STA 2101/442 Assignment Ten[1]

Please bring printouts of your complete SAS log and list files for Question 3 to the quiz; PDF output counts as a list file. Note that the log and list files *must be from the same run of SAS*. The non-computer questions are just practice for the quiz, and are not to be handed in.

1. In a study comparing the effectiveness of different exercise programmes, volunteers were randomly assigned to one of three exercise programmes (A, B, C) or put on a waiting list and told to work out on their own. Aerobic capacity is the body's ability to process oxygen. Aerobic capacity was measured before and after 6 months of participation in the program (or 6 months of being on the waiting list). The response variable was improvement in aerobic capacity. The explanatory variables were age (a covariate) and treatment group.

   (a) First consider a regression model with an intercept, and no interaction between age and treatment group.

      i. Make a table showing how you would set up indicator dummy variables for treatment group. Make Waiting List the reference category

      ii. Write the regression model. Please use x for age, and make its regression coefficient β_1.

      iii. In terms of β values, what null hypothesis would you test to find out whether, allowing for age, the three exercise programmes differ in their effectiveness?

      iv. Write the null hypothesis for the preceding question as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Just give the $\mathbf{L}$ matrix.

      v. In terms of β values, what null hypothesis would you test to find out whether Programme B was better than the waiting list?

      vi. In terms of β values, what null hypothesis would you test to find out whether Programmes A and B differ in their effectiveness?

      vii. Suppose you wanted to estimate the difference in average benefit between programmes A and C for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

      viii. Is it safe to assume that age is independent of the other explanatory variables? Answer Yes or No and briefly explain.

   (b) Now consider a regression model with an intercept and the interaction (actually a set of interactions) between age and treatment.

      i. Write the regression model. Make it an extension of your earlier model.

      ii. Suppose you wanted to know whether the slopes of the 4 regression lines were equal. In terms of β values, what null hypothesis would you test?

      iii. Suppose you wanted to know whether any differences among mean improvement in the four treatment conditions depends on the participant's age. In terms of β values, what null hypothesis would you test?

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

iv. Write the null hypothesis for the preceding question as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Just give the $\mathbf{L}$ matrix. It is $r \times p$. What is r? What is p?

v. Suppose you wanted to know whether the difference in effectiveness between Programme A and the Waiting List depends on the participant's age. In terms of β values, what null hypothesis would you test?

vi. Suppose you wanted to *estimate* the difference in average benefit between programmes A and C for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

(c) Now consider a regression model *without* an intercept, but *with* possibly unequal slopes. Make a table to show how the dummy variables could be set up, and write the regression model. Again, please use x for age and make its regression coefficient β_1. For each treatment condition, what is the conditional expected value of Y? The answer is in terms of x and the β values. Please put these values as the last column of your table.

i. Suppose you wanted to know whether the slopes of the 4 regression lines were equal. In terms of β values, what null hypothesis would you test?

ii. Suppose you wanted to know whether any differences among mean improvement in the four treatment conditions depends on the participant's age. In terms of β values, what null hypothesis would you test?

iii. Write the null hypothesis for the preceding question as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Just give the $\mathbf{L}$ matrix. It is $r \times p$. What is r? What is p?

iv. Suppose you wanted to know whether the difference in effectiveness between Programme A and the Waiting List depends on the participant's age. In terms of β values, what null hypothesis would you test?

v. Suppose you wanted to estimate the difference in average benefit between programmes A and C for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

2. 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 (this takes you back all the way to slide 25 of Likelihood Part One), what is $\widehat{\beta}_0$ in terms of $\widehat{\beta}'$ quantities? Assume ϵ_i is normal.

(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 is normal.

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. 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?

3. The Birth weight data set contains the following information on a sample of mothers who recently had babies.

   Identification code

   indicator of birth weight less than 2.5k

   Mother's age in years

   Mother's weight in pounds at last menstrual period

   Mother's race (1 = white, 2 = black, 3 = other)

   Smoking status during pregnancy

   Number of previous premature labours

   History of hypertension

   Presence of uterine irritability

   Number of physician visits during the first trimester

   Birth weight of baby in grams

   For this question, we will use just Mother's weight, Mother's race and Baby's birth weight.

   (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 state 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 estimated 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?

4. In the following regression model, the explanatory variables X_1 and X_2 are random variables. The true model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i,$$

independently for $i = 1, \ldots, n$, where $\epsilon_i \sim N(0, \sigma^2)$.

The mean and covariance matrix of the explanatory variables are given by

$$E\left[\begin{array}{c} X_{i,1} \\ X_{i,2} \end{array}\right] = \left[\begin{array}{c} \mu_1 \\ \mu_2 \end{array}\right] \quad \text{and} \quad Var\left[\begin{array}{c} X_{i,1} \\ X_{i,2} \end{array}\right] = \left[\begin{array}{cc} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{array}\right]$$

Unfortunately $X_{i,2}$, which has an impact on Y_i and is correlated with $X_{i,1}$, is not part of the data set. Since $X_{i,2}$ is not observed, it is absorbed by the intercept and error term, as follows.

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \epsilon_i \\ &= (\beta_0 + \beta_2\mu_2) + \beta_1 X_{i,1} + (\beta_2 X_{i,2} - \beta_2\mu_2 + \epsilon_i) \\ &= \beta_0' + \beta_1 X_{i,1} + \epsilon_i'. \end{aligned}$$

The primes just denote a new β_0 and a new ϵ_i. It was necessary to add and subtract $\beta_2\mu_2$ in order to obtain $E(\epsilon_i') = 0$. And of course there could be more than one omitted variable. They would all get swallowed by the intercept and error term, the garbage bins of regression analysis.

(a) What is $Cov(X_{i,1}, \epsilon_i')$?

(b) Calculate the variance-covariance matrix of $(X_{i,1}, Y_i)$ under the true model. Is it possible to have non-zero covariance between $X_{i,1}$ and Y_i when $\beta_1 = 0$?

(c) Suppose we want to estimate β_1. The usual least squares estimator is

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)(Y_i - \overline{Y})}{\sum_{i=1}^n (X_{i,1} - \overline{X}_1)^2}.$$

You may just use this formula; you don't have to derive it. Is $\widehat{\beta}_1$ a consistent estimator of β_1 if the true model holds? Answer Yes or no and show your work. You may use the consistency of the sample variance and covariance without proof.

(d) Are there *any* points in the parameter space for which $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ when the true model holds?

5. Consider simple regression through the origin in which the explanatory variable values are random variables rather than fixed constants. But you can't see the explanatory variable. It is a *latent* variable. Instead, all you see is the explanatory variable plus a piece of random noise. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= X_i\beta + \epsilon_i \\ W_i &= X_i + e_i, \end{aligned} \tag{1}$$

where

- X_i has expected value μ_x and variance σ^2_x,
- e_i has expected value 0 and variance σ^2_e
- ϵ_i has expected value 0 and variance σ^2_ϵ
- X_i, ϵ_i and e_i are all independent.

The value of the explanatory variable X_i, like ϵ_i and e_i, is not observable. All we can see are the pairs (W_i, Y_i) for $i = 1, \ldots, n$.

(a) Following common practice, we ignore the measurement error and apply the usual regression estimator with W_i in place of X_i. The parameter β is estimated by

$$\widehat{\beta}_{(1)} = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i^2}$$

Is $\widehat{\beta}_{(1)}$ a consistent estimator of β? Answer Yes, No or Impossible to determine. Show your work.

(b) Consider instead the estimator

$$\widehat{\beta}_{(2)} = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n W_i}.$$

Is $\widehat{\beta}_{(2)}$ a consistent estimator of β? Answer Yes, No or Impossible to determine. Show your work. Does the value of μ matter?

(c) Suppose X_i, ϵ_i and e_i are normally distributed. What is the joint distribution of (W_i, Y_i)? Calculate the vector of expected values and the covariance matrix.

(d) Using the invariance principle, obtain explicit formulas for the MLE of $\boldsymbol{\theta} = (\beta, \mu_x, \sigma^2_x, \sigma^2_e, \sigma^2_\epsilon)'$ without differentiating anything. You may use without proof the fact that the MLE if a general multivariate normal is $(\overline{\mathbf{D}}, \widehat{\boldsymbol{\Sigma}})$, where

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^n (\mathbf{D}_i - \overline{\mathbf{D}})(\mathbf{D}_i - \overline{\mathbf{D}})'.$$

Use symbols like $\widehat{\sigma}_{xw}$ for the sample variances and covariances.