

Dan LaDue 2/27/11 8:14 PM

Comment: Score 62/65

Part I:

In this part, we will be using social economic status (SES) to predict student achievement in reading and mathematics (fltxcomp). (Using NELS:88 dataset; see appendix for more detail)

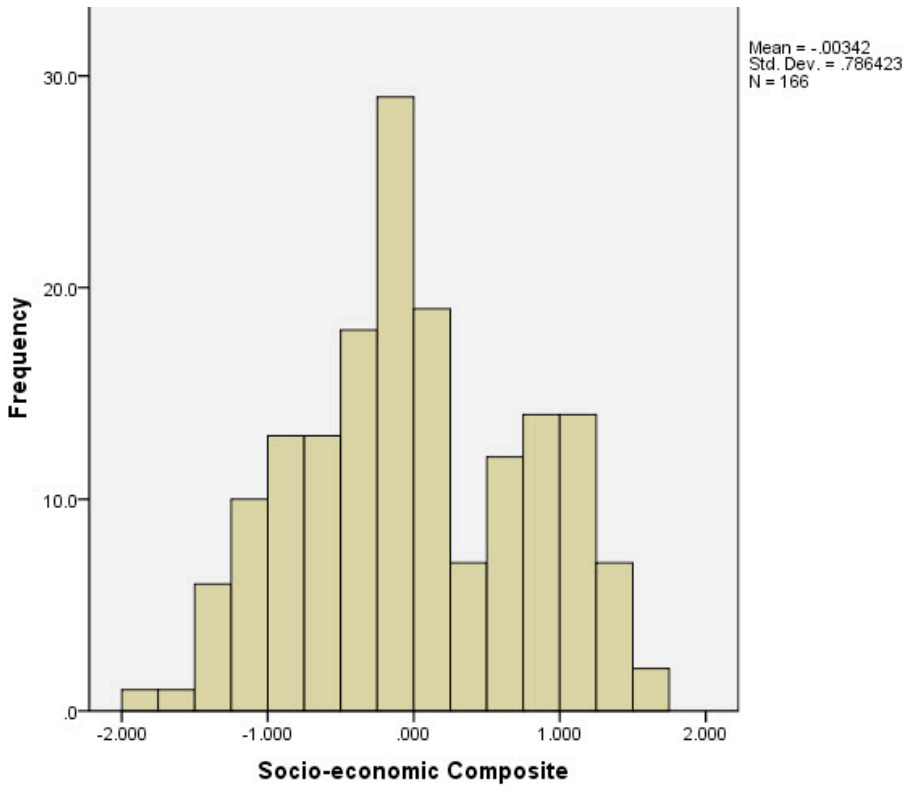
A. Descriptive, Graphical

1. Examine the distribution of each variable separately by considering the descriptive statistics (mean, standard deviation, minimum, and maximum) and a histogram. Describe the shape of each distribution. Do the values of the variable seem sensible (for example, it would not be sensible to have a negative value for someone's height)?

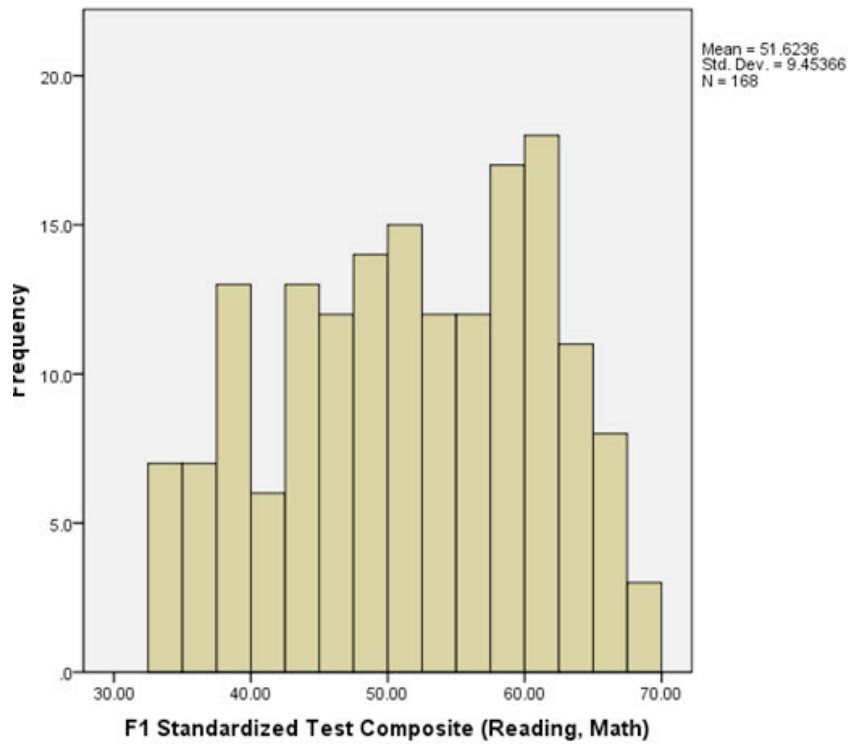
Dan LaDue 2/26/11 6:58 PM

Comment: 3/3

SES: mean = -.00342, standard deviation = 0.786, minimum = -1.862, maximum = 1.638



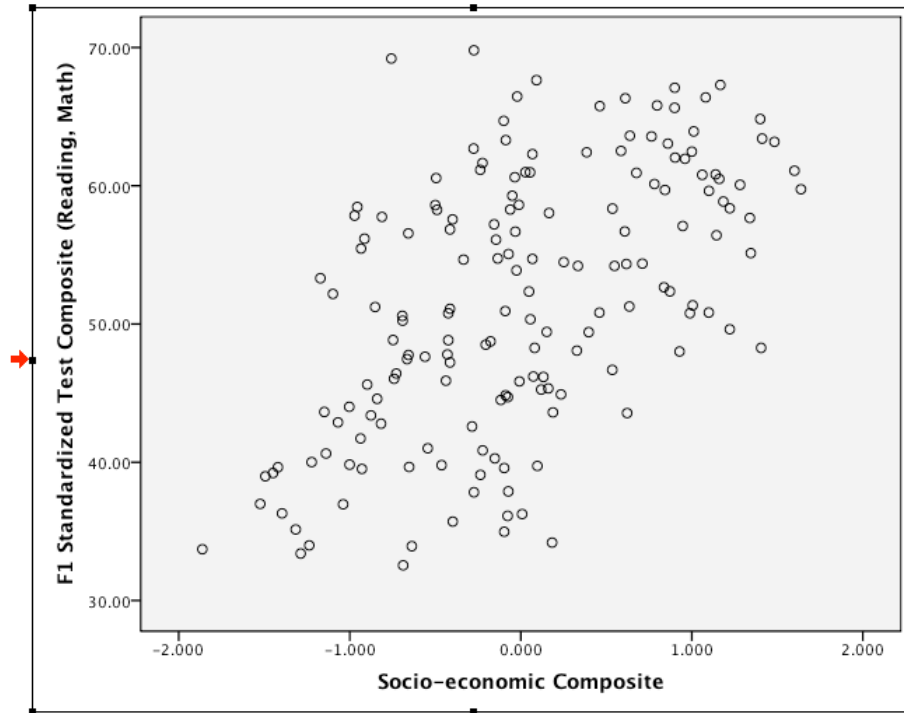
Student achievement in reading and math (fltxcomp): mean = 51.6236, standard deviation = 9.45366, minimum = 32.54, maximum = 69.8



The values of the variables seem sensible. In general, the histograms seem normally distributed. SES had some slight negative skewness, while student achievement was slightly positively skewed. However, the variable seems to reflect what one would expect based on a common understanding of that variable.

2. Generate a scatter plot representing the relationship between social economic status (SES) and student achievement (fltxcomp). What kind of a relationship do you see? (consider the following aspects of a relationship: linear vs. non-linear, positive vs. negative, strength of relationship). Be sure the graph is included in your answer.

Dan LaDue 2/26/11 6:59 PM
Comment: 2/2



Looks like a fairly strong, positive linear relationship. As SES increases, so do the Composite Reading and Math scores.

3. Do you think it is sensible to use regression to estimate the nature of the relationship between the two variables? Explain your answer.

Dan LaDue 2/26/11 7:00 PM
 Comment: 2/2

Regression models assume a linear relation between the continuous outcome and each prediction. In this case it looks like there's a fairly strong, positive linear relationship between SES and reading & math scores on standardized tests. Since regression models predict an outcome based on a variable, the relationships between variables is important. At the very least, these two variables are correlated and thus are good candidates to submit to a regression model. Based on this data, a reasonable prediction can be made about a student's test scores based on the independent variable (SES), but there would have to be some error factored in.

B. Population Model and Statistical Estimation

1. Write down the population model for the regression. Define each term in the model. Identify the population parameters to be estimated, and tell us in words what they

Dan LaDue 2/26/11 7:07 PM
 Comment: 2/2

represent about the relationship between social economic status (ses) and student achievement (fltxcomp).

In this case, we are estimating the population parameters based on the sample. In order to indicate that we are estimating from a sample, we use either $\hat{\beta}$ or b when referring to an individual and sample.

Outcome = continuous predictors + error

$$Y_i = b_0 + b_1X_{1i} + \varepsilon_i$$

Outcome = intercept + slope for predictor + residual error for individual

- This equation indicates how the slope changes for each increase in Y
- Error is difference between predicted and actual values; the part of the Y's that can't be literally explained by the X's

Interpreted as:

$Y_i = fltxcomp$ (student achievement for i th participant

b_0 = intercept (where the model exists in space)

b_1 = slopes for predictors (SES) X_{1i}

X_1 = SES variable

ε_i = residual for the i th participant

2. Use SPSS to obtain estimates of the parameters, and include a table that includes the values of the coefficients, standard errors, t-statistic, and p-value.

Dan LaDue 2/26/11 7:08 PM

Comment: 1.5/2

You need to identify the values of β_0 and β_1 as well as including the tables. i.e. β_0 : 51.708 β_1 : 6.725

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.559 ^a	.313	.308	7.86713

a. Predictors: (Constant), Socio-economic Composite

ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4614.717	1	4614.717	74.561	.000 ^a
	Residual	10150.245	164	61.892		
	Total	14764.962	165			

a. Predictors: (Constant), Socio-economic Composite

b. Dependent Variable: F1 Standardized Test Composite (Reading, Math)

Coefficients^a

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	51.708	.611		84.682	.000
	Socio-economic Composite	6.725	.779	.559	8.635	.000

a. Dependent Variable: F1 Standardized Test Composite (Reading, Math)

3. Provide an interpretation for each parameter estimate. That is, tell us what each parameter estimate means in terms of the relationship between social economic status (SES) and student achievement (f1txcomp), and use the actual values of the estimates in your interpretation.

Dan LaDue 2/26/11 7:20 PM

Comment: 1/2 You don't seem to identify β_0 and β_1
This question is asking only for the parameters from the previous question (β_0 and β_1)

Each parameter estimate tells us about the strength of the predictive power of the Independent Variable, in this case SES, to predict the outcome of student achievement.

One parameter is R Square and Adjusted R square, which help to determine the amount of variance within the model that is predicted by the independent variable, SES. In this table, R Square= 0.313, which means that 31% of the outcome variable (standardized test scores) is explained by SES. Therefore 69% of the variance within the model is not explained by the model at this time. Since adjusted R Square (0.308) is a measure of shrinkage when we account for if the model had been derived from population which the sample was taken. Because the values for R Square and Adjusted R Square are a similar, then we can assume that the predictive model of the sample is generalizable to the population.

Dan LaDue 2/26/11 7:13 PM

Comment: Not needed for this question

The b-values tell us about the relationship between student achievement and the predictor of SES. If there is a positive number, there is a positive relationship. In this case, the b value for SES is 6.725, indicating that there is a positive relationship between SES and student achievement, a statistic that was visually represented in the earlier scatterplot.

Dan LaDue 2/26/11 7:14 PM

Comment: Is this β_0 or β_1

The t-statistic tests whether the b-value is significantly different from 0. Because this is a simple linear regression with only one predictor variable, it is easy to interpret that the t-value is indeed significantly different from 0: the t-statistic is 8.635 and is also significant at $p < 0.01$. Therefore we can conclude that the predictor is making a significant contribute to the model: in other words, SES fits as a predictor of student achievement.

In looking at the standard errors, the $\beta = .559$. This value indicates that as SES increases by one standard deviation, student achievement will increase by .559 standard deviations, again reiterating the positive relationship between the predictor variable of SES and outcome variable of student achievement.

C. Hypothesis Testing

1. Pose the null and alternative hypothesis regarding the slope in the model. Specify these hypotheses in words first and then by using symbols.

The null hypothesis states that all the slopes are equal to zero in the population. It also holds that none of the independent variables are a good predictor of the outcome Y.

$$H_0 = \beta_1 = \beta_2 = \dots \beta_p = 0 \text{ or} \\ H_0 : = 0$$

The alternative hypothesis states that at least one of the independent variables is a non-zero in the population.

$$H_a: \beta_j \neq 0$$

2. What test statistic is appropriate for testing your null hypothesis?

The F-test is the ratio of the average variability that a given model can explain to the average variability unexplained by that same model. It is used to test the overall fit of the model in regression.

3. What is the probability of obtaining the slope estimate you've obtained (or one even more extreme) under the null hypothesis?

A good model will be a have a large F because that means the model is explaining more of the variance than is unaccounted for. If the F is greater than 1, then we know we have a good model. Using the tables for the critical values of the F-distribution to compare with the SPSS output of the F statistic as 74.51. The probability of getting this F-statistic under the null hypothesis is $p < 0.01$.

4. Make a decision about the null hypothesis (e.g., do you reject or retain the null hypothesis?). What does this tell you about the relationship between social economic status (SES) and student achievement (fltxcomp)?

It is safe to reject the null hypothesis based on the F-test. The probability of getting an F-statistic of 74.51 under the null hypothesis is $p < .01$ and can be rejected. At

Dan LaDue 2/26/11 7:15 PM

Comment: Not needed

Dan LaDue 2/26/11 7:21 PM

Comment: 2/2

Dan LaDue 2/26/11 7:21 PM

Comment: 2/2

Dan LaDue 2/26/11 7:21 PM

Comment: 2/2

Dan LaDue 2/26/11 7:22 PM

Comment: 2/2

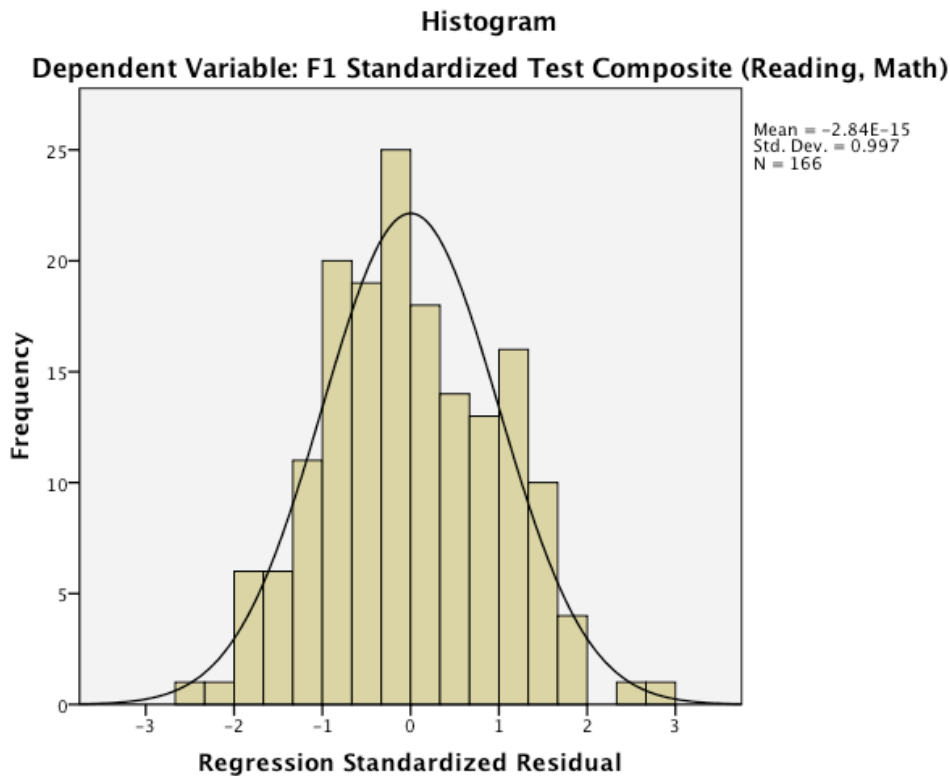
this point, based on this test, we are able to say that a good deal of the variance in student achievement (f1txcomp) can be explained by the predictor variable (SES) in this model.

5. Evaluate the assumption of normally distributed errors by visually inspecting a histogram of the residuals. Include relevant SPSS output. What is your assessment of this assumption?

Dan LaDue 2/26/11 7:23 PM
Comment: 3/3

Analyze> regression > linear regression. In the dialogue box choose “plot.” Put the ZPRED on the x axis (standardized predicted values y-hat that are stand according to the mean and sd of all y-hats) and put ZRESID (standardized residuals= the diff between the person’s actual data value for y and their predicted value aka y-hat) on the y axis.

Upon examination of the histogram it looks like the residuals are normally distributed. This would indicate that the assumption has been met that the errors are normally distributed.



Looking at the scatterplot, there appears to be homogeneity of variance due to the large cluster.

Part II

A. Multiple Regression: When we interpret our model we also assume that the variables are measured accurately and that the model has been specified correctly. In the case of model specification, there may be other variables which are related to the predictor social economic status (SES) and to student achievement (fltxcomp).

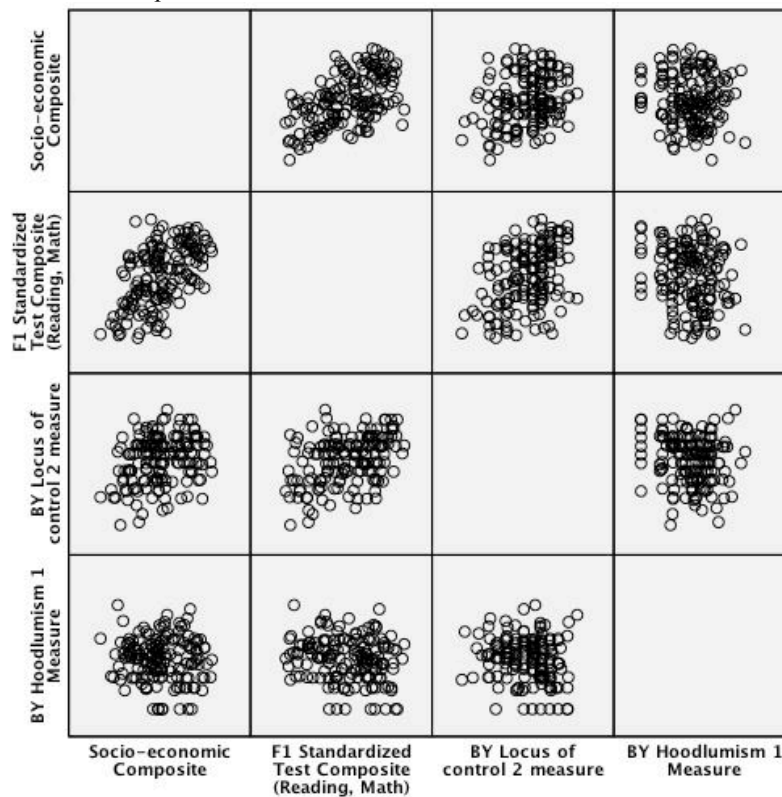
1. Identify at least one other variable in the data set that might have collinearity with social economic status (SES). Explain your reasoning.

Dan LaDue 2/26/11 7:24 PM

Comment: 2/2

We decided to execute a matrix scatterplot to look at not only the collinearity between SES and another variable, but that whichever variable we chose also had a predictive relationship with student achievement.

Here is the matrix scatterplot:



It looks to us like the locus of control measure both shows a positive linear relationship with the outcome variable and therefore would be a good candidate for a collinear relationship with SES. We looked at locus of control and hoodlumism because of previous knowledge that poverty is correlated with both of those attributes in students.

2. Write a population model that includes social economic status (SES) and the predictor you listed in question (1) to predict student achievement. Be sure to define each term in the model.

Dan LaDue 2/26/11 7:25 PM
 Comment: 3/3

Outcome = continuous predictors + error

$$Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + \epsilon_i$$

Outcome = intercept + slope for predictor + residual error for individual

- This equation indicates how the slope changes for each increase in Y
- Error is difference between predicted and actual values; the part of the Y's that can't be literally explained by the X's

Interpreted as:

$Y_i = f_{ltxcomp}$ (student achievement for ith participant)

b_0 = intercept (where the model exists in space)

b_1 = slopes for predictors (SES) X_{1i}

X_1 = SES variable

b_2 = slopes for predictor (bylocus2) X_{2i}

X_2 = bylocus2 variable

ϵ_i = residual for the ith participant

3. Use SPSS to obtain estimates of the parameters you specified in question (2) and report them in a table that includes the values of the coefficients, standard errors, t-statistics, and p-values.

Dan LaDue 2/26/11 7:27 PM
 Comment: 2.5/3
 You need to identify the values of β_0 , β_1 , and β_2 as well as including the tables.

Here's the output: Analyze > Regression > Linear

Coefficients ^a								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	(Constant)	51.317	.604		84.909	.000		
	BY Locus of control 2 measure	4.469	1.102	.266	4.054	.000	.925	1.081
	Socio-economic Composite	5.804	.801	.475	7.243	.000	.925	1.081

a. Dependent Variable: F1 Standardized Test Composite (Reading, Math)

4. State the null and alternative hypotheses in words and symbols for the “slope” associated with social economic status for the model you defined in question (2).

$$H_0 = \beta_1 = \beta_2 = \dots \beta_p = 0 \text{ or}$$
$$H_0 : = 0$$

In other words, the slopes for each independent variable would be equal to 0.

Dan LaDue 2/26/11 7:28 PM

Comment: 1/2
Alternative hypothesis?

5. Do you reject the null hypothesis? Why?

According to our SPSS output table above, we would reject the null hypothesis. The slope for SES = 5.804 and the slope for our next variable (Locus of Control bylocus2) = 4.469. Both are significant according to the t-test considering the standardized values and the degrees of freedom. The p-value < .001. All of this information would indicate that we can reject the null hypothesis.

Dan LaDue 2/26/11 7:28 PM

Comment: 3/3

6. Explain conceptually, in terms of overlapping variances, how multiple regression controls for a covariate such as the variable that you chose when estimating the effect of social economic status on student achievement.

Multiple regression controls for a covariate through partial slopes. Partial slopes are adjusted to account for the correlation between the two independent variables, in this case SES and locus of control. The square of the correlation represents the variance in one independent variable that is shared with the other. This shared variance reflects their redundancy when they are used to model Y.

Dan LaDue 2/26/11 7:29 PM

Comment: 3/3

7. Has the standard error for the slope associated with social economic status changed from the model you estimate in Part I B.2 to the model you estimated in Part II A.3? Explain any differences.

Yes, the standard error has changed slightly. In the first model, the standard error was .779 and in the new model with the additional independent variable, the standard error has increased slightly to .801. This is because now the model has become more complex and one would assume, given the additional variable, that the standard error might increase to account for the model including the additional variable.

Dan LaDue 2/26/11 7:30 PM

Comment: 2/2

8. Has the p-value for the test of the slope of social economic status changed from the model you estimate in Part I B.3 to the model you estimated in Part II A.3? Explain any differences.

No, the p-value has remained consistent for each test of the slope in both models.

Dan LaDue 2/26/11 7:30 PM

Comment: 2/2

9. Report the VIF for socioeconomic status as well as the variable you chose. Would you consider these VIFs to be ‘high’?

Dan LaDue 2/26/11 7:31 PM

Comment: 2/2
You should show the VIF in output from SPSS

The VIF for SES = 1.081
 The VIF for bylocus2= 1.081

These VIFs are slightly larger than 1, which may indicate some multicollinearity with SES, but are generally not of cause for concern. These VIFs would not be considered high.

Dan LaDue 2/26/11 7:31 PM
 Comment: ???

B. Pick an additional independent variable to add to the multiple regression model you specified in Part II, A.3.

Adding in Hoodlumism (byhood)

1. Using SPSS, compute the partial and semi-partial correlations for each of the three independent variables.

Dan LaDue 2/26/11 7:32 PM
 Comment: 2/2

Coefficients ^a									
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	54.132	1.533		35.319	.000			
	Socio-economic Composite	5.711	.798	.474	7.156	.000	.550	.501	.455
	BY Locus of control 2 measure	4.087	1.128	.243	3.625	.000	.394	.281	.231
	BY Hoodlumism 1 Measure	-1.378	.705	-.126	-1.954	.052	-.193	-.156	-.124

a. Dependent Variable: F1 Standardized Test Composite (Reading, Math)

SES: partial = .501; semi-partial = .455
 Locus of Control: partial = .281; semi-partial = .231
 Hoodlumism: partial = -.156; semi-partial = -.124

2. Provide an interpretation in words (and using the values reported by SPSS) of the squared partial correlation of each of the independent variables.

Dan LaDue 2/26/11 7:33 PM
 Comment: 2/2

SES: squared partial = .251; Locus of control: squared partial = .078; Hoodlumism: squared partial = .024.

Approximately 25% of the variation in the outcome variable not explained by locus of control or hoodlumism is explained by SES. Approximately 8% of the variation in the outcome variable not explained by SES or hoodlumism is explained by locus of control. Approximately 2% of the variation in the outcome variable not explained by SES or locus of control is explained by hoodlumism.

3. What does the squared semi-partial correlation suggest about the addition of the third independent variable that you added to the regression model?

Dan LaDue 2/26/11 7:34 PM
 Comment: 3/3

SES: squared semi-partial = .207; Locus of control: squared semi-partial = .053
 Hoodlumism: squared semi-partial = .015.

SES explains approximately 21% of the variance in the outcome variable once we have accounted for locus of control and hoodlumism. Locus of control explains approximately 5% of the variance in outcome once we have accounted for SES and hoodlumism. Hoodlumism explains roughly 2% of the variance in the outcome variable once we have accounted for SES and locus of control. The squared semi-partial correlation suggests that the addition of the third independent variable doesn't really add that much more to the explanation of the model. Our first two independent variables have significant overlap with hoodlumism.

4. Perform a statistical test of the differences between multiple Rs for this model and the model without the third independent variable. Specify the null and alternative hypotheses, compute the test statistic, and determine whether you reject the null based on an alpha of 0.05.

Dan LaDue 2/26/11 7:35 PM
 Comment: 3/3

$$\text{Model 2: Student achievement} = \beta_0 + \beta_{1SES} + \beta_{2Locus}$$

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.605 ^a	.366	.358	7.59097

a. Predictors: (Constant), Socio-economic Composite, BY Locus of control 2 measure

Model 2 Summary:

$$\text{Model 1: Student achievement} = \beta_0 + \beta_{1SES} + \beta_{2Locus} + \beta_{3Hoodlumism}$$

Model 1 Summary:

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.617 ^a	.381	.369	7.48959

a. Predictors: (Constant), BY Hoodlumism 1 Measure, Socio-economic Composite, BY Locus of control 2 measure

$$F = \frac{(R_1^2 - R_2^2)/(p_1 - p_2)}{(1 - R_1^2)/(n - p_1 - 1)} = \frac{(.37 - .36)/(2 - 1)}{(1 - .37)/200 - 2 - 1}$$

$$F_{1,198} = 31.25 \text{ and at } \alpha = .05$$

Using the table, this is significant, and we can reject the null hypothesis. Hoodlumism does add to the model in terms of explaining the outcome variable.

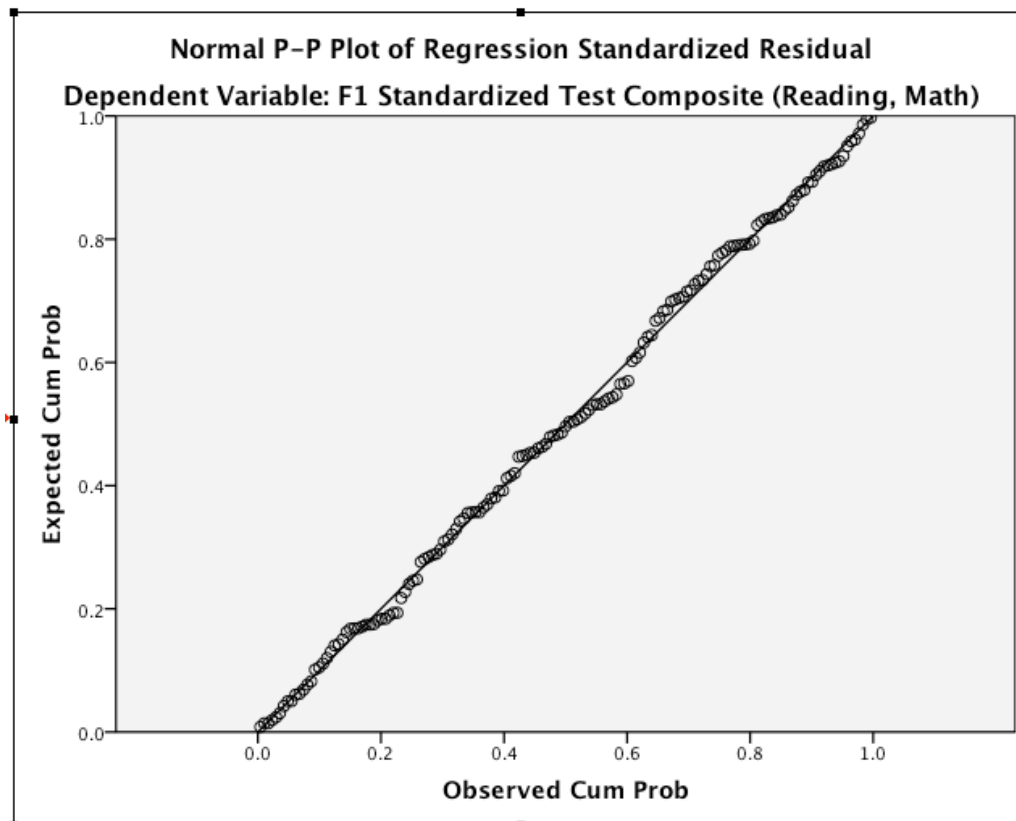
C. Using the model you specified in Part II, B., you will examine the tenability of the assumptions of multiple regression and evaluate the model for misfit.

1. Generate a P-P plot of the standardized residuals. Comment on what this plot suggests about the normality of the residuals.

There isn't a whole lot of stray in the residuals. Therefore, these residuals could be reasonably assumed to be normal.

Dan LaDue 2/26/11 7:36 PM

Comment: 2/2

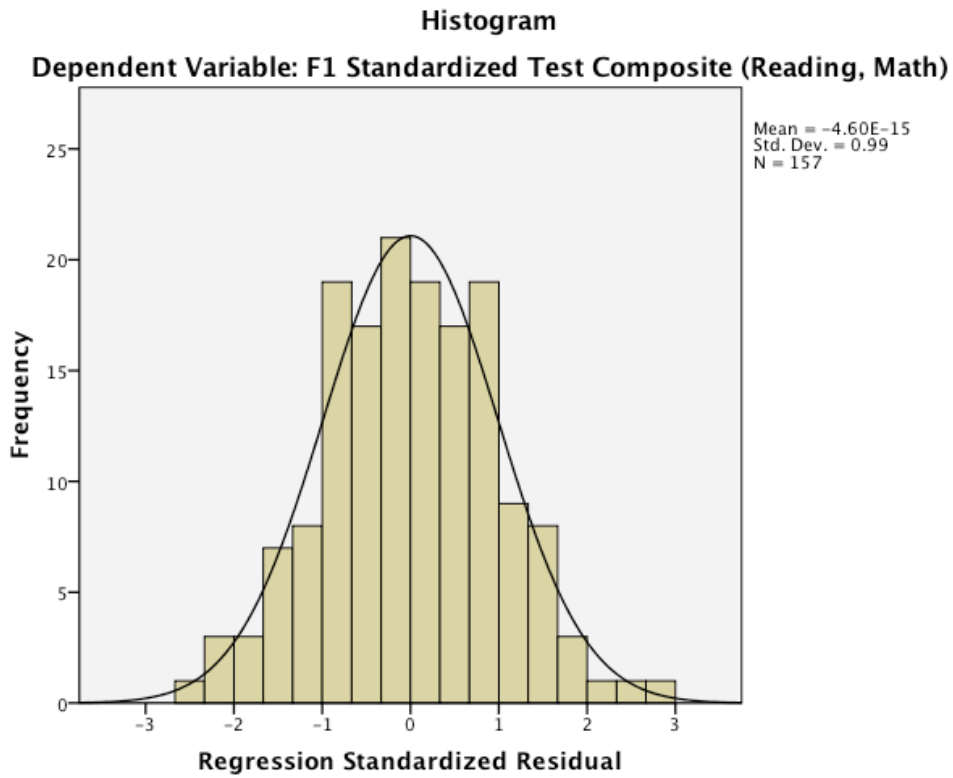


2. Examine a histogram of the standardized residuals. Comment on what this plot suggests about the normality of the residuals.

For this regression the histogram of the residuals is symmetric around zero. The residuals have a distribution that can reasonably be assumed to be normal.

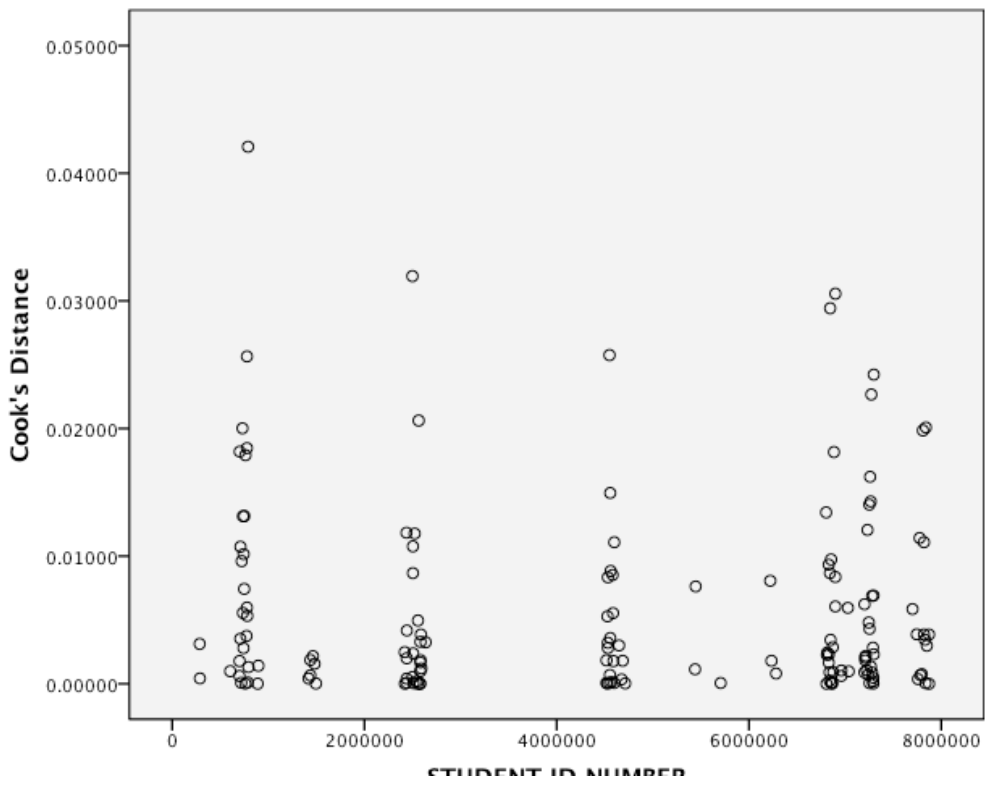
Dan LaDue 2/26/11 7:36 PM

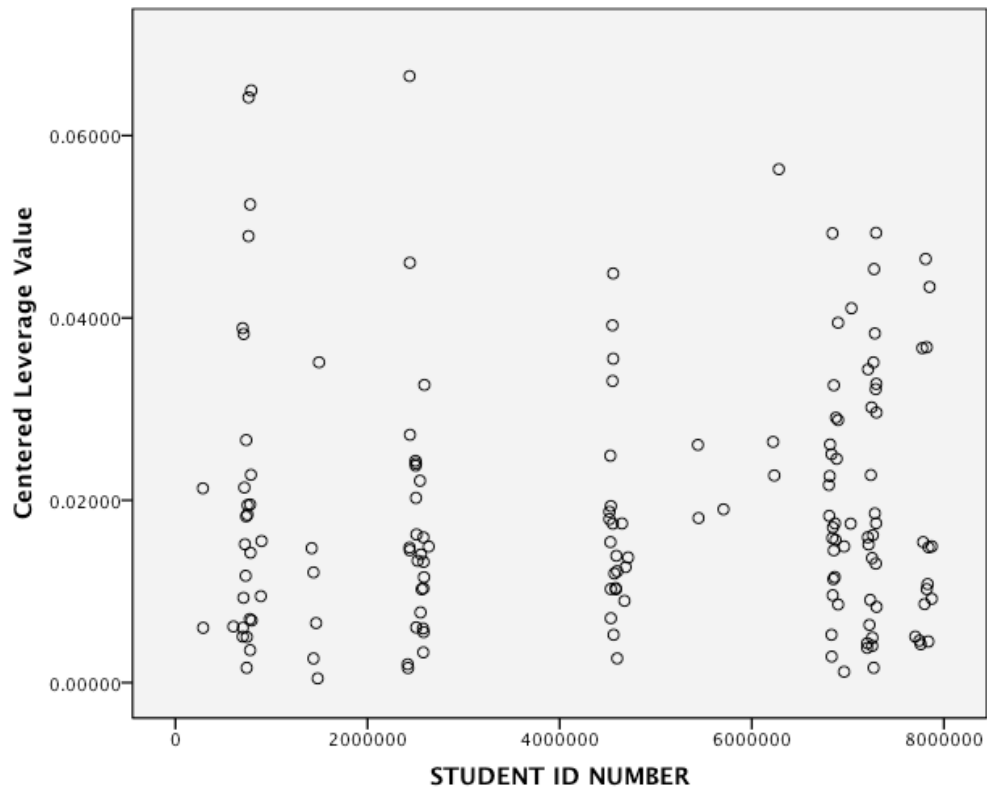
Comment: 2/2



3. Generate a plot of Cook's D and a plot of leverage. What do these plots suggest about the presence of any outliers?

Dan LaDue 2/26/11 7:38 PM
Comment: 2/2





Cook's Distance is a measure of the overall influence of a case on the model. Research suggests that values greater than one are of concern (Cook and Weisberg, 1982). The values for Cook's D above are all smaller than 1, therefore indicating that the outliers do not influence the effectiveness of the model as a whole. Leverage gauges the influence of the observed value of the outcome variable over the predicted value. Leverage: if $h > 2(k+1)/n$; in this case if $h > 8/157 = .0505$ if we see a data point that has a leverage that is greater than .0505, it's an indication that it may be an outlier.

4. Given what you found in Part II, C.1 – C.3, what do you think about the usefulness of the regression model?

Dan LaDue 2/26/11 7:40 PM

Comment: 3/3

There are two main aspects to a regression model. The first is to make sure that the model fits the data. The second is to be sure that the model is generalizable to the population. The steps we've taken through this homework to analyze the data have allowed us to define a model and then test the assumptions about that model to ensure generalizability.

In this particular case, this regression model seems to be a useful one; the predictor variables do a good job of predicting the outcome variable and fit the data. The three independent variables selected, SES, Locus of Control, and Hoodlumism each predict the

Dan LaDue 2/26/11 7:40 PM

Comment: Would you include hoodlumism even though it contributes so little to the independent variable?

outcome variable. Taken together, they represent three variables that go a long way to accounting for the variance in the outcome variable.

As we discussed earlier in the homework, multiple regression controls for a covariate through partial slopes. Partial slopes are adjusted to account for the correlation between the two independent variables, in this case SES and locus of control. The square of the correlation represents the variance in one independent variable that is shared with the other. This shared variance reflects their redundancy when they are used to model Y . With the addition of hoodlumism, this is taken a step further, and the slopes adjust accordingly.

In the end, we have examined a number of assumptions to gauge the generalizability of this particular regression model. We have identified variables in a model that have linear relationships to the outcome variable, we have tested for multicollinearity, and we have examined variance and determined the impact of outliers. The greatest weakness of this regression model is that there are more variables that remain unexamined that perhaps could, if added or traded for the independent variables we've identified here, improve the model. Model specification assumes that all the important predictors are in the model.