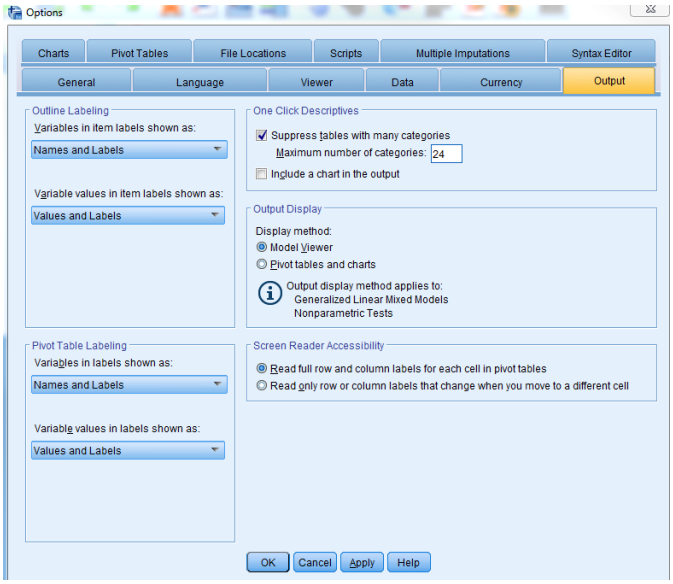
# CORRELATION AND OLS REGRESSION

First thing: I like work with variable names and see both names and label in the output so the first thing I do is go under the Options command and change the **General Tab** to display variable names and under the **Output Tab** to display both labels and names

EDIT /OPTIONS /OUTPUT




**ASIDE:** OLS handles dichotomous independent variables very well but they MUST be coded "0" and "1." If they are coded any other values, like "1" and "2," it will be impossible to interpret the coefficients. Other categorical variables like race/ethnicity are somewhat more complicated to use in a regression format. For example, suppose we have a 5-category race variable as follows:

1=White, non-Hispanic
2=Black, non-Hispanic
3=Hispanic of any race
4=Asian
5=Other

If we wanted to determine the extent to which there were significant differences between each of these race categories, we would need to create new dichotomous variables for each of the values. For example, one variable would be coded "1" for white, non-Hispanic and "0" for all others; another variable would be coded "1" for Black, non-Hispanic and "0" for all others, and so on. When entered into a multiple regression model, one of these variables would be left out and that would be the comparison for which all other dichotomous variables were made. This is a bit complicated for this class – it is something you will learn in 614. For this class, we typically stick with dichotomous independent variables!

Back to our regression exercise!

**PROBLEM 1**

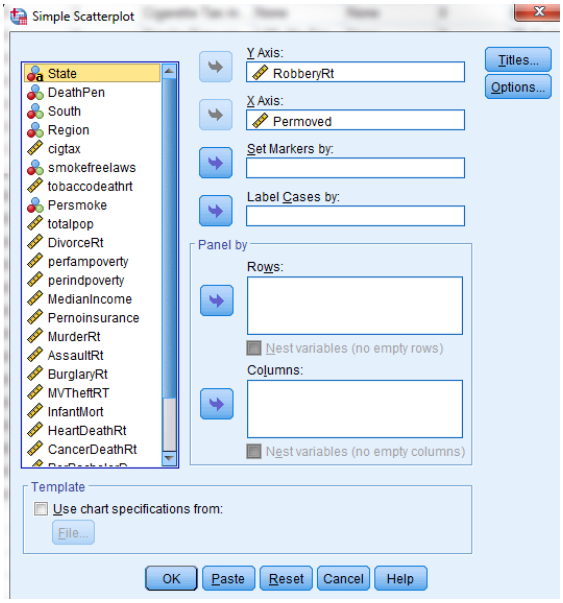
Motor vehicle theft rates vary from state-to-state in the United States. What factors are related to murder rates at the aggregate state level? Let's use STATE2012 to find out.

Stated very simply, social disorganization theory predicts that communities that are in flux are less able to control resident's behavior because there will be low collective efficacy. One indicator of social disorganization is the migration in and out of communities. One question the U.S. Census asks residents is whether they have moved in the past 5 years and this is frequently used as an indicator of social disorganization.

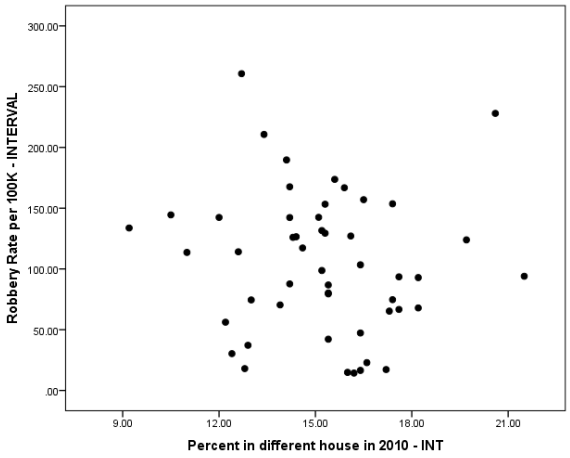
For this exercise, the independent variable will be **Permoved,** which measures the percent of a state's population who moved in the past 5 years. The dependent variable will be **RobberyRT,** which is the state motor vehicle theft rate.

Let's first look at the scatterplot of Permoved and RobberyRT.

GRAPHS /LEGACY DIALOGUES /SCATTER/DOT /SIMPLE SCATTER

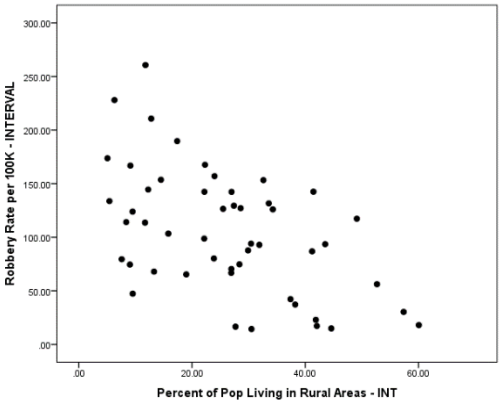



Since robbery rate is the dependent variable it is placed on the Y axis and Permoved is placed on the X axis:




From this, it is difficult to discern the direction in the relationship. If I had to draw a line through the bivariate scatter of dots, it would be hard to know whether the line would be ascending indicating positive relationship, descending indicating a negative relationship, or essentially flat, indicating very little relationship.

Does rural population affect robbery rates? Let's take a look at a scatterplot.




The direction of the relationship from this scatterplot appears to be negative. That is, on average, states with higher percentages of its population residing in rural areas tend to have lower rates of robbery.

Let's examine the correlation coefficients between these variables.

ANALYZE /CORRELATE /BIVARIATE

Enter all of the above variables into the dialogue "Variables" box.

You obtain the following results:

**Correlations**

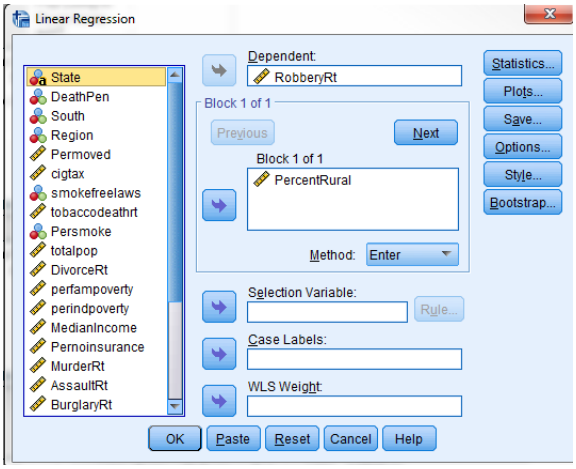
| | | Permoved Percent in different house in 2010 - INT | PercentRural Percent of Pop Living in Rural Areas - INT | RobberyRt Robbery Rate per 100K - INTERVAL |
|---|---|---|---|---|
| Permoved Percent in different house in 2010 - INT | Pearson Correlation | 1 | -.017 | -.091 |
| | Sig. (2-tailed) | | .907 | .529 |
| | N | 50 | 50 | 50 |
| PercentRural Percent of Pop Living in Rural Areas - INT | Pearson Correlation | -.017 | 1 | -.533** |
| | Sig. (2-tailed) | .907 | | .000 |
| | N | 50 | 50 | 50 |
| RobberyRt Robbery Rate per 100K - INTERVAL | Pearson Correlation | -.091 | -.533** | 1 |
| | Sig. (2-tailed) | .529 | .000 | |
| | N | 50 | 50 | 50 |

**. Correlation is significant at the 0.01 level (2-tailed).

From this, we can see that the correlation between mobility (permoved) and robbery rates in states is negative but almost zero ($r$ = -.091), and using a calculator, if we square this calculate the value of r2, we will see that mobility explains less than 1% of the variation in robbery rates. In fact, the obtained significance is .52 indicating that we cannot reject the null hypothesis at the alpha .05 level; we must conclude that social disorganization, at least not using the indicator of mobility, does not affect rates of robbery in states. However, we see that the relationship between rurality of a state and robbery is negatively and moderately related ($r$ = -.533) and if we square the r value, we will see that rurality explains 28.4% of the variation in robbery rates, which is a significant amount according to the alpha of .0001. **I have to note here that SPSS only displays 3 decimal places, so there is a 1 out there somewhere. We are not risking 0% error in rejecting this null hypothesis. In fact, if you double click on the significance value you will see that the actual significance is equal to .000069.**

Let's find out the exact impact of rurality on rates of robbery using OLS regression:

ANALYZE /REGRESSION /LINEAR




From this command, you will get the following output:

## Regression

**Variables Entered/Removed[a]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | PercentRural Percent of Pop Living in Rural Areas - INT[b] | . | Enter |

a. Dependent Variable: RobberyRt Robbery Rate per 100K - INTERVAL
b. All requested variables entered.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .533[a] | .284 | .269 | 49.37229 |

a. Predictors: (Constant), PercentRural Percent of Pop Living in Rural Areas - INT

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 46329.242 | 1 | 46329.242 | 19.006 | .000[b] |
| | Residual | 117005.926 | 48 | 2437.623 | | |
| | Total | 163335.167 | 49 | | | |

a. Dependent Variable: RobberyRt Robbery Rate per 100K - INTERVAL
b. Predictors: (Constant), PercentRural Percent of Pop Living in Rural Areas - INT

**Coefficients[a]**

| Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constant) | 160.866 | 14.677 | | 10.960 | .000 |
| | PercentRural Percent of Pop Living in Rural Areas - INT | -2.159 | .495 | -.533 | -4.360 | .000 |

a. Dependent Variable: RobberyRt Robbery Rate per 100K - INTERVAL

You can see that the correlation coefficient is the same, although it does not reflect the negative sign, which I noted earlier. The OLS regression equation from the output is:

**Robbery Rates ($y$) = 160.866 + -2.159 ($x_{\%rural}$)**

The interpretation of the slope indicates that as percent of the population residing in rural locations in states increases by 1 unit, robbery rates decrease by 2.159 units. The significance of the t test associated with this coefficient indicates that this is a significant relationship so we can conclude that on average, states with higher percentages of rural population tend to have lower rates of robbery.

To illustrate the effect of rurality on robbery rates, let's use this equation to predict robbery rates at high and low values of rurality.

The predicted robbery rate for a state with a very low percentage of the population living in rural areas, let's say about 6%, would be:

**Robbery Rates ( $\hat{y}$ ) = 160.866 + -2.159 (6)**

**Robbery Rates ( $\hat{y}$ ) = 160.866 + -12.95 = 147.91**

The predicted robbery rate for a state with a high percentage of the population living in rural areas, let's say about 50% (We know the highest is 60%), would be:

**Robbery Rates ( $\hat{y}$ ) = 160.866 + -2.159 (50)**

**Robbery Rates ( $\hat{y}$ ) = 160.866 + -107.95 = 52.91**

There is one thing I want to note here about the unstandardized slope coefficients. You CANT compare them to other unstandardized coefficients because their magnitude is based on the measurement units of both x and y. As such, you may have a very large unstandardized slope coefficient that is not significant and a very small one that is very significant. They are like standard deviations in that way. I will talk more about this when we talk about multiple regression in December.

**Now let's examine the homicide defendant data set, where the units of analysis are a sample of homicide defendants from the 75 largest SMSAs in the country.**

The dependent variable we are going to examine here is the incarceration sentence length in days (**PrisonTime)** for those convicted defendants.

The first research we will examine is whether the defendant's age effects their incarceration sentence. Our dependent variable is called **DefAge** and it is an interval/ratio variable measuring the defendant's age at the time of the offense.

ANALYZE /REGRESSION /LINEAR

The SPSS output we obtained for this is:

## Regression

**Variables Entered/Removed[a]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | Age in yr a/o offn[b] | . | Enter |

a. Dependent Variable: Incarceration term in days
b. All requested variables entered.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .012[a] | .000 | -.001 | 14628.885 |

a. Predictors: (Constant), Age in yr a/o offn

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 29687377.200 | 1 | 29687377.200 | .139 | .710[b] |
| | Residual | 211222226800.000 | 987 | 214004282.500 | | |
| | Total | 211251914200.000 | 988 | | | |

a. Dependent Variable: Incarceration term in days
b. Predictors: (Constant), Age in yr a/o offn

**Coefficients[a]**

| Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constant) | 16554.689 | 1376.774 | | 12.024 | .000 |
| | Age in yr a/o offn | -16.684 | 44.794 | -.012 | -.372 | .710 |

a. Dependent Variable: Incarceration term in days

The regression equation is:

**Sentence Length (*y*) = 16554.69 + -16.68 ($x_{defage}$)**

As you can see from the model summary statistics, r indicates that very little of the variation in sentence length is explained by the defendant's age. The slope indicates that for every one unit increase in defendant's age, sentence length decreases by 16.68 units (days), however, this is not significant. We cannot conclude that defendant's age effects the severity of the sentence they receive.

One legal factor that should affect sentence length is victim provocation or precipitation. We have a variable that measures whether victim provocation (**Victimprovoke**) was claimed in the case, coded 1 for those defendants who claimed victim provocation and 0 for those who did not. Using this variable as the independent variable to predict sentence length, we get the following results:

```
ANALYZE   /REGRESSION        /LINEAR
```

## Regression

**Variables Entered/Removed[a]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | victim provocation[b] | . | Enter |

a. Dependent Variable: Incarceration term in days
b. All requested variables entered.

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .233[a] | .054 | .053 | 14079.670 |

a. Predictors: (Constant), victim provocation

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 10063915490.000 | 1 | 10063915490.000 | 50.767 | .000[b] |
| | Residual | 176034562200.000 | 888 | 198237119.600 | | |
| | Total | 186098477700.000 | 889 | | | |

a. Dependent Variable: Incarceration term in days
b. Predictors: (Constant), victim provocation

**Coefficients[a]**

| Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constant) | 17125.995 | 505.105 | | 33.906 | .000 |
| | victim provocation | -10100.172 | 1417.548 | -.233 | -7.125 | .000 |

a. Dependent Variable: Incarceration term in days

The OLS regression equation is

**Sentence Length ($y$) = 17129.995 + -10100.17 ($x_{vicprov}$)**

The model summary statistics that this is a relatively weak relationship, explaining just over 5% of the variation in sentence length, but again, for individual level data, this is fairly good.

The slope tells us that compared to defendants who claimed no victim provocation, sentence length decreased by 10,100.17 days, or about 27.6 years, for defendants who claimed some type of victim precipitation. We can see this if we use the regression equation for prediction:

The predicted sentence for a defendant who did not claim victim provocation:

**Sentence Length ( $\hat{y}$ ) = 17129.995 + -10100.17 (0) = 17129.995 days or about 46.9 years**

The predicted sentence for a defendant who did claim victim provocation:

**Sentence Length ( $\hat{y}$ ) = 17129.995 + -10100.17 (1) = 7029.82 days or about 19.2 years**

**PROBLEM 2 – What affects alcohol/pot use for youth residing in rural America?**

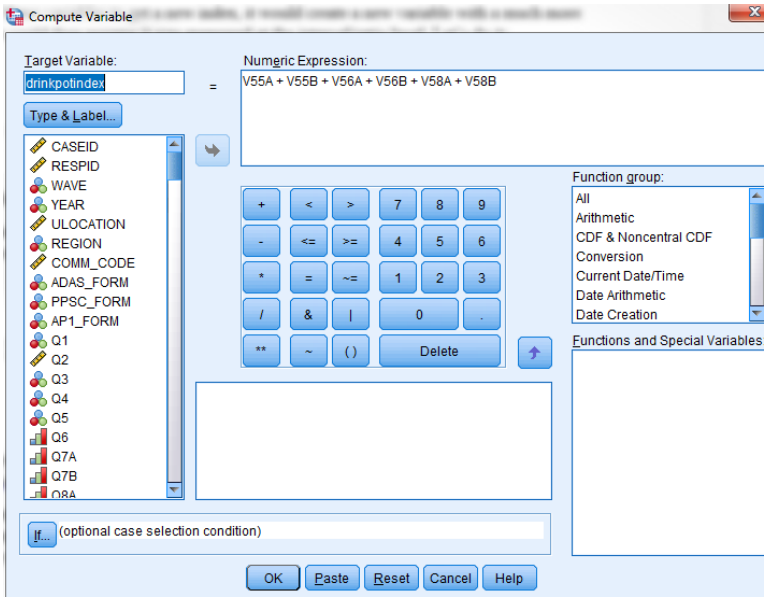
Data: **Rural.alcoholandruguse.SMALL**

The dependent variable we want to create is a how frequently students plan/want to use alcohol and pot in the next month. There are several variables that measure this is three ways: their estimate of how often they 1) plan to use, 2) want to use, 3) the average student will use. We are going to create a simple additive index for all of these variables. They are V55A, V55B, V56A, V56B, V58A, V58B. Let's first get a frequency of each of these variables.

ANALYZE /DESCRIPTIVE STATISTICS /FREQUENCIES

If we add each of these variables to get a new index, it would create a new variable with a much more variability and we could then assume it was measured at the interval/ratio level. Let's do it: Call your target variable **"drinkpotindex"**

TRANSFORM /COMPUTE VARIABLE



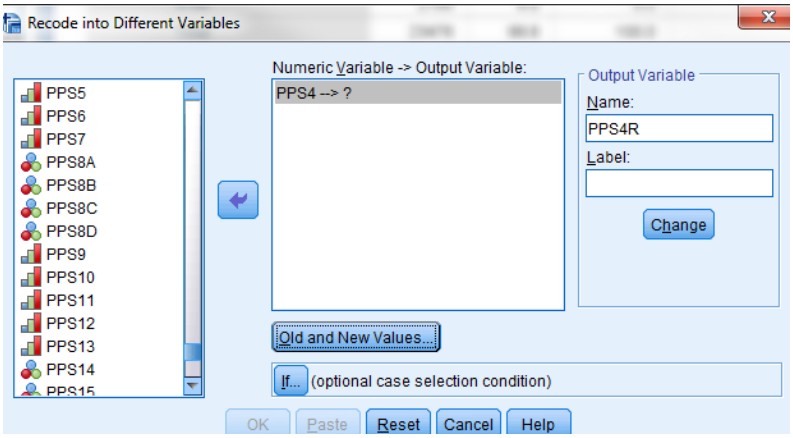

Run a frequency distribution – the new variable should range from 6 to 30 and the original missing values should have remained missing.

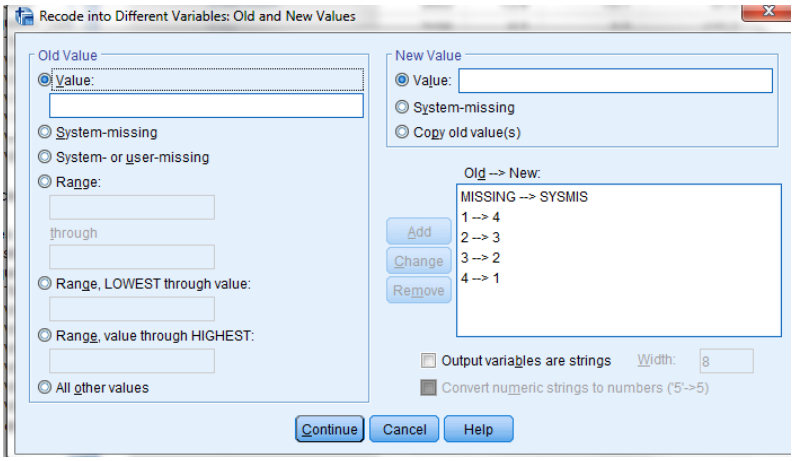
We now need to focus on our independent variables. One of the things we know affects drinking and drug use is attachment to school/teachers. There are several variables that measure this including whether respondents like school, perceive their teachers like them, they like their teachers, and think school is fun. Let's look at each of these variable distributions: PPS4 PPS5 PPS6 PPS7.

ANALYZE /DESCRIPTIVE STATISTICS /FREQUENCIES

Notice that for these variables, high scores indicate less attachment and low scores indicate more attachment. If we simply added these up, the interpretation of the variable would be confusing. To remedy this, we can reverse code them so that high scores indicate high attachments. To this, each of these variable must be recoded into a different variable so that 1=4, 2=3, 3=2, and 4=1. System missing should remain as system missing. As we did before, call them the same variable name with the addition of R to indicate it is a recoded variable.

TRANSFORM /RECODE INTO DIFFERENT VARIABLE

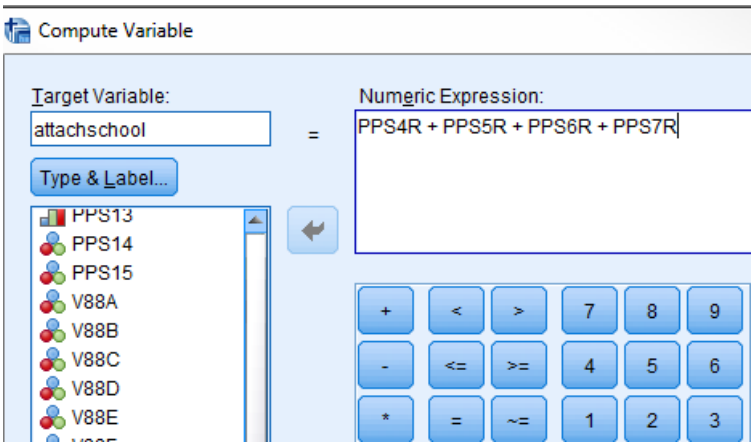



Remember, once you set the old>new values dictionary one, you simply have to change the variable names box, and hit “change” on each new variable to create the new variables. Once you have them all recoded, run frequency distributions for each to make sure they have been recoded correctly.

With these new variables, we can now create an additive index of school attachment where high scores indicate more attachment. Let’s call our new variable **“attachschool”**

TRANSFORM /COMPUTE VARIABLE




Let's now run an OLS regression using the drinkpotindex and the DV and the attachschool as the IV:

ANALYZE /REGRESSION /LINEAR

**Variables Entered/Removed[a]**

| Model | Variables Entered | Variables Removed | Method |
|---|---|---|---|
| 1 | attachschool[b] | . | Enter |

**Model Summary**

| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
|---|---|---|---|---|
| 1 | .178[a] | .032 | .032 | 5.59840 |

**ANOVA[a]**

| Model | | Sum of Squares | df | Mean Square | F | Sig. |
|---|---|---|---|---|---|---|
| 1 | Regression | 5013.798 | 1 | 5013.798 | 159.970 | .000[b] |
| | Residual | 153137.538 | 4886 | 31.342 | | |
| | Total | 158151.335 | 4887 | | | |

**Coefficients[a]**

| Model | | Unstandardized Coefficients B | Unstandardized Coefficients Std. Error | Standardized Coefficients Beta | t | Sig. |
|---|---|---|---|---|---|---|
| 1 | (Constant) | 16.376 | .300 | | 54.581 | .000 |
| | attachschool | -.328 | .026 | -.178 | -12.648 | .000 |

Dv=drinkpotindex

The resulting regression equation is as follows:

Y=a+bx

drinkpotindex = 16.376 + -.328 (attachschool)

Interpret the slope for school attachments and generalize the null hypothesis test to the population.