Assigment 2 Answers

Chapter 3

1 (Question 3.7):

a. See R output for the plot. The fitted regression model is:

$$\hat{y} = 14.93 + 1.92x_1 + 7.00x_2 + 0.15x_3 + 2.72x_4 + 2.01x_5$$
$$- 0.41x_6 - 1.40x_7 - 0.04x_8 + 1.56x_9$$

b. H_0: The regression is not significant

H_1: The regression is significant

P-value $= 0.000185 < 0.05$ (See R output)

Conclusion: Reject H_0 and the regression is significant.

c. At $\alpha = 0.05$, none of the t-tests are significant since all P-values are greater than α.

d. x_3: lot size

x_4: living space

$H_0: \beta_3 = \beta_4 = 0$ vs H_1: One or both of β_3 and β_4 are not zero.

Extra Sum of Squares due to x_3, x_4:

$$SSE(\text{reduced model}) - SSE(\text{full model}) = 127.36 - 121.75 = 5.61$$

Partial F test:

$$F = \frac{5.61/2}{121.75/14} = 0.322 \quad ; \quad \text{P-Value} = (F_{2,14} > 0.322) = 0.730$$

P-value > 0.05, do not reject H_0. That is, given all other variables are included in the model, the contribution from lot size and living space is not significant.

e. From the R output, $VIF(x_6) = 11.71 > 10$. Thus multicollinearity is a potential problem.

2 (Question 3.11)

a. See R output for the plots. The fitted model is:

$$\hat{y} = 52.08 + 0.056\,x_1 + 0.28\,x_2 + 0.13\,x_3 + 0.00\,x_4 - 16.06\,x_5$$

Pressure　Temperature　Moisture　flowrate　Particle size.

b. H_0: the regression is not significant

H_1: The regression is significant

R output: $F = 29.86$, P-value $= 1.055 \times 10^{-5} < 0.05$

Reject H_0 and conclude regression is significant.

c. $H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$ $j = 1, 2, 3, 4, 5$

Coefficient	t statistic (df = 10)	P-value	
β_1	1.86	0.093	
β_2	4.90	0.001	< 0.05
β_3	0.31	0.763	
β_4	0.00	1.000	
β_5	−11.03	0.000	< 0.05

Regressors x_2 and x_5 appear to contribute significantly to the model.

d. Full model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$

$R^2 = 93.72\%$, $R^2_{adj} = 90.58\%$

Reduced model: $Y = \beta_0 + \beta_2 x_2 + \beta_5 x_5 + \varepsilon$

$$R^2 = 91.49\% \,,\quad R^2_{adj} = 90.18\%$$

The two models have approximately the same R^2 values.

e. $\hat{\beta}_2 \pm t_{0.025, df}\, s.e.(\hat{\beta}_2)$

df = 10

Full model: $0.282 \pm 2.228\,(0.0576) = (0.154, 0.410)$

Reduced model: $0.282 \pm 2.160\,(0.0588) = (0.155, 0.409)$

df = 13

The CI's are almost the same.

3. (Question 3.18 [4th] or 3.22 (5th])

$H_0: \beta_1 = \beta_2 = \cdots = \beta_K = 0$ vs

H_1: At least one $\beta_j \neq 0$, $j = 1, 2, \cdots, K$.

$$F_0 = \frac{R^2 (n-p)}{K(1-R^2)}$$

$$= \frac{SS_R}{SS_T}(n-p) \Big/ K\left(1 - \frac{SS_R}{SS_T}\right)$$

$$= \frac{SS_R / K}{(SS_T - SS_R)/(n-p)}$$

$$= \frac{MS_R}{MS_E} \sim F_{K, n-p} \text{ under } H_0$$

P-value $= P(F_{K,n-p} > F_0) < P(F_{K,n-p} > F_{\alpha,K,n-p}) = \alpha$ if $F_0 > F_{\alpha,K,n-p}$. Thus we reject H_0 when this happens.

4 (Question 3.21 [4th] or 3.25 [5th]).

Model: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$

$H_0: T\underline{\beta} = 0$ vs $H_1: T\underline{\beta} \neq 0$ $\boxed{\text{Rank}(T) = r}$

Test statistic: $F = \dfrac{(T\hat{\underline{\beta}})^T (T(X^TX)^{-1}T^T)\, T\hat{\underline{\beta}} / r}{SS_E(\text{Full model}) / (n-p)}$

$\sim F_{r, n-p}$ under H_0

a) $T = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & 0 & -1 \end{pmatrix}$, $r = 3$

b) $T = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$, $r = 2$

c) $T = \begin{pmatrix} 0 & +1 & -2 & -4 & 0 \\ 0 & 1 & 2 & 0 & 0 \end{pmatrix}$, $r = 2$

5 (Question 3.23 [4th] or 3.27 [5th])

$\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{Y}$

$\therefore \hat{\underline{Y}} = X\hat{\underline{\beta}} = X(X^TX)^{-1}X^T\underline{Y} = H\underline{Y}$, $H = X(X^TX)^{-1}X^T$

$\therefore V(\hat{\underline{Y}}) = V(H\underline{Y}) = HV(\underline{Y})H^T = H\sigma^2 I H^T = \sigma^2 H^2 = \sigma^2 H$

Or $V(\hat{\underline{Y}}) = V(X\hat{\underline{\beta}}) = XV(\hat{\underline{\beta}})X^T = X\sigma^2(X^TX)^{-1}X^T = \sigma^2 H$

6 (Question 3.27 [4th] or 3.31 [5th])

$$\underline{e} = \underline{y} - \hat{\underline{y}} = X\underline{\beta} + \underline{\varepsilon} - H\underline{y}$$

$$= X\underline{\beta} + \underline{\varepsilon} - H(X\underline{\beta} + \underline{\varepsilon})$$

$$= (X - HX)\underline{\beta} + (I - H)\underline{\varepsilon}.$$

Now, $HX = \left[X(X^TX)^{-1}X^T\right]X = X \quad \therefore \quad X - HX = 0$

$$\therefore \quad \underline{e} = (I - H)\underline{\varepsilon}$$

Or $\quad \underline{e} = \underline{y} - \hat{\underline{y}} = \underline{y} - H\underline{y}$

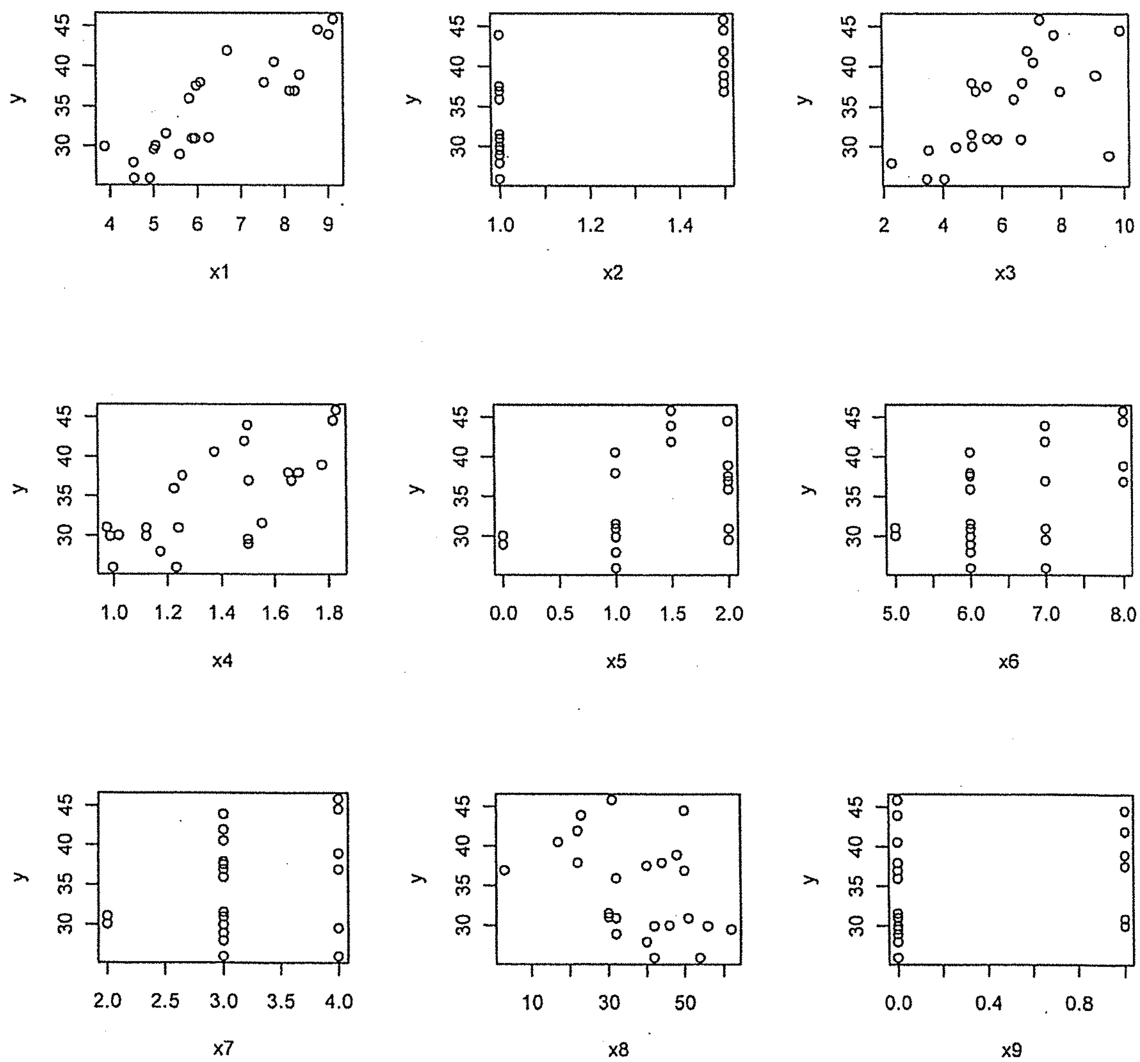
$$= (I - H)\underline{y}$$

$$= (I - H)(X\underline{\beta} + \underline{\varepsilon})$$

$$= (X - HX)\underline{\beta} + (I - H)\underline{\varepsilon}$$

$$= (I - H)\underline{\varepsilon}$$

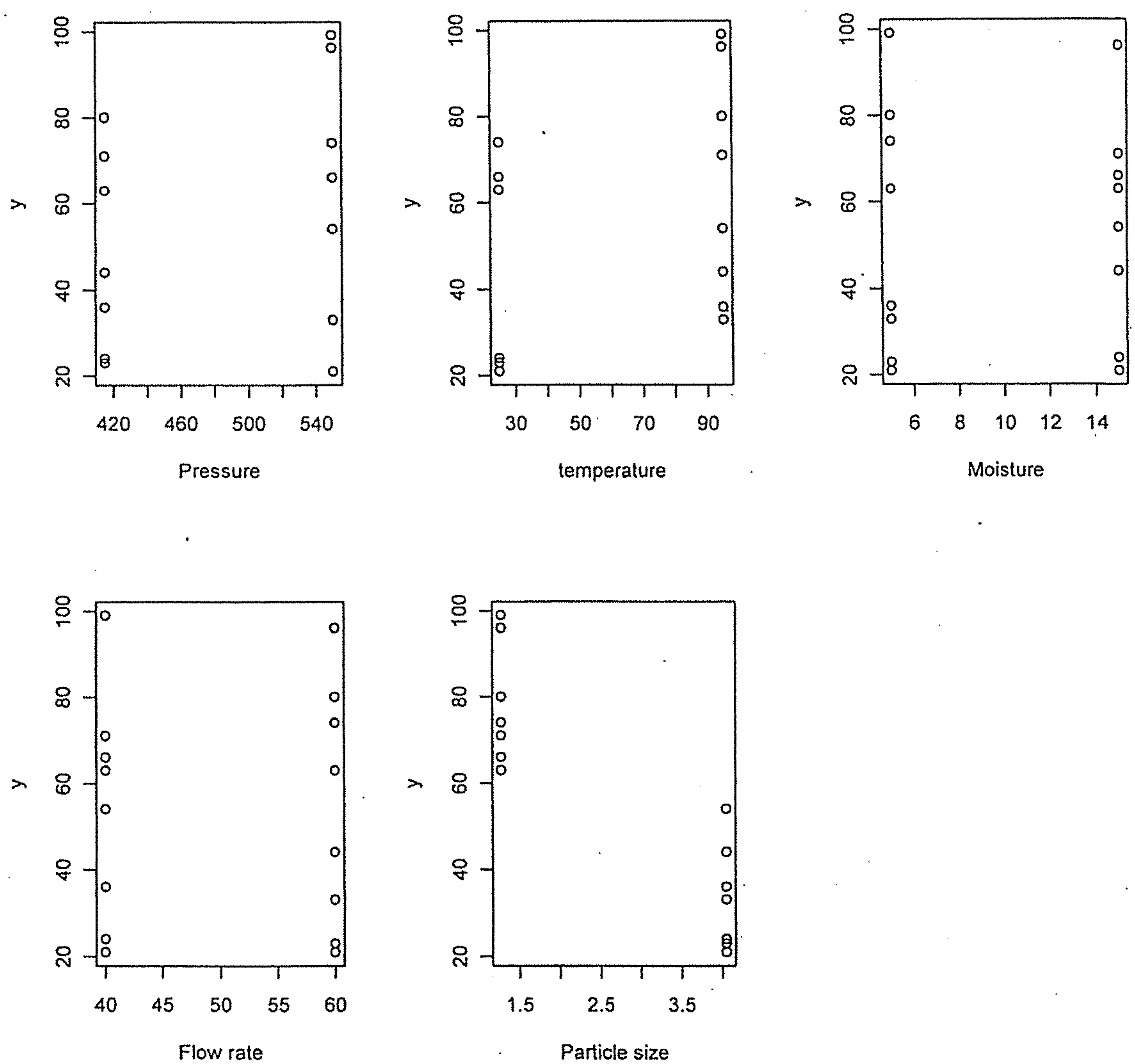
Question 3.7



```
#Question 3.7

#Read the data in R
data1=read.table(file="data-table-B4.prn", header=TRUE)
attach(data1)
print(data1)

#Plot the data
par(mfrow=c(3,3))
plot(x1,y,main="Question 3.7")
plot(x2,y)
plot(x3,y)
plot(x4,y)
plot(x5,y)
plot(x6,y)
plot(x7,y)
plot(x8,y)
plot(x9,y)

#Fit a multiple linear regression model

Model1=lm(y~x1+x2+x3+x4+x5+x6+x7+x8+x9)

summary(Model1)

#Fit another model without x3 and x4

Model2=lm(y~x1+x2+x5+x6+x7+x8+x9)

summary(Model2)

anova.lm(Model1, Model2)

#Multicollinearity?

library(car)
vif(Model1)     #load package: car


##################################################

output:
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9)

Residuals:
   Min     1Q Median     3Q    Max
-3.720 -1.956 -0.045  1.627  4.253

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14.92765    5.91285   2.525   0.0243 *
x1           1.92472    1.02990   1.869   0.0827 .
x2           7.00053    4.30037   1.628   0.1258
x3           0.14918    0.49039   0.304   0.7654
x4           2.72281    4.35955   0.625   0.5423
x5           2.00668    1.37351   1.461   0.1661
x6          -0.41012    2.37854  -0.172   0.8656
x7          -1.40324    3.39554  -0.413   0.6857
x8          -0.03715    0.06672  -0.557   0.5865
x9           1.55945    1.93750   0.805   0.4343
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.949 on 14 degrees of freedom
Multiple R-squared: 0.8531,     Adjusted R-squared: 0.7587
F-statistic: 9.037 on 9 and 14 DF,  p-value: 0.0001850
```

```
##################################

Call:
lm(formula = y ~ x1 + x2 + x5 + x6 + x7 + x8 + x9)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.4869 -1.9005 -0.2178  1.9221  4.1852 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 15.27482    5.54301   2.756   0.0141 *
x1           2.26902    0.86603   2.620   0.0186 *
x2           7.85962    3.70415   2.122   0.0498 *
x5           1.80882    1.28655   1.406   0.1789  
x6          -0.42813    2.26694  -0.189   0.8526  
x7          -0.89946    3.19175  -0.282   0.7817  
x8          -0.04113    0.06321  -0.651   0.5245  
x9           1.73134    1.73572   0.997   0.3334  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 2.821 on 16 degrees of freedom
Multiple R-squared: 0.8464,     Adjusted R-squared: 0.7792 
F-statistic: 12.59 on 7 and 16 DF,  p-value: 1.909e-05 

>
> anova.lm(Model1, Model2)
Analysis of Variance Table

Model 1: y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8 + x9
Model 2: y ~ x1 + x2 + x5 + x6 + x7 + x8 + x9
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     14 121.75                           
2     16 127.36 -2   -5.6083 0.3225 0.7296

##########################################

> vif(Model1)    #load package: car
       x1        x2        x3        x4        x5        x6        x7        x8 
 7.021036  2.835413  2.454907  3.836477  1.823605 11.710101  9.722663  2.320887 
       x9 
 1.942494 

##########################################################
```

Question 3.11



```
#Question 3.11

 Read the data in R
data1=read.table(file="data-table-B7.prn", header=TRUE)
attach(data1)
print(data1)

#Plot the data
par(mfrow=c(2,3))
plot(x1,y, xlab="Pressure",main="Question 3.11")
plot(x2,y,xlab="temperature")
plot(x3,y,xlab="Moisture")
plot(x4,y,xlab="Flow rate")
plot(x5,y,xlab="Particle size")

#Fit a multiple linear regression model

Model1=lm(y~x1+x2+x3+x4+x5)

summary(Model1)

#Fit another model with x2 and x5 only

Model2=lm(y~x2+x5)

summary(Model2)

###########################################
Output:

Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5)

Residuals:
    Min      1Q  Median      3Q     Max
-12.250  -4.438   0.125   5.250   9.500

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.208e+01  1.889e+01   2.757 0.020218 *
x1           5.556e-02  2.987e-02   1.860 0.092544 .
x2           2.821e-01  5.761e-02   4.897 0.000625 ***
x3           1.250e-01  4.033e-01   0.310 0.762949
x4           1.214e-16  2.016e-01   0.000 1.000000
x5          -1.606e+01  1.456e+00 -11.035  6.4e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.065 on 10 degrees of freedom
Multiple R-squared: 0.9372,     Adjusted R-squared: 0.9058
F-statistic: 29.86 on 5 and 10 DF,  p-value: 1.055e-05

###########################################

Call:
lm(formula = y ~ x2 + x5)

Residuals:
    Min      1Q  Median      3Q     Max
-15.375  -4.188  -0.875   3.438  12.625

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 80.13461    5.69146  14.080 3.01e-09 ***
x2           0.28214    0.05883   4.796 0.000349 ***
x5         -16.06498    1.48659 -10.807 7.26e-08 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.236 on 13 degrees of freedom
Multiple R-squared: 0.9149,     Adjusted R-squared: 0.9018
F-statistic: 69.89 on 2 and 13 DF,  p-value: 1.107e-07

################################################
```