Simple Linear Regression

Lecture 1

August 28, 2007

Applied Correlation and Regression Methods in Education

Today’s Lecture

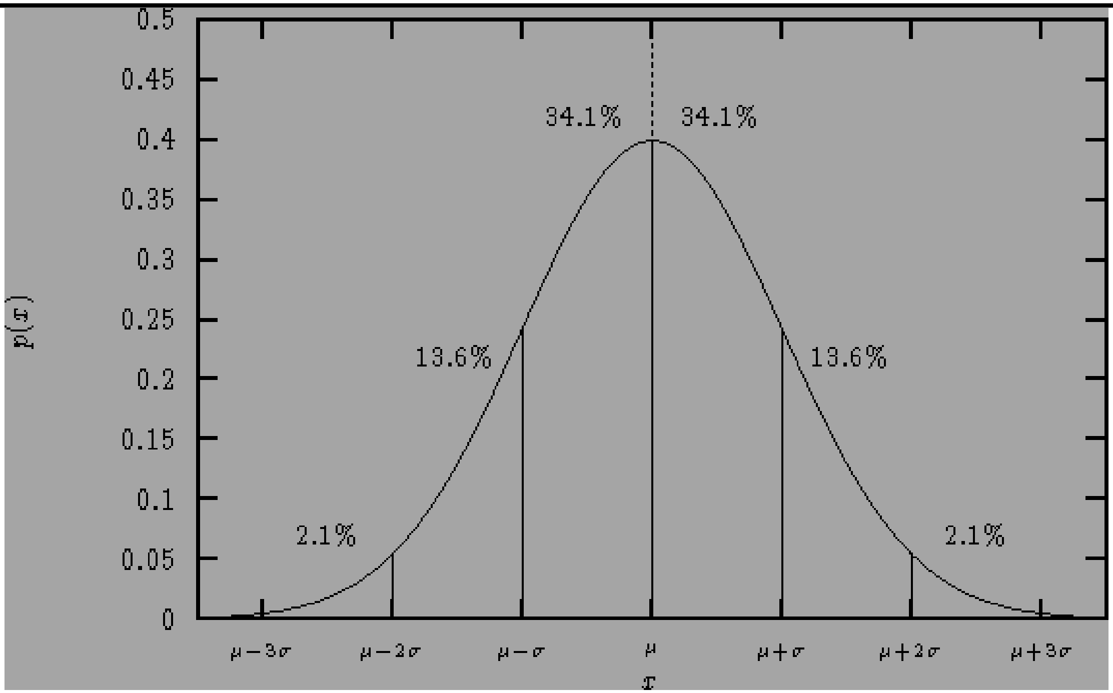
- Brief review of basic concepts:
  - Univariate statistical terminology.
  - Bivariate statistical terminology.
  - Simple linear regression.
- Simple linear regression.
- Partitioning the sum of squares.
- Tests of significance.
- Assumptions.
- Regression diagnostics (introduction).
- When X is random.
- How to do today’s topics in SPSS.

Univariate statistics

A single random variable can be described by several characteristics of it’s distribution:

- Measures of central tendency (or location).
  - Mean (a.k.a. expected value).
  - Median (point where 50% of distribution falls above or below)
  - Mode (most likely value)
- Measures of spread.
  - Variance and/or standard deviation.
  - Range(s).
- Measures that most normal people don’t think about.
  - Skewness - a measure of the symmetry of a distribution.
  - Kurtosis - a measure of the peakedness of a distribution.

Our Two Favorites



The mean:

$$\bar{X} = \sum_{i=1}^{N} \frac{X_i}{N}$$

The variance:

$$s_x^2 = \frac{\sum_{i=1}^{N}(X - \bar{X})^2}{N-1} = \frac{\sum_{i=1}^{N} X^2 - \frac{(\Sigma_{i=1}^{N} X)^2}{N}}{N-1}$$

Additional Notes

- Units associated with variance are squared (e.g., a variance of a distribution of height is given in $feet^2$).
- More understandable is something where units aren't squared: standard deviation:

$$s_x = \sqrt{s_x^2}$$

- We divide the sum of squares by $N - 1$ to obtain an unbiased estimate of the variance, or by N to obtain a "maximum likelihood" estimate of the variance (if a variable is normally distributed).
- Variance/Standard Deviation are positive numbers $[0, \infty)$.

Bivariate Distributions

Overview

Basic Statistics
- Univariate Stats
- Bivariate Statistics
- Putting It All Together

Simple Linear Regression

Partitioning the Sum of Squares

Tests of Significance

Assumptions

Diagnostics

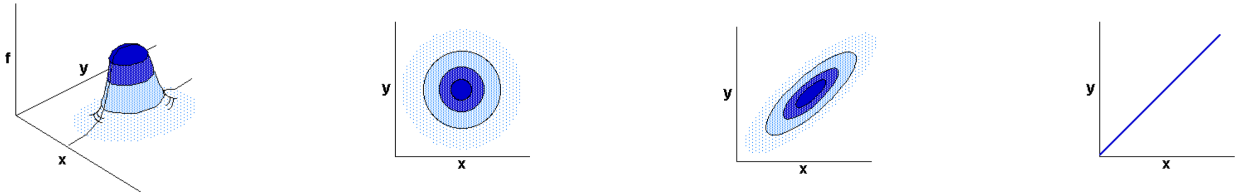
Random X

Full Example

SPSS

Wrapping Up

Each of these images shows the distribution of a pair of variables:



Bivariate Descriptive Statistics

Overview

Basic Statistics
- Univariate Stats
- Bivariate Statistics
- Putting It All Together

Simple Linear Regression

Partitioning the Sum of Squares

Tests of Significance

Assumptions

Diagnostics

Random X

Full Example

SPSS

Wrapping Up

- Covariance: the joint covariation of two sets of variables from their respective means.

$$s_{xy} = \frac{\sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})}{N-1} = \frac{\sum_{i=1}^{N} XY - \frac{(\Sigma_{i=1}^{N} X)(\Sigma_{i=1}^{N} Y)}{N}}{N-1}$$

  - Units are products of units used to create measure (e.g., the covariance of a distribution of height and weight might be reported in *foot-pounds*).
  - Covariance of a variable and itself is the variance.
  - Covariance can take the value of any real number $(-\infty, \infty)$.

Bivariate Descriptive Statistics

- Because products of units can be hard to interpret, consider the correlation:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

  - Correlation is “unitless.”
  - Range of correlation is from $[-1, 1]$.

A Single Hand Calculated Example

Consider some scores of students on Algebra and Geometry tests:

Person	Algebra (X)	Geometry (Y)	X^2	Y^2	XY
A	11	11	121	121	121
B	13	10	169	100	130
C	18	17	324	289	306
D	12	13	144	169	156
E	16	14	256	196	224
N=5	70	65	1014	875	937

$\sum X = 70$ $\quad$ $\sum Y = 65$ $\quad$ $\sum XY = 937$

$\sum X^2 = 1014$ $\quad$ $\sum Y^2 = 875$

$\bar{X} = 14$ $\quad$ $\bar{Y} = 13$

$s_X^2 = 8.5$ $\quad$ $s_Y^2 = 7.5$

$s_{XY} = 6.75$ $\quad$ $r_{XY} = 0.845$

A Single Hand Calculated Example

Overview

Basic Statistics
- Univariate Stats
- Bivariate Statistics
- Putting It All Together

Simple Linear Regression

Partitioning the Sum of Squares

Tests of Significance

Assumptions

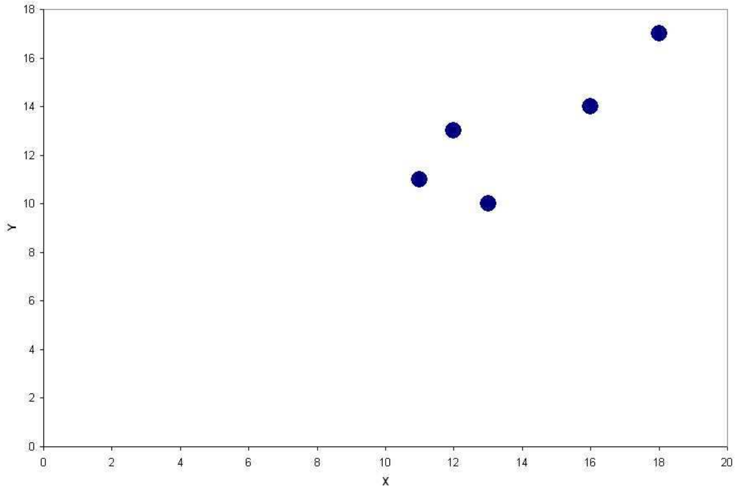
Diagnostics

Random X

Full Example

SPSS

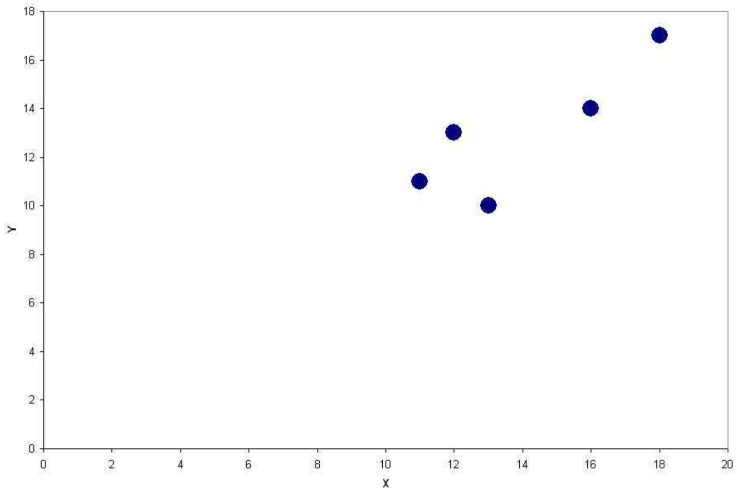
Wrapping Up



$$r_{XY} = 0.845$$

Previously...

Recall the hand calculated sample problem with the following scatterplot:



$$r_{XY} = 0.845$$

- Correlation shows the degree of association between X and Y...
    - ...but does so without scale.
- What if we wanted to predict Y given a value of X?

The Basics

- Assume (for now) X is fixed at pre-determined levels in an experiment - independent variable.
    - For example, we have an experiment where subjects are given X cups of coffee.
    - Jon the Mad Scientist decides that subjects should be randomly assigned to a group drinking either 1, 2, 3, 4,or 5 cups of coffee.
- The we want to estimate the *linear* effect of the independent variable X on the dependent variable Y.
    - For our example, we want to see how coffee drinking affects blood pressure.
    - Blood pressure = Y = dependent variable.

The Basics

- The linear regression model (for observation $i = 1, \ldots, N$):

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

- Don't be confused by the Greek alphabet, this is simply the equation for a line ($y = a + bx$).
- α is the mean of the population when X is zero...the Y intercept.
- β is the slope of the line, the amount of increase in Y brought about by a unit increase ($X' = X + 1$) in X.
- ϵ_i is the random error, specific to each observation.

Parameter Estimates

- Chapter 2 parameterizes the linear regression model as:

$$Y = a + bX + e$$

- To find estimates for a and b consider several possible choices:
  - So that $\sum\limits_{i}^{N} |e|$ is minimized.
  - So that $\sum\limits_{i}^{N} e^2$ is minimized.
  - From some guy in the hallway.

And The Winner Is...

- Finding a and b that minimize:

$$\sum_{i}^{N} e^2$$

- Using calculus, these happen to be:

$$b = \frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} = r_{xy}\frac{s_y}{s_x} = \frac{\sum xy}{\sum x^2}$$

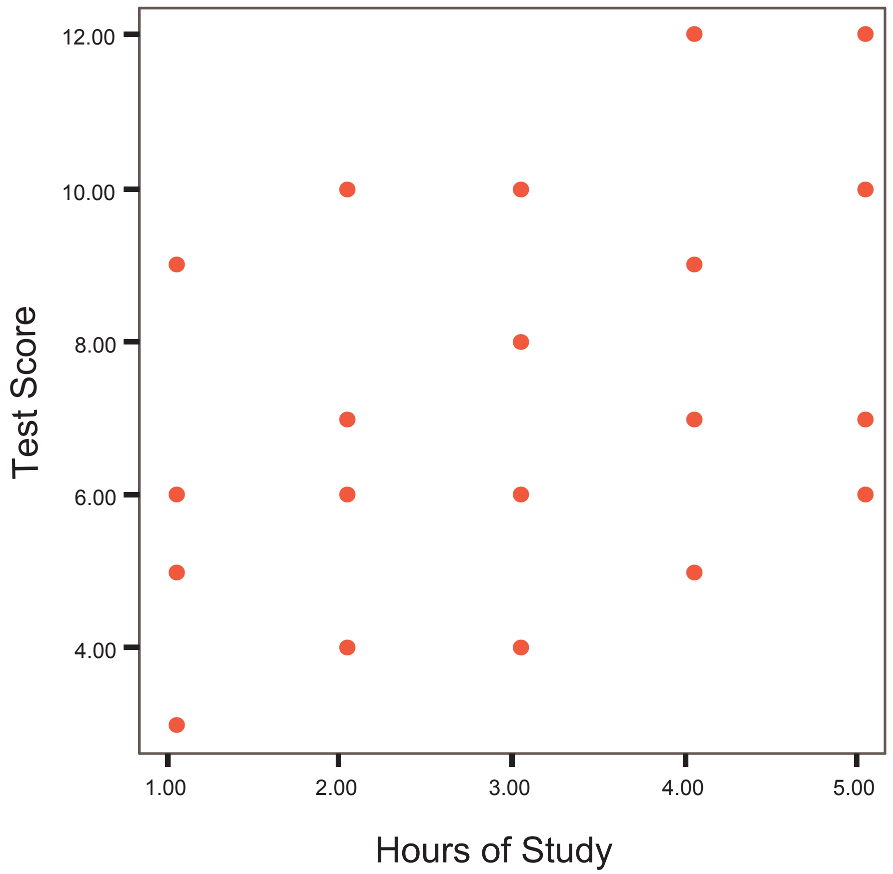
- And:

$$a = \bar{Y} - b\bar{X}$$

- LS estimates are considered BLUE: Best Linear Unbiased Estimators.

An Example of Simple Linear Regression

- Table 2.1 in Pedhazur lists data from an experiment where X was the number of hours given for study, and Y is the score on a test.



Example (continued)

- We can tell that:
  - $\sum(X-\bar{X})^2 = \sum x^2 = 40$
  - $\sum(X-\bar{X})(Y-\bar{Y}) = \sum xy = 30$
  - $\bar{X} = 3.0$
  - $\bar{Y} = 7.3$
- So:

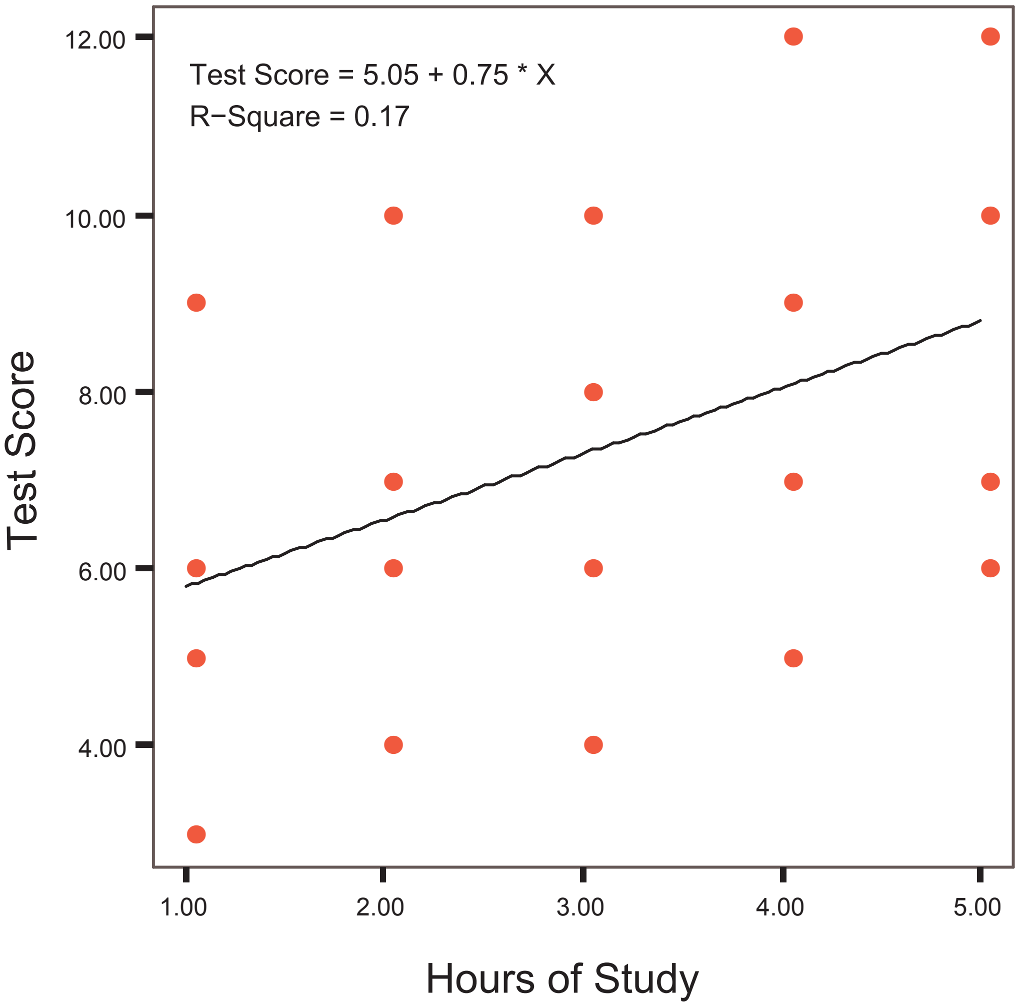
$$b = \frac{\sum xy}{\sum x^2} = \frac{30}{40} = 0.75$$

$$a = \bar{Y} - b\bar{X} = 7.3 - (0.75 \times 3.0) = 5.05$$

- Given these estimates, the linear regression line is given by:

$$Y' = 5.05 + 0.75X$$

Example (continued)



Properties of the LS Estimates

- Notice:

$$Y' = a + bX = (\bar{Y} - b\bar{X}) + bX = \bar{Y} + b(X - \bar{X})$$

- This means when $X = \bar{X}$, $Y' = \bar{Y}$.
- Furthermore, this implies that for every simple linear regression, the point $(\bar{X}, \bar{Y})$ will always fall on the regression line.
- Furthermore, if $b = 0$ (no relation between X and Y), the “best” guess (in a LS sense) for a value of Y would be $\bar{Y}$.

How “Good” Is Our Line?

- You’ve fit the model...you have your estimates...now what?
- We need to determine how well Y is predicted by X.
- Specifically, how much variability in Y can be accounted for by the linear regression where X is a predictor?
- A math adventure:

$$Y = \bar{Y} + (Y' - \bar{Y}) + (Y - Y')$$

$$Y - \bar{Y} = (Y' - \bar{Y}) + (Y - Y')$$

$$\sum(Y - \bar{Y})^2 = \sum(Y' - \bar{Y})^2 + \sum(Y - Y')^2$$

$$\sum(Y - \bar{Y})^2 = ss_{reg} + ss_{res}$$

Sums of Squares

- The Sums of Squares for our dependent variable, Y, is denoted in Pedhazur by $\sum y^2$.
- As shown before, in a regression, these can be partitioned into two components:
    - Part due to the regression line: ss_{reg}
    - Part due to error: ss_{res}
- Dividing both components by $\sum y^2$ gives the proportion of sum of squares due accounted for by the regression and the proportion accounted for by error.

Example Computation

- In our studying v. test score example from before, we know that:
  - $\sum y^2 = 130.2$
  - $b = 0.75$
  - $\sum xy$= 30
- To compute ss_{reg}, one of several formulas can be used:

$$ss_{reg} = \frac{(\sum xy)^2}{\sum x^2} = b \sum xy = b^2 \sum x^2 = 0.75 \times 30 = 22.5$$

- $ss_{res} = \sum y^2 - ss_{reg} = 130.2 - 22.5 = 107.7$

Example Continued

- Dividing both by $\sum y^2$ gives:
  - Proportion explained by regression: 0.173
  - Proportion explained by error: 0.827

Variance Accounted For

- Instead of approaching fit via Sums of Squares, one could look at it as a spinoff of the Pearson correlation:

$$r_{xy}^2 = \frac{(\sum xy)^2}{\sum x^2 \sum y^2}$$

$$ss_{reg} = r_{xy}^2 \sum y^2$$

Variance Accounted For - Example

- From our studying v. test score example from before, we know that:

$$r_{xy} = 0.416$$

- So, $r^2_{xy} = 0.173$...
- Which is the percent of variance accounted for by the regression.
- Therefore, $1 - r^2_{xy} = 0.827$; the percent of variance due to error, or unexplained variance.

Statistical Significance

- Statistical significance is literally the likelihood of the null hypothesis in a hypothesis test.

$$\begin{aligned} H_0 &: \quad b = 0 \\ H_1 &: \quad b \neq 0 \end{aligned}$$

- Statistical significance is not related to the meaningfulness of the result.
    - Take $b = 0.0001$.
    - With $se(b) = 0.0000001$.
    - b is “statistically significant” but in reality, there is a negligible relationship between X and Y.

Hypothesis Tests in Simple Linear Regression

- There are several types of hypothesis tests that can be used in simple linear regression.
- Most of these tests are all related: a test for a non-zero relationship between two variables, X and Y.
- Today, we have covered three different measures of association:
    - b -ss_{reg}
    - r^2
    - r_{xy}

Hypothesis Tests in Simple Linear Regression

- Each of the following tests evolve statistically from placing distributional assumptions on the disturbance term.

$$Y = a + bX + e$$

- Typically, we say that errors follow a normal distribution:

$$e \sim N(0, \sigma_e^2)$$

- By placing distributional assumptions
  - Sampling distributions are defined.
  - Hypothesis tests are enabled.
  - Assumptions must be recognized and checked for validity.

Testing the Entire Regression

- A test of all meaningful regression parameters is as follows:

$$\begin{aligned} H_0 : &\quad b = 0 \\ H_1 : &\quad b \neq 0 \end{aligned}$$

- Because we only have one predictor variable, this is very straight forward.
- The test statistic is F-distributed, and is computed by:

$$F = \frac{\frac{ss_{reg}}{df_{reg}}}{\frac{ss_{res}}{df_{res}}} = \frac{\frac{ss_{reg}}{k}}{\frac{ss_{res}}{N-k-1}}$$

- F-table gives significance for a given Type I error rate α, OR USE THE COMPUTER.

Testing the Entire Regression - Example

- From our example, the test of the entire regression is:

$$\begin{aligned} H_0: &\quad b = 0 \\ H_1: &\quad b \neq 0 \end{aligned}$$

$$F = \frac{\frac{ss_{reg}}{df_{reg}}}{\frac{ss_{res}}{df_{res}}} = \frac{\frac{ss_{reg}}{k}}{\frac{ss_{res}}{N-k-1}} = \frac{\frac{22.5}{1}}{\frac{107.7}{18}} = 3.76$$

- Typing “=fdist(3.76,1,18)” into MS Excel gives a p-value of 0.0683.
- We could infer that there is no statistical evidence to suggest b is different from 0.
- This means there is no statistical relationship between hours of study and the score on a test.

Testing the VAC

- Because of the relationship of r^2_{xy} with ss_{reg} the previous test could be phrased as:

$$F = \frac{\frac{r^2_{xy}}{df_{reg}}}{\frac{1-r^2_{xy}}{df_{res}}} = \frac{\frac{r^2_{xy}}{k}}{\frac{1-r^2_{xy}}{N-k-1}} = \frac{\frac{0.1728}{1}}{\frac{1-0.1728}{18}} = 3.76$$

- Notice this is identical to the F we received before. Therefore our conclusions are the same...

Testing the Regression Coefficient

- Finally, the regression coefficient between X and Y can tested for difference from zero:

$$H_0: \quad b = 0$$
$$H_1: \quad b \neq 0$$

- The variance of Y conditional on X is given by:

$$s^2_{y.x} = \frac{\sum(Y - Y')}{N-k-1} = \frac{ss_{res}}{N-k-1}$$

- The standard error of the regression coefficient, s_b, is based on the conditional variance of $Y|X$:

$$s_b = \sqrt{\frac{s^2_{y.x}}{\sum x^2}}$$

Testing the Regression Coefficient

- Finally, the test statistic for the following hypothesis test is:

$$\begin{aligned} H_0 &: \quad b = 0 \\ H_1 &: \quad b \neq 0 \end{aligned}$$

$$t = \frac{b}{s_b}$$

- This statistic is t distributed, with N-k-1 degrees of freedom (which is the denominator of $s^2_{y.x}$).

Testing the Regression Coefficient - Example

- From our example, the test of the regression coefficient is:

$$\begin{aligned} H_0 : &\quad b = 0 \\ H_1 : &\quad b \neq 0 \end{aligned}$$

$$s^2_{y.x} = \frac{\sum(Y - Y')}{N - k - 1} = \frac{ss_{res}}{N - k - 1} = \frac{107.7}{18} = 5.983$$

$$\sum x^2 = 40$$

$$t = \frac{b}{\sqrt{\frac{s^2_{y.x}}{\sum x^2}}} = \frac{0.75}{\sqrt{\frac{5.983}{40}}} = 1.94$$

- Typing in “=tdist(1.94,18,2)” (2 for a two-tailed test), gives a p-value of 0.0683 - the same as before.
- Recall $F = t^2$ from somewhere before...

Testing the Regression Coefficient

- In our example, we tested against zero.
- In reality, any value can be tested against:

$$\begin{aligned} H_0 &: \quad b = B \\ H_1 &: \quad b \neq B \end{aligned}$$

$$t = \frac{b - B}{\sqrt{\frac{s^2_{y.x}}{\sum x^2}}}$$

Key Assumptions Behind Simple Linear Regression

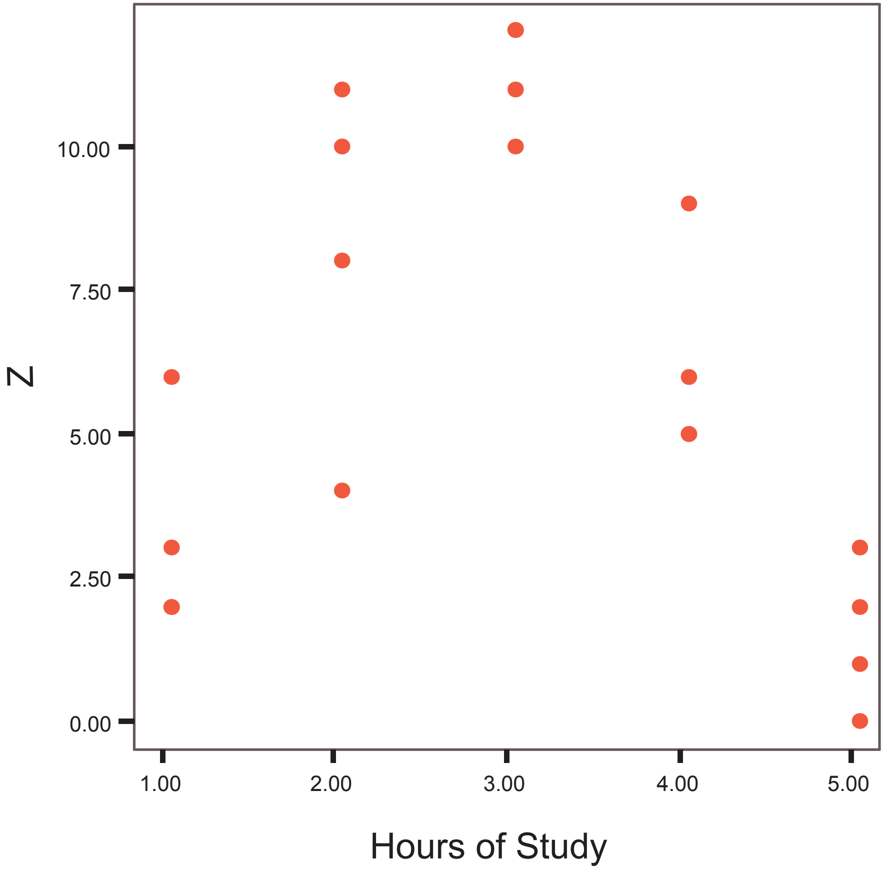
- Basic assumptions:
  - X is fixed - measured without error.
  - The mean of $Y|X$ has a linear relationship with X
- Error assumptions:
  - Mean of errors is zero.
  - All errors independent.
  - The variance of the errors does not change as a function of X - homoscedasticity.
  - All errors are uncorrelated with X.
  - Statistically speaking, errors are IID: $e \sim N(0, \sigma_e^2)$.

Diagnostics: Ways of Checking Assumptions

- An easy way to check many error based assumptions a visualization:
  - A plot of the standardized residual versus the predicted value.
- The next class will focus on other diagnostic measures.

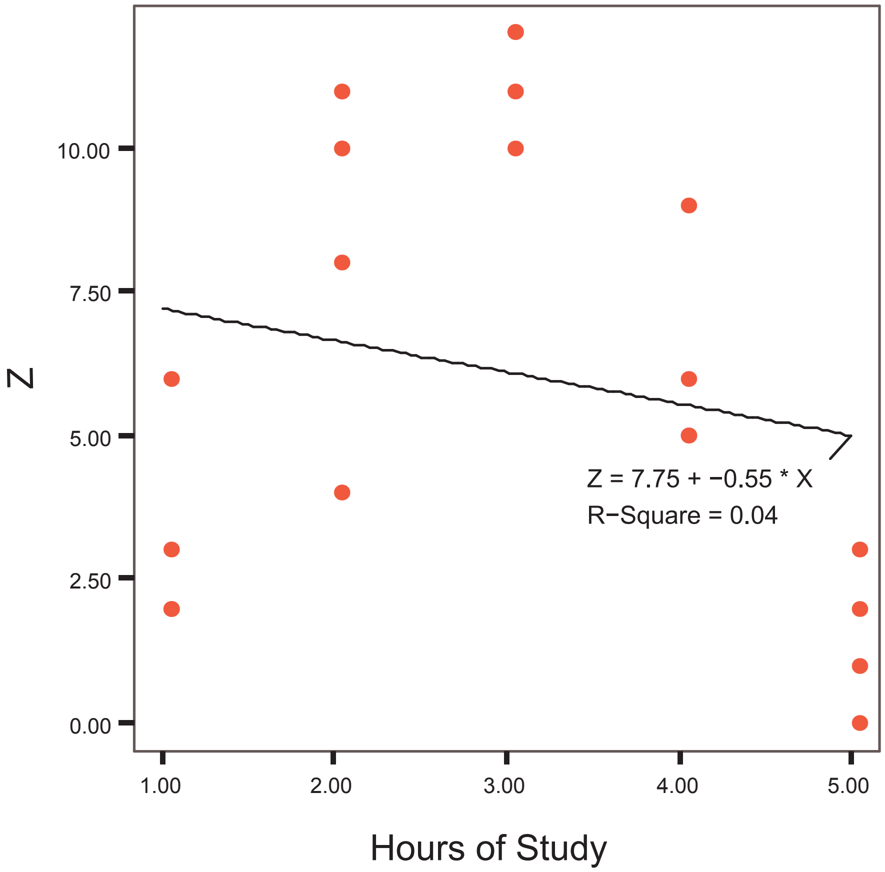
Curvilinear Example

What if we were to fit a linear regression to these data?



Curvilinear Example

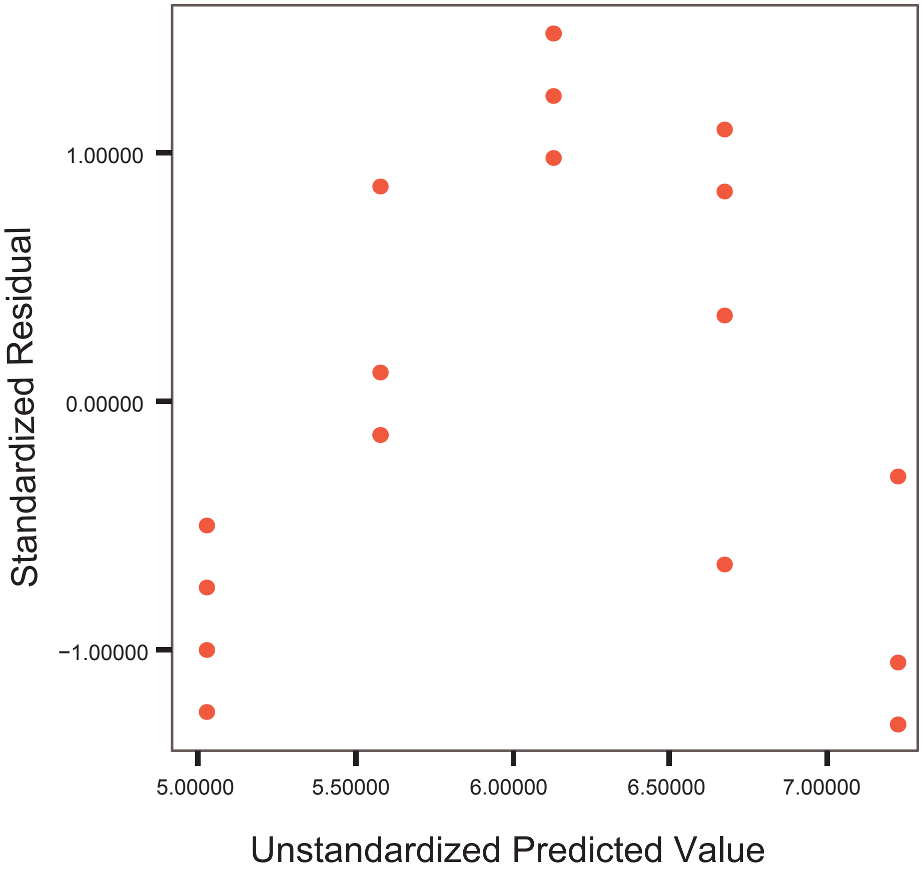
Fitting the simple linear regression model gives:



Not good?

Curvilinear Example

Let's take a look at a plot of the standardized residuals versus the predicted values:



When X is Random

- Commonly, we have no control over our predictor variable.
- Statistically speaking, having X being random adds:
  - The potential for measurement error in X - reliability.
  - Limitations to the causal inferences that can be drawn from a regression analysis.
- All test statistics and measures are used equivalently.
- We can predict Y from X or X from Y.
- Correlations are more common place: common variance rather than causal influence.

Wrap Up Example

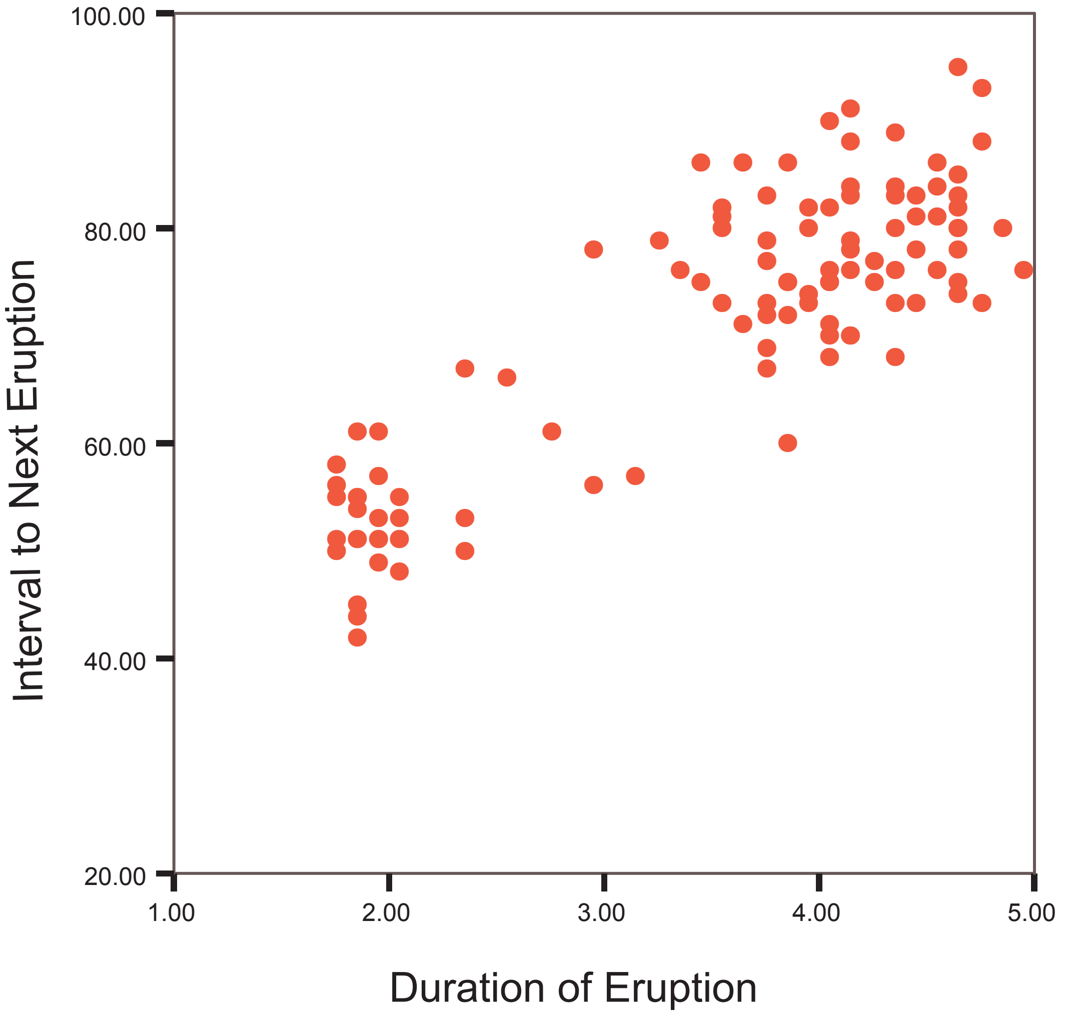
Example taken from Weisberg (1985, p. 230):

“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).”

Old Faithful Postcard

Old Faithful Scatterplot



Descriptive Statistics

Variable	N	Mean	SD
x - Duration	107	3.4607	1.0363
y - Interval	107	71.000	12.967

$$r_{xy} = 0.858$$
$$r^2_{xy} = 0.736$$
$$\sum x^2 = 113.84$$
$$\sum y^2 = 17,822.00$$
$$\sum xy = 1,222.70$$

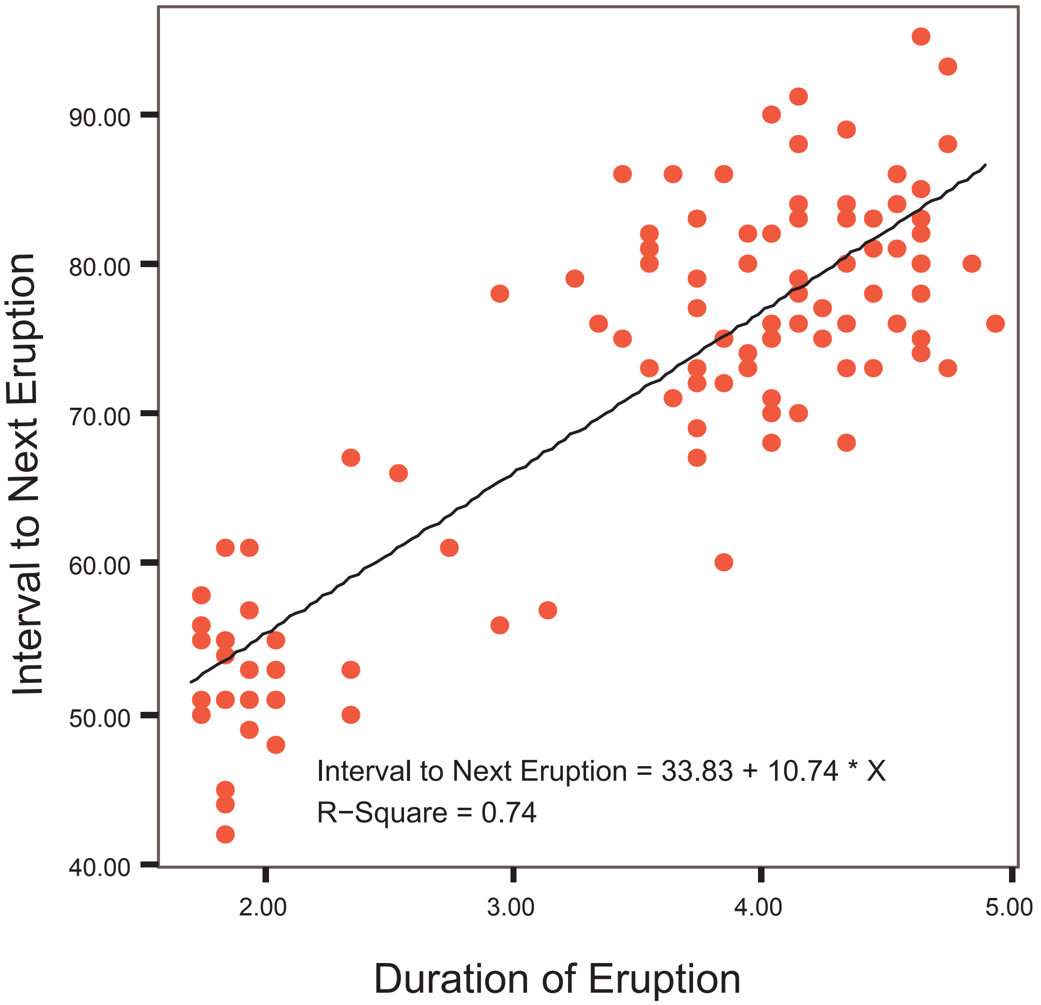
Regression Line

$$b = \frac{\sum xy}{\sum x^2} = \frac{1,222.70}{113.84} = 10.74$$

$$a = \bar{Y} - b\bar{X} = 71.0 - (10.74) \times 3.4607 = 33.84$$

$$Y' = 33.84 + 10.74X$$

Regression Line



Interval to Next Eruption = 33.83 + 10.74 * X

R−Square = 0.74

Hypothesis Tests

$$ss_{reg} = b\sum xy = 10.74 \times 1,222.70 = 13,132.99$$

$$ss_{res} = \sum y^2 - ss_{reg} = 17,822.0 - 13,132.99 = 4,689.01$$

$$F = \frac{\frac{ss_{reg}}{k}}{\frac{ss_{res}}{N-k-1}} = \frac{13,132.99/1}{4,689.01/(107-1-1)} = 294.084$$

Residual Plot

Overview

Basic Statistics

Simple Linear Regression

Partitioning the Sum of Squares

Tests of Significance

Assumptions

Diagnostics

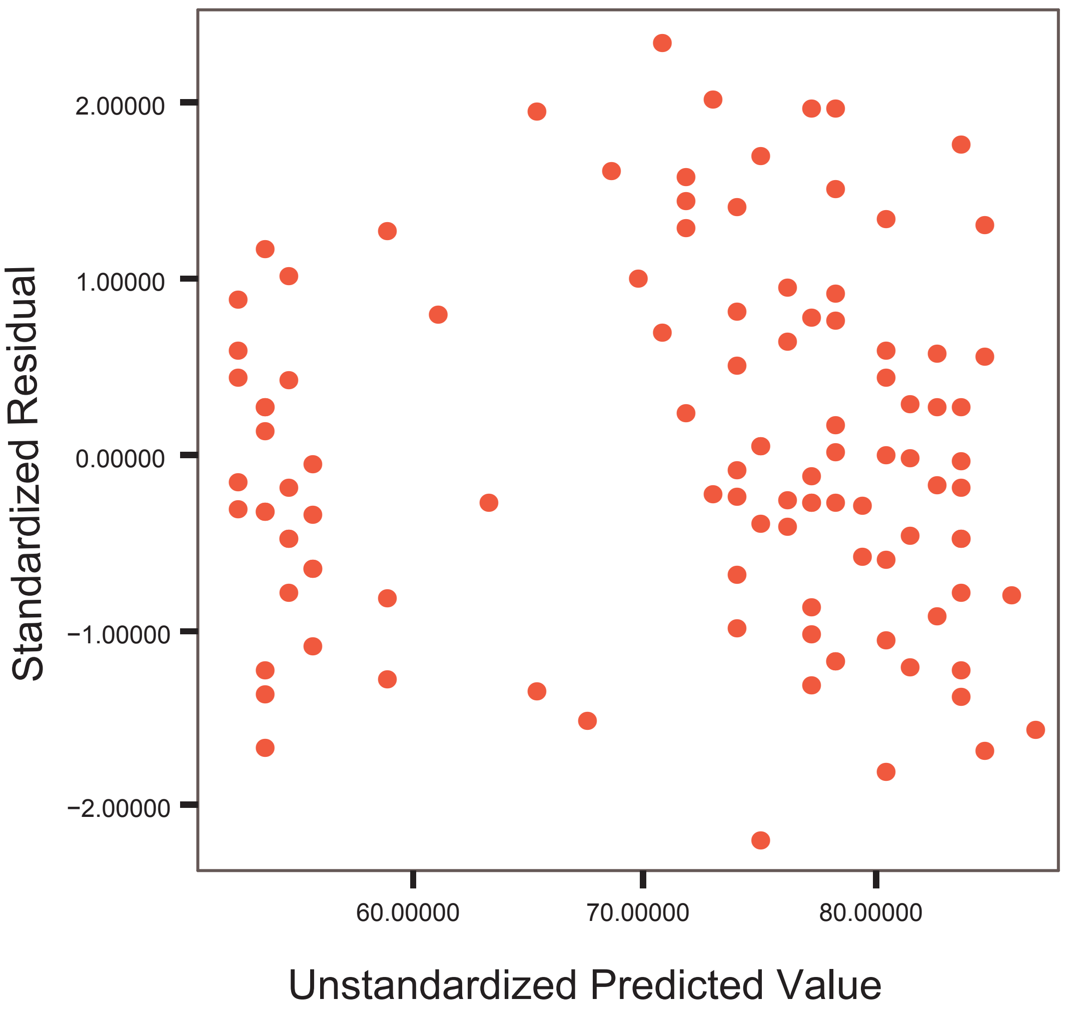
Random X

Full Example

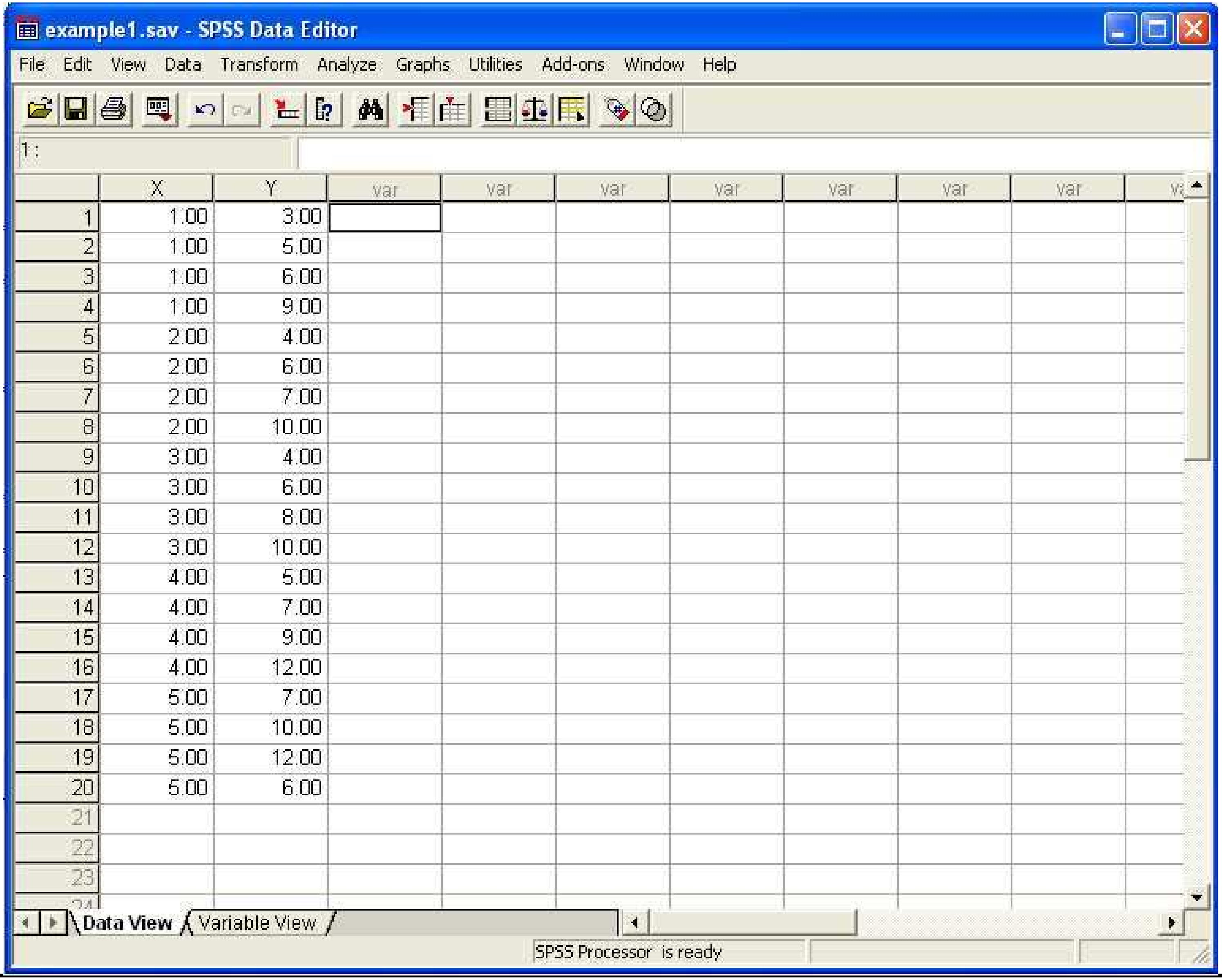
- ● Old Faithful
- ● Scatterplot
- ● Descriptives
- ● Regression Line
- ● Hypothesis Tests
- ○ Diagnostics

SPSS

Wrapping Up

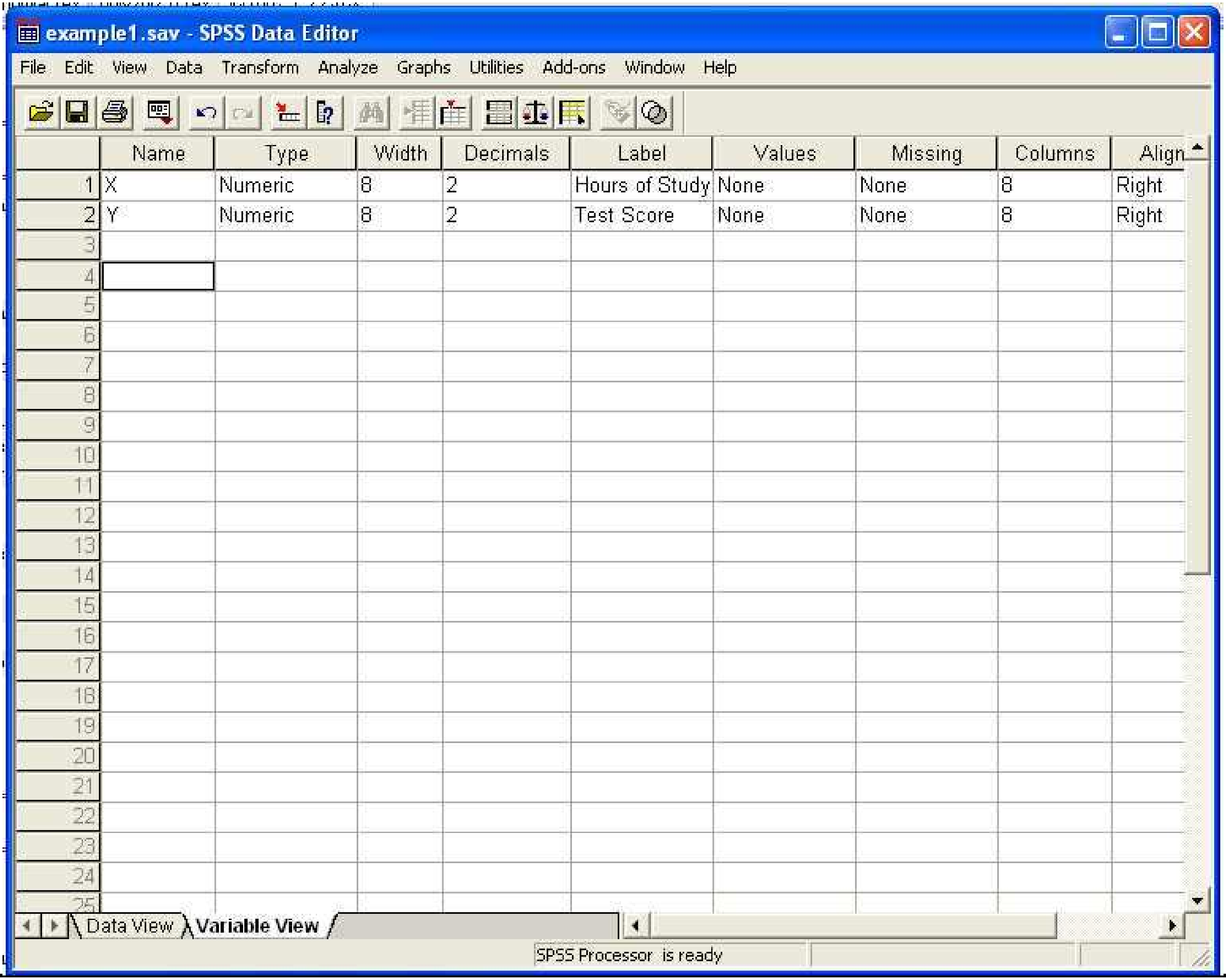


Data


example1.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
1 :
Data View
Variable View
SPSS Processor is ready

	X	Y	var	var	var	var	var	var	var
1	1.00	3.00							
2	1.00	5.00							
3	1.00	6.00							
4	1.00	9.00							
5	2.00	4.00							
6	2.00	6.00							
7	2.00	7.00							
8	2.00	10.00							
9	3.00	4.00							
10	3.00	6.00							
11	3.00	8.00							
12	3.00	10.00							
13	4.00	5.00							
14	4.00	7.00							
15	4.00	9.00							
16	4.00	12.00							
17	5.00	7.00							
18	5.00	10.00							
19	5.00	12.00							
20	5.00	6.00							
21									
22									
23									

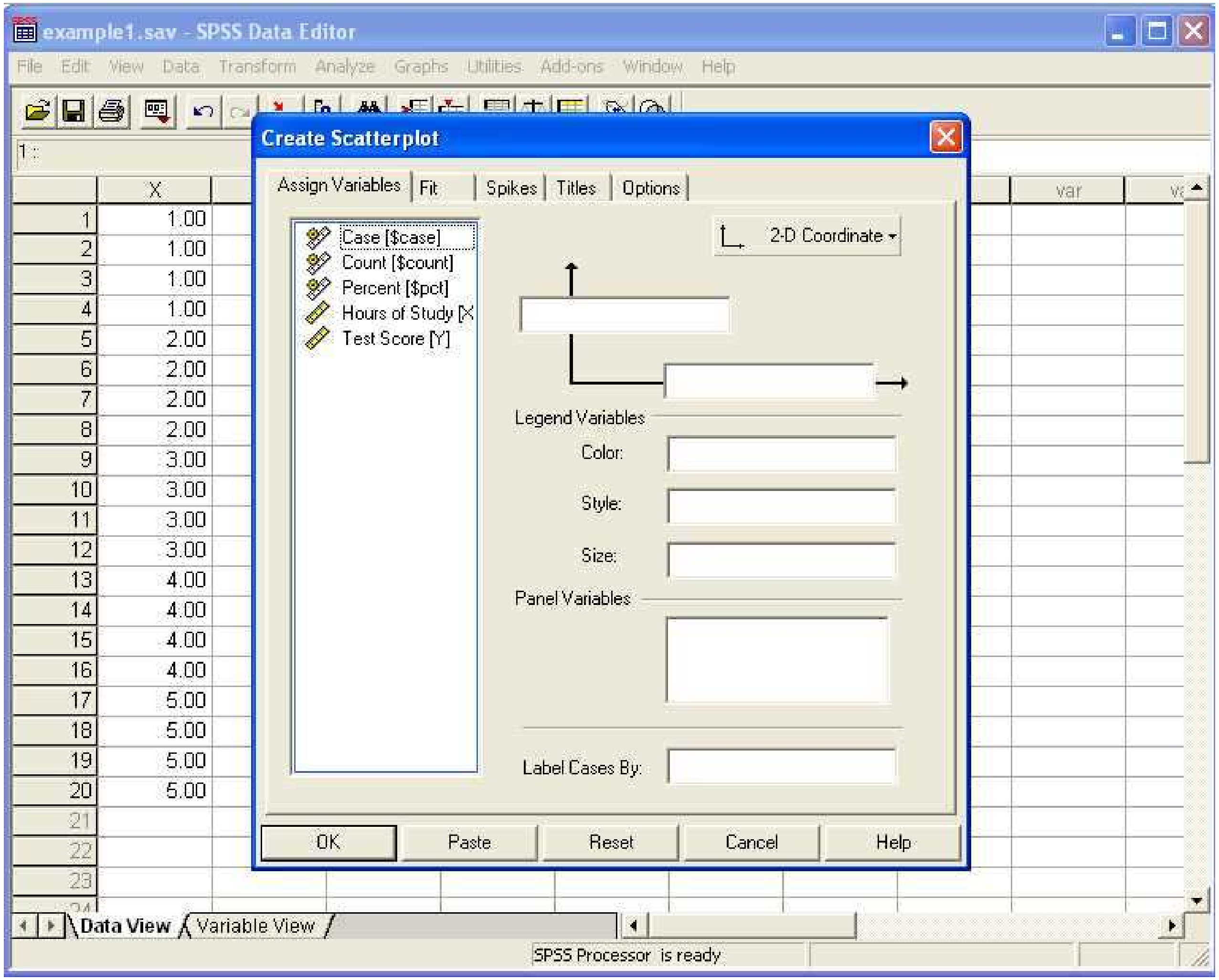
Data


example1.sav - SPSS Data Editor
File Edit View Data Transform Analyze Graphs Utilities Add-ons Window Help
Data View
Variable View
SPSS Processor is ready

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align
1	X	Numeric	8	2	Hours of Study	None	None	8	Right
2	Y	Numeric	8	2	Test Score	None	None	8	Right

Plots from Today

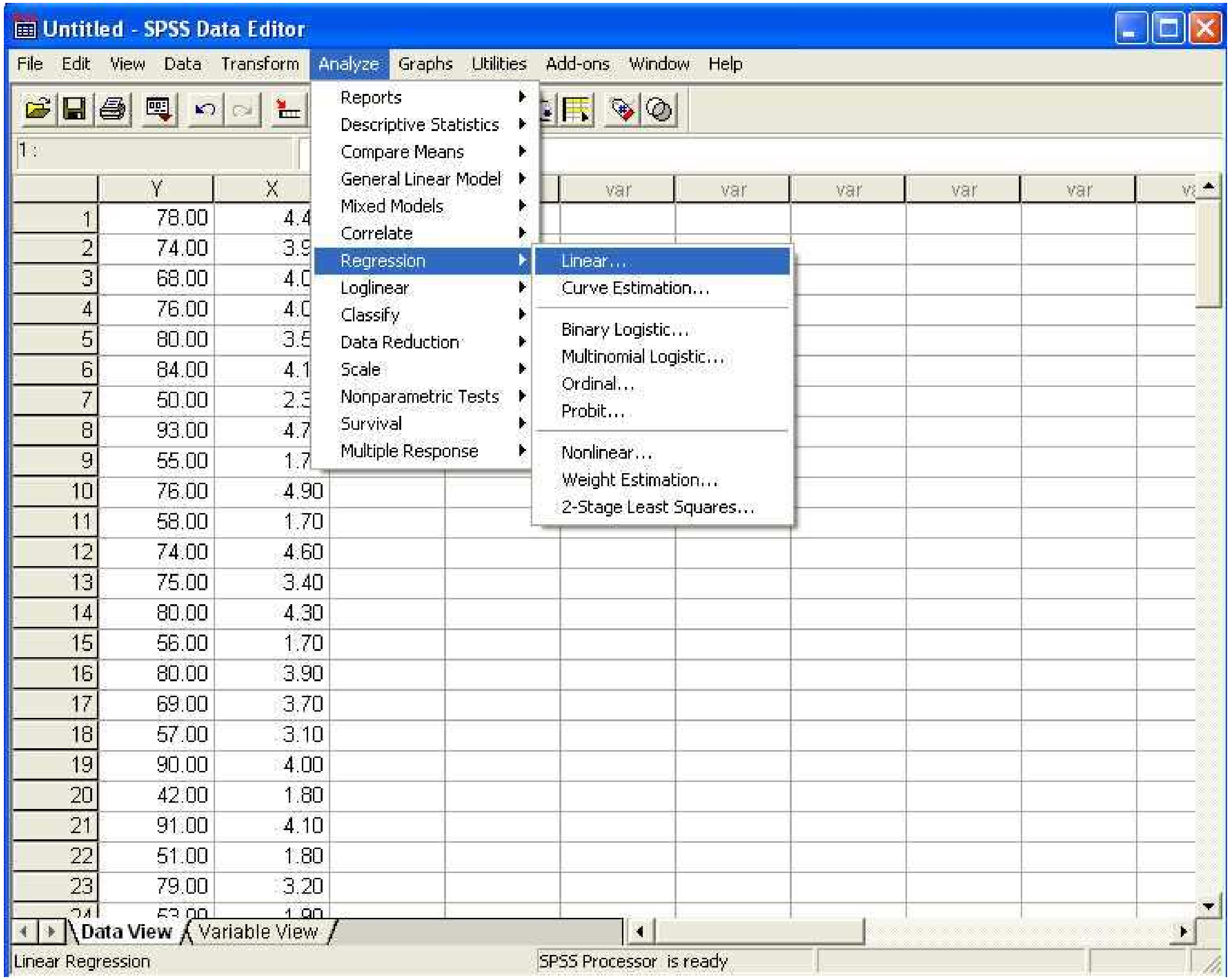
Scatterplots were made with the interactive graphs: Graphs...Interactive...Scatterplot.



	X
1	1.00
2	1.00
3	1.00
4	1.00
5	2.00
6	2.00
7	2.00
8	2.00
9	3.00
10	3.00
11	3.00
12	3.00
13	4.00
14	4.00
15	4.00
16	4.00
17	5.00
18	5.00
19	5.00
20	5.00
21	
22	
23	

Simple Linear Regression

Analyze...Regression...Linear.



	Y	X
1	78.00	4.4
2	74.00	3.9
3	68.00	4.0
4	76.00	4.0
5	80.00	3.5
6	84.00	4.1
7	50.00	2.3
8	93.00	4.7
9	55.00	1.7
10	76.00	4.90
11	58.00	1.70
12	74.00	4.60
13	75.00	3.40
14	80.00	4.30
15	56.00	1.70
16	80.00	3.90
17	69.00	3.70
18	57.00	3.10
19	90.00	4.00
20	42.00	1.80
21	91.00	4.10
22	51.00	1.80
23	79.00	3.20

Linear Regression

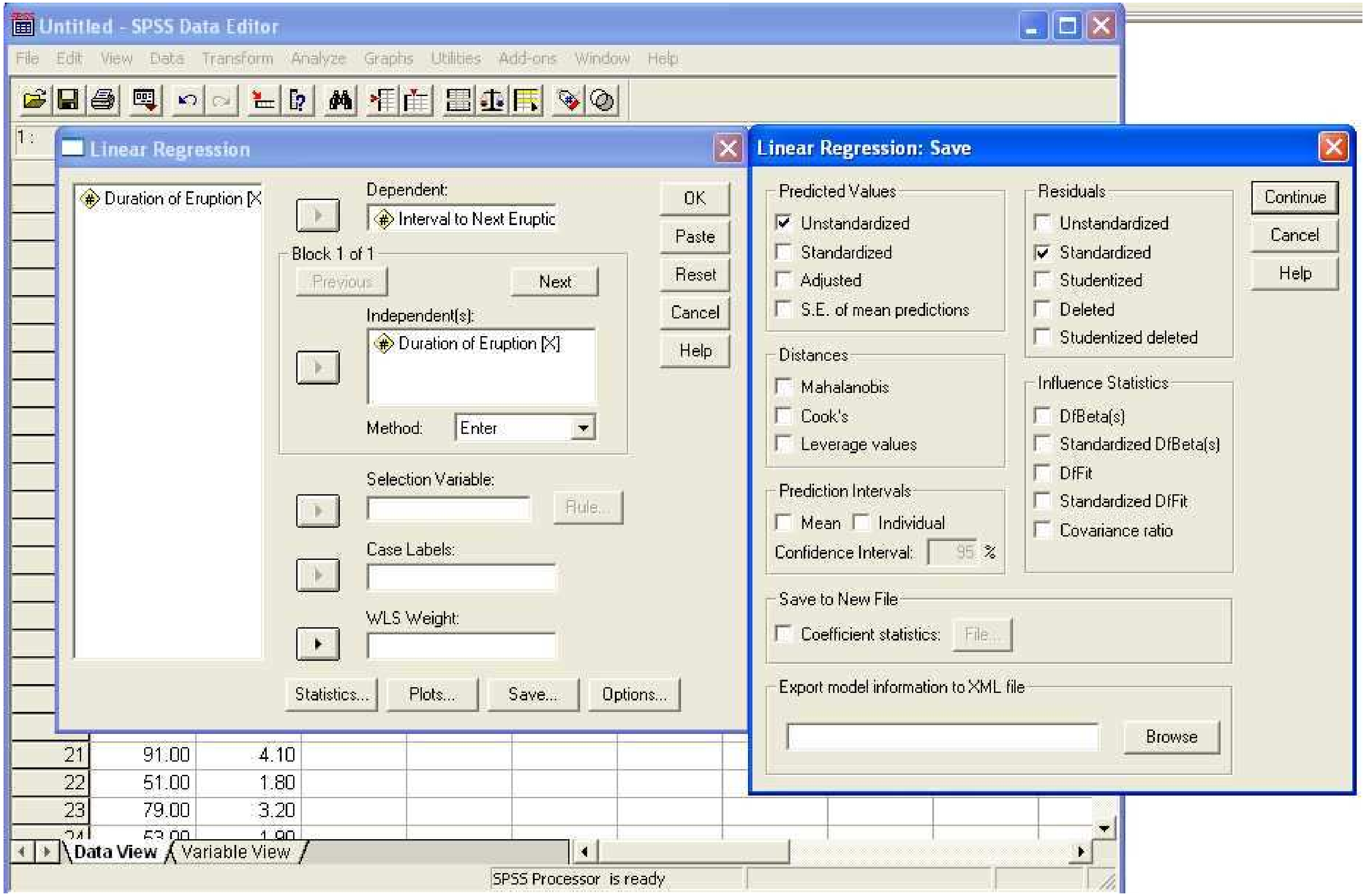
SPSS Processor is ready

Saving Residuals

Overview
Basic Statistics
Simple Linear Regression
Partitioning the Sum of Squares
Tests of Significance
Assumptions
Diagnostics
Random X
Full Example
SPSS
- Data
- More Data
- Scatterplots
- Regression
- **Saving Residuals**

Wrapping Up

Analyze...Regression...Linear → Save Button.



Duration of Eruption [X

Dependent: Interval to Next Eruptic

Independent(s): Duration of Eruption [X]

Method: Enter

Selection Variable:

Case Labels:

WLS Weight:

21	91.00	4.10
22	51.00	1.80
23	79.00	3.20
24	53.00	1.90

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

Final Thought

- Simple linear regression can be simple and straight forward.
- Need to understand difference between statistical significance and practical relevance.

- The tests of significance and model estimates are only valid if all assumptions are met.
- Regression is fairly robust to violations of assumptions, but...
- Usually when assumptions are violated, something else is wrong with the model.
- Listen to your data, they are trying to tell you something.

Next Time

- Regression diagnostics - Chapter 3.
- More SPSS.
- More practice with fitting models.
- QUESTION: Do you have any data you are willing to share with class?