REGRESSION DIAGNOSTICS
CHAPTER 3

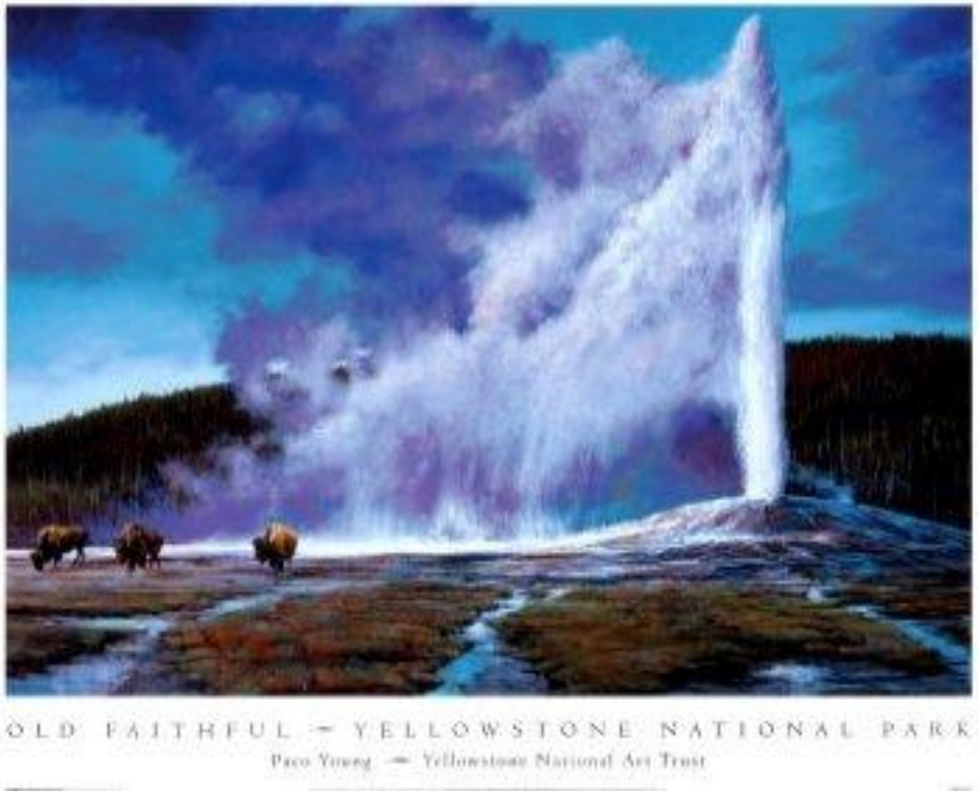
ERSH 8320 • Lecture 5 • August 30, 2007

Today's Class

- Regression diagnostics.
  - All the little things you need to look at before and after you run a regression to determine if you can interpret your results in a meaningful way.
- Outliers

Example Data For Today's Class

- To help introduce the concepts discussed today, we make use of data seen previously in ERSH 8320:


OLD FAITHFUL — YELLOWSTONE NATIONAL PARK
Paco Young — Yellowstone National Art Trust

Old Faithful Data

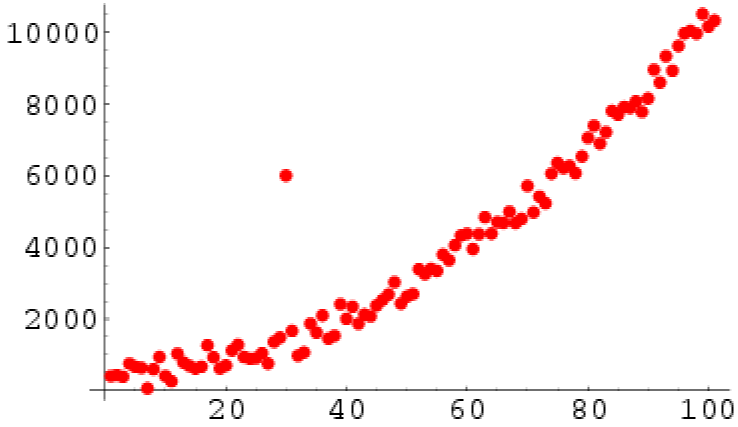
- Example taken from Weisberg (1985, p. 230):

``Perhaps the most famous geyser is Old Faithful, in Yellowstone National Park, Wyoming. The intervals of eruptions of Old Faithful range from about 30 to 90 minutes. Water shoots to heights generally over 35 meters, with eruptions lasting from 1 to 5.5 minutes.

``Data on Old Faithful has been collected for many years by ranger/naturalists in the park, using a stopwatch. The duration measurements have been rounded to the nearest 0.1 minute or 6 seconds, while intervals reported are to the nearest minute. The National Park Service uses x (values of the duration of an eruption) to predict y (the interval to the next eruption)."

Outliers

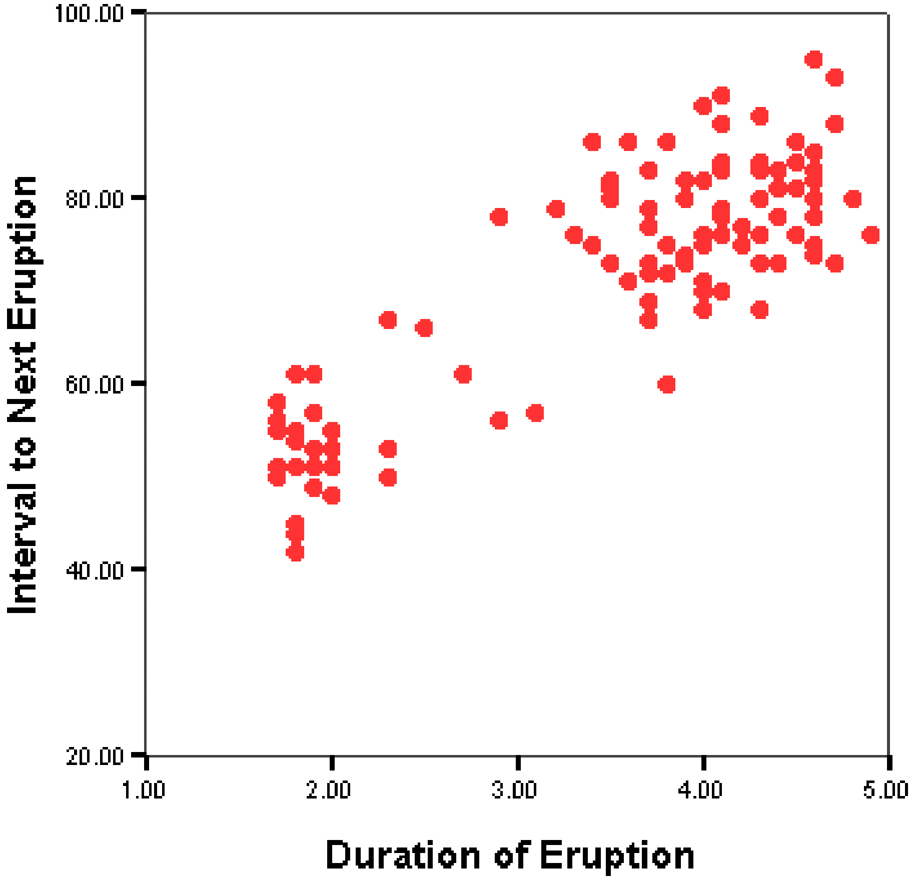
- You may notice that some of your data deviate from the rest on your scatterplot.
  - These points are called outliers:



Outliers

- There may be several reasons for outliers to occurs. These include:
  - Measurement error
  - Input error
  - Malfunction of instrument
  - Subjects inappropriately trained
- Some outliers can be removed for the above reason (i.e. you can correct the data), however true outliers are not a result of errors.
  - It may be beneficial to determine the cause of the outlier.

Old Faithful Scatterplot

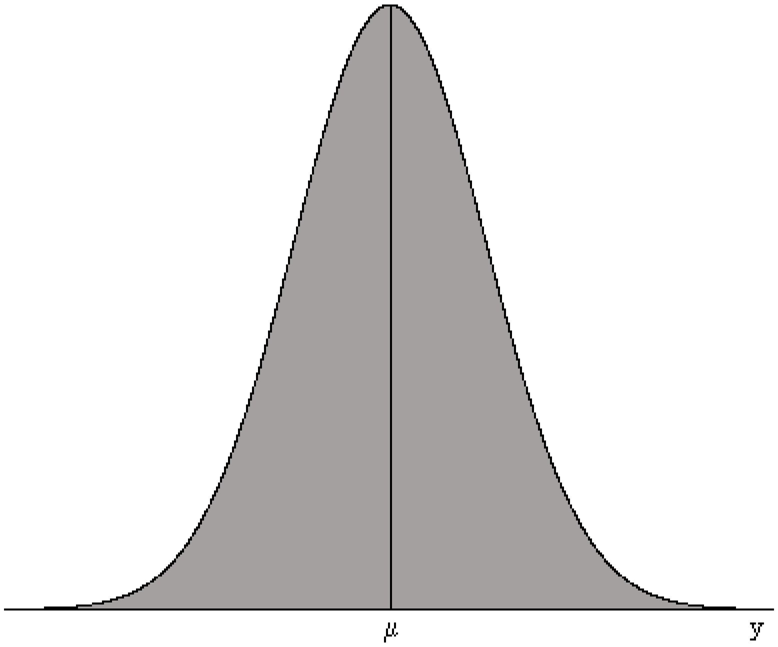


Detection of Outliers

- The most common way to detect an outlier is by evaluation of residuals .
  - extreme residual = extreme observation
- Definition of residual: $y - \hat{y}$
- 3 common residual analyses
  - Standardized Residuals (ZRESID)
  - Studentized Residuals (SRESID)
  - Studentized Deleted Residuals (SDRESID)

Standardized Residuals (ZRESID)

- All the residuals put into a normal distribution format

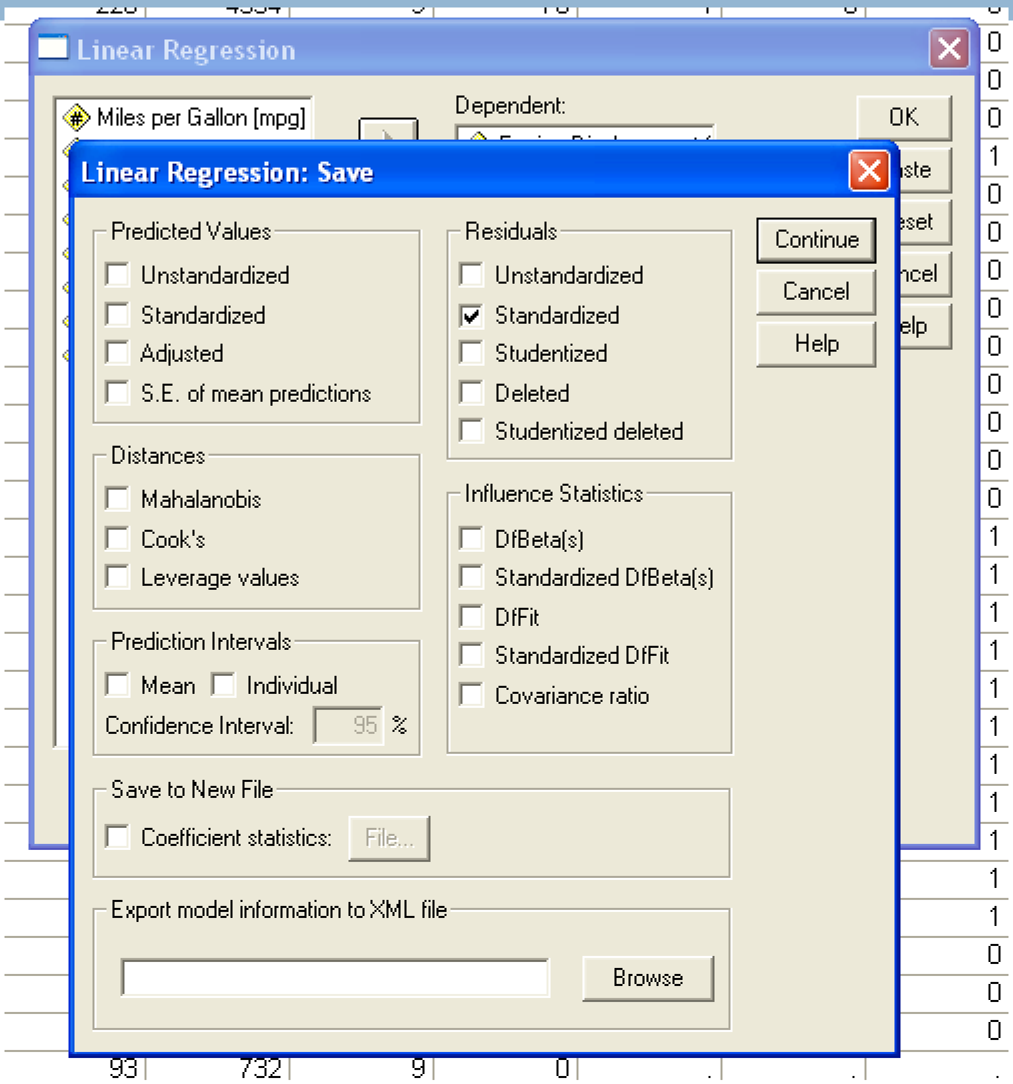


Standardized Residuals (ZRESID)

$$zresid = \frac{y - \hat{y}}{s_{y.x}}, \text{ where } s_{y.x} \text{ is the standard error of the estimate}$$

- ☐ To normalize, we take the variable, subtract the mean, and divide by the standard deviation.
- ☐ Here, we are normalizing the residual, the mean is 0 and the standard deviation is $s_{y.x}$.

Standardized Residuals (ZRESID)



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

Save to New File
- ☐ Coefficient statistics:

Export model information to XML file

Studentized Residuals (SRESID)

- The previous calculation (ZRESID) was based on the assumption that all residuals have the same variance.
  - This assumption is not usually valid.

- To correct for this, we use the studentized residuals (SRESID). Instead of using $s_{y.x}$, we will use a rather long corrected standard deviation.

Studentized Residuals (SRESID)

- The rather long corrected standard deviation:

$$s_{e_i} = s_{y.x}\sqrt{1 - \left[\frac{1}{N} + \frac{X_i - \bar{X}^{\;2}}{\sum x^2}\right]}$$

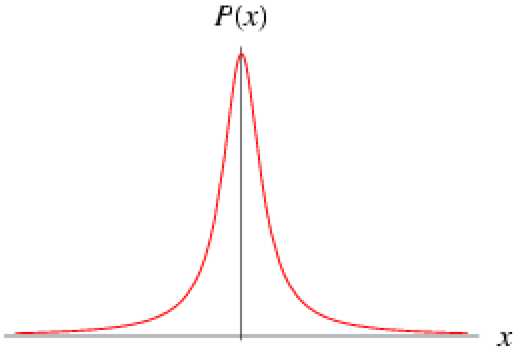
- So, what is this long formula doing?
  - We will see later that the term in brackets is really the leverage of the observation.
  - So, it is correcting the standard deviation by a penalty factor.
  - The higher the leverage (or pull of the observation on the regression line), the larger the corrected standard deviation.

Studentized Residuals (SRESID)

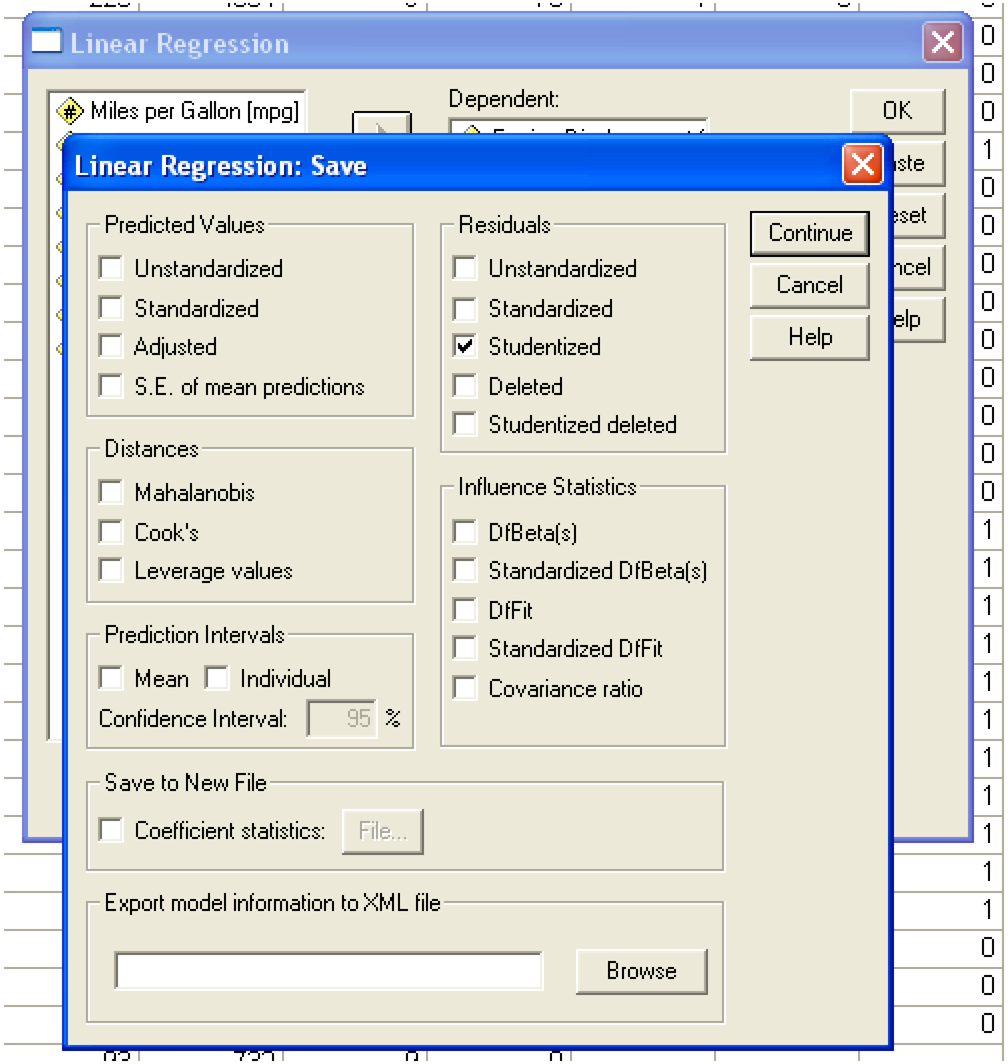
- After the correction, we use the same basic formula:

$$sresid = \frac{y - \hat{y}}{s_{e_i}}$$

- These studentized residuals follow a student's t distribution with df = N-k-1 , where N = sample size and k = # of independent variables



Studentized Residuals (SRESID)



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

Save to New File
☐ Coefficient statistics: File...

Export model information to XML file

Studentized Deleted Residuals (SDRESID)

- These are fairly similar to the previous residuals (SRESID).
- Instead of correcting in the way we did before, we correct in an even more complicated way!
- This time, we are going to delete the observation in question from the analysis, find the standard error of the estimate, then correct in the same way we did before.

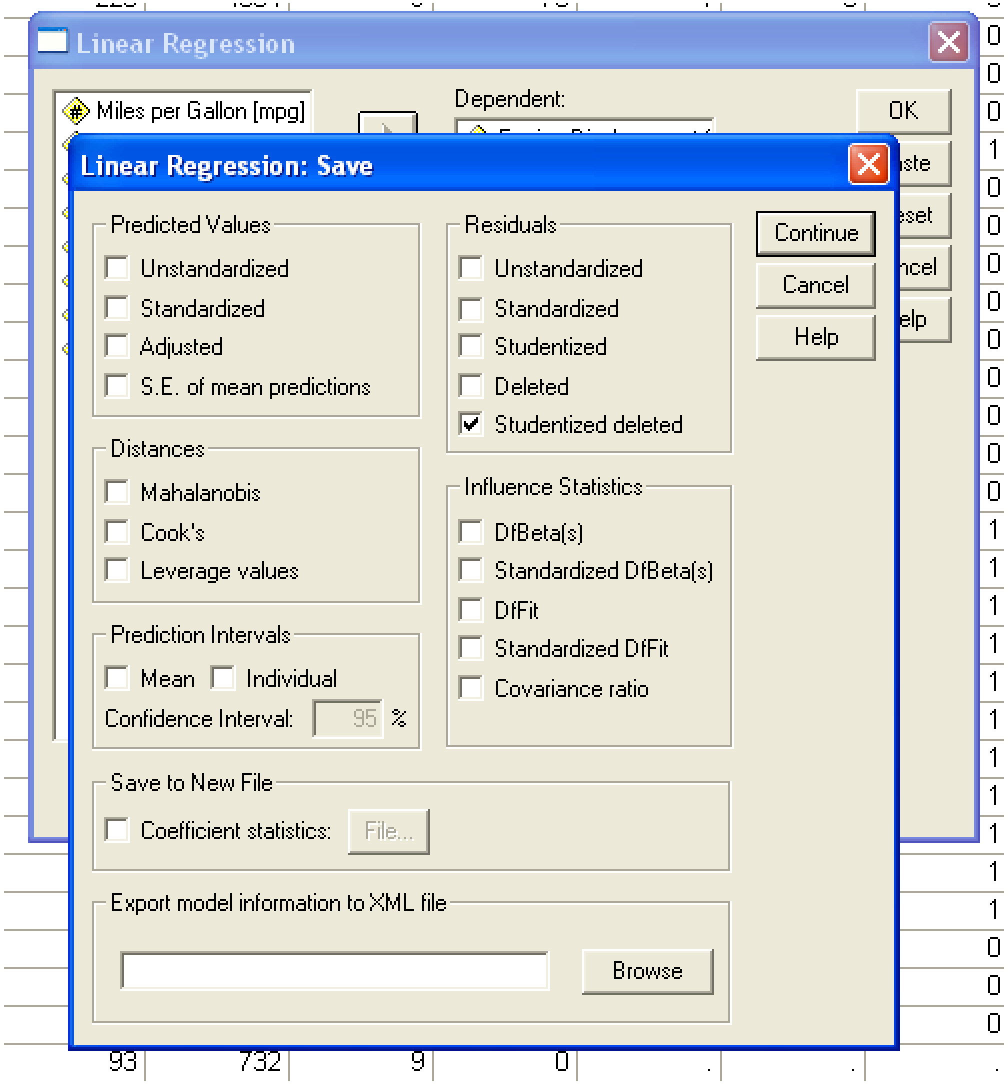
Studentized Deleted Residuals (SDRESID)

$$s_{e_i} = s_{y.x_i}\sqrt{1-\left[\frac{1}{N}+\frac{{X_i-\bar{X}}^{2}}{\sum x^{2}}\right]}$$

- So, the (i) means that the ith observation is deleted, then the standard deviation is calculated.
- Again, this is distributed t(N-k-2)

$$sdresid = \frac{y-\hat{y}}{s_{e_i}}$$

Studentized Deleted Residuals (SDRESID)



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

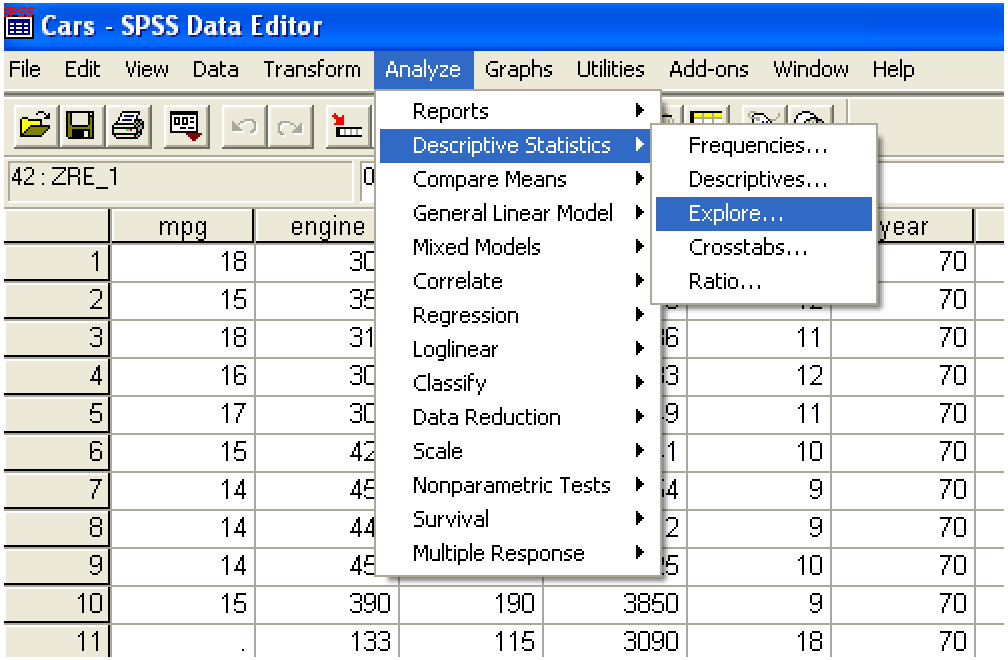
Influence Statistics
☐ DfBeta(s)
☐ Standardized DfBeta(s)
☐ DfFit
☐ Standardized DfFit
☐ Covariance ratio

Save to New File
☐ Coefficient statistics: File...

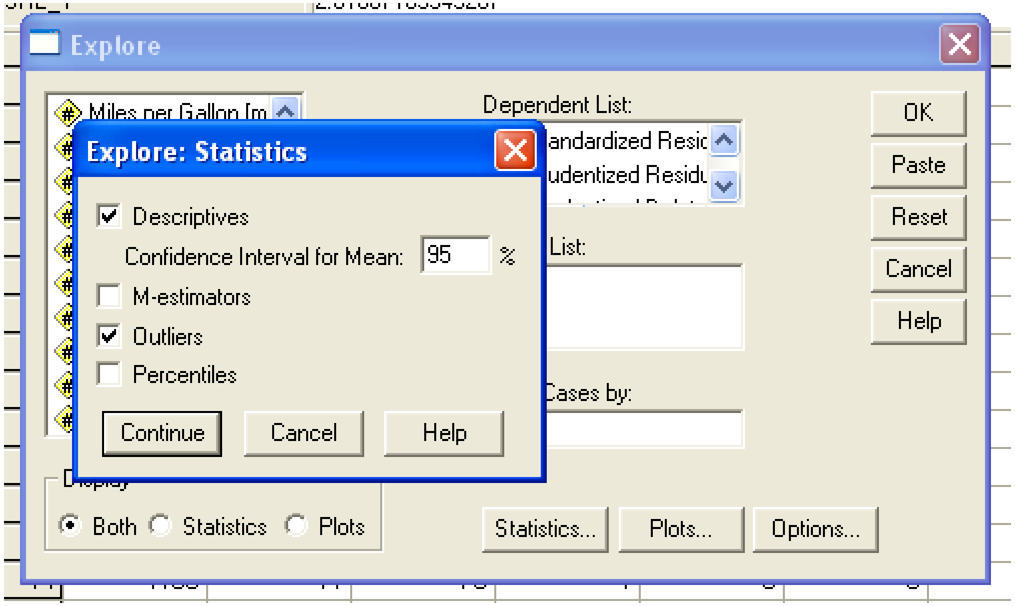
Export model information to XML file

Exploring the Distribution of the Residuals

- A good way to look at the distribution of the residuals:
  - Go to Analyze
  - Descriptive Statistics
  - Explore



Exploring Residuals



Miles per Gallon [m

Dependent List:
andardized Resic
udentized Residu

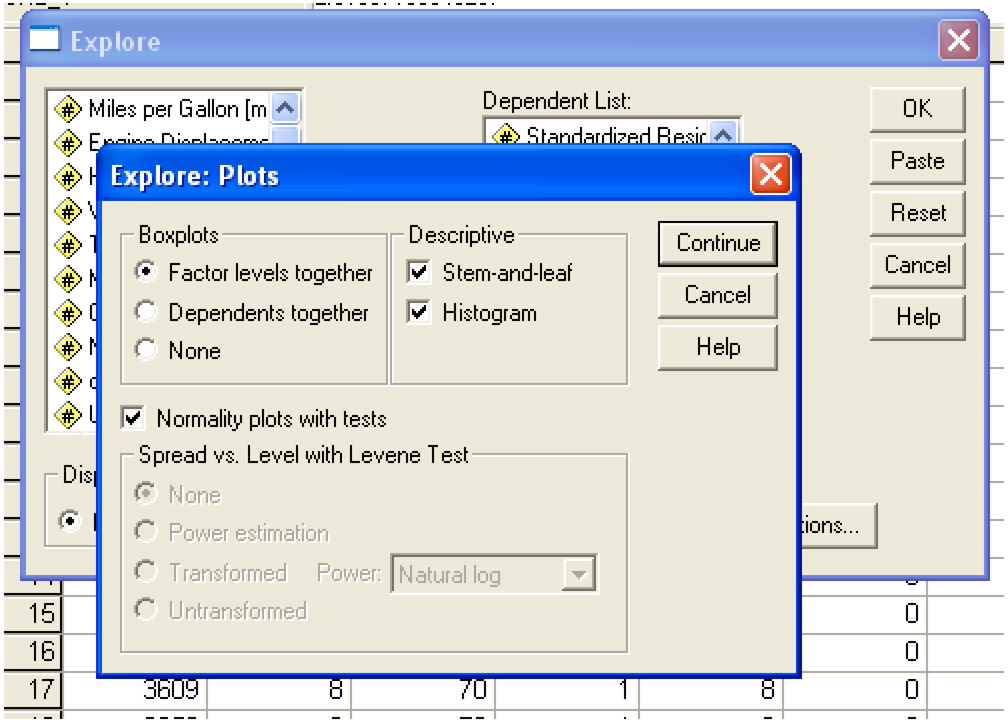
List:

Cases by:

Display
● Both ○ Statistics ○ Plots

☑ Descriptives
Confidence Interval for Mean: 95 %
☐ M-estimators
☑ Outliers
☐ Percentiles

Exploring Residuals



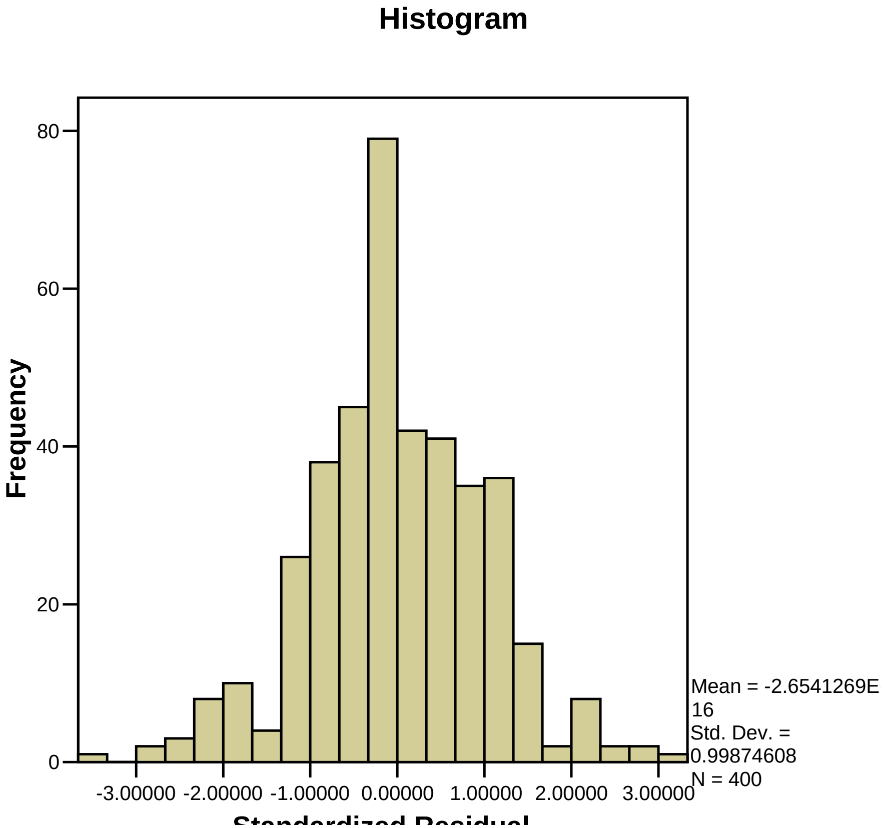
Boxplots
● Factor levels together
○ Dependents together
○ None

Descriptive
☑ Stem-and-leaf
☑ Histogram

☑ Normality plots with tests

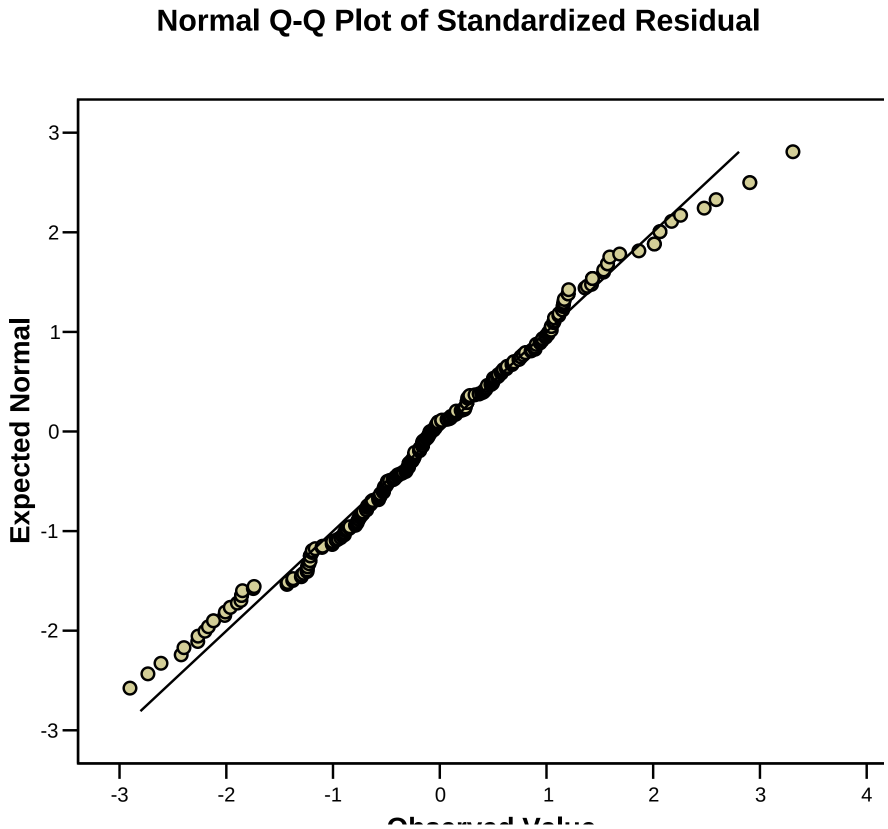
Spread vs. Level with Levene Test
● None
○ Power estimation
○ Transformed Power: Natural log
○ Untransformed

Distribution of Residuals

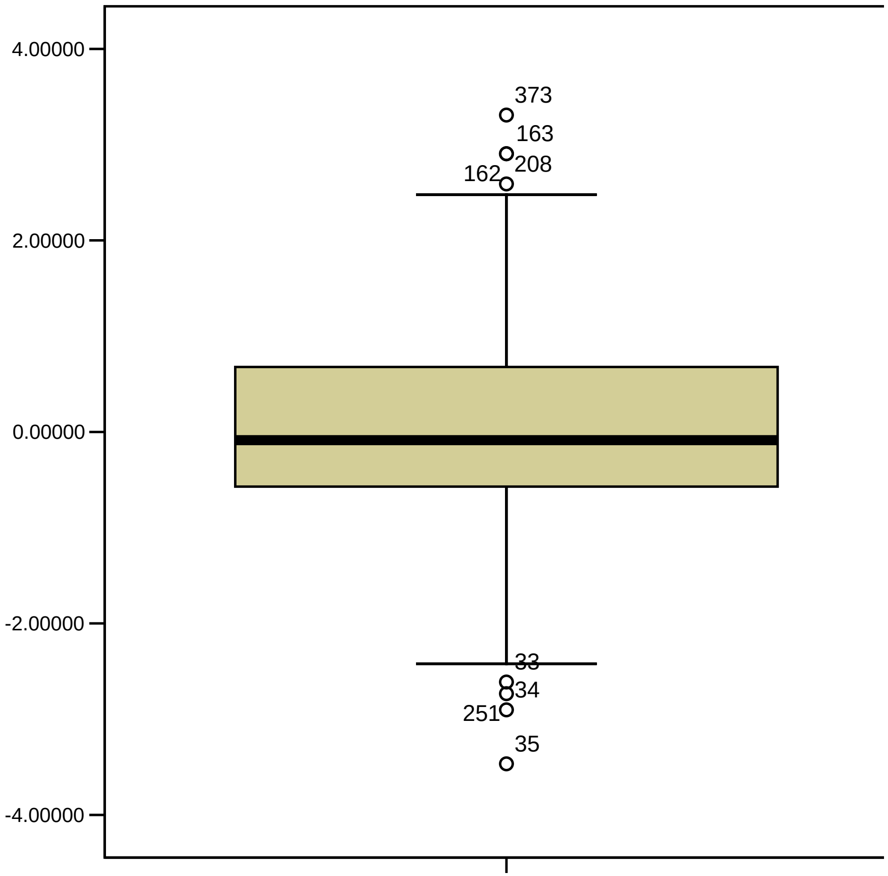


Q-Q Plot of Residuals

Expected
Value

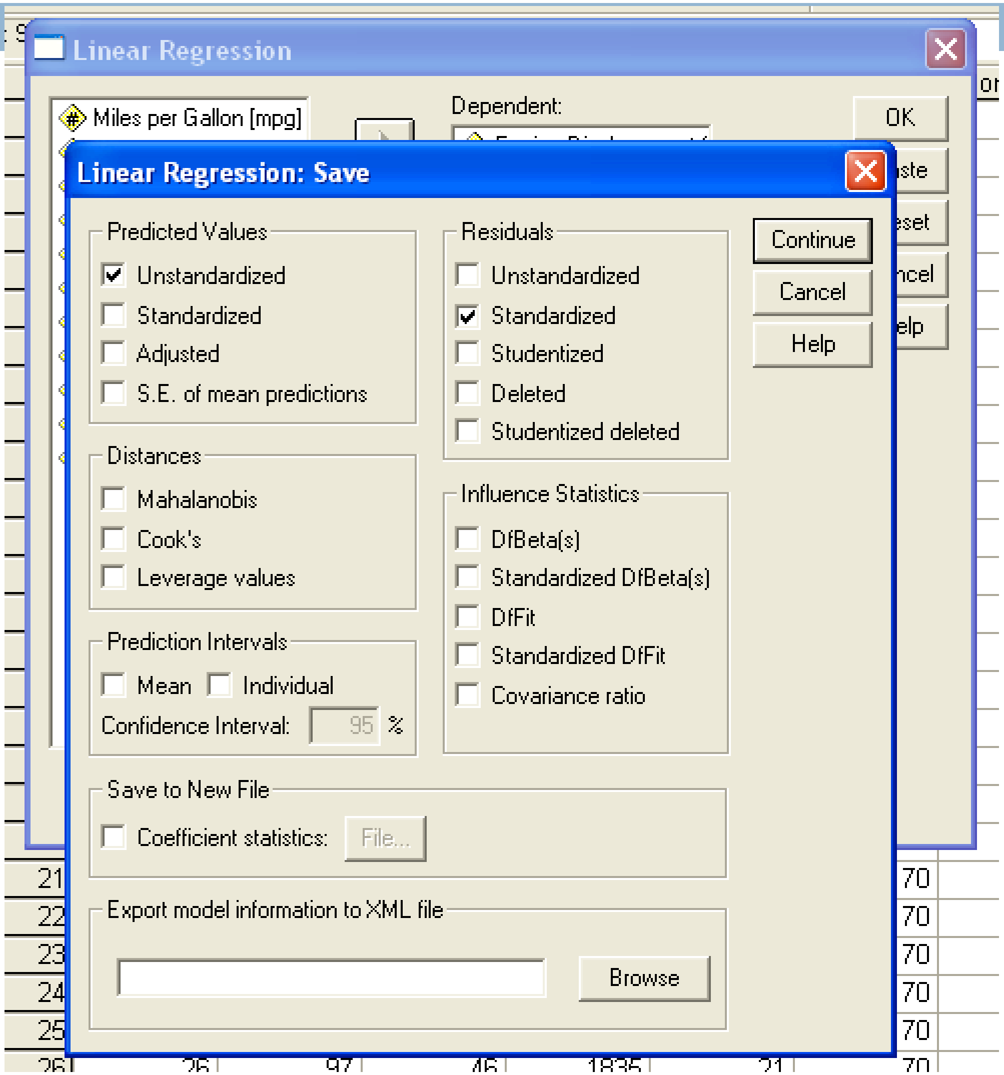


Boxplot of Residuals



Making a Residual Plot

- First, calculate the residual and the predicted value
- These will be saved in your SPSS data file, so make sure you save your data set before you close the window



Predicted Values
☑ Unstandardized
☐ Standardized
☐ Adjusted
☐ S.E. of mean predictions

Residuals
☐ Unstandardized
☑ Standardized
☐ Studentized
☐ Deleted
☐ Studentized deleted

Distances
☐ Mahalanobis
☐ Cook's
☐ Leverage values

Influence Statistics
☐ DfBeta(s)
☐ Standardized DfBeta(s)
☐ DfFit
☐ Standardized DfFit
☐ Covariance ratio

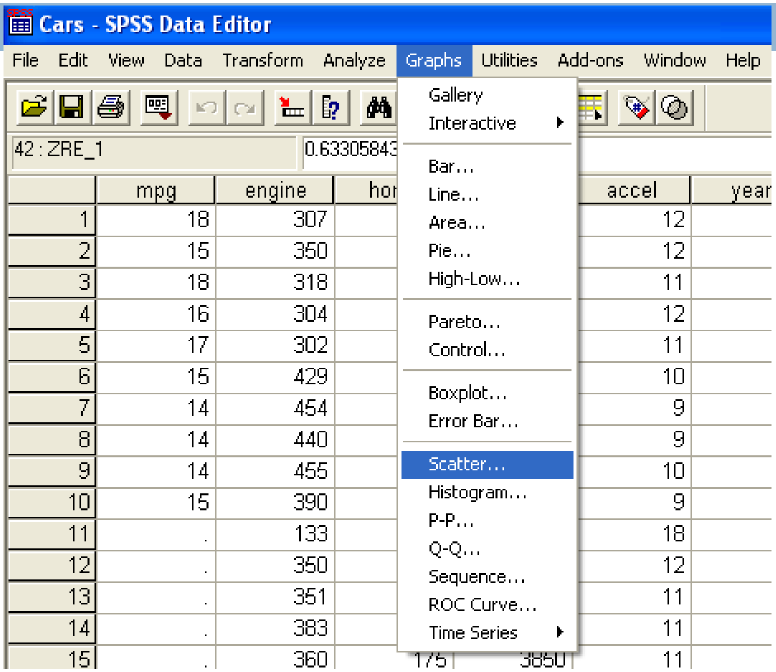
Prediction Intervals
☐ Mean ☐ Individual
Confidence Interval: 95 %

Save to New File
☐ Coefficient statistics:

Export model information to XML file

Making a Residual Plot

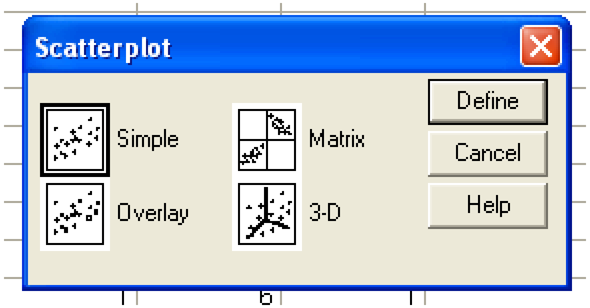
- Next, go to
  - Graphs
  - Scatterplot



42 : ZRE_1 0.63305843

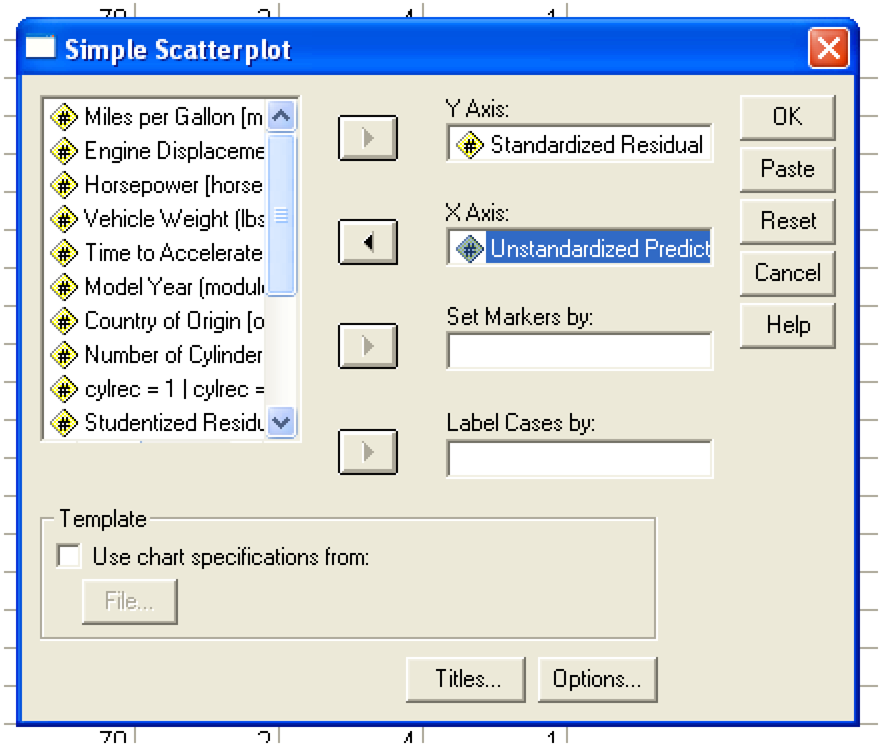
	mpg	engine	accel
1	18	307	12
2	15	350	12
3	18	318	11
4	16	304	12
5	17	302	11
6	15	429	10
7	14	454	9
8	14	440	9
9	14	455	10
10	15	390	9
11	.	133	18
12	.	350	12
13	.	351	11
14	.	383	11
15	.	360	11

  - Click on Simple then Define



Making a Residual Plot

- Move Residual to Y Axis
- Move Predicted Value to X Axis



Y Axis: Standardized Residual

X Axis: Unstandardized Predict

Set Markers by:

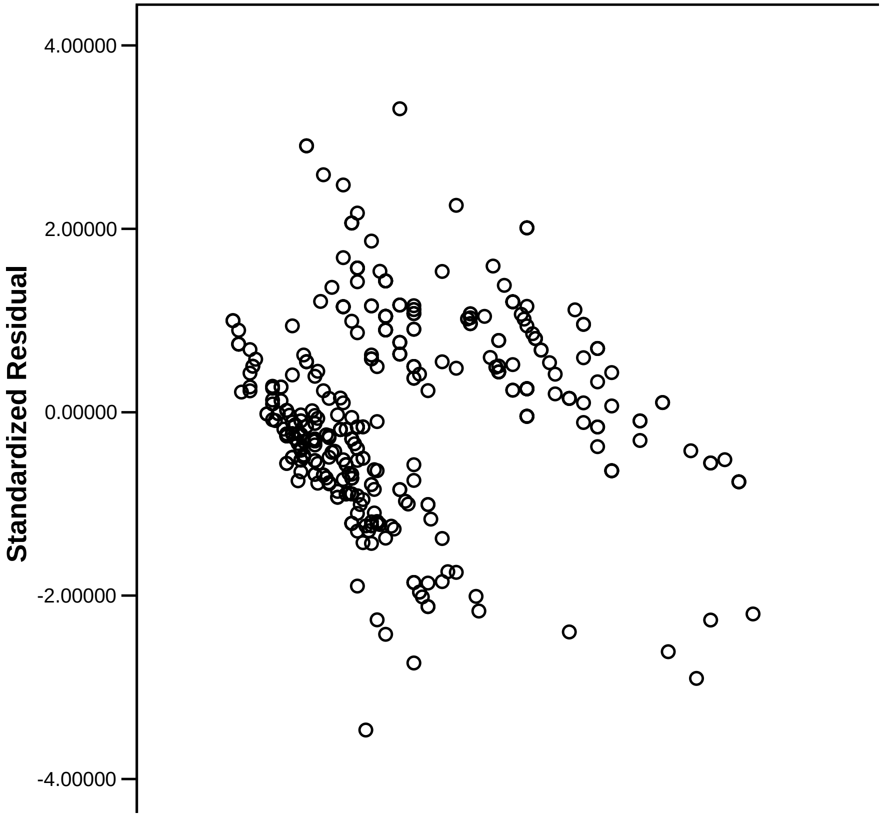
Label Cases by:

Template

☐ Use chart specifications from:

Residual Plot

Standardized
Residuals



Influence Analysis

- These statistics help us to determine the influence each observation has on the entire regression.

- Ideally, each observation should have the same influence on the regression analysis. If an observation has significantly greater influence than the rest, it can bias the results.

Some ways to determine Influence

- Leverage
  - "pull power" of the observation on the regression line
- Cook's D
  - Measures how much other residuals would change if observation was excluded from analysis
- DFBETA
  - Calculates the change in Beta if observation was excluded from analysis
- Standardized DFBETA
  - Same as DFBETA, except these values are standardized (made to fit normal curve)

Note: Larger Values indicate more influence

Leverage

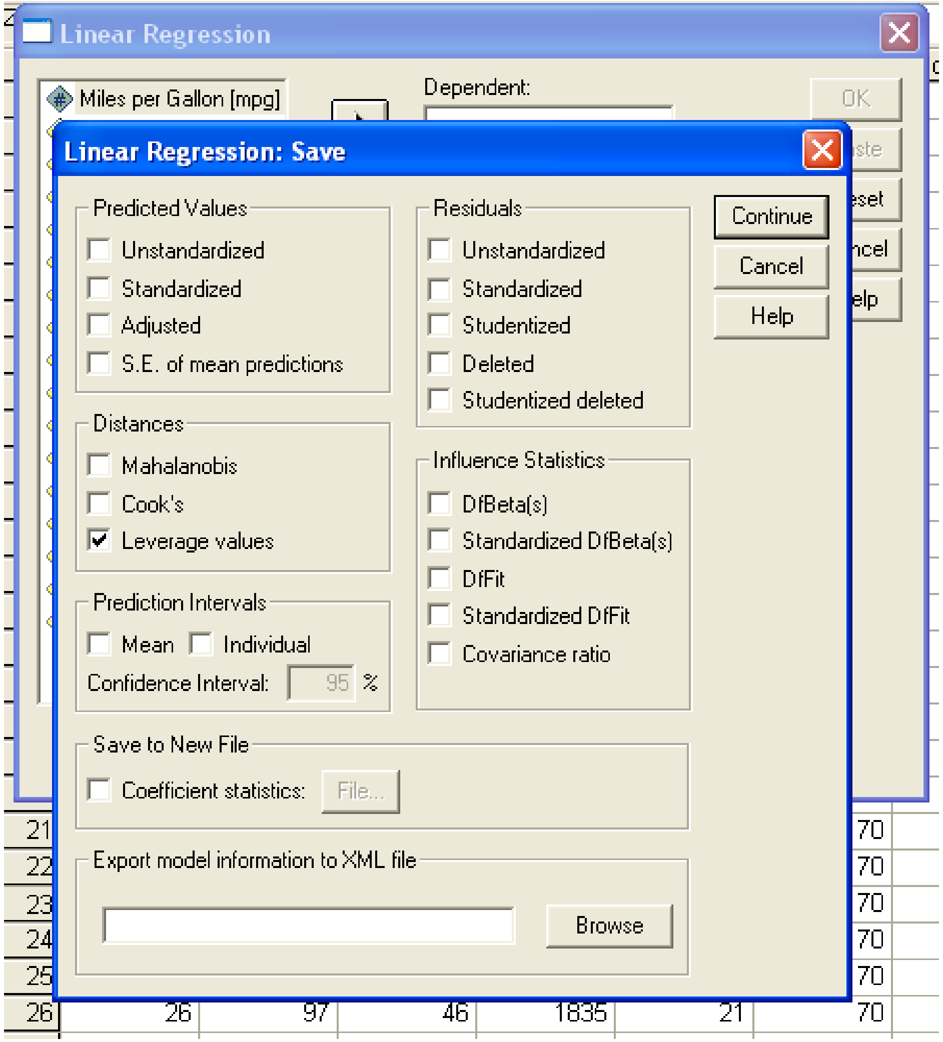
$$h_i = \frac{1}{N} + \frac{X - \bar{X}\;^{2}}{\sum x^{2}}$$

- The range of Leverage is between 1/N and 1.
- The larger the leverage, the more influence the observation has on the regression line.

Leverage

- What does it mean?
  - The larger the leverage, the more influence the single observation has on the regression
- Leverage only detects outliers as a function of the independent variable
- Rule of Thumb
  - $h_i > 2(k+1)/N$ are considered high

Leverage



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

Save to New File
☐ Coefficient statistics:

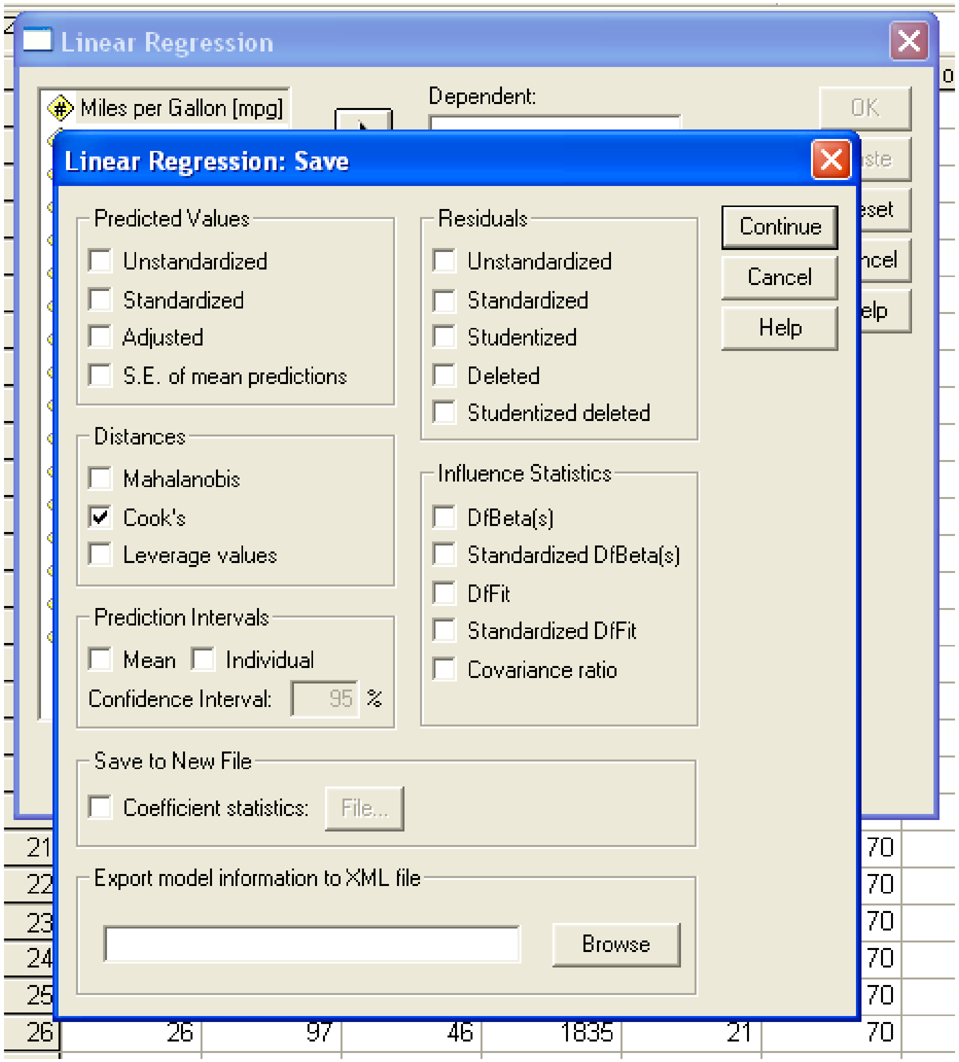
Export model information to XML file

Cook's D

$$D_i = \left[\frac{SRESID_i^2}{k+1} \right] \left[\frac{h_i}{1-h_i} \right]$$

- What does it mean?
  - Looks at the influence of an observation related to both the independent and dependent variables
- Look for large values as compared to the other observations

Cook’s D



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

Save to New File
- ☐ Coefficient statistics:

Export model information to XML file

DFBETA

Formula for a (intercept):

$$DFBETA_{a_i} = a - a_{(i)} = \left[\left(\frac{\sum X^2}{N\sum X^2 - \left(\sum X\right)^2} \right) + \left(\frac{-\sum X}{N\sum X^2 - \left(\sum X\right)^2} \right) X_i \right] \frac{e_i}{1 - h_i}$$

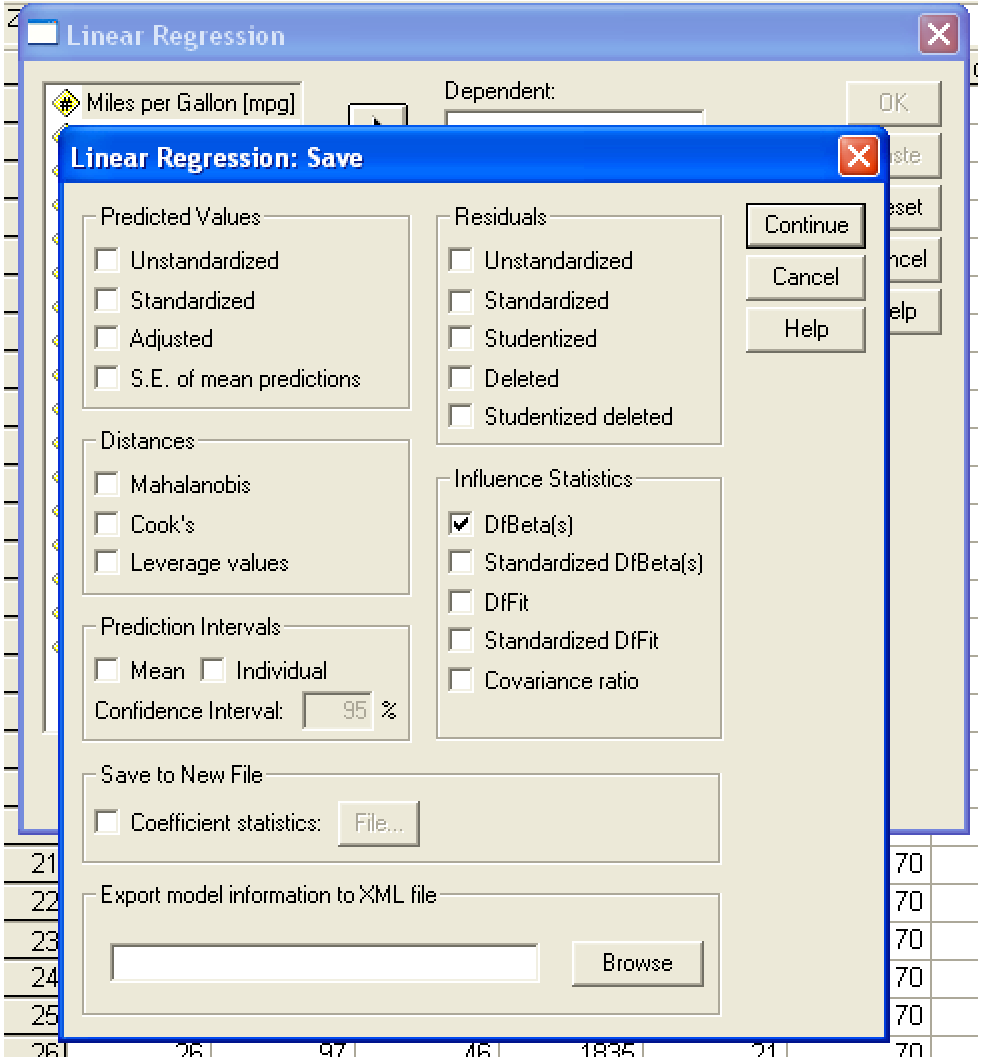
Formula for b (slope):

$$DFBETA_{b_i} = b - b_{(i)} = \left[\left(\frac{-\sum X}{N\sum X^2 - \left(\sum X\right)^2} \right) + \left(\frac{N}{N\sum X^2 - \left(\sum X\right)^2} \right) X_i \right] \frac{e_i}{1 - h_i}$$

DFBETA

- What does it mean?
  - This looks at the change in either the slope (b) or the intercept (a) when the individual value is removed
- Larger values indicate that the observation plays a large role in calculation of the regression equation (outlier)
- Problem: how large is large?

DFBETA



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

Save to New File
- ☐ Coefficient statistics: File...

Export model information to XML file

Standardized DFBETA

Formula for a (intercept):

$$DFBETAS_{a\,i} = \frac{DFBETA_{a\,i}}{\sqrt{MSR_{i}\left[\frac{\sum X^{2}}{N\sum X^{2} - \sum X^{2}}\right]}}$$

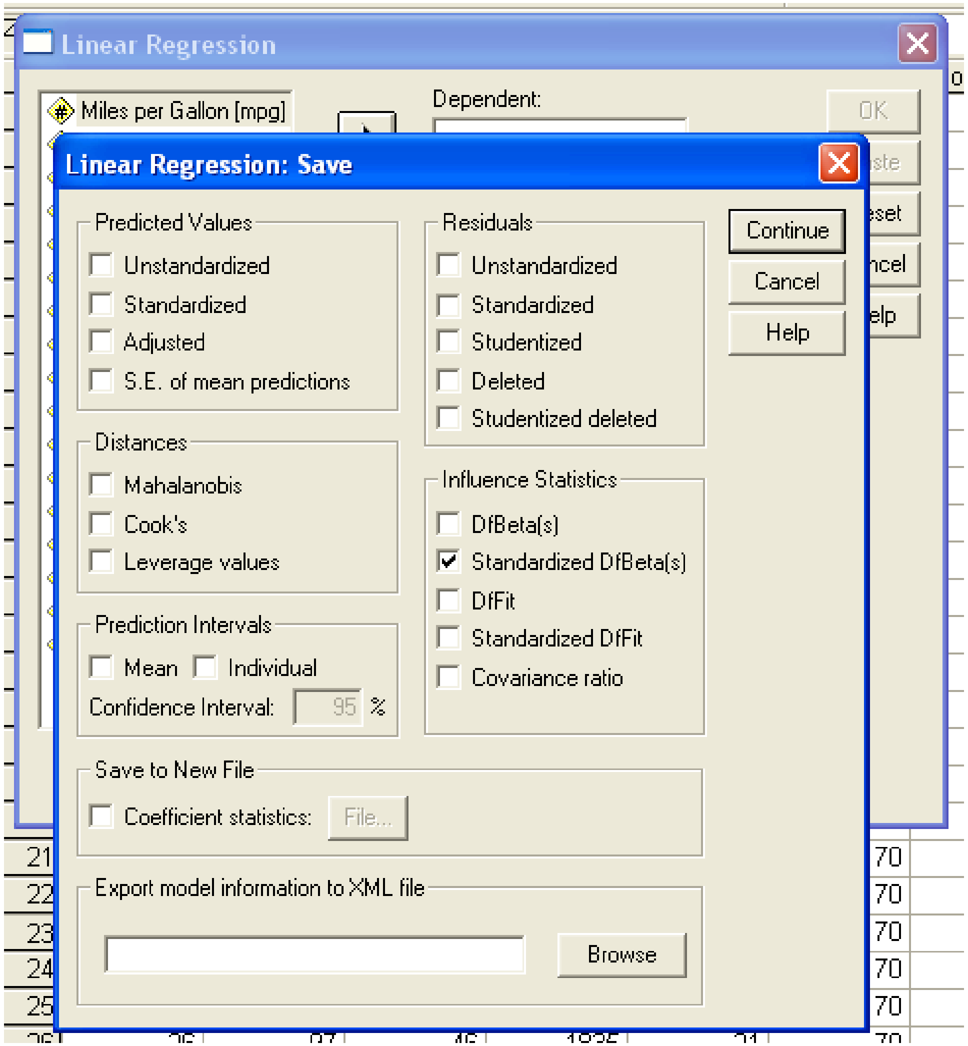
Formula for b (slope):

$$DFBETAS_{b\,i} = \frac{DFBETA_{b\,i}}{\sqrt{MSR_{i}\left[\frac{N}{N\sum X^{2} - \sum X^{2}}\right]}}$$

Standardized DFBETA

- We can standardize the DFBETA, this will allow us to easily determine how large is too large.
- Standardization places the values on a normal distribution

Standardized DFBETA



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

Save to New File
☐ Coefficient statistics:

Export model information to XML file

How does it all add up?

- First, you can look at the residuals to indicate which values are potential outliers.
- Next, examine leverage and Cook's D to determine if they have any pull on the regression line.
- Lastly, investigate the values of DFBETA and DFBETAS to see if the parameters change significantly.

When to get rid of outliers?

- You want to avoid getting rid of any data.
- If you find that there is a value that you cannot account for by measurement error alone and has large values of all the statistics we talked about today, you may want to delete the observation.
- The observation will throw off all your analysis otherwise.
- Just make sure you document the deletion and see if you can determine *why* this observation was irregular.

Short example of these values

- Regression equation

  $Y = -61.44 + 2.449*X$

1)	ZRESID	SRESID	SDRESID
	5.98	6.01	6.29
2)	Leverage	Cook’s D	
	.0093	.21622	
3)	DFBETA	DFBETAS	
	a:4.61	a:.68	
	b:-.037	b:-.61	

Final Thoughts

- Regression analyses rest on a set of assumptions.
  - These assumptions must be met for the results of the analysis to have validity.
- Regression diagnostics are one way to check the assumptions of the analysis.
- Make sure you look for outliers.
- Don’t spend too much time on it, but it can often help you find input errors that you wouldn’t otherwise have noticed.