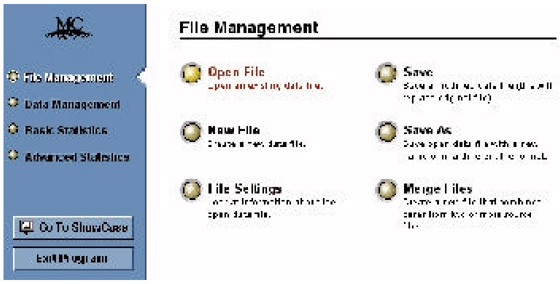
LEARNING TO USE THE MicroCase DATA ANALYSIS SOFTWARE: EXPLORING AMERICANS' BELIEFS ABOUT THE THEORY OF EVOLUTION

MicroCase is a collection of software and databases that will allow you to test a wide range of hypotheses for either this course or most others in the social sciences. The data in these databases are the same resources used by sociologists and political scientists all over the country.

We have, for instance, the near-annual **General Social Surveys** (GSS) of the National Opinion Research Corporation (NORC), which are among the most important sources of social scientific data in existence. The 2000 survey (they have been conducted since 1972), for example, contains the responses of a random sample of 2,817 Americans 18 years of age and older on over 800 variables. We also have the **National Election Studies**, surveying Americans during voting years from 1952 through 2000. For those interested in religion, our archives include the 1991 and 1998 **International Social Survey Religion Surveys**, the latter with responses of over 39,000 individuals from seventeen nations of the world, and **CUNY's 1989-90 National Survey of Religious Identification** (n=113,723).

In addition to survey data, we have ecological data sets with information on census tract, county, state, and national levels. These data sets come with maps, allowing one to see how a variable (i.e., divorce or suicide rates, percent of population that is church-going, or 65 years of age or older) varies across political units.

GETTING STARTED: SURVEY DATA USING TABULAR ANALYSES

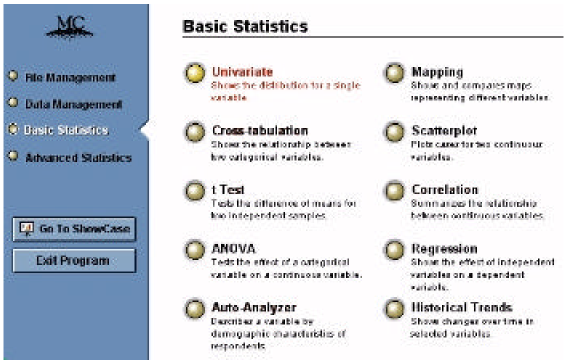


MicroCase only runs on a Windows platform. As demonstrated in class, you can drag the MicroCase icon on to your desktop from **n:\mkearl\microcase2001**. Click the icon and you will see the "File Management" screen. Click the "Open File" button to see a listing of all directories of available data sets for analysis. Double click on the directory and then the file that you want (or highlight the file and press the "Open" button). For this exercise open the **NORC GENERAL SOCIAL SURVEYS** directory and then file **GSS7296r.MC4**. You will next see a "File Setting" box with information about the file, including the number of cases, number of variables, and variable values that are to be automatically converted to "missing data" (e.g., DK, Don't Know, NA or No Answer). Click on the "OK" button on the upper right.

Next click the "Basic Statistics" button. This will take you to the menu page where one enters the statistical commands.

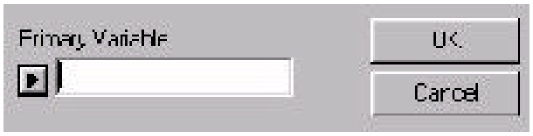
The "BASIC STATISTICS" Menu

Most of the time you will be working with the options on the page on the right. By pressing the **F3** key you will see a listing of all of the variables in this data set. With the down arrow key (or by pulling down the bar) move the highlight so that variable 2501, or SCI.TEST4, is highlighted. The exact wording of the question and the response options can be seen on the right. Press the "OK" button to make the variable window disappear.

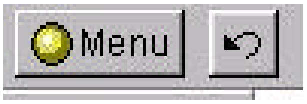

MC
File Management
Data Management
Basic Statistics
Advanced Statistics
Go To ShowCase
Exit Program
Basic Statistics
Univariate
Shows the distribution for a single variable
Mapping
Shows and compares maps representing different variables
Cross-tabulation
Shows the relationship between two categorical variables.
Scatterplot
Plots cases for two continuous variables.
t Test
Tests the difference of means for two independent samples
Correlation
Summarizes the relationship between continuous variables
ANOVA
Tests the effect of a categorical variable on a continuous variable.
Regression
Shows the effect of independent variables on a dependent variable.
Auto-Analyzer
Describes a variable by demographic characteristics of respondents
Historical Trends
Shows changes over time in selected variables

UNIVARIATE

Let's begin by seeing what percentage of American adults think that it is definitely or probably true or false that "human beings developed from earlier species of animals."



1. Push the **UNIVARIATE** button and you will again see the list of variables with variable 2501 still highlighted. Either double-click on the variable name or press the "Primary Variable" button and then the "OK" button (one can also simply type in "2501" or "SCI.TEST4"). You will note that 14% of Americans said it's definitely true that humans developed from earlier species of animals and that nearly one-half said it is probably or definitely not true. These percentages associated with the categories of a variable are called the variable's **marginals.**



2. What percent of fundamentalist Protestants over 35 years of age believe that humans evolved from earlier species? Press the circular arrow key next to the "Menu" key, which will return you to the command box just used so that one can change the variables. Note SCI.TEST4 remains as your primary variable. On the variable list highlight variable 3014, RELSPECTRM, and press the "Optional Subset Variables" button. What appears are the categories of this variable. Click the box next to Fundamentalist Protestant (FUND PROT) and push the OK button. Next highlight variable 42, or AGE, and either double-click it or again push the

"Optional Subset Variables" button. Here enter "35" as the lower limit and enter "95" as the upper (note the AGE value of 98=DK, or don't know, and 99=no answer). Press the OK button on the subset window. Push the OK button on the Univariate button. Presto: Only 26% of fundamentalist Protestants think it is definitely or probably true that humans descended from earlier animal species.

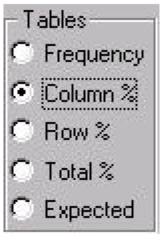
CROSSTABULATION

It's hypothesis time. What factors account for people believing or disbelieving in evolution? We've already seen an illustration of how religious faith does. Does belief increase or decrease with education? with age? And if, for instance, belief increases with education and decreases with age, what are the beliefs of highly educated older persons compared to relatively uneducated younger people?

Let's say our hypothesized relationships look like this: **AGE ---> EDUC ---> EVOLUTION.** The older one is the less education one has; the older one is the less likely one believes in evolution; the less education one has the less likely one believes in evolution. AGE and EDUC are **independent variables** which together determine the **dependent variable** or belief in EVOLUTION. Let's first look at the AGE-SCI.TEST 4 relationship.

1. Push the "Menu" button the return to the Basic Statistics page and then press the **Cross-Tabulation** button.
2. Enter SCI.TEST 4 as the row variable (as you did in Univariate step 1 above)and AGE7 (or 3013, which is the variable number of this **recoded** variable) as the column variable. Press the OK button and you will see the contingency table below.

```
                          SCI.TEST 4      by      AGE7
                          Weight Variable:    OVERSAMP

              18-29  30-39  40-49 50-59  60-69  70-79  80+ Missing TOTAL
DEFINIT.TR     97    102     99     48     29     12      5      1    392

PROB. TRUE    172    231    190    118     59     54     27      0    851

PROB.NOT T     76    126     82     49     37     41     13      0    424

DEFINIT.NO    141    206    161    141    124     89     45      3    907

CAN'T CHOS     34     44     46     30     30     41      6      2    231

Missing      7617   7090   5431   4468   3981   2746   1020    120  32473

TOTAL         520    709    578    386    279    237     96    126   2805
```



Tables

○ Frequency

● Column %

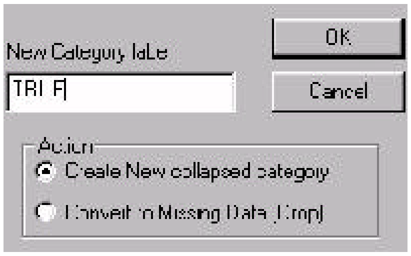
○ Row %

○ Total %

○ Expected

Each column is made up of individuals of different age groups. From the TOTAL row on the bottom observe that there are 520 people 18-29, 709 between the ages of 30 and 39, and so forth. You've have already seen figures near those the TOTAL column on the right: 1243 people believe the theory of evolution is definitely or probably true, 424 believe it is probably and 907 believe it is definitely false.   The number 97 is a **cell frequency**: there are 97 individuals between the ages of 18 and 29 who believe the theory of evolution is definitely true.

Observe the menu selection on the left-hand side of the screen.
Since each category of our independent variable AGE appears as a column, select the "column %" option (so that each column will add up to 100% and we can find out what percent of each age group believes in evolution).



New Category label: TRUE

Action

● Create New collapsed category

○ Convert to Missing Data (Emp)

Let's simplify the table by first combining those who believe that human evolution is definitely or probably true. To do this, click on the "definit.tr" and "prob.true" labels and observe how their two rows become shaded in green. Press the "Collapse" button on the lower left. A "Collapse Categories" box appears. Give your new category the label "TRUE" and make sure that under "Action" the "Create New collapsed category" option is selected. Press the OK button.

Next let's combine the probably not true, definitely not true, and can't choose categories and label it "FALSE/DK." Getting the hang of it? While we're at it let's also collapse the age categories into 18-39, 40-59, 60+.

Select the "Total %" Tables option on the left. You should see the following:

```
                   SCI.TEST 4      by      AGE7
                   Weight Variable:    OVERSAMP

                  18-39  40-59    60+  Missing       TOTAL
     TRUE            602    455    186       1        1243
                  21.5%  16.2%   6.6%                44.3%
     FALSE/DK        627    509    426       5        1562
                  22.4%  18.1%  15.2%                55.7%
    Missing       14706   9899   7748     120        32473

     TOTAL          1229   964    612     126         2805
```

The 16.2% here means that of all American adults, 16% believe the theory of human evolution to be true AND are 40 to 59 years of age.

Again select the "Column %" table option. Observe how the older one is the less likely one sees the theory to be true, from 49% of those 18 to 39 years of age to 30% of those 60 and older. There are several ways in which we can describe this relationship. First we can say that there is a **percentage difference** of 19 percentage points (49% - 30%) between the youngest and oldest groups. Or we can say that those in the youngest group are 63% (49%/30%) more likely than the oldest group to think the theory of evolution is definitely or probably true.

There are two additional statistics that we can consider and we'll let MicroCase do the computing. Press the "Statistics" button to reveal statistics about the table. Below "Nominal Statistics" you will see:

Chi-Square: 62.177 DF: 2 (Prob. = 0.000)

Chi-square is a statistic that tells us whether or not our observed cell frequencies could statistically have occurred by chance (these values, by the way, can be found by selecting the "Expected" Table option). If chance was at work, we would expect about 44% of each age group would think the theory of evolution to be true, matching the COLUMN TOTALS value. **If the chi-square probability figure is less than or equal to .05** (which is the probability that the observed frequencies did occur by chance we can say that the observed relation between AGE and SCI.TEST 4 was not due to chance. Since the probability here is 0.000 (**normally reported as < .001**), we can say the relationship is not due chance, that there is a **statistically significant relationship.**

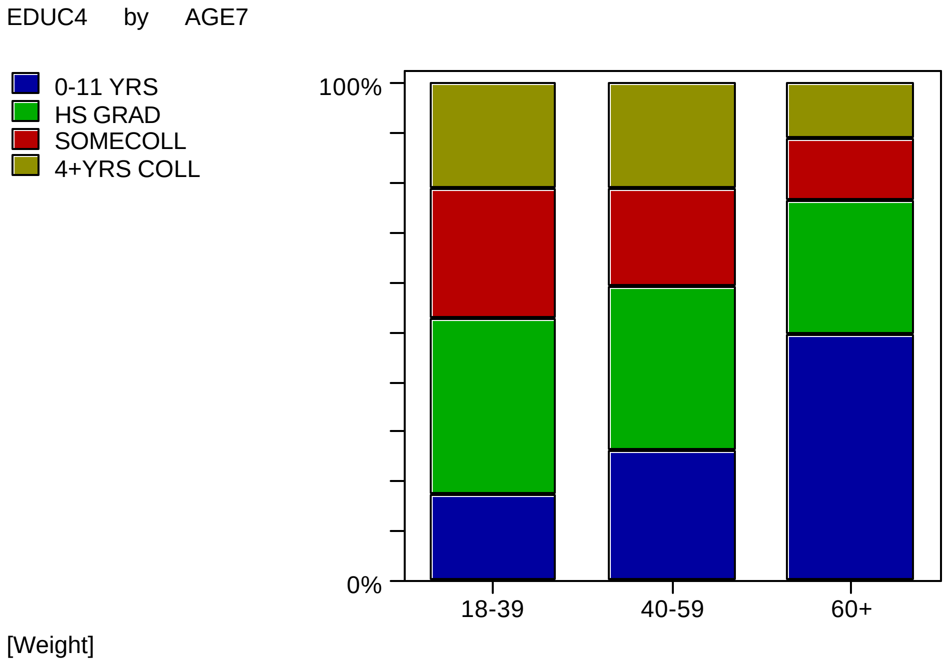
Okay, we have a statistically significant relationship but how big is it? Here we will use the **Gamma** statistic, located under "Ordinal Statistics." Gamma is a statistic that goes from zero if there is no linear relationship to plus or minus 1.0 if the relationship is perfect (a -.30 gamma correlation is the same strength of relation as a +.30). My rule of thumb is if chi-square has a significant probability and gamma is greater than the absolute value of .10, we can say there is a monotonically increasing or decreasing relationship. (If there is a curvilinear relationship--as when age is related to increasing belief in some attitude until the age of 45, from which age on the older one is the less one's belief--the relationship can be statistically significant even though gamma is not.) In the present example gamma=.21 (**round off to the second decimal**): as their age increases, the percent of Americans saying the theory of evolution is false consistently increases as well.

Repeating the steps above, let's consider the relationship between EDUC4 (variable #3012) and SCI.TEST 4, collapsing SCI.TEST 4 as we did before.

```
SCI.TEST 4      by      EDUC4
Weight Variable:    OVERSAMP

          0-11 YRS      HS GRAD       SOMECOLL      4+YRS COLL    Missing        TOTAL
TRUE        175         318           288           461            2           1242
          33.4%         37.0%         41.6%         63.1%                        44.3%
FALSE/DK     349        541           404           270            3           1564
           66.6%        63.0%         58.4%         36.9%                        55.7%
Missing     9239        10559         6706          5871          98          32473

TOTAL        524        859           692           731            103         2806
```

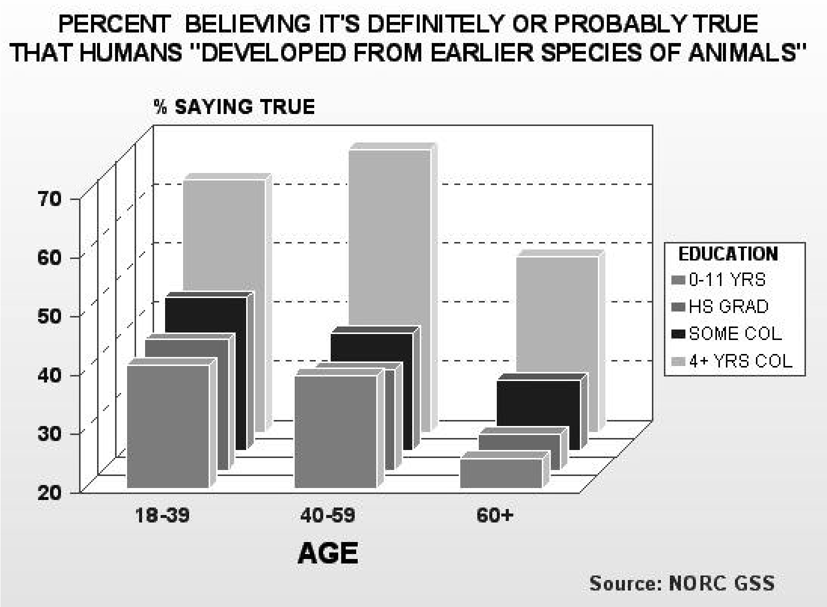
Observe how with increasing education individuals are increasingly likely to believe in evolution. However, this relationship is minimal for the lowest three levels of education. The most educated are 30 percentage points more likely than the least educated to believe the theory is true, or more than 90 percent more likely (63/33) to believe. Looking at the statistics, the chi-square probability indicates significance and gamma = -.31: the more education the less likely individuals say the theory is false.

Again repeating the above steps we can examine the relationship between AGE (again recoded 18-39, 40-59, 60+) and EDUC. This relationship is also significant (e.g., the chi-square probability is less than .0001), with a gamma of -.29: the youngest age group is nearly twice as likely as the oldest to have four or more years of college. Those in the oldest group are nearly three times as likely as those in the youngest to be high school dropouts. Are you more of a visual person? Press the "Bar Stack" graph option and you will see the following:



It is time to consider the effects of our two independent variables (age, education) simultaneously on beliefs in evolution. Repeat steps 1 and 2 on page 3, entering SCI.TEST 4 as the row variable and AGE7 as the dependent. This time, however, enter EDUC4 as the **control variable**. What this gives you are four AGE-SCI.TEST 4 tables, one for each level of education. To move forward through them simply press the arrow buttons on the lower left just above the "Collapse" button.

If you look at the associated statistics for each table (again with AGE and SCI.TEST 4 collapsed as we did before), observe how the **conditional gamma** statistics (i.e., the AGE-SCI.TEST 4 gamma relationship given the condition that EDUC = 0-11 years) consistently declines with education, from .27 for those with the least education to .04 for those with four or more years of college. Want to get visual again?



Source: NORC GSS

PART II: USING MEANS TESTS & REGRESSION WITH ECOLOGICAL DATA SETS

As mentioned earlier, in addition to surveys of individuals Trinity also has numerous quality data sets where city census tracts, counties, states, and countries are the units of analyses. To illustrate what one can do with these ecological data sets let's begin with a file where the units of analysis are not people but rather states. For each state we have such information as its age structure, poverty rates, rates of homicide and suicide, proportion of all births to unwed mothers, voting preferences in all Presidential elections since 1860, magazine subscription rates (which state has the greatest proportion of subscribers to *Soldier of Fortune*, *Playboy*, or *Muscle*?), number of psychiatrists per 100,000 population, proportion of all births that are Caesarians, gallons of beer consumed per capita per year, and number of deaths per

100,000 due to cirrhosis of the liver. Because the variables in these files are **interval** -- that is, instead of having categories like "agree somewhat" they have precise, meaningful numeric values like population totals or percent of residents who have college degrees--allowing for a different set of statistics, such as means and regressions.

Return to the "File Management" page, push the "Open file" button, open **the States & Counties** directory and select the **STATES (combined)** file. You will see in the description of this file that it has fifty cases (one for each state, duh) and over 1100 variables. To illustrate what one can do with this kind of data let's investigate something that might be relevant to you: education.

As the last Presidential election of the century heated up, the issue of education again became a topic of political discourse. What should the nation be investing in its "people resources" to guarantee its competitiveness in the world economic system of twenty-first century? In his acceptance speech to the Republican convention, candidate Dole singled out teachers unions as a target for criticism. Remember the line: too much paid for unacceptable outcomes. Candidate Clinton, on the other hand, talked of even greater financial investments in education.

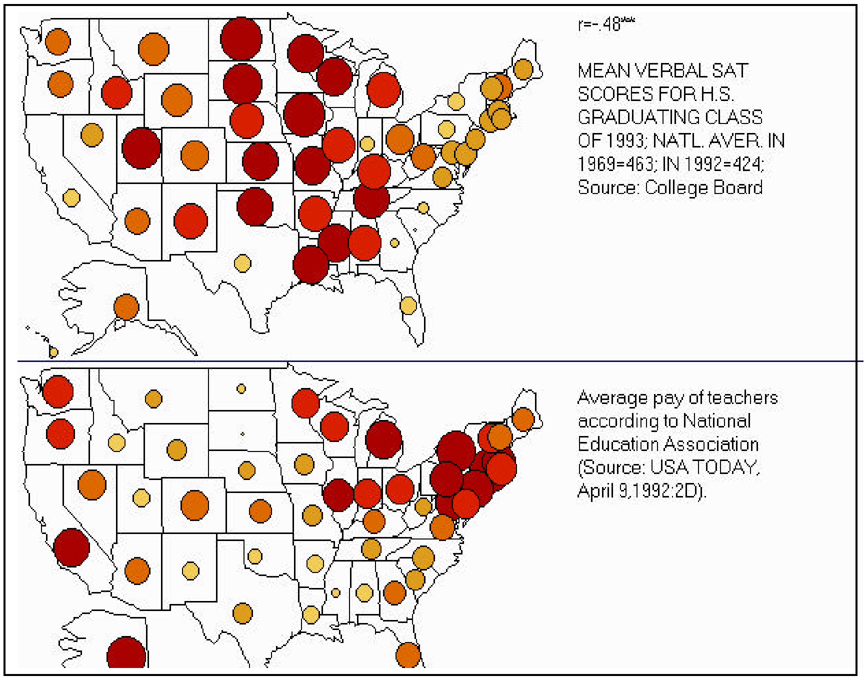
One measure of educational outcome that we may consider is the mean verbal SAT score for each state. To see what they are, select the **Mapping** command. Since this state-level data file has so many variables, let's let MicroCase locate the SAT variables for us. Press the "Search" button, type SAT, and then press the OK button. In the window appears the variables with SAT either in their variable labels or descriptors. Highlight variable 484 (SAT-V93), double-click on it (or push the arrow button under Variable 1) and press the OK button.

What appears is a shaded map of the U.S. To see what values are associated with what color click the "Legend" box under the Display options. Observe that the darker the state's color the higher its mean SAT verbal scores. If you put the cursor on a state and press the left mouse key you will see the state's name, its value on the variable, and its rank (1-50) on it. To view all of the states' values on the variable click the "List: Rank" bubble on the screen's left. And to view the relative values of the variable with proportionately-sized dots on the map, select the "Spot Fill" display option.

So what is to be made of this SAT distribution? Why are the highest scores in the Midwest? Is there something to living adjacent to the Mississippi River or might there be more sociological variables at work? (Does the same proportion of high school seniors--say the top two-thirds academically--take the SATs in all states?) Can we say that there are significant differences in the <u>mean</u> SAT scores of different regions in the U.S.? To find out, select the **ANOVA** command and enter variable

484 (SAT-V93) as the dependent variable and variable 14 (NORCregion) as the independent variable. (Note: *Variable NORCregion categorizes the states into the same regions as does the REGION variable in the NORC General Social Surveys. This allows you to, for instance, look at the mean abortion rates by region with the states file and then compare these results to survey data where you correlate region by people's attitudes toward the morality of abortion. Is public support for abortion greatest where, in fact, abortion rates are the highest?*) You will next see a graph with the overall mean (in green) and the means and standard deviations of verbal SAT scores for each of the regions. Select the "Means" option under the "Statistics" command. Observe that the mean SAT scores are significantly different across the regions of the country.

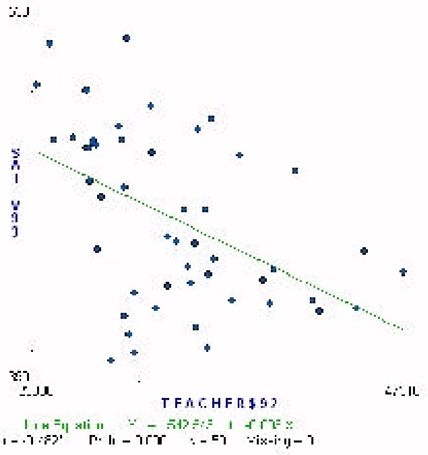
Might there be a relationship between student test scores and what teachers are paid? You will find there is variable 466-TEACHER$92, teachers' average pay. While you still have the map image on the screen click the change variables arrow on the upper left and enter either 466 or TEACHER$92 as "Optional: Variable 2" and repress the OK button. What you should see is illustrated on the right. Interesting. Where salaries are lowest it looks as though SAT scores are highest. The -.48 statistic on the middle right is another measure of association (like gamma), which will be explained shortly. For purposes here note that it is starred (**), which means there is a statistically significant negative relationship: the higher the salaries, the lower the SAT scores.


r=-.48**

MEAN VERBAL SAT SCORES FOR H.S. GRADUATING CLASS OF 1993; NATL. AVER. IN 1969=463; IN 1992=424; Source: College Board

Average pay of teachers according to National Education Association (Source: USA TODAY, April 9,1992:2D).

Press the Menu button to return you to the Basic Statistics page. To examine this teacher salary-SAT score relationship further, select the **SCATTERPLOT** command. Enter SAT-V93 as the dependent variable. This is the dependent variable because we may believe teachers' salaries in some way <u>cause</u> board scores and not the other way around. Enter



TEACHER$92 or 466 as the independent variable and press the OK button .

What appears looks like bacterial colonies in a Petri dish. This is a graph where the vertical axis is the dependent variable and the horizontal axis is the independent variable. Each point represents a state and its scores on the two variables. When you select the "Reg.Line" option you will see the **regression line**. The regression line, as you will learn in your calculus class, is that one line that minimizes the sum of squared distances from each observed point . If the line is perfectly horizontal it would mean that there is no relationship between the variables; the best predictor of the dependent variable is its overall mean (i.e., if we're trying to predict adults' IQs from their shoe sizes, we would expect the mean IQ of all Americans to be the same mean IQ for each foot size). If there is a perfect relationship, all observed points would fall exactly on the regression line.

To measure the adequacy of fit, the amount of scatter of the observed points around the line, the correlation coefficient (called Pearson's r) is used. If r = 0 there is random scatter and hence no relationship (and the regression line would be parallel to the X-axis). If r = 1.0 (a positively sloped line) or = -1.0 (a negatively sloped line like the one above) the fit is perfect as all points lie exactly on the line. This is a **perfect correlation** (r-squared is, in fact, the percent of variance of the dependent variable that is attributable to changes in the independent variable). To the right of the "r" statistic on the bottom of the screen is a PROB or probability statistic. This is the probability that the correlation could, in fact, actually be zero. Hence, the smaller this number the better. Any probability <.05 is said to be **statistically significant**.

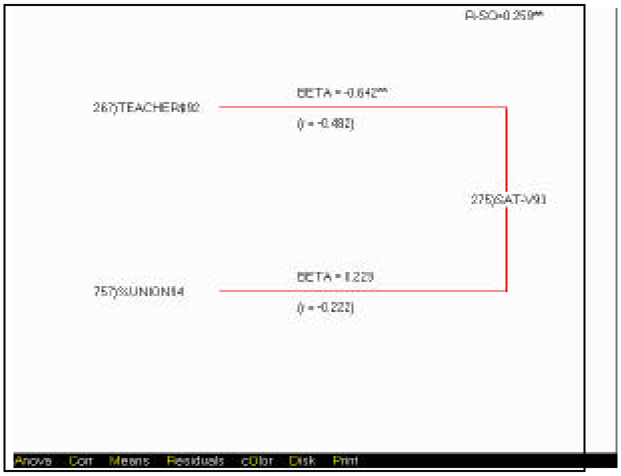
Above these statistics is a green "Line Equation." Ours, for instance, is Y = 542.54 -.003X. Y is the dependent variable, the mean verbal SAT score, and X is the mean teacher salary. According to this equation, if you don't pay your teachers you can expect a 543 mean SAT score. And pay your teachers $180,000 a year and you can bring your SAT scores down to zero! Candidate Bush, it seems, needed to consider other factors affecting SAT scores than just teacher salaries.

Clicking on any of the points in the graph will give you its associated state and its values on the two variables. To locate a particular state click on the "Case" option in the "Find" commands on the left. For instance, try locating Texas. After pressing the OK button note how one of the dots becomes framed in a red box That's the Lone Star state! How accurate is our model in predicting the verbal SAT scores of Texas students given the state's mean teachers' salary? Observe Texas SAT scores are lower than what we would predict with this model.

The world, unfortunately, is not made up of **bivariate** (two variable) relationships. Relationships change when other variables are taken into account. Certainly cutting teachers' salaries is not going to improve SAT scores. What other factors may be at work?

Although this file does not currently have information on the percentage of state teachers who are unionized, we do have the percentage of workers who are unionized (variable 1078, %UNION94). Let's assume that the more unionized the state is the more likely its teachers are to be unionized. If you correlate (using either the SCATTERPLOT or CORRELATION commands) %UNION94 with TEACHER$92 you will find the r- correlation is a significant .70--the more unionized the state the higher teachers salaries. You might also note that there is a non-significant r-correlation of -.22 between %UNION94 and SAT-V93--SAT scores go down somewhat as unionization increases. To see what happens when SAT-V93 is predicted with these two independent variables, select the **Regression** command. Enter SAT-V93 as the dependent variable, TEACHER$92 as the first independent variable and %UNION94 as the second.

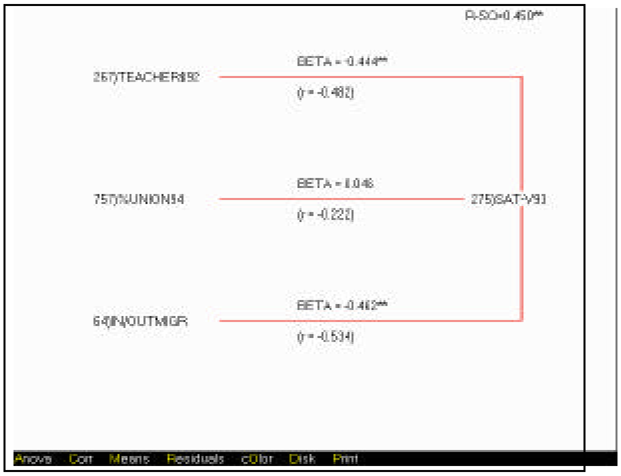
You will see a diagram of the independent variables as they affect the dependent variable. There are two statistics associated with each independent variable's effect: the r statistic is that variable's singular relationship with the dependent variable (e.g., what you would get using the SCATTERPLOT option, with only that independent variable run against the dependent one); and the **beta** statistic, which is the **partial correlation**. (*For the more statistically-inclined, the beta value is equal to how many standard deviations change in the dependent variable results from a one-standard deviation change in the independent variable*.) A partial correlation is the singular influence of that independent variable taking into account the presence of the other independent variable(s) in your equation. One nice quality of beta is that it allows us to compare apples with oranges--that is, if one independent variable has a beta of, say, .4 and the other has a beta of .2, one can say that the former is twice as predictive as the latter. In addition, on the upper right of the screen you will see an **R-SQ** statistic. Note this is a capital R (indicating more than one independent variable) and it is squared. This is the **explained variance** of the dependent variable taking into account all of the independent variables simultaneously.



Two rules of thumb when using the REGRESSION analysis:
1. disregard all betas of .15 or less
2. never use more than three independent variables at a time.

In the present example, observe that about one-quarter of the variance of SAT scores are explained by teacher salaries and unionization. Observe that %UNION94 not only has no significant effect (its statistics are not starred) but that when including teachers' salaries in the equation, its effect on SAT scores flip from being negative to positive: taking into account teachers' salaries, the greater the states' unionization the higher the SAT scores (albeit not significantly).

So what other variables might explain state SAT scores? There is, in fact, the possibility that the TEACHER$92-SATV93 is **spurious**--that is, when picking the "right" additional independent variable, the teacher salary-SAT beta goes to zero. A classic example of a spurious relationship is the relationship between the number of fire trucks on the scene and the amount of damage done. Hey, the two are related! However, when taking into account the size of fire (which determines both vehicle number and damage), the truck-damage relation disappears.



Your suspects in this SAT caper? Candidates might include such factors as immigration rates (foreign immigrants don't due as well as natives on SAT verbal tests, and immigrants are more likely to enter the country in the coastal states and states bordering Mexico), poverty rates, urbanization, etc. In the chart to the right we see the effect of variable 65 IN/OUTMIG, the ratio of those entering versus leaving the state. Note that we improved the R-SQ from .25 to .45, indicating that this is a better model. Also observe that this migration measure is a tad more potent than teachers' salaries in predicting SAT verbal scores.

Experiment on your own to see if you can build an even more predictive model. Wondering about what variables to use? This is why social scientists employ **theories**: to guide them in their selection of predictive variables.

Addendums

1. Take advantage of the "Help" command--there's a mini-statistics course built into the index.
2. Open the **Historic Trends** directory and then the TRENDS file. In the "Basic Statics" page, experiment with the "Historical Trends" command. Note how several trends can be plotted on the same graph.
3. On MicroCase's "Data Management" page are several useful features. Want your own recoding of variable AGE in the NORC GSS? Try out the "Collapse Variables" command. Need to save a file of only those variables that you are working with and not the hundreds of others (so that your data set fits on one floppy to take home to the dorm to work with)? Experiment with the "Subset File" command and select your dependent variable as the subset variable.