biomath 204 hw1

Benjamin Chu

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

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

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

Finally, to show min variance, let $\widetilde{b_0}$ be another estimator so that

$$\widetilde{b_0} = \sum Y_i r_i$$

For some r_i. We need to show that $V(\widetilde{b_0}) \geq V(b_0)$.

$$V(\widetilde{b_0}) = \sum V(Y_i) r_i^2 = \sigma^2 \sum r_i^2.$$

From here, following the strategy in lecture, define $d_i = r_i - c_i$, so that

$$\sigma^2 \sum r_i^2 = \sigma^2 \sum (d_i + c_i)^2 = \sigma^2 \sum \left[d_i^2 + 2d_i c_i + c_i^2\right]$$

Because $d_i^2 \geq 0$, if we could show $d_i c_i = 0$, then we are done, because $V(b_0) = \sigma^2 \sum c_i^2$. To show this, do a bunch of algebra:

$$\sum d_i c_i = \sum (r_i - c_i) c_i = \sum r_i c_i - \sum c_i^2 \tag{2}$$

Now because $\widetilde{b_0}$ is another estimator, the r_i's must satisfy the following properties that c_i's from b_0 satisfy:

$$\sum c_i = \sum (\frac{1}{n} - k_i \bar{X}) = 1 - \bar{X} \sum k_i = 1 - 0 = 1$$

$$\sum c_i k_i = \sum (\frac{1}{n} - k_i \bar{X}) k_i = \sum \frac{k_i}{n} - \bar{X} \sum k_i^2 = 0 - \frac{\bar{X}}{\sum (x - \bar{X})^2}$$

By these two, we can evaluate the two terms from eq(2) as follows:

$$\sum r_i c_i = \sum r_i (1/n - k_i \bar{X}) = \frac{1}{n} \sum r_i - \bar{X} \sum r_i k_i = 1 + \frac{\bar{X}}{\sum (X - \bar{X})^2}$$

$$\sum c_i^2 = \sum (\frac{1}{n} - k_i \bar{X})^2 = \sum (\frac{1}{n^2} - \frac{2 k_i \bar{X}}{n} + \bar{X}^2 k_i^2) = \frac{1}{n} + \frac{\bar{X}^2}{\sum (x - \bar{x})^2}$$

Now putting everything together:

$$\sum r_i c_i - \sum c_i^2 = \frac{1}{n} + \frac{\bar{X}^2}{\sum (X - \bar{X})^2} - \frac{1}{n} - \frac{\bar{X}^2}{\sum (X - \bar{X})^2} = 0$$

Therefore $V(b_0) \leq V(\widetilde{b_0})$ and we have proven the gauss-markov theorem completely.

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 $Y_i = b_0 + b_1 X_i$.

Here we are asked to prove that if $\bar{X}$ was the input, then $\bar{Y}$ must be the output. 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. If error is minimized, then the line roughly goes the the center of all data's, i.e. it passes through the mean.

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 (i.e. input $\bar{X}$ spits out $\bar{Y}$). Because $\frac{\partial^2 Q}{\partial b_0^2} = 2$, the function is concave upwards, so this point is indeed a minimum.

problem 3

Given the regression model $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$ show that the mean square error

$$MSE = \frac{\sum (Y_i - \hat{Y}_i)^2}{n-2}$$

is an unbiased estimator.

Couldn't say I made good progress, but here's my two cents.

From lecture, we received hints on doing this with the hat matrix, implying that we should work in vector form. So set our model as

$$Y = Xb + \epsilon, \quad b = (X'X)^{-1}X'Y$$
$$\epsilon = Y - Xb = Y - X(X'X)^{-1}X'Y.$$

Here Y is a $n \times 1$ vector, X is $n \times 2$, b is 2×1, and ϵ is $n \times 1$. From this we must show that

$$E(\sum (Y_i - \hat{Y}_i)^2) = E(\epsilon'\epsilon) = (n-2)\sigma^2.$$

Now, apply the well-known formula $V(x) = E(x^2) - E(x)^2$:

$$E(\epsilon'\epsilon) = V(\epsilon) + E(\epsilon)^2$$

We know that $V(\epsilon) = \sigma^2$ from assumption. To evaluate $E(\epsilon)^2 = E(\epsilon')E(\epsilon)$, substitute the definition above:

$$\begin{aligned} E(\epsilon')E(\epsilon) &= E(Y - Xb)E[(Y - Xb)'] \\ &= E(Y - X(X'X)^{-1}X'Y)E[(Y - X(X'X)^{-1}X'Y)'] \end{aligned} \tag{3}$$

and then I got stuck.

attemp 2: evaluate $E(\epsilon'\epsilon)$ in terms of Y's:

$$\begin{aligned} E(\epsilon'\epsilon) &= E[(Y - \hat{Y})'(Y - \hat{Y})] \\ &= E[Y'Y] - 2E[Y'\hat{Y}] + E[\hat{Y}'\hat{Y}] \end{aligned} \tag{4}$$

Now recall we were also told to take advantage of the hat matrix $H = X(X'X)^{-1}X'$. We have $\hat{Y} = HY$, so we can rewrite the above equation to

$$E[Y'Y] - 2E[Y'HY] + E[Y'H'HY] = E[Y'Y] - 2E[Y'HY] + E[Y'HY] = E[Y'Y] - E[Y'HY]$$

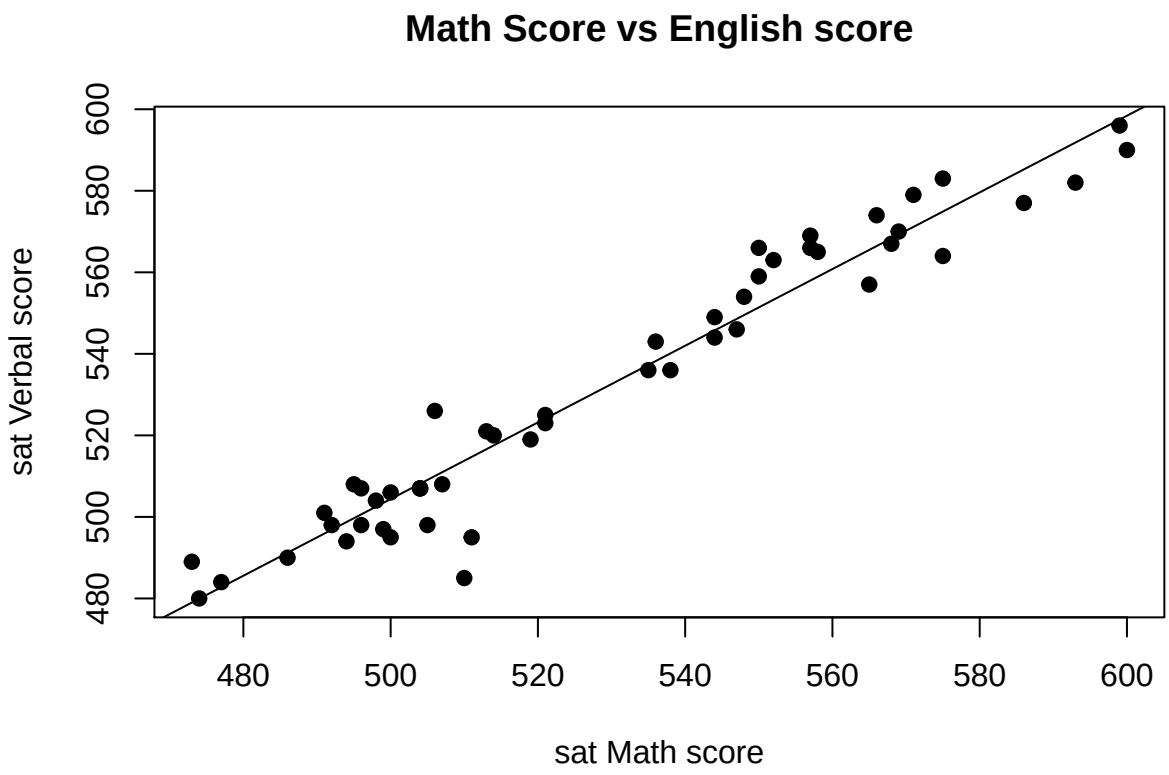
where the first equality is true because H is symmetric and idempotent.

If we now use the formula $V(X) = E(X^2) + E(X)^2$, we would run into the same trouble as in attempt 1. On the other hand, if I subsitute the model's definition for Y, $Y = Xb + \epsilon$ and $b = (X'X)^{-1}X'Y$, I wouldn't get much further. This is as far as I got.

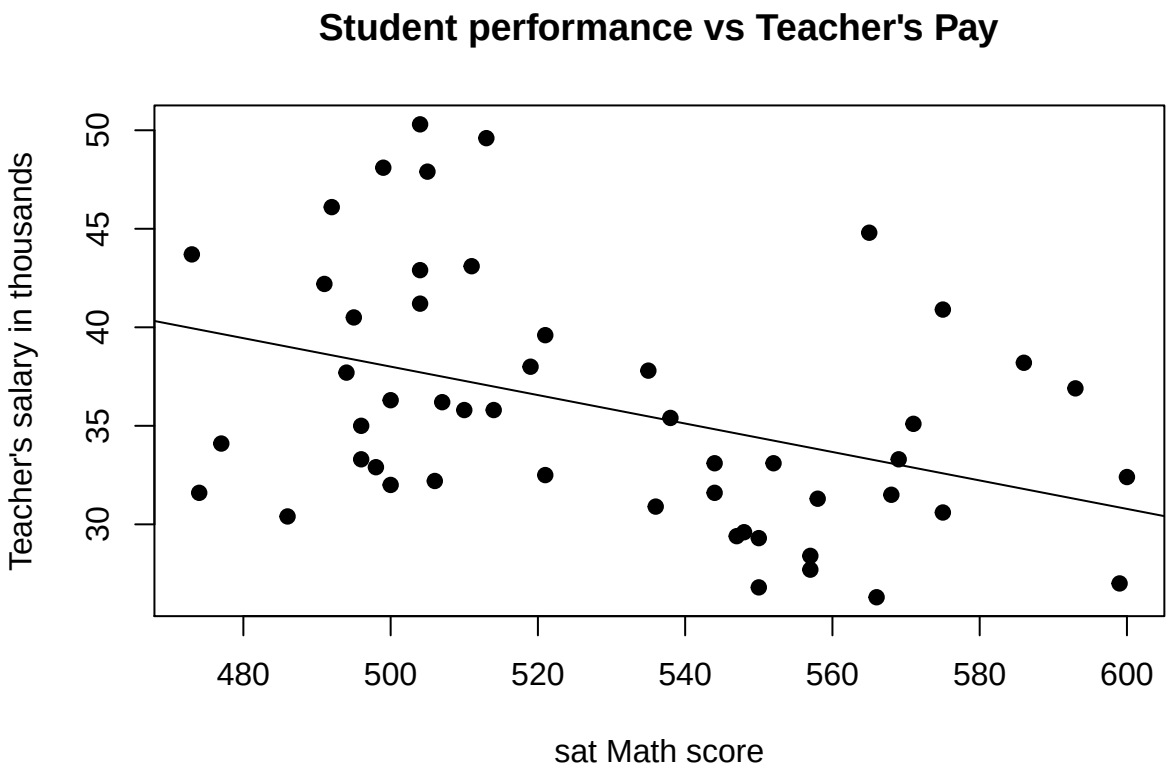
problem 4

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.

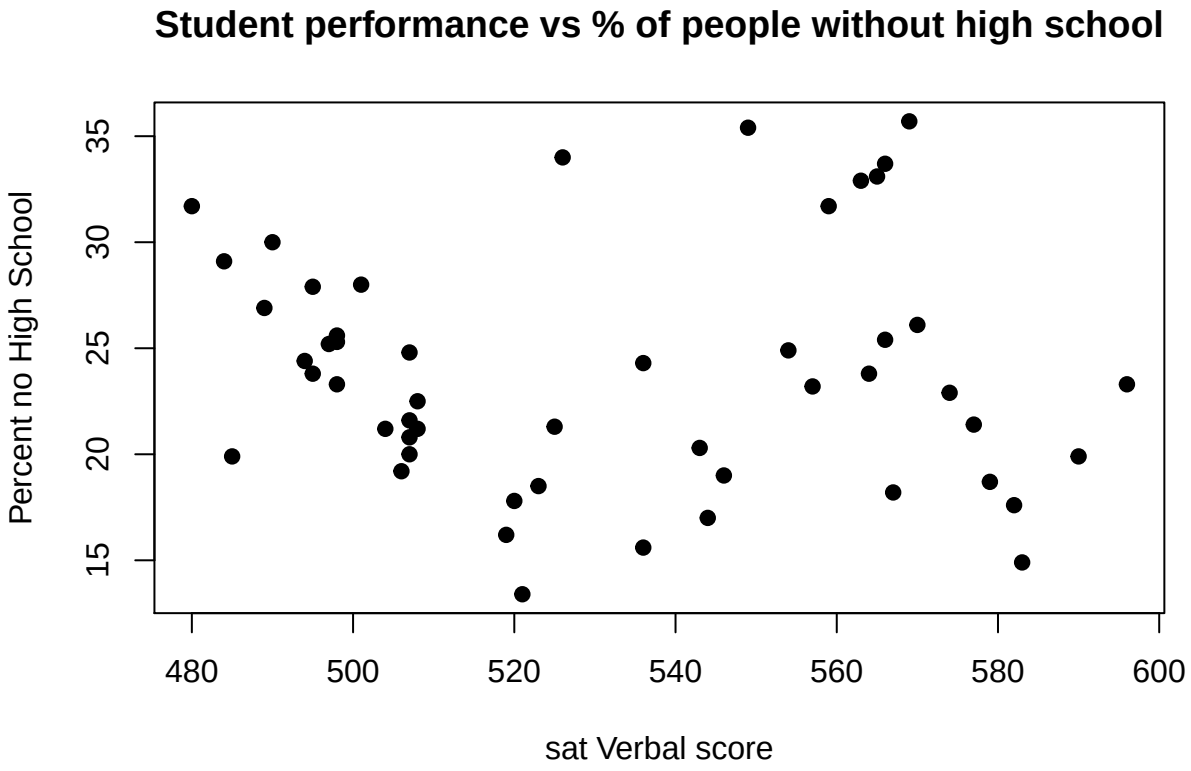
```r
mydata = read.table("C:/Users/biona002/Desktop/biomath_204-master/r_studio_stuff/States.txt")
noHS = mydata[[6]]
satVerbal = mydata[[3]]
satMath = mydata[[4]]
population = mydata[[2]]
pay = mydata[[7]]

plot(satMath, satVerbal, main="Math Score vs English score",
     xlab="sat Math score", ylab="sat Verbal score", pch=19)
math_verbal_cor = cor(satMath, satVerbal)
regression_line_1 = lm(satVerbal ~ satMath)
abline(regression_line_1)
```




```r
plot(satMath, pay, main="Student performance vs Teacher's Pay",
     xlab="sat Math score", ylab="Teacher's salary in thousands", pch=19)
regression_line_2 = lm(pay ~ satMath)
abline(regression_line_2)
```




```
plot(satVerbal, noHS, main="Student performance vs % of people without high school",
     xlab="sat Verbal score", ylab="Percent no High School", pch=19)
```



For the States.txt dataset, I compared the following: “sat Math vs english score,”" teachers pay vs student performance (math SAT)" and “percent of state population without high s chool education vs student performance (verbal SAT).” I really should have used a histogram to illustrate these graphs, but according to the textbook I should divide the 50 states into $2\sqrt{50} \approx 14$ bins so that the graph doesn't appear too overwhelming. However I'm 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.

First I wanted to determine how well a student's math ability can be used to predict his verbal ability. As seen in the first graph, the correlation is extremely good (0.9702879), so states that do well in one area tends to do well in the other.

Then I compared teacher's salary with student's SAT math scores. Intuitively we expect higher salary to reflect a higher qualification, so students receive a better education and thus scores better. However according to this dataset, the higher teacher's average salary, the worse students perform on their exams (cor = -0.4039747). This is quite unexpected and we may want to find out why this is.

Finally, 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 (cor = -0.04700939). 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.

The state scoring 510 on sat Math but 480 on sat Verbal is the point deviating the most away from the regression line in plot 1. Even so, 30 points is not a huge discrepency. Other two plots have such a high variance that we could say all are atypical values, or all are normal. As such, I'd say the only atypical trend is the part where some states have as much as 35 percent of its population without a high school diploma.

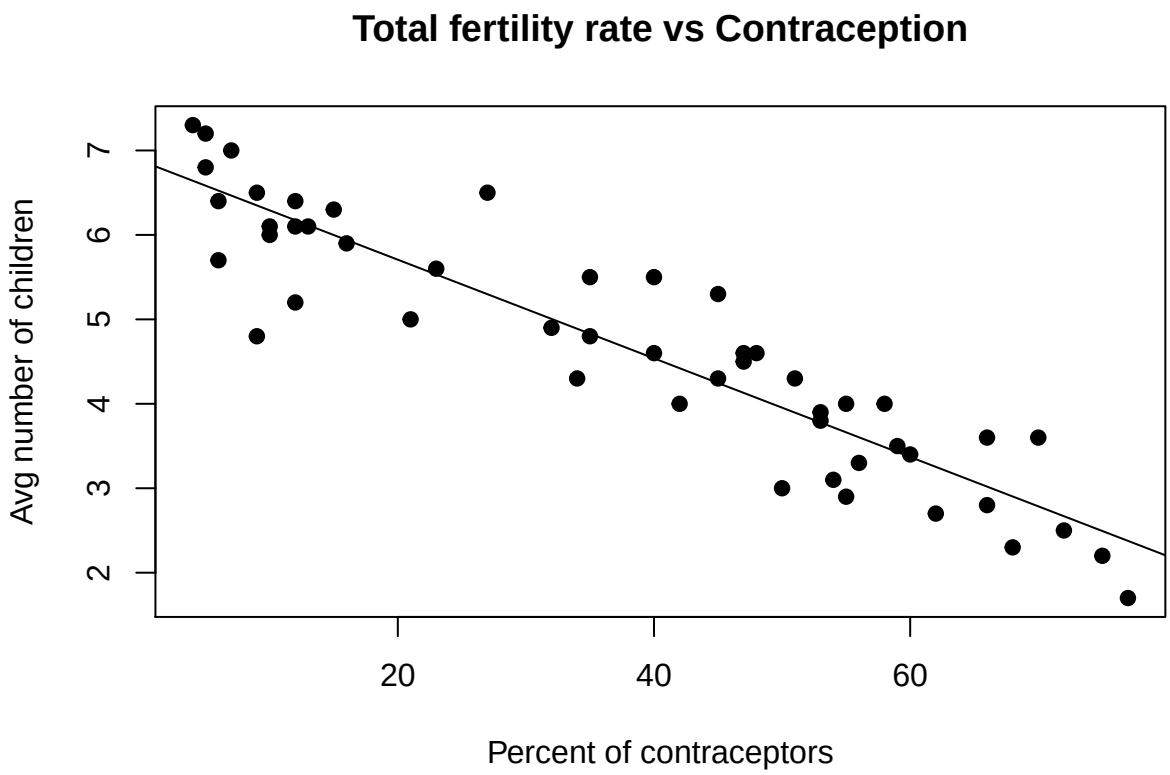
problem 5

Perform a least-squares regression of fertility on contraception from Robery et al.'s data in developing countries. Plot the data and the least-squares line. Does the line adequately summarize the relationship between these variables? Examine and interpret the values of $A, B, S_E,$, and r or r^2.

```
mydata = read.table("C:/Users/biona002/Desktop/biomath_204-master/r_studio_stuff/robey.txt")
total_fertility_rate = mydata[[2]]
contraception = mydata[[3]]

plot(contraception, total_fertility_rate, main="Total fertility rate vs Contraception",
     xlab="Percent of contraceptors", ylab="Avg number of children", pch=19)

lm(total_fertility_rate ~ contraception)
```

```
##
## Call:
## lm(formula = total_fertility_rate ~ contraception)
##
## Coefficients:
##   (Intercept)  contraception
##       6.87509       -0.05842
```

```
abline(lm(total_fertility_rate ~ contraception))
```



Here the X axis describes the percentage of women in childbearing age who use contraceptions, and the Y axis describes the average number of childrens a women gives birth to. Unlike the previous problem, this is of

no surprise because we know that contraception are, in principle, supposed to prevent successful fertilization, and hence a higher contraception rate should lead to a lower number of children.

The line of best fit has Y intercept 6.87509 with slope -0.058416. This means that we predict every unit increase in contraception will cause ≈ 0.05 decrease in the number of children. The correlation between the two variables is -0.9203109, meaning that the negative slope given by the line of best fit is a pretty good predictor.