CHAPTER 1

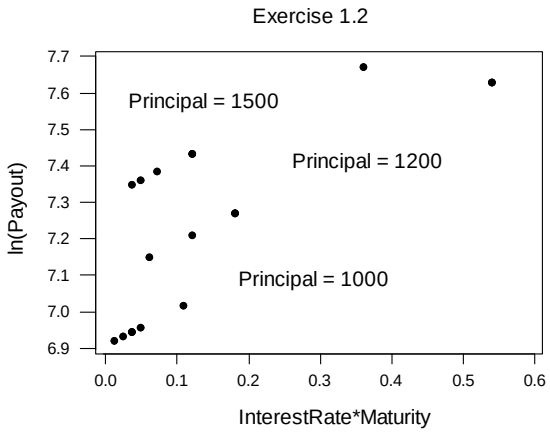
1.1 Tensile strength of an alloy can be expected to increase with increasing hardness and density of the stock. Bivariate scatter plots of tensile strength against hardness and of tensile strength against density of the stock are useful. Such scatter plots indicate whether the relationship is linear, or more complicated.

Bivariate scatter plots are unable to reveal 3-dimensional relationships. For that one needs to consider three-dimensional graphs. Alternatively, one can proceed as follows. If measurements on the tensile strengths of several different alloys of a given density but of changing values of hardness are given, one can plot tensile strength against hardness at this one fixed level of density. Furthermore, if tensile strength and hardness data for alloys of a second different density are available, one can construct a similar scatter plot for that other level of density. If the two scatter plots (scatter plots of tensile strength against hardness, at the two different levels of density) show different slopes, then the effect of hardness on tensile strength depends on the level of density. The factors hardness and density of the stock are said to interact in their effect on tensile strength.

Data from experiments are usually more informative as one can control the conditions under which the experimental runs are carried out. Experimentation is probably not possible in case (f). The relative humidity conditions in the plant can not be varied according to a fixed experimental plan. Instead, one takes measurements in the plant on the relative humidity, and at the same time on the output (performance) of the process. A danger with such data is that the relative humidity in the plant may be affected by unknown factors that also affect the output. The root cause is not the humidity of the plant, but these other "lurking" variables.

1.2 The graph given below indicates a linear relationship between ln(Payout) and the product of interest rate and maturity, with an intercept that depends on the invested principal. Note that the linear model in the transformed variables fits perfectly.

This is expected from the model $\text{Payout} = P\exp(RT)$. Taking the logarithm on both sides of the equation, leads to $\ln(\text{Payout}) = \ln(P) + RT$. The intercept changes with the logarithm of the invested principle; the regression coefficient of RT is one.



1.3 Selected examples are:

Exercise 2.9: MBA grade point average and GMAT score: observational study
Exercise 2.10: Fuel efficiency and car characteristics: observational study of 45 cars
Exercise 2.24: Thickness of egg shell and PCB: observational study on pelicans
Exercise 2.27: Absorption of chemical liquid; experimental data
Exercise 4.12: Amount of plant water usage: observational study
Exercise 4.14: Survival of bull semen: experimental data
Exercise 4.15: Toxic action of a certain chemical on silkworm larvae: experimental data
Exercise 4.21: Abrasion as function of hardness and tensile strength of rubber: experimental data
Exercise 6.14: Tear properties of paper: experimental data
Exercise 6.17: Rigidity, elasticity and density of timber: observational study
Exercise 8.1: Incumbent vote share in US presidential elections: observational study
Exercise 8.2: Height and weight of boys: observational study
Exercise 8.3: Soft drink sales: observational study

1.4 The response variable may be the breaking strength of a viscose fiber, and the explanatory variables may be the percentage of certain chemicals in the spin bath and the speed at which the liquid viscose is pressed through the nozzles into the spin bath. A designed experiment varies the explanatory variables (the design factors) according to a well thought-out plan and randomizes the arrangement of the experimental runs. The breaking strength of the resulting material is measured for each experimental run. In this case the data arise from a designed experiment.

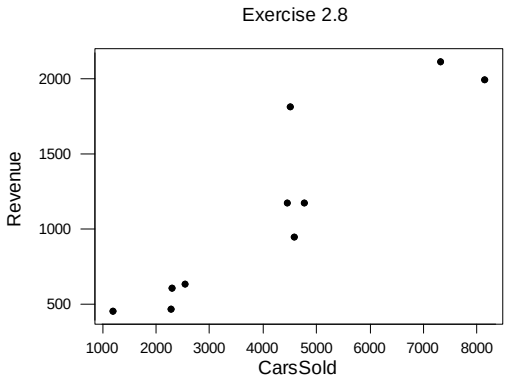
However, the data could also be obtained through an observational study. The plant manager may take readings on the process – measurements on the breaking strength of

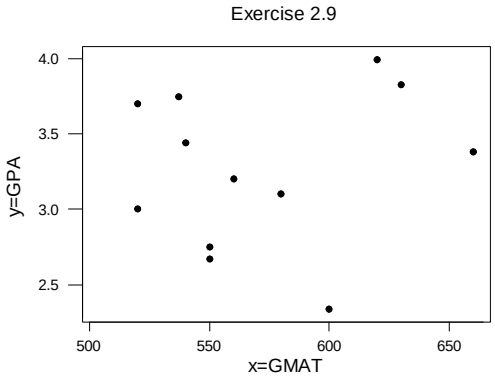
the fiber and on the chemicals present in the spin bath, as well as the speed of the process. The manager may do this every 4 hours, collect observational data, and construct a regression model relating the response to the explanatory variables. However, several problems may arise with such observational data. First, the variation in the explanatory factors may not be large enough to actually affect the response. Second, and more importantly, the response may be affected by other variables that one has failed to control and account for. For example, the relative humidity may play a role. With observational data such as these one is never sure whether a "lurking" variable may be present. With designed experiments, and proper randomization of the experimental runs, such problems are much smaller.

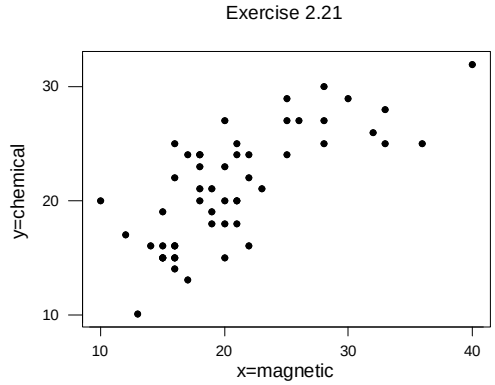
Monthly macroeconomic data on interest rates, GNP, and unemployment are examples of observational data. The data are given to the analyst who has no opportunity to affect the way the data are obtained.

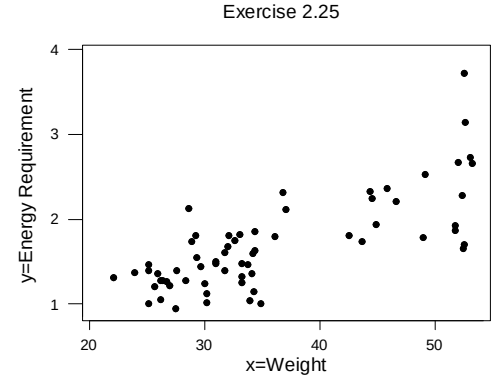
Survey data are other examples of observational data; for example, survey data that involve observations on brand choices and features of products. Alternatively, brand preferences can be assessed through designed experiments. Participants in such experiments are presented a sequence of brands with various characteristics, arranged in a random sequence, and their brand selections are measured. In this case the data arise from an experiment.

1.5 Scatter plots for the data in Exercises 2.8, 2.9, 2.21, and 2.25 are given below. We notice a linear relationship in Exercise 2.8. There is no strong (linear) relationship in Exercise 2.9. The relationship in Exercise 2.21 may involve a quadratic component; more information on the response when x is in the range from 30 to 40 would be helpful. We notice an approximate linear relationship in Exercise 2.25. However, note that the two responses between 3 and 4 at the high level of x are somewhat different from the rest.

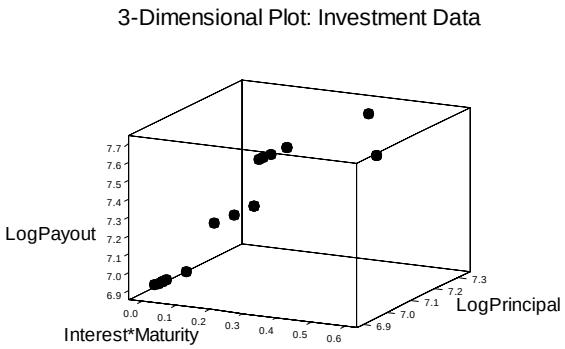


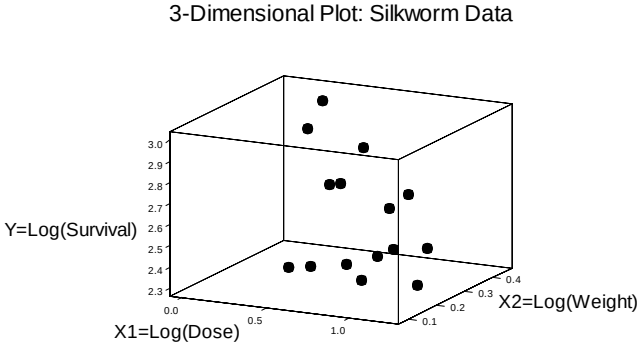

Exercise 2.9
y=GPA
x=GMAT
4.0
3.5
3.0
2.5
500
550
600
650


Exercise 2.21
y=chemical
x=magnetic
30
20
10
10
20
30
40


Exercise 2.25
y=Energy Requirement
x=Weight
4
3
2
1
20
30
40
50

1.6 Usually it is not very easy to spot relationships from 3-dimensional graphs; see the two examples shown below. The bivariate scatter plots for the silkworm data set are easier to interpret.





1.7 Consider models with a single explanatory variable *x*. The quadratic model,

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon \ ,$$

is nonlinear in the explanatory variable *x*, but linear in the three parameters β_0, β_1 and β_2. The polynomial model (with $p > 1$),

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_p x^p + \varepsilon \ ,$$

is nonlinear in the explanatory variable *x*, but linear in the parameters.

The quadratic model with two explanatory variables,

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} (x_1)^2 + \beta_{22} (x_2)^2 + \beta_{12} x_1 x_2 + \varepsilon \ ,$$

is nonlinear in x_1 and x_2, however it is linear in the parameters. The equation describes a quadratic function in two variables. For certain values of the parameters the expected response looks like a bowl with a unique minimum, an upside bowl with a unique maximum, or a saddle point.

1.8 Consider a response *y* and a single explanatory variable *x*. The following models are nonlinear in the parameters. You may want to consider one of these models and trace out the mean response for changing levels of *x*. For example, take the first model with $\alpha = 0.39$ and $\beta = 0.10$ and consider *x* values between 8 and 40. This particular model is studied in Chapter 9; *x* is the age of a chemical product in weeks, and the response *y* is its remaining chlorine.

$$y = \alpha + (0.49 - \alpha)\exp[-\beta(x-8)] + \varepsilon$$

$$\mathrm{y} = \beta_1 + \frac{\beta_2}{1+\exp[-\beta_3(\mathrm{x}-\beta_4]} + \varepsilon$$

$$y = \frac{\alpha}{1+\beta\exp(-\gamma x)} + \varepsilon \quad \alpha > 0, \beta > 0, \gamma > 0$$

$$y = \frac{\beta_1 x}{\beta_2 + x} + \varepsilon$$

1.9 Sales may increase linearly with time, but the variability may depend on the level (the mean) of sales. If sales are very small, one can not expect tremendous variability. Sales can not be negative, so the variability is automatically bounded from below. On the other hand there is more room for bigger variability if the level of the sales is high. It is useful to think in terms of percentages. One may expect a variability (expressed as a standard deviation) of ± 10 percent. If sales are at level 10, this implies an uncertainty of ± 1 units. On the other hand, if the level is at 1000, the uncertainty is ± 100 units. If the variability (standard deviation) is proportional to the level, one should analyze the logarithm of sales, and not the sales. You will learn in Chapter 6 that this transformation stabilizes the variance. In this situation the variability in the logarithms of sales does not depend on the level of the sales.

Another situation, where the variability of the response can be expected to depend on the explanatory variable is when measuring distance. Assume that we want to determine the distances between pairs of points (where some are close together, while others are far apart). We can expect that the error in measuring close distances is smaller than the error

in measuring points that are far apart. The variability in the measurements can be expected to increase with distance.

1.10 Economic "well-being" has an impact on people's decision to have children. During the post World War II period, a period characterized by rapid economic growth, many young Europeans affected by the war delayed their decision to have children. Economic activity of the post World War II period also had an impact on the breeding space for storks and led to a decrease in the number of storks. Considering annual numbers of births and annual numbers of storks, one can observe a strong positive correlation. However, no one - except young children - would interpret this correlation as a causal effect.

Poverty of a school district affects the number of students in subsidized lunch programs, with poorer districts having more children in these nationally subsidized programs. Poverty also affects the scholastic test scores in these districts. The strong positive correlation between the number of children in subsidized lunch programs and test achievement scores in these districts does not imply that there is a causal connection between subsidized lunch and test scores. It is poverty that is the driving causal factor.

High summer temperatures are related to high beer sales. High summer temperatures are also related to increased sales of suntan lotion. Daily sales of suntan lotion and beer sales are positively correlated. This, however, does not imply a causal connection. It is not that people who drink require more sun tan lotion.

1.11 Contact your state to obtain this information.

1.12 (a) Ignoring variability, we find that for the ith subject: $\text{RelativeRaise}_i = \beta\text{Performance}_i$. All points in the graph of RelativeRaise against Performance are on a straight line through the origin.

The absolute raise (that is, the raise in terms of dollars earned) can be written as

$$\text{AbsoluteRaise} = (R)(\text{PreviousSalary}) = (\beta\text{PreviousSalary})\text{Performance}$$

A graph of AbsoluteRaise against Performance does not exhibit a perfect linear association as the slope depends on the previous salary that changes from person to person. A regression of AbsoluteRaise on Performance may not provide the correct estimate of the parameter β. Take two workers; the previous salary of the first worker is half the salary of the second one, but the first worker is twice as productive. Their absolute raises are the same. The slope in the plot of AbsoluteRaise against Performance is zero, and not the desired parameter β.

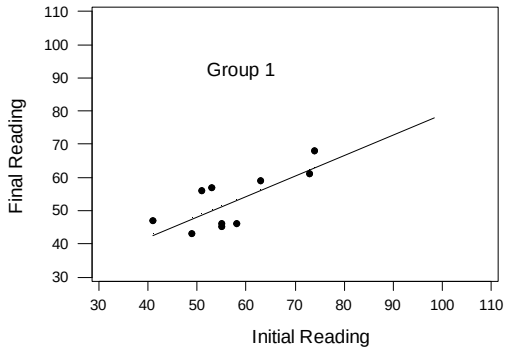
(b) Let R = Relative Raise, where R is a small number such as 0.03 (3 percent). The ratio

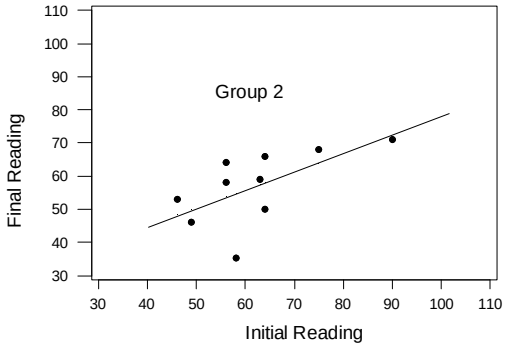
CurrentSalary/PreviousSalary = [(1 + R)PreviousSalary]/PreviousSalary = 1 + R . A first-order Taylor series expansion of $\ln(1 + R) \approx R$ is valid for small R . Hence $\ln(\text{CurrentSalary/PreviousSalary}) = \ln(1 + R) \approx R = \beta\text{Performance}$ is linearly related to Performance. A regression of ln(CurrentSalary/PreviousSalary) on Performance provides an estimate of β.

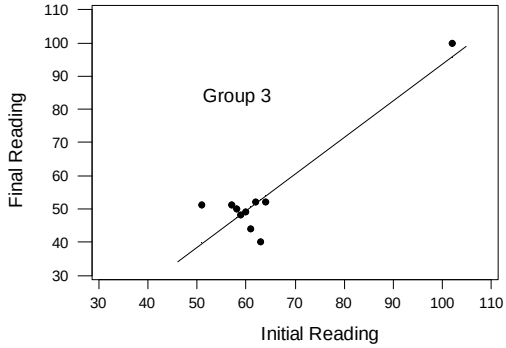
1.13 The five separate scatter plots of final reading *y* against initial reading *z*, one for each contraceptive group, are given below. The graphs have identical scales on both axes, and the "best fitting" straight lines have been added to the plots.

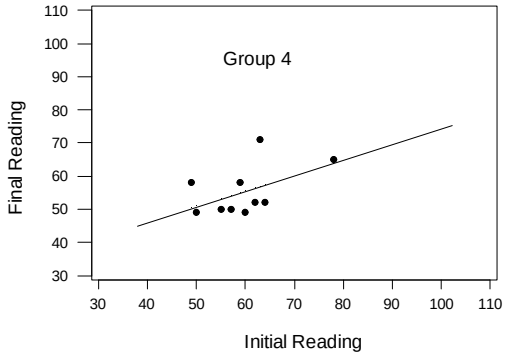
Model 1.8 assumes that the slopes in these five graphs are the same. The five graphs show that this may be a reasonable assumption. For the third group the slope is difficult to estimate. Apart from one subject with a very large initial reading ($z = 102$) there is little variation among the initial readings (all other *z*'s are between 50 and 65). It is difficult to pin down the value of the slope as the estimate is heavily influenced by the one subject with the high initial reading ($z = 102$) and response $y = 100$. Chapter 6 discusses influential observations in detail.

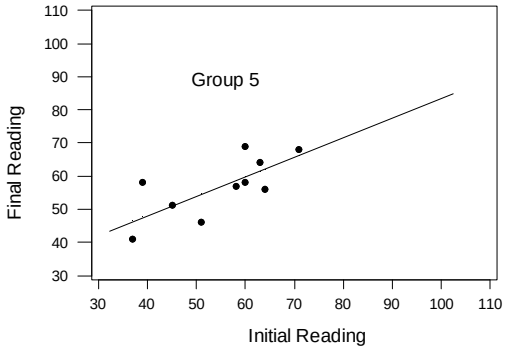
Assuming that $\alpha = 1$, one can look at the changes, $y - z$ = final reading - initial reading. This implies that we compare five groups (samples), with the objective to test whether the means of the changes are the same. That is, H_0: $\beta_1 = \beta_2 = \ldots = \beta_5$.




Group 2
Final Reading
Initial Reading
110
100
90
80
70
60
50
40
30
30 40 50 60 70 80 90 100 110


Group 3
Final Reading
Initial Reading
110
100
90
80
70
60
50
40
30
30 40 50 60 70 80 90 100 110


Group 4
Final Reading
Initial Reading
110
100
90
80
70
60
50
40
30
30 40 50 60 70 80 90 100 110


Group 5
Final Reading
Initial Reading
110
100
90
80
70
60
50
40
30
30
40
50
60
70
80
90
100
110