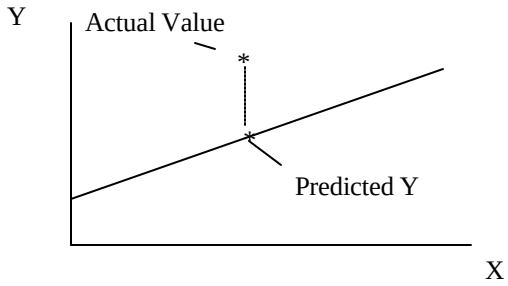
Regression Summary

ellis.godard@csun.edu

Line of "Best Fit"

The goal of regression is an equation to summarize the data with the least total error. Between our summary line and the actual data, there are errors or residuals, as illustrated by the dotted line in this diagram:



The Sum of Squared Errors (SSE) is the sum of squares of the differences between the actual value of the dependent variable Y and the expected or predicted value of Y given values of X:

$$SSE = \Sigma (Y_i - \hat{Y})^2$$

The "line of best fit" (the line given by the regression equation) is the line that minimizes SSE, that is which involves the least total squared deviations from the actual data.

For the regression equation $Y = \alpha + Bx$:

B ("beta") is the population parameter slope
(the expected increase in Y for an increase of 1 in X)
which we estimate with "b", where

$$b = \frac{\Sigma (X_i - \bar{X})(Y_i - \bar{Y})}{\Sigma (X_i - \bar{X})^2}$$

α ("alpha") is the population parameter Y-intercept
(the expected or predicted Y when X=)
which we estimate with "a", where

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

These summary statistics are given by SPSS output as follows:

VARIABLE	B	SE B	Beta	T	Sig T
IV name	"b"	se of "b"		test-stat for b	p-value
(constant)	"a"	se of "a"		test-stat for a	p-value

These are the hypotest values (t and p).

These coefficients are standardized in units of standard deviation
The standardized Beta = $(\sigma_x / \sigma_y) * b$
These are *not* the coefficients you want this week.

These are the coefficients you want for the equation $Y = a + bx$ (estimating $Y = \alpha + Bx$).

b = change, on average, in Y for each increase of 1 in X.
a = predicted value of Y when X=0 (but *don't infer beyond observed range of X)*

Null and Alternative Hypotheses

Ho: B=0 (X and Y are independent)

Ha: B = 0 (X and Y are dependent)

or B>0 (the relationship between X and Y is positive)

or B<0 (the relationship between X and Y is negative)

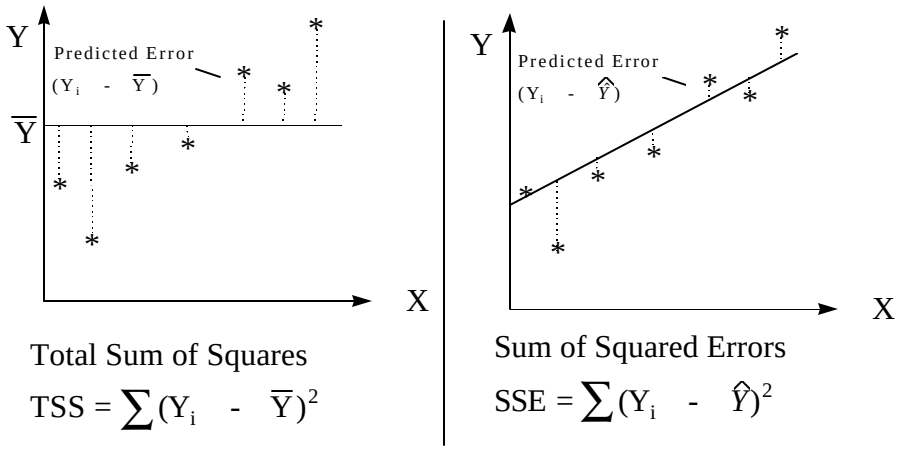
For the test statistic, $t = b / \sigma_b$ where we estimate σ_b (est. standard error of the slope) with

$$\sigma_b = \sigma \; / \; \sqrt{\Sigma (X_i - X)^2}$$

and we estimate σ (standard deviation of the conditional distribution of y) as:

$$\sigma_{y \mid x} = \sqrt{SSE / (n-2)}$$

r^2 reminders:



Total Sum of Squares

$TSS = \sum (Y_i - \overline{Y})^2$

Sum of Squared Errors

$SSE = \sum (Y_i - \hat{Y})^2$

$$r^2 = \frac{TSS - SSE}{TSS}$$

measures the % of variance in Y (dependent) which is explained by the variance in X (independent)

also PRE (proportional reduction in error - % of errors reduced in predicting Y by knowing X)

also used as a measure of the strength of the model (or, if only 2 variables, of the relationship)

ranges from 0-1:

0 is less explanation, 0% PRE, and weakest possible

1 is total explanation, 100% PRE, and strongest possible

0 - 0.1 is “weak”; 0.1 - 0.5 is “moderate”; 0.5 to 1.0 is “strong”

(note that an r^2 of 0.5 means an r of just above 0.7 (since 0.7 squared is 0.49)