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

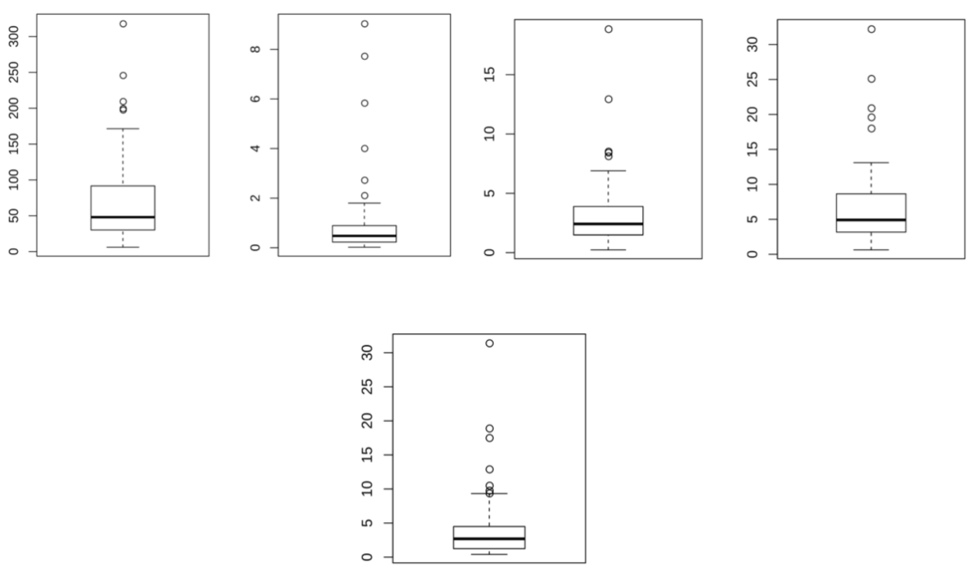


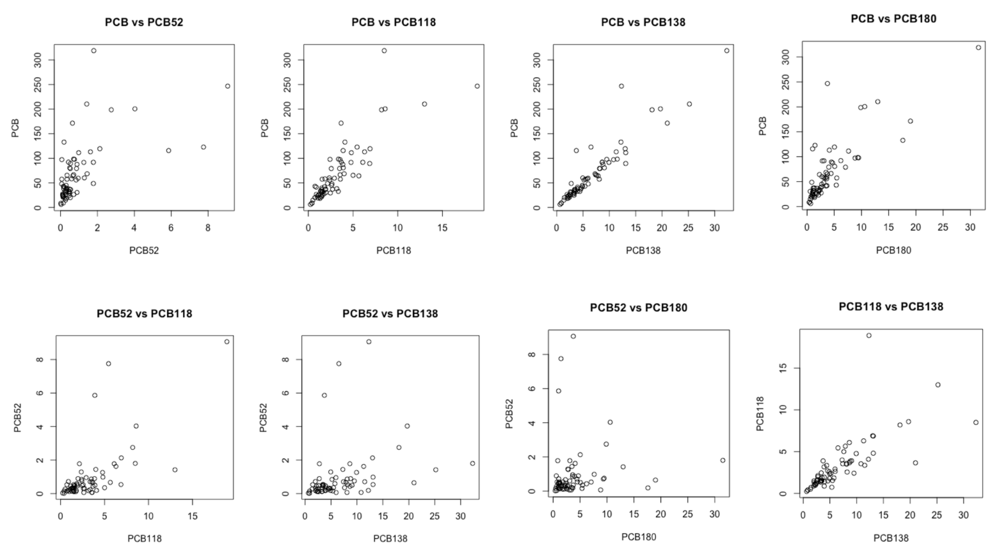
Figure 1: Boxplots of PCB, PBC52, PCB118, PCB138 and PCB180

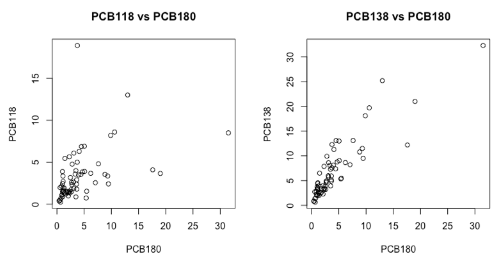
Figure 1 shows that the distribution of PCB and PCB180 is right skewed with about six outliers for both, while all the distribution of others are right skewed with about five outliers.

(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 $\mathrm{R}^2 = 0.989$
  - The residual SE $= 6.249$

Test 1

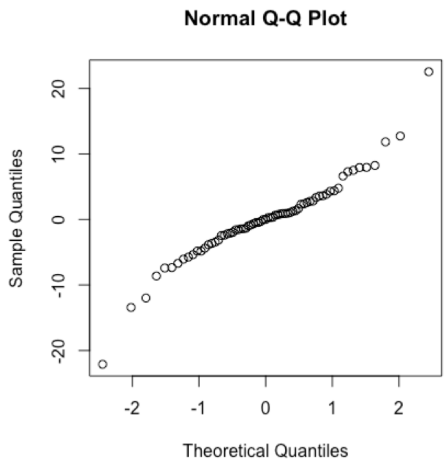
$\mathrm{H}_0 : \beta_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
$\mathrm{H}_1 : \beta_0 \neq 0 \vee \beta_1 \neq 0 \vee \beta_2 \neq 0 \vee \beta_3 \neq 0 \vee \beta_4 \neq 0$
Since there is at least one $\beta_\mathrm{n} \neq 0$, we reject H_0

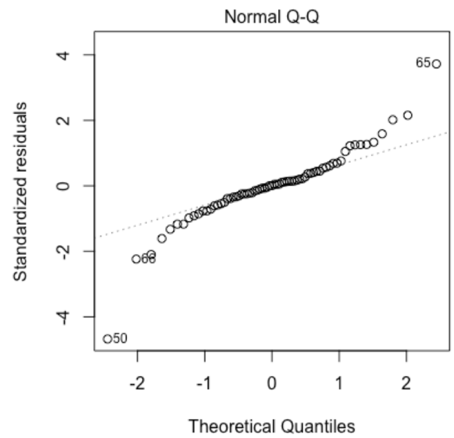
Test 2

$\mathrm{H}_0 = \beta_\mathrm{j} = 0$, $j = 0, 1, 2, 3$
$\mathrm{H}_1 = \beta_\mathrm{j} \neq 0$

All regression coefficients are significantly different from 0 with the except of 0.94. We found that $\mathrm{R}^2 = 0.989$, 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?



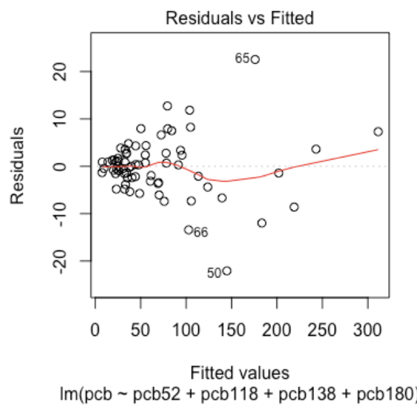


According to the graphs, the residuals shows two clear outliers and shows that the residuals are approximately normal. Rhere are no other patterns in the explanatory variables of note.

11.44 Adjusting the analysis for potential outliers.

The examination of 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.

(a) 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 specimen 50 and 65 are the two data points that are outliers. Specimen 65 corresponds to an overestimate of PCB due to its higher residual value.

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

```
(Intercept)       pcb52      pcb118      pcb138      pcb180
   1.627718   14.442021    2.599636    4.054061    4.108575
> summary(lm2)
Call:
lm(formula = pcb ~ pcb52 + pcb118 + pcb138 + pcb180, data = subdf2)

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

> anova(lm2)
Analysis of Variance Table
Response: pcb
          Df Sum Sq Mean Sq F value    Pr(>F)
pcb52      1  84307   84307  4062.7 < 2.2e-16 ***
pcb118     1  68740   68740  3312.6 < 2.2e-16 ***
pcb138     1  61670   61670  2971.9 < 2.2e-16 ***
pcb180     1   3476    3476   167.5 < 2.2e-16 ***
Residuals 62   1287      21
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The residual standard error has been decreased without the suspected outliers, from 6.382 to 4.555. R^2 has also increased from 0.989 to 0.994, meaning the predictions with this dataset become more accurate.

11.45 More on predicting the total amount of PCB.

Run a regression to predict PCB using the variables PCB52, PCB118, and PCB138. Note that this is similar to the analysis that you did in Exercise 11.43, with the change that PCB 180 is not included as an explanatory variable.

(a) Summarize the results.

```
> coef(lm3)
(Intercept)       pcb52      pcb118      pcb138
 -1.0183987  12.6441934   0.3131051   8.2545867
> summary(lm3)
Call:
lm(formula = pcb ~ pcb52 + pcb118 + pcb138, data = subdf3)

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

> anova(lm3)
Analysis of Variance Table
Response: pcb
          Df Sum Sq Mean Sq F value    Pr(>F)
pcb52      1  85302   85302  862.48 < 2.2e-16 ***
pcb118     1  85429   85429  863.77 < 2.2e-16 ***
pcb138     1  62693   62693  633.88 < 2.2e-16 ***
Residuals 65   6429      99
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can get the following values from the results of the regression:

- The squared multiple correlation coeffiicient $R^2 = 0.973$
- The residual standard error SE = 9.942

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

- Using a significance level $\alpha = 0.05$, Specimen PCB118 has a regression coefficient = 0.313 and P-value = 0.708
- Significance Test: $0.708 > 0.05$ (Reject when $P > \alpha$)
- P-value is much larger than the significance level. Therefore, we reject the null hypothesis.

(c) Find the estimate of the coefficient for PCB118 and the associated *P*-value for the model analyzed the Ecercise 11.43.

- Using a significance level $\alpha = 0.05$, Specimen PCB118(from Exercise 11.43) has a regression coefficient = 3.7611 and *P*-value = 0.000
- Significance Test: $0.000 < 0.05$ (Reject when $P > \alpha$)
- *P*-value is much smaller than the significance level. Therefore, we don't reject the null hypothesis.

(d) 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.

As parts (b) and (c) of this exercise show, the statistical significance of another variable is changed entirely, just by removing one explanatory variable. In the case above, removing the explanatory variable PCB180 made another explanatory variable PCB118 no longer statistically significant, along with drastically changing the variables corresponding regression coefficient and P-value.

11.46 Multiple regression model for total TEQ

(a) Consider using a multiple regression to predict TEQ using the tree components TEQPCB, TEQDIOXIN, and TEQFURAN as explanatory variables. Write the multiple regression model in the form: TEQ $= \beta_0 + \beta_1$TEQPCB $+ \beta_2$TEQDIOXIN $+ \beta_3$TEQFURAN $+ \epsilon$. Give numerical values for the parameters β_0, β_1, β_2, and β_3.

$\beta_0 = 0$, $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 1$
TEQ = 0 + 1 * TEQPCB + 1 * TEQDIOXIN + 1 * TEQFURAN

(b) 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 = s = 7.95e-6

(c) Use software to run this regression and summarize the results.

```
> lm4 <- lm(teq~teqpcb+teqdioxin+teqfuran, data=df)
> coef(lm4)
 (Intercept)       teqpcb    teqdioxin     teqfuran
3.425522e-07 1.000001e+00 1.000000e+00 1.000001e+00
```

```
> summary(lm4)

Call:
lm(formula = teq ~ teqpcb + teqdioxin + teqfuran, data = df)

Residuals:
       Min         1Q     Median         3Q        Max
-5.638e-06 -2.844e-06 -1.680e-06 -1.130e-06  3.714e-05

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

> anova(lm4)
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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- We gathered the following values from the results of the regression:
  - Multiple R-squared $R^2 = 1$
  - Residual standard error SE = 7.95e-06 ≈ 0

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 exception of the constant $R^1 = 1$, meaning 100% of TEQ is explained by TEQPCB, TEQDIOXIN and TEQFURAN.

11.47 Multiple regression model for total TEQ, cont.

```
Call:
lm(formula = teq ~ pcb52 + pcb118 + pcb138 + pcb180, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-1.6655 -0.6000 -0.1814  0.5162  2.7025

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

> summary(aov(lm5))
            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
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

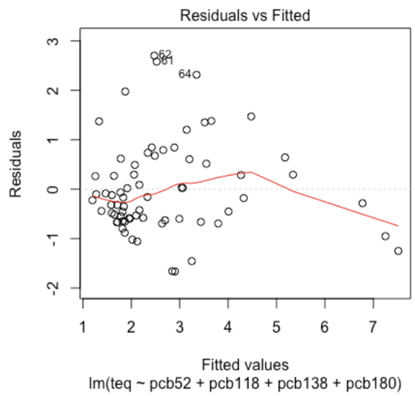
The regression equation used:

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

- Multiple R-squared $R^2 = 0.6772$
- Residual standard error SE = 0.9571

Significance Test:

- $H_0 = \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$
- H_a : one or more $\beta \neq 0$
- The *P*-value of both PCB118 and constant are close to 0, but still significantly different, therefore we reject null hypothesis.



When plotting the residuals, the data is skewed right but does not include any other obvious patterns.