STA 2101/442 Assignment Seven[1]

Please bring your R printouts to the quiz. *Your printouts must* not *contain answers to the non-computer parts of this assignment*, The non-computer questions are just practice for the quiz, and will not be handed in.

1. 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.

2. Consider a linear regression model with $n > p$, which is always the case in practice. Since the vector of residuals $\mathbf{e} \sim N(\mathbf{0}, \sigma^2(\mathbf{I}-\mathbf{H}))$, it is tempting to write $\frac{1}{\sigma^2}\mathbf{e}^\top(\mathbf{I}-\mathbf{H})^{-1}\mathbf{e} \sim \chi^2(n)$. Please locate support for this idea on the formula sheet. But it only works if the $n \times n$ matrix $\mathbf{I}-\mathbf{H}$ has an inverse.

   (a) The rank of a product is the minimum of ranks. Why does this tell you that that the hat matrix $\mathbf{H}$ has no inverse? If you don't remember what the rank of a matrix is, look it up.

   (b) But to me, it's not so obvious for $\mathbf{I}-\mathbf{H}$. Calculate $(\mathbf{I}-\mathbf{H})\,\mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}$. At this point the answer will be clear to some people, but not everybody. Continue by assuming that $(\mathbf{I}-\mathbf{H})^{-1}$ exists. If it does, you arrive at a conclusion that is impossible. Complete the proof.

3. For the general linear model (see formula sheet),

   (a) What is the distribution of $\mathbf{L}\widehat{\boldsymbol{\beta}}$? Note $\mathbf{L}$ is $r \times p$.

   (b) If $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$ is true, what is the distribution of $\frac{1}{\sigma^2}(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})^\top(\mathbf{L}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{L}^\top)^{-1}(\mathbf{L}\widehat{\boldsymbol{\beta}}-\mathbf{h})$? Please locate support for your answer on the formula sheet. For full marks, don't forget the degrees of freedom.

   (c) What other facts on the formula sheet allow you to establish the F distribution for the general linear test? The distribution is *given* on the formula sheet, so of course you can't use that. How do you know numerator and denominator are independent?

4. Suppose you wish to test the null hypothesis that a *single* linear combination of regression coefficients is equal to zero. That is, you want to test $H_0 : \mathbf{a}^\top\boldsymbol{\beta} = 0$. Referring to the formula sheet, verify that $F = T^2$. Show your work.

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

5. Suppose you fit (estimate the parameters of) a regression model, obtaining $\widehat{\boldsymbol{\beta}}$, $\widehat{\mathbf{Y}}$ and $\mathbf{e}$. Call this Model One.

   (a) Then just for fun, you fit a second regression model, using $\widehat{\mathbf{Y}}$ from Model One as the response variable, and exactly the same $\mathbf{X}$ matrix as Model One. Call this Model Two.

      i. What is $\widehat{\boldsymbol{\beta}}$ for Model Two? Show your work and simplify.
      ii. What is $\widehat{\mathbf{Y}}$ for Model Two? Show your work and simplify.
      iii. What is $\mathbf{e}$ for Model Two? Show your work and simplify.
      iv. What is MSE for Model Two?

   (b) Now you fit a *third* regression model, this time using $\mathbf{e}$ from Model One as the response variable, and again, exactly the same $\mathbf{X}$ matrix as Model One. Call this Model Three.

      i. What is $\widehat{\boldsymbol{\beta}}$ for Model Three? Show your work and simplify.
      ii. What is $\widehat{\mathbf{Y}}$ for Model Three? Show your work and simplify.
      iii. What is $\mathbf{e}$ for Model Three? Show your work and simplify.

6. For the usual multiple linear regression model with normal error terms, you already know that $\mathbf{e} \sim N(\mathbf{0}, \sigma^2(\mathbf{I} - \mathbf{H}))$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$. Let $\mathbf{Z} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\,\mathbf{e}$.

   (a) Find the distribution of $\mathbf{Z}$. The answer is remarkable, so keep simplifying! Cite facts from the formula sheet when you use them.

   (b) To see what happened, simplify the expression for $\mathbf{Z}$ itself as much as possible. Once you have the answer, look at the footnote[2].

   (c) What is $M_{\mathbf{Z}}(\mathbf{t})$?

7. 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 σ^2 is the same at the two locations.

   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.

8. This question will be a lot easier if you remember that if $X \sim \chi^2(\nu)$, then $E(X) = \nu$ and $Var(X) = 2\nu$. You don't have to prove this; just use it. You can also use things you already know about ordinary linear regression with normal errors.

[2] This is related to the concept of "degrees of freedom." Because p linear combinations of the residuals equal exactly zero, you can see that p of them are linear combinations of the others. Thus, only $n - p$ are free to vary. Mr. Fisher was a very smart guy.

For the usual linear regression model with normal errors, σ^2 is usually estimated with MSE.

(a) Show that MSE is an unbiased estimator of σ^2.

(b) Show that MSE is a consistent estimator of σ^2.

(c) Under the usual regression model what is the joint distribution of $\epsilon_1, \ldots, \epsilon_n$?

(d) Let $T_n = \frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$. What is $E(T_n)$?

(e) How do you know that $T_n \xrightarrow{p} \sigma^2$?

(f) Show that $Var(T_n) < Var(MSE)$.

(g) So it would appear that T_n is a better estimator of σ^2 than MSE is, since they are both unbiased and the variance of T_n is lower. So why do you think MSE is used in regression analysis instead of T_n?

9. Ordinary linear regression is often applied to data sets where the independent variables are best modeled as random variables. 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 $\mathbf{X}_i$?

10. 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.

11. In a study comparing the effectiveness of different exercise programmes, volunteers were randomly assigned to one of three exercise programmes (A, B, C) or put on a waiting list and told to work out on their own. Aerobic capacity is the body's ability to process oxygen. Aerobic capacity was measured before and after 6 months of participation in the program (or 6 months of being on the waiting list). The response variable was improvement in aerobic capacity. The explanatory variables were age (a covariate) and treatment group.

    (a) First consider a regression model with an intercept, and no interaction between age and treatment group.

    i. Make a table showing how you would set up indicator dummy variables for treatment group. Make Waiting List the reference category

    ii. Write the regression model. Please use x for age, and make its regression coefficient β_1.

    iii. In terms of β values, what null hypothesis would you test to find out whether, allowing for age, the three exercise programmes differ in their effectiveness?

    iv. Write the null hypothesis for the preceding question as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Just give the $\mathbf{L}$ matrix.

v. In terms of β values, what null hypothesis would you test to find out whether Programme B was better than the waiting list?

vi. In terms of β values, what null hypothesis would you test to find out whether Programmes A and B differ in their effectiveness?

vii. Suppose you wanted to estimate the difference in average benefit between programmes A and C for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

viii. Is it safe to assume that age is independent of the other explanatory variables? Answer Yes or No and briefly explain.

(b) Now consider a regression model with an intercept and the interaction (actually a set of interactions) between age and treatment.

i. Write the regression model. Make it an extension of your earlier model.

ii. Suppose you wanted to know whether the slopes of the 4 regression lines were equal. In terms of β values, what null hypothesis would you test?

iii. Suppose you wanted to know whether any differences among mean improvement in the four treatment conditions depends on the participant's age. In terms of β values, what null hypothesis would you test?

iv. Write the null hypothesis for the preceding question as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Just give the $\mathbf{L}$ matrix. It is $r \times p$. What is r? What is p?

v. Suppose you wanted to know whether the difference in effectiveness between Programme A and the Waiting List depends on the participant's age. In terms of β values, what null hypothesis would you test?

vi. Suppose you wanted to *estimate* the difference in average benefit between programmes A and C for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

(c) Now consider a regression model *without* an intercept, but *with* possibly unequal slopes. Make a table to show how the dummy variables could be set up, and write the regression model. Again, please use x for age and make its regression coefficient β_1. This model needs to have the *same number of regression coefficients as the model of Question 11b*, so you have to think about this a little.

For each treatment condition, what is the conditional expected value of Y? The answer is in terms of x and the β values. Please put these values as the last column of your table.

i. Suppose you wanted to know whether the slopes of the 4 regression lines were equal. In terms of β values, what null hypothesis would you test?

ii. Suppose you wanted to know whether any differences among mean improvement in the four treatment conditions depends on the participant's age. In terms of β values, what null hypothesis would you test?

iii. Write the null hypothesis for the preceding question as $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{0}$. Just give the $\mathbf{L}$ matrix. It is $r \times p$. What is r? What is p?

iv. Suppose you wanted to know whether the difference in effectiveness between Programme A and the Waiting List depends on the participant's age. In terms of β values, what null hypothesis would you test?

v. Suppose you wanted to estimate the difference in average benefit between programmes A and C for a 27 year old participant. Give your answer in terms of $\widehat{\beta}$ values.

12. The data file `CensusTract.data`, comes from *Applied Linear Statistical Models* (1996), by Neter et al.. The data are used here without permission. There is a link on the course home page in case the one in this document does not work.

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, fit 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 1.

Bring your printout to the quiz.

(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 each other's relationship when each 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 . . .

Bring your printout to the quiz.
