Correlation, Regression, Prediction

Simple Linear Correlation:

Simple --- Two Variables (X & Y)

Linear --- Straight Line ($Y = mX + b$)

Correlation --- Mathematical Relationship

If significant correlation exists between the two variables,
then the regression equation $\hat{Y} = \beta_0 + \beta_1 X$ is valid for predicting values of Y for given values of X.

Assumptions:

1. Both X and Y are random variables.

2. Data pairs (X, Y) are Bivariate Normal Distributions.
   For any fixed value of X, the Y values are normally distributed.
   For any fixed value of Y, the X values are normally distributed.

Properties of r:

1. The value of r is a measure of linear relationship only.

2. $-1 < r < +1$
   For $r = -1$, perfect *negative* correlation.
   For $r = +1$, perfect *positive* correlation.
   If $r = 0$, then zero *linear* correlation.

Common Errors:

1. Significant linear correlation does NOT provide proof of Cause & Effect.

2. Lack of significant LINEAR correlation does not imply there is no other mathematical relationship.

3. Tests for correlation should not based on rates or averages.

4. Do not use the regression equation for predicting if there is no significant correlation.

5. When using the regression equation for predicting, stay within the range of the X variable.

6. A regression equation based on old data is not necessarily valid for current situations.

7. A regression equation based on current data is not necessarily valid for future situations.

8. Do not use the regression equation to make predictions about a population that is different from the population from which the sample data were drawn.

Correlation and Regression

Correlation Coefficient

$$r = \frac{n\sum XY - \sum X \sum Y}{\sqrt{\left[n\sum X^2 - \left(\sum X\right)^2 \right] \left[n\sum Y^2 - \left(\sum Y\right)^2 \right]}}$$

t_{test}

$$t_{test} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Degrees of Freedom = n - 2

Slope

$$\beta_1 = \frac{n\sum XY - \sum X \sum Y}{n\sum X^2 - \left(\sum X\right)^2}$$

Intercept

$$\beta_0 = \bar{Y} - \beta_1 \bar{X}$$

Prediction Equation

$$\hat{Y} = \beta_0 + \beta_1 X$$

Standard Error of the Estimate

$$S_e = \sqrt{\frac{\sum Y^2 - \beta_0 \sum Y - \beta_1 \sum XY}{n-2}}$$

Confidence Interval

$$Y = \hat{Y} \pm t_{\alpha/2} S_e \sqrt{1 + \frac{1}{n} + \frac{n(X_0 - \bar{X})^2}{n\sum X^2 - \left(\sum X\right)^2}}$$

Alternate Forms

$$S_{XX} = \sum (X_i - \overline{X})^2 = \sum X_i^2 - \frac{\left(\sum X_i\right)^2}{n}$$

$$S_{YY} = \sum (Y_i - \overline{Y})^2 = \sum Y_i^2 - \frac{\left(\sum Y_i\right)^2}{n}$$

$$S_{XY} = \sum (X_i - \overline{X})(Y_i - \overline{Y})$$

$$S_{XY} = \sum X_i (Y_i - \overline{Y}) = \sum Y_i (X_i - \overline{X})$$

$$S_{XY} = \sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} S_{YY}}}$$

$$\beta_1 = r \sqrt{S_{YY} / S_{XX}} = \frac{S_{XY}}{S_{XX}}$$

$$\beta_0 = \overline{Y} - \beta_1 \overline{X}$$

$$Y = \beta_0 + \beta_1 X$$

Regression Analysis Using the ANOVA Table

Source	Sum Squares	df	Mean Square	F Test
Model(Regression)	_____	1	_____	_____
Error	_____	n - 2	_____	
Total	_____	n - 1		

Null Hypothesis: $H_0: \beta_1 = 0$ (i.e., No Correlation)

Critical Value: F Distribution Table F 0.05, dfnum, dfdenom

If Reject the Hull,
conclude significant correlation (Okay to Use the Regression Equation).

If Fail to Reject,
conclude no significant correlation (Do Not Use the Regression Equation).