STA 312F2007 Assignment 2

Do this review assignment in preparation for the quiz on Friday, Sept. 28th. The problems are practice for the quiz, and are not to be handed in.

1. 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 independent variables are random, not fixed. As discussed in lecture, if the model holds *conditionally* upon the values of the independent variables, then all the usual results hold, again conditionally upon the particular values of the independent 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) We know that the least-squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is conditionally unbiased. Show that it is also unbiased unconditionally. This is quick.

   (b) 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 independent variable values. That is, $P(F > f_c | \mathbf{X} = \mathbf{x}) = \alpha$. Find the *unconditional* probability of a Type I error. This is pretty quick too. It is okay to assume that the independent variables are jointly continuous, so you can write an integral.

2. Let $X = T + e$, where X is an observed measurement, T is the true measurement, and e is random error. Assume $E(T) = \mu$, $Var(T) = \sigma^2_T$, $E(e) = 0$, $Var(e) = \sigma^2_e$, and that T and e are independent.

   (a) What is $Var(X)$?

   (b) In classical psychometric theory, the *reliability* of a test is the squared correlation between the observed score and the true score. Recalling the definition $Corr(V, W) = \frac{Cov(V,W)}{SD(V)SD(W)}$, calculate the reliability of X using this definition. You will find the answer in your lecture notes, though we did not do the calculation in class.

   (c) Suppose we make the measurement twice, in such a way that the errors of measurement are independent on the two occasions. We have

$$\begin{aligned} X_1 &= T + e_1 \\ X_2 &= T + e_2, \end{aligned}$$

where $E(T) = \mu$, $Var(T) = \sigma^2_T$, $E(e_1) = E(e_2) = 0$, $Var(e_1) = Var(e_2) = \sigma^2_e$, and T, e_1 and e_2 are all independent. Calculate $Corr(X_1, X_2)$. Compare your answer to 2b. This is the basis of *test-retest reliability*, in which the reliability of an educational or psychological test is estimated by the sample correlation between two independent administrations of the test.

3. Recall the simulation study of inflated Type I error when independent variables are measured with error but one ignores it and uses ordinary regression anyway. We needed to produce correlated (latent, that is unobservable) independent variables ξ_1 and ξ_2. Here's how we did it.

   (a) It is easy to simulate a collection of independent random variables from any distribution, and then standardize them to have expected value zero and variance one. Let $E(X) = \mu$ and $Var(X) = \sigma^2$. Now define $Z = \frac{X-\mu}{\sigma}$. Find

      i. $E(Z)$

      ii. $Var(Z)$

   (b) Okay, now let R_1, R_2 and R_3 be independent random variables from any distribution you like, but standardized to have expected value zero and variance one. Now let

$$\begin{aligned}\xi_1 &= \sqrt{1-\phi}\,R_1 + \sqrt{\phi}\,R_3 \text{ and}\\ \xi_2 &= \sqrt{1-\phi}\,R_2 + \sqrt{\phi}\,R_3.\end{aligned}$$

   Find

      i. $Cov(\xi_1, \xi_2)$

      ii. $Corr(\xi_1, \xi_2)$

   (c) This one is more efficient. Let R_1 and R_2 be independent random variables with expected value zero and variance one. Now let

$$\begin{aligned}\xi_1 &= \sqrt{\frac{1+\phi}{2}}\,R_1 + \sqrt{\frac{1-\phi}{2}}\,R_2\\ \xi_2 &= \sqrt{\frac{1+\phi}{2}}\,R_1 - \sqrt{\frac{1-\phi}{2}}\,R_2\end{aligned}$$

   Find

      i. $Cov(\xi_1, \xi_2)$

      ii. $Corr(\xi_1, \xi_2)$

   In an upcoming lecture, we will see that if the R variables are normal and $\phi = 0$, both methods yield ξ_1 and ξ_2 independent. But if the Rs are non-normal, then $\phi = 0$ only implies independence for the first method.