**SAT Taking & Scoring Behavior**

The data for this project are in the file satscores.xls on the class webpage. The dataset, which is from the National Longitudinal Survey of the High School Class of 1972 (NLS-72), includes the following variables:

sat combined verbal and math SAT score

satobs dummy variable ( =1 if respondent’s SAT score is observed; =0 otherwise)

rank class rank (percentile)

mlhs =1 if respondent’s mother did not complete high school; =0 otherwise

mcol =1 if respondent’s mother completed one or more years of college; =0 otherwise

flhs =1 if respondent’s father did not complete high school; =0 otherwise

fcol =1 if respondent’s father completed one or more years of college; =0 otherwise

black =1 if respondent is African American; =0 otherwise

hisp =1 if respondent is Hispanic; =0 otherwise

asian =1 if respondent is Asian-American; =0 otherwise

female =1 if respondent is female; =0 otherwise

The NLS respondents were given tests in inductive reasoning, math, memory, perception, reading comprehension, and vocabulary.

rdsc scaled high school reading test score (mean=50, SD=10)

vocab scaled high school vocabulary test score

pict scaled high school picture test score

lgsc scaled high school letter groups test score

matsc scaled high school mathematics test score

mosaic scaled high school mosaic comparison test score

nsib number of respondent’s siblings

1. Of the variables listed above, which are potential dependent variables and which are potential explanatory variables?
2. Sort the data in descending order by SAT score (highest to lowest). Insert two new worksheets; copy the data for those who took the SAT into one worksheet and the data for those who didn’t take the SAT into the other worksheet. Fill out the table below.

|  |  |  |  |
| --- | --- | --- | --- |
| **Average** | **Entire Sample** | **SAT-Takers** | **Non-SAT-Takers** |
|  SAT Score |  |  |  |
|  Class Rank |  |  |  |
|  # Siblings |  |  |  |
| **Proportion** |  |  |  |
|  Taking the SAT |  |  |  |
|  Male |  |  |  |
|  Female |  |  |  |
|  White |  |  |  |
|  Black |  |  |  |
|  Hispanic |  |  |  |
|  Asian |  |  |  |
| **Number of Obs. (N)** |  |  |  |

HINT: Start by filling out the row for the Entire Sample.

1. Using regression to find the means for various groups
	1. In your worksheet with data on SAT test-takers, insert a column to the right of the column with the variable ‘female’ and create a new variable called ‘male’.
	2. Run a regression with ‘sat’ as the dependent variable and ‘female’ and ‘male’ as the explanatory variables. **Be sure to check the box labeled ‘Constant is Zero’**. The regression you are running is  (no *β0*) and the coefficient estimates for *β1* and *β2* are the average SAT scores for females and males, respectively, who take the SAT test.

**IMPORTANT NOTE: Because everyone *must* be either a female or a male, these two variables cannot be included in the same regression *unless* you set the intercept (*β0*) to zero.**

* 1. Follow a similar procedure to find the average *class rank* for females and males who take the SAT test. 
	2. Follow a similar procedure to find the average SAT score among SAT test-takers for each race category. 
1. Assume that the dependent variable is SAT score and that we are only working with the subsample of respondents that actually took the SAT test. Explain how you expect race, gender, class rank, parents’ education, family size, and other test scores will affect respondents’ SAT scores (*i.e.,* I’m asking you for the expected sign on the regression coefficient of each variable). *Why* do you have these expectations?
2. Again, working with the subsample of respondents that actually took the SAT test, run a regression with ‘sat’ as the dependent variable and the ‘flhs’ and ‘fcol’ dummy variables as explanatory variables (be sure to include a constant term).
	1. What is the ‘reference group’ in this regression?
	2. What is the average SAT score for respondents in the reference group?
	3. What is the average SAT score for respondents with fathers who have at least one year of college?
	4. What is the interpretation of the coefficient estimate on ‘fcol’?
	5. Does this simple model do a good job of explaining the SAT scores that we observe in the data? What other variables would you include? Run another regression with these additional variables and determine which ones are statistically significant in explaining variation in SAT scores.
3. Go back to the original data (on test-takers and non-test-takers). Run a regression with ‘satobs’ as the dependent variable and ‘rank’ as the explanatory variable (be sure to include a constant term). 
	1. In this regression, you are examining the determinants of the decision to *take* the SAT test. Based on your regression results, what is the predicted probability that a person ranked in the 75th percentile in her class will take the SAT test? What about for a person ranked in the 76th percentile?
	2. What is the interpretation of the estimated value of *β1*?
	3. What other explanatory variables might you include in this regression and what signs to you expect on each coefficient estimate? Run a regression with these additional explanatory variables and interpret the coefficient estimates.