MORE ABOUT BIVARIATE REGRESSION

Regression Toward the Mean

Whenever prediction is not perfect, the best prediction is always less 'extreme' (i.e., is closer to its mean in standard deviation units) than the score it is predicted from.

Example 1. Using pretest scores to predict posttest scores

- Students with extremely low scores on the first test will often tend to show some improvement in standing on the second test.
- Students with extremely high scores on the first test will often tend to show some decline in standing on the second test.

Example 2. Sports

- Teams who perform exceptionally well in the previous year will probably perform relatively less well in their current year.
- Teams who perform exceptionally poorly in the previous year will probably perform relatively better in their current year.

Regression toward the mean would occur if $|r|\frac{S_Y}{S_X} < 1$.

Egression away from the mean could occur if S_Y is greater than S_X. Specifically, egression from the mean would occur if

$$|r|\frac{S_Y}{S_X} > 1$$

Inference in Linear Regression

The Normal Regression Model

Linear Regression Equation:

$$\hat{Y}_i = b_0 + b_1 X_i$$
$$Y_i = \hat{Y}_i + e_i$$

where

Y_i = person's score on the dependent variable

b_0 = Y intercept, the value of Y when $X = 0$.

b_1 = regression coefficient in the population, slope of the line, $\Delta Y / \Delta X$

X_i = person's score on the independent variable (predictor)

e_i = prediction error for the i$^{\text{th}}$ person.

$\hat{Y}_i$ = predicted value for the i$^{\text{th}}$ person.

Assumptions Made When Testing the Linear Regression Model

- The residuals, e_i, are normally distributed with a mean of zero.
- The variance of the residuals, is the same for all values of the predictor, *X*. This is called homoscedasticity.
- The covariance between residuals is zero. This means that the residuals are independent of each other.

These assumptions imply:
- *Y* is linearly related to *X*.
- *X* is a fixed variable—if we replicated the study, exactly the same values of *X* would be used.
- *X* is measured without error (i.e., reliability is 1).

The general format for a confidence interval about the population value is:

$$\hat{\theta} \pm t_{crit}\left(s\tan darderror_{\hat{\theta}}\right)$$

The general format for testing a hypothesis about the population value is:

$$t = \frac{\hat{\theta} - \theta_{hyp}}{s\tan darderror_{\hat{\theta}}}$$

Parent	Child
15	12
17	15
17	16
11	10
14	9
16	18
18	10
26	19
15	10
11	11

Descriptive Statistics for Parent Data

Score	Score minus the Mean	(Score minus the Mean)^2	
15	15 – 16 = -1	1	**Parent Mean**
17	17 – 16 = 1	1	**16**
17	17 – 16 = 1	1	
11	11 – 16 = -5	25	
14	14 – 16 = -2	4	**Parent Variance**
16	16 – 16 = 0	0	**162/9 = 18.00**
18	18 – 16 = 2	4	
26	26 – 16 = 10	100	
15	15 – 16 = -1	1	**Parent Standard Deviation**
11	11 – 16 = -5	25	4.2426
Mean 16		SS_{Parent} 162	

Descriptive Statistics for Child Data

Score	Score minus the Mean	(Score minus the Mean)^2	**Child Mean**
12	12 – 13 = -1	1	**13**
15	15 – 13 = 2	4	
16	16 – 13 = 3	9	
10	10 – 13 = -3	9	**Child Variance**
9	9 – 13 = -4	16	**122/9 = 13.5556**
18	18 – 13 = 5	25	
10	10 – 13 = -3	9	
19	19 – 13 = 6	36	**Child Standard Deviation**
10	10 – 13 = -3	9	**3.6818**
11	11 – 13 = -2	4	
Mean 13		SS_{Child} 122	

Covariance and Correlation Between Parent and Child Data

(Parent Score – Parent Mean)(Child Score – Child Mean)
-1(-1) = 1
1(2) = 2
1(3) = 3
-5(-3) = 15
-2(-4) = 8
0(5) = 0
2(-3) = -6
10(6) = 60
-1(-3) = 3
-5(-2) = 10
$SS_{Parent,Child}$
96

Covariance Between
Parent and Child Data

$$S_{XY} = \frac{SS_{Parent,Child}}{N-1} =$$

96/9

= **10.66666**

Correlation between Parent and Child Data

$$r = \frac{\text{Covariance}_{\text{Parent,Child}}}{\text{Standard Deviation}_{\text{Parent}} * \text{Standard Deviation}_{\text{Child}}} = \frac{10.666667}{4.2426(3.6818)} = .6829$$

Regression Coefficient, Slope

$$b_1 = r\frac{S_Y}{S_X} = .6829\frac{3.6818}{4.2426} = 0.5926$$

Intercept

$$b_0 = \overline{Y} - b_1\overline{X} = 13 - .5926(16) = 3.5184$$

Linear Regression Equation

Predicted Child's Score = 3.5184+0.5926(Parent's Score)

Coefficient of Determination, *r^2*

$.6829^2 = .4664$

$$\textbf{\textit{MSE}} = \frac{(1-r^2)SS_Y}{N-2} = \frac{(1-.4664)122}{10-2} = 8.1374$$

Standard Error = $\sqrt{MSE} = \sqrt{8.1374} = 2.8526$

Testing the Significance of the Entire Regression Model

Step 1. Write the Null and Alternative Hypotheses; Write the Full and Reduced Models.

Null Hypothesis:

Reduced Model:

Alt. Hypothesis:

Full Model:

Step 2. Determine the Alpha Level and the Critical Value for the *F* Test.

Alpha Level =
Numerator Degrees of Freedom **(*p*)** =
Denominator Degrees of Freedom **(*N - p - 1*)** =
F Critical Value:

Step 3. Calculate SSR and SSE for the regression model. (See *Relations Between Quantitative Variables Chapter.*)

$$\boldsymbol{SS_{Regression}} = r^2 SS_Y \qquad \boldsymbol{SS_{Residual}} = \left(1 - r^2\right) SS_Y$$

Step 4. Calculate the Statistical Test.

$$F = \frac{SS_{\mathrm{Regression}} / p}{SS_{\mathrm{Residual}} / (N - p - 1)} = \frac{MS_{\mathrm{Regression}}}{MSE}$$

Step 5. Determine the significance of the statistical test by comparing *F* to the critical value.

Step 6. Write a sentence that explains your results. For example,

- Results of the linear regression analysis indicate that age was not a significant predictor of IQ scores (F(1,30) = 2.01, MSE = 88.25, p > .05, $\boldsymbol{R^2_{Adj}} = .04$).
- Results of the linear regression analysis indicate that age was a significant predictor of IQ scores (F(1,30) = 8.29, MSE = 67.15, p < .05, $\boldsymbol{R^2_{Adj}} = .14$). See Table 1 for the regression model.

Testing the Significance of a Single Regression Coefficient

Step 1. Write the Null and Alternative Hypotheses; Write the Full and Reduced Models.

Null Hypothesis:

Reduced Model:

Alt. Hypothesis:

Full Model:

Step 2. Determine the Alpha Level and the Critical Value for the *t* Test.

Alpha Level =
Degrees of Freedom ***(N - p - 1)*** =
t Critical Values:

Step 3. Calculate the Statistical Test.

$$t = \frac{b_1 - \beta_{1hyp}}{\sqrt{\frac{MSE}{SS_X}}}$$

Step 5. Determine the significance of the statistical test by comparing *t* to the critical values.

Step 6. Write a sentence that explains your results. For example,

- Age was not a significant predictor of IQ scores after adjusting for other predictors in the model, so it was dropped from the final model.
- Age was a significant predictor of IQ scores after adjusting for other predictors in the regression model.

Instead of conducting a test of the Null Hypothesis, you could calculate the Confidence Interval for an Estimated Regression Coefficient…

$$\begin{aligned} \beta_1 &= b_1 \pm t_{crit}\left(\text{Estimated Standard Error of the Regression Coefficient}\right) \\ &= b_1 \pm t_{crit}\sqrt{MSE/SS_X} \end{aligned}$$

Testing the Significance of the Intercept of a Regression Model

Step 1. Write the Null and Alternative Hypotheses; Write the Full and Reduced Models.

Null Hypothesis:

Reduced Model:

Alt. Hypothesis:

Full Model:

Step 2. Determine the Alpha Level and the Critical Value for the *t* Test.

Alpha Level =
Degrees of Freedom **(*N - p - 1*)** =
t Critical Values:

Step 3. Calculate the Statistical Test.

$$t = \frac{b_0 - \beta_{0Hyp}}{\sqrt{MSE\left(\frac{1}{N} + \frac{\overline{X}^2}{SS_X}\right)}}$$

Step 5. Determine the significance of the statistical test by comparing *t* to the critical values.

Step 6. Write a sentence that explains your results.

This is rarely done because it usually makes no sense.

Instead of conducting a test of the Null Hypothesis, you could calculate the Confidence Interval for an Estimated Regression Coefficient…

$$\beta_0 = b_0 \pm t_{crit}(\text{Estimated Standard Error of the Regression Intercept})$$
$$= b_0 \pm t_{crit}\sqrt{MSE\left(\frac{1}{N} + \frac{\overline{X}^2}{SS_X}\right)}$$

Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
PARENT	16.0000	4.24264	10
CHILD	13.0000	3.68179	10

Correlations

		PARENT	CHILD
PARENT	Pearson Correlation	1	.683*
	Sig. (2-tailed)	.	.030
	N	10	10
CHILD	Pearson Correlation	.683*	1
	Sig. (2-tailed)	.030	.
	N	10	10

*. Correlation is significant at the 0.05 level (2-tailed).

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	PARENT[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: CHILD

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.683[a]	.466	.400	2.85287

a. Predictors: (Constant), PARENT

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	56.889	1	56.889	6.990	.030[a]
	Residual	65.111	8	8.139		
	Total	122.000	9			

a. Predictors: (Constant), PARENT

b. Dependent Variable: CHILD

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	3.519	3.698		.951	.369	-5.009	12.046
	PARENT	.593	.224	.683	2.644	.030	.076	1.109

a. Dependent Variable: CHILD

Correlations

Descriptive Statistics

	Mean	Std. Deviation	N
GPA	2.5830	.84833	30
IQ	97.2667	13.27932	30

Correlations

		GPA	IQ
GPA	Pearson Correlation	1	.702**
	Sig. (2-tailed)	.	.000
	N	30	30
IQ	Pearson Correlation	.702**	1
	Sig. (2-tailed)	.000	.
	N	30	30

**. Correlation is significant at the 0.01 level

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	GPA[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: IQ

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.702[a]	.492	.474	9.62840

a. Predictors: (Constant), GPA

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2518.095	1	2518.095	27.162	.000[a]
	Residual	2595.772	28	92.706		
	Total	5113.867	29			

a. Predictors: (Constant), GPA

b. Dependent Variable: IQ

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	68.894	5.721		12.04	.000	57.176	80.613
	GPA	10.984	2.108	.702	5.212	.000	6.667	15.301

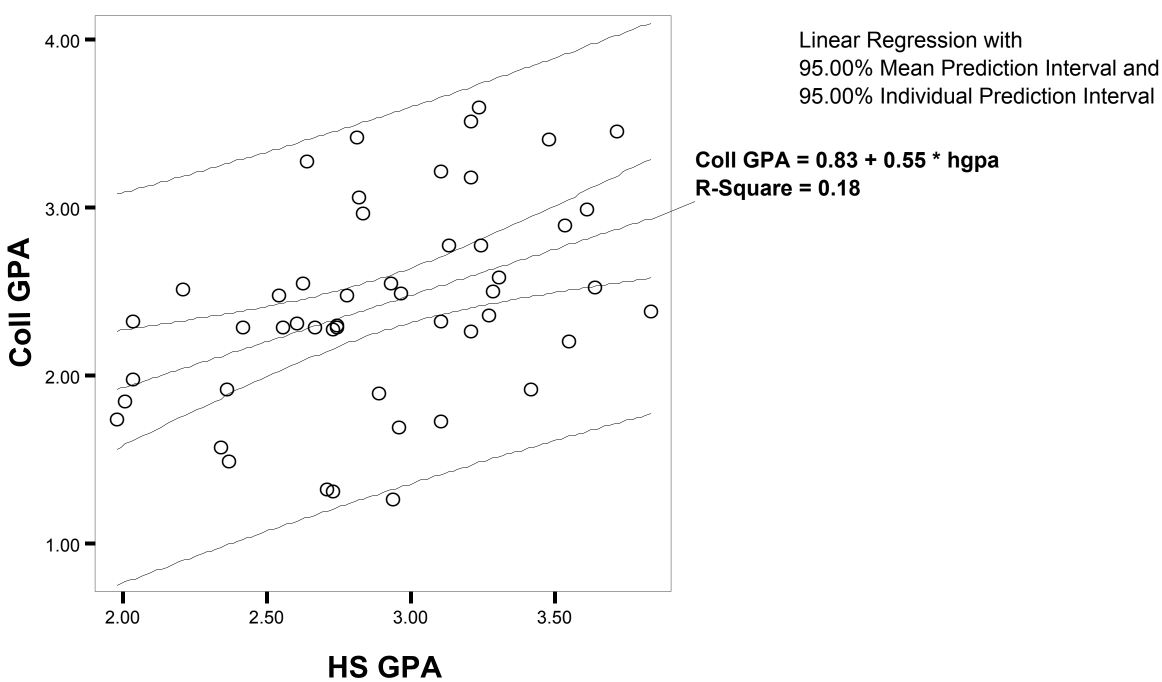
a. Dependent Variable: IQ

Total Sample Size Needed to Have a Specified Power with Alpha = .05 for a Test that all of the regression coefficients are zero (i.e., $R^2 = 0$)—Anova *F* Test.

# of	*Power = .70*			*Power = .80*			*Power = .90*		
Predictors	Small	Med	Large	Small	Med	Large	Small	Med	Large
2	389	55	26	485	68	31	636	88	40
3	444	63	30	550	77	36	713	99	45
4	489	70	33	602	85	40	776	108	50
5	529	76	36	647	92	43	830	116	53
6	564	81	39	688	98	46	878	123	57
7	596	86	42	725	103	49	922	130	60
8	626	91	44	759	109	52	962	136	63

Small: $R^2 = .0196$ and $R = .14$; **Medium:** $R^2 = .1304$ and $R = .36$; **Large:** $R^2 = .2592$ and $R = .51$

Note. **GPOWER** software was used to obtain a priori sample size estimates. See Erdfelder, E., Faul, F., & Buckner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Comptuers, 28,* 1-11. It can be downloaded for free from the web. Use **gpower** as your search term to obtain the most current website.



95% CI for the Conditional Mean Of Y at a specific X value

Example: A university wants to predict the mean college gpa of students based on their high school gpa. As well, they want the 95% confidence interval for the mean college gpa of all students who have a high school GPA of 3.0.

The predicted mean college gpa is calculated using a high school gpa of 3.0.

The standard error of the predicted mean value is:

$$SE(\hat{\mu}_{Y.X_j}) = \sqrt{MSE\left(\frac{1}{N} + \frac{(X_j - \overline{X})^2}{SS_X}\right)}$$

Therefore, the 95% confidence interval for the predicted mean value would be calculated as follows:

$$\hat{Y} \pm t_{crit} SE(\hat{\mu}_{Y.X_j})$$

95% CI for an Individual Y Score at a specific X value

Example: An academic advisor wants to predict the college gpa of a student who had a high school gpa of 3.0. As well, they want the 95% confidence interval for the predicted college gpa of the person who has a high school gpa of 3.0.

The predicted college gpa is calculated using a high school gpa of 3.0.

The standard error of the predicted value is:

$$SE(\hat{Y}_{newj}) = \sqrt{MSE\left(1 + \frac{1}{N} + \frac{(X_j - \overline{X})^2}{SS_X}\right)}$$

Therefore, the 95% confidence interval for the predicted mean value would be calculated as follows:

$$\hat{Y}_{newj} \pm t_{crit} SE(\hat{Y}_{newj})$$

The more X is deviant from the mean of X, the larger the standard error. This makes senses because a more deviant a score, the more prone it is to error.

The standard errors allow for prediction error due to differences in Y and differences in X.

Confidence intervals for predicting an individual score are much larger than confidence intervals for predicting a mean of numerous scores.

Regression Analysis in Nonexperimental Research

Linear regression can be used when *X* is a random variable (rather than fixed). The same equations and statistical tests are conducted.

But, <u>additional</u> assumptions about the data must be used for linear regression to be appropriate.

- *Y* is drawn from a normal distribution with mean $\mu_{Y.X} = \beta_0 + \beta_1 X$ and constant variance (homoscedasticity).
- The predictor, *X*, and the residuals, *e*, are not correlated with each other.

Regression When X is Subject to Random Error

As stated previously, the impact of unreliability (measurement error) in the predictor is to reduce the size of the slope—when there is only one predictor in the regression model.

The impact of measurement error is more ambiguous when there are multiple predictors in the regression model.

Checking for Violations of Assumptions

Checking Assumptions by Using Residuals

The residuals can provide useful information about whether there are violations of the assumptions of linearity, homogeneity of variance (homoscedasticity), normality and independence of error.

SPSS will generate 2 graphs for evaluating normality.

- Histogram of the residuals with a normal distribution superimposed on the graph.
- Probability plot of the residuals.

<u>How to Read a Probability Plot</u>.

- If the residuals are normally distributed then the points fall on a straight line.

- If the sample data are positively skewed compared to the normal distribution, then some data points will lie to the right of the line. If the sample data are negatively skewed compared to the normal distribution, then some data points will lie to the left of the line.

Violations of the linearity and homogeneity of variance assumptions may cause the residuals to depart from normality, so that generally the linearity and homogeneity of variance assumptions should be checked before looking for violations of the normality assumption.

Residual-Predicted Plots (SDRESID as the Y and ZPRED as the X)

- Indicate whether error variance increases as a function of the dependent variable (i.e., heteroskedasticity).

Residual-Predictor Plots (SDRESID as the Y against each X predictor)

These would be most easily obtained by saving the SDRESID to the data file and then using a Scattterplot Matrix to create all the graphs. SDRESID and each of the predictors would be included in the matrix.

Indicate whether there is a systematic relationship between error variance and a particular *X*.

Partial Regression Plots

The objective of data analysis for several variables is typically to investigate *partial* relationships (between pairs of variables, "controlling" statistically for other variables). The partial regression plots show the partial relationships between Y and each of the X's.

- Show the partial relationship between Y and each of the X predictors.

- Determine whether nonlinearity is a concern, but are not as useful for locating a transformation.

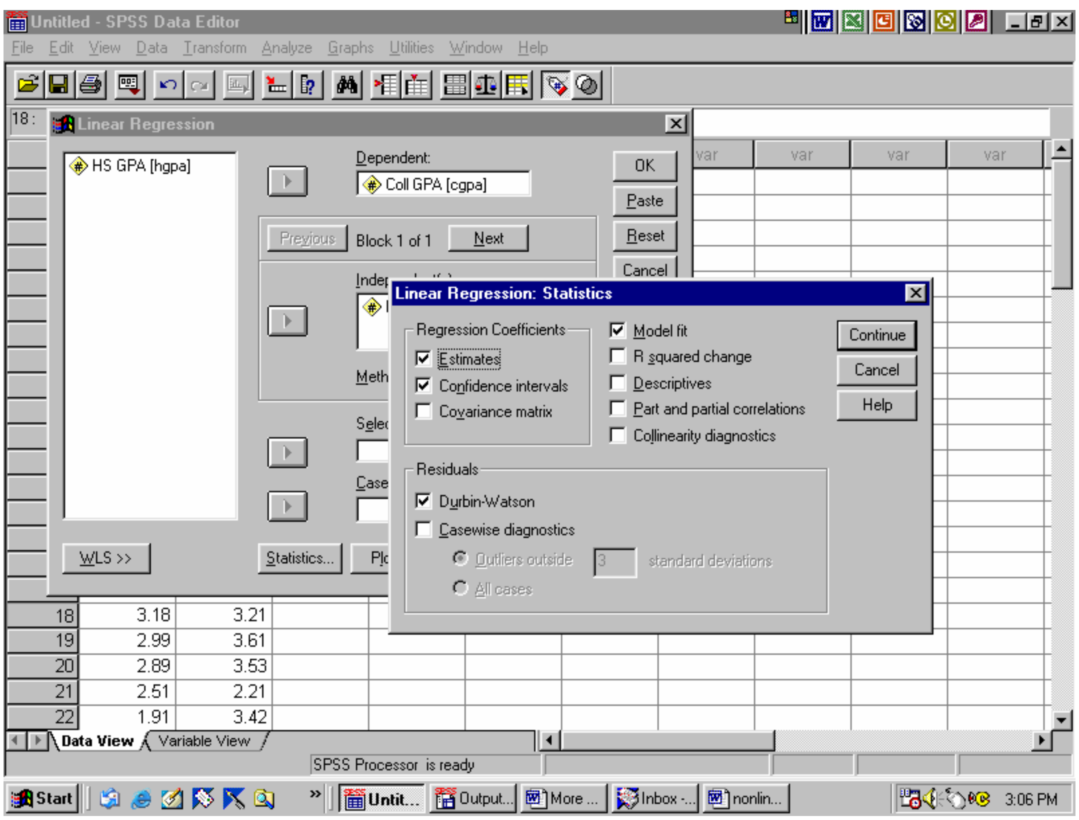
Checking for Independence among the Residuals

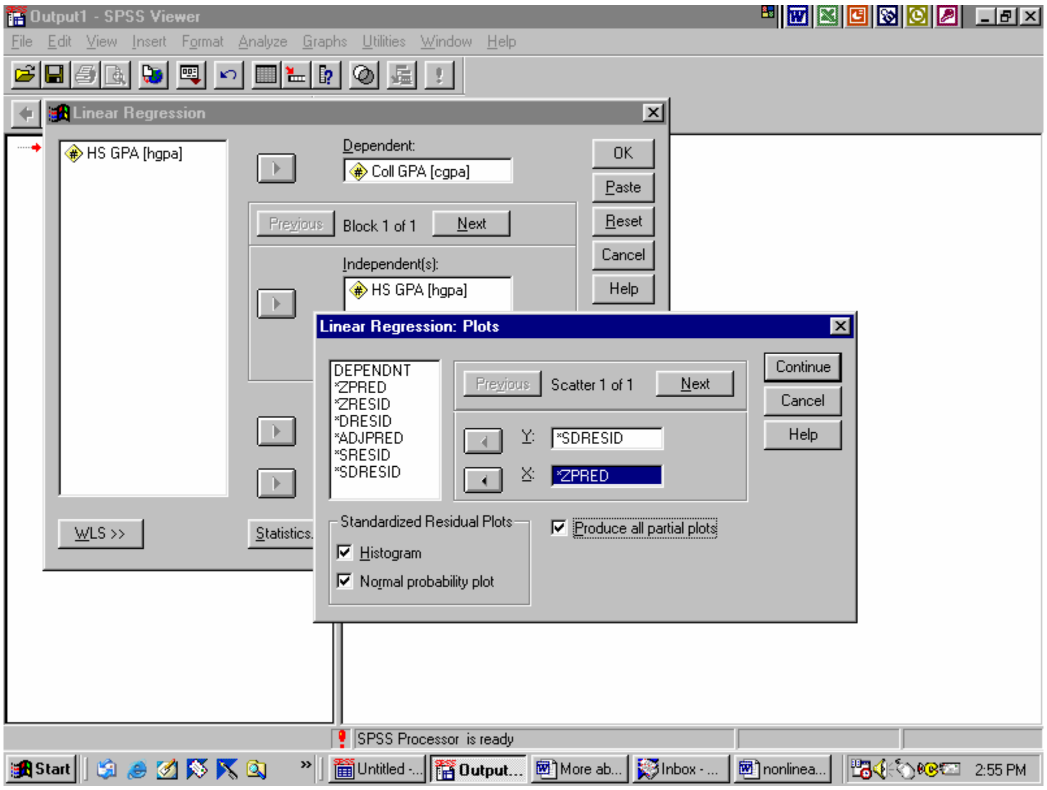
If the residuals are positively correlated, perhaps because of the omission of some important variables from the model, standard errors calculated using OLS procedures may underestimate the true standard deviations of the regression coefficients, and the confidence intervals and hypothesis tests will be too liberal.

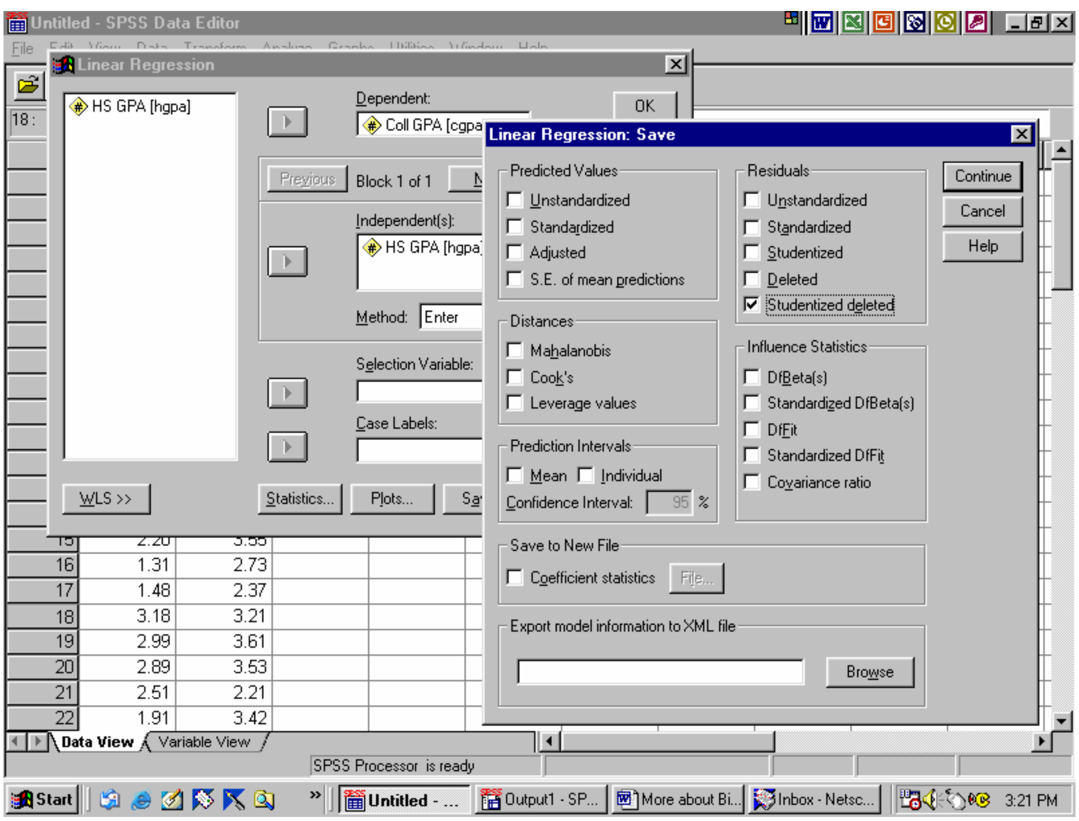
If the residuals are negatively correlated, the confidence intervals and hypothesis tests will be too conservative.

It is possible to test for correlation among the residuals using the Durbin-Watson statistic.

- *D* is approximately equal to 2(1 – autocorrelation).
- *D* can range from 0 to 4, with larger deviations from 2 providing stronger evidence of serial correlation.
- *D* will be smaller than 2 when residuals are positively correlated.
- *D* will be larger than 2 when residuals are negatively correlated.
- Draper and Smith (1998) or Mendenhall and Sincich (1996) provide tables for determining whether the autocorrelation is significantly different from zero. It also is possible to use SPSS or SAS to test this. A complete discussion of this is beyond the scope of this course.




Output1 - SPSS Viewer
File Edit View Insert Format Analyze Graphs Utilities Window Help
Linear Regression
HS GPA [hgpa]
Dependent:
Coll GPA [cgpa]
OK
Paste
Reset
Cancel
Help
Previous
Block 1 of 1
Next
Independent(s):
HS GPA [hgpa]
WLS >>
Statistics...
Linear Regression: Plots
DEPENDNT
*ZPRED
*ZRESID
*DRESID
*ADJPRED
*SRESID
*SDRESID
Previous
Scatter 1 of 1
Next
Continue
Cancel
Help
Y: *SDRESID
X: *ZPRED
Standardized Residual Plots
Histogram
Normal probability plot
Produce all partial plots
SPSS Processor is ready
Start
2:55 PM


Untitled - SPSS Data Editor
Linear Regression
HS GPA [hgpa]
Dependent:
Coll GPA [cgpa]
OK
Previous
Block 1 of 1
Independent(s):
HS GPA [hgpa]
Method: Enter
Selection Variable:
Case Labels:
WLS >>
Statistics...
Plots...
Linear Regression: Save
Predicted Values
Unstandardized
Standardized
Adjusted
S.E. of mean predictions
Distances
Mahalanobis
Cook's
Leverage values
Prediction Intervals
Mean
Individual
Confidence Interval: 95 %
Residuals
Unstandardized
Standardized
Studentized
Deleted
Studentized deleted
Influence Statistics
DfBeta(s)
Standardized DfBeta(s)
DfFit
Standardized DfFit
Covariance ratio
Continue
Cancel
Help
Save to New File
Coefficient statistics
File...
Export model information to XML file
Browse
16 1.31 2.73
17 1.48 2.37
18 3.18 3.21
19 2.99 3.61
20 2.89 3.53
21 2.51 2.21
22 1.91 3.42
Data View
Variable View
SPSS Processor is ready
Start
3:21 PM

Regression

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	HS GPA[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: Coll GPA

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin-Watson
1	.428[a]	.183	.166	.55013	2.054

a. Predictors: (Constant), HS GPA

b. Dependent Variable: Coll GPA

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	3.192	1	3.192	10.547	.002[a]
	Residual	14.224	47	.303		
	Total	17.416	48			

a. Predictors: (Constant), HS GPA

b. Dependent Variable: Coll GPA

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	.829	.496		1.670	.102	-.169	1.827
	HS GPA	.548	.169	.428	3.248	.002	.209	.888

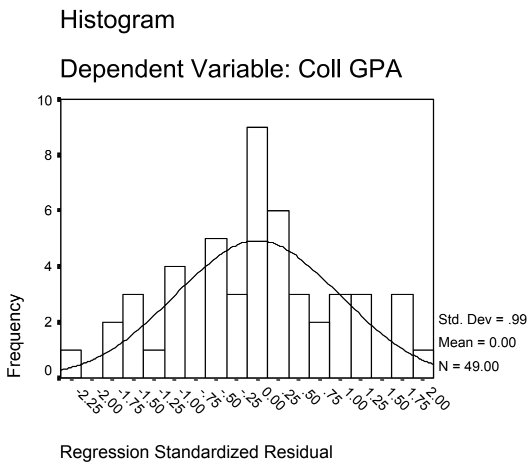
a. Dependent Variable: Coll GPA

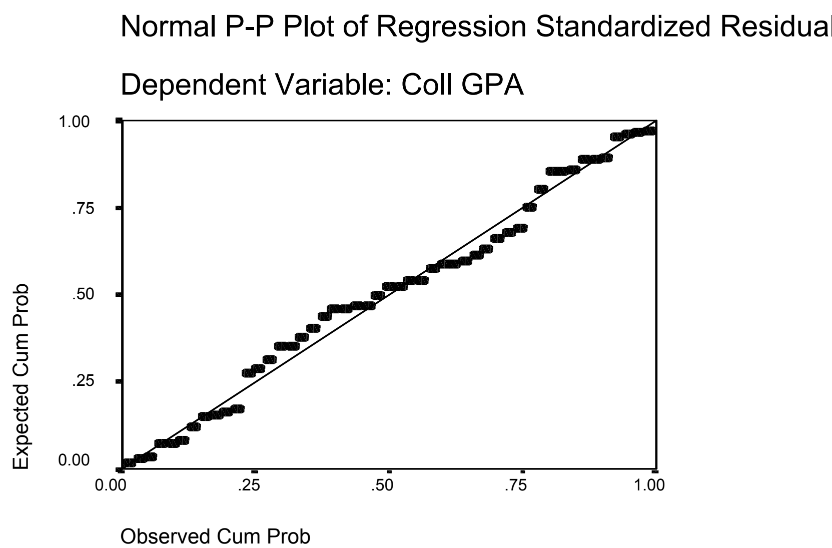
Residuals Statistics[a]

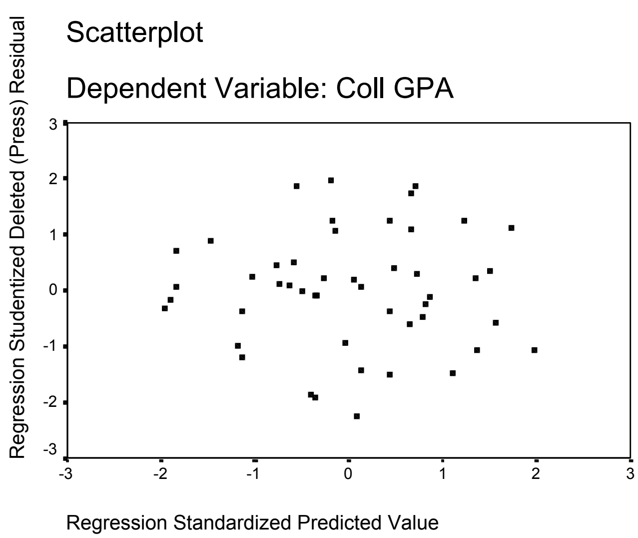
	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1.9142	2.9307	2.4204	.25788	49
Std. Predicted Value	-1.963	1.979	.000	1.000	49
Standard Error of Predicted Value	.07863	.17569	.10715	.02983	49
Adjusted Predicted Value	1.9066	2.9937	2.4207	.26019	49
Residual	-1.1778	1.0371	.0000	.54437	49
Std. Residual	-2.141	1.885	.000	.990	49
Stud. Residual	-2.163	1.905	.000	1.007	49
Deleted Residual	-1.2025	1.0595	-.0004	.56348	49
Stud. Deleted Residual	-2.255	1.962	-.001	1.024	49
Mahal. Distance	.001	3.916	.980	1.165	49
Cook's Distance	.000	.064	.018	.019	49
Centered Leverage Value	.000	.082	.020	.024	49

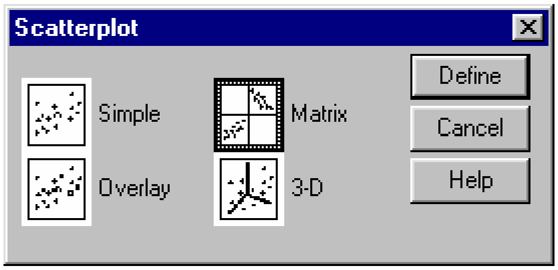
a. Dependent Variable: Coll GPA

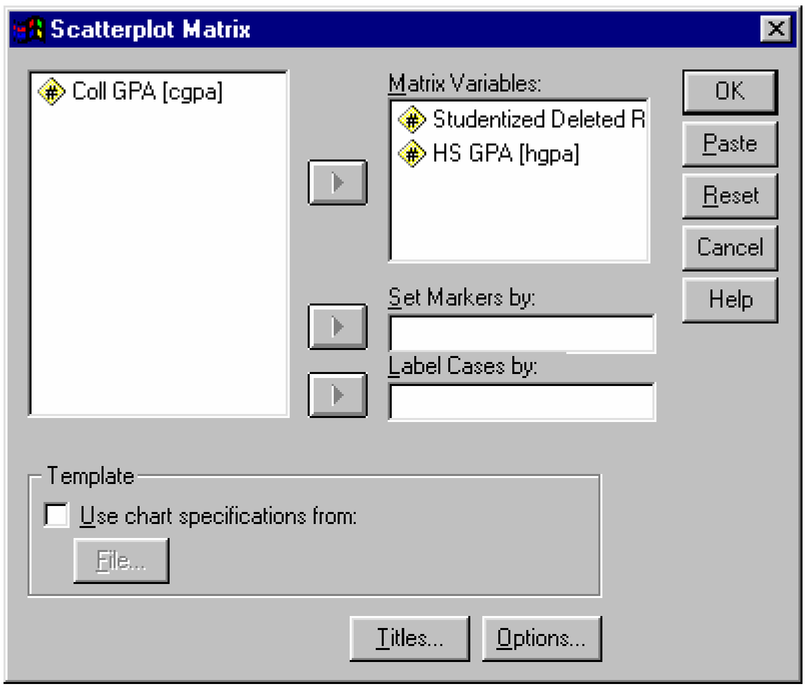
Charts

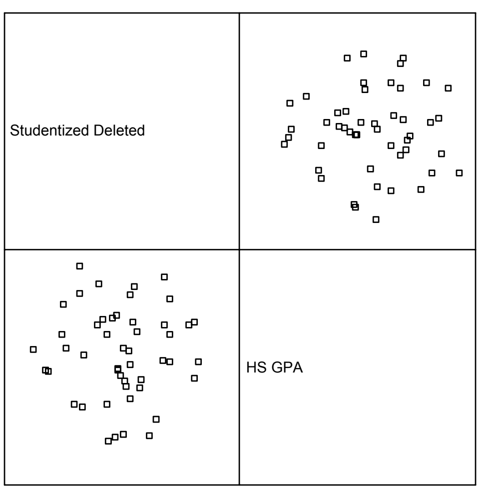

Histogram
Dependent Variable: Coll GPA
Frequency
Std. Dev = .99
Mean = 0.00
N = 49.00
Regression Standardized Residual


Normal P-P Plot of Regression Standardized Residual
Dependent Variable: Coll GPA
Expected Cum Prob
Observed Cum Prob


Scatterplot
Dependent Variable: Coll GPA
Regression Studentized Deleted (Press) Residual
Regression Standardized Predicted Value


Scatterplot
Simple
Matrix
Define
Cancel
Help
Overlay
3-D


Scatterplot Matrix
Coll GPA [cgpa]
Matrix Variables:
Studentized Deleted R
HS GPA [hgpa]
OK
Paste
Reset
Cancel
Help
Set Markers by:
Label Cases by:
Template
Use chart specifications from:
File...
Titles...
Options...


Studentized Deleted
HS GPA

Locating Outliers and Influential Data Points

An outlier among a set of residuals is much larger than the rest in absolute value, perhaps lying three or more standard deviations from the mean of the residuals.

An outlier may indicate
- recording errors
- you have an unusual subpopulation in your data.
- the regression estimates are distorted
- Model misspecification

Scientific judgment is more important here than statistical tests, once influential observations have been flagged.

Simple Approaches to Diagnosing Problems in Data

You should be familiar with each of the following aspects of your data:
- The type of subject (e.g., male humans, cars, snakes)
- The procedure for collecting data
- The unit of measurement for each variable (e.g., kilograms, inches, likert scaling, IQ)
- A plausible range of values and a typical value for each variable

First, list the five largest and five smallest values for every variable. This allows you to detect data recording errors, format errors in computer input, and some outliers.

> The mere fact that an observation appears to be unusual when compared with the rest of the data does not automatically mean that is should be dropped.

Second, calculate descriptive statistics.
- For continuous variables: mean, standard deviation, minimum and maximum values, graphs
- For variables with limited values: frequency tables.

Third, conduct an analysis of residuals and other regression diagnostic procedures because they provide the most refined and accurate evaluation of model assumptions.

Types of Residuals

1. A residual is defined as $\boldsymbol{e_i = Y_i - \hat{Y}_i}$

   - It represents the amount of discrepancy between the observed and predicted values from the regression model.

   The residuals are influenced by the scale of the dependent variable and the independent variables.

2. Standardized Residual…analogous to a Z-score. Standardized Residuals have a mean of zero and a variance of 1.

$$Z_i = \frac{residual}{\sqrt{MSE}} = \frac{Y_i - \hat{Y}_i}{\sqrt{MSE}}$$

•Reflects the discrepancy between the predicted and observed Y values.

3. Studentized Residual…analogous to a *t* score with *N-k-1* degrees of freedom. The h_i have values between 0 and 1. ***Studentized Residuals*** have a mean near zero with a variance slightly larger than 1.

$$r_i = \frac{residual}{\sqrt{MSE(1-h_i)}} = \frac{Y_i - \hat{Y}_i}{\sqrt{MSE(1-h_i)}}$$

where h_i is the leverage of the i^{th} observation in determining the model fit.

- Reflects the discrepancy between the predicted and observed Y values.
- Reflects observations that have undue influence on the regression model because of their unusual values on the independent variables.

4. Studentized Deleted (a.k.a., Jackknife Residual) is a studentized residual with the effect of the i^{th} observation deleted from the ***MSE***. The studentized deleted residuals have a mean near zero and a variance slightly greater than 1.

$$r_{(-i)} = r_i \frac{\sqrt{MSE}}{\sqrt{MSE_{(-i)}}} = \frac{residual}{\sqrt{MSE_{(-i)}(1-h_i)}} = \frac{Y_i - \hat{Y}_i}{\sqrt{MSE_{(-i)}(1-h_i)}}$$

A large studentized deleted residual may be due to one or more of three reasons.

- **The person has an unusual score on the dependent variable**. The numerator, $e_i = Y_i - \hat{Y}_i$ reflects the extremeness of the i^{th} observation with respect to the dependent variable compared to the predicted score on the dependent variable.

- **The person has an unusual influence on the fit of the model.** The ratio of the two MSE's reflects the degree to which the i^{th} observation affects the fit of the model.

- **The person has an unusual score on one or more of the independent variables.** The h_i represents the **leverage** of the i^{th} observation…it is a measure of the geometric distance of the i^{th} predictor point $(X_{i1}, X_{i2}, \ldots, X_{ik})$ from the center point $(\bar{X}_{i1}, \bar{X}_{i2}, \ldots, \bar{X}_{ik})$ of the predictor space. It represents outliers in the predictors.

Influence Diagnostics

Cook's Distance

It is a measure of the *influence* of an observation on the **regression estimates**. It is the extent to which the regression coefficients change when the particular observation in question is deleted.

Cook's d_i may be large because:

- the observation is extreme in the predictor space (***h***), or
- the observation has a large studentized residual.

DFBETA's (standardized)

Whereas Cook's D can be viewed as a measure of the simultaneous change in the parameter estimates when an observation is deleted from the analysis, **DFBETA's measure the change in the parameter estimate for each individual predictor.**

Using Residual Diagnostics to Detect Outlying Observations and Assumption Violation

1. **Save the following statistics to the SPSS data file.** Request that Cook's D, Leverage Values, Studentized Deleted Residuals, Standardized DFBeta(s), and Standardized DFfit statistics be saved in the data file.

2. **Create Scatter Plots of the Residuals.** The residual plots allow us to check for violation of the model assumptions (e.g., nonlinearity, heteroskedasticity, model misspecification) and outliers.

   - Plot of studentized deleted residuals versus predicted Y values.
   - Plot of studentized deleted residuals versus each independent variable.

3. **Inspect the Residual and Influence Diagnostics for extreme values.**

   When viewing the studentized deleted residuals, compare the actual values to a *t* critical value with *N-k-2* degrees of freedom. Use an alpha of .05/*N* to control for the *N* statistical tests you perform. If the actual values are larger in absolute value than the critical value, consider the observation to be an outlier.

   To find the *t critical value*, calculate by hand:

   df = *N*-*k*-2
   alpha = .05/*N*.
   half of alpha = *alpha/2*.

Example: *N = 74* and *k = 3*. *df =69* and *alpha = .000675676* Thus, *half of alpha = .000337838*

   Then, use SPSS to obtain the exact *t* critical value.

   The general format is:
   jackcrit = IDF.T(1-half of alpha, df)

Jackcrit = IDF.T(1-.000337838, 69) = 3.5605

When viewing the leverage values, check for values greater than $[2(k + 1)]/N$. If the value is greater, consider the observation to be an outlier.

In our example, if a leverage value is greater than $0.1081 = \frac{2(3+1)}{74}$ then it is an outlier.

When viewing Cook's Distance, go to Table A10 in the KKM text and find the tabled critical value for your N and k. Divide the tabled critical value by $(N\text{-}k\text{-}1)$. Then, compare your Cook's Distance values to the divided critical value.

For our example, $N = 74$ and $k = 3$, so I will use $N \approx 50$. The tabled value is 17.93 for an alpha of .05. The transformed tabled value is 17.93/(50-3-1) = 0.38978. If the Cook's Distance is more extreme than 0.38978 then that observation is an outlier.

When viewing standardized DFBETA's,

- Belsley et al. (1980, p. 28) suggest an "absolute cutoff" of 2.
- **Belsley et al. (1980, p. 28) suggest a "size-adjusted cutoff" of $2/\sqrt{N}$ for small samples.**
- Neter et al. (1989, p. 403) suggested $2/\sqrt{N}$ be used for large data sets whereas 1 serve as a cutoff for "small to medium data sets."
- Mason et al. (1989, p. 520) suggest $3/\sqrt{N}$ be used as a general cutoff.

For our example, if a DFBeta has a value that is larger than $.2325 = \frac{2}{\sqrt{74}}$ then it is an outlier.

Please note,

1. I would focus on the impact that observations have on the estimates (i.e., DFBETAs and Cook's D).
2. An observation would be considered an outlier if it exceed the "rules of thumb" in either direction (+ or -).

What do you do once you have identified influential observations?

If you have outliers due to recording errors, etc. you must correct the recording error. If correction is not possible, you must delete the score.

If you have outliers that are not due to recording errors, etc. then you may

(1) Delete the outliers,
(2) Run two sets of analyses…one with all the data and one with the outliers removed, or
(3) Use an alternative to linear regression that is not so easily distorted by outliers (e.g., robust regression or Weighted Least Squares Regression).