STA 431s15 Assignment Four[1]

The non-computer questions on this assignment are for practice, and will not be handed in. For the SAS part of this assignment (Question 14) please bring your log file and your output file to the quiz. There may be one or more questions about them, and you may be asked to hand the printouts in with the quiz.

1. 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?

2. Continuing Exercise 1, 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 even when ν_1 and ν_2 are non-zero?

3. 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.

   (a) Make a path diagram of this model.

(b) Prove that the squared correlation between W and G is strictly less than the reliability of W. 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 reliability of the ordinary measurement. If we were to estimate such a squared correlation by the corresponding squared *sample* correlation, we would be 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.

4. In this continuation of Exercise 3, show what happens when you calculate the squared *sample* correlation between a usual measurement and an imperfect gold standard and let $n \to \infty$. It's just what you would think.

5. 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 $E(X) = \mu$, $Var(X) = \sigma_x^2$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma_e^2$, 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.

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

(c) What is the reliability of $\overline{W}_n = \frac{1}{n}\sum_{i=1}^{n} W_i$? Show your work.

(d) What happens to the reliability of S and $\overline{W}_n$ as the number of measurements $n \to \infty$?

6. Consider the two equivalent measurements at the start of Question 5. It is easy to imagine omitted variables that would affect both observed scores. For example, if W_1 and W_2 are two questionnaires about eating habits, some people will probably mis-remember or lie the same way on both questionnaires. Since e_1 and e_2 represent "everything else," this means that e_1 and e_2 will have non-zero covariance. Furthermore, this covariance will be positive, since the omitted variables (there could be dozens of them) will tend to affect the two measurements in the same way. Accordingly, in the initial model of Question 5, let $Cov(e_1, e_2) = \kappa > 0$.

(a) Draw a path diagram of the model.

(b) Show that $Corr(W_1, W_2)$ is strictly *greater* than the reliability.

This means that in practice, omitted variables will result in over-estimates of reliability. And there are always omitted variables.

[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.

7. Let $X_1, \ldots, X_n$ be a random sample from a continuous distribution with density

$$f(x;\theta) = \frac{1}{\theta^{1/2}\sqrt{2\pi}}\, e^{-\frac{x^2}{2\theta}},$$

where the parameter $\theta > 0$. Propose a reasonable estimator for the parameter θ, and use the Law of Large Numbers to show that your estimator is consistent.

8. Let $X_1, \ldots, X_{n_1}$ be a random sample from a distribution with expected value μ and variance σ_x^2. Independently of $X_1, \ldots, X_n$, let $Y_1, \ldots, Y_{n_2}$ be a random sample from a distribution with the same expected value μ and variance σ_y^2. Let $T_n = \alpha\overline{X}_n + (1-\alpha)\overline{Y}_n$, where $0 \leq \alpha \leq 1$.

   (a) Is T_n an unbiased estimator of μ for any value of $\alpha \in [0,1]$? Answer Yes or No and show your work.

   (b) Is T_n a consistent estimator of μ for any value of $\alpha \in [0,1]$? Answer Yes or No and show your work.

   (c) Find the value of α that minimizes the variance of the estimator T_n.

9. Let $X_1, \ldots, X_n$ be a random sample from a Gamma distribution with $\alpha = \beta = \theta > 0$. That is, the density is

$$f(x;\theta) = \frac{1}{\theta^\theta \Gamma(\theta)} e^{-x/\theta} x^{\theta-1},$$

for $x > 0$. Let $\widehat{\theta} = \overline{X}_n$. Is $\widehat{\theta}$ a consistent estimator of θ? Answer Yes or No and prove your answer. Hint: If X has a Gamma distribution with parameters α and β, $E(X) = \alpha\beta$.

10. Let $X_1, \ldots, X_n$ be a random sample from a distribution with mean μ and variance σ^2. Prove that the sample variance $S^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}$ is consistent for σ^2.

11. Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be a random sample from a bivariate distribution with $E(X_i) = \mu_x$, $E(Y_i) = \mu_y$, $Var(X_i) = \sigma_x^2$, $Var(Y_i) = \sigma_y^2$, and $Cov(X_i, Y_i) = \sigma_{xy}$. Show that the sample covariance $S_{xy} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$ is a consistent estimator of σ_{xy}.

12. Independently for $i = 1, \ldots, n$, let

$$Y_i = \beta X_i + \epsilon_i,$$

where $E(X_i) = \mu$, $E(\epsilon_i) = 0$, $Var(X_i) = \sigma_x^2$, $Var(\epsilon_i) = \sigma_\epsilon^2$, and ϵ_i is independent of X_i. The variables X_i and Y_i are both observable.

   (a) Let

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^n Y_i}{\sum_{i=1}^n X_i}.$$

      i. Is $\widehat{\beta}_1$ a consistent estimator of β? Answer Yes or No and justify your answer.

      ii. Does it matter if $\mu = 0$?

   (b) Let

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}.$$

      i. Is $\widehat{\beta}_2$ a consistent estimator of β? Answer Yes or No and prove your answer.

      ii. Does it matter if $\mu = 0$?

13. The Laws of Large Numbers we are using in this class assume that independent observations are being averaged.

    (a) Does our Law of Large Numbers apply to $\overline{W}_n$ of Question 5? Answer Yes or No and *say why.*

    (b) Do we have $\overline{W}_n \stackrel{a.s.}{\to} \mu$, or $\overline{W}_n \stackrel{a.s.}{\to} X$? Prove your answer. Hint: If $X_n \stackrel{a.s.}{\to} X$ and $Y_n \stackrel{a.s.}{\to} Y$, the the vector $(X_n, Y_n)^\top$ converges almost surely to $(X, Y)^\top$.

14. Before the beginning of the Fall term, students in a first-year Calculus class took a diagnostic test with two parts: Pre-calculus and Calculus. Their High School Calculus marks and their marks in University Calculus were also available. In order, the variables in the data file are: Identification code, Mark in High School Calculus, Score on the Pre-calculus portion of the diagnostic test, Score on the Calculus portion of the diagnostic test, and mark in University Calculus[3]. Data are available in the file `mathtest.data.txt`.

    Using SAS `proc calis`, carry out an unconditional regression in which the explanatory variables are Mark in High School Calculus, Score on the Pre-calculus portion of the diagnostic test and Score on the Calculus portion of the diagnostic test. The response variable is mark in University Calculus. All the variables are observable.

    You are fitting just one model, and it is saturated – meaning its parameters are one-to-one with those of an unrestricted multivariate normal model. 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 are lots of "t-tests" (actually, Z-tests). Know what null hypotheses they all are testing. **There must be no error messages, and no notes or warnings about invalid data on your log file.**

[3]Thanks to Cleo Boyd for premission to use these original data.