Regression Analysis

Nikos Comoutos PhD

- Regression analysis is useful when you want to predict the scores of one dependent variable from the scores of one independent variable (Simple Regression)

- Or more independent variables (Multiple Regression).

Note!!

- Note that a significant prediction does not prove that the predictor (independent) variables have a causal effect on the predicted (dependent) variable.
- A significant prediction merely indicates that changes in the scores of the dependent variables can by predicted by the independent variables (remember correlation…)

Multiple regression is not just one technique but a family of techniques (e.g., hierarchical, stepwise) that can be used to explore the relationship between one continuous dependent variable and a number of independent variables or predictors - usually continuous

Multiple regression is based on correlation but allows a more sophisticated exploration of the interrelationship among a set of variables

You should have a sound theoretical or conceptual reason for the analysis and, in particular; the order of variables entering the equation.

- It can tell you how well a set of variables is able to predict a particular outcome

For example, you may be interested in exploring how well a set of subscales on achievement goals (e.g., mastery approach, performance approach, etc..) is able to predict positive self-talk.

Multiple regression will provide you with information about the model as a whole (all subscales) and the relative contribution of each of the variables that make up the model (subscales).

As an extension of this, multiple regression will allow you to test whether adding a variable (e.g., anxiety) contributes to the predictive ability of the model, over and above those variables already included in the model

Types of multiple regression

- There are a number of different types of multiple regression analyses that you can use, depending on the nature of the question you wish to address. The three main types of multiple regression analyses are:
- standard or simultaneous;
- hierarchical or sequential
- stepwise

Standard multiple regression

- In standard multiple regression, all the independent (or predictor) variables are entered into the equation simultaneously. Each independent variable is evaluated in terms of its predictive power, over and above that offered by all the other independent variables.
- This is the most commonly used multiple regression analysis.

When to use it

- You would use this approach if you had a set of variables (e.g. various motivational subscales) and wanted to know how much variance in a dependent variable (e.g. negative self-talk) they were able to explain as a group or block.
- This approach would also tell you how much unique variance in the dependent variable each of the independent variables explained.

Hierarchical multiple regression

- **In** hierarchical regression (also called sequential regression), the independent variables are entered into the equation in the order specified by the researcher based on theoretical grounds. Variables or sets of variables are entered in steps (or blocks), with each independent variable being assessed in terms of what it adds to the prediction of the dependent variable, after the previous variables have been controlled for.

When to use it

- For example, if you wanted to know how well extraversion and goal setting predict distractability, after the effect of extraversion is controlled for.
- You would enter extraversion in Block 1 and then goal-setting in Block 2. Once all sets of variables are entered, the overall model is assessed in terms of its ability to predict the dependent measure (distractability). The relative contribution of each block of variables is also assessed.

Stepwise multiple regression

- In stepwise regression, the researcher provides SPSS with a list of independent variables and then allows the program to select which variables it will enter and in which order they go into the equation, based on a set of statistical criteria.
- There are three different versions of this approach:
- forward selection
- backward deletion
- and stepwise regression.

ASSUMPTIONS OF MULTIPLE REGRESSION

- It makes a number of assumptions about the data, and it is not all that forgiving if they are violated.
- Sample size
- Multicollinearity and singularity
- Outliers
- Normality, linearity, homoscedasticity, independence of residuals

Sample size

- The ratio of participants to independent variables should be at least 5:1 and ideally 20:1.
- Stevens (1996, p. 72) recommends that 'for social science research, about 15 subjects per predictor are needed for a reliable equation'.
- If the *stepwise* method is used (see below), the ratio should be 40:1. This is due to the possibility that with small sample sizes this method can produce results which do not generalize to other samples.
- Make sure you have enough cases (participants) in the data file, as this analysis deletes all cases with missing values. If there are not enough cases, you may need to replace the missing values with the variable mean

Multicollinearity and singularity

- This refers to the relationship among the independent variables. Multicollinearity exists when the independent variables are highly correlated (r=.9 and above).
- Singularity occurs when one independent variable is actually a combination of other independent variables (e.g. when both subscale scores and the total score of a scale are included).
- Multiple regression doesn't like multicollinearity or singularity, and these certainly don't contribute to a good regression model, so always check for these problems before you start.
- To test for multicollinearity or singularity, use the *Collinearity diagnostics* in *Statistics.*

Outliers

- Multiple regression is very sensitive to outliers (very high or very low scores).
- Checking for extreme scores should be part of the initial data screening process
- You should do this for all the variables, both dependent and independent, that you will be using in your regression analysis.
- Outliers can either be deleted from the data set or, alternatively, given a score for that variable that is high but not too different from the remaining cluster of scores.
- You can also use the *Casewise diagnostics* in *Statistics* (see below). To identify outliers in the values of the dependent variable create a scatterplot of its *standardised residuals* (use *Save* below to save these residuals in the data file). To detect multivariate outliers among the independent variables, that is, cases with extreme values on a combination of variables, use the *Mahalanobis distance* or the *Leverage value* (see *Save* below).

How to test this in SPSS

- *Mahalanobis distance* is a measure of how much the value of a case differs in the independent variables from the average of all other cases.
- Large *Mahalanobis distances* signify potential outlier cases.
- Tabachnick and Fidell (2007, p. 128) define outliers as those with standardised residual values above about 3.3 (or less than -3.3).

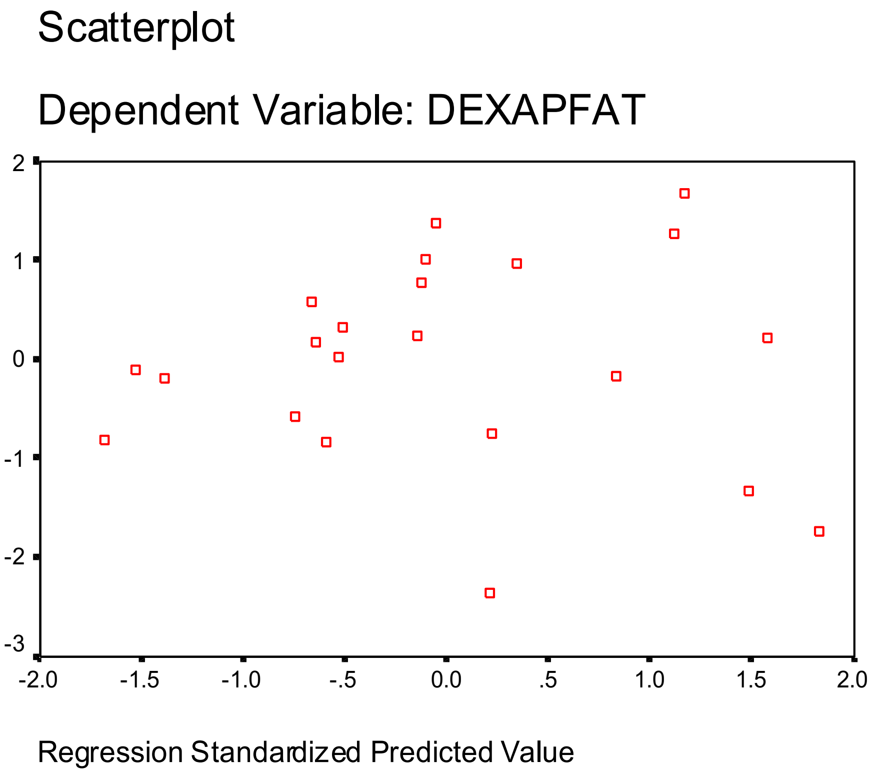
Normality, linearity, homoscedasticity, independence of residuals

- These all refer to various aspects of the distribution of scores and the nature of the underlying relationship between the variables. These assumptions can be checked from the *residuals* scatterplots which are generated as part of the multiple regression procedure. Residuals are the differences between the obtained and the predicted dependent variable (DV) scores. The residuals scatterplots allow you to check:

Normality, linearity, homoscedasticity, independence of residuals

- *normality:* the residuals should be normally distributed about the predicted DV scores;
- *linearity:* the residuals should have a straight-line relationship with predicted DV scores;
- *homoscedasticity:* the variance of the residuals about predicted DV scores should be the same for all predicted scores.

Checking the assumptions of regression analyses

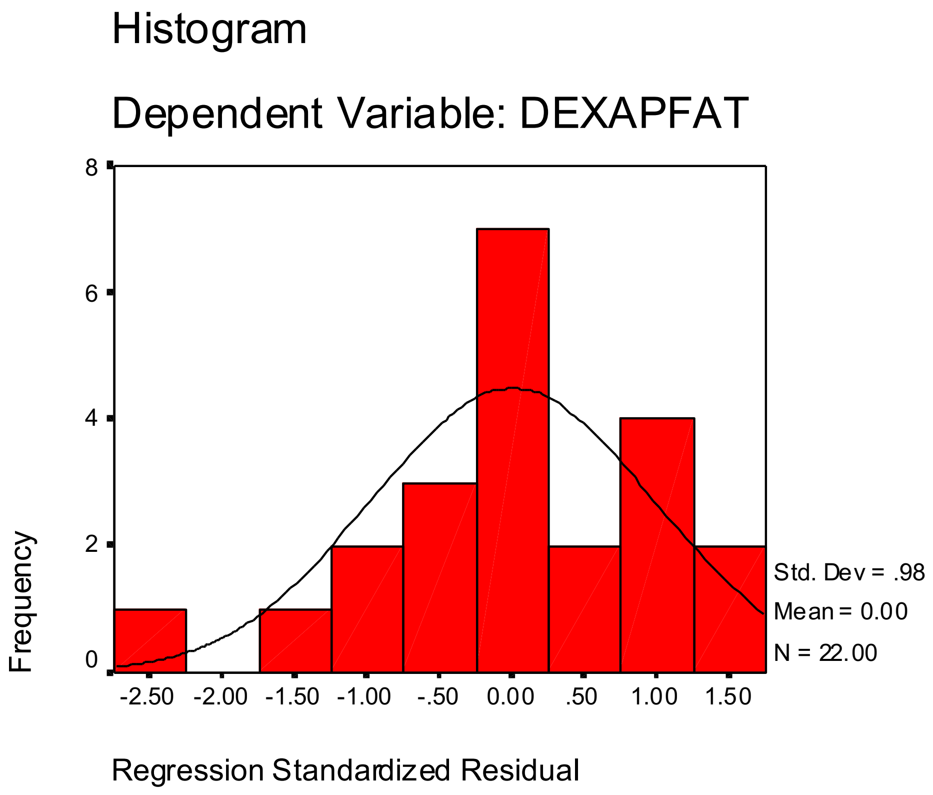


To check for assumptions of homoscedasticity, plot studentized residuals (SRESID) against standardised predicted values of the dependent variable (ZPRED)

If the linearity assumption is met, such plots should not show any pattern.

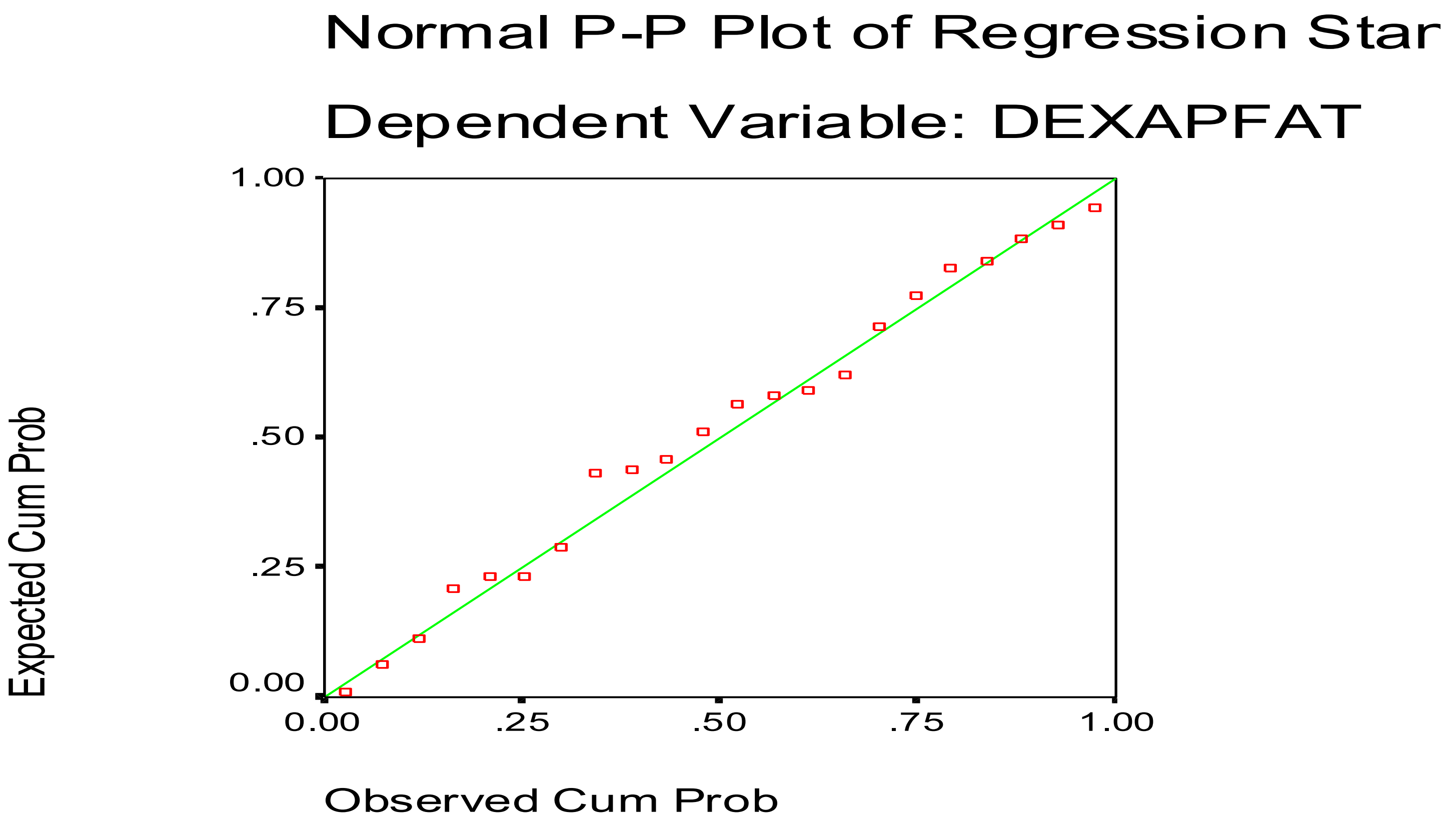
Residuals can be saved in the '*save*' function to be plotted in *simple scatter plot Graphs* menu

Checking the assumptions of regression analyses

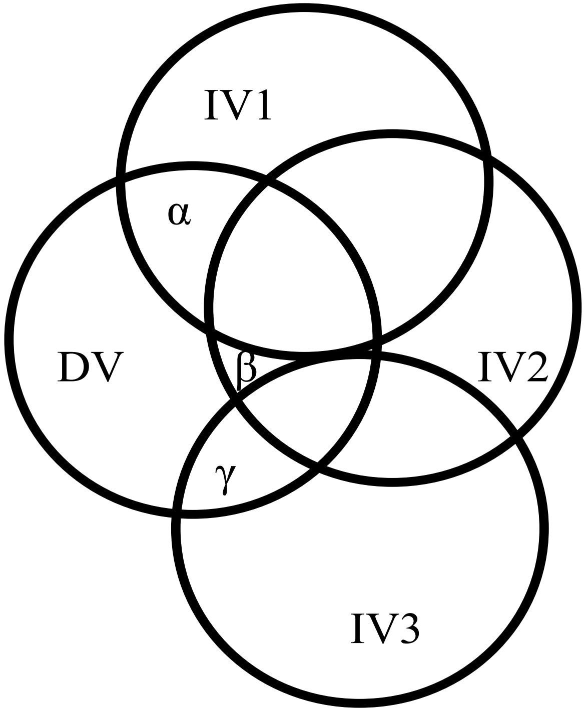


The histogram shows that the residuals are normally distributed.

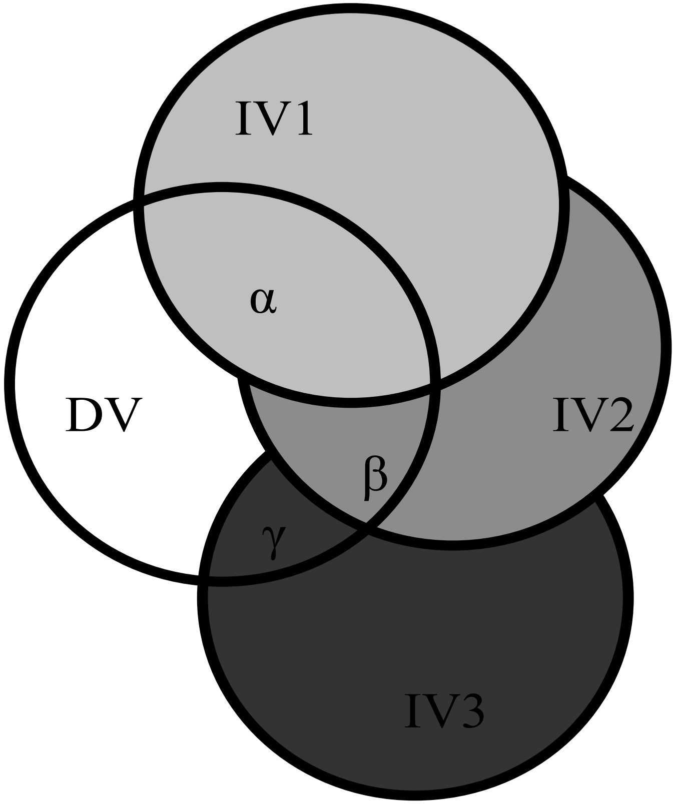
Checking the assumptions of regression analyses



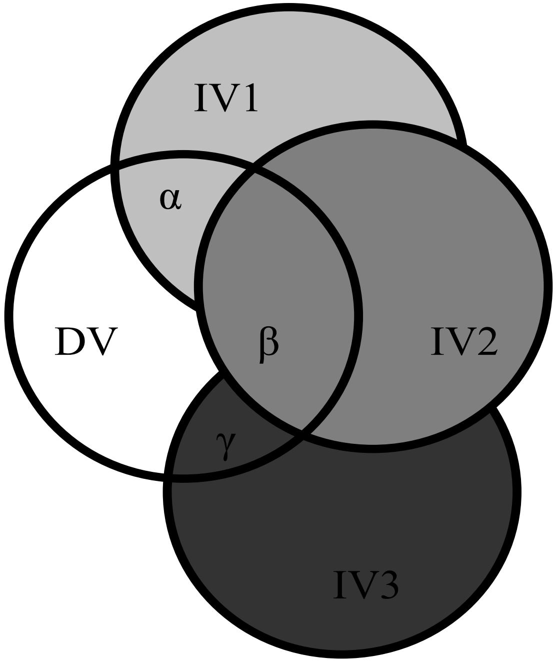
The Normal Probability Plot shows the distribution of the standardised residuals against a normal distribution. Graph shows that distribution is more or less normal as points cluster around straight line.



Standard regression



Hierarchical regression



Stepwise regression

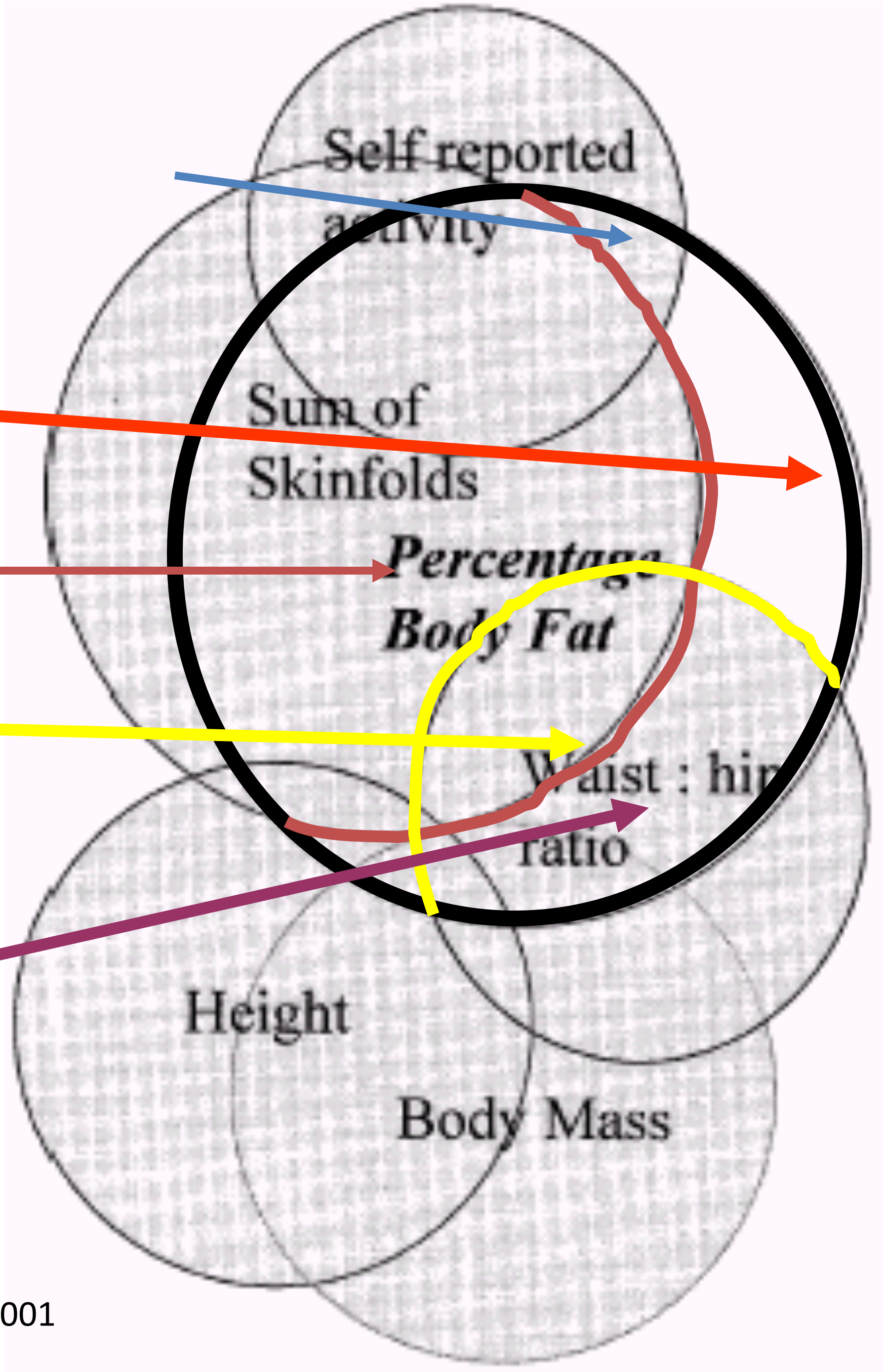
Self-reported activity accounts for only a very small proportion of unique variance

Unexplained variance

Largest %variance accounted for by skinfolds

Next largest %variance accounted for by waist:hip ratio

Waist:hip ratio accounts for some unique' variance



The concept of multiple regression

Winter Eston Lamb J Sports Sci 2001

Acrobat Document

Selection of variables in hierarchical regression

- This means that we will be entering our variables in steps or blocks in a predetermined order (not letting the computer decide, as would be the case for stepwise regression). In the first block, we will 'force' the variable(s) that we want to control. This has the effect of statistically controlling for these variables. The difference this time is that the possible effect of this variable (s) has been 'removed' and we can then see whether our block of independent variables are still able to explain some of the remaining variance in our dependent variable.

Selection of variables for stepwise Analysis

Make available for selection at each analysis

a) all skinfolds, weight and height,

   To do: Draw a Venn diagram to show proportion of variance explained by pectoral and thigh (clue use R^2 and correlation matrix to help you draw it)

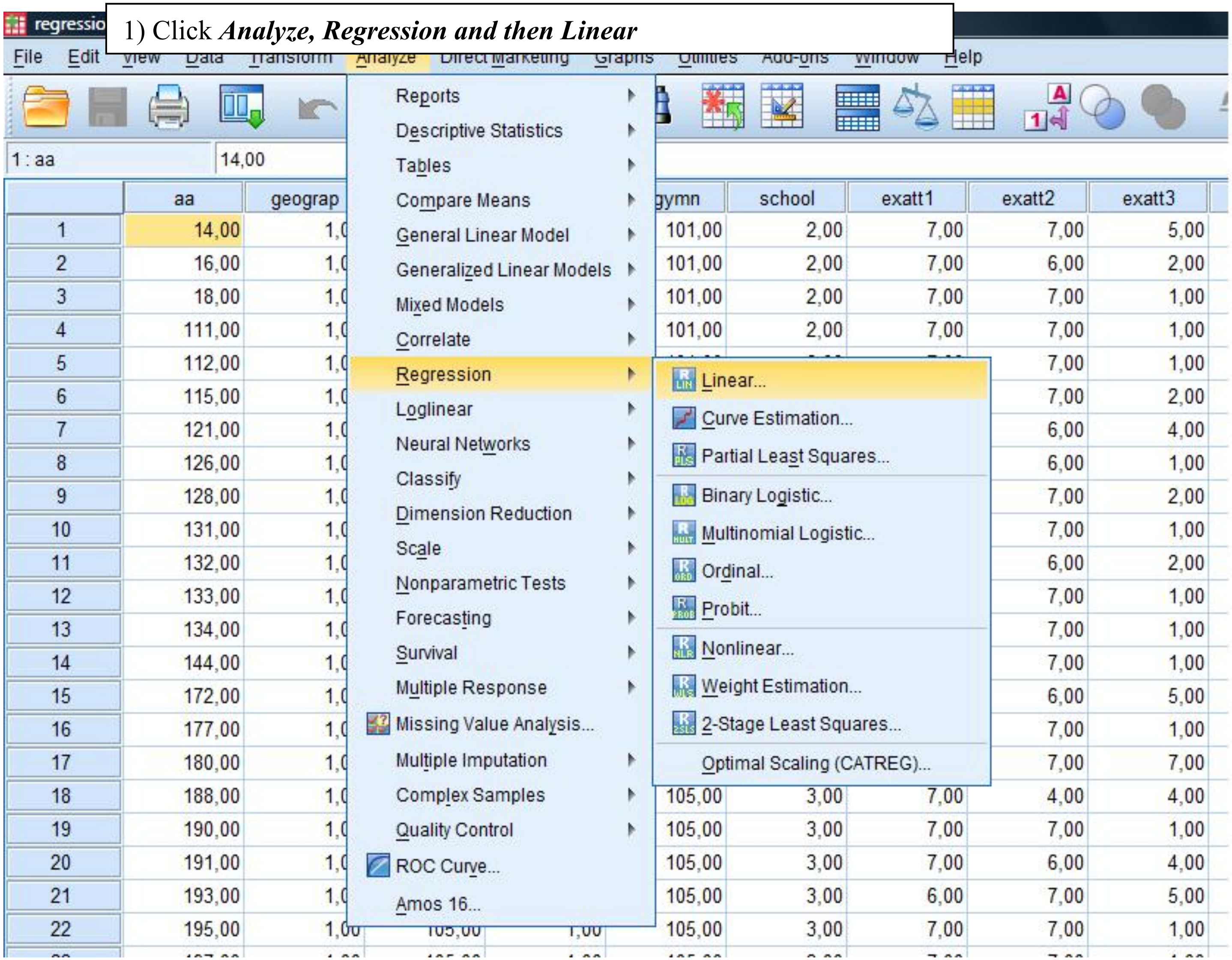
b) You may argue that from a theoretical perspective, that the BMI would account for some additional variance in %fat. You can force this into the analysis AFTER entering the best predictors.

Example of standard multiple regression

- *Question* 1: How well do the two measures of exercise behaviour (intention to exercise-exint, behaviour of exercise-bexbeh) **predict frequency of exercise? How much variance in frequency of exercise can be** explained by scores on these two scales?
- *Question* 2: Which is the best predictor of frequency of exercise: exint or bexbeh?

- This involves all of the independent variables being entered into the equation at once.
- The results will indicate how well this set of variables is able to predict frequency of exercise; and it will also tell us how much *unique* variance each of the independent variables explains in the dependent variable, *over and above* the other independent variables included in the set. For each of the procedures,

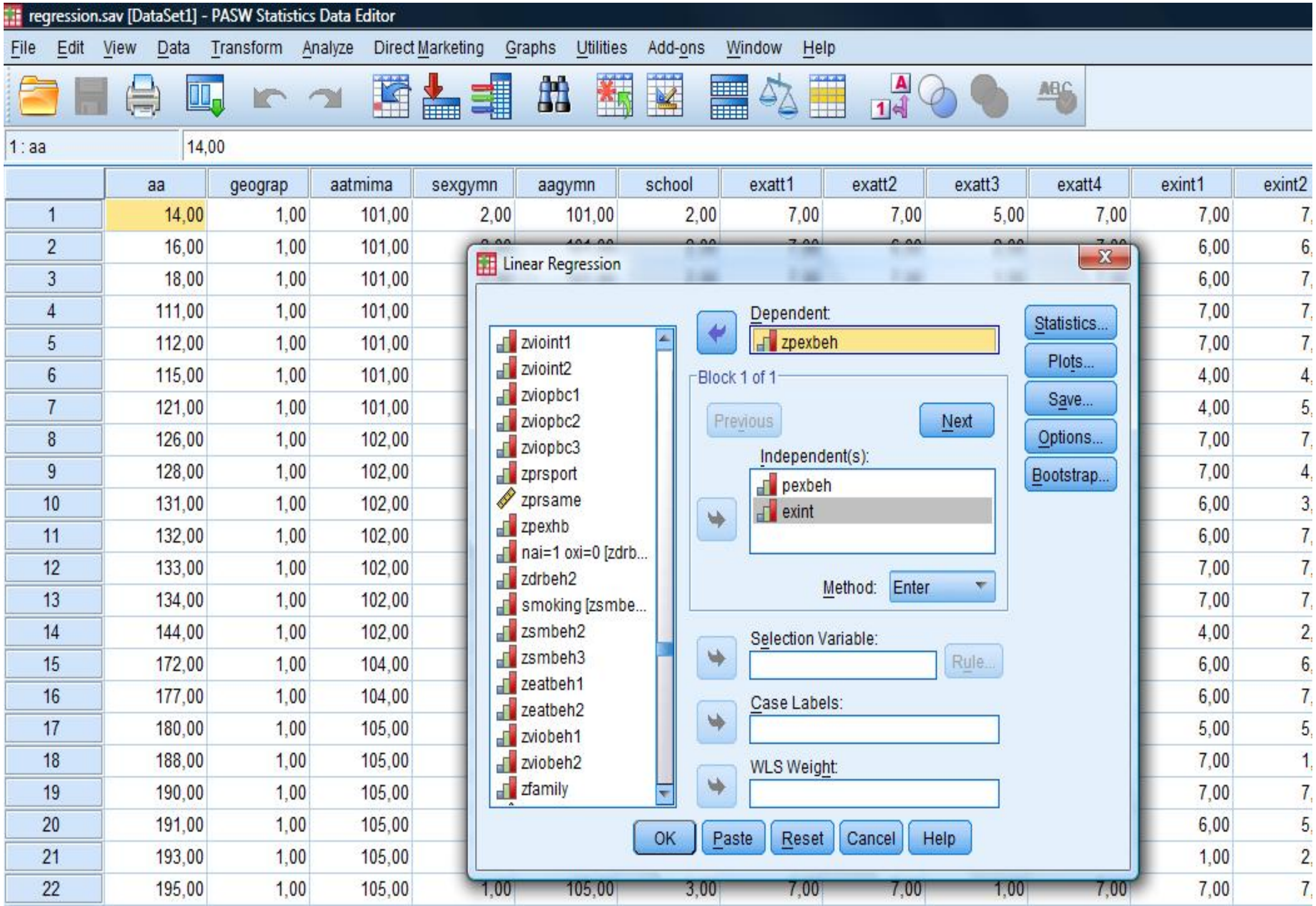
1) Click ***Analyze, Regression and then Linear***



	aa	geograp	gymn	school	exatt1	exatt2	exatt3
1	14,00	1,0	101,00	2,00	7,00	7,00	5,00
2	16,00	1,0	101,00	2,00	7,00	6,00	2,00
3	18,00	1,0	101,00	2,00	7,00	7,00	1,00
4	111,00	1,0	101,00	2,00	7,00	7,00	1,00
5	112,00	1,0				7,00	1,00
6	115,00	1,0				7,00	2,00
7	121,00	1,0				6,00	4,00
8	126,00	1,0				6,00	1,00
9	128,00	1,0				7,00	2,00
10	131,00	1,0				7,00	1,00
11	132,00	1,0				6,00	2,00
12	133,00	1,0				7,00	1,00
13	134,00	1,0				7,00	1,00
14	144,00	1,0				7,00	1,00
15	172,00	1,0				6,00	5,00
16	177,00	1,0				7,00	1,00
17	180,00	1,0				7,00	7,00
18	188,00	1,0	105,00	3,00	7,00	4,00	4,00
19	190,00	1,0	105,00	3,00	7,00	7,00	1,00
20	191,00	1,0	105,00	3,00	7,00	6,00	4,00
21	193,00	1,0	105,00	3,00	6,00	7,00	5,00
22	195,00	1,00	105,00	3,00	7,00	7,00	1,00

Make sure that **enter** is selected this will give you standard multiple regression

2) Dependent variable (zpexbeh) in the *Dependent* box then the independent or predictors (pexbeh, exint) in the *Independent (s):* box and then click ***Statistics***



1 : aa 14,00

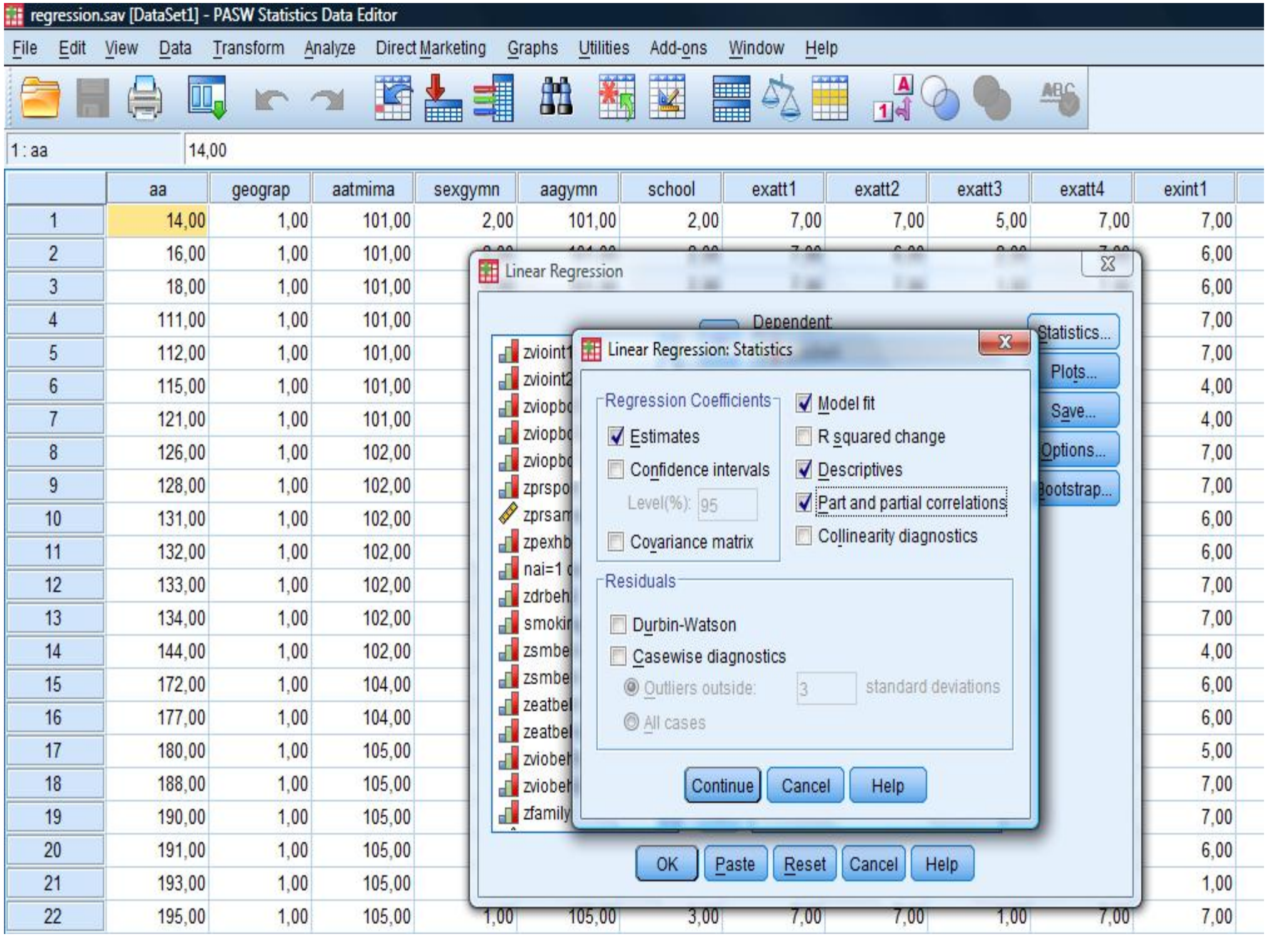
	aa	geograp	aatmima	sexgymn	aagymn	school	exatt1	exatt2	exatt3	exatt4	exint1	exint2
1	14,00	1,00	101,00	2,00	101,00	2,00	7,00	7,00	5,00	7,00	7,00	7,
2	16,00	1,00	101,00								6,00	6,
3	18,00	1,00	101,00								6,00	7,
4	111,00	1,00	101,00								7,00	7,
5	112,00	1,00	101,00								7,00	7,
6	115,00	1,00	101,00								4,00	4,
7	121,00	1,00	101,00								4,00	5,
8	126,00	1,00	102,00								7,00	7,
9	128,00	1,00	102,00								7,00	4,
10	131,00	1,00	102,00								6,00	3,
11	132,00	1,00	102,00								6,00	7,
12	133,00	1,00	102,00								7,00	7,
13	134,00	1,00	102,00								7,00	7,
14	144,00	1,00	102,00								4,00	2,
15	172,00	1,00	104,00								6,00	6,
16	177,00	1,00	104,00								6,00	7,
17	180,00	1,00	105,00								5,00	5,
18	188,00	1,00	105,00								7,00	1,
19	190,00	1,00	105,00								7,00	7,
20	191,00	1,00	105,00								6,00	5,
21	193,00	1,00	105,00								1,00	2,
22	195,00	1,00	105,00	1,00	105,00	3,00	7,00	7,00	1,00	7,00	7,00	7,

Dependent: zpexbeh

Independent(s): pexbeh, exint

Method: Enter

3) Click ***Descriptives*** and ***Part and partial correlations*** and then ***Continue***



	aa	geograp	aatmima	sexgymn	aagymn	school	exatt1	exatt2	exatt3	exatt4	exint1
1	14,00	1,00	101,00	2,00	101,00	2,00	7,00	7,00	5,00	7,00	7,00
2	16,00	1,00	101,00								6,00
3	18,00	1,00	101,00								6,00
4	111,00	1,00	101,00								7,00
5	112,00	1,00	101,00								7,00
6	115,00	1,00	101,00								4,00
7	121,00	1,00	101,00								4,00
8	126,00	1,00	102,00								7,00
9	128,00	1,00	102,00								7,00
10	131,00	1,00	102,00								6,00
11	132,00	1,00	102,00								6,00
12	133,00	1,00	102,00								7,00
13	134,00	1,00	102,00								7,00
14	144,00	1,00	102,00								4,00
15	172,00	1,00	104,00								6,00
16	177,00	1,00	104,00								6,00
17	180,00	1,00	105,00								5,00
18	188,00	1,00	105,00								7,00
19	190,00	1,00	105,00								7,00
20	191,00	1,00	105,00								6,00
21	193,00	1,00	105,00								1,00
22	195,00	1,00	105,00	1,00	105,00	3,00	7,00	7,00	1,00	7,00	7,00

Linear Regression: Statistics

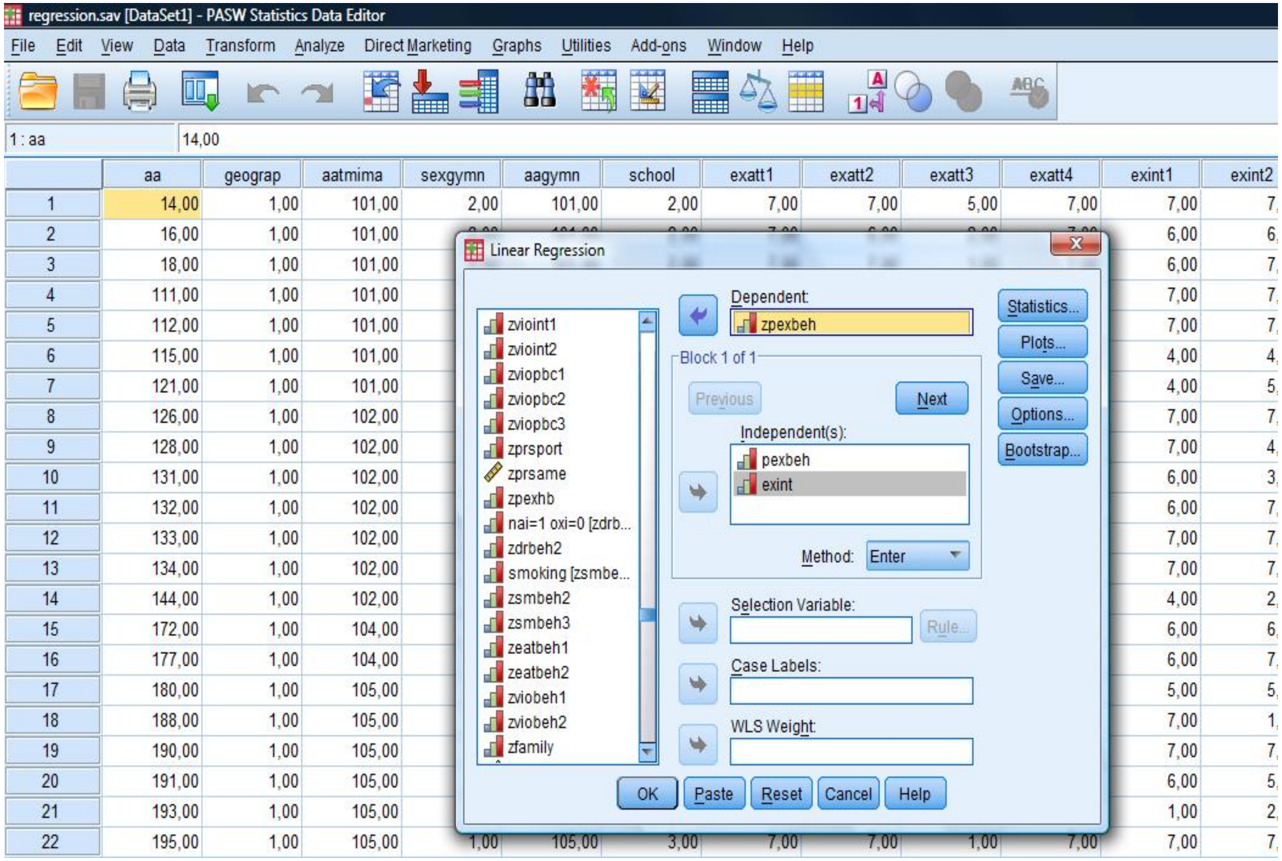
Regression Coefficients
- ☑ Estimates
- □ Confidence intervals
  - Level(%): 95
- □ Covariance matrix

- ☑ Model fit
- □ R squared change
- ☑ Descriptives
- ☑ Part and partial correlations
- □ Collinearity diagnostics

Residuals
- □ Durbin-Watson
- □ Casewise diagnostics
  - ● Outliers outside: 3 standard deviations
  - ○ All cases

Continue | Cancel | Help

4) Click ***Ok***



	aa	geograp	aatmima	sexgymn	aagymn	school	exatt1	exatt2	exatt3	exatt4	exint1	exint2
1	14,00	1,00	101,00	2,00	101,00	2,00	7,00	7,00	5,00	7,00	7,00	7,
2	16,00	1,00	101,00								6,00	6,
3	18,00	1,00	101,00								6,00	7,
4	111,00	1,00	101,00								7,00	7,
5	112,00	1,00	101,00								7,00	7,
6	115,00	1,00	101,00								4,00	4,
7	121,00	1,00	101,00								4,00	5,
8	126,00	1,00	102,00								7,00	7,
9	128,00	1,00	102,00								7,00	4,
10	131,00	1,00	102,00								6,00	3,
11	132,00	1,00	102,00								6,00	7,
12	133,00	1,00	102,00								7,00	7,
13	134,00	1,00	102,00								7,00	7,
14	144,00	1,00	102,00								4,00	2,
15	172,00	1,00	104,00								6,00	6,
16	177,00	1,00	104,00								6,00	7,
17	180,00	1,00	105,00								5,00	5,
18	188,00	1,00	105,00								7,00	1,
19	190,00	1,00	105,00								7,00	7,
20	191,00	1,00	105,00								6,00	5,
21	193,00	1,00	105,00								1,00	2,
22	195,00	1,00	105,00	1,00	105,00	3,00	7,00	7,00	1,00	7,00	7,00	7,

zvioint1
zvioint2
zviopbc1
zviopbc2
zviopbc3
zprsport
zprsame
zpexhb
nai=1 oxi=0 [zdrb...
zdrbeh2
smoking [zsmbe...
zsmbeh2
zsmbeh3
zeatbeh1
zeatbeh2
zviobeh1
zviobeh2
zfamily

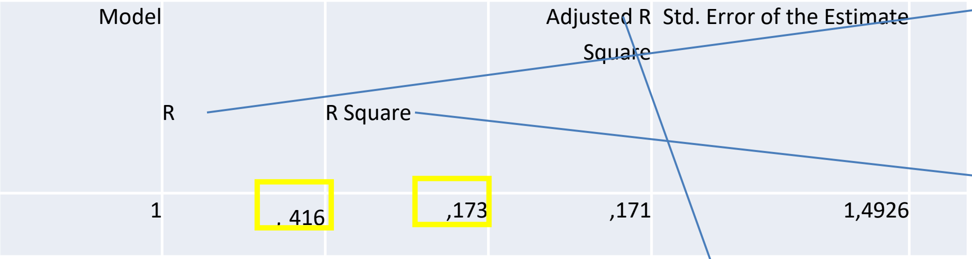
Dependent: zpexbeh

Independent(s):
pexbeh
exint

Method: Enter

Multiple R is .42.

Overall 17% of the variance of frequency of exercise (ZPEXBEH) explained by the two independent variables (EXINT, PEXBEH)



Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	. 416	,173	,171	1,4926

R represents the total correlation between the 2 independent variables and the dependent variable.

R squared is R which has been squared. The square of a correlation is the same as a proportion of variance – it represents the total amount of variance accounted for in the dependent variable by the independent variables

A reduced value of R squared which attempts to make an estimate of the value of R squared in the population rather than the sample

The significance of the value F (called sig. in the table) is the probability associated with R squared. This probability can be thought of as a significance value for the whole model or equivalently, a significance value of R squared

Statistical significant, F (2, 829)= 86.74, p < .001.

ANOVA[b]

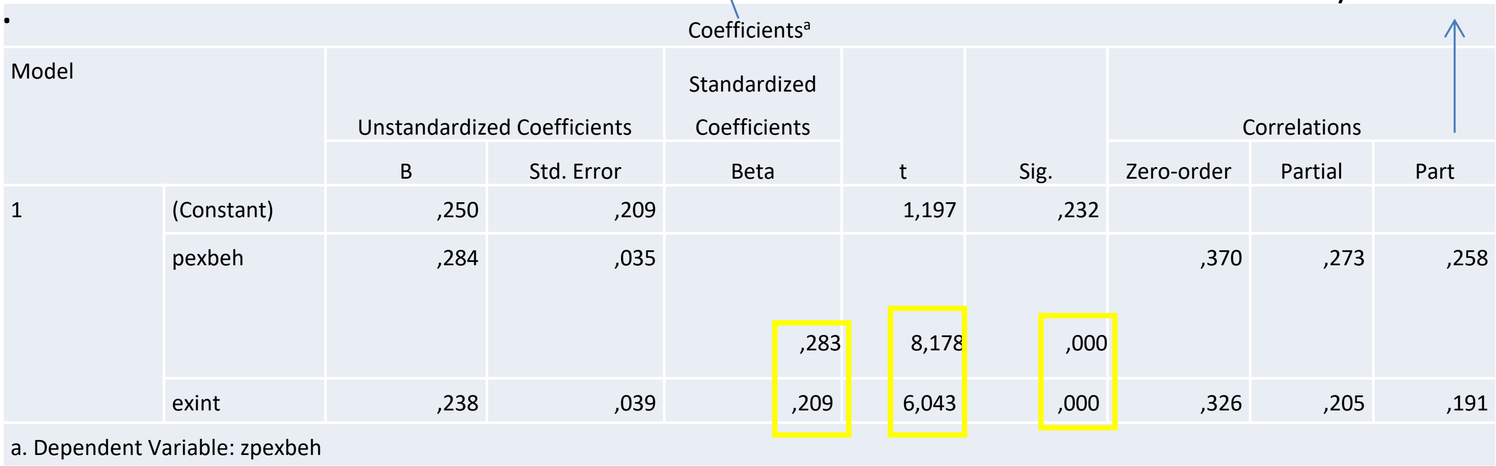
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	386,451	2	193,225	86,737	,000[a]
	Residual	1846,780	829	2,228		
	Total	2233,231	831			

a. Predictors: (Constant), exint, pexbeh

b. Dependent Variable: zpexbeh

The final part of the printout is the coefficients. To find out how well each of the variables contributes to the final equation, we need to look in the Coefficients. This summarizes the results, with *all* the variables entered into the equation. Scanning the Sig. column, both the two variables make a statistically significant contribution (less than .05). In order of importance, they are: pexbeh (beta = .28) and exint (beta = .21). Remember, these beta values represent the unique contribution of each variable, when the overlapping effects of all other variables are statistically removed. In different equations, with a different set of independent variables, or with a different sample, these values would change. Standardised regression coefficients range from -1 to 1. The higher the standardised regression coefficient (in absolute terms), the better the prediction of the dependent variable.

If you square this value, you get an indication of the contribution of that variable to the total R square. In other words, it tells you how much of the total variance in the dependent variable is uniquely explained by that variable and how much R square would drop if it wasn't included in your model.



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations		
		B	Std. Error	Beta			Zero-order	Partial	Part
1	(Constant)	,250	,209		1,197	,232			
	pexbeh	,284	,035	,283	8,178	,000	,370	,273	,258
	exint	,238	,039	,209	6,043	,000	,326	,205	,191

a. Dependent Variable: zpexbeh

PRESENTING THE RESULTS FROM MULTIPLE REGRESSION

- As a minimum, you should indicate what type of analysis was performed (standard or hierarchical), standardised (beta) values if the study was theoretical, or unstandardised (B) coefficients (with their standard errors) if the study was applied (usually we use the standardised). If you performed a hierarchical multiple regression, you should also provide the R square change values for each step and associated probability values.

Example

- Multiple regression was used to predict frequency of exercise from intention to exercise and exercise behaviour. Preliminary analyses were conducted to ensure no violation of the assumptions of normality, linearity, multicollinearity and homoscedasticity. The total variance explained by the model as a whole was 17%, *F* (2, 829) = 86.74, *P* < .001. In the final model both measures were statistically significant, with the behaviour of exercise recording a higher beta value (t = 8.18, beta = .28, *p*< .001) than the intention of exercise (t = 6.04, beta = .21, *p< .001).*