Ch. 11: The R.M.S. Error for Regression

The regression line is used to predict y from x, but the actual values won't generally fall exactly on the line.

A _____ or _____ measures how far off the actual value is from the predicted value:

residual = actual y-value − predicted y-value

Ex: (see Chapter 10, page 81)

If someone scored 60 points in the Quiz, what is their predicted score in the Final?

If someone scored 40 points in the Quiz, what is their predicted score in the Final?

In fact, the student with 60 points in the Quiz obtained 110 points in the Final. The residual is:

One student with 40 points in the Quiz obtained 103 points in the Final. The residual is:

Another student with 40 points in the Quiz obtained 55 points in the Final. The residual is:

We can measure the average size of our errors by using the _____:

$$\text{r.m.s. error} = \sqrt{\text{average of (residuals)}^2}$$

The *r.m.s. error of the regression line* means the same for the regression line as the SD for the average.

In many cases, about ___ % of data points will lie within 1 r.m.s. error from the regression line.

About ___ % of data points will lie within 2 r.m.s. errors from the regression line.

Computing the R.M.S. Error

We have two ways of computing the r.m.s. error of the regression line:

1. By definition:
    - Calculate the regression line.
    - Calculate the predicted values.
    - Calculate the residuals.
    - Calculate the r.m.s. of the residuals.
2. By the following shortcut:

$$\text{r.m.s. error} = \sqrt{1 - r^2} \times SD_y$$

Calculating r.m.s. errors

Ex: Predicting Final score from Quiz score:

quiz-avg (x) = 44 final-avg (y) = 85

$SD_{quiz} = 12.5$ $SD_{final} = 25$ r = 0.75

Ex: Predicting son's height from father's height (see Chapter 8 and page 170 in book):

Fathers: $avg_x = 68''$, $SD_x = 2.7''$

Sons: $avg_y = 69''$, $SD_y = 2.7''$

$r = 0.5$

Plotting Residuals

We usually plot residuals against our explanatory (x) variable or any other known variable.

The residuals should have an average of 0 and the residual plot should show no trend.

If there is a strong pattern in the residual plot, the regression line is not appropriate.

Ch. 12: The Regression Line

Equation of the regression line:

$$y = \text{intercept} + \text{slope} \cdot x$$

where

slope =

intercept =

= predicted y when $x = 0$

Ex: In the "Performance of Stats Students" data, we have

quiz-avg (x) = 44 final-avg (y) = 85

$SD_{quiz} = 12.5$ $SD_{final} = 25$ $r = 0.75$

We calculate the regression line as

Using this equation, we can predict the Final scores for Quiz scores:

Quiz	Final
40	
60	

The slope says that _____ with each extra point in the Quiz, there is an increase of _____ points in the Final, on average.

We cannot fully rely on the slope to predict y when the data originates from an _____.

There are two major confounding factors here:

- _____

- _____

Remember, association is not the same as causation.

Ex: HANES men 18–24 years:

height-avg (x) = 70"	weight-avg (y) = 162 lb
SD_x = 3"	SD_y = 30 lb
r = 0.47	

Regression equation for predicting weight based on height:

Regression equation is:

Note: Sometimes the intercept does not make sense; it may be negative when we would expect it to be zero or positive.

The Method of Least-Squares

Among all possible lines one can draw on a scatterplot, the regression line has the smallest possible sum of the squared residuals, i.e., the smallest r.m.s. error. For this reason, the regression line is also called "least squares" line.