Part One. Statistical Report

Part Two. Textbook Exercises

11.42 Relationships among PCB congeners

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

(a) Using numerical and graphical summaries, describe the distribution of each of these variables.

Table 1: Numerical Summaries

Variable	Min	1st Qu.	Median	Mean	3rd Qu.	Max
PCB	6.10	30.18	47.96	68.47	91.63	318.70
PCB52	0.020	0.228	0.477	0.958	0.892	9.060
PCB118	0.236	1.490	2.420	3.256	3.890	18.900
PCB138	0.640	3.180	4.920	6.827	8.650	32.300
PCB180	0.395	1.240	2.690	4.158	4.490	31.500

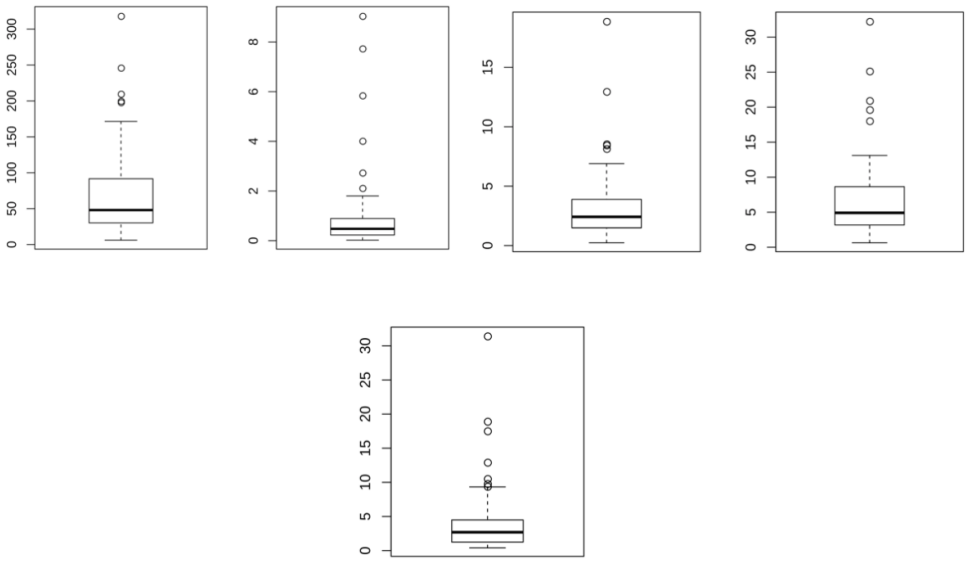


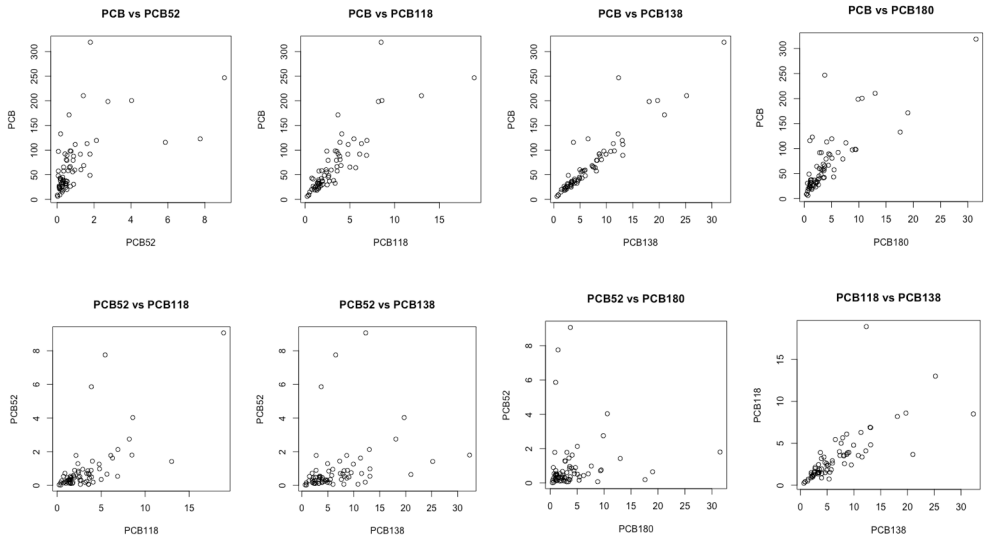
Figure 1: Boxplots of PCB, PBC52, PCB118, PCB138 and PCB180

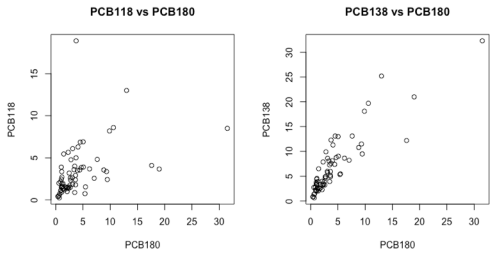
Figure 1 shows that the distribution of PCB and PCB180 is right skewed with about six outliers for both, while all the distribution of others are right skewed with about five outliers.

(b) Using numerical and graphical summaries, describe the relationship between each pair of variables.

Table 2: Correlations

Variable 1	Variable 2	Correlation
PCB	PCB52	0.5963572
PCB	PCB118	0.843298
PCB	PCB138	0.9288353
PCB	PCB180	0.8008549
PCB52	PCB118	0.6849073
PCB52	PCB138	0.3008983
PCB52	PCB180	0.08692971
PCB118	PCB138	0.7293792
PCB118	PCB180	0.4374443
PCB138	PCB180	0.8823022





11.43 Predictiong the total amount of PCB

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

(a) Write the statistical model for this analysis. Include all assumptions.

The multiple linear regression model for the data with 69 observations:

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + i$ *for* $i = 1, 2, \ldots, 69$

We assume that the residuals are independent and are normally distributed.

(b) Run the regression and summarize the results.

Multiple regression analyses were conducted to examine the relationship between PCB and four congeners. Running the multiple regression model in R with the four congeners produced the following:

```
subdf <- subset(df, select = c("pcb", "pcb52", "pcb118", "pcb138", "pcb180"))
> lm1 = lm(pcb~pcb52 + pcb118 + pcb138 + pcb180, data=subdf)
> coef(lm1)
(Intercept)       pcb52      pcb118      pcb138      pcb180
  0.9369203  11.8726953   3.7610694   3.8842264   4.1823010
> summary(lm1)

Call:
lm(formula = pcb ~ pcb52 + pcb118 + pcb138 + pcb180, data = subdf)

Residuals:
     Min       1Q   Median       3Q      Max
-22.0864  -2.4554   0.0278   2.7726  22.5487

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

Residual standard error: 6.382 on 64 degrees of freedom
Multiple R-squared:  0.9891,Adjusted R-squared:  0.9885
F-statistic:  1456 on 4 and 64 DF,  p-value: < 2.2e-16

> anova(lm1)
Analysis of Variance Table

Response: pcb
          Df Sum Sq Mean Sq  F value    Pr(>F)
pcb52      1  85302   85302 2094.273 < 2.2e-16 ***
pcb118     1  85429   85429 2097.405 < 2.2e-16 ***
pcb138     1  62693   62693 1539.202 < 2.2e-16 ***
pcb180     1   3822    3822   93.834  3.64e-14 ***
Residuals 64   2607      41
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We gathered the following from the results of the regression:
  - The multiple $R^2 = 0.989$
  - The residual SE = 6.249

Test 1

$H_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

$H_1 : \beta_0 \neq 0 \vee \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0 \vee \beta_4 \neq 0$

Since there is at least one $\beta_n \neq 0$, we reject H_0

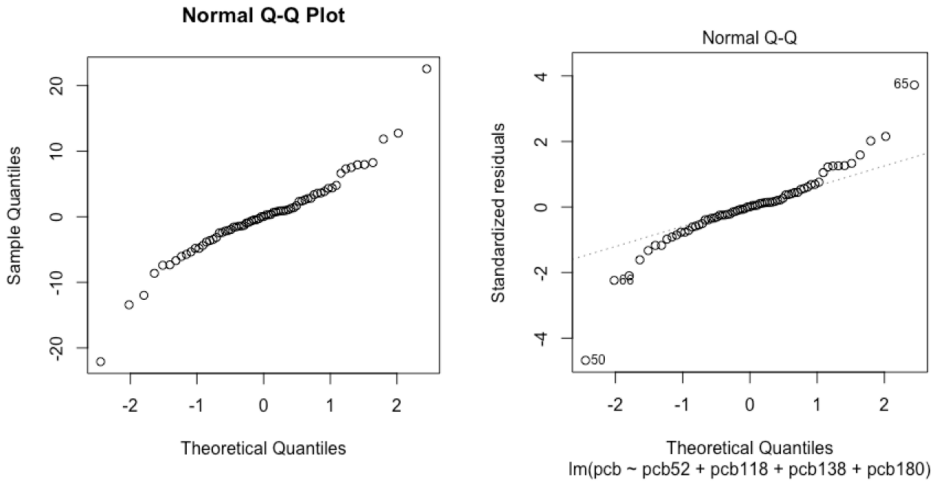
Test 2

$H_0 = \beta_j = 0$, *j = 0, 1, 2, 3*

$H_1 = \beta_j \neq 0$

All regression coefficients are significantly different from 0 with the except of 0.94. We found that $R^2 = 0.989$, meaning that 98.9% of variation in PCB is from PCB52, PCB118, PCB138 and PCB180.

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



According to the graphs, the residuals shows two clear outliers and shows that the residuals are approximately normal. Rhere are no other patterns in the explanatory variables of note.

11.44 Adjusting the analysis for potential outliers.

The examination of