STA 431s13 Assignment Three[1]

For the SAS question, please bring your log and list files to the quiz. Do not write anything on the printouts except your name and student number. The other questions are just practice for the quiz on Feb. 1st, and are not to be handed in. The first part of Chapter 0 from the text is now posted. It may be a helpful supplement to lecture material.

1. Everybody knows that $Var(Y_i) = \sigma^2$ for a regression model, but that's really a conditional variance. Independently for $i = 1, \ldots, n$, let

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

   where $\epsilon_1, \ldots \epsilon_n$ are independent random variables with expected value zero and common variance σ^2, $E(X_{i,1}) = \mu_1$, $Var(X_{i,1}) = \sigma_1^2$, $E(X_{i,2}) = \mu_2$, $Var(X_{i,2}) = \sigma_2^2$, and $Cov(X_{i,1}, X_{i,2}) = \sigma_{12}$. Calculate $Var(Y_i)$; show your work.

2. The usual univariate multiple regression model with independent normal errors is

   $$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

   where $\mathbf{X}$ is an $n \times p$ matrix of known constants, $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants, and $\boldsymbol{\epsilon}$ is multivariate normal with mean zero and covariance matrix $\sigma^2 \mathbf{I}_n$, with $\sigma^2 > 0$ an unknown constant. But of course in practice, the explanatory variables are random, not fixed. Clearly, if the model holds *conditionally* upon the values of the explanatory variables, then all the usual results hold, again conditionally upon the particular values of the explanatory variables. The probabilities (for example, p-values) are conditional probabilities, and the F statistic does not have an F distribution, but a conditional F distribution, given $\mathbf{X} = \mathbf{x}$.

   (a) Show that the least-squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is conditionally unbiased.

   (b) Show that $\widehat{\boldsymbol{\beta}}$ is also unbiased unconditionally.

   (c) A similar calculation applies to the significance level of a hypothesis test. Let F be the test statistic (say for an extra-sum-of-squares F-test), and f_c be the critical value. If the null hypothesis is true, then the test is size α, conditionally upon the explanatory variable values. That is, $P(F > f_c | \mathbf{X} = \mathbf{x}) = \alpha$. Find the *unconditional* probability of a Type I error. Assume that the explanatory variables are discrete, so you can write a multiple sum.

3. Ordinary least squares is often applied to data sets where the independent variables are best modeled as random variables. In what way does the usual conditional linear regression model imply that (random) independent variables have zero covariance with the error term? Hint: Assume $\mathbf{X}_i$ as well as ϵ_i continuous. What is the conditional distribution of ϵ_i given $\mathbf{X}_i = \mathbf{x}_i$?

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

4. Show that $E(\epsilon_i|X_i = x_i) = 0$ for all x_i implies $Cov(X_i, \epsilon_i) = 0$, so that a standard regression model without the normality assumption still implies zero covariance (though not necessarily independence) between the error term and explanatory variables.

5. In a study of diet and health, suppose we want to know how much snack food each person eats, and we "measure" it by asking a question on a questionnaire. Surely there will be measurement error, and suppose it is of a simple additive nature. But we are pretty sure people under-report how much snack food they eat, so a model like $W = X + e$ with $E(e) = 0$ is hard to defend. Instead, let

$$W = \nu + X + e,$$

where $E(X) = \mu$, $E(e) = 0$, $Var(X) = \sigma^2_X$, $Var(e) = \sigma^2_e$, and $Cov(X, e) = 0$ The unknown constant ν could be called *measurement bias*. Calculate the reliability of W for this model. Is it the same as the expression for reliability given in the text and lecture, or does $\nu \neq 0$ make a difference?

6. Continuing Exercise 5, suppose that two measurements of W are available.

$$\begin{aligned} W_1 &= \nu_1 + X + e_1 \\ W_2 &= \nu_2 + X + e_2, \end{aligned}$$

where $E(X) = \mu$, $Var(X) = \sigma^2_T$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma^2_e$, and X, e_1 and e_2 are all independent. Calculate $Corr(W_1, W_2)$. Does this correlation still equal the reliability?

7. Let X be a latent variable, $W = X + e_1$ be the usual measurement of X with error, and $G = X + e_2$ be a measurement of X that is deemed "gold standard," but of course it's not completely free of measurement error. It's better than W in the sense that $0 < Var(e_2) < Var(e_1)$, but that's all you can really say. This is a realistic scenario, because nothing is perfect. Accordingly, let

$$\begin{aligned} W &= X + e_1 \\ G &= X + e_2, \end{aligned}$$

where $E(X) = \mu$, $Var(X) = \sigma^2_X$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = \sigma^2_1$, $Var(e_2) = \sigma^2_2$ and that X, e_1 and e_2 are all independent of one another. Prove that the squared correlation between W and G is strictly less than the reliability. Show your work.

The idea here is that the squared *population* correlation[2] between an ordinary measurement and an imperfect gold standard measurement is strictly less than the actual

[2]When we do Greek-letter calculations, we are figuring out what is happening in the population from which a data set might be a random sample.

reliability of the ordinary measurement. If we were to estimate such a squared correlation by the corresponding squared *sample* correlation, all we would be doing is estimating a quantity that is not the reliability. On the other hand, we would be estimating a lower bound for the reliability — and this could be reassuring if it is a high number.

8. Suppose we have two equivalent measurements with uncorrelated measurement error:

$$\begin{aligned} W_1 &= X + e_1 \\ W_2 &= X + e_2, \end{aligned}$$

where $E(X) = \mu$, $Var(X) = \sigma^2_X$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma^2_e$, and X, e_1 and e_2 are all independent. What if we were to measure the true score X by adding the two imperfect measurements together? Would the result be more reliable?

   (a) Let $S = W_1 + W_2$. Calculate the reliability of S. Is there any harm in assuming $\mu = 0$?

   (b) Suppose you take k independent measurements (in psychometric theory, these would be called equivalent test items). What is the reliability of $S = \sum_{i=1}^{k} W_i$? Show your work.

   (c) What happens as the number of measurements $k \to \infty$?

9. Let X be a latent variable, $W = X + e_1$ be the usual measurement of X with error, and $G = X + e_2$ be a measurement of X that is deemed "gold standard," but of course it's not completely free of measurement error. It's better than W in the sense that $0 < Var(e_2) < Var(e_1)$, but that's all you can really say. This is a realistic scenario, because nothing is perfect. Accordingly, let

$$\begin{aligned} W &= X + e_1 \\ G &= X + e_2, \end{aligned}$$

where $E(X) = \mu$, $Var(X) = \sigma^2_X$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = \sigma^2_1$, $Var(e_2) = \sigma^2_2$ and that X, e_1 and e_2 are all independent of one another. Prove that the squared correlation between W and G is strictly less than the reliability. Show your work.

The idea here is that the squared *population* correlation[3] between an ordinary measurement and an imperfect gold standard measurement is strictly less than the actual reliability of the ordinary measurement. If we were to estimate such a squared correlation by the corresponding squared *sample* correlation, all we would be doing is

[3]When we do Greek-letter calculations, we are figuring out what is happening in the population from which a data set might be a random sample.

estimating a quantity that is not the reliability. On the other hand, we would be estimating a lower bound for the reliability — and this could be reassuring if it is a high number.

10. In this continuation of Exercise 9, show what happens when you calculate the squared *sample* correlation between a usual measurement and an imperfect gold standard. It's just what you would think.

11. Suppose we have two equivalent measurements with uncorrelated measurement error:

$$\begin{aligned} W_1 &= X + e_1 \\ W_2 &= X + e_2, \end{aligned}$$

where $E(X) = \mu$, $Var(X) = \sigma^2_X$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma^2_e$, and X, e_1 and e_2 are all independent. What if we were to measure the true score X by adding the two imperfect measurements together? Would the result be more reliable?

(a) Let $S = W_1 + W_2$. Calculate the reliability of S. Is there any harm in assuming $\mu = 0$?

(b) Suppose you take k independent measurements (in psychometric theory, these would be called equivalent test items). What is the reliability of $S = \sum_{i=1}^{k} W_i$? Show your work.

(c) What happens as the number of measurements $k \to \infty$?

12. In the United States, admission to university is based partly on high school marks and recommendations, and partly on applicants performance on a standardized multiple choice test called the Scholastic Aptitude Test (SAT). The SAT has two sub-tests, Verbal and Math. A university administrator selected a random sample of 200 applicants, and recorded the Verbal SAT, the Math SAT and first-year university Grade Point Average (GPA) for each student. The data are given in the file `opensat.data`. There is a link on the course web page in case the one in this document does not work.

    Using SAS, calculate means and standard deviations for all the variables. That's it. Bring your log file and your list file to the quiz. You may be asked for numbers from your printouts, and you may be asked to hand them in. **There must be no warnings, error messages or notes about missing data on your log file.**
