

# **MATU9D2 : PRACTICAL STATISTICS**

**Spring 2017**

## **PRACTICAL 9**

In this Practical you will learn to

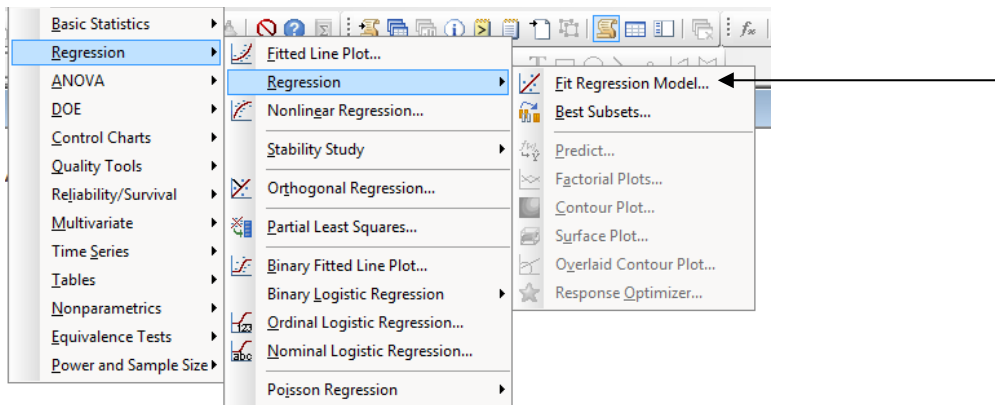
1. Perform multiple linear regression
2. Split data into a Training Set and Test Set
3. Find the Best Model using Forward Stepping, Backward Stepping and Best Subset Regression.
4. Validate the model.

**THERE ARE TWO SECTIONS IN THESE NOTES:**

1. **INSTRUCTIONS ON HOW TO PERFORM TASKS USING MINITAB.**
2. **A LIST OF EXERCISES TO DO USING THE ABOVE COMMANDS**

Using Minitab to examine relationships between Quantitative variables –

Having plotted the data and used Simple Linear Regression to examine the relationship in Practical 8, we will use the following menus in this Practical to use Multiple Linear Regression.



## PART 1 : MULTIPLE LINEAR REGRESSION

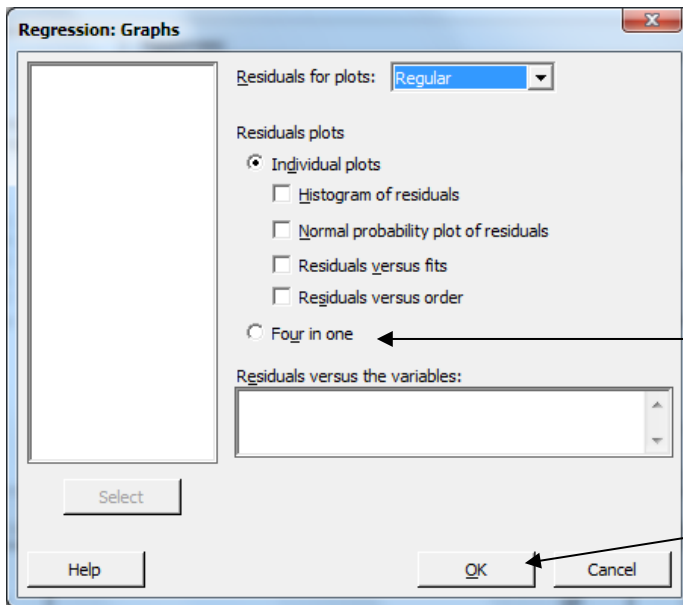
In many cases we will not only be interested in the association between variables but we will want to define the relationship. This will allow us to use the model to predict future values.

You should always validate the assumptions of normality and constant variance using the Residual Plots.

Select the Stat menu then Regression then Fit Regression Model (as above)  
Then the following dialogue box will appear

The screenshot shows the 'Regression' dialog box in Minitab. On the left is a list of variables: C1 Age, C2 Years, C3 Weight, C4 Height, C5 Chin, C6 Forearm, C7 Calf, C8 Pulse, C9 Systol, and C10 Diastol. The 'Responses' field contains 'Systol'. The 'Continuous predictors' field contains 'Age-Pulse'. The 'Categorical predictors' field is empty. At the bottom are buttons for 'Model...', 'Options...', 'Coding...', 'Stepwise...', 'Graphs...', 'Results...', 'Storage...', 'Select', 'Help', 'OK', and 'Cancel'. Four numbered annotations with arrows point to specific elements: 1. 'Select the Y' points to 'Systol' in the Responses field. 2. 'Select the X variable Or X variables (if more than one)' points to 'Age-Pulse' in the Continuous predictors field. 3. 'Always draw Residual plots (see first dialogue box on next page)' points to the 'Graphs...' button. 4. 'Click OK' points to the 'OK' button.

1. Select the Y
2. Select the X variable  
**Or X variables**  
(if more than one)
3. Always draw Residual plots (see first dialogue box on next page)
4. Click OK

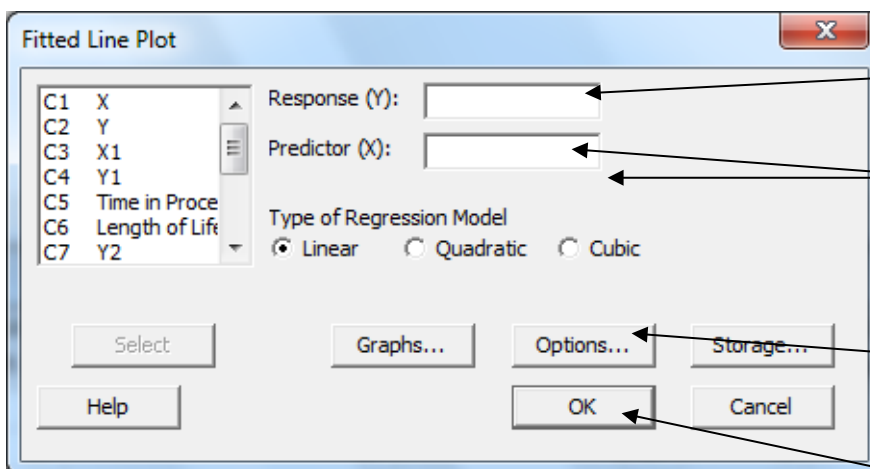


3.1 Select Four in One

3.2 Click OK

## Graph including Fitted Line

Select the Stat menu then Regression then Fitted Line Plot – the following dialogue box will appear:



1. Select Y variable

2. Select X variable

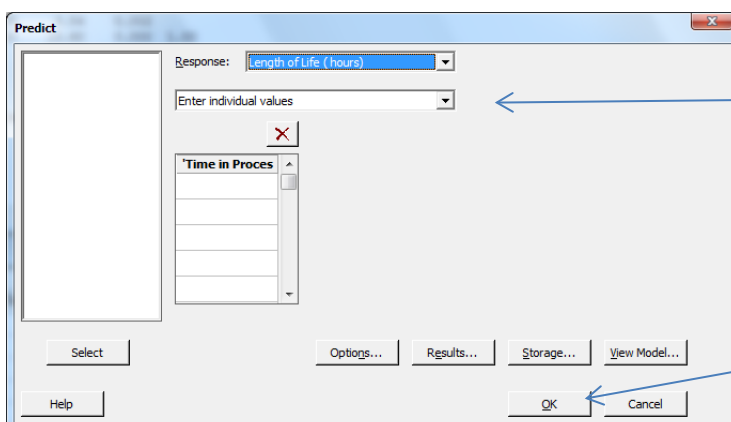
3. Select if require CI's or PI's

4. Click OK

## Prediction using Regression Model

Firstly, 'do the regression' then

Select the Stat menu then Regression then Predict then enter the individual values or column of values for which you want a prediction.



1. Choose either individual values or column of

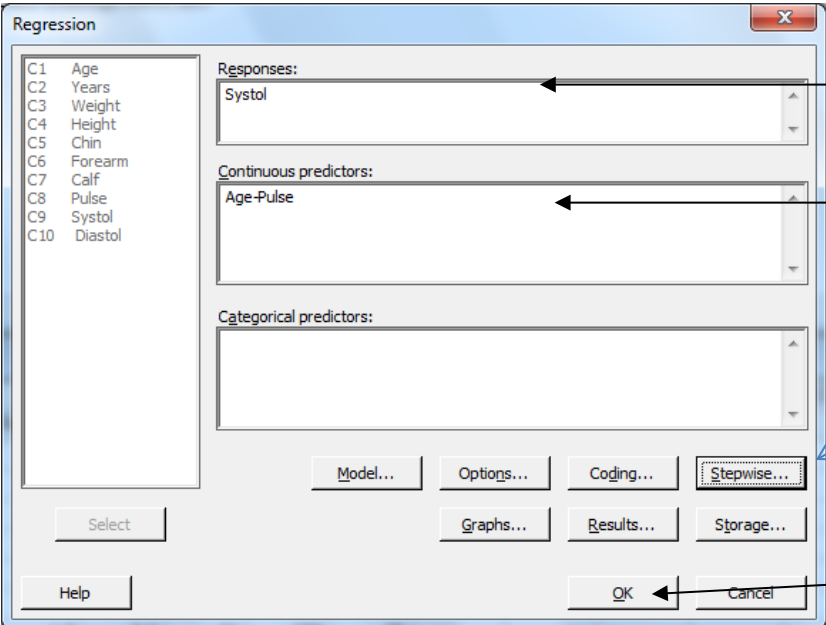
2. Click OK

## PART 2 : FINDING THE 'BEST' MODEL

### Technique 1. Stepwise Regression

Select the Stat menu then Regression, Fit Regression Model, then choose Stepwise – the following dialogue box will appear:

If you run the procedure as below (no changes to the default settings), Minitab performs Forward Stepping then when all X variables that can be included are included then performs Backward Stepping in case any can be omitted. - DO THIS!!

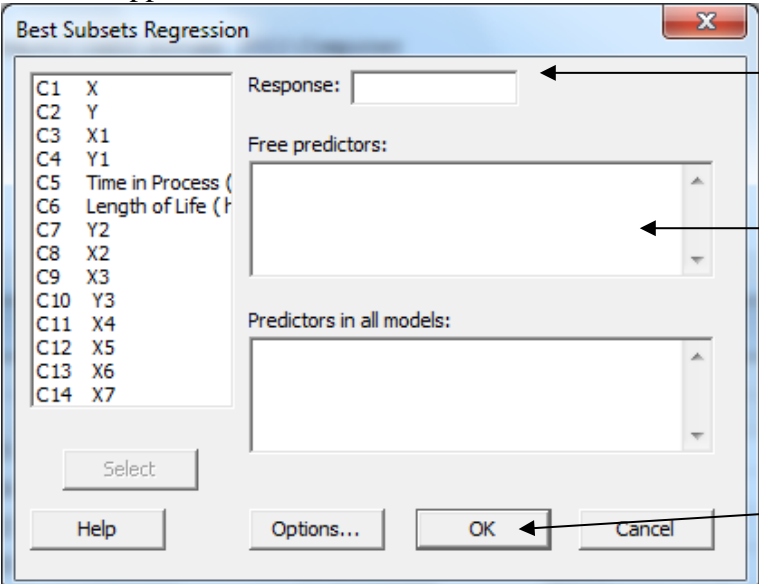


The image shows the Minitab Regression dialog box. On the left is a list of variables: C1 Age, C2 Years, C3 Weight, C4 Height, C5 Chin, C6 Forearm, C7 Calf, C8 Pulse, C9 Systol, C10 Diastol. The 'Responses:' field contains 'Systol'. The 'Continuous predictors:' field contains 'Age-Pulse'. The 'Categorical predictors:' field is empty. At the bottom are buttons for 'Model...', 'Options...', 'Coding...', 'Stepwise...', 'Graphs...', 'Results...', 'Storage...', 'OK', 'Cancel', 'Help', and a 'Select' button. Four numbered callouts point to specific elements: 1. 'Select the Y variable' points to the 'Responses:' field. 2. 'Select the possible X variables' points to the 'Continuous predictors:' field. 3. 'Click Stepwise Choose Method Stepwise or Backward or Forward' points to the 'Stepwise...' button. 4. 'Click OK' points to the 'OK' button.

1. Select the Y variable
2. Select the possible X variables
3. Click Stepwise Choose Method Stepwise or Backward or Forward
4. Click OK

### Technique 2. Best Subset Regression

Select the Stat menu then Regression then Regression then Best Subsets.... – the following dialogue box will appear:



The image shows the Minitab Best Subsets Regression dialog box. On the left is a list of variables: C1 X, C2 Y, C3 X1, C4 Y1, C5 Time in Process (, C6 Length of Life (, C7 Y2, C8 X2, C9 X3, C10 Y3, C11 X4, C12 X5, C13 X6, C14 X7. The 'Response:' field is empty. The 'Free predictors:' field is empty. The 'Predictors in all models:' field is empty. At the bottom are buttons for 'Select', 'Help', 'Options...', 'OK', and 'Cancel'. Three numbered callouts point to specific elements: 1. 'Select the Y variable' points to the 'Response:' field. 2. 'Select the possible X variables' points to the 'Free predictors:' field. 3. 'Click OK' points to the 'OK' button.

1. Select the Y variable
2. Select the possible X variables
3. Click OK

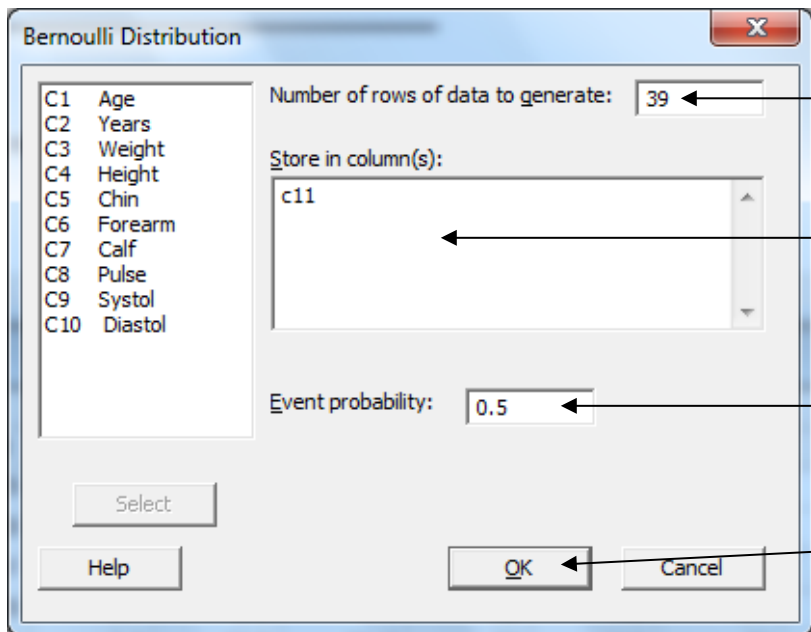
## PART 3 : SPLITTING THE DATA INTO A TRAINING SET & TEST SET

### Step 1 Divide the Data into a Training Set & Test Set

- Generate Random Numbers

The following will generate a column with 0 and 1 like tossing a coin would produce heads and tails!!

Choose the Calc Menu → Random Data → Bernoulli Distribution



1. Number of rows needed  
e.g. 39 for Peru data  
as 39 rows of data

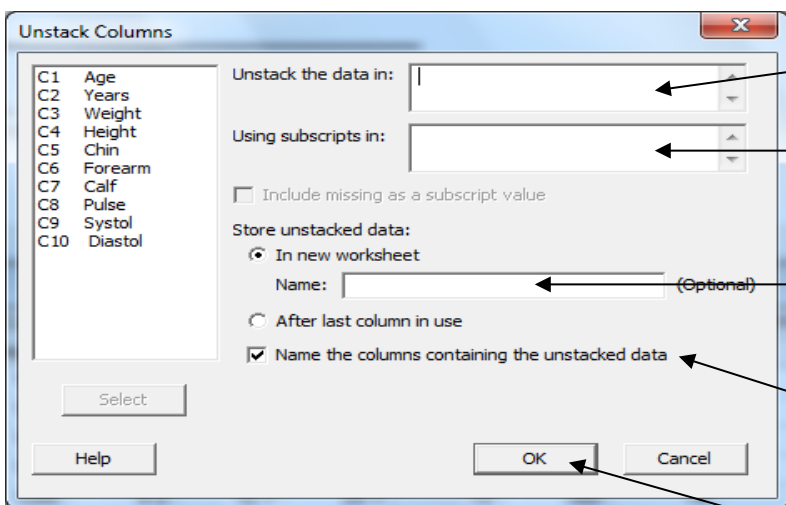
2. Enter Empty column  
for this NEW data

3. 0.5 to split the data  
roughly in half

4. Click OK

- Split the data into two – a Training Set and a Test Set  
This method automatically splits the data into 2 groups of columns using the 0/1 data column (in new worksheet if used as below)

Choose Data Menu then Unstack the following dialogue box will appear:



1. Select columns  
containing the data

2. Select column with  
0/1.

3. New worksheet to  
keep separate from  
original

4. Name columns -  
useful

5. Click OK

**Step 2.** Use Stepwise Regression or Best Subset Regression to find best model for Training Set (1's)

**Step 3.** Check the assumptions and find the equation for the Best Model using Regression  
Note down the equation i.e. X variables and coefficients (intercept and slopes).

**Step 4.** Validate this model on the Test Set (0's)

See how good this model is by comparing the predictions for the rest of the data.

Use the best model from the Training Set with the Test Set data to predict SBP then compare by looking at differences between Observed and Predicted.

1. Use the Calculator to calculate the predicted values
2. Use the Calculator to calculate the differences between the Observed and Predicted
3. Calculate the Descriptive Statistics for the differences
4. Draw an appropriate plot of the differences
5. Perform a paired t test on the Observed and Predicted to see whether there is a statistically significant difference.
6. Relate the descriptive and t test results to discuss whether we have a 'good' model.

## EXERCISES

1. The data is in a Minitab project file on Succeed under Practicals and the file is called 'Practical 8 MLR data'.

Y	10	12	15	17	19	22	24	27	29	30
X <sub>2</sub>	1	1	2	2	3	4	4	5	5	6
X <sub>3</sub>	10	9	8	7	6	5	4	3	2	1

- (a) Use the data given above to find the regression relationship between Y and X<sub>2</sub>.
- (b) Use the data given above to find the regression relationship between Y and X<sub>3</sub>.
- (c) Use the data given above to find the regression relationship between Y and X<sub>2</sub> and X<sub>3</sub>. (Put both X<sub>2</sub> and X<sub>3</sub> in the 'Predictors' box)
- (d) From your answer to part (c) test each of the coefficients to find if they are non-zero.
- (e) Discuss your results.

## 2. Finding the 'Best Model'

We want to find the best model to predict an 'Aroma' score for wine. The data set contains 37 Pinot Noir wine samples, each described by 17 elemental concentrations (Cd, Mo, Mn, Ni, Cu, Al, Ba, Cr, Sr, Pb, B, Mg, Si, Na, Ca, P and K) and a score on the wine's aroma from a panel of judges.

The data is in a Minitab project file on Succeed under Practicals and the file is called 'Practical 8 Wine data'.

- (i) Divide the data into a Training Set and a Test Set
- (ii) Use Stepwise Regression and the Training Set to develop the 'Best Model' for Aroma
- (iii) Validate the assumptions for this 'Best Model' using Regression and Residual Plots
- (iv) Note the equation for this Best Model for Aroma
- (v) Go to the Test Set. Use Calc → Calculator and the model in (iv) to calculate the Predicted Aroma
- (vi) Draw appropriate plots to subjectively see if the model is satisfactory.
- (vii) Perform an appropriate test to answer whether the model is satisfactory, on average.
- (viii) Repeat (ii) – (vii) above for Forward Stepping, Backward Stepping and Best Subset Regression.
- (ix) Compare your answers in (viii)