36-401 Modern Regression HW #8 Solutions

DUE: 11/10/2017 at 3PM

Problem 1 [25 points]

(a)

This is still a linear regression model—the simplest possible one. That being the case, the solution we derived before

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

still holds. The design matrix is

$$X = \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix},$$

so

$$\begin{aligned} \widehat{\beta} &= \underbrace{(X^T X)^{-1}}_{1/n} \underbrace{X^T Y}_{=\sum Y_i} \\ &= \overline{Y}. \end{aligned}$$

(b)

$$\begin{aligned} H &= X(X^T X)^{-1} X^T \\ &= \begin{pmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{pmatrix} (n)^{-1} \begin{pmatrix} 1 & 1 & \cdots & 1 \end{pmatrix} \\ &= \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & & \frac{1}{n} \\ \vdots & & \ddots & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \end{aligned}$$

So

$$h_{ii} = \frac{1}{n} \quad \text{for all } i = 1, \ldots, n.$$

(c)

$$\begin{aligned}\widehat{Y} &= HY \\ &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^{n} Y_i \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} Y_i \end{pmatrix} \\ &= \begin{pmatrix} \overline{Y} \\ \vdots \\ \overline{Y} \end{pmatrix}\end{aligned}$$

(d)

$$\begin{aligned} t_i &= \frac{Y_i - \widehat{Y}_{i(-i)}}{s_i} \\ &= \frac{Y_i - \frac{1}{n-1}\sum\limits_{\substack{1\le j\le n \\ j\neq i}} Y_i}{\sqrt{\mathrm{Var}(Y_i - \widehat{Y}_{i(-i)})}} \\ &= \frac{Y_i - \frac{1}{n-1}\left(\sum_{j=1}^{n} Y_j - Y_i\right)}{\sqrt{\mathrm{Var}(Y_i) + \mathrm{Var}(\widehat{Y}_{i(-i)})}} \\ &= \frac{Y_i - \frac{1}{n-1}\left(n\overline{Y} - Y_i\right)}{\sqrt{\sigma^2 + \frac{\sigma^2}{n-1}}} \\ &= \frac{Y_i - \frac{1}{n-1}\left(n\overline{Y} - Y_i\right)}{\sqrt{\frac{n\sigma^2}{n-1}}} \\ &= \frac{\frac{n}{n-1}(Y_i - \overline{Y})}{\sqrt{\frac{n\sigma^2}{n-1}}} \end{aligned}$$

(e)

$$\begin{aligned} \lim_{Y_i\to\infty} t_i &= \lim_{Y_i\to\infty} \frac{\frac{n}{n-1}(Y_i - \overline{Y})}{\sqrt{\frac{n\sigma^2}{n-1}}} \\ &= \frac{\frac{n}{n-1}(\lim_{Y_i\to\infty} Y_i - \overline{Y})}{\sqrt{\frac{n\sigma^2}{n-1}}} \\ &= \infty \end{aligned}$$

Problem 2 [20 points]

$$H = HH$$

$$\begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & & \ddots & \vdots \\ h_{n1} & \cdots & \cdots & h_{nn} \end{pmatrix} = \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & & \ddots & \vdots \\ h_{n1} & \cdots & \cdots & h_{nn} \end{pmatrix} \begin{pmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & & \ddots & \vdots \\ h_{n1} & \cdots & \cdots & h_{nn} \end{pmatrix}$$

So for each diagonal element of H we have,

$$h_{ii} = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2.$$

We have expressed h_{ii} as a sum of squares, so

$$h_{ii} \geq 0.$$

Furthermore,

$$h_{ii} \geq h_{ii}^2,$$

so

$$h_{ii} \leq 1.$$

Problem 3 [20 points]

Comments omitted. One of the points in data set four produces NaN's because it has leverage 1. See equation (2) of Lecture Notes 20.

```
attach(anscombe)
names(anscombe)
```

```
## [1] "x1" "x2" "x3" "x4" "y1" "y2" "y3" "y4"
```

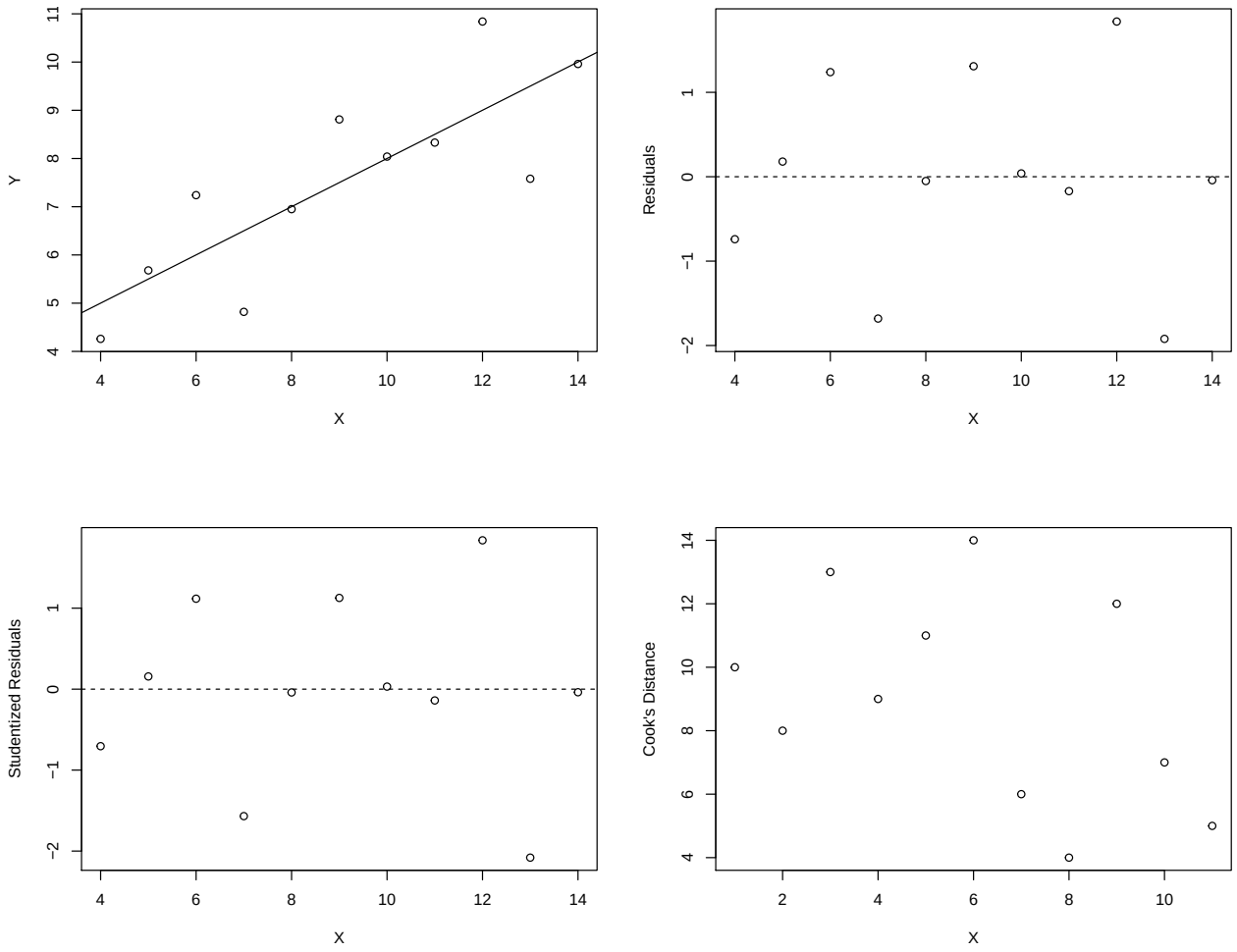


Figure 1: Data Set 1

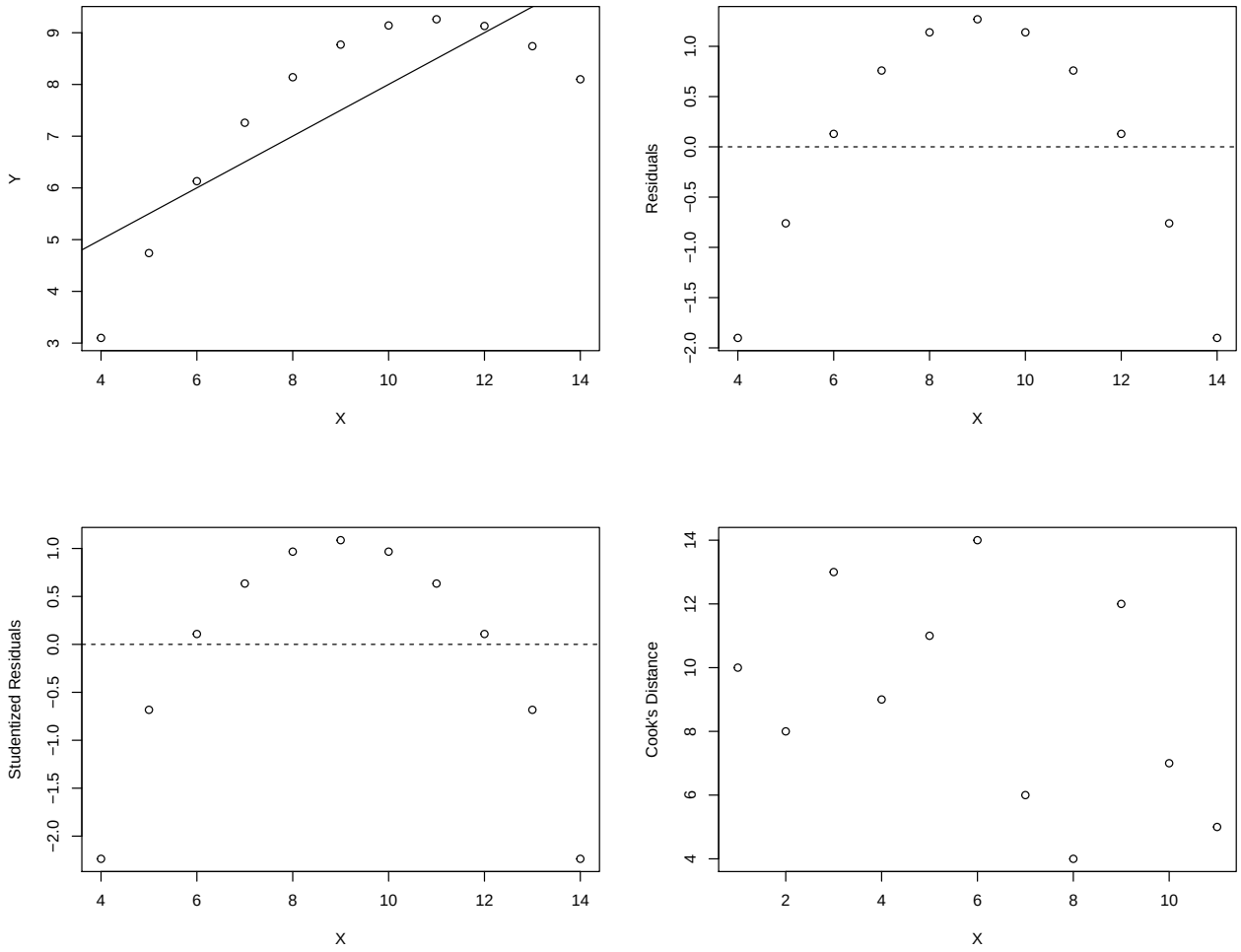


Figure 2: Data Set 2

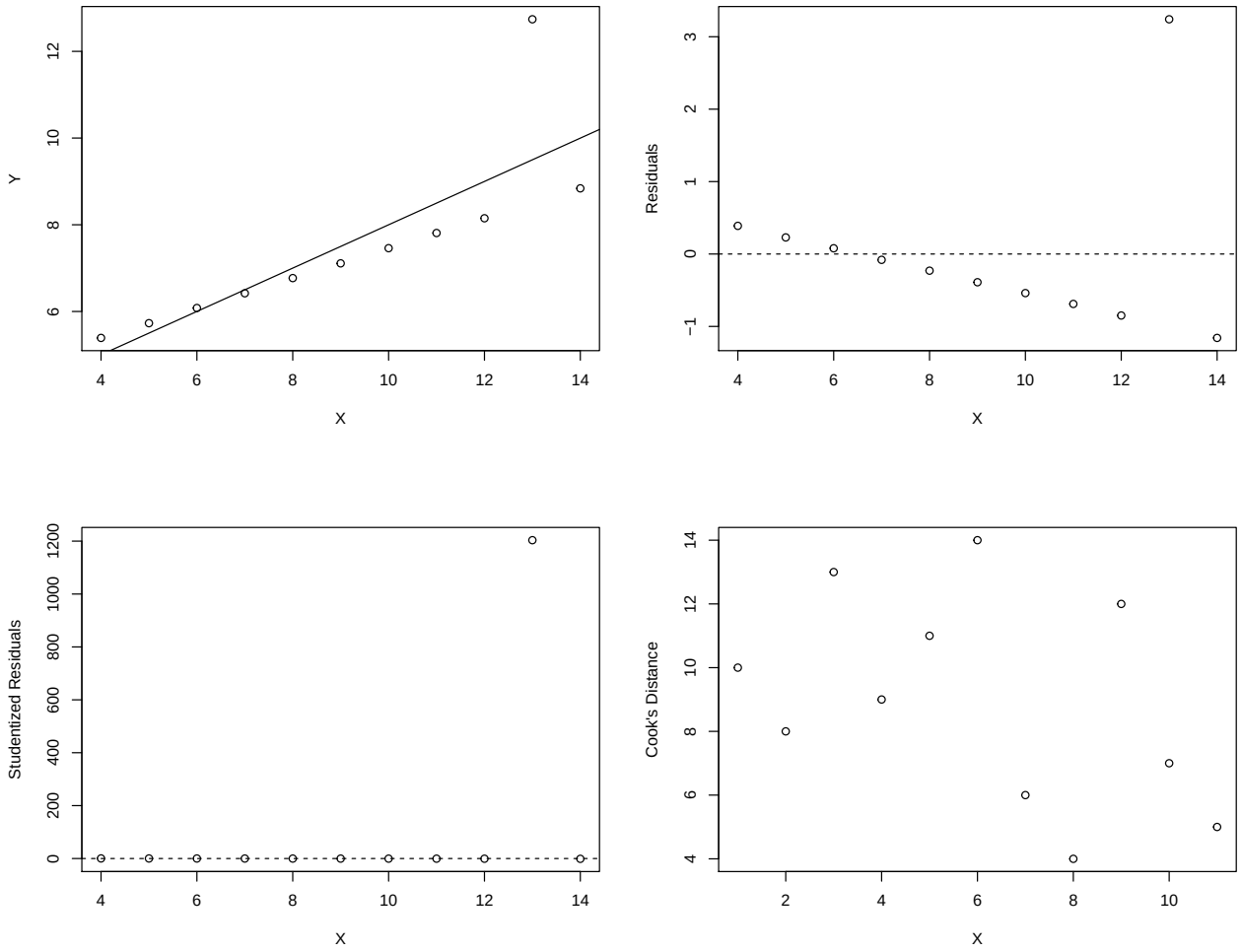


Figure 3: Data Set 3

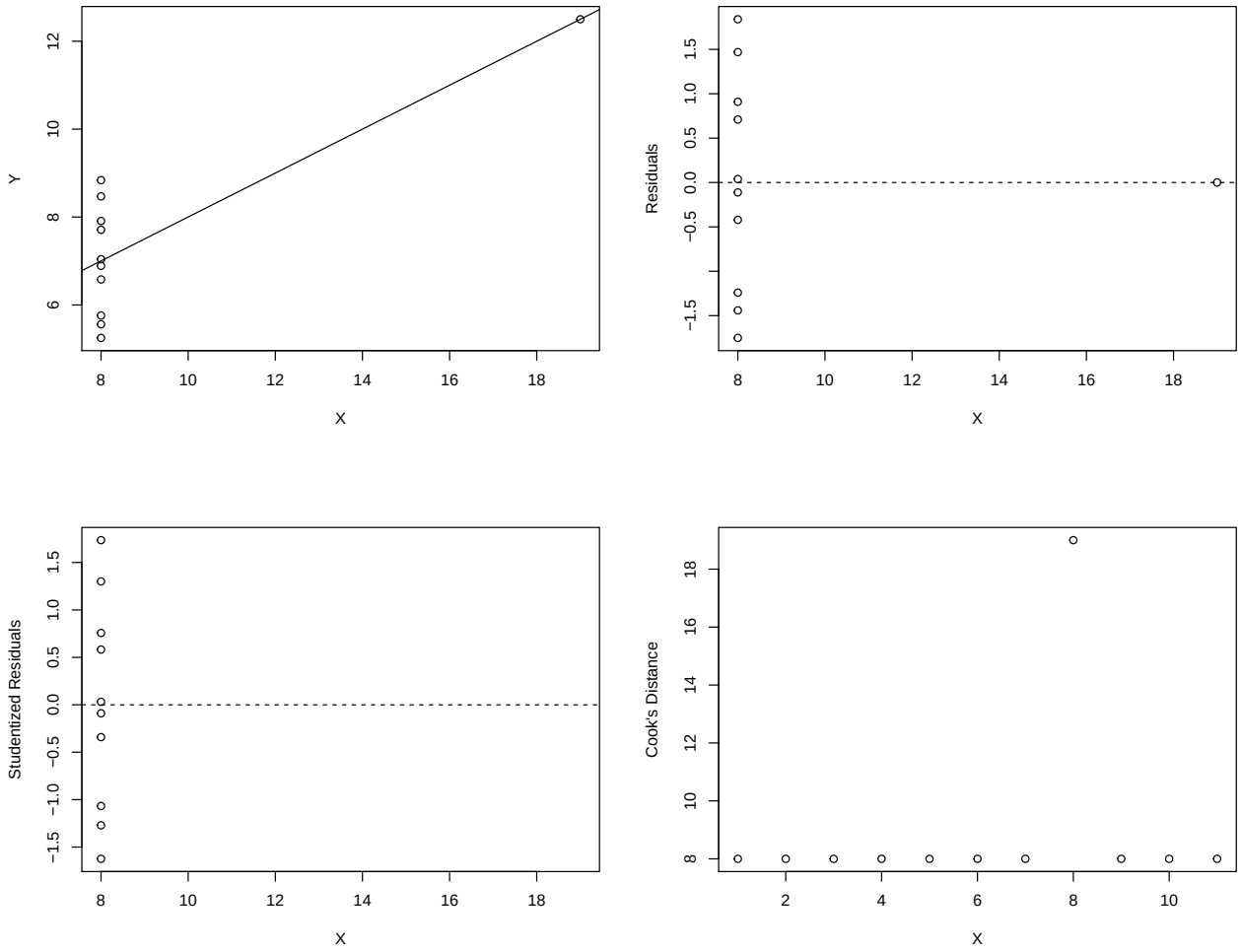
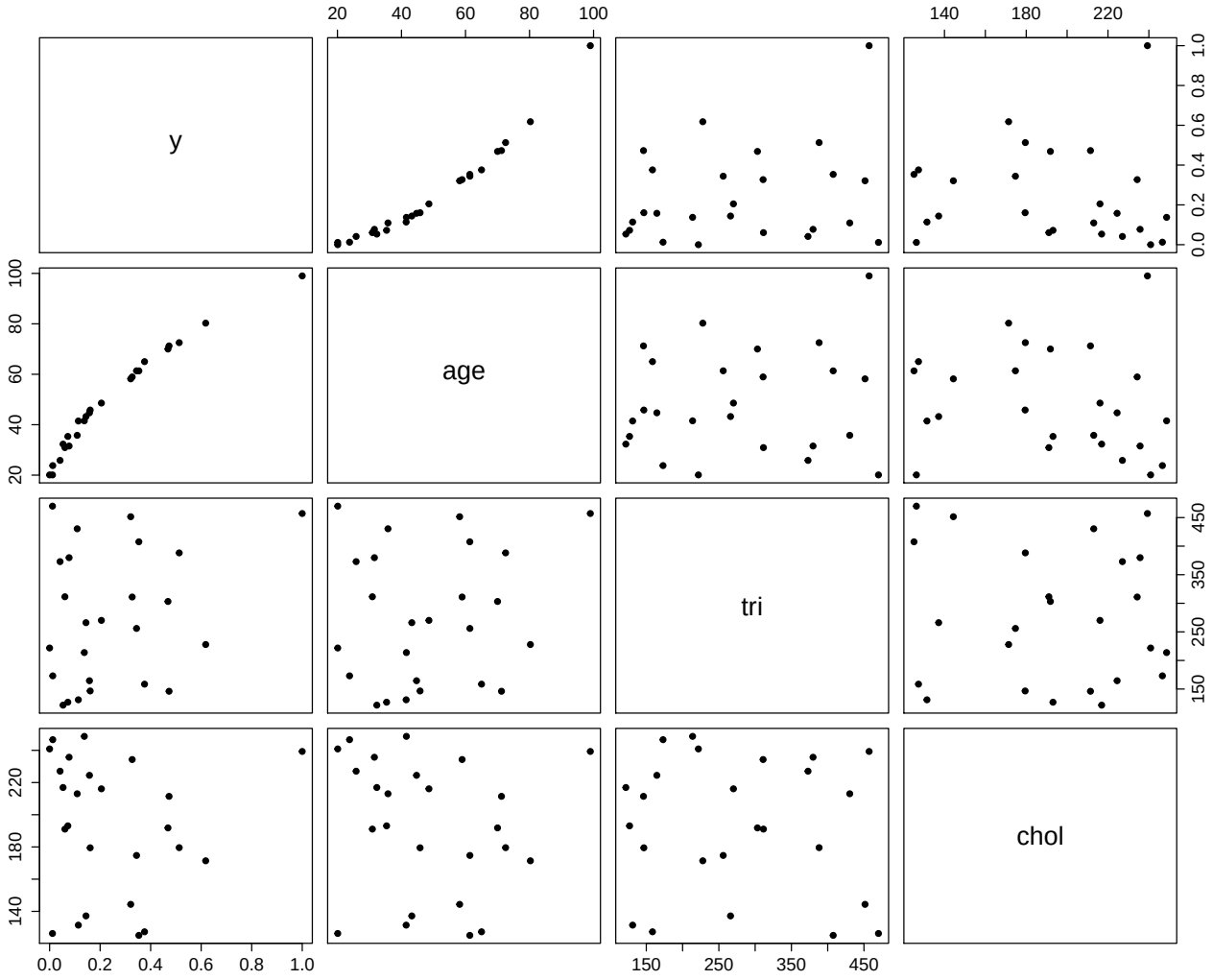


Figure 4: Data Set 4

Problem 4 [25 points]

Discussions omitted

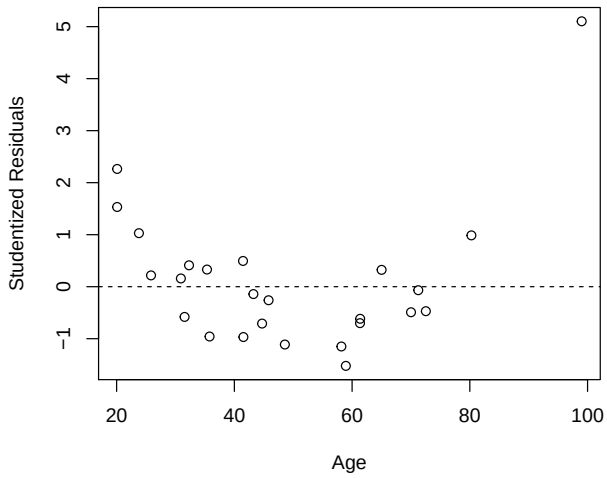
(i)

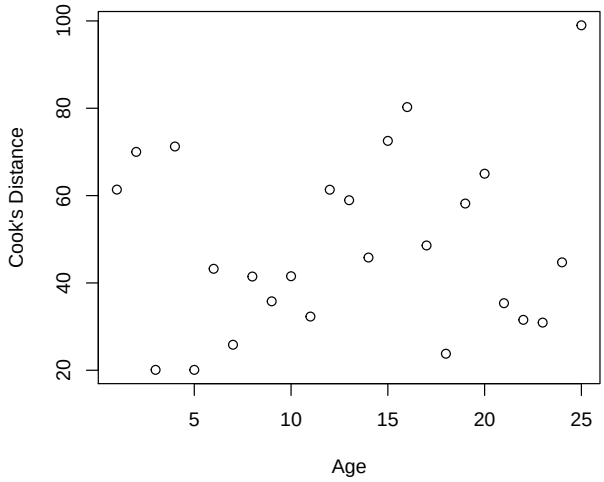


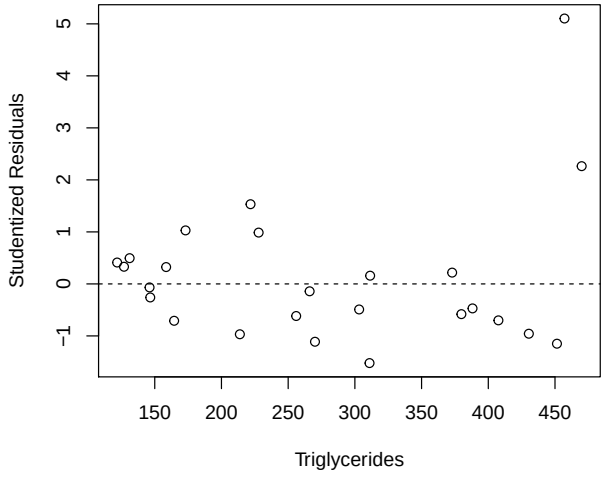
(ii)

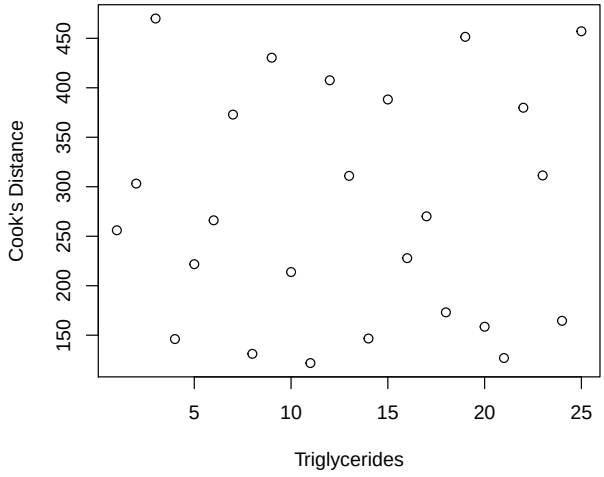
```
model <- lm(y ~ age + tri + chol, data = data)
```

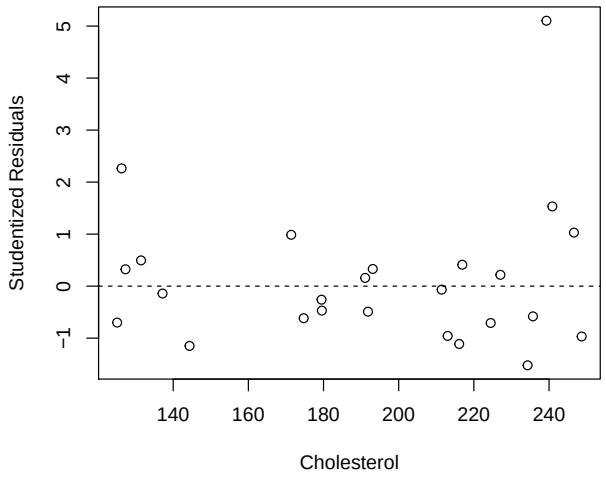
(iii)

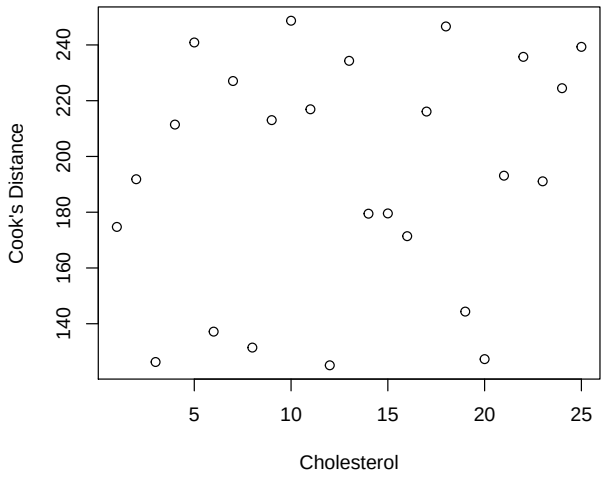

Studentized Residuals
Age


Cook's Distance
Age


Studentized Residuals
Triglycerides

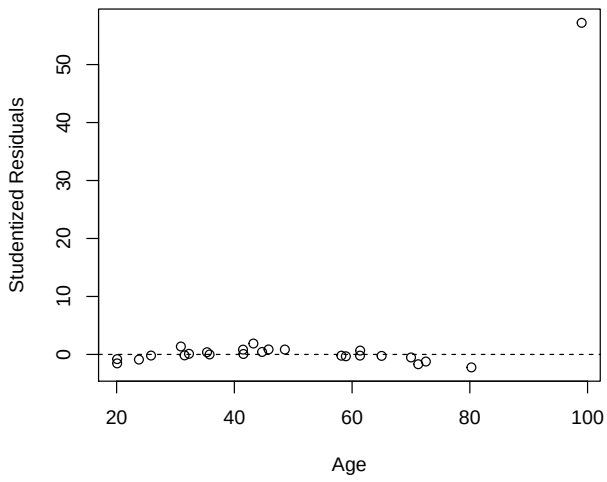

Cook's Distance
Triglycerides

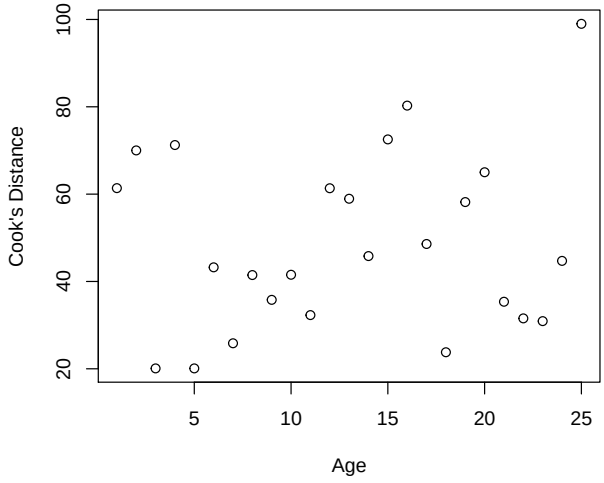

Studentized Residuals
Cholesterol

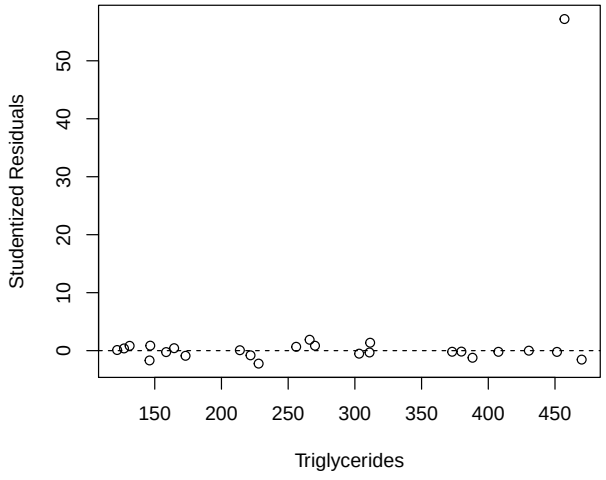

Cook's Distance
Cholesterol

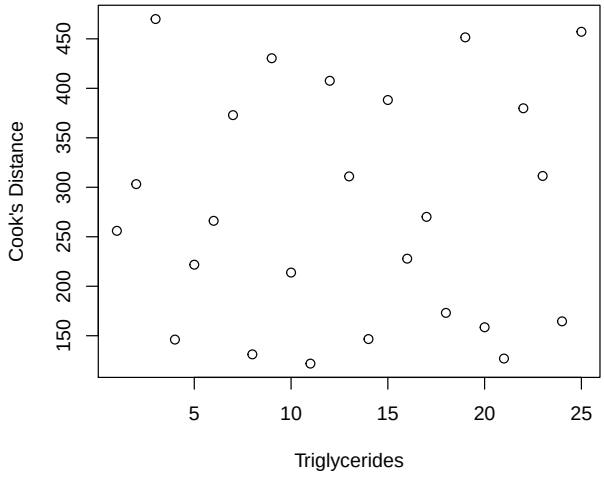
(iv)

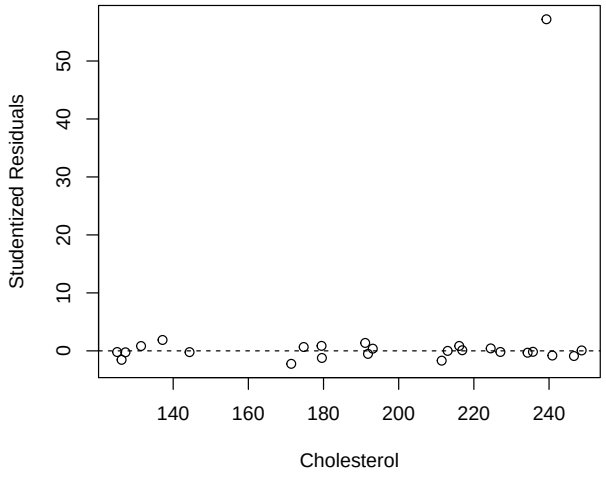
```
model <- lm(y ~ age + tri + chol + I(age^2) + I(tri^2) + I(chol^2), data = data)
```

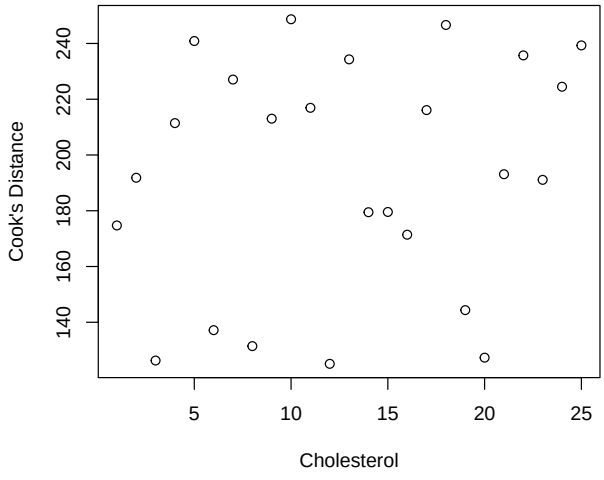












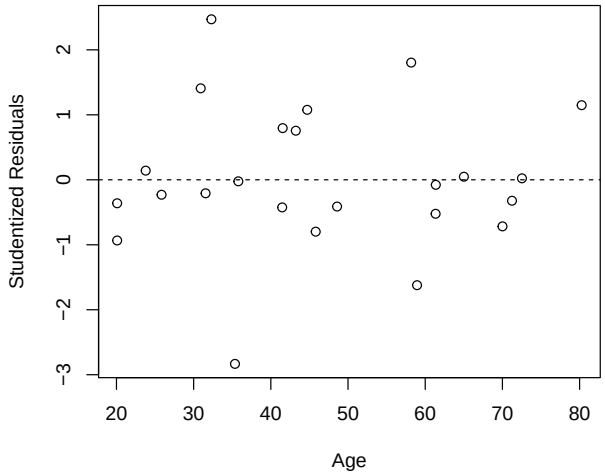
```r
which(rstudent(model) > 30)
```

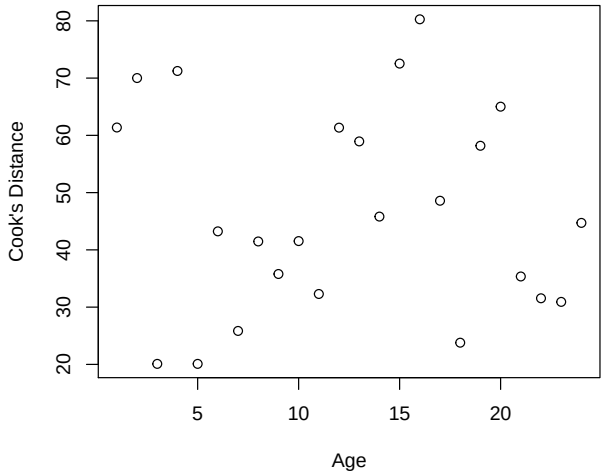
```
## 25
## 25
```

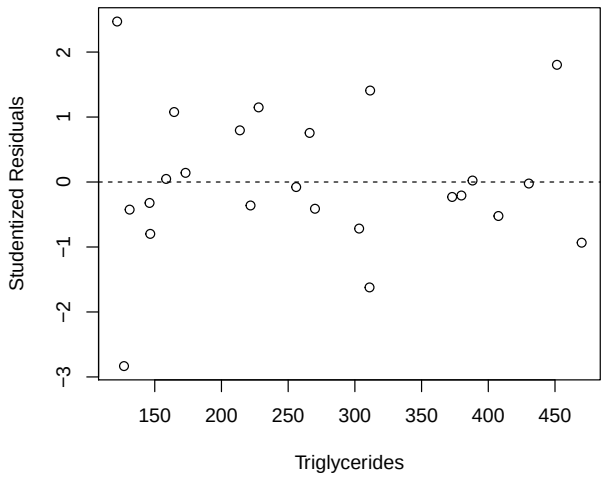
(v)

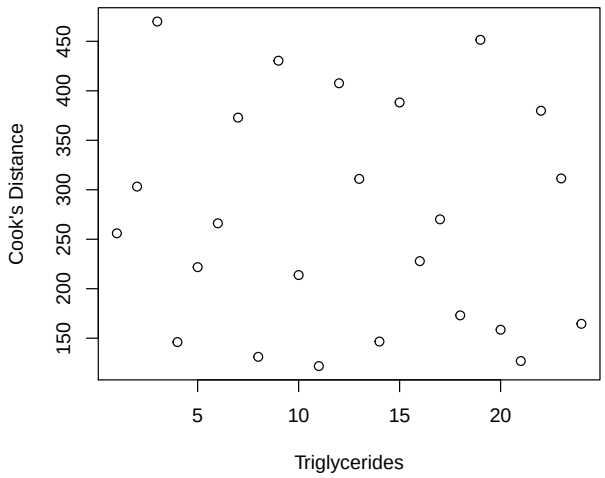
```r
data2 <- data[-25,]
```

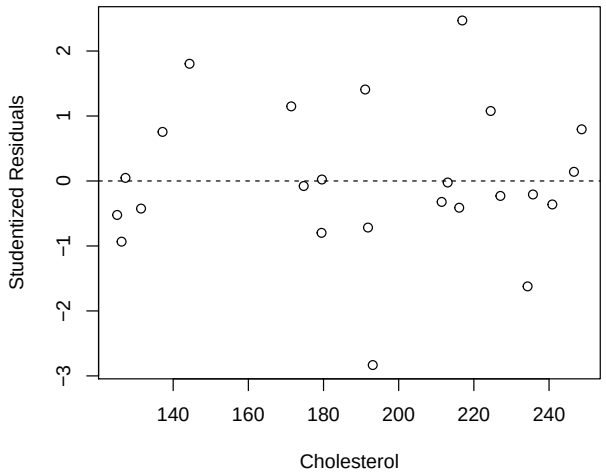
```r
model <- lm(y ~ age + tri + chol + I(age^2) + I(tri^2) + I(chol^2), data = data2)
```

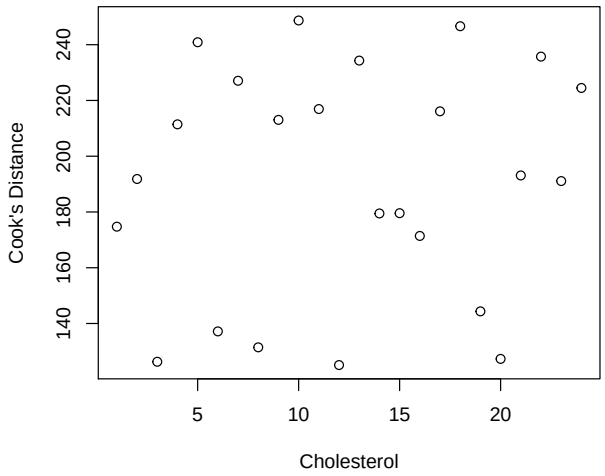













```r
library(knitr)
kable(summary(model)$coefficients)
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.0953012	0.0001105	-862.4135881	0.0000000
age	0.0000417	0.0000015	28.0644703	0.0000000
tri	0.0001036	0.0000003	385.7563999	0.0000000
chol	0.0001242	0.0000012	103.1582206	0.0000000
I(age^2)	0.0001032	0.0000000	6932.5827948	0.0000000
I(tri^2)	0.0000000	0.0000000	-1.1114373	0.2818524
I(chol^2)	0.0000000	0.0000000	-0.2724321	0.7885711

```r
kable(confint(model, parm = 2:6, level = 1 - 0.05 / 6))
```

	0.417 %	99.583 %
age	0.0000373	0.0000462
tri	0.0001028	0.0001044
chol	0.0001206	0.0001278
I(age^2)	0.0001032	0.0001033
I(tri^2)	0.0000000	0.0000000

Problem 5 [10 points]

```
d = read.table("http://stat.cmu.edu/~larry/secretdata.txt")
y = d[,1]
x1 = d[,2]
x2 = d[,3]
x3 = d[,4]
x4 = d[,5]
```

(a)

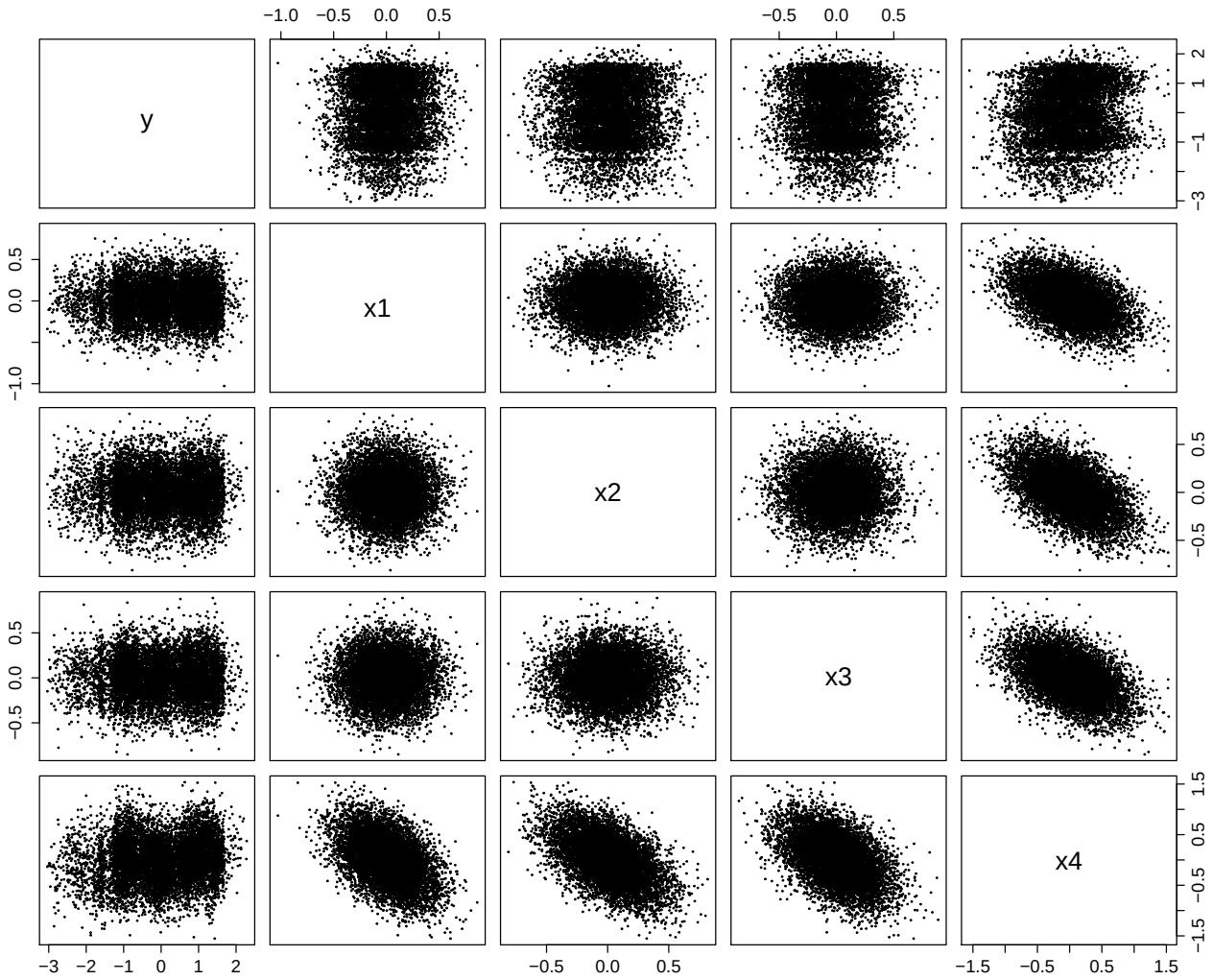


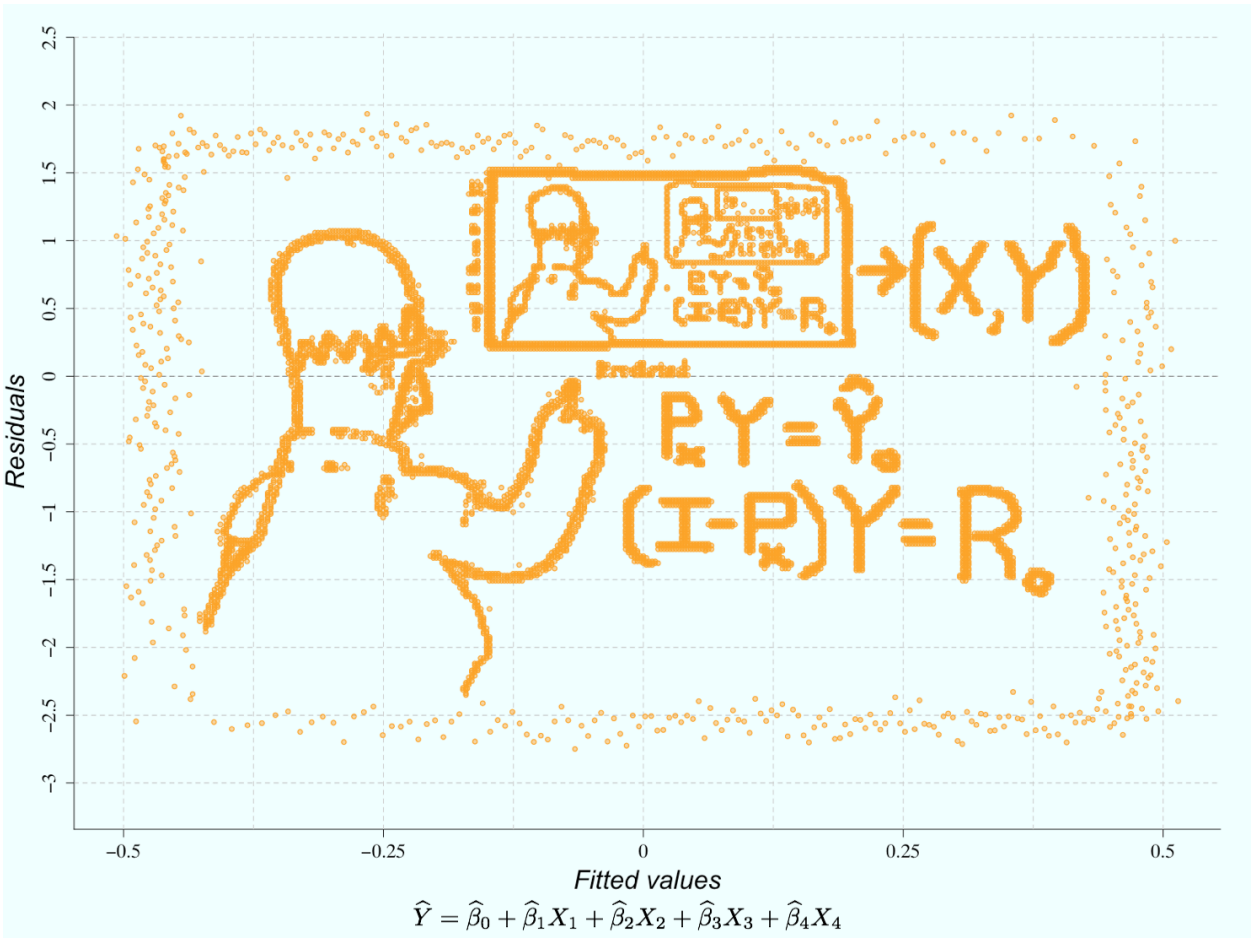
Figure 5: Secret Data

(b)

```
model5 <- lm(y ~ x1 + x2 + x3 + x4)
```

Summarize the fitted model.


Linear Regression Residuals vs. Predictors
Residuals
X_1
X_2
X_3
X_4


Residuals
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
−3
−0.5
−0.25
0
0.25
0.5
Fitted values
$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \widehat{\beta}_3 X_3 + \widehat{\beta}_4 X_4$
(X,Y)
$P_X Y = \hat{Y}$
$(I - P_X) Y = R$

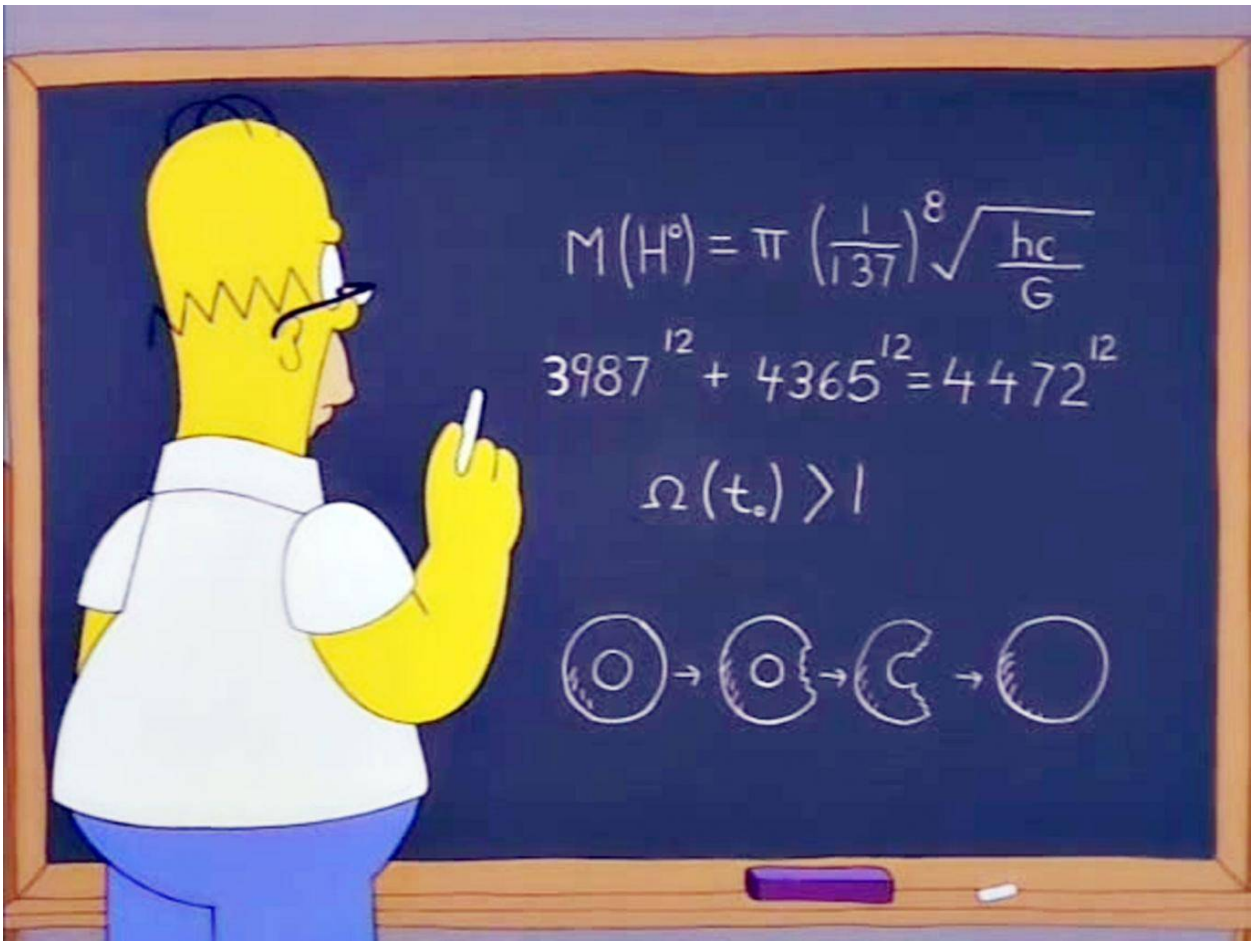


Figure 6: The above residual plot is a spoof of this scene from The Simpsons

Appendix

```
quartz(height = 12, width = 16)
par(mar = c(4,4.5,2,2) + 0.1)
par(oma = c(1.5,1,2,1) + 0.1)
par(mfrow=c(2,2))
par(bg = "azure")

out <- lm(y ~ x1+x2+x3+x4)

plot(x1, residuals(out), col = NA, pch = 19, axes = FALSE, ylab = "Residuals",
     font.lab = 3, xlab = "", xlim = range(x4), cex.lab=2)
axis(side = 1, at = seq(-2.5,2.5,0.25), labels = as.character(seq(-2.5,2.5,0.25)),
     font = 5, cex.axis = 1.25)
axis(side = 2, at = seq(-3,2.5,0.5), labels = as.character(seq(-3,2.5,0.5)),
     font = 5, cex.axis = 1.25)
abline(h = seq(-3,2.5,0.5), col = "gray75", lty = 2)
abline(v = seq(-2.5,2.5,0.25), col = "gray80", lty = 2)
```

```
abline(0,0, lty = 2, col = "gray45")
points(x1, residuals(out), col = addTrans("orange",120), pch = 19)
points(x1, residuals(out), col = "orange")
panel.smooth(x1, residuals(out), col = "orange",cex = 1, lwd = 1.2,
             col.smooth = "seagreen", span = 2/3, iter = 3)

plot(x2, residuals(out), col = NA, pch = 19, axes = FALSE, ylab = "Residuals",
     font.lab = 3, xlab = "", xlim = range(x4), cex.lab=2)
axis(side = 1, at = seq(-2.5,2.5,0.25), labels = as.character(seq(-2.5,2.5,0.25)),
     font = 5, cex.axis = 1.25)
axis(side = 2, at = seq(-3,2.5,0.5), labels = as.character(seq(-3,2.5,0.5)),
     font = 5, cex.axis = 1.25)
abline(h = seq(-3,2.5,0.5), col = "gray75", lty = 2)
abline(v = seq(-2.5,2.5,0.25), col = "gray80", lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(x2, residuals(out), col = addTrans("orange",120), pch = 19)
points(x2, residuals(out), col = "orange")
panel.smooth(x2, residuals(out), col = "orange",cex = 1, lwd = 1.2,
             col.smooth = "seagreen", span = 2/3, iter = 3)

plot(x3, residuals(out), col = NA, pch = 19, axes = FALSE, ylab = "Residuals",
     font.lab = 3, xlab = "", xlim = range(x4), cex.lab=2)
axis(side = 1, at = seq(-2.5,2.5,0.25), labels = as.character(seq(-2.5,2.5,0.25)),
     font = 5, cex.axis = 1.25)
axis(side = 2, at = seq(-3,2.5,0.5), labels = as.character(seq(-3,2.5,0.5)),
     font = 5, cex.axis = 1.25)
abline(h = seq(-3,2.5,0.5), col = "gray75", lty = 2)
abline(v = seq(-2.5,2.5,0.25), col = "gray80", lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(x3, residuals(out), col = addTrans("orange",120), pch = 19)
points(x3, residuals(out), col = "orange")
panel.smooth(x3, residuals(out), col = "orange",cex = 1, lwd = 1.2,
             col.smooth = "seagreen", span = 2/3, iter = 3)

plot(x4, residuals(out), col = NA, pch = 19, axes = FALSE, ylab = "Residuals",
     font.lab = 3, xlab = "", xlim = range(x4), cex.lab=2)
axis(side = 1, at = seq(-2.5,2.5,0.25), labels = as.character(seq(-2.5,2.5,0.25)),
     font = 5, cex.axis = 1.25)
axis(side = 2, at = seq(-3,2.5,0.5), labels = as.character(seq(-3,2.5,0.5)),
     font = 5, cex.axis = 1.25)
abline(h = seq(-3,2.5,0.5), col = "gray75", lty = 2)
abline(v = seq(-2.5,2.5,0.25), col = "gray80", lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(x4, residuals(out), col = addTrans("orange",120), pch = 19)
points(x4, residuals(out), col = "orange")
panel.smooth(x4, residuals(out), col = "orange",cex = 1, lwd = 1.2,
             col.smooth = "seagreen", span = 2/3, iter = 3)
mtext("Linear Regression Residuals vs. Predictors", side = 3, line = -1.2,
      outer = TRUE, font = 3, cex = 2)
quartz.save(file = "residuals.png", type = "png")
graphics.off()
```

```
quartz(height = 12, width = 16)
par(mar = c(7,4.5,2,2) + 0.1)
par(bg = "azure")
par(mfrow=c(1,1))
plot(out, which = 1, col = NA, pch = 19, axes = FALSE,add.smooth = FALSE,
     caption = "", font.lab = 3, sub.caption = "", cex.lab = 2, labels.id = NA)
axis(side = 1, at = seq(-3,2.5,0.25), labels = as.character(seq(-3,2.5,0.25)),
     font = 5, cex.axis = 1.5)
axis(side = 2, at = seq(-3.5,2.5,0.5), labels = as.character(seq(-3.5,2.5,0.5)),
     font = 5, cex.axis = 1.5)
abline(h = seq(-3,2.5,0.5), col = "gray75", lty = 2)
abline(v = seq(-2.5,2.5,0.125), col = "gray80", lty = 2)
abline(0,0, lty = 2, col = "gray45")
points(fitted(out), residuals(out), col = addTrans("orange",120), pch = 19,
       cex = 0.8)
points(fitted(out), residuals(out), col = "orange", cex = 0.8)
quartz.save(file = "homer.png", type = "png")
```