Jane Downer
Math 484-01

Homework 3

Problem 1

5.3 Perform a thorough analysis of the Ski Sales data in Table 5.11 using the ideas presented in Section 5.6.

4 Quarters $\longrightarrow$ $4-1=3$ indicator variables

$$\begin{array}{cc} \begin{array}{ccc} Z_1 & Z_2 & Z_3 \end{array} & \\ \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} & \begin{array}{c} Q_1 \\ Q_2 \\ Q_3 \\ Q_4 \end{array} \end{array}$$

Reduced model: $S_t = \beta_0 + \beta_1 \cdot PDI_t + \varepsilon_t$

Full model: $S_t = \beta_0 + \beta_1 \cdot PDI_t + \beta_2 \cdot Z_1 + \beta_3 \cdot Z_2 + \beta_4 \cdot Z_3 + \varepsilon_t$

$$\begin{cases} H_0: \beta_2 = \beta_3 = \beta_4 = 0 \\ H_a: \text{at least one not zero} \end{cases}$$

F-test

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \bigg/ \frac{SSE(F)}{df_F} = \frac{47.245 - 346.433}{3} \bigg/ \frac{47.245}{35} = 73.881$$

$$\left.\begin{aligned} SSE &= Y^T(I - H)Y \\ H &= X(X^TX)^{-1}X^T \end{aligned}\right] \text{calculated in Python}$$

$F_{(p-q,\, n-p-1)} = F_{(3,35)} = 2.874$

when $\alpha = 0.05$

$F^* > F_{critical}$

$\downarrow$

$\boxed{\text{We reject the null hypothesis and favor the full model}}$

Problem 2

5.4 Perform a thorough analysis of the Education Expenditures data in Tables 5.12, 5.13, and 5.14 using the ideas presented in Section 5.7.

1960 1970 1975

$T_1 = \begin{cases} 1 & \text{if time} = 1960 \\ 0 & \text{otherwise} \end{cases}$

$T_2 = \begin{cases} 1 & \text{if time} = 1970 \\ 0 & \text{otherwise} \end{cases}$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 T_1 + \beta_5 T_2 + \beta_6 T_1 \cdot X_1 + \beta_7 T_1 \cdot X_2 + \beta_8 T_1 \cdot X_3 + \beta_9 T_2 \cdot X_1 + \beta_{10} T_2 \cdot X_2 + \beta_{11} T_2 \cdot X_3 + \varepsilon$$

1960: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 + \beta_6 X_1 + \beta_7 X_2 + \beta_8 X_3 + \varepsilon$

$= (\beta_0 + \beta_4) + (\beta_1 + \beta_6) X_1 + (\beta_2 + \beta_7) X_2 + (\beta_3 + \beta_8) X_3 + \varepsilon$

1970: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_5 + \beta_9 X_1 + \beta_{10} X_2 + \beta_{11} X_3 + \varepsilon$

$= (\beta_0 + \beta_5) + (\beta_1 + \beta_9) X_1 + (\beta_2 + \beta_{10}) X_2 + (\beta_3 + \beta_{11}) X_3 + \varepsilon$

1975: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

$H_0: \beta_4 = \beta_5 = \beta_6 = \beta_7 = \beta_8 = \beta_9 = \beta_{10} = \beta_{11} = 0$

H_a: at least one of these ↑ is non-zero

$\left[\begin{array}{l} SSE = Y^T (I - H) Y \\ H = X(X^T X)^{-1} X^T \end{array}\right]$ calculated in Python

F-test

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \bigg/ \frac{SSE(F)}{df_F} = 9.422$$

$\left[\begin{array}{l} SSE(R) = 185764.329 \rightarrow n - 3 - 1 = 146 \\ SSE(F) = 120140.176 \rightarrow n - 11 - 1 = 138 \end{array}\right]$

$F_{(p-q,\, n-p-1)} = F_{(8,138)} = 2.006$

↑ when $\alpha = 0.05$

$F^* > F_{critical}$

$\boxed{\text{reject } H_0 \text{ and favor the full model}}$

Problem 3

5.7 Three types of fertilizer are to be tested to see which one yields more corn crop. Forty similar plots of land were available for testing purposes. The 40 plots are divided at random into four groups, 10 plots in each group. Fertilizer 1 was applied to each of the 10 corn plots in Group 1. Similarly, Fertilizers 2 and 3 were applied to the plots in Groups 2 and 3, respectively. The corn plants in Group 4 were not given any fertilizer; it will serve as the control group. Table 5.17 gives the corn yield y_{ij} for each of the 40 plots.

(a) Create three indicator variables F_1, F_2, F_3, one for each of the three fertilizer groups.

(b) Fit the model $y_{ij} = \mu_0 + \mu_1 F_{i1} + \mu_2 F_{i2} + \mu_3 F_{i3} + \varepsilon_{ij}$.

(c) Test the hypothesis that, on the average, none of the three types of fertilizer has an effect on corn crops. Specify the hypothesis to be tested, the test used, and your conclusions at the 5% significance level.

(e) Which of the three fertilizers has the greatest effects on corn yield?

(a) $F_1 = \begin{cases} 1 & \text{if group 1} \\ 0 & \text{otherwise} \end{cases}$ $\quad F_2 = \begin{cases} 1 & \text{if group 2} \\ 0 & \text{otherwise} \end{cases}$ $\quad F_3 = \begin{cases} 1 & \text{if group 3} \\ 0 & \text{otherwise} \end{cases}$

(b)
$$\hat{\beta} = (X^T X)^{-1} X^T Y$$
$$SSE = Y^T (I - H) Y$$
$$H = X(X^T X)^{-1} X^T$$

calculated in Python

$$\hat{\beta} = \begin{bmatrix} 29.8 \\ 6.8 \\ 0.1 \\ 5.1 \end{bmatrix}$$

$$\boxed{\hat{Y} = 29.8 + 6.8\mu_1 + 0.1\mu_2 + 5.1\mu_3}$$

(c)
$SSE_F = 845.8$
$SSE_R = 1208.4$
$df_F = 36$
$df_R = 39$

$$F^* = \frac{SSE(R) - SSE(F)}{df_R - df_F} \Big/ \frac{SSE(F)}{df_F} = 5.144$$

$$F_{(p-q,\, n-p-1)} = F_{(3, 36)} = 2.866$$

when $\alpha = 0.05$

$F^* > F_{critical}$

reject H_0 and favor the full model

(e) Fertilizer 1 has the highest coefficient, and therefore the highest effect.

Problem 4

4.6 The following graphs are used to verify some of the assumptions of the ordinary least squares regression of Y on $X_1, X_2, \cdots, X_p$:

1. The scatter plot of Y versus each predictor X_j.
2. The scatter plot matrix of the variables $X_1, X_2, \cdots, X_p$.
3. The normal probability plot of the internally standardized residuals.
4. The residuals versus fitted values.
5. The potential-residual plot.
6. Index plot of Cook's distance.
7. Index plot of Hadi's influence measure.

For each of these graphs:

(a) What assumption can be verified by the graph?

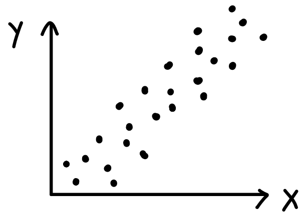
(b) Draw an example of the graph where the assumption does not seem to be violated.

(c) Draw an example of the graph which indicates the violation of the assumption.

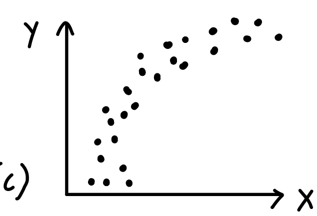
1. Y vs. X scatter plot

(a) linearity

(b)



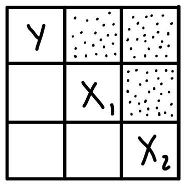
(c)



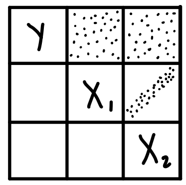
2. $X_1, \ldots, X_p$ scatterplot matrix

(a) linear independence

(b)



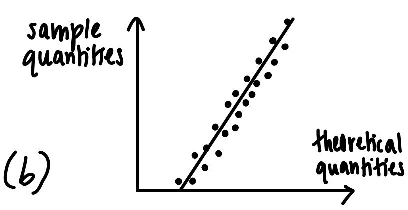
(c)



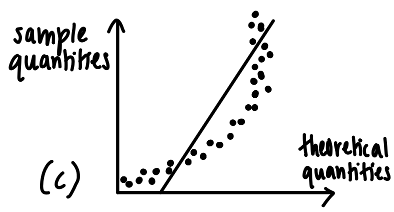
3. normal probability plot of internally-standardized residuals

(a) normal distribution of error terms

(b)



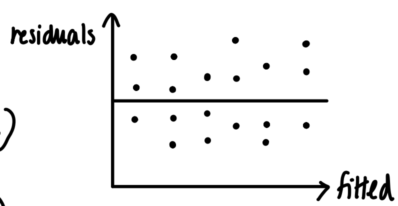
(c)



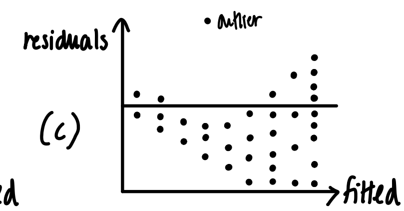
4. residuals vs. fitted values

(a) linearity
equal variance
(can also help detect outliers)

(b)



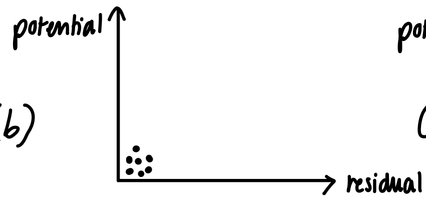
(c)



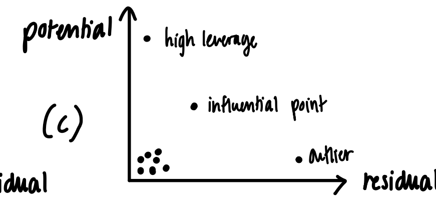
5. potential-residual plot

(a) can detect { outliers, high-leverage points, infuential observations }

(b)



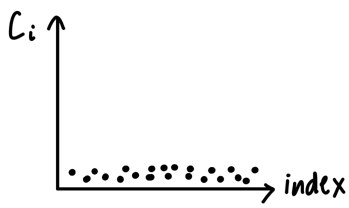
(c)



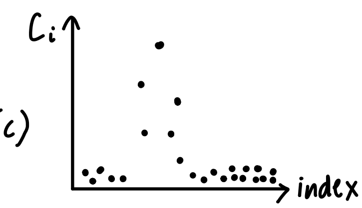
6. index plot - Cook's Distance

(a) can detect influence in the data

(b)



(c)

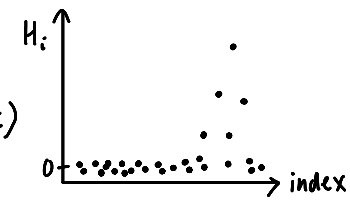


7. index plot - Hadi's influence

(a) can detect influence in the data

(b)

(c)



Problem 5

4.8 Consider again the Examination Data used in Exercise 3.3 and given in Table 3.10:

(a) For each of the three models, draw the P-R plot. Identify all unusual observations (by number) and classify as outlier, high-leverage point, and/or influential observation.

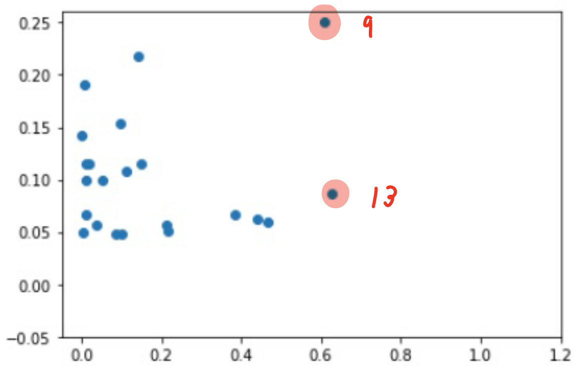
$$d_i = \frac{e_i}{\sqrt{SSE}}$$

$$\text{potential} = \frac{h_{ii}}{1-h_{ii}}$$

$$\text{residual} = \left(\frac{p+1}{1-h_{ii}}\right)\left(\frac{d_i^2}{(1-d_i)^2}\right)$$

No single observation has a value of $|r_i^*| > 3$, but highlighted in red below are the observations with values of $|r_i^*| > 2$.

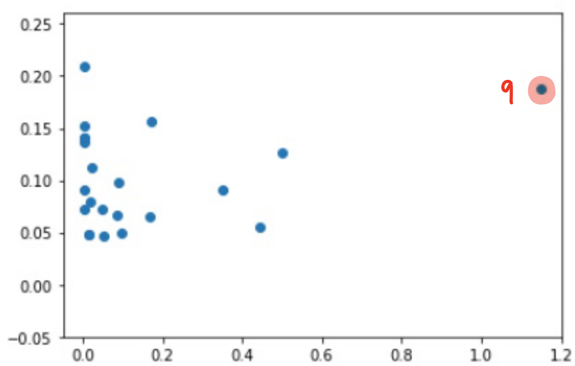
Model 1



$|r^*|$

	Potential	Residual	r_
1	0.062699	0.440821	1.682428
2	0.107420	0.112028	0.843515
3	0.048218	0.101113	0.800462
4	0.100110	0.052788	0.578140
5	0.066098	0.011404	0.268438
6	0.100110	0.009389	0.243574
7	0.048218	0.086418	0.739833
8	0.142857	0.000637	0.063426
9	0.250000	0.610012	2.015275
10	0.218027	0.142723	0.954699
11	0.114827	0.012415	0.280110
12	0.057082	0.036958	0.483450
13	0.086957	0.626116	2.016679
14	0.153403	0.097401	0.786696
15	0.066098	0.384801	1.570537
16	0.114827	0.018035	0.337657
17	0.059322	0.467592	1.733227
18	0.114827	0.150157	0.977654
19	0.049318	0.004634	0.171087
20	0.051525	0.217850	1.177337
21	0.190476	0.005980	0.194383
22	0.057082	0.211667	1.160593

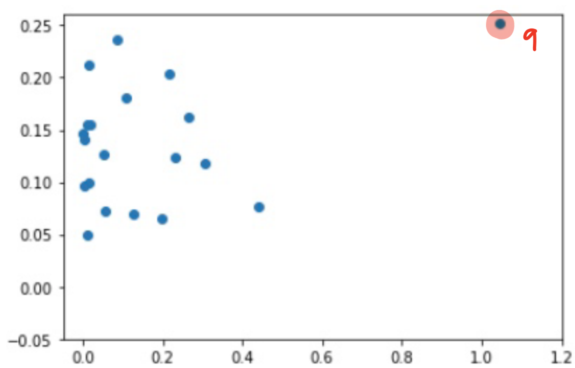
Model 2



$|r^*|$

	Potential	Residual	r_
1	0.097695	0.087439	0.744673
2	0.071811	0.002656	0.129513
3	0.054852	0.442875	1.685524
4	0.140251	0.002290	0.120284
5	0.072961	0.047983	0.551030
6	0.090513	0.001545	0.098771
7	0.126126	0.499537	1.801082
8	0.152074	0.001639	0.101745
9	0.187648	1.147900	2.808099
10	0.136364	0.001610	0.100828
11	0.066098	0.085543	0.736238
12	0.078749	0.018817	0.344871
13	0.048218	0.014931	0.307157
14	0.156069	0.170586	1.043553
15	0.064963	0.167864	1.032947
16	0.140251	0.002290	0.120284
17	0.112347	0.022031	0.373230
18	0.090513	0.352175	1.503413
19	0.048218	0.015753	0.315504
20	0.047120	0.051501	0.570804
21	0.209190	0.003601	0.150819
22	0.049318	0.096104	0.780332

Model 3



$|r^*|$

	Potential	Residual	r_
1	0.123596	0.232790	1.220263
2	0.126126	0.053373	0.581448
3	0.076426	0.441358	1.684982
4	0.146789	0.000087	0.023385
5	0.072961	0.056092	0.595875
6	0.100110	0.014216	0.299745
7	0.447178	0.224346	1.206774
8	0.154734	0.011657	0.271433
9	0.251564	1.045974	2.692027
10	0.236094	0.085944	0.739335
11	0.154734	0.016484	0.322823
12	0.096491	0.003351	0.145487
13	0.203369	0.216184	1.177936
14	0.161440	0.263085	1.299997
15	0.481481	0.001564	0.099411
16	0.140251	0.004016	0.159285
17	0.180638	0.109663	0.835376
18	0.117318	0.304112	1.397102
19	0.049318	0.010693	0.259921
20	0.069519	0.127936	0.901161
21	0.212121	0.015377	0.311817
22	0.064963	0.196477	1.118128