MA 542 Homework 11

- **Question 1. Chapter 10. Page 415. Problem 10.8**
- **Part a**

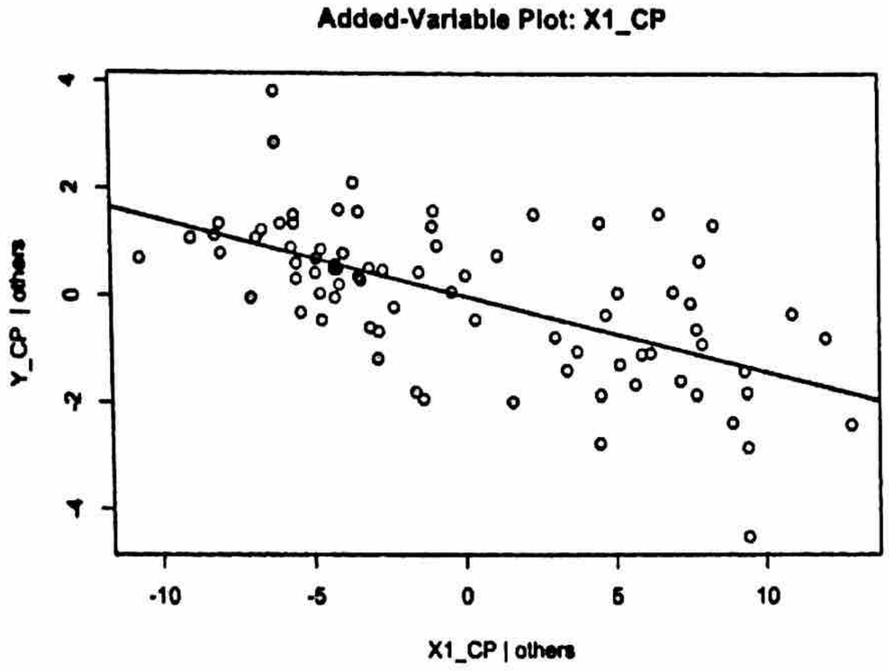


Figure 1 Added Variable Plot for X1

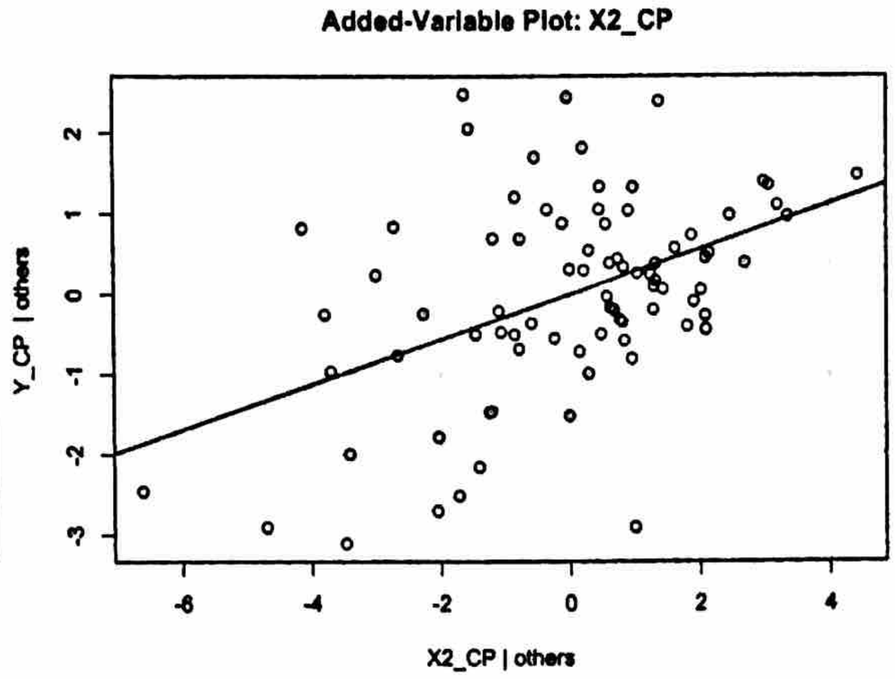


Figure 2 Added Variable Plot for X2

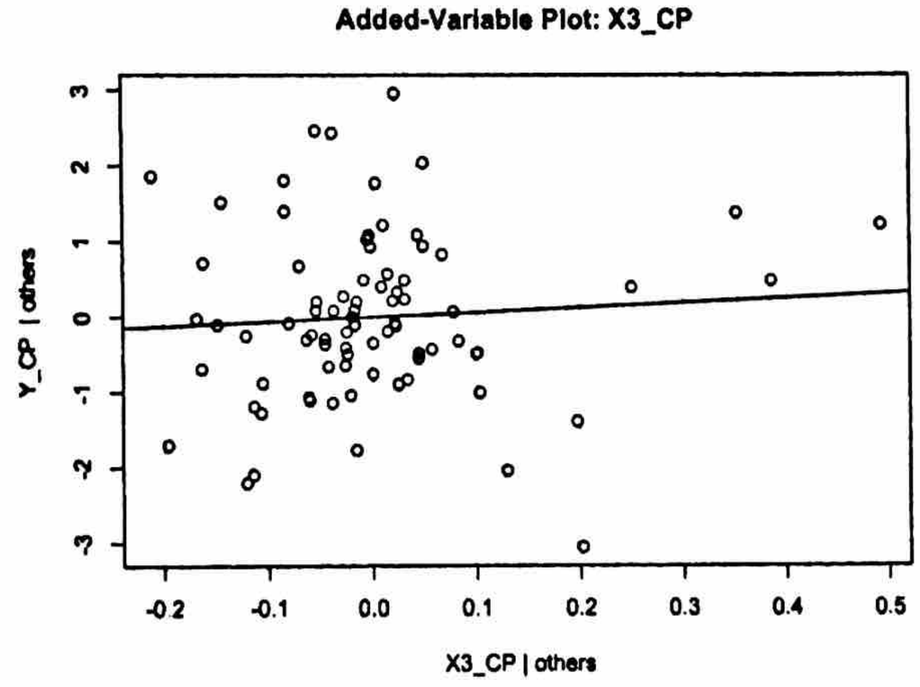


Figure 3 Added Variable Plot for X3

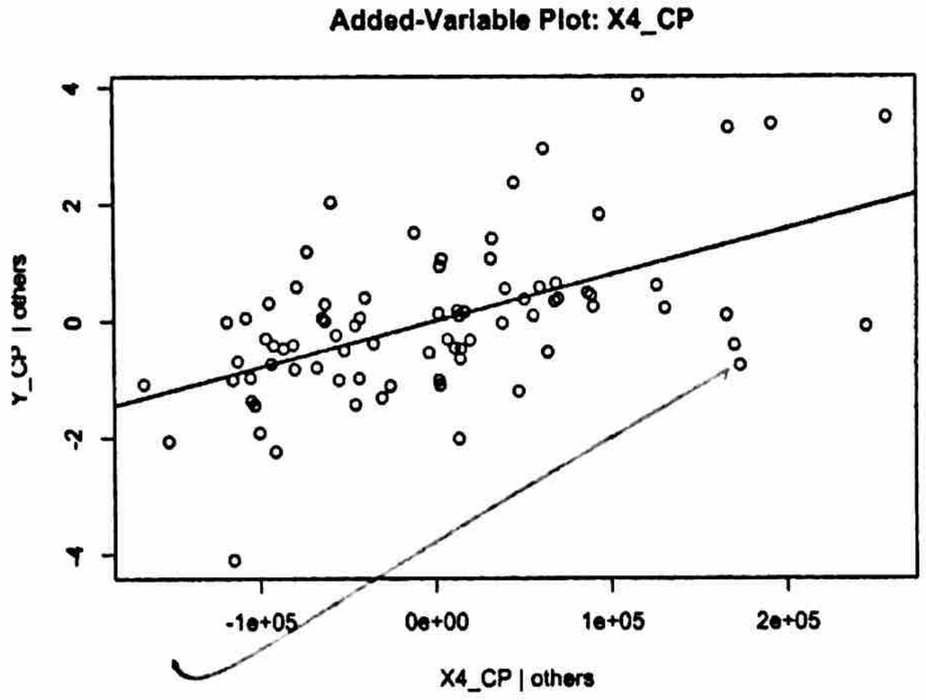


Figure 4 Added Variable Plot for X4

* Figure 1 shows a linear band with a negative nonzero slope, which indicates that a linear term with negative relationship in X_1 may be a helpful addition to the regression model already containing X_2, X_3 and X_4.
* Figure 2 shows a linear band with a positive nonzero slope, which indicates that a linear term with positive relationship in X_2 may be a helpful addition to the regression model already containing X_1, X_3 and X_4.
* Figure 3 shows a band with an almost zero slope, which indicates that X_3 contains no additional information useful for predicting Y beyond that contained in X_1, X_2 and X_4, so that it is not helpful to add X_3 the the regression model here.
* Figure 4 shows a linear band with a positive nonzero slope, which indicates that a linear term with positive relationship in X_4 may be a helpful addition to the regression model already containing X_1, X_2 and X_3.

- **Part b**

The plots in part a suggest that predictor variable X_1 has a negative relationship to Y, X_2 and X_4 have positive

relationship to Y, while X3 do not have significant influence on Y. These results are consistent with the results in 6.18 (part c), which show that the coefficient of X_1 is negative, coefficients of X2 and X4 are positive, the P-value of X3 is 0.57, greater than 0.05.

• Question 2. Chapter 10. Page 415. Problem 10.10

• Part a

I obtained the studentized deleted residuals as following:

```
         1          2          3          4          5          6          7
-0.22408721  1.22548992 -0.17058919 -0.38465341  0.59079235  0.19612400 -1.11220889
         8          9         10         11         12         13         14
-1.20529288 -0.97317127  2.03651735  0.93459504 -0.23775602 -0.41516264 -1.57563552
        15         16         17         18         19         20         21
 0.16177699 -0.94585526  1.27571152 -0.53946529  0.76695364 -1.30688316 -0.37866869
        22         23         24         25         26         27         28
 0.09348668 -0.86199405  0.58204372  0.34737815 -0.16427703  0.55395252 -0.41694578
        29         30         31         32         33         34         35
 0.43222636 -0.16381815 -0.98793509 -1.99766626  1.58402985  1.70041630 -1.66686255
        36         37         38         39         40         41         42
-0.48548055 -0.98726017  2.11878566 -0.77401405  2.17827155  0.20535369  0.59660176
        43         44         45         46         47         48         49
 0.88556619  0.43246953  0.27680517 -0.56690322 -1.04682225 -0.23443686  0.01909696
        50         51         52
 1.63020438 -1.37470496  0.45278953
```

Figure 5 Studentized Deleted Residuals

We shall use the Bonferroni simultaneous test procedure with a family significance level $\alpha = 0.05$. We therefore require: $t(1-\alpha/2n,n-p-1) = t(0.9995,47) = 3.551$

```
> for (i in 1:length(ti_GR))
+ {
+   if (abs(ti_GR[i]) > 3.551)
+     print(i)
+ }
> |
```

Figure 6 Results of Test

Since the there is no point whose $|t^*| > t(1-\alpha/2n,n-p-1) = t(0.9995,47) = 3.551$, then there does not exist any outlier here.

• Part b

```
> lev_GR = hat(model.matrix(fit_GR))
> lev_GR
 [1] 0.02258497 0.06179963 0.21887726 0.05297322 0.20632818 0.02712212 0.02861964 0.05635264
 [9] 0.04017169 0.04826901 0.03011634 0.04977033 0.02761134 0.06047246 0.03756448 0.25542493
[17] 0.03324965 0.05104935 0.02561758 0.02491881 0.19360472 0.25771995 0.05677233 0.07959049
[25] 0.05613301 0.02189441 0.02697280 0.06097409 0.03684681 0.04174658 0.03663401 0.09602318
[33] 0.04193292 0.02517837 0.04621057 0.06622225 0.03108517 0.03204566 0.04903249 0.03210502
[41] 0.04373193 0.12395571 0.28685861 0.22002363 0.11050577 0.03159426 0.06494377 0.28177664
[49] 0.02446692 0.03420197 0.10278142 0.02754093
```

Figure 7 Diagonal Element of the Hat Matrix

```
> lev_GR = hat(model.matrix(fit_GR))
> frame_GR[lev_GR > (2 * mean(lev_GR)),]
   Y_GR  X1_GR X2_GR X3_GR
3  4317 317164  4.61     0
5  4945 265518  8.61     1
16 4833 321773  5.82     1
21 4816 245674  7.72     1
22 4867 211944  6.45     1
43 5045 369989  9.65     1
44 4469 472476  8.20     0
48 4993 442782  7.61     1
```

Figure 8 Outliers by Using Diagonal Element of Hat Matrix

Here shows that we have 8 outliers in out dataset, which are observation 3, 5, 16, 21, 22, 43, 44 and 48.

- **Part c**

```
> X_new_GR = c(300000,7.2,0)
> h_new_GR = t(X_new_GR) %*% solve(t(X_GR) %*% X_GR) %*% X_new_GR
> plot(X1_GR,X2_GR,col = "blue",pch=16)
> points(X_new_GR[1],X_new_GR[2],col ="red",pch=4,lwd=5)
```

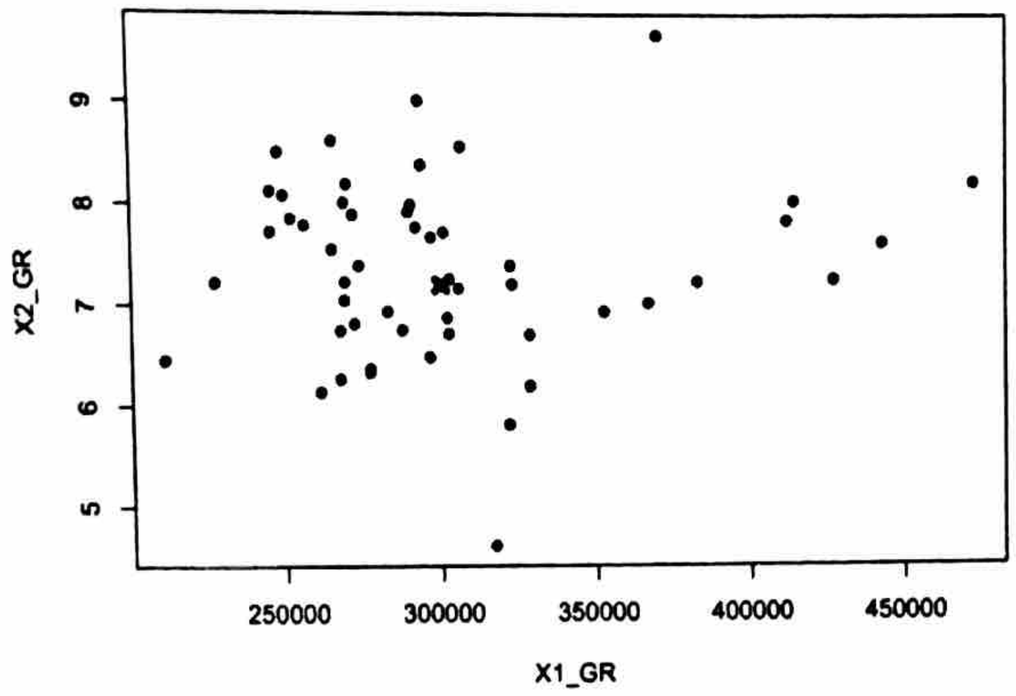


Figure 9 A Scatter Plot of X2 Against X1

From the plot above, we can determine visually that this prediction does not involves an extrapolation beyond the range of the data.

```
> h_new_GR
           [,1]
[1,] 0.02091887
```

Figure 10 Leverage of New Prediction

Since the leverage of new prediction is 0.02091887, which is not within the range of leverage values h_{ii} (0.02189441, 0.2868586) but is also not much larger the leverage values for the cases in data set, there does not exist an extrapolation, which agrees with the result of the first method.

- **Part d**

Cases Number	EFFITS	Cook's Distance	DFBETAS
16	-0.553990262	0.07689508	-0.2476886725, -0.0597817143 3.248149e-01, -0.4521004239
22	0.055085831	0.0007746088	0.0304231943, -0.0252871156, -1.870175e-02, 0.0446455176
43	0.561651861	0.07921931	-0.3577973374, 0.1338420534, 3.261797e-01, 0.3566275889
48	-0.146841460	0.005498867	0.0449857976, -0.0938422885, 9.013941e-03, -0.1022153800
10	0.458632975	0.04935012	0.3640749489, -0.1044031737, -3.141587e-01, -0.0633462565
32	-0.651077059	0.09975974	0.4095414971, 0.0913417314, -5.708322e-01, 0.1652061576
38	0.385517659	0.03463803	-0.0996147941,-0.0827384689, 2.083647e-01, -0.1270086765
40	0.396720295	0.03649915	0.0737987619, -0.2120590640, 9.325265e-02, -0.1110471738

```
> dffits_GR = ti_GR * sqrt(lev_GR / (1 - lev_GR))
> frame_GR[abs(dffits_GR) > (2 * sqrt(4/52)),]
   Y_GR  X1_GR X2_GR X3_GR
32 3998 293225  9.01     0
43 5045 369989  9.65     1
> dffits_GR[43]
       43
0.5616518
> dffits_GR[32]
       32
-0.651077
```

Figure 11 Result of DFFITS

* From DFFITS, we can see that the DFFITS values that exceeds our guideline for a large size data set is for case 43 and case 32, whose DFFITS are 0.5616518 and -0.651077, respectively. The absolute values of these cases are somewhat larger than our guideline of 0.5547002. However, the values are close enough to 0.5547002 that these two cases may not be influential enough to require remedial action.

```
> cook_GR = cooks.distance(fit_GR)
> max(cook_GR)
[1] 0.09975974
```

Figure 12 Results of Cook's Distance

* The largest value of Cook's Distance is about 0.09975974, belonging to case 32, and the second largest value is about 0.07921931, belonging to case 43. I now refer to the corresponding F distribution, namely, F (p, n-p) = F(4, 48). I find that 0.09975 is the 1.8th percentile of this distribution. Since 0.01798 is much less than 0.2, we can see that there does not exist any outlier that can have an influence on all fitted values.

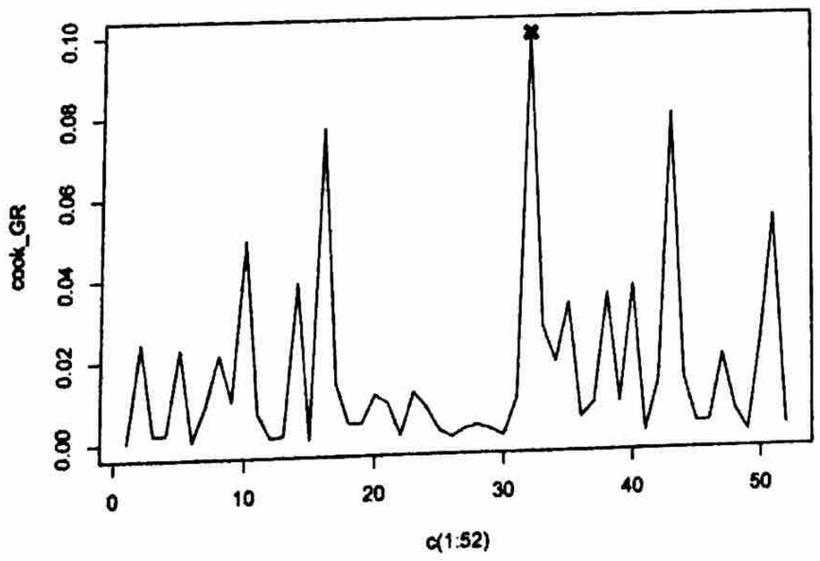
```
> dfbetas_GR = dfbetas(fit_GR)
```

* Our guideline here for DFBETAS is 14.42221, since there does not exist any outlier whose values of DFBETAS are greater than 14.42221, we can conclude that the outliers do not have any influence on the regression coefficients.

* In conclusion, our outliers obtained actually do not have any influence on all fitted values and regression coefficients. Even though the DFFITS values for case 43 and case 32 are greater than our guideline, they do not exceed it by very much so that these two cases may not be so influential as to require remedial action.

- **Part f**

```
> plot(c(1:52),cook_GR,type = 'l')
```




```
> frame_GR[pf(cook_GR, 4, 48, lower.tail = TRUE, log.p = FALSE) > 0.2,]
[1] Y_GR  X1_GR X2_GR X3_GR
<0 rows> (or 0-length row.names)
```

This plot identifies the most influential case as case 32. But I find that 0.09975, the Cook's Distance value

of case 32, is the 1.8th percentile of this distribution. Since 0.01798 is much less than 0.2, we can see that there does not exist any outlier that can have an influence on all fitted values.

• Question 3. Chapter 10. Page 416. Problem 10.12

• Part a

```
> fit_CP <- lm(Y_CP~X1_CP+X2_CP+X3_CP+X4_CP)
> hat_CP <- lm.influence(fit_CP)$hat
>
> res_CP <- residuals(fit_CP)
> ti_CP <- res_CP * sqrt((81-5-1) / (98.231 * (1-hat_CP) - res_CP^2))
> ti_CP
          1           2           3           4           5           6           7
-0.939935919 -1.392592890 -0.577010169 -0.119075270  0.278240526 -3.072097391 -0.481686005
          8           9          10          11          12          13          14
 0.231248210  1.875607547  0.093764592  0.020986487 -0.301081674  0.639838123 -0.353609629
         15          16          17          18          19          20          21
-0.188875057 -0.744639428  0.090403548 -1.615500296 -1.107546259 -0.562848673 -0.327338652
         22          23          24          25          26          27          28
 0.258288681 -0.083334978  0.208326124 -0.772289365 -1.978293666  0.416596628 -0.511756498
         29          30          31          32          33          34          35
-0.956026829 -0.176356204 -1.005351102 -0.154726948 -0.924213792 -0.081555166  0.192082646
         36          37          38          39          40          41          42
 0.697626905  0.984130563 -1.993206071 -0.164751874 -1.018157594 -0.011515870  2.323012605
         43          44          45          46          47          48          49
-1.486141918  0.837829080  0.354116376  0.184403680  0.742465844  1.473703722  0.580425311
         50          51          52          53          54          55          56
 0.444802696  0.187431071 -0.028642401  1.124078667  0.217241212 -0.958383659  0.950754405
         57          58          59          60          61          62          63
-0.234535542  0.921075391 -0.307614644  0.181886679  0.967197690  2.784056545  2.279873507
         64          65          66          67          68          69          70
 1.748671236  1.376414028 -0.435478762 -0.677251216  1.841757370  0.871777709  0.008824738
         71          72          73          74          75          76          77
 1.602835333 -0.415805815 -0.456318050 -0.094082356  1.101173471 -0.232388769 -0.562615261
         78          79          80          81
 0.830177070 -0.492368590 -1.923157010 -0.809546409
```

We shall use the Bonferroni simultaneous test procedure with a family significance level $\alpha = 0.01$. We therefore require: $t(1-\alpha/2n, n-p-1) = t(0.9999, 75) = 3.46$

```
> frame_CP <- data.frame(Y_CP,X1_CP,X2_CP,X3_CP,X4_CP)
> frame_CP[abs(ti_CP) > 3.46,]
[1] Y_CP  X1_CP X2_CP X3_CP X4_CP
<0 rows> (or 0-length row.names)
```

Since the there is no point whose $|t^*| > t(1-\alpha/2n, n-p-1) = t(0.9999, 75) = 3.46$, then there does not exist any outlier here.

• Part b

```
> lev_CP = hat(model.matrix(fit_CP))
> frame_CP[lev_CP > (2 * mean(lev_CP)),]
    Y_CP X1_CP X2_CP X3_CP  X4_CP
3  10.50    16  3.00  0.00  39998
8  16.50     1  6.62  0.60 248172
53 17.00     1  5.99  0.57 220000
61 16.50     1  4.99  0.73 210000
65 19.25    13 12.70  0.04 484290
```

Here shows that we have 5 outliers in out dataset, which are cases 3, 8, 53, 61 and 65.

• Part c

```
> X_new_CP = c(10,12.00,0.05,350000)
> h_new_CP = t(X_new_CP) %*% solve(t(X_CP) %*% X_CP) %*% X_new_CP
> h_new_CP
            [,1]
[1,] 0.05203178
```

Since the leverage of new prediction is 0.05203178, which is within the range of leverage values h_{ii}

(0.02419885, 0.3036714), there does not exist an extrapolation.

- **Part d**

Case Number	DFFITS	Cook's Distance	DFBETAS
61	0.638719423	0.08166217	-0.0554152848 0.0242485298 -0.0076084289 0.5457127012 0.0038197889
8	0.116413401	0.002744609	-0.0142102315 -0.0071978923 0.0030142503 0.0955192720 0.0125990640
3	-0.284279865	0.0163062	-0.2317857240 -0.1553283208 0.2364136428 0.1007804144 -0.0114939468
53	0.525225345	0.05498189	-0.0196280259 -0.0239835334 -0.0243404423 0.4179638413 0.0489678626
6	-0.873546780	0.1373665	0.1951154628 -0.5648515358 -0.1767222512 -0.6171913680 0.4481729286
62	0.690330294	0.08753589	0.2758146886 -0.3334961755 -0.2594703696 0.0627288004 0.4050781439

```
> dffits_CP = ti_CP * sqrt(lev_CP / (1 - lev_CP))
> frame_CP[abs(dffits_CP) > (2 * sqrt(5/81)),]
    Y_CP X1_CP X2_CP X3_CP  X4_CP
6  10.50    15  9.45  0.24 101385
9  17.50     1  6.20  0.00 215000
26 12.50     1  5.00  0.33 120000
38 13.00    14  8.53  0.03 315000
42 15.50    15  8.32  0.00  73521
43 12.00     1  4.00  0.00  50000
53 17.00     1  5.99  0.57 220000
61 16.50     1  4.99  0.73 210000
62 19.25     0  7.33  0.22 240000
63 17.75    18 12.11  0.00 281552
64 18.75    16 12.86  0.00 421000
65 19.25    13 12.70  0.04 484290
80 15.25    11 11.27  0.03 434746
```

* From DFFITS, we can see that the DFFITS values that exceeds our guideline for a large size data set is for case 6, 9, 26, 38, 42, 43, 53, 61, 62, 63, 64, 65 and 80, whose DFFITS are -0.873546780, 0.628366974, -0.559302358, -0.586957331, 0.540649449, -0.518207407, 0.525225345, 0.638719423, 0.690330294, 0.517812667, 0.560330071, 0.529948778 and -0.666372181, respectively. The absolute values of these cases are somewhat larger than our guideline of 0.496904. However, the values of case 9, 26, 38, 42, 43, 53, 61, 63, 64, 65 and 80 are close enough to 0.496904 that these two cases may not be influential enough to require remedial action. But we can see that the absolute values of case 6 and 62 are much larger than our guideline, thus, we can say that these two cases have an influence on single fitted value.

```
> cook_CP = cooks.distance(fit_CP)
> cook_CP
           1            2            3            4            5            6            7
1.172148e-02 3.082585e-02 1.630620e-02 1.170536e-04 5.024783e-04 1.373665e-01 2.122748e-03
           8            9           10           11           12           13           14
2.744089e-03 7.643086e-02 4.991388e-05 6.245098e-06 8.024768e-04 2.874841e-03 1.562689e-03
          15           16           17           18           19           20           21
3.972186e-04 9.538360e-03 5.881950e-05 3.447867e-02 1.943961e-02 1.714589e-03 1.080543e-03
          22           23           24           25           26           27           28
6.506810e-04 6.621591e-05 3.387530e-04 7.710837e-03 6.025418e-02 1.553887e-03 1.906941e-03
          29           30           31           32           33           34           35
6.454024e-03 2.620572e-04 7.694798e-03 1.767081e-04 7.250896e-03 7.138187e-05 3.316295e-04
          36           37           38           39           40           41           42
2.708006e-03 1.277470e-02 6.631823e-02 1.819756e-04 1.386215e-02 1.742666e-06 5.526375e-02
          43           44           45           46           47           48           49
5.286726e-02 7.617833e-03 1.394657e-03 8.658181e-05 8.668267e-03 3.877847e-02 2.535423e-03
          50           51           52           53           54           55           56
2.236103e-03 4.768873e-04 7.098224e-06 5.498189e-02 1.190856e-03 5.755538e-03 7.572763e-03
          57           58           59           60           61           62           63
5.091541e-04 5.499839e-03 6.552548e-04 3.160344e-04 8.166217e-02 8.753588e-02 5.882155e-02
          64           65           66           67           68           69           70
6.116864e-02 5.551595e-02 2.233224e-03 4.066889e-03 3.328982e-02 9.767999e-05 6.608475e-07
          71           72           73           74           75           76           77
2.120843e-02 1.737285e-03 1.049483e-03 4.448182e-05 1.856576e-02 4.036388e-04 3.085828e-03
          78           79           80           81
1.106132e-02 2.108266e-03 8.570547e-02 4.580458e-03
```

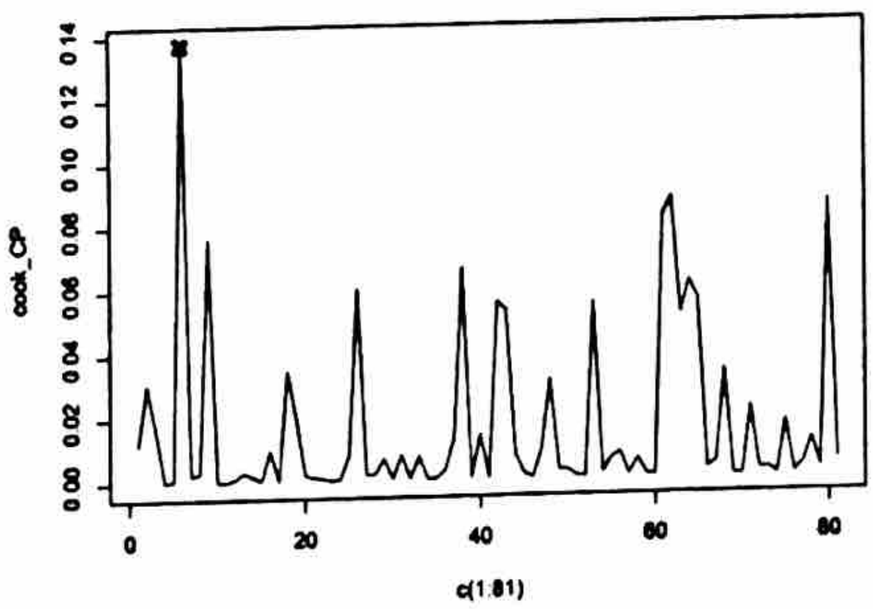
* The largest value of Cook's Distance is about 0.1373665, belonging to case 6, and the second largest value is about 0.08753589, belonging to case 62. I now refer to the corresponding F distribution, namely, F (p, n-p) = F(5, 76). I find that 0.1373665 is the 1.6874^{th} percentile of this distribution. Since 0.0168742296839 is much less than 0.2, we can see that there does not exist any outlier that can have an influence on all fitted values.

* Our guideline here for DFBETAS is 18, since there does not exist any outlier whose values of DFBETAS are greater than 18, we can conclude that the outliers do not have any influence on the regression coefficients.

* In conclusion, our outliers obtained actually do not have any influence on all fitted values and regression coefficients. However, for case 6 and 62, their absolute values of DFFITS are much larger than the guideline, these two cases have influence on their single fitted values.

- **Part f**

```
> plot(c(1:81),cook_CP,type = 'l')
```

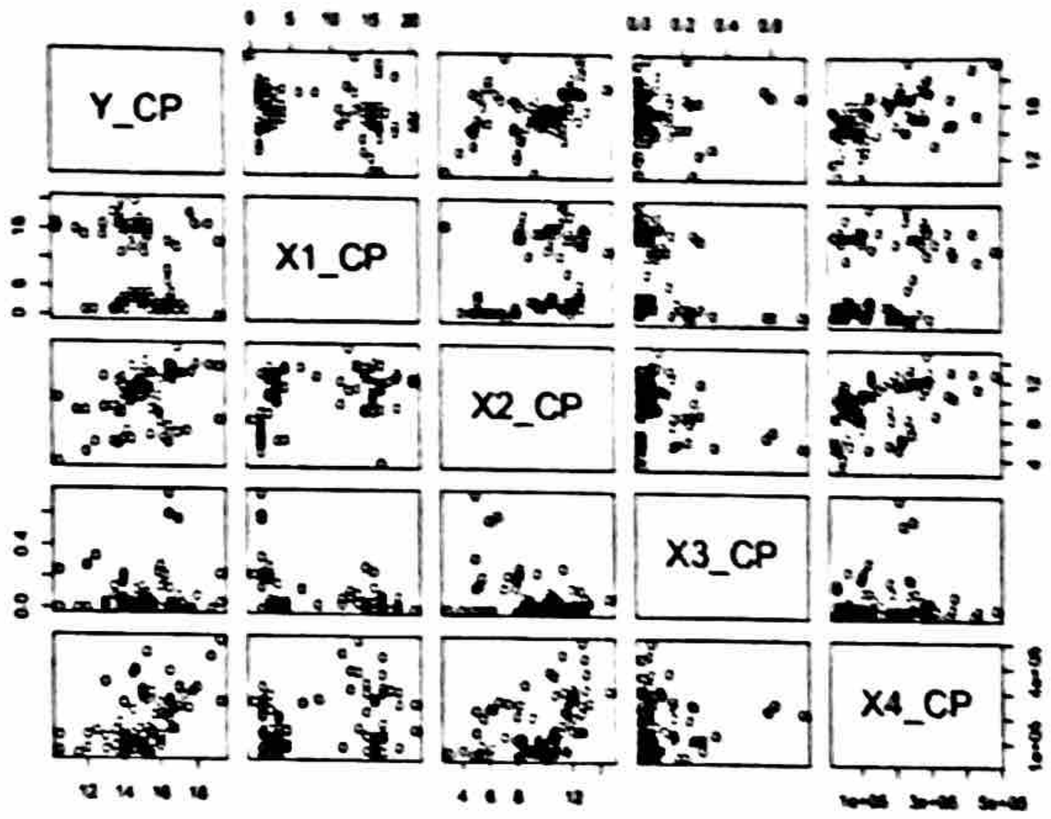



```
> frame_CP[pf(cook_CP, 5, 76, lower.tail = TRUE, log.p = FALSE) > 0.2,]
[1] Y_CP  X1_CP X2_CP X3_CP X4_CP
<0 rows> (or 0-length row.names)
```

This plot identifies the most influential case as case 6. But I find that 0.1373665, the Cook's Distance value of case 6, is the 1.6874^{th} percentile of this distribution. Since 0.016874 is much less than 0.2, we can see that there does not exist any outlier that can have an influence on all fitted values.

- **Question 4. Chapter 10. Page 417. Problem 10.18**
- **Part a**

```
> plot(frame_CP)
```




```
> cor(frame_CP)
             Y_CP       X1_CP      X2_CP       X3_CP      X4_CP
Y_CP    1.00000000 -0.2502846  0.4137872  0.06652647 0.53526237
X1_CP -0.25028456  1.0000000  0.3888264 -0.25266347 0.28858350
X2_CP  0.41378716  0.3888264  1.0000000 -0.37976174 0.44869713
X3_CP  0.06652647 -0.2526635 -0.3797617  1.00000000 0.08061073
X4_CP  0.53526237  0.2885835  0.4486971  0.08061073 1.00000000
```

From the scatter plot matrix, we can see that the relations between the age (X_1) and Vacancy rate (X_3), between X1 and X2, between X1 and X4, between X3 and X4, between X2 and X3 are week, but there is a high correlation between X_2 and X_4.
We also can confirm our finding from the correlation matrix.

- **Part b**

```
> vif(fit_CP)
   X1_CP    X2_CP    X3_CP    X4_CP
1.240348 1.648225 1.323552 1.412722
```

Since there is no VIF larger than 10, we can conclude that there does not exist a serious multicollinearity problem.