CSCI E-106: Data Modeling

Assignment 1

Due: February, 4 2019 at 7:19 pm EST

Instructions: Students should submit their reports on Canvas. The report needs to clearly state what question is being solved, step-by-step walk-through solutions, and final answers clearly indicated. Please solve by hand where appropriate.

Please submit either scanned hand-written solution or typed solutions and two files: (1) a R Markdown file (.Rmd extension) and (2) a PDF document generated using knitr where appropriate.

All questions are coming from Kutner, M. *et al*: Applied Linear Statistical Models, Fifth Edition.

1. (1.32) Derive the equation for b_1 in $b_1 = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sum(X_i - \bar{X})^2}$.
2. (1.33) Refer to the regression model $Y_i = \beta_0 + \varepsilon_i$. Derive the least squares estimator β_0 of for this model.
3. (1.40) In fitting regression model (1.1), it was found that observation Y_i fell directly on the fitted regression line (i.e., $Y_i = \hat{Y}_i$). If this case were deleted, would the least square regression line fitted to the remaining $n - 1$ cases be changed? [*Hint:* What is the contribution of case i to the least squares criterion Q in (1.8) of the book]
4. (1.42)
5. (1.43)

1.32 Derive b_1 in $b_1 = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$

Using: Eqn. 1.9a $\Sigma Y_i = nb_0 + b_1 \Sigma X_i$

1.9b $\Sigma X_i Y_i = b_0 \Sigma X_i + b_1 \Sigma X_i^2$

Via 1.9a: $b_0 = \frac{\Sigma Y_i - b_1 \Sigma X_i}{n}$

Via 1.9b: $b_0 = \frac{\Sigma X_i Y_i - b_1 \Sigma X_i^2}{\Sigma X_i}$

$b_0 = b_0$

$$\frac{\Sigma Y_i - b_1 \Sigma X_i}{n} = \frac{\Sigma X_i Y_i - b_1 \Sigma X_i^2}{\Sigma X_i}$$

$$\frac{\Sigma Y_i - b_1 \Sigma X_i}{n} + \frac{b_1 \Sigma X_i^2}{\Sigma X_i} = \frac{\Sigma X_i Y_i}{\Sigma X_i}$$

$$\frac{(\Sigma Y_i - b_1 \Sigma X_i)\Sigma X_i + b_1 \Sigma X_i^2 (n)}{n \Sigma X_i} = \frac{\Sigma X_i Y_i}{\Sigma X_i}$$

$$\frac{\Sigma X_i \Sigma Y_i - b_1 (\Sigma X_i)^2 + n b_1 \Sigma X_i^2}{n \Sigma X_i} = \frac{\Sigma X_i Y_i}{\Sigma X_i}$$

$(\times n \Sigma X_i)$

$(\times n \Sigma X_i)$

$$b_1(-(\Sigma X_i)^2 + n \Sigma X_i^2) = n \Sigma X_i Y_i - \Sigma X_i \Sigma Y_i$$

$$b_1 = \cancel{n \Sigma X_i \Sigma Y_i} \; \frac{n \Sigma X_i Y_i - \Sigma X_i Y_i}{-(\Sigma X_i)^2 + n \Sigma X_i^2}$$

Final Step: $b_1 = \frac{\Sigma X_i Y_i - \frac{\Sigma X_i Y_i}{n}}{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}} = \frac{\Sigma(X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2}$

(1.33) Regression Model: $Y_i = \beta_0 + \varepsilon_i$. Derive β_0 (least-squares estimator)

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \varepsilon_i)^2 \qquad \text{Note: } \sum_{i=1}^{n} \varepsilon_i = 0$$

$$= \sum_{i=1}^{n} (Y_i - \beta_0)^2$$

$$= \sum_{i=1}^{n} (Y_i^2 - 2Y_i\beta_0 + \beta_0^2)$$

$\sum_{i=1}^{n} \beta_0^2 = n \cdot \beta_0^2$ as β_0 is constant for each term of i

$$= \sum_{i=1}^{n} Y_i^2 - 2\beta_0 \sum_{i=1}^{n} Y_i + n\beta_0^2$$

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n} Y_i + 2n\beta_0$$

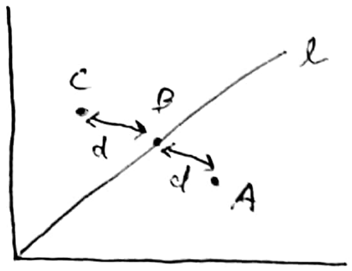
Set $\frac{\partial Q}{\partial \beta} = 0$

$$0 = -2\sum_{i=1}^{n} Y_i + 2n\beta_0$$

$$\beta_0 = \frac{\cancel{2}\sum_{i=1}^{n} Y_i}{\cancel{2}n} = \boxed{\frac{\Sigma Y_i}{n} = \beta_0}$$

$$(= \bar{Y})$$

(1.40) Thinking of a simple example:



→ Line l is the least squares regression line for points C and A.

→ If point B is found and added, it would not change line l as least squares regression line.

(No)

→ Reverse is true as well: if we start with points A, B, C, taking away B won't change line l.

(1.42)

a) $L(\beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left[-\frac{1}{2\sigma^2}(Y_i - \beta_0 - \beta_1 X_i)^2\right]$

$$\boxed{\begin{gathered} L(\beta_0, \beta_1, 16) = \prod_{i=1}^{6} \frac{1}{(32\pi)^{1/2}} \exp\left[-\frac{1}{32}(Y_i - \beta_1 X_i)^2\right] \\ n = 6, \\ Y_i = \beta_1 X_i \end{gathered}}$$

Assignment 1

Yinan Kang

R Markdown

Problem 1.42(b)

```r
data <- data.frame( x = c(7,12,4,14,25,30), y = c(128,213,75,250,446,540))

llhd.fn <- function(data, b1) {
sol <- 1
  for (i in 1:nrow(data)) {

    sol.temp <- (1/(sqrt(32*pi))) * exp((-1/32)*((data[i,2] - b1*data[i,1]))^2)

    sol <- sol*sol.temp

  }
return(sol)
}

llhd.fn(data, 17)
```

```
## [1] 9.45133e-30
```

```r
llhd.fn(data, 18)
```

```
## [1] 2.649043e-07
```

```r
llhd.fn(data, 19)
```

```
## [1] 3.047285e-37
```

Evaluating likelihood function for B1 = 17: 9.45133e-30
Evaluating likelihood function for B1 = 18: 2.649043e-07
Evaluating likelihood function for B1 = 19: 3.047285e-37

B1 = 18 results in the largest likelihood function value.

##1.42(c)

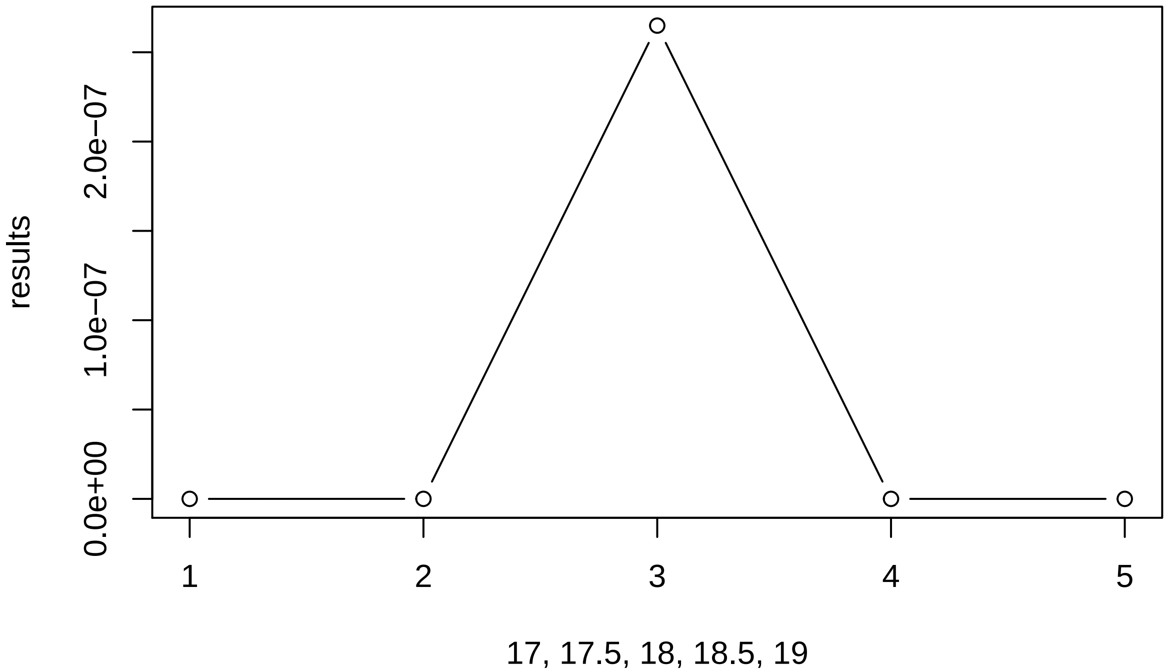
```r
numer <- 0
for (i in 1:nrow(data)) {
  numer.temp <- data[i,1]*data[i,2]
  numer <- numer + numer.temp
}
denom <- 0
for (i in 1:nrow(data)) {
  denom.temp <- (data[i,1])^2
  denom <- denom+denom.temp
}

b1 <- numer/denom
b1
```

```
## [1] 17.9285
```

The b1 estimate is: 17.9285. As it rounds to 18, it matches result in (b)

##1.42(d)

```r
results <- c(llhd.fn(data,17),llhd.fn(data,17.5),llhd.fn(data,18),llhd.fn(data,18.5),llhd.fn(data,19))
plot(results,type = "b", xlab = "17, 17.5, 18, 18.5, 19")
```



Yes, as is shown in plot above, the likelihood function's maximum is at B1 = 18, consistent with result in (c)

Problem 1.43

```r
data1 <- read.csv("CDI.csv")
head(data1,3)
```

```
##   ID      County State Land.Area Total.Population X.Pop.aged.18.34
## 1  1 Los_Angeles    CA      4060          8863164             32.1
## 2  2        Cook    IL       946          5105067             29.2
## 3  3      Harris    TX      1729          2818199             31.3
##   X.Pop.aged.65.and.over Number.Active.Physicians Number.of.Hospital.Beds
## 1                    9.7                    23677                   27700
## 2                   12.4                    15153                   21550
## 3                    7.1                     7553                   12449
##   Total.Serious.Crimes X.High.school.grad X.Bachelor.s.degrees
## 1               688936               70.0                 22.3
## 2               436936               73.4                 22.8
## 3               253526               74.9                 25.4
##   X.Below.Poverty.level X.Unemployment Per.Capita.Income
## 1                  11.6            8.0             20786
## 2                  11.1            7.2             21729
## 3                  12.5            5.7             19517
```

```
##   Total.personal.income Geographic.Region
## 1                184230                 4
## 2                110928                 2
## 3                 55003                 3
```

```
# Running linear regressions:

reg.pop <- lm(data1$Number.Active.Physicians ~ data1$Total.Population)
reg.beds <- lm(data1$Number.Active.Physicians ~ data1$Number.of.Hospital.Beds)
reg.pIncome <- lm(data1$Number.Active.Physicians ~ data1$Total.personal.income)

reg.pop
```

```
##
## Call:
## lm(formula = data1$Number.Active.Physicians ~ data1$Total.Population)
##
## Coefficients:
##            (Intercept)  data1$Total.Population
##             -1.106e+02               2.795e-03
```

```
reg.beds
```

```
##
## Call:
## lm(formula = data1$Number.Active.Physicians ~ data1$Number.of.Hospital.Beds)
##
## Coefficients:
##                   (Intercept)  data1$Number.of.Hospital.Beds
##                      -95.9322                         0.7431
```

```
reg.pIncome
```

```
##
## Call:
## lm(formula = data1$Number.Active.Physicians ~ data1$Total.personal.income)
##
## Coefficients:
##                 (Intercept)  data1$Total.personal.income
##                    -48.3948                       0.1317
```

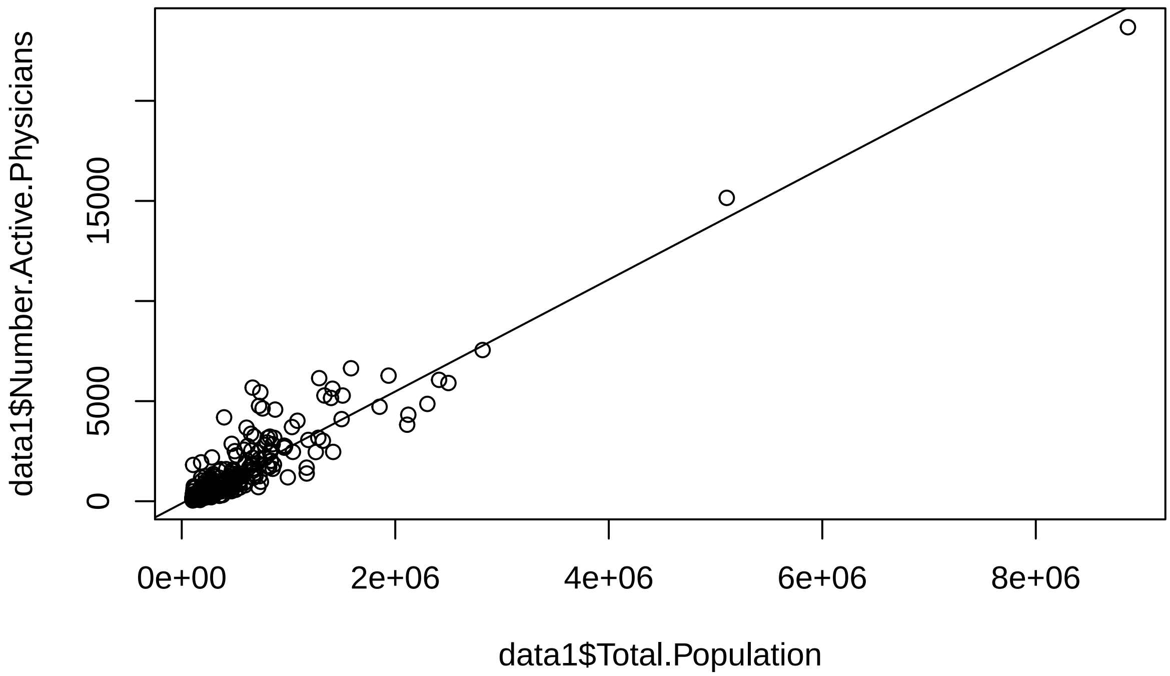
1.43(a)

Regressing when Y = Number of Active Physicians:

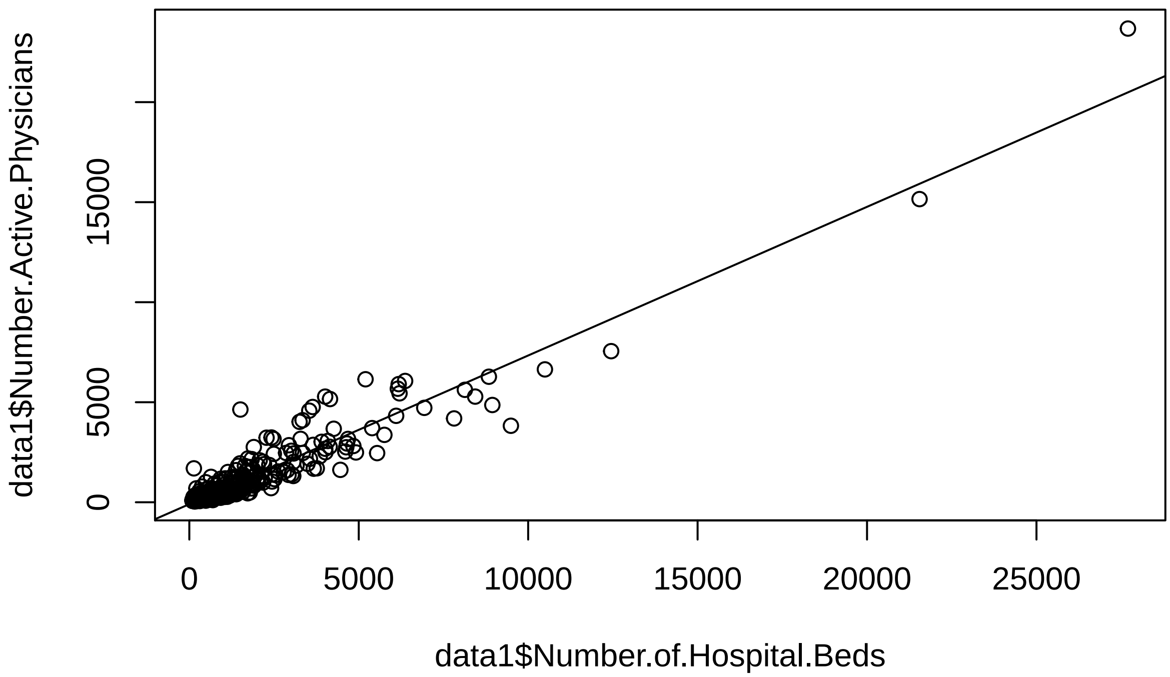
Using X = Total Population, Y = -1.106e+02 + 2.795e-03X
*Using X = Number of Hospital Beds, Y = -95.9322 + 0.7431*X
Using X = Total Personal Income, Y = -48.3948 + 0.1317*X

1.43(b)

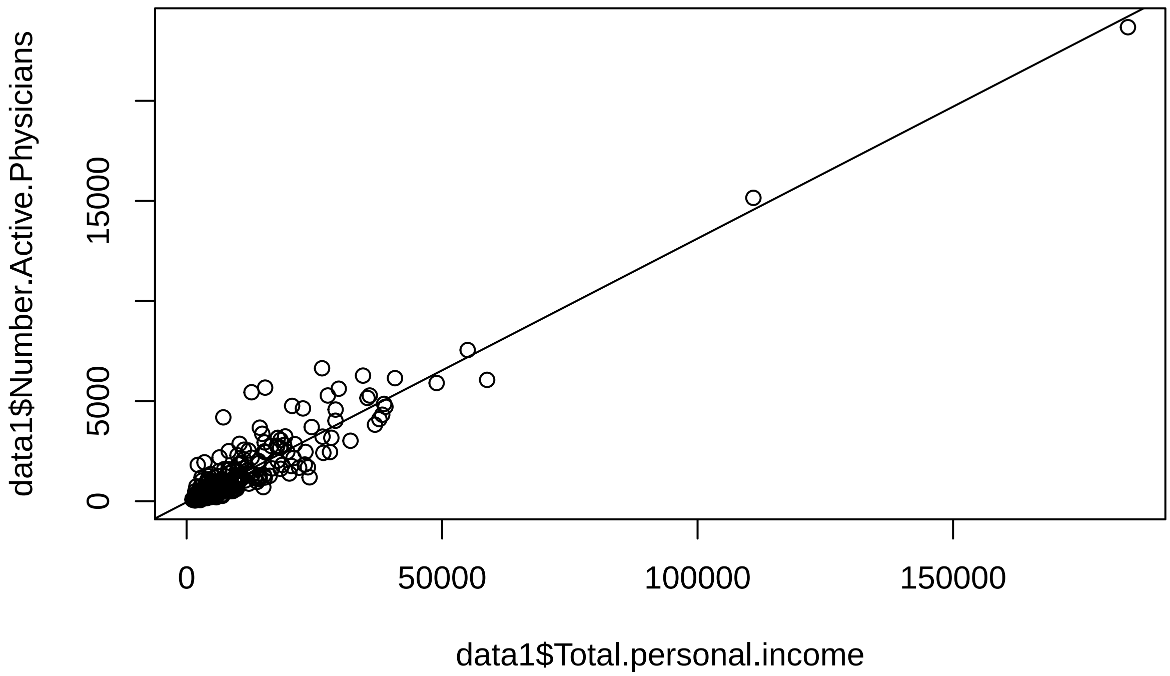
```
plot(data1$Number.Active.Physicians ~ data1$Total.Population)
abline(lm(data1$Number.Active.Physicians ~ data1$Total.Population))
```




```r
plot(data1$Number.Active.Physicians ~ data1$Number.of.Hospital.Beds)
abline(lm(data1$Number.Active.Physicians ~ data1$Number.of.Hospital.Beds))
```




```r
plot(data1$Number.Active.Physicians ~ data1$Total.personal.income)
abline(lm(data1$Number.Active.Physicians ~ data1$Total.personal.income))
```



From a glance at the plots, yes, a linear regression seems to be a good fit (though it is difficult to see as the outliers "zoom-out" the plot significantly… but there seem to be a cluster of observations near the lower portions of the line, so it seems to be a decent fit).

##1.43(c) Calculating Mean Squared Error

```
mean(reg.pop$residuals^2)
```

```
## [1] 370511.7
```

```
mean(reg.beds$residuals^2)
```

```
## [1] 308781.9
```

```
mean(reg.pIncome$residuals^2)
```

```
## [1] 323064.2
```

Assuming no bias, # Hospital beds leads to smallest variability around regression line