

Linear Regression

James H. Steiger



Regression – The General Setup

- # You have a set of data on two variables, X and Y , represented in a scatter plot.
 - # You wish to find a simple, convenient mathematical function that comes close to most of the points, thereby describing succinctly the relationship between X and Y .
-

Linear Regression

- # The straight line is a particularly simple function.
- # When we fit a straight line to data, we are performing *linear regression analysis*.

Linear Regression

The Goal

- Find the “best fitting” straight line for a set of data. Since every straight line fits the equation

$$Y = bX + c$$

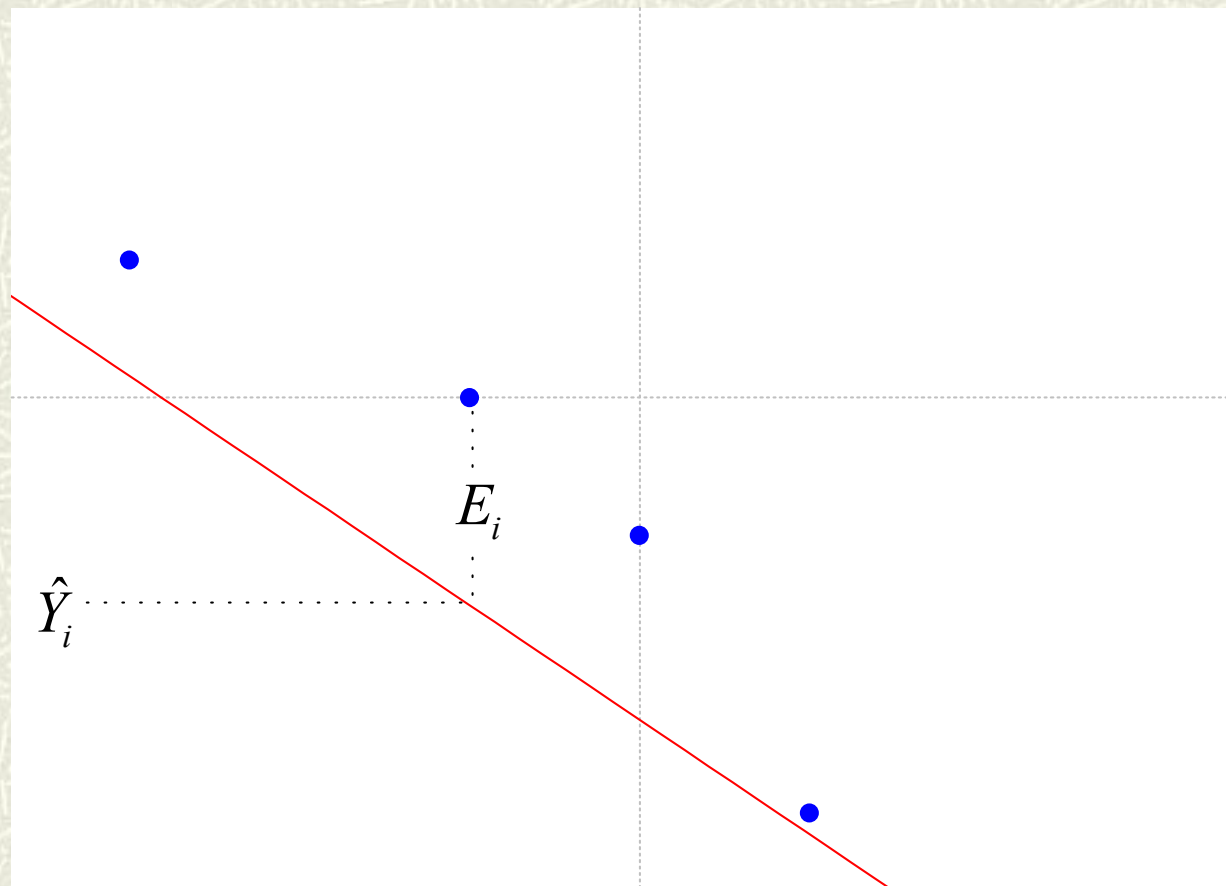
with slope b and Y -intercept c , it follows that our task is to find a b and c that produce the best fit.

- There are many ways of operationalizing the notion of a “best fitting straight line.” A very popular choice is the “Least Squares Criterion.”
-

The Least Squares Criterion

- # For every data point X_i, Y_i , define the “predicted score” \hat{Y}_i to be the value of the straight line evaluated at X_i . The “regression residual,” or “error of prediction” is the distance from the straight line to the data point in the up-down direction.

The Least Squares Criterion



The Least Squares Criterion

- # The *least squares criterion* operationalizes the notion of best fitting line as that choice of b and c that minimizes the sum of squared “residuals,” or errors. Solving this problem is an elementary problem in differential calculus. The method of solution need not concern us here, but the result is famous, important, and should be memorized.
-

The Least Squares Solution

$$b = r_{YX} \frac{S_y}{S_X}$$

$$c = \bar{Y}_\bullet - b\bar{X}_\bullet$$

The Standardized Least Squares Solution

Suppose X and Y are both in Z score form.
What will the equations for the least squares coefficients b and c reduce to? (C.P.)

Variance of Predicted and Error Scores

- When you compute a regression line, you can “plug in” all the X scores into the equation for the regression line, and obtain predicted and error scores for each person. If you do this, it is well known that the variances of the predicted and error scores are

$$S_{\hat{Y}}^2 = r_{YX}^2 S_Y^2 \quad S_E^2 = (1 - r_{YX}^2) S_Y^2$$

- Moreover, the predicted and error scores are always precisely uncorrelated, that is,

$$S_{\hat{Y},E} = 0$$

Variance of Predicted and Error Scores

- # Can you prove the results on the preceding page, using linear transformation and linear combination theory? (C.P.)