STA 302f20 Assignment Eleven[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. Be ready for R questions similar to the ones asked in this assignment. You will not be asked to hand in your complete answers to the R parts of this assignment, but you might be asked to do something similar on the quiz.

1. In *generalized least squares*, the regression model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\mathbf{V})$. The matrix $\mathbf{V}$ is a *known* symmetric positive definite matrix.
   (a) Let $\widehat{\boldsymbol{\beta}}$ denote the ordinary least squares estimate. Is $\widehat{\boldsymbol{\beta}}$ still an unbiased estimator of $\boldsymbol{\beta}$ for this problem?
   (b) What is $cov(\widehat{\boldsymbol{\beta}})$ for this problem?
   (c) Multiply $\mathbf{y} = \mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}$ on the left by $\mathbf{V}^{-1/2}$, obtaining $\mathbf{y}^* = \mathbf{X}^*\boldsymbol{\beta}+\boldsymbol{\epsilon}^*$. What is the distribution of $\boldsymbol{\epsilon}^*$?
   (d) Substituting $\mathbf{X}^*$ and $\mathbf{y}^*$ into the usual formula for $\widehat{\boldsymbol{\beta}}$, obtain the generalized least squares estimate on page 165 of the textbook. The textbook just calls it $\widehat{\boldsymbol{\beta}}$, but we will call it $\widehat{\boldsymbol{\beta}}_{gls}$.
   (e) What is the distribution of $\widehat{\boldsymbol{\beta}}_{gls}$? Show your work.
   (f) Suppose you guessed wrong about $\mathbf{V}$, and the true distribution of $\boldsymbol{\epsilon}$ is $N_n(\mathbf{0}, \sigma^2\boldsymbol{\Omega})$, where $\boldsymbol{\Omega}$ is symmetric and positive definite. Is $\widehat{\boldsymbol{\beta}}_{gls}$ still an unbiased estimator of $\boldsymbol{\beta}$?
   (g) In terms of the parameters of the original model, please express the following in terms of the original model.
      i. $\mathbf{H}^*$. Is $\mathbf{H}^*$ symmetric and idempotent?
      ii. $\widehat{\mathbf{y}}^*$.
      iii. $\widehat{\boldsymbol{\epsilon}}^*$.
      iv. Show that $SSE^* = (\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{gls})^\prime\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}}_{gls})$.
      v. Write the F statistic for the general linear test of $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$. Use the notation $\widehat{\boldsymbol{\beta}}_{gls}$; otherwise it's too ugly.

[1] This assignment was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. The LaTeX source code is available from the course website: http://www.utstat.toronto.edu/~brunner/oldclass/302f20

2. In *weighted least squares*, the model is the same as in Question 1, except that the matrix $\mathbf{V}$ is diagonal. In scalar form, $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_k x_{i,k} + \epsilon_i$, where the ϵ_i are independent $N(0, \sigma^2 v_i)$. The constants $v_1, \ldots, v_n$ are known, and strictly greater than zero. As usual, σ^2 is unknown. This means that the variances of the y_i are unequal, but known up to a proportionality constant.

   (a) What is the distribution of y_i?

   (b) Multiply both sides of the regression equation by $\frac{1}{\sqrt{v_i}}$, obtaining $y_i^* = \beta_0 x_{i,0}^* + \beta_1 x_{i,1}^* + \cdots + \beta_k x_{i,k}^* + \epsilon_i^*$.

      i. What is $x_{i,0}^*$?

      ii. What is the joint distribution of $\epsilon_1^*, \ldots, \epsilon_n^*$?

      iii. What is the distribution of y_i^*? Are they independent?

   (c) The R help file for `lm` claims that the weighted least squares estimates of $\beta_0, \ldots, \beta_k$ are obtained by minimizing `sum(w*e^2)`. In our notation, they mean minimizing $\sum_{i=1}^n w_i \widehat{\epsilon}_i^2$ over $\widehat{\beta}_0, \ldots, \widehat{\beta}_k$. Show that this yields the least squares estimates for our starred model. In our notation, what are the "weights" $w_1, \ldots, w_n$?

3. In lecture, there was an example in which $y_{ij} \stackrel{i.i.d.}{\sim} ?(\mu, \sigma^2)$, and we observe $\overline{y}_1, \ldots, \overline{y}_m$ along with $n_1, \ldots, n_m$. Verify that the estimator of μ from lecture, $\frac{\sum_{j=1}^m n_j \overline{y}_j}{\sum_{j=1}^m n_j}$, is equal to $\widehat{\boldsymbol{\beta}}_{gls}$.

4. Independently for $i = 1, \ldots, n$, let $y_i = \beta x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2 x_i^2)$.

   (a) Based on the approach of Problem 2, obtain the weighted least squares estimate of β. The answer is a formula for computing $\widehat{\beta}$. There's a way that uses calculus (minimizing the quantity of Question 2c), and another way that relies on known results. Either one is fine.

   (b) Verify that the answer is a special case of $\widehat{\boldsymbol{\beta}}_{gls}$.

5. Data from the example of Question 4 are available here. The URL is `http://www.utstat.toronto.edu/~brunner/data/legal/TinyWLS.data.txt` .

   (a) Estimate β by ordinary least squares, which is the default for the `lm()` function. What is $\widehat{\beta}$? Can you reject $H_0 : \beta = 0$ at $\alpha = 0.05$?

   (b) Now estimate β by weighted least squares. What is $\widehat{\beta}_{wls}$? This time, can you reject $H_0 : \beta = 0$ at $\alpha = 0.05$?

   (c) Verify that $\widehat{\beta}_{wls}$ agrees with your answer to Question 4.

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. The uncentered model for y_i equals the centered model for y_i, as follows.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\ &= \alpha_0 + \alpha_1 (x_i - \overline{x}) + \epsilon_i \end{aligned}$$

   (a) Give α_0 and α_1 in terms of β_0 and β_1. Show your work.

   (b) What does the $\mathbf{X}$ matrix look like for the uncentered model? What does the corresponding matrix look like for the centered model?

   (c) Centering can be accomplished by matrix multiplication. For this simple regression example, give the matrix $\mathbf{A} = (a_{ij})$ such that $\mathbf{XA} = \mathbf{W}$, where $\mathbf{W}$ is the explanatory variable matrix for the centered model.

   (d) Verify that

$$\left(\begin{array}{c} \alpha_0 \\ \alpha_1 \end{array} \right) = \left(\begin{array}{cc} a_{11} & a_{12} \\ a_{21} & a_{22} \end{array} \right)^{-1} \left(\begin{array}{c} \beta_0 \\ \beta_1 \end{array} \right).$$

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. Find the least squares estimate of β_0 by differentiating; show your work. Don't forget the second derivative test.

8. For the general regression model with k predictor variables, centering is also a one-to-one linear transformation. In general,

$$\begin{aligned} & \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \Longleftrightarrow \quad & \mathbf{y} = \mathbf{XAA}^{-1}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \Longleftrightarrow \quad & \mathbf{y} = \mathbf{W}\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \end{aligned}$$

where $\mathbf{A}$ is a $(k+1) \times (k+1)$ matrix, $\mathbf{W} = \mathbf{XA}$ and $\boldsymbol{\alpha} = \mathbf{A}^{-1}\boldsymbol{\beta}$.

   (a) Denoting the least-squares estimate of $\boldsymbol{\alpha}$ by $\widehat{\boldsymbol{\alpha}}$, find a formula for $\widehat{\boldsymbol{\alpha}}$. Simplify. What is its connection to $\widehat{\boldsymbol{\beta}}$?

   (b) What is the vector of predicted y values for the transformed model? How does it compare to $\widehat{\mathbf{y}}$ from the original model?

   (c) Give a null hypothesis equivalent to $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$, but in terms of the transformed model. It's $H_0 : \mathbf{C}_2\boldsymbol{\alpha} = \mathbf{t}$. What is $\mathbf{C}_2$?

   (d) Compare the F^* statistics for testing $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$ and $H_0 : \mathbf{C}_2\boldsymbol{\alpha} = \mathbf{t}$. One would hope they are the same. Are they? Show your work.

9. This question will re-use the `sales` data from Assignment Nine. For reference, here is the description of the data again.

   Telephone sales representatives use computer software to help them locate potential customers, answer questions, take credit card information and place orders. Twelve sales representatives were randomly assigned to each of three new software packages the company was thinking of purchasing. The data for each sales representative include the software package (1, 2 or 3), sales last quarter with the old software, and sales this quarter with one of the new software packages. Sales are in number of units sold. The response variable is sales this quarter.

   The data are available here. The URL is

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

   The beautiful plot on slide 13 of lecture unit 22 (The Centered Model) is actually a plot of these data. The effect of software package on sales this quarter certainly seems to depend on sales last quarter. So, all the regression models for this question should include interactions.

   (a) Fit an uncentered model. Test for differences in expected sales for the three software packages for sales representatives whose sales last quarter were average – that is, exactly at the overall sample mean. Obtain the F statistic, the p-value, and so on. What do you conclude?

   (b) Now carry out the same test on a centered model in which Sales Last Quarter is transformed by subtracting off the sample mean. Note that the product terms in your model ar products of the dummy variable and *centered* sales last quarter. Is the F statistic equal to what you got for the uncentered model?

   (c) The coloured scatterplot suggests that Software Package Two might be the best for sales reps with below average performance last quarter, and worst for sales reps with above average performance last quarter. But is it significant? Accordingly, test for differences among the three packages when Sales Last Quarter = 85. You have to admit that nothing in the theory depends on the centering being done at exactly $\overline{x}$, so do this job by "centering" Sales Last Quarter at 85. What do you conclude?

   (d) Two of the three pairwise comparisons at $x = 85$ are available from the output of `summary`. Carry out the third pairwise comparison. With a Bonferroni correction for all pairwise comparisons and in plain, non-statistical language, what do you conclude?