INTERACTIVE 3-DIMENSIONAL DIAGRAMS FOR TEACHING MULTIPLE REGRESSION

Doug Stirling
Massey University, New Zealand
d.stirling@massey.ac.nz

Many concepts in simple linear regression can be explained or illustrated on scatterplots. Similar diagrams for regression with two explanatory variables require 3-dimensional scatterplots. Appropriate colouring and dynamic rotation on a computer are needed to effectively show their 3-dimensional nature. Concepts such as multicollinearity, sequential sums of squares and interaction have no analogue in simple linear regression, so it is particularly helpful to illustrate them graphically. This paper gives several examples of concepts in multiple regression that can be illustrated well with 3-dimensional diagrams.

INTRODUCTION

Most introductory statistical concepts can be explained with diagrams in addition to, or even replacing, mathematical formulae and proofs. Dynamic and interactive computer-based diagrams are usually more effective than static paper-based ones, not only for their explanatory power, but also for their ability to keep student interest and because they are more memorable.

In this paper, we show that dynamic and interactive diagrams are equally useful for teaching more advanced topics such as multiple regression. In a static medium, it is difficult to explain to readers why interactive diagrams are more effective than static ones, so it is suggested that readers also examine the interactive versions of the diagrams from this paper in Release 3.1 of *CAST* (Stirling, 2006). All diagrams in this paper are taken from *CAST*.

SIMPLE LINEAR REGRESSION

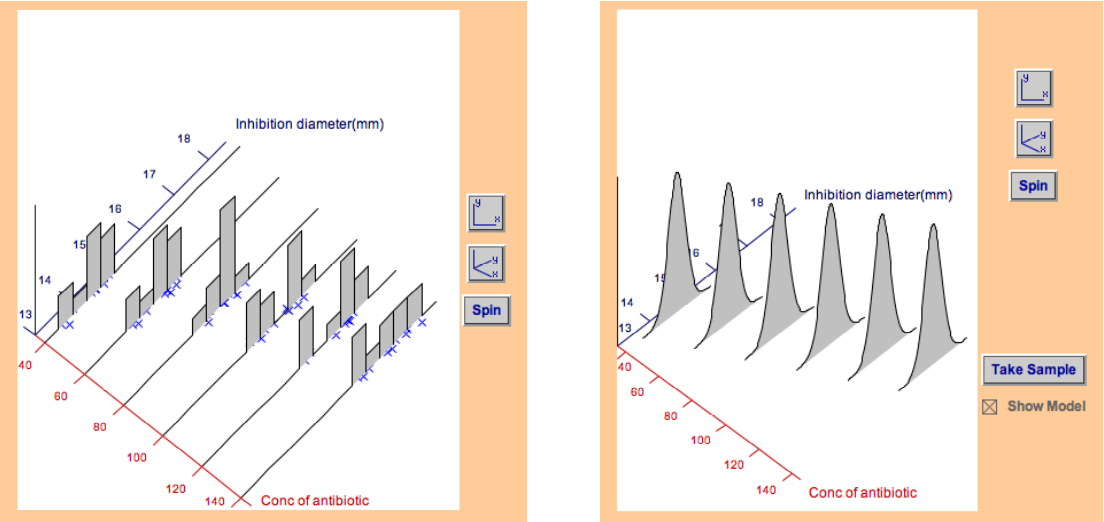
Many concepts in regression can be explained or illustrated much more effectively with diagrams than with proofs and formulae. For example, in simple linear regression, the representation of residuals as vertical lines from the data points on a scatterplot to the least squares line is easier to understand for most students than the equation

$$e_i = y_i - b_0 - b_1 x_i$$

Static diagrams can explain many concepts in simple linear regression, but others are more effectively illustrated with interactive diagrams on a computer. For example, a square can be drawn for each data point with one side being the vertical line representing the residual. Dragging the line to minimize the total area of these squares illustrates the principle of least squares (Finzer *et al*., 1998). Dragging one point in a scatterplot to make it an outlier is also an effective way to demonstrate the concepts of leverage and influence (Lock, 2002).

Figure 1(a) represents a data set with repeated response measurements at each *x* as a set of histograms in a 3-dimensional diagram, motivating the corresponding normal linear model for the data in Figure 1(b). Both diagrams are easiest to understand when the diagram can be dynamically rotated; dragging the centre of each diagram in *CAST* rotates it. For example, the ideas of linearity and constant variance can be explained for both the data and model by rotating to make all histograms or normal curves coincide as closely as possible.

Finally, simulations are effective ways to demonstrate the sampling distributions of the least squares coefficients and the properties of related confidence intervals and p-values. These are easiest to understand by students if interaction is involved by clicking a button to generate each sample and build up the relevant sampling distribution.



(a) Sample distributions of *y* at each *x*

(b) Possible normal linear regression model

Figure 1: Diagrams motivating the simple linear regression model

LEAST SQUARES WITH TWO EXPLANATORY VARIABLES

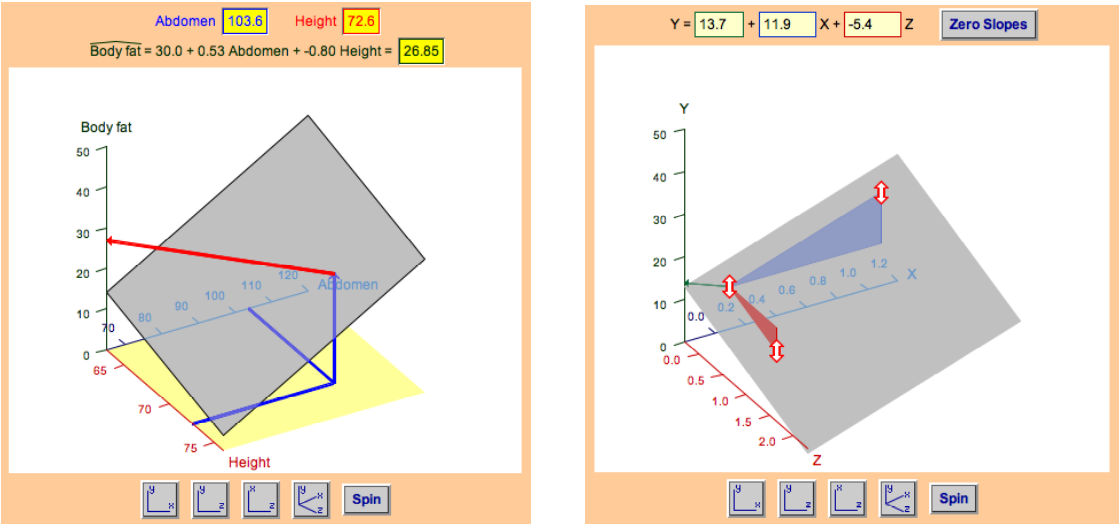
Simple linear regression is relatively easy to teach, in part because most concepts can be illustrated so well with 2-dimensional diagrams. Many students find it much harder to extend these ideas to multiple regression, in part because it is much harder to illustrate the concepts with diagrams; many textbooks abandon graphical illustrations in favour of formulae and equations. However although multiple regression data sets and models cannot be represented well graphically if there are three or more explanatory variables, graphical displays are possible for models with two explanatory variables. Many of the additional complications when moving from one explanatory variable to many are present when there are only two explanatory variables, so ideas such as multicollinearity can be illustrated effectively using two explanatory variables.

Most diagrams illustrating regression concepts are based on a 3-dimensional scatterplot of the response, *y*, against *x* and *z*. The third dimension must be effectively represented and this can only be done well in computer-based interactive diagrams where the scatterplot can be rotated, either by dragging with the mouse or automatically spinning the diagram. It is best to restrict the rotations to prevent the y-axis from pointing left or right since completely free rotations do not help to explain regression concepts. Appropriate shading of elements in the plot, such as dimming axes and crosses behind the regression plane, helps to reinforce the 3-dimensional nature of the diagrams.

As a preliminary to multiple regression, students must be taught that the linear equation

$$y \;=\; \beta_0 \;+\; \beta_1 x \;+\; \beta_2 z$$

corresponds to a plane in three dimensions. In Figure 2(a), students can click anywhere in the yellow *x*-*z* plane and the diagram will show the value of *y* generated by the equation to verify that the predictions lie on a plane. In Figure 2(b), the three arrows can be dragged up and down to adjust the intercept and two slope parameters, showing in particular how the sign and magnitude of the slope parameters affect the position of the regression plane.



Abdomen 103.6 Height 72.6

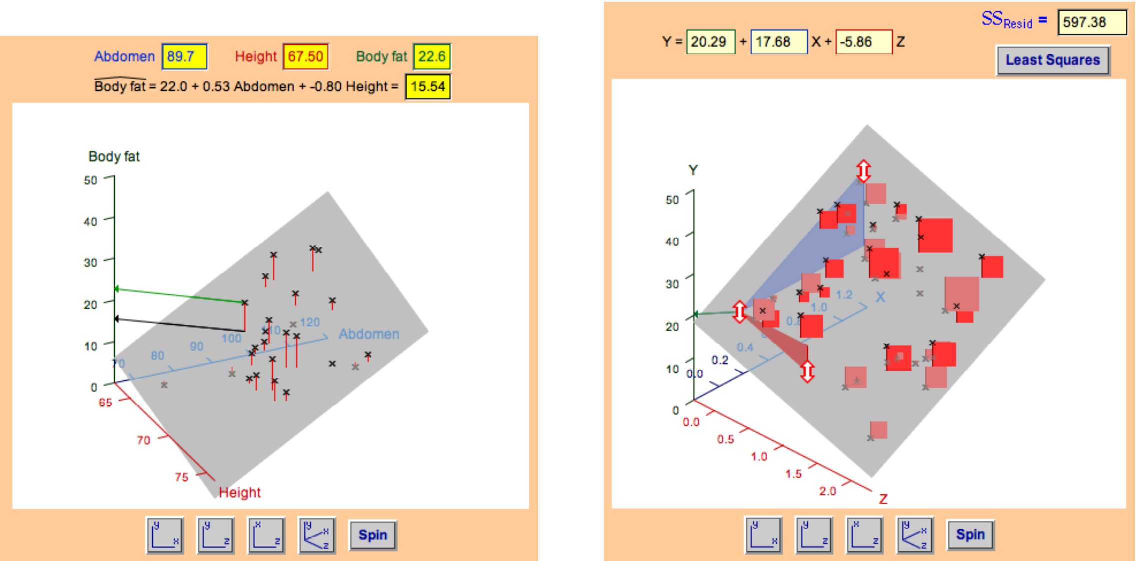
$\widehat{\text{Body fat}}$ = 30.0 + 0.53 Abdomen + -0.80 Height = 26.85

Y = 13.7 + 11.9 X + -5.4 Z

(a) Using a linear equation to predict *y* from *x* and *z* (b) Meaning of the intercept and slope parameters

Figure 2: Diagrams representing a linear equation as a plane

The representation of residuals as vertical distances from the data points to the regression plane is shown in Figure 3(a). Clicking on a cross displays the actual and fitted response. Figure 3(b) shows squares representing the squared residuals and illustrates how the parameters can be adjusted (by dragging the arrows) to minimise the residual sum of squares. As in the other diagrams shown here, dynamic rotation and shading helps to support the 3-dimensional nature of the diagrams.



Abdomen 89.7 Height 67.50 Body fat 22.6

$\widehat{\text{Body fat}}$ = 22.0 + 0.53 Abdomen + -0.80 Height = 15.54

SS_{Resid} = 597.38

Y = 20.29 + 17.68 X + -5.86 Z

(a) Fitted values and residuals for a plane (not the least squares plane) (b) Adjusting parameters to minimise residual sum of squares

Figure 3: Diagrams showing residuals and illustrating the method of least squares

REGRESSION MODEL AND INFERENCE

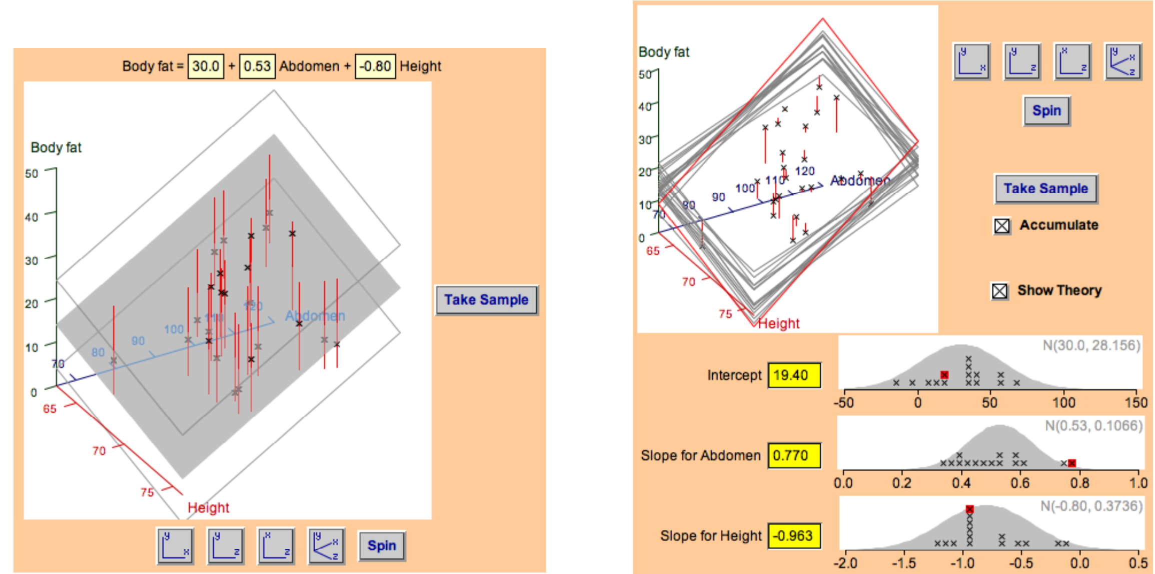
Figure 1(b) cannot be extended to represent the normal linear model with two explanatory variables,

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 z_i + \varepsilon_i \quad \text{where} \quad \varepsilon_i \sim \text{normal}(0, \sigma)$$

A simpler display of the normal distribution is needed to replace the normal density curve. Figure 4(a) does this with a line extending 2σ on each side of the response mean at each (x_i, z_i) combination. Planes 2σ on each side of the regression plane are also shown in outline. The button

"Take Sample" in the diagram simulates response values from the model and shows the data as crosses. Such samples demonstrate that approximately 95% of the crosses are within the bands at ±2σ.

Figure 4(b) simulates data from the same model and shows the sampling variability of the least squares plane and the individual parameter estimates.



Body fat = 30.0 + 0.53 Abdomen + -0.80 Height

Intercept 19.40

Slope for Abdomen 0.770

Slope for Height -0.963

☒ Accumulate

☒ Show Theory

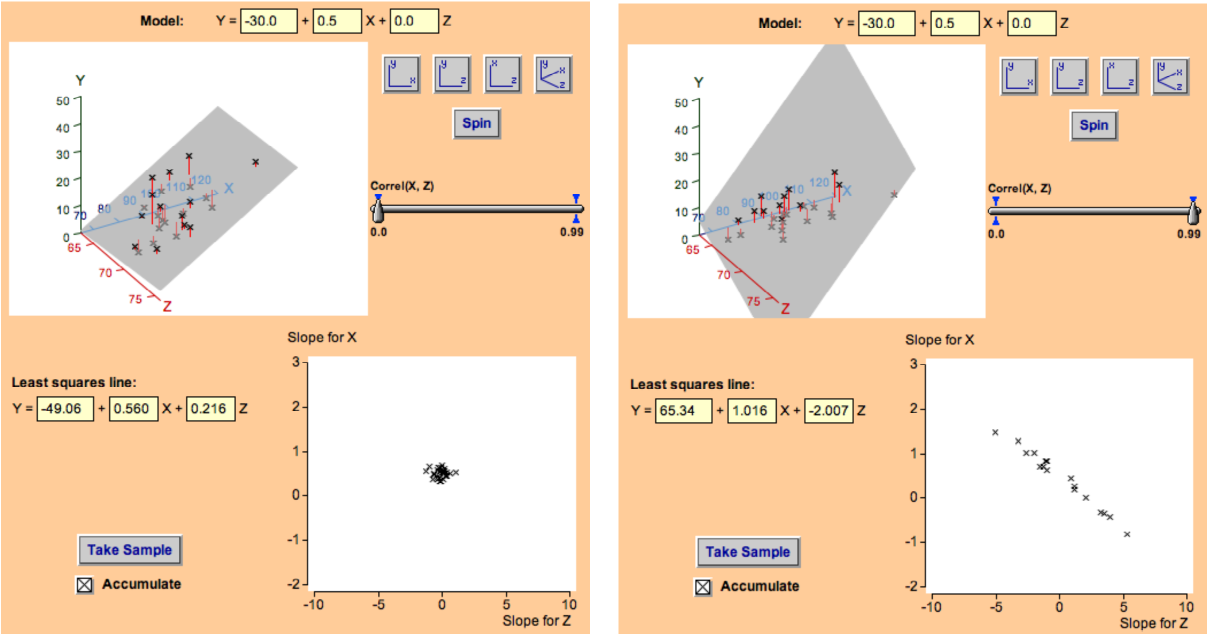
(a) Normal regression model with bands 2σ on each side of the mean plane

(b) Sampling distribution of least squares planes and coefficients

Figure 4: Representation of normal regression model and the resulting sampling variability

MULTICOLLINEARITY

Figures 2 to 4 all explain or illustrate concepts in multiple regression that mirror ones in simple linear models. The effects of correlated explanatory variables are harder to explain since there is no analogue in the simple linear model. The simulation in Figure 5 demonstrates the higher standard errors of the two slope estimates when the explanatory variables are correlated.



Model: Y = -30.0 + 0.5 X + 0.0 Z

Correl(X, Z) 0.0 0.99

Least squares line: Y = -49.06 + 0.560 X + 0.216 Z

☒ Accumulate

Model: Y = -30.0 + 0.5 X + 0.0 Z

Correl(X, Z) 0.0 0.99

Least squares line: Y = 65.34 + 1.016 X + -2.007 Z

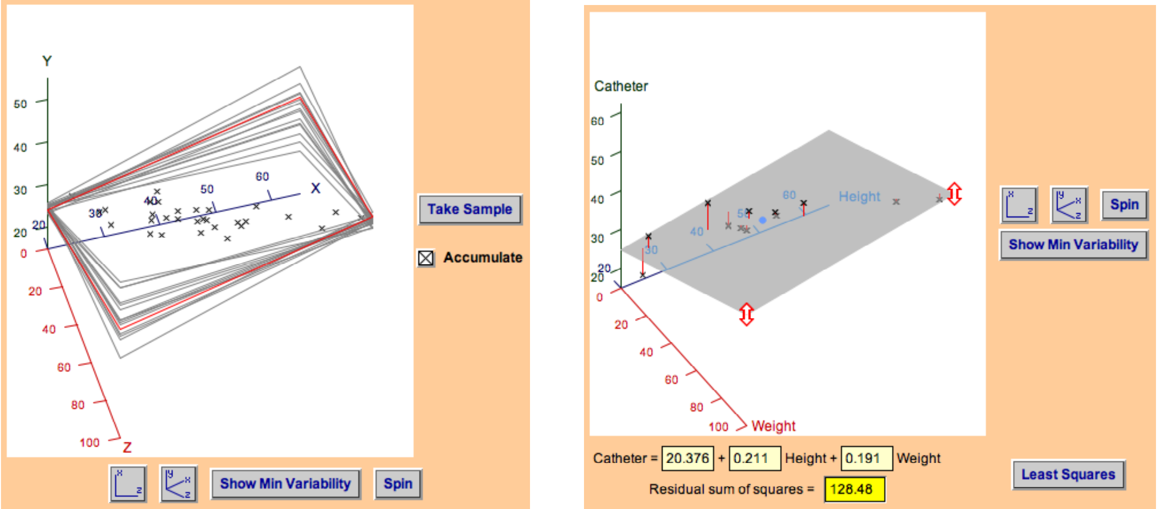
☒ Accumulate

(a) When *x* and *z* are uncorrelated

(b) When *x* and *z* are highly correlated

Figure 5: Simulations showing how multicollinearity increases the standard errors of the slopes

The simulation in Figure 6(a) shows the sampling variability of the least squares planes for multicollinear data. Rotating the diagram to look down the ‘cylinder of data’ shows the extra variability in two corners as ‘flapping wings.’ In Figure 6(b), the plane can be tilted with the two arrows to show that the residual sum of squares is relatively insensitive to moving two of the four corners.



Catheter = 20.376 + 0.211 Height + 0.191 Weight

Residual sum of squares = 128.48

(a) Sampling distribution of LS planes

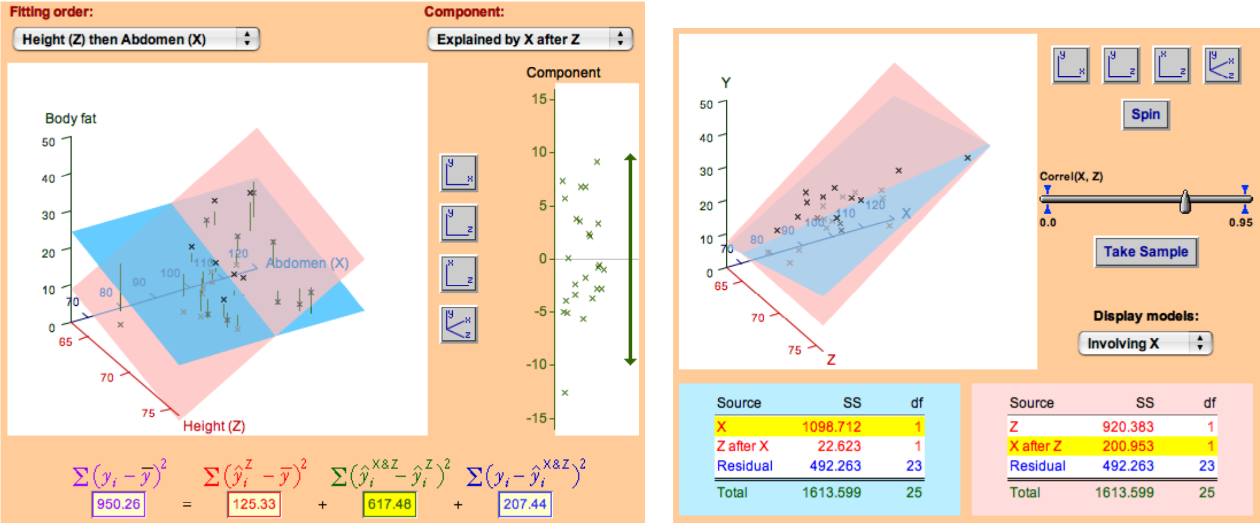
(b) Sensitivity of residual sum of squares to moving the plane

Figure 6: Diagrams explaining some consequences of multicollinearity

PARTITIONING VARIABIILITY

As a final example of the use of 3-dimensional diagrams to illustrate concepts in regression models with two explanatory variables, note that the sequential sums of squares in analysis of variance are sums of squared differences between the fitted values for different models. The components that are squared can be represented as distances between regression planes, as in Figure 7(a).

Figure 7(b) shows the sequential sums of squares for a data set in which the correlation between *x* and *z* can be altered, illustrating that the two sets of sequential sums of squares are only equal when the explanatory variables are uncorrelated.



Fitting order: Height (Z) then Abdomen (X)

Component: Explained by X after Z

$$\sum(y_i-\bar{y})^2 = \sum(\hat{y}_i^{Z}-\bar{y})^2 + \sum(\hat{y}_i^{X\&Z}-\hat{y}_i^{Z})^2 + \sum(y_i-\hat{y}_i^{X\&Z})^2$$

950.26 = 125.33 + 617.48 + 207.44

Display models: Involving X

Source	SS	df
X	1098.712	1
Z after X	22.623	1
Residual	492.263	23
Total	1613.599	25

Source	SS	df
Z	920.383	1
X after Z	200.953	1
Residual	492.263	23
Total	1613.599	25

(a) Representation of sums of squares as vertical lines between planes

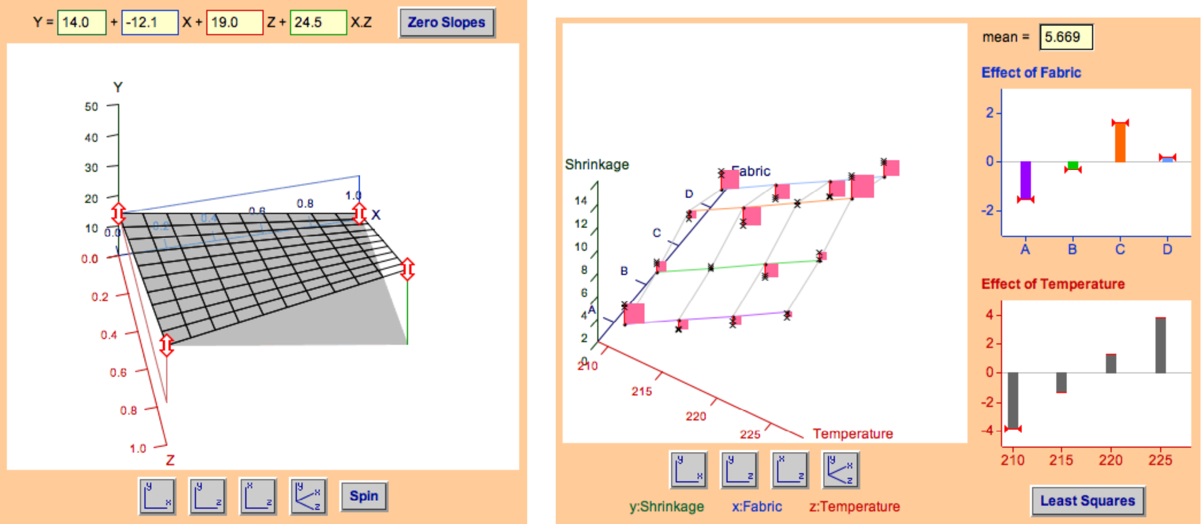
(b) Sums of squares for different orders of adding *x* and *z*

Figure 7: Diagrams about analysis of variance

INTERACTION AND FACTORS

Although most multiple regression models with three or more explanatory variables cannot be displayed in rotating plots, a few extensions to the two-explanatory-variable model can represented effectively. Figure 8(a) represents a model with a linear-by-linear interaction term as a grid. The corresponding model without interaction is shown in grey. In the diagram, the four arrows can be used to drag the four parameters of the model. The model with a quadratic term in either variable could be shown in a similar way as a curved surface.

Figure 8(b) represents a linear model with one numerical variable and a factor. The user-interface is not as obvious as in Figure 8(a), but the parameters of the model can be adjusted by dragging the four bars in the upper bar-chart and the leftmost bar in the lower bar chart. This diagram also shows a data set and the squared residuals, so the parameters can be adjusted to minimise the residual sum of squares.



Y = 14.0 + -12.1 X + 19.0 Z + 24.5 X.Z

mean = 5.669

(a) Linear-by-linear interaction

(b) Linear variable and a factor

Figure 8: Some extensions to the regression model with two explanatory variables

CONCLUSION

Most concepts that can be illustrated on a scatterplot for simple linear regression can be illustrated in a similar way for regression with two explanatory variables by replacing the 2-dimensional scatterplot with a rotating 3-dimensional one. Concepts such as multicollinearity, sequential sums of squares and interaction have no counterpart in simple linear regression, so it is particularly useful to illustrate their effects graphically.

The interactive diagrams that are shown in this paper are only a selection of these about simple and multiple regression in *CAST*; many further concepts and points can be effectively illustrated with interactive diagrams. *CAST* is an online web application that is available at no charge (after registration) from the reference below. Real data sets are used wherever possible.

REFERENCES

Stirling, W.D. (2006). *CAST 3.1*, http://cast.massey.ac.nz.

Finzer, W. F. and Erickson, T. E. (1998). *Dataspace*—A computer learning environment for data analysis and statistics based on dynamic dragging, visualization, simulation, and networked collaboration. In L. Pereria-Mendoza, L. S. Kea, T. W. Kee, and W-K. Wong (Eds.), *Statistical Education - Expanding the Network: Proceedings of the Fifth International Conference on Teaching Statistics*, Singapore, (pp. 827-831). Voorburg: The Netherlands: International Statistical Institute.

Lock, R. H. (2002). Using Fathom to promote interactive explorations of statistical concepts. In B. Phillips (Ed.), *Proceedings of the Sixth International Conference on Teaching of Statistics*, Cape Town. Voorburg, The Netherlands: International Statistical Institute.