

Data set DSP2 consist of $n=224$ observations on

- Grade point average (GPA) after three semesters (the dependent variable).
- High school Math grades (HSM)
- High school Science grades (HSS)
- High school English grades (HSE)
- SAT Math (SATM)
- SAT Verbal (SATV)
- Gender (1 = male, 2 = female)

1. In this project you will illustrate some of the ideas described in Chapter 7 of the text related to the extra sum 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.Take the difference between the SSEs for the two analyses and construct the F statistic (i.e., general linear model 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 on the model statement. Use the SS1 and SS2 options on the model statement.
 - (d) 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 sums 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.
 - (e) 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:
- SATM as the explanatory variable
 - SATV as the explanatory variable
 - HSM as the explanatory variable
 - HSS as the explanatory variable
 - HSE as the explanatory variable
 - SATM and SATV as the explanatory variables
 - SAT=SATM+SATV as the explanatory variable
 - HSM, HSS, and HSE as the explanatory variables
 - HS=HSM+HSS+HSE as the explanatory variable
 - SATM, SATV, HSM, HSS, and HSE as the explanatory variables
 - SAT and HS as the explanatory variables
 - HSM, HSS, HSE, SATM, SATV, HSM*HSE and SATM*SATV.

(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. In a DATA step, create a new variable GI that has values 1 for women and 0 for men. Run a regression to predict GPA using the explanatory variables HSM, HSS, HSE, SATM, SATV, and GI (Do not include any interaction terms).
- Give the fitted regression line for women.
 - Give the fitted regression line for men.
5. Use the C_p criterion to select the best subset of variables for this problem (i.e. use the options “ / selection = cp b;”). Use only the original six explanatory variables, not HS or SAT, and use GI and not Gender. Summarize the results and explain your choice of the best model.
6. Check the assumptions of this “best” model using all the usual plots (you know what they are by now). Do there appear to be any problems such as multicollinearity? Explain.