STA 312f12 Assignment Seven[1]

Please bring your R printout for the last question to the quiz. The non-computer questions are practice for the quiz on Friday Nov. 2nd, and are not to be handed in.

1. 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).

   (a) 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.

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

(b) 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.

(c) 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.

(d) 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.

(e) 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.

2. 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?

3. If two events have equal probability, the odds ratio equals ____.

4. For a multiple logistic regression model, if the value of the kth independent variable is increased by c units and everything else remains the same, the odds of Y=1 are ____ times as great. Prove your answer.

5. For a multiple logistic regression model, let $P(Y_i = 1|x_{i,1}, \ldots, x_{i,p-1}) = \pi(\mathbf{x}_i)$. Show that a linear model for the log odds is equivalent to

$$\pi(\mathbf{x}_i) = \frac{e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}}{1+e^{\beta_0+\beta_1 x_1+\ldots+\beta_{p-1}x_{p-1}}} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1+e^{\mathbf{x}_i'\boldsymbol{\beta}}}$$

6. Write the log likelihood for the last question, and simplify it as much as possible.

7. A logistic regression model with no independent variables has just one parameter, β_0. It also the same probability $\pi = P(Y = 1)$ for each case.

(a) Write π as a function of β_0; show your work.

(b) The *invariance principle* of maximum likelihood estimation says the MLE of a function of the parameter is that function of the MLE. It is very handy. Now, still considering a logistic regression model with no independent variables,

i. Suppose $\overline{y}$ (the sample proportion of $Y = 1$ cases) is 0.57. What is $\widehat{\beta}_0$? Your answer is a number.

ii. Suppose $\widehat{\beta}_0 = -0.79$. What is $\overline{y}$? Your answer is a number.

8. Consider a logistic regression in which the cases are newly married couples with both people from the same religion, the independent variable is religion (A, B, C and None – let's call "None" a religion), and the dependent variable is whether the marriage lasted 5 years (1=Yes, 0=No).

   (a) Make a table with four rows, showing how you would set up indicator dummy variables for Religion, with None as the reference category.

   (b) Add a column showing the odds of the marriage lasting 5 years. The *symbols* for your dummy variables should not appear in your answer, because they are zeros and ones, and different for each row. But of course your answer contains β values.

   (c) What is the ratio of the odds of a marriage lasting 5 years or more for Religion C to the odds of lasting 5 years or more for No Religion? Answer in terms of the β symbols of your model.

   (d) What is the ratio of the odds of lasting 5 years or more for religion A to the odds of lasting 5 years or more for Religion B? Answer in terms of the β symbols of your model.

   (e) You want to test whether Religion is related to whether the marriage lasts 5 years. State the null hypothesis in terms of one or more β values.

   (f) You want to know whether marriages from Religion A are more likely to last 5 years than marriages from Religion C. State the null hypothesis in terms of one or more β values.

   (g) You want to test whether marriages between people of No Religion have a 50-50 chance of lasting 5 years. State the null hypothesis in terms of one or more β values.

9. This question uses an R data set called `birthwt`. In R's Packages and Data menu, Select Package Manager, and make sure MASS is checked. Then, `help(birthwt)` will tell you about the data set. It's often used for logistic regression, but this time we're just going to do ordinary regression. The response variable will be the child's birth weight and the explanatory variables will be Mother's age, Mother's weight, and Race.

   (a) For each of the following questions, be able to give the null hypothesis in symbols. Give the value of the test statistic (t or F), the p-value, and whether you reject H_0 at $\alpha = 0.05$. How would you state the conclusion in plain language, with *no* statistical terminology? (You can say "allowing for" instead of "controlling for.")

      i. Controlling for mother's age and weight, do White and Black mothers differ in the mean weight of their babies? If one is more (meaning H_0 is rejected), which one race has heavier babies on average and how can you tell?

      ii. Controlling for mother's age and weight, do White and Other mothers differ in the mean weight of their babies? If one is more (meaning H_0 is rejected), which one race has heavier babies on average and how can you tell?

iii. Controlling for mother's age and weight, do Black and Other mothers differ in the mean weight of their babies? If one is more (meaning H_0 is rejected), which one race has heavier babies on average and how can you tell?

iv. Controlling for mother's weight and race, is the mother's age related to her baby's weight?

v. Controlling for mother's age and race, is the mother's weight related to her baby's weight?

vi. Controlling for mother's age and weight, is the mother's race related to her baby's weight? This is one test.

vii. Controlling for mother's race, are the mother's age and/or weight related to her baby's weight? This is one test.

viii. Are any of the explanatory variables related to baby's birth weight? This is one test.

(b) It's helpful to be able to do a general linear test of $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$ directly in R. To do it, you need to download a package. There is more than one possibility, but the `car` package is okay. In R's Packages and Data menu, Select Package Installer. Click on Get List. With Install Dependencies checked, select `car` and click Install Selected. I had to quit and restart R, then again in the Package Manager I had to *select* car. Then `help(linearHypothesis)`.

i. Repeat the test of race controlling for age and weight, just to verify that you can get the same test statistic.

ii. With White as the reference category, repeat the test comparing Black to Other controlling for age and weight.
