STA 302f20 Assignment Four[1]

The following problems are not to be handed in. They are preparation for the Quiz on Oct. 8th during tutorial, and for the final exam. Please try them before looking at the answers. Use the formula sheet.

1. Independently for $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, where the β_j are unknown constants, the x_{ij} are known, observable constants, and the ϵ_i are unobservable random variables with expected value zero. Of course, values of the dependent variable y_i are observable. Start deriving the least squares estimates of β_0, β_1 and β_2 by minimizing the sum of squared differences between the y_i and their expected values. I say *start* because you don't have to finish the job. Stop when you have three linear equations in three unknowns, arranged so they are clearly the so-called "normal" equations $\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \mathbf{X}'\mathbf{y}$.

2. Assuming $(\mathbf{X}'\mathbf{X})^{-1}$ exists, solve the normal equations for the general case of k predictor variables, obtaining $\widehat{\boldsymbol{\beta}}$.

3. For the regression model $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$ etc.,
   (a) Differentiate and simplify to obtain the first normal equation.
   (b) Realizing that the least-squares estimates must satisfy this equation, put hats on the β_j parameters.
   (c) Defining "predicted" y_i as $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik}$, show that $\sum_{i=1}^n \widehat{y}_i = \sum_{i=1}^n y_i$.
   (d) The *residual* for observation i is defined by $\widehat{\epsilon}_i = y_i - \widehat{y}_i$. Show that the sum of residuals equals exactly zero.
   (e) What is $\widehat{y}$ when $x_1 = \overline{x}_1, x_2 = \overline{x}_2, \ldots, x_k = \overline{x}_k$? Show your work.
   (f) Thus, the least squares plane passes through the point $(\overline{x}_1, \overline{x}_2, \ldots, \overline{x}_k, \_\_\_)$. Fill in the blank. You have shown that predicted y for average x values is exactly average y, and this fact does not depend upon the data at all.

4. For the general regression model of Question 3, show that $SST = SSR + SSE$; see the formula sheet for definitions. I find it helpful to switch to matrix notation partway through the calculation.

5. It is possible to think of the total variation in the y_i not as variation around $\overline{y}$, but as variation around zero. This would make sense if the y_i were differences, like weight loss or increase in profits. Then, variation of y_i around zero can be split into variation of y_i around $\overline{y}$, plus variation of $\overline{y}$ around zero.
   (a) Prove $\sum_{i=1}^n (y_i - 0)^2 = \sum_{i=1}^n (y_i - \overline{y})^2 + \sum_{i=1}^n (\overline{y} - 0)^2$.
   (b) Propose a version of R^2 for this setting.

6. In the *centered* linear regression model, sample means are subtracted from the explanatory variables, so that values above average are positive and values below average are negative. Here is a version with one explanatory variable. For $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \epsilon_i$, where the x_i values are fixed constants, and so on.

   (a) Find the least squares estimates of β_0 and β_1. The answer is a pair of formulas. Show your work.

   (b) Because of the centering, it is possible to verify that the solution actually *minimizes* the sum of squares, using only single-variable second derivative tests. Do this part too.

   (c) In an x, y scatterplot, centering x just slides the cloud of points over to the left or right. Should the slope of the least squares line be affected? Comparing your answer to Question 6a to the formula for $\widehat{\beta}_1$ for the uncentered model on the formula sheet, what do you see?

7. Consider the centered multiple regression model

$$y_i = \beta_0 + \beta_1(x_{i,1} - \overline{x}_1) + \cdots + \beta_k(x_{i,k} - \overline{x}_k) + \epsilon_i$$

   with the usual details.

   (a) What is the least squares estimate of β_0? Show your work.

   (b) What is the connection to Problem 3?

8. For the general linear regression model in matrix form,

   (a) Show (there is no difference beween "show" and "prove") that the matrix $\mathbf{X}'\mathbf{X}$ is symmetric. You may use without proof the fact that the transpose of an inverse is the inverse of the transpose.

   (b) Show that $\mathbf{X}'\mathbf{X}$ is non-negative definite.

   (c) Show that if the columns of $\mathbf{X}$ are linearly independent, then $\mathbf{X}'\mathbf{X}$ is positive definite.

   (d) Show that if $\mathbf{X}'\mathbf{X}$ is positive definite, then $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

   (e) Show that if $(\mathbf{X}'\mathbf{X})^{-1}$ exists, then the columns of $\mathbf{X}$ are linearly independent.

   This is a good problem because it establishes that the least squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ exists if and only if the columns of $\mathbf{X}$ are linearly independent, meaning that no predictor variable is a linear combination of the other ones.

9. For the general linear regression model in matrix form with the columns of $\mathbf{X}$ linearly independent as usual, show that $(\mathbf{X}'\mathbf{X})^{-1}$ is positive definite. You may use the existence and properties of $\boldsymbol{\Sigma}^{-1/2}$ without proof.

10. In the matrix version of the general linear regression model, $\mathbf{X}$ is $n \times (k+1)$ and $\mathbf{y}$ is $n \times 1$.

   (a) What are the dimensions of the hat matrix $\mathbf{H}$? Give the number of rows and the number of columns.
   (b) Show that $\mathbf{H}$ is symmetric.
   (c) Show that $\mathbf{H}$ is idempotent, meaning $\mathbf{H} = \mathbf{H}^2$
   (d) Using $tr(\mathbf{AB}) = tr(\mathbf{BA})$, find $tr(\mathbf{H})$.
   (e) Show that if $\mathbf{H}$ has an inverse, $\mathbf{H} = \mathbf{I}$.
   (f) Assuming that the column of $\mathbf{X}$ are linearly independent (and we always do), what is the rank of $\mathbf{H}$?
   (g) Show that $\widehat{\mathbf{y}} = \mathbf{Hy}$.
   (h) Show that $\widehat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$.
   (i) Show that $\mathbf{I} - \mathbf{H}$ is symmetric.
   (j) Show that $\mathbf{I} - \mathbf{H}$ is idempotent
   (k) What is $tr(\mathbf{I} - \mathbf{H})$?
   (l) Show $(\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$.

11. Prove that $\mathbf{X}'\widehat{\boldsymbol{\epsilon}} = \mathbf{0}$. If the statement is false (not true in general), explain why it is false when $k > 2$.

12. In all practical applications, the sample size is larger than the number of regression coefficients: $n > k+1$. But suppose for once that $n = k+1$ and the columns of $\mathbf{X}$ are still linearly independent. This means that $\mathbf{X}^{-1}$ could be obtained by elementary row reduction, proving that $\mathbf{X}^{-1}$ exists. So, for this weird case,

   (a) What is $\widehat{\boldsymbol{\beta}}$?
   (b) What is $\mathbf{H}$?
   (c) What is $\widehat{\mathbf{y}}$?
   (d) What is $\widehat{\boldsymbol{\epsilon}}$?
   (e) How do you know that all the points are exactly on the best-fitting plane?
   (f) For simple regression with an intercept, what is n?
   (g) Are all the points exactly on the least squares line?

13. Returning to the matrix version of the linear model and writing $Q(\boldsymbol{\beta}) = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$,

   (a) Show that $Q(\boldsymbol{\beta}) = \widehat{\boldsymbol{\epsilon}}'\,\widehat{\boldsymbol{\epsilon}} + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\mathbf{X}'\mathbf{X})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$.
   (b) Why does this imply that the minimum of $Q(\boldsymbol{\beta})$ occurs at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$?
   (c) The columns of $\mathbf{X}$ are linearly independent. Why does linear independence guarantee that the minimum is unique?

14. "Simple" regression is just regression with a single predictor variable. The model equation is $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$. Fitting this simple regression problem into the matrix framework of the general linear regression model,

    (a) What is the $\mathbf{X}$ matrix?

    (b) What is $\mathbf{X}'\mathbf{X}$?

    (c) What is $\mathbf{X}'\mathbf{y}$?

    (d) What is $(\mathbf{X}'\mathbf{X})^{-1}$?

15. Show that for simple regression, the proportion of explained sum of squares is the square of the correlation coefficient. That is, $R^2 = \frac{SSR}{SST} = r^2$.

16. In Question 14, the model had an intercept and one predictor variable. But suppose the model has no intercept. This is called simple *regression through the origin*. The model equation would be $y_i = \beta_1 x_i + \epsilon_i$.

    (a) What is the $\mathbf{X}$ matrix?

    (b) What is $\mathbf{X}'\mathbf{X}$?

    (c) What is $\mathbf{X}'\mathbf{y}$?

    (d) What is $(\mathbf{X}'\mathbf{X})^{-1}$?

    (e) What is $\widehat{\boldsymbol{\beta}}$?

17. There can even be a regression model with an intercept but no predictor variables. In this case the model equation is $y_i = \beta_0 + \epsilon_i$.

    (a) Find the least squares estimator $\widehat{\beta}_0$ with calculus.

    (b) Find the least squares estimator $\widehat{\beta}_0$ without calculus, using Problem 13 as a model.

    (c) What is the $\mathbf{X}$ matrix?

    (d) What is $\mathbf{X}'\mathbf{X}$?

    (e) What is $\mathbf{X}'\mathbf{y}$?

    (f) What is $(\mathbf{X}'\mathbf{X})^{-1}$?

    (g) Verify that your expression for $\widehat{\beta}_0$ agrees with $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.

    (h) What is $\widehat{\mathbf{y}}$? What are its dimensions?

18. For the general linear regression model,

    (a) Show that $s^2 = \frac{\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}}{n-k-1}$ is an unbiased estimator of σ^2.

    (b) What is the connection of this s^2 to the usual s^2?