STA 108: Homework 1
Hangshi Jin
913142686
Christiana Drake
Section A01

Problem 1.7

(a)we cannot state the exact probability because we do not know if $\epsilon_i \sim N(0, \sigma^2)$.

(b)

$$E[Y] = 100 + 20 \cdot 5 = 200.$$

Since we assumed $\epsilon_i \sim N(0, \sigma^2)$,

$$z_1 = \frac{195 - 200}{5} = -1, z_2 = \frac{205 - 200}{5} = 1.$$

$$P(-1 \le z \le 1) = 0.8413 - 0.1587 = 0.6826.$$

Problem 1.21

(a)

$$\overline{X} = \frac{10}{10} = 1, \overline{Y} = \frac{142}{10} = 14.2.$$

$$\beta_1 = E[b_1] = \frac{40}{10} = 4, \beta_0 = E[b_0] = 14.2 - b_1 = 10.2.$$

$$\Rightarrow Y_i = 10.2 + 4X_i.$$

$$SSE = 17.6, SSTO = 177.6, \Rightarrow R^2 = 1 - \frac{17.6}{177.6} = 0.9009$$

Since R^2 is 0.9009, the linear regression function appear to give a good fit here.

(b)

$$Y_i = 10.2 + 4 = 14.2$$

The expected number of broken ampules when X=1 transfer is made is 14.2.

Problem 1.25

(a)

$$e_1 = Y_1 - \hat{Y}_1 = 16 - 14.2 = 1.8$$

ϵ_1 is the difference between true mean and the observed value, whereas e_1 is the difference between the observed value and the predicted value based on the linear regression function.

(b)

$$\sum e_i^2 = SSE = 17.6$$

$$MSE = \frac{SSE}{10 - 2} = 2.2$$

MSE estimates the variance of e_i.

Problem 1.38

(1)

$$\hat{Y}_i = \begin{bmatrix} 12 & 9 & 15 & 9 & 18 & 12 & 9 & 12 & 15 & 9 \end{bmatrix}$$
$$\Rightarrow e_i = \begin{bmatrix} 4 & 0 & 2 & 3 & 4 & 1 & -1 & 3 & 4 & 2 \end{bmatrix}$$
$$\sum e_i^2 = 76 > 17.6$$

(2)

$$\hat{Y}_i = \begin{bmatrix} 16 & 11 & 21 & 11 & 26 & 16 & 11 & 16 & 21 & 11 \end{bmatrix}$$
$$\Rightarrow e_i = \begin{bmatrix} 0 & -2 & -4 & 1 & -4 & -3 & -3 & -1 & -2 & 0 \end{bmatrix}$$
$$\sum e_i^2 = 60 > 17.6$$

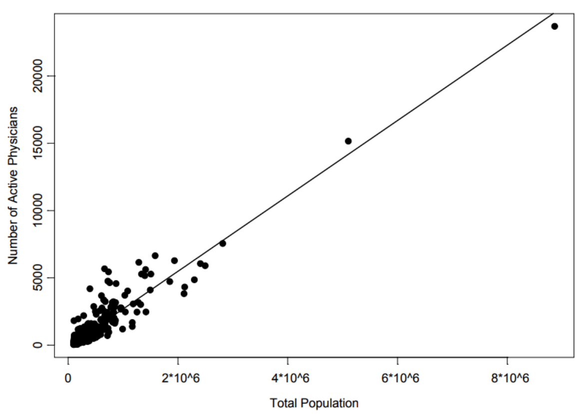
The criterion is larger for these estimates than for the least squares estimates.

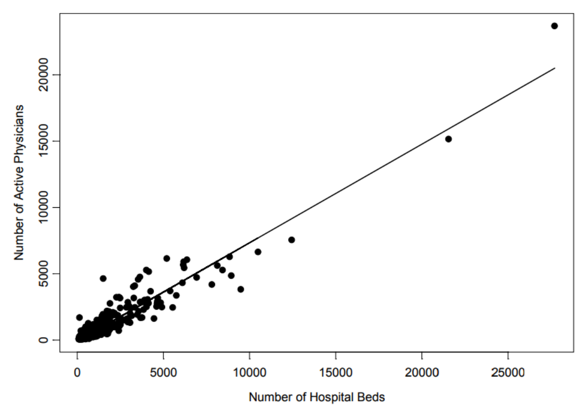
Problem 1.43

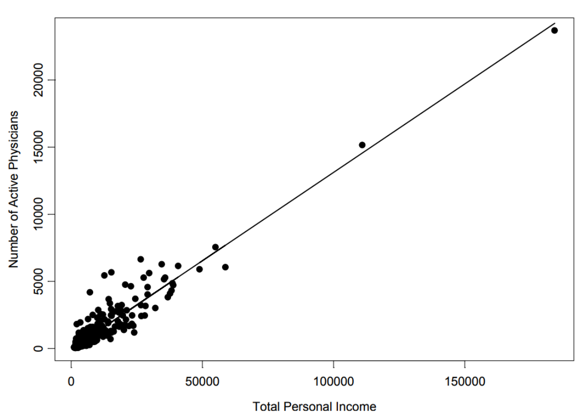
(a)Let Y be the number of active physicians, X_1 be the total population,X_2 be the number of hospital beds, X_3 be the total personal income.

$$\hat{Y}_i = -110.635 + 0.003X_{1i}$$
$$\hat{Y}_i = -95.932 + 0.743X_{2i}$$
$$\hat{Y}_i = -48.395 + 0.132X_{3i}$$

(b)







The linear regression has a good fit for each one, but there are two points that are out of scale.

(c)

Predictors	MSE
X_1	372204
X_2	310192
X_3	324539

We can see that regression function with X_2 has the smallest variation around the fitted regression line.

Problem 2.1

The conclusion is warranted. The level of siginificance is 0.05.

Because applying the regression model, the program may ignore the fact that sale cannot be negative.

Problem 2.6

(a)

$$t(0.975, 8) = 2.306, s\{b_1\} = 0.469 \Rightarrow 4 - 2.306 \cdot 0.469 \leq \beta_1 \leq 4 + 2.306 \cdot 0.469 \Rightarrow 2.918 \leq \beta_1 \leq 5.082.$$

We are 95% confident that the increase of number of ampules found to be broken upon arrival lies somewhere between 2.918 and 5.082 per unit of number of times the carton was transfered.

(b)Alternatives:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

$$t^* = \frac{b_1}{s\{b_1\}} = 8.528$$

Decision rule:

$$\text{If } |t^*| \leq t, \text{ conclude } H_0$$

$$\text{If } |t^*| > t, \text{ conclude } H_a$$

Since $|t^*| = 8.528 > 2.306$, we conclude H_a.

$$P\{t(8) > t^*\} < 0.0005 \Rightarrow \text{p-value} < 0.001$$

(c)

$$t(0.975, 8) = 2.306, s\{b_0\} = 0.663 \Rightarrow 10.2 - 0.663 \cdot 2.306 \leq \beta_0 \leq 10.2 + 0.663 \cdot 2.306 \Rightarrow 8.671 \leq \beta_0 \leq 11.729$$

We are 95% confident that the increase of number of ampules found to be broken upon arrival lies somewhere between 8.671 and 11.729 when number of times the carton was transfered is 0.

(d)Alternatives:

$$H_0 : \beta_0 > 9$$

$$H_a : \beta_0 \leq 9$$

Decision rule:

$$\text{If } t^* \leq t, \text{ conclude } H_0$$

$$\text{If } t^* > t, \text{ conclude } H_a$$

$$t(0.975; 8) = 2.306, b_0 = 10.2, t^* = \frac{b_0 - \beta_0}{s\{b_0\}} = 1.810 < 2.306$$

We conclude that the mean number of broken ampules will exceed 9.0 when no transfers are made.

$$\text{p-value} < 0.001$$

(e)

$$\delta = \frac{|2-0|}{0.5} = 4 \Rightarrow \text{Power of test in part (b)} = 0.94$$

$$\delta = \frac{|11-9|}{0.75} = 2.667 \Rightarrow \text{Power of test in part (d)} = 0.42 + \frac{2.667-2}{3-2}(0.75-0.42) = 0.640$$

Problem 2.10

(a) A prediction interval for the observation on the humidity level in this greenhouse tomorrow is appropriate, because temperature is predictor and we want to find the interval humidity falls as an observation.

(b)Confidence interval is appropriate here, because we try to want the average dispense on meals ouside.

(c)Prediction interval is appropriate here, because assumed that the index business activity remains at its present level we try to find the interval which electricity usage as an observation falls.

Problem 2.15

(a)

$$\text{For} X = 2, \hat{Y_h} = 18.2, t(0.995; 8) = 3.355$$

$$s\{\hat{Y_h}\} = \sqrt{2.2(\frac{1}{10} + \frac{1}{10})} = 0.663$$

$$\Rightarrow 18.2 - 3.355 \cdot 0.663 \leq E\{Y_h\} \leq 18.2 + 3.355 \cdot 0.663 \Rightarrow 15.976 \leq E\{Y_h\} \leq 20.424$$

We are 99% confident that the number of mean breakage for 2 transfers lies between 15.976 and 20.424.

$$\text{For} X = 4, \hat{Y_h} = 26.2, t(0.995; 8) = 3.355$$

$$s\{\hat{Y_h}\} = \sqrt{2.2(\frac{1}{10} + \frac{9}{10})} = 1.483$$

$$\Rightarrow 26.2 - 3.355 \cdot 1.483 \leq E\{Y_h\} \leq 26.2 + 3.355 \cdot 1.483 \Rightarrow 21.225 \leq E\{Y_h\} \leq 31.175$$

We are 99% confident that the number of mean breakage for 4 transfers lies between 21.225 and 31.175.

(b)

$$\text{For} X = 2, \hat{Y_h} = 18.2, t(0.995; 8) = 3.355$$

$$s\{\text{pred}\} = \sqrt{2.2 + 0.44} = 1.625$$

$$\Rightarrow 18.2 - 3.355 \cdot 1.625 \leq Y_h \leq 18.2 + 3.355 \cdot 1.625 \Rightarrow 12.748 \leq Y_h \leq 23.652$$

With confidence coefficient .99, we predict that the number of breakage for the next shipment of two transfers will be somewhere between 12.748 and 23.652.

(c)

$$s\{\text{predmean}\} = \sqrt{\frac{2.2}{3} + 0.44} = 1.083$$

$$\Rightarrow 18.2 - 3.355 \cdot 1.083 \leq \overline{Y}_h \leq 18.2 + 3.355 \cdot 1.083 \Rightarrow 14.567 \leq \overline{Y}_h \leq 21.833$$

$$3 \cdot 14.567 \leq Total(Y) \leq 3 \cdot 21.833 \Rightarrow 43.701 \leq Total(Y) \leq 65.499$$

(d)

$$F(.99; 2, 8) = 8.649 \Rightarrow W = \sqrt{2 \cdot 8.649} = 4.159$$

For $X = 2, 18.2 - 4.159 \cdot 0.663 \leq \beta_0 + \beta_1 X_h \leq 18.2 + 4.159 \cdot 0.663 \Rightarrow 15.443 \leq \beta_0 + \beta_1 X_h \leq 20.957$

For $X = 4, 26.2 - 4.159 \cdot 1.483 \leq \beta_0 + \beta_1 X_h \leq 26.2 + 4.159 \cdot 1.483 \Rightarrow 20.032 \leq \beta_0 + \beta_1 X_h \leq 32.368$

The confidence band is a bit wider because it approximates the regression line. In this case, the approximation is fairly precise.

Problem 2.25

(a)

Source	df	SS	MS	F
Regression	1	160	160	72.727
Error	8	17.6	2.2	
Total	9	177.6		

$$SSE + SSR = SSTO, df_{SSE} + df_{SSR} = df_{SSTO}$$

(b)

$$F(.95; 1, 8) = 5.318$$

Alternatives:

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

$$F^* = \frac{MSR}{MSE} = 72.727$$

Decision rule:

$$\text{If } F^* \leq F(.95; 1, 8), \text{ conclude } H_0$$

$$\text{If } F^* > F(.95; 1, 8), \text{ conclude } H_a$$

Since $F^* = 72.727 > 5.318$, we conclude there is a linear association between the number of times a carton is transferred and the number of broken ampules.

(c)

$$t^* = 8.528 \Rightarrow (t^*)^2 = F^*$$

(d)

$$R^2 = 1 - \frac{17.6}{177.6} = 0.9009$$

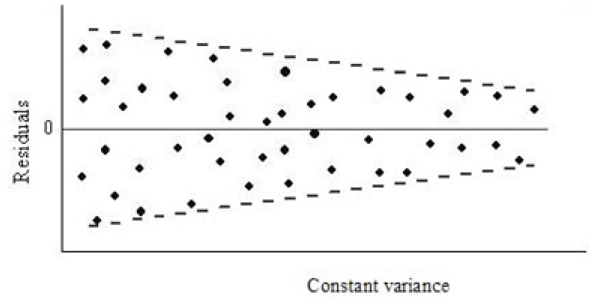
Since b_1 is positive,

$$r = \sqrt{R^2} = 0.9492$$

Thus, 90.09% of the variation in Y is acoounted for by introducing X into the regression model.

Problem 3.2

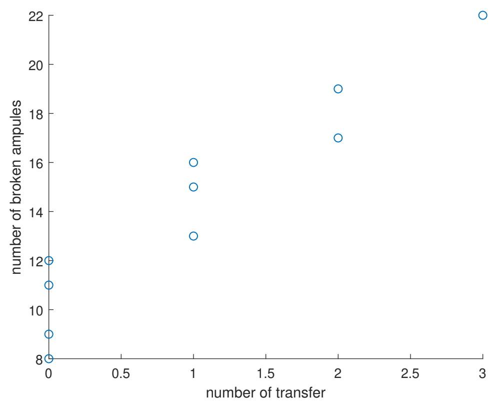
(a)



(b)

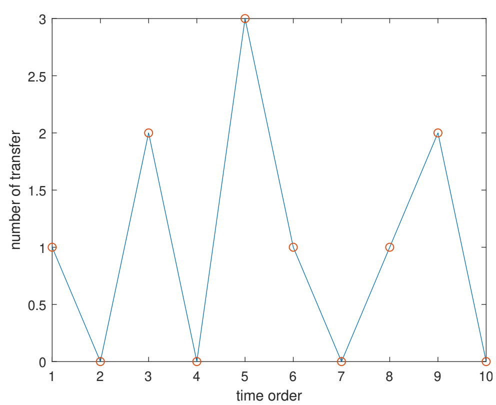
Problem 3.5

(a)



The distribution of number of transfers appear to be asymmetrical.

(b)



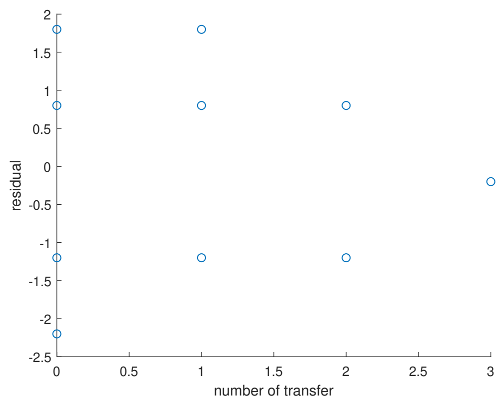
The distribution of number of transfers appear to be asymmetrical.

(c)

$$e_i = \begin{pmatrix} 1.8 \\ -1.2 \\ -1.2 \\ 1.8 \\ -0.2 \\ -1.2 \\ -2.2 \\ 0.8 \\ 0.8 \\ 0.8 \end{pmatrix} \quad \begin{array}{r|c} \text{-2} & 2 \\ \text{-1} & 222 \\ \text{-0} & 2 \\ 0 & 888 \\ 1 & 88 \end{array}$$

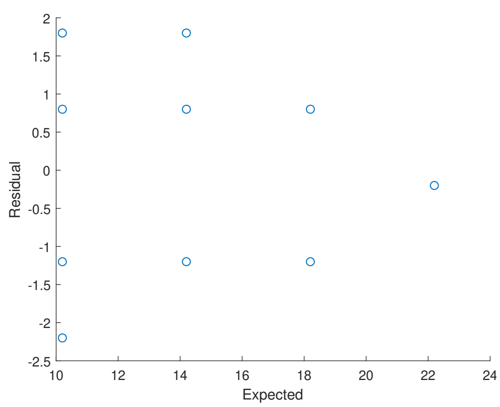
The stem and leaf plot has | where the decimal point locates. This plot provides the frequency of each residual value.

(d)



As the number of transfers increases, the residual decreases. In conclusion, the linear regression function has a good fit here.

(e)

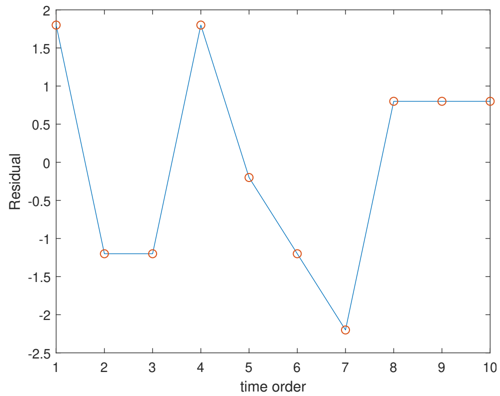


As the number of transfers increases, the residual decreases. In conclusion, the linear regression function has a good fit here.

The critical value for n=10 = 0.879 < 0.949

Since the observed coefficient exceeds this level, we conclude that the distribution of the error terms does not depart substantially from a normal distribution.

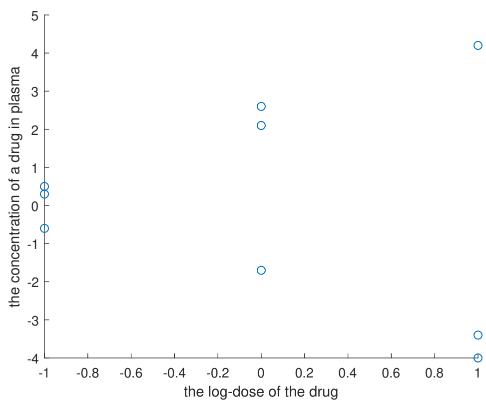
(f)



As the number of transfers increases, the residual decreases. In conclusion, the linear regression function has a good fit here.

Problem 3.11

(a)



As the number of transfers increases, the residual decreases. In conclusion, the linear regression function has a good fit here.

Problem 3.23

$$\text{Full model:} Y_{ij} = \overline{Y_j} + \epsilon_{ij}, df_F = 10$$

$$\text{Reduced model:} Y_{ij} = \beta_1 X_j + \epsilon ij, df_R = 18$$

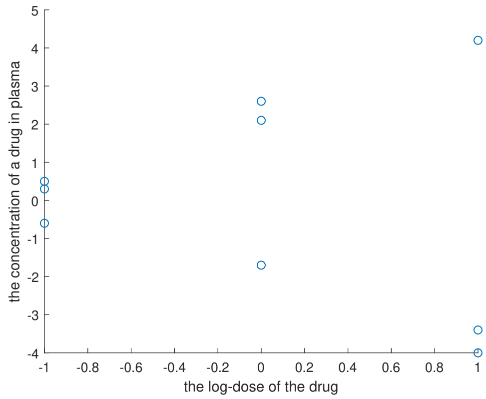
Problem 3.24

(a)

$$\overline{X} = 9.25, \overline{Y} = 70.25.$$

$$\beta_1 = E[b_1] = \frac{7}{3}, \beta_0 = E[b_0] = \frac{146}{3}$$

$$\Rightarrow Y_i = \frac{146}{3} + \frac{7}{3} X_i.$$



There is a peak of blood pressure in the middle of ages between 5 and 13. But there is case which does not follow.

(b)

$$\overline{X} = 8.8571, \overline{Y} = 67.4286.$$

$$\beta_1 = E[b_1] = \frac{167}{103}, \beta_0 = E[b_0] = \frac{5466}{103}$$

$$\Rightarrow Y_i = \frac{5466}{103} + \frac{167}{103}X_i.$$

The regression funtion is more accurate.

(c)

$$\overline{Y_h} = 72.5243, t(0.995; 5) = 4.032$$

$$s\{\text{pred}\} = \sqrt{6.9981(\frac{8}{7} + \frac{(12 - 8.8571)^2}{58.8571})} = 3.0286$$

$$\Rightarrow 72.5243 - 4.032 \cdot 3.0286 \leq Y_h \leq 72.5243 + 4.032 \cdot 3.0286 \Rightarrow 60.3130 \leq Y_h \leq 84.7356$$

Y_7 falls outside this prediction interval meaning this observation is an outlier.

Problem 4.2

The family confidence coefficient for this set ensures at least 90% that intercept and slope fall into interval estimates.

Problem 4.4

(a)Opposite directions, since $\overline{X} = 1 > 0$ and $cov(b_0, b_1) < 0$.

(b)

$$B = t(.9975; 8) = 3.8325$$

$$s\{b_0\} = 0.663 \Rightarrow 10.2 - 0.663 \cdot 3.8325 \leq \beta_0 \leq 10.2 + 0.663 \cdot 3.8325 \Rightarrow 7.6491 \leq \beta_0 \leq 12.7409$$

The family confidence coefficient for this set ensures at least 99% that the increase of number of ampules found to be broken upon arrival lies somewhere between7.6491 and 12.7409 when number of times the carton was transfered is 0.

$$s\{b_1\} = 0.469 \Rightarrow 4 - 3.8325 \cdot 0.469 \leq \beta_1 \leq 4 + 3.8325 \cdot 0.469 \Rightarrow 2.2026 \leq \beta_1 \leq 5.7974.$$

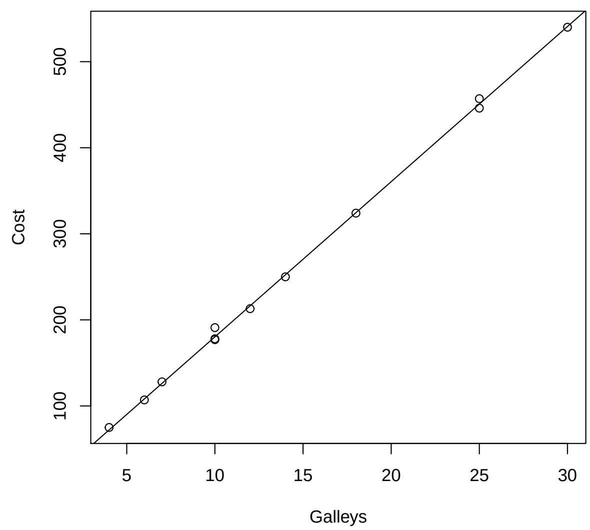
The family confidence coefficient for this set ensures at least 99% that the increase of number of ampules found to be broken upon arrival lies somewhere between 2.2026 and 5.7974 per unit of number of times the carton was transfered.

Problem 4.12

(a)

$$\overline{Y_h} = 18.03X$$

(b)



The regression line has a good fit.

(c)Alternatives:

$$H_0 : E[Y] = \beta_1 = 17.5$$

$$H_a : E[Y] \neq \beta_1 = 17.5$$

$$17.8123 \leq E[Y_h] \leq 19.2444$$

Decision rule:

$$\text{If } 17.8123 \leq \beta_1 \leq 19.2444, \text{ conclude } H_0$$

$$\text{If } \beta_1 < 17.8123 \text{ or } \beta_1 > 19.2444, \text{ conclude } H_a$$

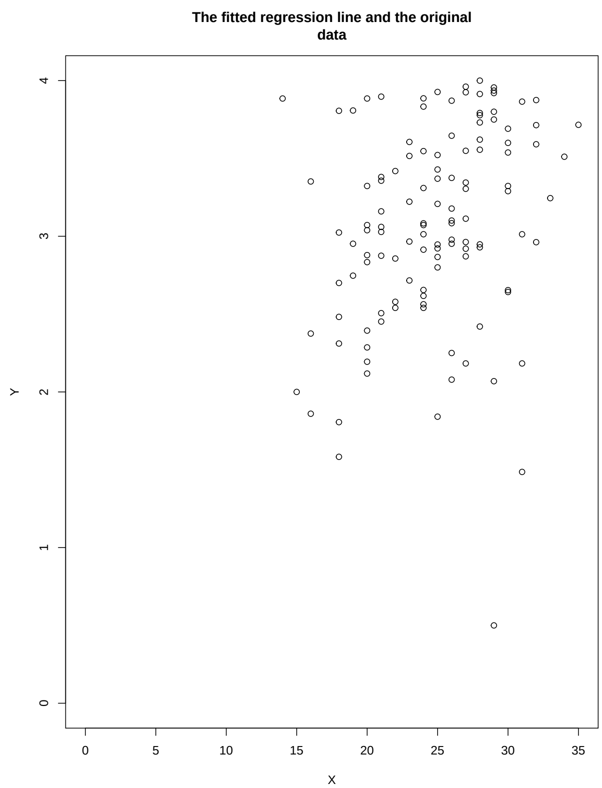
Since $17.5 < 17.8123$, conclude H_a. (d)

$$\hat{Y_h} = 180.283, s\{\text{pred}\} = 4.5068$$

$$167.8441 \leq Y_h \leq 192.722$$

Problem 4.15

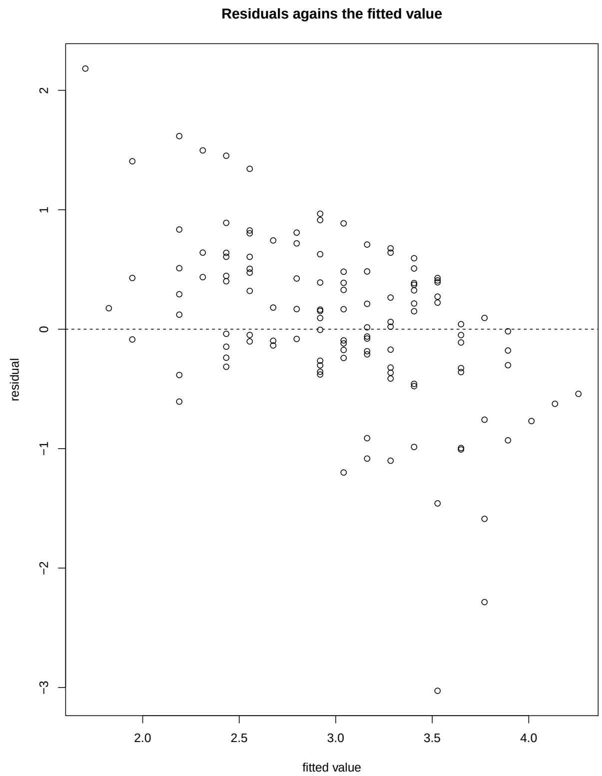
(a)



The regression line does not have good fit since the points do not spread randomly on two sides of the line.

(b)

$$\sum e_i = 7.9715 \neq 0$$



Residuals take more positive values when $\hat{Y}$is small compared with when is large. There is a decreasing linear trend between residuals and $\hat{Y}$.

(c)Alternatives:

$$H_0 : Y = \beta_1 X + \epsilon$$

$$H_a : Y = \beta_0 + \beta_1 X + \epsilon$$

Decisions:

$$\text{If } p < 0.005, \text{ conclude } H_a$$

$$\text{If } p \geq 0.005, \text{ conclude } H_0$$

$$F = 43.4 \Rightarrow p = 1.304 \cdot 10^{-9} < 0.005$$

We conclude that regression model with intercept is more appropriate.

Problem 4.23

Let

$$f(b_1) = \sum (y_i - b_1 x_i)^2 = \sum (y_i^2 + b_1^2 x_i^2 - 2b_1 x_i y_i)$$

$$f'(b_1) = \sum (2b_1 x_i^2 - 2x_i y_i) = -2 \sum x_i (y_i - b_i x_i) = -2 \sum x_i e_i$$

In order to get the least square regression line, $f'(b_1) = 0 \Rightarrow \sum x_i e_i = 0$.