STA 302f16 Assignment Eleven[1]

Except for Question 1, these questions are preparation for the quiz in tutorial on Thursday December 1st, and are not to be handed in. Please bring your printout for Problem 1 to the quiz.

1. 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 data are in `sales.data.txt`. Get the data with

   ```
   sales = read.table("http://www.utstat.toronto.edu/~brunner/data/legal/sales.data.txt",header=T).
   ```

   The independent and dependent variables are what you would think.

   (a) Fit a full model in which the slopes and intercepts of the regression lines relating sales last quarter to sales this quarter might depend on the kind of software the sales representatives are using.

   (b) Carry out an ordinary F-test to determine whether the effect of software type on sales depends on the representative's performance last quarter. Be able to state your conclusion in plain, non-statistical language.

   (c) Estimate the slopes of the three regression lines. Make sure these numbers are on your printout. I don't see how you can do this without making a table.

   (d) Carry out tests to answer these questions. If they are already on the output of `summary`, use that.

      i. Are the slopes for Software 1 and 2 different?
      ii. Are the slopes for Software 1 and 3 different?
      iii. Are the slopes for Software 2 and 3 different?

      Protecting the three tests with a Bonferroni correction at the joint 0.05 significance level, what do you conclude? Plain language is not necessary, but you should say what happened.

   (e) The average (sample mean) performance last quarter was 76.56 (please use exactly this number). We are interested in whether the three software packages differ in their effectiveness for sales representatives with average performance last quarter.

      i. Estimate expected performance this quarter for sales representatives with average performance last quarter. These three numbers should appear on your printout.
      ii. State the null hypothesis in symbols.
      iii. Carry out the F-test.
      iv. In plain language, what do you conclude?

[1] Copyright information is at the end of the last page.

2. This question explores the practice of centering quantitative independent variables in a regression by subtracting off the mean. Geometrically, this should not alter the configuration of data points in the multi-dimensional scatterplot. All it does is shift the axes. Thus, the intercept of the least squares plane should change, but the slopes should not.

   (a) Consider a simple experimental study with an experimental group, a control group and a single quantitative covariate. Independently for $i = 1, \ldots, n$ let

   $$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \epsilon_i,$$

   where x_i is the covariate and d_i is an indicator dummy variable for the experimental group. If the covariate is "centered," the model can be written

   $$y_i = \beta_0^* + \beta_1^*(x_i - \overline{x}) + \beta_2^* d_i + \epsilon_i,$$

   where $\overline{x} = \frac{1}{n}\sum_{i=1}^n x_i$.

      i. Express the β^* quantities in terms of the original β quantities.

      ii. Let's generalize this. For the general linear model in matrix form suppose $\boldsymbol{\beta}^* = \mathbf{A}\boldsymbol{\beta}$, where $\mathbf{A}$ is a square matrix with an inverse. This makes $\boldsymbol{\beta}^*$ a one-to-one function of $\boldsymbol{\beta}$. Of course X is affected as well. Show that $\mathbf{b}^* = \mathbf{A}\mathbf{b}$.

      iii. Give the matrix $\mathbf{A}$ for this $p = 3$ model.

      iv. If the data are centered, what is $E(y|x)$ for the experimental group, and what is $E(y|x)$ for the control group?

   (b) In the following model, there are k quantitative independent variables. The un-centered version is

   $$y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_k x_{i,k} + \epsilon_i,$$

   and the centered version is

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

   where $\overline{x}_j = \frac{1}{n}\sum_{i=1}^n x_{i,j}$ for $j = 1, \ldots, k$.

      i. What is β_0^* in terms of the β quantities?

      ii. What is β_j^* in terms of the β quantities?

      iii. What is $\widehat{\beta}_0$ in terms of the $\widehat{\beta}^*$ quantities?

      iv. Using $\sum_{i=1}^n \widehat{y}_i = \sum_{i=1}^n y_i$, show that $\widehat{\beta}_0^* = \overline{y}$.

   (c) Now consider again the study with an experimental group, a control group and a single covariate. This time the interaction is included.

   $$y_i = \beta_0 + \beta_1 x_i + \beta_2 d_i + \beta_3 x_i d_i + \epsilon_i$$

   The centered version is

   $$y_i = \beta_0^* + \beta_1^*(x_i - \overline{x}) + \beta_2^* d_i + \beta_3^*(x_i - \overline{x}) d_i + \epsilon_i$$

      i. Express the β^* quantities from the centered model in terms of the β quantities from the un-centered model. Is the correspondence one to one?

ii. For the un-centered model, what is the difference between $E(y|X = \overline{x})$ for the experimental group and $E(y|X = \overline{x})$ for the control group?

iii. What is the difference between intercepts for the centered model? Compare this to your answer to Question 2(c)ii.

3. In the regression model $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, let $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2\Omega)$, with Ω a *known* symmetric positive definite matrix.

   (a) Is $\mathbf{b}$ still an unbiased estimator of $\boldsymbol{\beta}$ for this problem?

   (b) What is $cov(\mathbf{b})$ for this problem?

   (c) Multiply $\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$ on the left by $\Omega^{-1/2}$, obtaining $\mathbf{y}^* = X^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$. What is the distribution of $\boldsymbol{\epsilon}^*$? (Note that the meaning of the "*" symbol is different from its meaning in Question 2, except that in both cases it refers to a transformed version..)

   (d) Substituting X^* and $\mathbf{y}^*$ into the formulas for $\mathbf{b}$, obtain the generalized least squares estimate $\mathbf{b}_{gls} = (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}\mathbf{y}$ on page 133 of the textbook. If you look in the textbook, I think you will appreciate the notation we are using.

   (e) What is the distribution of $\mathbf{b}_{gls}$? Show your work. All you have to do is calculate the expected value and covariance matrix, but *why*? What specific fact on the formula sheet are you using?

   (f) Just realize that there's more. You could obtain formulas for the starred versions of H, $\widehat{\mathbf{y}}$, $\mathbf{e}$, and F statistic for the general linear test.

4. For a very simple aggregated data set, our data are a collection of sample means $\overline{y}_1, \ldots, \overline{y}_n$. Data values in the *unaggregated* data set come from a distribution with common mean μ and common variance σ^2. Sample mean i is based on m_i observations, so that (approximately by the Central Limit Theorem), $\overline{y}_i \sim N(\mu, \frac{\sigma^2}{m_i})$.

   (a) One could estimate μ with the arithmetic mean of the sample means. Is this estimator unbiased? What is its variance?

   (b) Start with the regression-like equation $\overline{y}_i = \mu + \epsilon_i$, where $\epsilon_i \sim N(\mu, \frac{\sigma^2}{m_i})$. Multiply both sides by $\sqrt{n_i}$, obtaining a starred version of the regression equation. Give the generalized (weighted) least squares estimate of μ.

   (c) If you had access to the unaggregated data, how would you estimate μ? What is the connection of this statistic to the weighted least squares estimate?

Please bring your printout for Question 1 to the quiz. **Your printout should show *all* R input and output, and *only* R input and output**. Do not write anything on your printouts except your name and student number.
