STA 302f14 Assignment Ten[1]

Please bring your printout for Question 1 to the quiz.

1. This question uses the data file `sat.data` from lecture unit 10 ("Inference with R: Part One"). There are links from the course home page in case the one in this document does not work.

   R's `confint` function gives confidence intervals for individual regression coefficients; see `help(confint)`. But it would be nicer to have a function that would calculate the confidence interval for a *linear combination* of regression coefficients. Your job is to write such a function.

   Before proceeding, please read these rules. The rules are part of this question, so you cannot pretend that you did not know about them.

   - You must write the function yourself. This means working by yourself a good part of the time.
   - Do not look at anyone else's R code, and do not show your R code to anyone except possibly me or Charles. This means don't even let someone glance at it.
   - If you have a "tutor" from in or outside the class who tells you in detail what to do on this question, you are guilty of an academic offence.
   - Suppose a "tutor" or anyone else writes a function like the one requested in this question, just to give you an idea of how to proceed. If you even look at what this person done, you are guilty of an academic offence.
   - Almost surely, people have written functions like this in the past, and some of them could be posted on the Internet. If you even look at these functions (much less copying the code), you are guilty of an academic offence.

   - What I want you to do has a lot in common with my `ftest` function for the general linear test. There is a link from the course home page in case the one in this document does not work. So take a look at that, and use it as a model. You can discuss `ftest` with anybody. Charles and I will be happy to answer questions about `ftest`. We will be more restrained in answering questions about the confidence interval function you are writing.
   - One final comment is that you should not be disturbed if you don't know how to do this question at first (though it will be easy for programmers). You are supposed to think about it and figure out what to do.

   After all this introduction, the actual question starts on the next page.

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

Here is what you need to do.

(a) Write a function that computes a confidence interval for a linear combination of regression coefficients. Output is an estimate of the linear combination, a lower confidence limit and an upper confidence limit. Label the output.

Input to the function should be

- An `lm` model object.
- A vector of constants for the linear combination. (What is this? Look on the formula sheet; there's only one possibility.)
- A confidence level, like 0.95 for a 95% confidence interval, 0.99 for a 99% confidence interval, and so on.

The printout you bring to the quiz *must* include a listing of your function.

(b) For the SAT data, use the built-in `confint` function to calculate 95% confidence intervals for β_0, β_1 and β_2.

(c) Now use your function to calculate 95% confidence intervals for β_0, β_1 and β_2. This will tell you if your function works.

(d) Use your function to calculate a 95% confidence interval for $\beta_1 - \beta_2$.

(e) Use your function to calculate a 99% confidence interval for $\beta_1 - \beta_2$.

2. In the usual univariate multiple regression model, the $\mathbf{X}$ is an $n \times (k+1)$ matrix of known constants. But of course in practice, the independent variables are often random, not fixed. Clearly, 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) Show that the least-squares estimator $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is conditionally unbiased. You've done this before.

   (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 F-test comparing full and reduced models), 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. Assume that the independent variables are discrete, so you can write a multiple sum.

3. Consider the following model with random independent variables. Independently for $i = 1, \ldots, n$,

$$\begin{aligned} Y_i &= \alpha + \beta_1 X_{i1} + \cdots + \beta_k X_{ik} + \epsilon_i \\ &= \alpha + \boldsymbol{\beta}' \mathbf{X}_i + \epsilon_i, \end{aligned}$$

where

$$\mathbf{X}_i = \left(\begin{array}{c} X_{i1} \\ \vdots \\ X_{ik} \end{array} \right)$$

and $\mathbf{X}_i$ is independent of ϵ_i.

Here, the symbol α is represents the intercept of an uncentered model; and $\boldsymbol{\beta}$ does not include the intercept. The "independent" variables $\mathbf{X}_i = (X_{i1}, \ldots, X_{ik})'$ are not statistically independent. They have the symmetric and positive definite $k \times k$ covariance matrix $\boldsymbol{\Sigma}_x = [\sigma_{ij}]$, which need not be diagonal. They also have the $k \times 1$ vector of expected values $\boldsymbol{\mu}_x = (\mu_1, \ldots, \mu_k)'$.

(a) Let $\boldsymbol{\Sigma}_{xy}$ denote the $k \times 1$ matrix of covariances between Y_i and X_{ij} for $j = 1, \ldots, k$. Calculate $\boldsymbol{\Sigma}_{xy} = C(\mathbf{X}_i, Y_i)$, obtaining $\boldsymbol{\Sigma}_{xy} = \boldsymbol{\Sigma}_x \boldsymbol{\beta}$.

(b) Solve the equation above for $\boldsymbol{\beta}$ in terms of $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_{xy}$.

(c) Using the expression you just obtained and letting $\widehat{\boldsymbol{\Sigma}}_x$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$ denote matrices of *sample* variances and covariances, what would be a reasonable estimator of $\boldsymbol{\beta}$ that you could calculate from sample data?

(d) Let $\mathbf{X}_c$ denote the $n \times k$ matrix of *centered* independent variable values. To see that your "reasonable" (Method of Moments) estimator from Question 3c is actually the usual one, first verify that the matrix $\frac{1}{n-1}\mathbf{X}_c'\mathbf{X}_c$ is a sample variance-covariance matrix. Show some calculations. What about $\frac{1}{n-1}\mathbf{X}_c'\mathbf{Y}_c$?

(e) In terms of $\widehat{\boldsymbol{\Sigma}}_x$ and $\widehat{\boldsymbol{\Sigma}}_{xy}$, what is $\widehat{\boldsymbol{\beta}} = (\mathbf{X}_c'\mathbf{X}_c)^{-1}\mathbf{X}_c'\mathbf{Y}_c$?

4. In the following regression model, the independent 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 independent variables are given by

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

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.

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

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 $E(\epsilon_i^\prime) = 0$. 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 $Cov(X_{i,1}, \epsilon_i^\prime)$?

(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. You may also use the fact that like sample means, sample variances and covariances converge to the corresponding Greek-letter versions as $n \rightarrow \infty$ (except possibly on a set of probability zero) like ordinary limits, and all the usual rules of limits apply. So for example, defining $\widehat{\sigma}_{xy}$ as $\frac{1}{n-1}\sum_{i=1}^n (X_{i,1} - \overline{X}_1)(Y_i - \overline{Y})$, we have $\widehat{\sigma}_{xy} \rightarrow Cov(X_i, Y_i)$.

So finally, here is the question. As $n \rightarrow \infty$, does $\widehat{\beta}_1 \rightarrow \beta_1$? Show your work.

Please bring your printout for Question 1 to the quiz. The other questions are just practice for the quiz, and are not to be handed in.

This assignment was prepared by Jerry Brunner, Department of Statistical Sciences, University of Toronto. The LaTeX source code is available from the course website: `http://www.utstat.toronto.edu/~brunner/oldclass/302f14`