

Research Design & Analysis:

Lecture 6: Regression

Recap: Regression

- Research questions:
 - Which of a **range** of variables are most strongly related to a specific variable of interest?
 - Do these variables do better than chance or simple mean scores at **predicting** scores on a variable of interest?
 - How might these variables **combine** to predict a variable?
- Multiple variables:
 - Could just carry out many simple correlations
 - Provide several Bivariate associations (2 variables correlated)
 - Still only a non-causal correlation produced
 - Will not indicate our overall success with all variables used

Recap: Regression

- Important extension of 'regression' over 'correlation':
 - **Combined** association effect of the predictor variables
 - **Multiple** regression
 - Treating the predictors (separate variables) as a combined variable in relation to the dependent variable
 - How 'associated' they are as a whole
 - How 'predictive' of the DV they are as a whole
 - (nb. Will also look at each in isolation)

Assumptions

- Continuous data (NB: dummy variables)
- Normal distribution of data
- Linear relationships between IVs and DVs
- Homogeneity of variance

Regression

- The concept of collinearity is also important here:
 - The correlation between the **independent** variables
 - If two are highly correlated, the effect of the second variable is greatly reduced
 - The amount of variance the second variable shares with the DV is already explained by the first variable
 - One variable may be 'redundant' in the model

Regression

- Example: What factors might predict maths ability in children?
 - Study might take following measures
 - Measure of current maths ability
 - Measures of IQ, spoken language development, memory, musical ability etc.
 - Can then examine which of these latter variables are most strongly correlated to the measure of maths achievement
 - Then can examine which of these individual variables is the best **predictor** of maths scores
 - How predictive or explanatory are the predictor variables **combined** in relation to maths achievement

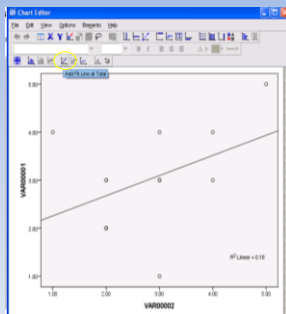
Regression

- Terminology check:
 - The variable we are trying to predict from scores of other variables:
 - The 'dependent' variable (maths achievement score)
 - The variables whose association (and predictive power) with the DV is being investigated:
 - 'Predictor' or 'independent' variables (Measures of IQ, spoken language development, memory, musical ability)

Correlation and Regression

- Regression is an extension of the approach that correlation adopts
- In SPSS add a 'line of best fit' (fit-line) to our scatterplot
- What can the fit-line show?
 - The best prediction for each variable
 - Based upon the relation between the variables
 - Precise amount of change in one variable based on change in the other variable

Adding a fit (regression) line



Double click on your scatterplot to add the 'fit-line' in the chart editor

Click the button indicated (yellow circled) in the menu above the scatterplot

Use 'linear' option (default option) then close chart editor

Line sloping up = positive correlation

Line sloping down = negative correlation

Regression

- Regression can be used to analyse a more 'causal' relation between variables
- Theory suggests a dependent variable and independent (or explanatory) variable(s)
 - Can also use more than one explanatory variable (IV)
- Regression analyses amount of change in the dependent variable based on change in the explanatory/independent variables
- Can be used to make **predictions** for the DV based on scores for the IV

Assessing goodness of fit

- R^2 - expressed as a percentage
- The proportion of the variability in your DV that can be explained by its relationship with:
 - The combined model (all IVs) together
 - Each individual IV (unique variance explained)
 - *Predicting the variance explained*

Regression

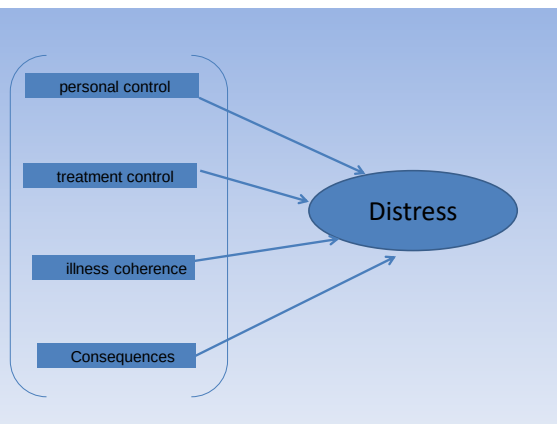
- Multiple Regression statistics of interest:
 - Multiple regression indicates how good a 'fit' all the predictors are in explaining the DV
 - Can assess the degree to which the DV can be predicted from the set of predictors
 - R = multiple correlation coefficient
 - (between 0 and 1)
 - The greater R value, the less residual values
 - As with r , of more use can be R^2
 - The % of variance in the DV that 'model' explains
 - Adjusted R^2 - adjusted for sample size
 - Beta values represent amount of contribution (how good a fit) for each of the independent variables

Example....

Do illness beliefs predict psychological outcomes?

Regression design:

- Do illness representations predict distress in people undergoing cancer genetic risk assessment?
- Which ones are most predictive?
- Illness representations
 - personal control,
 - treatment control
 - illness coherence
 - Consequences



Reporting regression

- **Descriptive statistics**

- More useful for comparisons
- But still generally useful for correlation/regression (e.g. mean academic score, mean income etc)
- Mean, standard deviation score for all IVs and DV

Descriptive Statistics			
	Mean	Std. Deviation	N
toties2	15.84	15.816	200
t1 consequences	19.4185	4.34063	173
t1 personal control	18.3341	3.98288	173
t1 illness coherence	8.6293	2.69804	174
t1 treatment control	14.0900	2.46651	174

Reporting regression

- **Correlations analysis**

- Needed to assess correlations between IVs
- The **correlations** analysis shows the simple bivariate correlations between all variables.
- analyse and report whether any of the independent variables are correlated with one another.

		Correlations					
		totes2	t1 consequences	t1 personal control	t1 illness coherence	t1 treatment control	totes1
Pearson Correlation	totes2	1.000	.496	.007	-.143	.078	.670
	t1 consequences	.496	1.000	-.066	-.178	.119	.389
	t1 personal control	.007	-.066	1.000	.057	.450	-.031
	t1 illness coherence	-.143	-.178	.057	1.000	.100	-.146
	t1 treatment control	.078	.119	.450	.100	1.000	.002
Sig. (1-tailed)	totes1	.670	.389	-.031	-.146	.002	1.000
	totes2	.000	.000	.465	.030	.153	.000
	t1 consequences	.000		.193	.009	.060	.000
	t1 personal control	.465	.193		.230	.000	.345
	t1 illness coherence	.030	.009	.230		.095	.029
N	t1 treatment control	.153	.060	.000	.095		.489
	totes1	.000	.000	.345	.029	.489	
	totes2	.200	.173	.173	.174	.174	.195
	t1 consequences	.173	.173	.172	.173	.173	.168
	t1 personal control	.173	.172	.173	.173	.173	.168
	t1 illness coherence	.174	.173	.173	.174	.174	.169
	t1 treatment control	.174	.173	.173	.174	.174	.169
	totes1	.195	.168	.168	.169	.169	.196

Reporting regression

• Main elements

– Model Summary table (model = all IVs included)

- Adjusted R squared value = amount of variance the model explains for the DV (as a %)

– ANOVA table

- Whether model performs better than mean DV guess
- If significant $p < 0.05$ if non-significant $p > 0.05$

Model Summary ^a									
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics				
					R Square Change	F Change	df1	df2	Sig. Change
1	.501 ^a	.251	.233	13.862	.251	13.982	4	167	.000
a. Predictors: (Constant), t1 treatment control, t1 illness coherence, t1 consequences, t1 personal control									
b. Dependent Variable: totes2									
ANOVA ^a									
Model		Sum of Squares	df	Mean Square	F		Sig.		
1	Regression	10731.942	4	2682.986	13.982		.000 ^b		
	Residual	32045.619	167	191.890					
	Total	42777.561	171						
a. Dependent Variable: totes2									
b. Predictors: (Constant), t1 treatment control, t1 illness coherence, t1 consequences, t1 personal control									

Adjusted R Square: Accounting for sample size (1.0 = 100% explained)

Model explains 23.3% of variance for the DV (in this example)

ANOVA – is the model significantly better than simply making predictions using the mean score for the DV ($p < 0.05$ if significantly better)

Reporting regression

- Coefficients table (can be reported as a table)
 - Analysis for individual IVs in model
 - Beta values: how much each IV change equates to change in the DV
 - Larger the Beta value = more change in DV associated
 - Negative (minus sign) = negative association (correlation)
 - (Not the same as NO association!)
 - Significance value for each IV
 - $P < 0.05$ equals significant as an IV
 - IV can still contribute to model if $p > 0.05$
 - Beta value will specify how much influence in model

Coefficients ^a										
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
(Constant)	-19.230	8.754		-2.197	.029					
1 t1 consequences	1.775	.253	.487	7.020	.000	.496	.477	.470	.932	1.073
t1 personal control	.153	.301	.039	.510	.611	.007	.039	.034	.782	1.278
t1 illness coherence	-.344	.402	-.059	-.856	.393	-.143	-.066	-.057	.953	1.049
t1 treatment control	.055	.491	.009	.112	.911	.078	.009	.007	.765	1.308

a. Dependent Variable: tories2

Coefficients ^a										
Model	Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
	B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
(Constant)	-19.230	8.754		-2.197	.029					
t1 consequences	1.775	.253	.487	7.020	.000	.496	.477	.470	.932	1.073
t1 personal control	.153	.301	.039	.510	.611	.007	.039	.034	.782	1.278
t1 illness coherence	-.344	.402	-.059	-.856	.393	-.143	-.066	-.057	.953	1.049
t1 treatment control	.055	.491	.009	.112	.911	.078	.009	.007	.765	1.308

a. Dependent Variable: tories2

Values for the individual predictor variables:

Beta values: how much change in DV observed with change in IV (in Standard Deviations)

Higher the Beta score = more of a predictor for the dependent variable (better fit)

Note: can be a minus value (negative correlation)

p-value (sig.) also shown for each predictor variable

Hierarchical Multiple Regression... a quick example

Are some illness beliefs stronger predictors than others?

Multiple regression on those completing time 1 and time 2 assessments
(n=201)

- Step one:
 - Illness representations (personal control, treatment control, illness coherence, consequences)
- Step two:
 - Illness representations
 - Causal beliefs
- Step three:
 - Illness representations
 - Causal beliefs
 - Baseline distress/PA/NA

Predicting cancer-specific distress at Time 2

		β	Adjusted R^2	R^2 change	F
Step 1	Consequences	.47***	.25		13.47***
	Personal control	-.03			
	Treatment control	.04			
	Illness coherence	-.06			
Step 2	Consequences	.46***	.26	.01	5.47***
	Personal control	-.06			
	Treatment control	.04			
	Illness coherence	-.04			
	Psychological attributions	.05			
	Risk factors	-.01			
	Harm/Neglect	.03			
	Hereditary	.04			
	Chance/bad luck	.05			
	Accident/injury	.03			
Step 3	Consequences	.25***	.53	.27	15.73***
	Personal control	-.06			
	Treatment control				
	Illness coherence				
	Psychological attributions				
	Risk factors				
	Harm/Neglect				
	Hereditary				
	Chance/bad luck				
	Accident/injury				
	Baseline distress				
	*** p<.001				

Stronger beliefs in the serious consequences of being found to be at increased risk of cancer significantly predicted higher levels of cancer specific distress at follow-up

Predicting Positive Affect at Time 2

		β	Adjusted R^2	R^2 change	F
Step 1	Consequences	.11	.12	.12	5.41***
	Personal control	-.05			
	Treatment control	.33***			
	Illness coherence	.01			
Step 2	Consequences	.10	.16	.04	3.00***
	Personal control	-.04			
	Treatment control	.29***			
	Illness coherence	.03			
	Psychological attributions	-.05			
	Risk factors	.05			
	Harm/Neglect	.00			
	Hereditary	.13			
	Chance/bad luck	-.11			
	Accident/injury	.14			
Step 3	Consequences	.11	.16	.00	2.72***
	Personal control	-.04			
	Treatment control				
	Illness coherence				
	Psychological attributions				
	Risk factors				
	Harm/Neglect				
	Hereditary				
	Chance/bad luck				
	Accident/injury				
	Baseline distress				

Stronger beliefs in screening/surgery helping reduce chances of getting cancer significantly predicted higher levels of positive affect at follow-up

Summary of results section write up:

- Report in text the main statistics from these:
 - Descriptives + correlations
 - Overall model:
 - Adjusted R squared value (% explained)
 - ANOVA significance value
 - Beta and significance values for each IV (coefficient)
 - Refer clearly to which IV each of these values relates
 - Can use a table format for coefficients

Example of reporting main regression statistics in text

"The regression analysis showed that as an overall model the set of independent variables (a,b,c) produced an adjusted R-squared value of 0.23. The combined independent variables therefore explained 23% of the variance in the dependent variable (dv) scores.

The model performed significantly better at predicting DV scores than the mean dependent variable score, as shown by the ANOVA table significance value, $p < 0.05$.

Table 1 shows the coefficients for the separate independent variables in the analysis. The variable (a) was a significant factor ($p < 0.05$) with a Beta value of -1.35. This indicated a negative direction of correlation, with a decrease of 1.35 standard deviations for the DV found for every one standard deviation increase in (a) scores."

Note: you must clearly indicate (name) which variables are being referred to in your report for both independent and dependent variables

Discussion sections

- Have you supported or not supported hypotheses?
 - Overall model significant?
 - How much variance did the model explain
 - (Even if not significant in ANOVA table!)
 - How does this finding fit with previous lit?
 - Discussion of each IV separately
 - Which showed more/less contribution to model?
 - Which were significant individually?
 - How do these findings fit with previous lit?
- Could the model make good predictions?
- Is the model (or elements) a good explanation?

Discussion sections

Nb: Marks not dependent on 'significance' of tests

- Correctly interpret non-significant model!
- Discuss in relation to whether support/contradict previous literature
- Any evidence to suggest better models for future studies?
- Balanced methodological critique

Reminder: Abstracts!

- Abstract:
 - Though you should place this at start/front of report is often best written last (because you then have all the rest)
- Condensed summary of entire report
 - Area/theory being investigated
 - Methodology
 - Results (not actual statistics, just summary)
 - Conclusions that could be supported

e.g., a good abstract

Psychological responses to genetic risk assessment are complex and not always linked to objective risk status (Phelps et al., 2007). This study explored whether Leventhal's (2003) Self-Regulation Model (SRM) was a useful predictor of psychological responses to cancer genetic risk assessment. Questionnaire data from an adapted version of Moss Morris et al.'s Illness Perceptions Questionnaire was analysed from 210 individuals who had undergone cancer genetic risk assessment to explore associations between baseline psychological variables and psychological distress one month later. Multiple regression analyses indicated that illness perceptions explained 21.2% of the variance in cancer-specific distress ($F(1,145) = 5.17, p > .001$) with strong beliefs about the consequences of being at risk of cancer being the only variable to uniquely contribute to psychological distress ($p < .001$). Overall, these findings support Kelly et al. (2008) who also suggested that illness perceptions based around the SRM could help explain individuals responses to genetic testing.
