Solution to HW 1

Guanhua FANG

September 20, 2017

1.29

a. $\beta_1 = 0.0388$ and $\beta_0 = 2.11$. $Y = 2.11 + 0.0388X$.

b. See figure 1, the estimated line fits data not to bad.

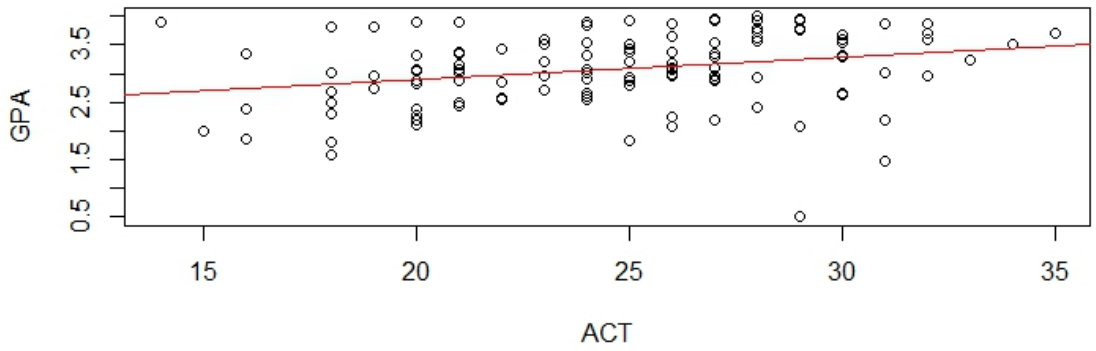


Figure 1: Ch1.19

c. $\hat{Y}_{|30} = 3.27$

d. Changed GPA is 0.0388 by increasing one point of ACT.

1.29

The regression line goes through the origin.

1.30

X and Y are independent, the regression line is almost flat.

1.33

$\hat{\beta}_0 = \frac{1}{n}\sum_i Y_i$.

1.34

$\mathrm{E}\hat{\beta}_0 = \mathrm{E}\{\frac{1}{n}\sum_i Y_i\} = n\frac{1}{n}\beta_0 = \beta_0$. It is unbiased by definition.

1.39

a. Suppose six points are $(5, y_1)$, $(5, y_2)$, $(10, y_3)$, $(10, y_4)$, $(15, y_5)$, $(15, y_6)$. Denote the $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimated coefficients by these six points. Similarly, denote $\hat{\beta}_0^{'}$ and $\hat{\beta}_1^{'}$ are the estimated coefficients by three mid points. What we need to show are $\hat{\beta}_0 = \hat{\beta}_0^{'}$ and $\hat{\beta}_1 = \hat{\beta}_1^{'}$. In the all following, superscript $'$ means the estimated value by three mid points.

Recall that

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_i (x_i - x)(y_i - y)}{\sum_i (x_i - \bar{x})^2} \\
&= \frac{(5-10)(y_1 - \bar{y}) + (5-10)(y_2 - \bar{y}) + (15-10)(y_5 - \bar{y}) + (15-10)(y_6 - \bar{y})}{2[(5-10)^2 + (15-10)^2]} \\
&= \frac{(5-10)(\frac{y_1+y_2}{2} - \bar{y}) + (5-10)(\frac{y_5+y_6}{2} - \bar{y})}{[(5-10)^2 + (15-10)^2]} \\
&= \hat{\beta}_1^{'}.
\end{aligned}$$

here, we use the fact that $\bar{y} = \bar{y}^{'}$. We the recall that

$$\hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 = \bar{y}^{'} - \bar{x}^{'}\hat{\beta}_1^{'} = \hat{\beta}_0^{'}.$$

b. Recall that MSE $= \frac{\text{SSE}}{n-2} = \frac{\sum_{i=1}^{6}(y_i - \hat{y}_i)^2}{6-2}$. By using the fact of simple algebra that

$$\begin{aligned}
\sum_{i=1}^{6}(y_i - \hat{y}_i)^2 &= (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2 + (y_3 - \hat{y}_3)^2 + (y_4 - \hat{y}_4)^2 + (y_5 - \hat{y}_5)^2 + (y_6 - \hat{y}_6)^2 \\
&= (y_1 - \bar{y}_1 + \bar{y}_1 - \hat{y}_1)^2 + (y_2 - \bar{y}_1 + \bar{y}_1 - \hat{y}_2)^2 + (y_3 - \bar{y}_2 + \bar{y}_2 - \hat{y}_3)^2 \\
&+ (y_4 - \bar{y}_2 + \bar{y}_2 - \hat{y}_4)^2 + (y_5 - \bar{y}_3 + \bar{y}_3 - \hat{y}_5)^2 + (y_6 - \bar{y}_3 + \bar{y}_3 - \hat{y}_6)^2 \\
&= (y_1 - \bar{y}_1)^2 + (y_2 - \bar{y}_1)^2 + (y_3 - \bar{y}_2)^2 + (y_4 - \bar{y}_2)^2 + (y_5 - \bar{y}_3)^2 + (y_6 - \bar{y}_3)^2 \\
&+ 2(\bar{y}_1 - \hat{y}_1^{'})^2 + 2(\bar{y}_2 - \hat{y}_2^{'})^2 + 2(\bar{y}_3 - \hat{y}_3^{'})^2. \qquad (1)
\end{aligned}$$

In the above display, we use the fact $\hat{y}_{2i-1} = \hat{y}_{2i} = \hat{y}_i^{'}, i = 1, 2, 3$ and $\bar{y}_{2i-1} = \bar{y}_{2i}, i = 1, 2, 3$. From the above equation, we can see that y_i and $\bar{y}_i$ have nothing to do with regression. What we only need to consider is how to derive $\hat{y}_i^{'}$ which represents the fitted value by three points at each $X_i = 5, 10, 15$. Thanks to the fact that $2 * 10 = 5 + 15$, the symmetry of the points, we have that $\hat{y}_2^{'} = \bar{y}$, $\hat{y}_1^{'} = (\bar{y}_1 - \bar{y}_3)/2 + \bar{y}$ and $\hat{y}_3^{'} = (\bar{y}_3 - \bar{y}_1)/2 + \bar{y}$. It means that the left hand side of (1) does not require calculating the regression line. Hence, we can estimate the MSE with out fitting regression line.

1.41

a. $\hat{\beta}_1 = \frac{\sum_i x_i y_i}{\sum_i x_i^2}$

b. $\hat{\beta}_1' = \arg\max_\beta - \exp\{[\sum_i (y_i - x_i\beta)^2]/2\} + \text{const}$. It is equivalent to minimize $\sum_i (y_i - x_i\beta)^2$. Hence, by the definition of least square estimates, we know that $\hat{\beta}_1 = \hat{\beta}_1'$.

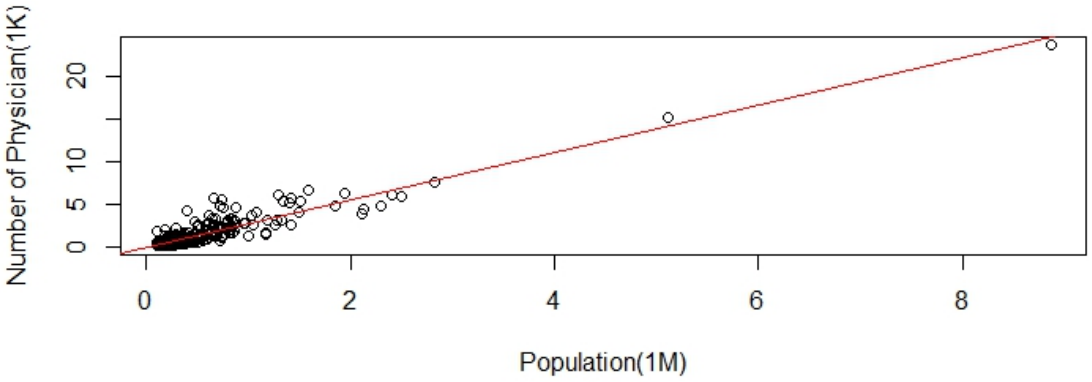
c. $\mathrm{E}\hat{\beta}_1 = \mathrm{E}[\frac{\sum_i x_i y_i}{\sum_i x_i^2}] = \sum_i \frac{x_i \mathrm{E} y_i}{\sum_i x_i^2} = \sum_i \frac{x_i \beta_1 x_i}{\sum_i x_i^2} = \beta_1$.

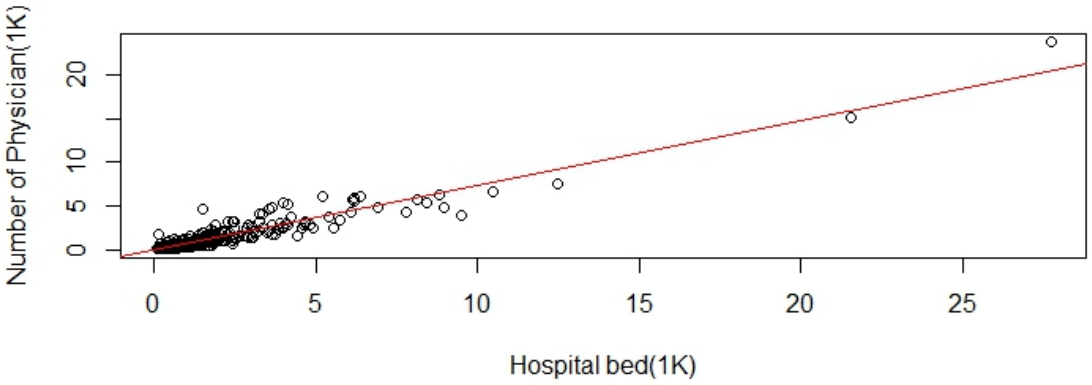
1.43

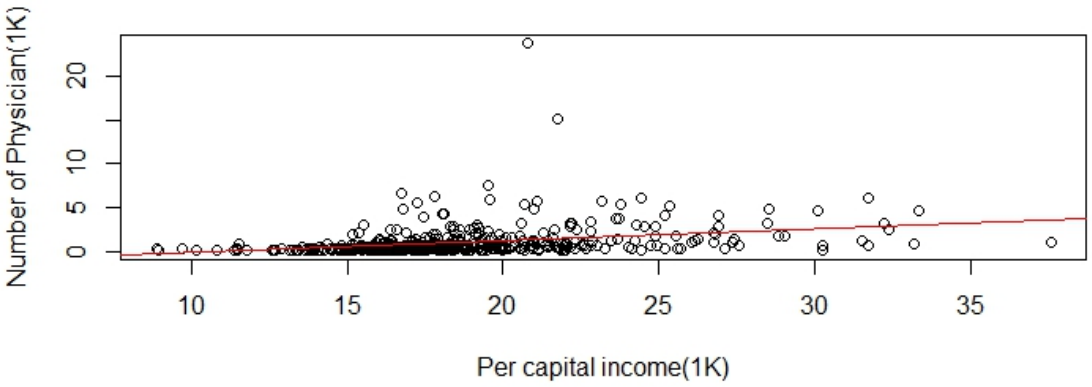
a.

$$\begin{aligned} \text{fit1}: &\quad Y = -0.11 + 2.79X \\ \text{fit2}: &\quad Y = -0.0959 + 0.743X \\ \text{fit3}: &\quad Y = -1.59 + 0.139X \end{aligned}$$

b. From graph, we can see the number of active physician is quite linearly correlated with total population and number of hospital beds. But it seems not so linearly correlated with total personal income.







c.

The MSE of the first regression is the 0.37, the second is 0.31, the third one is 2.9. The number of hospital bed leads to the smallest variability of regression fit.

1.44

a.

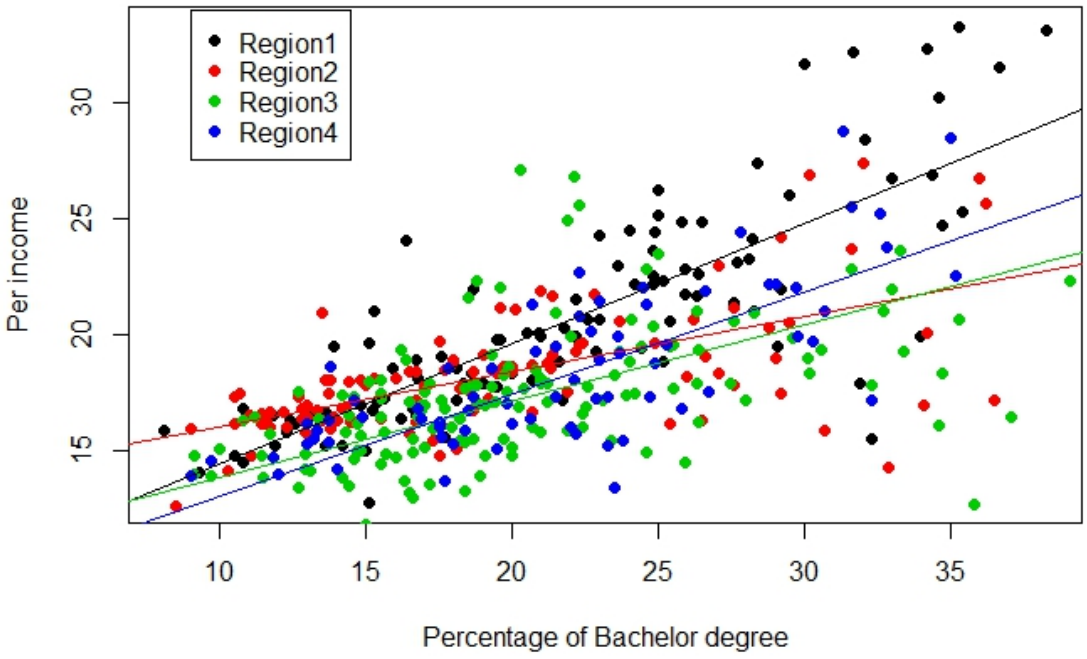
$$\begin{aligned} \text{region1}: &\quad Y = 9.22 + 0.52X; \\ \text{region2}: &\quad Y = 13.58 + 0.24X; \\ \text{region3}: &\quad Y = 10.5 + 0.33X; \\ \text{region4}: &\quad Y = 8.61 + 0.44X. \end{aligned}$$

b.

From graph, the four regions have roughly same relationship between personal income and percentage of bachelor degree.

c.

MSE for the first region is 7.3. MSE for the second region is 4.4. MSE for the third region is 7.5. MSE for the fourth region is 8.21. So region 1,3,4 have similar MSE value. Region 2 has smaller MSE value.


Region1
Region2
Region3
Region4
Per income
15
20
25
30
10
15
20
25
30
35
Percentage of Bachelor degree