STA 302f14 Assignment Nine[1]

Please bring your printout for Question 6 to the quiz. The other questions are just practice for the quiz, and are not to be handed in.

1. Look at the general linear model in scalar form on the formula sheet. Suppose that each observation Y_i is actually the mean of n_i independent observations with common variance σ^2. For example, Y_i could be the average customer satisfaction rating at a bank branch, but different numbers of customers took the survey at each branch.

   (a) What is $Var(Y_i)$?

   (b) Because this is a regression model, let's make this the variance of ϵ_i, so that it's also the variance of Y_i. Do the ϵ_i all have equal variance now?

   (c) Multiply both sides of the regression model equation by a constant c_i (different for each i), obtaining

   $$Y_i^* = \beta_0^* + \beta_1 x_{i1}^* + \cdots + \beta_k x_{ik}^* + \epsilon_i^*.$$

   Call this the "transformed model." Choose the constants c_i so that the variances of all the ϵ_i^* are equal. What is c_i?

   (d) Remember that the least squares problem is to minimize

   $$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2$$

   Write Q for the transformed model in terms of the n_i, and simplify (factor something out). This is a simple but very useful version of "weighted least squares," in which some observations get more weight than others in determining $\widehat{\boldsymbol{\beta}}$. What are the "weights?"

2. Here is a generalization of Question 1. In the general linear regression model, let $cov(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{V}$, where $\mathbf{V}$ is a *known* symmetric and positive definite matrix. As usual, σ^2 is an unknown constant.

   (a) What is the $cov(\mathbf{Y})$ for this unequal variance model?

   (b) Multiply both sides of $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ by $\mathbf{V}^{-1/2}$, obtaining the "transformed" model $\mathbf{Y}^* = \mathbf{X}^*\boldsymbol{\beta} + \boldsymbol{\epsilon}^*$. Notice that $\boldsymbol{\beta}$ is the same for the original model and the transformed model.

      i. What is the variance-covariance matrix of $\boldsymbol{\epsilon}^*$?

      ii. What is the matrix $\mathbf{V}$ for Question 1?

   (c) Write down and simplify a formula for $\widehat{\boldsymbol{\beta}}^*$.

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

(d) Is $\widehat{\boldsymbol{\beta}}^*$ unbiased given the unequal variance model? Answer Yes or No and show your work.

(e) Is $\widehat{\boldsymbol{\beta}}$ unbiased given the unequal variance model? Answer Yes or No and show your work.

(f) Given the unequal variance model, which has the smaller variance, $\mathbf{a}'\widehat{\boldsymbol{\beta}}^*$, or $\mathbf{a}'\widehat{\boldsymbol{\beta}}$? Why?

(g) If $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{V})$, what is the distribution of $\widehat{\boldsymbol{\beta}}^*$? Show your work.

3. The Wisconsin Power and Light Company studied the effectiveness of two devices for improving the efficiency of gas home-heating systems. The electric vent damper (EVD) reduces heat loss through the chimney when the furnace is in the off cycle by closing off the vent. It is controlled electrically. The thermally activated vent damper (TVD) is the same as the EVD except it is controlled by the thermal properties of a set of bimetal fins set in the vent. Ninety test houses were randomly assigned to have a free vent damper installed; 40 received EVDs and 50 received TVDs. For each house, energy consumption was measured for a period of several weeks with the vent damper active ("vent damper in") and for an equal period with the vent damper not active ("vent damper out". Here are the variables:

   House Identification Number

   Type of furnace (1=Forced air 2=Gravity 3=Forced water 4=Steam)

   Chimney area

   Chimney shape (1=Round 2=Square 3=Rectangular)

   Chimney height in feet

   Type of Chimney liner (0=Unlined 1=Tile 2=Metal)

   Type of house (1=Ranch 2=Two-story 3=tri-level 4=Bi-level 5=One and a half stories)

   House age in yrs

   Type of damper (1=EVD 0=TVD)

   Energy consumpt with damper active (in)

   Energy consumpt with damper inactive (out)

   Consider a model in which the response variable (Y) is average energy consumption with vent damper in and vent damper out, and the explanatory variables are age of house (X_1), chimney area (X_2) and furnace type (4 categories). There should be no interactions in your model.

   (a) Write $E[Y|\mathbf{X}]$ for your model. This would be the *full* model for any F-test that uses the full versus reduced approach.

(b) Make a table with four rows, one for each type of furnace. Make columns showing how your dummy variables are defined, and include one wider column at the end, showing $E[Y|\mathbf{X}]$ for each furnace type.

(c) You want to test whether, controlling for age of house and chimney area, average energy consumption depends on furnace type.

i. Give the null hypothesis in terms of the βs.

ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

(d) You want to test whether, controlling for furnace type and chimney area, average energy consumption depends on age of house.

i. Give the null hypothesis in terms of the βs.

ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

(e) You want to test whether, controlling for age of house and chimney area, average energy is different for Forced air furnaces and Gravity furnaces.

i. Give the null hypothesis in terms of the βs.

ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

(f) You want to test whether, controlling for age of house and chimney area, average energy consumption is different for Forced air and Forced water furnaces.

i. Give the null hypothesis in terms of the βs.

ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

(g) You want to test whether, controlling for age of house and chimney area, average energy consumption is for Steam furnaces is different from the average of Forced air and Forced water furnaces. (You are comparing an expected value with the mean of two expected values.)

i. Give the null hypothesis in terms of the βs.

ii. Give $E[Y|\mathbf{X}]$ for the reduced model.

4. High School History classes from across Ontario are randomly assigned to either a discovery-oriented or a memory-oriented curriculum in Canadian history. At the end of the year, the students are given a standardized test and the median score of each class is recorded. Please consider a regression model with these variables:

   X_1 Equals 1 if the class uses the discovery-oriented curriculum, and equals 0 the class it uses the memory-oriented curriculum.

   X_2 Average parents' education for the classroom

   X_3 Average parents' income for the classroom

   X_4 Number of university History courses taken by the teacher

   X_5 Teacher's final cumulative university grade point average

   Y Class median score on the standardized history test.

   The full regression model has $E[Y|\mathbf{X}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$. Give $E[Y|\mathbf{X}]$ for the reduced model you would use to answer each of the following questions. Don't re-number the variables. Also, for each question please give the null hypothesis in terms of β values.

   (a) If you control for parents' education and income and for teacher's university background, does curriculum type affect test scores? (And why is it okay to use the word "affect?")

   (b) Controlling for parents' education and income and for curriculum type, is teacher's university background (two variables) related to their students' test performance?

   (c) Controlling for teacher's university background and for curriculum type, are parents' education and income (considered simultaneously) related to students' test performance?

   (d) Controlling for curriculum type, teacher's university background and parents' education, is parents' income related to students' test performance?

5. In a study of recovery from spinal cord injury, patients were randomly assigned to four different physical therapy programmes, which will be called A, B, C and D. The dependent variable is "mobility" (basically how well the patients can move around on their own) after two months, and severity of the initial injury is a covariate. Call the covariate x, and call the dummy variables p_j for $j = 1, \ldots, ?$.

   (a) Write the equation for a regression model that includes the possibility of regression lines that are not parallel.

   (b) Make a table with columns showing how the dummy variables are defined. Make D the reference category. Include a wider column in which you show $E(Y|x)$ for each treatment programme.

(c) In terms of the β coefficients of your model, what null hypothesis would you test to answer each of the following questions?

i. Are the four regression lines parallel?

ii. Are the slopes for treatments A, B and C equal?

iii. Are the slopes for treatments A, B and D equal?

iv. Is there an interaction between treatment programme an initial severity of the injury?

v. Holding initial severity of the injury constant at $x = 5$ (the definition of a "moderate" injury), do the treatments differ in their effectiveness?

vi. Holding initial severity of the injury constant at $x = 5$, which is more effective, treatment A or treatment C?

(d) Write the last three null hypotheses in matrix form as $H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{t}$.

6. Please return to the Census Tract data again. Fit a 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. If you can't remember what the regions are during the quiz, nobody will tell you.

Based on this model,

(a) What is k? The answer is a number.

(b) What is $\widehat{\beta}_4$? The answer is a number.

(c) 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.

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

ii. $H_0 : \beta_7 = 0$

iii. $H_0 : \beta_0 = 0$

(d) What proportion of the variation in crime rate is explained by the independent variables in this model? The answer is a number.

(e) What is the smallest value of $\widehat{\epsilon}_i$? The answer is a number.

(f) What is the largest value of $\widehat{\epsilon}_i$? The answer is a number.

(g) 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.

(h) 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.

(i) Controlling for all the other variables in the model, is percent High School graduates related to crime rate?

i. Give the null hypothesis in symbols.

ii. Give the value of the test statistic. The answer is a number from your printout.

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

iv. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

v. Allowing for other variables, census regions with a higher percentage of High School graduates tend to have _____ crime rates.

(j) Controlling for all the other variables in the model, is number of physicians related to crime rate?

i. Give the null hypothesis in symbols.

ii. Give the value of the test statistic. The answer is a number from your printout.

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

iv. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

v. Is there enough evidence to conclude that allowing for other variables, number of physicians is related to crime rate?

(k) Controlling for all the other variables in the model, is there a difference in crime rate between the Northeast and North Central regions?

i. Give the null hypothesis in symbols.

ii. Give the value of the test statistic. The answer is a number from your printout.

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

iv. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

v. State the conclusion in plain, non-statistical language. If there is a difference, say which region has a higher average crime rate!

(l) Controlling for all the other variables in the model, is there a difference in crime rate between the Northeast and South regions?

i. Give the null hypothesis in symbols.

ii. Give the value of the test statistic. The answer is a number from your printout.

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

iv. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

v. State the conclusion in plain, non-statistical language. If there is a difference, say which region has a higher average crime rate!

(m) Controlling for all the other variables in the model, is there a difference in crime rate between the South and West regions?

i. Give the null hypothesis in symbols.

ii. Give the value of the test statistic. The answer is a number from your printout.

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

iv. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

v. State the conclusion in plain, non-statistical language. If there is a difference, say which region has a higher average crime rate!

(n) 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?

(o) 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.

i. Give the null hypothesis in symbols.

ii. Give the value of the test statistic. The answer is a number from your printout.

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

iv. Do you reject the null hypothesis at $\alpha = 0.05$? Answer Yes or No.

v. What proportion of the remaining variation do these variables explain?

vi. 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?

(p) To be continued ...

Bring your printout to the quiz.
