STA 2053 Assignment 4 (Double measurement regression, surrogate models)[1]

Questions 1 and 3 are not to be handed in. They are practice for the quiz on November 13th. Please bring your complete R input and output for Question 2 to the quiz.

1. Question 2 (the R part of this assignment) will use the *Pig Birth Data*. As part of a much larger study, farmers filled out questionnaires about various aspects of their farms. Some questions were asked twice, on two different questionnaires several months apart. Buried in all the questions were

   - Number of breeding sows (female pigs) at the farm on June 1st
   - Number of sows giving birth later that summer.

   There are two readings of these variables, one from each questionnaire. We will assume (maybe incorrectly) that because the questions were buried in a lot of other material and were asked months apart, that errors of measurement are independent between the two questionnaires. However, errors of measurement might be correlated within a questionnaire.

   (a) Propose a reasonable *surrogate* centered model for these data, using the usual notation. Give all the details. You may assume normality if you wish.

   (b) Make a path diagram of the model you have proposed.

   (c) Calculate the covariance matrix of one observable data vector $\mathbf{d}_i$.

   (d) Even though you have a general result that applies to this case, prove that all the parameters in the covariance matrix are identifiable.

   (e) If there are any equality constraints on the covariance matrix, say what they are.

   (f) Based on your answer to the last question, how many degrees of freedom should there be in the chi-squared test for model fit?

   (g) Give a consistent estimator of β that is *not* the MLE, and explain why it's consistent. You may use the consistency of sample variances and covariances without proof. Your estimator must not be a function of any unknown parameters.

2. The Pig Birth Data are given in the file `openpigs2.data.txt`. There are $n = 114$ farms; please verify that you are reading the correct number of cases.

   (a) Start by reading the data and then using the `var` function to produce a sample covariance matrix of all the observable variables. Don't worry about n versus $n-1$.

   (b) Use `lavaan` to fit your model. Look at `summary`. If you experience numerical problems you are doing something differently from the way I did it. When I fit a good model everything was fine. When I fit a poor model there was trouble. Just to verify that we are fitting the same model, my estimate of the variance of the latent exogenous variable is 357.145.

(c) Does your model fit the data adequately? Answer Yes or No and give three numbers: a chi-squared statistic, the degrees of freedom, and a p-value. Do the degrees of freedom agree with your answer to Question 1f?

(d) If the number of breeding sows present in September increases by one, what happens to the estimated number giving birth that summer? You answer is based on a single number from the output of `summary`. It is not an integer.

(e) Using your answer to Question 1g and the output of `var`, give a *numerical* version of your consistent estimate of β. How does it compare to the MLE?

(f) Give a large-sample confidence interval for your answer to 2d. I used `parameterEstimates` to do it the easy way.

(g) Recall that reliability of a measurement is the proportion of its variance that does *not* come from measurement error. What is the estimated reliability of number of breeding sows? There are two numbers, one for questionnaire one and another for questionnaire two. You could do this with a calculator and the output of `summary`, but I did it with `:=` in the model string.

(h) It would be inconvenient at best to get confidence intervals for reliability with a calculator. Obtain confidence intervals for the reliabilities in Question 2g. Try `parameterEstimates`. It uses the delta method.

(i) Is there evidence of correlated measurement error within questionnaires? Answer Yes or No and give some numbers from the results file to support your conclusion.

(j) The answer to that last question was based on two separate tests. Though it is already pretty convincing, conduct a *single* Wald (not likelihood ratio) test of the two null hypotheses simultaneously. Give the Wald chi-squared statistic, the degrees of freedom and the p-value. What do you conclude? Is there evidence of correlated measurement error, or not?

3. Of course the model of Questions 1 and 2 is a surrogate model.

(a) Give the model equations for the original model, in centered form. It's still double measurement.

(b) Calculate the covariance matrix of the observable data for the original model. Do you have the same equality constraint?

(c) Re-parameterizing by a change of variables (actually, two changes of variables), obtain the surrogate model of Question 1. What is β' in terms of the parameters of the original model?