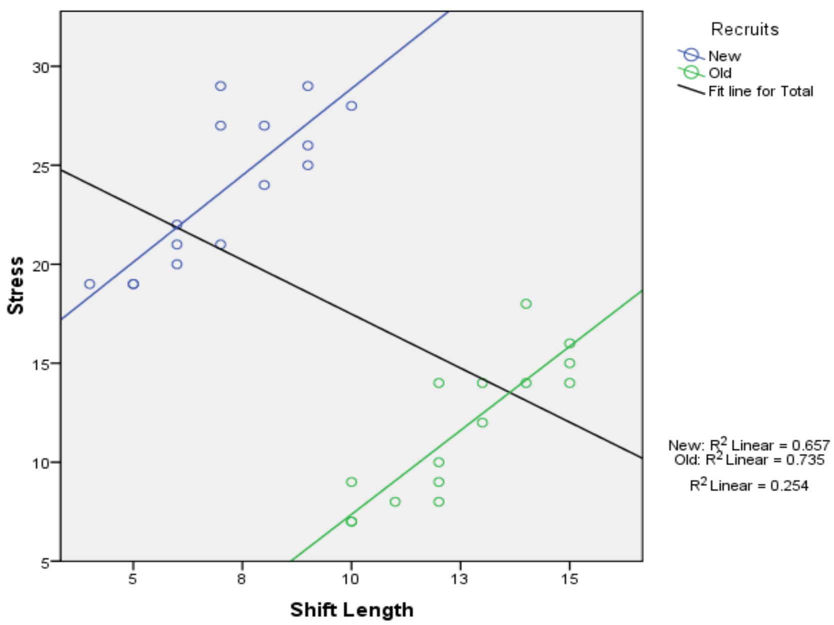
Correlation and Regression

Correlations

- Correlations assume relationships are linear
- Correlations are range specific
- Correlations assume data is homogenous
- Outliers can have large effects
- Normality only assumed when significance testing



Example of heterogenous subsamples deflating the overall r value.

Some examples of linear and non linear relationships.

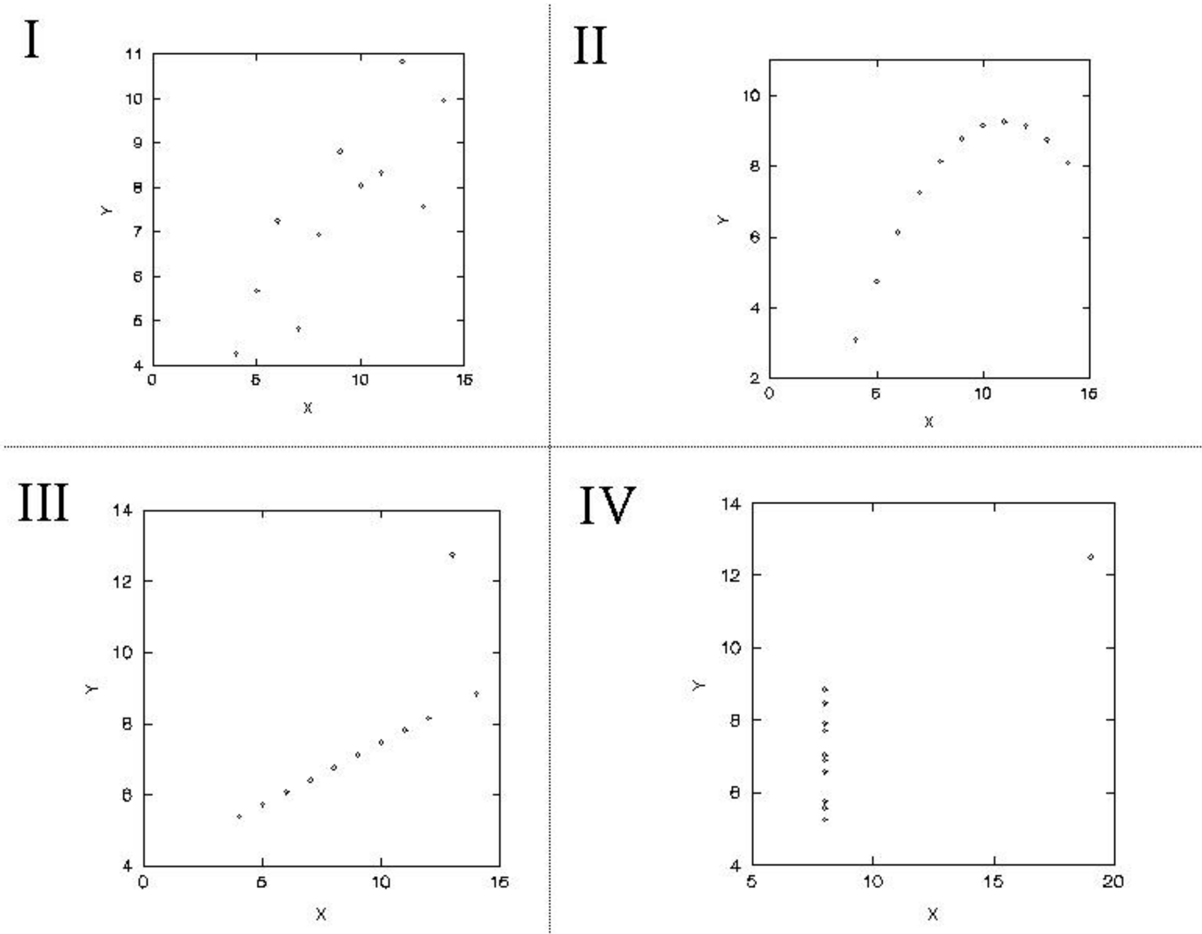
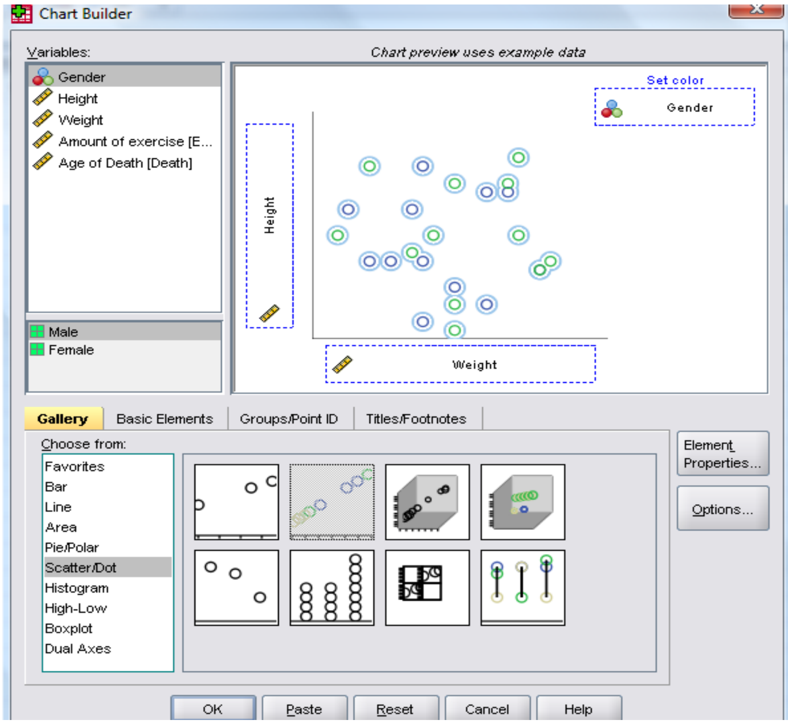


Chart builder for scatter plots



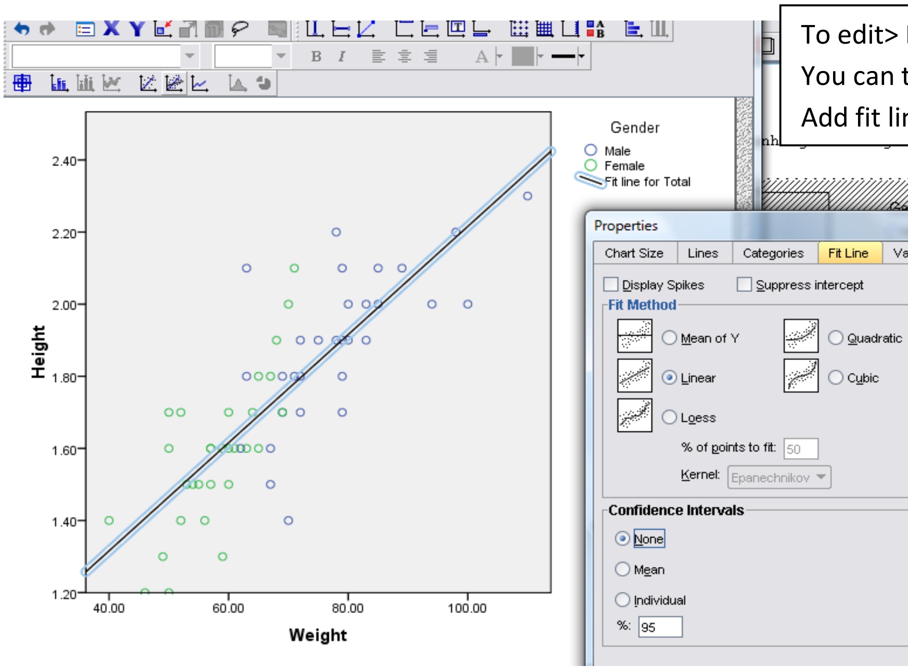
Graphs> Chart Builder > Highlight Scatter/dot

Select either (simple scatter)

Or (for if you have a grouping variable)

Place your variables in the axes boxes

And (if appropriate grouping variable in 'set color'



To edit> Double click on graph for chart editor

You can then change colors/ weightings of lines

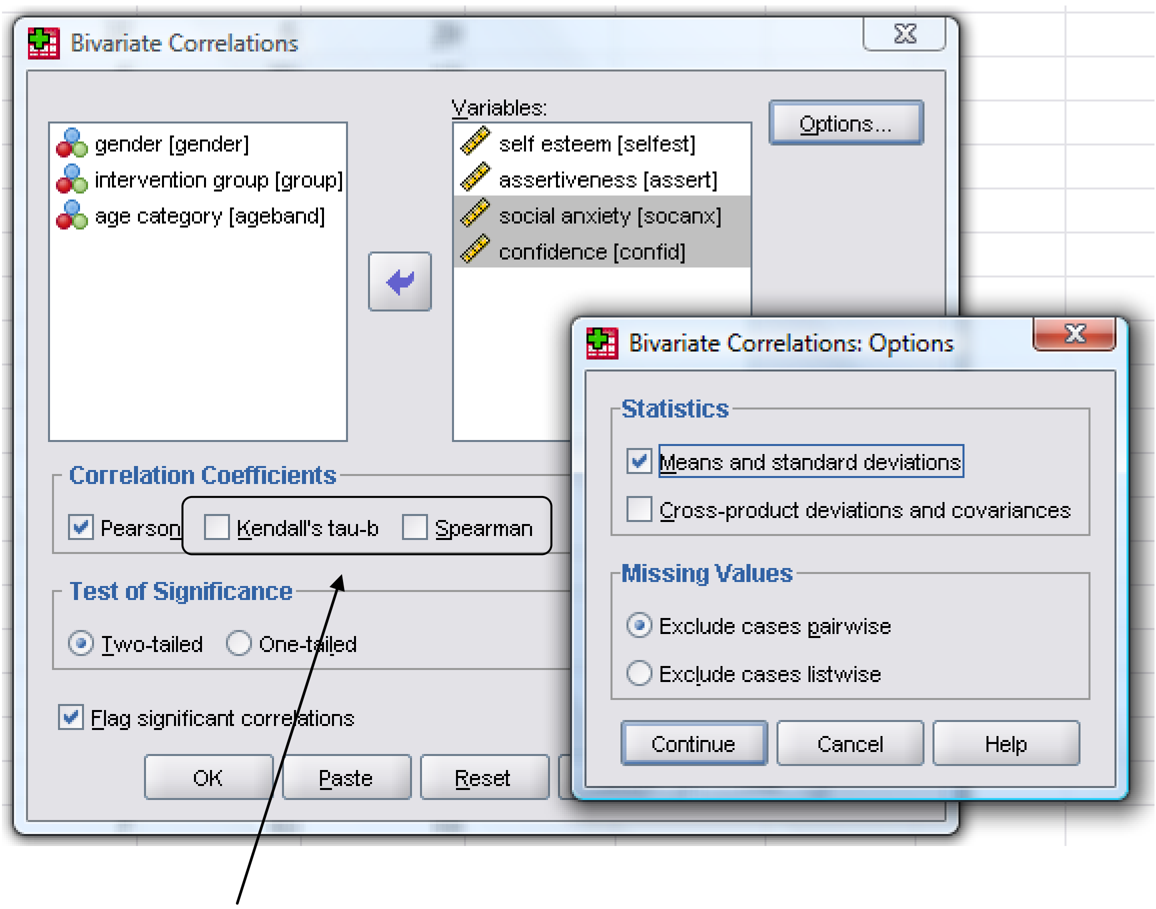
Add fit lines for whole group and subgroups

Running the correlation

Analyze > Correlate > Bivariate

Select the variables of interest

You can ask for descriptive statistics by clicking on OPTIONS



If you would like to assess the relationship in non parametric data you can simply select Kendalls Tau-b or Spearman

Main output

Correlations

```
[DataSet1] C:\Users\Roz\Documents\MsCstuff.sav
```

Descriptive statistics for the variables which is needed for your write up

The top and bottom of the table are mirror images you will only need to write up one half

Descriptive Statistics

	Mean	Std. Deviation	N
self esteem	15.22	7.861	82
assertiveness	16.94	8.408	82
social anxiety	10.23	7.139	82
confidence	24.70	6.878	82

Correlations

		self esteem	assertiveness	social anxiety	confidence
self esteem	Pearson Correlation	1	.745**	-.603**	.727**
	Sig. (2-tailed)		.000	.000	.000
	N	82	82	82	82
assertiveness	Pearson Correlation	.745**	1	-.376**	.723**
	Sig. (2-tailed)	.000		.000	.000
	N	82	82	82	82
social anxiety	Pearson Correlation	-.603**	-.376**	1	-.471**
	Sig. (2-tailed)	.000	.000		.000
	N	82	82	82	82
confidence	Pearson Correlation	.727**	.723**	-.471**	1
	Sig. (2-tailed)	.000	.000	.000	
	N	82	82	82	82

**. Correlation is significant at the 0.01 level (2-tailed).

**** = significant**

Report r, p and N (if it differs in the differing correlations)

The write up:

In a sample of 82 participants bivariate correlations indicate positive significant relationships between self esteem and assertiveness: r = .745, p <0.001; self esteem and confidence: r = .727, p < 0.001; and a negative relationship between self esteem and confidence: r = -603, p< 0.001

For this many variables I would create a correlation table using the lower triangle

Table 1:

	Self Esteem	Assertiveness	Social Anxiety	Confidence
Self Esteem				
Assertiveness	.745**			
Social Anxiety	-.603**	-.376**		
Confidence	.727**	.723**	-.471**	
Mean	15.22	16.94	10.23	24.70
SD	7.86	8.41	7.14	6.88

Partial Correlations

If we would like to focus on the association between confidence and assertiveness we can see from Table 1 that this association is highly significant: $r = .727$, $p < 0.001$. However, perhaps this relationship is explained by a third variable and is thus a redundant relationship (or a spurious finding) If we were to run a partial correlation (Analyze > correlate > partial) between Confidence (X) and Assertiveness (Y) whilst controlling for Self Esteem (Z) the relationship between X and Y changes when we control for Z. The relationship decreases in significance although continues to be significant: $r = .395$, $p < 0.001$.

Correlations

Control Variables			assertiveness	confidence
self esteem	Assertiveness	Correlation	1.000	.395
		Significance (2-tailed)	.	.000
		df	0	79
	Confidence	Correlation	.395	1.000
		Significance (2-tailed)	.000	.
		df	79	0

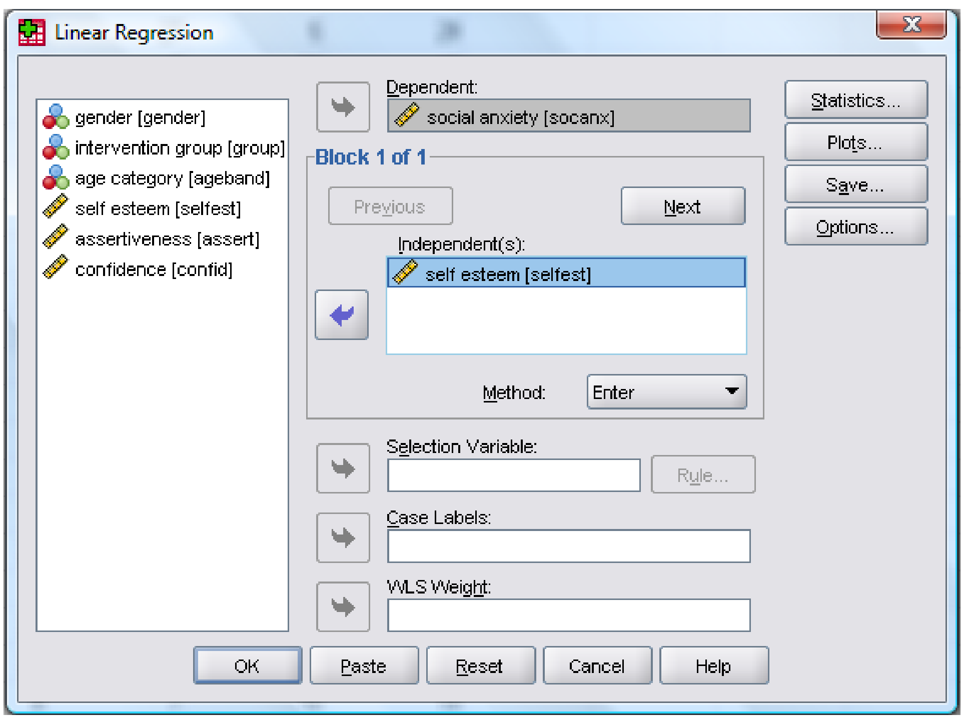
Linear Regression

Before conducting any regression you should run a correlation first to see which variables are significantly related to one another – if they are not related there is not much point in running a regression.

Additionally you should ensure that none of the predictor variables are too highly correlated with one another – this will control for multicollinearity

Linear regression

Analyze > Regression > Linear



For simple linear regression>

Place your IV and DV in their boxes

Leave method as Enter

OK

The output

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.603[a]	.364	.356	5.731

a. Predictors: (Constant), self esteem

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1500.988	1	1500.988	45.699	.000[a]
	Residual	2627.610	80	32.845		
	Total	4128.598	81			

a. Predictors: (Constant), self esteem

b. Dependent Variable: social anxiety

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	18.565	1.386		13.397	.000
	self esteem	-.548	.081	-.603	-6.760	.000

a. Dependent Variable: social anxiety

The model summary gives you the r^2 – the amount of shared variance.

The ANOVA provides you with the goodness of fit of the statistical model – i.e. if this is significant ten you have a good fit of model to the data points.

The Coefficients gives you the gradient (b) and the constant (a) and the significance of these. Essentially the t-tests assess whether your gradient is significantly different from 0.

Multiple Regression

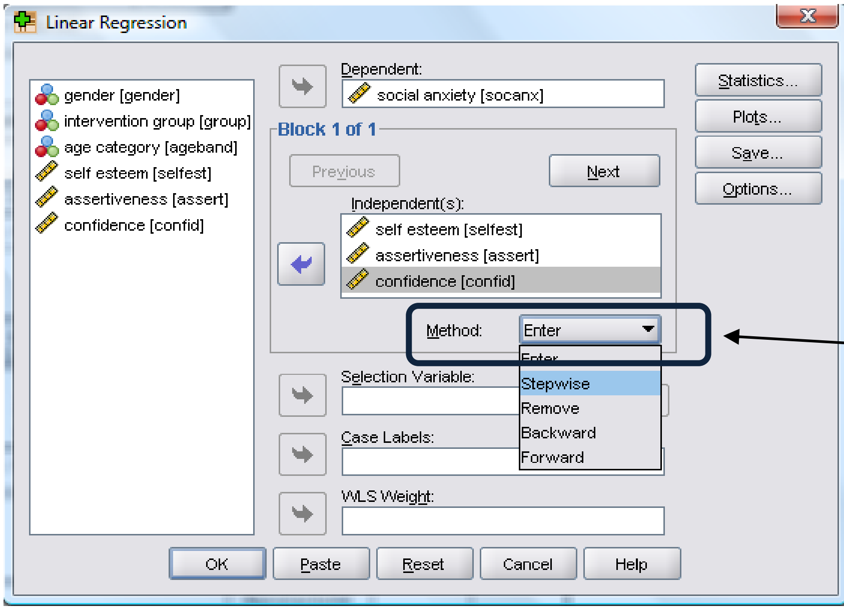
Most of the time we do not try to predict an outcome variable from one predictor variable... we often have several predictors and thus would adopt multiple regression analysis. Multiple regression shows us both the separate effects and the combined effects of these predictors on a dependent variable. The separate effect of each predictor on a dependent variable is equivalent to different simple linear regressions estimated for each predictor.

There are several different methods for running a multiple regression dependent on your particular question, hypothesis and on the basis of previous literature.

- Setwise Method: Tests only one equation including all possible predictors.
- Hierarchical Selection: ("blocked", "blockwise") Enters the predictors in the equation following some theoretical considerations.
- Stepwise Selection: ("enter" or "standard"). Calculates the equation that maximises the explained variance with minimum number of predictors.

Setwise

For the set wise model simply place all the variables of interest into the independents box and leave the Method box on its default of Enter – this will give you an overall model and R2 although you can still assess from the coefficients box which of the variables is having a greater effect and perhaps which ones that are not predicting anything at all.

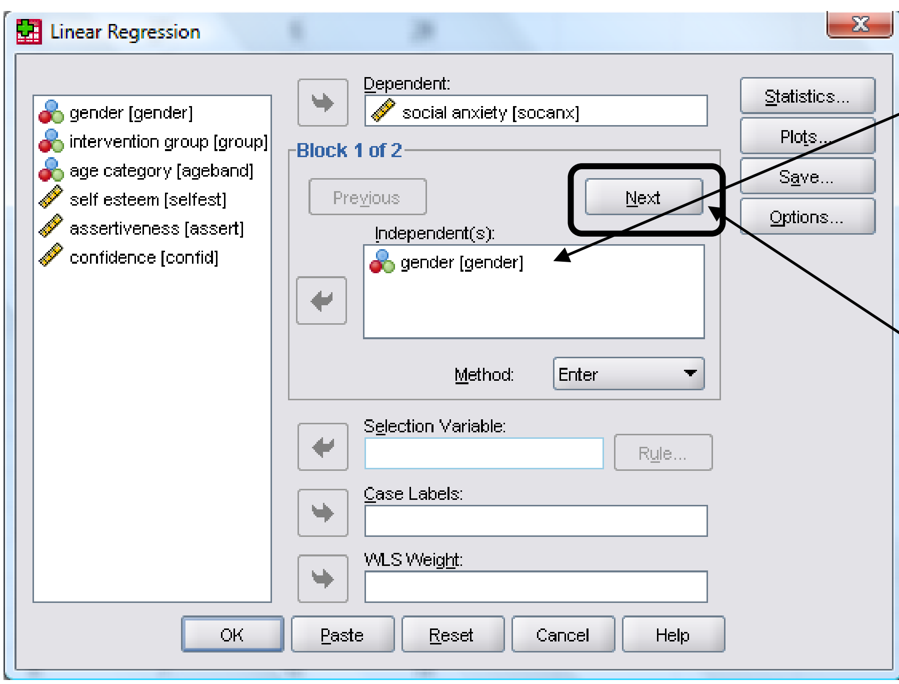


Setwise:

Leave method on its default of ENTER

Hierachical Selection

This model is based on some theoretical assumptions- therefore you as a researcher set the order in which you enter your variables in 'blocks'. For example for much of my research I would like to control for time one variables and would enter these variables first.



Hierachical:

Place your first theoretically driven variable across

Leave the method as enter (unless you would like to select stepwise for more than one variable)

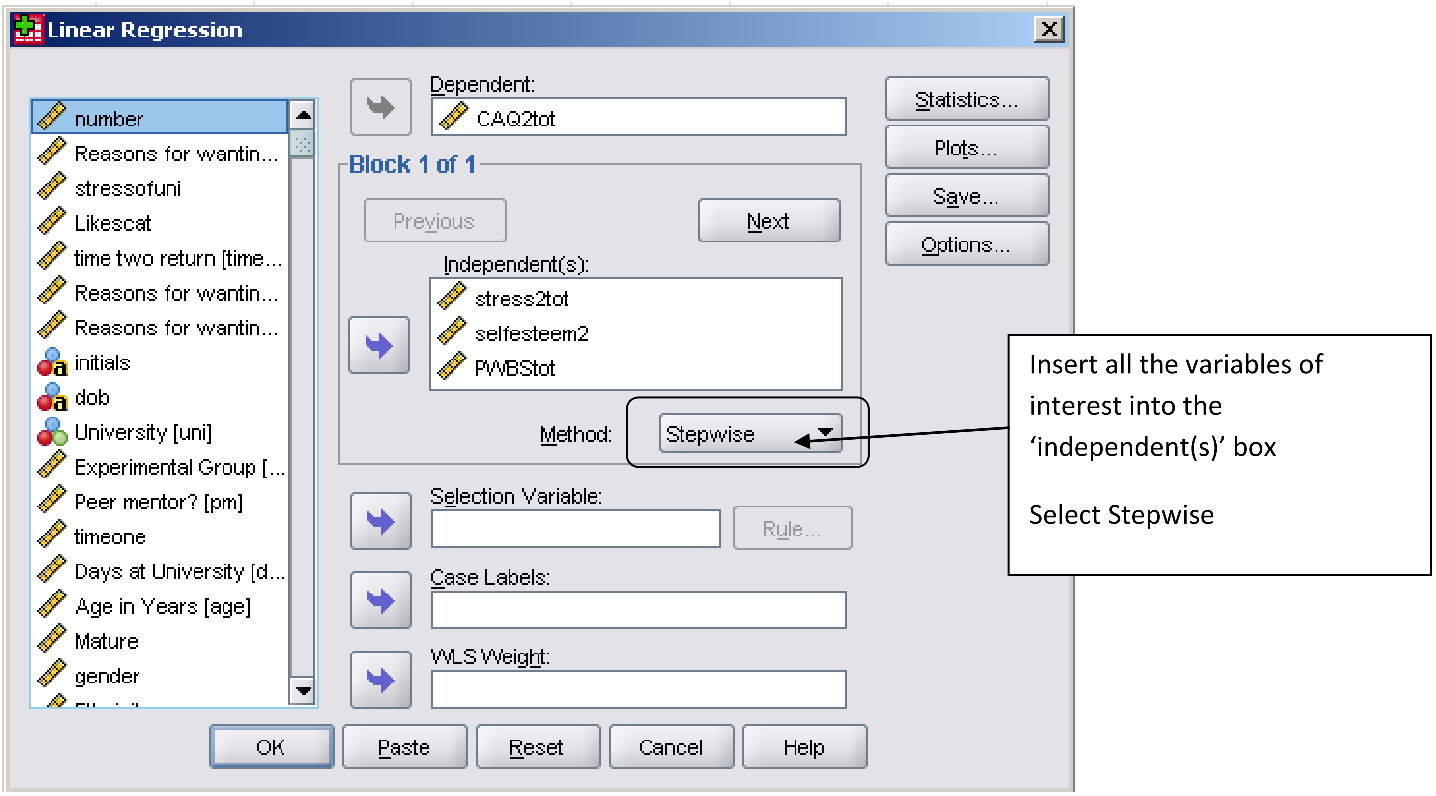
Press '**next'** which will open up a new window

Place your next variables into this window.

You can mixed Hierachical and stepwise within the same regression.

Stepwise.

This model is often used as an exploratory model i.e. when it is unknown which variable is going to be the greater predictor from the set.



Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	PWBStot	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	stress2tot	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

a. Dependent Variable: CAQ2tot

This output box informs you which variables have been entered into the equation as significant predictors.

It also tells you which one is a greater predictor.

In this case wellbeing was the greatest predictor of college adaptation with stress adding significant weight to the equation.

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.636[a]	.405	.402	14.445	.405	116.399	1	171	.000	
2	.668[b]	.447	.440	13.970	.042	12.827	1	170	.000	1.756

a. Predictors: (Constant), PWBStot
b. Predictors: (Constant), PWBStot, stress2tot
c. Dependent Variable: CAQ2tot

ANOVA[c]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24287.232	1	24287.232	116.399	.000[a]
	Residual	35679.959	171	208.655		
	Total	59967.191	172			
2	Regression	26790.485	2	13395.242	68.638	.000[b]
	Residual	33176.706	170	195.157		
	Total	59967.191	172			

a. Predictors: (Constant), PWBStot
b. Predictors: (Constant), PWBStot, stress2tot
c. Dependent Variable: CAQ2tot

These boxes are all similar to the ones you have seen before.

The model summary now has an R^2 for the first variable that enters the equation as well as an R^2 change for the second variable and an overall R^2 for the two variables together.

When writing this up you will need to provide the coefficients' of each stage.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Tolerance	VIF
1	(Constant)	27.938	5.706		4.896	.000	16.676	39.201		
	PWBStot	.929	.086	.636	10.789	.000	.759	1.099	1.000	1.000
2	(Constant)	69.408	12.827		5.411	.000	44.088	94.728		
	PWBStot	.766	.095	.525	8.076	.000	.579	.953	.770	1.298
	stress2tot	-.727	.203	-.233	-3.581	.000	-1.128	-.326	.770	1.298

a. Dependent Variable: CAQ2tot

The coefficients box also provides you with your tolerance statistics. There are several guidelines for these. If the largest VIF is greater than 10 then there is cause for concern, if the average VIF is substantially greater than 1 then the regression may be biased (Bowerman & O'Connell, 1990; Myers, 1990)
Tolerance below 0.1 indicates a serious problem, tolerance below 0.2 indicates a potential problem (Menard, 1995)

Excluded Variables[c]

Model		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics Tolerance	VIF	Minimum Tolerance
1	stress2tot	-.233[a]	-3.581	.000	-.265	.770	1.298	.770
	selfesteem2	.042[a]	.571	.569	.044	.651	1.535	.651
2	selfesteem2	-.045[b]	-.605	.546	-.046	.583	1.717	.583

a. Predictors in the Model: (Constant), PWBStot

b. Predictors in the Model: (Constant), PWBStot, stre

c. Dependent Variable: CAQ2tot

This box simply tells you the excluded variables at each step and includes tolerance tests as well. As can be seen Self Esteem has no predictive utility when looking at college adaptation

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	PWBStot	stress2tot
1	1	1.981	1.000	.01	.01	
	2	.019	10.294	.99	.99	
2	1	2.954	1.000	.00	.00	.00
	2	.042	8.410	.00	.39	.15
	3	.004	25.711	1.00	.60	.85

a. Dependent Variable: CAQ2tot

For this test of multicollinearity high variances should be proportioned across all variables for the low eigenvalues (bottom rows) in this case dimension 3 has equal proportions across PWBS and Stress indicating a possible problem

Casewise Diagnostics[a]

Case Number	Std. Residual	CAQ2tot	Predicted Value	Residual
343	-3.044	33	75.53	-42.529

a. Dependent Variable: CAQ2tot

This box tells you of any case numbers that are a significant outlier.

Case Summaries[a]

	Mahalanobis Distance	Cook's Distance	Centered Leverage Value
1	.18008	.00005	.00103
2	1.57012	.00000	.00897
3	1.45468	.01173	.00831
4	.40522	.00028	.00232
5	.33726	.00074	.00193
6	.51844	.00988	.00296
7	.08194	.00252	.00047
8	.41567	.00002	.00238
9	3.54524	.00222	.02026
10	5.24767	.01236	.02999
11	.57342	.00008	.00328
12	2.08133	.01159	.01189
13	.41567	.00107	.00238
14	.86887	.00820	.00496
15	5.86826	.04288	.03353
16	1.81653	.00125	.01038
17	1.20717	.00005	.00690
18	3.39520	.00155	.01940
19	.50321	.00083	.00288
20	1.73619	.01247	.00992
21	.90655	.00080	.00518
22	3.18390	.00238	.01819
23	.89477	.00253	.00511
24	.74128	.00569	.00424
25	1.17516	.01466	.00672

Cooks distance:None have a cooks distance greater than 1 and so none of the cases are having undue influence on the model.

The average leverage can be calculated as (k+1/n) = 4/ 359 = 0.01 and so we are looking for values either twice as large as this (0.02) or three times as large (0.03) dependent on the statistician… There are a couple of cases that are 0.03 which may be problematic.

Guidelines for Mahalanobis distance are with a sample of 100 and three predictors, values greater than 15 are problematic. We have 3 predictors and a larger sample size so the value is a conservative cut off, yet none of the cases come close to exceeding this.

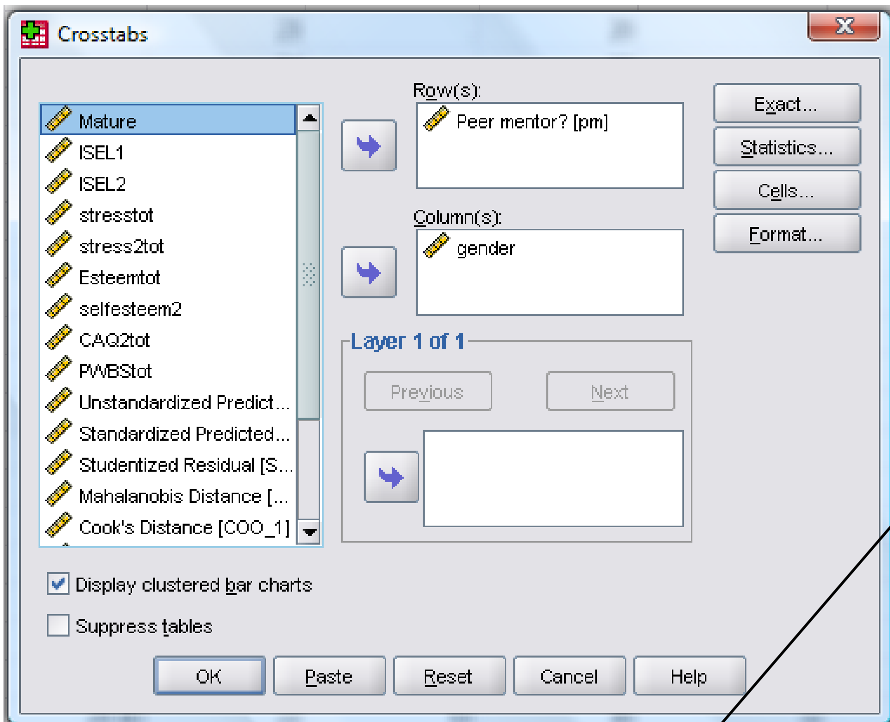
The evidence suggests that 15 may be problematic on one measure only. Regarding the rest of the data there is appears to be no influencing cases.

Chi Square Test of Independence

This will test the association between two categorical variables.

Analyze > Descriptive Statistics > Crosstabs

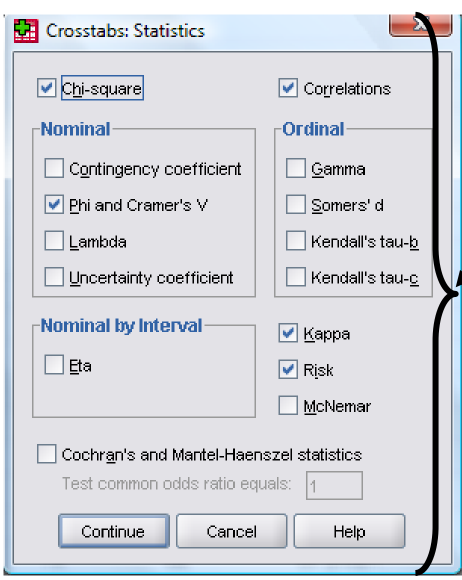
Place across the variables you are interested in.

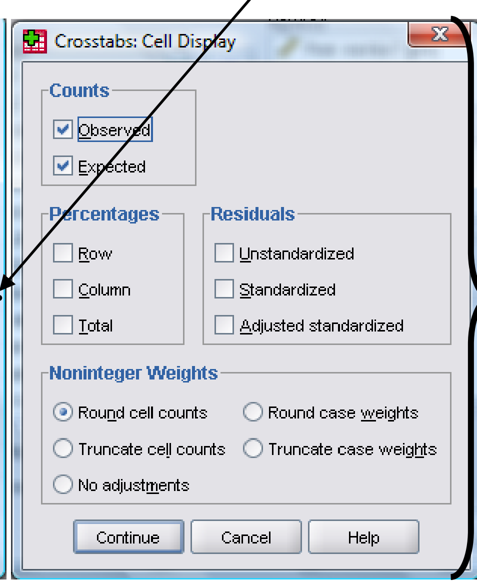


Once you have placed the variables of interest across you can select Display Clustered Bar Charts (you can also do this via Chart Builder)

You need to select statistics and tell SPSS you would like it to calculate the chi-square. Additionally I would select Phi and Cramer's V (this is your effect size).

Additionally if you have a 2 x 2 chi model (and it is appropriate) you can ask for Risk – this is an odds ratio.





Additionally I would select Cells and tick expected – this allows you to compare what you have and what should be expected so you can see where your data deviates.

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
Peer mentor? * gender	319	88.9%	40	11.1%	359	100.0%

Peer mentor? * gender Crosstabulation

			gender		Total
			Male	Female	
Peer mentor?	Has PM	Count	13	93	106
		Expected Count	19.3	86.7	106.0
	Has no PM	Count	45	168	213
		Expected Count	38.7	174.3	213.0
Total		Count	58	261	319
		Expected Count	58.0	261.0	319.0

This table gives you what scores are expected if the two variables are truly independent and your observed values. As can be seen within the table males are slightly less likely than expected to have a peer mentor in comparison to females

Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)	Exact Sig. (2-sided)	Exact Sig. (1-sided)
Pearson Chi-Square	3.737[a]	1	.053		
Continuity Correction[b]	3.165	1	.075		
Likelihood Ratio	3.945	1	.047		
Fisher's Exact Test				.064	.035
Linear-by-Linear Association	3.725	1	.054		
N of Valid Cases	319				

a. 0 cells (.0%) have expected count less than 5. The minimum expected count is 19.27.

b. Computed only for a 2x2 table

The main analysis box shows you that there is only an approaching significance: χ^2 (1) 3.737, p = 0.053.

You must always look to the bottom of the table – if this is ≥ 16% then you have violated the assumption of normality for chi square and you should be reporting the Fishers Exact Ratio instead.

Symmetric Measures

		Value	Asymp. Std. Error[a]	Approx. T[b]	Approx. Sig.
Nominal by Nominal	Phi	-.108			.053
	Cramer's V	.108			.053
Interval by Interval	Pearson's R	-.108	.051	-1.938	.053[c]
Ordinal by Ordinal	Spearman Correlation	-.108	.051	-1.938	.053[c]
Measure of Agreement	Kappa	-.100	.047	-1.933	.053
N of Valid Cases		319			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

c. Based on normal approximation.

Risk Estimate

	Value	95% Confidence Interval	
		Lower	Upper
Odds Ratio for Peer mentor? (Has PM / Has no PM)	.522	.268	1.017
For cohort gender = Male	.581	.328	1.028
For cohort gender = Female	1.112	1.007	1.229
N of Valid Cases	319		

This box provides your odds ratio and the 95%CI of that ratio.

In this case females are 1.112 times more likely to have a mentor than their male counterparts.