**Investigating Bivariate Measurement Data using iNZight**

**Statistics Teachers Day**

**22 November 2012**

**Ross Parsonage**

**Basic principles**

1. Each component of the cycle is communicated
2. Use context
3. Refer to visual aspects

**Commenting on features**

* **Trend**

Refer to the graph

Linear or non-linear

Use descriptions of variables

* **Association (Nature)**

Refer to the graph

Use descriptions of variables

Use a contextual description

Use “tends to”

Can use terms such as positive, negative association

* **Finding a model**

Say why they are fitting this model

Discuss appropriateness

* **Strength**

Must refer to visual aspects (degree of scatter or closeness of points to the fitted model)

Use terms such as strong, moderate or weak

If linear, could refer to the correlation coefficient r, but not at the expense of visual aspects

* **Unusual points or other features**

Refer to the graph

Refer to data points where appropriate

**Higher level considerations**

* Justify
* Extend
* Reflect

Posing an appropriate relationship question using a given multivariate data set

* Consider several sensible pairs of variables
* Reflect on pairs of variables before deciding on those to investigate
* Extend the investigation to at least two questions

Selecting and using appropriate displays

* Justify placement of variables on axes

Identifying features in data (includes describing the nature and strength of the relationship and relating this to the context)

* Consider contextual reasons for features
* Discuss relevance to a wider population
* The existence of a statistical relationship does not necessarily imply causation
* Acknowledge that other factors (which should be specifically identified) could influence the response variable

Finding an appropriate model

* Consider alternative models, if appropriate
* Consider improving the model by removal of outliers (as long as this is justifiable) and repeating the analysis
* Analyse separate subsets
* Take account of the number of data points

Using the model to make a prediction

* Justify choice of the variable to use for predictions (This can be done if different investigations use the same response variable but different explanatory variables)
* Justify the value of the *x*-value used
* Discuss relevance to a wider population
* Discuss precision of predictions

General

* State any assumptions and discuss the effect on the validity of the analysis

**Other issues I have been pondering**

* The use or articles or reports to assist contextual understanding
* The effect of outliers on a model
* The process used for fitting a line to data that has a linear trend
* Residuals and residual plots
* Transforming variables

**Where to get iNZightVIT?**

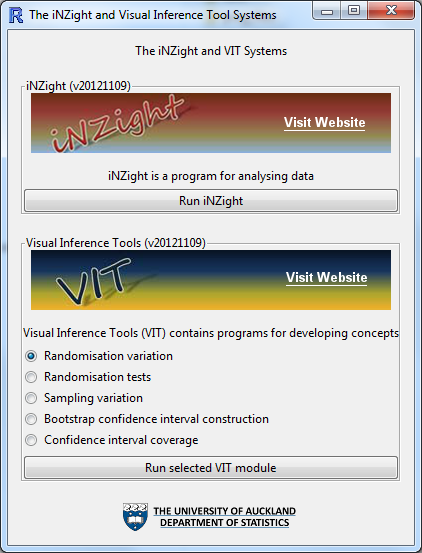
[**http://www.stat.auckland.ac.nz/~wild/iNZight/dlw.html**](http://www.stat.auckland.ac.nz/~wild/iNZight/dlw.html)

**Using iNZight to start a bivariate measurement data analysis**

Open the iNZight folder

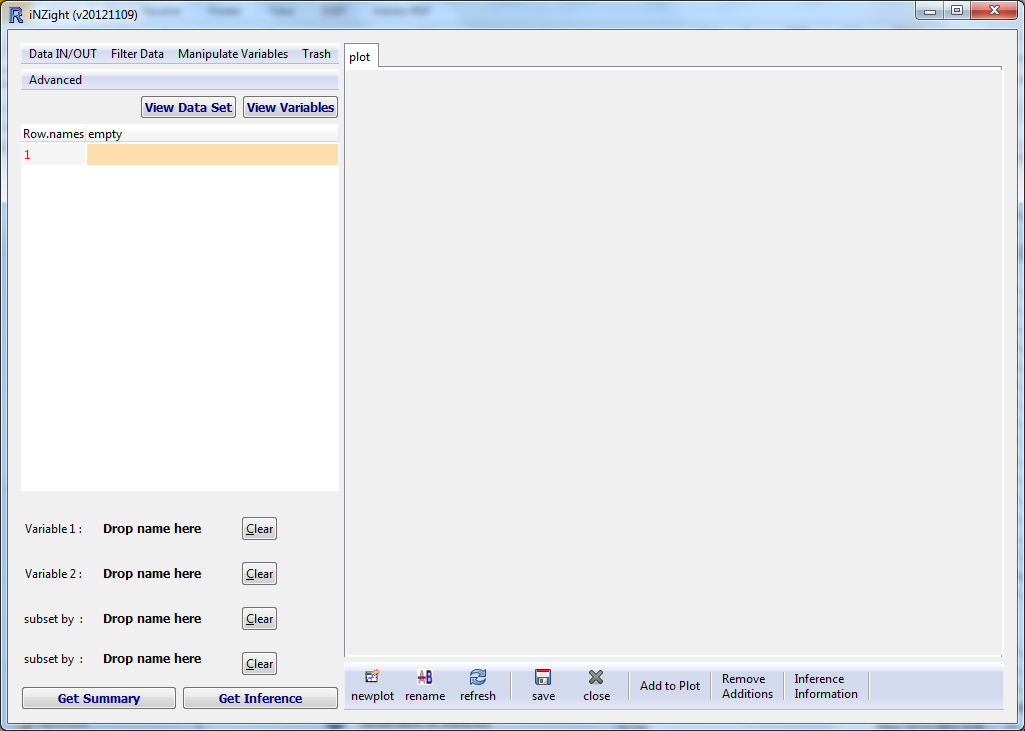
Select the START-iNZightVIT.bat (You may get an Open File – Security Warning. If you do, select Run)

After a short time this will appear:



Click on Run iNZight

This will appear:



Click on Data IN/OUT

From the drop-down menu, select Import Data

In the File Browser box, click on the ‘browse’ button to find the csv file to import

Click OK

The data set now appears.

Drag the variable name of the explanatory variable to Variable 1.

Drag the variable name for the response variable to Variable 2.

The scatter plot will appear in the display screen.

**Data sources**

**CensusAtSchool NZ**

<http://www.censusatschool.org.nz/resources/data-analysis-tools/>

**StatisticsNZ SURF**

<http://www.stats.govt.nz/searchresults.aspx?q=SURF>

These may not have enough quantitative variables

**OzDASL – Australasian Data and Story Library**

<http://www.statsci.org/data/>

The Multiple Regression and Multiple Regression with Factors datasets are probably more useful than the First Course in Statistics

<http://www.statsci.org/data/multiple.html>

**DASL**

<http://lib.stat.cmu.edu/DASL/>

**ConnectMV**

<http://datasets.connectmv.com/>

Some are simulated data

**StatLib JASA Data Archive (17 data sets)**

<http://lib.stat.cmu.edu/modules.php?op=modload&name=PostWrap&file=index&page=jasadata/>

**StatLib Dataset Archive (111 data sets)**

<http://lib.stat.cmu.edu/modules.php?op=modload&name=PostWrap&file=index&page=datasets/>

**StatLib Data Expo Archive (9 data sets)**

<http://lib.stat.cmu.edu/modules.php?op=modload&name=Downloads&file=index&req=viewsdownload&sid=30>

<http://www.mvstats.com/Resources/page3_datasets_greatideas.htm>

Has links to 8 sources of statistical datasets

Some links are broken

There are no data sets in the UCLA Case Studies

The UCI Machine Learning Repository looks worthwhile

<http://www.models.life.ku.dk/datasets>

Quality and technology based from Department of Food Science, University of Copenhagen

Many are in Matlab format

**From the Department of Statistics, University of Munich**

<http://www.stat.uni-muenchen.de/service/datenarchiv/welcome_e.html>

**NationMaster**

<http://www.nationmaster.com/index.php>

Huge database of statistics from many countries