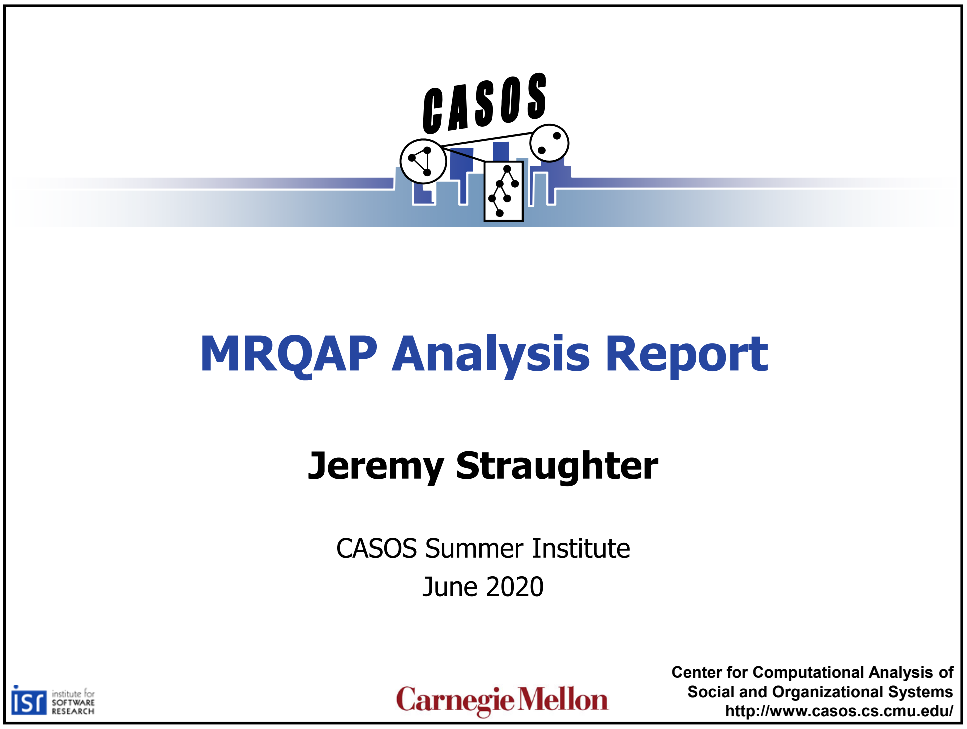
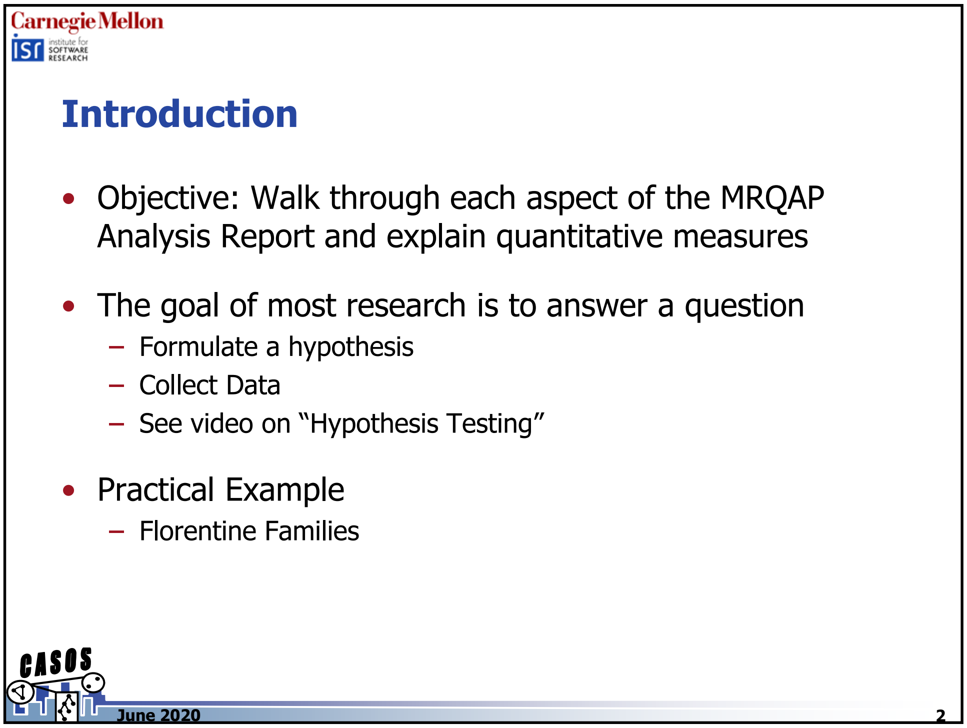


MRQAP Analysis Report

Jeremy Straughter

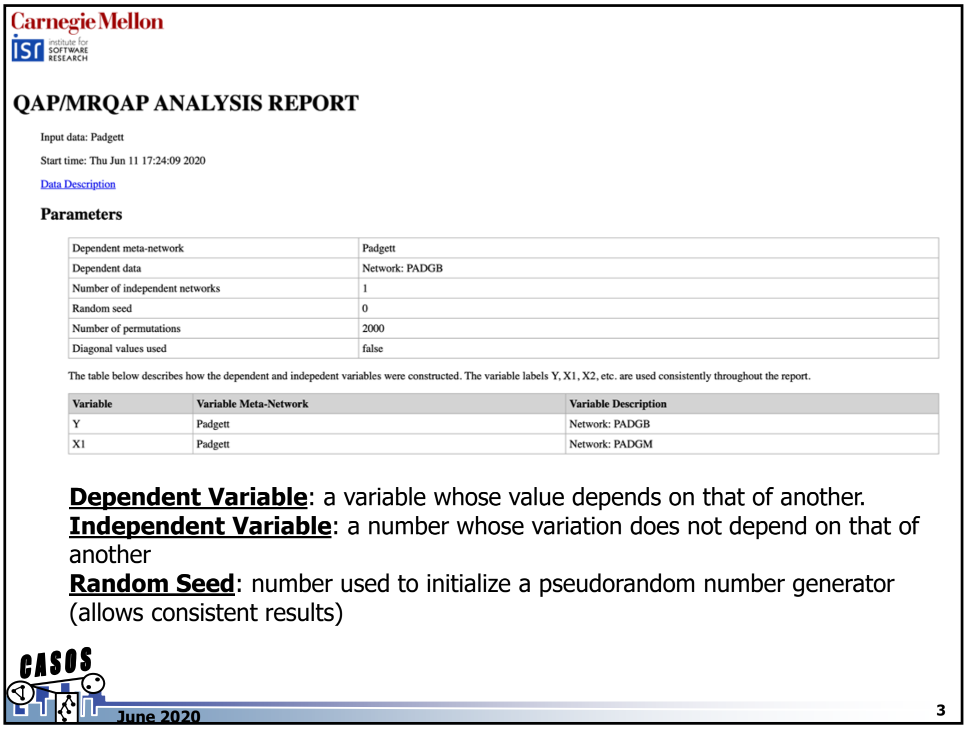
CASOS Summer Institute
June 2020

Center for Computational Analysis of
Social and Organizational Systems
http://www.casos.cs.cmu.edu/



Introduction

- Objective: Walk through each aspect of the MRQAP Analysis Report and explain quantitative measures
- The goal of most research is to answer a question
  - Formulate a hypothesis
  - Collect Data
  - See video on "Hypothesis Testing"
- Practical Example
  - Florentine Families



QAP/MRQAP ANALYSIS REPORT

Input data: Padgett

Start time: Thu Jun 11 17:24:09 2020

Data Description

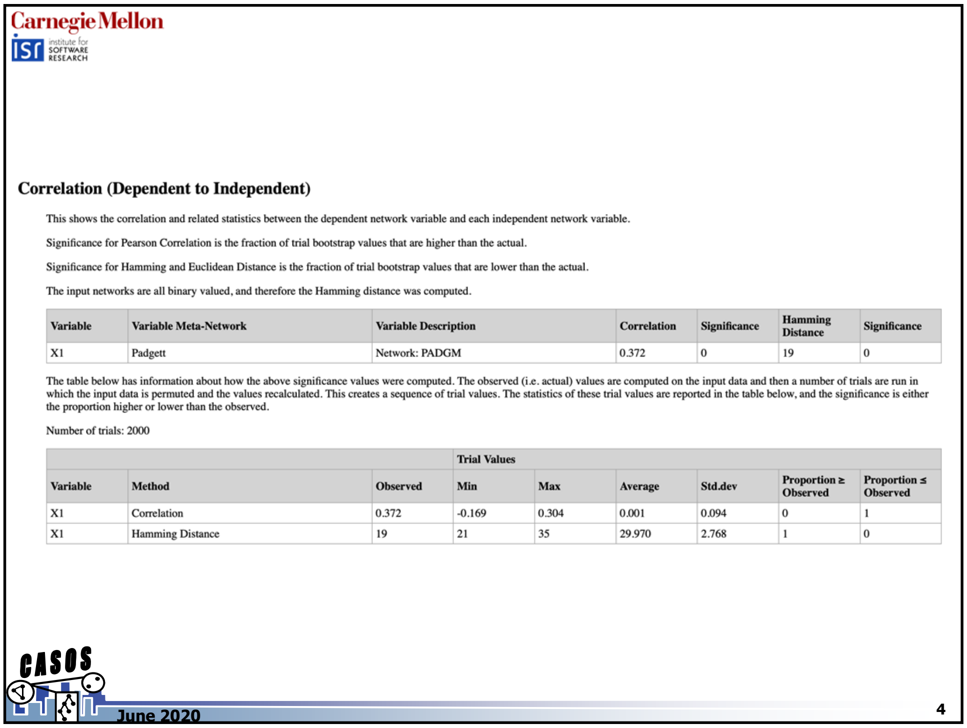
Parameters

Dependent meta-network	Padgett
Dependent data	Network: PADGB
Number of independent networks	1
Random seed	0
Number of permutations	2000
Diagonal values used	false

The table below describes how the dependent and indepedent variables were constructed. The variable labels Y, X1, X2, etc. are used consistently throughout the report.

Variable	Variable Meta-Network	Variable Description
Y	Padgett	Network: PADGB
X1	Padgett	Network: PADGM

Dependent Variable: a variable whose value depends on that of another.
Independent Variable: a number whose variation does not depend on that of another
Random Seed: number used to initialize a pseudorandom number generator (allows consistent results)



Correlation (Dependent to Independent)

This shows the correlation and related statistics between the dependent network variable and each independent network variable.

Significance for Pearson Correlation is the fraction of trial bootstrap values that are higher than the actual.

Significance for Hamming and Euclidean Distance is the fraction of trial bootstrap values that are lower than the actual.

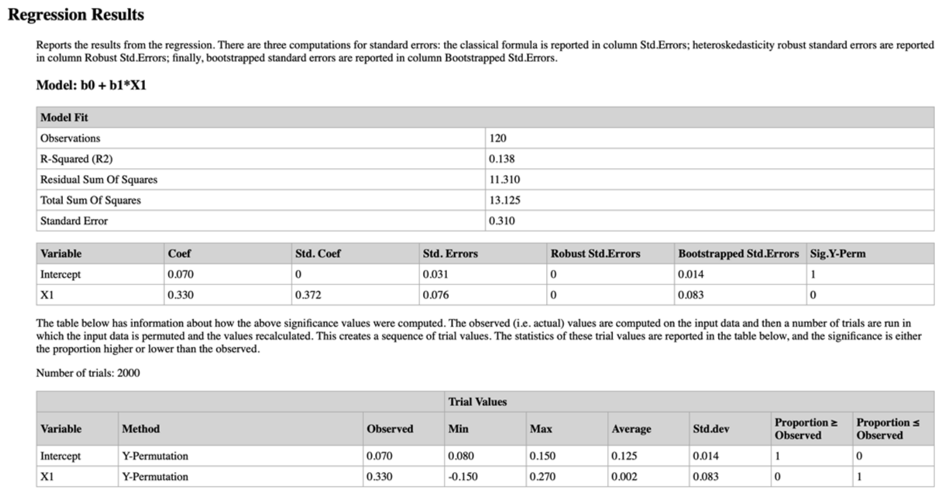
The input networks are all binary valued, and therefore the Hamming distance was computed.

Variable	Variable Meta-Network	Variable Description	Correlation	Significance	Hamming Distance	Significance
X1	Padgett	Network: PADGM	0.372	0	19	0

The table below has information about how the above significance values were computed. The observed (i.e. actual) values are computed on the input data and then a number of trials are run in which the input data is permuted and the values recalculated. This creates a sequence of trial values. The statistics of these trial values are reported in the table below, and the significance is either the proportion higher or lower than the observed.

Number of trials: 2000

			Trial Values					
Variable	Method	Observed	Min	Max	Average	Std.dev	Proportion ≥ Observed	Proportion ≤ Observed
X1	Correlation	0.372	-0.169	0.304	0.001	0.094	0	1
X1	Hamming Distance	19	21	35	29.970	2.768	1	0



Regression Results

Reports the results from the regression. There are three computations for standard errors: the classical formula is reported in column Std.Errors; heteroskedasticity robust standard errors are reported in column Robust Std.Errors; finally, bootstrapped standard errors are reported in column Bootstrapped Std.Errors.

Model: b0 + b1*X1

Model Fit	
Observations	120
R-Squared (R2)	0.138
Residual Sum Of Squares	11.310
Total Sum Of Squares	13.125
Standard Error	0.310

Variable	Coef	Std. Coef	Std. Errors	Robust Std.Errors	Bootstrapped Std.Errors	Sig.Y-Perm
Intercept	0.070	0	0.031	0	0.014	1
X1	0.330	0.372	0.076	0	0.083	0

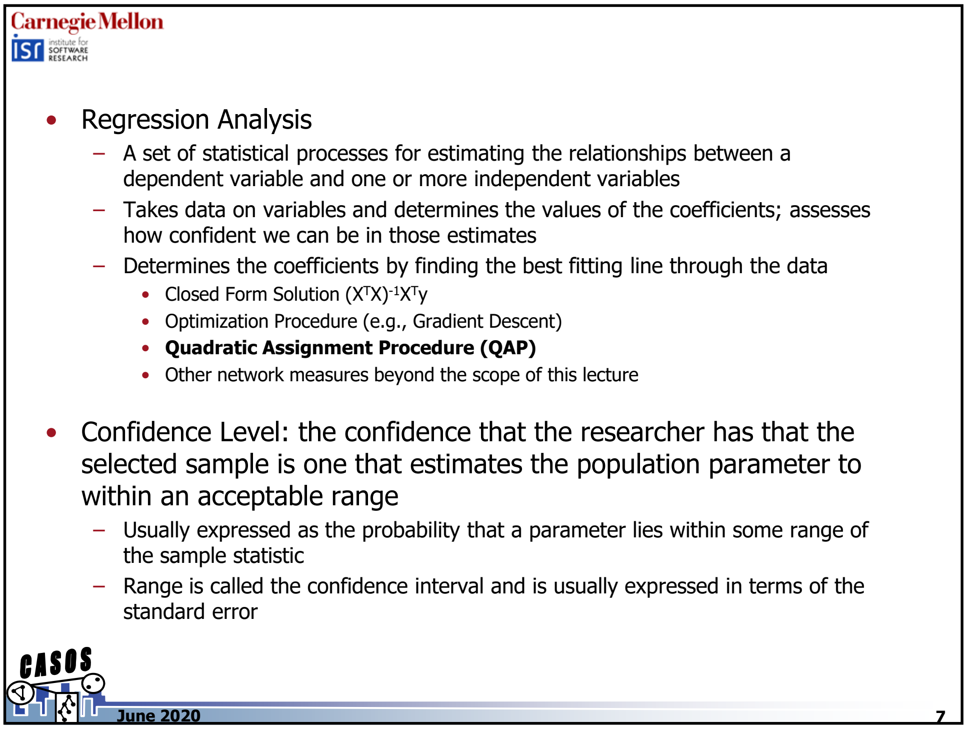
The table below has information about how the above significance values were computed. The observed (i.e. actual) values are computed on the input data and then a number of trials are run in which the input data is permuted and the values recalculated. This creates a sequence of trial values. The statistics of these trial values are reported in the table below, and the significance is either the proportion higher or lower than the observed.

Number of trials: 2000

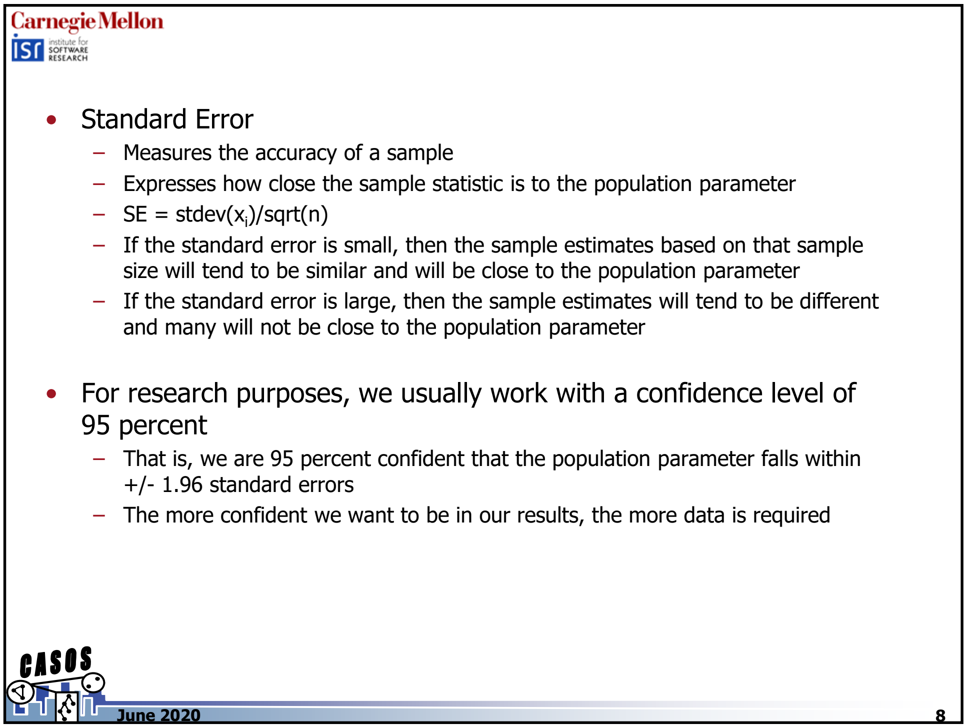
			Trial Values					
Variable	Method	Observed	Min	Max	Average	Std.dev	Proportion ≥ Observed	Proportion ≤ Observed
Intercept	Y-Permutation	0.070	0.080	0.150	0.125	0.014	1	0
X1	Y-Permutation	0.330	-0.150	0.270	0.002	0.083	0	1

Useful Terminology

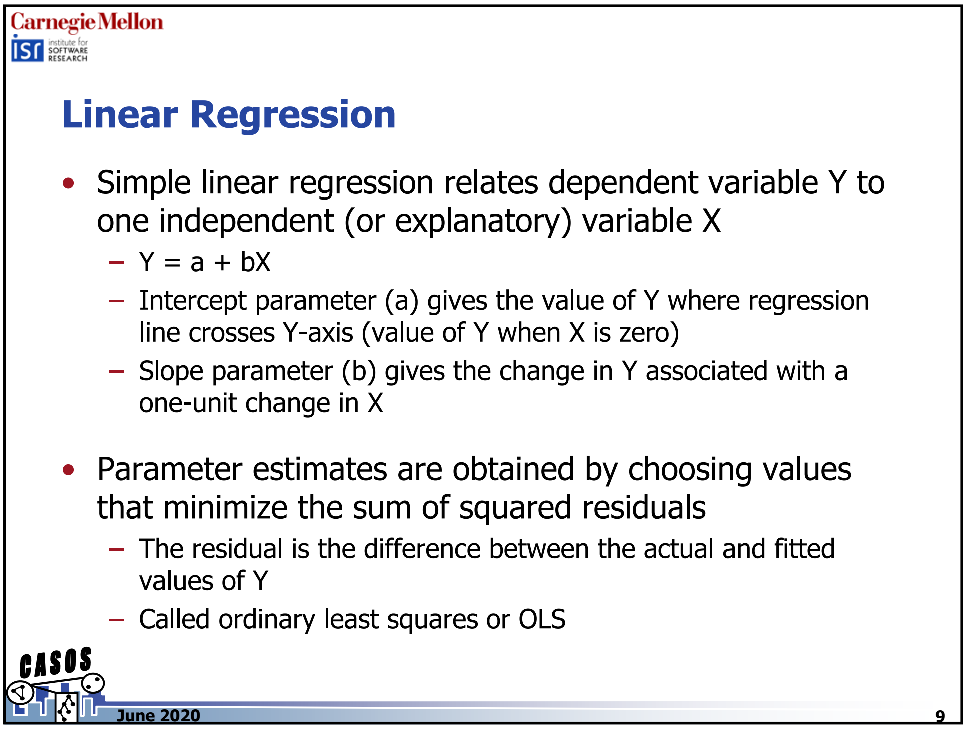
- Correlation: measures the strength of the relationship between two variables
  - Definition: $r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} * \sqrt{\sum_i (y_i - \bar{y})^2}}$
  - Bounded from -1,1
  - Positive values indicate a direct (positive) relationship
  - Negative values indicate an inverse (negative) relationship
  - The farther from zero, the stronger the relationship
- Parameters: coefficients that determine the mathematical relationship among the variables
  - Example: Y = a + bX1 + cX2 + dX3
  - You have data for the Y and X variables
  - Coefficients/Parameters describe the relationship



- Regression Analysis
  - A set of statistical processes for estimating the relationships between a dependent variable and one or more independent variables
  - Takes data on variables and determines the values of the coefficients; assesses how confident we can be in those estimates
  - Determines the coefficients by finding the best fitting line through the data
    - Closed Form Solution $(X^TX)^{-1}X^Ty$
    - Optimization Procedure (e.g., Gradient Descent)
    - **Quadratic Assignment Procedure (QAP)**
    - Other network measures beyond the scope of this lecture
- Confidence Level: the confidence that the researcher has that the selected sample is one that estimates the population parameter to within an acceptable range
  - Usually expressed as the probability that a parameter lies within some range of the sample statistic
  - Range is called the confidence interval and is usually expressed in terms of the standard error



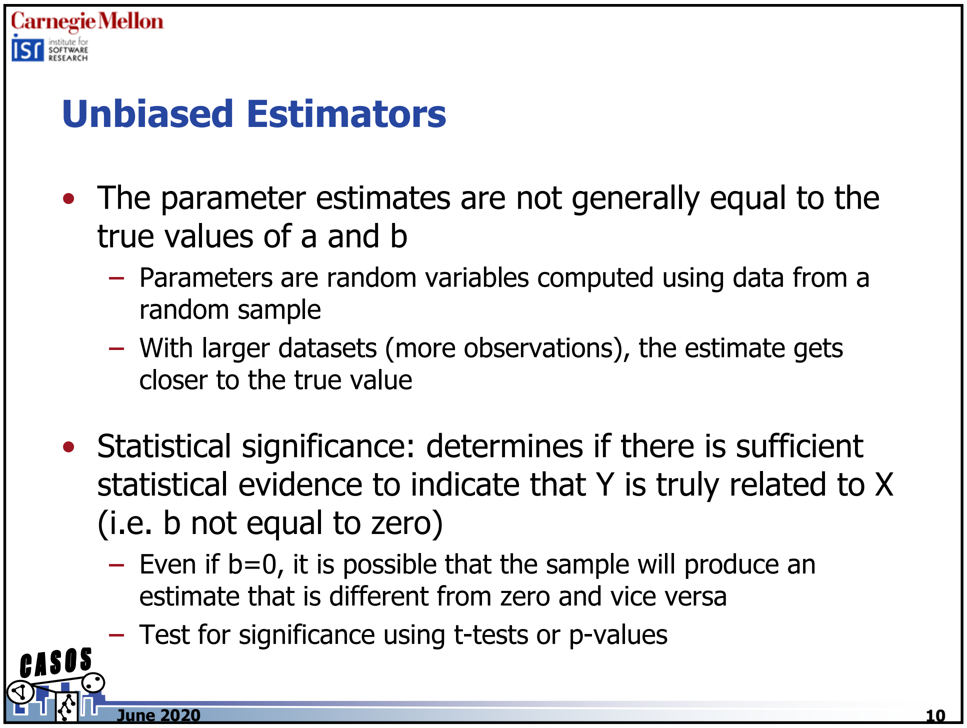
- Standard Error
  - Measures the accuracy of a sample
  - Expresses how close the sample statistic is to the population parameter
  - SE = stdev(x_i)/sqrt(n)
  - If the standard error is small, then the sample estimates based on that sample size will tend to be similar and will be close to the population parameter
  - If the standard error is large, then the sample estimates will tend to be different and many will not be close to the population parameter
- For research purposes, we usually work with a confidence level of 95 percent
  - That is, we are 95 percent confident that the population parameter falls within +/- 1.96 standard errors
  - The more confident we want to be in our results, the more data is required



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Linear Regression

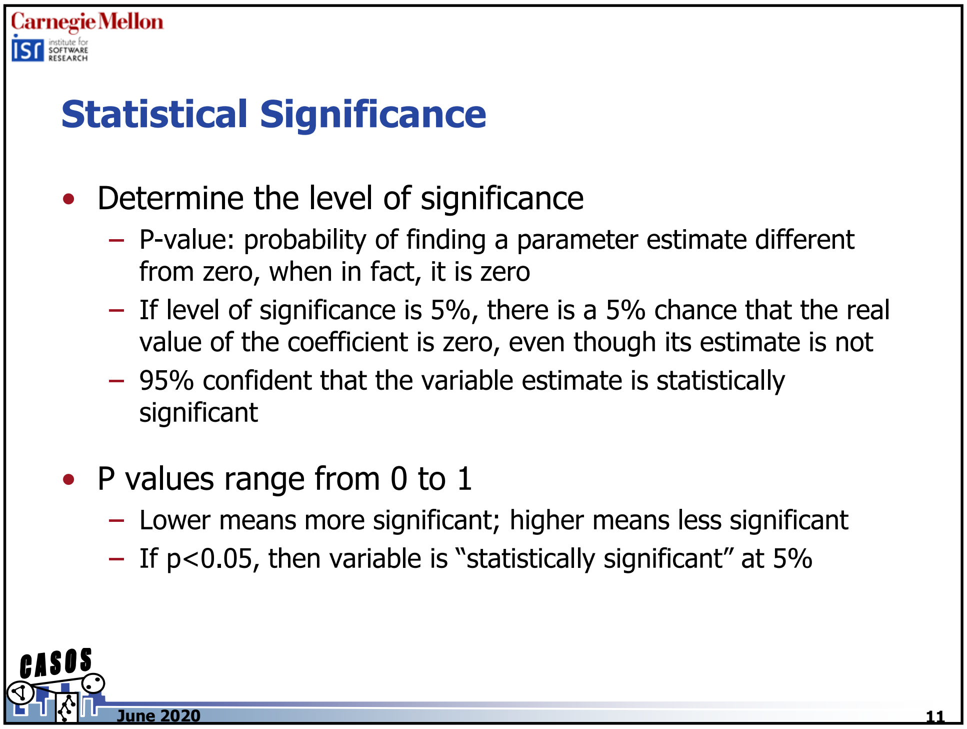
- Simple linear regression relates dependent variable Y to one independent (or explanatory) variable X
  - Y = a + bX
  - Intercept parameter (a) gives the value of Y where regression line crosses Y-axis (value of Y when X is zero)
  - Slope parameter (b) gives the change in Y associated with a one-unit change in X
- Parameter estimates are obtained by choosing values that minimize the sum of squared residuals
  - The residual is the difference between the actual and fitted values of Y
  - Called ordinary least squares or OLS



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Unbiased Estimators

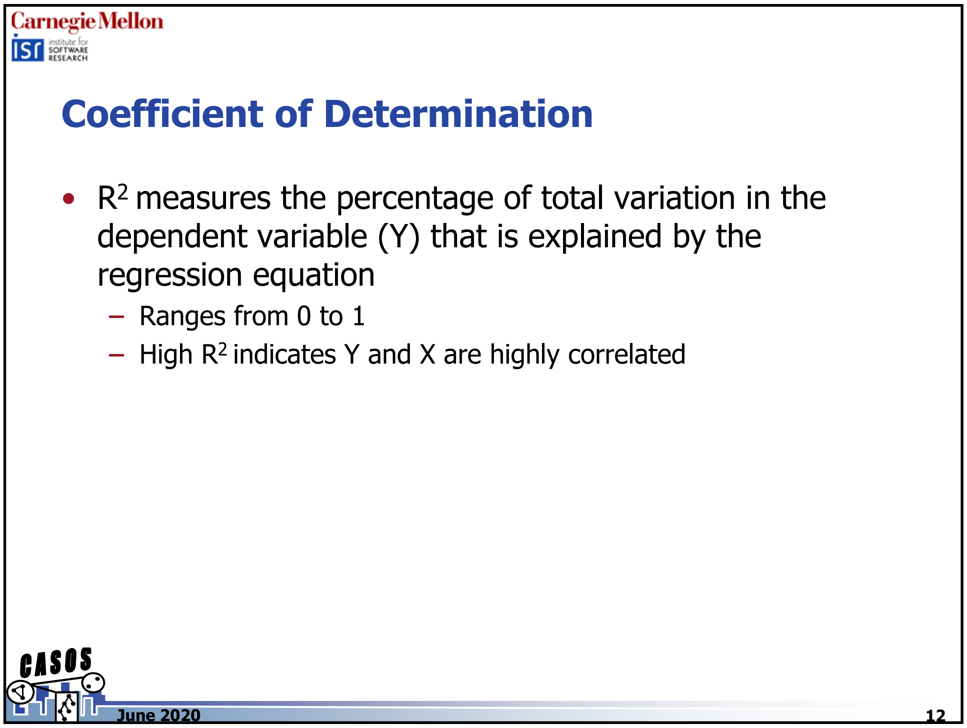
- The parameter estimates are not generally equal to the true values of a and b
  - Parameters are random variables computed using data from a random sample
  - With larger datasets (more observations), the estimate gets closer to the true value
- Statistical significance: determines if there is sufficient statistical evidence to indicate that Y is truly related to X (i.e. b not equal to zero)
  - Even if b=0, it is possible that the sample will produce an estimate that is different from zero and vice versa
  - Test for significance using t-tests or p-values



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Statistical Significance

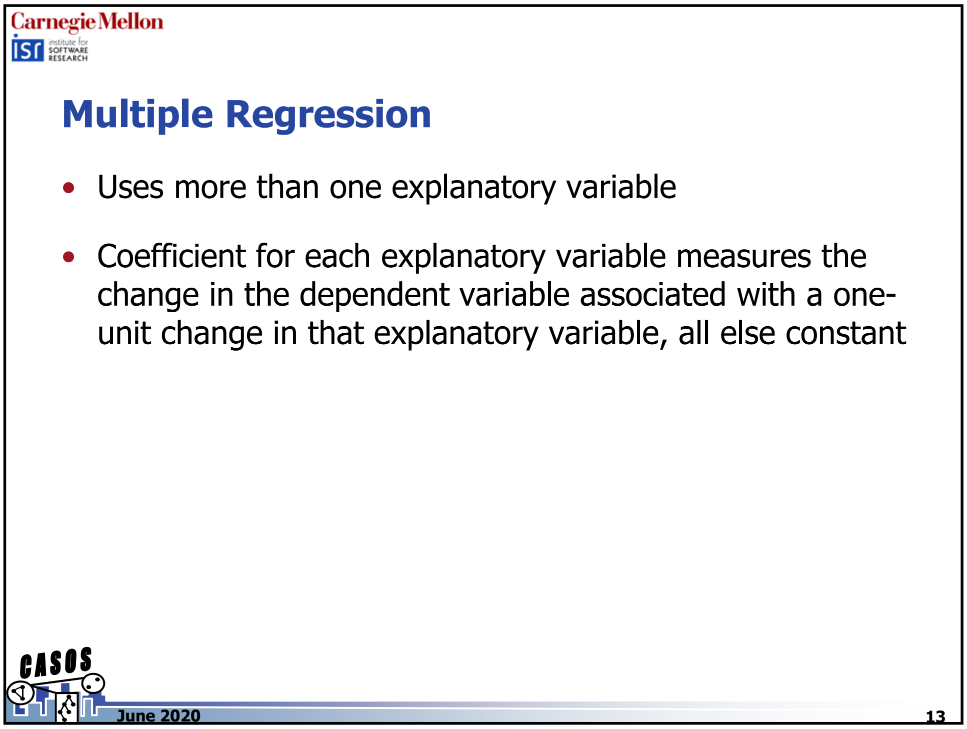
- Determine the level of significance
  - P-value: probability of finding a parameter estimate different from zero, when in fact, it is zero
  - If level of significance is 5%, there is a 5% chance that the real value of the coefficient is zero, even though its estimate is not
  - 95% confident that the variable estimate is statistically significant
- P values range from 0 to 1
  - Lower means more significant; higher means less significant
  - If p<0.05, then variable is "statistically significant" at 5%



Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Coefficient of Determination

- R^2 measures the percentage of total variation in the dependent variable (Y) that is explained by the regression equation
  - Ranges from 0 to 1
  - High R^2 indicates Y and X are highly correlated

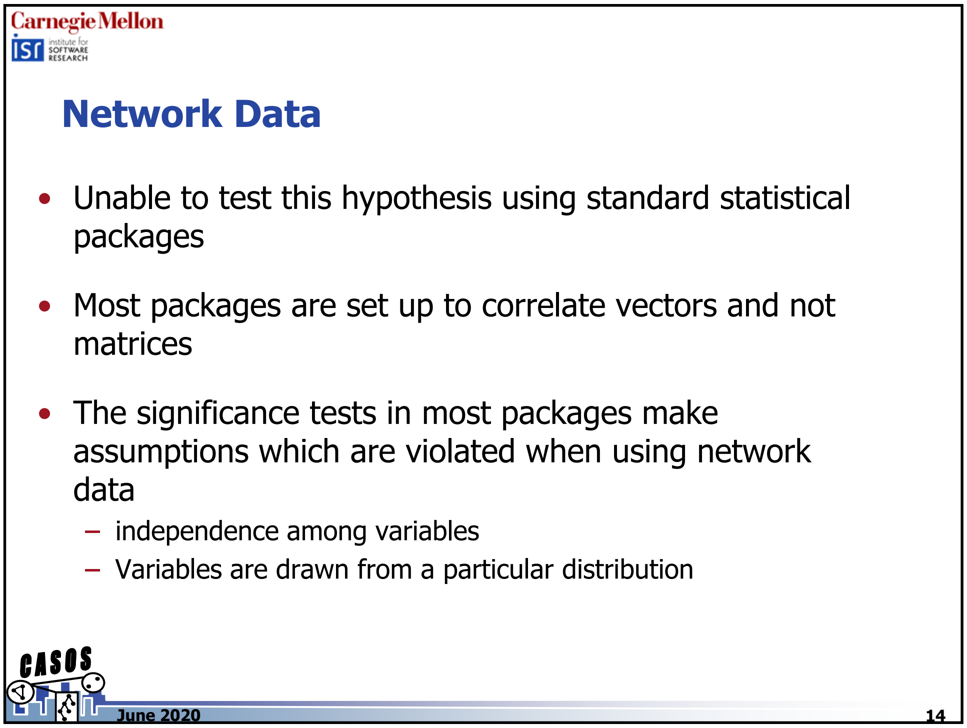


Carnegie Mellon

isr institute for SOFTWARE RESEARCH

Multiple Regression

- Uses more than one explanatory variable
- Coefficient for each explanatory variable measures the change in the dependent variable associated with a one-unit change in that explanatory variable, all else constant

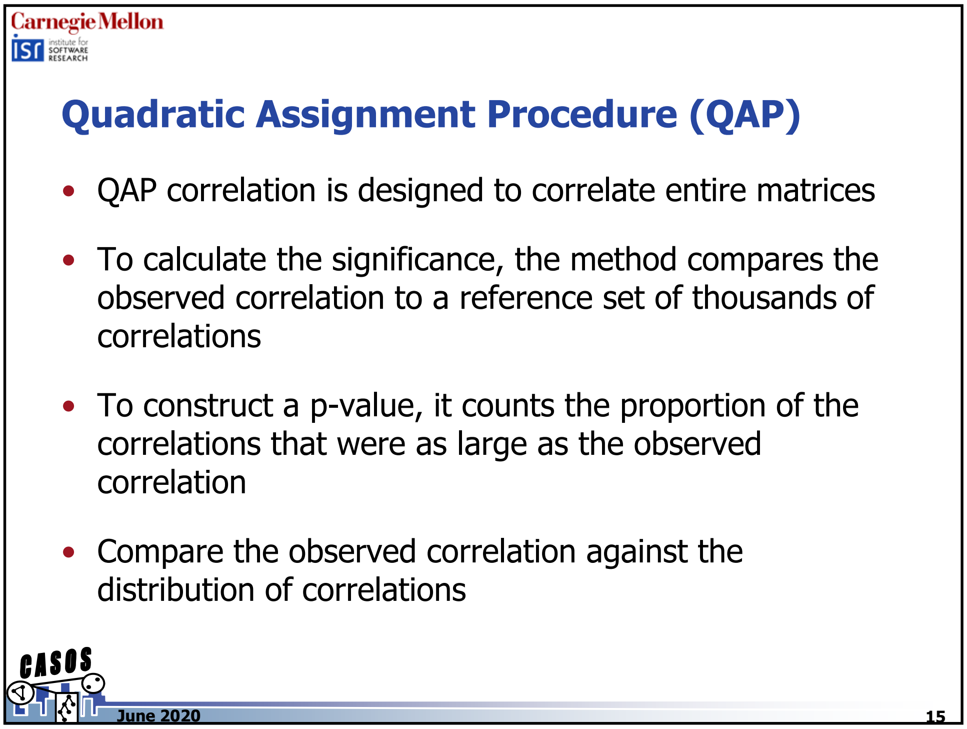


Carnegie Mellon

isr institute for SOFTWARE RESEARCH

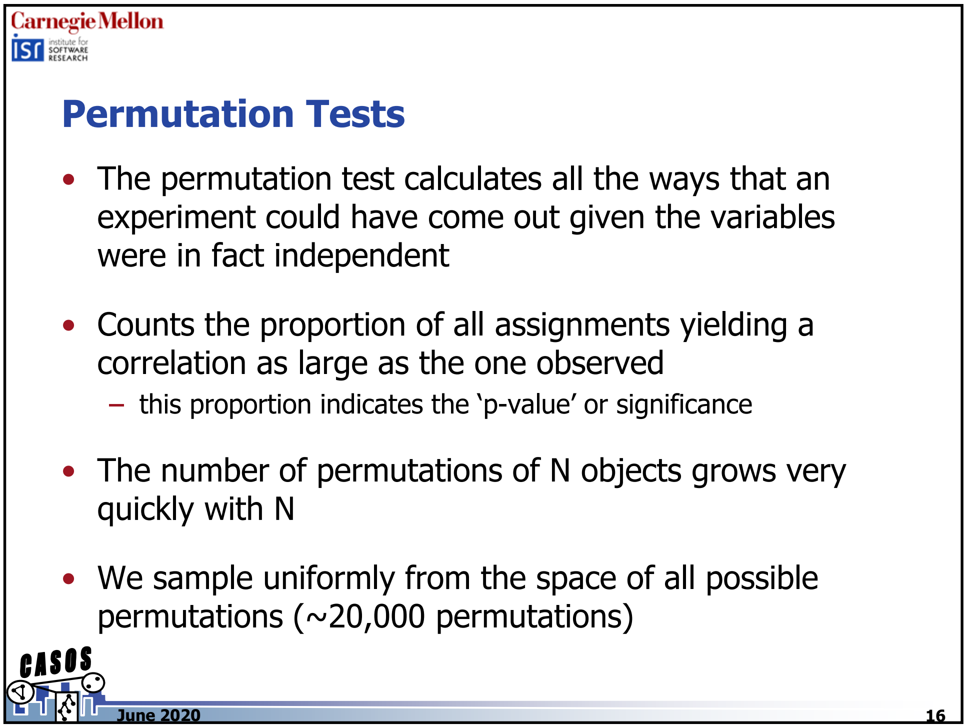
Network Data

- Unable to test this hypothesis using standard statistical packages
- Most packages are set up to correlate vectors and not matrices
- The significance tests in most packages make assumptions which are violated when using network data
  - independence among variables
  - Variables are drawn from a particular distribution



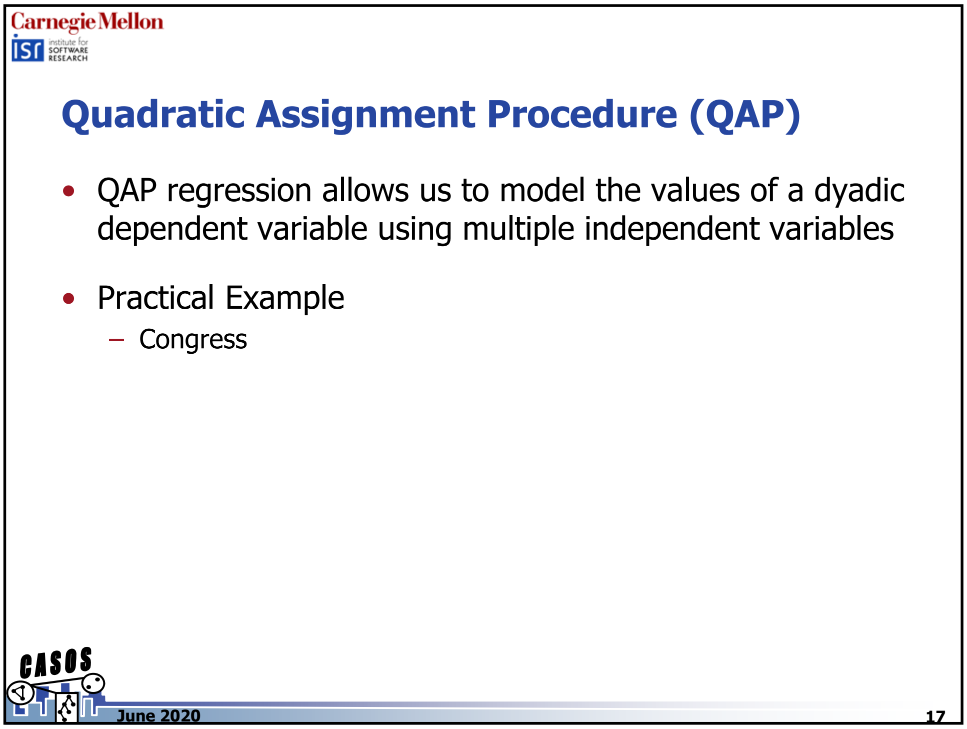
Quadratic Assignment Procedure (QAP)

- QAP correlation is designed to correlate entire matrices
- To calculate the significance, the method compares the observed correlation to a reference set of thousands of correlations
- To construct a p-value, it counts the proportion of the correlations that were as large as the observed correlation
- Compare the observed correlation against the distribution of correlations



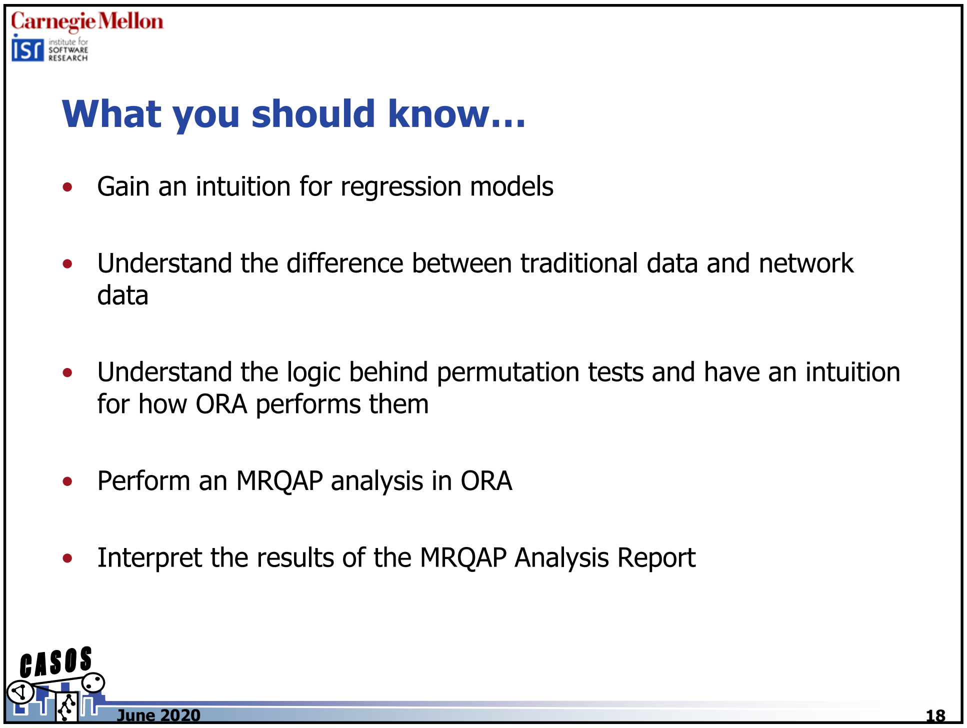
Permutation Tests

- The permutation test calculates all the ways that an experiment could have come out given the variables were in fact independent
- Counts the proportion of all assignments yielding a correlation as large as the one observed
  - this proportion indicates the ‘p-value’ or significance
- The number of permutations of N objects grows very quickly with N
- We sample uniformly from the space of all possible permutations (~20,000 permutations)



Quadratic Assignment Procedure (QAP)

- QAP regression allows us to model the values of a dyadic dependent variable using multiple independent variables
- Practical Example
  - Congress



What you should know…

- Gain an intuition for regression models
- Understand the difference between traditional data and network data
- Understand the logic behind permutation tests and have an intuition for how ORA performs them
- Perform an MRQAP analysis in ORA
- Interpret the results of the MRQAP Analysis Report