**Statistics 512: Solution to Homework#5**

For the following 6 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`;

   **Solution:** Notice the analysis of variance with respect to the next problem.

   Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value |
|---|---|---|---|---|
| Model | 3 | 27.71233 | 9.23744 | 18.86 |
| Error | 220 | 107.75046 | 0.48977 | |
| Corrected Total | 223 | 135.46279 | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.69984 | R-Square | 0.2046 |
| Dependent Mean | 2.63522 | Adj R-Sq | 0.1937 |
| Coeff Var | 26.55711 | | |

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

   **Solution:** The analysis of variance table is as follows:

   Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value |
|---|---|---|---|---|
| Model | 4 | 27.88746 | 6.97187 | 14.19 |
| Error | 219 | 107.57533 | 0.49121 | |
| Corrected Total | 223 | 135.46279 | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.70086 | R-Square | 0.2059 |
| Dependent Mean | 2.63522 | Adj R-Sq | 0.1914 |
| Coeff Var | 26.59603 | | |

   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?

   **Solution:** Using the definition of $SSE$,

$$\begin{aligned} SSM(sat|hsm, hss, hse) &= SSE(hsm, hss, hse) - SSE(sat, hsm, hss, hse) \\ &= (107.750 - 107.575) = 0.175 \end{aligned}$$

   or, using the $SSM$ definition,

$$\begin{aligned} SSM(sat|hsm, hss, hse) &= SSM(sat, hsm, hss, hse) - SSM(hsm, hss, hse) \\ &= (27.887 - 27.712) = 0.175 \end{aligned}$$

The degrees of freedom for the numerator are $(220 - 219) = (4 - 3) = 1$.
Thus, the $F$ statistic (tested against an $F$ distribution with $(1, 219)$ degrees of freedom) is

$$\begin{aligned} F &= \frac{SSM(sat|hsm, hss, hse)/1}{MSE(full)} \\ &= \frac{0.175/1}{0.491} = 0.356. \end{aligned}$$

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

**Solution:** The SAS output gives the same $F$ ratio 0.36, which gives a $p$-value of 0.5511, thus failing to reject the null hypothesis. There is no evidence of a linear relationship with SAT when HSM, HSE, and HSS are in the model.

Test nosat Results for Dependent Variable gpa

| Source | DF | Mean Square | F Value | Pr > F |
|---|---|---|---|---|
| Numerator | 1 | 0.17513 | 0.36 | 0.5511 |
| Denominator | 219 | 0.49121 | | |

(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.

**Solution:** The $t$ statistic value is 0.60, which is consistent with the other test where $F = t^2 = 0.36$ . The $p$-values for the two tests are the same (0.5511). They are both testing the same null hypothesis: namely, that the coefficient of SAT is equal to zero when the three high school scores are included in the model.

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|---|
| Intercept | 1 | 0.43685 | 0.39053 | 1.12 | 0.2645 |
| sat | 1 | 0.00019948 | 0.00033408 | 0.60 | 0.5511 |
| hsm | 1 | 0.16243 | 0.03700 | 4.39 | <.0001 |
| hss | 1 | 0.03245 | 0.03774 | 0.86 | 0.3908 |
| hse | 1 | 0.04508 | 0.03875 | 1.16 | 0.2460 |

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.

**Solution:** The required SAS output is as follows:

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 5 | 28.64364 | 5.72873 | 11.69 | <.0001 |
| Error | 218 | 106.81914 | 0.49000 | | |
| Corrected Total | 223 | 135.46279 | | | |

| | | | |
|---|---|---|---|
| Root MSE | 0.70000 | R-Square | 0.2115 |
| Dependent Mean | 2.63522 | Adj R-Sq | 0.1934 |
| Coeff Var | 26.56311 | | |

Parameter Estimates

| Variable | DF | Parameter Estimate | Standard Error | t Value | Pr > \|t\| | Type I SS | Type II SS |
|---|---|---|---|---|---|---|---|
| Intercept | 1 | 0.32672 | 0.40000 | 0.82 | 0.4149 | 1555.54591 | 0.32691 |
| satm | 1 | 0.00094359 | 0.00068566 | 1.38 | 0.1702 | 8.58293 | 0.92800 |
| satv | 1 | -0.00040785 | 0.00059189 | -0.69 | 0.4915 | 0.00090549 | 0.23265 |
| hsm | 1 | 0.14596 | 0.03926 | 3.72 | 0.0003 | 17.72647 | 6.77243 |
| hse | 1 | 0.05529 | 0.03957 | 1.40 | 0.1637 | 1.89119 | 0.95680 |
| hss | 1 | 0.03591 | 0.03780 | 0.95 | 0.3432 | 0.44214 | 0.44214 |

(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.

**Solution:** The sum of the Type I $SS$ is 28.64, while the sum of the Type II $SS$ is 9.33 .
The Type I $SS$ sum to the model $SS$ because the Type I $SS$ are a hierarchical extra SS in the sense that

$$\begin{aligned} SSM(satm, satv, hsm, hse, hss) &= SSM(satm) + SSM(satv|satm) + \\ &\quad SSM(hsm|satm, satv) + SSM(hse|satm, satv, hsm) + \\ &\quad SSM(hss|satm, satv, hsm, hse). \end{aligned}$$

This is true because the extra sum of squares is defined as

$$SSM(X_k|X_1, \ldots, X_{k-1}) = SSM(X_1, \ldots, X_k) - SSM(X_1, \ldots, X_{k-1}).$$

The Type I and II $SS$ are the same for hss since it is last in the list, so that they both represent $SSM(hss|satm, satv, hsm, hse)$.

(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`;

**Solution:** The SAS output for this model is as follows:

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 2 | 8.58384 | 4.29192 | 7.48 | 0.0007 |
| Error | 221 | 126.87895 | 0.57411 | | |
| Corrected Total | 223 | 135.46279 | | | |

ii. predict `GPA` using `SATM`.

**Solution:** The SAS output for this model is as follows:

Analysis of Variance

| Source | DF | Sum of Squares | Mean Square | F Value | Pr > F |
|---|---|---|---|---|---|
| Model | 1 | 8.58293 | 8.58293 | 15.02 | 0.0001 |
| Error | 222 | 126.87986 | 0.57153 | | |
| Corrected Total | 223 | 135.46279 | | | |

**Solution:** The Type I $SS$ is

$$SSM(satv|satm) = 0.00090549 \approx 0.00091 \text{ (from part a)}$$

We also have that the extra $SS$ is

$$SSM(satm, satv) - SSM(satm) = 8.58384 - 8.58293 = 0.00091$$

or

$$SSE(satm) - SSE(satm, satv) = 126.87986 - 126.87895 = 0.00091.$$

These are the same except for roundoff error.

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 given models.

**Solution:** The table of the $R^2$ is given below.

| Variables | $R^2$ |
|---|---|
| satm | 0.0632 |
| satv | 0.0131 |
| hsm | 0.1905 |
| hss | 0.1085 |
| hse | 0.0835 |
| satm, satv | 0.0634 |
| sat | 0.0446 |
| hsm, hss, hse | 0.2046 |
| hs | 0.1803 |
| satm, satv, hsm, hss, hse | 0.2115 |
| sat, hs | 0.1842 |