LINEAR REGRESSION MODELS: Homework #1

Due on September 18, 2017

Professor Jingchen Liu

Fan Yang
UNI: fy2232

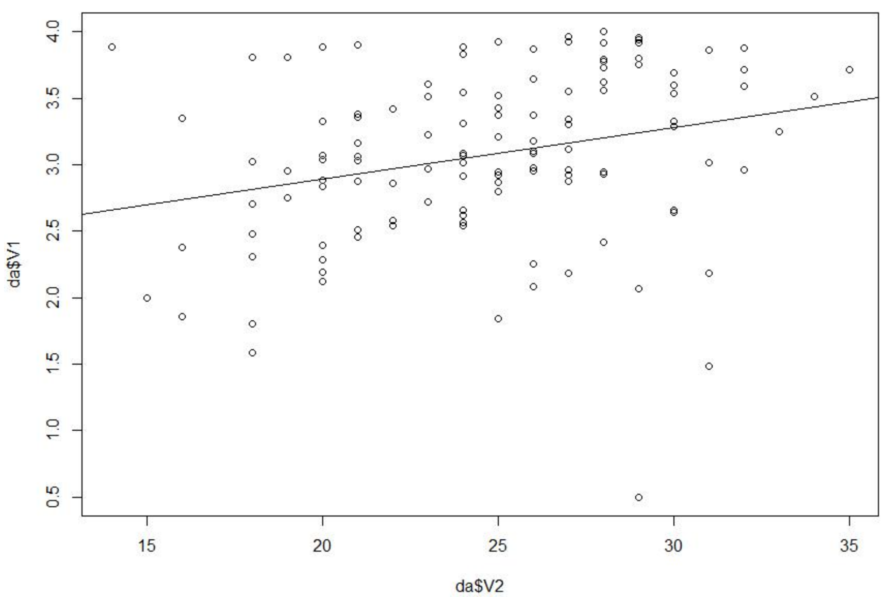
Problem 1

a.

```
da = read.table("CH01PR19.txt")
lm.pr19<-lm(da$V1~da$V2)
summary(lm.pr19)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.11405    0.32089   6.588  1.3e-09 ***
da$V2        0.03883    0.01277   3.040  0.00292 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6231 on 118 degrees of freedom
Multiple R-squared:  0.07262,   Adjusted R-squared:  0.06476
F-statistic:  9.24 on 1 and 118 DF,  p-value: 0.002917
```

$$therefore\ \widehat{y} = 2.44105 + 0.03883x$$



b.

when a x is fixed, there could be some different possible y. so the relation between x and y could. not be linear. We can draw conclusion from the plot above that the function does not fit the data.

c.

When X=30, we apply the function in (a)

$$\widehat{y} = \beta_0 + \beta_1 * x$$
$$= 2.11405 + 0.03883 * 30 = 3.27895$$

d.

$$\Delta y = y_1 - y_2 = (2.11405 + 0.03883 * x) - (2.11405 + 0.03883 * (x + 1)) = 0.03883$$

Problem 2

When β_0 is 0,we know that the regression model only determines by β_1,and the function line goes through origin point and is a linear line which only depends on the slope.

Problem 3

When β_1 is 0,the response variable is a constant and no longer related to explanatory variable. And Y_i always equal to β_0. The regression now becomes horizontal.

Problem 4

the goal of least squares estimator is to minimize $\sum_i^n (y_i - \beta_0)^2$

$$\sum_i^n (y_i - \beta_0)^2 = \sum_i^n (y_i - \overline{y} + \overline{y} - \beta_0)^2 = \sum_i^n (y_i - \overline{y})^2 - 2\sum_i^n (y_i - \overline{y})(\overline{y} - \beta_0) + \sum_i^n (\overline{y} - \beta_0)^2$$

$$= \sum_i^n (y_i - \overline{y})^2 + \sum_i^n (\overline{y} - \beta_0)^2$$

in order to minimize the above statement, β_0 should be equal to $\overline{y}$.

Therefore,least squares estimator of β_0 is $\widehat{\beta_0} = \overline{y}$

Problem 5

$$\begin{aligned} E(\widehat{\beta}_0) &= E(\overline{y}) = E(\frac{1}{n}\sum y_i) \\ &= \frac{1}{n}E(\sum y_i) = \frac{1}{n}\sum E(y_i) = \frac{1}{n}\sum E(\beta_0) = \frac{1}{n} * n * \beta_0 \\ &= \beta_0 \end{aligned}$$

so that $\widehat{\beta}_0$ is unbiased

Problem 6

a.

We use the following conclusion without proof

$\widehat{\beta}_1 = \dfrac{\sum(x_i - \overline{x})(y_i - \overline{y})}{(\sum x_i - \overline{x})^2}$

$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$

$\overline{x} = 10$

denote $\overline{Y}$ as the mean of the 6 observations(also the mean of 3 means of observations)

1) in the 3 points regression

$$\widehat{\beta}_1^1 = \frac{(5-10)(\overline{Y}_1 - \overline{Y}) + (10-10)(\overline{Y}_2 - \overline{Y}) + (15-10)(\overline{Y}_3 - \overline{Y})}{(5-10)^2 + (10-10)^2 + (15-10)^2}$$

$$= \frac{-5(\overline{Y}_1 - \overline{Y}) + 5(\overline{Y}_3 - \overline{Y})}{50} = \frac{\overline{Y}_3 - \overline{Y}_1}{10}$$

2) in the 6 points regression

$\widehat{\beta}_1^2 =$

$$\frac{(5-10)[(Y_{11} - \overline{Y}) + (Y_{12} - \overline{Y})] + (10-10)[(Y_{21} - \overline{Y}) + (Y_{22} - \overline{Y})] + (15-10)[(Y_{32} - \overline{Y}) + (Y_{33} - \overline{Y})]}{2*(5-10)^2 + 2*(10-10)^2 + 2*(15-10)^2}$$

$$= \frac{-5(Y_{11} - \overline{Y} + Y_{12} - \overline{Y}) + 5(Y_{31} - \overline{Y} + Y_{32} - \overline{Y})}{100}$$

$$= \frac{-5(2\overline{Y}_1 - 2\overline{Y}) + 5(2\overline{Y}_3 - 2\overline{Y})}{100}$$

$$= \frac{\overline{Y}_3 - \overline{Y}_1}{10}$$

$$= \widehat{\beta}_1^1$$

that's to say, the $\widehat{\beta}_1$ in two models are identical

Besides, $\overline{x}$ and $\overline{y}$ are same in two models.

according to $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$ we know $\widehat{\beta}_0$ are same.

therefore, the two regression lines are identical.

b.

$\widehat{\sigma}^2 = \frac{\sum(y_i - \widehat{y}_i)^2}{n-2} = \frac{\sum(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2}{n-2}$
$= (\overline{Y}_1 - \widehat{\beta}_0 - \widehat{\beta}_1 X_1)^2 + (\overline{Y}_2 - \widehat{\beta}_0 - \widehat{\beta}_1 X_2)^2 + (\overline{Y}_3 - \widehat{\beta}_0 - \widehat{\beta}_1 X_3)^2$
$= (\overline{Y}_1 - \widehat{\beta}_0 - 5\widehat{\beta}_1)^2 + (\overline{Y}_2 - \widehat{\beta}_0 - 10\widehat{\beta}_1)^2 + (\overline{Y}_3 - \widehat{\beta}_0 - 15\widehat{\beta}_1)^2$
from a we know $\widehat{\beta}_1 = \frac{\overline{Y}_3 - \overline{Y}_1}{10}$ and $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$
then $\widehat{\sigma}^2 = (\overline{Y}_1 - \overline{Y})^2 + (\frac{(\overline{Y}_1 + 4\overline{Y}_2 - 5\overline{Y}_3}{6})^2 + (\overline{Y}_1 - \overline{Y})^2$
$= 2(\frac{2\overline{Y}_1 - \overline{Y}_2 - \overline{Y}_3}{3})^2 + (\frac{(\overline{Y}_1 + 4\overline{Y}_2 - 5\overline{Y}_3}{6})^2$
Thus, we only need to apply $\overline{Y}_1$ $\overline{Y}_2$ $\overline{Y}_3$ to above equation, then we will get the estimator of σ^2 without fitting a regression line

Problem 7

a.

We use the following conclusion without proof
$\widehat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{(\sum x_i - \overline{x})^2}$
$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x}$
since we know that $\beta_0 = 0$
now our goal is to minimize $\sum_i^n (y_i - \beta_1 x_i)^2$

$$\sum_i^n (y_i - \beta_1 x_i)^2 = \sum_i^n [y_i - \overline{y} + \beta_1 \overline{x} - \beta_1 x_i]^2$$
$$= \sum_i^n (y_i - \overline{y})^2 + 2\sum_i^n (y_i - \overline{y})(\beta_1 \overline{x} - \beta_1 x_i) + \sum_i^n (\beta_1 \overline{x} - \beta_1 x_i)^2$$
$$= \sum_i^n (y_i - \overline{y})^2 + \sum_i^n (\beta_1 \overline{x} - \beta_1 x_i)^2$$
$$= \sum_i^n (y_i - \overline{y})^2 + \beta_1^2 \sum_i^n (\overline{x} - x_i)^2$$
$$= \sum_i^n (x_i - \overline{x})^2 \left(\beta_1 - \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}\right)^2 + \sum_i^n (y_i - \overline{y})^2 - \frac{(\sum(x_i - \overline{x})(y_i - \overline{y}))^2}{\sum(x_i - \overline{x})^2}$$
$$\Rightarrow \widehat{\beta}_1 = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sum(x_i - \overline{x})^2}$$

b.

$$\varepsilon_i \sim N(0, \sigma^2), pdf = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{x^2}{2\sigma^2}\right)$$

$$L(\beta_1, \sigma) = \prod_i^n \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(Y_i - \beta_1 X_i)^2}{2\sigma^2}\right)$$

$$= \frac{1}{(\sigma\sqrt{2\pi})^n} exp\left(-\frac{\sum_i^n (Y_i - \beta_1 X_i)^2}{2\sigma^2}\right)$$

in order to maximize $L(\beta_1, \sigma)$, we simply only need to minimize $\sum_i^n (Y_i - \beta_1 X_i)^2$

now it becomes the same problem as least square estimate, therefore the two estimator of β_1 are identical.

$$\widehat{\beta}_1 = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2}$$

c.

$$\widehat{\beta}_1 = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2}$$

$$E(\widehat{\beta}_1) = E(\frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2}) = \frac{\sum_i^n (x_i - \overline{x})E(y_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} = \frac{\sum_i^n (x_i - \overline{x})(E(y_i) - \overline{y})}{\sum_i^n (x_i - \overline{x})^2}$$

because $E(y_i) = \beta_1 x_i$ *and* $\overline{y} = \beta_1 x_i$;

$$E(\widehat{\beta}_1) = \frac{\sum_i^n (x_i - \overline{x})(\beta_1 x_i - \overline{y})}{\sum_i^n (x_i - \overline{x})^2} = \frac{\sum_i^n (x_i - \overline{x})(\beta_1 x_i - \beta_1 \overline{x})}{\sum_i^n (x_i - \overline{x})^2} = \frac{\beta_1 \sum_i^n (x_i - \overline{x})(x_i - \overline{x})}{\sum_i^n (x_i - \overline{x})^2}$$

$$= \beta_1$$

therefore $\widehat{\beta}_1$ is unbiased.

Problem 8

Firstly we get the observation data. Let d be the raw data. Y is number of active physicians and X is the combination of the three predictor variables.

```
Y <- d$`Number of active physicians`
X <- cbind(d$`Total population`, d$`Number of hospital beds`, d$`Total personal income`)
```

(1)

a. Regress the number of active physicians on **total population.**

```
lm.cdi1 <- lm(Y~X[,1])
summary(lm.cdi1)
```

```
Call:
lm(formula = Y ~ X[, 1])

Residuals:
    Min      1Q  Median      3Q     Max 
-1969.4  -209.2   -88.0    27.9  3928.7 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept) -1.106e+02  3.475e+01  -3.184  0.00156 ** 
X[, 1]       2.795e-03  4.837e-05  57.793  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 610.1 on 438 degrees of freedom
Multiple R-squared:  0.8841,	Adjusted R-squared:  0.8838 
F-statistic:  3340 on 1 and 438 DF,  p-value: < 2.2e-16
```

we get the regression function $Y = -110.6 + 2.795 \times 10^{-3} X$

b. Regress the number of active physicians on **number of hospital beds.**

```
lm.cdi2 <- lm(Y~X[,2])
summary(lm.cdi2)
```

```
Call:
lm(formula = Y ~ X[, 2])

Residuals:
    Min      1Q  Median      3Q     Max 
-3133.2  -216.8   -32.0    96.2  3611.1 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -95.93218   31.49396  -3.046  0.00246 ** 
X[, 2]        0.74312    0.01161  63.995  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 556.9 on 438 degrees of freedom
Multiple R-squared:  0.9034,	Adjusted R-squared:  0.9032 
F-statistic:  4095 on 1 and 438 DF,  p-value: < 2.2e-16
```

we get the regression function $Y = -95.932 + 0.743X$

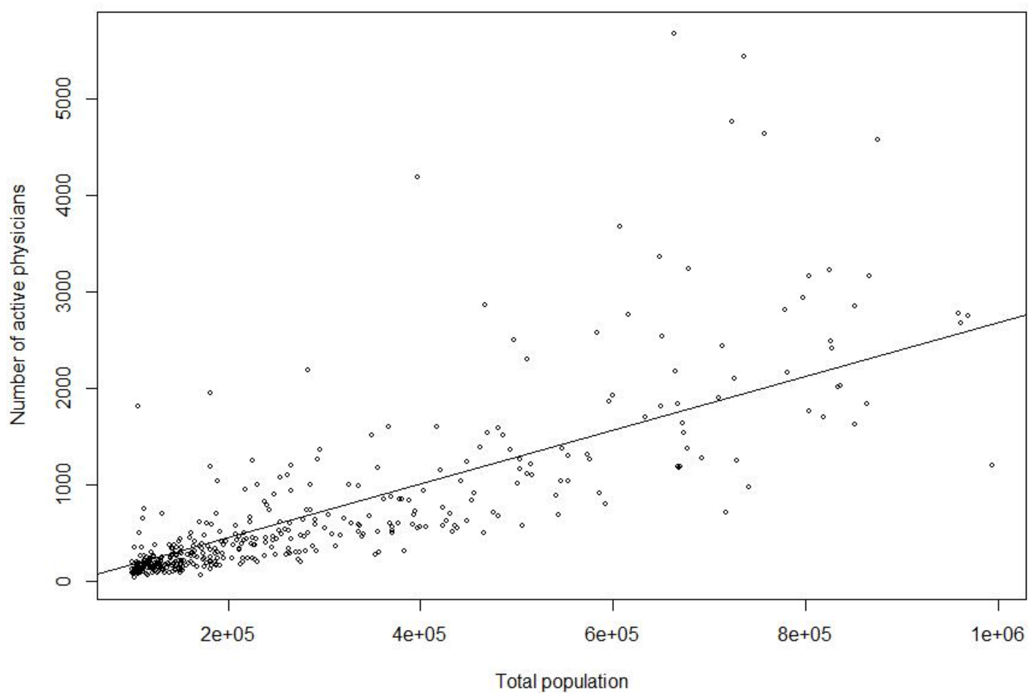
c. Regress the number of active physicians on **total personal income.**

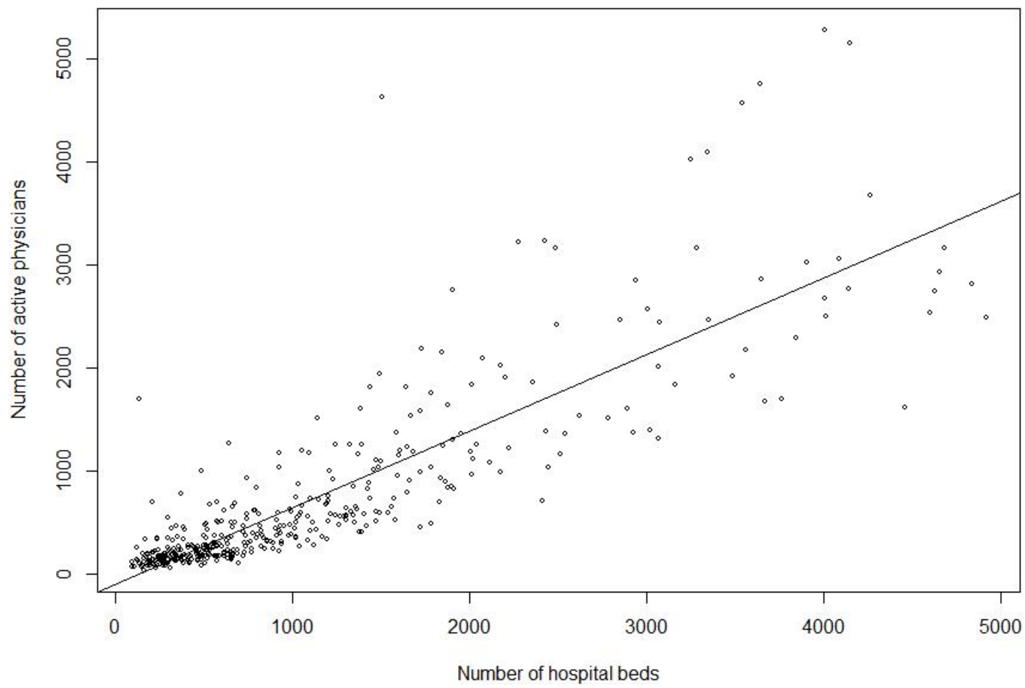
```
lm.cdi3 <- lm(Y~X[,3])
summary(lm.cdi3)
```

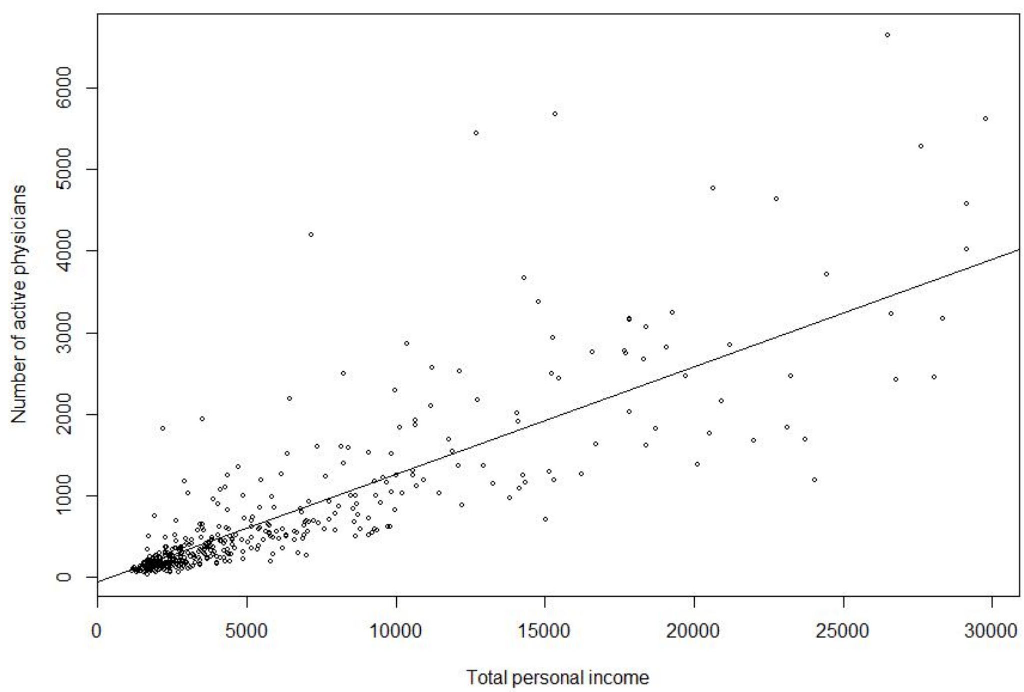
```
Call:
lm(formula = Y ~ X[, 3])

Residuals:
    Min      1Q  Median      3Q     Max 
-1926.6  -194.5   -66.6    44.2  3819.0 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -48.39485   31.83333   -1.52    0.129    
X[, 3]        0.13170    0.00211   62.41   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 569.7 on 438 degrees of freedom
Multiple R-squared:  0.8989,	Adjusted R-squared:  0.8987 
F-statistic:  3895 on 1 and 438 DF,  p-value: < 2.2e-16
```

we get the regression function $Y = -48.395 + 0.132X$

(2)







from 3 plots above, we find that simple linear regression line somehow depicts the relation between X and Y. Besides the P-values in each model are less than 0.001, so the regression lines seem to be a good fit for each of the predictor variables

(3)

```
MSE1 <- sum(lm.cdi1$residuals^2)/(440-2)
```

```
[1] 372203.5


MSE2 <- sum(lm.cdi2$residuals^2)/(440-2)
[1] 310191.9


MSE3 <- sum(lm.cdi3$residuals^2)/(440-2)
[1] 324539.4
```

from the results above, we can conclude that the variable **number of hospital beds** leads to the smallest variability.

Problem 9

Let Yi be per capita income and Xi be the percentage of individuals having bachelor's degree in i^{th} region.

```
Y1<-d$‘Per capita income‘[d$‘Geographic region‘==1]
X1<-d$‘Percent bachelor’s degrees‘[d$‘Geographic region‘==1]
Y2<-d$‘Per capita income‘[d$‘Geographic region‘==2]
X2<-d$‘Percent bachelor’s degrees‘[d$‘Geographic region‘==2]
Y3<-d$‘Per capita income‘[d$‘Geographic region‘==3]
X3<-d$‘Percent bachelor’s degrees‘[d$‘Geographic region‘==3]
Y4<-d$‘Per capita income‘[d$‘Geographic region‘==4]
X4<-d$‘Percent bachelor’s degrees‘[d$‘Geographic region‘==4]
```

(1)

Regress the **per capita income** on **total population** for the **first** region

```
lm.cdi1<-lm(Y1~X1)
summary(lm.cdi1)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9223.82     851.77   10.83   <2e-16 ***
X1            522.16      37.13   14.06   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

we get the regression function $Y = 9223.82 + 522.16X$

Regress the **per capita income** on **total population** for the **second** region

```
lm.cdi2<-lm(Y2~X2)
summary(lm.cdi2)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13581.41     575.14  23.614  < 2e-16 ***
X2            238.67      27.23   8.765 3.34e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

we get the regression function $Y = 13581.41 + 238.67X$

Regress the **per capita income** on **total population** for the **third** region

```
lm.cdi3<-lm(Y3~X3)
summary(lm.cdi3)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 10529.79     612.48   17.19   <2e-16 ***
X3            330.61      27.13   12.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

we get the regression function $Y = 10529.79 + 330.61X$

Regress the **per capita income** on **total population** for the **forth** region

```
lm.cdi4<-lm(Y4~X4)
summary(lm.cdi4)
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8615.05    1052.20   8.188 5.24e-12 ***
X4            440.32      45.37   9.705 6.86e-15 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

we get the regression function $Y = 8615.05 + 440.32X$

(2)

Let R1 = 1 if in region 1; otherwise 0; Let R2 = 1 if in region 2; otherwise 0;

Let R3 = 1 if in region 3; otherwise 0;

Let X,Y be the **per capita income** and **total population** ,respectively.

Then we apply full model

$Y = \beta_0 + \beta_1 X + \beta_2 R1X + \beta_3 R2X + \beta_3 R3X + \varepsilon$

```
> lm.full <- lm(Y~X+R1*X+R2*X+R3*X)
```

If there is no region effect, then the reduced model is

$Y = \beta_0 + \beta_1 X + \varepsilon$

```
> lm.reduce <- lm(Y~X)
```

For the full model,we have

```
> sum(lm.full$residuals^2)
[1] 2945664166
> lm.full$df.residual
[1] 432
```

$$SSE = 3496250017 \qquad DF = 432$$

For the reduced model,we have

```
> sum(lm.reduce$residuals^2)
[1] 3735858256
> lm.reduce$df.residual
[1] 438
```

$$SSE = 3735858256 \qquad DF = 438$$

Thus

$$\begin{aligned} F^* &= \frac{(3735858256 - 2945664166)/(438 - 432)}{2945664166/432} \\ &= 19.31448 > F_{95\%}(6, 432) = 2.12 \end{aligned}$$

Therefore, region matters, which means different region have different regression functions.

(3)

```
MSE1 <- sum(lm.cdi1$residuals^2)/lm.cdi1$df.residual
[1] 7335008


MSE2 <- sum(lm.cdi2$residuals^2)/lm.cdi2$df.residual
[1] 4411341


MSE3 <- sum(lm.cdi3$residuals^2)/lm.cdi3$df.residual
[1] 7474349


MSE4 <- sum(lm.cdi4$residuals^2)/lm.cdi4$df.residual
[1] 8214318
```

from the results above, we can conclude that the variable around the fitted regression line differs from each other. But the variable for region #1 and #3 are approximately the same.