Regression Equations

A linear regression with 1 predictor.

Equations:

Theory or Observed
\[ y_i = \beta_0 + \beta_1 x_i + \epsilon_i \]

Model or Prediction
\[ \hat{y} = \beta_0 + \beta_1 x \]

The mean of \( y \) ( \( \bar{y} \) ) is the best guess for \( y \) when no other information (e.g., the value of \( x \)) is used to predict.

Model aka fitted values include the intercept ( \( \beta_0 \) aka \( \alpha \)) and coefficients (\( \beta_1 \)) for each predictor variable (\( x \)).

Residuals (\( \epsilon_i \)) should represent random error. To check, create a scatterplot for predicted vs residual values. Across each predicted value, residuals should be unbiased (mean of 0) and homoscedastic (equal variance).

Sum of Squares (SS)

\[ SSM = \sum (\hat{y}_i - \bar{y})^2 \quad \text{Explained Variation (also SS Explained)} \]
\[ SSE = \sum (y_i - \hat{y}_i)^2 = \sum \epsilon_i^2 \quad \text{Unexplained Variation, Prediction Error} \]
\[ SST = \sum (y_i - \bar{y})^2 \quad \text{Total Variation, \( SST = \sum \epsilon_i^2 \)} \]

Mean Squared (MS)

\[ MS = \frac{SS}{df} \quad \text{SS divided by the degrees of freedom} \]
\[ MSM = \frac{SSM}{p-1} \quad \text{Average Explained Variation per predictor} \]
\[ MSE = \frac{SSE}{n-p} \quad \text{Unbiased Estimate of Residuals Variance} \]
\[ MST = \frac{SST}{n-1} \quad \text{Unbiased Estimate of Total Variance} \]

\[ F = \frac{MSM}{MSE} \quad \text{Test statistic for the null hypothesis (H0) that all coefficients (e.g., \( \beta_1 \)) equal zero} \]

\[ SEE = \sqrt{MSE} \quad \text{Standard Deviation of the Residuals} \]

R-Squared (R²)

\[ R^2 = r^2 = 1 - \left( \frac{SSE}{SST} \right) = \frac{SSM}{SST} \quad \text{where} \ r \ \text{is the correlation between observed and predicted \( y \)’s} \]

\[ \text{Adj R}^2 = 1 - \left( \frac{MSE}{MST} \right) \quad \text{for multiple regression; penalizes for added predictors that do not decrease MSE} \]
Null Hypothesis Significance Testing

Test Statistics

Because the purpose of each test is different as are the types of values, each has a different way of calculating the test statistic.

**Test Statistic** \[ \frac{\text{Signal}}{\text{Noise}} = \frac{\text{Explained Variation}}{\text{Unexplained Variation}} \]

- \( Z = \frac{\text{Difference Between Population and Sample Means}}{\text{Population Standard Error}} \) (1-Sample Z-Test)
- \( t = \frac{\text{Difference between Group Means}}{\text{Pooled Standard Error}} \) (Independent Samples t-Test)
- \( F = \frac{\text{MS Model}}{\text{MS Error}} \) (Regression) = \( \frac{\text{MS Between Groups}}{\text{MS Within Groups}} \) (ANOVA)
- \( X^2 = \frac{\sum (\text{Observed} - \text{Expected})^2}{\text{Expected}} \) (Chi-Square)

Critical Values

http://graphpad.com/support/faq/plotting-t-z-f-or-chi-square-distributions-with-prism/

Rejecting the Null Hypothesis

Before computers, Test Statistics and Critical Values were crucial because the exact p-value was very difficult to calculate by hand. Now, computers will give the exact p-value so these are less necessary.

Reject Null if

- **OLD SCHOOL** Test Statistic \( \geq \) Critical Value
- **COMPUTER ERA** p-value \( \leq \) alpha [level]

*From your Data Decision Criteria*