

# Fitting a Linear Least Squares Model

## 1 What is a linear least squares model?

Many problems in analyzing data involve describing how variables are related. The simplest of all models describing the relationship between two variables is a linear, or straight-line, model. The simplest method of fitting a linear model is to “eye-ball” a line through the data on a plot, but a more elegant, and conventional method is that of least squares, which finds the line minimizing the sum of distances between observed points and the fitted line.

## 2 Objectives

The lesson is designed to illustrate aspects of linear model fitting by least squares. By the end of the lesson students should:

- Realize that fitting the “best” line by eye is difficult, especially when there is a lot of residual variability in the data.
- Know that there is a simple connection between the numerical coefficients in the regression equation and the slope and intercept of regression line.
- Know that a single summary statistic like a correlation coefficient or  $r^2$  does not tell the whole story. A scatterplot is an essential complement to examining the relationship between the two variables.

## 3 Startup Instructions

The startup instructions are given using the confidence interval module as the example. Substitute the other module names to run another.

On a Unix workstation

```
% ls_module
```

On a PC

- Click on the Lisp-Stat icon in the program manager window
- Click on the `ls.lsp` icon in the Lisp-Stat window

On a Macintosh

- Start up `xlispstat`, by clicking on the `XLispStat` icon
- Pull down the `File` menu and select `Load`
- Select the folder `Teach`
- Select `ls.lsp`

## 4 The module interface

The module allows the user to change the slope and intercept of a line using sliders to see how this affects the way the line is drawn. The user can try to “eye-ball” the best line for several data sets, and then compare the results with the line generated by the Least Squares method. Each data set has different features which can affect the ease of eye-balling the best

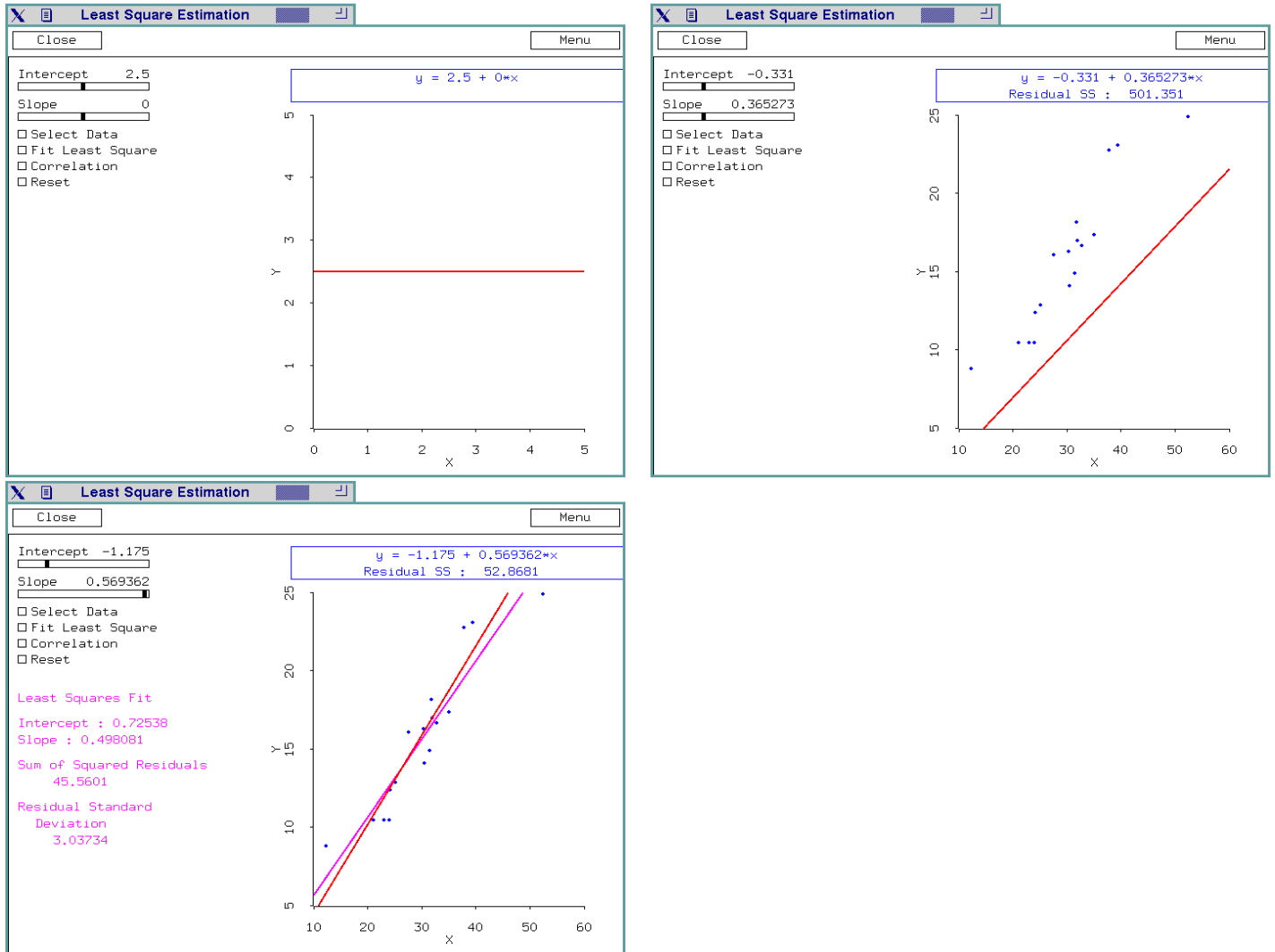


Figure 1: *Least Squares Teaching Module (a) On startup, (b) With Snake River data, (c) including fitted least squares line.*

line. The SNAKE-RIVER data ( $X$  =Water-content,  $Y$  =Water-yield) and the OAK-SEEDING data ( $X$  =Degree-hours,  $Y$  =Shoot-elongation) both small have little variability, that is high correlation. The TENSILE-STRENGTH data ( $X$  =Hardness,  $Y$  =Tensile-strength) contains a lot of variability. The PETROCHEMICAL-PLANT data ( $X$  =Plant Size,  $Y$  =Weight of particle) contains one large value. The CRAB data ( $X$  =Weight,  $Y$  =Length) contains some curvature.

## 5 Warmups

To gain some familiarity with the module, try the following:

1. Select a data set by clicking and holding on **Select Data** square, and dragging the cursor down.
2. Shift the red line around by using the scrollbars (by clicking to the right or left of black bar, or dragging it along with the mouse) to change the intercept and slope parameters.
3. Generating the Least Squares fitted line by clicking on **Fit Least Square**.
4. Compute the correlation between  $X$  and  $Y$  by clicking on **Correlation**.

## 6 Exercises

1. Snake River data
  - (a) Select the **SNAKE-RIVER** data. (You will see a window like Figure 1(b).)
  - (b) Increase the intercept parameter. What happens to the line?
  - (c) Decrease the slope parameter. What happens to the line?
  - (d) Compute the correlation between  $X$  and  $Y$ . What does the correlation tell you about the relationship between the two variables? Does it give a good summary for this data?
  - (e) By eye try to fit the best possible line to the data.
  - (f) See how well your line compares with the Least Square fitted line. Compare the Residual SS (SSE) of your fitted line with the least squares line. Record the values in Table 1.
  - (g) What do the numbers in the Least Squares fitted line equation for the Snake River data  $\hat{y} = -1.75 + 0.569362 * x$  mean?

DATA SET	Correlation	Your Resid SS (1)	LS SS (2)	Diff (1)-(2)
SNAKE-RIVER				
OAK-SEEDLING				
TENSILE-STRENGTH				
CRAB				
PETROCHEMICAL-PLANT				

Table 1: Comparing eye-ball fits with Least Squares fits.

2. Write the numbers 2,3,4,5 on separate pieces of paper, jumble the pieces up and select one. The number 2 represents the CRAB data, 3 represents the OAK-SEEDLING data, 4 represents the PETROCHEMICAL-PLANT data and 5 represents the TENSILE-STRENGTH. the

3. Repeat steps 1(d),(e),(f) with the data represented by your number choice.
4. Select a new number (without replacing the last) and repeat steps 1(d),(e),(f) with the data represented by your number choice.
5. Select a new number (without replacing the last) and repeat steps 1(d),(e),(f) with the data represented by your number choice.
6. Repeat steps 1(d),(e),(f) on the remaining data set.
7. On which of the datasets did you do better at approximating the Least Squares line?
8. If this exercise was conducted with a class compile statistics (frequency, mean, std dev) on all individuals performance. Is there a relationship between the correlation between  $X$  and  $Y$  for the data set and performance in eye-balling (that is, differences between eye-balling and least squares residual SS)?

## 7 Solutions to Exercises

1.b. Increasing the intercept causes the line to move vertically upwards.

1.c. Decreasing the slope causes the line to get flatter, that is, less steep.

1.d Correlation is a measure of the linear relationship between the two variables which is related. If the data is linearly related it is a reasonable summary of the relationship, as is the case with the SNAKE-RIVER data. For the PETROCHEMICAL-PLANT correlation is affected by the one extreme value and is not a reasonable summary of the relationship between the other points. For the CRAB data the relationship is close to cubic ( $Wgt \equiv Volume = Length^3$ ) and correlation does not give any information about this non-linear relationship.

1.g. -1.75 is the value on the y-axis where the Least Squares line crosses; For every unit increase in x, the increase in y is 0.569362.

Results in Table 1. One would expect that if the data is very closely linearly related that eye-balling a line would work reasonably well.