Biomath 204 Homework 1

Benjamin Chu

January 23, 2017

Problem 1. Prove the Gauss-Markov theorem for β_0 in the following simple linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

assuming $E(\epsilon_i) = 0, Var(\epsilon_i') = \sigma^2, Cov(\epsilon_i, \epsilon_j) = 0$.

Proof. In class we derived that $b_0 = \bar{Y} - b_1\bar{X}$. To show b_0 is unbiased, note:

$$E(\bar{Y}) = \frac{1}{n}E(\sum_i Y_i) = \frac{1}{n}\sum_i[\beta_0 + \beta_1 X_i] = \frac{1}{n}n\beta_0 + \frac{1}{n}\beta_1\sum_i X_i = \beta_0 + \beta_1\bar{X}$$

On the other hand, $E(\beta_1 X_i) = \beta_1\bar{X}$, so

$$E(b_0) = E(\bar{Y} - \beta_1 X_i) = E(\bar{Y}) - E(\beta_1 X_i) = \beta_0 + \beta_1\bar{X} - \beta_1\bar{X} = \beta_0.$$

To show linearity in Y, recall in lecture we showed

$$b_1 = \sum_i k_i Y_i, \quad k_i = \frac{(X_i - \bar{X})}{\sum_i (X_i - \bar{X})^2}.$$

Using this, we have

$$\begin{aligned} b_0 &= \bar{Y} - b_1\bar{X} \\ &= \bar{Y} - \big(\sum_i k_i Y_i\big)\bar{X} \\ &= \frac{1}{n}\sum Y_i - \frac{1}{n}\sum k_i Y_i X_i \\ &= \frac{1}{n}\sum[Y_i - k_i Y_i X_i] \\ &= \frac{1}{n}\sum Y_i(1 - k_i X_i) \end{aligned} \tag{1}$$

Finally, to show min variance, note

□

Problem 2. Given b_0, b_1 are least-square estimators for the above regression model, show that the point $(\bar{X}, \bar{Y})$ always falls on the line $Y_i = b_0 + b_1 X_i$.

Proof. Intuitively, if we have a line that we know best estimates a set of data, then that line should be placed so that the sum of squared error is minimized. To minimize error, we must have this line pass through the mean of the data, because otherwise there would be as associated bias away from the mean. If there is bias then this line cannot be the best estimate of the data.

The formal proof has already been given in lecture, though. Let

$$Q = \sum_i \epsilon_i^2 = \sum_i [Y_i - b_0 - b_1 X_i]^2$$

$$\frac{\partial Q}{\partial b_0} = -2 \sum_i [Y_i - b_0 - b_1 X_i]$$

Setting the above expression equal to zero (i.e. finding the minimum or maximum), we have

$$\sum_i Y_i - n b_0 - b_1 \sum_i X_i = 0 \iff \bar{Y} - b_0 - b_1 \bar{X} = 0$$

$$\Rightarrow \bar{Y} = b_0 + b_1 \bar{X}$$

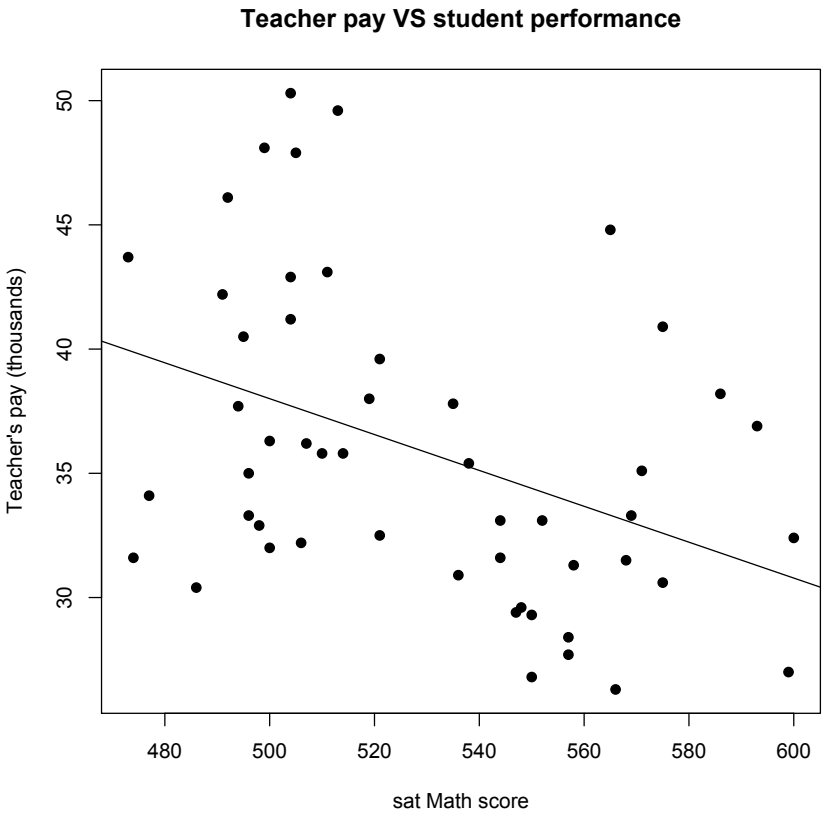
Thus the point $(\bar{X}, \bar{Y})$ is on the regression line defined by b_0 and b_1. Because $\frac{\partial^2 Q}{\partial b_0^2} = 2$, the function is concave upwards, so this point is indeed a minimum. □

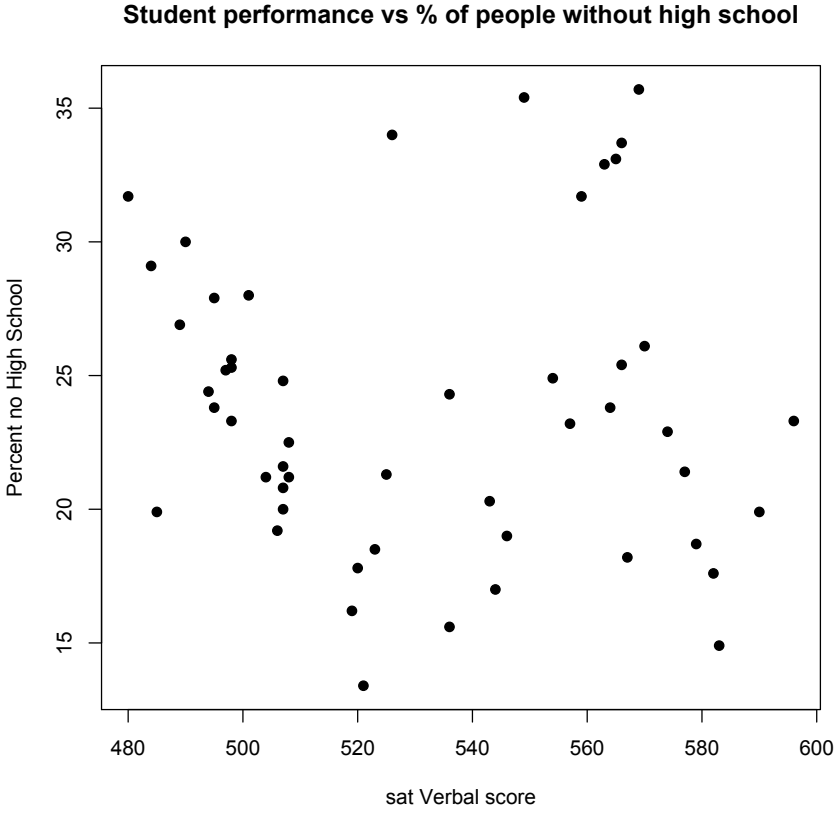
Problem 4. Using methods described in section 3.1, examine the quantitative variables of "States.txt". Characterize the distribution of the variables in terms of symmetry or skewness; non-normality or apparently normality, number of modes, and presence/absence of unusual values.

Proof. For the States.txt dataset, I compared the following: "teachers pay vs student performance (math SAT)" and "percent of state population without high school education vs student performance (verbal SAT)." I really should have used a histogram to illustrate these graphs, but then according to the textbook I should divide the 50 states into $2\sqrt{50} \approx 14$ bins so that the graph doesn't appear to overwhelming. However I'm very new to R and had a hard time figuring out how to do that since the data were given in terms of the 50 separate states, so I just plotted everything with scatter plot in the hope that it's more illustrative than 50 bars (see next page).

For the top graph, interestingly, the higher the teacher's salary the lower their student's SAT scores (i.e. these two variables have negative covariance). Intuitively we expect higher salary to reflect a higher qualification and hence teaching ability, but apparently that isn't the case. On the other hand, the bottom graph shows that while a state's population without high school education could vary considerably (from 13 to 35), that does not have any effect on student's verbal SAT score. This is another rather strange phenomenon since we would expect a more educated state to treat SAT more seriously and hence be more successful at it.

□


Teacher pay VS student performance
Teacher's pay (thousands)
30
35
40
45
50
480
500
520
540
560
580
600
sat Math score


Student performance vs % of people without high school
Percent no High School
15
20
25
30
35
480
500
520
540
560
580
600
sat Verbal score