STA 2101/442 Assignment 8[1]

Questions 4 and 5 use R. Please bring your printouts to the quiz on Friday November 3d. The non-computer questions on this assignment are practice for the quiz, and are not to be handed in. Please do the problems using the formula sheet as necessary. A copy of the formula sheet will be distributed with the quiz. As usual, you may use anything on the formula sheet unless you are directly asked to prove it.

1. Here is a question you may have already done (except that Quiz 2 was cancelled). You are asked to do it again as warm-up. Let $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ as on the formula sheet. Show $\mathbf{X}^\top \mathbf{e} = \mathbf{0}$.

2. Question 1 tells you that if a regression model has an intercept, the residuals add to zero. This yields $SST = SSR + SSE$, and makes R^2 is meaningful. It turns out that the residuals also add up to zero for some models that do not have intercepts. Again, this is attractive because in that case R^2 is meaningful.

   Here is an easy condition to check. Let $\mathbf{1}$ denote an $n \times 1$ column of ones. Show that if there is a $p \times 1$ vector of constants $\mathbf{v}$ with $\mathbf{X}\mathbf{v} = \mathbf{1}$, then $\sum_{i=1}^{n} e_i = 0$. Another way to state this is that if there is a linear combination of the columns of $\mathbf{X}$ that equals a column of ones, then the sum of residuals equals zero. Clearly this applies to a model with a categorical explanatory variable and cell means coding.

3. Suppose data for a regression study are collected at two different locations; n_1 observations are collected at location one, and n_2 observations are collected at location two. The same explanatory variables are used at each location. We need to know whether the error variance $\sigma^2 = Var(\epsilon_i)$ is the same at the two locations. For example, the locations might be different hospitals in a multi-center clinical trial, or two shopping malls in a market research study. Different error variances σ^2 might suggest we are dealing with different populations, or possibly that data collection was not carried out with the same care at the location with the larger variance. We are willing to assume normality.

   Recall the definition of the F distribution. If $W_1 \sim \chi^2(\nu_1)$ and $W_2 \sim \chi^2(\nu_2)$ are independent, then $F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F(\nu_1, \nu_2)$. Suggest a statistic for testing $H_0 : \sigma_1^2 = \sigma_2^2$. Using facts from the formula sheet, show it has an F distribution when H_0 is true. Don't forget to state the degrees of freedom. Assume that data coming from the two locations are independent.

4. People who raise large numbers of birds inhale potentially dangerous material, especially tiny fragments of feathers. Can this be a risk factor for lung cancer, controlling for other possible risk factors? Which of those other possible risk factors are important? Here are the variables in the file http://www.utstat.utoronto.ca/~brunner/data/illegal/birdlung.data.txt. These data are from a textbook called the *Statistical Sleuth* by Ramsey and Schafer, and are used without permission.

Variable	Values
Lung Cancer	1=Yes, 0=No
Gender	1=Female, 0=Male
Socioeconomic Status	1=High, 0=Low
Birdkeeping	1=Yes, 0=No
Age	
Years smoked	
Cigarettes per day	

If you look at `help(colnames)`, you can see how to add variable names to a data frame. It's a good idea, because if you can't remember which variables are which during the quiz, you're out of luck.

First, make tables of the binary variables using `table`, Use `prop.table` to find out the percentages. What proportion of the sample had cancer. Any comments?

There is one primary issue in this study: Controlling for all other variables, is birdkeeping significantly related to the chance of getting lung cancer? Carry out a likelihood ratio test to answer the question.

(a) In symbols, what is the null hypothesis?

(b) What is the value of the likelihood ratio test statistic G^2? The answer is a number.

(c) What are the degrees of freedom for the test? The answer is a number.

(d) What is the p-value? The answer is a number.

(e) What do you conclude? Presence of a relationship is not enough. Say what happened.

(f) For a non-smoking, bird-keeping woman of average age and low socioeconomic status, what is the estimated probability of lung cancer? The answer (a single number) should be based on the full model.

(g) Obtain a 95% confidence interval for that last probability. Your answer is a pair of numbers. There is an easy way and a hard way. Do it the easy way.

(h) Your answer to the last question made you uncomfortable. Why? Another approach is to start with a confidence interval for the log odds, and then use the fact that the function $p(x) = \frac{e^x}{1+e^x}$ is strictly increasing in x. Get the confidence interval this way. Again, your answer is a pair of numbers. Which confidence interval do you like more?

(i) Naturally, you should be able to interpret all the Z-tests too. Which one is comparable to the main likelihood ratio test you have just done?

(j) Controlling for all other variables, are the chances of cancer different for men and women?

(k) Also, are *any* of the explanatory variables related to getting lung cancer? Carry out a single likelihood ratio test. You could do it from the default output with a calculator, but use R. Get the p-value, too.

(l) Now please do the same as the last item, but with a Wald test. Of course you should display the value of W_n, the degrees of freedom and the p-value.

(m) Finally and just for practice, fit a simple logistic regression model in which the single explanatory variable is number of cigarettes per day.

   i. When a person from this population smokes ten more cigarettes per day, the odds of lung cancer are multiplied by r (odds ratio). Give a point estimate of r. Your answer is a number.

ii. Using the `vcov` function and the delta method, give an estimate of the asymptotic variance of r. Your answer is a number.

Please bring your R printout for this question to the quiz. Also, this question requires some paper and pencil work, and it would be fair to ask for something like that on the quiz too.

5. Men and women are calling a technical support line according to independent Poisson processes with rates λ_1 and λ_2 per hour. Data for 144 hours are available, but unfortunately the sex of the caller was not recorded. All we have is the number of callers for each hour, which is distributed Poisson$(\lambda_1 + \lambda_2)$. The data are available in the file http://www.utstat.toronto.edu/~brunner/data/legal/poisson.data.txt.

   (a) The parameter in this problem is $\boldsymbol{\theta} = (\lambda_1, \lambda_2)^\top$. Try to find the MLE analytically. Show your work. Are there any points in the parameter space where both partial derivatives are zero?

   (b) Now try to find the MLE numerically with R's `nlm` function[2]. The Hessian is interesting; ask for it. Try two different starting values. Compare the minus log likelihoods at your two answers. What seems to be happening here?

   (c) Try inverting the Hessian to get the asymptotic covariance matrix. Any comments?

   (d) Why did estimation fail for this fairly realistic model?

   Please bring your R printout for this question to the quiz. Your printout must *not* contain answers to the non-computer parts of this question. That is, it must contain only numerical answers.

6. Ordinary linear regression is often applied to data sets where the independent variables are best modeled as random variables: write $y_i = \mathbf{X}_i^\top \boldsymbol{\beta} + \epsilon_i$. In what way does the usual conditional linear regression model with normal errors imply that random explanatory 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 $\mathbf{X}_i$?

7. For a model with just one (random) explanatory variable, 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.

8. In the following regression model, the explanatory 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 explanatory variables are given by

$$E\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \quad \text{and} \quad Var\begin{pmatrix} X_{i,1} \\ X_{i,2} \end{pmatrix} = \begin{pmatrix} \phi_{11} & \phi_{12} \\ \phi_{12} & \phi_{22} \end{pmatrix}$$

   The explanatory variables $X_{i,1}$ and $X_{i,2}$ are independent of ϵ_i.

[2]When I did this, I got lots of warning messages with some starting values, when the search repeatedly left the parameter space and then bounced back in. For this problem, you don't need to worry about the warnings as long as the exit code is one.

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.

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

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

(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. Is $\widehat{\beta}_1$ a consistent estimator of β_1 if the true model holds? Answer Yes or no and show your work. You may use the consistency of the sample variance and covariance without proof.

(d) Are there *any* points in the parameter space for which $\widehat{\beta}_1 \xrightarrow{p} \beta_1$ when the true model holds?

9. Independently for $i = 1, \ldots, n$, let $Y_i = \beta X_i + \epsilon_i$, where $X_i \sim N(\mu, \sigma_x^2)$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. Because of omitted variables that influence both X_i and Y_i, we have $Cov(X_i, \epsilon_i) = c \neq 0$.

(a) The least squares estimator of β is $\frac{\sum_{i=1}^n X_i Y_i}{\sum_{i=1}^n X_i^2}$. Is this estimator consistent? Answer Yes or No and prove your answer.

(b) Give the parameter space for this model. There are some constraints on c.

(c) First consider points in the parameter space where $\mu \neq 0$. Give an estimator of β that converges almost surely to the right answer for that part of the parameter space. If you are not sure how to proceed, try calculating the expected value and covariance matrix of (X_i, Y_i).

(d) What happens in the rest of the parameter space — that is, where $\mu = 0$? Is a consistent estimator possible there? So we see that parameters may be identifiable in some parts of the parameter space but not all.

10. We know that omitted explanatory variables are a big problem, because they induce non-zero covariance between the explanatory variables and the error terms ϵ_i. The residuals have a lot in common with the ϵ_i terms in a regression model, though they are not the same thing. A reasonable idea is to check for correlation between explanatory variables and the ϵ_i values by looking at the correlation between the residuals and explanatory variables.

Accordingly, for a multiple regression model with an intercept so that $\sum_{i=1}^n e_i = 0$, calculate the sample correlation r between explanatory variable j and the residuals $e_1, \ldots, e_n$. Use this formula for the correlation: $r = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}$. Simplify. What can the sample correlations between residuals and x variables tell you about the correlation between ϵ and the x variables?

11. This question explores the consequences of ignoring measurement error in the explanatory variable when there is only one explanatory variable. Independently for $i = 1, \ldots, n$, let

$$\begin{aligned} Y_i &= \beta X_i + \epsilon_i \\ W_i &= X_i + e_i \end{aligned}$$

where all random variables are normal with expected value zero, $Var(X_i) = \phi > 0$, $Var(\epsilon_i) = \sigma^2_\epsilon > 0$, $Var(e_i) = \sigma^2_e > 0$ and ϵ_i, e_i and X_i are all independent. The variables W_i and Y_i are observable, while X_i is latent (unobservable, like true number of calories eaten). Error terms are never observable.

(a) What is the parameter vector $\boldsymbol{\theta}$ for this model?

(b) Denote the variance-covariance matrix of the observable variables by $\boldsymbol{\Sigma} = [\sigma_{ij}]$. The distribution of the observable data is completely determined by $\boldsymbol{\Sigma}$. Calculate the $\boldsymbol{\Sigma}$, expressed as a function of the model parameters.

(c) Here, identifiability means that the parameter can be recovered from $\boldsymbol{\Sigma}$ – that is, one can express the parameter as a function of the σ_{ij} values. Are there any points in the parameter space where the parameter β is identifiable? Are there infinitely many, or just one point?

(d) The naive estimator of β is $\widehat{\beta}_n = \frac{\sum_{i=1}^n W_i Y_i}{\sum_{i=1}^n W_i^2}$. Is $\widehat{\beta}_n$ a consistent estimator of β? Why can you answer this question without doing any calculations?

(e) Go ahead and do the calculation. To what does $\widehat{\beta}_n$ converge?

(f) Are there any points in the parameter space for which $\widehat{\beta}_n$ converges to the right answer? Compare your answer to the set of points where β is identifiable.