STA 302f20 Assignment Five[1]

The following problems are not to be handed in. They are preparation for the Quiz in tutorial and the final exam. Please try them before looking at the answers. Use the formula sheet. Please remember that the R parts (Questions 16 and 17 are *not group projects*. You may compare numerical answers, but do not show anyone your code or look at anyone else's.

1. For the general linear regression model in matrix form, find $E(\mathbf{y})$ and $cov(\mathbf{y})$. Show your work.
2. What are the dimensions of the random vector $\widehat{\boldsymbol{\beta}}$? Give the number of rows and the number of columns.
3. Is $\widehat{\boldsymbol{\beta}}$ an unbiased estimator of $\boldsymbol{\beta}$? Answer Yes or No and show your work.
4. Calculate $cov(\widehat{\boldsymbol{\beta}})$ and simplify. Show your work.
5. What are the dimensions of the random vector $\widehat{\mathbf{y}}$?
6. What is $E(\widehat{\mathbf{y}})$? Show your work.
7. What is $cov(\widehat{\mathbf{y}})$? Show your work.
8. What are the dimensions of the matrix $\widehat{\boldsymbol{\epsilon}}$?
9. What is $E(\widehat{\boldsymbol{\epsilon}})$? Show your work. Is $\widehat{\boldsymbol{\epsilon}}$ an unbiased estimator of $\boldsymbol{\epsilon}$? This is a trick question, and requires thought.
10. What is $cov(\widehat{\boldsymbol{\epsilon}})$? Show your work. It is easier if you use $\mathbf{I} - \mathbf{H}$.
11. This is the simplest case of the Gauss-Markov Theorem. Let $Y_1, \ldots, Y_n$ be independent random variables with $E(Y_i) = \mu$ and $Var(Y_i) = \sigma^2$ for $i = 1, \ldots, n$.
    (a) Write down $E(\overline{Y})$ and $Var(\overline{Y})$.
    (b) Let $c_1, \ldots, c_n$ be constants and define the linear combination L by $L = \sum_{i=1}^{n} c_i Y_i$.Recall that L unbiased means for μ that $E(L) = \mu$ for *all* real μ. Show that L unbiased for μ implies $\sum_{i=1}^{n} c_i = 1$.
    (c) Is $\overline{Y}$ a special case of L? If so, what are the c_i values?
    (d) What is $Var(L)$?
    (e) Now show that $Var(\overline{Y}) \leq Var(L)$ for every unbiased L, with equality when $L = \overline{Y}$. Hint: $\sum_{i=1}^{n} c_i^2 = \sum_{i=1}^{n} \left(c_i - \frac{1}{n} + \frac{1}{n}\right)^2$.
12. For the general linear model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, suppose we want to estimate the linear combination $\boldsymbol{\ell}'\boldsymbol{\beta}$ based on sample data. The Gauss-Markov Theorem tells us that the most natural choice is also (in a sense) the best choice. This question leads you through an alternative proof of the Gauss-Markov Theorem, modelled on Question 11.

(a) What is the most natural choice for estimating the linear combination $\boldsymbol{\ell}'\boldsymbol{\beta}$?

(b) Show that this estimate is unbiased.

(c) The natural estimator is a *linear* unbiased estimator of the form $\mathbf{c}_0'\mathbf{y}$. What is the $n \times 1$ vector $\mathbf{c}_0$?

(d) Of course there are lots of other possible linear unbiased estimators of $\boldsymbol{\ell}'\boldsymbol{\beta}$. They are all of the form $\mathbf{c}'\mathbf{y}$; the natural estimator $\mathbf{c}_0'\mathbf{Y}$ is just one of these. The best one is the one with the smallest variance, because its distribution is the most concentrated around the right answer. What is $Var(\mathbf{c}'\mathbf{y})$? Show your work.

(e) We insist that $\mathbf{c}'\mathbf{y}$ be unbiased. Show that if $E(\mathbf{c}'\mathbf{y}) = \boldsymbol{\ell}'\boldsymbol{\beta}$ for *all* $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$, we must have $\mathbf{X}'\mathbf{c} = \boldsymbol{\ell}$.

(f) So, the task is to minimize $Var(\mathbf{c}'\mathbf{y})$ by minimizing $\mathbf{c}'\mathbf{c}$ over all $\mathbf{c}$ subject to the constraint $\mathbf{X}'\mathbf{c} = \boldsymbol{\ell}$. As preparation for this, show $(\mathbf{c} - \mathbf{c}_0)'\mathbf{c}_0 = 0$.

(g) Using the result of the preceding question, show

$$\mathbf{c}'\mathbf{c} = (\mathbf{c} - \mathbf{c}_0)'(\mathbf{c} - \mathbf{c}_0) + \mathbf{c}_0'\mathbf{c}_0.$$

(h) Since the formula for $\mathbf{c}_0$ has no $\mathbf{c}$ in it, what choice of $\mathbf{c}$ minimizes the preceding expression? How do you know that the minimum is unique?

The conclusion is that $\mathbf{c}_0'\mathbf{y} = \boldsymbol{\ell}'\widehat{\boldsymbol{\beta}}$ is the Best Linear Unbiased Estimator (BLUE) of $\boldsymbol{\ell}'\boldsymbol{\beta}$.

13. The model for simple regression through the origin is $y_i = \beta x_i + \epsilon_i$, where the x_i are known constants and $\epsilon_1, \ldots, \epsilon_n$ are independent with expected value 0 and variance σ^2. In previous homework, you found the least squares estimate of β to be $\widehat{\beta} = \frac{\sum_{i=1}^n x_i y_i}{\sum_{i=1}^n x_i^2}$.

(a) What is $Var(\widehat{\beta})$?

(b) Let $\widehat{\beta}_2 = \frac{\overline{y}_n}{\overline{x}_n}$.

i. Is $\widehat{\beta}_2$ an unbiased estimator of β? Answer Yes or No and show your work.

ii. Is $\widehat{\beta}_2$ a linear combination of the y_i variables, of the form $L = \sum_{i=1}^n c_i y_i$? Is so, what is c_i?

iii. What is $Var(\widehat{\beta}_2)$?

iv. How do you know $Var(\widehat{\beta}) \leq Var(\widehat{\beta}_2)$? No calculations are necessary.

v. Under what circumstances are the two variances equal?

(c) Let $\widehat{\beta}_3 = \frac{1}{n}\sum_{i=1}^n \frac{y_i}{x_i}$.

i. Is $\widehat{\beta}_3$ an unbiased estimator of β? Answer Yes or No and show your work.

ii. Is $\widehat{\beta}_3$ a linear combination of the y_i variables, of the form $L = \sum_{i=1}^n c_i y_i$? Is so, what is c_i?

iii. What is $Var(\widehat{\beta}_3)$?

iv. How do you know $Var(\widehat{\beta}) \leq Var(\widehat{\beta}_3)$? No calculations are necessary.

v. Under what circumstances are the two variances equal?

14. The set of vectors $\mathcal{V} = \{\mathbf{v} = \mathbf{Xa} : \mathbf{a} \in \mathbb{R}^{k+1}\}$ is the subset of $\mathbb{R}^n$ consisting of linear combinations of the columns of $\mathbf{X}$. That is, $\mathcal{V}$ is the space *spanned* by the columns of $\mathbf{X}$. The least squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ was obtained by minimizing $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ over all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$. Thus, $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$ is the point in $\mathcal{V}$ that is *closest* to the data vector $\mathbf{y}$. Geometrically, $\widehat{\mathbf{y}}$ is the *projection* (shadow) of $\mathbf{y}$ onto $\mathcal{V}$. The hat matrix $\mathbf{H}$ is a *projection matrix*. It projects the image on any point in $\mathbb{R}^n$ onto $\mathcal{V}$. Now we will test out several consequences of this idea.

    (a) The shadow of a point already in $\mathcal{V}$ should be right at the point itself. Show that if $\mathbf{v} \in \mathcal{V}$, then $\mathbf{Hv} = \mathbf{v}$.

    (b) The vector of differences $\widehat{\boldsymbol{\epsilon}} = \mathbf{y} - \widehat{\mathbf{y}}$ should be perpendicular (at right angles) to each and every basis vector of $\mathcal{V}$. How is this related to the formula $\mathbf{X}'\widehat{\boldsymbol{\epsilon}} = \mathbf{0}$?

    (c) Show that the vector of residuals $\widehat{\boldsymbol{\epsilon}}$ is perpendicular to any $\mathbf{v} \in \mathcal{V}$.

    (d) The picture on Slide 27 of Lecture Unit Eight (More Least Squares) suggests that the closest point to $\widehat{\boldsymbol{\epsilon}}$ in $\mathcal{V}$ should be $\mathbf{0}$. Is this true? Answer Yes or No and show your work.

    (e) In the proof of the Gauss-Markov Theorem (see Question 12), $\mathbf{c}$ is a point in $\mathbb{R}^n$. Show that if $E(\mathbf{c}'\mathbf{y}) = \boldsymbol{\ell}'\boldsymbol{\beta}$ for all $\boldsymbol{\beta} \in \mathbb{R}^{k+1}$, then the closest point to $\mathbf{c}$ in $\mathcal{V}$ is $\mathbf{c}_0$.

15. For the general linear regression model, suppose the error terms ϵ_i are independent and normally distributed.

    (a) In this case, what is the distribution of y_i? Just write down the answer without proof.

    (b) Show that the maximum likelihood estimates of $\beta_0, \ldots, \beta_k$ are identical to the least squares estimates. This is an important result.

    (c) Find the maximum likelihood estimate of σ^2. How does it compare to the general version of s^2 on the formula sheet? Is the MLE unbiased?

16. In the United States, admission to university is based partly on high school marks and recommendations, and partly on applicants' performance on a standardized multiple choice test called the Scholastic Aptitude Test (SAT). The SAT has two sub-tests, Verbal and Math. A university administrator selected a random sample of 200 applicants, and recorded the Verbal SAT, the Math SAT and first-year university Grade Point Average (GPA) for each student. The data are available at

    http://www.utstat.toronto.edu/~brunner/data/legal/openSAT.data.txt.

    We seek to predict GPA from the two test scores. Please use R with *matrix operations* to do the following. I found the `as.matrix` function to be useful, since I did not use `attach`.

    (a) Calculate $\widehat{\boldsymbol{\beta}}$. Display the answer, a set of three numbers.

    (b) Predict first-year GPA for a student with a Verbal SAT of 600 and a Math SAT of 700. The answer is a number. Display it on your output.

    (c) Calculate $\widehat{\mathbf{y}}$. You don't have to display all the numbers because there are 200 of them, but calculate and display the sample mean of the $\widehat{y}_i$. Compare $\overline{y}$, which you should also display.

(d) Calculate $\widehat{\boldsymbol{\epsilon}}$. Don't display them all, but compute their sample mean and display that. It is not *exactly* what you expected. Why?

(e) Calculate and display the inner product of $\widehat{\mathbf{y}}$ and $\widehat{\boldsymbol{\epsilon}}$.

(f) Calculate and display the inner product of $\widehat{\boldsymbol{\epsilon}}$ with total SAT score, which is the sum of Verbal SAT and Math SAT. How did you know what to expect?

You can check your answers using the `lm` function. It would be good to prepare a PDF with your complete R input and output, in case you need to hand it in with the quiz.

17. In Faraway's *Linear models with R*, read Chapter One. The main lesson is that there is more to data analysis than the technical material we will cover in STA302. Then read Chapter 2, skipping Section 2.9 on Identifiability. The author seemingly was in a bit of a hurry when he wrote that part. The coverage of projections and the Gauss-Markov Theorem is unexpected in an R book, but by now you should be able to follow it. There are some handy methods for extracting information from R objects, say on page 22.

    Then, using the `lm` function, do Exercise 1 on page 25 and display the answers. The description of the data set is sketchy. The variables are

    - `sex`: 0=male, 1=female
    - `status`: Socioeconomic status score based on parents' occupation
    - `income`: in pounds per week
    - `verbal`: verbal score in words out of 12 correctly defined
    - `gamble`: expenditure on gambling in pounds per year

    For part (f), consider the *difference* between a female and a male with identical values on all the other x variables.

    There currently seems to be a problem with the `faraway` R package. You can obtain the teen gambling data from

    `http://www.utstat.toronto.edu/~brunner/data/legal/teengamb.data.txt`.

    As in Question 16, please prepare a PDF with your complete R input and output, in case you need to hand it in with the quiz.