Regression Equations

A **linear regression** with 1 predictor.

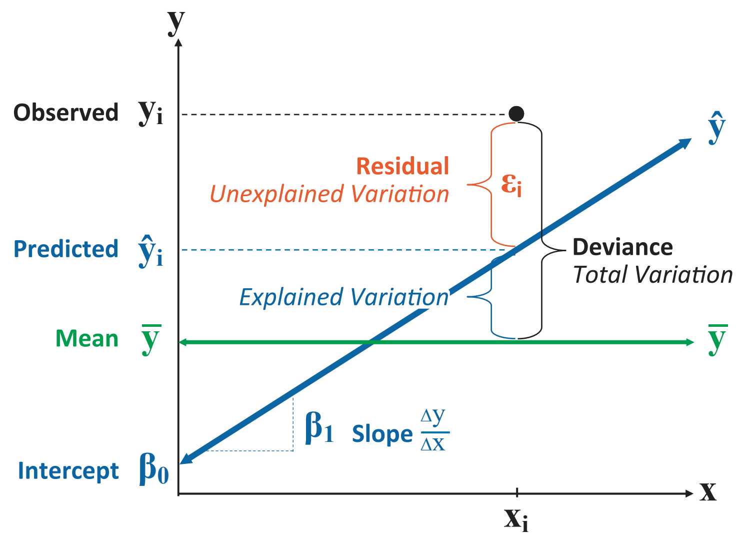
Equations:

Theory or Observed

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

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* (β_0 *aka* α) and *coefficients* (β_1) for each predictor variable (x).

Residuals (ε_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 **S**quares (SS)

Sum of the **S**quared Deviations

$\sum$ means to add the values for each *observation* (i)

SSM $= \sum (\hat{y}_i - \bar{y})^2$	*Explained Variation* (*also* **SS** ***E****xplained*)	**SS** ***M****odel* *aka* **SS** ***R****egression*
SSE $= \sum (y_i - \hat{y}_i)^2 = \sum \varepsilon_i^2$	*Unexplained Variation, Prediction Error*	**SS** ***E****rror* *aka* **SS** ***R****esidual*
SST $= \sum (y_i - \bar{y})^2$	*Total Variation,* **SST = SSE + SSM**	**SS** *Total*

Mean **S**quared (MS)

SS divided by the **d**egrees of **f**reedom

MS = **SS** / **df**

n = sample size; **p** = # of modeled values (coefficients + intercept)

MSM = **SSM** / **p** - 1	*Average Explained Variation per predictor*	**MS** ***M****odel* *aka* **MS** ***R****egression*
MSE = **SSE** / **n** - **p**	*Unbiased Estimate of Residuals Variance*	**MS** ***E****rror* *aka* **MS** ***R****esidual*
MST = **SST** / **n** - 1	*Unbiased Estimate of Total Variance*	**MS** *Total*

F = **MSM** / **MSE** — Test statistic for the null hypothesis (H_0) that all coefficients (e.g., β_1) equal zero

SEE = $\sqrt{\text{MSE}}$ — *Standard Deviation of the Residuals* — **S**td **E**rror of the **E**stimate *aka* **R**oot **MSE**

R-Squared (R^2)

Proportion of variation in y's explained by model; perfect prediction = 1 or 100%

aka **C**oefficient **o**f **D**etermination *aka* **P**roportional **R**eduction in **E**rror

R^2 = r^2 = 1 - (SSE / SST) = SSM / SST *where* **r** *is the correlation between* ***observed*** *and* ***predicted*** y's

Adj R^2 = 1 - (MSE / MST) *for multiple regression; penalizes for added predictors that do not decrease MSE*

© 2015 by Debby Kermer, Mason Library Data Services

This work is licensed under the © Attribution-NonCommercial-ShareAlike International License: http://creativecommons.org/licenses/by-nc-sa/4.0/

Null **H**ypothesis **S**ignificance **T**esting

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

$$\textit{Test Statistic} = \frac{\textit{Signal}}{\textit{Noise}} = \frac{\textit{Explained Variation}}{\textit{Unexplained Variation}}$$

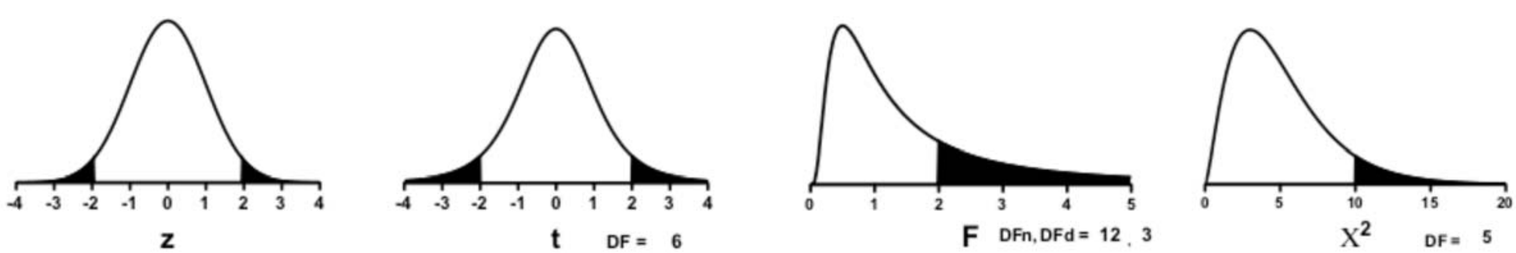
$$Z = \frac{\text{Difference Between Population and Sample Means}}{\text{Population Standard Error}} \quad \text{(1-Sample Z-Test)}$$

$$t = \frac{\text{Difference between Group Means}}{\text{Pooled Standard Error}} \quad \text{(Independent Samples t-Test)}$$

$$F = \frac{\text{MS Model}}{\text{MS Error}} \ \text{(Regression)} = \frac{\text{MS Between Groups}}{\text{MS Within Groups}} \ \text{(ANOVA)}$$

$$X^2 = \frac{\sum(\text{Observed} - \text{Expected})^2}{\text{Expected}} \quad \text{(Chi-Square)}$$

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	≥	Critical Value
COMPUTER ERA	p-value	≤	alpha [level]
	From your Data		*Decision Criteria*