Homework2

2021年10月3日 1:29

MATH34...
HW2

Assignment #2 — Due Thu, 14 Oct.

*This homework covers Chapter 3 (*Problem* 1-6) and Chapter 4 (*Problem* 7-11). Submit your homework on Canvas or send it to our TA, Mr. LYU Zhongyuan (zlyuab@connect.ust.hk).

*No late homework will be accepted for credit.

*Append the R codes you used to your submission. *If the problem does not need R or is not explicitly stated to complete in R, then you should just do it by hand with a calculator.*

*In case of rounding error, keep 3 figures after the decimal point.

Problem 1 Using the supervisor data, verify that the coefficient of X_1 in the fitted equation $\hat{Y} = 15.3276 + 0.7803X_1 - 0.0502X_2$ in (3.12), *equation on lecture slides of Chapter 3*, can be obtained from a series of simple regression equations, as outlined in Section 3.2 for the coefficient of X_2.

Problem 2 (*use* R) The following Table 3.10 shows the scores in the final examination F and the scores in two preliminary examinations P_1 and P_2 for 22 students in a statistics course. The data is given in the file *Examination_Data.txt*.

Table 3.10 Examination Data: Scores in Final (F), First Preliminary (P_1), and Second Preliminary (P_2) Examinations

Row	F	P_1	P_2	Row	F	P_1	P_2
1	68	78	73	12	75	79	75
2	75	74	76	13	81	89	84
3	85	82	79	14	91	93	97
4	94	90	96	15	80	87	77
5	86	87	90	16	94	91	96
6	90	90	92	17	94	86	94
7	86	83	95	18	97	91	92
8	68	72	69	19	79	81	82
9	55	68	67	20	84	80	83
10	69	69	70	21	65	70	66
11	91	91	89	22	83	79	81

(a) Fit each of the following models of the data:

$$\text{Mode 1}: F = \beta_0 + \beta_1 P_1 + \varepsilon$$
$$\text{Model 2}: F = \beta_0 + \beta_2 P_2 + \varepsilon$$
$$\text{Model 3}: F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \varepsilon$$

(b) Test whether $\beta_0 = 0$ in each of the three models.

(c) Which variable individually, P_1 or P_2, is a better predictor of F? Why?

(d) Which of the three models would you use to predict the final examination scores for a student who scored 78 and 85 on the first and second preliminary examinations, respectively? What is your prediction interval in this case?

Problem 3 Table 3.11 shows the regression output, with some numbers erased, when a simple regression model relating a response variable Y to a predictor variable X_1 is fitted based on 20 observations. Complete the 13 missing numbers, then compute the *sample variances* $\mathrm{Var}(Y)$ and $\mathrm{Var}(X_1)$.

Table 3.11 Regression Output When Y is Regressed on X_1 for 20 Observations

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	1848.76	–	–	–
Residuals	–	–	–	
Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	−23.4325	12.74	–	0.0824
X_1	–	0.1528	8.32	< 0.0001
$n =$ –	$R^2 =$ –	$R_a^2 =$ –	$\hat{\sigma} =$ –	df = –

Problem 4 (*use* R) Using the Supervisor Performance data *Supervisor.txt*, test the hypothesis $H_0 : \beta_1 = \beta_3 = 0.5$ in each of the following models:

(a) $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$

(b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

Problem 5 Table 3.14 shows the regression output of a multiple regression model relating the beginning salaries in dollars of employees in a given company to the following predictor variables:

Gender : An indicator variable 1=man and 0=woman
Education : Years of schooling at the time of hire
Experience : Number of months of pervious work experience
Months : Number of months with the compnay

Table 3.14 Regression Output When Salary is Related to Four Predictor Variables

ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	23665352	4	5916338	22.98
Residuals	22657938	88	257477	
Coefficients Table				
Variable	Coefficient	s.e.	t-Test	p-value
Constant	3526.4	327.7	10.76	0.000
Gender	722.5	117.8	6.13	0.000
Education	90.02	24.69	3.65	0.000
Experience	1.2690	0.5877	2.16	0.034
Months	23.406	5.201	4.50	0.000
$n = 93$	$R^2 = 0.515$	$R_a^2 = 0.489$	$\hat{\sigma} = 507.4$	df = 88

In (a)-(b) below, specify the null and alternative hypothesis, the test used, and your conclusion using a 5% level of significance. For (c)-(e), only point prediction is desired.

(a) Conduct the F-test for the overall fit of the regression

(b) Is there a *positive* linear relationship between salary and experience, after accounting for the effect of the variables Gender, Education and Months?

(c) What salary would you forecast for a man with 12 years of education, 10 months of experience, and 15 months with the company?

(d) What salary would you forecast, on average, for men with 12 years of education, 10 months of experience, and 15 months with the company?

(e) What salary would you forecast, on average, for women with 12 years of education, 10 months of experience, and 15 months with the company?

Problem 6 Consider the regression model that generated the output in Table 3.14 to be a full model. Now consider the reduced model in which Salary is regressed on only Education. The ANOVA table obtained when fitting the model is shown in Table 3.15. Conduct a single test to compare the full and reduced models. What conclusion can be drawn from the result of the test? (Use $\alpha = 0.05$.)

Table 3.15 ANOVA Table When the Beginning Salary is Regressed on Education

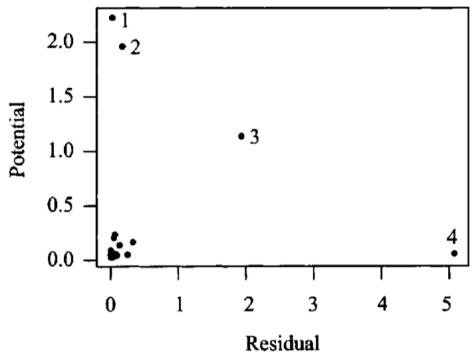
ANOVA Table				
Source	Sum of Squares	df	Mean Square	F-Test
Regression	7862535	1	7862535	18.60
Residuals	38460756	91	422646	

Problem 7 (*use* R) Consider the computer repair problem discussed in Chapter 2. In a second sampling period, 10 more observations on the variables Minutes and Units were obtained. Since all observations were collected by the same method from a fixed environment, all 24 observations were pooled to form one dataset. The data appear in Table 4.6 and stored in the file *Computer_Repair.txt*.

(a) Fit a linear regression model relating Minutes to Units.

(b) Check each of the standard regression assumptions and indicate which assumptions seems to be violated.

Problem 8 In an attempt to find unusual points in a regression data set, a data analyst examines the *potential-Residual* plot. Classify each of the unusual points on this plot according to the type.



Problem 9 Name one or more graphs that can be used to validate each of the following assumptions. For each graph, sketch an example where the corresponding assumption is valid and an example where the assumption is clearly invalid.

(a) There is a linear relationship between the response and predictor variables.

(b) The observations are independent of each other

(c) The error terms have constant variance

(d) The error terms are uncorrelated

(e) The error terms are normally distributed

(f) The observations are equally influential on least squares results

Problem 10 Consider again the Examination data used in *Problem 2* and given in Table 3.10.

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

(b) What model would you use to predict the final score F?

Problem 11 Identify unusual observations for the data set in Table 4.7.

Table 4.7

Row	Y	X	Row	Y	X
1	8.11	0	7	9.60	19
2	11.00	5	8	10.30	20
3	8.20	15	9	11.30	21
4	8.30	16	10	11.40	22
5	9.40	17	11	12.20	23
6	9.30	18	12	12.90	24

Problem 1:

Step 1: Fit the simple regression model that relates Y to X_1.

$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} y_i = 64.6333$

$\bar{X_1} = \frac{1}{n}\sum_{i=1}^{n} x_{1i} = 66.6$

$\sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x}_1) = 3879.6$

$\sum_{i=1}^{n} (x_{1i} - \bar{x}_1)^2 = 5141.2$

$\hat{\beta}_{1(X_1)} = \frac{\sum_{i=1}^{n} (y_i - \bar{y})(x_{1i} - \bar{x}_1)}{} = \frac{3879.6}{}$

$$\hat{\beta}_{1(Y,X_1)} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_{1i}-\bar{x}_1)}{\sum_{i=1}^{n}(x_{1i}-\bar{x}_1)^2} = \frac{3879.6}{5141.2} = 0.754682$$

$$\hat{\beta}_{0(Y,X_1)} = \bar{y} - \hat{\beta}_1\bar{x} = 64.6333 - 0.754682(66.6) = 14.376394$$

⇒ The fitted regression equation of Y to X_1 is:

$$\hat{Y} = 14.3763 + 0.754610 X_1$$

Step 2: Fit a simple regression model that relates X_2 to X_1: (treating X_2 temporarily as response variable)

$$\bar{X}_2 = \frac{1}{n}\sum_{i=1}^{n} X_{2i} = 53.1333$$

$$\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(X_{1i}-\bar{X}_1) = 2637.6$$

$$\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)^2 = 5141.2$$

$$\hat{\beta}_{1(X_1,X_2)} = \frac{2637.6}{5141.2} = 0.513031$$

$$\hat{\beta}_{0(X_1,X_2)} = \bar{X}_2 - \hat{\beta}_1\bar{X}_1 = 53.1333 - 0.513031(66.6)$$

$$= 18.96547$$

∴ The fitted regression equation of X_2 relates to X_1 is:

$$\hat{X}_2 = 18.9654 + 0.513032 X_1$$

Denote $e_{(Y,X_1)}$, $e_{(X_2,X_1)}$ as the residuals from the regression model that relates Y to X_1, X_2 to X_1 respectively.

Step 3: Fit a simple regression model the relates $e_{(Y,X_1)}$ and

$$e_{(Y,X_1)i} = Y_i - (\hat{\beta}_{0(Y,X_1)} + \hat{\beta}_{1(Y,X_1)} X_{1i})$$

$$e_{(Y,X_1)i} = Y_i - (\hat{\beta}_{0(Y,X_1)} + \hat{\beta}_{1(Y,X_1)} X_{1i})$$

$$\sum_{i=1}^{n} e_{(Y,X_1)i} = -3.33955 \times 10^{-13} \approx 0$$

$$\overline{e_{(Y,X_1)}} = \frac{1}{n}\sum_{i=1}^{n} e_{(Y,X_1)i} = -1.11318 \times 10^{-14} \approx 0$$

$$e_{(X_2,X_1)i} = X_{2i} - (\hat{\beta}_{0(X_2,X_1)} + \hat{\beta}_{1(X_2,X_1)} X_{1i})$$

$$e_{(X_2,X_1)i} = X_{2i} - (18.9634 + 0.513032 X_{1i})$$

$$\sum_{i=1}^{n} e_{(X_2,X_1)i} = 2.84217 \times 10^{-14} \approx 0$$

$$\overline{e_{(X_2,X_1)}} = \frac{1}{n}\sum_{i=1}^{n} e_{(X_2,X_1)i} = 9.4739 \times 10^{-16} \approx 0$$

$$\sum_{i=1}^{n} \left(e_{(X_2,X_1)i} - \overline{e_{(X_2,X_1)}}\right)\left(e_{(Y,X_1)i} - \overline{e_{(Y,X_1)}}\right) = -149.8921$$

$$\sum_{i=1}^{n} \left(e_{(X_2,X_1)i} - \overline{e_{(X_2,X_1)}}\right)^2 = 298.2932$$

$$\hat{\beta}_{1(e(Y,X_1)\ vs\ e(X_2,X_1))} = \frac{-149.8921}{298.2932} = -0.502598$$

$$\hat{\beta}_{0(e(Y,X_1)\ vs\ e(X_2,X_1))} = \overline{e_{(Y,X_1)}} - \hat{\beta}_1 \overline{e_{(X_2,X_1)}}$$

$$= -1.10843 \times 10^{-14} \approx 0$$

∴ The fitted regression line between $e_{(Y,X_1)}$ and $e_{(X_2,X_1)}$ is:

$$\hat{e}_{(Y,X_1)} = 0 - 0.0502\, e_{(X_2,X_1)}$$

⇒ -0.0502 is the coefficient of X_2 in multiple regression model of Y, X_1, and X_2.

Y, X_1, and X_2.

Now, apply the similar steps to X_1 again:

Step 1: fit the regression line that Y relates to X_2.

$$\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)(Y_i-\bar{Y}) = 1840.46667$$

$$\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2 = 4341.46667$$

$$\hat{\beta}_1(Y,X_2) = \frac{1840.46667}{4341.46667} = 0.4239274$$

$$\hat{\beta}_0(Y,X_2) = \bar{Y} - \hat{\beta}_1(Y,X_2)\bar{X}_2 = 42.10866$$

∴ The fitted regression equation of Y versus X_2 is:

$$\hat{Y} = 42.10866 + 0.4239274 X_2$$

Step 2: Treating X_1 as response variable, fit regression line of X_1 relates to X_2.

$$\sum_{i=1}^{n}(X_{1i}-\bar{X}_1)(X_{2i}-\bar{X}_2) = 2637.6$$

$$\sum_{i=1}^{n}(X_{2i}-\bar{X}_2)^2 = 4341.46667$$

$$\hat{\beta}_1(X_2,X_1) = \frac{2637.6}{4341.46667} = 0.6075366 2$$

$$\hat{\beta}_0(X_2,X_1) = \bar{X}_2 - \hat{\beta}_1(X_2,X_1)\bar{X}_1 = 12.671394$$

∴ The fitted regression equation of X_1 versus X_2 is:

$\hat{}$

X_2 is:

$$\hat{X}_1 = 12.671394 + 0.607537 X_2$$

Step 3: Fit a regression line of $e(Y, X_2)$ versus $e(X_1, X_2)$

$$\sum_{i=1}^{n} (e(Y,X_2)_i - \overline{e(Y,X_2)})(e(X_1,X_2)_i - \overline{e(X_1,X_2)})$$

$$= 2761.4491$$

$$\sum_{i=1}^{n} (e(X_1,X_2)_i - \overline{e(X_1,X_2)})^2 = 3538.761402$$

$$\hat{\beta}_1(e(Y,X_2), e(X_1,X_2)) = \frac{2761.4491}{3538.761402} = 0.7803434$$

$$\hat{\beta}_0(e(Y,X_2), e(X_1,X_2)) = \overline{e(Y,X_2)} - \hat{\beta}_1 \overline{e(X_1,X_2)}$$

$$= -9.4282 \times 10^{-15} \approx 0$$

∴ The fitted regression equation is

$$\hat{e}(Y,X_2) = 0 + 0.7803434\, \hat{e}(X_1,X_2)$$

⇒ 0.7803 is the coefficient of X_1 in multiple linear regression.

Let $\hat{\beta}_0(Y,X_1,X_2)$ be the intercept term in multiple linear regression of Y relates to X_1, X_2.

$$\hat{\beta}_0(Y,X_1,X_2) = \overline{Y} - \hat{\beta}_1(e(Y,X_2), e(X_1,X_2))\overline{X_1} - \hat{\beta}_1(e(Y,X_1), e(X_2,X_1))\overline{X_2}$$

$= 64.6333 - 0.7863434(66.6) - (-0.0502)(53.1333)$

$= 15.3276202$

⇒ The intercept term is 15.3276

⇒ Multiple linear regression is

$\hat{Y} = 15.3276 + 0.7863X_1 - 0.0502X_2$

Problem 2:

(a) Fit each of the following models of the data:

$$\text{Mode 1}: F = \beta_0 + \beta_1 P_1 + \varepsilon$$
$$\text{Model 2}: F = \beta_0 + \beta_2 P_2 + \varepsilon$$
$$\text{Model 3}: F = \beta_0 + \beta_1 P_1 + \beta_2 P_2 + \varepsilon$$

(b) Test whether $\beta_0 = 0$ in each of the three models.

(c) Which variable individually, P_1 or P_2, is a better predictor of F? Why?

(d) Which of the three models would you use to predict the final examination scores for a student who scored 78 and 85 on the first and second preliminary examinations, respectively? What is your prediction interval in this case?

a).
Model 1: F_hat = -22.3424 + 1.2605*P1
Model 2: F_hat = -1.85355 + 1.00427*P2
Model 3: F_hat = -14.5005 + 0.4883*P1 + 0.6720*P2

b). Using a significant level of 0.05,
We do a t-test on null hypothesis of beta0 = 0 versus alternative hypothesis of beta0 =/= 0.
From the summary in output from R,
Model 1:

```
> regression1 <- lm(F ~ P1, data = q2data)
> summary(regression1)

Call:
lm(formula = F ~ P1, data = q2data)

Residuals:
   Min     1Q Median     3Q    Max
-8.844 -2.020 -0.587  4.043  7.938

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -22.3424    11.5640  -1.932   0.0676 .
P1            1.2605     0.1399   9.008 1.78e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.081 on 20 degrees of freedom
Multiple R-squared:  0.8023,    Adjusted R-squared:  0.7924
F-statistic: 81.14 on 1 and 20 DF,  p-value: 1.779e-08
```

t-value = -1.932, t(0.05,20) = 2.086.
Since |t| < tc, We do not reject the null hypothesis, therefore beta0 in model 1 is statistically

not significant different from 0.

Model 2:

```
> regression2 <- lm(F ~ P2, data = q2data)
> summary(regression2)

Call:
lm(formula = F ~ P2, data = q2data)

Residuals:
     Min       1Q   Median       3Q      Max
-10.4323  -1.5027   0.5421   2.2580   7.5165

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.85355    7.56181  -0.245    0.809
P2           1.00427    0.09059  11.086 5.44e-10 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.275 on 20 degrees of freedom
Multiple R-squared:   0.86,     Adjusted R-squared:  0.853
F-statistic: 122.9 on 1 and 20 DF,  p-value: 5.442e-10
```

t-value = -0.245, t(0.05,20) = 2.086.
Since |t| < tc, we do not reject the null hypothesis, therefore beta0 in model 2 is statistically not significant different from 0.

Model 3:

```
> regression3 <- lm(F ~ ., data = q2data)
> summary(regression3)

Call:
lm(formula = F ~ ., data = q2data)

Residuals:
    Min      1Q  Median      3Q     Max
-8.7328 -2.1703  0.3938  2.6443  6.3660

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -14.5005     9.2356  -1.570  0.13290
P1            0.4883     0.2330   2.096  0.04971 *
P2            0.6720     0.1793   3.748  0.00136 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.953 on 19 degrees of freedom
Multiple R-squared:  0.8863,    Adjusted R-squared:  0.8744
F-statistic: 74.07 on 2 and 19 DF,  p-value: 1.069e-09
```

t-value = -1.570, t(0.05, 19) = 2.093.
Since |t| < tc, we do not reject the null hypothesis, therefore beta0 in model3 is statistically not significant different from 0.

c). We can use Multiple R-squared to compare Model 1 and Model 2 and give a conclusion on whether P1 or P2 is a better predictor of F.
Multiple R^2 of Model 1: 0.8023
Multiple R^2 of Model 2: 0.86
Therefore Model 2 has a higher multiple R^2 and thus P2 individually is a better predictor of F.
d). We can use adjusted R-squared to compare goodness of fit of three models.
Adjusted R^2 of Model 1: 0.7924
Adjusted R^2 of Model 2: 0.853
Adjusted R^2 of Model 3: 0.8744
Therefore, we will use model 3 to predict the final examination score.
Using predict() function in R,
Prediction: F_hat = -14.5005 + 0.4883 * 78 + 0.6720 * 85 = 80.71282
Prediction interval: [71.79724, 89.6284]

Table 3.11 Regression Output When Y is Regressed on X_1 for 20 Observations

ANOVA Table

Source	Sum of Squares	df	Mean Square	F-Test
Regression	1848.76	1	⑧	③
Residuals	⑥	18	⑤	

Coefficients Table

Variable	Coefficient	s.e.	t-Test	p-value
Constant	−23.4325	12.74	①	0.0824
X_1	②	0.1528	8.32	< 0.0001
n = 20	R^2 = ⑨	R_a^2 = ⑩	$\hat{\sigma}$ = ⑦	df = 18

① : $\frac{-23.4325}{12.74} = -1.84$

② : $8.32 \times 0.1528 = 1.271$

③ : $F = (t_1(df_1))^2 = (8.32)^2 = 69.222$

④ : $MSR = \frac{SSR}{p} = \frac{1848.76}{1} = 1848.76$

⑤ : $F = \frac{MSR}{MSE}$

$69.22 = \frac{1848.76}{MSE}$

$MSE = 26.70846576$

$MSE = 26.708$

⑥ : $SSE = MSE \times (n-p-1)$

$SSE = 26.708 \times 18$

$SSE = 480.744$

⑦ : $\hat{\sigma} = \sqrt{\frac{SSE}{n-p-1}} = \sqrt{MSE} = \sqrt{26.708} = 5.168$

⑧ : $R^2 = \frac{SSR}{SSR+SSE} = \frac{1848.76}{1848.76+480.78} = 0.794$

⑨ : $R_a^2 = 1 - \frac{n-1}{n-p-1}(1-R^2) = 1 - \frac{19}{18}(1-0.794) = 0.783$

$Var(Y) = \frac{SST}{} \quad 122.605 \quad Var(X_1) = [Cov(X,Y)]^2 Var(Y)$

$Var(Y) = \frac{SST}{n-1} = 122.605$ $Var(X_1) = \frac{[Cov(X,Y)]^2 Var(Y)}{\hat{\beta}_1^2}$

$= 60.261$

Table 3.11 Regression Output When Y is Regressed on X_1 for 20 Observations

ANOVA Table

Source	Sum of Squares	df	Mean Square	F-Test
Regression	1848.76	1	1848.76	69.222
Residuals	480.736	18	26.708	

Coefficients Table

Variable	Coefficient	s.e.	t-Test	p-value
Constant	−23.4325	12.74	-1.84	0.0824
X_1	1.271	0.1528	8.32	< 0.0001
$n = 20$	$R^2 = 0.794$	$R_a^2 = 0.783$	$\hat{\sigma} = 5.168$	df = 18

Problem 4.

Problem 4 (*use* R) Using the Supervisor Performance data *Supervisor.txt*, test the hypothesis $H_0 : \beta_1 = \beta_3 = 0.5$ in each of the following models:

(a) $Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$

(b) $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$

a). It can be divided in 2 steps, first: test whether beta1 = beta3 = beta1', second: test whether beta1' = 0.5

First,

H0 : beta1 = beta3 = 0.5 vs H1: Y = beta0 + beta1*X1 + beta3*X3 + epsilon

So Full model is:

Y = beta0 + beta1X1 + beta3X3 + eps

Under the null hypothesis, where beta1 = beta3 = 0.5, the reduced model is

Y = beta0' + 0.5*(X1+X3) + eps

Create a new variable, W = X1 + X3, and the model becomes:

Y = beta0' + 0.5*W + eps

Y - 0.5* W = beta0' + eps

```
#q4
#a).
q4data <- read.table("~/Downloads/3424/HW2/Supervisor.txt", head = TRUE)
q4y <- q4data$Y
q4x1 <- q4data$X1
q4x2 <- q4data$X2
q4x3 <- q4data$X3
q4w <- q4x1 + q4x3
q4regression1 <- lm(q4y ~ q4x1 + q4x3, data = q4data)
q4regression2 <- lm(q4y-0.5*q4x1-0.5*q4x3 ~ 1, data = q4data)
#anova(q4regression1, q4regression2)
n <- 30
p <- 2
k <- 1
q4a_F_val <- ((anova(q4regression2)[1,2] - anova(q4regression1)[3,2])/(p+1-k))/(anova(q4regression1)[3,2]/(n-p-1))
q4a_F_val
q4a_F_crit <- qf(p=0.05, df1=2, df2=27, lower.tail = FALSE)
q4a_F_crit

#(b)
p <- 3
k <- 2
q4regression3 <- lm(q4y ~ q4x1 + q4x2 + q4x3, data = q4data)
q4regression4 <- lm(q4y - 0.5*q4x1 - 0.5*q4x3 ~ q4x2, data = q4data)
q4b_F_val <- ((anova(q4regression4)[2,2] - anova(q4regression3)[4,2])/(p+1-k))/(anova(q4regression3)[4,2]/(n-p-1))
q4b_F_val
q4b_F_crit <- qf(p=0.05, df1=2, df2=26, lower.tail = FALSE)
q4b_F_crit
```

R code is shown above and output is shown below:

```
> q4data <- read.table("~/Downloads/3424/HW2/Supervisor.txt", head = TRUE)
> q4y <- q4data$Y
> q4x1 <- q4data$X1
> q4x2 <- q4data$X2
> q4x3 <- q4data$X3
> q4w <- q4x1 + q4x3
> q4regression1 <- lm(q4y ~ q4x1 + q4x3, data = q4data)
> q4regression2 <- lm(q4y-0.5*q4x1-0.5*q4x3 ~ 1, data = q4data)
> #anova(q4regression1, q4regression2)
> n <- 30
> p <- 2
> k <- 1
> q4a_F_val <- ((anova(q4regression2)[1,2] - anova(q4regression1)[3,2])/(p+1-k))/(anova(q4regression1)[3,
2]/(n-p-1))
> q4a_F_val
[1] 2.3126
> q4a_F_crit <- qf(p=0.05, df1=2, df2=27, lower.tail = FALSE)
> q4a_F_crit
[1] 3.354131
>
> #(b)
> p <- 3
> k <- 2
> q4regression3 <- lm(q4y ~ q4x1 + q4x2 + q4x3, data = q4data)
> q4regression4 <- lm(q4y - 0.5*q4x1 - 0.5*q4x3 ~ q4x2, data = q4data)
> q4b_F_val <- ((anova(q4regression4)[2,2] - anova(q4regression3)[4,2])/(p+1-k))/(anova(q4regression3)[4,
2]/(n-p-1))
> q4b_F_val
[1] 1.975316
> q4b_F_crit <- qf(p=0.05, df1=2, df2=26, lower.tail = FALSE)
> q4b_F_crit
[1] 3.369016
```

F_val = ((SSE_rm - SSE_fm) / (p + 1 - k)) / (SSE_fm / (n - p - 1))
F_val is used for testing if beta1 = beta3.
F_val = 2.3126 . The tabulated value of F(2,27,0.05) = 3.3541. Therefore, the resulting F is not significant, thus null hypothesis is not rejected at a significant level of 0.05

b).
H0 : beta1 = beta3 = 0.5 vs H1: Y = beta0 + beta1*X1 + beta2 * X2 + beta3 * X3 + epsilon
So the full model is:
Y = beta0 + beta1X1 + beta2X2 + beta3X3 + eps
Under the null hypothesis, where beta1 = beta3 = 0.5, the reduced model is:
Y = beta0 + beta2X2 + beta1'(X1+X3) + eps
Create a new variable, W = X1 + X3, and the reduced model becomes:
Y = beta0' + beta2X2 + 0.5*W + eps
Y - 0.5W = beta0' + beta2X2 + eps
From the R result, F_val is 1.975316 and F(0.05,2,26) = 3.369
Therefore, the resulting F is not significant, thus null hypothesis is not rejected at a significant level of 0.05

Problem 5

Problem 5 Table 3.14 shows the regression output of a multiple regression model relating the beginning salaries in dollars of employees in a given company to the following predictor variables:

Gender: An indicator variable 1=man and 0=woman
Education: Years of schooling at the time of hire
Experience: Number of months of pervious work experience
Months: Number of months with the compnay

Table 3.14 Regression Output When Salary is Related to Four Predictor Variables

ANOVA Table

Source	Sum of Squares	df	Mean Square	F-Test
Regression	23665352	4	5916338	22.98
Residuals	22657938	88	257477	

Coefficients Table

Variable	Coefficient	s.e.	t-Test	p-value
Constant	3526.4	327.7	10.76	0.000
Gender	722.5	117.8	6.13	0.000
Education	90.02	24.69	3.65	0.000
Experience	1.2690	0.5877	2.16	0.034
Months	23.406	5.201	4.50	0.000
$n = 93$	$R^2 = 0.515$	$R_a^2 = 0.489$	$\hat{\sigma} = 507.4$	df = 88

In (a)-(b) below, specify the null and alternative hypothesis, the test used, and your conclusion using a 5% level of significance. For (c)-(e), only point prediction is desired.

(a) Conduct the F-test for the overall fit of the regression

(b) Is there a *positive* linear relationship between salary and experience, after accounting for the effect of the variables Gender, Education and Months?

(c) What salary would you forecast for a man with 12 years of education, 10 months of experience, and 15 months with the company?

(d) What salary would you forecast, on average, for men with 12 years of education, 10 months of experience, and 15 months with the company?

(e) What salary would you forecast, on average, for women with 12 years of education, 10 months of experience, and 15 months with the company?

a). F-test is used to assess the overall fit of a regression,
H0 = true values of all of the regression coefficients are 0
H1 = true values of some or all of regression coefficients are non-zero
Formula =

$$F = \frac{[\text{SSE(RM)} - \text{SSE(FM)}]/(p+1-k)}{\text{SSE(FM)}/(n-p-1)}.$$

Value of test statistics is given, which is equal to 22.98
F(0.05,4,88) critical value = 2.475277
F > Fc, so reject the null hypothesis.

b). T-test is used to test true value of the regression coefficient of Experience variable
H0 = true value of the regression coefficient of Experience is 0
H1 = true value of the regression coefficient of Experience is greater than 0
This is a one-sided test

$$t_j = \frac{\hat{\beta}_j}{\text{s.e.}(\hat{\beta}_j)},$$

T test statistics is given in the table, which is equal to 2.16
T(0.05, 93-4-1) critical value = 1.662354
T > Tc so we reject the null hypothesis

c). The forecasted salary should be 3526.4 + 1*722.5 + 12*90.02 + 10*1.2690 + 15*23.406 = 5692.92
d). The value of the forecasted mean salary is the same as above(5692.92), with a smaller standard error.
e). The forecasted salary should be 3526.4 + 0*722.5 + 12*90.02 + 10*1.2690 + 15*23.406 = 4970.42

Q6.

Problem 6 Consider the regression model that generated the output in Table 3.14 to be a full model. Now consider the reduced model in which Salary is regressed on only Education. The ANOVA table obtained when fitting the model is shown in Table 3.15. Conduct a single test to compare the full and reduced models. What conclusion can be drawn from the result of the test? (Use $\alpha = 0.05$.)

Table 3.15 ANOVA Table When the Beginning Salary is Regressed on Education

ANOVA Table

Source	Sum of Squares	df	Mean Square	F-Test
Regression	7862535	1	7862535	18.60
Residuals	38460756	91	422646	

We use the following formula to test the null hypothesis that the reduced model involving only Education is adequate, versus
in this case, p = 4 and k = 2.
F = ([SSE(RM) - SSE(FM)] / (p+1-k))/(SSE(FM) / (n-p-1)) = ((38460756 - 22657938)/(4+1-2))/((22657938)/(93 - 4 - 1)) = 20.4586
Fcritical = F(0.05, 3, 88) = 2.708
F > Fcritical so we reject the null hypothesis, concluding that some or all of the other predictors are needed.

Q7.

Problem 7 (*use* R) Consider the computer repair problem discussed in Chapter 2. In a second sampling period, 10 more observations on the variables Minutes and Units were obtained. Since all observations were collected by the same method from a fixed environment, all 24 observations were pooled to form one dataset. The data appear in Table 4.6 and stored in the file *Computer_Repair.txt*.

(a) Fit a linear regression model relating Minutes to Units.

(b) Check each of the standard regression assumptions and indicate which assumptions seems to be violated.

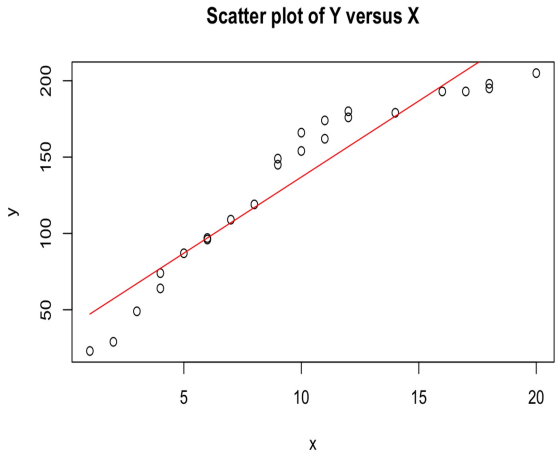
7a).

```
q7data <- read.table("~/Downloads/3424/HW2/Computer_Repair.txt", head = TRUE)
q7lm <- lm(Minutes ~ Units, data = q7data)
summary(q7lm)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  37.2127     7.9853    4.66  0.00012 ***
Units         9.9695     0.7218   13.81 2.56e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 18.75 on 22 degrees of freedom
Multiple R-squared:  0.8966,    Adjusted R-squared:  0.8919
F-statistic: 190.7 on 1 and 22 DF,  p-value: 2.556e-12
```

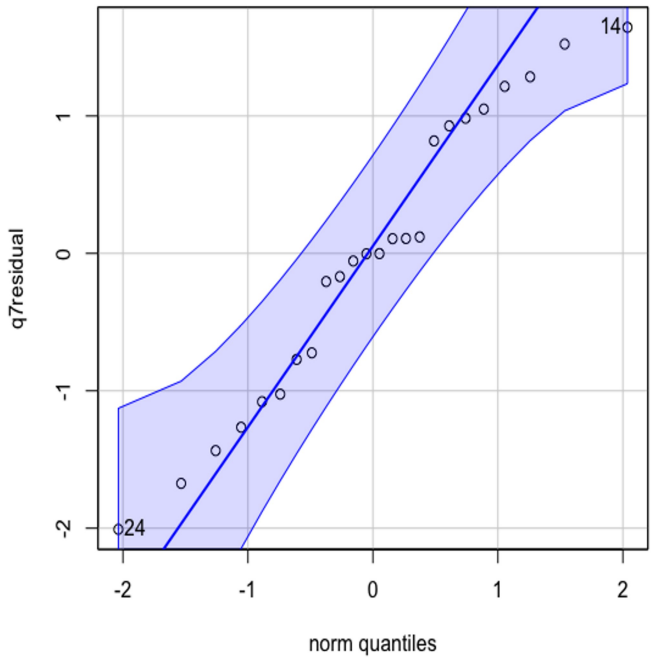
Fitted regression line is Minutes = 37.2127 + 9.9695*Units

B). 1. linearity, we can check the scatter plot of Y against X. A linear scatter plot ensures linearity.

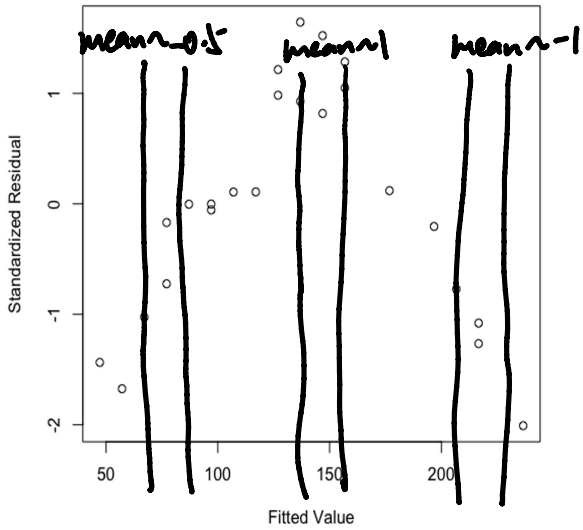


We can see that the scatter plot of Y vs X shows a good linearity. Therefore the linearity assumption is fulfilled.

2.1 For the assumption of errors being normally distributed, we can check the Q-Q plot, ordered residuals versus normal scores,
and we observed that most of the residuals lied near the line y=x, so this assumption is held.

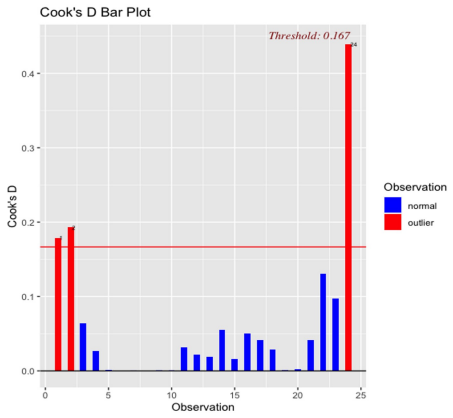


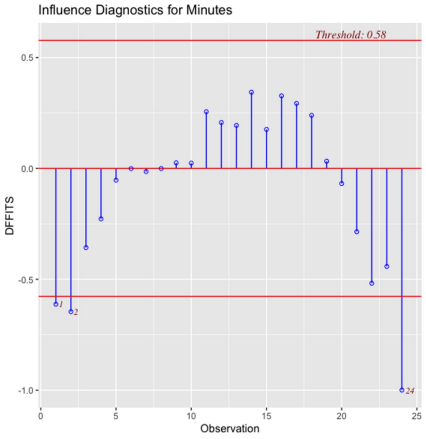
2.2) The mean of errors are 0, we can use the scatter plot of standardized residual versus fitted value to test this assumption. For most of the fitted values, the average of residual at that point is not 0, so this assumption is violated.



4. For the assumption of observations having equal influence, we can check the index plot of Cook's distance or DFITS.
   From the index plot of Cook's distance, we can see that #1,#2 and #24 are influential points.

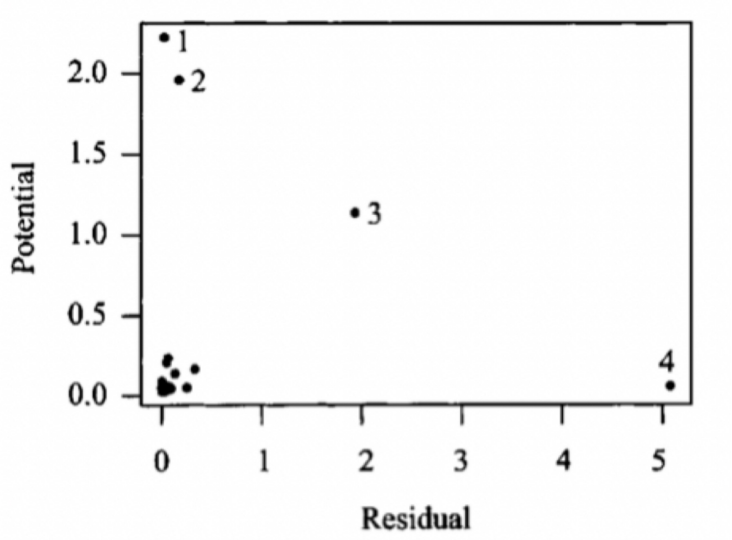
   From the index plot of DFITS, we can see that #1, #2 and #24 are influential points.
   So the assumption is violated, as there are 3 influential points.





q8).

Problem 8 In an attempt to find unusual points in a regression data set, a data analyst examines the *potential-Residual* plot. Classify each of the unusual points on this plot according to the type.



observation 1 and 2 are points of high potential with small residual, they are outliers in X-space
Observation 3 is of high potential and high residual, it is influential point.
Observation 4 is of high residual with small potential, it is an outlier in Y-space.

Q9).

Problem 9 Name one or more graphs that can be used to validate each of the following assumptions. For each graph, sketch an example where the corresponding assumption is valid and an example where the assumption is clearly invalid.

(a) There is a linear relationship between the response and predictor variables.

(b) The observations are independent of each other

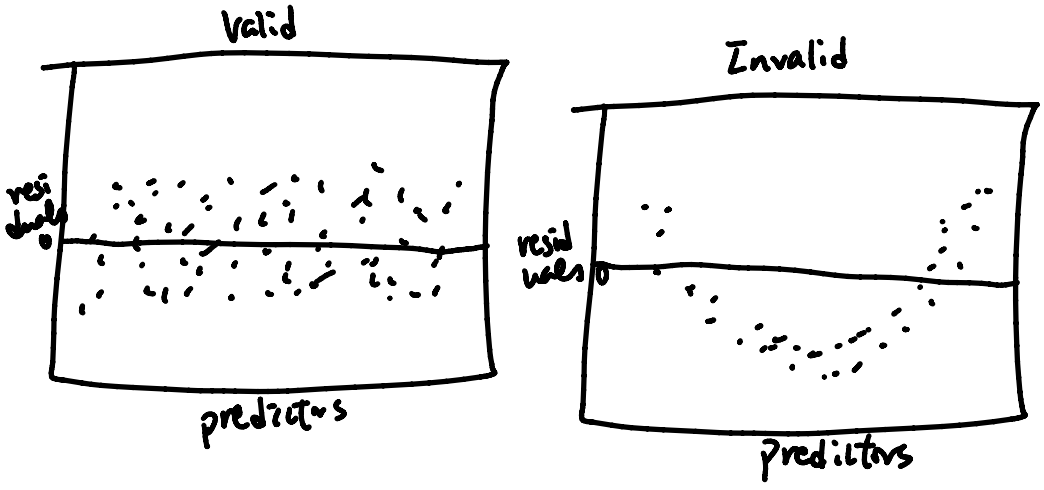
(c) The error terms have constant variance

(d) The error terms are uncorrelated

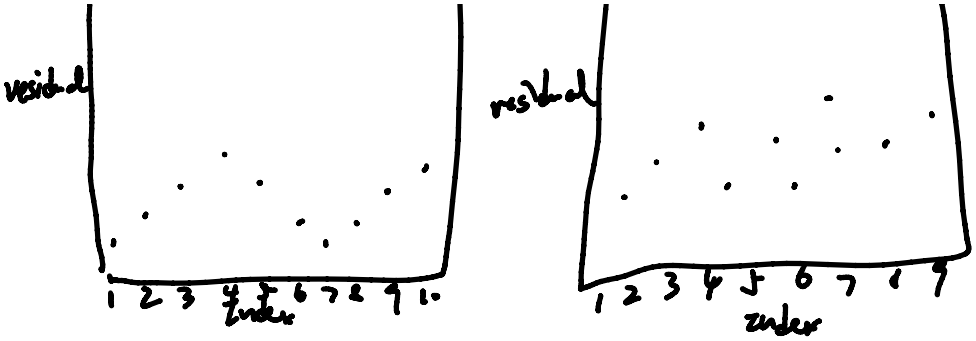
(e) The error terms are normally distributed

(f) The observations are equally influential on least squares results

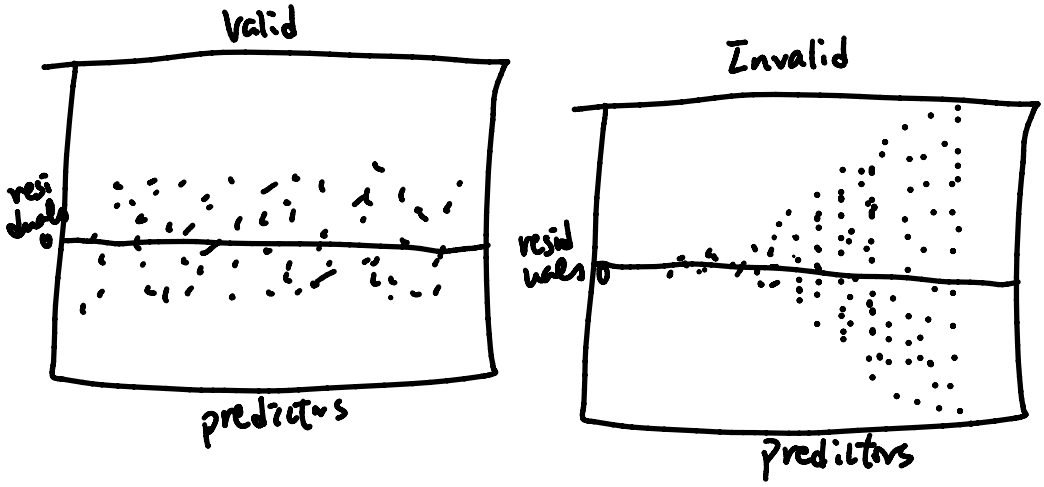
a). Scatterplots of the internally studentized residuals against the predictors can be used. If the mean of residuals is roughly 0 at every fitted value, the linearity assumption is valid.



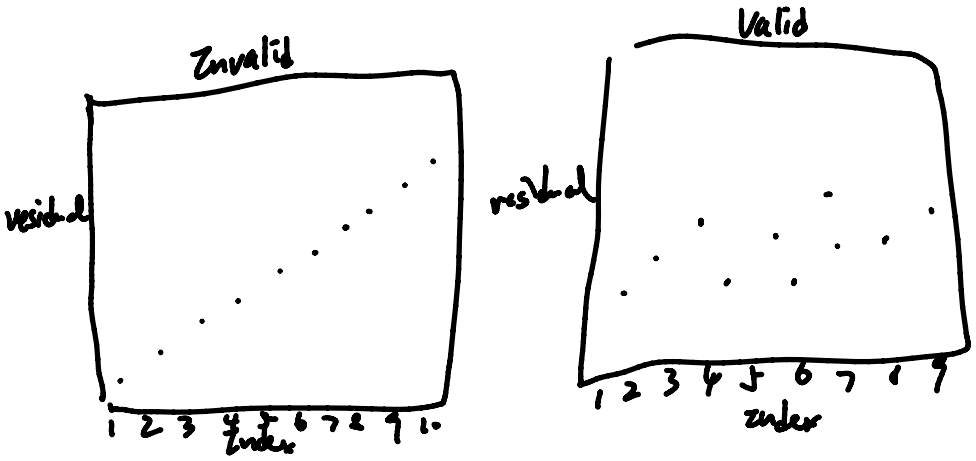
b). For the assumption of errors are independent of each other, index plot of standardized residuals can be used to test this assumption. If the plot shows some specific patterns(e.g. sine curve or a straight line), we know that this assumption is violated.



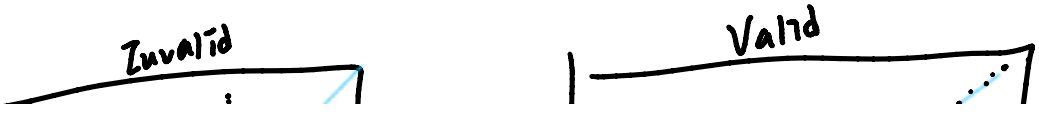
c).Error terms have constant variance:
Scatter plots of internally studentized residuals against the fitted values or
Scatter plots of internally studentized residuals against the predictors
can be used to test the assumption of constant variance
If the spread of residuals is kept constant/similar at every fitted value, the assumption of constant variance of error terms is valid.

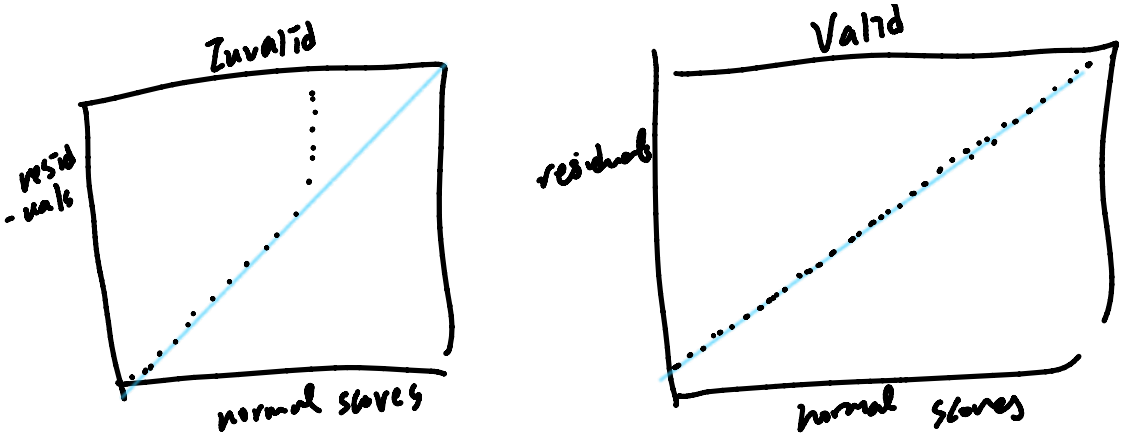


d). For the assumption of errors are uncorrelated, index plot of standardized residuals can be used to test this assumption. If the plot shows some specific patterns(e.g. a straight line or a quadratic function), we know that this assumption is violated as they should have correlated inherited.

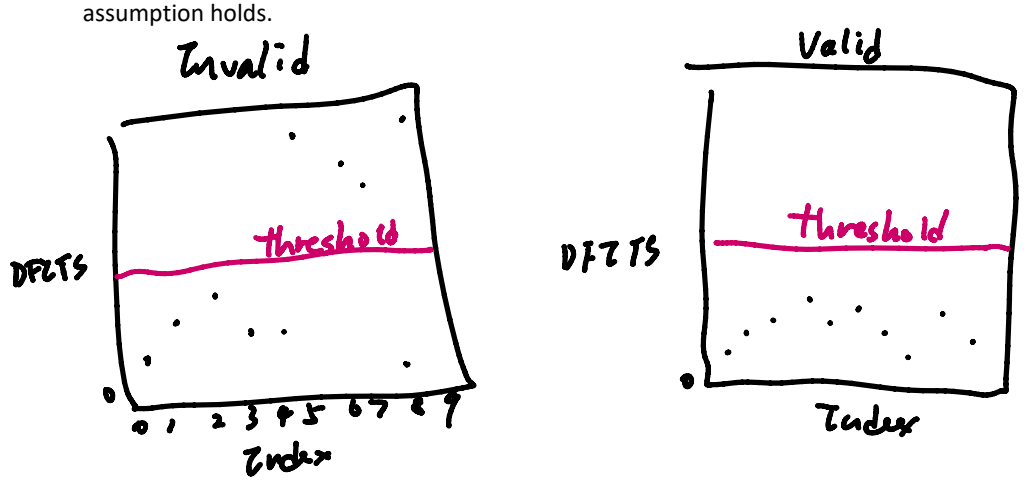


e). For assumption of errors are normal distributed, Q-Q plot (Scatter plot of ordered Residuals versus normal scores) can be used to examine this assumption. If the points are lied on the line of y=x, the assumption can be held. Otherwise, this assumption is violated.





f). For the assumption that the observations are equally influential on least squares results, we can use index plot of DFITS to test this assumption. If all data points are below the threshold of 2*sqrt((p+1)/(n-p-1)), with p = numbers of regression coefficient and n is the sample size, this assumption holds.



Problem 10 Consider again the Examination data used in *Problem 2* and given in Table 3.10.

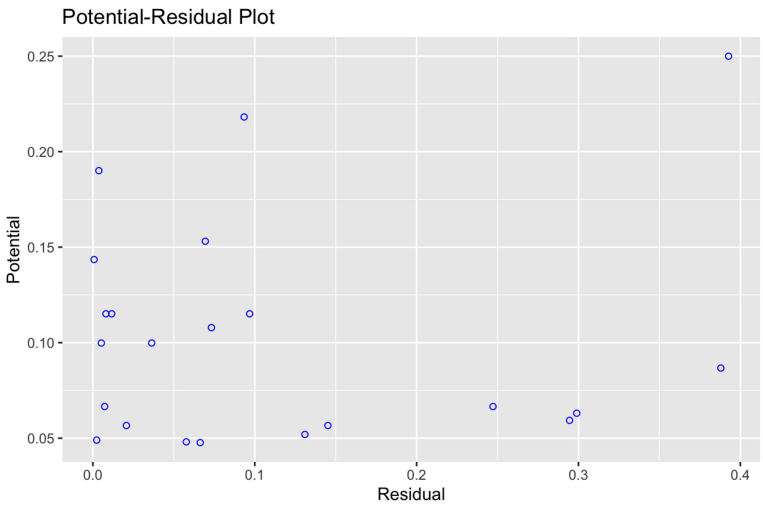
(a) For each of the three models, draw the *potential-Residual* plot. Identify all unusual observations (by number) and classify as outliers, high-leverage point, and/or influential observation.

(b) What model would you use to predict the final score F?

10a). We use threshold for outlier in Y-space with $|r_i| > 3$, outlier in X space with $p_{ii} > 2*(p+1)/n$.
In model 1 & 2, the threshold for leverage is 4/22 and 6/22 in model 3.
We use threshold for influential point by $|DFITS| > 2*\sqrt{((p+1)/(n-p+1))}$. In model 1 & 2, the DFITS threshold is 0.632, where in model 3, this threshold is 0.795.
We can also consider Cook's distance, where the threshold is

For the Potential-Residual plot,

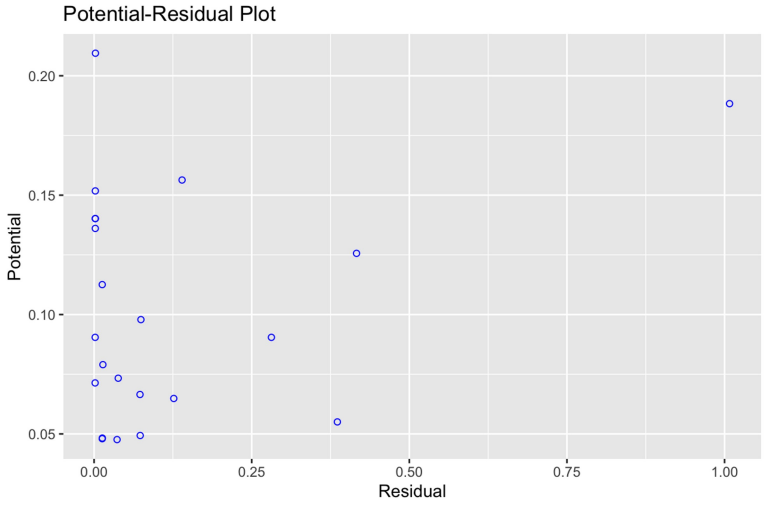
Model 1:



Using the result from R, points 9 is an outlier in X-space.
no points are outlier in Y-space.
point 9 is influential points.

```
[1] "Finding outliers & influential points in model 1:"
> print(hatvalues(q2regression1) < 2*2/22)
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   17    18    19    20    21    22
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> print(abs(rstandard(q2regression1)) < 3)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  21   22
TRUE TRUE
> print(abs(dffits(q2regression1)) < 2*sqrt((1+1)/(22-1-1)))
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   17    18    19    20    21    22
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> ols_plot_dffits(q2regression1)
```

Model 2:

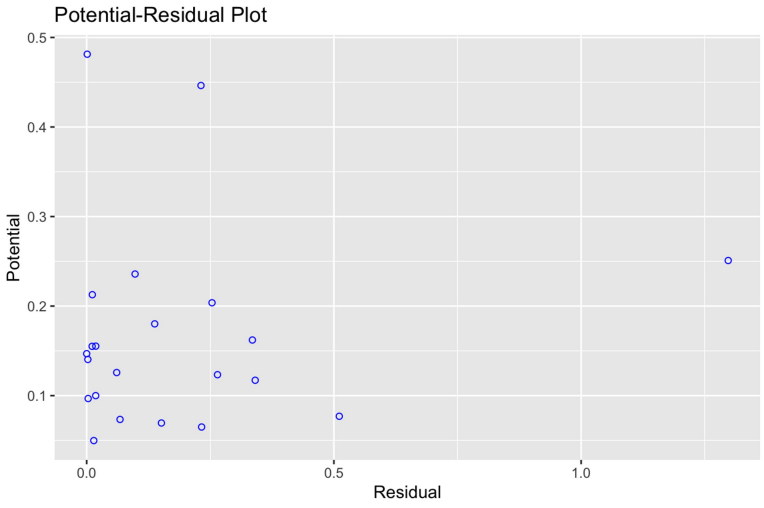


Using the result from R, no points are outlier in X-space.
No points are outlier in Y-space.

point 7 and 9 are influential points.

```
[1] "Finding outliers & influential points in model 2:"
> print(hatvalues(q2regression2) < 2*2/22)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  21   22
TRUE TRUE
> print(abs(rstandard(q2regression2)) < 3)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  21   22
TRUE TRUE
> print(abs(dffits(q2regression2)) < 2*sqrt((1+1)/(22-1-1)))
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   17    18    19    20    21    22
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> ols_plot_dffits(q2regression2)
```

Model 3:



Using the result from R, point 7 and 15 are outlier in X-space.
No points are outliers in Y-space.
point 7 and 9 are an influential point.

```
> print("Finding outliers & influential points in model 3:")
[1] "Finding outliers & influential points in model 3:"
> print(hatvalues(q2regression3) < 2*3/22)
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE
   17    18    19    20    21    22
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> print(abs(rstandard(q2regression3)) < 3)
   1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19   20
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
  21   22
TRUE TRUE
> print(abs(dffits(q2regression3)) < 2*sqrt((2+1)/(22-2-1)))
    1     2     3     4     5     6     7     8     9    10    11    12    13    14    15    16
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE FALSE  TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
   17    18    19    20    21    22
 TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```

10b). I would choose model 1 because the potential and residuals are lower than other two models and the influence is more balanced across all observations.

Problem 11 Identify unusual observations for the data set in Table 4.7.

Table 4.7

Row	Y	X	Row	Y	X
1	8.11	0	7	9.60	19
2	11.00	5	8	10.30	20
3	8.20	15	9	11.30	21
4	8.30	16	10	11.40	22
5	9.40	17	11	12.20	23
6	9.30	18	12	12.90	24

Q11.
We can use leverage to identify outlier in X-space: pii = 1/n + (xi-xmean)^2/sum((xi-xmean)^2)
Threshold for leverage is pii >= 2 * (p+1)/n = 1/3
We can find that 1st observation having leverage of 0.5650289 > 1/3, so it is an outlier in X-space

We can use standardized residual to identify outlier in Y-space: |ri| < 3, and according to R, no outliers are found.

We can use DFITS to identify influential points, |DFITS| < sqrt((1+1)/(12-1+1)) = 0.408
According to R, 2nd observation is found as influential point.

```
> q11X <- c(0,5,15,16,17,18,19,20,21,22,23,24)
> q11Y <- c(8.11,11.00,8.20,8.30,9.40,9.30,9.60,10.30,11.30,11.40,12.20,12.90)
> q11regression <- lm( q11Y ~ q11X, data = data.frame(q11X + q11Y))
> q11pii <- 1/12 + (q11X - mean(q11X))^2/sum((q11X - mean(q11X))^2)
> print(q11pii < 1/3)
 [1] FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
> print(abs(rstandard(q11regression)) < 3)
   1    2    3    4    5    6    7    8    9   10   11   12
TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
> print(abs(dffits(q11regression)) < 2*sqrt((1+1)/(12-1-1)))
    1     2     3     4     5     6     7     8     9    10    11    12
 TRUE FALSE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE  TRUE
```