STA 2101/442 Assignment Six[1]

The questions are just practice for the quiz, and are not to be handed in. Use R for Question 5, and bring your printout to the quiz. **Your printout should show *all* R input and output, and *only* R input and output**. Do not write anything on your printouts except your name and student number.

1. Suppose you wish to test the null hypothesis that a *single* linear combination of regression coefficients is equal to zero. You can test either $H_0 : \mathbf{a}^\top \boldsymbol{\beta} = 0$ with a two-sided t-test, or $H_0 : \mathbf{L}\boldsymbol{\beta} = 0$ with an F-test. Referring to the formula sheet, verify that $F = t^2$. Show your work.

2. The exact way that you express a linear null hypothesis does not matter. Let $\mathbf{A}$ be an $r \times r$ nonsingular matrix (meaning $\mathbf{A}^{-1}$ exists), so that $\mathbf{L}\boldsymbol{\beta} = \mathbf{h}$ if and only if $\mathbf{AL}\boldsymbol{\beta} = \mathbf{Ah}$. This is a useful way to express a logically equivalent null hypothesis, because any matrix that is row equivalent to $\mathbf{L}$ can be written as $\mathbf{AL}$. Show that the general linear test statistic F for testing $H_0 : (\mathbf{AL})\boldsymbol{\beta} = \mathbf{Ah}$ is the same as the one for testing $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$. Use the fact that if the inverses exist, the inverse of a matrix product is the product of inverses, in reverse order.

3. Also, all those dummy variable coding schemes are equivalent. Let $\mathbf{A}$ be a $p \times p$ nonsingular matrix (it's a different $\mathbf{A}$ from the one in Question 2). Note that $\mathbf{X}^* = \mathbf{XA}$ is a one-to-one linear transformation of the explanatory variables, and

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \;\Leftrightarrow\; \mathbf{y} = \mathbf{XA}\,\mathbf{A}^{-1}\boldsymbol{\beta} + \boldsymbol{\epsilon} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}.$$

   This is already interesting, because it shows how transforming the explanatory variables changes the meaning of the regression coefficients. Refer to $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ as the "original" model, and $\mathbf{y} = \mathbf{X}^*\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$ as the "transformed" model.

   (a) Just to make this more concrete, suppose you have a 3-category explanatory variable and a quantiative covariate. $Y_i = \beta_0 + \beta_1 d_{i,1} + \beta_2 d_{i,2} + \beta_3 x_i + \epsilon_i$, where $d_{i,1}$ and $d_{i,2}$ are indicator dummy variables for the first two groups. You want to switch to cell means coding, so that $Y_i = \beta_1^* g_{i,1} + \beta_2^* g_{i,2} + \beta_3^* g_{i,3} + \beta_4^* x_i + \epsilon_i$. Note that $\beta_4^* = \beta_3$. Give the matrix $\mathbf{A}$; you can make tables if that helps.

   (b) Write down the least squares estimate $\widehat{\boldsymbol{\beta}}^*$ for the transformed model, and simplify. How is $\widehat{\boldsymbol{\beta}}^*$ related to $\widehat{\boldsymbol{\beta}}$?

   (c) Compare the vector of predicted values from the two models.

   (d) Compare the vector of residuals from the two models.

   (e) Which is greater, SSE or SSE^*?

   (f) Suppose you want to test $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$. Give the equivalent null hypothesis for the transformed model. That is, what are matrices $\mathbf{L}^*$, $\boldsymbol{\beta}^*$ and $\mathbf{h}^*$ in $H_0 : \mathbf{L}^*\boldsymbol{\beta}^* = \mathbf{h}^*$?

   (g) Compare the F statistic for $H_0 : \mathbf{L}^*\boldsymbol{\beta}^* = \mathbf{h}^*$ to the F statistic for $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$.

4. Question 3 suggests that if a regression model with no intercept is equivalent to one with an intercept, then the residuals will add to zero. This is good to know, because it means $SST = SSR + SSE$, and R^2 is meaningful; so is a, the proportion of remaining variation. 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 $\mathbf{Xv} = \mathbf{1}$, 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 cell means coding.)

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

5. The data file CensusTract.data.txt, comes from *Applied Linear Statistical Models* (1996), by Neter et al.. The data are used here without permission. You can get the data with

```
census = read.table("http://www.utstat.toronto.edu/~brunner/data/illegal/CensusTract.data.txt").
```

The cases (there are n cases) are a sample of census tracts in the United States. For each census tract, the following variables are recorded.

`area`	Land area in square miles
`pop`	Population in thousands
`urban`	Percent of population in cities
`old`	Percent of population 65 or older
`docs`	Number of active physicians
`beds`	Number of hospital beds
`hs`	Percent of population 25 or older completing 12+ years of school
`labor`	Number of persons 16+ employed or looking for work
`income`	Total Total before tax income in millions of dollars
`crimes`	Total number of serious crimes reported by police
`region`	Region of the country: 1=Northeast, 2=North Central, 3=South, 4=West

(a) First, use R to fit a regression model with `crimes` as the response variable and just one explanatory variable: `pop`.

i. In plain, non-statistical language, what do you conclude from this analysis? The answer is something about population size and number of crimes.

ii. What proportion of the variation in number of crimes is explained by population size? The answer is a number between zero and one.

(b) Based on that last analysis, we will create a new response variable called crime *rate*, defined as number of crimes divided by population size. Now fit a new regression model in which crime rate is a function of `area`, `urban`, `old`, `docs`, `beds`, `hs`, `labor`, `income` and `region` of the country. There are no interactions for now. This is the *full model* in all the analyses that follow.

Just so we will be doing things the same way, please make `region` a factor, and look at help to see how to use the `labels=` option. It really helps.

Based on this model,

i. What is p? The answer is a number.

ii. What is $\widehat{\beta}_4$? The answer is a number.

iii. Give the test statistic, the degrees of freedom and the p-value for each of the following null hypotheses. The answers are numbers from your printout.

A. $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{11} = 0$

B. $H_0 : \beta_7 = 0$

C. $H_0 : \beta_0 = 0$

iv. What proportion of the variation in crime rate is explained by the explanatory variables in this model? The answer is a number.

v. What is the smallest value of e_i? The answer is a number.

vi. What is the largest value of e_i? The answer is a number.

vii. Look at the output of `summary`. For the first entry under "`t value`" (that's `1.502`), what is the null hypothesis? The answer is a symbolic statement involving one or more Greek letters.

viii. Look at the F test at the end of the `summary` output. What is the null hypothesis? The answer is a symbolic statement involving one or more Greek letters.

ix. Controlling for all the other variables in the model, is percent High School graduates related to crime rate?
   A. Give the null hypothesis in symbols.
   B. Give the value of the test statistic. The answer is a number from your printout.
   C. Give the p-value. The answer is a number from your printout.
   D. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.
   E. Allowing for other variables, census regions with a higher percentage of High School graduates tend to have _____ crime rates.

x. Controlling for all the other variables in the model, is number of physicians related to crime rate?
   A. Give the null hypothesis in symbols.
   B. Give the value of the test statistic. The answer is a number from your printout.
   C. Give the p-value. The answer is a number from your printout.
   D. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.
   E. Is there enough evidence to conclude that allowing for other variables, number of physicians is related to crime rate?

xi. Controlling for all the other variables in the model, is there a difference in crime rate between the Northeast and North Central regions?
   A. Give the null hypothesis in symbols.
   B. Give the value of the test statistic. The answer is a number from your printout.
   C. Give the p-value. The answer is a number from your printout.
   D. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.
   E. State the conclusion in plain, non-statistical language. If there is a difference, say which region has a higher average crime rate!

xii. Controlling for all the other variables in the model, is there a difference in crime rate between the Northeast and South regions?
   A. Give the null hypothesis in symbols.
   B. Give the value of the test statistic. The answer is a number from your printout.
   C. Give the p-value. The answer is a number from your printout.
   D. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.
   E. State the conclusion in plain, non-statistical language. If there is a difference, say which region has a higher average crime rate!

xiii. I think it's remarkable that only one variable apart from region seems to make a difference once you allow for the others. Which one is it?

xiv. But the other variables may be masking one another's relationship to the response variable when each one is controlled for all the others. Please test them all at once, with a view to maybe dropping them and obtaining a simpler model.
   A. Give the null hypothesis in symbols.
   B. Give the value of the test statistic. The answer is a number from your printout.
   C. Give the p-value. The answer is a number from your printout.
   D. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.
   E. Is there evidence that, once we control for region and percent High School graduates, that any of these variables is related to the crime rate?

xv. To be continued ...

Please bring your printout to the quiz. **Your printout should show *all* R input and output, and *only* R input and output**. Do not write anything on your printouts except your name and student number.

6. 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.

7. 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 parameter in this problem is $\boldsymbol{\theta} = (\lambda_1, \lambda_2)$. Try to find the MLE analytically. Show your work. Are there any points in the parameter space where both partial derivatives are zero? Why did estimation fail for this fairly realistic model?

8. Show that for a simple regression with an intercept and one explanatory variable, $R^2 = r^2$, where $r = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2} \sqrt{\sum_{i=1}^n (Y_i - \overline{Y})^2}}$ is the ordinary sample correlation coefficient. You may also use the formulas $\widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(Y_i - \overline{Y})}{\sum_{i=1}^n (x_i - \overline{x})^2}$ and $\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{x}$. It helps to start with the formula for R^2, and then substitute for $\widehat{Y}_i$ right away.

9. 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. In class, somebody suggested checking 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 satisfying the condition of Question 4 (so that $\sum_{i=1}^n e_i = 0$), calculate the sample correlation r between explanatory variable j and the residuals $e_1, \ldots, e_n$. The final answer is a number. Does this suggest what *plots* of the residuals versus explanatory variables should look like if the model is okay?

10. Again for a model for which $\sum_{i=1}^n e_i = 0$, calculate the sample correlation between the residuals and predicted values $\widehat{Y}_i$. Does this suggest what a *plot* of the residuals versus predicted values should look like if the model is okay?

11. Still for a model for which $\sum_{i=1}^n e_i = 0$, show that the squared correlation between the predicted and observed response variable values is equal to R^2. Hint: As preparation, verify that $\mathbf{y}^\top \widehat{\mathbf{y}} = \widehat{\mathbf{y}}^\top \widehat{\mathbf{y}}$.

    Thus, a scatterplot of Y versus $\widehat{Y}$ gives a picture of how well the explanatory variables are doing their job. How do you know that the correlation is always non-negative?
