STA 2101/442 Assignment Eight[1]

Please bring printouts of your SAS log and list files from Question 6 to the quiz. The non-computer questions are just practice for the quiz, 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). In case you know what an interaction is, there should be no interactions in your model. We haven't gotten to interactions yet.

   (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.

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

(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.

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. 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.

   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?

4. 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 $\mathbf{X}_i$?

5. For a model with just one 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.

6. This question uses the `sat.data` file you first saw in Assignment 1. There is a link on the course web page in case the one in this document does not work. This is just to get your feet wet with SAS and unix, in case you have not used these tools before. Using SAS, get sample sizes, means and standard deviations for all three variables. That's it. Bring your log file and your list file to the quiz.

   One little problem is that SAS is unable to use the variable names on the first line of the data file — or more precisely, I don't know how to make SAS use them. You can skip the first line by putting `firstobs=2` right after the name of the data file, on the `infile` statement.
