STA 2101/442f12 Assignment Ten[1]

Please bring your log and list files for Question 3 to the quiz. The non-computer parts are just practice for the quiz, and are not to be handed in. Any necessary formulas will be provided.

1. This question explores the practice of "centering" quantitative explanatory variables in a regression by subtracting off the mean.

   (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 β quantities.
   ii. If the data are centered, what is $E(Y|x)$ for the experimental group compared to $E(Y|x)$ for the control group?
   iii. By the invariance principle, what is $\widehat{\beta}_0$ in terms of $\widehat{\beta}'$ quantities? Assume ϵ_i normal if you wish.

   (b) In this model, there are $p-1$ quantitative explanatory variables. The un-centered version is

   $$Y_i = \beta_0 + \beta_1 x_{i,1} + \ldots + \beta_{p-1} x_{i,p-1} + \epsilon_i,$$

   and the centered version is

   $$Y_i = \beta_0' + \beta_1'(x_{i,1} - \overline{x}_1) + \ldots + \beta_{p-1}'(x_{i,p-1} - \overline{x}_{p-1}) + \epsilon_i,$$

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

   i. What is β_0' in terms of the β quantities?
   ii. What is β_j' in terms of the β quantities?
   iii. By the invariance principle, what is $\widehat{\beta}_0$ in terms of the $\widehat{\beta}'$ quantities? Assume ϵ_i normal if you wish.
   iv. Show that $\widehat{\beta}_0' = \overline{Y}$. Hint: Differentiate the log likelihood.

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

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

ii. What is the difference between intercepts for the centered model?

(d) Suppose that in the study with an experimental group, a control group and a single covariate, the response variable is binary and we are doing a logistic regression.

i. Under the un-centered model, if there is no interaction, the odds of $Y = 1$ are ___ times as great for the experimental group, for any fixed value of x.

ii. Under the *centered* model, if there is no interaction, the odds of $Y = 1$ are ___ times as great for the experimental group, for any fixed value of x.

iii. If there *is* an interaction and $x = \overline{x}$, the odds of $Y = 1$ for the experimental group are ___ times as great. Express the answer in terms of β values, and also in terms of β' values.

2. This question will be a lot easier if you remember that if $X \sim \chi^2(\nu)$, then $E(X) = \nu$ and $Var(X) = 2\nu$. You don't have to prove this; just use it. You can also use things you already know about ordinary linear regression with normal errors.

For the usual linear regression model with normal errors, σ^2 is usually estimated with MSE.

(a) Show that MSE is an unbiased estimator of σ^2.

(b) Show that MSE is a consistent estimator of σ^2.

(c) Under the usual regression model what is the joint distribution of $\epsilon_1, \ldots, \epsilon_n$?

(d) Let $T_n = \frac{1}{n}\sum_{i=1}^n \epsilon_i^2$. What is $E(T_n)$?

(e) How do you know that $T_n \stackrel{p}{\rightarrow} \sigma^2$?

(f) Show that $Var(T_n) < Var(MSE)$.

(g) So it would appear that T_n is a better estimator of σ^2 than MSE is, since they are both unbiased and the variance of T_n is lower. So why do you think MSE is used in regression analysis instead of T_n?

3. In Question 9 of Assignment 7, you analyzed the Birth Weight data with R. This time we will use SAS. There is a link to the data from our Data Sets page. In Assignment 7 the mother's age did not do much, so the variables we will use this time are

   - Mother's weight in pounds at her last period (`lwt`)
   - Mother's race (`race`: 1 = white, 2 = black, 3 = other)
   - Baby's birth weight in grams (`bwt`)

   (a) First, fit a model with parallel regression lines for the three racial groups. For all the hypothesis tests, be able to give the value of the test statistic, the p-value, whether you reject H_0 at $\alpha = 0.05$, and stte the conclusion in plain, non-statistical language.

      i. What proportion of the variation in baby's weight is explained by the mother's weight and race together?

      ii. Controlling for mother's weight, is mother's race related to baby's weight?

      iii. If the answer to the last question is Yes, carry out Bonferroni-corrected pairwise comparisons and draw a plain language conclusion.

      iv. Controlling for mother's race, is mother's weight related to baby's weight? If the answer is Yes, be able to say *how* it's related.

      v. For every one pound increase in the mother's weight, the baby's weight (increases, decreases) by ________ grams.

   (b) Now test whether race differences in baby's birth weight *depend* on the mother's weight. In plain language, what do you conclude?

   (c) Before proceeding with the data analysis, let's do a little thinking about Studentized deleted residuals. As discussed in lecture, the Studentized deleted residuals have a t distribution under the assumption that the observation in question comes from the same population as the other $n-1$. Thus, each Studentized deleted residual may be treated as the test statistic of a t-test. Get any requested numbers with `proc iml`.

      i. For this data set, what is the critical value at $\alpha = 0.05$? Please don't do any adjustments for multiple tests, yet.

      ii. How many (absolute valued) Studentized deleted residuals would you expect to fall beyond this critical value just by chance if the model is correct for all n observations? The answer is a number (not an integer). Look at the next question to see the reasoning.

      iii. Write the number of Studentized deleted residuals beyond the critical value as a sum of random variables, then take the expected value. This shows that the non-independence of these random variables has no effect on the *expected* number beyond the critical value.

iv. And indeed the random variables are not independent. There is one for each hypothesis test, and the test statistics have almost the same $\widehat{\beta}$ and MSE. How many test statistics are there? The answer is a number.

v. Suppose we want to protect all the tests against Type I error at *joint* significance level 0.05 with a Bonferroni correction? What critical values of t should we use? The answer is a number – well, a pair of numbers.

vi. If the model is correct, the probability of getting *any* Studentized deleted residuals beyond the Bonferroni critical value can be no more than _______. That's better! It's helpful to think of detecting outliers as a multiple comparison problem.

(d) Based on the results of Question 3b, you will choose either a model with interactions or without interactions. For that model, generate the Studentized deleted residuals.

i. List all the Studentized deleted residuals that are beyond the the Bonferroni critical values. Is this cause for serious concern?

ii. How about approximate normality of the residuals? Base your assessment on tests using the usual $\alpha = 0.05$ significance level.