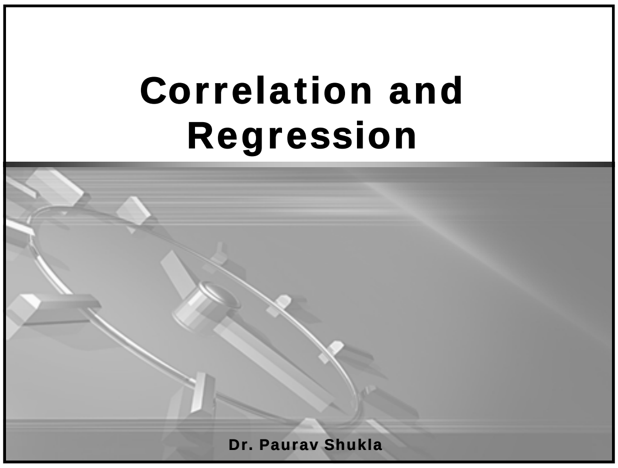
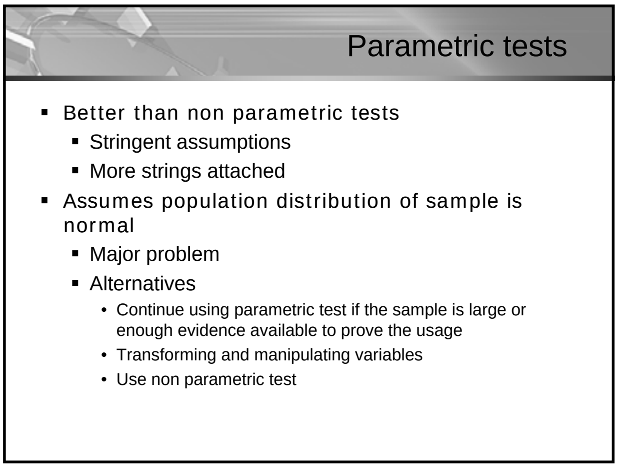


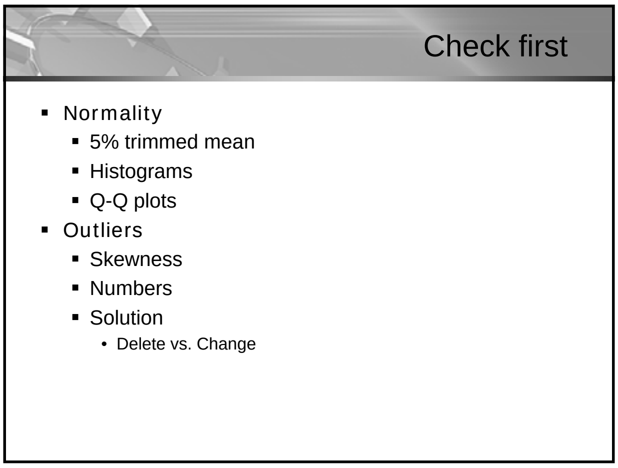
Correlation and Regression

Dr. Paurav Shukla



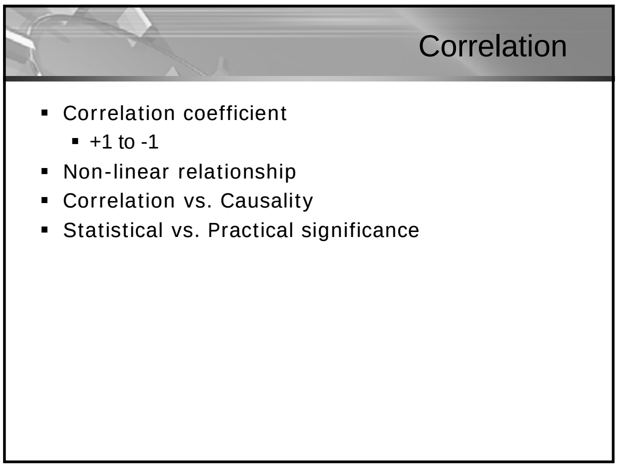
Parametric tests

- Better than non parametric tests
  - Stringent assumptions
  - More strings attached
- Assumes population distribution of sample is normal
  - Major problem
  - Alternatives
    - Continue using parametric test if the sample is large or enough evidence available to prove the usage
    - Transforming and manipulating variables
    - Use non parametric test



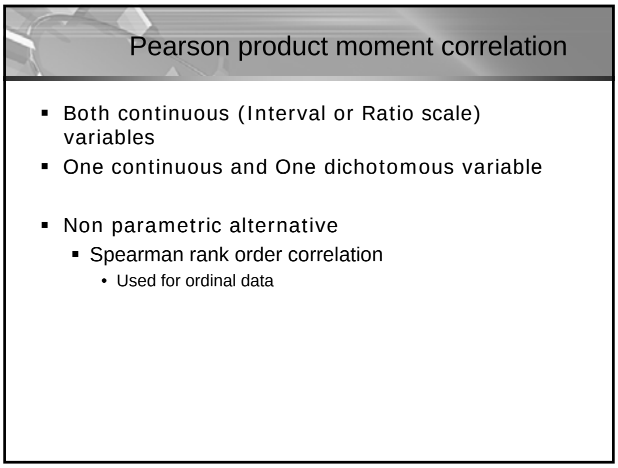
Check first

- Normality
  - 5% trimmed mean
  - Histograms
  - Q-Q plots
- Outliers
  - Skewness
  - Numbers
  - Solution
    - Delete vs. Change



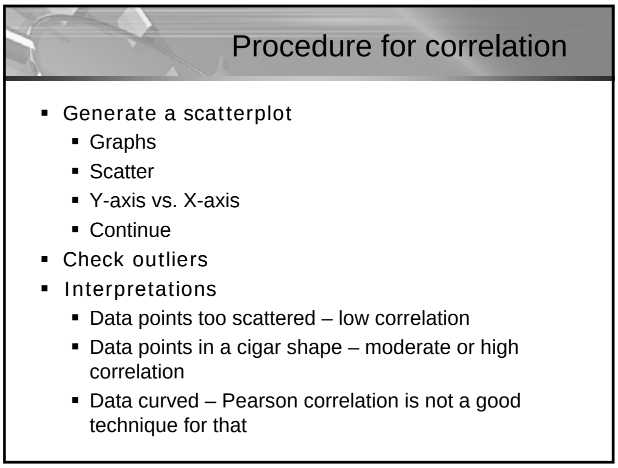
Correlation

- Correlation coefficient
  - +1 to -1
- Non-linear relationship
- Correlation vs. Causality
- Statistical vs. Practical significance



Pearson product moment correlation

- Both continuous (Interval or Ratio scale) variables
- One continuous and One dichotomous variable
- Non parametric alternative
  - Spearman rank order correlation
    - Used for ordinal data



Procedure for correlation

- Generate a scatterplot
  - Graphs
  - Scatter
  - Y-axis vs. X-axis
  - Continue
- Check outliers
- Interpretations
  - Data points too scattered – low correlation
  - Data points in a cigar shape – moderate or high correlation
  - Data curved – Pearson correlation is not a good technique for that

Procedure for correlation

- Analyze
- Correlate
- Bivariate
  - Pearson
    - One tail (for specific direction) vs. Two tail tests (no specific direction)
  - Options
    - Exclude cases pairwise

Results of correlation

- Direction of correlation
- Strength of correlation (Cohen, 1988)
  - r = -+ .10 to -+ .29 small (weak)
  - r = -+ .30 to -+ .49 medium (moderate)
  - r = -+ .50 to -+ 1.0 large (strong)
- Coefficient of correlation (r^2)
- Significance level
  - Desired <0.05
  - Avoid interpretation as its very messy area when you have large sample

Cohen, J. (1988). Statistical power analysis for the behavioural sciences (2nd Ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

How to present correlation results?

- The relationship between (variable X) and (variable Y) was investigated using Pearson product-moment correlation coefficient. Preliminary analyses were performed to ensure no violation of the assumptions of normality and linearity. There was a (weak, moderate or strong) (positive or negative) correlation between the two variables (r = ______; n = ______; p<_____), with _____ (high, moderate or low) level of variable X associated with (high, moderate or low) level of variable Y.

Comparing groups for correlation

- Data editor mode is important
  - Data
  - Split file
  - Compare groups
- Don’t forget to get it back to the Analyse all cases, don’t create groups mode again after the analysis

Testing statistical significance of the diff between correlation coefficient

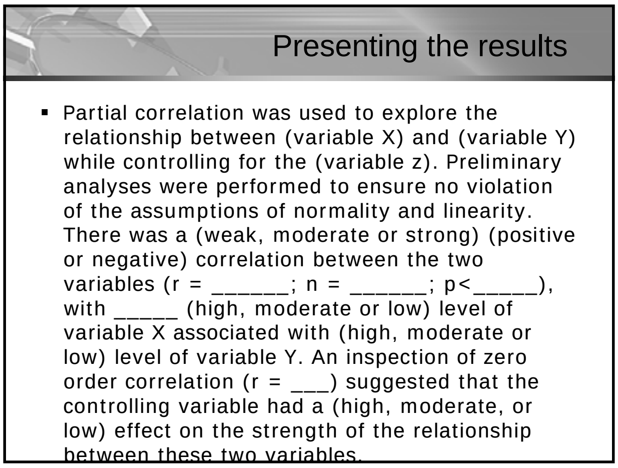
- Assumption
  - Random sample
  - Two groups of cases are independent
- Convert r value to z value
- Z observed

$$Z_{obs} = \frac{Z1 - Z2}{\sqrt{\frac{1}{N1-3} + \frac{1}{N2-3}}}$$

- Z observed > + - 1.96 (z value for 95% confidence level)
  - Statistically significant difference

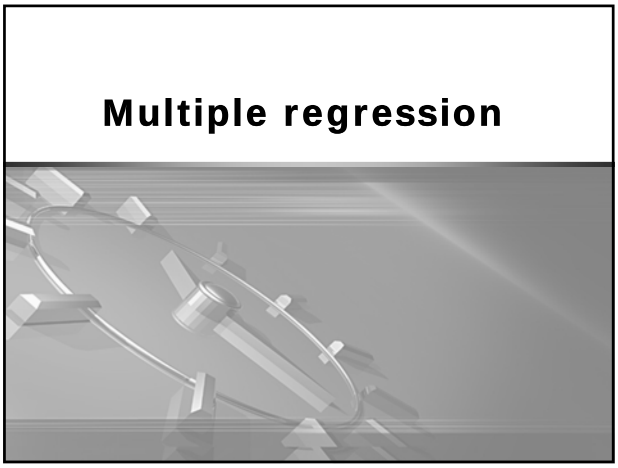
Partial correlation

- Involving a control variable
- Procedure
  - Analyze
  - Correlate
  - Partial
    - Options
      - Missing values
      - Exclude cases pairwise
    - Statistic
      - Zero order correlation
- Observation
  - If there is a significant difference in correlation ‘r’ the control variable is important

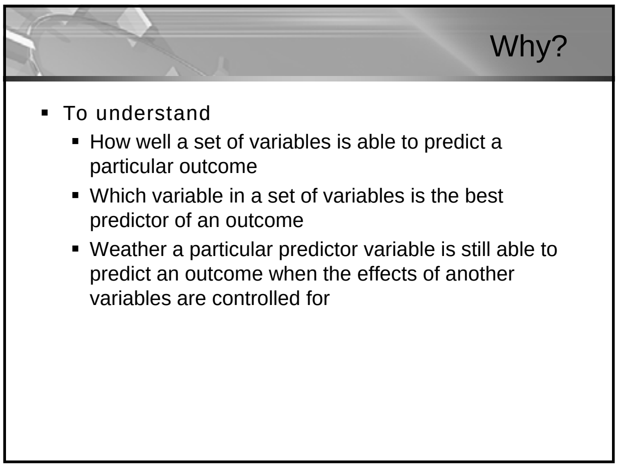


Presenting the results

- Partial correlation was used to explore the relationship between (variable X) and (variable Y) while controlling for the (variable z). Preliminary analyses were performed to ensure no violation of the assumptions of normality and linearity. There was a (weak, moderate or strong) (positive or negative) correlation between the two variables (r = ______; n = ______; p<_____), with _____ (high, moderate or low) level of variable X associated with (high, moderate or low) level of variable Y. An inspection of zero order correlation (r = ___) suggested that the controlling variable had a (high, moderate, or low) effect on the strength of the relationship between these two variables.

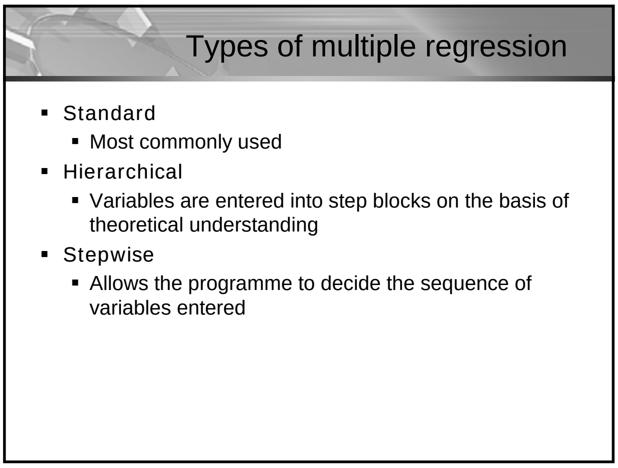


Multiple regression



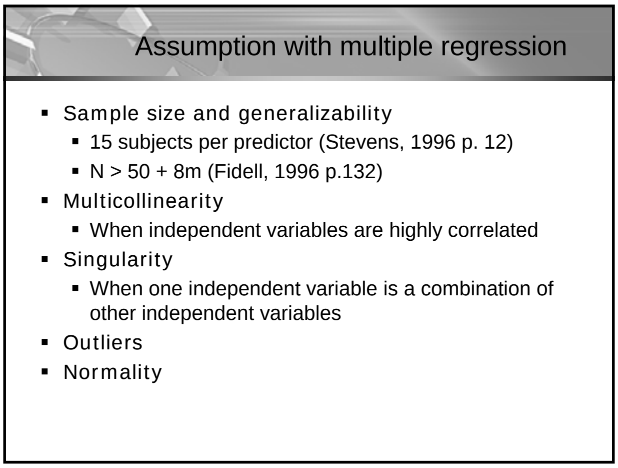
Why?

- To understand
  - How well a set of variables is able to predict a particular outcome
  - Which variable in a set of variables is the best predictor of an outcome
  - Weather a particular predictor variable is still able to predict an outcome when the effects of another variables are controlled for



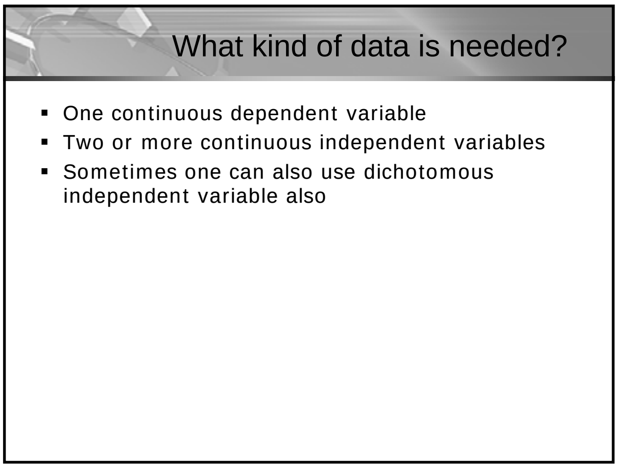
Types of multiple regression

- Standard
  - Most commonly used
- Hierarchical
  - Variables are entered into step blocks on the basis of theoretical understanding
- Stepwise
  - Allows the programme to decide the sequence of variables entered



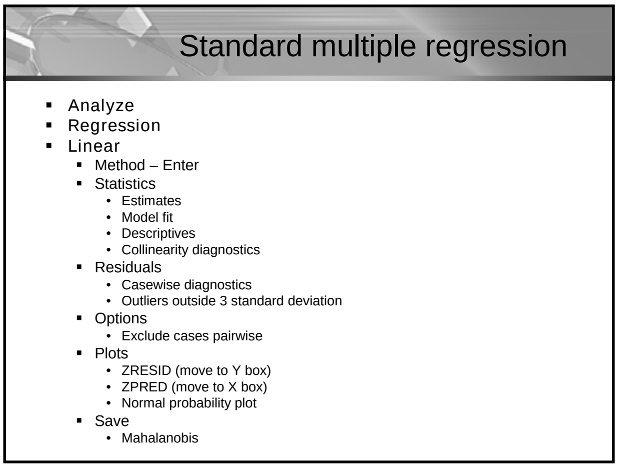
Assumption with multiple regression

- Sample size and generalizability
  - 15 subjects per predictor (Stevens, 1996 p. 12)
  - N > 50 + 8m (Fidell, 1996 p.132)
- Multicollinearity
  - When independent variables are highly correlated
- Singularity
  - When one independent variable is a combination of other independent variables
- Outliers
- Normality



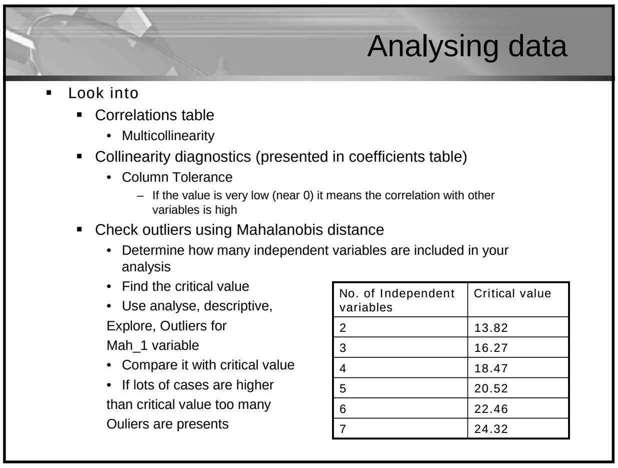
What kind of data is needed?

- One continuous dependent variable
- Two or more continuous independent variables
- Sometimes one can also use dichotomous independent variable also



Standard multiple regression

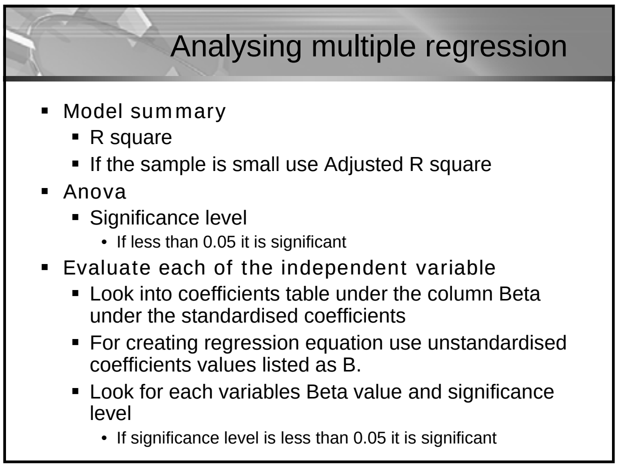
- Analyze
- Regression
- Linear
  - Method – Enter
  - Statistics
    - Estimates
    - Model fit
    - Descriptives
    - Collinearity diagnostics
  - Residuals
    - Casewise diagnostics
    - Outliers outside 3 standard deviation
  - Options
    - Exclude cases pairwise
  - Plots
    - ZRESID (move to Y box)
    - ZPRED (move to X box)
    - Normal probability plot
  - Save
    - Mahalanobis



Analysing data

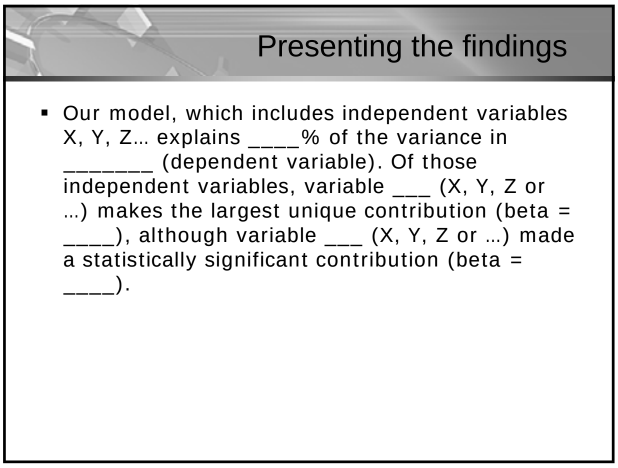
- Look into
  - Correlations table
    - Multicollinearity
  - Collinearity diagnostics (presented in coefficients table)
    - Column Tolerance
      - If the value is very low (near 0) it means the correlation with other variables is high
  - Check outliers using Mahalanobis distance
    - Determine how many independent variables are included in your analysis
    - Find the critical value
    - Use analyse, descriptive, Explore, Outliers for Mah_1 variable
    - Compare it with critical value
    - If lots of cases are higher than critical value too many Ouliers are presents

No. of Independent variables	Critical value
2	13.82
3	16.27
4	18.47
5	20.52
6	22.46
7	24.32



Analysing multiple regression

- Model summary
  - R square
  - If the sample is small use Adjusted R square
- Anova
  - Significance level
    - If less than 0.05 it is significant
- Evaluate each of the independent variable
  - Look into coefficients table under the column Beta under the standardised coefficients
  - For creating regression equation use unstandardised coefficients values listed as B.
  - Look for each variables Beta value and significance level
    - If significance level is less than 0.05 it is significant



Presenting the findings

- Our model, which includes independent variables X, Y, Z… explains ____% of the variance in _______ (dependent variable). Of those independent variables, variable ___ (X, Y, Z or …) makes the largest unique contribution (beta = ____), although variable ___ (X, Y, Z or …) made a statistically significant contribution (beta = ____).