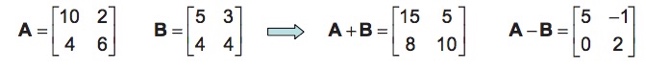
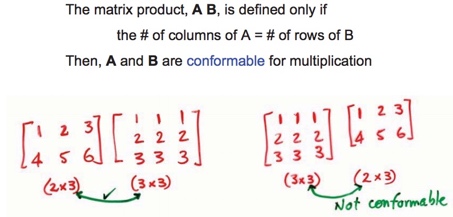
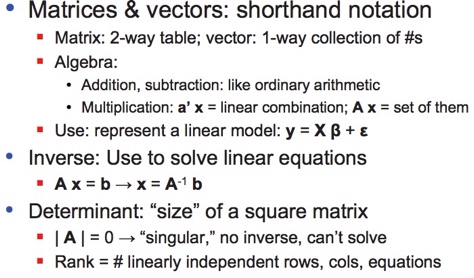
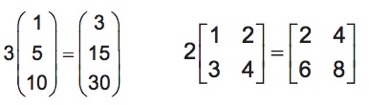
**Matrix/Linear Algebra**

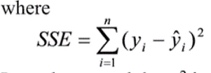
* A matrix is a rectangular array of numbers, with r rows and c columns. Simple way to express linear combinations of variables and general solutions of equations AND Linear statistical models (regression, ANOVA) generalize to any # of predictors & responses. AND transposed matrices just interchange rows and columns. **Goal**: a reading knowledge of matrix expressions to aid in understanding statistical concepts. A **vector** is just a one column matrix, sometimes written in transposed (row) form to save space. **Special Matrices**: unit, zero, contrast, square, symmetric, diagonal, identity.
* **Addition & subtraction**: add corresponding elements. Must have same shape. **Scalar multiplication**: multiply each element by a scalar:





**Y_i = \beta_0 + \beta_1 X_{i1} +Multiple Linear Regression**

Regression is a special case of CCA with only one dependent variable.

../../Screen%20Shot%202016-02-09%20at%201.03.53%20PM.jpeg**Model assumptions**: random error (e) has a normal prob. distribution with mean = 0 and variance σ2 AND random errors are independent. **Fit the model** (method of least squares): uses the plane of best fit that minimize SSE (sum of squared errors), min-max normalization. **Estimate variance σ2**: the variance of random error will be unknown and must be estimated. **Estimate and test the hypothesis about β parameters**: properties of the sampling distributions of beta’s, need all mean equal, want small variance for estimators. **Check the utility of the model**: use a global test, R2 = SSE / SSyy; hypothesis test (one-tail or two-tail), F-Statistic. **Using model for estimation & prediction**: confidence interval for average & prediction interval for a single value (larger/wider than CI). **Validating the model**: assumptions about population 1) linearity 2) normality 3) heteroscedascity 4) auto-correlation. **Pitfalls & Issues**: Over-specification [ # of samples ≥ 5(# of vars + 2)], Extrapolation beyond data range, multicolinearity. **How many variables to choose**: Selection Algos. (Forward, Backwards, Stepwise)

**Canonical Correlation Analysis (2 sets of variables)**

• How to describe/find relationships between multiple **dependent** and **independent** variables.

• Linear combination of the DVs and IVs. The number of CC variates you can have depends the number of variables in the smaller set. The linear combo is that it adds up a multiple of each variable’s value and thus results in a single value; which gives the *score* of the latent variable.

• **Goal**: to model a *latent* (unobserved variable) on each side so that each *latent* (never measured directly and are weighted sums of the real variables specified by the coefficients of the original variables) variable on each side has maximum correlation.

• **Correlation** is just the normalized covariance. Sample variance measures the average of the square differences from the sample mean. ( *E*[ x – xbar)2]), **covariance** does that for two different variables rather than x with itself ( *E*[ x – xbar)( y - ybar])

**Principal Component Analysis (1 set of variables)**

* How to describe data by reducing the number of features in a dataset. PCA can be used in CCA on variables in the dataset to remove correlation before running CCA.
* The variables’ coefficients in the principal components are sometimes called the *scores*, e.g. in SPSS, one of the outputs is the “component score coefficient matrix”
* The ***loadings*** are the correlation value between the *real* values and the *latent* values
* The PCA vectors come from **eigenvectors** of the covariance matrix, which is the matrix of the covariances of all pairs of variables and is thus m x m. PC1 is the first eigenvector that captures the largest variance in the dataset.