Statistics 512: Homework 5

Divisions 4 and 5: Due 11:59pm, Monday, March 27, 2017

A reminder – Please do not hand in any unlabeled or unedited SAS output. Include in your write-up only those results that are necessary to present a complete solution. In particular, questions must be answered in order (including graphs), and all graphs must be fully labeled (main title should include question number, and all axes should be labeled). Don't forget to put all necessary information (see course policies) on the first page. Include the SAS input for all questions at the very end of your homework. You will often be asked to continue problems on successive homework assignments. So save all your SAS code.

Important Note *– Every graph or plot you create should have your name printed as a subtitle. Consequently, any graph with no name will result in a* **20% points off** *on that question. Also, please attach your code at the end; any homework with no code provided will result in a* **50% points off** *on the entire assignment, NO EXCEPTIONS.*

For the following problems use the computer science data that we have been discussing in class. You can get a copy of the data set `csdata.dat` from the class website. The variables are: `id`, a numerical identifier for each student; `GPA`, the grade point average after three semesters; `HSM`; `HSS`; `HSE`; `SATM`; `SATV`, which were all explained in class; and `GENDER`, coded as 1 for men and 2 for women.

1. In this exercise you will illustrate some of the ideas described in Chapter 7 of the text related to the extra sums of squares.

   (a) Create a new variable called `SAT` which equals `SATM + SATV` and run the following two regressions:

   i. predict `GPA` using `HSM`, `HSS`, and `HSE`;

   ii. predict `GPA` using `SAT`, `HSM`, `HSS`, and `HSE`.

   Calculate the extra sum of squares for the comparison of these two analyses. Use it to construct the F-statistic – in other words, the general linear test statistic – for testing the null hypothesis that the coefficient of the `SAT` variable is zero in the model with all four predictors. What are the degrees of freedom for this test statistic?

   (b) Use the `test` statement in `proc reg` to obtain the same test statistic. Give the statistic, degrees of freedom, p-value and conclusion.

   (c) Compare the test statistic and p-value from the `test` statement with the individual t-test for the coefficient of the `SAT` variable in the full model. Explain the relationship.

2. Run the regression to predict `GPA` using `SATM`, `SATV`, `HSM`, `HSE`, and `HSS`. Put the variables in the order given above in the `model` statement. Use the `SS1` and `SS2` options in the `model` statement.

   (a) Add the Type I sums of squares for the five predictor variables. Do the same for the Type II sums of squares. Do either of these sum to the model sum of squares? Are there any predictors for which the two sums of squares (Type I and Type II) are the same? Explain why.

(b) Verify (by running additional regressions and doing some arithmetic with the results) that the Type I sum of squares for the variable `SATV` is the difference in the model sum of squares (or error sum of squares) for the following two analyses:

i. predict `GPA` using `SATM`, `SATV`;

ii. predict `GPA` using `SATM`.

3. Create an additional variable called `HS` that is the sum of the three high school scores (`HSE + HSS + HSM`). Run the regression to predict `GPA` using a variety of variables, including `HS` and `SAT`, as described below. Summarize the results by making a table giving the percentage of variation explained (R^2) by each of the following models:

(a) `SATM` as the explanatory variable

(b) `SATV` as the explanatory variable

(c) `HSM` as the explanatory variable

(d) `HSS` as the explanatory variable

(e) `HSE` as the explanatory variable

(f) `SATM` and `SATV` as the explanatory variables

(g) `SAT` (`=SATM+SATV`) as the explanatory variable

(h) `HSM`, `HSS`, and `HSE` as the explanatory variables

(i) `HS` (`=HSM+HSS+HSE`) as the explanatory variable

(j) `SATM`, `SATV`, `HSM`, `HSS`, and `HSE` as the explanatory variables

(k) `SAT` and `HS` as the explanatory variables

(Please do not include the SAS output for all these models. Only the R^2 value is needed. Note that you can run `proc reg` with multiple `model` statements to save typing.)

4. A data set contains 50 observations. There are 4 explanatory variables: A, B, C, and D. Use the following results:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1.3972 & -1.8892 \times 10^{-3} & -3.6060 \times 10^{-3} & -1.3523 \times 10^{-3} & -2.9728 \times 10^{-2} \\ -1.8892 \times 10^{-3} & 5.0363 \times 10^{-5} & -6.8773 \times 10^{-6} & -3.9875 \times 10^{-6} & -5.0387 \times 10^{-6} \\ -3.6060 \times 10^{-3} & -6.8773 \times 10^{-6} & 4.9685 \times 10^{-5} & -8.6113 \times 10^{-6} & 5.4578 \times 10^{-5} \\ -1.3523 \times 10^{-3} & -3.9875 \times 10^{-6} & -8.6113 \times 10^{-6} & 4.7933 \times 10^{-5} & -1.3931 \times 10^{-5} \\ -2.9728 \times 10^{-2} & -5.0387 \times 10^{-6} & 5.4578 \times 10^{-5} & -1.3931 \times 10^{-5} & 8.0975 \times 10^{-4} \end{bmatrix}$$

$$\mathbf{b} = \begin{bmatrix} 469.7658 \\ -2.4148 \\ 3.3341 \\ -4.3285 \\ 0.9546 \end{bmatrix}$$

$$MSE = 1963.48714$$

(a) Obtain a 94% confidence interval for β_4 (the coefficient for D).

(b) You wish to test $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$. That is, you wish to determine if variable A provides significant power for Y when variables B, C, and D are already in the model. Obtain the test statistic for this hypothesis test and determine if you would accept or reject the null hypothesis ($\alpha = 0.05$). You should give either a critical value or a p-value to support your conclusion.

(c) Obtain a 94% confidence interval for the mean (expected) response when $A = 30$, $B = 30$, $C = 40$, and $D = 40$. *hint:* $\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h = .0751$ when $\mathbf{X}_h' = (1, 30, 30, 40, 40)$.

(d) Obtain a 94% prediction interval for a single response when $A = 30$, $B = 30$, $C = 40$, and $D = 40$.