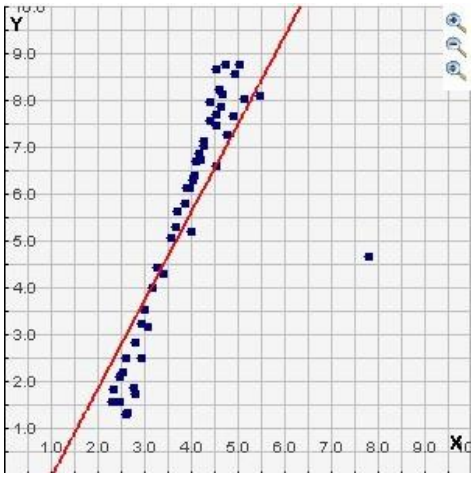
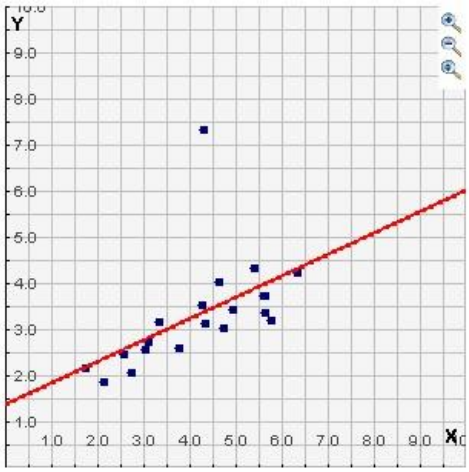
STAT 3008 Solution of Homework 1

1.

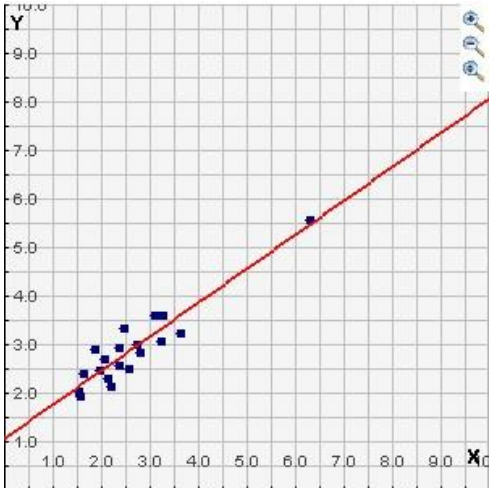
a) x-separated point and outlier



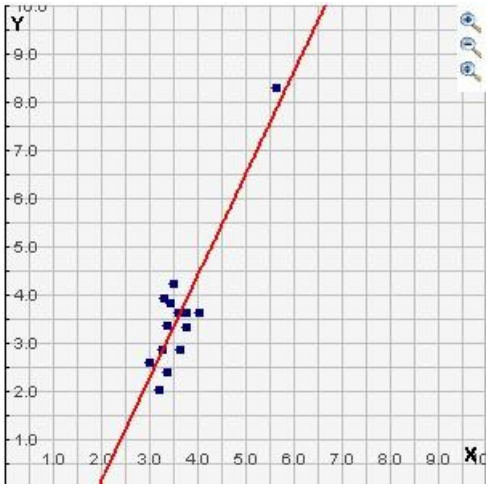
b) y-separated point and outlier



c) x-separated point and NOT outlier

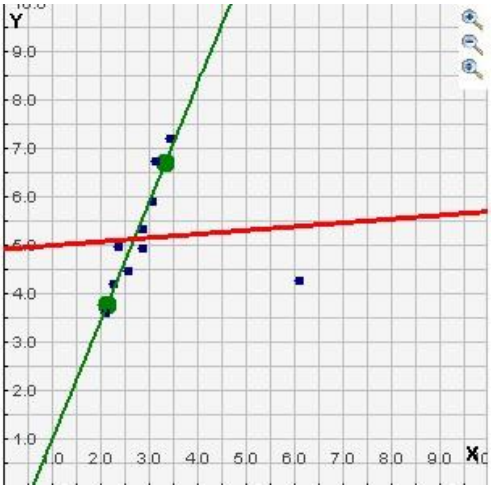


d) y-separated point and NOT outlier

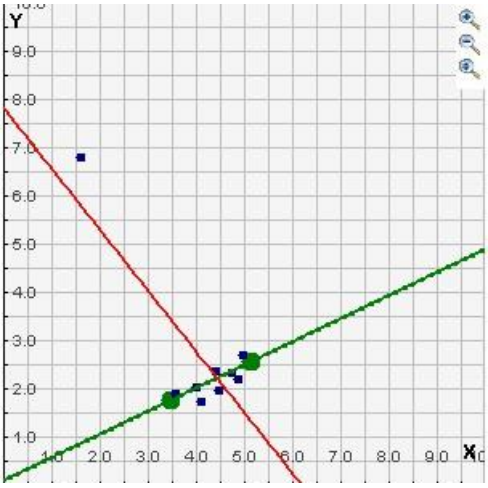


In the following plots, the green line is the regression line without the specific point, and the red line is the one with the specific point.

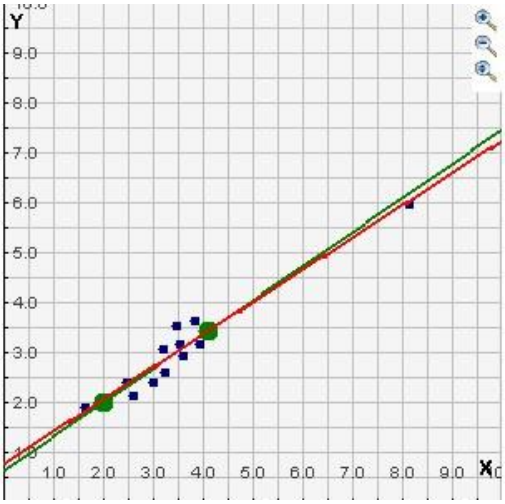
e) x-separated and influential point



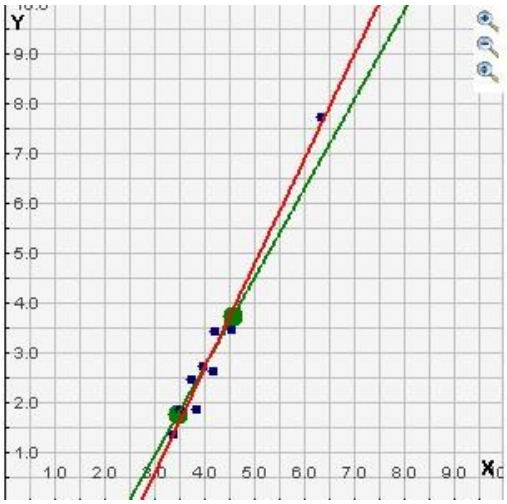
f) y-separated and influential point



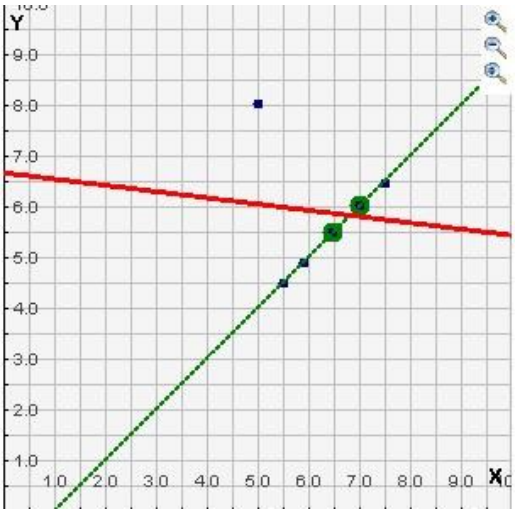
g) x-separated and NOT influential point



h) y-separated and NOT influential point

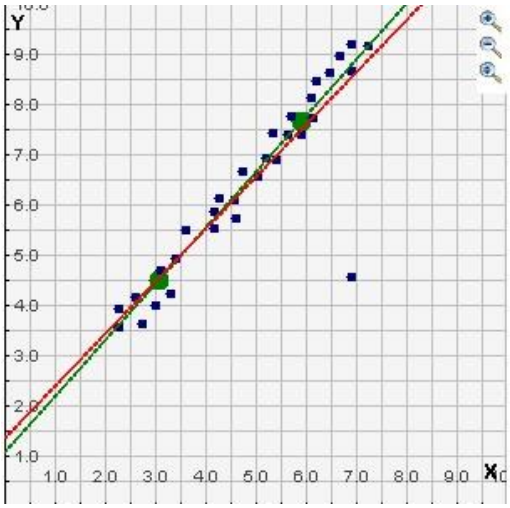


i) Outlier and influential point

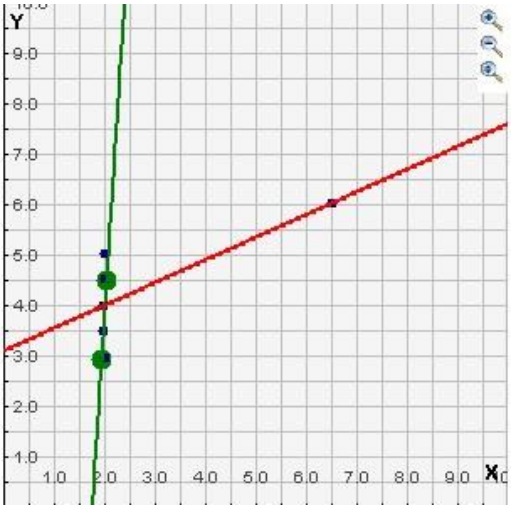


Points:
(7.5, 6.5), (5.5, 4.5), (6.0, 5.0), (6.5, 5.5), (7.0, 6.0)
Influential point but not outlier: (5.0, 8.0)

j) Outlier but NOT influential point



k) Influential point but NOT Outlier



Points:
(2.0, 4.0), (2.0, 3.5), (2.0, 5.0), (2.0, 4.5), (2.0, 3.0)
Influential point but not outlier: (6.5, 6.0)

2.

i)

$\mathrm{RSS}(\beta) = \sum_{i=1}^{n}(y_i - \beta x_i^3)^2$ and $\frac{\mathrm{d}\,\mathrm{RSS}(\beta)}{d\beta} = -2\sum_{i=1}^{n}(y_i - \beta x_i^3)\,x_i^3$

Set $\frac{\mathrm{d}\,\mathrm{RSS}(\beta)}{d\beta} = 0$, we have

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i^3 y_i}{\sum_{j=1}^{n} x_j^6}$$

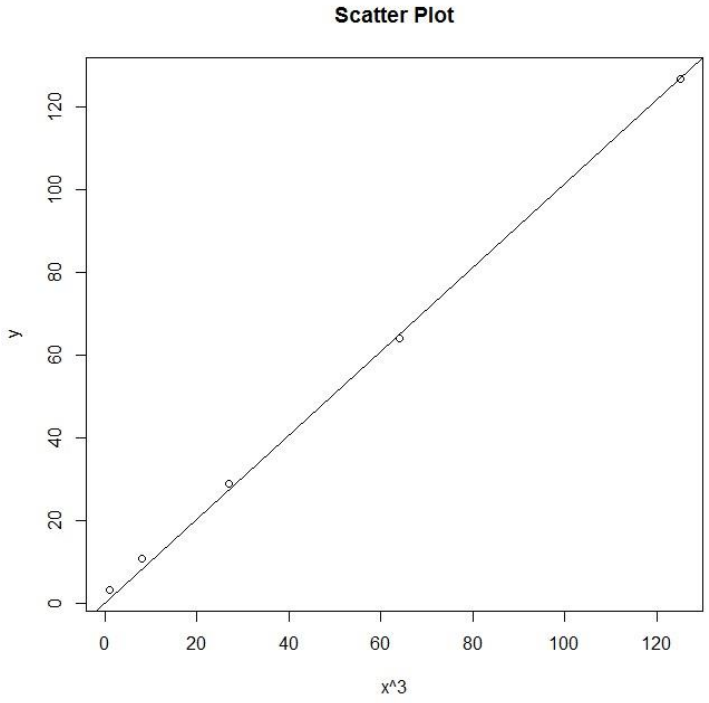
$$\widehat{\sigma^2} = \frac{\mathrm{RSS}(\hat{\beta})}{\mathrm{n}-1} = \frac{\sum_{i=1}^{n}(y_i - \hat{\beta}x_i^3)^2}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \frac{\sum_{s=1}^{n} x_s^3 y_s}{\sum_{j=1}^{n} x_j^6} x_i^3)^2}{n-1}$$

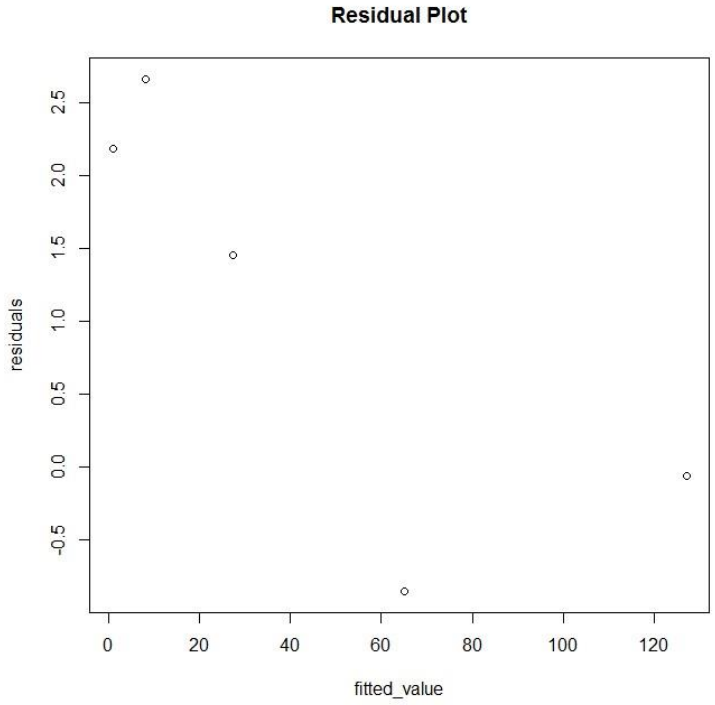
ii)

$$\hat{\beta} = \frac{\sum_{i=1}^{5} x_i^3 y_i}{\sum_{j=1}^{5} x_j^6} = 1.01651$$

$$\widehat{\sigma^2} = \frac{\sum_{i=1}^{5}(y_i - 1.01651x_i^3)^2}{5-1} = 3.6845$$

iii)





iv)

$\bar{X} = 3$ and $\bar{Y} = 46.82$,

$\hat{\beta}\bar{X}^3 = 1.01651 * 3^3 = 27.44577 \neq 46.82 = \bar{Y}$

The fitted regression line does not pass through $(\bar{X},\ \bar{Y})$.

$\sum_{i=1}^{n} \widehat{e_i} = \sum_{i=1}^{5}(y_i - \hat{\beta}x_i^3) = 5.38525 \neq 0$

The sum of residuals does not equal to zero.

v)

a) $\text{SXX} = \sum_i (x_i - \bar{x})^2 = 10$, $\text{SXY} = \sum_i (x_i - \bar{x})(y_i - \bar{y}) = 301$,

$\text{SYY} = \sum_i (y_i - \bar{y})^2 = 10252.17$,

$\therefore \widehat{\alpha_1} = \frac{SXY}{SXX} = 30.1$, $\text{se}(\widehat{\alpha_1}) = \frac{\hat{\sigma}}{\sqrt{SXX}} = \frac{\sqrt{\left(SYY - \frac{SXY^2}{SXX}\right)/3}}{\sqrt{SXX}} = 6.3036.$

$\therefore$ the 95% confidence interval for α_1 is

$[\widehat{\alpha_1} - t(0.025,3)se(\widehat{\alpha_1}), \widehat{\alpha_1} + t(0.025,3)se(\widehat{\alpha_1})] = [10.0391,50.1609].$

The p-value of testing α_1 against $\alpha_1 \neq 0$ is

$\mathrm{P}\left(|t| > \frac{\widehat{\alpha_1}}{se(\widehat{\alpha_1})}\right) = P(|t| > 4.7750) = 0.0175$, where t is a t distributed random variable.

b) Compare the model $y_i = \alpha_0 + e_i$ with the model $y_i = \alpha_0 + \alpha_1 x + e_i$.

$\mathrm{NH}: y_i = \alpha_0 + e_i \leftrightarrow AH: y_i = \alpha_0 + \alpha_1 x + e_i$.

$\mathrm{P}\left(F > \frac{SSreg/1}{RSS/3}\right) = P\left(F > \frac{SXY^2/SXX}{(SYY - SXY^2/SXX)/3}\right) = P(F > 22.8009) = 0.0175$, where F is a F distributed random variable.

So the p-value of testing NH against AH is 0.0175, which is smaller than 0.05. Hence we should reject the null hypothesis.