Spring Term 2003
Research Methods II

Using SPSS for Multiple Regression

For each subject you will have three data points: Their BJW score (the dependent variable) and their scores on the other two tests that you used (the independent variables). As before, enter the data using one column for each variable and one row for each subject.

Looking at the data

First, have a look at the distributions of each of your variables with histograms and check that your variables are roughly normally distributed without outliers. To generate histograms click **Graphs > Histogram**; and click on **Display normal curve** to get an idea of how closely your data approximate normality.

Next have a look at the scatterplots of each of your variables against each other: **Graphs > Scatter** (leave simple as default) **> Define,** choosing two of your variables. Check that the relationship between each pair of variables is linear and without outliers. Note that linear does not mean there has to be a positive or negative relationship - it could be that a horizontal straight line is the best description of the relationship (if the data looks like "plum pudding" - a ovalish mass of points - a horizontal straight line is the best fit). It just means that there is not systematic curvature in the relationship.

Finally, you might like to make a 3D plot of your variables: **Graphs > Scatter > 3D > Define**. Enter your three variables for the x, y, and z dimensions. Click on OK, and the graph will appear. To see the graph from different perspectives, double click on the chart. Then on the tool bar just above the graph, click on the last icon - if you move the cursor to it, text will briefly appear saying "set/exit spin mode". A new box will appear, with icons for different rotations. If you click on one of the rotation icons, and keep the mouse pressed down, the graph will rotate and you may be able to see what your data would look like in 3D. Does it look like a flat plane would be able to fit the data? Are there any outliers?

Performing the multiple regression

- Pull down the menu items **Analyze > Regression > Linear.**
- Highlight and enter BJW into the Dependent box, and the other variables into the Independent(s) box.
- Click on the Statistics box: leave all the defaults as they are and click Descriptives (this will give you the mean, SD and N for each variable, and all pairwise correlations and their one-tailed significance); click on the Continue button and you will be back in the Linear Regression box
- Click on the Save box. In the "predicted values" section, click on unstandardized; in the "residuals" section also click on unstandardized. This will create two new columns - a column of fits (predicted values) and a column of residuals (difference between actual BJW scores and fits). Click on the Continue button and you will be back in the Linear Regression box. Click on OK to run the analysis.

Checking assumptions

· Produce a scatter plot of BJW against predicted values (SPSS will have called this column "pre_1"). Regression assumes that the relationship is linear (remembering that a horizontal straight line is still a straight line), with the points scattered about the line (in the vertical direction) roughly normally with the same variance for each predicted value. You can create the best fitting line by double-clicking on the graph, which will open a "Chart Editor" box. Click on "Chart", "Options" then on "Fit Line Total" and then OK. Check by eye. Are the points scattered about this straight line, or do they systematically curve around it? Are the points scattered about the line roughly normally with roughly equal variance for each predicted value? Are there any outliers?

· Produce a histogram of the residuals (SPSS will have called this column "res_1"). Regression assumes that the residuals are distributed roughly normally. Check by eye that they are.

SPSS output

[my comments in brackets]

Regression

Descriptive Statistics

	Mean	Std. Deviation	N
BJW	8.2500	2.9890	20
REL	5.7500	2.0995	20
AUTH	7.8000	1.6416	20

[Have a careful look at this table of correlations. What is the correlation between each IV and the DV? Also check the correlation between the IVs. If they are correlated, the regression slope for an IV may indicate a different relationship with BJW than the correlation between that IV and BJW- because of controlling for the other IV in the regression solution.]

Correlations

		BJW	REL	AUTH
Pearson Correlation	BJW	1.000	.891	.493
	REL	.891	1.000	.397
	AUTH	.493	.397	1.000
Sig. (1-tailed)	BJW	.	.000	.014
	REL	.000	.	.042
	AUTH	.014	.042	.
N	BJW	20	20	20
	REL	20	20	20
	AUTH	20	20	20

Variables Entered/Removed[b]

Model	Variables Entered	Variables Removed	Method
1	AUTH, REL[a]	.	Enter

a. All requested variables entered.

b. Dependent Variable: BJW

[The Multiple R given below - just called "R" - is the correlation between the predicted values of BJW and the actual values of BJW. It tells you how well you can predict BJW based on your IVs.
As explained in the first handout for this module, R Square tells you the proportion of variance (in BJW) explained (by the IVs). To see what "Adjusted R Square" means consider what would happen if you ran only two subjects with one DV and one IV. You would have two data points to fit a line to - and you can always fit a line perfectly to two points! So R Square (and R) would be one even if the IV in the population bore no relation to the DV (i.e. population R could be zero). Similarly, if you had two IVs and three subjects, you would be fitting a plane to three points - this is always possible, so once again R=1, regardless of the true population relationship. So to get an unbiased estimate of what the actual population correlation between the DV and the fits, the R square value needs to be "shrunk" by a precise amount that depends on the number of subjects you've run relative to the number of IVs. This is what adjusted R square is.)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.904[a]	.817	.796	1.3510

a. Predictors: (Constant), AUTH, REL

b. Dependent Variable: BJW

[The following Analysis of Variance Table is testing whether the Multiple R is significantly different from zero; i.e. do both IVs taken together allow significant prediction of BJW?]

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	138.722	2	69.361	38.003	.000[a]
	Residual	31.028	17	1.825		
	Total	169.750	19			

a. Predictors: (Constant), AUTH, REL

b. Dependent Variable: BJW

[Now we come to the interesting bit. The Bs are the raw regression slopes - the effect of each variable on BJW controlling for the other variable. The next column gives the standard error of each B. The t value (meaning value of the t-test) is B/SE. "Sig" tells you whether the B is significantly different from zero. The degrees of freedom for the t-tests are those given in the residuals row in the ANOVA Table above (i.e. 17 in this case). The Beta values are the standardized regression slopes; i.e. the slopes you would get if you normalized each variable (expressed each subject as a difference from the mean in SD units) and then did the regression analyses (see last handout for further explanation). The constant is the intercept - not of any interest to you, because the scales have arbitrary and probably meaningless zero points.]

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.860	1.531		-.562	.582
	REL	1.175	.161	.825	7.305	.000
	AUTH	.302	.206	.166	1.467	.161

a. Dependent Variable: BJW

Tips for writing up:

Introduction

Start general, and then progressively focus on the precise question addressed by the study. State your hypotheses for how your IVs will predict BJW, and make sure you provide justification for each hypothesis. Consider the difference between hypotheses just about the correlations between the IVs and BJW and the regression slopes. Commonly on this module, students just state hypotheses about correlations; but then why do a multiple regression, why not just produce a correlation table? In other words, consider whether controlling one IV will alter the relation between the other IV and BJW. Maybe you predict the regression slopes and correlations should both show a relation, and there will be no effect of controlling each IV on the other IVs, because each IV is hypothesized to have an independent effect on BJW: that's fine, just say so. Or maybe you think that there should be a correlation (like between shoe size and intelligence) but the apparent relationship will disappear when another variable is controlled for (like when the regression slope for shoe size goes to zero when age is controlled for). Either way, say why you make the predictions that you do: In terms of your theory of how the variables are related, why does controlling one variable change or not change the relation between the other two?

In developing your thinking about these issues, bear in mind the correlations will at most be of a moderate size. This means that if you postulate a relationship between e.g. empathy and political views, that relationship will be far from perfect. Thus, even though the more right wing one is, the less empathic you think they will be (say), there will still be some people who

are very right ring yet relatively high in empathy. Thus, you can consider people of a certain constant level of political belief (say, right wing), and then think about how if you took one person who is right wing and high empathy and another who is right wing and low empathy, how would BJW vary? You need to consider this to work out your predictions for the regression slope for empathy (i.e. the effect of empathy on BJW holding political view constant).
The moderate values of your correlations also mean that no two correlations determine the third. For example, just because you postulate that trust will correlate positively with BJW and also religiousness will correlate positively with BJW, that does not mean you can infer trust and religiousness must correlate positively. They might do, but they might not. Consider: If you took some dancers and gave them a surprise written test on the theory of dance, you might find that dancing ability correlated positively with the test (the more they had danced, the more they understood about it), and also performance on a history test correlated positively with the test (because ability to do well on a written test partly depends on general writing skills). But the history test and dancing ability might not be not correlated at all, or even be negatively correlated (because the people who were good at dancing had missed their history lessons). In sum, you must carefully consider your reasons for each prediction you make for the correlations and multiple regression slopes.

Method section

· In the materials section, fully cite any questionnaires you have based yours on. Indicate why and how you have changed any questions. You should do this in some detail because you are changing a published instrument. If you have constructed your own questionnaires, describe how you construe whatever the construct is (i.e. "optimism") - do not just give the reader a label, give the reader a feeling of the sorts of questions you constructed (e.g. do they fall into different categories?). How many questions were there? Were there different orderings of questions for different subjects?

· Fully indicate the response options open to the subjects: yes/no? A six point scale? How were the points defined? How many questions indicating a high score on the construct demanded a yes or no answer (or a 1 or 6 answer)?

· How were the questionnaires scored? E.g. converting "1"s to "6"s

· The design is called a correlational design, because the experimenter does not determine the level a subject has on an independent variable, the subject's level just has to be observed. Just call it correlational, not between-subjects nor within-subjects.

Results section

· Give a table of means and SDs for each variable. State what the maximum and minimum possible scores are for each variable either here or in the materials section.

· Give a table of all pairwise correlations between the variables. State whether the variables satisfied the assumptions of correlation; i.e. each variable normally distributed; scatterplots

showing linear relationships between each pair of variables. Indicate any significant correlations (using a two tailed test, unless you predicted a direction of the correlation; note SPSS gives 1-tailed probabilities; double these to get the 2-tailed probabilities).

· State you conducted a multiple regression, giving the DV and IVs. State whether the assumptions of multiple regression were satisfied; i.e. linear relation between BJW and fits, with normal distribution of BJW about the line and equal variance of BJW about the line; normal distribution of residuals.

· Decide whether you want to report Bs (raw slopes) or Betas (standardized slopes). The raw values on your questionnaires probably wouldn't mean much to the reader - what meaning does "one unit" have on your questionnaires?? So on those grounds the standardized slopes would probably be more meaningful. Report the full regression equation: e.g. bjw = -0.83*authoritarian personality + 0.17*religiousness. (Make sure you state clearly whether the slopes are raw or standardized.)

· Report the adjusted R square, and its test of significance; e.g "Adjusted R Square = .80, $F(2,17) = 38.00$, $p < .005$."

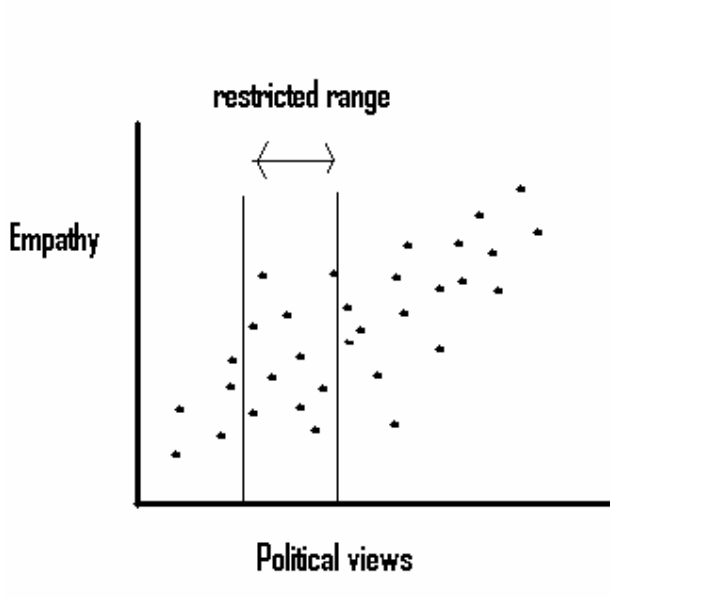
· Report the tests of significance for each regression slope; e.g. "The effect of religiousness was significant, $t(17) = 7.31$, $p < .0005$."

Discussion section

Restate the aims of the investigation and any predictions you made. Use the correlations and regression slopes to interpret the relationships of the variables with each other, and state clearly which predictions were supported. Indicate how you appreciate that correlations and regression slopes give you different information (i.e. the latter controls for other variables) - how does this apply in your particular case? Also bear in mind that a correlational design does not show causality between variables - only an experimental design does that. It could be used to distinguish different causal theories that made different predictions about how the causes would produce different associations, but it couldn't establish that the associations were specifically due to direct causation. Discuss how you might assess the validity (and reliability) of the questionnaires you used. Finally, make some suggestions for how research might carry on from where you left off.

Some people find it difficult to know what to discuss if all or some of their correlations and slopes are non-significant. If you predicted a correlation between two variables and your result is non-significant, here are four possibilities to consider:
1) There really is not a population correlation and you need to re-think your reasons for believing there would be one.
2) There really is a population correlation, but it is relatively small, and you did not run enough subjects to have a good chance of detecting it.
3) There really is a population correlation between the two constructs, e.g. empathy and political views, but your questionnaires were not reliable (did you ask enough questions?) or not valid measures of those constructs.

4) There really is a population correlation (in a sufficiently varied population), and your questionnaires were both reliable and valid, but your sample was restricted in its values of one or more of the constructs. For example, if you tested your friends, your friends might have very similar political views as yourself, bringing about a restriction in the range of political views tested. Consider the graph below of empathy versus political views below. Overall, there is a strong correlation. But if you consider just the points in the small range between the lines, there is no correlation at all.



To see if there is a range restriction problem, consider the means and standard deviations of your scores. Can you compare them to previous studies that investigated those variables? Or maybe you can simply make a plausible judgement based on your means and SDs as to what extent your sample varied over an interesting range of values.