STA 302f17 Assignment Seven[1]

Except for Problem 2, these problems are preparation for the quiz in tutorial on Thursday November 9th, and are not to be handed in. As usual, sometimes you may be asked to prove things that are false. Please bring your printout for Problem 2 to the quiz. Do not write anything on the printout in advance of the quiz, except possibly your name and student number.

1. For the general linear regression model, assume that $n > k + 1$ and that the columns of X are linearly independent, so that $(X'X)^{-1}$ exists and $\mathbf{b}$ is well defined. Starting from the definition on the formula sheet, prove that $\mathbf{e} = \mathbf{0}$..

2. To read the `statclass` data used in Assignment 5, at the R prompt type

```
statclass = read.table("http://www.utstat.utoronto.ca/~brunner/data/legal/LittleStatclassdata.txt")
```

   Fit a regression model in which the dependent variable is mark on the final exam, and the independent variables are Quiz Average, Computer Average, and mark on the Midterm test. Please use this variable order in your R program.

   (a) What is the predicted Final Exam score for a student with a Quiz average of 8.5, a Computer average of 5, and a Midterm mark of 60%? The answer is a number. Be able to do this kind of thing on the quiz with a calculator from the output of `summary`.

   (b) For any fixed Quiz Average and Computer Average, a score one point higher on the Midterm yields a predicted mark on the Final Exam that is _____ higher.

   (c) For any fixed Quiz Average and Midterm score, an average one point higher on the Computer Average yields a predicted mark on the Final Exam that is _____ higher. Or is it lower?

   (d) What is b_3? The answer is a number from your printout.

   (e) For each of the following null hypotheses, give the value of the test statistic and the p-value. These are numbers from your printout. Also state whether you reject H_0 at $\alpha = 0.05$.

H_0	Test Statistic	p-value	Reject H_0?
$\beta_1 = \beta_2 = \beta_3 = 0$			
$\beta_0 = 0$			
$\beta_1 = 0$			
$\beta_2 = 0$			
$\beta_3 = 0$			

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

(f) For each of the following questions, give the null hypothesis you tested to answer the question, and also a conclusion expressed in plain, non-statistical language. Remember the rules: No statistical terminology, draw a directional conclusion if you can, be guided by $\alpha = 0.05$ but never mention it, and don't accept H_0.

i. Controlling for quiz average and computer average, is mark on the midterm test related to mark on the final exam?

ii. Allowing for mark on the midterm test and quiz average, is computer average a useful predictor of mark on the final exam?

iii. Taking into account mark on the midterm test and computer average, is quiz average related connected to mark on the final exam?

iv. Are any of the predictor variables useful?

(g) Controlling for mark on the midterm tests, are the other two variables (either or both) related to mark on the Final exam?

i. State the null hypothesis in terms of scalar β values.

ii. State the null hypothesis in matrix terms. That is, give the matrices C, $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$.

iii. Write the reduced model. Please do not re-number the variables are β_j parameters.

iv. Give the value of the test statistic F. It is a number from your printout, but not part of the `summary` output.

v. Give the p-value. The answer is a number from your printout.

vi. Do you reject H_0 at $\alpha = 0.05$? Answer Yes or No.

vii. Are the results statistically significant at the $\alpha = 0.05$ level? Answer Yes or No.

viii. Allowing for mark on the midterm test, what proportion of the remaining variation in final exam score is explained by computer average and quiz average?

ix. State your conclusions (if any) in plain, non-statistical language.

(h) What is the largest e_i in absolute value? The answer is on your printout.

(i) What is k for this problem? You can get it from the output of `summary`.

(j) What is n for this problem? You can calculate it from the output of `summary` without a calculator.

(k) What are the dimensions of the X matrix? The answer is a pair of numbers, number of rows and number of columns. You can calculate them from the output of `summary` without a calculator.

(l) What are the dimensions of $\mathbf{b}$? The answer is a pair of numbers, number of rows and number of columns. You can obtain them from the output of `summary` without a calculator.

(m) What are the dimensions of $\mathbf{e}$? The answer is a pair of numbers, number of rows and number of columns. You can obtain them from the output of `summary` without a calculator.

(n) What are the dimensions of $\mathbf{e}'\mathbf{e}$?

(o) What are the dimensions of the $\widehat{\mathbf{y}}$ matrix? The answer is a pair of numbers, number of rows and number of columns.

(p) What are the dimensions of the hat matrix H? The answer is a pair of numbers, number of rows and number of columns.

(q) What is $\mathbf{e}'\mathbf{e}$? You can calculate this number from the output of `summary` using a calculator, using the fact that `Residual standard error` from your printout is the square root of $s^2 = \mathbf{e}'\mathbf{e}/(n-k-1)$.

(r) What is SST? The answer is a single number. You can check your work with R, but calculate the number based just on the output of `summary` and the formula sheet. First show your work (there is some algebra), and then obtain the result with a calculator. Circle your final answer.

(s) The tests and confidence intervals based on the t distribution all use $t_{\alpha/2}$. By default we are using $\alpha = 0.05$, so $t_{\alpha/2}$ is the point cutting off the top 2.5% of the t distribution with $n-k-1$ degrees of freedom. Obtain this number with R and make sure it is included in your printout.

(t) With a calculator (or using R as a calculator) calculate a 95% confidence interval for β_3. You can get the numbers you need from the output of `summary`. You don't need `vcov` for this one.

(u) For this question, first use the `attach` function to make the variables conveniently available for calculation. See the *Least squares with R* handout. Then calculate the means of all the independent variables. You might as well calculate $\overline{y}$ as well.

i. First, give a point estimate of $E(y|x_1 = \overline{x}_1, x_2 = \overline{x}_2, x_3 = \overline{x}_3)$. There is an easy way and a hard way. You decide: the easy way, the hard way, or both because you like to double-check everything.

ii. Give a 95% confidence interval for $E(y|x_1 = \overline{x}_1, x_2 = \overline{x}_2, x_3 = \overline{x}_3)$. For this, you will need to use `vcov`. Your answer is a pair of numbers. You should do this with R and it should be on your printout.

3. The U.S. Census Bureau divides the United States into small pieces called census tracts; lots of information is collected about each census tract. The census tracts are grouped into four geographic regions: Northeast, North Central, South and West. In one study, the cases were census tracts, the explanatory variables were Region and average income, and the response variable was crime rate, defined as the number of reported serious crimes in a census tract, divided by the number of people in the census tract.

(a) Write $E(y|x)$ for a regression model with *no intercept* and parallel regression lines. You do not have to say how your dummy variables are defined. You will do that in the next part.

(b) Make a table showing how your dummy variables are set up. There should be one row for each region, and a column for each dummy variable. Add a wider column on the right, in which you show $E(Y|x)$. Note that the *symbols* for your dummy variables will not appear in this column. There are examples of this format in the lecture slides.

(c) For each of the following questions, give the null hypothesis in terms of the β parameters of your regression model. We are not doing one-tailed tests, regardless of how the question is phrased.

i. Controlling for average income, does average crime rate differ by geographic region?

ii. Allowing for average income, is average crime rate different in the Northeast and North Central regions?

iii. Controlling for average income, is average crime rate different in the Northeast and Western regions?

iv. Correcting for average income, is the crime rate in the South more than the average of the other three regions?

v. Holding average income constant, is the average of the crime rates in the Northeast and North Central regions different from the average of the crime rates in the South and West?

vi. Controlling for geographic region, is crime rate connected to average income?

4. Now please re-do Problem 3, using a model with an intercept. Make North Central the reference category; that's what R would do, since it's alphabetically first.

5. You know that if a regression model has an intercept, the residuals add to zero. This yields $SST = SSR + SSE$, and makes R^2 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 $(k+1) \times 1$ vector of constants $\mathbf{v}$ with $X\mathbf{v} = \mathbf{1}$, 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 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.

6. It was suggested in lecture that using a different dummy variable coding scheme is just a linear transformation of the X matrix: $W = XA$, where A is a $(k+1) \times (k+1)$ matrix with an inverse, and W is the new X matrix. Suppose you want to switch from cell means coding to indicators with an intercept. Consider the specific case of a single categorical independent variable with three categories, and a single quantitative

independent variable. Making the last category the reference category, there is a 4×4 matrix A such that

$$\begin{pmatrix} 1 & 0 & 0 & x_1 \\ 0 & 1 & 0 & x_2 \\ 0 & 0 & 1 & x_3 \\ 1 & 0 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & x_n \end{pmatrix} A = \begin{pmatrix} 1 & 1 & 0 & x_1 \\ 1 & 0 & 1 & x_2 \\ 1 & 0 & 0 & x_3 \\ 1 & 1 & 0 & x_4 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 1 & x_n \end{pmatrix}$$

Give the matrix A. It is a matrix of specific numbers.

7. Linear transformations of the X matrix are not limited to switching dummy variable schemes. In general,

$$\begin{aligned} & \mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \Leftrightarrow \quad & \mathbf{y} = XAA^{-1}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\ \Leftrightarrow \quad & \mathbf{y} = W\boldsymbol{\alpha} + \boldsymbol{\epsilon}, \end{aligned}$$

where A is a $(k+1) \times (k+1)$ matrix, $W = XA$ and $\boldsymbol{\alpha} = A^{-1}\boldsymbol{\beta}$.

   (a) Denoting the least-squares estimate of $\boldsymbol{\alpha}$ by $\mathbf{a}$, find a formula for $\mathbf{a}$. Simplify. What is its connection to $\mathbf{b}$?

   (b) What is the vector of predicted y values for the transformed model? How does it compare to $\widehat{\mathbf{y}}$ from the original model?

   (c) Give a null hypothesis equivalent to $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$, but in terms of the transformed model. It's $H_0 : C_2\boldsymbol{\alpha} = \boldsymbol{\gamma}$. What is C_2?

   (d) Compare the F^* statistics for testing $H_0 : C\boldsymbol{\beta} = \boldsymbol{\gamma}$ and $H_0 : C_2\boldsymbol{\alpha} = \boldsymbol{\gamma}$. One would hope they are the same. Are they? Show your work.
