

Table of Contents

Practical 1 Exploratory stage	3
Practical 1 Outline Solution and Genstat Output	6
Practical 2 Simple linear regression: fitting straight lines and curves.....	9
Practical 2 Outline Solution and GenStat Output.....	13
Practical 3 Model Checking.....	18
Practical 3 Outline Solution and Genstat Output	23
Practical 4 Comparison of regression lines	29
Practical 4 Outline Solution and Genstat output	34
Practical 5 Multiple Linear Regression	39
Practical 5 Outline Solution and Genstat Output	43
Practical 6 Case study: use of the logarithmic transformation	53
Practical 6 Outline Solution and Genstat Output	58
Practical 7 Modelling Curved Relationships	64
Practical 7 Outline Solution and Genstat Output	69
Practical 8 Issues in Linear Regression	74
Practical 8 Outline Solution and Genstat Output	79



Practical 1 Exploratory Stage



Practical 1

Exploratory stage

GENSTAT is a powerful statistical software package that provides a wide range of data analysis capabilities. It is particularly strong on the analysis of designed experiments and on regression analysis.

It has a large number of commands called 'directives' that can either be accessed via the menu system or they can be typed into, and run from the Input window.

Load GENSTAT by double-clicking on the GENSTAT icon. In GENSTAT you view a file or work on it in a spreadsheet. Three windows will open automatically on starting; these are:

- Input Log - records all the commands that have been executed
- Output - for displaying output (except high resolution graphics)
- Event log - for displaying error messages.

A fourth window becomes active only when you produce a plot and each new plot is appended to the same active **Graphics** window.

Importing Data from Excel

Genstat can easily import the contents of Excel workbooks one sheet at a time.

For example, click on the **Excel Import Wizard** icon on the toolbar and select **EDA.xls** stored on the **C:\user** folder. Select the **Single predictor** sheet and click Finish then OK. This action will import the data into a spreadsheet window that inherits the name 'EDA.xls'.



Exploratory Data Analysis

Aim: To gain familiarity with visual exploratory tools in order to choose a suitable model and to propose a type of summary equation.

I. Straight lines and curves with a single predictor

Plot y versus x in the sequence indicated and propose a summary equation for the empirical relationship that you observe in the graphs. For this purpose, use the sequence **Graphics > 2D Scatter plot**, and fill in the resulting dialog box as appropriate to produce one plot at a time.

Also indicate if the equation you propose is likely to be a good summary and why.

y	x	name of model	equation	good summary?
hours	lot_size			
steroid_level	female_age			
gas_yield	device_pressure			

NOTE: Each new graph is shown in a new page of a single Graphics window. You can cycle between pages by pressing the icons with the light blue arrows.

Now draw the line of best 'fit' across the points on all plots following the procedure below.

Try the menu sequence **Stats > Regression analysis > Linear Models**. Select the first pair of response and explanatory variable and click **Run**. Next click the **Further Output** and the **Fitted model** buttons in succession. In the **Graphics** window the straight line model appears superimposed on the observed datapoints.

Repeat for all plots. Does the software give any warnings saying your model is not appropriate, especially in the case of **gas_yield** plotted against **device_pressure**?

How plausible is the slope of this line?

Now clear all data and output by using **Run > Restart Session**.



IF YOU HAVE TIME, TRY THE FOLLOWING:

2. A mix of numerical and qualitative predictors

Now use the data stored in the sheet **Variate and factor** in the workbook **EDA.xls**.

Import the dataset by clicking again the **Excel Import Wizard** icon on the menu toolbar.

Make sure to convert **location** to a **factor** column.

Now plot **time_0** vs **lotsize** by **location**, i.e. using a different symbol for each group.

- Select **Graphics > 2D Scatter plot**
- Select **time_0** on the y-axis, **lotsize** on the x-axis and **location** as **Grouping factor**.
- Click **OK**.

Does it look like **time_0** depends on **lotsize**?

Does **location** make a difference to the relationship?

Propose a plausible summary model, bearing in mind that **lotsize** is a continuous variable (a variate) and **location** is a discrete variable (a factor). How would you write it down?

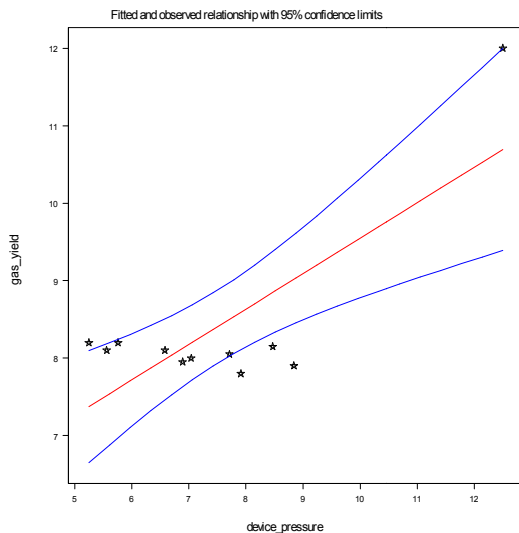
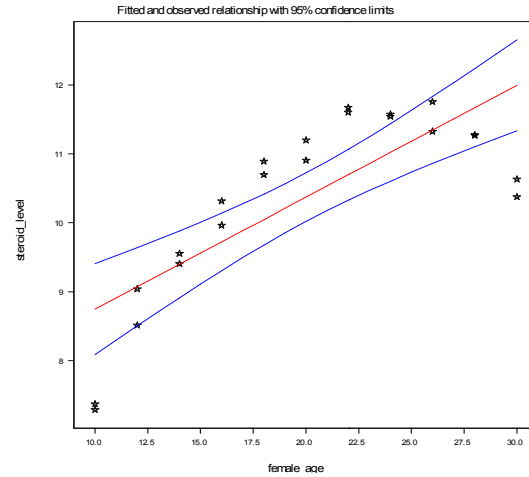
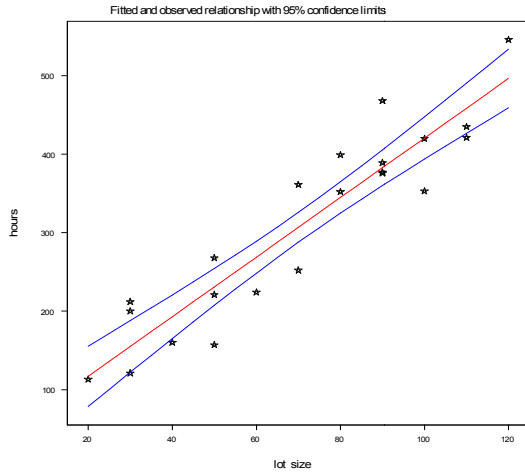
Hint: This looks like a comparison of regressions with a common slope but different intercepts.



Practical 1

Outline Solution and Genstat Output

1. Straight lines and curves with a single predictor



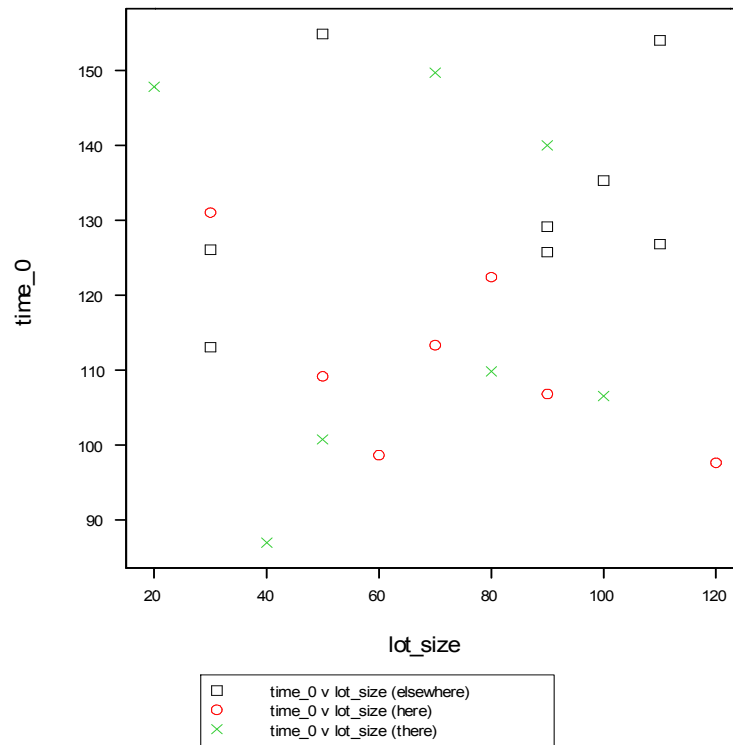
← this fitted straight line is not an adequate summary model

Y	X	name of model	equation	good summary?
hours	Lot_size	Straight line regression	$y = \beta_0 + \beta_1 x$	yes
steroid_level	female_age	Quadratic curve regression	$y = \beta_0 + \beta_1 x + \beta_2 x^2$	yes
gas_yield	device_pressure	Horizontal line when omitting the largest y value	$y = \beta_0$	Cannot say, need to re-plot on new scale

The statistical package gives no warning about the fitting of obviously inadequate summary models.



2. A mix of numerical and qualitative predictors



Both **lotsize** (x) and **location** (f) seem to be associated with **time_0** (y).

The summary model proposed for this trend is that of a set of 3 parallel straight lines:

$$y = \beta_0 + \beta_1 x + f_j$$

Where $j=1,2,3$