Seminar 6
How to conduct and interpret the standard multiple regression

Quantitative Research Methods
Dr. Naser Alziyadat

Murdoch UNIVERSITY

REGRESSION

Regression allows the researcher to make *predictions* of the likely values of the dependent variable Y

from known values of independent variable X in a **simple linear regression**,

or from known values of a combination of independent variables D, E, and F in **multiple linear regression.**

Use of Regression

Murdoch UNIVERSITY

We often need to determine such issues as:

- the relationship between decrease in pollutant emissions and a factory's annual expenditure on pollution abatement devices. If we spend more on abatement can we decrease emissions even more?
- how investment varies with interest rates. Can we predict how much more investment occurs with a 1% interest rise?
- how unemployment varies with inflation. What level will unemployment reach if inflation increases by 5% this year ?

Examples of simple and multiple linear regression questions

- In simple linear regression – does the number of customers predict value of sales - variations in one IV predicting variations in one DV

- In multiple linear regression – does maximising value of sales depend on a particular combination of the number of customers, price variations, number of sales outlets, number of salespersons…etc.

REGRESSION

- Regression therefore investigates relationships between IV and DV in terms of the predictive ability of the IV to predict (estimate) the DV

- It is therefore closely linked to correlation and shares many of the assumptions of 'r', e.g.

  * the relationships should be linear

  * the measurements of both the IV and DV variables must be interval or ratio (scale data)

IMPORTANT CONCEPTS

- ***Predictor variable.***

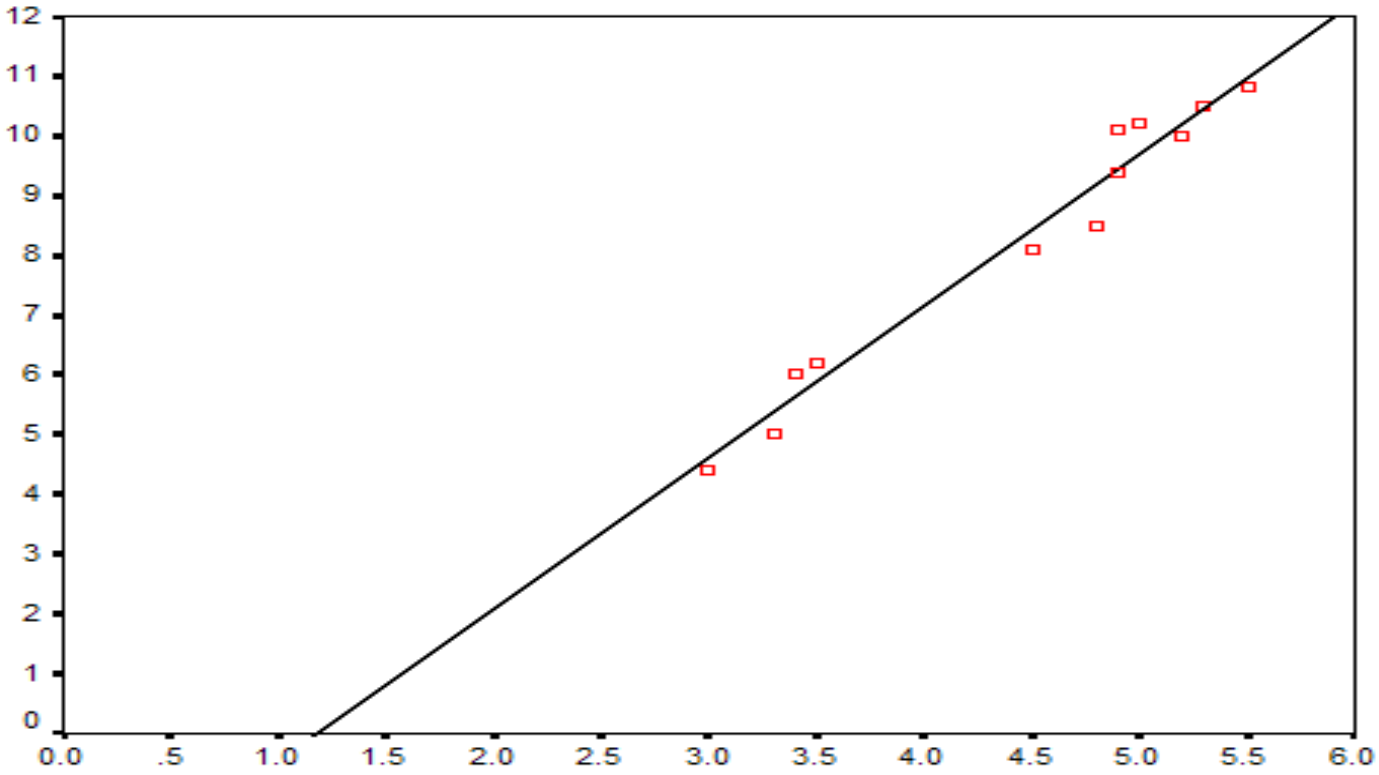
  A variable (IV) from which a value is used to estimate a value on another variable (DV)

- ***Criterion variable.***

  A variable (DV) a value of which is estimated from a value of the predictor variable (IV)

USING A LINE TO SUMMARISE DATA

When we look at two variables together we summarise that relationship with a straight line – the scattergraph showing the intersection of paired X and Y values.



USING A LINE TO SUMMARISE DATA

A straight line may be represented graphically (previous slide) or as an equation. The general form of the equation of a straight line is:

$$Y = b_0 + b_1X$$

This is the **Regression Equation** and defines the **Line of Best Fit**

Where:

Y is the variable on the vertical axis

X is the variable on the horizontal axis

$\mathbf{b_0}$ is the value of Y where the line of best fit intercepts the Y axis (also called **the Constant**)

$\mathbf{b_1}$ is the slope of the line (the bigger b, the steeper the line)

LINE OF BEST FIT

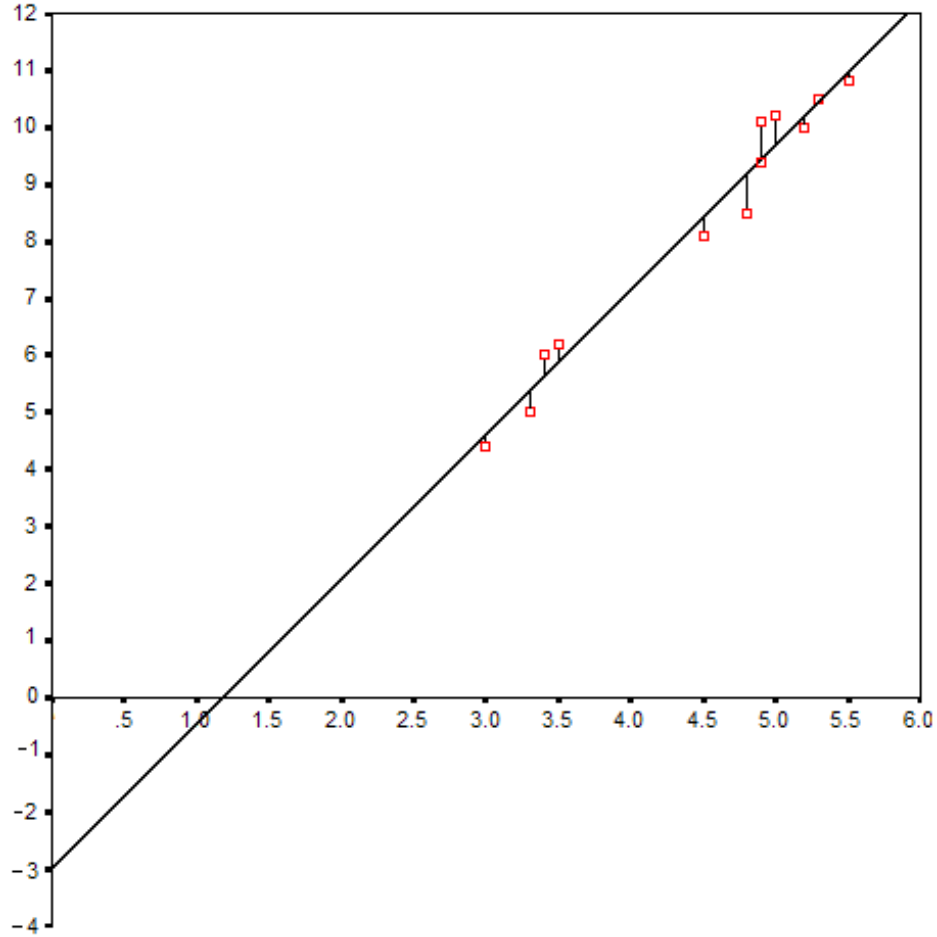
- This is the straight line on a scattergraph that ‘fits’ the scatter points best i.e. as closely possible

- This **line of best fit** minimises the deviations from the line of all the points on a scattergraph

- It make errors of prediction of Y as small as possible.

Line of Best Fit and Least Squares Solution

If our line was perfect, each datapoint would lie on the line. We can examine the error in terms of the vertical discrepancy between each datapoint and the line



We can think of these vertical distances as the "**error**" or "**residual**".

By squaring these distances we have the "**squared residuals**".

The "**best**" line is the line which has the **smallest value** of these **squared residuals**.

The method which determines the line of best fit is the "**least squares**" method.

The Least Squares Solution

Murdoch UNIVERSITY

- This is the model that minimises the sum of the squared deviations from each point to the regression line

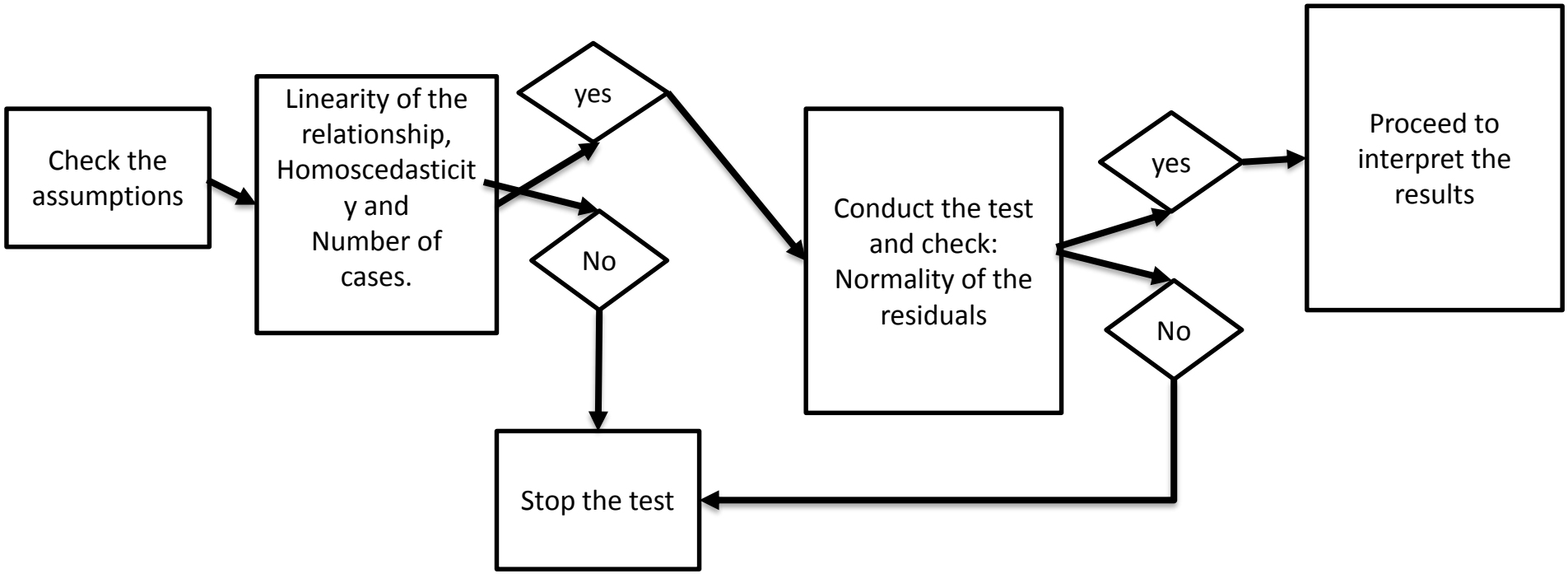
- The regression line defined by the least squares model is the line of best fit.

ASSUMPTIONS OF REGRESSION

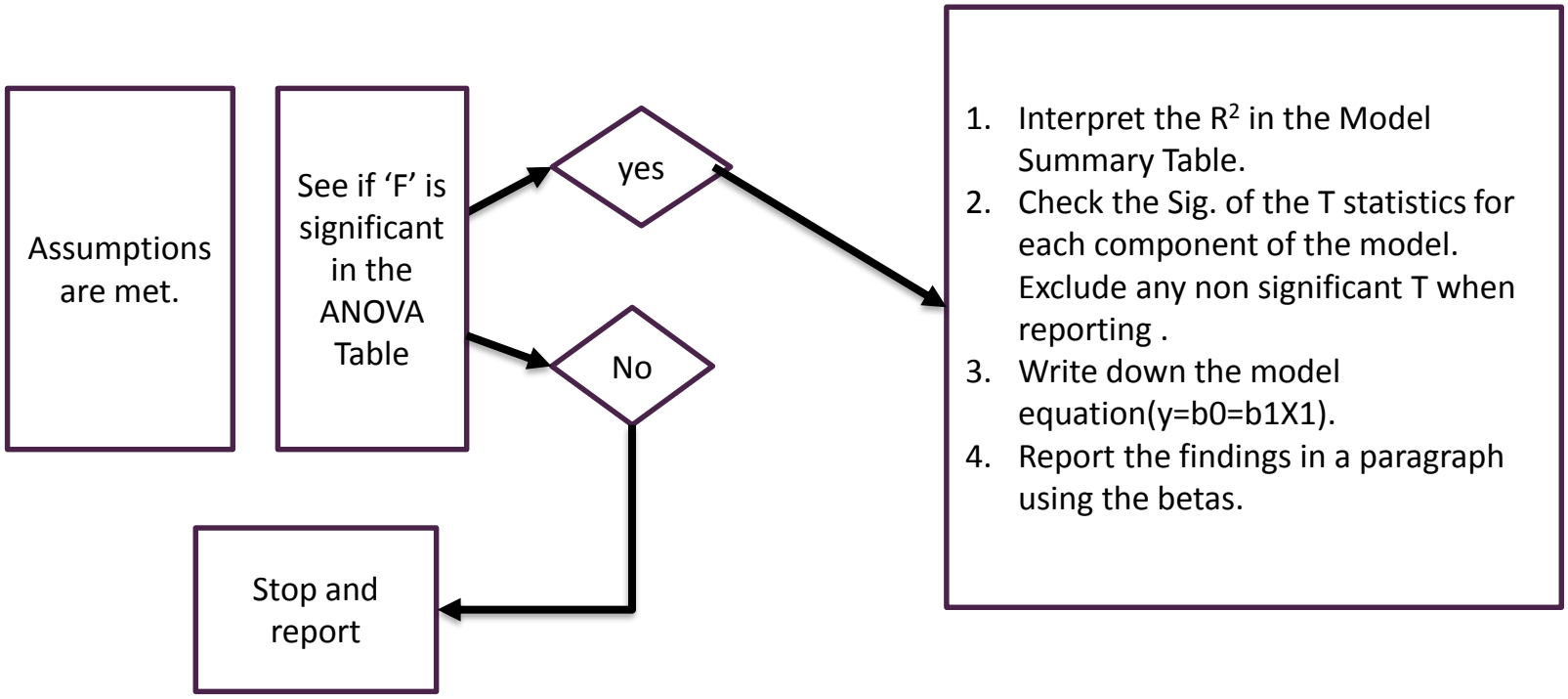
Murdoch UNIVERSITY

- A minimum requirement is to have at least 15 times more cases than IV's i.e. with 3 IV's - a minimum of 45 cases. This assumption should be checked before we proceed with the test.
- Outliers should be removed. One extremely low or high value distorts the prediction by changing the angle of slope of the regression line. We do not practice this in this unit because of time restrictions.
- Differences between obtained and predicted DV values should be normally distributed. We will test this after we obtain the results.
- variance of residuals the same for all predicted scores (*homoscedasticity*). We will test this after we obtain the results.
- Regression procedures assume that the dispersion of points is linear. This is needed to be tested before we proceed in simple liner regression and is assumed in multiple linear regression.
- There is no implication that an increase in *X causes an* increase in Y. Simultaneous increase in X and Y may have been caused by an unknown third variable excluded from the study.

Simple Linear Regression



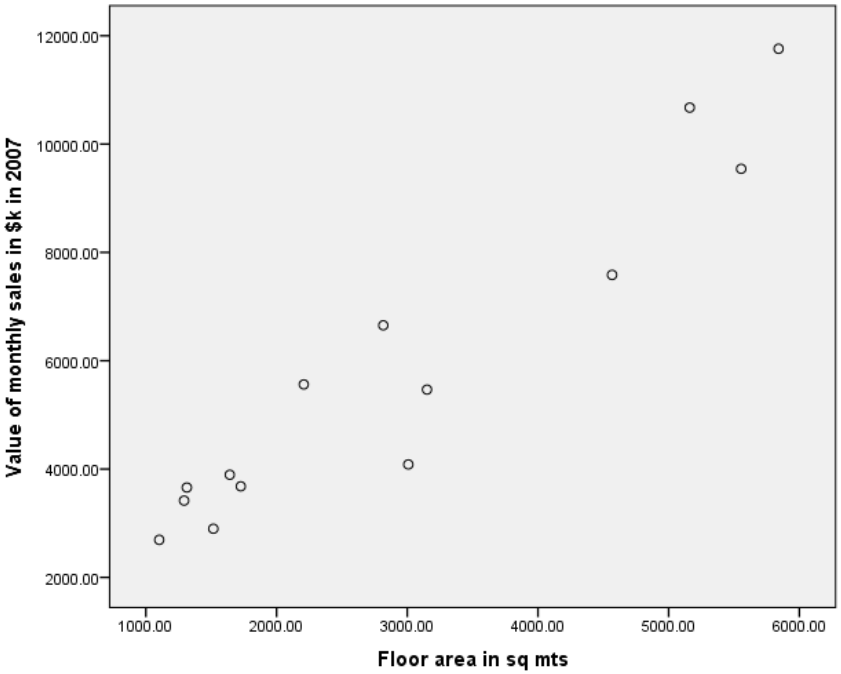
Simple Linear Regression



Example

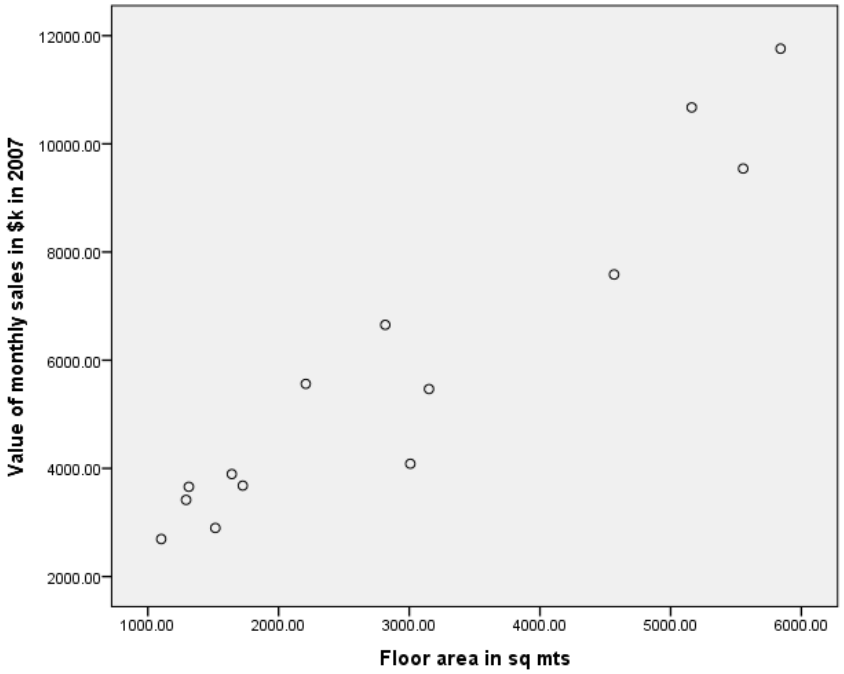
- Assume we want to predict 'value of monthly sales' from knowledge of 'floor area in sq m.' Access the data file Chapter 16 C from the LMS.
- First, test the assumptions of linearity and homoscedasticity by generating a scatter graph for the two variables

Testing linearity

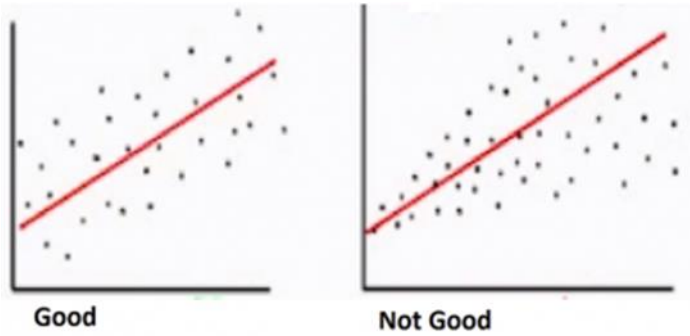


Is this a linear relationship? Yes, we can draw a line that can represent the direction of the relationship and is as close as to all the data points.

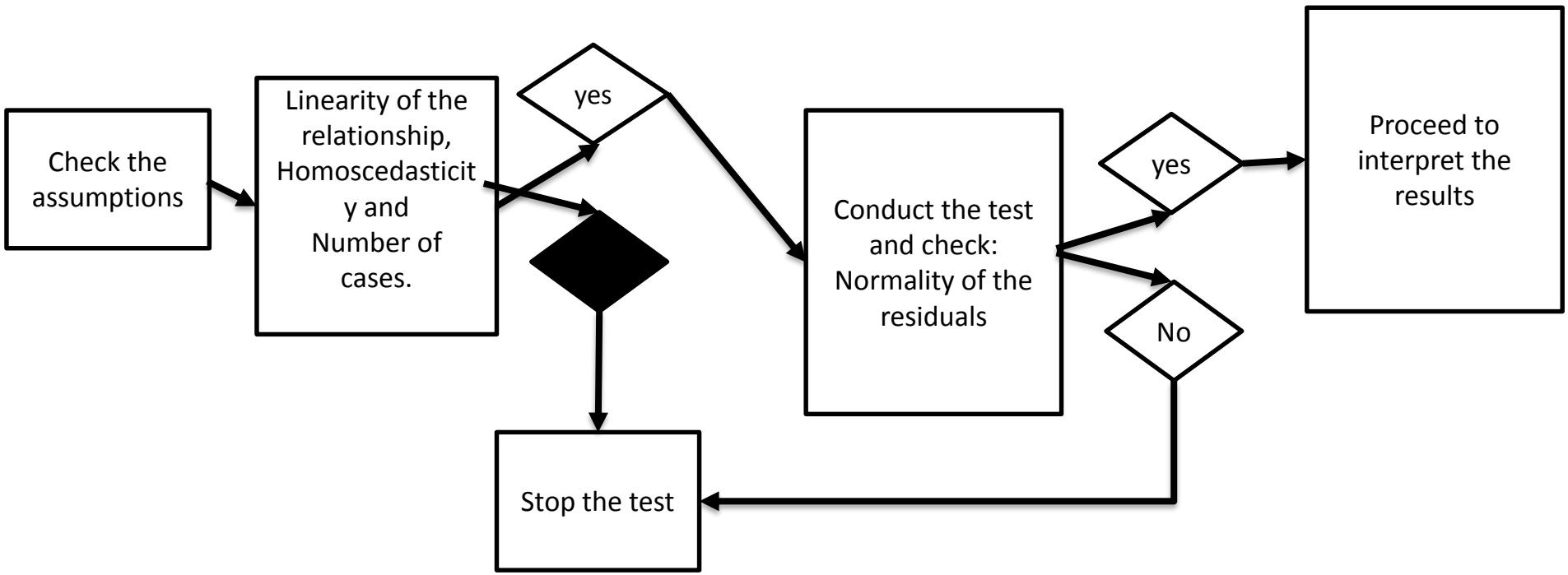
Testing homoscedasticity



Homoscedasticity makes the data take the shape of a funnel or a cigar. Our variables do not breach homoscedasticity.

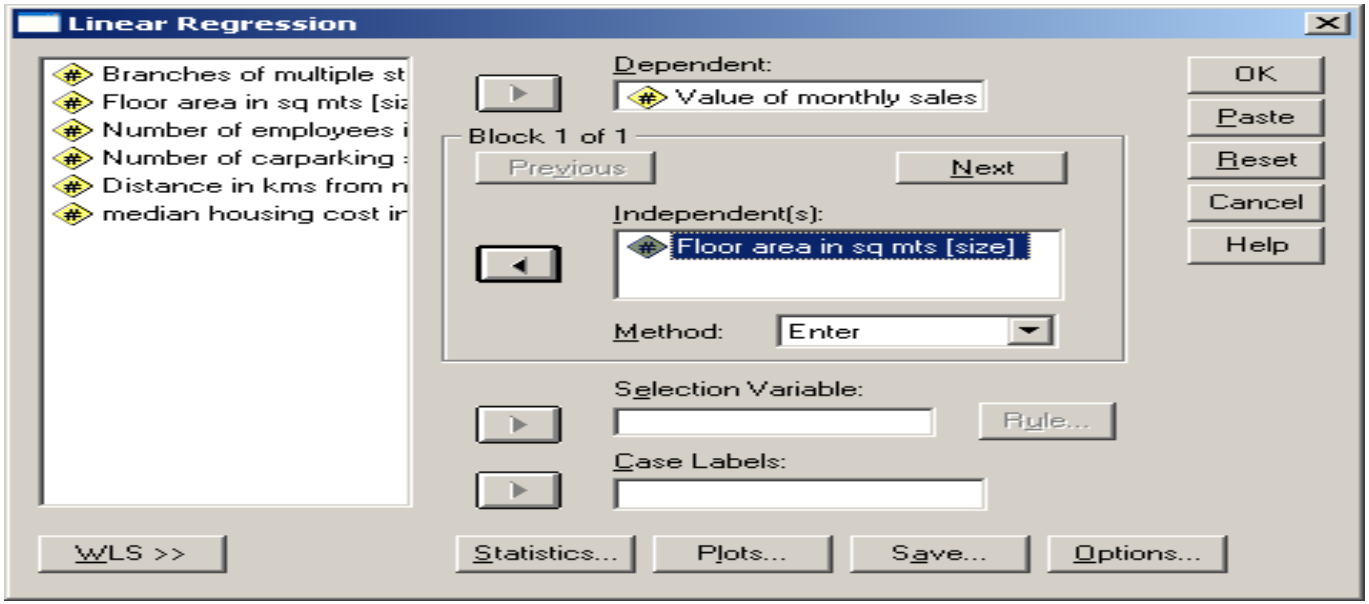


Remember


Check the assumptions
Linearity of the relationship, Homoscedasticity and Number of cases.
yes
Conduct the test and check: Normality of the residuals
yes
Proceed to interpret the results
No
Stop the test

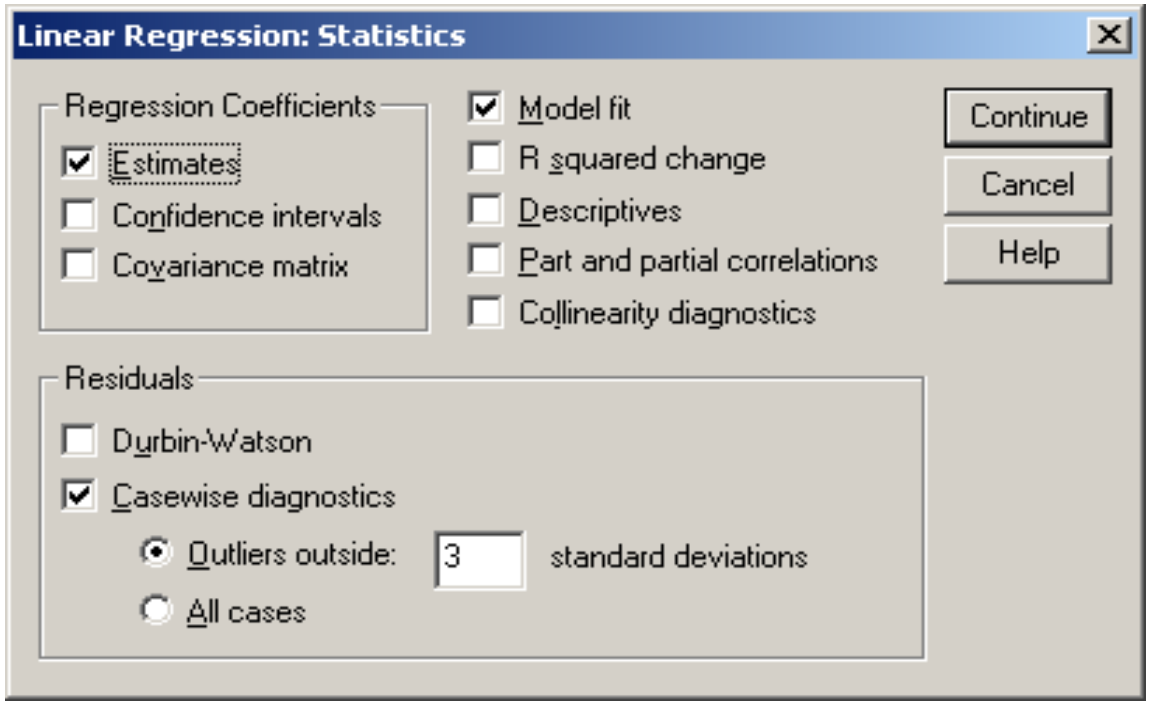
Simple Linear Regression - SPSS

1. Assume we want to predict 'value of monthly sales' from knowledge of 'floor area in sq m.' Access the data file Chapter 16 C from the LMS.
2. Select ***Analyse*** and then ***Regression***.
3. Choose ***Linear*** to open the ***Linear Regression*** dialogue box
4. Click on the dependent variable (*value of sales per month*) and place it in the ***Dependent*:** box.
5. Select the independent variable*(floor area in sq mts*) and move it into ***Independent [s]*:** box
6. In ***Method*** box ensure ***Enter*** is selected.



Simple Linear Regression - SPSS

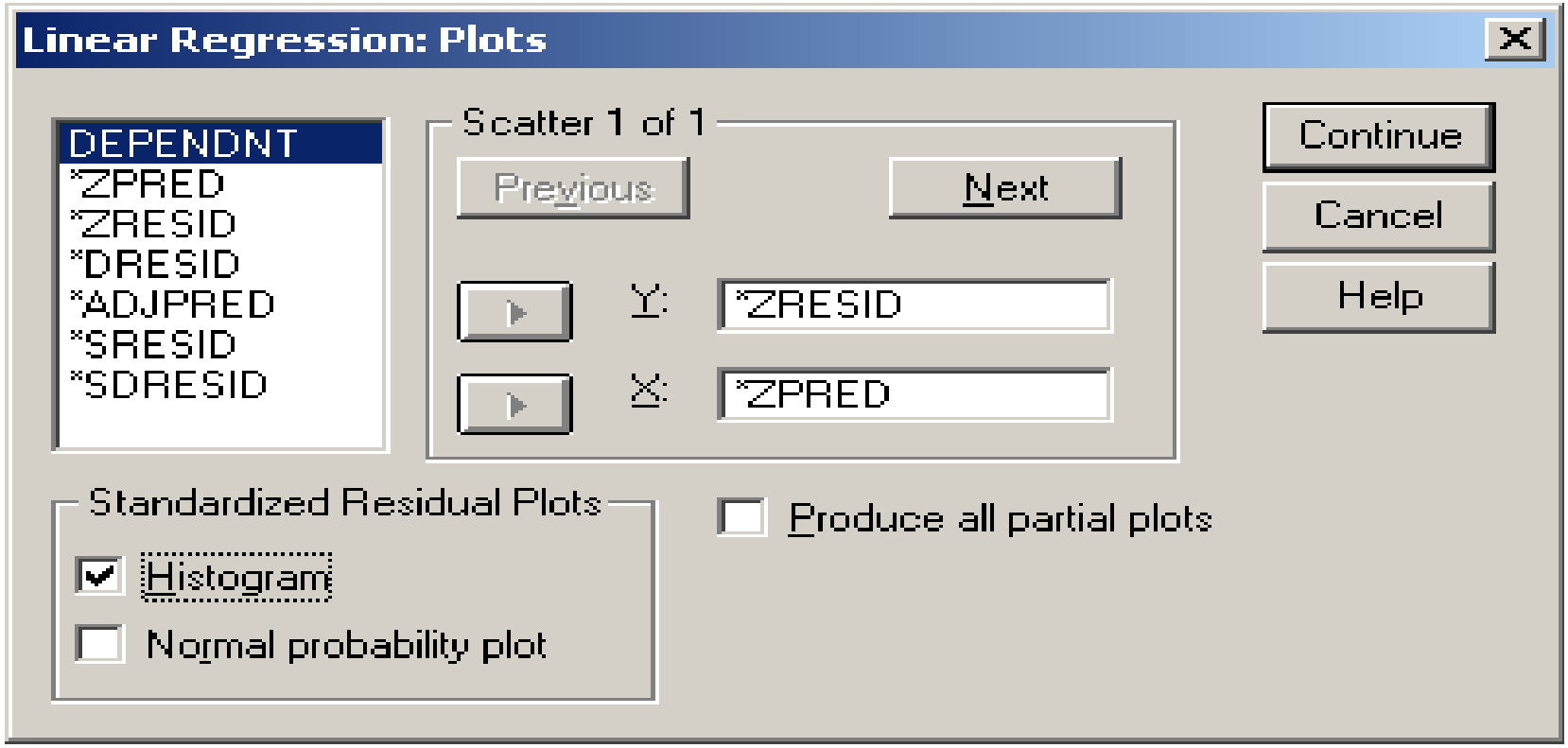
7. Select ***Statistics*** to obtain ***the Linear Regression: Statistics*** dialogue box.
8. Choose ***Estimates***, and ***Model fit***. Check ***Casewise diagnostics*** box to detect outliers and accept default value of ***3*** sd's



Simple Linear Regression - SPSS

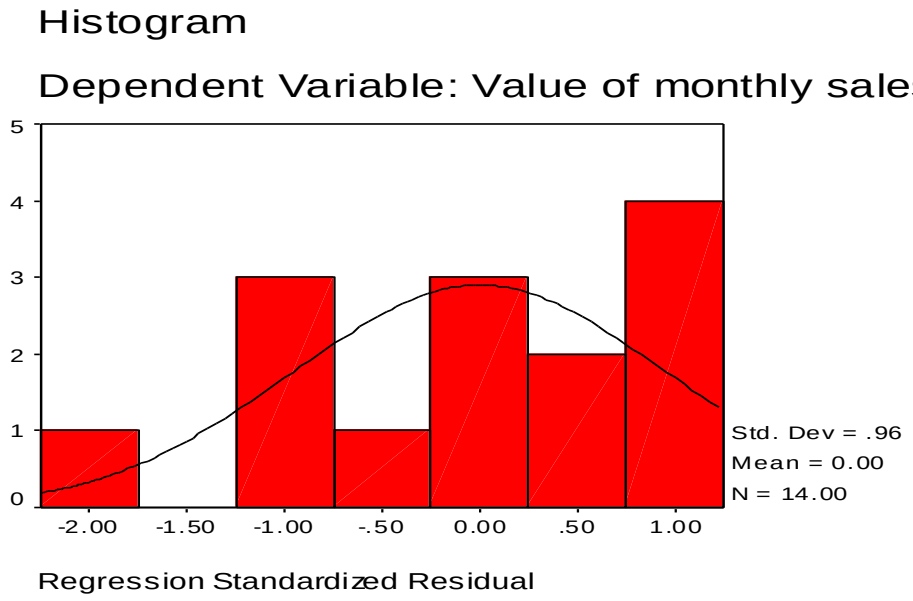
Murdoch UNIVERSITY

9. Next click ***Plots***. Place ***ZRESID*** into ***Y*** box and ***ZPRED*** into ***X*** box, finally select ***histogram***.
10. ***Continue*** and finally ***OK***

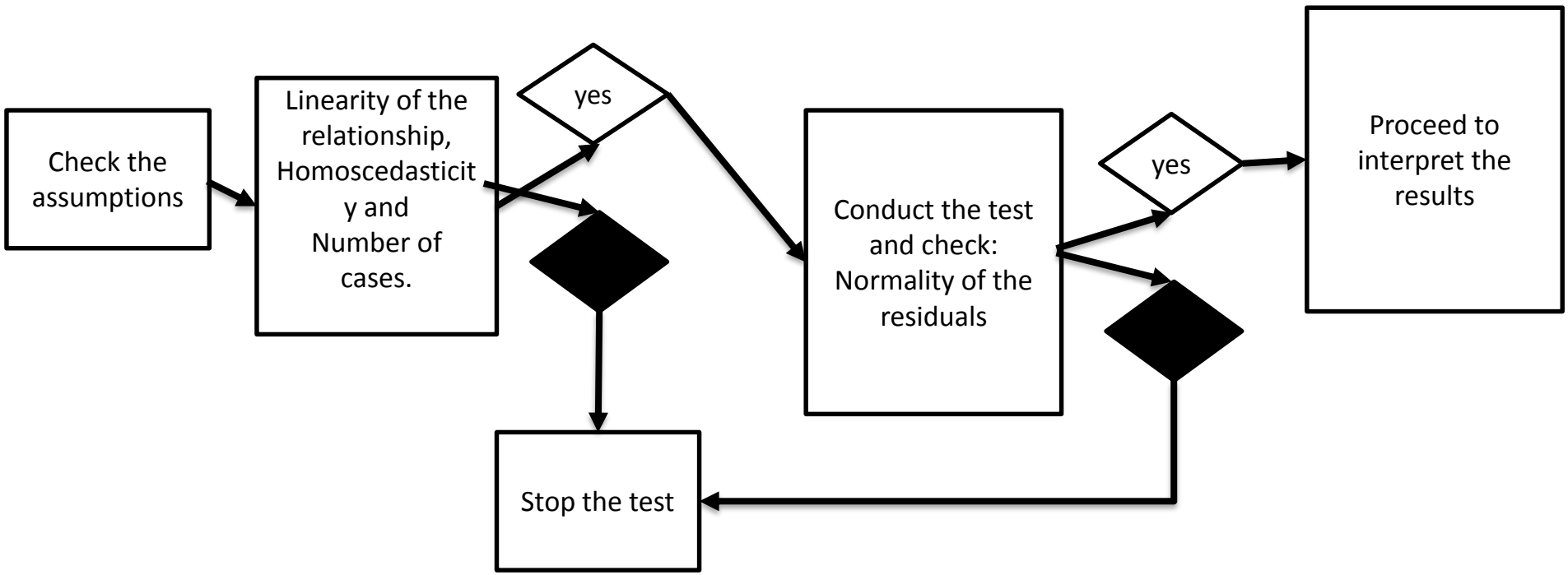


Check the normality of the

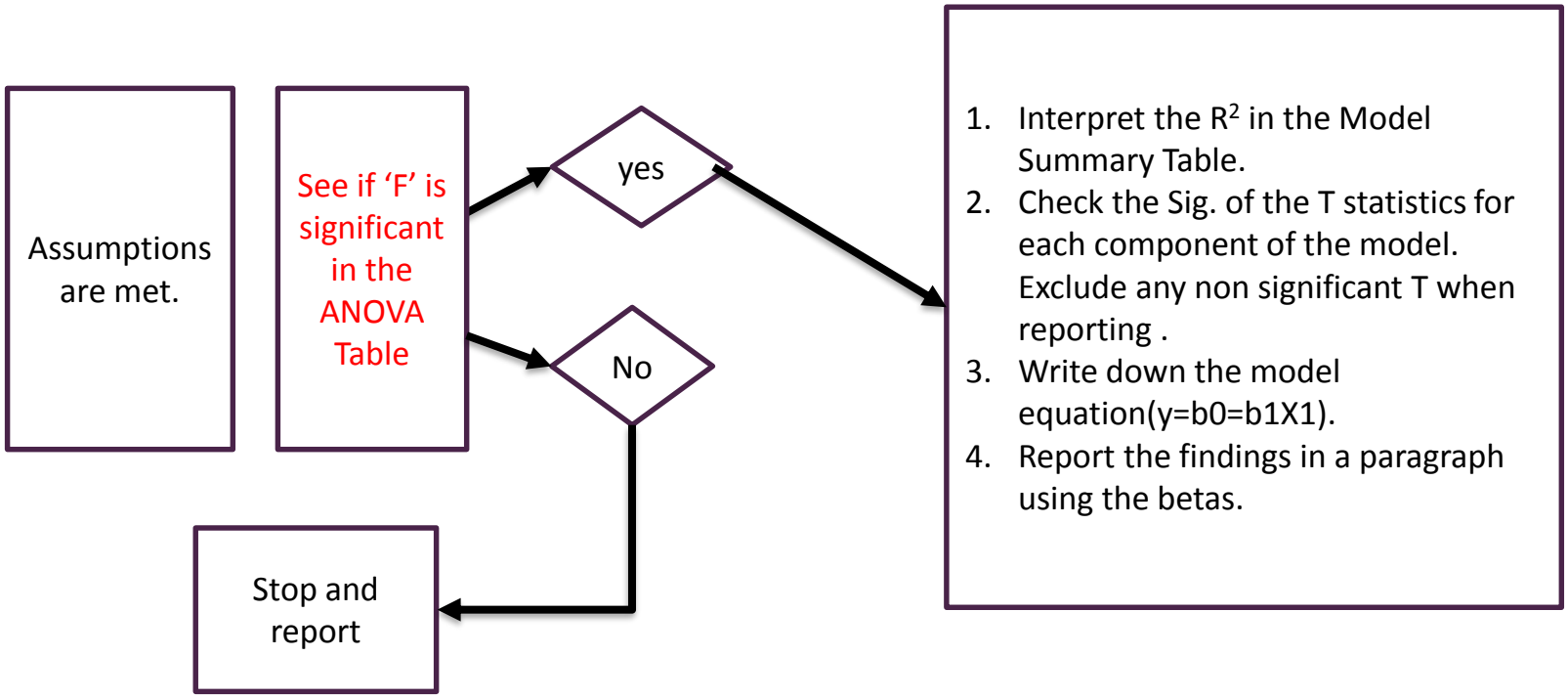
- A histogram and standardized residual scattergraph are usually obtained because they are essential to address the issue of whether major assumptions for linear regression were met.
- The histogram assesses normality and reveals no definite skewness or extreme outliers.



Remember


Check the assumptions
Linearity of the relationship, Homoscedasticity and Number of cases.
yes
Conduct the test and check: Normality of the residuals
yes
Proceed to interpret the results
Stop the test

Simple Linear Regression



INTERPRETATION OF OUTPUT

Murdoch UNIVERSITY

ANOVA[a]

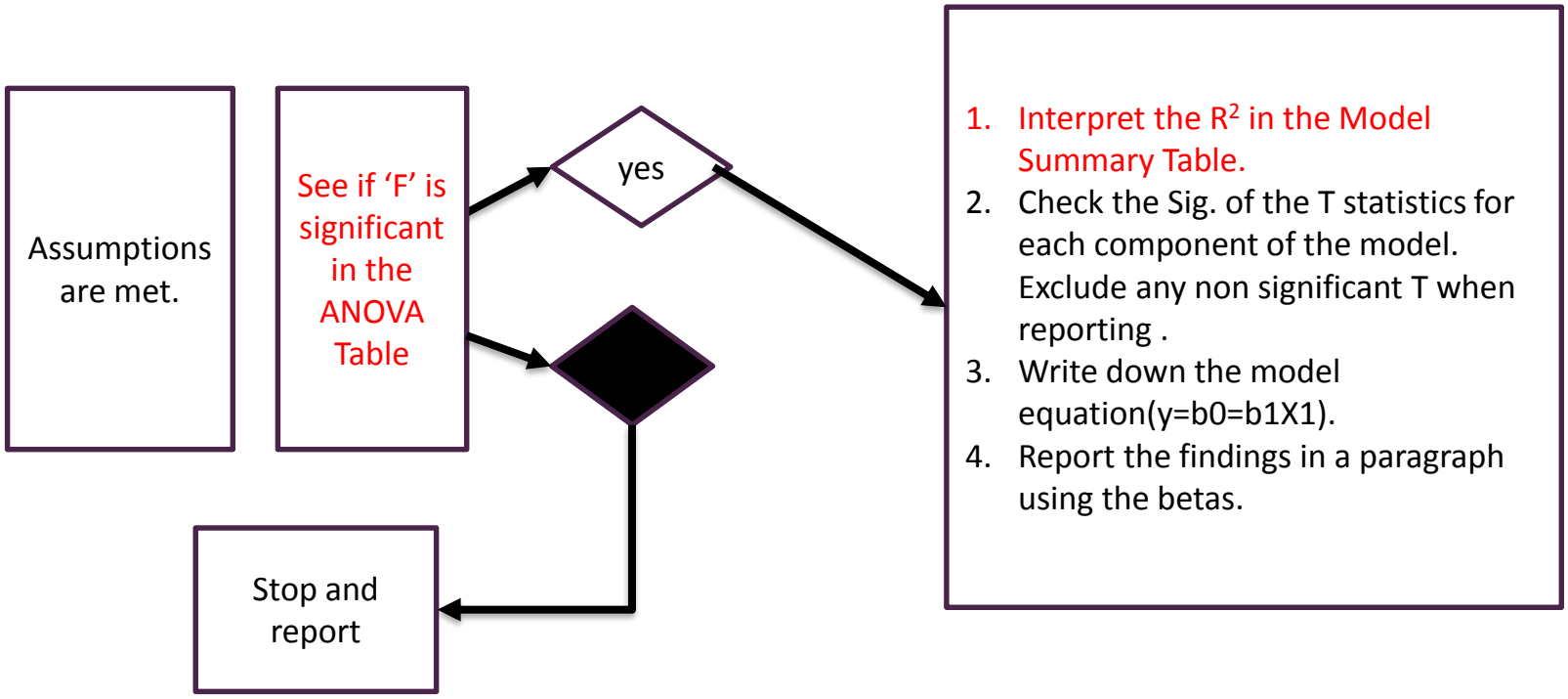
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	106208119.7	1	106208119.7	121.009	.000[b]
	Residual	10532255.24	12	877687.937		
	Total	116740374.9	13			

a. Dependent Variable: Value of monthly sales in $k in 2007

b. Predictors: (Constant), Floor area in sq mts

The ANOVA Table indicates that the regression equation is highly significant with an F = 121.009, p< .05. So in terms of variance explained and significance the regression equation ('model') is excellent. Should F not be significant then the regression as a whole has failed and no more interpretation is necessary.

Simple Linear Regression



INTERPRETATION OF OUTPUT

The Model Summary Table displays 'R' as +0.954 and adjusted R^2 as 0.902 which are very high. 90.2% of the variance in monthly sales value is explained by the variance in floor area. Adjusted R^2 is used as this refers to sample data.

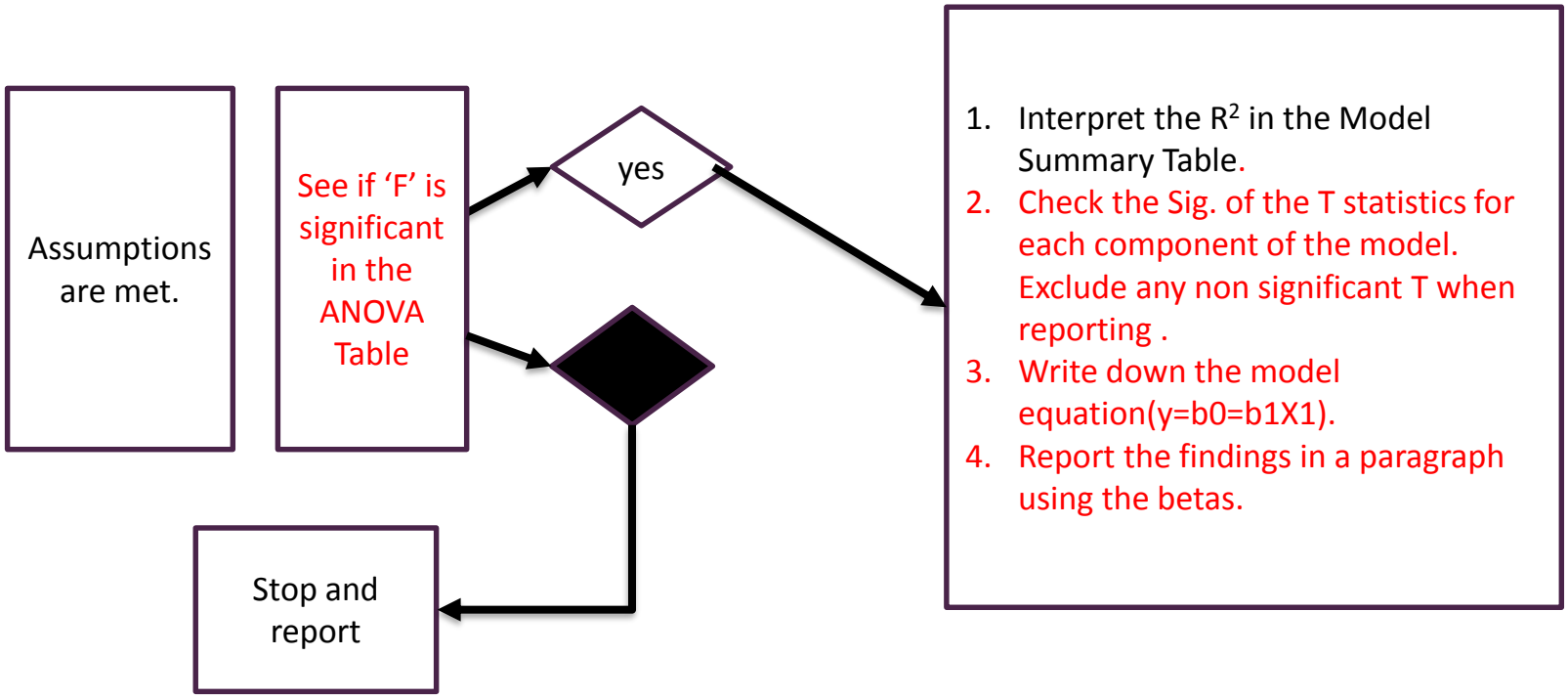
Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.954[a]	.910	.902	936.85001

a. Predictors: (Constant), Floor area in sq mts

b. Dependent Variable: Value of monthly sales in $k in 2007

Simple Linear Regression



Output for Simple Linear Regression

- The coefficients table is crucial and displays the values for constant and beta from which the regression equation can be derived.
- The constant (intercept), b_0 = 901.247, but it is not significant. The unstandardised or raw score regression coefficient or slope (b_1) displayed in SPSS under B as the second line =1.686. The t value for B was significant and implies that this variable (floor area) is a significant predictor.
- Think of B as the change in outcome associated with a unit change in the predictor. This means for every one unit rise (1 sq metre increase in floor space) in B, sales (the outcome) rise by $1686.
- Management can now determine whether the cost of increasing floor area (e.g. building, rental, staffing etc) will bring sufficient returns over a defined time span.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	901.247	513.023		1.757	.104
	Floor area in sq mts	1.686	.153	.954	11.000	.000

a. Dependent Variable: Value of monthly sales in $k in 2007

REGRESSION EQUATION

- The regression equation is:

Y = **b_0** + **b_1** **X**

Y = 0 +1.686 (X).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	901.247	513.023		1.757	.104
	Floor area in sq mts	1.686	.153	.954	11.000	.000

a. Dependent Variable: Value of monthly sales in $k in 2007

- It enables us to predict expected values of Y for any new case of X. For example, we can now ask and answer the question "What is the expected monthly sales for increasing floor area to 7000 sq m?"
- Y = 0 + 1.686(7000) = ?? (calculate it manually).

MULTIPLE REGRESSION

- Multiple Regression

  a technique for estimating the value of the criterion variable (Y) from values on two or more other predictor variables (X's)

- Multiple Correlation (R)

  a measure of the correlation of one dependent variable with a combination of two or more predictor variables.

  Coefficient of Multiple Determination is R^2

MULTIPLE REGRESSION

- So far we have focussed on simple linear regression in which one independent or predictor variable was used to predict the value of a dependent or criterion variable.
- But there can be many other potential predictors that might establish a better or more meaningful prediction. With more than one predictor variable we use multiple regression.
- E.g. the prediction of individual income may depend on a combination of education, job experience, gender, age, etc.
- Multiple regression employs the same rationale as simple regression and the formula is a logical extension of that for linear regression:

$$Y = b_0 + b_1X_1 + b_2 X_2 + b_3 X_3 + \ldots$$ etc

Assumptions of Multiple Regression

Murdoch UNIVERSITY

- Linearity: it is assumed when it comes to our work in this unit.
- Normality of the residuals: To be checked after we run the test as it will be part of the output.
- Homoscedasticity: To be checked after we run the test as it will be part of the output.
- *Multicollinearity:* Very high correlations between IV's should be avoided. To be checked after we run the test as it will be part of the output.

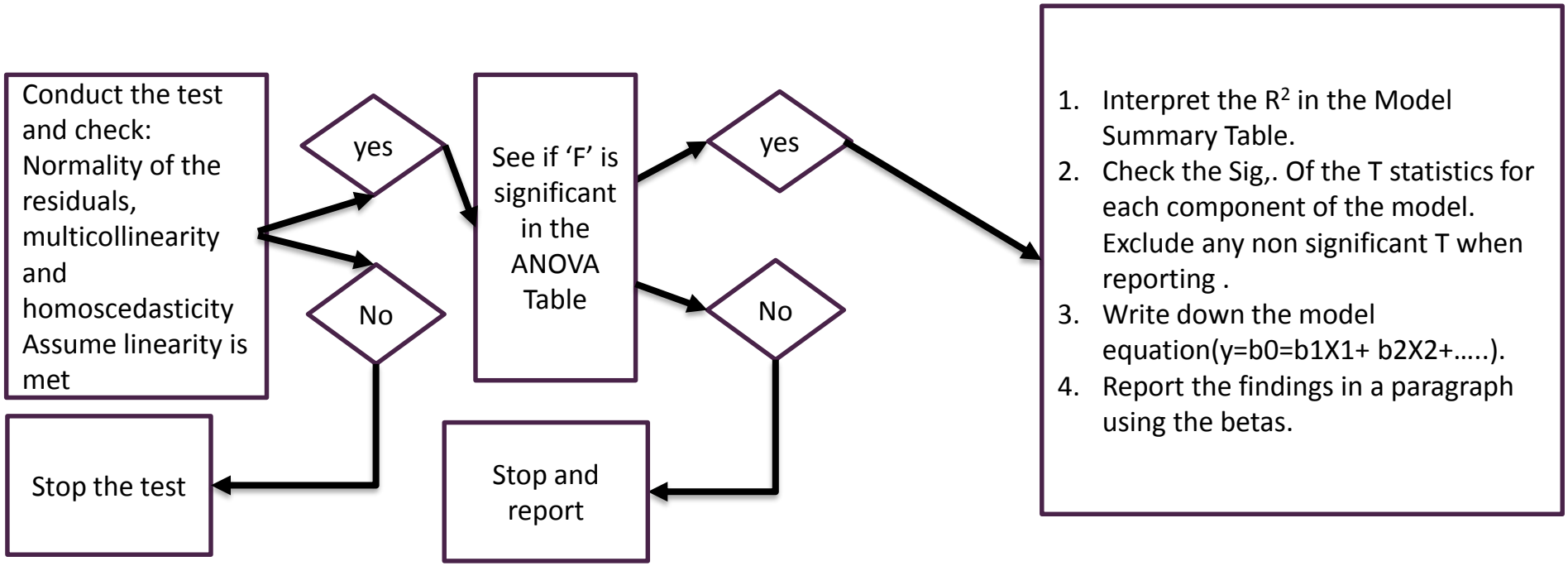
TYPES OF MULTIPLE REGRESSION

Murdoch UNIVERSITY

There are three types of Multiple Regression

1. Standard Multiple regression
2. Hierarchical Multiple Regression
3. Stepwise Multiple Regression

Multiple Linear Regression



Example

- We will be exploring the impact of respondents' perceptions of control on their levels of perceived stress. The literature in this area suggests that if people feel that they are in control (IVs) of their lives, they are less likely to experience 'stress' (DV) (tpstress)
- Control = IVs =
  - the Mastery scale, which measures the degree to which people feel they have control over the events in their lives (tmast)
  - the Perceived Control of Internal States Scale (PCOISS), which measures the degree to which people feel they have control over their internal states
- Example of research questions
  - 1. How well do the two measures of control (mastery, PCOISS) predict perceived stress? How much variance in perceived stress scores can be explained by scores on these two scales?
  - 2. Which is the best predictor of perceived stress: control of external events (Mastery scale), or control of internal states (PCOISS)?

Procedure for standard multiple regression

- 1. From the menu at the top of the screen click on: Analyze, then click on Regression, then on Linear.
- 2. Click on your continuous dependent variable (e.g. total perceived stress: tpstress) and move it into the **Dependent** box.
- 3. Click on your independent variables (total mastery: tmast; total PCOISS: tpcoiss) and move them into the **Independent** box.
- 4. For **Method**, make sure **Enter** is selected (this will give you standard multiple regression).
- 5. Click on the Statistics button:

  - Tick the box marked **Estimates, Confidence Intervals, Model fit, Descriptives, Part and partial correlations and Collinearity diagnostics.**
  - In the **Residuals section** tick the **Casewise diagnostics and Outliers outside 3 standard deviations.**
  - Click on **Continue**.

Procedure for standard multiple regression

Murdoch UNIVERSITY

6. Click on the **Options** button. In the **Missing Values section** click on **Exclude cases pairwise**.

7. Click on the **Plots** button.
   - Click on ***ZRESID** and the arrow button to move this into the Y box.
   - Click on ***ZPRED** and the arrow button to move this into the X box.
   - In the section headed Standardized Residual Plots, tick the **Normal probability plot** option.
   - Click on **Continue**.

8. Click on the Save button.

- • In the section labelled **Distances** tick the **Mahalanobis** box (this will identify multivariate outliers for you) and **Cook's**, then Click on **Continue**.

9. Click **OK**

Check for the assumption of **Multicollinearity**

* Go for the Correlation table.
* Check that your independent variables show at least some relationship with your dependent variable (above 0.3 preferably).
* Also check that the correlation between each of your independent variables is not too high (no more than 0.7).
* In cases with more than (0.7), you need to omit one of the variables.
* Go for table “collinearity diagnostics” .
* A condition index greater than 15 indicates a possible problem
* An index greater than 30 suggests a serious problem with collinearity.

Correlations

		Total perceived stress	Total PCOISS	Total Mastery
Pearson Correlation	Total perceived stress	1.000	-.581	-.612
	Total PCOISS	-.581	1.000	.521
	Total Mastery	-.612	.521	1.000
Sig. (1-tailed)	Total perceived stress	.	.000	.000
	Total PCOISS	.000	.	.000
	Total Mastery	.000	.000	.
N	Total perceived stress	433	426	433
	Total PCOISS	426	430	429
	Total Mastery	433	429	436

Collinearity Diagnostics[a]

Model	Dimension	Eigenvalue	Condition Index	Variance Proportions		
				(Constant)	Total PCOISS	Total Mastery
1	1	2.965	1.000	.00	.00	.00
	2	.019	12.502	.62	.80	.01
	3	.016	13.780	.38	.20	.99

a. Dependent Variable: Total perceived stress

Check for the assumption of **Multicollinearity** >>> Cont.

Murdoch UNIVERSITY

Coefficients[a]

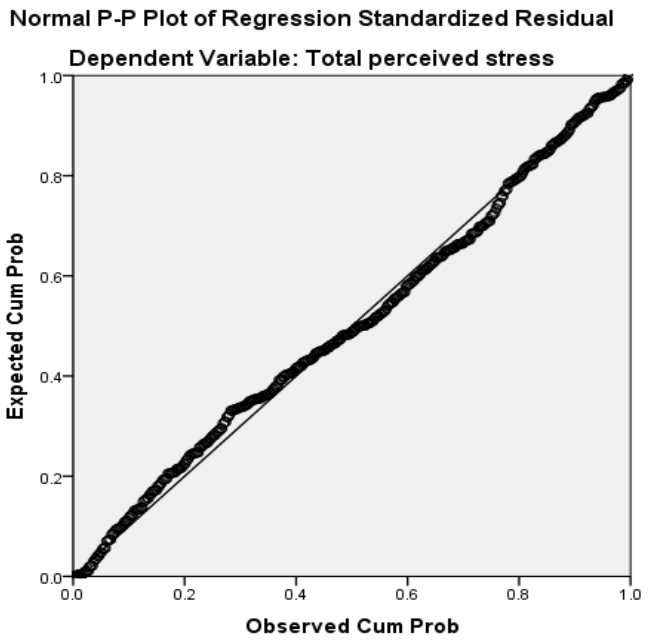
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	50.971	1.273		40.035	.000	48.469	53.474					
	Total PCOISS	-.175	.020	-.360	-8.660	.000	-.215	-.136	-.581	-.388	-.307	.729	1.372
	Total Mastery	-.625	.061	-.424	-10.222	.000	-.745	-.505	-.612	-.445	-.362	.729	1.372

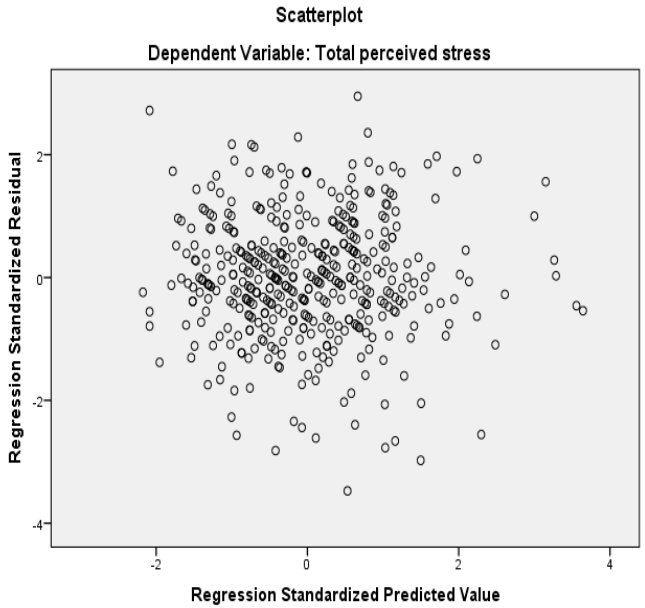
a. Dependent Variable: Total perceived stress

is an indicator of how much of the variability of the specified independent is not explained by the other independent variables in the model and is calculated using the formula 1–R2 for each variable. If this value is very small (less than .10), it indicates that the multiple correlation with other variables is high, suggesting the possibility of multicollinearity.

The other value given is the VIF (Variance inflation factor), which is just the inverse of the Tolerance value (1 divided by Tolerance). VIF values above 10 would be a concern here, indicating multicollinearity.

Check for the assumption of **normality and homoscedasticity**

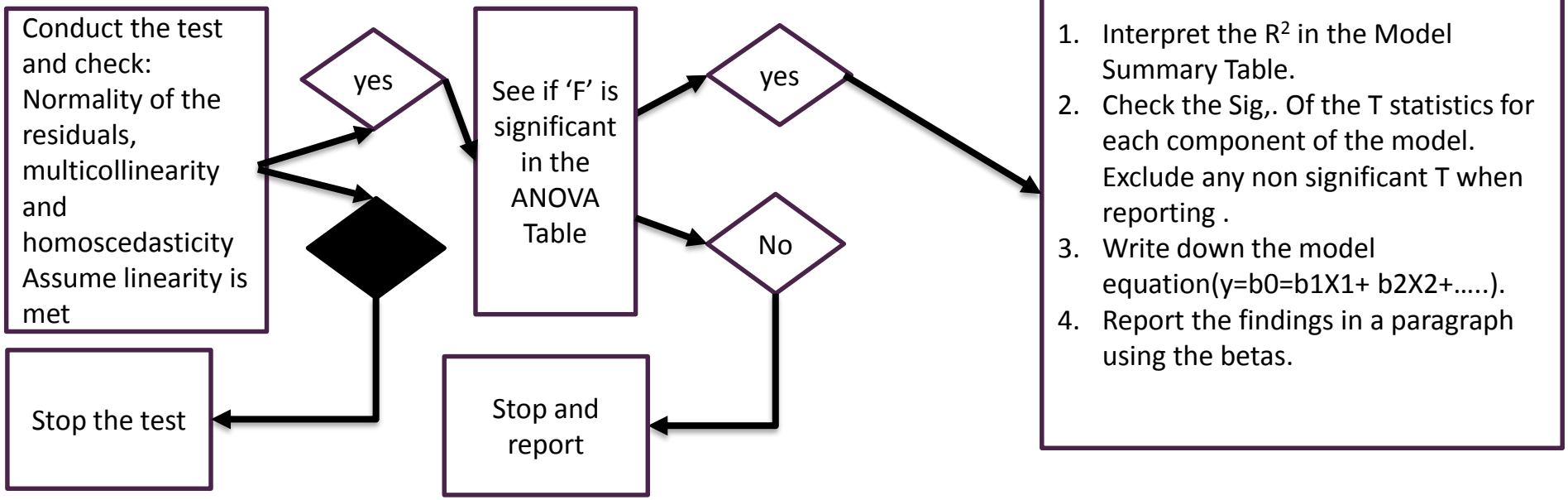




In the Normal Probability Plot you are hoping that your points will lie in a reasonably straight diagonal line from bottom left to top right.

In the Scatterplot of the standardised Residuals you are hoping that the residuals will be roughly rectangularly distributed, with most of the scores concentrated in the centre (along the 0 point).

Multiple Linear Regression



Interpretation of the output

- Evaluating the Model

ANOVA[a]

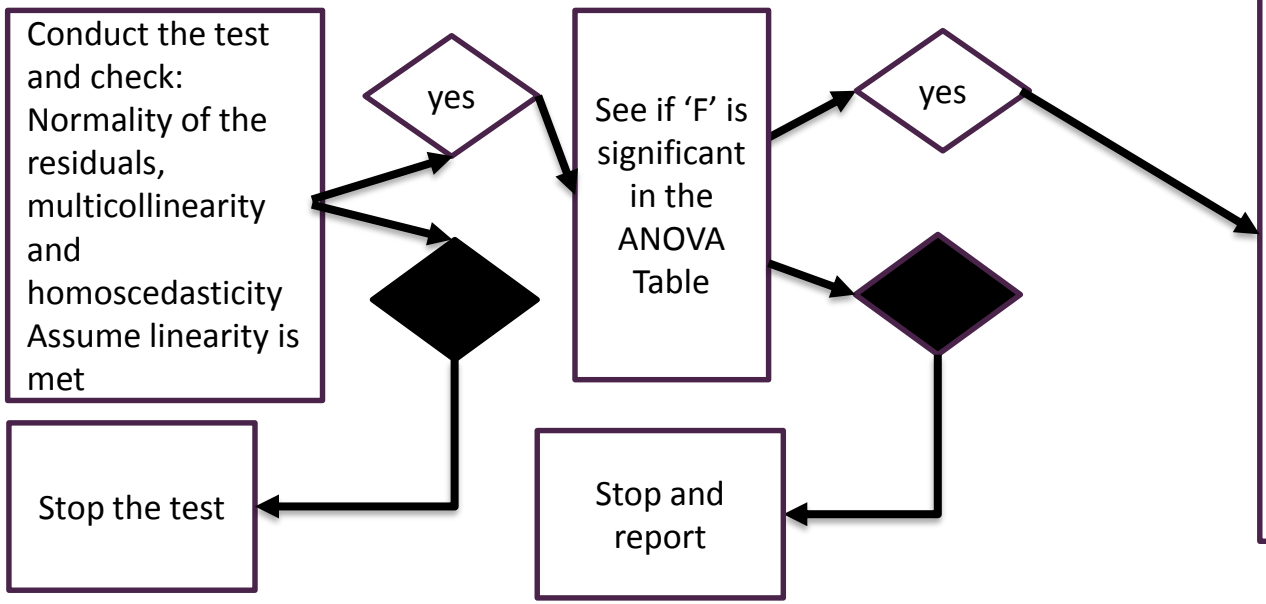
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6806.728	2	3403.364	186.341	.000[b]
	Residual	7725.756	423	18.264		
	Total	14532.484	425			

a. Dependent Variable: Total perceived stress

b. Predictors: (Constant), Total Mastery, Total PCOISS

(Sig = .000, this really means p<.0005).

Multiple Linear Regression



1. Interpret the R^2 in the Model Summary Table.
2. Check the Sig,. Of the T statistics for each component of the model. Exclude any non significant T when reporting .
3. Write down the model equation(y=b0=b1X1+ b2X2+…..).
4. Report the findings in a paragraph using the betas.

Interpretation of the output

- Evaluating the Model

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.684[a]	.468	.466	4.274

a. Predictors: (Constant), Total Mastery, Total PCOISS

b. Dependent Variable: Total perceived stress

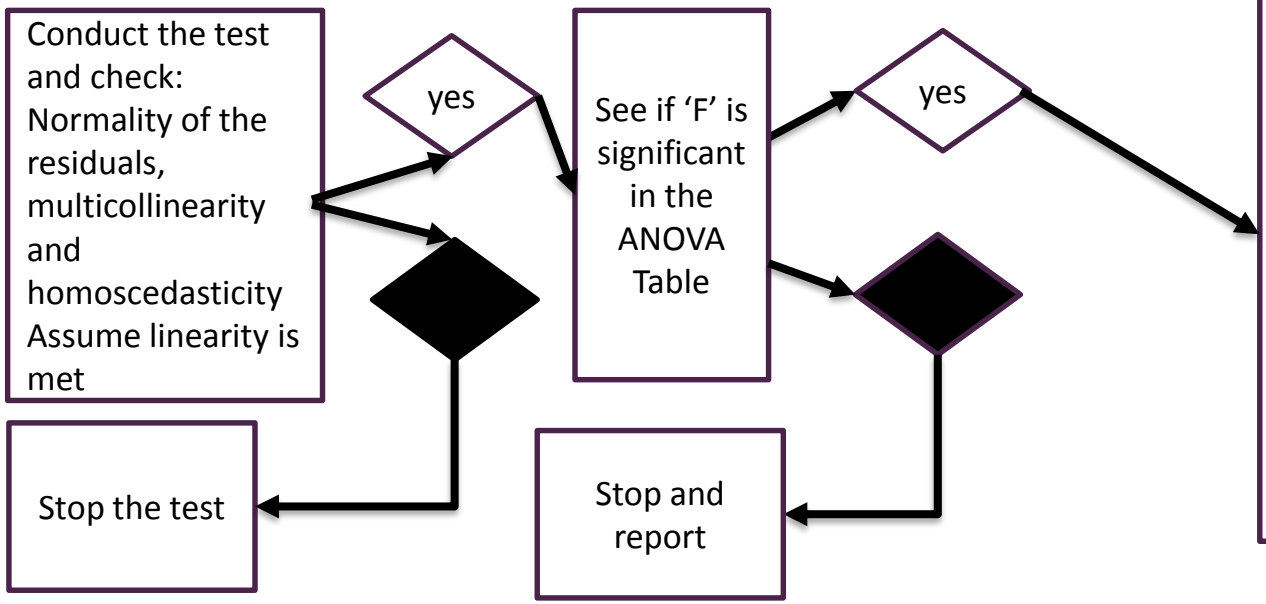
How much of the variance in the DV (stress) is explained by the model.

When a small sample is involved, the R square value in the sample tends to be a rather optimistic overestimation of the true value in the population. The Adjusted R square statistic 'corrects' this value to provide a better estimate of the true population value. If you have a small sample you may wish to consider reporting this value, rather than the normal R Square value.

Research question

1. How well do the two measures of control (mastery, PCOISS) predict perceived stress? How much variance in perceived stress scores can be explained by scores on these two scales?

Multiple Linear Regression



1. Interpret the R^2 in the Model Summary Table.
2. Check the Sig,. Of the T statistics for each component of the model. Exclude any non significant T when reporting .
3. Write down the model equation(y=b0=b1X1+ b2X2+…..).
4. Report the findings in a paragraph using the betas.

Interpretation of the output

- Evaluating each of the independent variables

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	50.971	1.273		40.035	.000	48.469	53.474					
	Total PCOISS	-.175	.020	-.360	-8.660	.000	-.215	-.136	-.581	-.388	-.307	.729	1.372
	Total Mastery	-.625	.061	-.424	-10.222	.000	-.745	-.505	-.612	-.445	-.362	.729	1.372

a. Dependent Variable: Total perceived stress

Research question
2. Which is the best predictor of perceived stress: control of external events (Mastery scale), or control of internal states (PCOISS)?

If you square this value (whatever it is called) you get an indication of the unique contribution of that variable to the total R squared

(Sig = .000, this really means p<.0005) for each IV

Standardized coefficients refer to how many standard deviations a dependent variable will change, per standard deviation increase in the predictor variable. Standardization of the coefficient is usually done to answer the question of which of the independent variables have a greater effect on the dependent variable in a multiple regression analysis, when the variables are measured in different units of measurement (for example, income measured in dollars and family size measured in number of individuals).

Interpretation of the output

- More interpretations

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	50.971	1.273		40.035	.000	48.469	53.474					
	Total PCOISS	-.175	.020	-.360	-8.660	.000	-.215	-.136	-.581	-.388	-.307	.729	1.372
	Total Mastery	-.625	.061	-.424	-10.222	.000	-.745	-.505	-.612	-.445	-.362	.729	1.372

a. Dependent Variable: Total perceived stress

This relationship is in the original units (scores of PCOISS, and scores of Mastery). This is useful for predicting things in the real world, but it is difficult to compare different predictors. Predictors might have large B values just because they are measured on a larger scale (compare minutes to hours in the above example).

Interpreting Estimated Coefficient

tpstress = 50.971 + **(**-.175 × tPCOISS) + (-.625 × t Mastery)

This is the model of prediction of the tpstress by the two variables.

Reporting the results (APA style)

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6806.728	2	3403.364	186.341	.000[b]
	Residual	7725.756	423	18.264		
	Total	14532.484	425			

a. Dependent Variable: Total perceived stress

b. Predictors: (Constant), Total Mastery, Total PCOISS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.684[a]	.468	.466	4.274

a. Predictors: (Constant), Total Mastery, Total PCOISS

b. Dependent Variable: Total perceived stress

Multiple regression analysis was used to test if the two measures of control (mastery, PCOISS) significantly predicted the perceived stress predicted. The results of the regression indicated the two predictors explained 46.8% of the variance (R^2=.468, $F(2,423)=186.341$, $p<.05$). It was found that Mastery significantly predicted total perceived stress (β= -.424, p<.05), as did PCOISS (β= -.36, p<.05).

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B		Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Lower Bound	Upper Bound	Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	50.971	1.273		40.035	.000	48.469	53.474					
	Total PCOISS	-.175	.020	-.360	-8.660	.000	-.215	-.136	-.581	-.388	-.307	.729	1.372
	Total Mastery	-.625	.061	-.424	-10.222	.000	-.745	-.505	-.612	-.445	-.362	.729	1.372

a. Dependent Variable: Total perceived stress