SAT Taking & Scoring Behavior

ANSWER KEY

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
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

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

1. Of the variables listed above, which are potential dependent variables and which are potential explanatory variables?
   Possible dependent variables include sat, satobs, rank; potential explanatory variables include rank, mlhs, mcol, flhs, fcol, black, hisp, asian, female, rdsc, vocab, pict, lgsc, matsc, mosaic, nsib

2. Taking averages of all the variables allows you to fill in the first column of the table (note that the average of a dummy variable – like satobs, male, female, white, black, hisp, and asian – yields the proportion of the sample for whom the dummy variable equals 1.

Average	**Entire Sample**	**SAT-Takers**	**Non-SAT-Takers**
SAT Score	343.31	961.28	0
Class Rank	57.59	69.48	50.98
# Siblings	3.07	2.60	3.33
Proportion			
Taking the SAT	35.71%	100%	0%
Male	46.81%	49.63%	45.25%
Female	53.19%	50.37%	54.75%
White	85.19%	90.07%	82.49%
Black	9.25%	6.59%	10.72%
Hispanic	4.49%	1.63%	6.08%
Asian	1.07%	1.71%	0.71%
Number of Obs. (N)	6370	2275	4095

3. Using regression to find the means for various groups
    a. 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'.
    Formula is 1-female

    b. 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 $sat_i = \beta_1 female_i + \beta_2 male_i + \varepsilon_i$ (no β_0) and the coefficient estimates for β_1 and β_2 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 (β_0) to zero.

    β_1-hat = 936.07 (average SAT score for females taking the SAT test)
    β_2-hat = 986.87 (average SAT score for males taking the SAT test)
    Conduct a test to see if these two means are statistically different from one another… in Excel:

    c. Follow a similar procedure to find the average *class rank* for females and males who take the SAT test. $rank_i = \beta_1 female_i + \beta_2 male_i + \varepsilon_i$

    β_1-hat = 73.8 (average class rank for females taking the SAT test)
    β_2-hat = 65.1 (average class rank for males taking the SAT test)
    Conduct a test to see if these two means are statistically different from one another… in Excel:

    d. Follow a similar procedure to find the average SAT score among SAT test-takers for each race category. $sat_i = \beta_1 white_i + \beta_2 black_i + \beta_2 hisp_i + \beta_2 asian_i + \varepsilon_i$

    Formula for white is 1-black-hisp-asian.

    β_1-hat = 980.9 (average SAT score for whites taking the SAT test)
    β_2-hat = 733.7 (average SAT score for blacks taking the SAT test)
    β_3-hat = 769.2 (average SAT score for Hispanics taking the SAT test)
    β_4-hat = 986.4 (average SAT score for Asians taking the SAT test)

4. 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?

Generally speaking, any variable that is associated with better academic ability and family affluence will probably also be positively associated with SAT score. So, I would expect positive coefficient estimates (β's) on class rank, mcol, fcol, white or asian (depending on which race variables you include), and any of the other test scores (particularly reading and math scores). I would expect negative coefficient estimates on mlhs, flhs, black, Hispanic, and nsibs. All that a negative coefficient estimate means is that individuals with parents who dropped out of high school or who belong to traditionally underrepresented groups in college are likely to have lower SAT scores than their peers with parents who have high school diplomas and who are white or Asian. These expectations probably come from the human capital production function theory… better inputs yield better outputs. Some of these variables (like race and parental education) are just proxies for the inputs we'd prefer to include (like attributes of your neighborhood or school).

5. 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).

SUMMARY OUTPUT **SAT Score Regressed on Father's Education**

Regression Statistics	
Multiple R	0.256444648
R Square	0.065763857
Adjusted R Square	0.064941466
Standard Error	195.0878234
Observations	2275

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	6086943.454	3043471.727	79.96665777	2.74645E-34
Residual	2272	86470636.08	38059.25884		
Total	2274	92557579.54			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	919.0752864	7.892408547	116.4505462	0	903.5982053	934.5523676
flhs	-34.24601812	12.45462056	-2.749663706	0.006012882	-58.66963632	-9.822399923
fcol	87.75884437	9.624988251	9.117813142	1.6396E-19	68.8841595	106.6335292

a. What is the 'reference group' in this regression?
The omitted category is fathers with exactly a high school education (no more, no less).

b. What is the average SAT score for respondents in the reference group?
Comes from the estimate of the intercept… 919.1.

c. What is the average SAT score for respondents with fathers who have at least one year of college?
87.8 points higher than the reference group… 919.1+87.8 = 1006.9.

d. What is the interpretation of the coefficient estimate on 'fcol'?
The precise interpretation of the coefficient estimate on fcol is the marginal effect of having a father with some college education on an individual's SAT score.

e. 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.
With an adjusted R^2 of only 0.065, only 6.5% of the variation in SAT scores in our data is explained by variation in fathers' level of education. Clearly, we could do a lot better by including more variables in our regression. For example, if we include class 'rank' and both mothers' and fathers' education dummy variables, the adjusted R^2 jumps to 0.355!

6. 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).
$satobs_i = \beta_0 + \beta_1 rank_i + \varepsilon_i$

SUMMARY OUTPUT **REGRESSION OF SATOBS ON CLASS RANK**

Regression Statistics	
Multiple R	0.31990145
R Square	0.102336938
Adjusted R Square	0.102195973
Standard Error	0.45404938
Observations	6370

Note that all variables are statistically significant (*i.e.*, different from zero) because the t-stats are all bigger than 2 in absolute value terms.

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	149.6677718	149.6677718	725.9757574	1.6115E-151
Residual	6368	1312.832228	0.20616084		
Total	6369	1462.5			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	0.038620738	0.0131193	2.943810901	0.003253692	0.012902495	0.064338981
rank	0.005531164	0.000205284	26.9439373	1.6115E-151	0.005128738	0.00593359

This is called a Linear Probability Model… when your dependent variable is binary and you run OLS.

a. 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?
You can use the coefficient estimates from the above regression to predict the probability that someone would take the SAT test based on their class rank.
   - For someone with a percentile rank of 75, their predicted probability of taking the SAT would be 0.0386 + (0.0055*75) = .4511 or 45.11%.
   - For someone with a percentile rank of 76, the predicted probability of taking the SAT would be 0.0386 + (0.0055*76) = .4566 or 45.66%.

b. What is the interpretation of the estimated value of β_1?
Note that the difference between the two is 0.0055, which is the estimated coefficient on 'rank'. This is not a coincidence… it is the marginal effect of a 1 unit increase in 'rank' on the probability of taking the SAT. It doesn't matter whether that increase is from the 75th to the 76th percentile or from the 1st to the 2nd percentile, the marginal boost in the probability that they take the exam is still 0.0055 or 0.55%.

c. 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.
You might expect the same variables that are good predictors of individuals' SAT scores to also be good predictors of individuals' decisions to take the test. These additional explanatory variables would likely have the same signs that we anticipated in that exercise.