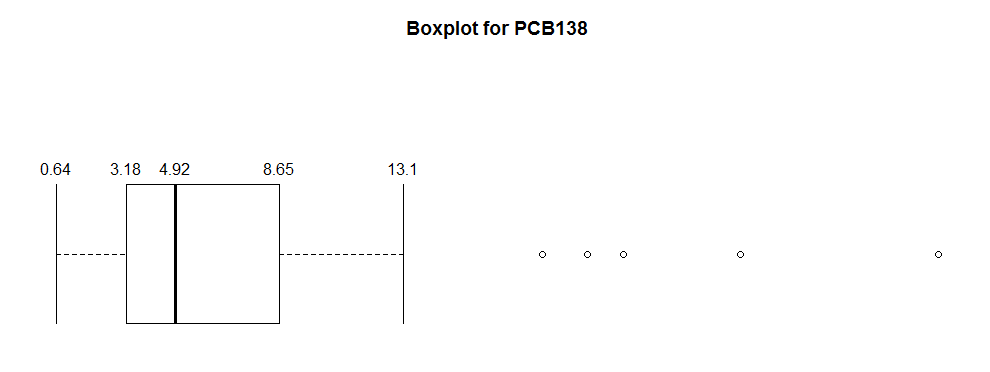
PCB is right skewed. The variance for the dataset is 3527, which is enormous.



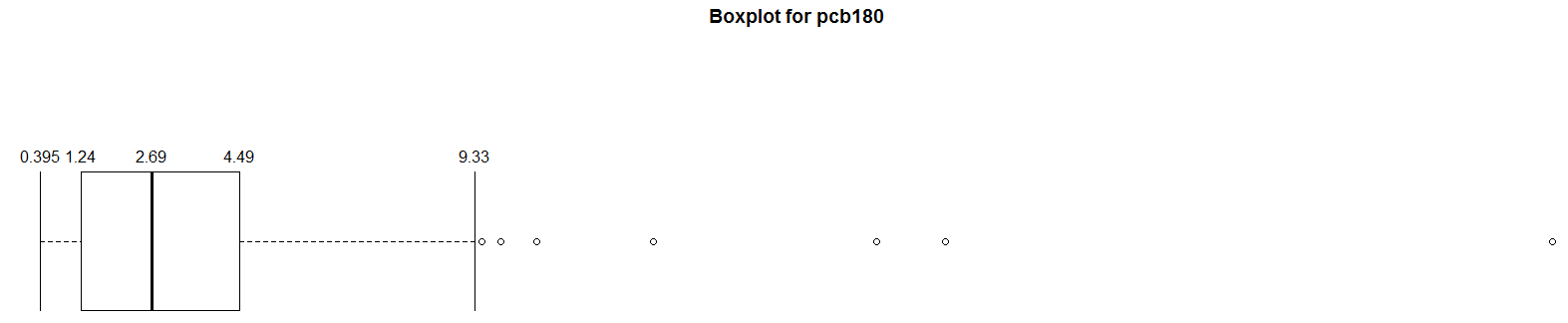
PCB52 is incredibly right skewed, with many spaced out outliers. This suggests a cluster of small numbers with a few very large values throughout.



PCB118 is right skewed but fairly normal in its distribution.

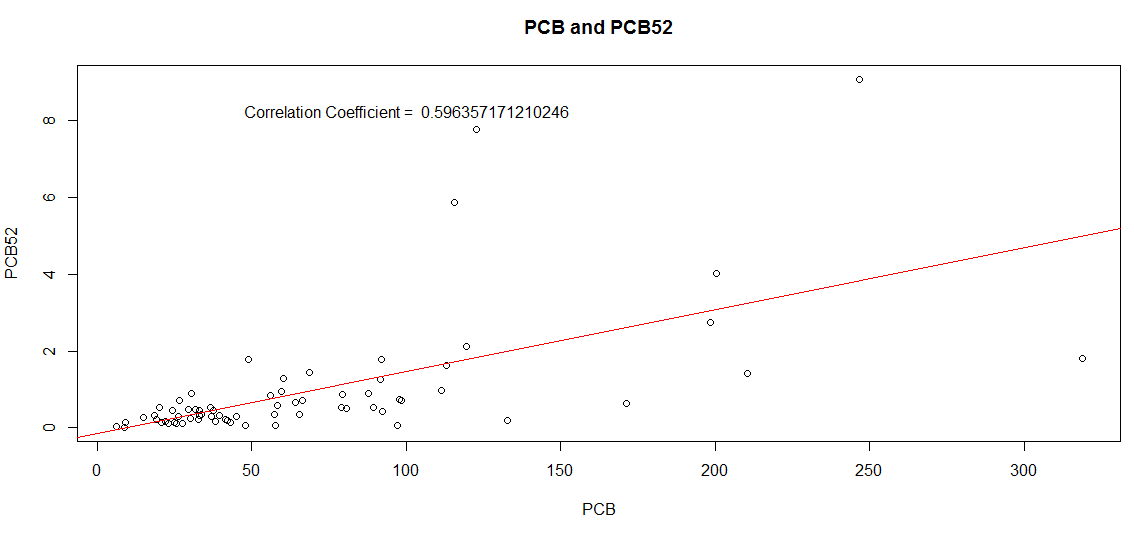


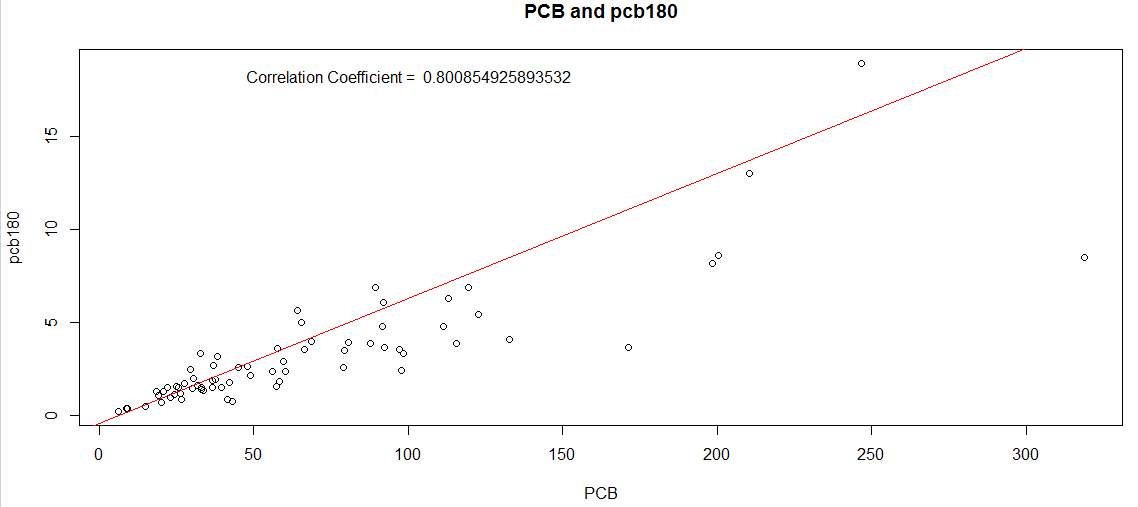
PCB138 is right skewed as well, with outliers that are not very severe.

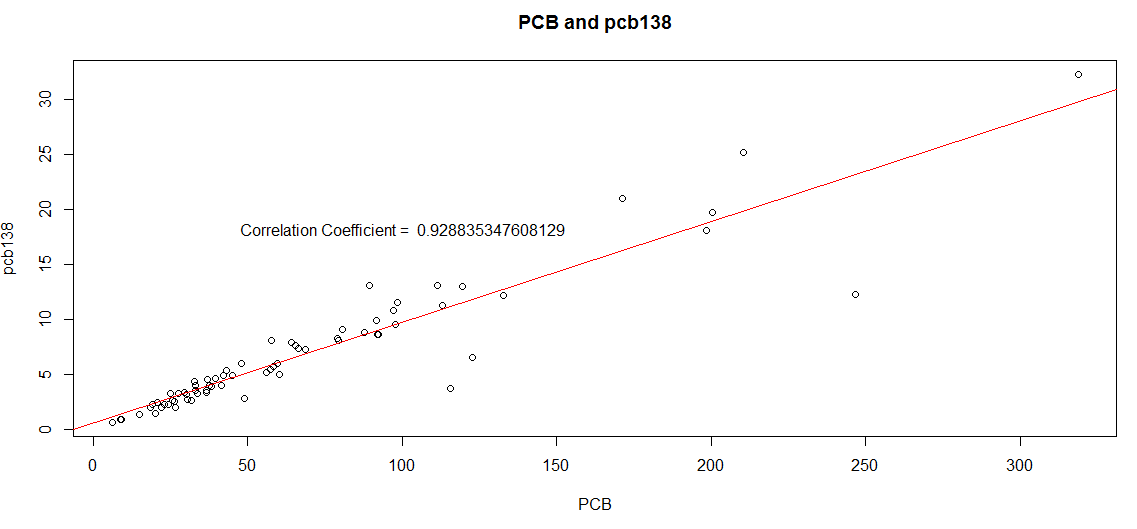


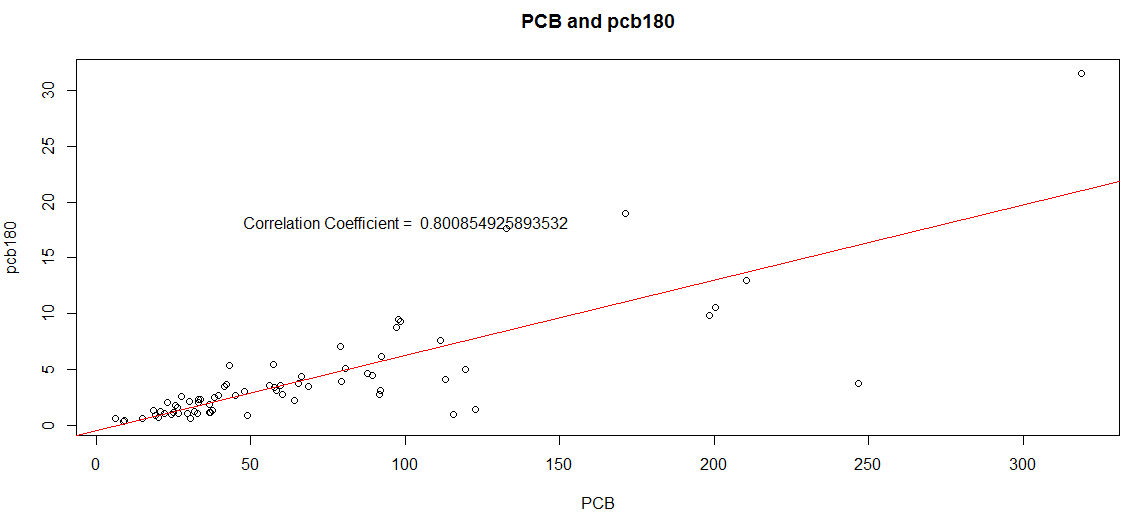
PCB180 is very right skewed, with one particularly large outlier.

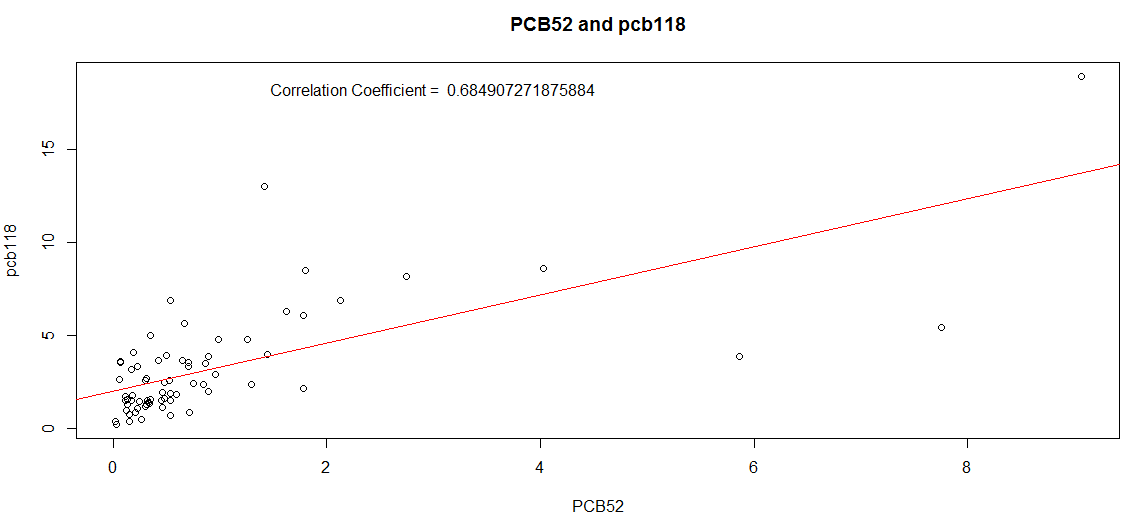
2. Using numerical and graphical summaries, describe the relationship between each pair of variables.

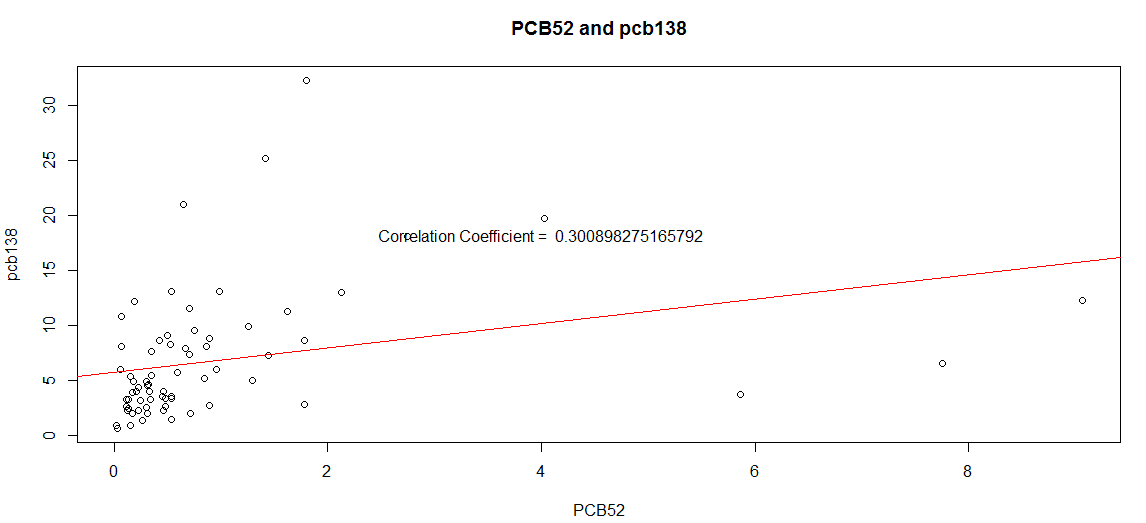
**Comparative Scatterplots:**

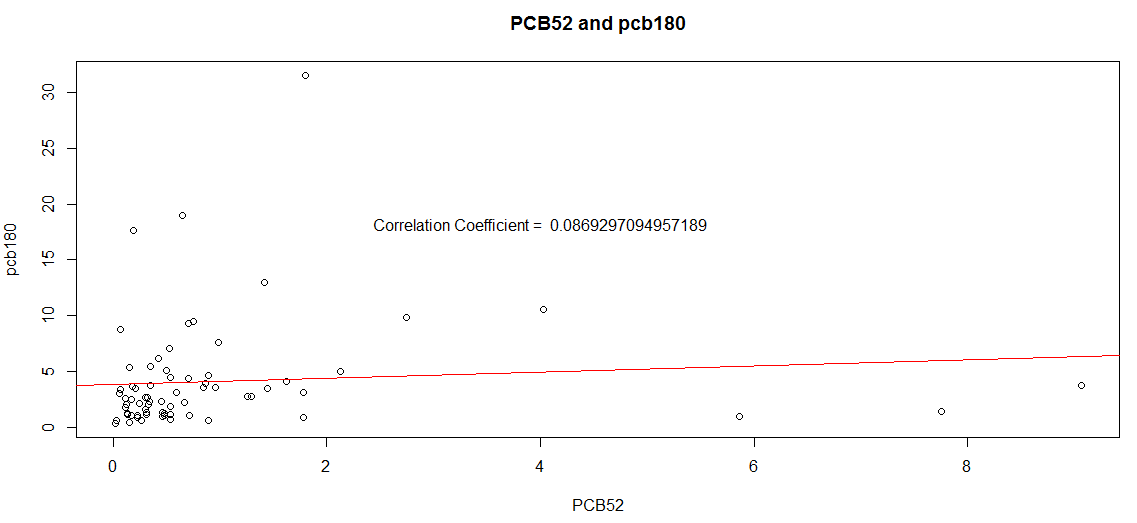




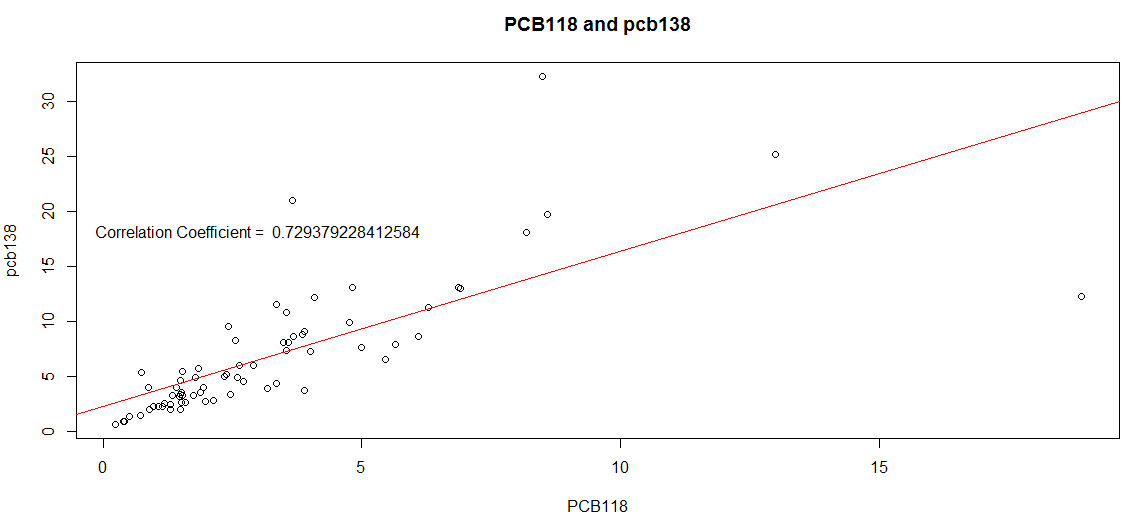


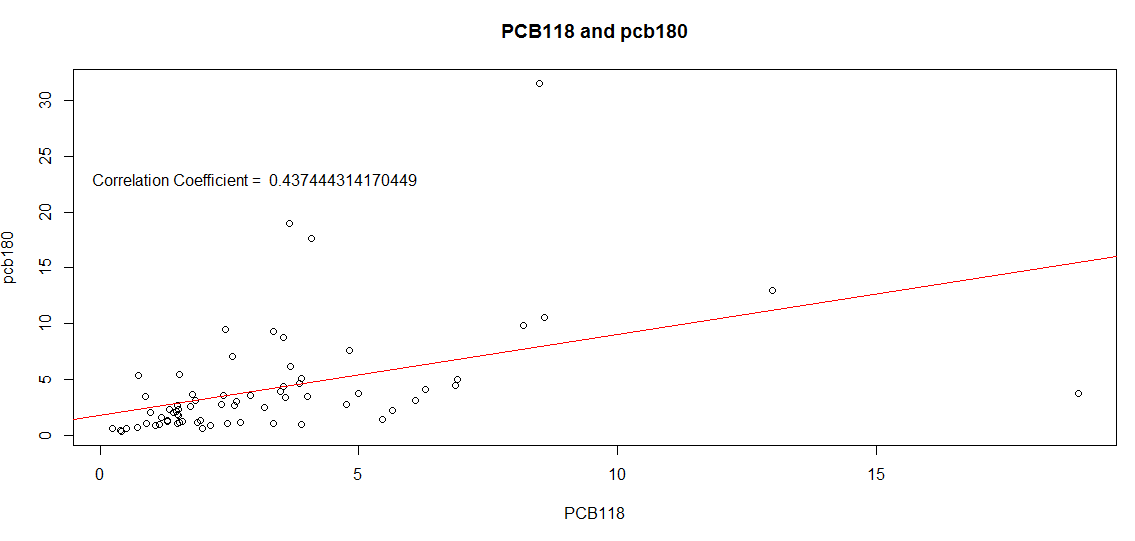


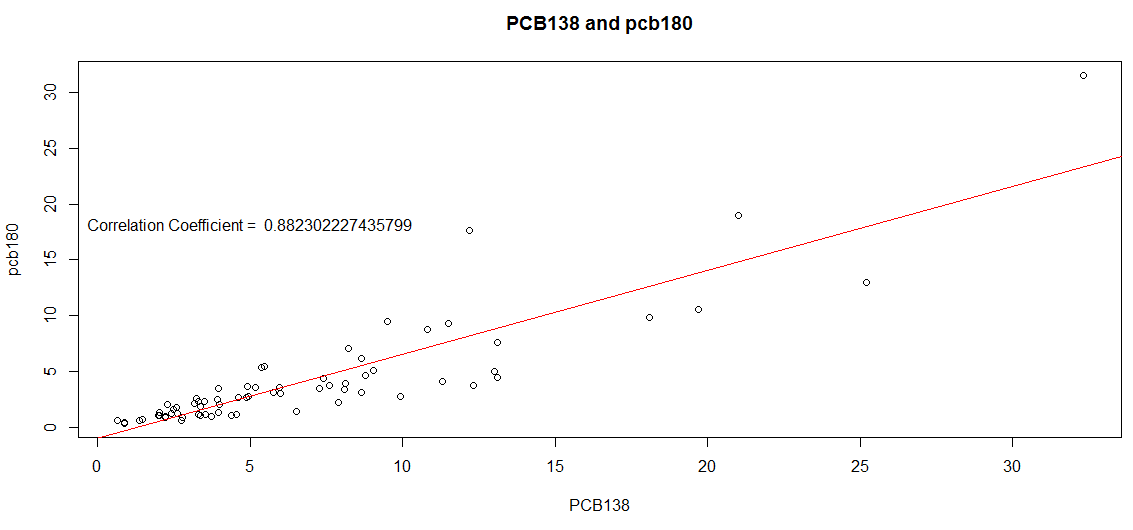




At 8% correlation, PCB52 and PCB180 are not at all correlated to each other.





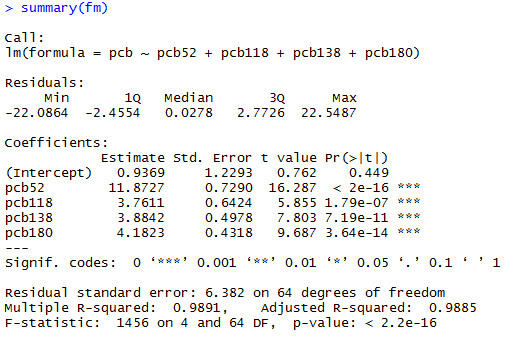


43. Use the four congeners PCB52, PCB118, PCB138, and PCB180 in multiple regression to predict PCB.

1. Write the statistical model for this analysis. Include all assumptions.

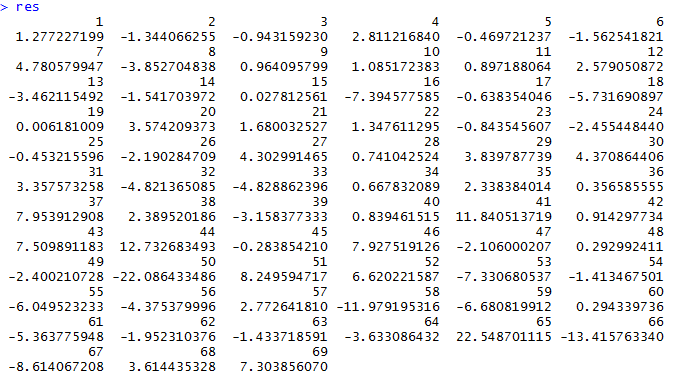
yi = ꞵ0 + ꞵ1xi1 + ꞵ2xi2 + ꞵ3xi3 + ꞵ4xi4 + εi, i from 1 to 69. εi are independent N(0, sigma) random normally distributed variables.

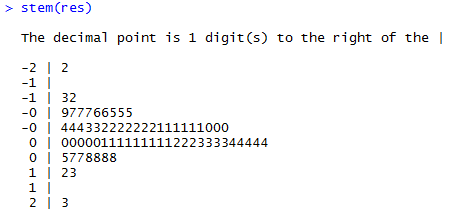
2. Run the regression and summarize the results.



The coefficients are 0.9369, 11.87, 3.76, 3.88, and 4.1. Most of these are significantly different from 0, except for 0.9369.

3. Examine the residuals. Do they appear to be approximately Normal? When you plot them versus each of the explanatory variables, are any patterns evident?





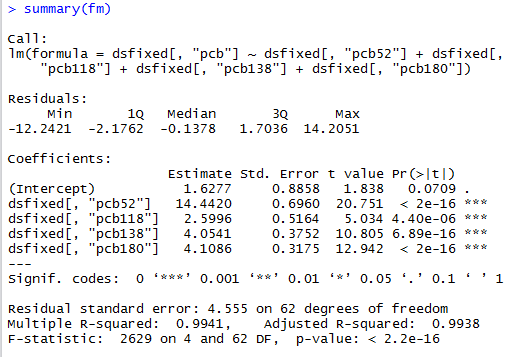
The patterns are roughly normal, as per the stem graph, though there appear to be two outliers, which are -22 and 23. No real patterns emerge.

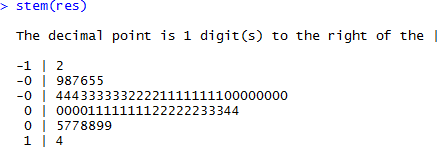
44. The examination of the residuals in part (c) of the previous exercise suggests that there may be two outliers, one with a high residual and one with a low residual.

1. Because of safety issues, we are more concerned about underestimating PCB in a specimen than about overestimating. Give the specimen number for each of the two suspected outliers. Which one corresponds to an overestimate of PCB?

The two specimens are #50 and #65. As specimen 65 is greatly higher residual than most other values, it corresponds to an overestimate of PCB.

2. Rerun the analysis with the two suspected outliers deleted, summarize these results, and compare them with those you obtained in the previous exercise.

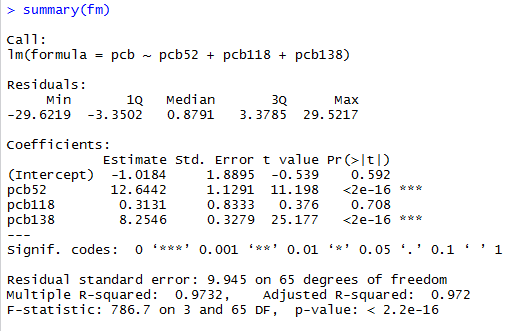




By removing the two outliers, we see the same multiple regression go from R-squared = 0.98 to R-squared = 0.99, making the model even better at predicting the value of the PCB dataset. The stems now make an almost perfect normal distribution.

45. Run a regression to predict PCB using the variables PCB52, PCB118, and PCB138. Note that this is similar to the analysis you did in Exercise 11.43, with the change that PCB180 is not included as an explanatory variable.

1. Summarize the results.



2. In this analysis, the regression coefficient for PCB118 is not statistically significant. Give the estimate of the coefficient and the associated P-value.

The estimate of the coefficient of PCB118 is 0.3131, which is significantly close to 0. The associated P value is seen in the table as well, which is 0.708.

3. Find the estimate of the coefficient for PCB118 and the associated P-value for the model analyzed in Exercise 11.43.

In exercise 11.43, the coefficient for PCB118 was 3.76, and its associated P-value was 1.79\*e^-7, which is significant at the 5%, or even 1% level.

4. Using the results in parts (b) and (c), write a short paragraph explaining how the inclusion of other variables in a multiple regression can have an effect on the estimate of a particular coefficient and the results of the associated significance test.

By removing variables from a multiple regression, it is possible to lose a piece of the bigger picture. We see a drop in the significance of other variables, that may be relevant to the whole dataset, but not a subset of it. As such, it is important to include as much data as possible for a more accurate multiple regression, coefficient, and significance test.

46. Dioxins and furans are other classes of chemicals that can cause undesirable health effects similar to those caused by PCB. Three types of chemicals are combined using toxic equivalent scores (TEQs), which attempt to measure the health effects on a common scale. The PCB data file contains TEQs for PCB dioxins, and furans. The variables are called TEQPCB, TEQDIOXIN, and TEQFURAN. The data file also includes the total TEQ, desired to be the sum of these three variables.

1. Consider using a multiple regression to predict TEQ using the three components TEQPCB, TEQDIOXIN, and TEQFURAN as explanatory variables. Write the multiple regression model in the form TEQ = ꞵ0 + ꞵ1TEQPCB + ꞵ2TEQDIOXIN + ꞵ3TEQFURAN + 𝜖. Give numerical values for the parameters ꞵ0, ꞵ1, ꞵ2, and ꞵ3.

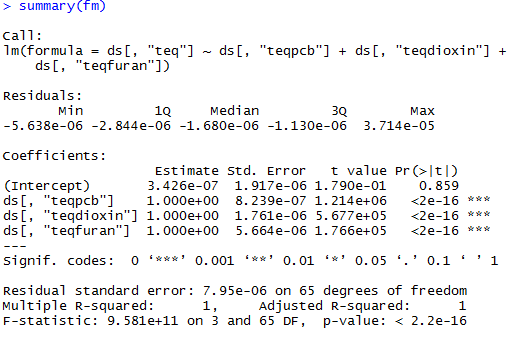
Model is: TEQ = 3.426\*e^-6 + (1)TEQPCB + (1)TEQDIOXIN + (1)TEQFURAN + 𝜖

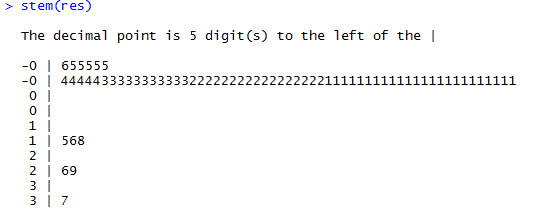
ꞵ0 almost equals 0, ꞵ1, ꞵ2, and ꞵ3 all equal 1.

2. The multiple regression model assumes that the 𝜖’s are Normal with mean zero and standard deviation 𝜎. What is the numerical value of 𝜎?

𝜎 = SE/sqrt(n) = 7.95e-06/sqrt(69) approximately = 0. As such, the value of 𝜎 is 0.

3. Use software to run this regression and summarize the results.

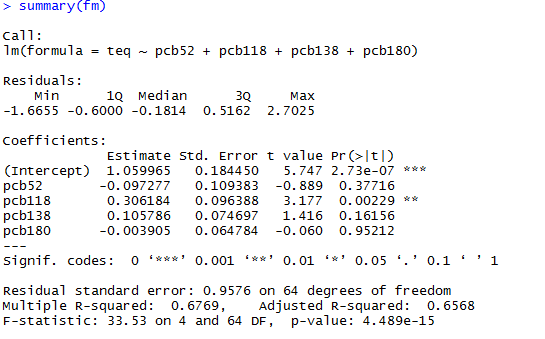




The R-squared value for this approximation is literally 1. As such, these values are able to perfectly approximate TEQ. The distribution of the residuals is a right-skewed normal distribution.

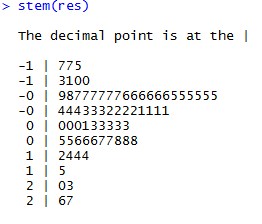
47. The information summarized in TEQ is used to assess and manage risks from these chemicals. For example, the World Health Organization (WHO) has established the tolerable daily intake (TD1) as 1 to 4 TEQs per kilogram of body weight per day. Therefore, it would be very useful to have a procedure for estimating TEQ using just a few variables that can be measured cheaply. Use the for PCB congeners PCB52, PCB118, PCB138, and PCB180 in a multiple regression to predict TEQ. give a description of the model and assumptions, summarize the results, examine the residuals and write a summary of what you have found.

The model is: TEQi = ꞵ0 + ꞵ1PCB52 + ꞵ2PCB118 + ꞵ3PCB138 + ꞵ4PCB180 + εi, i from 1 to 69. εi are independent N(0, sigma) random normally distributed variables.



The coefficients are -0.097, 0.306, 0.106, and -0.0039, making the equation:

TEQ = 1.06 – 0.097PCB52 + 0.306PCB118 +0.106PCB138 - 0.0039PCB180.



As per this stem plot of the residuals, they seem to be slightly right skewed. No clear patterns emerge.

48. Because distributions of variables such as PCB, the PCB congeners, and TEQ tend to be skewed researchers frequently analyze the logarithms of the measured variables. Create a data set that has the logs of each of the variables in the PCB data file. Note that zero is a possible value for PCB126. Most software packages will eliminate these cases when you request a log transformation.

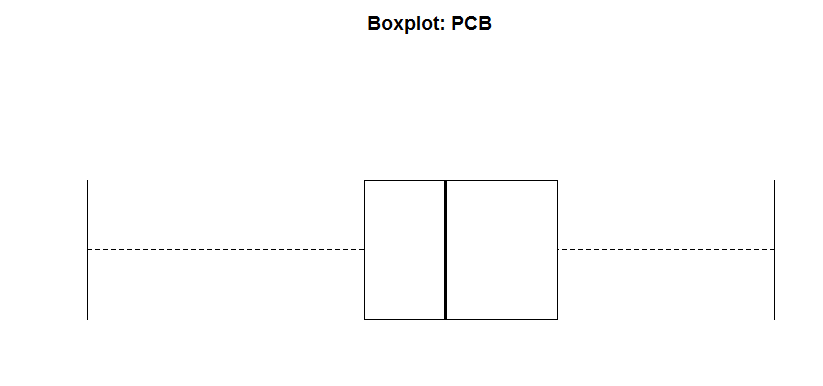
1. If you do not do anything about the 16 zero values of PCB126, what does your software do with these cases? Is there an error message of some kind?

R sets all of the zeroes in PCB126 to negative infinity; as this is the value of log(0). There is no error message upon setting this, though it could cause problems when using the data in equations and graphs.

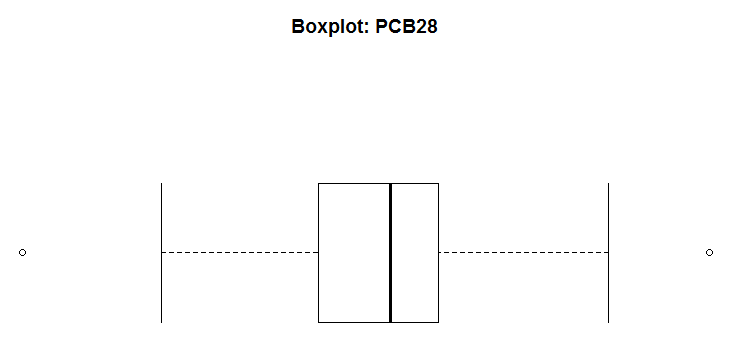
2. If you attempt to run a regression to predict the log of PCB using the log on PCB126 and the log of PCB52, are the cases with the zero values of PCB126 eliminated? Do you think that this is a good way to handle this situation?

These cases are not eliminated. As such, doing this throws an error with the software. I believe this is reasonable, as arbitrarily removing problematic datapoints without notifying the user could lead to confusion in certain cases. Notifying the user that something is wrong with the dataset, and leaving the solution up to them, is a much safer solution.

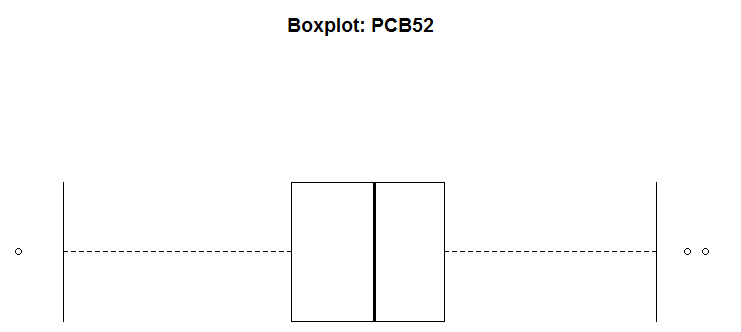
3. The smallest nonzero value of PCB126 is 0.0052. One common practice when taking logarithms of measured values is to replace the zeros by one-half of the smallest observed value. Create a logarithm data set using this procedure; that is, replace the 16 zero values of PCB126 by 0.0026 before taking logarithms. Use numerical and graphical summaries to describe the distributions of the log variables.

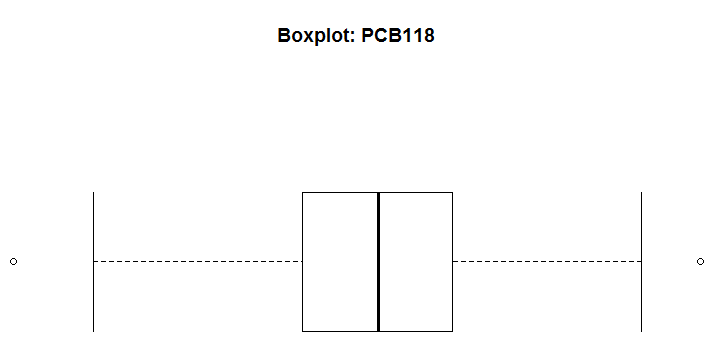




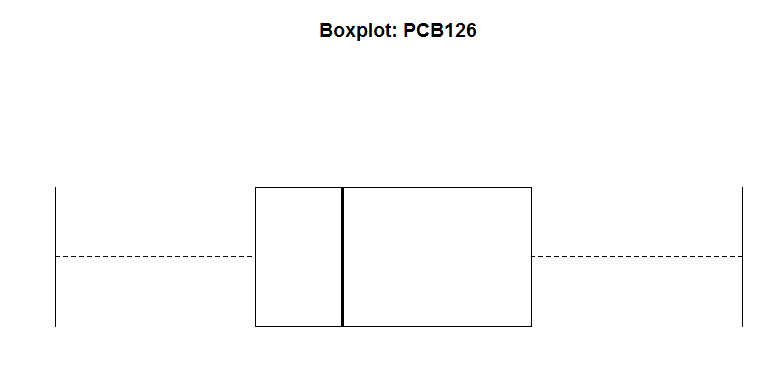




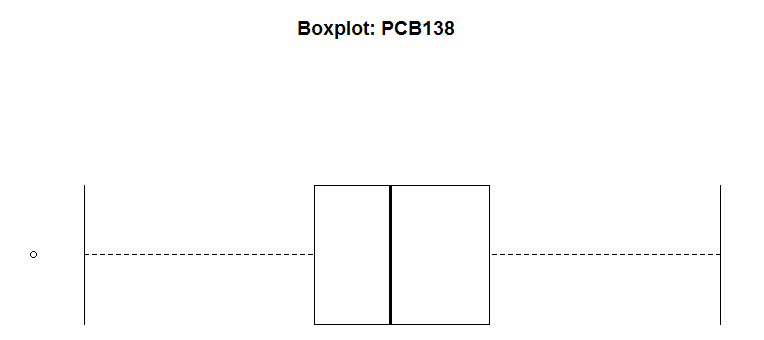




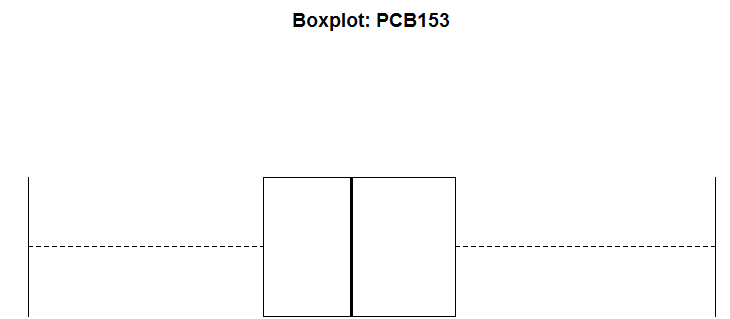




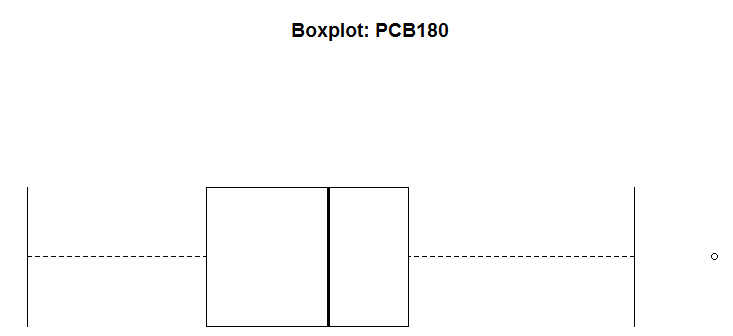










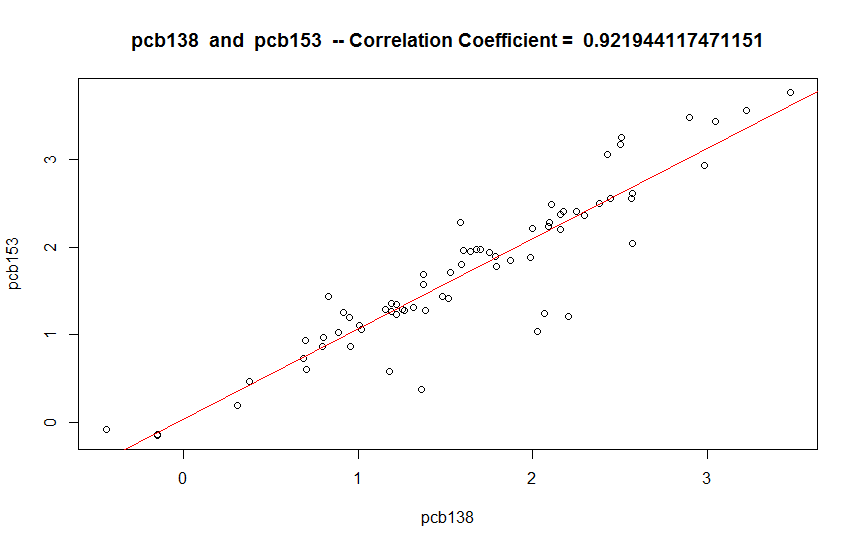
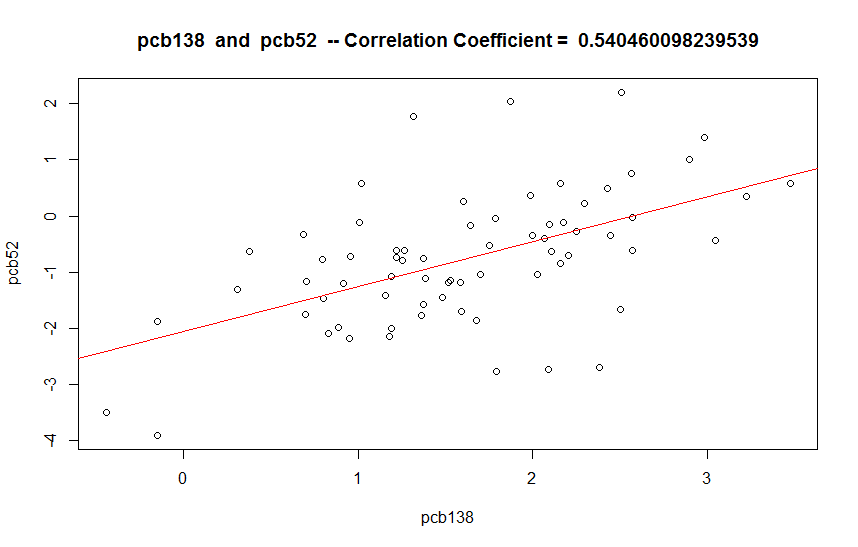
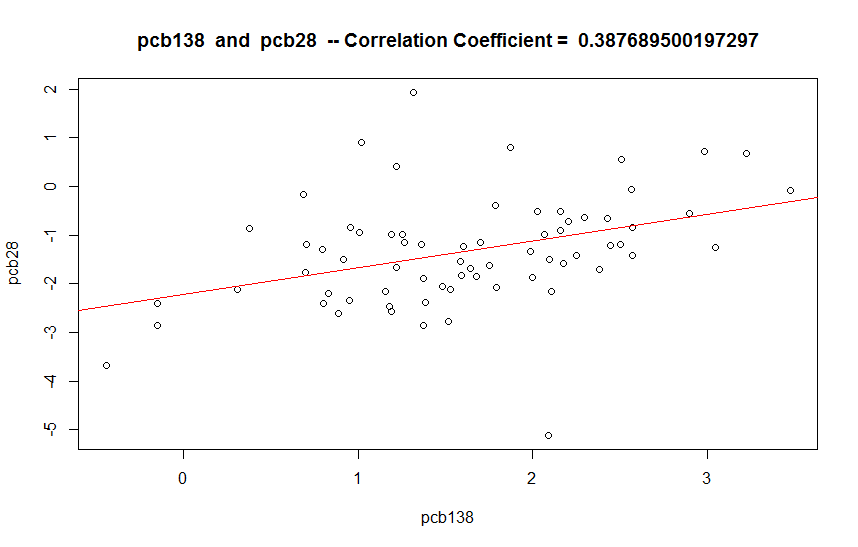
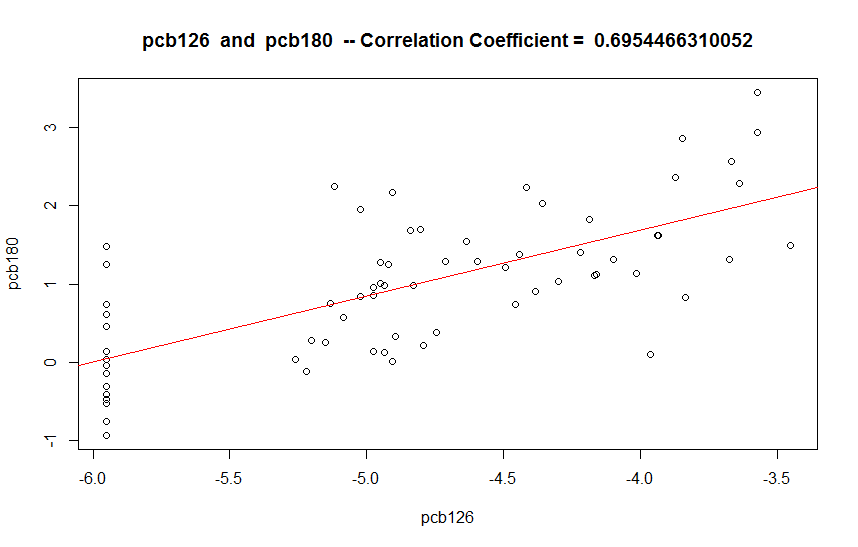
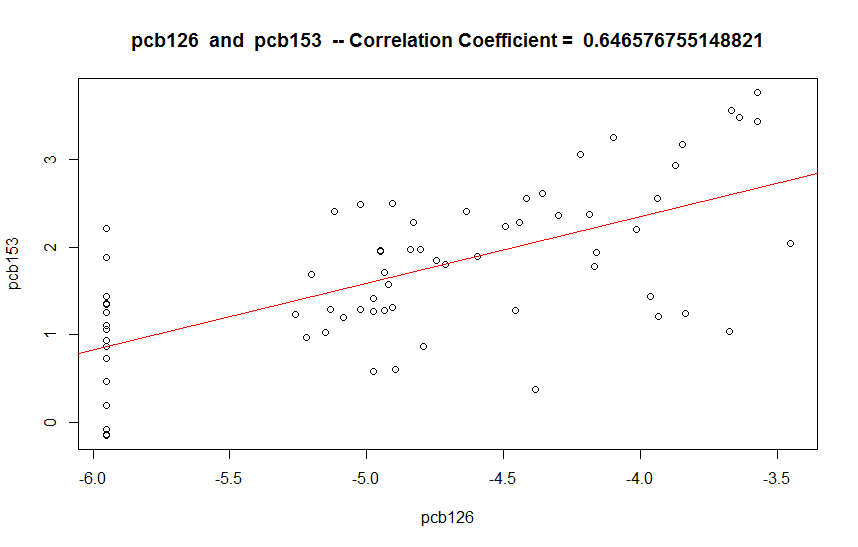
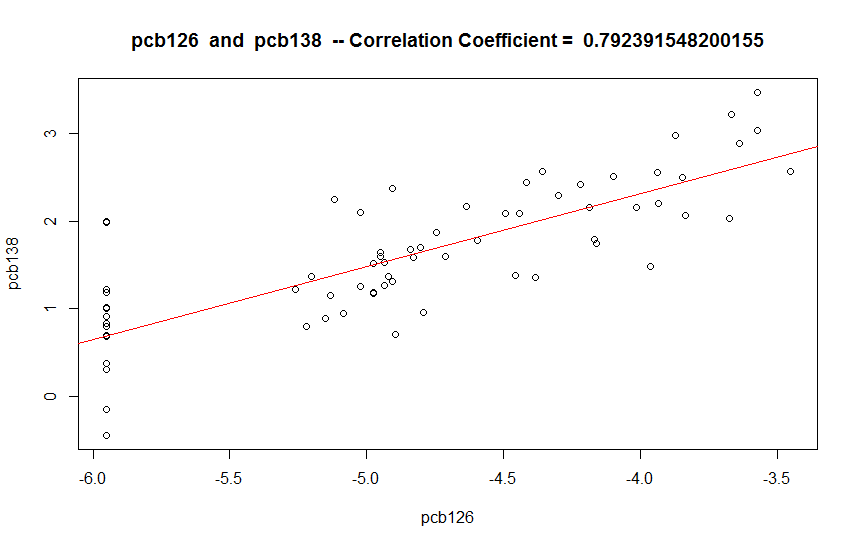
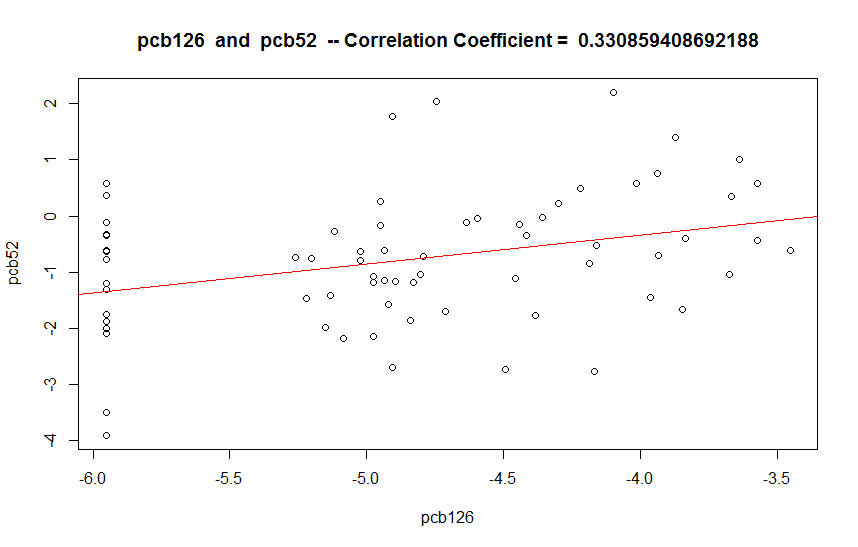
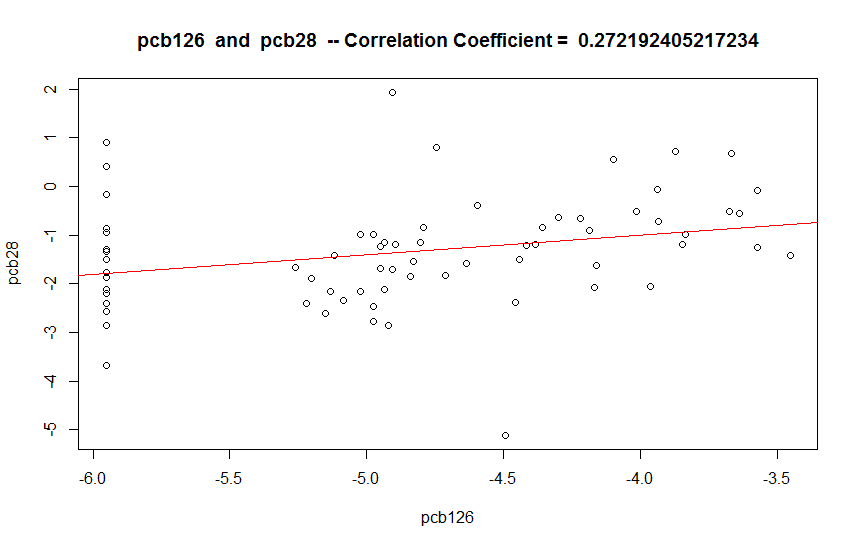
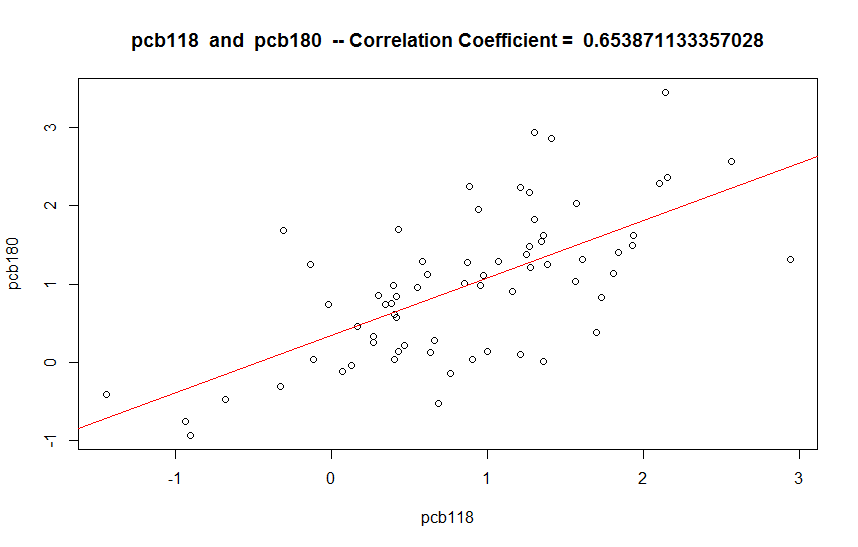
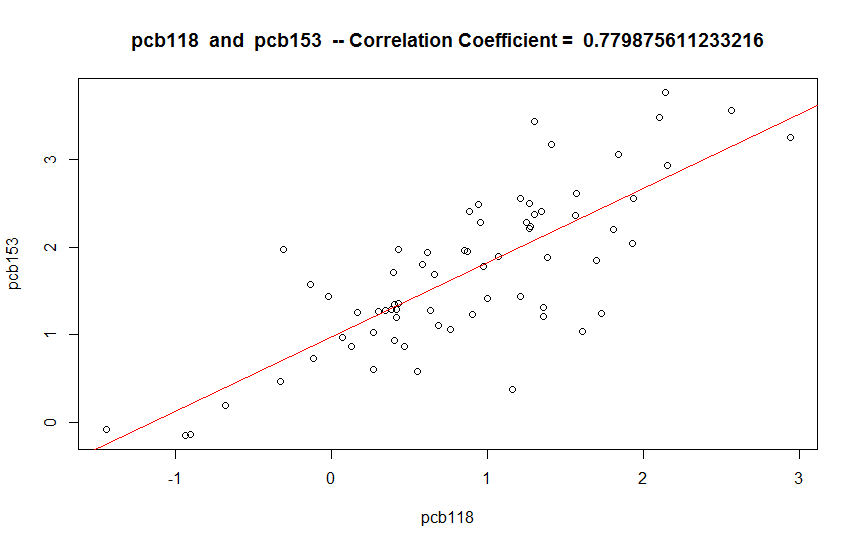
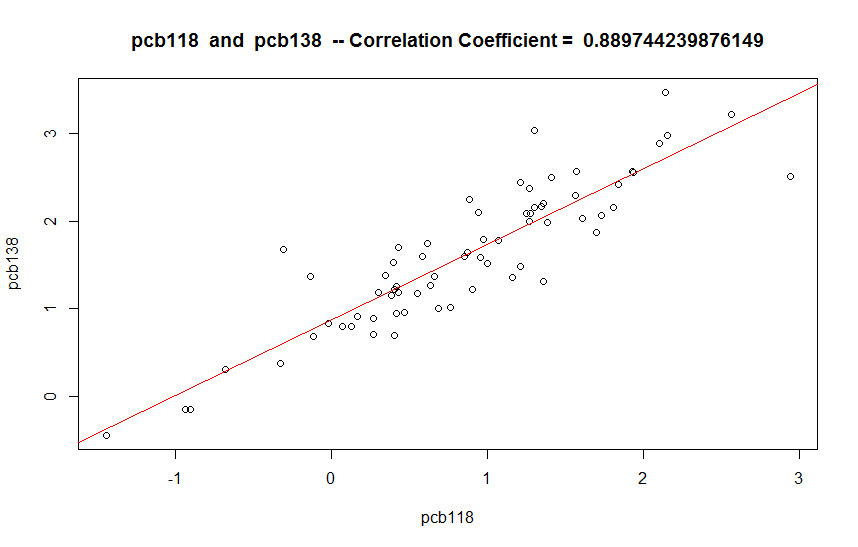
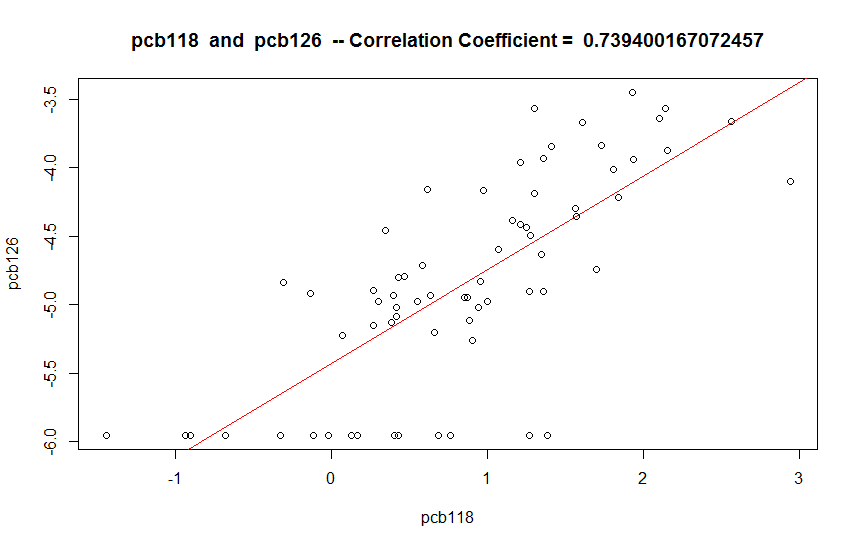
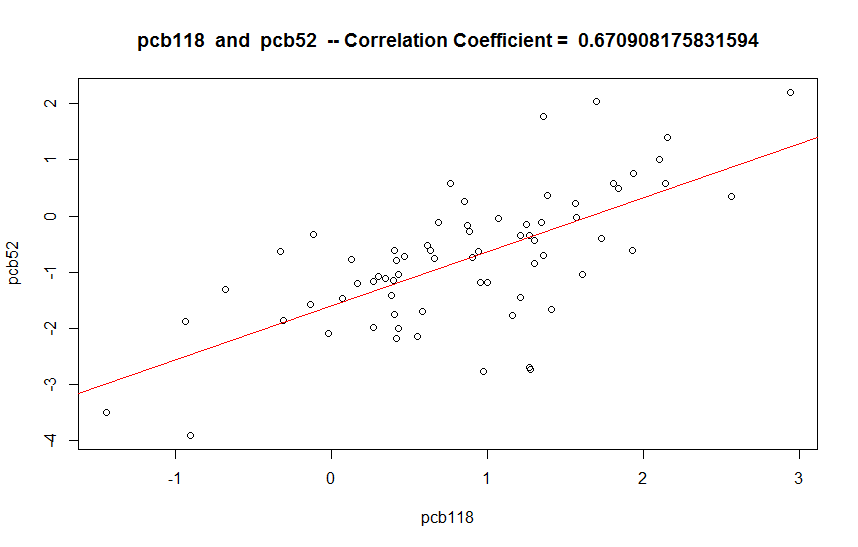
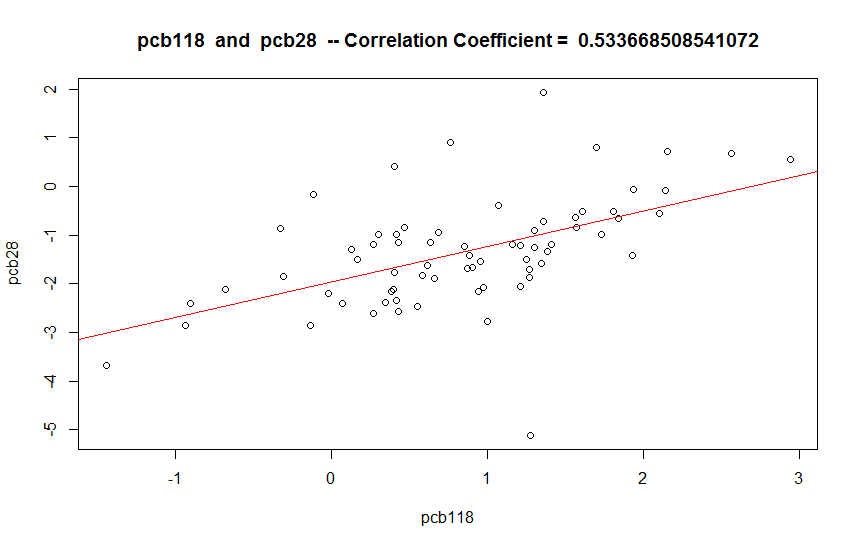
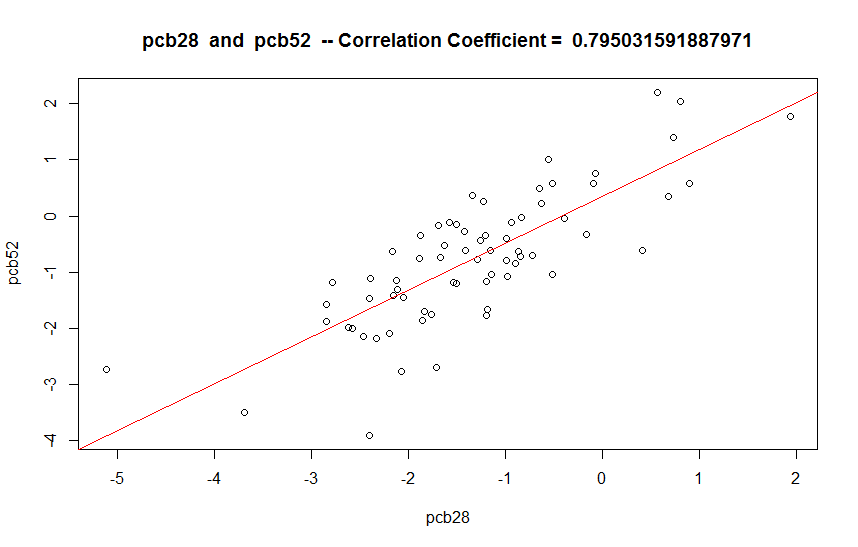
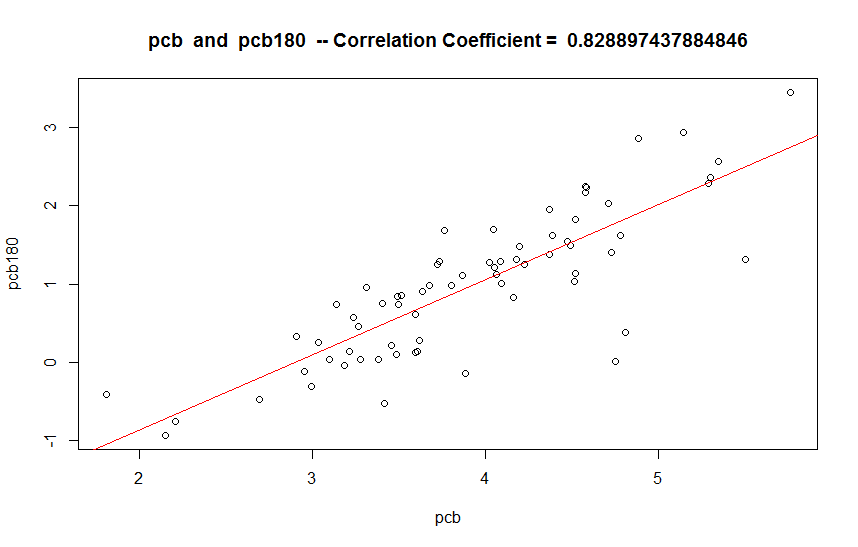
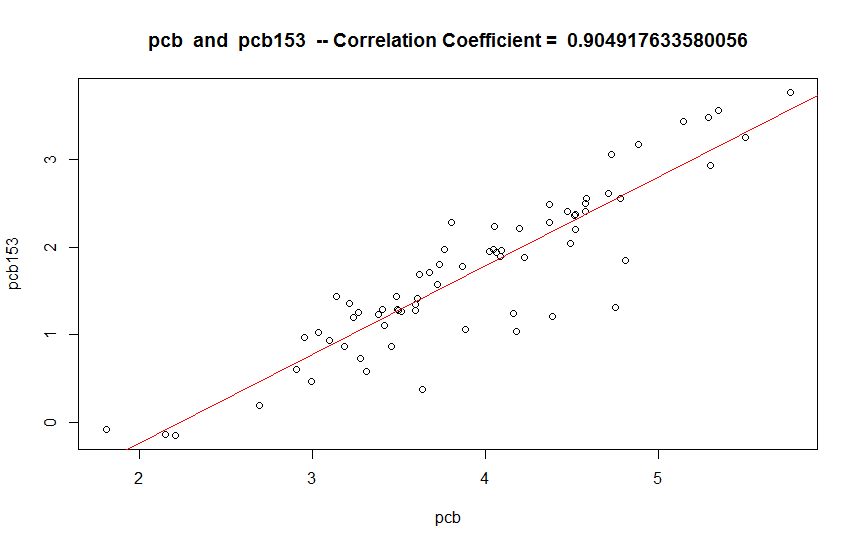
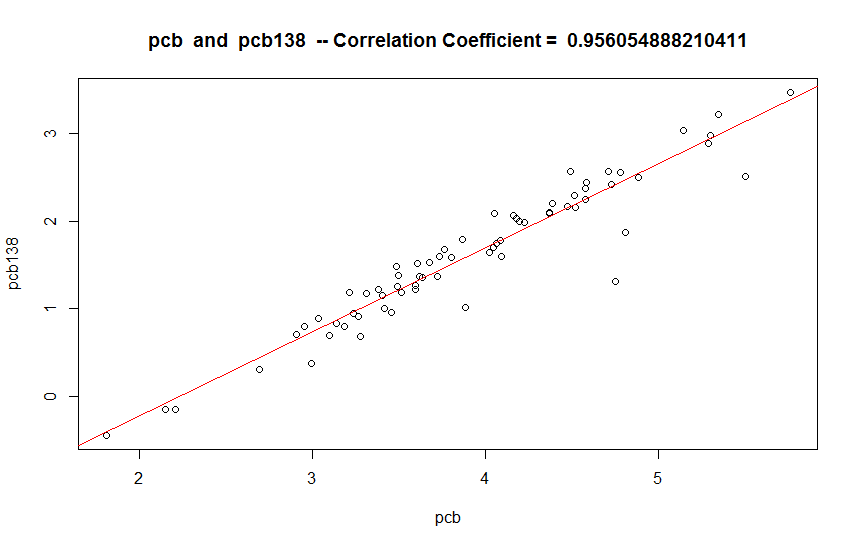
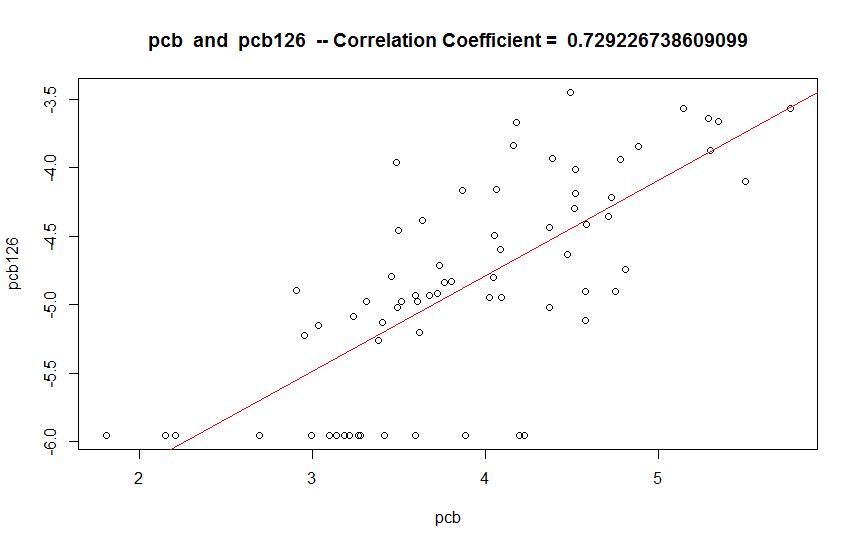
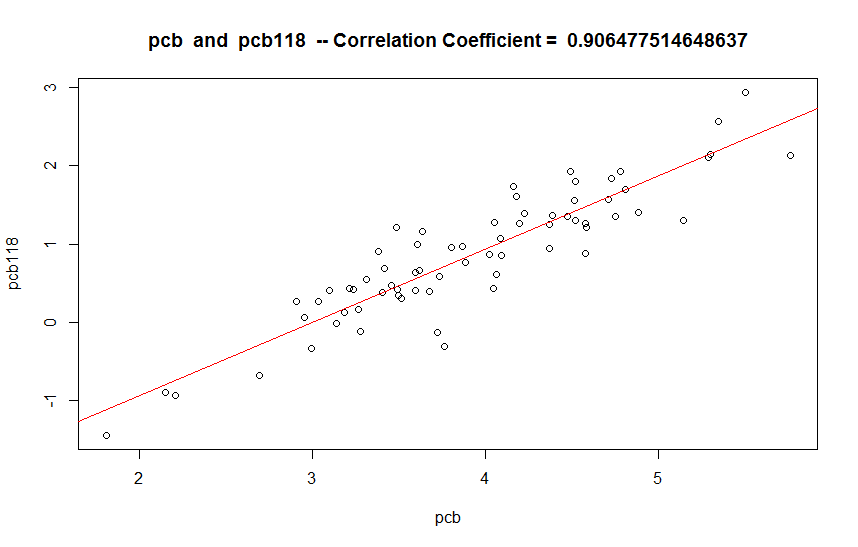
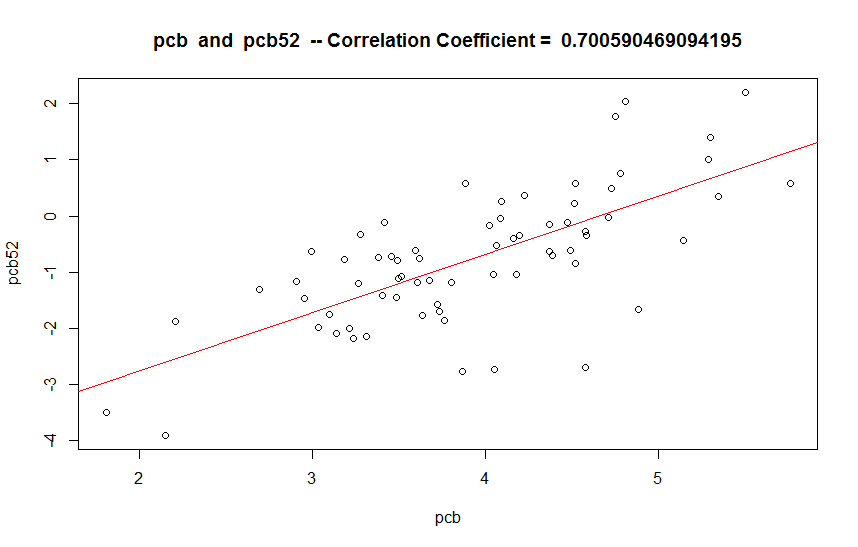
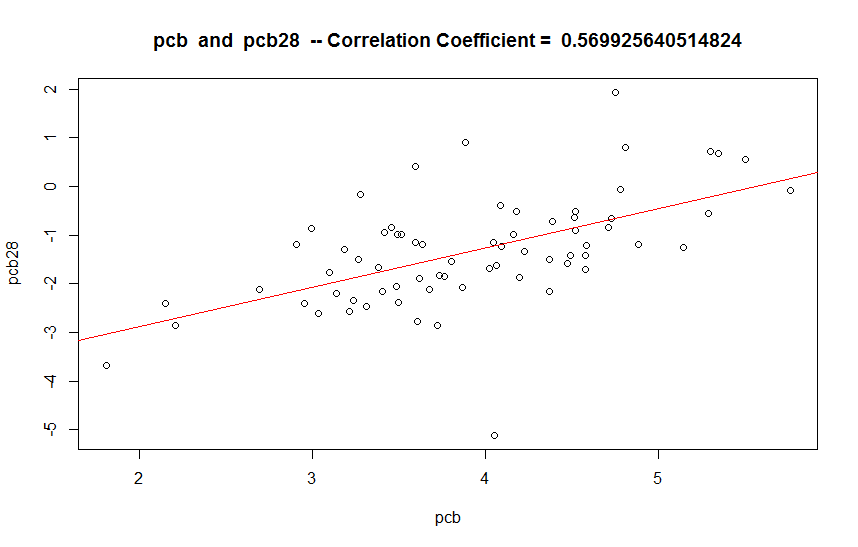
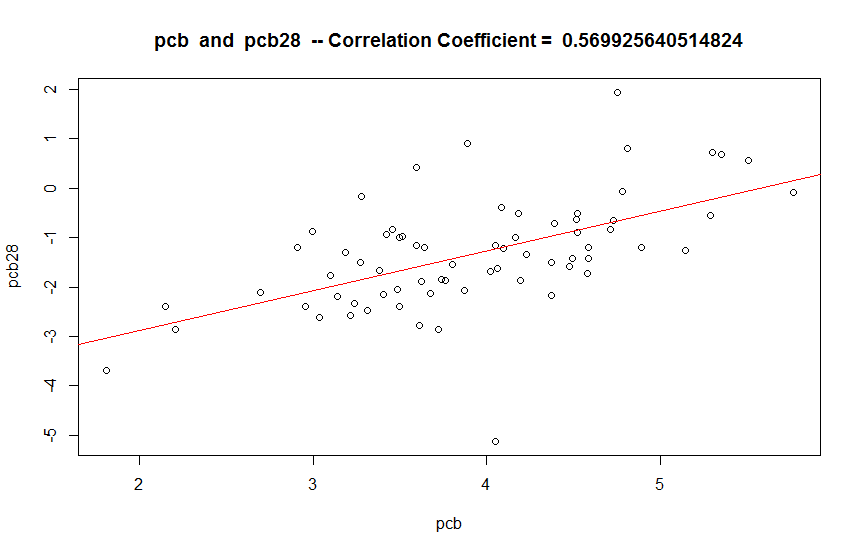


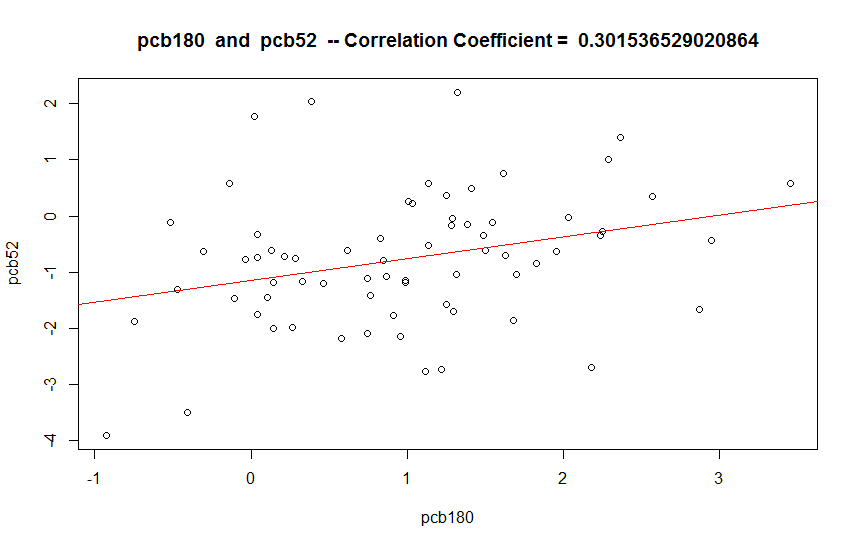
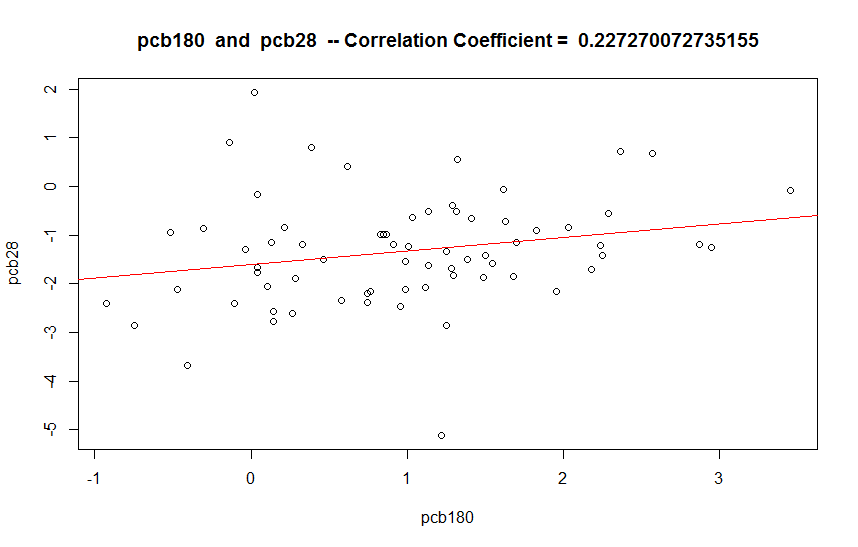
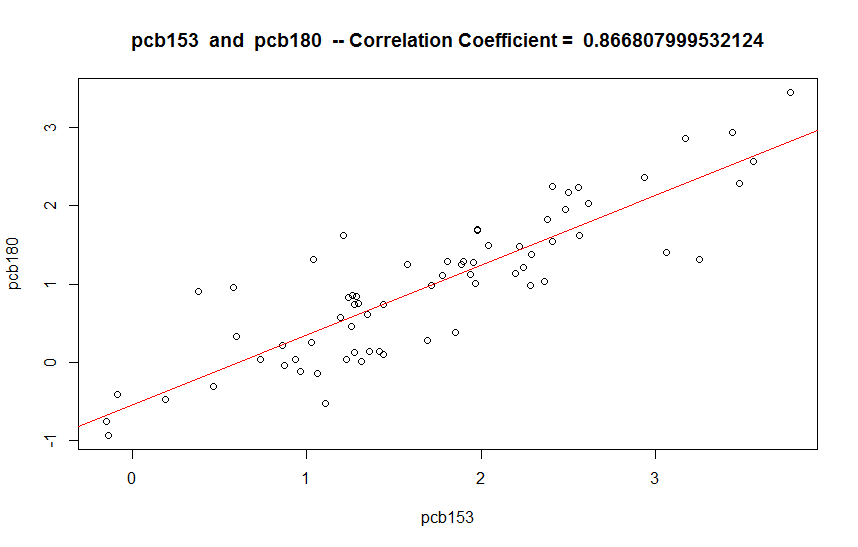
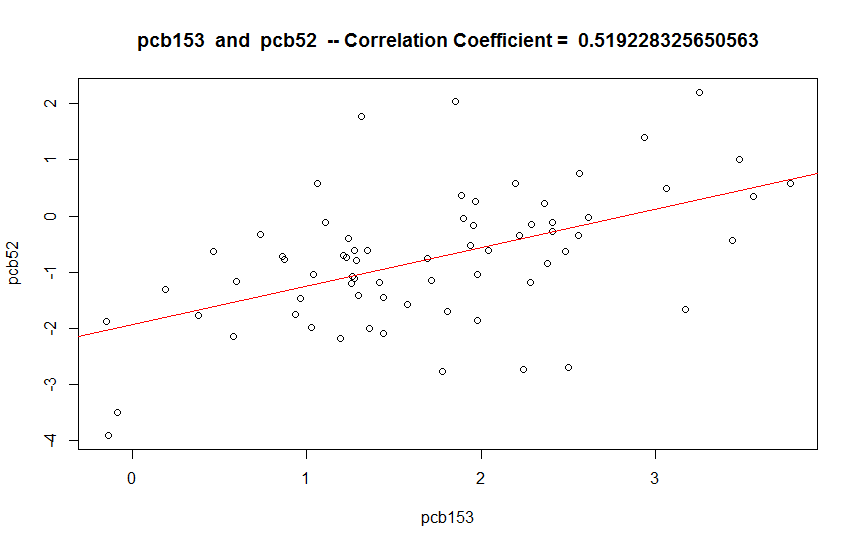
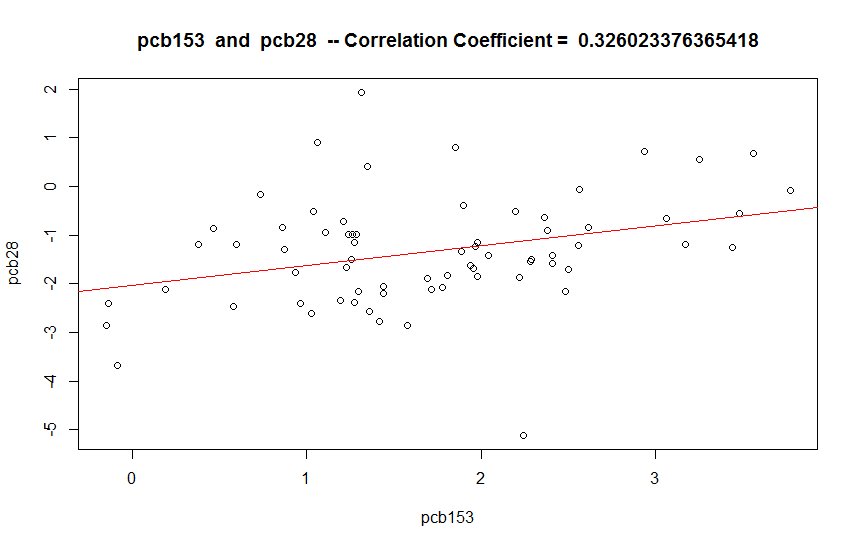
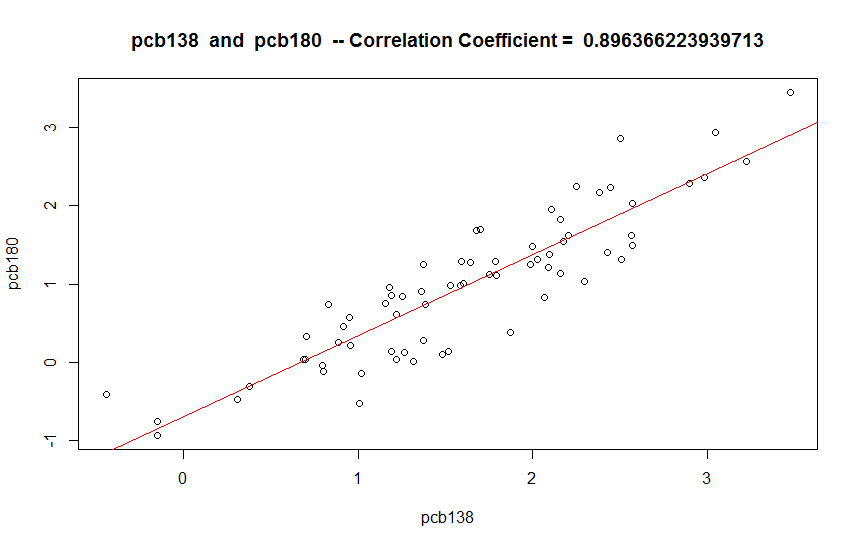


All distributions appear to be approximately normal.

49. Refer to the previous exercise.

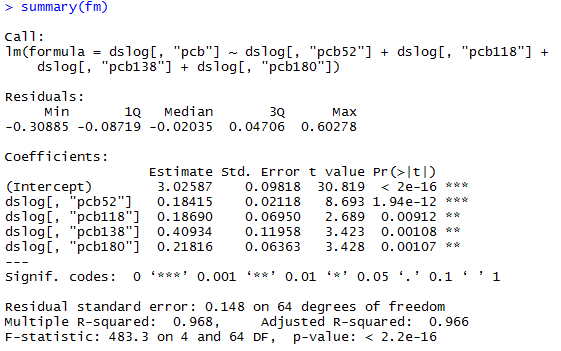
1. Use the numerical and graphical summaries to describe the relationships between each pair of log variables.

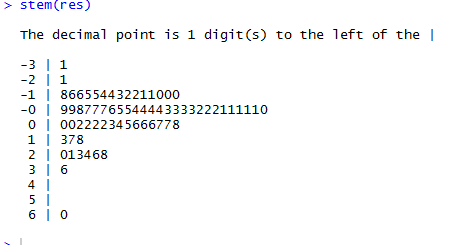




2. Compare these summaries with the summaries that you produced in Exercise 11.42 for the measured variables.  
The correlations found for the logarithmic comparisons are higher than the correlations found for the measured variables.

50. Use the log data set that you created in Exercise 11.48 to find a good multiple regression model for predicting the log of PCB. Use only log PCB variables for this analysis. Write a report summarizing your results.

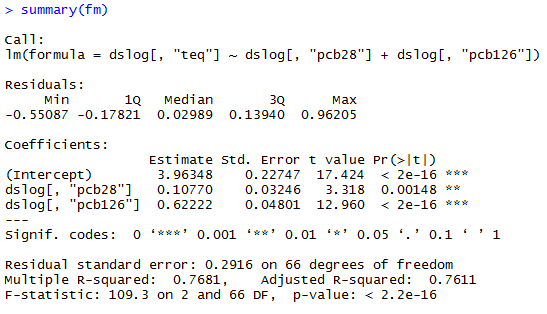


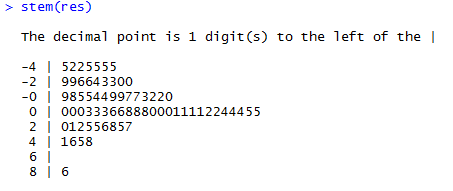


By taking the multiple regression of several permutations of different PCB congeners, I determined that using LOGpcb52, LOGpcb118, LOGpcb138, and LOGpcb180 yields an R-squared value of 0.968, among the highest of all tested permutations. As this was the same combination used earlier, I decided to stick with it.

As seen in the stem graph of the residuals, the distribution of the residuals is a slightly right-skewed normal distribution.

51. Use the log data set that you created in Exercise 11.48 to find a good multiple regression model for predicting the log of TEQ. Use only log PCB variables for this analysis. Write a report summarizing your results and comparing them with the results you obtained in the previous exercise.





Using previously selected rows yielded a fairly low R-squared value of < 0.55. By using the rows previously unused (pcb28 and pcb126), I got an R-squared value of 0.7681.

As before, the stem of the residuals is a slightly right-skewed normal distribution.

52. Use the results of your analysis of the log PCB data in Exercise 11.50 to write an explanation of how regression coefficients, Standard errors of regression coefficients, and tests of significance for explanatory variables can change depending on what other explanatory variables are included in the multiple regression analysis.

By cherry-picking datasets with minimal outliers (as seen when we removed two outliers earlier in this assignment), and datasets with distributions similar to the target dataset (when we picked 52, 118, 138, 180, while excluding 28 and 126), we get more accurate multiple regression predictions. These reduce the standard error, and increase the significance of the results we get when predicting datasets.

The most ideal set to use in multiple regression will, when averaged out, yield a similar distribution with a similar range to the target dataset. Conversely, the least ideal set would have a very different distribution and/or many outliers compared to the range of the target set.

These assumptions are fairly intuitive; similar datasets can be extrapolated to predict things in aggregate.

R-CODE FOLLOWS:

Adam Gincel

MA331

R Code for Final Project

ds = file.csv("pcb.csv")

#11.42

boxplot(ds[,"pcb"], main="Boxplot for pcb", horizontal=T, axes=F, staplewex=1)

text(x=boxplot.stats(ds[,"pcb"])$stats, labels=boxplot.stats(ds[,"pcb"])$stats, y=1.25)

mean(ds[,"pcb"])

sd(ds[,"pcb"])

var(ds[,"pcb"])

boxplot(ds[,"pcb52"], main="Boxplot for pcb52", horizontal=T, axes=F, staplewex=1)

text(x=boxplot.stats(ds[,"pcb52"])$stats, labels=boxplot.stats(ds[,"pcb52"])$stats, y=1.25)

mean(ds[,"pcb52"])

sd(ds[,"pcb52"])

var(ds[,"pcb52"])

boxplot(ds[,"pcb118"], main="Boxplot for pcb118", horizontal=T, axes=F, staplewex=1)

text(x=boxplot.stats(ds[,"pcb118"])$stats, labels=boxplot.stats(ds[,"pcb118"])$stats, y=1.25)

print(mean(ds[,"pcb118"]))

print(sd(ds[,"pcb118"]))

print(var(ds[,"pcb118"]))

boxplot(ds[,"pcb138"], main="Boxplot for PCB138", horizontal=T, axes=F, staplewex=1)

text(x=boxplot.stats(ds[,"pcb138"])$stats, labels=boxplot.stats(ds[,"pcb138"])$stats, y=1.25)

mean(ds[,"pcb138"])

sd(ds[,"pcb138"])

var(ds[,"pcb138"])

boxplot(ds[,"pcb180"], main="Boxplot for pcb180", horizontal=T, axes=F, staplewex=1)

text(x=boxplot.stats(ds[,"pcb180"])$stats, labels=boxplot.stats(ds[,"pcb180"])$stats, y=1.25)

print(mean(ds[,"pcb180"]))

print(sd(ds[,"pcb180"]))

print(var(ds[,"pcb180"]))

plot(pcb, pcb52, main="PCB and PCB52", xlab="PCB", ylab="PCB52")

abline(lm(pcb52~pcb), col="red")

corrCoeff <- cor(pcb,pcb52)

text(labels = paste("Correlation Coefficient = ", corrCoeff), x=100.25, y=8.25)

# ^ repeat for all values

#11.43

fm <- lm(pcb ~ pcb52 + pcb118 + pcb138 + pcb180)

fm

summary(fm)

res <- residuals(fm)

res

stem(res)

#11.45

fm <- lm(pcb ~ pcb52 + pcb118 + pcb138)

fm

summary(fm)

res <- residuals(fm)

res

stem(res)

#11.46

fm <- lm(ds[,"teq"] ~ ds[,"teqpcb"] + ds[,"teqdioxin"] + ds[,"teqfuran"])

fm

summary(fm)

res <- residuals(fm)

res

stem(res)

#11.47

fm <- lm(teq ~ pcb52 + pcb118 + pcb138 + pcb180)

fm

summary(fm)

res <- residuals(fm)

res

stem(res)

#11.48

boxplot(dslog[,"pcb118"], main="Boxplot: PCB118", horizontal=T, axes=F, staplewex=1)

print(summary(dslog[,"pcb118"]))

#repeat for each pcb value

#11.49

for (i in pcblist) {

for(j in pcblist) {

if(i < j)

{

corrCoeff <- cor(dslog[,i],dslog[,j])

plot(dslog[,i], dslog[,j], main=paste(i, " and ", j, " -- Correlation Coefficient = ", corrCoeff), xlab=i, ylab=j)

abline(lm(dslog[,j]~dslog[,i]), col="red")

}

}

}

#11.50

datasets <- c("pcb28", "pcb52", "pcb118", "pcb126", "pcb138", "pcb153", "pcb180")

permutations <- combn(datasets, 2) # use this to determine best multiple regression

for (i in 1:21)

{

print(paste(permutations[1,i], permutations[2,i]))

print(summary(lm(dslog[,"pcb"] ~ dslog[,permutations[1,i]] + dslog[,permutations[2,i]])))

}