Matt Sirota

MA 331

11 December 2017

Final Exam

***Section 1***

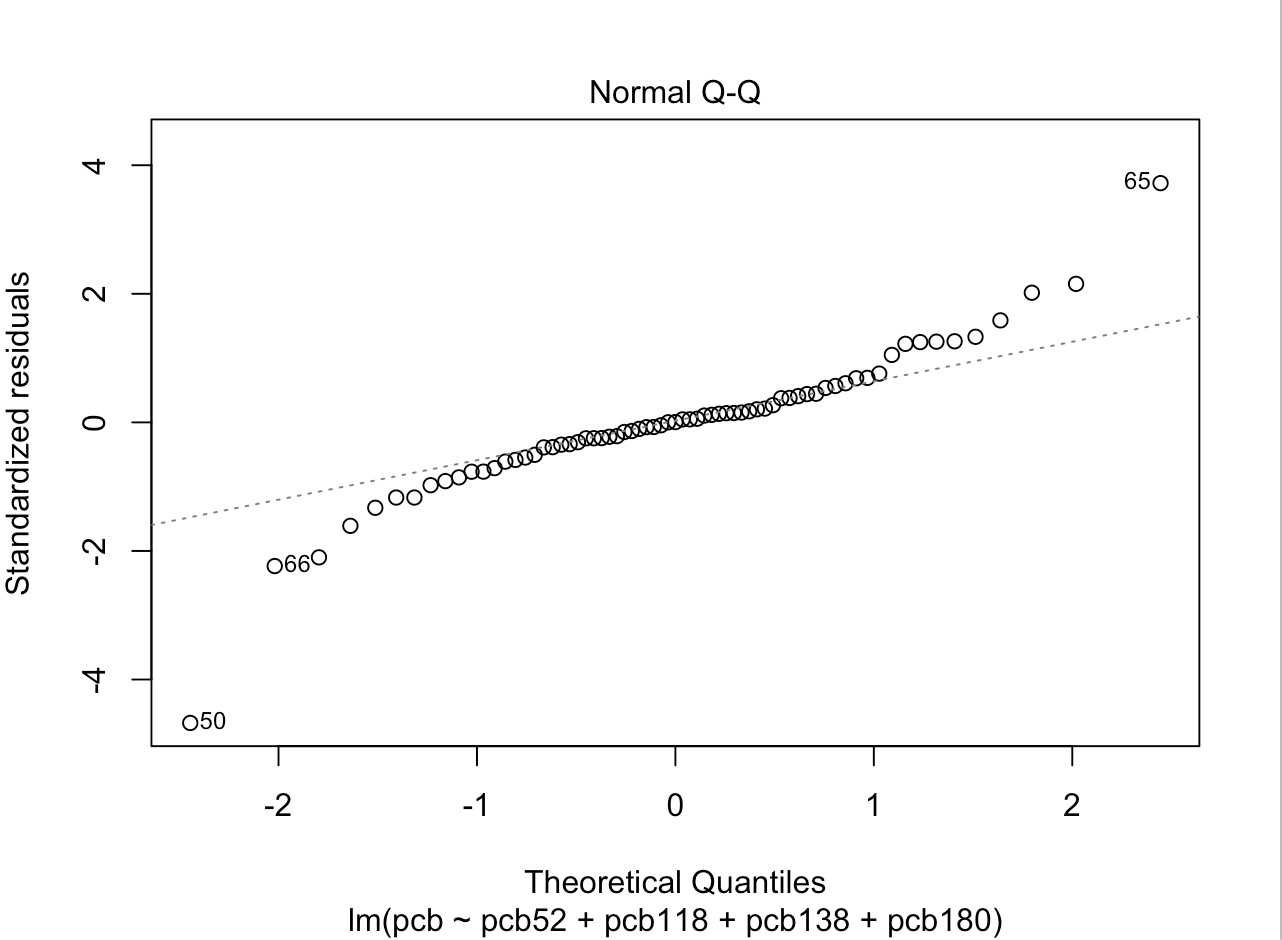
This study considers the multiple linear regressions and coorelations between Polychlorinated biphenyls (PCBs) found in a given specimen and the congener that makes up the PCB compound. PCBs are made up of synthetic compounds, known as congeners, that are toxic to living things, especially fetuses and children. From studying each PCB, we can find out many things about them. From the given sample data, it is evident that the different PCBs have a right skewed distribution as seen below, because the means for each is larger than the median.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | N | Min | Q1 | Q2 (median) | Mean | Q3 | Max |
| PCB | 69 | 6.10 | 30.18 | 47.96 | 68.47 | 91.63 | 318.70 |
| PCB52 | 69 | 0.020 | 0.228 | 0.477 | 0.958 | 0.892 | 9.060 |
| PCB118 | 69 | 0.236 | 1.490 | 2.420 | 3.256 | 3.890 | 18.900 |
| PCB138 | 69 | 0.640 | 3.180 | 4.920 | 6.827 | 8.650 | 32.300 |
| PCB180 | 69 | 0.395 | 1.240 | 2.690 | 4.158 | 4.490 | 31.500 |

There also seems to be a positive correlation between the variables, so if one PCB is found, there’s likely to be more inside that organism.

|  |  |  |
| --- | --- | --- |
| Data1 | Data2 | Correlation |
| PCB | PCB52 | 0.596 |
| PCB118 | 0.843 |
| PCB138 | 0.929 |
| PCB180 | 0.801 |
| PCB52 | PCB 118 | 0.685 |
| PCB138 | 0.301 |
| PCB180 | 0.0869 |
| PCB118 | PCB138 | 0.729 |
| PCB180 | 0.437 |
| PCB138 | PCB180 | 0.882 |

4 specific PCBs were chosen for regression analysis and the results can be seen below. There were outliers in this data that ended up being removed so the data would not appear as varied.

(co-ef denotes co-efficients and ESE denotes estimated standard error)

Variable co-ef ESE t value Pr(>|t|)

(Intercept) 0.9369 1.2293 0.762 0.449

pcb52 11.8727 0.7290 16.287 < 2e-16 \*\*\*

pcb118 3.7611 0.6424 5.855 1.79e-07 \*\*\*

pcb138 3.8842 0.4978 7.803 7.19e-11 \*\*\*

pcb180 4.1823 0.4318 9.687 3.64e-14 \*\*\*

All coefficients vary greatly from zero except 0.9369, and multiple R-squared, R2=0.989 so 98.9% of variation is due to explanatory variables and the rest is unexplained, 1.01%. The residual standard error is 6.384 on 64 degrees of freedom.

The significance of specific PCBs was also tested. PCB118, for example was shown not to have significance once PCB180 was removed from the regression analysis:

Variable co-ef ESE t value Pr(>|t|)

(Intercept) -1.0184 1.8895 -0.539 0.592

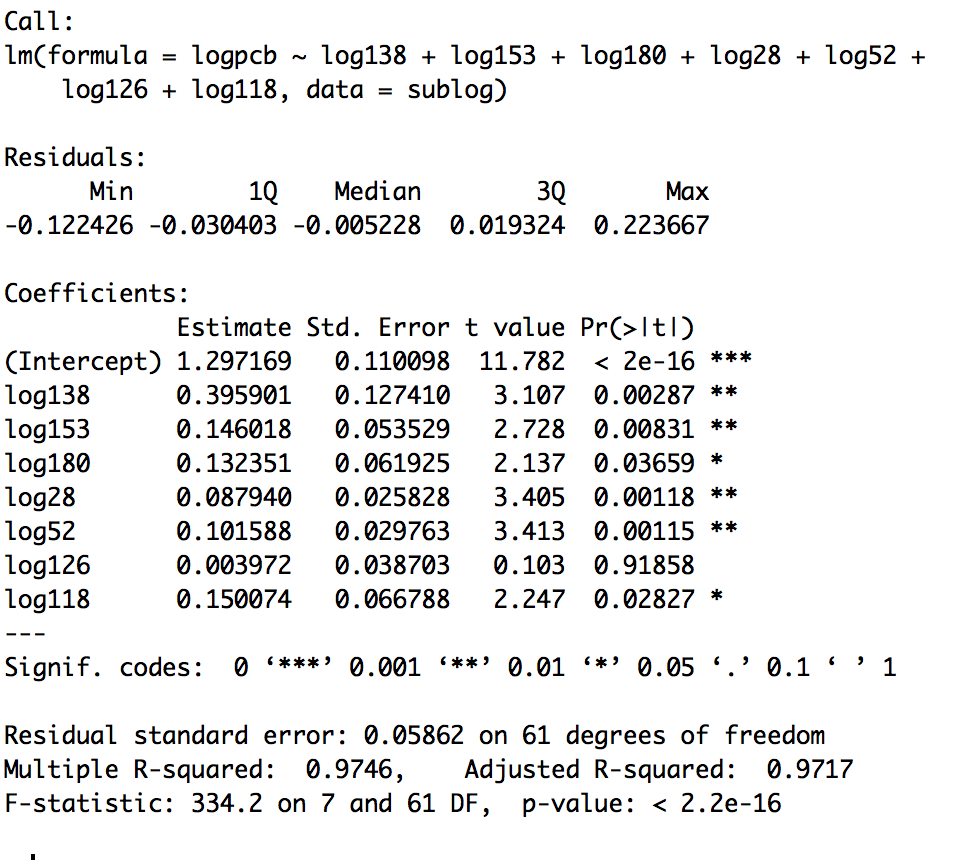
pcb52 12.6442 1.1291 11.198 <2e-16 \*\*\*

pcb118 0.3131 0.8333 0.376 0.708

pcb138 8.2546 0.3279 25.177 <2e-16 \*\*\*

Residual standard error: 9.945 on 65 degrees of freedom

Multiple R-squared: 0.9732

From this data, it shows that 97.32% of variance is from the explanatory variables. While we know where the variance comes from, we still do not know how to reduce the standard error. It turns out that this is actually achievable. To do this, we took the log of every PCB variable.

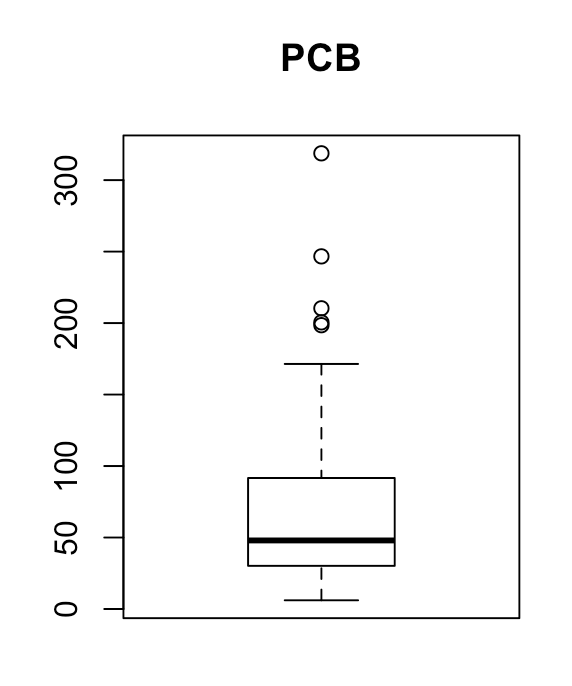
It is clear that by creating a regression analysis for LOGPCB, we were able to reduce the standard error. The error of this case is only 0.05862. This is much lower than the previous examples. In this case, we also note that 97.46% of variance is due to explanatory variables.

***Section 2***

**11.42**

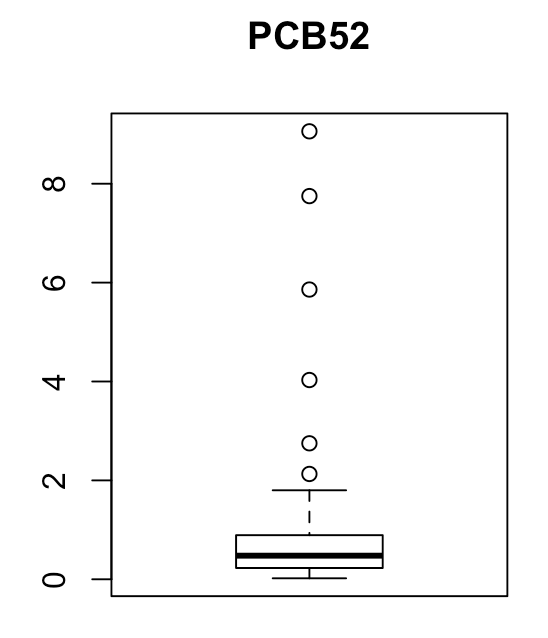
a.) Numerical Summary of Data

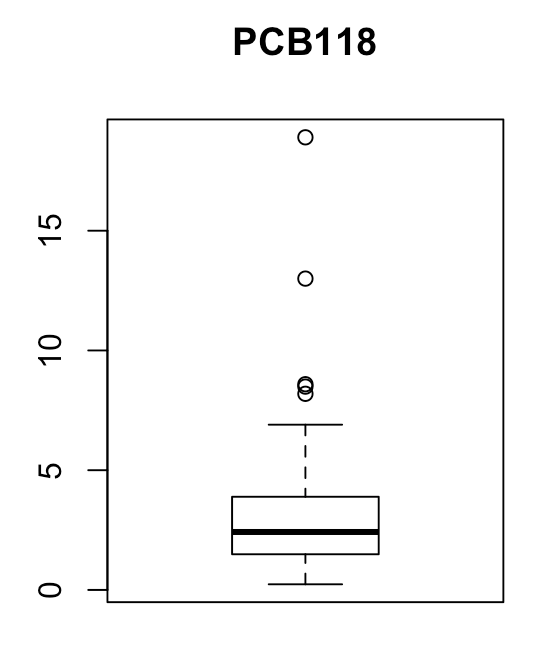
|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Variable | N | Min | Q1 | Q2 (median) | Mean | Q3 | Max |
| PCB | 69 | 6.10 | 30.18 | 47.96 | 68.47 | 91.63 | 318.70 |
| PCB52 | 69 | 0.020 | 0.228 | 0.477 | 0.958 | 0.892 | 9.060 |
| PCB118 | 69 | 0.236 | 1.490 | 2.420 | 3.256 | 3.890 | 18.900 |
| PCB138 | 69 | 0.640 | 3.180 | 4.920 | 6.827 | 8.650 | 32.300 |
| PCB180 | 69 | 0.395 | 1.240 | 2.690 | 4.158 | 4.490 | 31.500 |



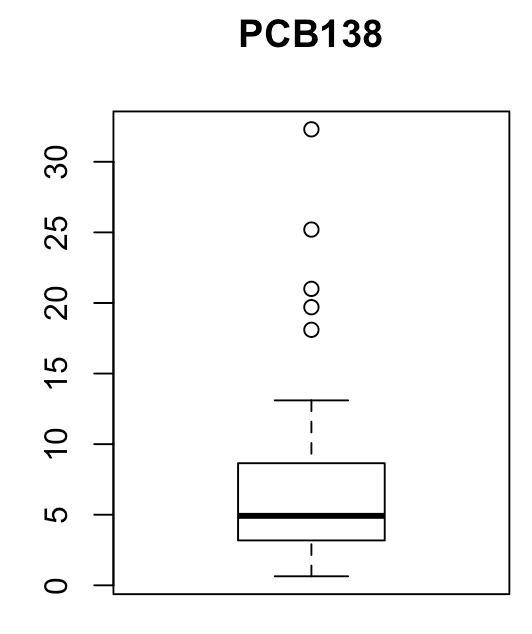
PCB appears to be skewed right and contains 5 outliers as

seen from the numerical summaries and the boxplot.

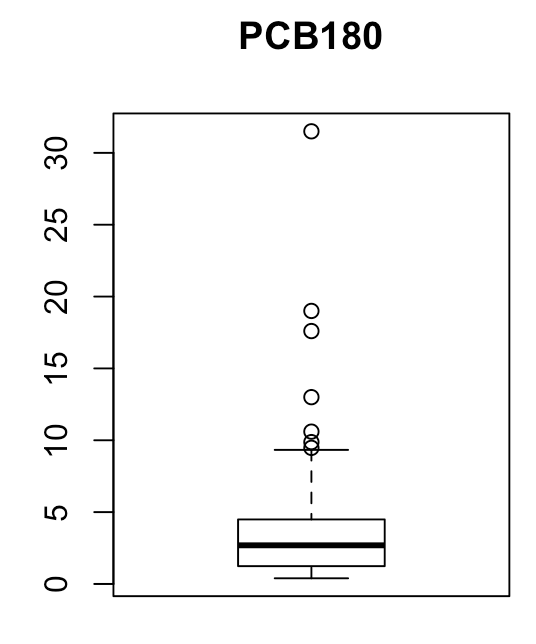
PCB52 appears to be right skewed and contains 6 outliers as seen from the numerical summaries and the boxplot.



PCB118 appears to be skewed right and contains 5 outliers as seen from the numerical summaries and the boxplot.



PCB138 appears to be skewed right and contains 5 outliers as seen from the numerical summaries and the boxplot.

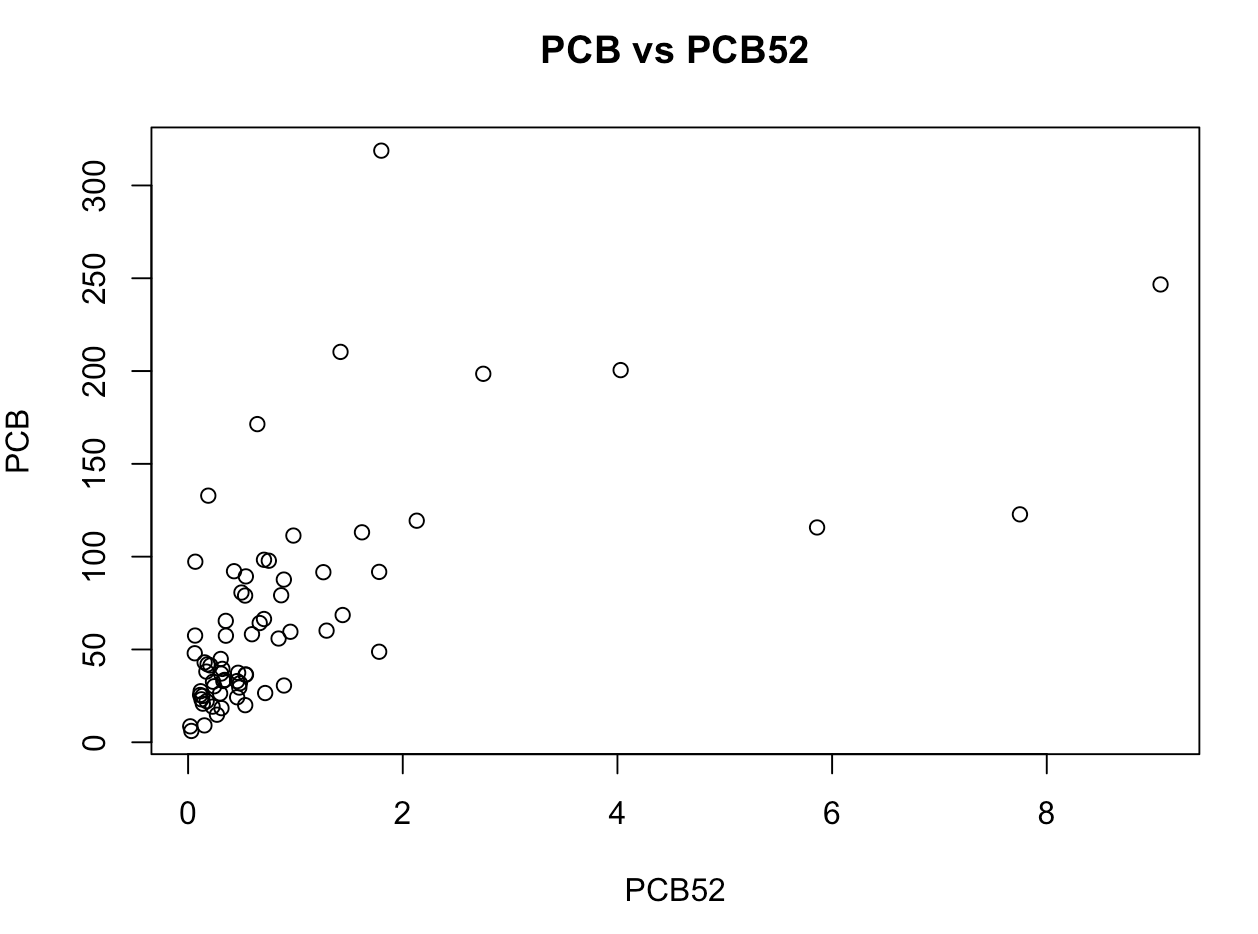
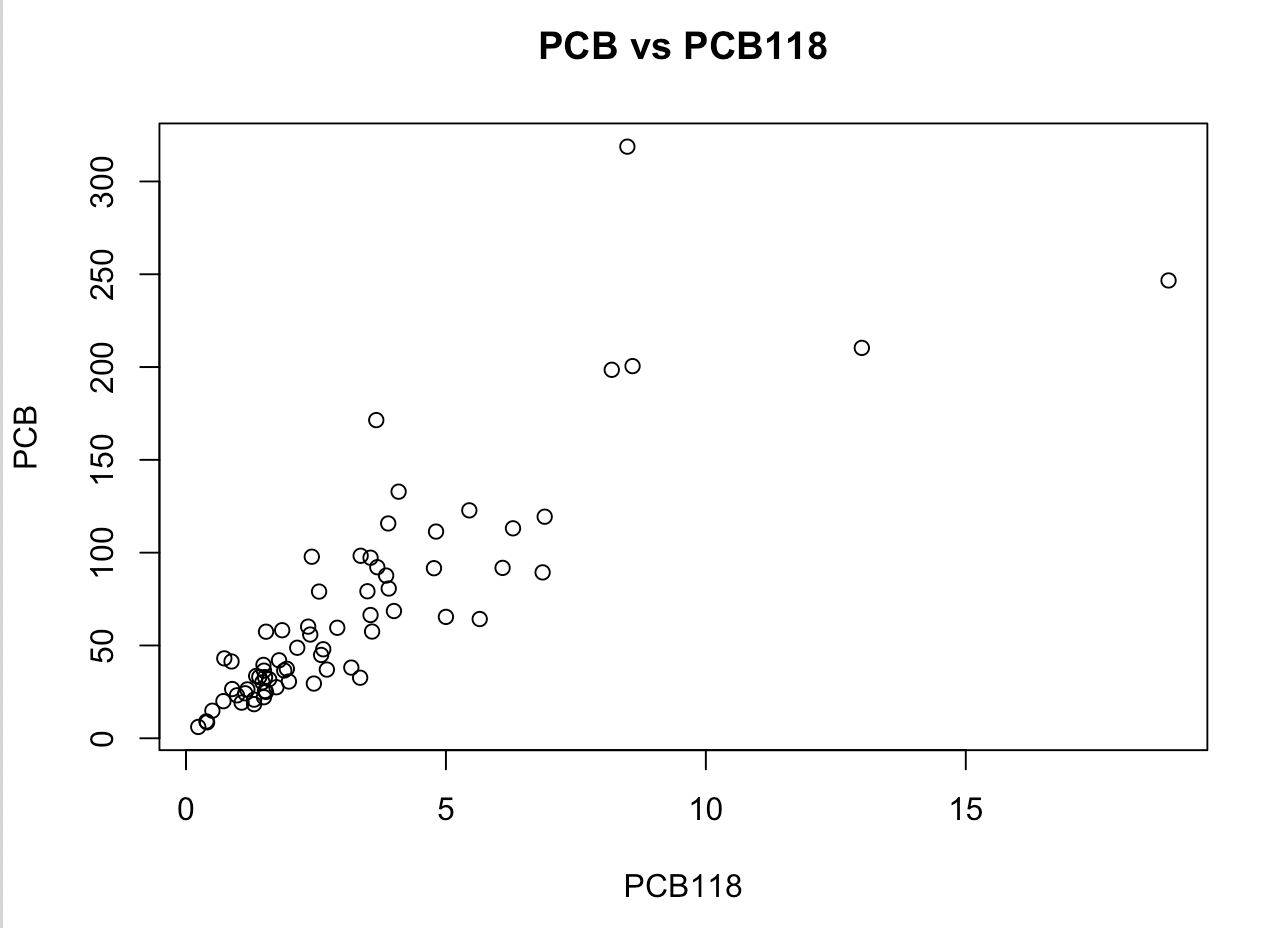


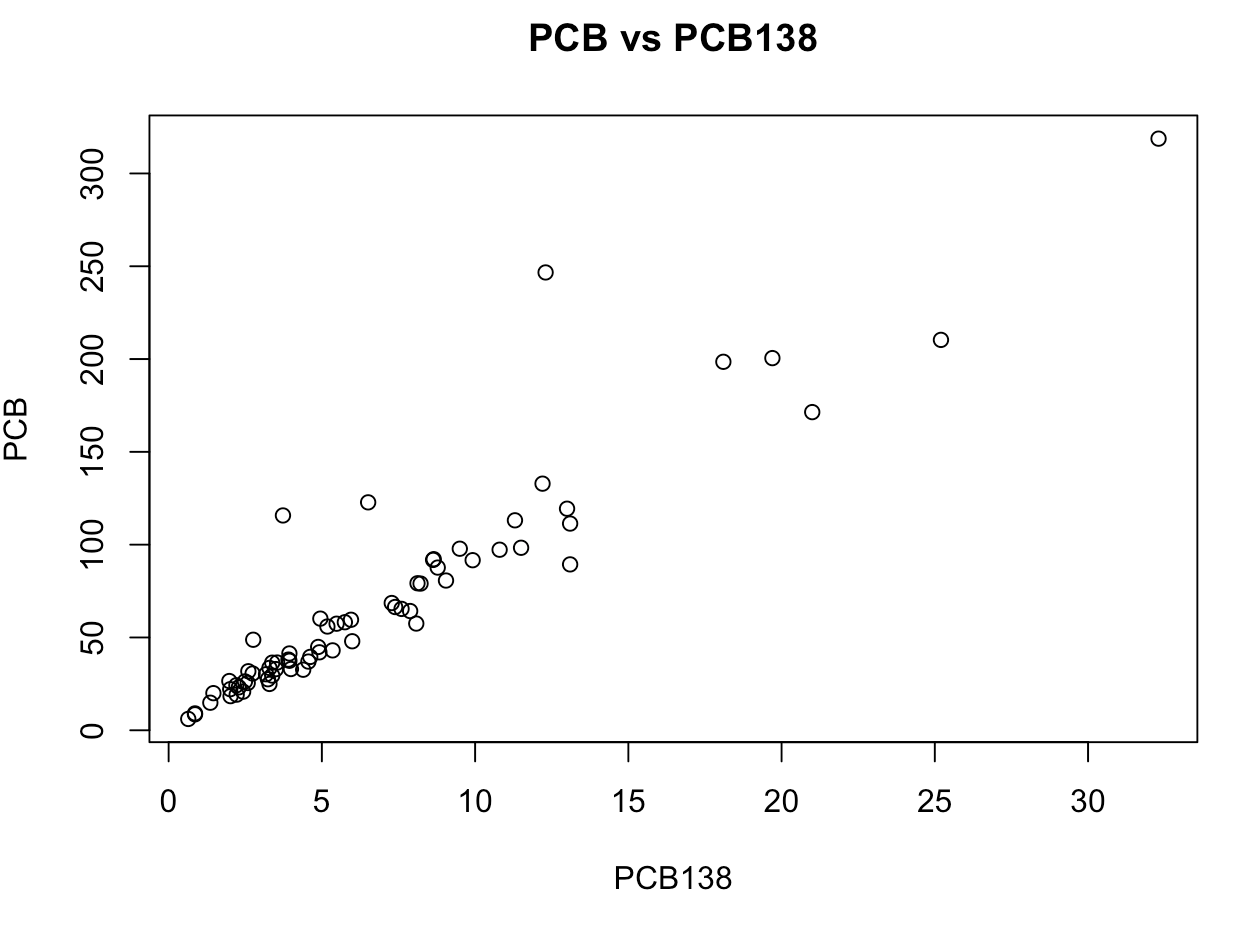
PCB180 appears to be skewed right and contains 6 outliers

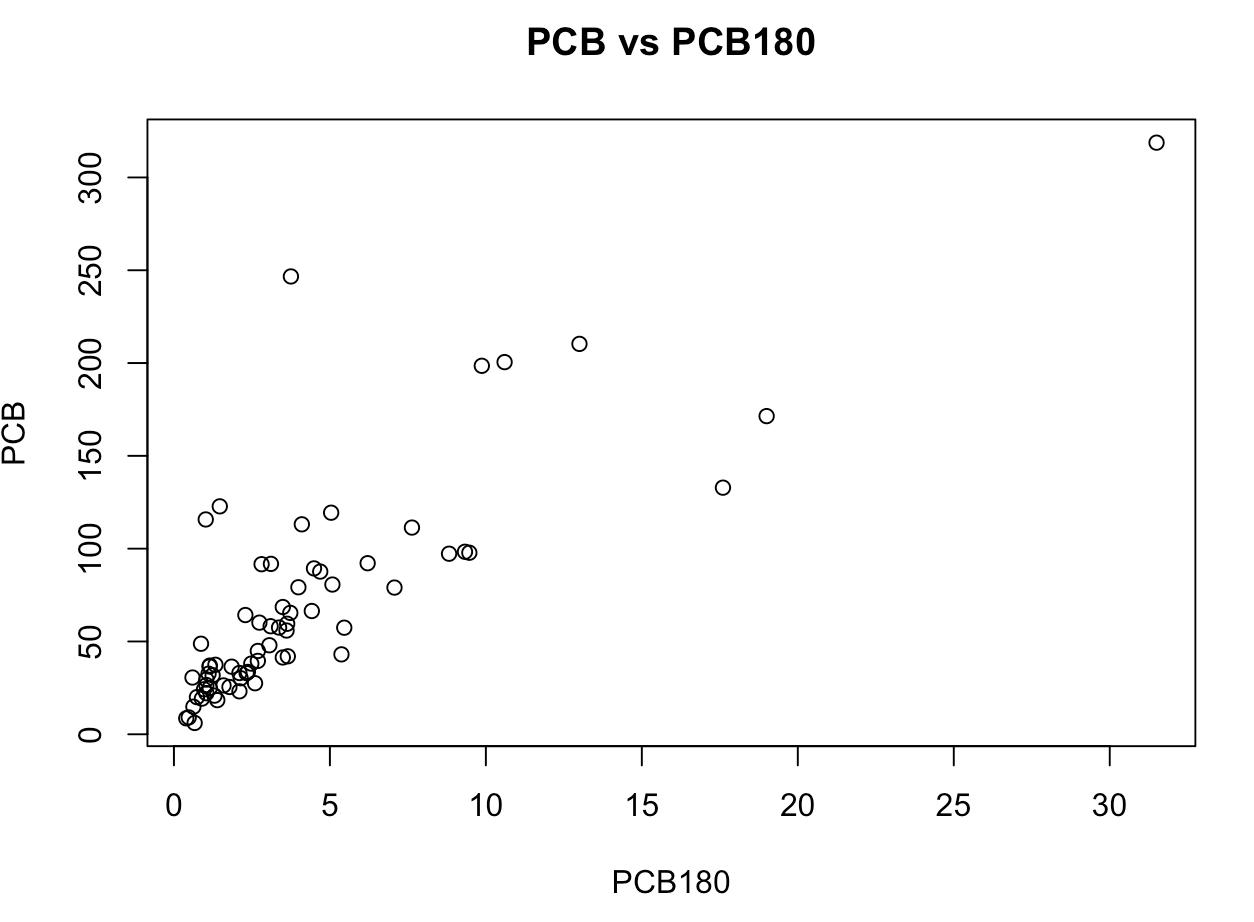
as seen from the numerical summaries and the boxplot.

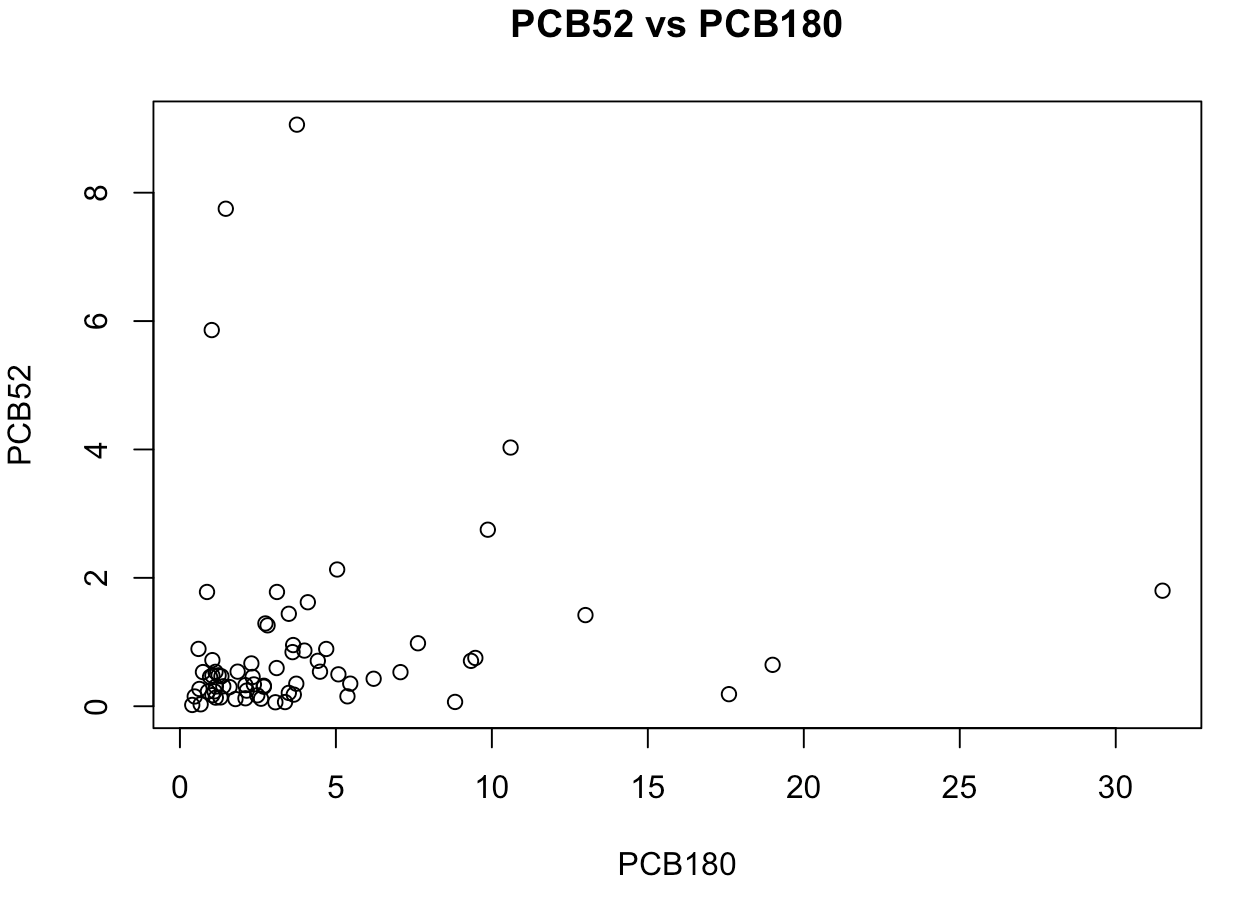
b.)

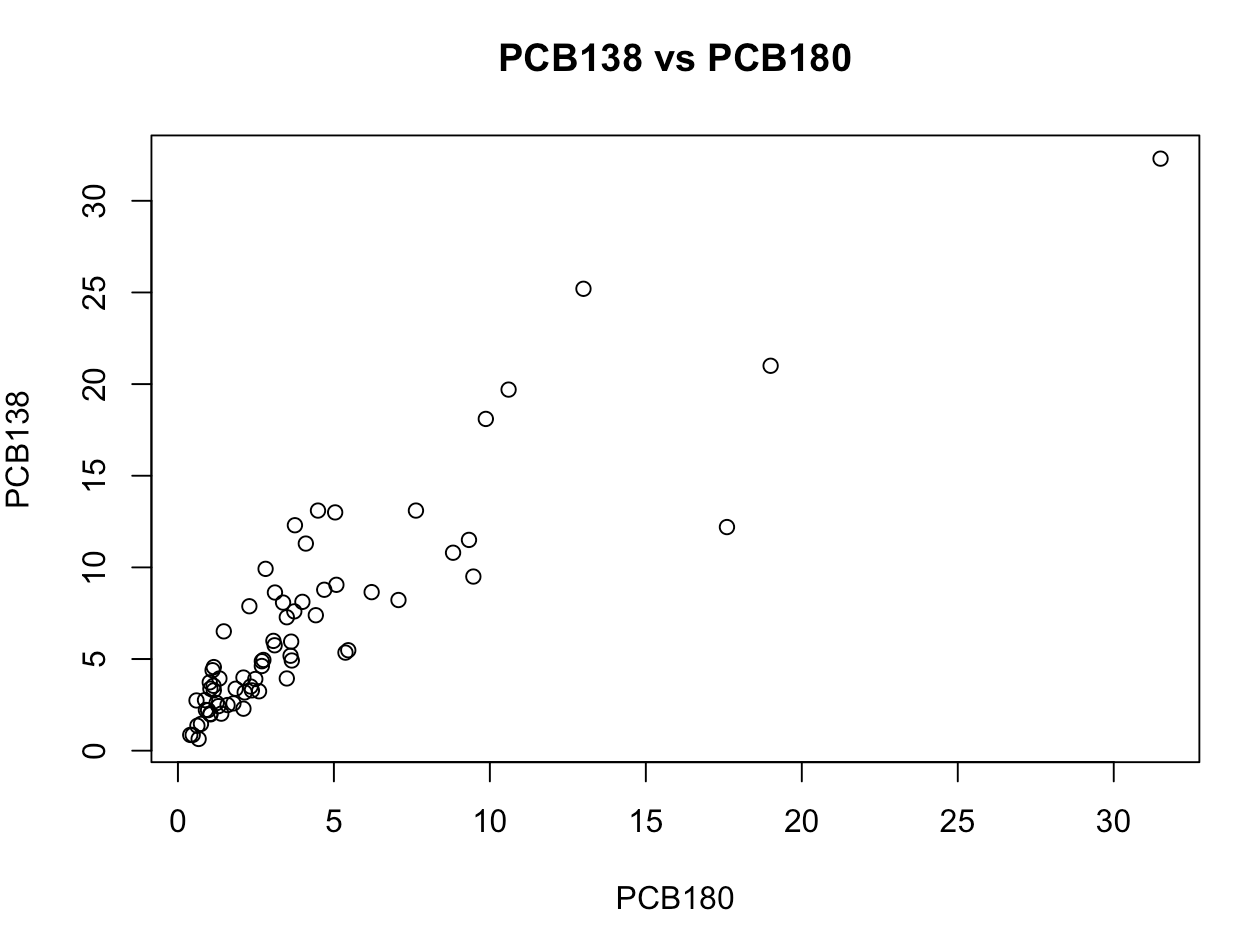
|  |  |  |
| --- | --- | --- |
| Data1 | Data2 | Correlation |
| PCB | PCB52 | 0.596 |
| PCB118 | 0.843 |
| PCB138 | 0.929 |
| PCB180 | 0.801 |
| PCB52 | PCB 118 | 0.685 |
| PCB138 | 0.301 |
| PCB180 | 0.0869 |
| PCB118 | PCB138 | 0.729 |
| PCB180 | 0.437 |
| PCB138 | PCB180 | 0.882 |

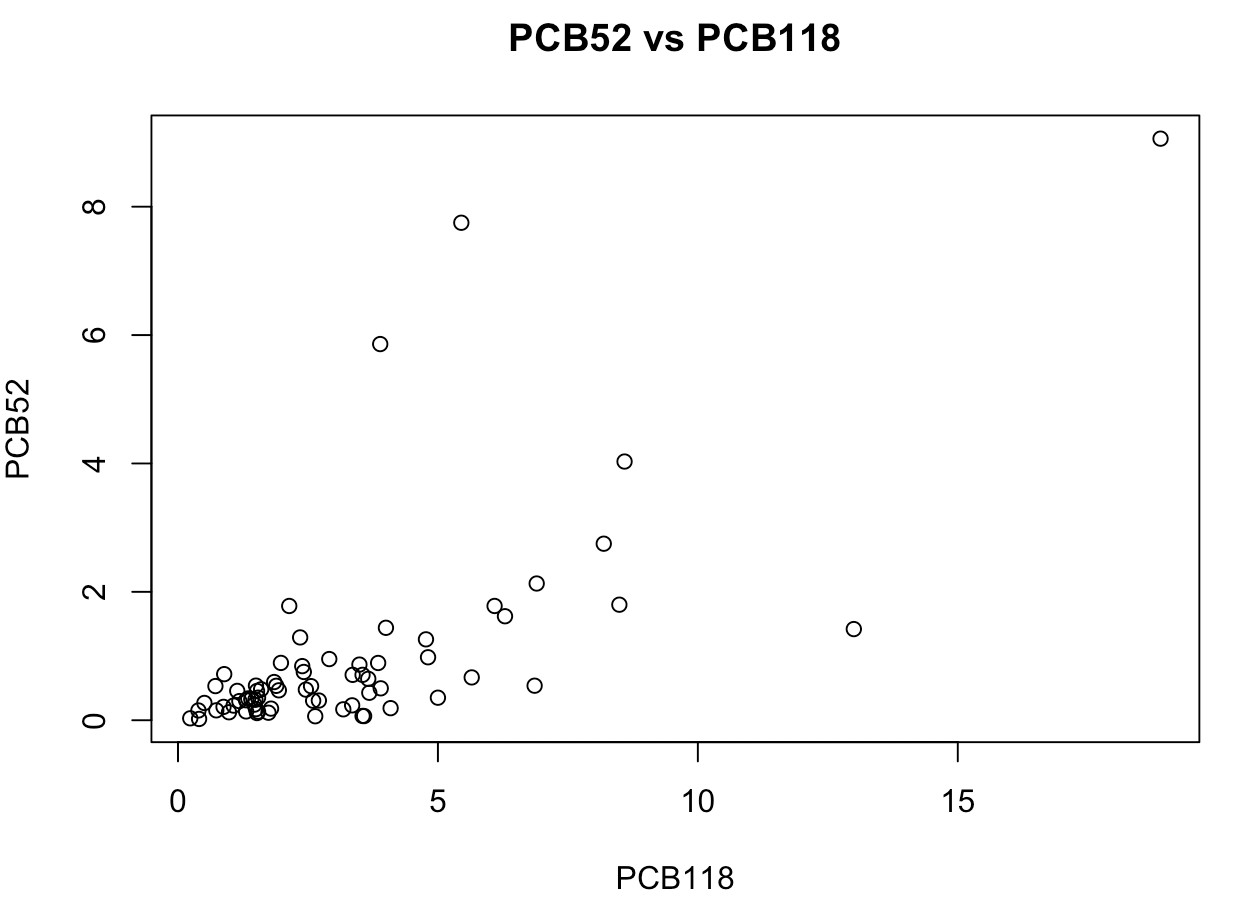


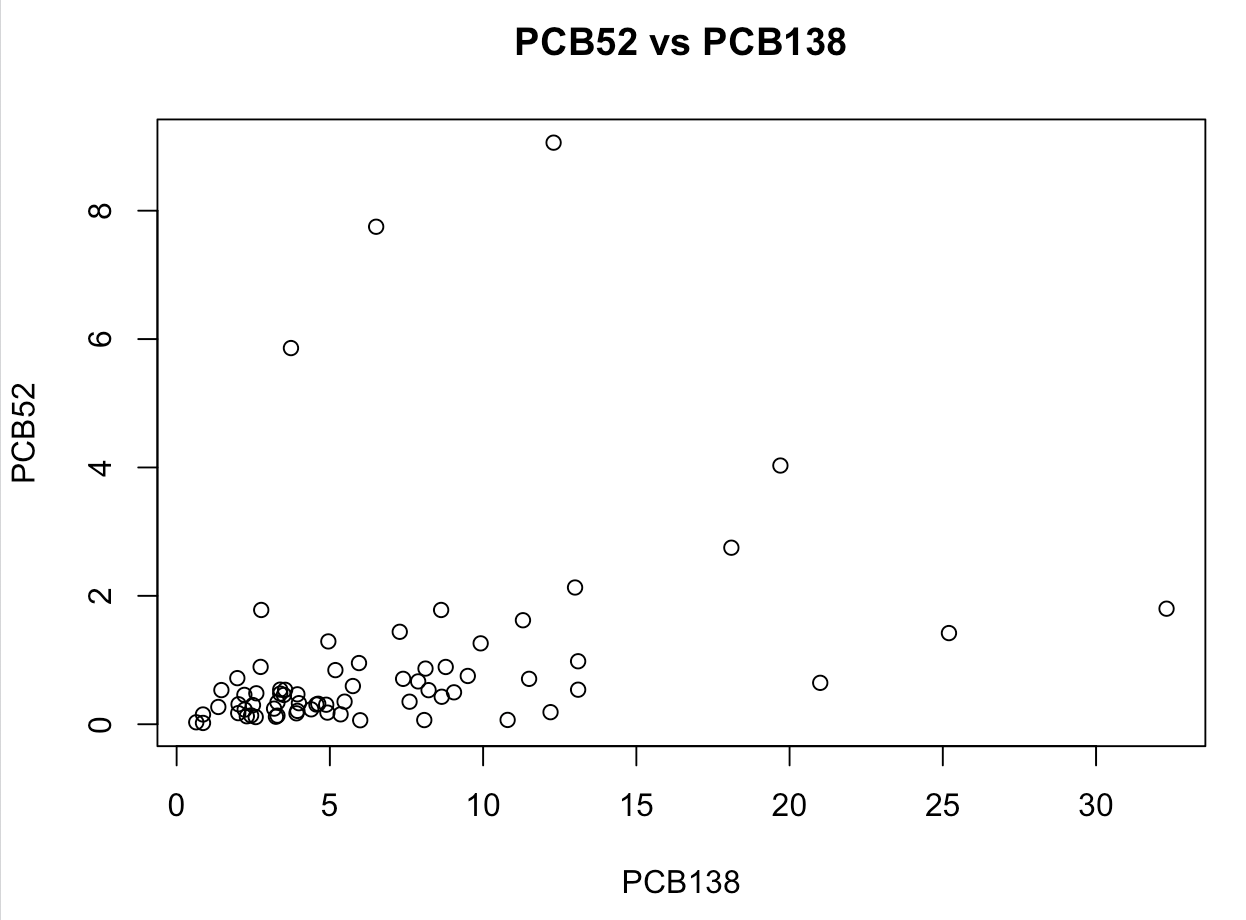


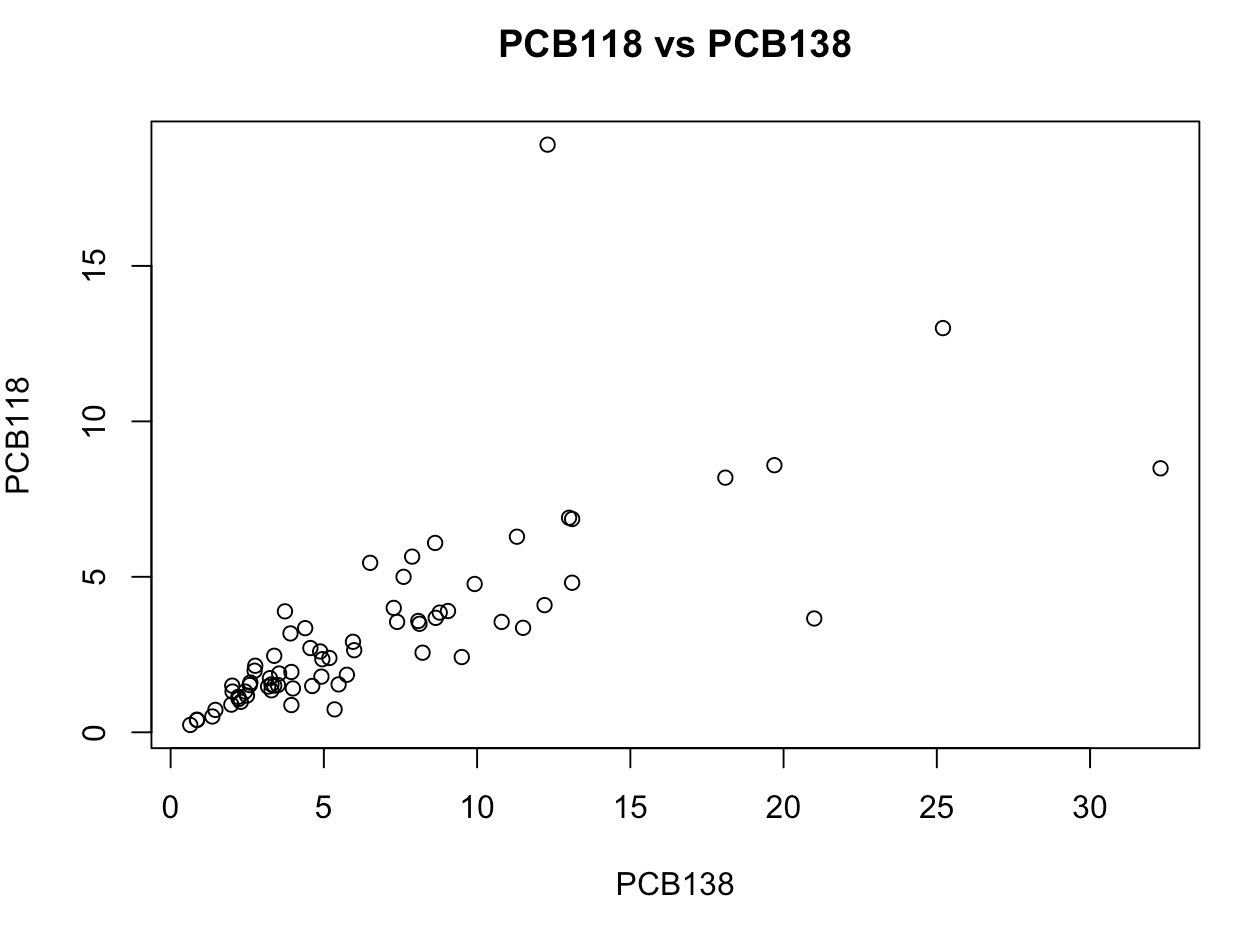
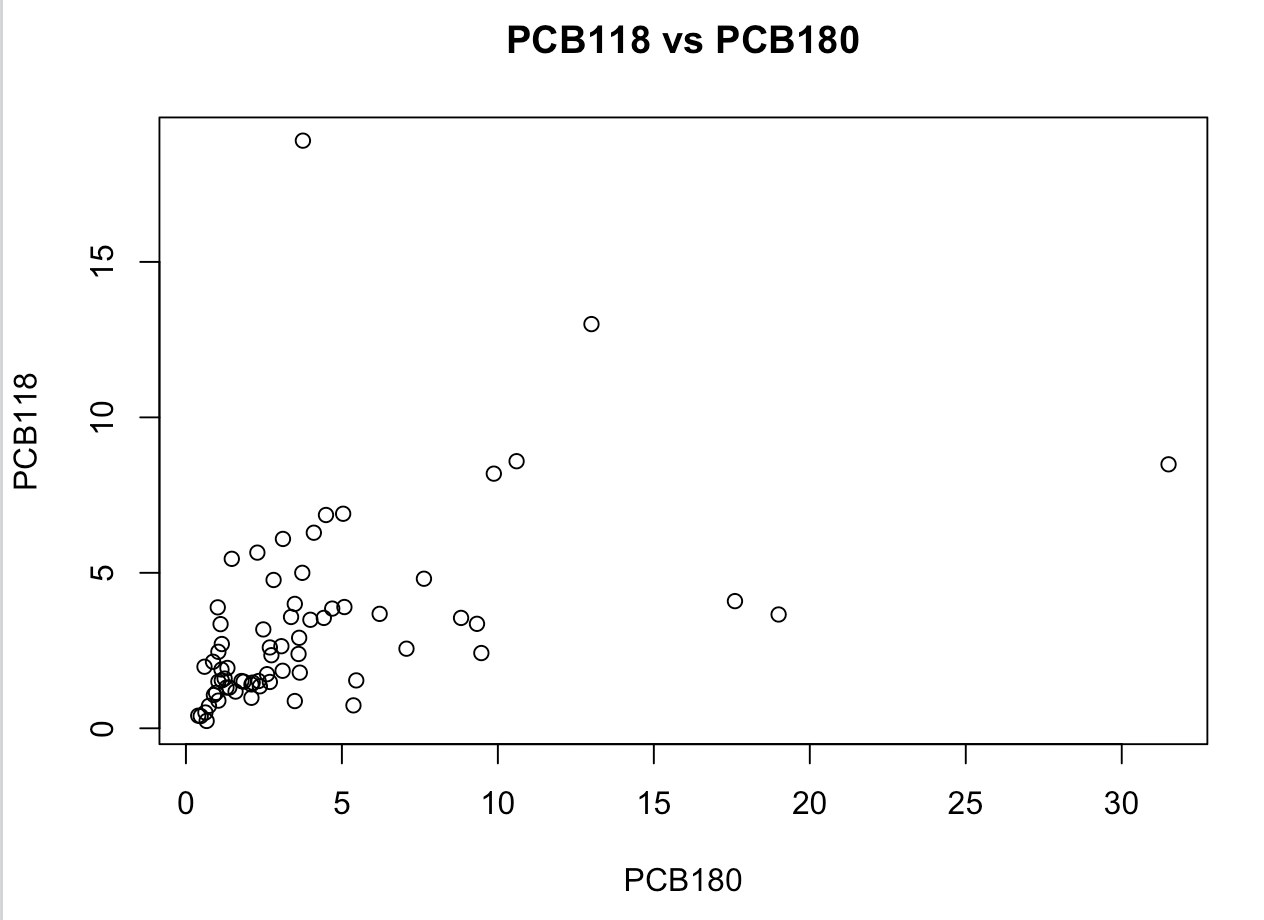












As seen by the scatterplots and the table of correlations, there is a positive correlation between each sets of data, although there are some that are much stronger like PCB vs PCB138 and weaker ones such ass PCB118 vs PCB180

**11.43**

a.) Statistical model for multiple linear regression with p=4 and n=69

yi = β0 + β1xi1 + β2xi2 + β3xi3 + β4xi4 + ∈i  with i=1,2,3,4........69

the assumption is that the deviations (∈i) about mean are independent and are normally distributed with a mean of zero and an unknown standard deviation (σ) which isn’t dependent on explanatory variables (x1,x2,x3x4). The parameters are Parameters are β0, β1, β2, β3, β4, σ.

b.) Response variable: PCB and explanatory variables PCB52, PCB118, PCB138, and PCB180

Test1

H0=β0 = β1 = β2 = β3 = β4 = 0

HA= at least one β not zero

Reject H0 because there is at least one β that is not zero, as seen from the equation previously stated.

Test2

H0=β0 = β1 = β2 = β3 = β4 = 0

HA=β0 , β1 , β2 , β3 ≠ 0

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.9369 1.2293 0.762 0.449

pcb52 11.8727 0.7290 16.287 < 2e-16 \*\*\*

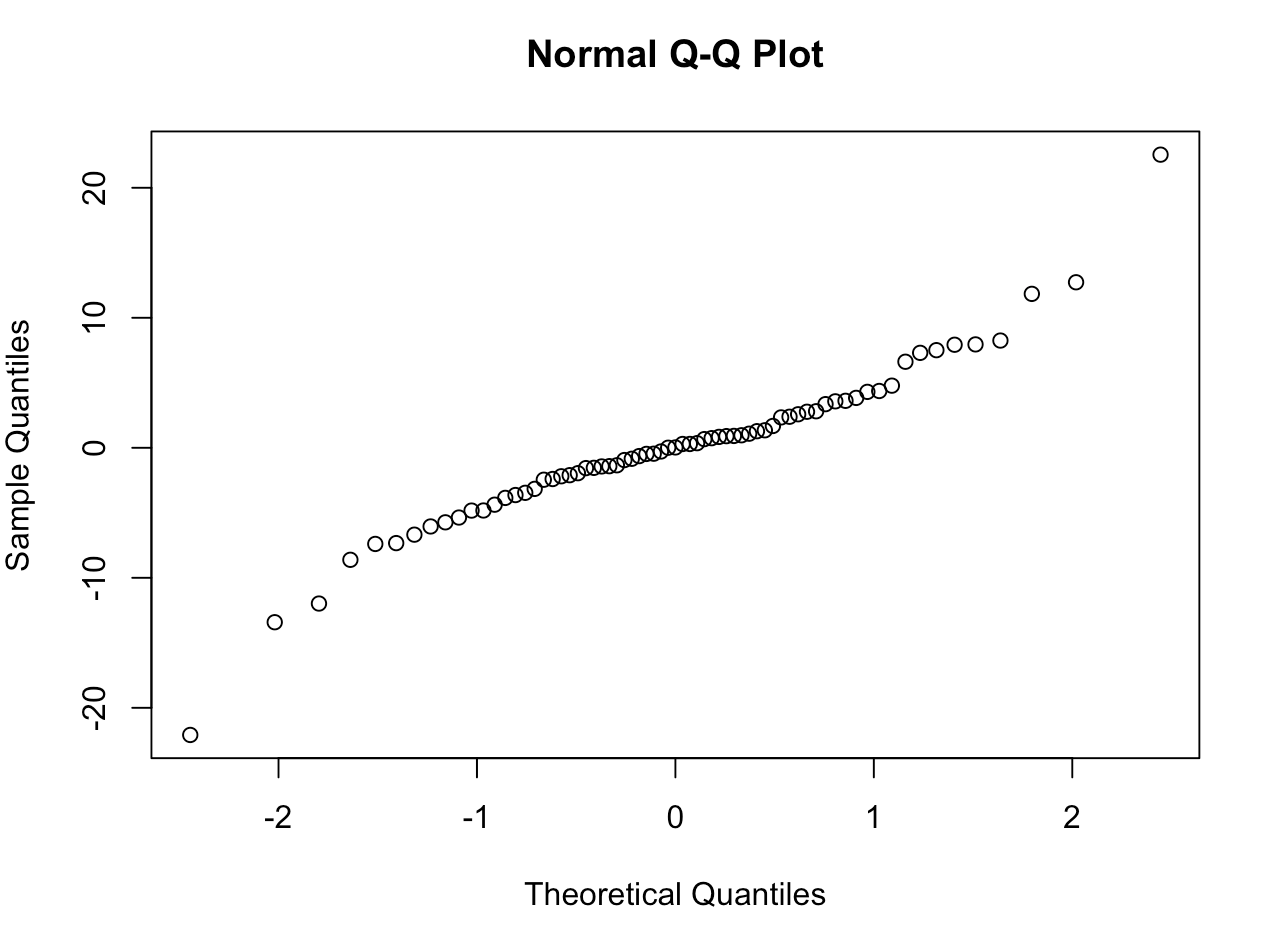
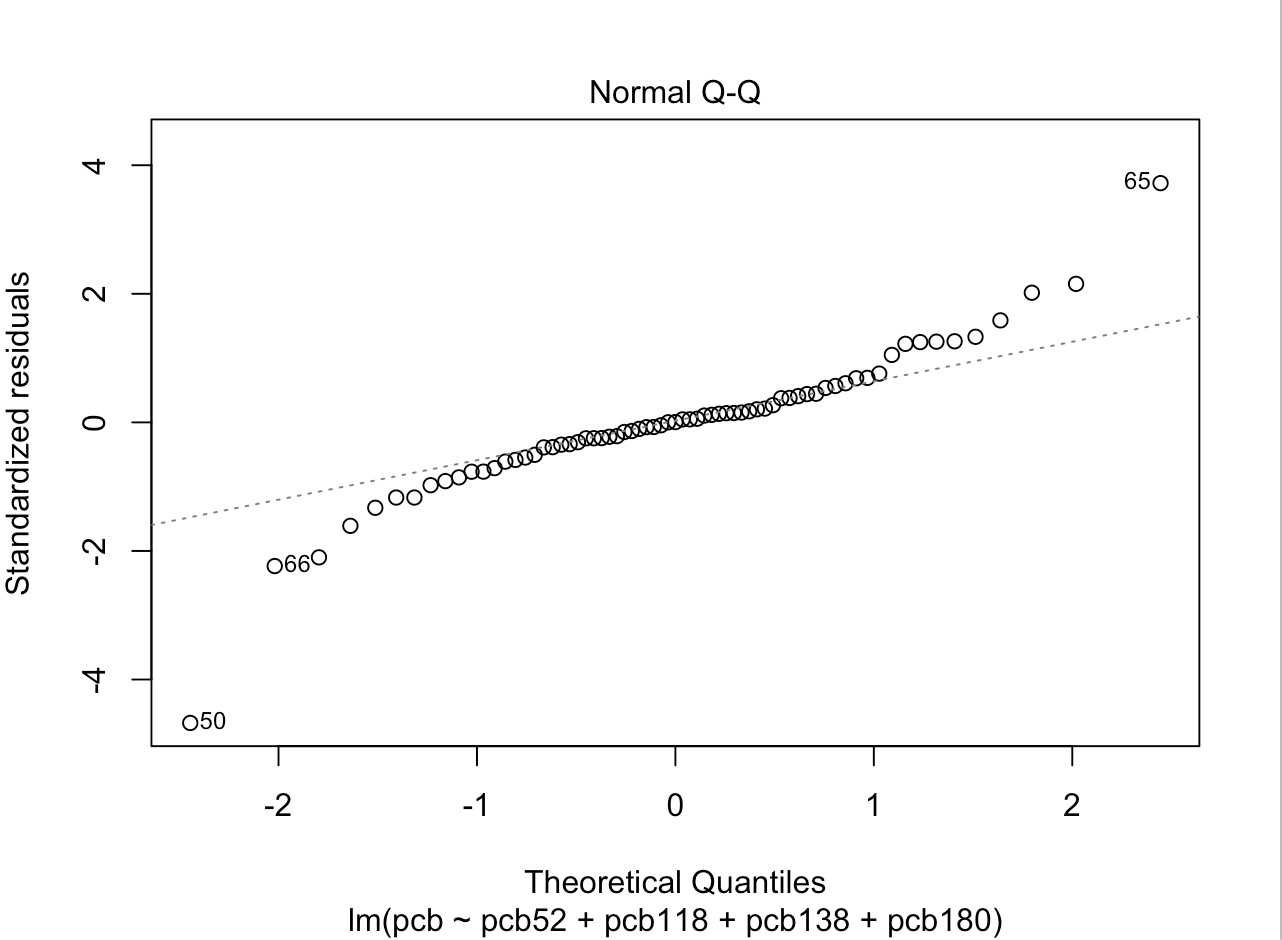
pcb118 3.7611 0.6424 5.855 1.79e-07 \*\*\*

pcb138 3.8842 0.4978 7.803 7.19e-11 \*\*\*

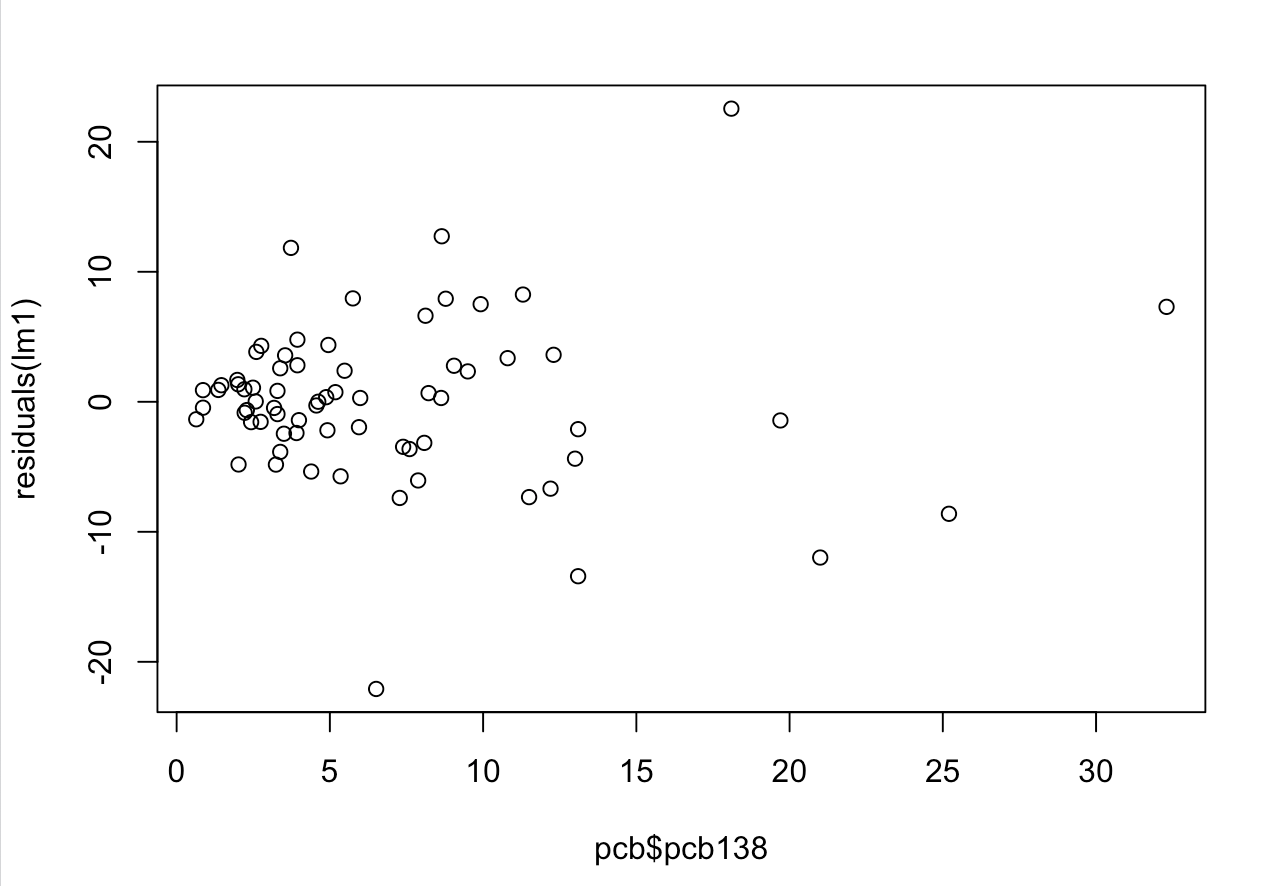
pcb180 4.1823 0.4318 9.687 3.64e-14 \*\*\*

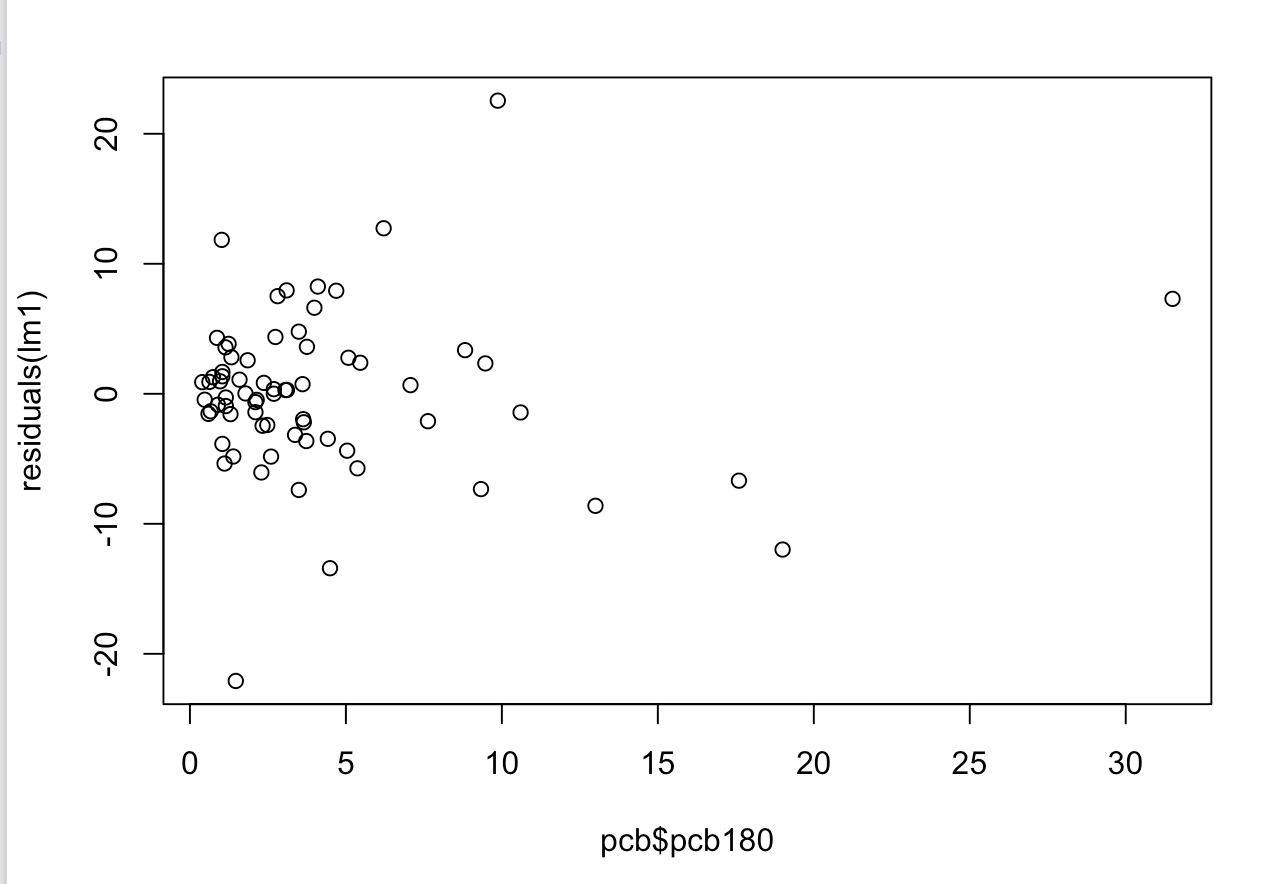
All coefficients vary greatly from zero except 0.9369, and multiple R-squared, R2=0.989 so 98.9% of variation is due to explanatory variables and the rest is unexplained, 1.01%. The residual standard error is 6.384 on 64 degrees of freedom.

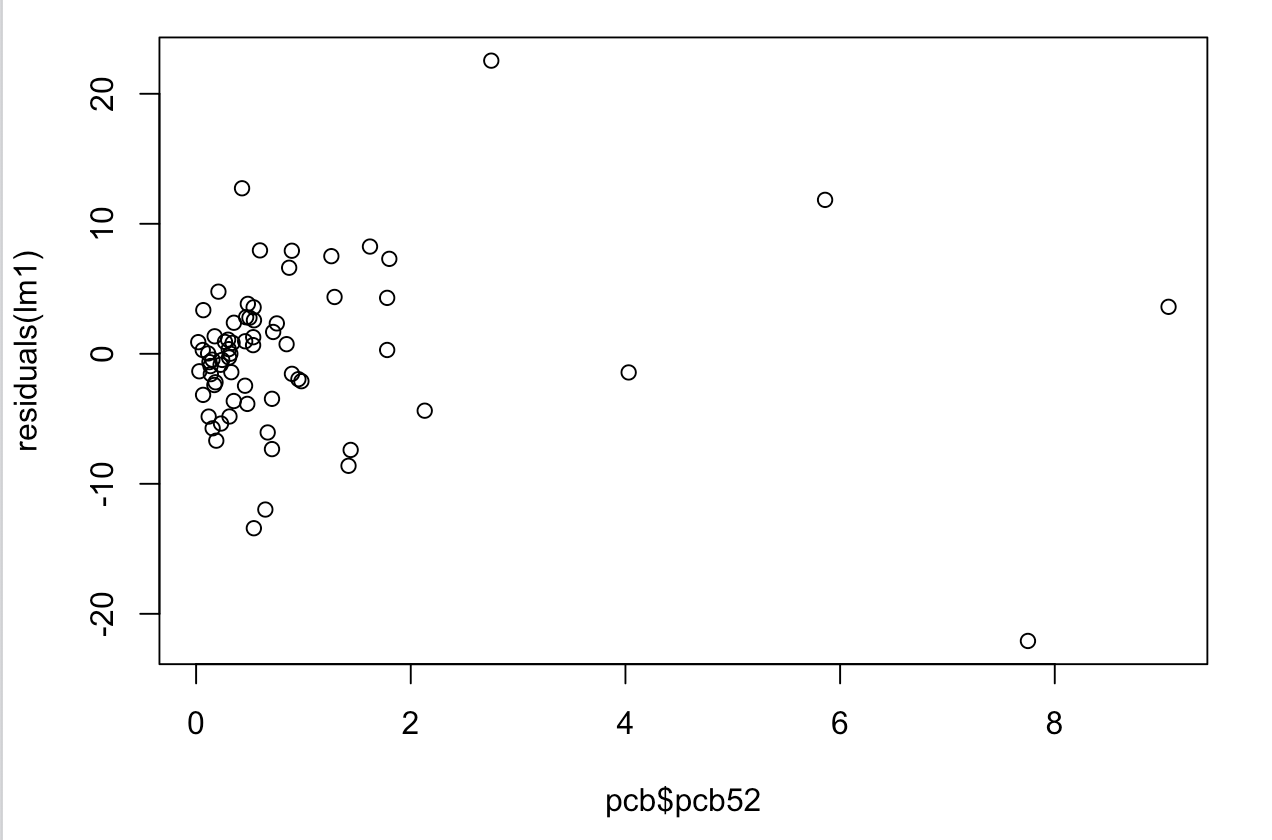
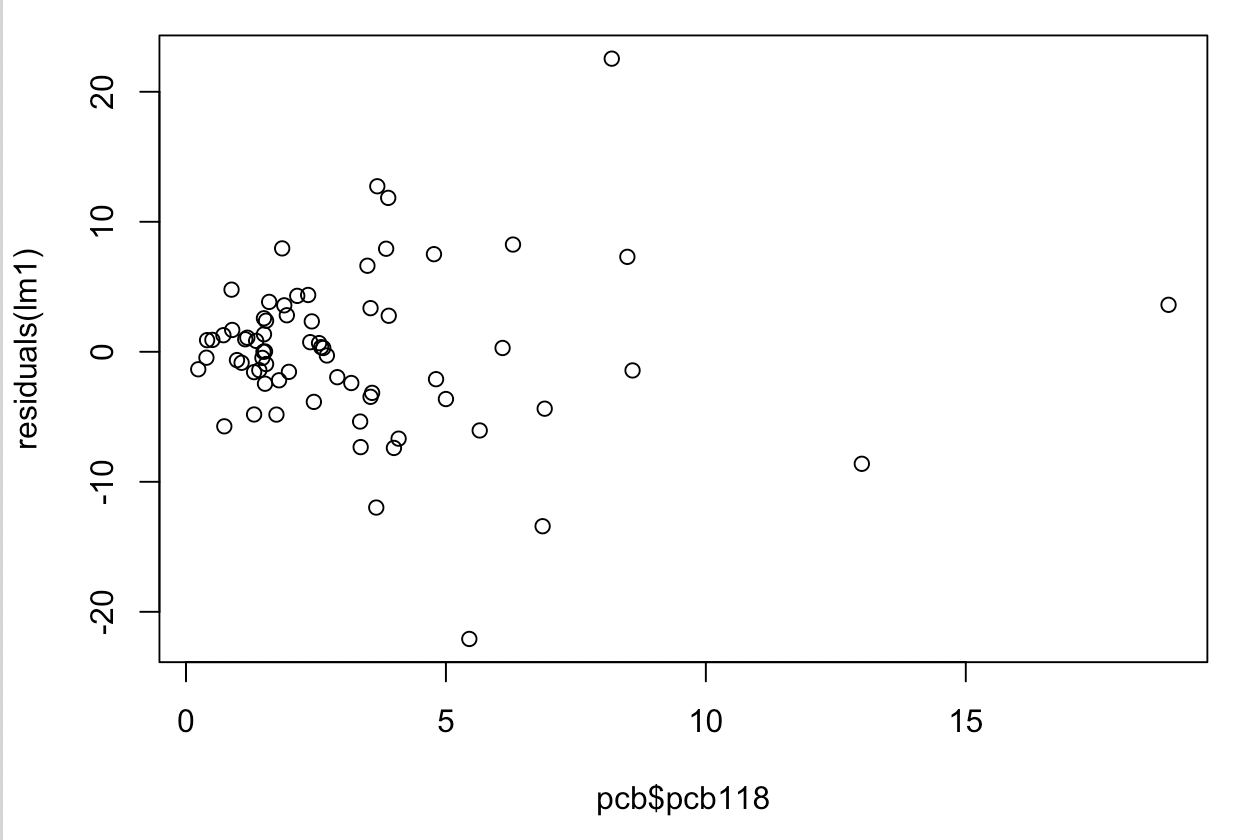
c.)



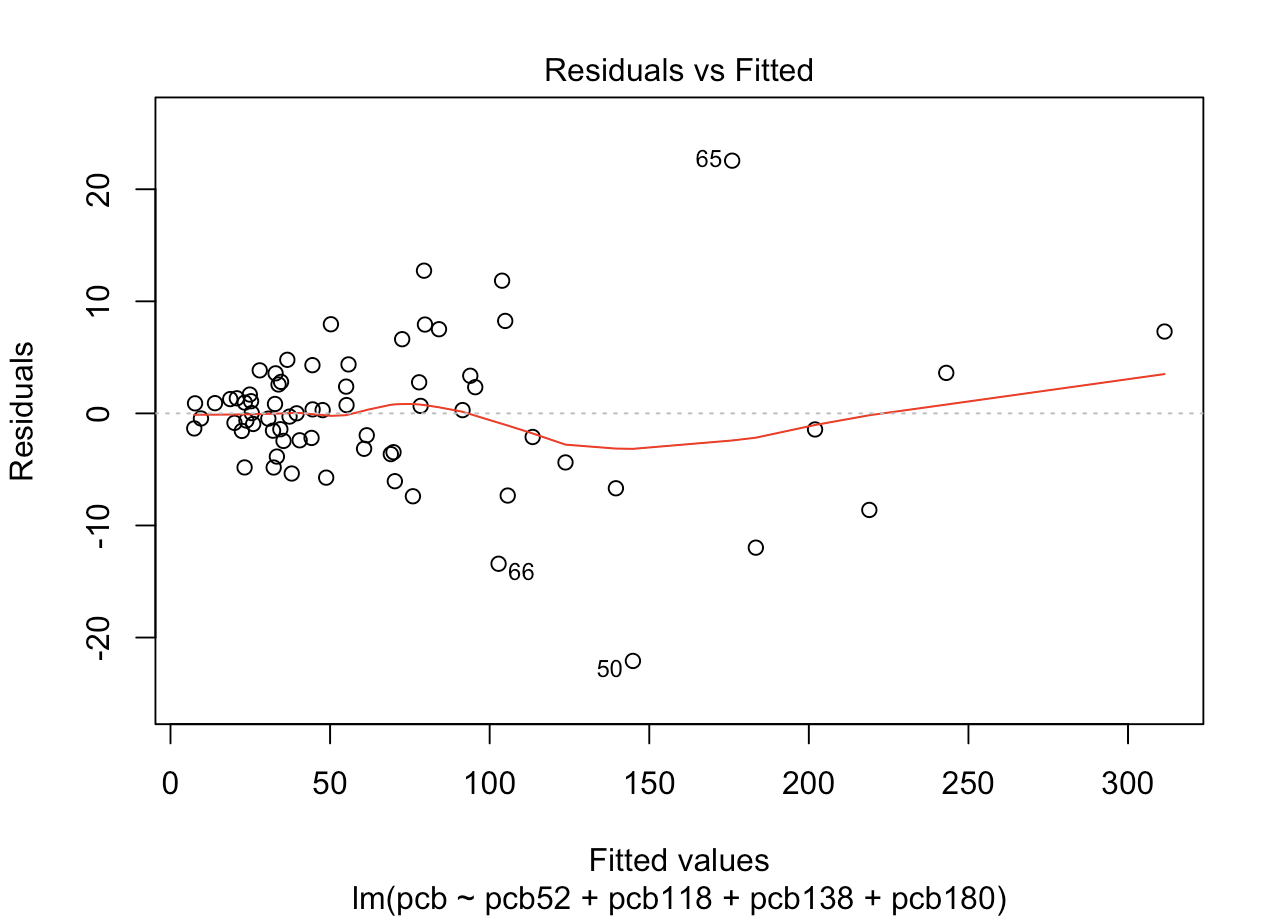
From the graphs, the residuals appear approximately normal and contain two outliers. The explanatory variables did not reveal a clear pattern, as seen below.







**11.44**

a.)

There is no pattern but there are two outliers, one

with a high residual (65) and one with a low

residual (50). Specimen 65 causes the overestimate

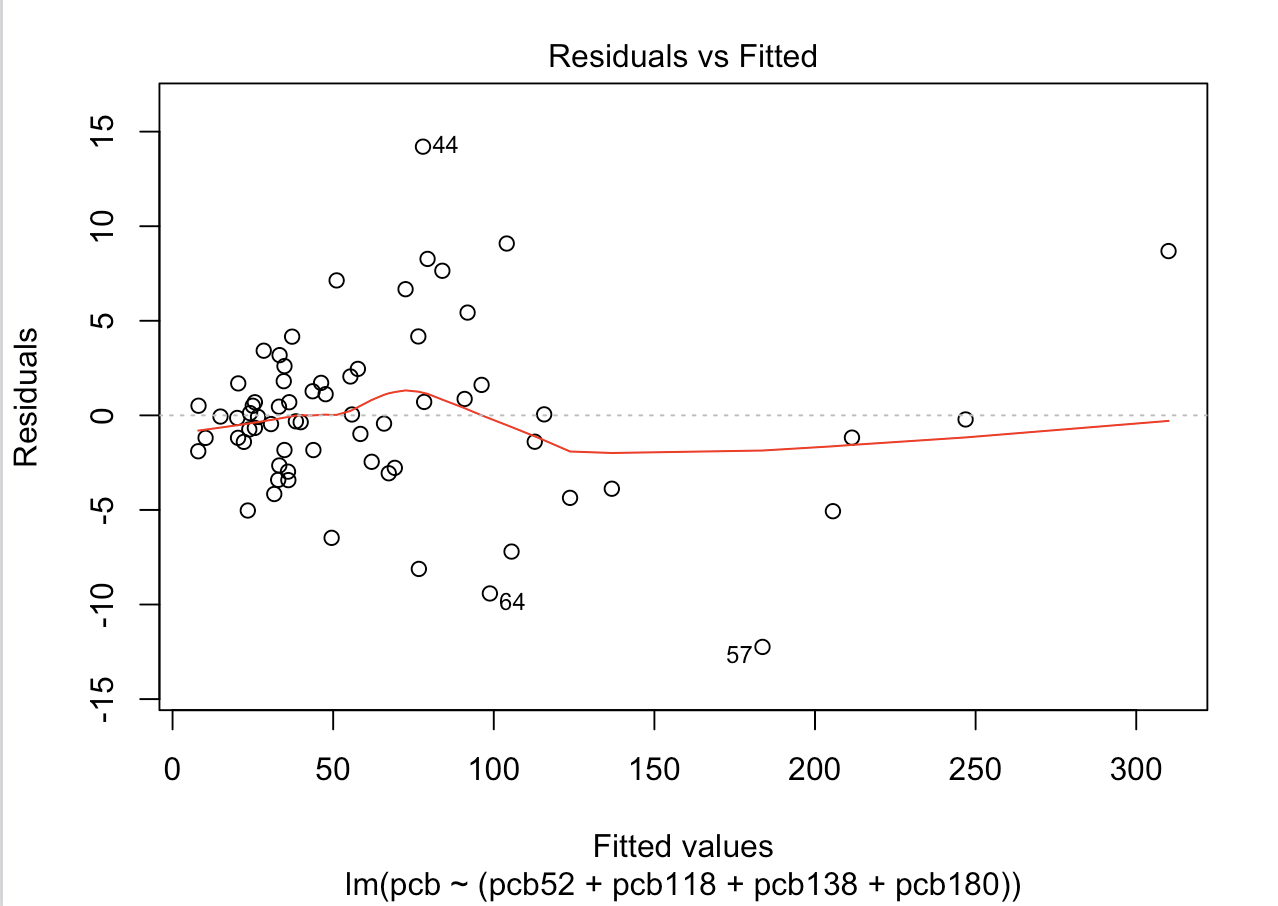
of PCB.

b.) New equation:

Call:

lm(formula = pcb ~ (pcb52 + pcb118 + pcb138 + pcb180), data = pcb2)

Residuals:

 Min 1Q Median 3Q Max

-12.2421 -2.1762 -0.1378 1.7036 14.2051

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.6277 0.8858 1.838 0.0709 .

pcb52 14.4420 0.6960 20.751 < 2e-16 \*\*\*

pcb118 2.5996 0.5164 5.034 4.40e-06 \*\*\*

pcb138 4.0541 0.3752 10.805 6.89e-16 \*\*\*

pcb180 4.1086 0.3175 12.942 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.555 on 62 degrees of freedom

Multiple R-squared: 0.9941, Adjusted R-squared: 0.9938

F-statistic: 2629 on 4 and 62 DF, p-value: < 2.2e-16

Upon the removal of outliers, Specimen 50 and 65, the residual standard error has gone down. It is now 4.555, whereas it was 6.382 before. Each variable’s estimated standard error has gone down as well. The data is approximately symmetric since this new residual plot does not show a pattern.

**11.45**

a.)

Call:

lm(formula = pcb ~ (pcb52 + pcb118 + pcb138), data = pcb)

Residuals:

Min 1Q Median 3Q Max

-29.6219 -3.3502 0.8791 3.3785 29.5217

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -1.0184 1.8895 -0.539 0.592

pcb52 12.6442 1.1291 11.198 <2e-16 \*\*\*

pcb118 0.3131 0.8333 0.376 0.708

pcb138 8.2546 0.3279 25.177 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.945 on 65 degrees of freedom

Multiple R-squared: 0.9732, Adjusted R-squared: 0.972

F-statistic: 786.7 on 3 and 65 DF, p-value: < 2.2e-16

Response variable=PCB, explanatory variables=PCB52,PCB118,PCB138

97.32% of PCB is explained by the explanatory variables.

b.)

α = 0.05

PCB118 coefficient=0.313

PCB118 P-value=0.708

Fail to reject null hypothesis if P-value > α

0.708 > 0.05, so fail to reject. This means that no statistically significant relationship exists.

c.)

From 11.43,

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.9369 1.2293 0.762 0.449

pcb52 11.8727 0.7290 16.287 < 2e-16 \*\*\*

pcb118 3.7611 0.6424 5.855 1.79e-07 \*\*\*

pcb138 3.8842 0.4978 7.803 7.19e-11 \*\*\*

pcb180 4.1823 0.4318 9.687 3.64e-14 \*\*\*

So, the coefficient for PCB118 is 3.761 and the P-value is 1.79e-07.

d.) By including or excluding one variable, there can be a drastic change. With the exclusion of PCB180 for example, PCB118’s coefficient went from3.76 to 0.313. PCB’s significance level changed drastically when PCB180 was removed and it’s P-value became not significant. It is now clear how big of a change excluding (or including) a specific variable can be.

**11.46**

a.)

ε

(β0 = 0, β1 = 1, β2 = 1, β3 =1)

b.)

c.)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 3.426e-07 1.917e-06 1.790e-01 0.859

teqpcb 1.000e+00 8.239e-07 1.214e+06 <2e-16 \*\*\*

teqdioxin 1.000e+00 1.761e-06 5.677e+05 <2e-16 \*\*\*

teqfuran 1.000e+00 5.664e-06 1.766e+05 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.95e-06 on 65 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 9.581e+11 on 3 and 65 DF, p-value: < 2.2e-16

Analysis of Variance Table

Response: teq

Df Sum Sq Mean Sq F value Pr(>F)

teqpcb 1 152.801 152.801 2.4174e+12 < 2.2e-16 \*\*\*

teqdioxin 1 26.903 26.903 4.2562e+11 < 2.2e-16 \*\*\*

teqfuran 1 1.970 1.970 3.1174e+10 < 2.2e-16 \*\*\*

Residuals 65 0.000 0.000

From running this regression, we learn the following:

The total TEQ is zero when all explanatory variables are zero. Also, the total TEQ will increase by 1 percent when TEQPCB, TEQDIOXIN, or TEQFURAN increase by 1.

The multiple R squared is 1 , so 100% of the response variable TEQ is explained by TEQPCB, TEQDIOXIN, and TEQFURAN (the explanatory variables).

So,

**11.47**

a.)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.059965 0.184450 5.747 2.73e-07 \*\*\*

pcb52 -0.097277 0.109383 -0.889 0.37716

pcb118 0.306184 0.096388 3.177 0.00229 \*\*

pcb138 0.105786 0.074697 1.416 0.16156

pcb180 -0.003905 0.064784 -0.060 0.95212

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9576 on 64 degrees of freedom

Multiple R-squared: 0.6769, Adjusted R-squared: 0.6568

F-statistic: 33.53 on 4 and 64 DF, p-value: 4.489e-15

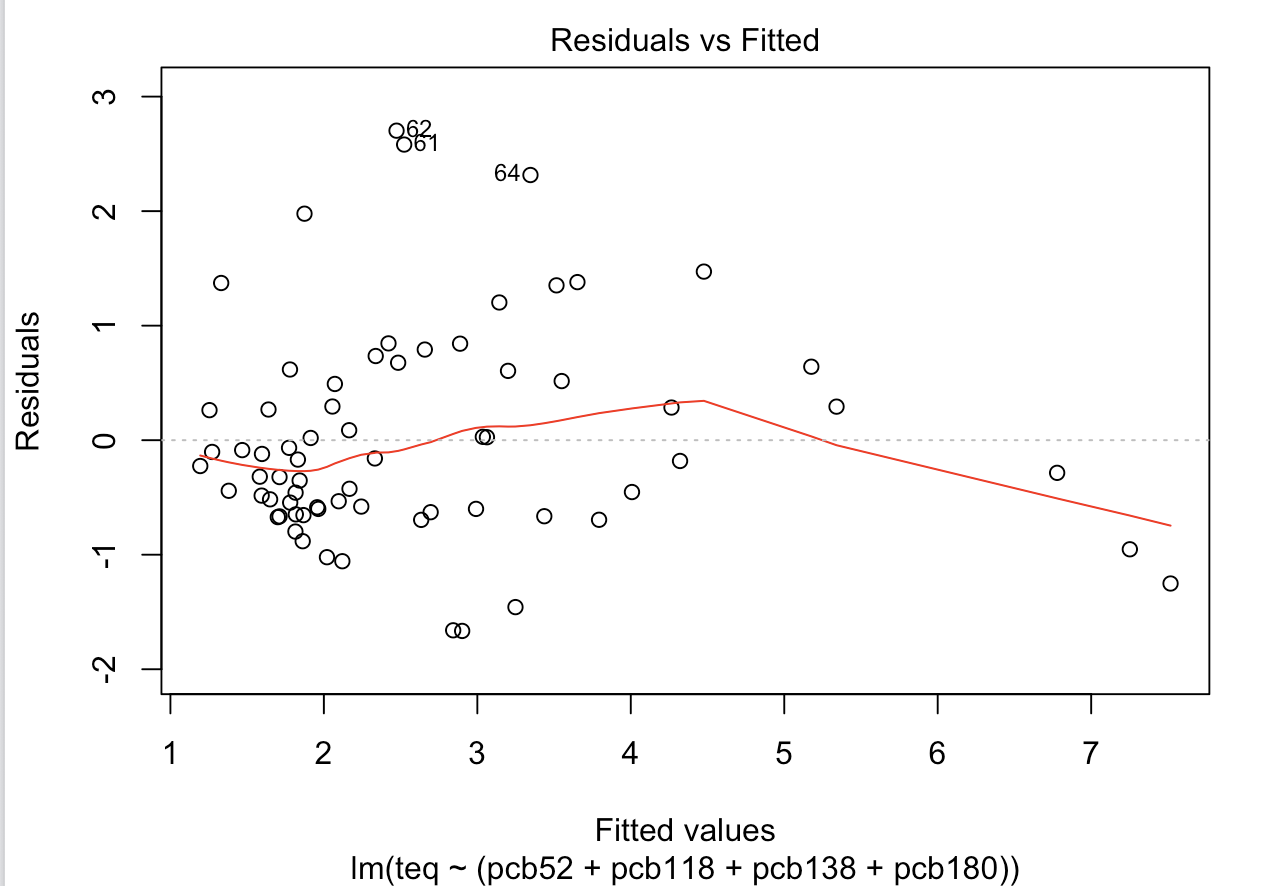
Df Sum Sq Mean Sq F value Pr(>F)

pcb52 1 29.85 29.85 32.553 3.21e-07 \*\*\*

pcb118 1 83.61 83.61 91.174 6.30e-14 \*\*\*

pcb138 1 9.52 9.52 10.378 0.00201 \*\*

pcb180 1 0.00 0.00 0.004 0.95212

Residuals 64 58.69 0.92

H0=β0 = β1 = β2 = β3 = β4 = 0

HA= at least one β not zero

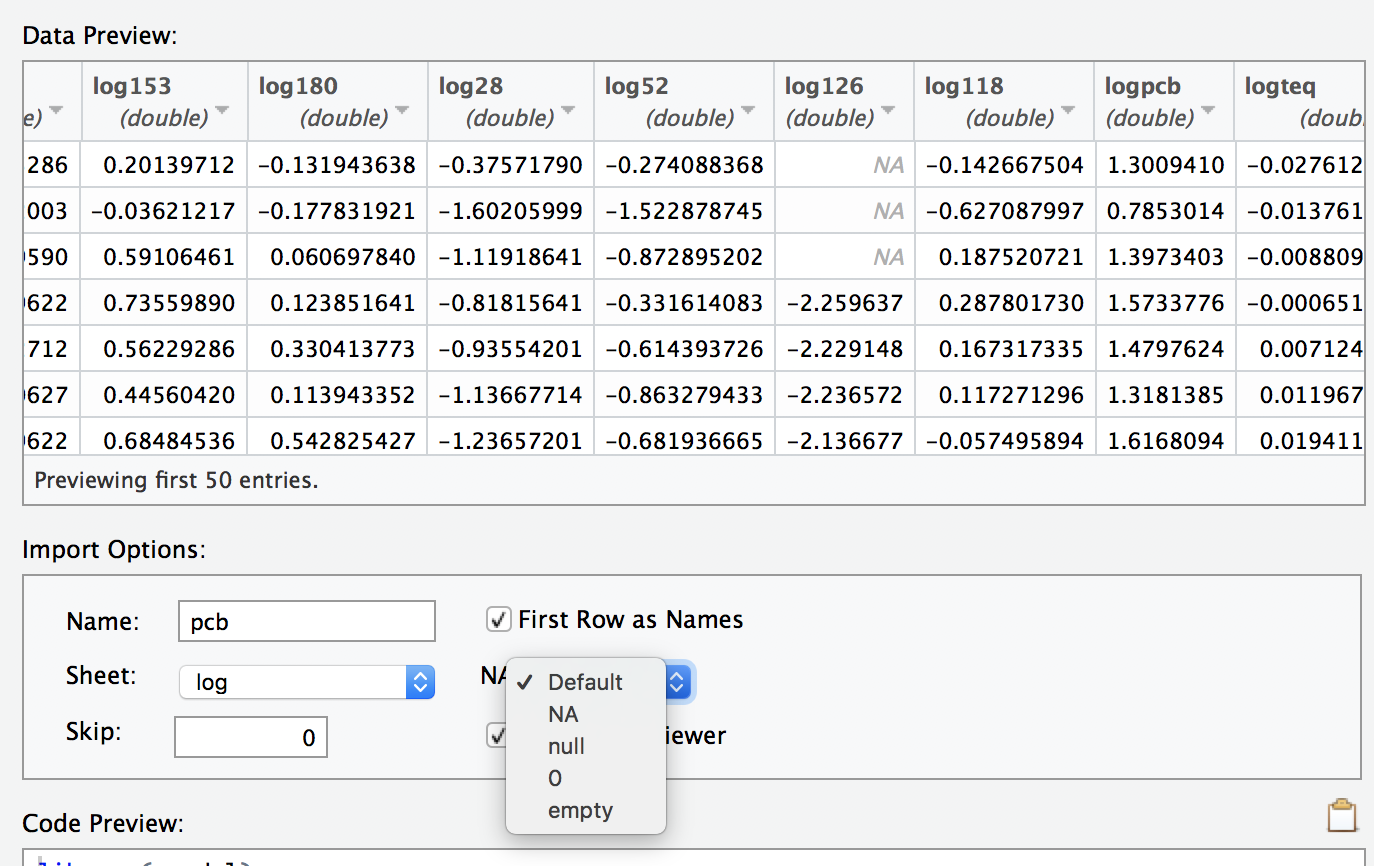
The residuals appear right skewed and do

not have a clear pattern.

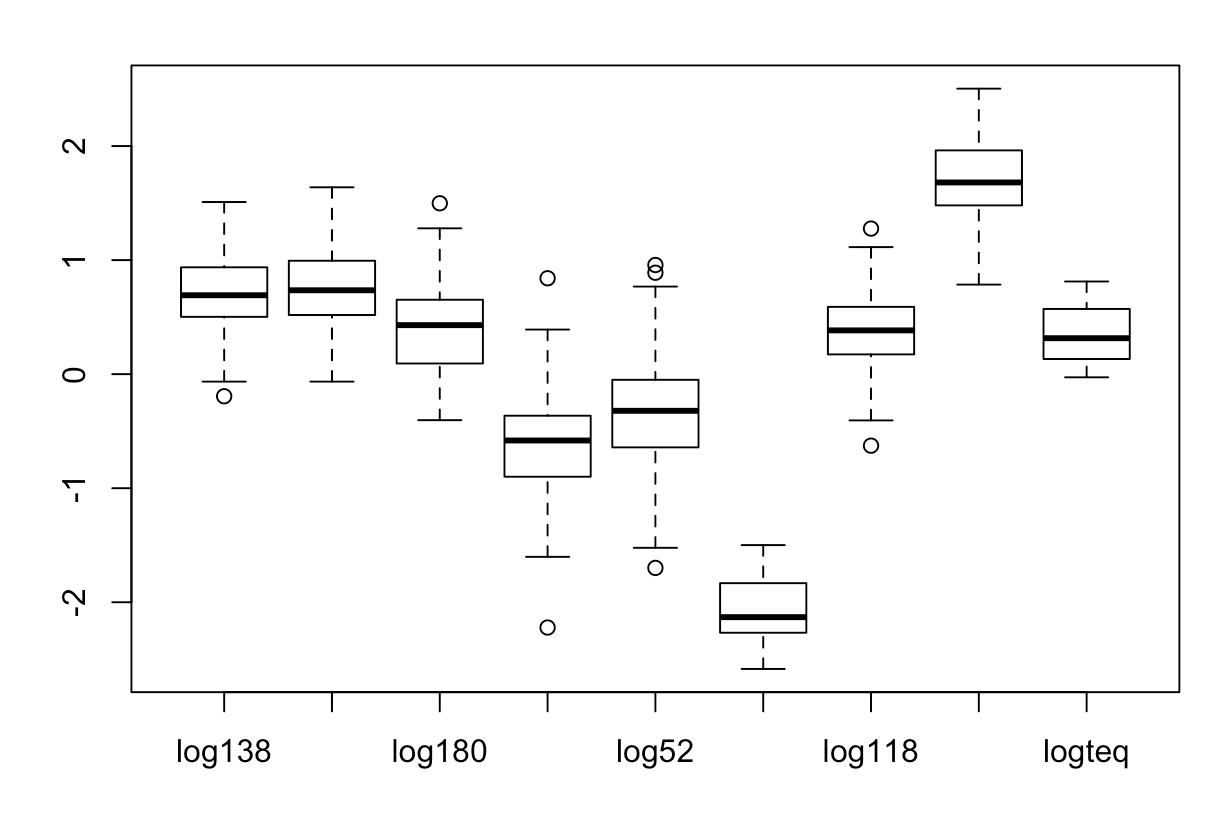
P-value for constant and PCB118 are very small which means that they are the only ones significantly different from 0.

**11.48**

a.) In excel, when the 16 zero values are logged, they become “#NUM!”. In R, it will be listed as NA which can be set to NA, null, 0, or empty.



b.) The software will not run regressions with cells that contain “NUM!” and there will be an error rather than ignoring it. This is a poor way of handling the situation because it is frustrating for the software to not just ignore them. If a data set is very large like this one (or possibly bigger) there could be many error values and being able to skip over them would be much more convenient for the user.

c.) Boxplots listed in order: 138, 153, 180, 28, 52, 126, 118, pcb, teq

**log138**  **log153** **log180** **log28** **log52**

Min. :-0.1938 Min. :-0.0655 Min. :-0.40340 Min. :-2.2218 Min. :-1.69897

1st Qu.: 0.5024 1st Qu.: 0.5185 1st Qu.: 0.09342 1st Qu.:-0.8996 1st Qu.:-0.64207

Median : 0.6920 Median : 0.7356 Median : 0.42975 Median :-0.5817 Median :-0.32148

Mean : 0.7009 Mean : 0.7397 Mean : 0.42354 Mean :-0.5793 Mean :-0.33537

3rd Qu.: 0.9370 3rd Qu.: 0.9943 3rd Qu.: 0.65225 3rd Qu.:-0.3645 3rd Qu.:-0.04964

Max. : 1.5092 Max. : 1.6385 Max. : 1.49831 Max. : 0.8407 Max. : 0.95713

**log126** **log118** **logpcb** **logteq**

Min. :-2.585 Min. :-0.6271 Min. :0.7853 Min. :-0.02761

1st Qu.:-2.268 1st Qu.: 0.1732 1st Qu.:1.4798 1st Qu.: 0.13274

Median :-2.131 Median : 0.3838 Median :1.6809 Median : 0.31534

Mean :-2.104 Mean : 0.3717 Mean :1.7011 Mean : 0.34950

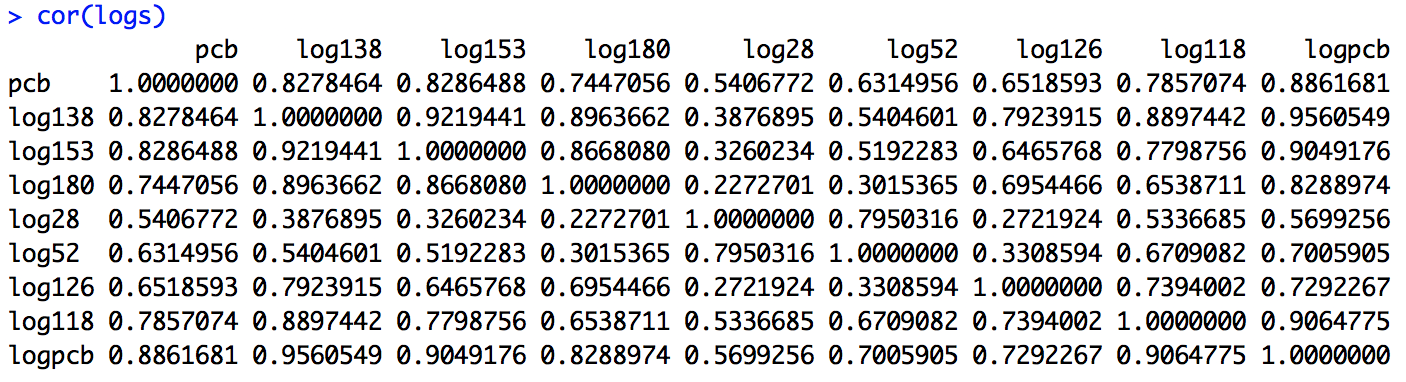
3rd Qu.:-1.833 3rd Qu.: 0.5899 3rd Qu.:1.9620 3rd Qu.: 0.57174

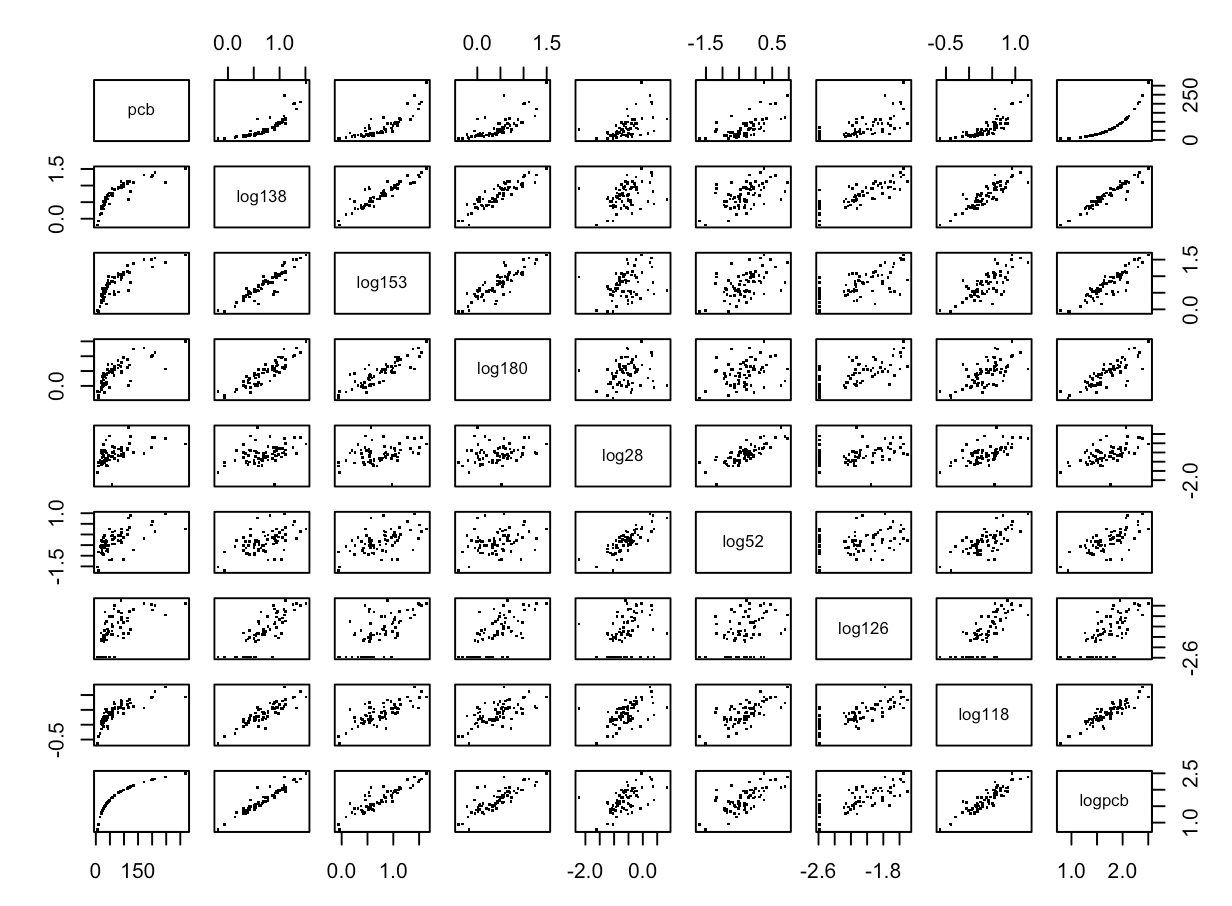
Max. :-1.499 Max. : 1.2765 Max. :2.5034 Max. : 0.81245

From these data, the logs of the variables appear to be approximately normal.

**11.49**

a.)

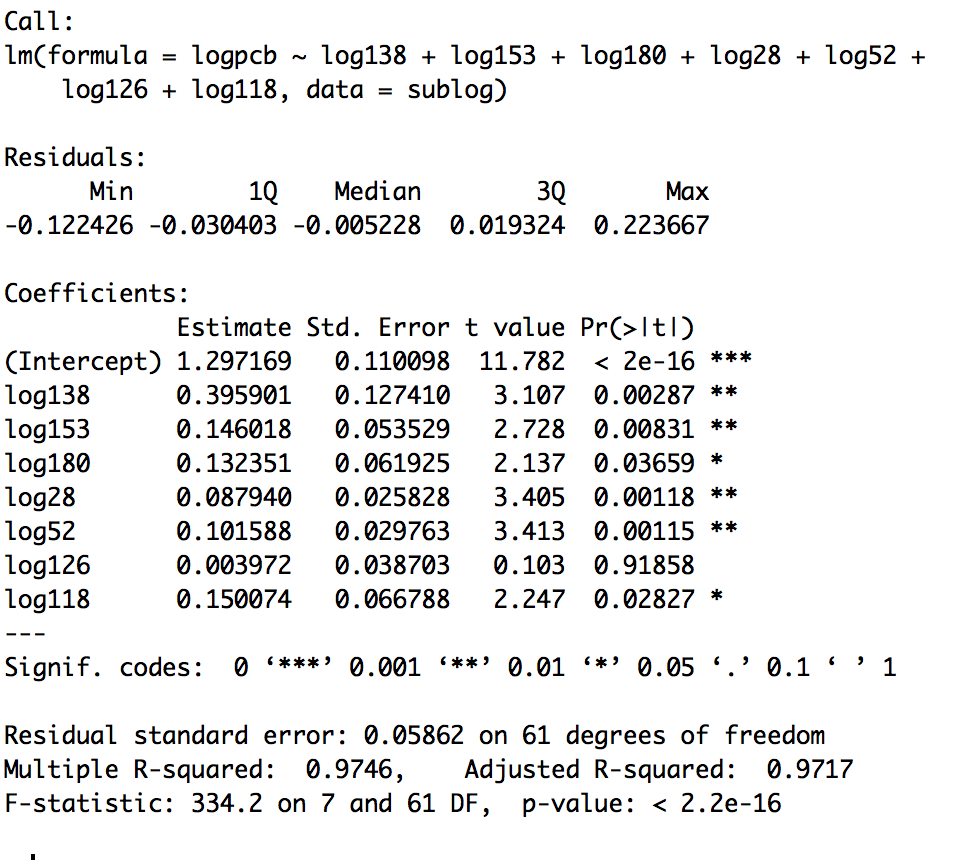
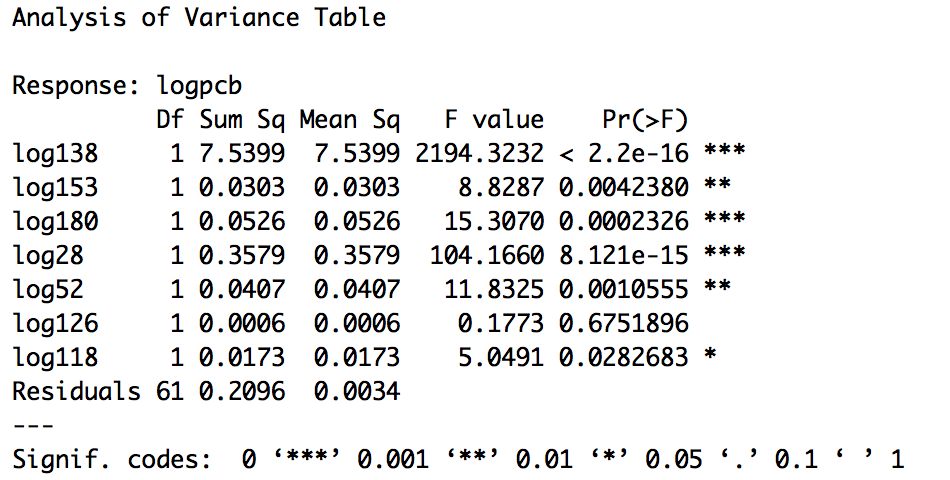




All pairs have positive correlation and logpcb28 has 1 outlier but other than that they are linearly correlated

b.) In both exercises, the pairs show a positive correlation but the log transformed variables have stronger/higher correlations when compared to the values in 11.42.

**11.50**

LOGPCB will increase by 0.396 percent when LOG138 increases by one, similarly LOGPCB will increase by 0.146 percent when LOG153 increases by one, and the same applies for the rest of the explanatory variables. R-squared is 0.9746 meaning that 97.46% of variation is due to explanatory variables and the rest is unexplained. The residual standard error is 0.05862 on 61 degrees of freedom.

Test1

H0=β0 = β1 = β2 = β3 = β4 = 0

HA= at least one β not zero

Reject H0 because there is at least one β that is not zero, as seen from the equation previously stated.

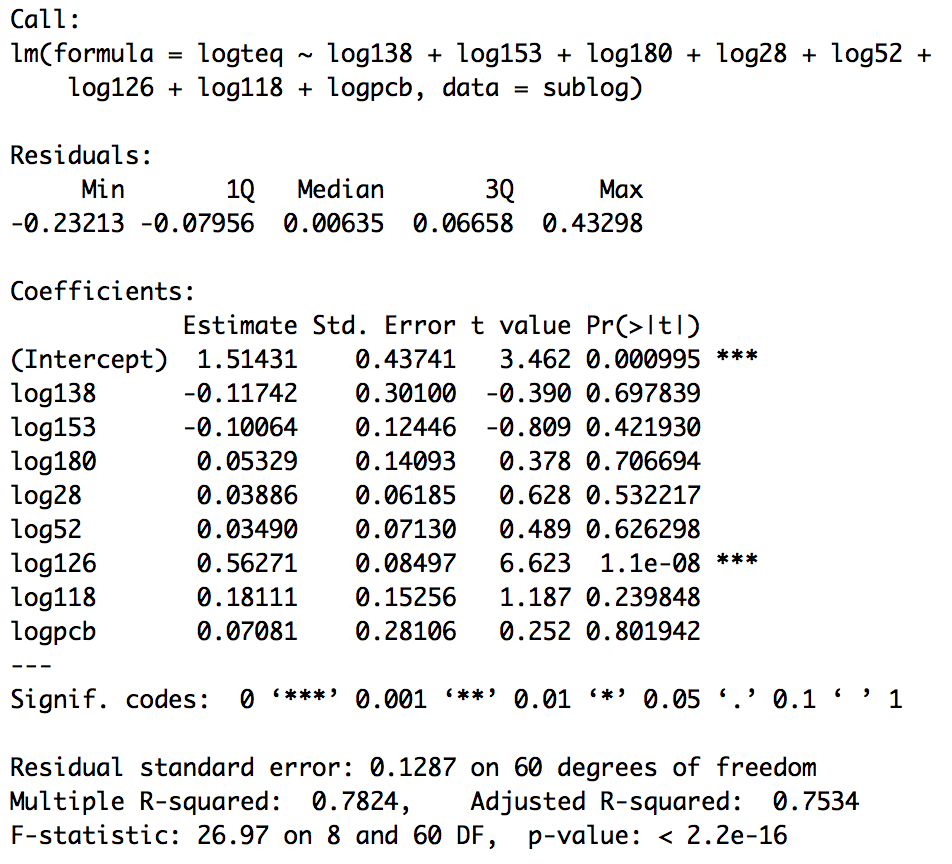
Test2

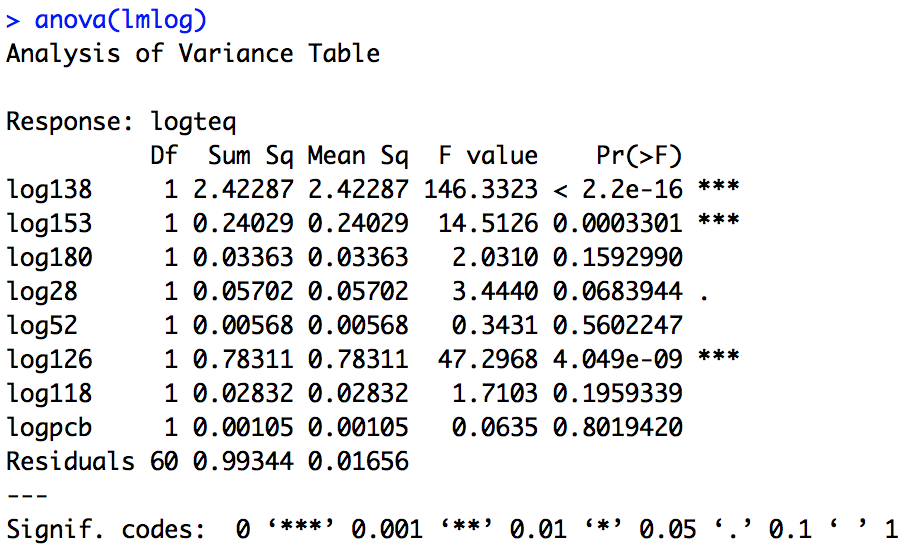
H0=β0 = β1 = β2 = β3 = β4........ = 0

HA=β0 , β1 , β2 , β3........ ≠ 0

The P-value is much smaller than 0 so we can reject null hypothesis and say that the coefficients are significantly different from 0.

**11.51**

****



LOGTEQ will decrease by 0.117 percent when LOG138 increases by one, similarly LOGTEQ will increase by 0.053 percent when LOG180 increases by one, and the same applies for the rest of the explanatory variables. R-squared is 0.7824 meaning that 78.24% of variation is due to explanatory variables and the rest is unexplained. The residual standard error is 0.1287 on 60 degrees of freedom.

Test1

H0=β0 = β1 = β2 = β3 = β4 = 0

HA= at least one β not zero

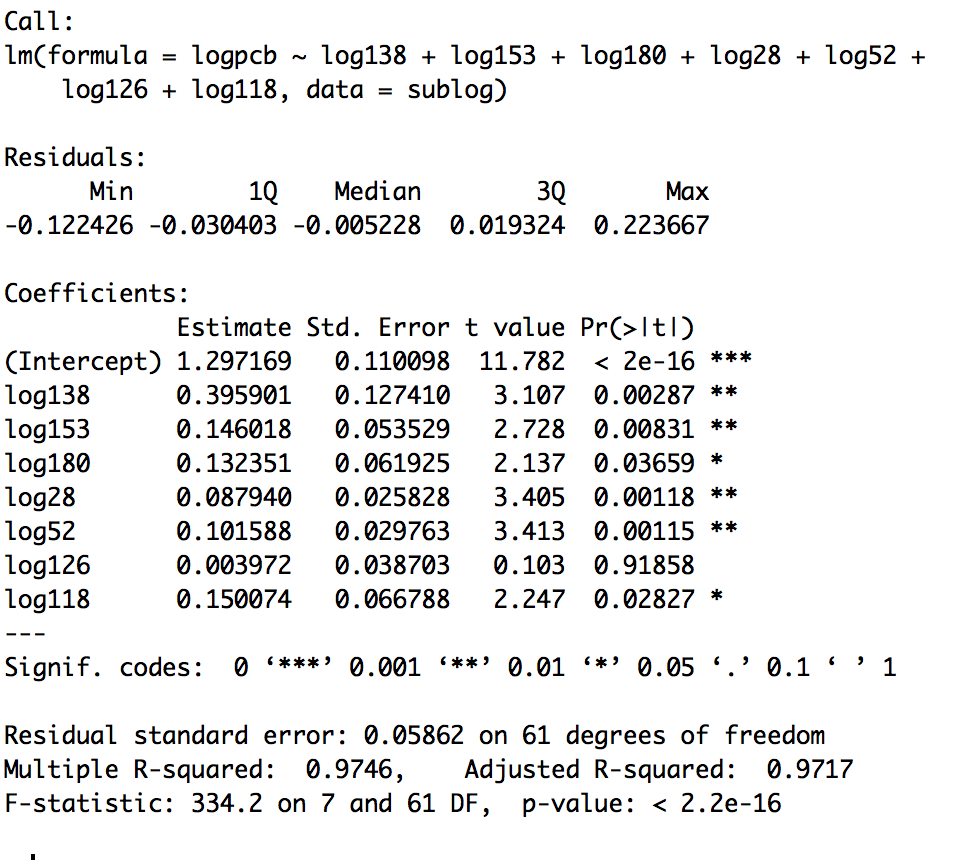
Reject H0 because there is at least one β that is not zero, as seen from the equation previously stated.

Test2

H0=β0 = β1 = β2 = β3 = β4........ = 0

HA=β0 , β1 , β2 , β3........ ≠ 0

The P-value is much smaller than 0 so we can reject null hypothesis and say that the coefficients are significantly different from 0.

**11.52**

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.6277 0.8858 1.838 0.0709 .

pcb52 14.4420 0.6960 20.751 < 2e-16 \*\*\*

pcb118 2.5996 0.5164 5.034 4.40e-06 \*\*\*

pcb138 4.0541 0.3752 10.805 6.89e-16 \*\*\*

pcb180 4.1086 0.3175 12.942 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.555 on 62 degrees of freedom

Multiple R-squared: 0.9941, Adjusted R-squared: 0.9938

F-statistic: 2629 on 4 and 62 DF, p-value: < 2.2e-16

Here we have two estimates for PCB, one using PCB52, PCB118, PCB138, and PCB180 (from 11.44) and the other an estimate for the log of PCB using the logs of all PCB variables (from 11.50). It is evident that using logs gives a better result, as the standard error from 11.50 is 0.05862 compared to 9.945 in 11.44. Also, all PCB variables were used in 11.50 where in 11.44 only 4 were used. By taking the log of all of these variables and using all of them, there is less error and is a better estimate for PCB than what was done in 11.44.