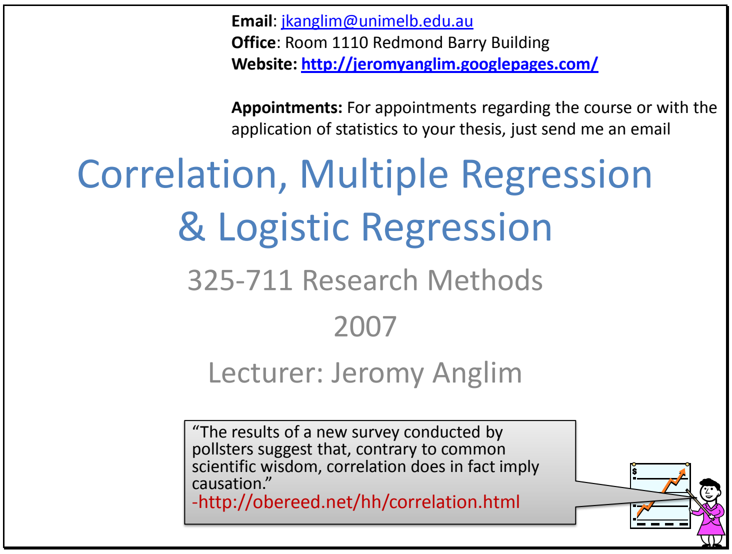
Slide 1



Email: jkanglim@unimelb.edu.au
Office: Room 1110 Redmond Barry Building
Website: http://jeromyanglim.googlepages.com/

Appointments: For appointments regarding the course or with the application of statistics to your thesis, just send me an email

Correlation, Multiple Regression & Logistic Regression

325-711 Research Methods

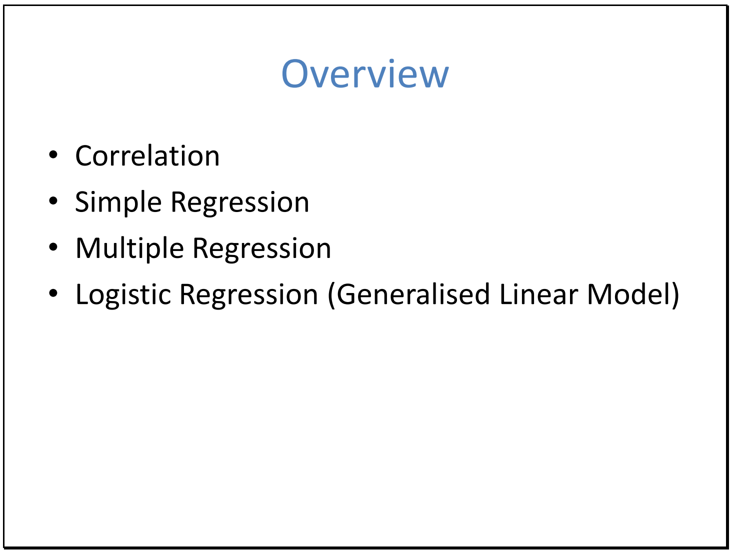
2007

Lecturer: Jeromy Anglim

"The results of a new survey conducted by pollsters suggest that, contrary to common scientific wisdom, correlation does in fact imply causation."
-http://obereed.net/hh/correlation.html

DESCRIPTION
This session will provide a brief overview of correlational analyses including correlations, multiple regression, and logistic regression. Regression equations are used when the researcher wants to use one or more independent variables to predict a metric scaled dependent variable. Logistic regression is used when the researcher wants to use one or more independent variables to predict a dichotomous dependent variable. We will review the assumptions of multiple and logistic regression, review the statistical output generated by regression analyses (in SPSS) in order to interpret the results, and discuss how to report regression analyses and results in a research article. The use of categorical independent variables and hierarchical regression will also be briefly discussed.

Slide 2

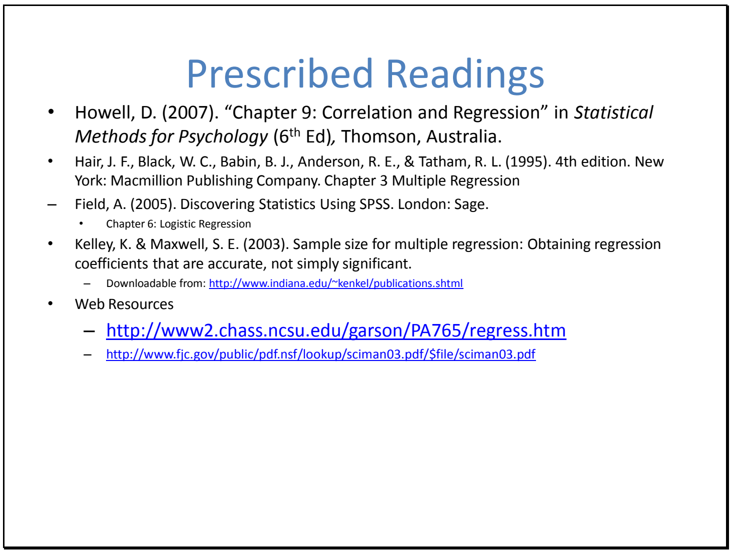


Overview

- Correlation
- Simple Regression
- Multiple Regression
- Logistic Regression (Generalised Linear Model)

Data analysis is a cumulative skill. Techniques like logistic regression (generalised linear model) are extensions of multiple regression (general linear model). Understanding multiple regression is based on an understanding of simple regression. All these techniques rely on an understanding of correlation, sums of squares, variance and other introductory statistics concepts.

Slide 3



Prescribed Readings

- Howell, D. (2007). “Chapter 9: Correlation and Regression” in *Statistical Methods for Psychology* (6th Ed)*,* Thomson, Australia.
- Hair, J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (1995). 4th edition. New York: Macmillion Publishing Company. Chapter 3 Multiple Regression

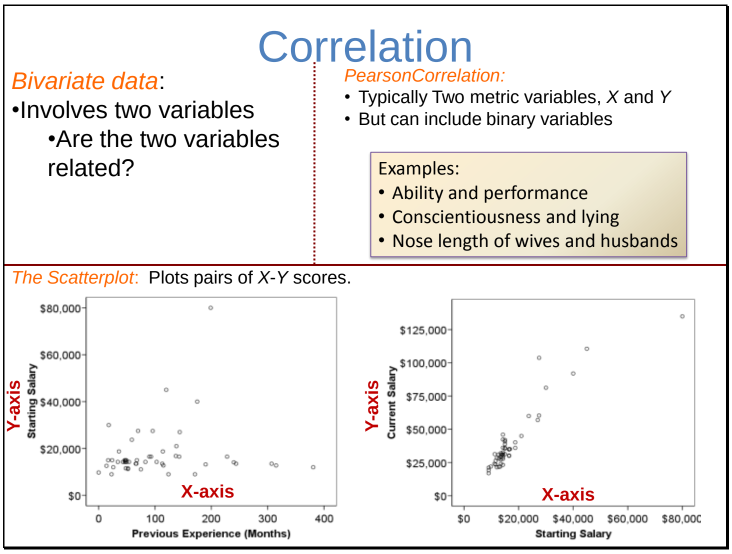
– Field, A. (2005). Discovering Statistics Using SPSS. London: Sage.

  - Chapter 6: Logistic Regression

- Kelley, K. & Maxwell, S. E. (2003). Sample size for multiple regression: Obtaining regression coefficients that are accurate, not simply significant.
  - Downloadable from: http://www.indiana.edu/~kenkel/publications.shtml
- Web Resources
  - http://www2.chass.ncsu.edu/garson/PA765/regress.htm
  - http://www.fjc.gov/public/pdf.nsf/lookup/sciman03.pdf/$file/sciman03.pdf

Howell, D. (2007): This reading provides an understanding of some of the mathematics behind multiple regression without describing matrix algebra.
Howell (2007): Howell provides a comprehensive introduction to correlation, covariance and simple regression. This text is particularly useful if you need to perform certain types of significance tests on correlations not readily found in statistical packages such as: 1) testing whether two correlations in different samples are different; 2) testing whether a correlation is significantly different from a specified correlation; 3) testing whether two correlations in a nonindependent sample (i.e., X1 with Y versus X2 with Y) are significantly different.
http://www2.chass.ncsu.edu/garson/PA765/regress.htm A comprehensive website on many statistical techniques including multiple regression. It also includes annotated SPSS output to many techniques including multiple regression.
http://www.fjc.gov/public/pdf.nsf/lookup/sciman03.pdf/$file/sciman03.pdf An interesting non-technical overview of multiple regression written for those trying to evaluate it as evidence in the court room.

Slide 4



Correlation

Bivariate data:

•Involves two variables

•Are the two variables related?

PearsonCorrelation:

- Typically Two metric variables, *X* and *Y*
- But can include binary variables

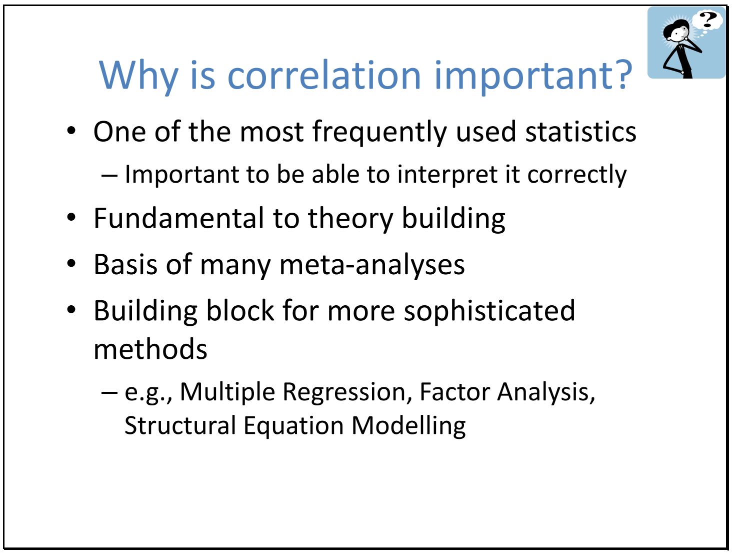
Examples:

- Ability and performance
- Conscientiousness and lying
- Nose length of wives and husbands

The Scatterplot: Plots pairs of *X*-*Y* scores.

We often ask research questions regarding the relationship between two variables. Even when we are interested in more complex relationships between more two variables, it is important not to lose site of bivariate relationships.
There are many different measures for describing association between two variables. The main focus of this seminar is on Pearson's correlation

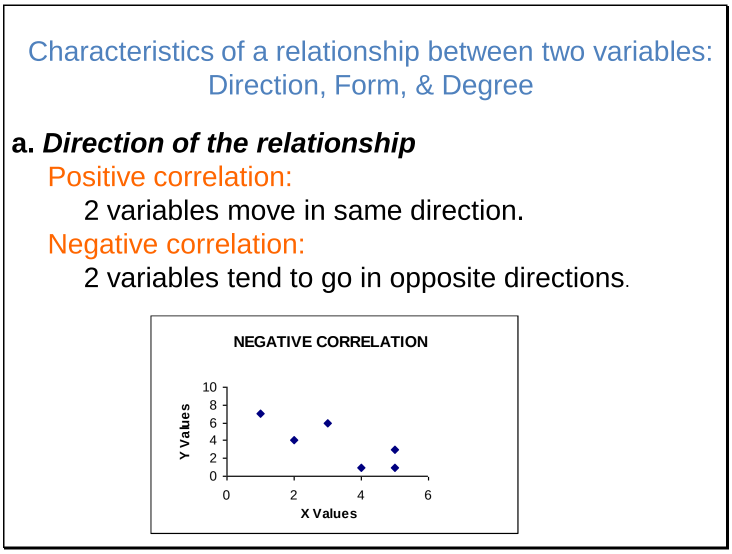
Slide 5



Why is correlation important?

- One of the most frequently used statistics
  – Important to be able to interpret it correctly
- Fundamental to theory building
- Basis of many meta-analyses
- Building block for more sophisticated methods
  – e.g., Multiple Regression, Factor Analysis, Structural Equation Modelling

Slide 6



Characteristics of a relationship between two variables: Direction, Form, & Degree

a. ***Direction of the relationship***

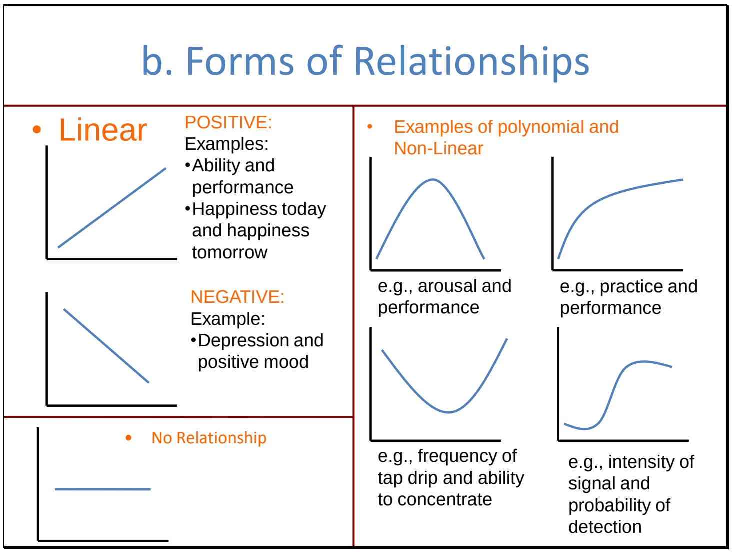
Positive correlation:

2 variables move in same direction.

Negative correlation:

2 variables tend to go in opposite directions.

Slide 7



b. Forms of Relationships

- Linear

POSITIVE:
Examples:
- Ability and performance
- Happiness today and happiness tomorrow

NEGATIVE:
Example:
- Depression and positive mood

- No Relationship

- Examples of polynomial and Non-Linear

e.g., arousal and performance

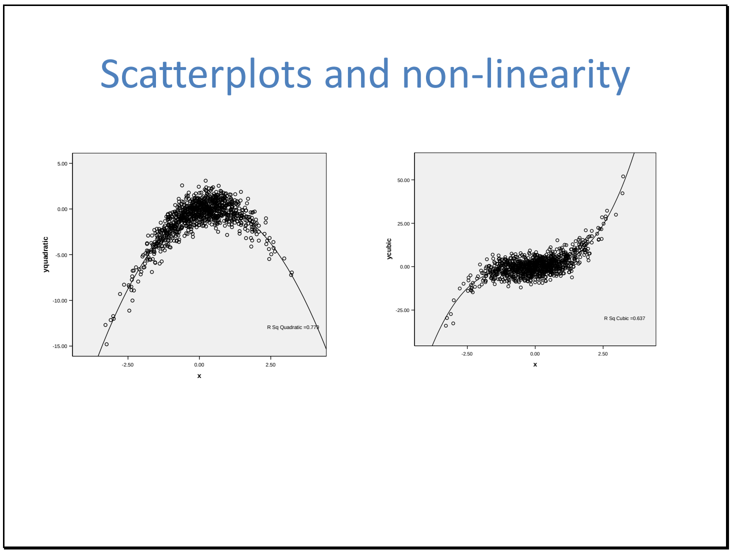
e.g., practice and performance

e.g., frequency of tap drip and ability to concentrate

e.g., intensity of signal and probability of detection

There are many different types of non-linear relationships. These include the higher order polynomial trends such as quadratic, cubic and quartic.

Slide 8

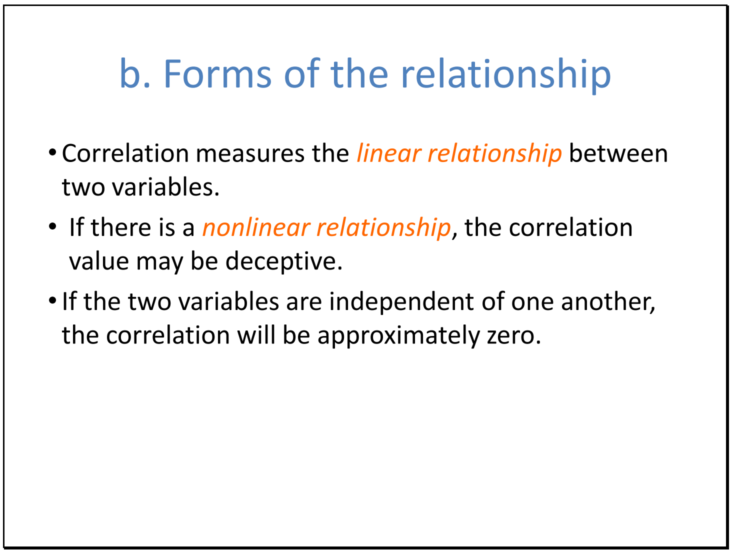


Scatterplots and non-linearity

The standard pearson's correlation describes the linear relationship between two variables. A linear relationship is one where an increase in oSne variable is associated with an increase or decrease in another variable.

It is important to check whether the scatterplots of the correlations you report to verify that the assumption holds, particularly if there are theoretical reasons to suspect a non-linear relationship. Above we see examples of quadratic and cubic relationships. These are somewhat idealised and often non-linear trends in the social sciences are somewhat more subtle. There are also statistical ways of examining for non-linear relationships using polynomial regression and non-linear regression.

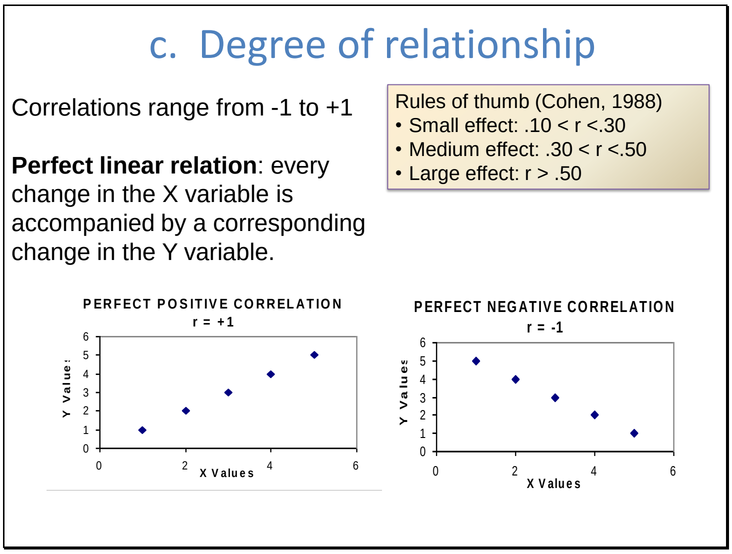
Slide 9



b. Forms of the relationship

- Correlation measures the *linear relationship* between two variables.
- If there is a *nonlinear relationship*, the correlation value may be deceptive.
- If the two variables are independent of one another, the correlation will be approximately zero.

Slide 10



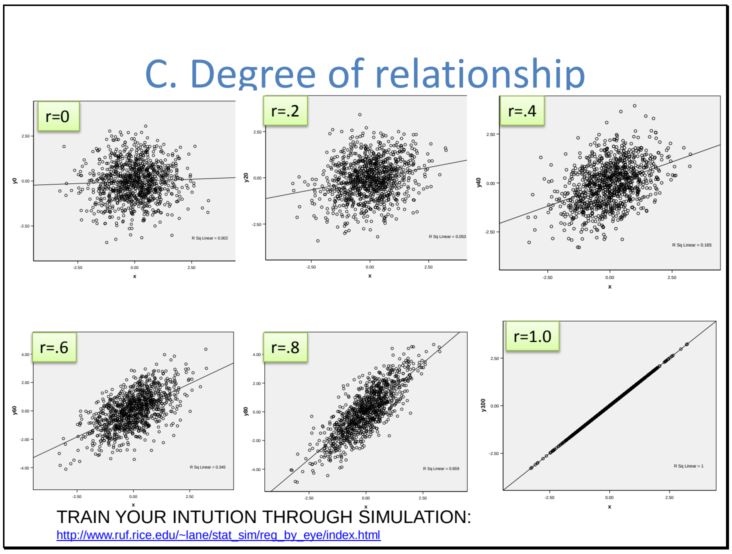
c. Degree of relationship

Correlations range from -1 to +1

Perfect linear relation: every change in the X variable is accompanied by a corresponding change in the Y variable.

Rules of thumb (Cohen, 1988)
- Small effect: .10 < r <.30
- Medium effect: .30 < r <.50
- Large effect: r > .50

Slide 11

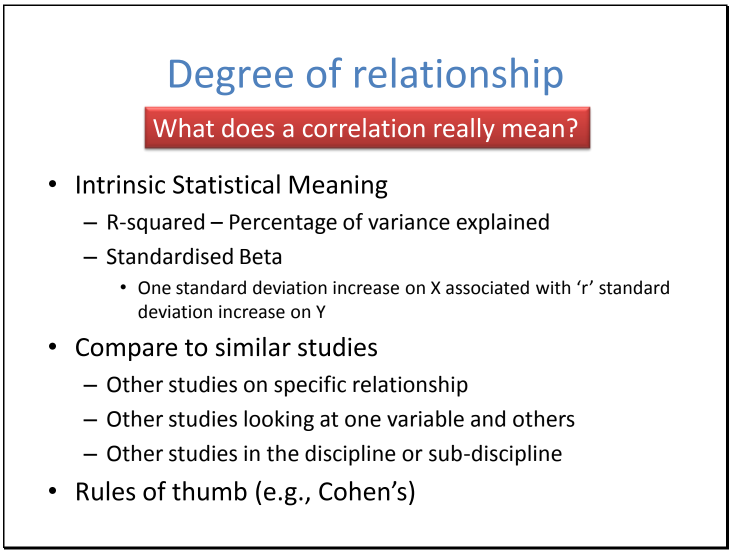


C. Degree of relationship

TRAIN YOUR INTUTION THROUGH SIMULATION:
http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html

It is useful to develop your intuition about the relationship between a scatterplot and a correlation. An excellent Simulation to train your intuition on the relationship between a scatter plot and a correlation coefficient is available at the following website http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html

Slide 12



Degree of relationship

What does a correlation really mean?

- Intrinsic Statistical Meaning
  - R-squared – Percentage of variance explained
  - Standardised Beta
    - One standard deviation increase on X associated with ‘r’ standard deviation increase on Y
- Compare to similar studies
  - Other studies on specific relationship
  - Other studies looking at one variable and others
  - Other studies in the discipline or sub-discipline
- Rules of thumb (e.g., Cohen’s)

There is currently a big push to start reporting effect sizes. However, merely reporting some form of effect size measures such as a correlation is only the beginning. It is more important to start “thinking” in terms of effect size. Thinking in binary terms of there is or is not a relationship between two variables is inadequate for most purposes. The relative size of a correlation is a theoretically meaningful statement. Knowing whether a correlation between two variables is .1, .2, .3, .5, .7, .8 or .9 is theoretically meaningful.

Example of the consequences: If we were trying to decide on a test for a selection and recruitment context, the difference in productivity gains and cost savings to an organisation of using a test with a .5 correlation as opposed to one with a .25 correlation could be massive.

THE CHALLENGE: So what does a correlation really mean? How can we align this mathematical parameter with our conceptual understanding of the world?

There are several broad approaches to this. The key theme here is training your intuition and integrating your conceptual thinking with knowledge of statistical parameters.

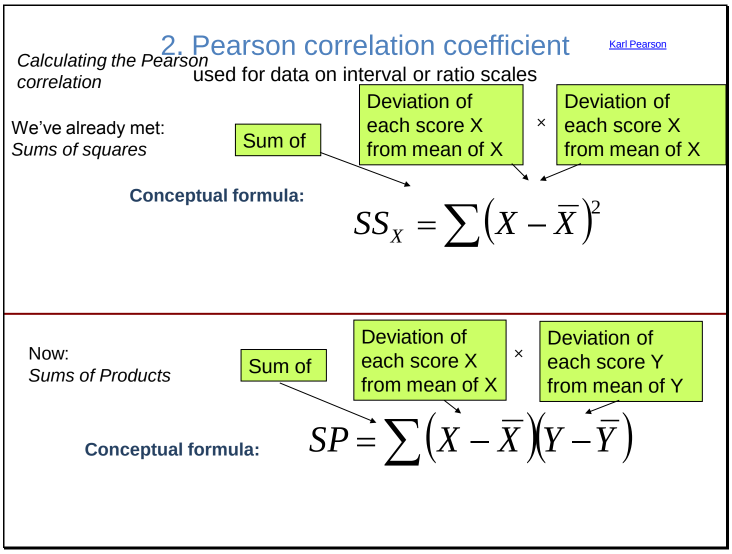
R-squared: Multiply the correlation by itself gives what is called the “coefficient of determination” or “r-squared”. It has the intuitive meaning of the percentage of variance in one variable explained by, or shared with, the other variable.

A simple example where we have a .8 correlation yields a r squared of .64.

- $r = 0.80$
- $r^2 = 0.80 \times 0.80 = 0.64$
- 64% of the variability in the Y scores can be predicted from the relation with X.

Standardised beta: A correlation is the same as a standardised beta in simple regression.

Slide 13



2. Pearson correlation coefficient

Karl Pearson

Calculating the Pearson correlation

used for data on interval or ratio scales

We've already met:
Sums of squares

Sum of

Deviation of each score X from mean of X

×

Deviation of each score X from mean of X

Conceptual formula:

$$SS_X = \sum \left(X - \bar{X}\right)^2$$

Now:
Sums of Products

Sum of

Deviation of each score X from mean of X

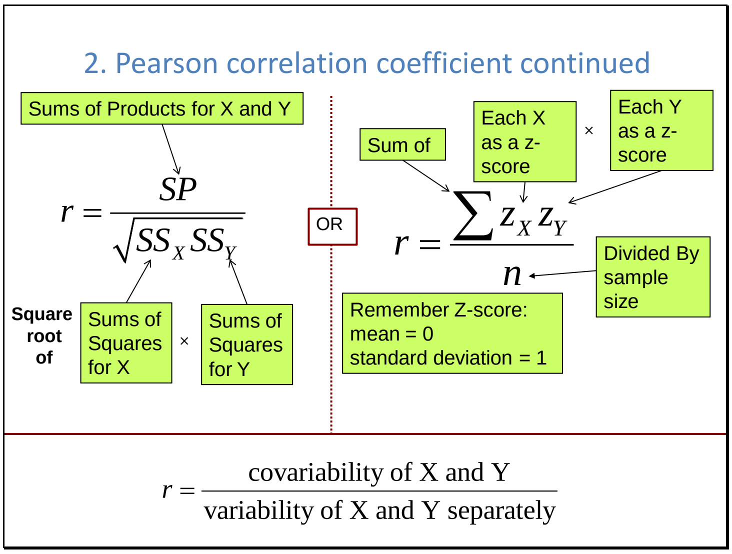
×

Deviation of each score Y from mean of Y

Conceptual formula:

$$SP = \sum \left(X - \bar{X}\right)\left(Y - \bar{Y}\right)$$

Slide 14



2. Pearson correlation coefficient continued

Sums of Products for X and Y

$$r = \frac{SP}{\sqrt{SS_X SS_Y}}$$

Square root of

Sums of Squares for X

×

Sums of Squares for Y

OR

Sum of

Each X as a z-score

×

Each Y as a z-score

$$r = \frac{\sum z_X z_Y}{n}$$

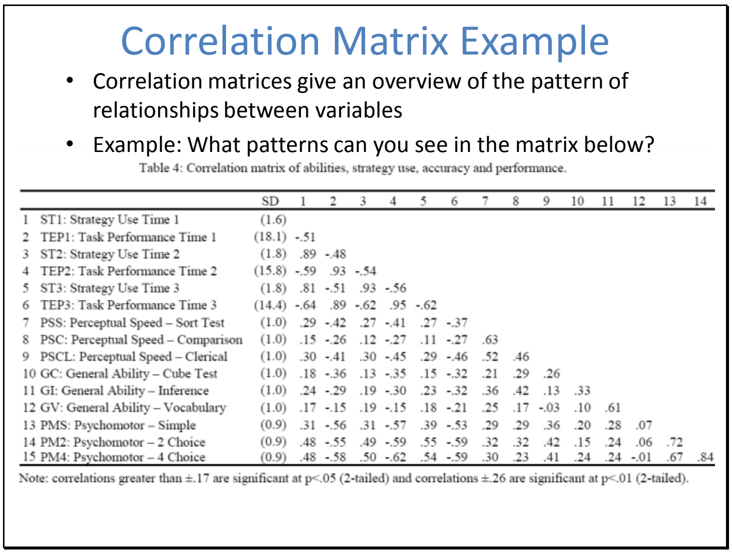
Divided By sample size

Remember Z-score:
mean = 0
standard deviation = 1

$$r = \frac{\text{covariability of X and Y}}{\text{variability of X and Y separately}}$$

To improve your intuition of these formulas, make sure you work through some practice questions; see how increasing one quantity effects the result.

Slide 15



Correlation Matrix Example

- Correlation matrices give an overview of the pattern of relationships between variables
- Example: What patterns can you see in the matrix below?

Table 4: Correlation matrix of abilities, strategy use, accuracy and performance.

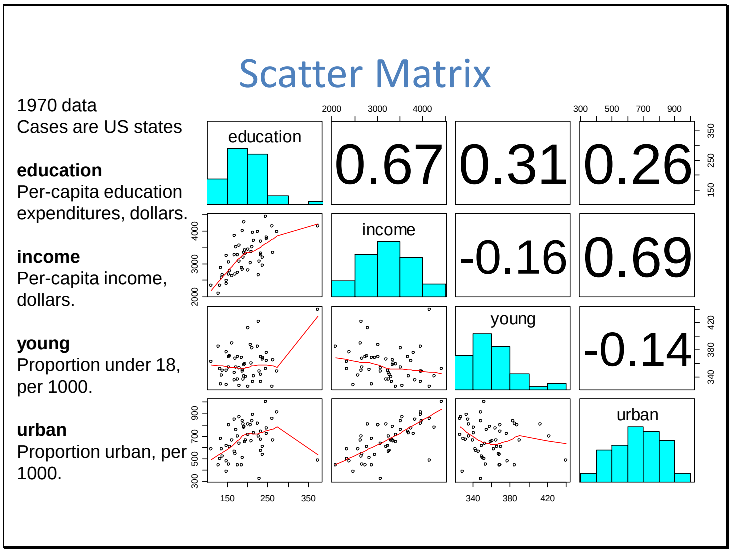
		SD	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	ST1: Strategy Use Time 1	(1.6)														
2	TEP1: Task Performance Time 1	(18.1)	-.51													
3	ST2: Strategy Use Time 2	(1.8)	.89	-.48												
4	TEP2: Task Performance Time 2	(15.8)	-.59	.93	-.54											
5	ST3: Strategy Use Time 3	(1.8)	.81	-.51	.93	-.56										
6	TEP3: Task Performance Time 3	(14.4)	-.64	.89	-.62	.95	-.62									
7	PSS: Perceptual Speed – Sort Test	(1.0)	.29	-.42	.27	-.41	.27	-.37								
8	PSC: Perceptual Speed – Comparison	(1.0)	.15	-.26	.12	-.27	.11	-.27	.63							
9	PSCL: Perceptual Speed – Clerical	(1.0)	.30	-.41	.30	-.45	.29	-.46	.52	.46						
10	GC: General Ability – Cube Test	(1.0)	.18	-.36	.13	-.35	.15	-.32	.21	.29	.26					
11	GI: General Ability – Inference	(1.0)	.24	-.29	.19	-.30	.23	-.32	.36	.42	.13	.33				
12	GV: General Ability – Vocabulary	(1.0)	.17	-.15	.19	-.15	.18	-.21	.25	.17	-.03	.10	.61			
13	PMS: Psychomotor – Simple	(0.9)	.31	-.56	.31	-.57	.39	-.53	.29	.29	.36	.20	.28	.07		
14	PM2: Psychomotor – 2 Choice	(0.9)	.48	-.55	.49	-.59	.55	-.59	.32	.32	.42	.15	.24	.06	.72	
15	PM4: Psychomotor – 4 Choice	(0.9)	.48	-.58	.50	-.62	.54	-.59	.30	.23	.41	.24	.24	-.01	.67	.84

Note: correlations greater than ±.17 are significant at p<.05 (2-tailed) and correlations ±.26 are significant at p<.01 (2-tailed).

In most situations, I would consider a correlation matrix optionally with means and standard deviations in the first two columns and reliability coefficients on the diagonal as one of the most important bits of output to report in any journal article or thesis. Before presenting more complex modeling efforts, it gives the reader a sense of the bivariate relationships between the core variables in your study.

Anglim, J., Langan-Fox, J., & Mahdavi, N. (2005). Modeling the Relationship between Strategies, Abilities and Skilled Performance. CogSci 2005, 27th Annual Meeting of the Cognitive Science Society, July 21-23 Stresa, Italy. web ref: www.psych.unito.it/csc/cogsci05/frame/poster/3/p465-anglim.pdf

Slide 16



Scatter Matrix

1970 data
Cases are US states

education
Per-capita education expenditures, dollars.

income
Per-capita income, dollars.

young
Proportion under 18, per 1000.

urban
Proportion urban, per 1000.

A scatter matrix is a way of showing the bivariate relationship between a set of metric variables.

SPSS has a graph called matrix-scatterplot which will shows the scatterplots for all pairs of a set of variables.

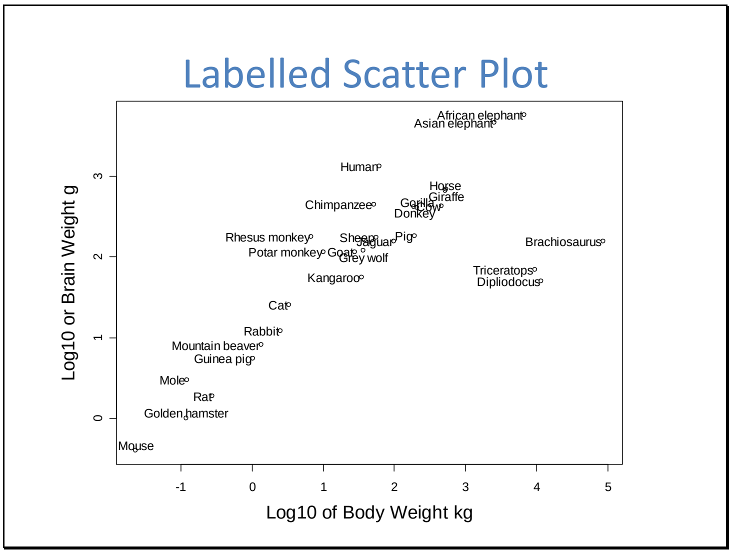
However, this particular graph has a lot more going on and is far superior to SPSS's matrix scatterplot. It shows the distribution through a histogram, the actual correlation coefficient, and loess regression line which assists in determining any non-linearity in the relationship. It answers so many important questions all at once. it was produced with R using this single command:

```
psych::pairs.panels(Anscombe)
```

"The Anscombe data frame has 51 rows and 4 columns. The observations are the U. S. states plus Washington, D. C. in 1970. "

"education Per-capita education expenditures, dollars.   income Per-capita income, dollars. young Proportion under 18, per 1000.   urban Proportion urban, per 1000." Data taken from the CAR package in R - C:\Program Files\R\R-2.5.0\library\car\html\Anscombe.html

Slide 17

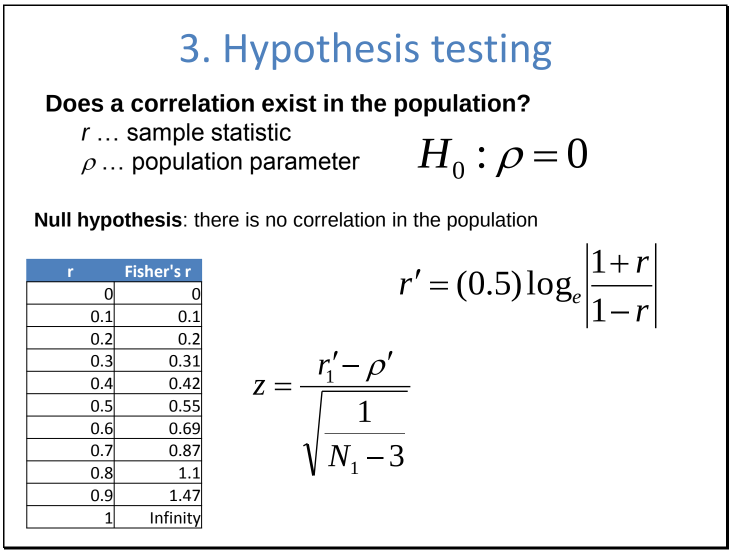


Labelled Scatter Plot

A labelled scatterplot can be particularly powerful for showing the relationship between two variables when the cases themselves are of intrinsic interest. It is sometimes used to show the results of factor analysis of multidimensional scaling solutions, where the points are variables (factor analysis) or other objects of interest (MDS). SPSS has a label cases option in its scatter plot dialog box.

The data shows average brain and body weights for 28 species of land animals. It was taken from the MASS package in R: which in turn reports obtaining the data from P. J. Rousseeuw and A. M. Leroy (1987) *Robust Regression and Outlier Detection.* Wiley, p. 57.

Slide 18



3. Hypothesis testing

Does a correlation exist in the population?

r … sample statistic

ρ … population parameter

$$H_0 : \rho = 0$$

Null hypothesis: there is no correlation in the population

r	Fisher's r
0	0
0.1	0.1
0.2	0.2
0.3	0.31
0.4	0.42
0.5	0.55
0.6	0.69
0.7	0.87
0.8	1.1
0.9	1.47
1	Infinity

$$r' = (0.5)\log_e \left| \frac{1+r}{1-r} \right|$$

$$z = \frac{r_1' - \rho'}{\sqrt{\dfrac{1}{N_1 - 3}}}$$

Greek letters to describe population parameters; ρ ('Rho') indicates population correlation

Formulas taken from Howell 6th Edition (pp 259-260)

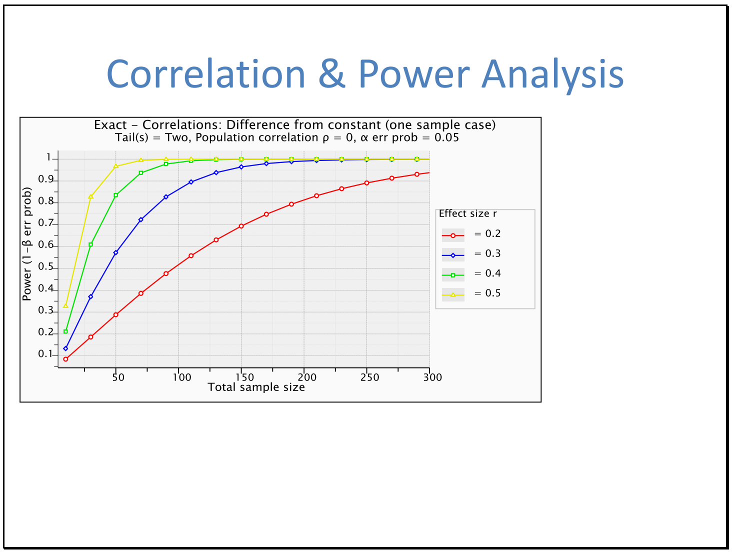
The process for testing for statistical significance of a correlation can be done using SPSS. However, there is some value in understanding how the test of statistical significance is derived. In particular knowledge of the process can assist when wanting to test more complex tests of statistical significance relating to correlations, which are not always readily available in statistical packages.

Step 1: convert r to Fisher's r using the formula half of the natural log of the absolute value of one plus r over one minus r.

Step 2: Work out the corresponding z-score based on the difference between the samle fisher's r and the null hypothesis fisher's r. The null hypothesis fisher's r will usually be zero. This is then divided by the square root of one divided by the sample size minus three.

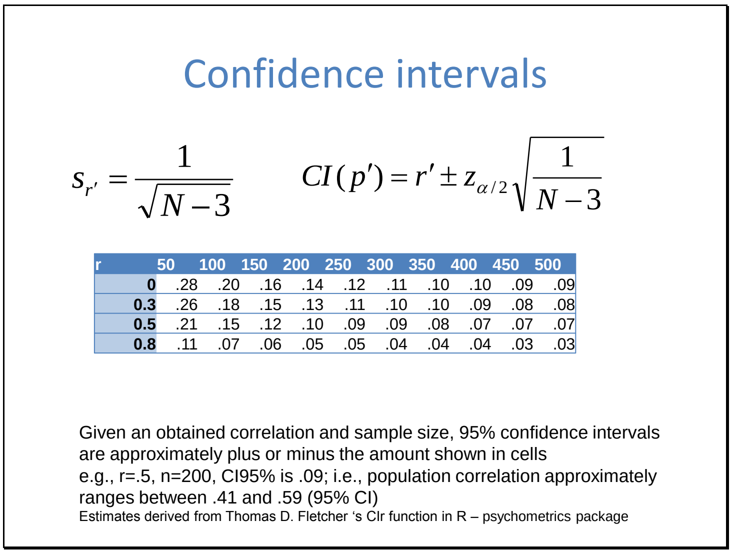
Step 3: Look up the obtained z value in a table of the normal distribution.

Slide 19



Correlation & Power Analysis

Slide 20



Confidence intervals

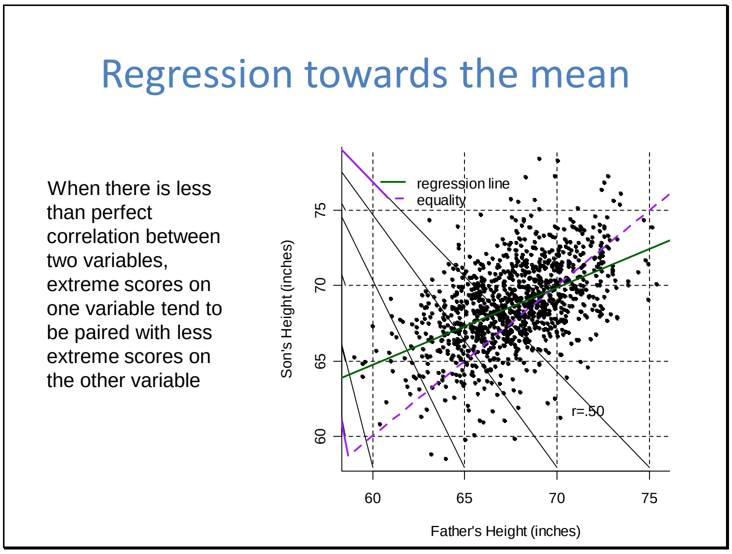
$$s_{r'} = \frac{1}{\sqrt{N-3}} \qquad CI(p') = r' \pm z_{\alpha/2}\sqrt{\frac{1}{N-3}}$$

r	50	100	150	200	250	300	350	400	450	500
0	.28	.20	.16	.14	.12	.11	.10	.10	.09	.09
0.3	.26	.18	.15	.13	.11	.10	.10	.09	.08	.08
0.5	.21	.15	.12	.10	.09	.09	.08	.07	.07	.07
0.8	.11	.07	.06	.05	.05	.04	.04	.04	.03	.03

Given an obtained correlation and sample size, 95% confidence intervals are approximately plus or minus the amount shown in cells
e.g., r=.5, n=200, CI95% is .09; i.e., population correlation approximately ranges between .41 and .59 (95% CI)
Estimates derived from Thomas D. Fletcher 's CIr function in R – psychometrics package

The Power Analysis approach has value in estimating our chance of detecting a statistically significant correlation. However, in many areas of study, research has moved on from the question of whether there is a relationship, to the question of what is the strength of the relationship. In this case, we may be more interested in getting a sample size that has sufficiently accurate confidence intervals around the correlation coefficient. The above table gives you a feel what kind of confidence you can expect given a particular correlation and sample size. This represents the use of confidence intervals as an a priori tool for deciding on an appropriate sample size.
From a post hoc perspective we can also use confidence intervals to estimate our confidence in knowing the population effect size from our sample.
THE FORMULA: The above formula can be used to determine confidence intervals on a correlation.

Slide 21



Regression towards the mean

When there is less than perfect correlation between two variables, extreme scores on one variable tend to be paired with less extreme scores on the other variable

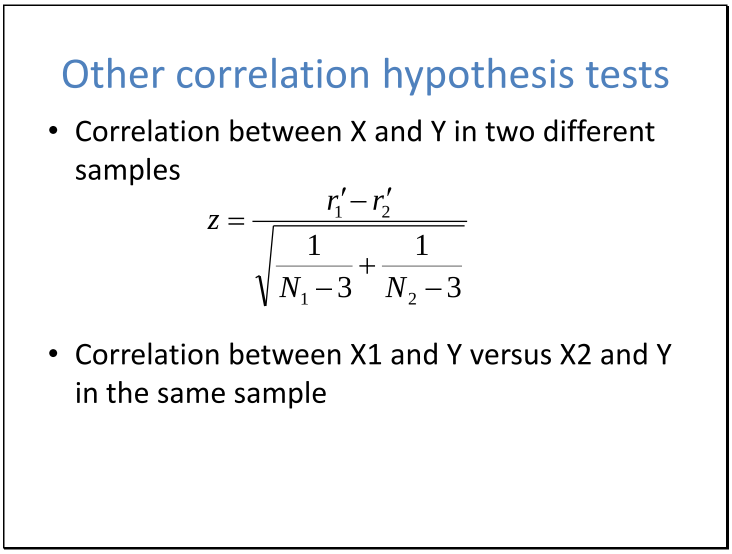
Dataset for Graph: Source: Recommendation from UsingR and from http://stat-www.berkeley.edu/users/juliab/141C/pearson.dat – 1078 measurements of father and son heights.

Examples:

Parent and Child Height: X=Dad's Height; Y=Son's Height; Correlation between X and Y is positive but imperfect. Assuming no overall increase in heights across generations, tall dads will tend to have sons who are shorter than them, but the sons will still tend to be taller than average. In the graph above the effect is not so great because in this sample the sons were on average taller than their dads, but even here, we see that Dad's that were taller than about 70 inches (6foot) tended to on average have sons that were shorter than them.

Pre-test Post-test studies: X=Ability at baseline; Y=Ability after intervention; Correlation between X and Y is positive but imperfect. Give ability test to 1000 children and assign bottom 20 to special program; bottom 20 children will tend to improve regardless of effectiveness of program

Slide 22



Other correlation hypothesis tests

- Correlation between X and Y in two different samples

$$z = \frac{r_1' - r_2'}{\sqrt{\frac{1}{N_1 - 3} + \frac{1}{N_2 - 3}}}$$

- Correlation between X1 and Y versus X2 and Y in the same sample

Correlation between X and Y in two different samples

Example: Do intelligence tests predict job performance equally well for different racial groups? Is the relationship between pay and performance the same for males and females?

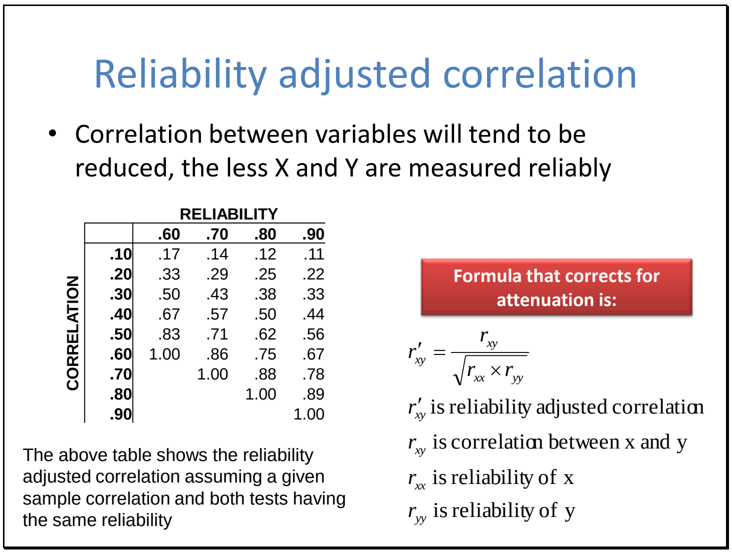
Formula taken from Howell 6th Edition (p.259)

Correlation between X1 and Y versus X2 and Y in the same sample

See Page 261-262 for a discussion of the situation where the two correlations come from the same sample

Examples: Do personality or ability tests correlate more with job performance?

Slide 23



Reliability adjusted correlation

- Correlation between variables will tend to be reduced, the less X and Y are measured reliably

		RELIABILITY			
		.60	.70	.80	.90
CORRELATION	.10	.17	.14	.12	.11
	.20	.33	.29	.25	.22
	.30	.50	.43	.38	.33
	.40	.67	.57	.50	.44
	.50	.83	.71	.62	.56
	.60	1.00	.86	.75	.67
	.70		1.00	.88	.78
	.80			1.00	.89
	.90				1.00

The above table shows the reliability adjusted correlation assuming a given sample correlation and both tests having the same reliability

Formula that corrects for attenuation is:

$$r'_{xy} = \frac{r_{xy}}{\sqrt{r_{xx} \times r_{yy}}}$$

r'_{xy} is reliability adjusted correlation

r_{xy} is correlation between x and y

r_{xx} is reliability of x

r_{yy} is reliability of y

Reliability Formula: See Murphy, K. R., & Davidshofer, C. O. (1998). Psychological Testing: Principles and Applications. Prentice Hall, New Jersey.

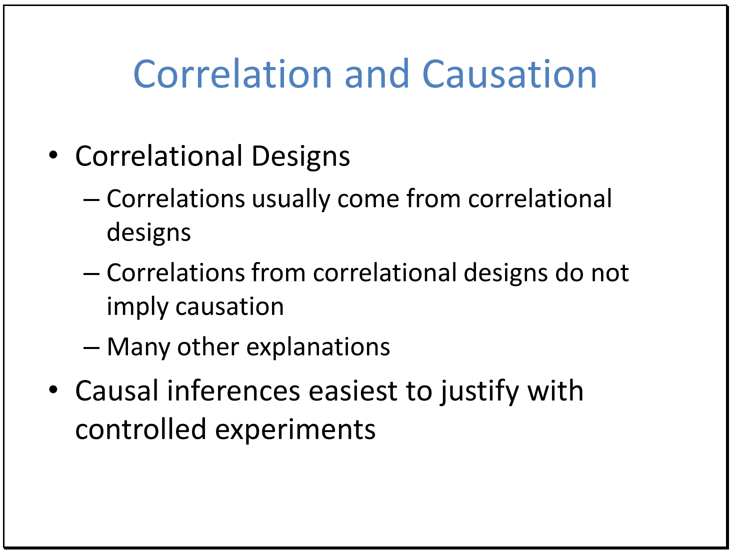
Most constructs measured on individuals in the social sciences are measured with less than perfect reliability. Even some of the best intelligence and personality tests often have reliabilities between .8 and .9. Similarly, most well constructed self-report measures in the social sciences tend to have reliabilities in the .7 to .9 range.

In **classical test theory** we distinguish between true scores and error. Observed scores are a function of true score and error. Our theories are built on trying to understand relationships between true scores, not observed scores. Thus, if we are interested in the relationship between job satisfaction and performance, we may be more interested in the correlation between underlying job satisfaction and performance and not the correlation that we obtain with our less than perfectly reliable measures of job satisfaction and performance.

This is one reason why it is important to use reliable tests.

SEM: This correction formula gives an initial insight into one of the appeals of structural equation modelling (SEM). SEM attempts to model the relationship between latent variables (e.g., True scores). Thus, SEM performs adjustments to relationships between latent variables based on estimates of reliability.

Slide 24



Correlation and Causation

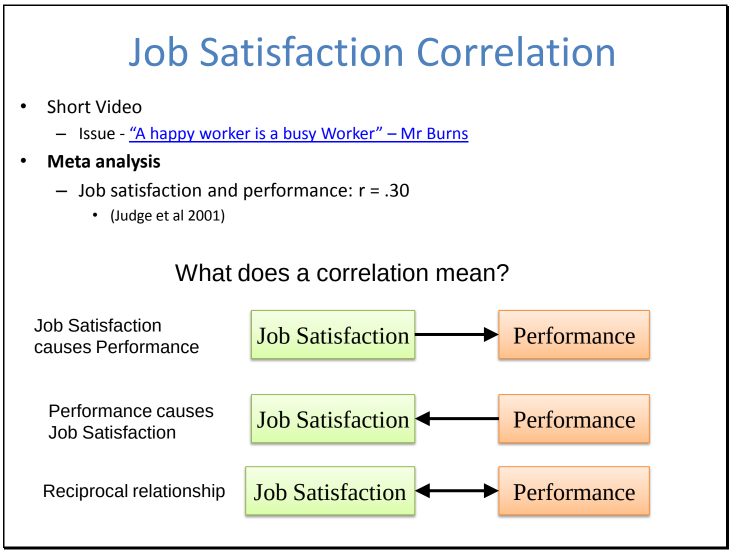
- Correlational Designs
  - Correlations usually come from correlational designs
  - Correlations from correlational designs do not imply causation
  - Many other explanations
- Causal inferences easiest to justify with controlled experiments

Humorous examples of non-causal correlations: Ice cream consumption and drowning; Reduction in pirates and global warming; Stock prices and reduction in skirt length

Note that at a deeper level , causation is best determined by experimental manipulation of an independent variable, random assignment of subjects to groups, and proper experimental control. Thus, technically correlation could imply causation if you measure the correlation of variables where one variable had been experimentally manipulated and participants had been randomly assigned to levels. It is not the statistical technique that determines (i.e., t-test, anova, correlation, regression) whether causation is a valid inference, it is the experimental design. This is an insight that is all too often forgotten.

For a further discussion of this issue and several others, read what is arguably the most important article on recommendations for statistical methods in psychology : There's a web copy here: http://www.loyola.edu/library/ref/articles/Wilkinson.pdf or the reference is: Wilkinson, L., and Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594–604.
The introductory chapter on Structural Equation Modeling in Hair eta al also has a discussion of the issue.

Slide 25



Job Satisfaction Correlation

- Short Video
  - Issue - “A happy worker is a busy Worker” – Mr Burns
- **Meta analysis**
  - Job satisfaction and performance: r = .30
    - (Judge et al 2001)

What does a correlation mean?

http://www.youtube.com/watch?v=7IGJnFKHPyc – to see how mistakes can be made in causal inference

The literature in I/O psychology is filled with correlations. Meta-analysis provides a way of getting a robust estimate of what the average correlation is between two variables.

As part of your revision, you may wish to list all the correlations mentioned in the course. What do these different correlations mean?

The following discussion is presented within the context of the relationship between job satisfaction and performance.

Correlation and causation

You have no doubt been taught that correlation does not equal causation. The classic example of the correlation between ice-cream consumption and drownings highlights the issue that a third variable could be at play such as temperature.

While correlation does not necessarily mean causation, it is worth thinking about what the relationship might be. Causation is one possible explanation for a correlation, but there are many others.

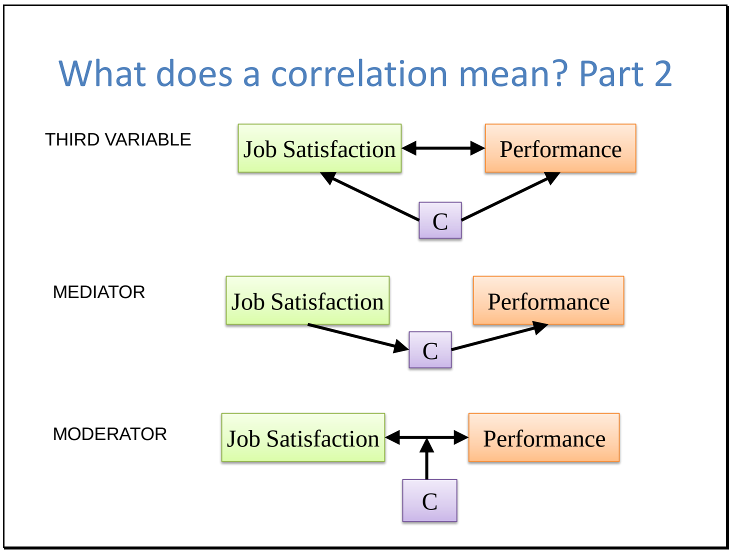
For some purposes correlation is enough when all you are concerned about is prediction. When you want to manipulate the environment, you want to know that you have found the right leverage point. For example, you want to know that an intervention targeted at increasing job satisfaction will improve organisational profit or some other antecedent such as performance, customer satisfaction or turnover. By thinking about the range of possible causal pathways between variables, interventions may be better targeted.

Job Satisfaction causes Performance: This is the classic and simplistic model that suggests that job satisfaction directly influences performance.

Performance causes Job Satisfaction: An alternative model sometimes mentioned is that performance causes job satisfaction. Such a relationship is usually explained in terms of several mediators such as status, pride, financial rewards, promotions, etc.

Reciprocal Relationship: An alternative argument is that both variables influence each other in complex reciprocal ways.

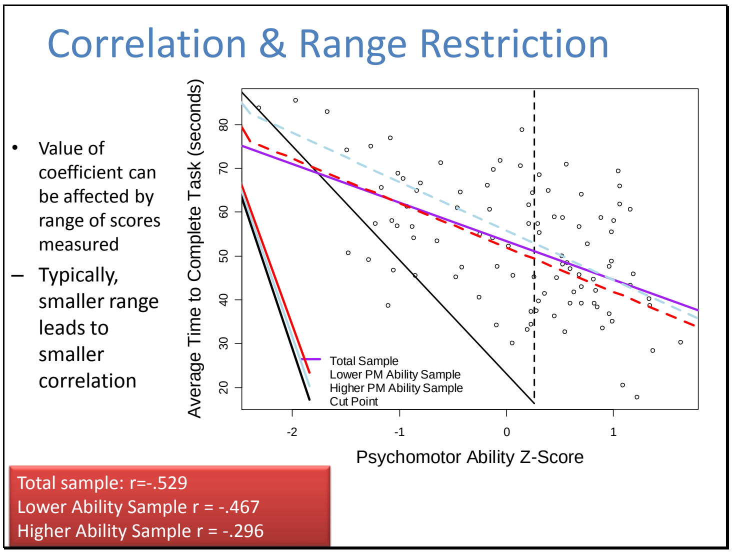
Slide 26



What does a correlation mean? Part 2

THIRD VARIABLE: The correlation between job satisfaction and performance might be explained by a third variable such as self-esteem.
See work by Barron & Kenny for a discussion of the mediator, moderator distinctions: http://davidakenny.net/cm/mediate.htm
MEDIATOR: The correlation between job satisfaction is mediated or explained by a third variable (e.g., possibly self-esteem, personality, intelligence, etc.).
MODERATOR: The relationship between job satisfaction and performance is moderated or changed by a third variable. E.g., the greater control over rewards the more performance would lead to rewards which would lead to job satisfaction.
The key message is: Think about what a correlation means. When you see a correlation think about what kinds of models might be operating. Accept that there are multiple possible explanations. Use theory, common sense, and knowledge of the experimental design to assess the plausibility of different models.

Slide 27



Correlation & Range Restriction

- Value of coefficient can be affected by range of scores measured
  - Typically, smaller range leads to smaller correlation

Total sample: r=-.529
Lower Ability Sample r = -.467
Higher Ability Sample r = -.296

Check out the **simulation** to make the idea of restricted range clearer: http://www.ruf.rice.edu/~lane/stat_sim/restricted_range/index.html

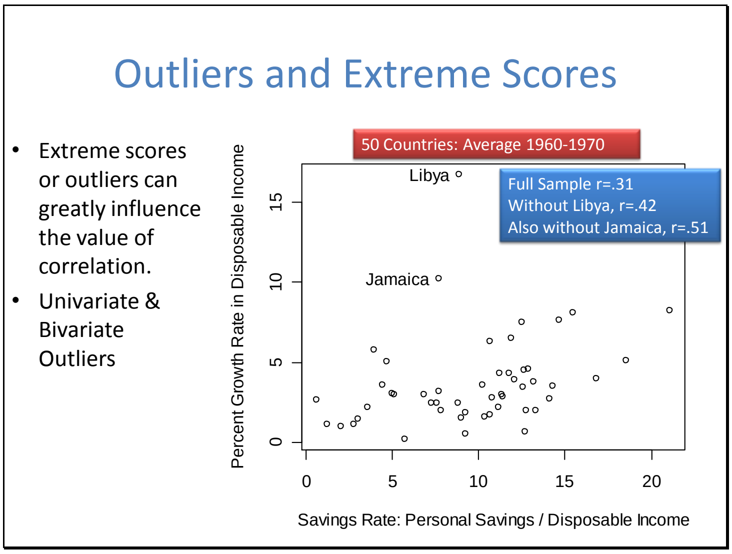
Example: Correlation between intelligence and job performance; Smaller correlation for those who get the job, because those who get the job are more intelligent (i.e., smaller range in intelligence). The example shown above was based on data I collected (Anglim, J., Langan-Fox, J., & Mahdavi, N. (2005). Modeling the Relationship between Strategies, Abilities and Skilled Performance. CogSci 2005, 27th Annual Meeting of the Cognitive Science Society, July 21-23 Stresa, Italy). In the experiment participants completed a series of tasks measuring reaction time which formed a measure, which we called psychomotor ability. Participants then completed a series of trials on a text editing task and the y axis shows average trial completion times. We see that in the overall sample psychomotor ability is correlated -.529 with task completion time (i.e., a strong relationship). However, if we arbitrarily split the sample in half as might happen if I was selecting applicants for a text editing job based on scores on this valid test, we would find that the correlation would be substantially attenuated in the higher ability sample (r=-.296)

Another common example I see is when people study extreme populations. For example, you might have a sample of high performing individuals and you want to know what makes them high performing. You might get a range of measures on these individuals and see how they correlate with their level of performance. But there's a big problem here that is often overlooked in practice. We are now seeing what differentiates high performers from very high performers. We have range restriction in performance which may well attenuate observed correlations. We might have been better to compare high performers with low performers on the measures of interest.

There are many other examples of people studying particular extreme populations and running into these problems: looking at people with psychological disorders or medical ailments, elite athletes, deviant employees, experts, etc.

MAIN POINT: Our sample is representative of a particular hypothetical population, based on where it was drawn from. Generalising the obtained correlation beyond the sampled population is difficult, particularly if the range of values for which the correlation was based on is different to the population that the generalisation is to be made.

Slide 28



Outliers and Extreme Scores

- Extreme scores or outliers can greatly influence the value of correlation.
- Univariate & Bivariate Outliers

50 Countries: Average 1960-1970

Full Sample r=.31
Without Libya, r=.42
Also without Jamaica, r=.51

Check out this simulation to better understand the effect of extreme scores: http://www.uvm.edu/~dhowell/SeeingStatisticsApplets/PointRemove.html

Univariate Outlier: a case that is particular high or low on one variable; rules of thumb suggest z-scores larger than plus or minus 2.5 or 3 or 3.5. In big samples, you are more likely to set the threshold higher.

Bivariate Outlier: A case that is far from the centre of the scatterplot either because it is particularly high one or both variables or because it has unusual combination of scores on the variables (e.g., high ability, but low performance)

Outliers are more influential, and therefore problematic, when sample sizes are small.

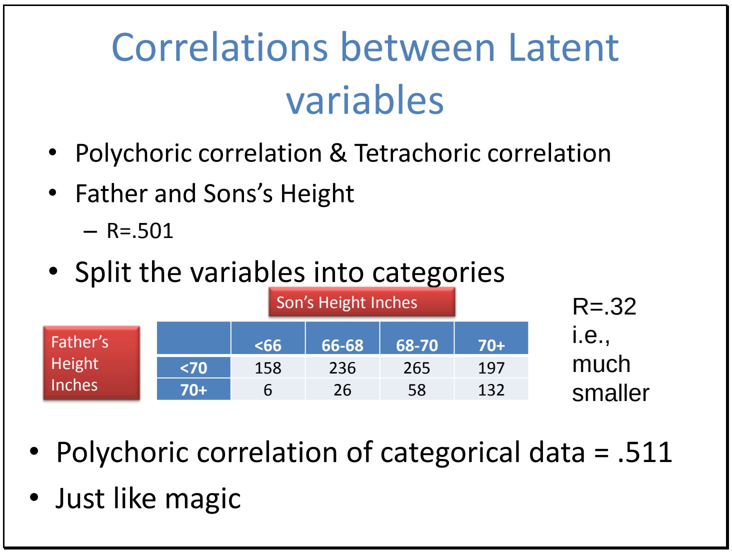
Example

The data represents savings rates and growth rates of disposable income in 50 countries averaged over the years 1960 to 1970.

The dataset was taken from the faraway package in R which in turn sourced the data from Belsley, D., Kuh. E. and Welsch, R. (1980) "Regression Diagnostics" Wiley.

The example highlights the points made above. Libya is a clear outlier having had a dramatic growth rate in disposable income over the period. Thus, if this was our data, what would we do? The first step is to understand the reason for the outliers. Doing a little research, the outlier status of Libya is likely due to the discovery of oil in 1959 in what was otherwise an impoverished country (wikipedia). Once we have understanding of the reasons for the outlier, we can think about whether it is something we want to model. We could say that the typical correlation between the two variables when no unusual economic events occur is about r=.5 and just note that unusual economic events such as the discovery of oil or other natural resources do occur with modest frequency and in such cases the usual relationship between savings rate and growth in disposable income does not hold.

Slide 29



Correlations between Latent variables

- Polychoric correlation & Tetrachoric correlation
- Father and Sons's Height
  - R=.501
- Split the variables into categories

Son's Height Inches

Father's Height Inches

	<66	66-68	68-70	70+
<70	158	236	265	197
70+	6	26	58	132

R=.32 i.e., much smaller

- Polychoric correlation of categorical data = .511
- Just like magic

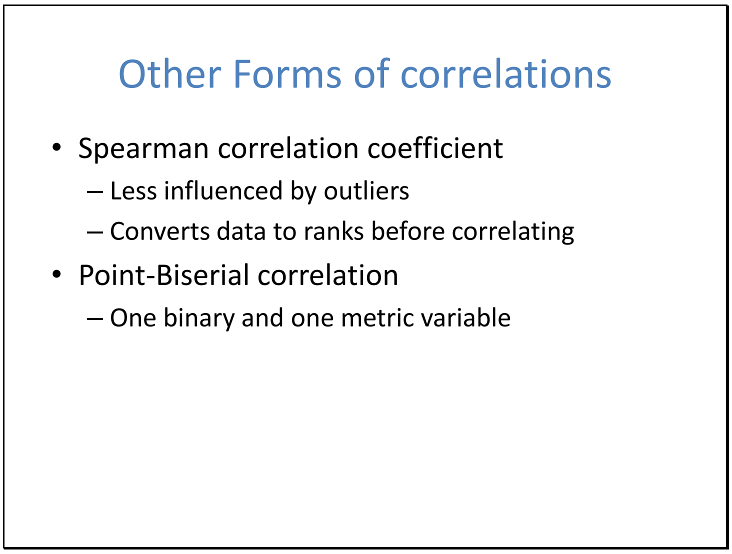
Sometimes we assume that there are latent dimensions underlying categorical response variables.

Polychoric correlation: Estimated Correlation of two continuous latent variables underlying two ordered categorical variables, or one ordered and one binary variable.

Tetrachoric correlation: Estimated Correlation of two continuous latent variables underlying binary variables.

Example: 1078 heights from fathers and sons http://stat-www.berkeley.edu/users/juliab/141C/pearson.dat
The original correlation between father and son height was r=.501. I then split the variable into a bunch of categories for both father's and sons. This new dataset is shown in the table. Now father's height has two points and son's height has four points. We see that if we correlate the variables in this form, the correlation is substantially smaller (r=.32). However, if we run a polychoric correlation on this table, we are able to get a pretty good estimate (r=.511) of the original correlation based on the continuous variables.
TAKE HOME MESSAGE: Often we find ourselves with response scales that do not capture the full spectrum of the continuous variable. This is particularly the case in survey research, market research, and most self-report measures.
SOFTWARE: Unfortunately SPSS does not implement polychoric correlations. For more information and a list of software, see: http://ourworld.compuserve.com/homepages/jsuebersax/tetra.htm . I used John Fox's polycor package in R for the above analyses.

Slide 30



Other Forms of correlations

- Spearman correlation coefficient
  - Less influenced by outliers
  - Converts data to ranks before correlating
- Point-Biserial correlation
  - One binary and one metric variable

Use when two variables are ordinal (ranks)

rs: not unduly affected by outliers.

Also use when relationship between the variables is consistent (ie they covary) but not necessarily linear

In this case, convert data to ranks

Examples:

Position of football teams on the ladder in 2006 with 2007

Liking chocolate (no, maybe, yes) with buying chocolate (never, sometimes, always)

Ranking

Raw times for 100 metres run of 10.4, 9.6, 9.9, and 10.2 seconds, become ranks of 4th, 1st, 2nd, 3rd).

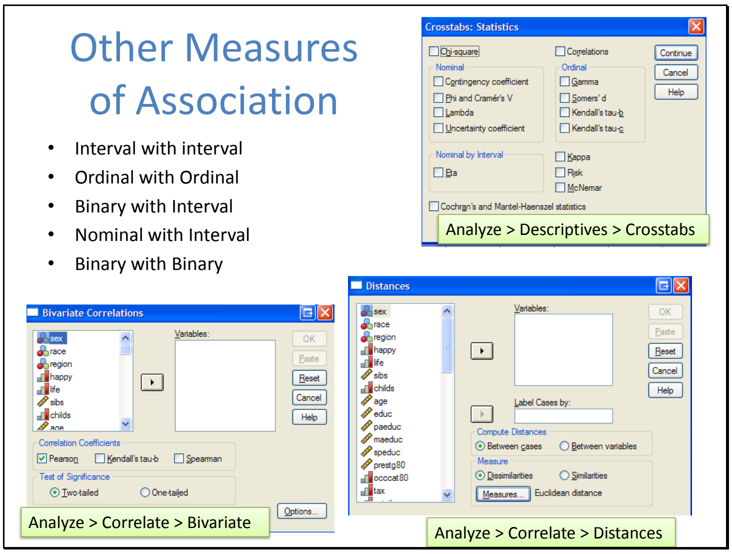
Point-Biserial correlation:

Used to describe the correlation between a binary variable and a metric (e.g., interval or ratio) variable.

Examples: Gender with intelligence; Political orientation (left vs right) with amount donated to charity

Note the similarities with independent groups t-test; applies to the same designs; significance testing gives the same p-value

Slide 31



Other Measures of Association

- Interval with interval
- Ordinal with Ordinal
- Binary with Interval
- Nominal with Interval
- Binary with Binary

Crosstabs: Statistics

□ Chi-square □ Correlations

Nominal: □ Contingency coefficient □ Phi and Cramér's V □ Lambda □ Uncertainty coefficient

Ordinal: □ Gamma □ Somers' d □ Kendall's tau-b □ Kendall's tau-c

Nominal by Interval: □ Eta

□ Kappa □ Risk □ McNemar

□ Cochran's and Mantel-Haenszel statistics

Analyze > Descriptives > Crosstabs

Bivariate Correlations

Correlation Coefficients: ☑ Pearson □ Kendall's tau-b □ Spearman

Test of Significance: ● Two-tailed ○ One-tailed

Analyze > Correlate > Bivariate

Distances

Compute Distances: ● Between cases ○ Between variables

Measure: ● Dissimilarities ○ Similarities

Measures... Euclidean distance

Analyze > Correlate > Distances

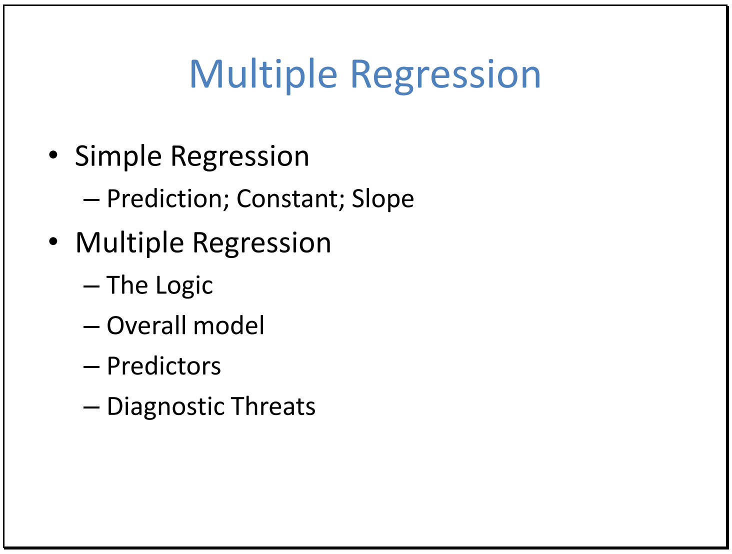
There are many ways to mathematically summarise the strength of the relationship between two variables.

There are different ways of categorising and thinking about these measures.

SPSS has a range of different tools.

When learning a new measure of association it is important to train your intuition in terms of what different values mean. Play around with formula and test it on different pairs of variables.

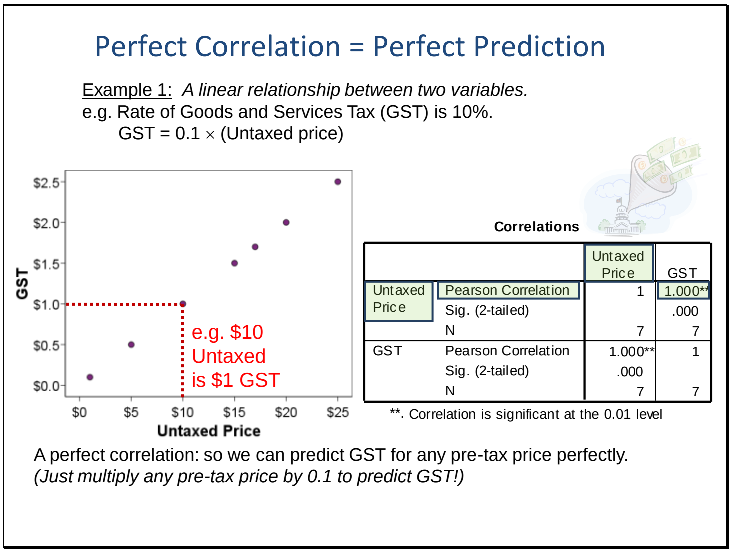
Slide 32



Multiple Regression

- Simple Regression
  - Prediction; Constant; Slope
- Multiple Regression
  - The Logic
  - Overall model
  - Predictors
  - Diagnostic Threats

Slide 33



Perfect Correlation = Perfect Prediction

Example 1: *A linear relationship between two variables.*
e.g. Rate of Goods and Services Tax (GST) is 10%.
GST = 0.1 × (Untaxed price)

Correlations

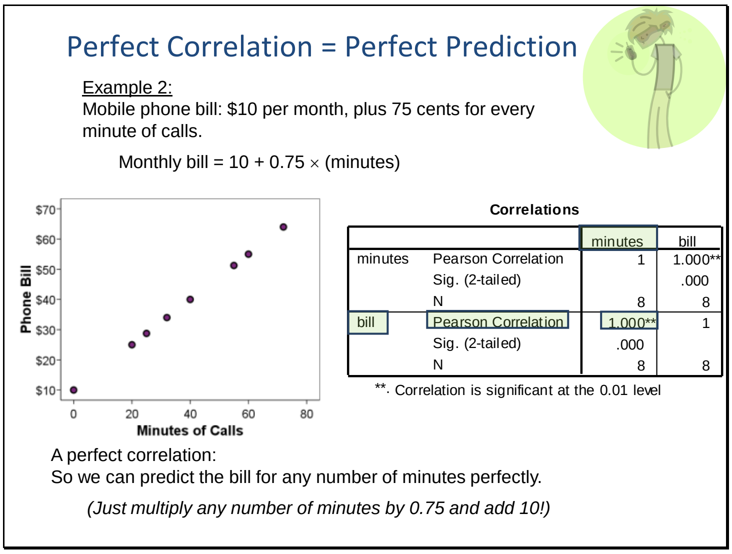
		Untaxed Price	GST
Untaxed Price	Pearson Correlation	1	1.000**
	Sig. (2-tailed)		.000
	N	7	7
GST	Pearson Correlation	1.000**	1
	Sig. (2-tailed)	.000	
	N	7	7

**. Correlation is significant at the 0.01 level

A perfect correlation: so we can predict GST for any pre-tax price perfectly.
(Just multiply any pre-tax price by 0.1 to predict GST!)

Using everyday examples it is clear that we have regression models in our head to describe particular relationships.

Slide 34



Perfect Correlation = Perfect Prediction

Example 2:
Mobile phone bill: $10 per month, plus 75 cents for every minute of calls.

Monthly bill = 10 + 0.75 × (minutes)

Correlations

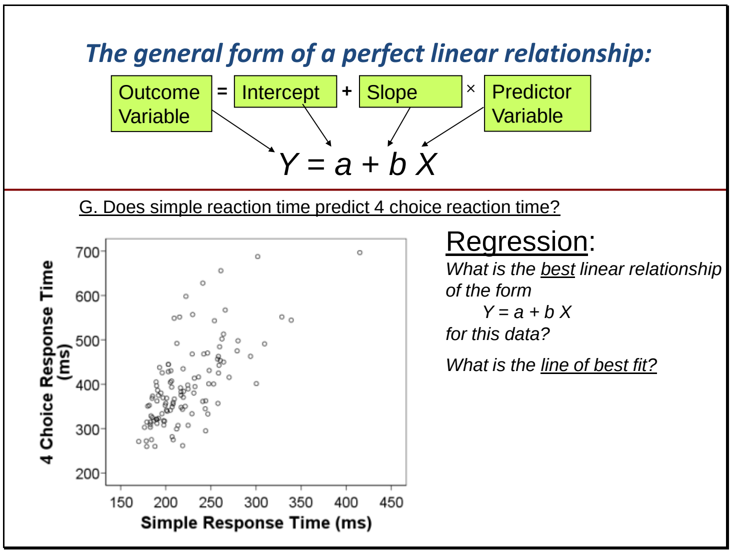
		minutes	bill
minutes	Pearson Correlation	1	1.000**
	Sig. (2-tailed)		.000
	N	8	8
bill	Pearson Correlation	1.000**	1
	Sig. (2-tailed)	.000	
	N	8	8

**. Correlation is significant at the 0.01 level

A perfect correlation:
So we can predict the bill for any number of minutes perfectly.

(Just multiply any number of minutes by 0.75 and add 10!)

Slide 35



The general form of a perfect linear relationship:

Outcome Variable = Intercept + Slope × Predictor Variable

$Y = a + b\,X$

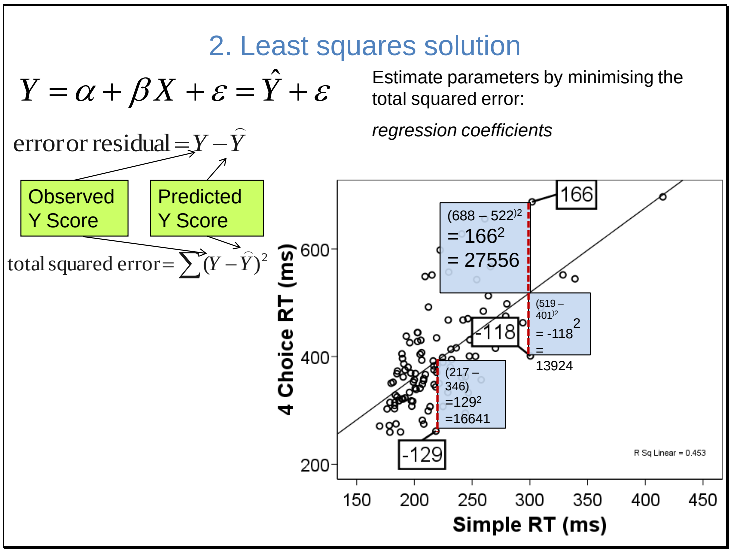
G. Does simple reaction time predict 4 choice reaction time?

Regression:
What is the best linear relationship of the form
Y = a + b X
for this data?

What is the line of best fit?

This simulation is particularly useful in improving your understanding here: Note that many lines will be quite good at minimising error (MSE), but only one will be "best". If you're curious about what MSE is in the simulation: standard error of the estimate = square root of (MSE)
http://www.ruf.rice.edu/~lane/stat_sim/reg_by_eye/index.html

Slide 36



2. Least squares solution

$$Y = \alpha + \beta X + \varepsilon = \hat{Y} + \varepsilon$$

Estimate parameters by minimising the total squared error:

regression coefficients

$$\text{error or residual} = Y - \widehat{Y}$$

Observed Y Score

Predicted Y Score

$$\text{total squared error} = \sum (Y - \widehat{Y})^2$$

ASSORTED NOTES

The regression line represents the predicted 4 Choice RT (Y) for any Simple RT (X)

The graph shows examples of error and squared error for three cases in the dataset

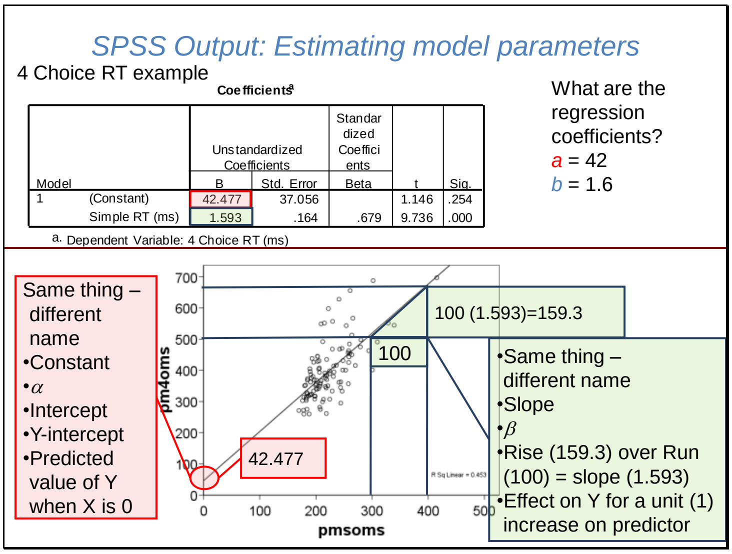
Note that when 4 choice RT (Y) is larger than predicted, error is positive, and when 4 choice RT is smaller than predicted, error is negative

Squared error removes the sign

The graph also attempts to make intuitive why it is called squared error. Squared error is error multiplied by error. Thus, the area taken up by a square of length ‘error’, represents the size of squared error.

It is called “**least squares**” because the line of best fit is such that the sum of **squared** error is “**least**”

Slide 37



SPSS Output: Estimating model parameters

4 Choice RT example

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	42.477	37.056		1.146	.254
	Simple RT (ms)	1.593	.164	.679	9.736	.000

a. Dependent Variable: 4 Choice RT (ms)

What are the regression coefficients?
a = 42
b = 1.6

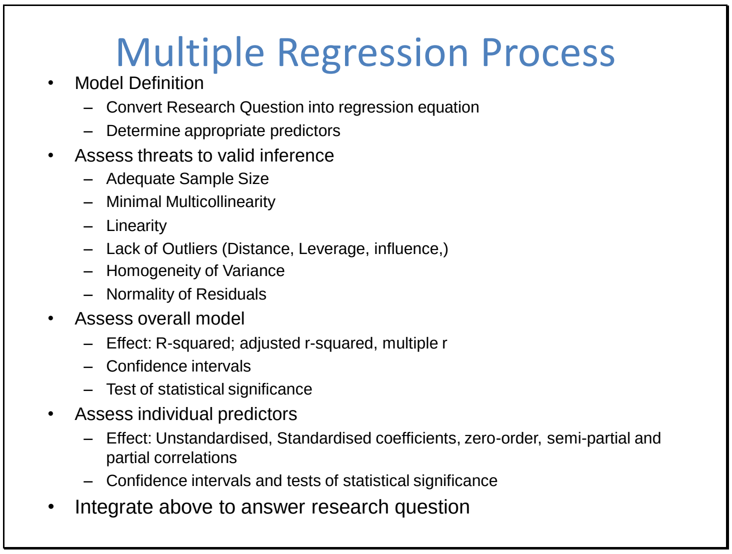
Same thing – different name
- Constant
- α
- Intercept
- Y-intercept
- Predicted value of Y when X is 0

- Same thing – different name
- Slope
- β
- Rise (159.3) over Run (100) = slope (1.593)
- Effect on Y for a unit (1) increase on predictor

Standardised beta will be the same as the correlation between the IV and the DV when it is simple regression.

The regression equation states that for each increase of 1 ms in Simple Reaction Time, the model predicts a 1.6 ms increase in 4 choice Reaction Time. This is consistent with theoretical expectations that it should take longer to press the right button out of four choices than it would to just press one button from one choice (Simple RT). In addition, because both tasks concern speed of responding to visual stimuli, it makes sense that they are fairly strongly related.

Slide 38

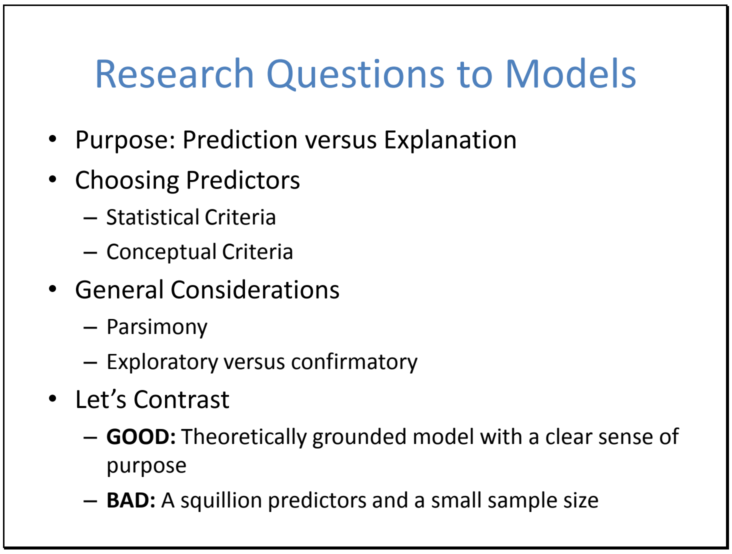


Multiple Regression Process

- Model Definition
  - Convert Research Question into regression equation
  - Determine appropriate predictors
- Assess threats to valid inference
  - Adequate Sample Size
  - Minimal Multicollinearity
  - Linearity
  - Lack of Outliers (Distance, Leverage, influence,)
  - Homogeneity of Variance
  - Normality of Residuals
- Assess overall model
  - Effect: R-squared; adjusted r-squared, multiple r
  - Confidence intervals
  - Test of statistical significance
- Assess individual predictors
  - Effect: Unstandardised, Standardised coefficients, zero-order, semi-partial and partial correlations
  - Confidence intervals and tests of statistical significance
- Integrate above to answer research question

Above sets out a set of steps that are involved in running a multiple regression from formulation of a research question to answering the research question. The above set of steps hopefully provides a useful checklist. However, a proper multiple regression analysis is rarely such a linear process and often involves going backwards and forwards between variable selection and model evaluation. Similarly, assessment threats to valid inferences can lead to decisions which change the model or the data and require the model to be re-run. Even the theoretical question can be redefined by the process of model building.

Slide 39

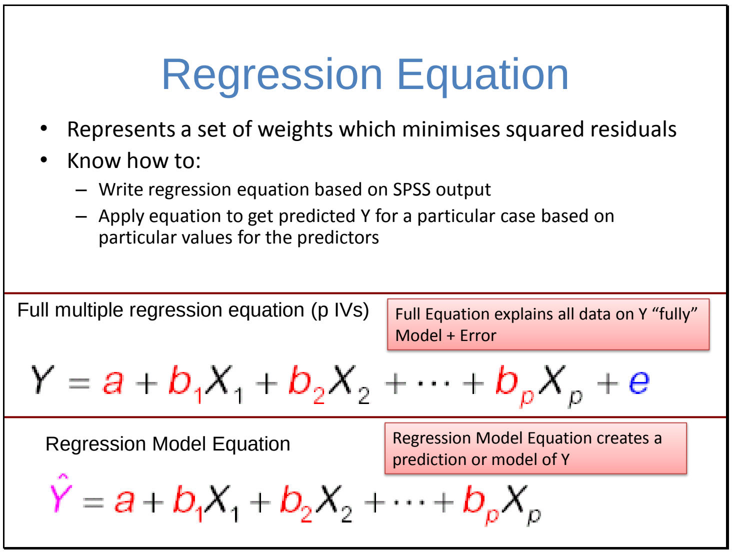


Research Questions to Models

- Purpose: Prediction versus Explanation
- Choosing Predictors
  - Statistical Criteria
  - Conceptual Criteria
- General Considerations
  - Parsimony
  - Exploratory versus confirmatory
- Let’s Contrast
  - **GOOD:** Theoretically grounded model with a clear sense of purpose
  - **BAD:** A squillion predictors and a small sample size

Purpose: It is important to consider the purpose of running a multiple regression. What is the research question? Who is going to use this model and how are they going to use it? Answers to these questions influence what is appropriate in model building and variable selection.
Prediction: Sometimes we are only concerned with accurately predicting some outcome as best as possible. A typical example would be in the context of selection and recruitment where we want to develop a regression equation that best predicts future job performance. While it is interesting and important to think about the reasons why some people perform better than others on the job, for the purpose of maximising financial gains from the recruitment process, recruiters mainly want to know what they need to measure and how these measurements should be combined to make an overall prediction.
Explanation: More commonly in the social sciences, we are trying to understand something about the systems and processes, causal or otherwise, that gave rise to the outcome variable. In this situation we are interested in the relative importance of various predictors. There is often a desire to develop a parsimonious explanation of the phenomena that could assist our understanding of the domain.
Choosing predictors: How do we decide which predictors to include in a model?

Slide 40



Regression Equation

- Represents a set of weights which minimises squared residuals
- Know how to:
  - Write regression equation based on SPSS output
  - Apply equation to get predicted Y for a particular case based on particular values for the predictors

Full multiple regression equation (p IVs)

Full Equation explains all data on Y "fully"
Model + Error

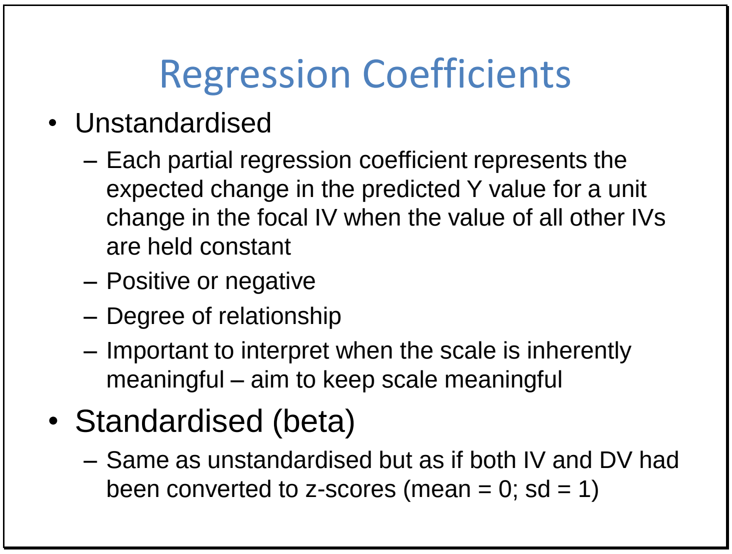
$$Y = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p + e$$

Regression Model Equation

Regression Model Equation creates a prediction or model of Y

$$\hat{Y} = a + b_1X_1 + b_2X_2 + \cdots + b_pX_p$$

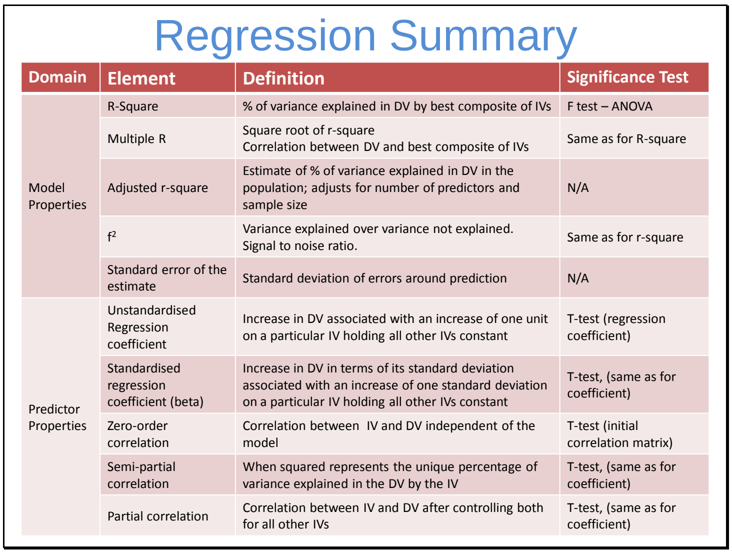
Slide 41



Regression Coefficients

- Unstandardised
  - Each partial regression coefficient represents the expected change in the predicted Y value for a unit change in the focal IV when the value of all other IVs are held constant
  - Positive or negative
  - Degree of relationship
  - Important to interpret when the scale is inherently meaningful – aim to keep scale meaningful
- Standardised (beta)
  - Same as unstandardised but as if both IV and DV had been converted to z-scores (mean = 0; sd = 1)

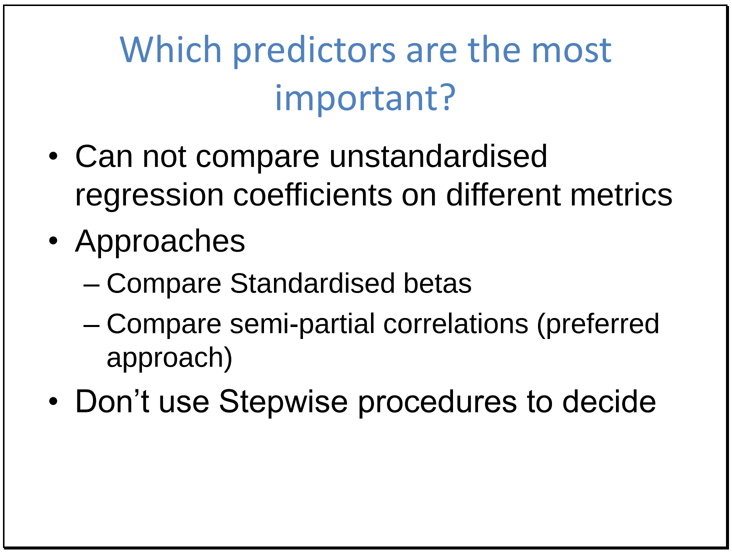
Slide 42



Regression Summary

Domain	Element	Definition	Significance Test
Model Properties	R-Square	% of variance explained in DV by best composite of IVs	F test – ANOVA
	Multiple R	Square root of r-square Correlation between DV and best composite of IVs	Same as for R-square
	Adjusted r-square	Estimate of % of variance explained in DV in the population; adjusts for number of predictors and sample size	N/A
	f^2	Variance explained over variance not explained. Signal to noise ratio.	Same as for r-square
	Standard error of the estimate	Standard deviation of errors around prediction	N/A
Predictor Properties	Unstandardised Regression coefficient	Increase in DV associated with an increase of one unit on a particular IV holding all other IVs constant	T-test (regression coefficient)
	Standardised regression coefficient (beta)	Increase in DV in terms of its standard deviation associated with an increase of one standard deviation on a particular IV holding all other IVs constant	T-test, (same as for coefficient)
	Zero-order correlation	Correlation between IV and DV independent of the model	T-test (initial correlation matrix)
	Semi-partial correlation	When squared represents the unique percentage of variance explained in the DV by the IV	T-test, (same as for coefficient)
	Partial correlation	Correlation between IV and DV after controlling both for all other IVs	T-test, (same as for coefficient)

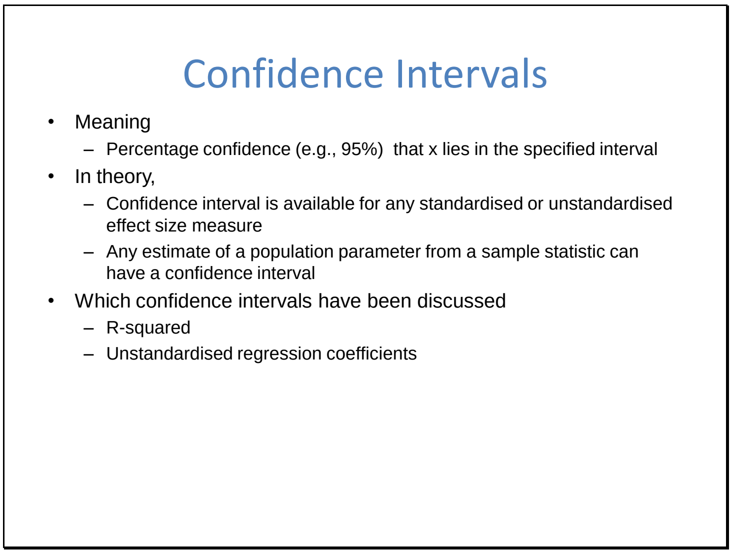
Slide 43



Which predictors are the most important?

- Can not compare unstandardised regression coefficients on different metrics
- Approaches
  - Compare Standardised betas
  - Compare semi-partial correlations (preferred approach)
- Don't use Stepwise procedures to decide

Slide 44

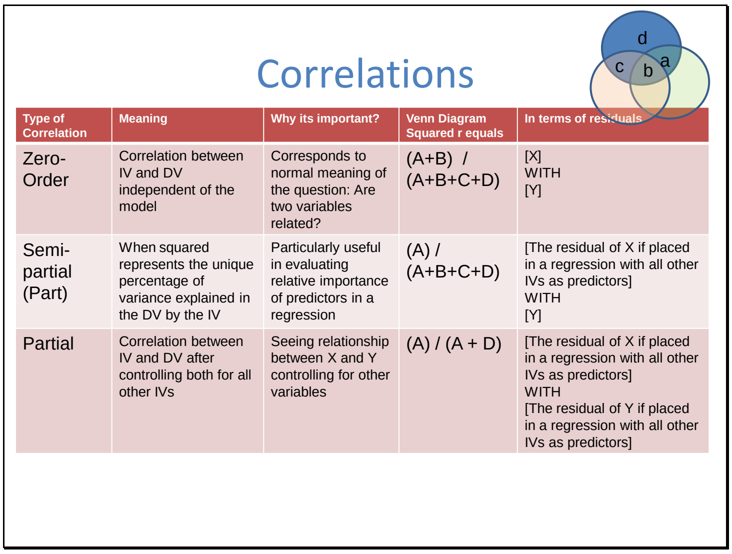


Confidence Intervals

- Meaning
  - Percentage confidence (e.g., 95%) that x lies in the specified interval
- In theory,
  - Confidence interval is available for any standardised or unstandardised effect size measure
  - Any estimate of a population parameter from a sample statistic can have a confidence interval
- Which confidence intervals have been discussed
  - R-squared
  - Unstandardised regression coefficients

Note that obtaining confidence intervals for r-squared can not readily be done in SPSS. There is a program called R2, and there is the MBESS package in R that provides this information.

Slide 45



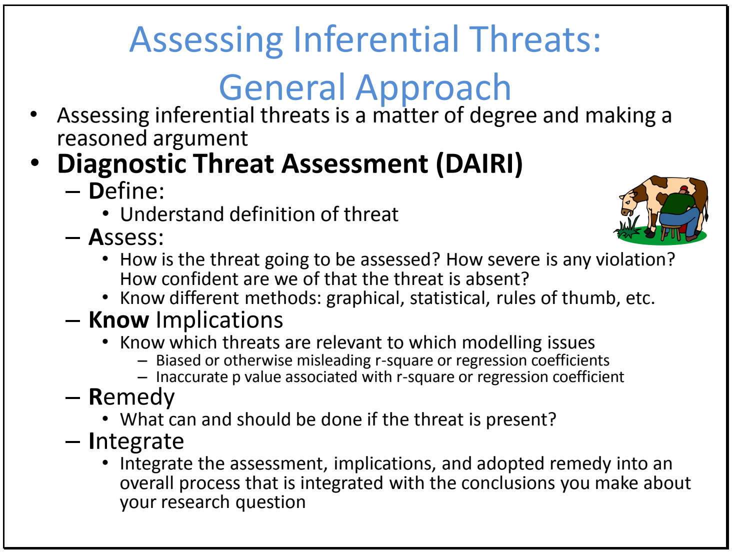
Correlations

Type of Correlation	Meaning	Why its important?	Venn Diagram Squared r equals	In terms of residuals
Zero-Order	Correlation between IV and DV independent of the model	Corresponds to normal meaning of the question: Are two variables related?	(A+B) / (A+B+C+D)	[X] WITH [Y]
Semi-partial (Part)	When squared represents the unique percentage of variance explained in the DV by the IV	Particularly useful in evaluating relative importance of predictors in a regression	(A) / (A+B+C+D)	[The residual of X if placed in a regression with all other IVs as predictors] WITH [Y]
Partial	Correlation between IV and DV after controlling both for all other IVs	Seeing relationship between X and Y controlling for other variables	(A) / (A + D)	[The residual of X if placed in a regression with all other IVs as predictors] WITH [The residual of Y if placed in a regression with all other IVs as predictors]

Memory aide: It’s “semi”-partial (only PART-ly partialling) because it only partials out the effect of the other IVs on X, but not Y

Partial correlation partials out the effect of the other IVs on both X and Y

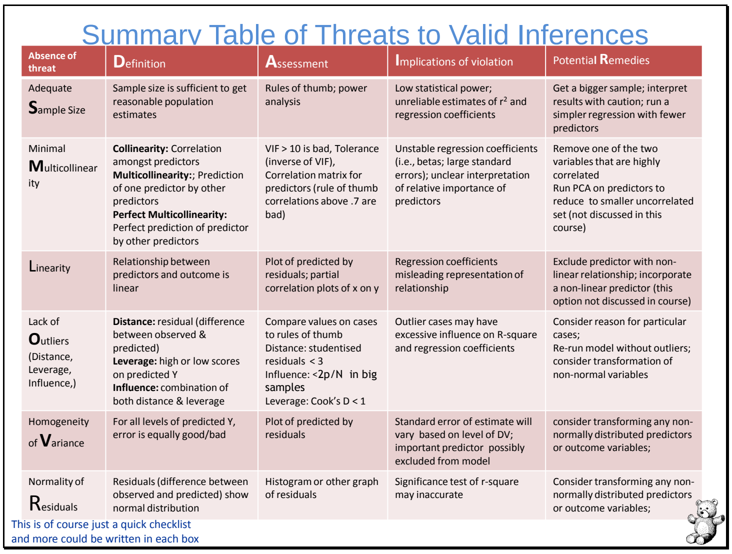
Slide 46



Assessing Inferential Threats: General Approach

- Assessing inferential threats is a matter of degree and making a reasoned argument
- **Diagnostic Threat Assessment (DAIRI)**
  - **D**efine:
    - Understand definition of threat
  - **A**ssess:
    - How is the threat going to be assessed? How severe is any violation? How confident are we of that the threat is absent?
    - Know different methods: graphical, statistical, rules of thumb, etc.
  - **Know** Implications
    - Know which threats are relevant to which modelling issues
      - Biased or otherwise misleading r-square or regression coefficients
      - Inaccurate p value associated with r-square or regression coefficient
  - **R**emedy
    - What can and should be done if the threat is present?
  - **I**ntegrate
    - Integrate the assessment, implications, and adopted remedy into an overall process that is integrated with the conclusions you make about your research question

Slide 47

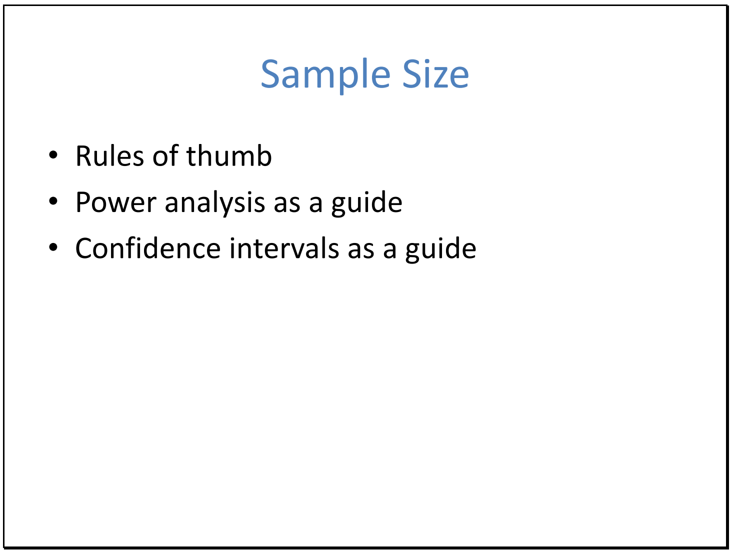


Summary Table of Threats to Valid Inferences

Absence of threat	Definition	Assessment	Implications of violation	Potential Remedies
Adequate Sample Size	Sample size is sufficient to get reasonable population estimates	Rules of thumb; power analysis	Low statistical power; unreliable estimates of r^2 and regression coefficients	Get a bigger sample; interpret results with caution; run a simpler regression with fewer predictors
Minimal Multicollinearity	**Collinearity:** Correlation amongst predictors **Multicollinearity:**; Prediction of one predictor by other predictors **Perfect Multicollinearity:** Perfect prediction of predictor by other predictors	VIF > 10 is bad, Tolerance (inverse of VIF), Correlation matrix for predictors (rule of thumb correlations above .7 are bad)	Unstable regression coefficients (i.e., betas; large standard errors); unclear interpretation of relative importance of predictors	Remove one of the two variables that are highly correlated Run PCA on predictors to reduce to smaller uncorrelated set (not discussed in this course)
Linearity	Relationship between predictors and outcome is linear	Plot of predicted by residuals; partial correlation plots of x on y	Regression coefficients misleading representation of relationship	Exclude predictor with non-linear relationship; incorporate a non-linear predictor (this option not discussed in course)
Lack of Outliers (Distance, Leverage, Influence,)	**Distance:** residual (difference between observed & predicted) **Leverage:** high or low scores on predicted Y **Influence:** combination of both distance & leverage	Compare values on cases to rules of thumb Distance: studentised residuals < 3 Influence: <2p/N in big samples Leverage: Cook's D < 1	Outlier cases may have excessive influence on R-square and regression coefficients	Consider reason for particular cases; Re-run model without outliers; consider transformation of non-normal variables
Homogeneity of Variance	For all levels of predicted Y, error is equally good/bad	Plot of predicted by residuals	Standard error of estimate will vary based on level of DV; important predictor possibly excluded from model	consider transforming any non-normally distributed predictors or outcome variables;
Normality of Residuals	Residuals (difference between observed and predicted) show normal distribution	Histogram or other graph of residuals	Significance test of r-square may inaccurate	Consider transforming any non-normally distributed predictors or outcome variables;

This is of course just a quick checklist and more could be written in each box

Slide 48

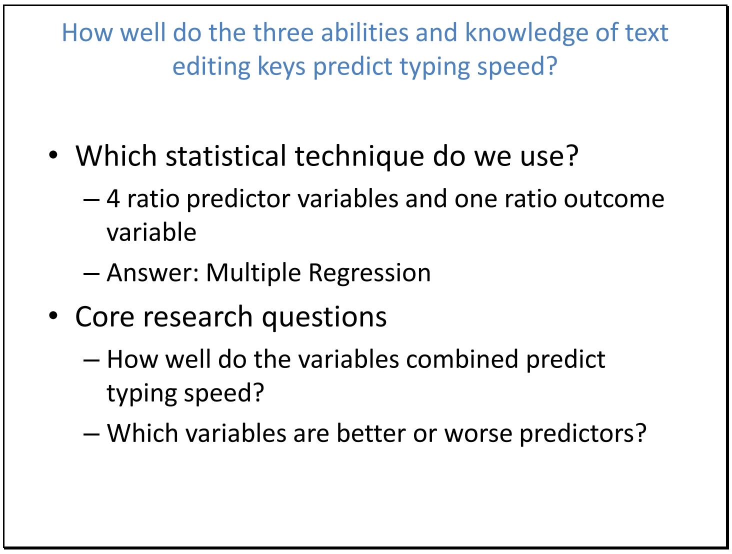


Sample Size

- Rules of thumb
- Power analysis as a guide
- Confidence intervals as a guide

Rules of thumb:
For testing individual regression coefficients:
N > 103 + m, where m = the number of independent variables (Tabachnick & Fiddel)
N > 20m, where m = the number of independent variables
For testing r-square:N >= 50 + 8m
Power Analysis: Power analysis is a more principled approach to determining what sample size is required. Power analysis is a function of the design, alpha, sample size and population effect sizes. Software such as G-power can be used to estimate the probability of obtaining a statistically significant result given
Confidence intervals: see Kelley & Maxwell (2003)

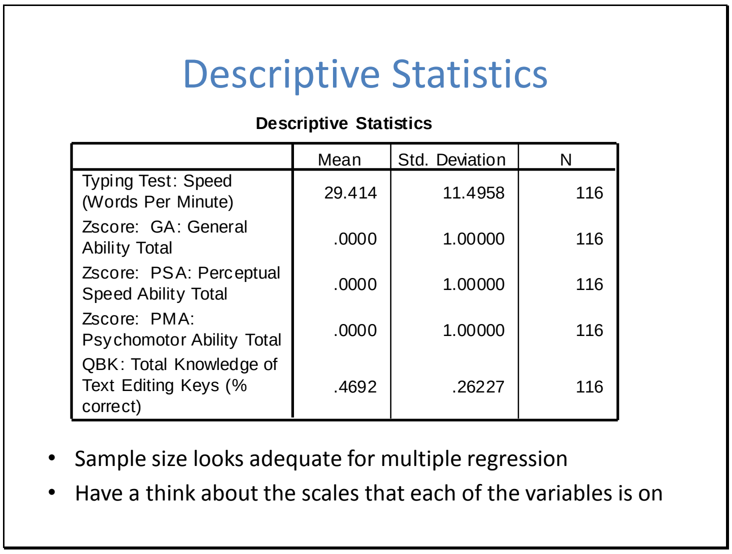
Slide 49



How well do the three abilities and knowledge of text editing keys predict typing speed?

- Which statistical technique do we use?
  - 4 ratio predictor variables and one ratio outcome variable
  - Answer: Multiple Regression
- Core research questions
  - How well do the variables combined predict typing speed?
  - Which variables are better or worse predictors?

Slide 50



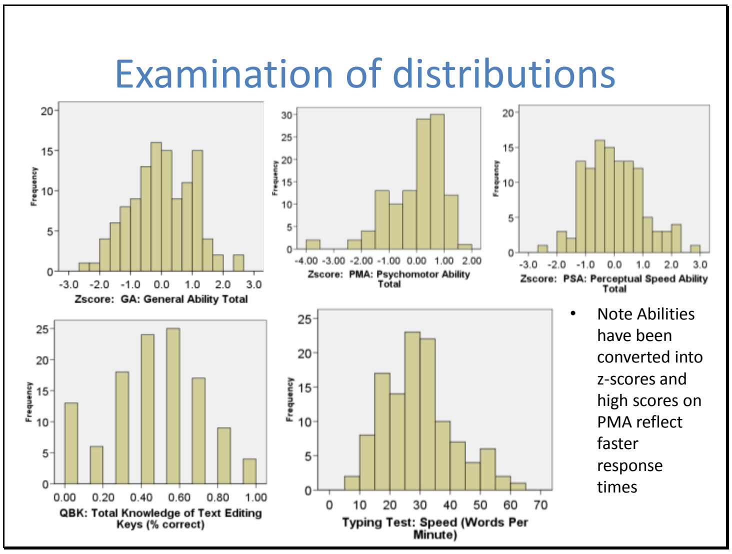
Descriptive Statistics

Descriptive Statistics

	Mean	Std. Deviation	N
Typing Test: Speed (Words Per Minute)	29.414	11.4958	116
Zscore: GA: General Ability Total	.0000	1.00000	116
Zscore: PSA: Perceptual Speed Ability Total	.0000	1.00000	116
Zscore: PMA: Psychomotor Ability Total	.0000	1.00000	116
QBK: Total Knowledge of Text Editing Keys (% correct)	.4692	.26227	116

- Sample size looks adequate for multiple regression
- Have a think about the scales that each of the variables is on

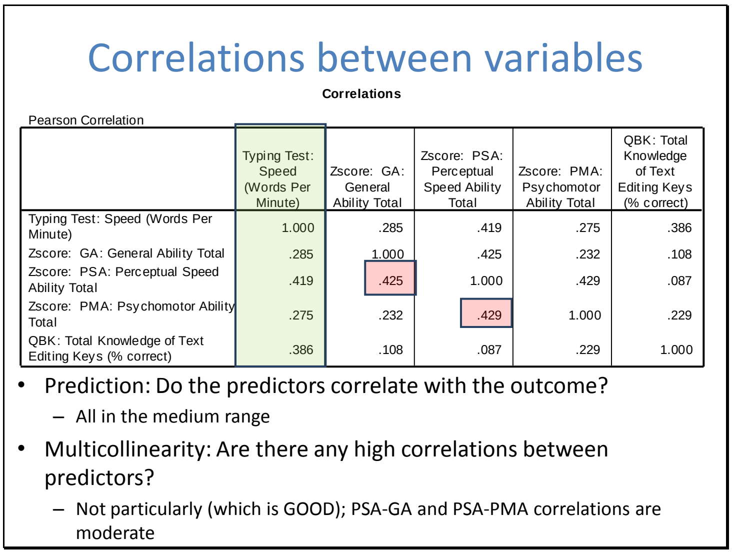
Slide 51



Examination of distributions

- Note Abilities have been converted into z-scores and high scores on PMA reflect faster response times

Slide 52



Correlations between variables

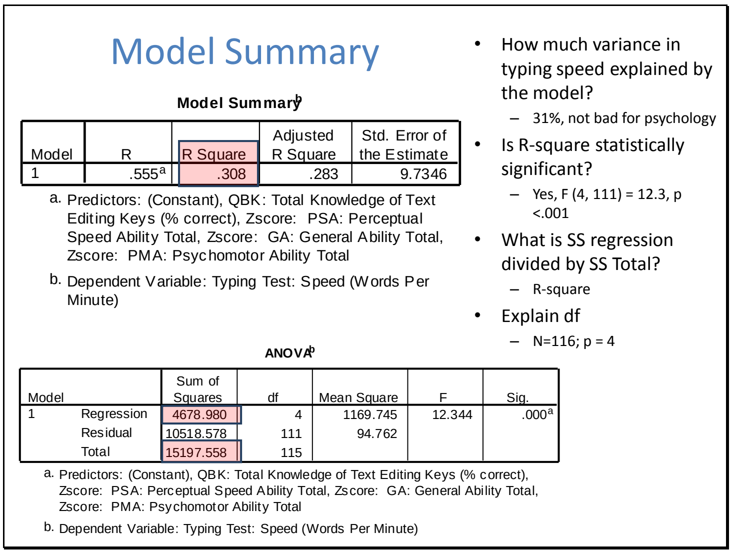
Correlations

Pearson Correlation

	Typing Test: Speed (Words Per Minute)	Zscore: GA: General Ability Total	Zscore: PSA: Perceptual Speed Ability Total	Zscore: PMA: Psychomotor Ability Total	QBK: Total Knowledge of Text Editing Keys (% correct)
Typing Test: Speed (Words Per Minute)	1.000	.285	.419	.275	.386
Zscore: GA: General Ability Total	.285	1.000	.425	.232	.108
Zscore: PSA: Perceptual Speed Ability Total	.419	.425	1.000	.429	.087
Zscore: PMA: Psychomotor Ability Total	.275	.232	.429	1.000	.229
QBK: Total Knowledge of Text Editing Keys (% correct)	.386	.108	.087	.229	1.000

- Prediction: Do the predictors correlate with the outcome?
  - All in the medium range
- Multicollinearity: Are there any high correlations between predictors?
  - Not particularly (which is GOOD); PSA-GA and PSA-PMA correlations are moderate

Slide 53



Model Summary

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.555[a]	.308	.283	9.7346

a. Predictors: (Constant), QBK: Total Knowledge of Text Editing Keys (% correct), Zscore: PSA: Perceptual Speed Ability Total, Zscore: GA: General Ability Total, Zscore: PMA: Psychomotor Ability Total

b. Dependent Variable: Typing Test: Speed (Words Per Minute)

- How much variance in typing speed explained by the model?
  - 31%, not bad for psychology
- Is R-square statistically significant?
  - Yes, F (4, 111) = 12.3, p <.001
- What is SS regression divided by SS Total?
  - R-square
- Explain df
  - N=116; p = 4

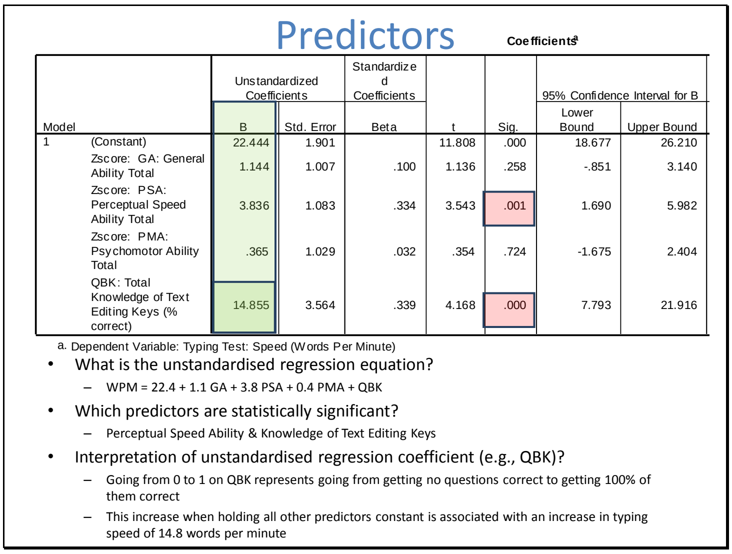
ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	4678.980	4	1169.745	12.344	.000[a]
	Residual	10518.578	111	94.762		
	Total	15197.558	115			

a. Predictors: (Constant), QBK: Total Knowledge of Text Editing Keys (% correct), Zscore: PSA: Perceptual Speed Ability Total, Zscore: GA: General Ability Total, Zscore: PMA: Psychomotor Ability Total

b. Dependent Variable: Typing Test: Speed (Words Per Minute)

Slide 54



Predictors

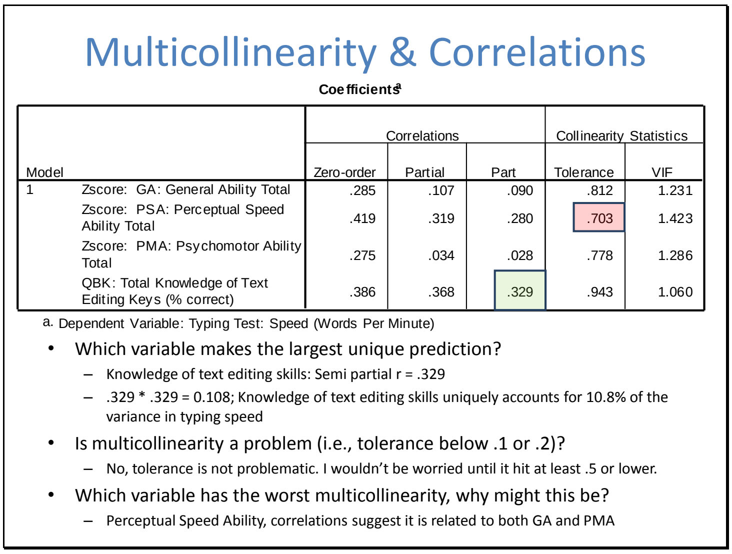
Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	22.444	1.901		11.808	.000	18.677	26.210
	Zscore: GA: General Ability Total	1.144	1.007	.100	1.136	.258	-.851	3.140
	Zscore: PSA: Perceptual Speed Ability Total	3.836	1.083	.334	3.543	.001	1.690	5.982
	Zscore: PMA: Psychomotor Ability Total	.365	1.029	.032	.354	.724	-1.675	2.404
	QBK: Total Knowledge of Text Editing Keys (% correct)	14.855	3.564	.339	4.168	.000	7.793	21.916

a. Dependent Variable: Typing Test: Speed (Words Per Minute)

- What is the unstandardised regression equation?
  - WPM = 22.4 + 1.1 GA + 3.8 PSA + 0.4 PMA + QBK
- Which predictors are statistically significant?
  - Perceptual Speed Ability & Knowledge of Text Editing Keys
- Interpretation of unstandardised regression coefficient (e.g., QBK)?
  - Going from 0 to 1 on QBK represents going from getting no questions correct to getting 100% of them correct
  - This increase when holding all other predictors constant is associated with an increase in typing speed of 14.8 words per minute

Slide 55



Multicollinearity & Correlations

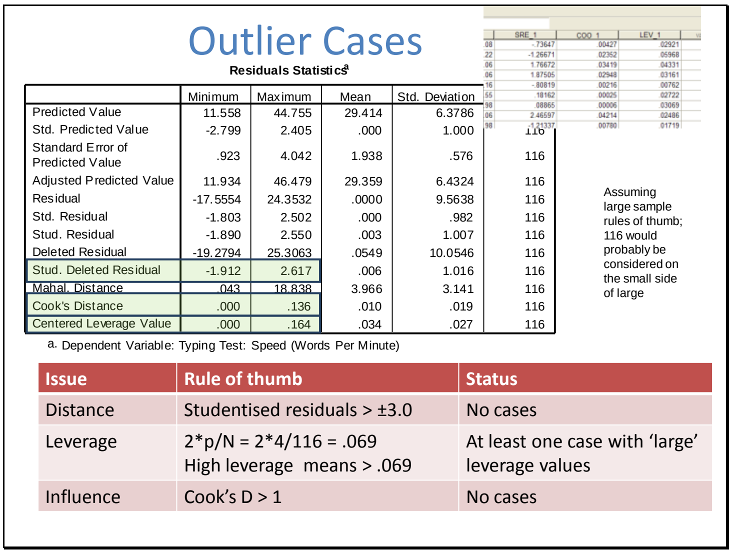
Coefficients[a]

Model		Correlations			Collinearity Statistics	
		Zero-order	Partial	Part	Tolerance	VIF
1	Zscore: GA: General Ability Total	.285	.107	.090	.812	1.231
	Zscore: PSA: Perceptual Speed Ability Total	.419	.319	.280	.703	1.423
	Zscore: PMA: Psychomotor Ability Total	.275	.034	.028	.778	1.286
	QBK: Total Knowledge of Text Editing Keys (% correct)	.386	.368	.329	.943	1.060

a. Dependent Variable: Typing Test: Speed (Words Per Minute)

- Which variable makes the largest unique prediction?
  - Knowledge of text editing skills: Semi partial r = .329
  - .329 * .329 = 0.108; Knowledge of text editing skills uniquely accounts for 10.8% of the variance in typing speed
- Is multicollinearity a problem (i.e., tolerance below .1 or .2)?
  - No, tolerance is not problematic. I wouldn't be worried until it hit at least .5 or lower.
- Which variable has the worst multicollinearity, why might this be?
  - Perceptual Speed Ability, correlations suggest it is related to both GA and PMA

Slide 56



Outlier Cases

Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	
Predicted Value	11.558	44.755	29.414	6.3786	
Std. Predicted Value	-2.799	2.405	.000	1.000	116
Standard Error of Predicted Value	.923	4.042	1.938	.576	116
Adjusted Predicted Value	11.934	46.479	29.359	6.4324	116
Residual	-17.5554	24.3532	.0000	9.5638	116
Std. Residual	-1.803	2.502	.000	.982	116
Stud. Residual	-1.890	2.550	.003	1.007	116
Deleted Residual	-19.2794	25.3063	.0549	10.0546	116
Stud. Deleted Residual	-1.912	2.617	.006	1.016	116
Mahal. Distance	.043	18.838	3.966	3.141	116
Cook's Distance	.000	.136	.010	.019	116
Centered Leverage Value	.000	.164	.034	.027	116

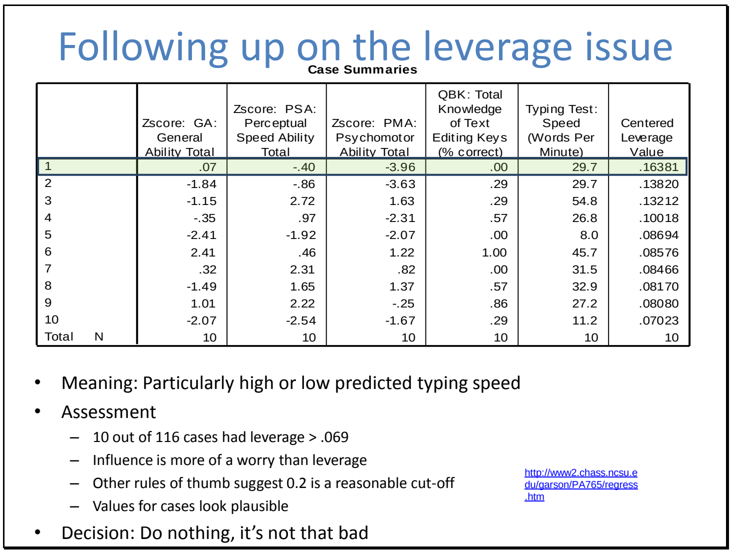
a. Dependent Variable: Typing Test: Speed (Words Per Minute)

	SRE_1	COO_1	LEV_1
.08	-.73647	.00427	.02921
.22	-1.26671	.02352	.05968
.06	1.76672	.03419	.04331
.06	1.87505	.02948	.03161
.16	-.80819	.00216	.00762
.55	.18162	.00025	.02722
.98	.08865	.00006	.03069
.06	2.46597	.04214	.02486
.98	-1.21337	.00780	.01719

Assuming large sample rules of thumb; 116 would probably be considered on the small side of large

Issue	Rule of thumb	Status
Distance	Studentised residuals > ±3.0	No cases
Leverage	2*p/N = 2*4/116 = .069 High leverage means > .069	At least one case with 'large' leverage values
Influence	Cook's D > 1	No cases

Slide 57



Following up on the leverage issue

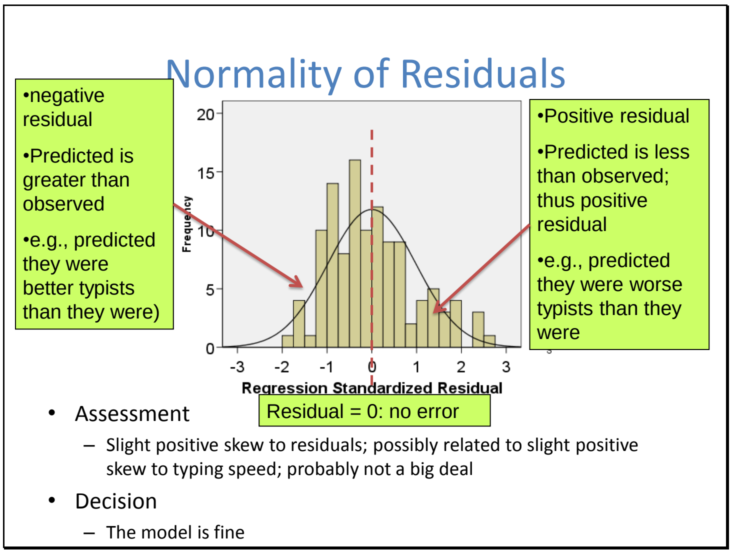
Case Summaries

		Zscore: GA: General Ability Total	Zscore: PSA: Perceptual Speed Ability Total	Zscore: PMA: Psychomotor Ability Total	QBK: Total Knowledge of Text Editing Keys (% correct)	Typing Test: Speed (Words Per Minute)	Centered Leverage Value
1		.07	-.40	-3.96	.00	29.7	.16381
2		-1.84	-.86	-3.63	.29	29.7	.13820
3		-1.15	2.72	1.63	.29	54.8	.13212
4		-.35	.97	-2.31	.57	26.8	.10018
5		-2.41	-1.92	-2.07	.00	8.0	.08694
6		2.41	.46	1.22	1.00	45.7	.08576
7		.32	2.31	.82	.00	31.5	.08466
8		-1.49	1.65	1.37	.57	32.9	.08170
9		1.01	2.22	-.25	.86	27.2	.08080
10		-2.07	-2.54	-1.67	.29	11.2	.07023
Total	N	10	10	10	10	10	10

- Meaning: Particularly high or low predicted typing speed
- Assessment
  - 10 out of 116 cases had leverage > .069
  - Influence is more of a worry than leverage
  - Other rules of thumb suggest 0.2 is a reasonable cut-off
  - Values for cases look plausible
- Decision: Do nothing, it’s not that bad

http://www2.chass.ncsu.edu/garson/PA765/regress.htm

Slide 58



Normality of Residuals

•negative residual

•Predicted is greater than observed

•e.g., predicted they were better typists than they were)

•Positive residual

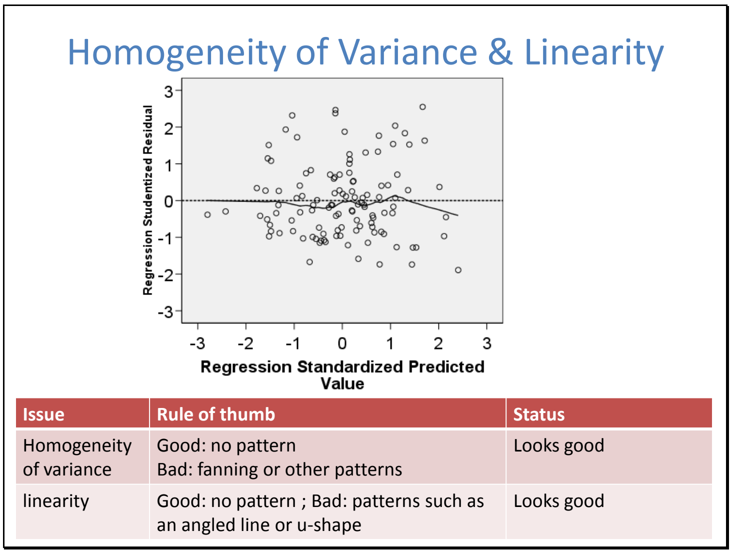
•Predicted is less than observed; thus positive residual

•e.g., predicted they were worse typists than they were

Residual = 0: no error

- Assessment
  - Slight positive skew to residuals; possibly related to slight positive skew to typing speed; probably not a big deal
- Decision
  - The model is fine

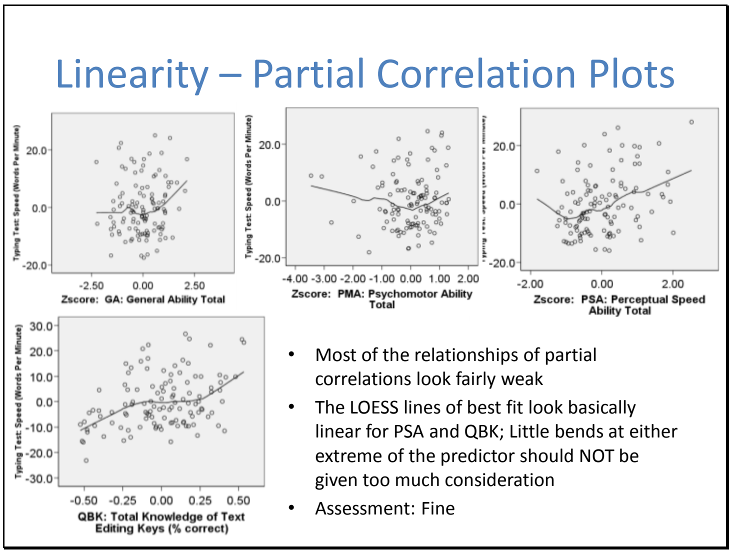
Slide 59



Homogeneity of Variance & Linearity

Issue	Rule of thumb	Status
Homogeneity of variance	Good: no pattern Bad: fanning or other patterns	Looks good
linearity	Good: no pattern ; Bad: patterns such as an angled line or u-shape	Looks good

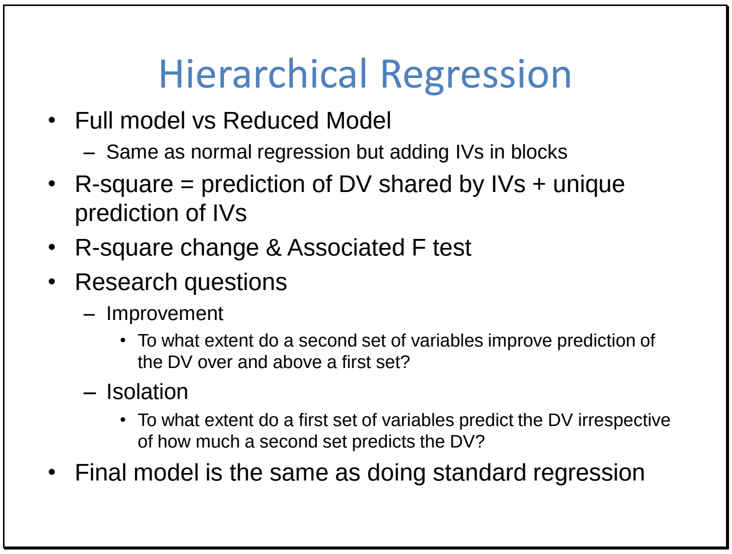
Slide 60



Linearity – Partial Correlation Plots

- Most of the relationships of partial correlations look fairly weak
- The LOESS lines of best fit look basically linear for PSA and QBK; Little bends at either extreme of the predictor should NOT be given too much consideration
- Assessment: Fine

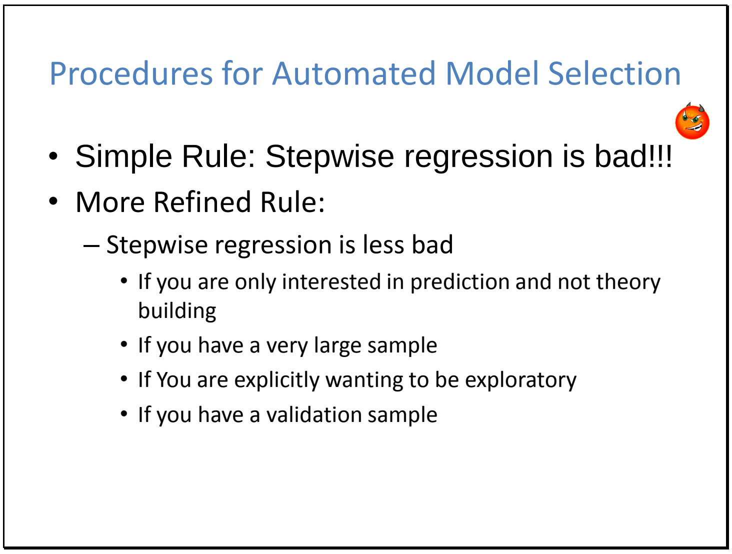
Slide 61



Hierarchical Regression

- Full model vs Reduced Model
  - Same as normal regression but adding IVs in blocks
- R-square = prediction of DV shared by IVs + unique prediction of IVs
- R-square change & Associated F test
- Research questions
  - Improvement
    - To what extent do a second set of variables improve prediction of the DV over and above a first set?
  - Isolation
    - To what extent do a first set of variables predict the DV irrespective of how much a second set predicts the DV?
- Final model is the same as doing standard regression

Slide 62

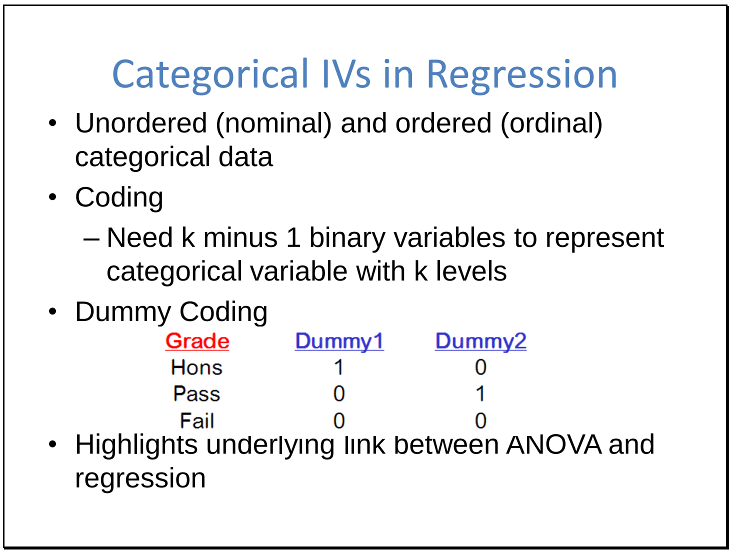


Procedures for Automated Model Selection

- Simple Rule: Stepwise regression is bad!!!
- More Refined Rule:
  - Stepwise regression is less bad
    - If you are only interested in prediction and not theory building
    - If you have a very large sample
    - If You are explicitly wanting to be exploratory
    - If you have a validation sample

There are many ways of automating the process of variable selection. Stepwise regression progressively adds predictors one at a time. At each step it adds the variable that increases r-square the most. It stops adding variables when none of the variables adds a statistically significant amount of variance, using the criteria of statistical significance set by the user.

Slide 63



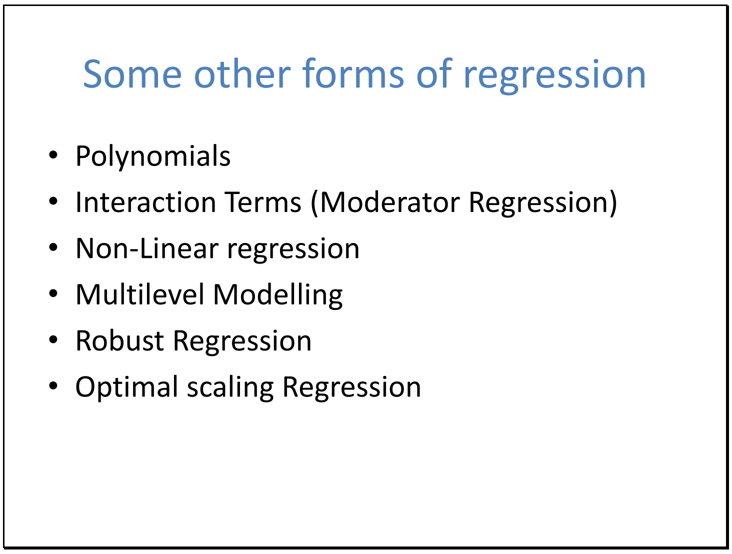
Categorical IVs in Regression

- Unordered (nominal) and ordered (ordinal) categorical data
- Coding
  - – Need k minus 1 binary variables to represent categorical variable with k levels
- Dummy Coding

Grade	Dummy1	Dummy2
Hons	1	0
Pass	0	1
Fail	0	0

- Highlights underlying link between ANOVA and regression

Slide 64



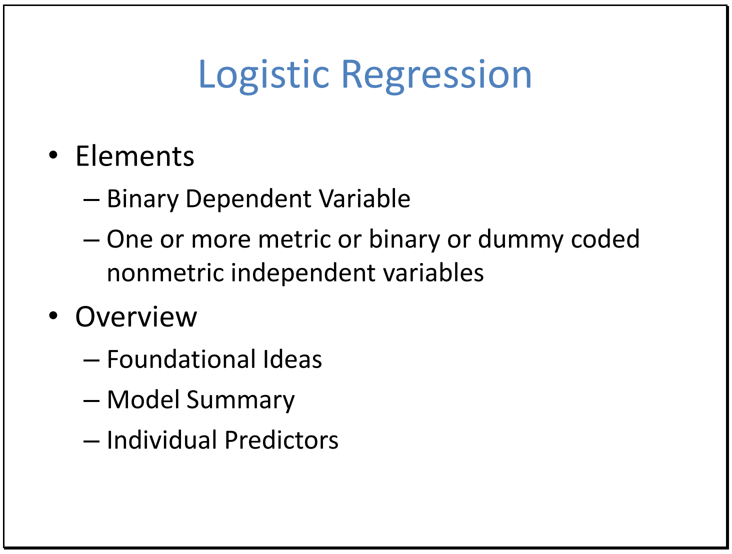
Some other forms of regression

- Polynomials
- Interaction Terms (Moderator Regression)
- Non-Linear regression
- Multilevel Modelling
- Robust Regression
- Optimal scaling Regression

Polynomials: Sometimes we want to see whether there is a quadratic, cubic or higher order trend in the relationship between two variables. In particular, a quadratic relationship is not uncommon in the social sciences. To incorporate polynomial predictors we simply have to raise the predictor to the power of the polynomial and incorporate this new variable into the model. For example, if we were predicting job performance from arousal and we thought there were both linear and quadratic effects (order two polynomial). We would include arousal as a predictor as well as a new variable which would be the square of arousal. We can optionally centre arousal by subtracting the mean from the variable before squaring. This reduces the issue of multicollinearity between linear and quadratic effects, which can aid interpretation. Usually the analyses are presented in the form of a hierarchical regression whereby the first step includes the linear term and then in a second step, the quadratic term is included. If the addition of the quadratic term leads to a significant r-square change, the quadratic term is retained.

Interaction Terms: Interaction terms test hypotheses about moderators. To include an interaction term we multiply the two variables together, optionally centring each variable (i.e., subtracting the mean) prior to multiplying. This variable is then included in the regression model in addition to the main effects.
Multilevel Modelling: This is a specialised form of regression often used to look at longitudinal data and data involving the nesting of participants in groups such as departments or organisations or teams. Popular software tools include HLM and MLWin. SPSS and SAS also have tools for dealing with these models.
Optimal Scaling Regression: This is the same as normal multiple regression except that the procedure is allowed to change the scaling of the predictor and outcome variables in ways consistent with the type of variables they are defined as. Variables are rescaled to maximise R-squared. This is implemented in SPSS using the Categories add-on module. This can yield interesting results, but there are many potential pitfalls, if you do not know what the procedure is doing.

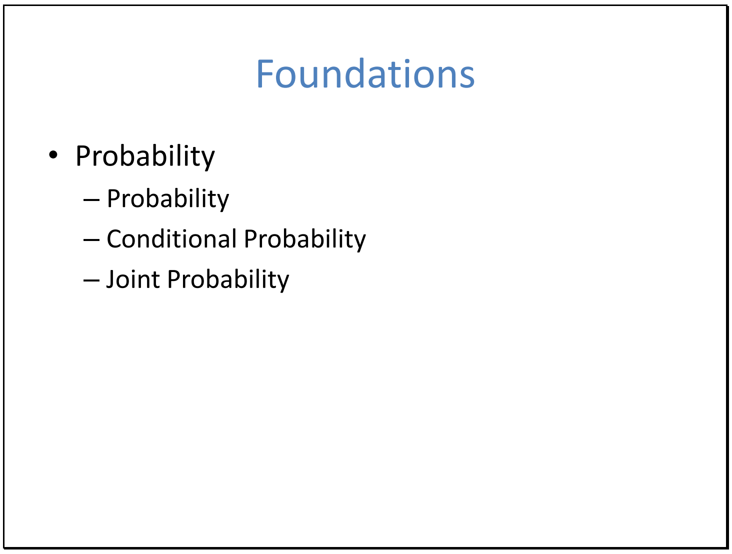
Slide 65



Logistic Regression

- Elements
  - Binary Dependent Variable
  - One or more metric or binary or dummy coded nonmetric independent variables
- Overview
  - Foundational Ideas
  - Model Summary
  - Individual Predictors

Slide 66

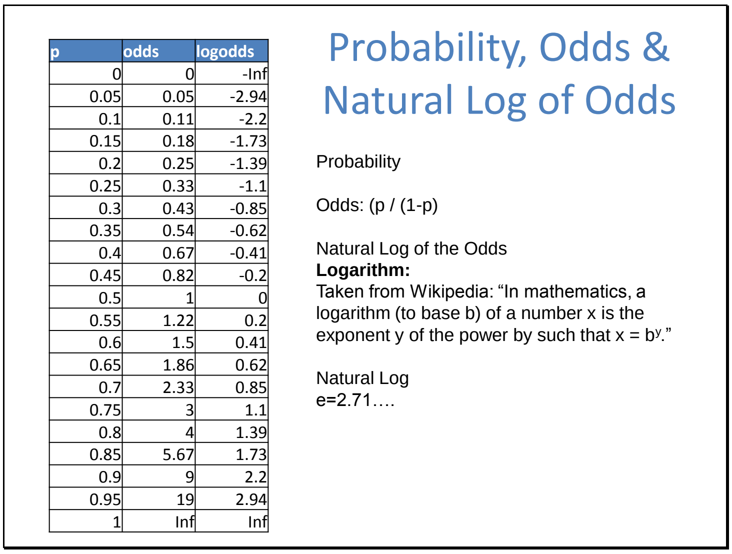


Foundations

- Probability
  - Probability
  - Conditional Probability
  - Joint Probability

Event: The general term “event” can be used to describe the occurrence of something. In the context of logistic regression this is often things like death, getting a disease, or answering a question correctly.
Probability: Probability represents the percentage chance that an event will occur. Probability ranges from 0 to 1. A value of 0 represents that there is no chance of the event occurring. A value of 1 represents that the event will definitely occur.
Conditional Probability: If something has already happened at time A, what is the probability that the event will occur at time B?
Joint Probability: chance of two events both occuring

Slide 67



p	odds	logodds
0	0	-Inf
0.05	0.05	-2.94
0.1	0.11	-2.2
0.15	0.18	-1.73
0.2	0.25	-1.39
0.25	0.33	-1.1
0.3	0.43	-0.85
0.35	0.54	-0.62
0.4	0.67	-0.41
0.45	0.82	-0.2
0.5	1	0
0.55	1.22	0.2
0.6	1.5	0.41
0.65	1.86	0.62
0.7	2.33	0.85
0.75	3	1.1
0.8	4	1.39
0.85	5.67	1.73
0.9	9	2.2
0.95	19	2.94
1	Inf	Inf

Probability, Odds & Natural Log of Odds

Probability

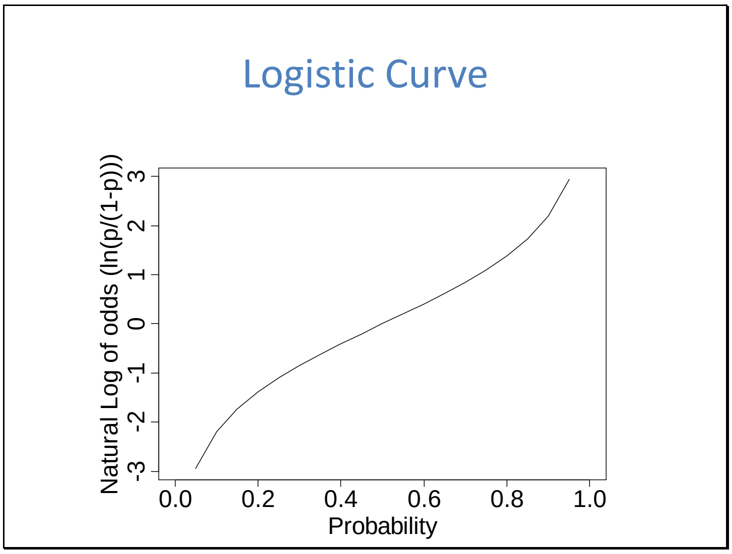
Odds: (p / (1-p)

Natural Log of the Odds
Logarithm:
Taken from Wikipedia: “In mathematics, a logarithm (to base b) of a number x is the exponent y of the power by such that x = b^y.”

Natural Log
e=2.71….

Probability: The likelihood of an event occurring in percentage terms.
Odds: The likelihood of an event occurring (p) divided by the chance of it not occurring (1-p). This is typically used in betting language, when they say the odds of a horse winning is 4 to 1. this means that probability is that for every one time it wins there are predicted to be 4 times it loses.

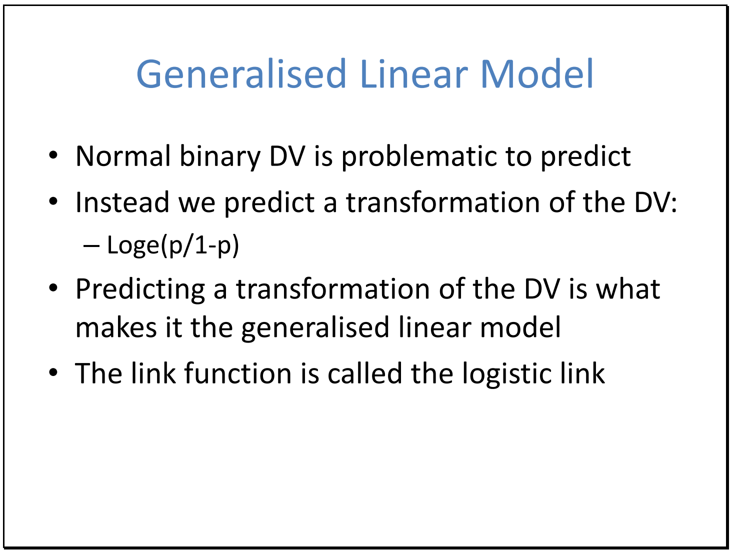
Slide 68



Logistic Curve

The logistic curve shows the relationship between probability and the logit.

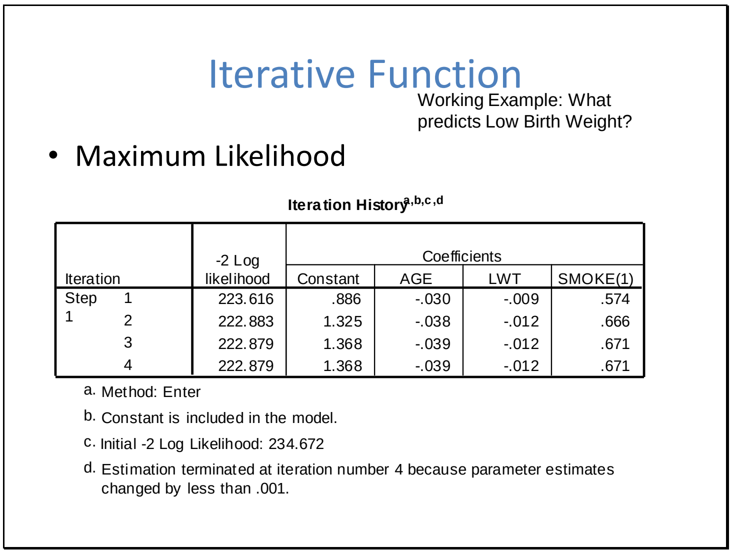
Slide 69



Generalised Linear Model

- Normal binary DV is problematic to predict
- Instead we predict a transformation of the DV:
  - Loge(p/1-p)
- Predicting a transformation of the DV is what makes it the generalised linear model
- The link function is called the logistic link

Slide 70



Iterative Function

Working Example: What predicts Low Birth Weight?

- Maximum Likelihood

Iteration History[a,b,c,d]

Iteration		-2 Log likelihood	Coefficients			
			Constant	AGE	LWT	SMOKE(1)
Step 1	1	223.616	.886	-.030	-.009	.574
	2	222.883	1.325	-.038	-.012	.666
	3	222.879	1.368	-.039	-.012	.671
	4	222.879	1.368	-.039	-.012	.671

[a] Method: Enter

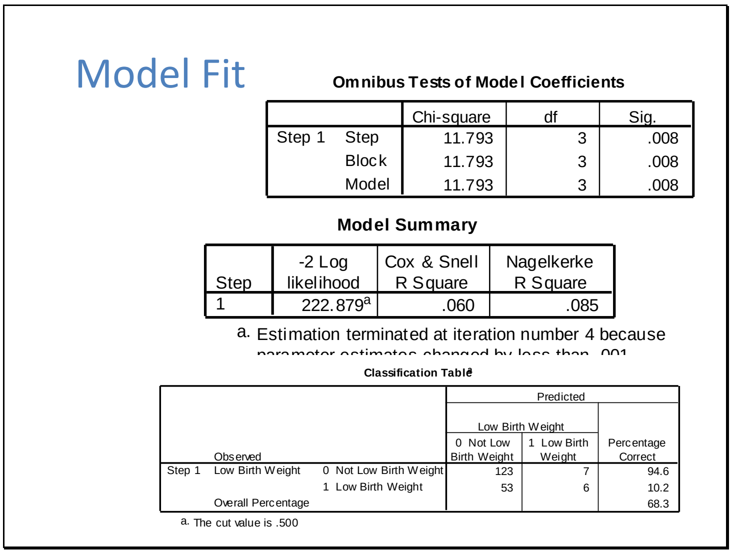
[b] Constant is included in the model.

[c] Initial -2 Log Likelihood: 234.672

[d] Estimation terminated at iteration number 4 because parameter estimates changed by less than .001.

Maximum Likelihood is an alternative to least squares for finding an optimal set of regression coefficients. It is the standard procedure for logistic regression.

Slide 71



Model Fit

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	11.793	3	.008
	Block	11.793	3	.008
	Model	11.793	3	.008

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	222.879[a]	.060	.085

[a] Estimation terminated at iteration number 4 because parameter estimates changed by less than .001

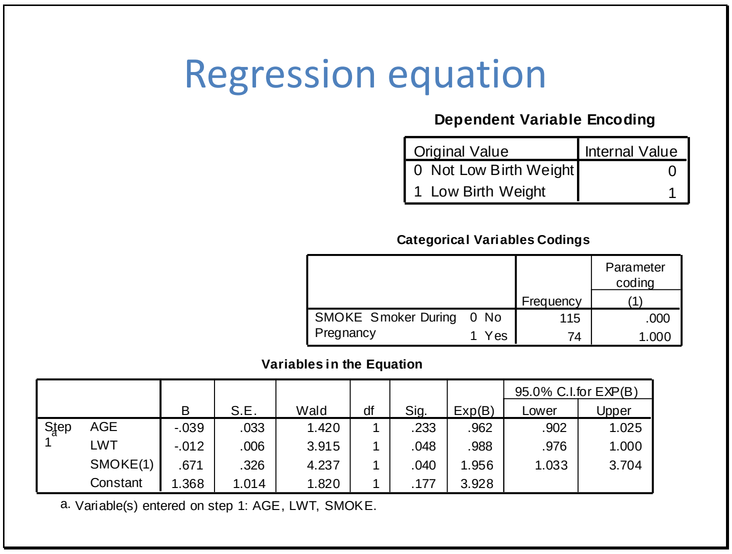
Classification Table[a]

			Predicted		
			Low Birth Weight		Percentage Correct
Observed			0 Not Low Birth Weight	1 Low Birth Weight	
Step 1	Low Birth Weight	0 Not Low Birth Weight	123	7	94.6
		1 Low Birth Weight	53	6	10.2
	Overall Percentage				68.3

[a] The cut value is .500

Chi-square test: This compares the prediction compared to another model (most commonly a model with only a constant). Larger chi-squares for a given df lead to smaller p values. A statistically significant chi-square in this context indicates that the model leads to a statistically significant improvement in prediction.
Model Summary: Smaller -2 Log liklihoods indicate better prediction. There a couple of measures that attempt to match r-squared in the multiple regression context, although interpretation is not as clear.
Classification Table: Classification is an important indicator of success of the model. Prediction can be compared to model with just the constant.

Slide 72



Regression equation

Dependent Variable Encoding

Original Value	Internal Value
0 Not Low Birth Weight	0
1 Low Birth Weight	1

Categorical Variables Codings

		Frequency	Parameter coding
			(1)
SMOKE Smoker During Pregnancy	0 No	115	.000
	1 Yes	74	1.000

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)	95.0% C.I.for EXP(B)	
								Lower	Upper
Step 1[a]	AGE	-.039	.033	1.420	1	.233	.962	.902	1.025
	LWT	-.012	.006	3.915	1	.048	.988	.976	1.000
	SMOKE(1)	.671	.326	4.237	1	.040	1.956	1.033	3.704
	Constant	1.368	1.014	1.820	1	.177	3.928		

[a] Variable(s) entered on step 1: AGE, LWT, SMOKE.

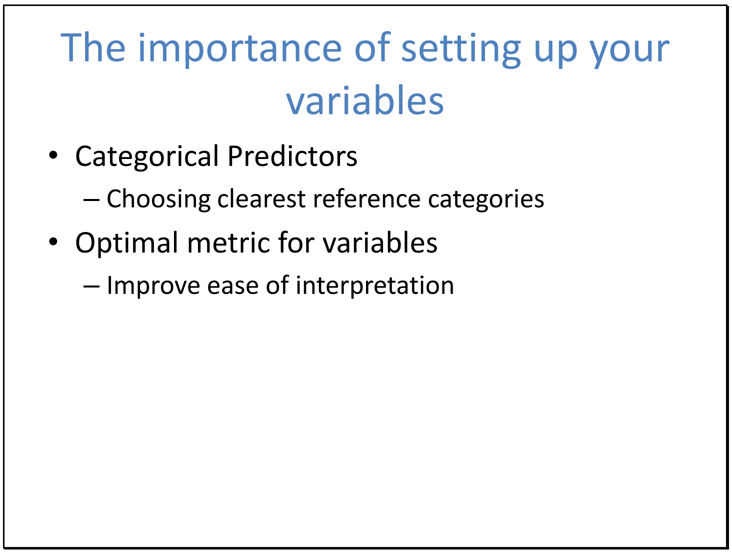
Interpretation of coefficients

B: Positive values indicate increases in the predictor are associated with greater likelihood of the event occuring. Specifically these are the regression coefficients used to predict the logit (i.e., natural log of the odds).

Exp(B): This is the inverse natural log of B. It has a very nice interpretation. An increase of one on the predictor indicates exp(B) change in the odds of the event occurring. For example, Smoking leads to 1.956 greater odds of the baby being born low birth weight.

Tests of statistical significance and 95% confidence intervals are also available. It is also important to check the interpretation of the categorical variables including the outcome variable.

Slide 73



The importance of setting up your variables

- Categorical Predictors
  – Choosing clearest reference categories
- Optimal metric for variables
  – Improve ease of interpretation