DOES MORE MONEY RESULT IN BETTER SAT SCORES?

As you are probably well aware, SAT scores are one measure of a school’s performance. Given a choice, most parents would probably prefer their children went to a school with high SAT scores, for example. A continuing debate in public policy is whether more money should be spent on public schools, or whether money should be spent more efficiently. (Of course, you might argue that both policies might be pursued at the same time, but for some reason this point rarely comes up in public debate.)

Of course this dataset will not settle this question. But it does provide an example of the types of data that public policy makers consider when making decisions and crafting arguments.

The variables in this data set, all aggregated to the state level, were extracted from the 1997 Digest of Education Statistics, an annual publication of the U.S. Department of Education. This data set contains variables that address the relationship between public school expenditures and academic performance, as measured by the SAT. The variables are: name of state, current expenditure per pupil (measured in thousands of dollars per average daily attendance in public elementary and secondary schools), average pupil/teacher ratio in public elementary and secondary schools during Fall 1994, estimated average annual salary of teachers in public elementary and secondary schools during 1994-95 (in thousands of dollars), percentage of all eligible students taking the SAT in 1994-95, average verbal SAT score in 1994-95, average math SAT score in 1994-95, and average total score on the SAT, 1994-95.

In today’s lab we will explore the relationship between state expenditures on public elementary and secondary education and the performance of students on the SAT exam.

How to handle data forms other than .dta Because data are not always in Stata’s .dta form, sometimes more code is needed to read in a data set. Often data are delimited by tabs or commas. When this is the case, the "insheet command is used. Try this "insheet command with today’s dataset and see what the results are with the "browse" command
. insheet using http://www.stat.ucla.edu/labs/datasets/sat.dat
. browse

Question 1: How many variables are present in the data set? What went wrong with the insheet command?

When you see something like this, it’s often because the columns of data are separated by spaces rather than tabs or commas. To read in data like this, we use the "in file" command. With this command, we can specify the names of the variables and read in the data all at one time. Clear out what you have in Stata now with the clear command, and read in the data properly.

. clear
. infile str15 state cost ratio salary percent verbal math total using http://www.stat.ucla.edu/labs/datasets/sat.dat

The "str15" part of the command lets Stata know that the first variable "state" is a string variable that takes up the first 15 positions.

Question 2: Now what do you see when you type the "browse" command?

It is good practice to annotate your dataset as much as possible. This is helpful if you return to working with the data after a break. Using the description of the dataset provided above, enter in descriptive labels for each of the variables. Example:

. label var salary "estimated avg teacher salaries in thousands for 94-95"

Correlations and causations

Correlations are frequently used to make a point about relationships. You’ll frequently hear politicians are newspaper columnists claim something along
the line of “State spending on education is positively correlated with SAT scores, and therefore we should increase our State’s spending.”

As you should know by now, correlations measure the strength of the linear relationship between two variables. The fact that there is a strong linear relationship, however, says nothing about the causal relationship. That is, it doesn’t mean that spending causes higher SAT scores.

Question 3: Take a guess: will the correlation between total SAT scores and cost per pupil be positively or negatively correlated? What value do you think the correlation will have?

To get Stata to calculate the correlation, type

. corr total cost

Question 4: Were you right? Are you surprised? Why or why not?

Question 5: Interpret this correlation for someone who has read it in a newspaper.

Question 6: What value would you get for the correlation if you typed the variables in reverse order? That is, if you typed corr cost total?

Question 7: Which variable do you think will have the greatest correlation (disregarding positive or negative signs) with total? Which do you think will have a correlation closest to zero?

Question 8: Which variables do you think will have a negative correlation
You can see all possible correlations at once by including more variables in
the `bf corr` command. This produces a correlation matrix.

```
. corr total math verbal cost ratio salary percent
```

This tells us, for example, that the correlation between the total SAT and the
verbal scores is .9915. And the correlation between verbal and math scores is
.9703. The diagonal of this matrix will always have 1.0000 along it, because
every variable is perfectly correlated with itself.

*Question 9: What is the correlation between ratio and cost? Explain why it is negative.*

---

**A Picture Paints 1000 Words**

You might have been surprised by the negative correlation between spend-
ing and SAT scores. But remember the correlation is a simple summary of
sometimes complex information. We can get a much better picture of what’s
going on by looking at a graph. In fact, a graph should always be our first
step. Calculating a correlation coefficient should occur only if we think the
relation is linear.

```
. graph total cost
```

*Question 10: Describe this plot. What is the trend? Does it look linear? Are there any unusual observations?*

*Stata* provides at least two different methods for checking up on particular
points. One is to label the points with their observation numbers:
. graph total cost, symbol([_n])

*Question 11:* Use the `browse` command to determine which states were entries 30 and 32.

Another method can be a bit confusing if you have too much data. Still, it is frequently useful.

. graph total cost, symbol([state])

*Question 12:* Which states have performance much like California?

You might have noticed that four states – Alaska, Connecticut, New Jersey and New York – look different from the others.

*Question 13:* If we remove these four points, will the trend look more linear? How do you think the correlation coefficient will change?

Removing points is a little tricky. Don’t worry too much about the inner logic of these commands, but if you’re interested we’ll explain at the end of this lab. For now, just type along.

. gen unusual = total if _n==2 | _n==7 | _n==30 | _n==32
. gen usethese = total if missing(unusual)==1
. graph usethese cost, symbol([state])

*Question 14:* What do you think the correlation between total SAT and cost will be for these 46 states?

. corr usethese cost
The correlation is only slightly closer to 0. But notice that it is difficult to
tell if there is a linear relation hidden in there somewhere.

**Using Regression to Summarize a Linear Relation**

Let’s again consider the entire dataset. Make a graph of total versus cost. Now we’ll compute the regression line, the line that “best fits” these data.

`regress total cost predict predtot graph total predtot cost, symbol(oi) connect(.l)`

note that the code is connect(.l) with a lowercase letter l, not a number 1

The first line of code here generates the regression summary information that you saw in your results window. The second line created a new variable called ”predtot” that you should now see in your variables window. The ”predtot” variable contains the predicted values of the total SAT score as predicted by your regression line. The last line creates the scatter plot with the overlaid regression line. The symbol(oi) tells Stata that we want to graph total against cost with a small circle and plot predtot against cost with an invisible marking. The connect(.l) indicates that we don’t want to connect the points in the the total vs. cost plotting, but want to connect our invisible predtot vs. cost points with a line.

*Question 15: What is the equation of the regression line?*

*Question 16: Interpret the slope.*

A shortcut method for creating this plot hsaves you the trouble of typing in three lines of code for the one nice graph. Try the shortcut ”scatter” command now to verify that the results are the same.

```
. scatter total cost
```
We have already noted that there are some outlier states. What effect do they have on the regression equation?

Question 17: Redo the regression without the four outlier states. What is the effect on the slope?

Notes

Aggregated Data. These data consisted of aggregated data. This means that rather than compare the actual amount spent at each school with the average SAT at each school, we instead considered all of the schools of the states aggregated together. In practice this can create pitfalls for interpretation in a regression context. One big effect is that often regression relationships look stronger than they actually are. Sometimes much stronger. The reason for this is that the aggregation hides a lot of the variation in the data. Keep in mind that each point on this graph might represent ten thousand schools, and that there might be greater differences between schools within California than there are between the average of all California schools and the average of all schools from any other state.

The “if” command used to remove points. Here’s a quick description of how these commands worked. Here are the commands, one by one:

. gen unusual = total if _n==1 | _n==2 | _n==30 | _n==32

This creates a new variable named “unusual.” Unusual will have the same values as the variable total, but only in the 1st, 2nd, 30th, and 32nd positions. Everywhere else it will be coded with the value “missing”. So unusual has only four values and the rest are missing.

. gen usethese = total if missing(unusual)==1

We next want to create a new variable that has the same values as total but is blank (or missing) for the four outlier states. This command creates a new variable named usethese and sets it equal to the values in total only at the locations where the variable unusual is missing values. The function
missing(varname) returns a 1 wherever varname is missing a value and a 0 everywhere else.

Take Home

Use these SAT data to answer the following questions:

Question 18: To attract good teachers, you need to offer high salaries. If you have good teachers, students should have higher SAT scores. Therefore, states with high teacher salaries should have high SAT scores. Examine the data. Do the data refute this reasoning, or are the other factors at play?