ACU
AUSTRALIAN CATHOLIC UNIVERSITY

PUBH620 Biostatistics

Week 6

Multiple Linear Regression

Lecture notes by Dr Brandon Cheong and Dr Mike Steele

ELECTRONIC WARNING NOTICE

Commonwealth of Australia
Copyright Act 1968
Form of notice for paragraph 135KA (a) of the *Copyright Act 1968*
Warning
This material has been copied and communicated to you by or on behalf of *Australian Catholic University under Part VA of the Copyright Act 1968 (the **Act).***
The material in this communication may be subject to copyright under the Act. Any further copying or communication of this material by you may be the subject of copyright or performers' protection under the Act.
Do not remove this notice.

In this lecture…

(clicking the links below will direct you to the topic page)

6.1 Multiple Linear Regression

- 6.1.1 Introduction to multiple linear regression
- 6.1.2 Assumptions of multiple linear regression
- 6.1.3 Multiple linear regression in SPSS
- 6.1.4 Interpreting the results of a multiple linear regression model
- 6.1.5 Using a backward elimination method

Topic Learning Objectives (TLOs)

- Understand the basis for multiple regression techniques, their interpretation and limitations

6.1 Multiple Linear Regression

Multiple Regression is an extension of Simple Linear Regression. In Multiple Regression, we assess whether there is a causal relationship between one dependent variable (DV) and two or more independent variables (IV).

In order to fully understand how Multiple Regression works, it is essential that you have a clear understanding of Simple Linear Regression and Correlation.

In the same way simple linear regression requires that we find a regression equation, multiple regression is the same but only this time we are expecting to have more coefficients in the equation.

The simple form of the regression equation which is sometimes reported in applied public health research is:

$$y = b_0 + b_1x_1 + b_2x_2 + \cdots + b_kx_k$$

6.1 Multiple Linear Regression

In practice, multiple regression is much more useful as there are often several factors that will be associated with the dependent (outcome) of interest.

Example: Framingham Study – prediction of systolic blood pressure based on several associated factors.

Firstly: simple linear regression – predictor = BMI

$$BP_{sys} = 108.28 + 0.67 \times BMI$$

6.1 Multiple Linear Regression

But in multiple regression: BMI, age, gender, hypertension treatment

- BMI and age: continuous variables (kg or years)
- Gender: dichotomised; female = 0, male = 1
- Hypertension treatment: also dichotomous, no treatment used = 0 (i.e. absence), receiving treatment = 1 (i.e. present)

$$BP_{sys} = 68.15 + (0.58 \times BMI) + (0.65 \times Age) + (0.95 \times Male) + (6.44 \times HTRx)$$

Prediction: female, 65 years, BMI = 30 and not treated for HT

$$68.15 + 17.40 + 42.25 + 0 + 0 = 127.80\ mmHg$$

6.1.1 Introduction to multiple linear regression

The aim of regression is to determine the values of b_i (the regression coefficients) which minimise the sum of the squared deviations between the predicted values of the DV and the observed values of the DV.

Although the calculation of b_i is important, in an applied statistics environment we are often interested in determining which, if any, variables are significant. In some situations we can then effectively ignore the other non-significant variables.

There are three (3) major types of multiple regression used in applied research. Unfortunately there is not one “correct” way to do it so it is up to the researcher and/or statistician to decide which is suitable.

6.1.1 Introduction to multiple linear regression

Three main types of regression methods:

1. Standard (*or Simultaneous*)

In Standard Multiple Regression all of the independent variables (IVs) are entered into the equation at the same time. So each independent variable contributes something to the regression model.

2. Sequential (*or Hierarchical*)

In Sequential Multiple Regression the independent variables are entered into the equation in an order defined by the researcher. This is often based on some theoretical belief on what independent variables are important or based on previous research.

6.1.1 Introduction to multiple linear regression

Three main types of regression methods:

3. Stepwise (*or Statistical*)

There are several types of Stepwise Multiple Regression. Two of the more common ones used in applied research are Backward Elimination and Forward Selection. In Backward Elimination all variables are entered into the equation and then sequentially removed (basically the variable with the smallest partial correlation with the dependent variable is removed until only the important variables are remaining). In Forward Selection the variables are sequentially entered into the model (basically the important variables are added sequentially until adding any further variables does not improve the model).

6.1.1 Introduction to multiple linear regression

Let's look at both simultaneous and statistical types in more detail.

Standard (Simultaneous) Multiple Linear Regression example

Eg. 100 patients were involved in a study identifying factors that lead to the risk of heart disease. Each patient was asked to complete a survey about the following variables:

Response variable: Risk of developing heart disease (measured on a scale of 0-100)

Predictor variables:

- Age (in years)
- Body mass index (BMI, in kg/m^2)
- Physical activity level (measured in MET-minutes/week)

The data set that will be used for this example is called MLRHeartRiskExample.xlsx

6.1.2 Assumptions of multiple linear regression

There are a number of assumptions for multiple regression. In applied problems it is rare that all of the assumptions are satisfied so some flexibility is required.

The main assumptions to consider are:

Dependent variable – A key assumption of the MLR is that the dependent (outcome) variable is continuous.

Outliers – Multiple regression is quite sensitive to outliers in the dependent and independent variables so you need to consider whether to include the outliers or not. Note that there is no correct way to do this. Obviously if you find an outlier that cannot happen (eg. You find someone in the data set with a negative age!) then you can certainly exclude that observation or, if possible, a better option is to try and find out the correct measurement and fix up the incorrect data set.

6.1.2 Assumptions of multiple linear regression

There are a number of assumptions for multiple regression. In applied problems it is rare that all of the assumptions are satisfied so some flexibility is required.

The main assumptions to consider are:

Multicollinearity – This is when independent variables are “highly correlated” (eg. $r \geq 0.9$). Keep in mind though that there is no practical difference between a correlation of 0.89 and 0.9 so care must be taken when evaluating this assumption. Multicollinearity can be detected with the Tolerance and VIF (variance inflation factor) statistics in SPSS. Note: You do not want multicollinearity!

Normality, Homoscedasticity, Linearity of Residuals –

Normality: The residuals should be approximately normally distributed about the predicted dependent variable values.

Homoscedasticity: The variance of the residuals about the predicted dependent variable values should be the same for all predicted values.

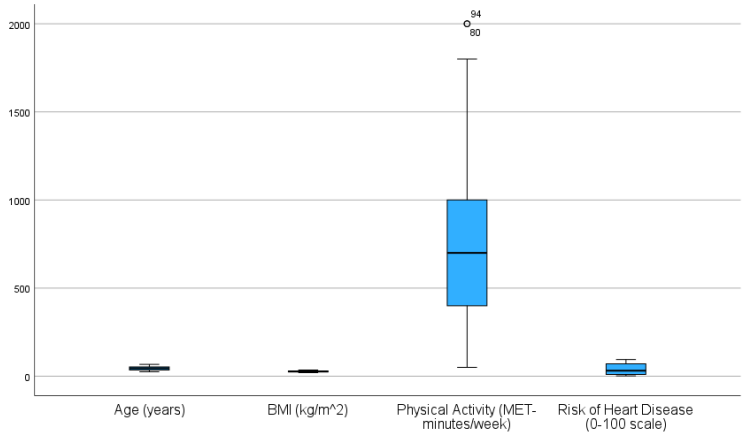
Linearity: The residuals should have a linear relationship with the predicted dependent variable values.

These three assumptions can be evaluated through graphical methods in SPSS.

6.1.2 Assumptions of multiple linear regression

Our dependent variable is a value between 0 and 100 and therefore is on the ratio/interval scale.

When looking for outliers, we can see that SPSS has identified two potential outliers:



These are rows 80 and 94 for the Physical Activity. Both of these values have 2000 MET-minutes per week. We need to test if these are outliers and should be removed.

6.1.2 Assumptions of multiple linear regression

We can calculate a z-score to determine how far these observations are away from the mean.

$$z = \frac{observation\ - mean}{standard\ deviation}$$

The condition is that **if the value is > 3.29 standard deviations** then it is considered an outlier and should be removed. We can get mean and SD from SPSS.

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Physical Activity (MET-minutes/week)	100	50	2000	738.10	460.041
Valid N (listwise)	100				

$$z = \frac{2000 - 738.10}{460.04} = 2.74$$

Therefore we can say that these points are not outliers as 2.74 < 3.29.

The remaining assumptions of multiple linear regression are done through doing the actual MLR procedure in SPSS.

6.1.2 Assumptions of multiple linear regression

Points to note:

The most important aspect of getting good data is having a large enough sample size.

Due to the subjectivity nature of hypothesis test assumptions, it is often common that not ALL assumptions are satisfied for multiple linear regression.

This is OK as long as the distributions do not completely deviate away from the assumption being tested.

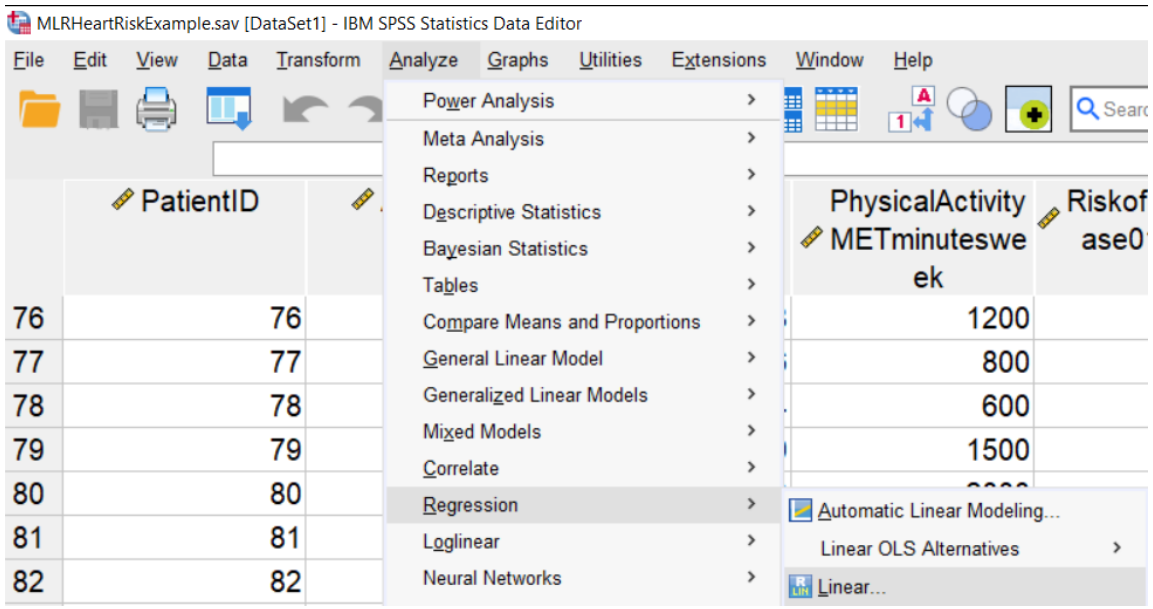
6.1.3 Multiple Linear Regression in SPSS

The remaining assumptions and standard multiple regression

<Analyze>

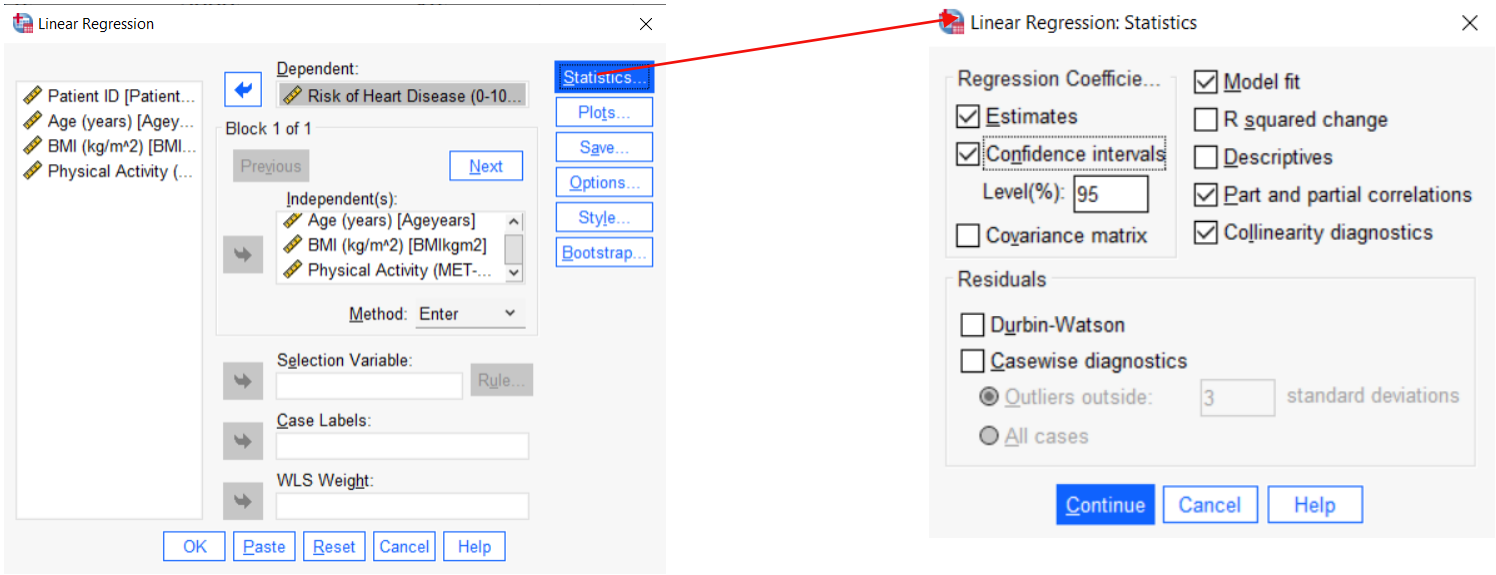
<Regression>

<Linear>



6.1.3 Multiple Linear Regression in SPSS

The remaining assumptions and standard multiple regression



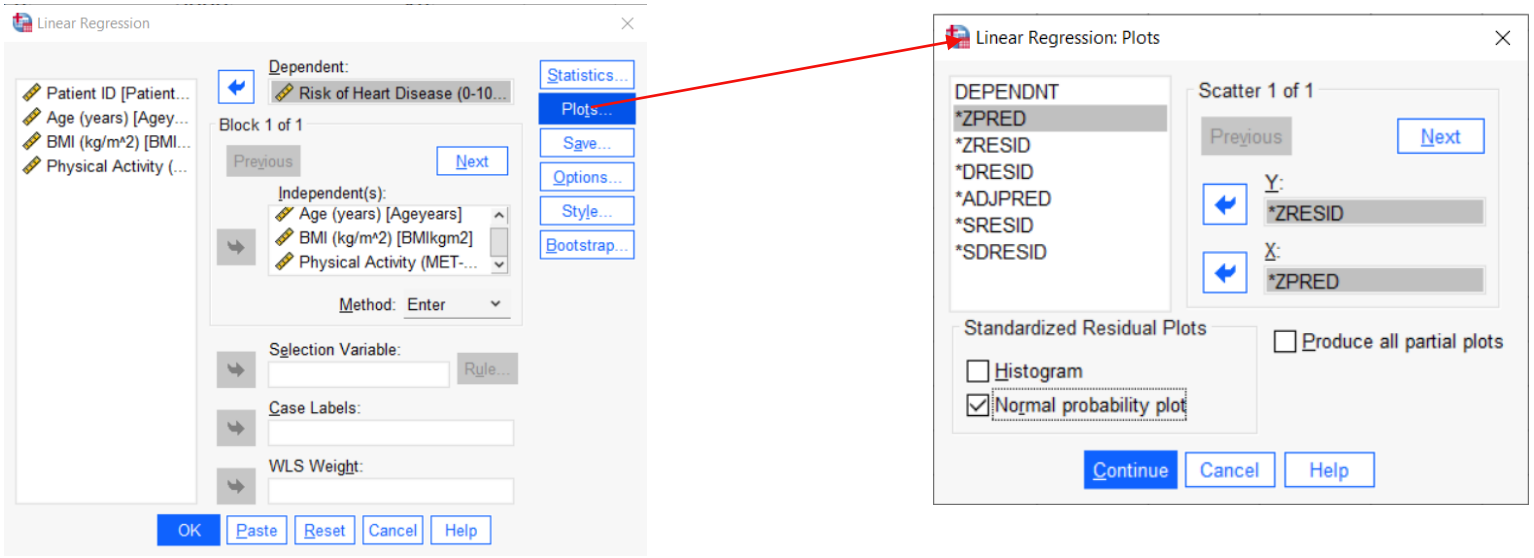
Estimates: provides information on each predictor in the model

Model fit: provides information to assess predictive utility of the model

Other options that are useful are: part and partial correlation and collinearity diagnostics. Used to help us assess the multicollinearity assumption of MLR.

6.1.3 Multiple Linear Regression in SPSS

The remaining assumptions and standard multiple regression



Placing *ZRESID in Y and *ZPRED in X as well as checking Normality probability plot will give us the information needed to assess the normality, linearity and homoscedasticity assumptions of the residuals.

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Physical Activity (MET-minutes/week), Age (years), BMI (kg/m^2)[b]	.	Enter

a. Dependent Variable: Risk of Heart Disease (0-100 scale)

b. All requested variables entered.

In a standard multiple linear regression, all predictor (independent) variables are entered in all at once. This means we only have one model to consider.

This table summarises the variables that have been entered into the model. As you can see these are the predictor variables of Physical Activity, Age and BMI.

The dependent variable is also quoted here to be Risk of Heart Disease. This is what we would expect as this is how we set up our MLR in SPSS.

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.687[a]	.472	.455	23.080

a. Predictors: (Constant), Physical Activity (MET-minutes/week), Age (years), BMI (kg/m^2)

b. Dependent Variable: Risk of Heart Disease (0-100 scale)

The Model Summary table includes the following statistics:

R – This is the correlation coefficient but in regression is often not used as R^2 is deemed more appropriate.

R^2 – This represents the proportion of variance in the risk of heart disease score that can be accounted for by the predictors. Here, 47.2% of the variance can be explained by the continuous predictor variables.

Adjusted R^2 – This is a more accurate measure of the R^2 value above. SPSS calculates this value in order to take in to account the fact that we have a multivariate model. It is often wise to quote the adjusted R^2 value when presenting statistical results for a MLR.

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45655.810	3	15218.603	28.571	<.001[b]
	Residual	51135.900	96	532.666		
	Total	96791.710	99			

a. Dependent Variable: Risk of Heart Disease (0-100 scale)

b. Predictors: (Constant), Physical Activity (MET-minutes/week), Age (years), BMI (kg/m^2)

The ANOVA table helps us to know if our regression model has predictive utility. In other words, are we able to use this model to make accurate predictions about our dependent variable?

If we look at the p-value (Sig.) of this table, we see that it is less than .001. What this is telling us is that the slope of our regression curve is not zero.

When writing up a MLR, it is always good to quote the test statistic ($F = 28.57$), df (regression) = 3, df (residual) = 96) and $p < .001$.

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-123.024	23.267		-5.287	<.001
	Age (years)	.709	.262	.253	2.703	.008
	BMI (kg/m^2)	4.879	.961	.497	5.079	<.001
	Physical Activity (MET-minutes/week)	-.001	.005	-.014	-.172	.864

a. Dependent Variable: Risk of Heart Disease (0-100 scale)

The first half of the coefficients table details the role of each predictor variable in the regression model. This table allows us to understand how each predictor impacts our dependent variable.

Column B represent the coefficients for each variable in our regression model. B indicates the predicted change in the dependent variable associated with a 1-unit change in the predictor, after controlling for the effects of all the other predictors. For example, the unstandardized regression coefficient for Age (years) is .709, which means after controlling for BMI and Physical Activity, it is predicted that there will be a .709-unit increase (due to positive sign) in risk of heart disease.

Standardized coefficients predict changes in standard deviations. So the standard regression coefficient for Physical Activity is -0.014, which means after controlling for all other predictors, a 1 SD increase in Physical Activity will result in a 0.014 SD decrease in risk of heart disease.

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-123.024	23.267		-5.287	<.001
	Age (years)	.709	.262	.253	2.703	.008
	BMI (kg/m^2)	4.879	.961	.497	5.079	<.001
	Physical Activity (MET-minutes/week)	-.001	.005	-.014	-.172	.864

a. Dependent Variable: Risk of Heart Disease (0-100 scale)

The t and Sig. columns represent the test statistic and P-value respectively and give us an indication of whether or not each predictor accounts for a significant proportion of unique variance in the dependent variable.

We see that Age (years) and BMI (kg/m^2) are significant, which indicates to us that they are good predictors of risk of heart disease.

Interestingly, Physical Activity for this data set is not seen as a good predictor of risk of heart disease!

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.

Correlations			Collinearity Statistics	
Zero-order	Partial	Part	Tolerance	VIF
.560	.266	.201	.627	1.596
.657	.460	.377	.575	1.740
-.254	-.018	-.013	.867	1.153

The correlations section of the coefficients table provides three statistics for each predictor.

Zero-order – Assuming only two variables, this is the Pearson's correlation coefficient for the predictor (independent) and outcome (dependent) variable.

Partial – the partial correlation between the predictor and outcome.

Part – the semi-partial correlation between the predictor and outcome.

Part correlation can be squared to give the proportion in variance in the outcome variable that can be uniquely explained by the predictor variable. Eg. Part correlation for Physical Activity is -0.013 indicating that 0.017% of the variance in Risk of Heart Disease can be uniquely attributed to Physical Activity score.

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.

Correlations			Collinearity Statistics	
Zero-order	Partial	Part	Tolerance	VIF
.560	.266	.201	.627	1.596
.657	.460	.377	.575	1.740
-.254	-.018	-.013	.867	1.153

Two measures of multicollinearity are provided in the collinearity statistics section of the coefficients table:

Tolerance - predictor variables with tolerances < 0.1 are multicollinear with one or more other predictors.

VIF – predictor variables with VIFs > 10 are cause for concern as they indicate multicollinearity with other predictor variables.

As can be seen, there are no issues with multicollinearity from looking at our SPSS table.

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions (Constant)	Age (years)	BMI (kg/m^2)	Physical Activity (MET-minutes/week)
1	1	3.711	1.000	.00	.00	.00	.01
	2	.257	3.803	.00	.02	.00	.72
	3	.028	11.521	.11	.78	.03	.12
	4	.005	28.666	.89	.20	.97	.15

a. Dependent Variable: Risk of Heart Disease (0-100 scale)

The two columns that are of most interest to us here are the **eigenvalues** and **condition index** columns.

Eigenvalues that are close to the value of zero indicate that we might have multicollinearity however, a closer inspection of the condition index column will help us know whether or not this is true.

The condition index is calculated by taking the square root of the ratio of dimension 1 and another dimension. Eg. The condition index for dimension 3 is $\sqrt{\frac{3.711}{.028}} = 11.51$ (note: there will always be some rounding error with SPSS). Condition index values above 15 are a cause for concern and indicate there are multicollinearity problems.

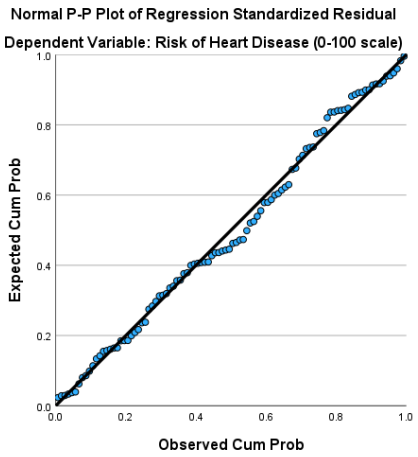
This is also evident if a value of 0.90 shows up **twice** or more in the variance proportion columns.

Dimension 4 shows a condition index 28.67, which is greater than 15, however, only a value greater than 0.90 shows up once (for BMI) so this is no cause for concern.

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.



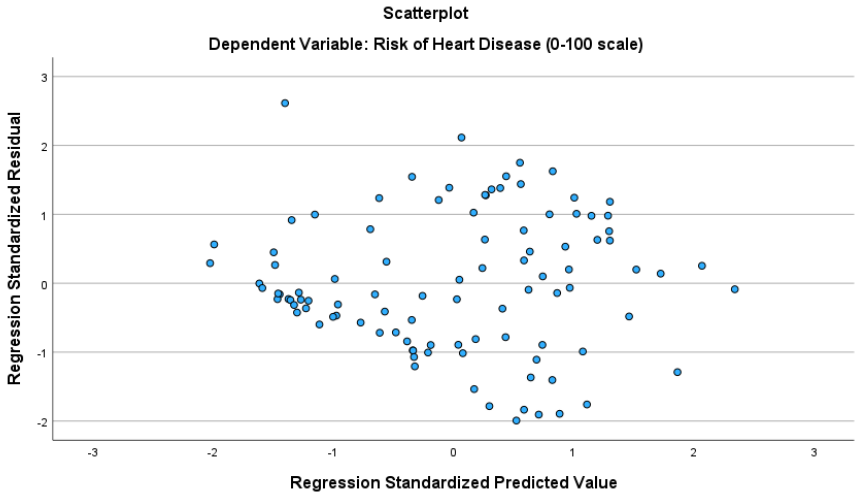
The normal P-P (Probability) Plot of Regression Standardized Residuals can be used to asses the assumption of normality distributed residuals.

If the data points are clustered closely to the diagonal line, the residuals are normally distributed.

6.1.3 Multiple Linear Regression in SPSS

Multiple linear regression SPSS output

SPSS will give you a number of different tables and graphs. Let's go through each one step by step.



The scatterplot of standardised residuals against standardised predicted values can be used to assess the assumptions of normality, linearity and homoscedasticity of residuals.

As the data points are spread out and do not show a particular shape (not arching or a cone shape), then the assumptions have been met.

6.1.4 Interpreting the results of a multiple linear regression model

A multiple linear regression was used to predict the risk of heart disease based on age, BMI and physical activity.

A significant regression equation was found, $F(3, 96) = 28.57$, $p < .001$, with an adjusted $R^2 = .455$. Risk of Heart Disease = -123.02 + .71 x age + 4.879 x BMI.

This indicates that risk of heart disease increased .71 likelihood with each unit increase in age.

OR

This indicates that risk of heart disease increased by 4.88 with each unit increase in BMI.

Note that each of these statements suggest that “all other predictors are held constant”.

(Cronk, 2012)

6.1.5 Using a backward elimination method

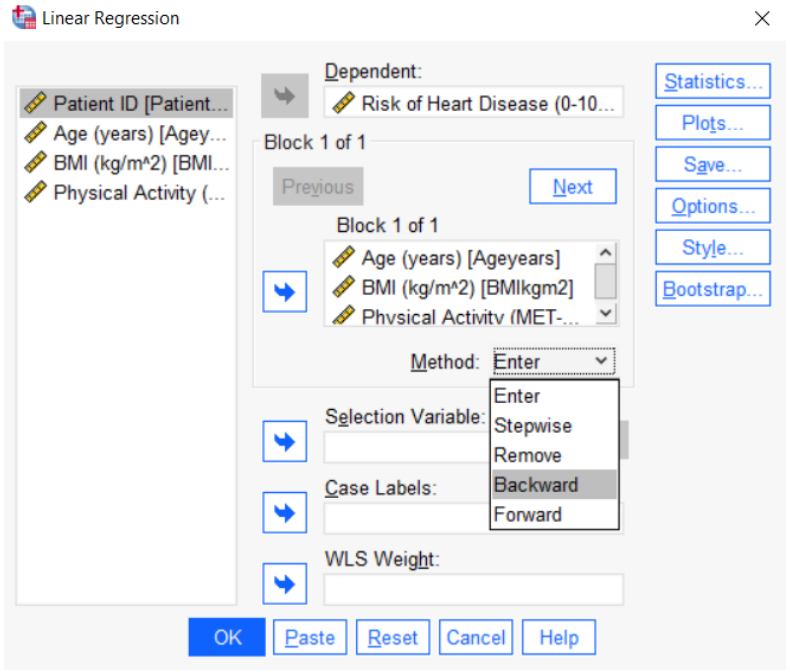
Let's now look at how we would do a statistical MLR in SPSS.

We previously used the "Enter" method.

But as can be seen, there are more methods.

The backward elimination is typically a preferred method.

Let's apply it and see what happens.



6.1.5 Using a backward elimination method

Backward elimination method

Variables Entered/Removed[a]

Model	Variables Entered	Variables Removed	Method
1	Physical Activity (MET-minutes/week), Age (years), BMI (kg/m^2)[b]	.	Enter
2	.	Physical Activity (MET-minutes/week)	Backward (criterion: Probability of F-to-remove >= .100).

a. Dependent Variable: Risk of Heart Disease (0-100 scale)

b. All requested variables entered.

Here we can see that in Model 1 all the predictor variables have been entered in but for each progressing model, one predictor is eliminated from the regression model.

The benefit of this approach enables us to find the best regression model by eliminating predictor variables that may not have a lot of relevance to us (ie. Do not show any statistical significance).

You can see that SPSS had decided that the best model is Model 2 where all non-significant predictor variables are eliminated from the model and only the significant ones remain (as we previously saw using the Enter method).

6.1.5 Using a backward elimination method

Backward elimination method

Model Summary[c]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.687[a]	.472	.455	23.080
2	.687[b]	.472	.461	22.964

a. Predictors: (Constant), Physical Activity (MET-minutes/week), Age (years), BMI (kg/m^2)

b. Predictors: (Constant), Age (years), BMI (kg/m^2)

c. Dependent Variable: Risk of Heart Disease (0-100 scale)

The Model Summary that gives the R, R square and Adjusted R Square values for each model.

6.1.5 Using a backward elimination method

Backward elimination method

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45655.810	3	15218.603	28.571	<.001[b]
	Residual	51135.900	96	532.666		
	Total	96791.710	99			
2	Regression	45640.106	2	22820.053	43.274	<.001[c]
	Residual	51151.604	97	527.336		
	Total	96791.710	99			

a. Dependent Variable: Risk of Heart Disease (0-100 scale)

b. Predictors: (Constant), Physical Activity (MET-minutes/week), Age (years), BMI (kg/m^2)

c. Predictors: (Constant), Age (years), BMI (kg/m^2)

Models 1 and 2 are statistically significant as P < .001.

But recall that Model 2 has eliminated non-significant predictors.

Which as a result simplifies our regression equation.

6.1.5 Using a backward elimination method

Backward elimination method

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	-123.024	23.267		-5.287	<.001
	Age (years)	.709	.262	.253	2.703	.008
	BMI (kg/m^2)	4.879	.961	.497	5.079	<.001
	Physical Activity (MET-minutes/week)	-.001	.005	-.014	-.172	.864
2	(Constant)	-125.023	20.044		-6.237	<.001
	Age (years)	.710	.261	.254	2.720	.008
	BMI (kg/m^2)	4.927	.915	.502	5.382	<.001

a. Dependent Variable: Risk of Heart Disease (0-100 scale)

Coefficients table for each model. Model 2 again showing the removal of the non-significant predictor (Physical Activity).

When reporting your results, you would use the last model that SPSS gives you i.e. Model 2. You would also state that you have used a “backward elimination” method or approach.