1
Linear Regression

1.1 Introduction

When we first meet Statistics, we encounter random quantities (random variables, in probability language, or variates, in statistical language) one at a time. This suffices for a first course. Soon however we need to handle more than one random quantity at a time. Already we have to think about how they are related to each other.

Let us take the simplest case first, of two variables. Consider first the two extreme cases.

At one extreme, the two variables may be independent (unrelated). For instance, one might result from laboratory data taken last week, the other might come from old trade statistics. The two are unrelated. Each is *uninformative* about the other. They are best looked at separately. What we have here are really *two* one-dimensional problems, rather than one two-dimensional problem, and it is best to consider matters in these terms.

At the other extreme, the two variables may be essentially the same, in that each is *completely informative* about the other. For example, in the Centigrade (Celsius) temperature scale, the freezing point of water is 0^o and the boiling point is 100^o, while in the Fahrenheit scale, freezing point is 32^o and boiling point is 212^o (these bizarre choices are a result of Fahrenheit choosing as his origin of temperature the lowest temperature he could achieve in the laboratory, and recognising that the body is so sensitive to temperature that a hundredth of the freezing-boiling range as a unit is inconveniently large for everyday,

N.H. Bingham and J.M. Fry, *Regression: Linear Models in Statistics*,
Springer Undergraduate Mathematics Series, DOI 10.1007/978-1-84882-969-5_1,
© Springer-Verlag London Limited 2010

non-scientific use, unless one resorts to decimals). The transformation formulae are accordingly

$$C = (F - 32) \times 5/9, \qquad F = C \times 9/5 + 32.$$

While both scales remain in use, this is purely for convenience. To look at temperature in both Centigrade and Fahrenheit together for scientific purposes would be silly. Each is *completely informative* about the other. A plot of one against the other would lie *exactly* on a straight line. While apparently a two–dimensional problem, this would really be only *one* one-dimensional problem, and so best considered as such.

We are left with the typical and important case: two–dimensional data, $(x_1, y_1), \ldots, (x_n, y_n)$ say, where each of the x and y variables is *partially but not completely informative about the other*.

Usually, our interest is on one variable, y say, and we are interested in what knowledge of the other – x – tells us about y. We then call y the *response variable*, and x the *explanatory variable*. We know more about y knowing x than not knowing x; thus knowledge of x explains, or accounts for, part but not all of the variability we see in y. Another name for x is the *predictor* variable: we may wish to use x to predict y (the prediction will be an uncertain one, to be sure, but better than nothing: there is information content in x about y, and we want to use this information). A third name for x is the *regressor*, or regressor variable; we will turn to the reason for this name below. It accounts for why the whole subject is called *regression*.

The first thing to do with any data set is to look at it. We subject it to exploratory data analysis (EDA); in particular, we plot the graph of the n data points (x_i, y_i). We can do this by hand, or by using a statistical package: Minitab®,[1] for instance, using the command `Regression`, or S-Plus/R® by using the command `lm` (for linear model – see below).

Suppose that what we observe is a scatter plot that seems roughly linear. That is, there seems to be a systematic component, which is linear (or roughly so – linear to a first approximation, say) and an error component, which we think of as perturbing this in a random or unpredictable way. Our job is to fit a line through the data – that is, to estimate the systematic linear component.

For illustration, we recall the first case in which most of us meet such a task – experimental verification of Ohm's Law (G. S. Ohm (1787-1854), in 1826). When electric current is passed through a conducting wire, the current (in amps) is proportional to the applied potential difference or voltage (in volts), the constant of proportionality being the inverse of the *resistance* of the wire

[1] Minitab®, Quality Companion by Minitab®, Quality Trainer by Minitab®, Quality. Analysis. Results® and the Minitab logo are all registered trademarks of Minitab, Inc., in the United States and other countries.

(in ohms). One measures the current observed for a variety of voltages (the more the better). One then attempts to fit a line through the data, observing with dismay that, because of experimental error, no three of the data points are exactly collinear. A typical schoolboy solution is to use a perspex ruler and fit by eye. Clearly a more systematic procedure is needed. We note in passing that, as no current flows when no voltage is applied, one may restrict to lines through the origin (that is, lines with zero intercept) – by no means the typical case.

1.2 The Method of Least Squares

The required general method – the Method of Least Squares – arose in a rather different context. We know from Newton's *Principia* (Sir Isaac Newton (1642–1727), in 1687) that planets, the Earth included, go round the sun in elliptical orbits, with the Sun at one focus of the ellipse. By cartesian geometry, we may represent the ellipse by an algebraic equation of the second degree. This equation, though quadratic in the variables, is *linear* in the coefficients. How many coefficients p we need depends on the choice of coordinate system – in the range from two to six. We may make as many astronomical observations of the planet whose orbit is to be determined as we wish – the more the better, n say, where n is large – much larger than p. This makes the system of equations for the coefficients grossly over-determined, *except* that all the observations are polluted by experimental error. We need to tap the information content of the large number n of readings to make the best estimate we can of the small number p of parameters.

Write the equation of the ellipse as

$$a_1x_1 + a_2x_2 + \ldots = 0.$$

Here the a_j are the *coefficients*, to be found or estimated, and the x_j are those of x^2, xy, y^2, x, y, 1 that we need in the equation of the ellipse (we will always need 1, unless the ellipse degenerates to a point, which is not the case here). For the ith point, the left-hand side above will be 0 if the fit is exact, but ϵ_i say (denoting the ith error) in view of the observational errors. We wish to keep the errors ϵ_i small; we wish also to put positive and negative ϵ_i on the same footing, which we may do by looking at the squared errors ϵ_i^2. A measure of the discrepancy of the fit is the sum of these squared errors, $\sum_{i=1}^n \epsilon_i^2$. The Method of Least Squares is to choose the coefficients a_j so as to minimise this sums of squares,

$$SS := \sum_{i=1}^n \epsilon_i^2.$$

As we shall see below, this may readily and conveniently be accomplished.

The Method of Least Squares was discovered independently by two workers, both motivated by the above problem of fitting planetary orbits. It was first

published by Legendre (A. M. Legendre (1752–1833), in 1805). It had also been discovered by Gauss (C. F. Gauss (1777–1855), in 1795); when Gauss published his work in 1809, it precipitated a priority dispute with Legendre.

Let us see how to implement the method. We do this first in the simplest case, the fitting of a straight line

$$y = a + bx$$

by least squares through a data set $(x_1, y_1), \ldots, (x_n, y_n)$. Accordingly, we choose a, b so as to minimise the *sum of squares*

$$SS := \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2.$$

Taking $\partial SS/\partial a = 0$ and $\partial SS/\partial b = 0$ gives

$$\begin{aligned} \partial SS/\partial a := -2\sum_{i=1}^{n} e_i &= -2\sum_{i=1}^{n} (y_i - a - bx_i), \\ \partial SS/\partial b := -2\sum_{i=1}^{n} x_i e_i &= -2\sum_{i=1}^{n} x_i (y_i - a - bx_i). \end{aligned}$$

To find the minimum, we equate both these to zero:

$$\sum_{i=1}^{n} (y_i - a - bx_i) = 0 \quad \text{and} \quad \sum_{i=1}^{n} x_i (y_i - a - bx_i) = 0.$$

This gives two simultaneous linear equations in the two unknowns a, b, called the *normal equations*. Using the 'bar' notation

$$\overline{x} := \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Dividing both sides by n and rearranging, the normal equations are

$$a + b\overline{x} = \overline{y} \quad \text{and} \quad a\overline{x} + b\overline{x^2} = \overline{xy}.$$

Multiply the first by $\overline{x}$ and subtract from the second:

$$b = \frac{\overline{xy} - \overline{x}.\overline{y}}{\overline{x^2} - (\overline{x})^2},$$

and then

$$a = \overline{y} - b\overline{x}.$$

We will use this bar notation systematically. We call $\overline{x} := \frac{1}{n}\sum_{i=1}^{n} x_i$ the *sample mean*, or average, of $x_1, \ldots, x_n$, and similarly for $\overline{y}$. In this book (though not all others!), the *sample variance* is defined as the average, $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$, of $(x_i - \overline{x})^2$, written s_x^2 or s_{xx}. Then using linearity of average, or 'bar',

$$s_x^2 = s_{xx} = \overline{(x - \overline{x})^2} = \overline{x^2 - 2x.\overline{x} + \overline{x}^2} = \overline{(x^2)} - 2\overline{x}.\overline{x} + (\overline{x})^2 = \overline{(x^2)} - (\overline{x})^2,$$

since $\overline{x.\overline{x}} = (\overline{x})^2$. Similarly, the *sample covariance* of x and y is defined as the average of $(x-\overline{x})(y-\overline{y})$, written s_{xy}. So

$$\begin{aligned} s_{xy} &= \overline{(x-\overline{x})(y-\overline{y})} = \overline{xy - x.\overline{y} - \overline{x}.y + \overline{x}.\overline{y}} \\ &= \overline{(xy)} - \overline{x}.\overline{y} - \overline{x}.\overline{y} + \overline{x}.\overline{y} = \overline{(xy)} - \overline{x}.\overline{y}. \end{aligned}$$

Thus the slope b is given by the *sample correlation coefficient*

$$b = s_{xy}/s_{xx},$$

the ratio of the sample covariance to the sample x-variance. Using the alternative 'sum of squares' notation

$$S_{xx} := \sum_{i=1}^{n} (x_i - \overline{x})^2, \qquad S_{xy} := \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}),$$

$$b = S_{xy}/S_{xx}, \qquad a = \overline{y} - b\overline{x}.$$

The line – the *least-squares line* that we have fitted – is $y = a + bx$ with this a and b, or

$$y - \overline{y} = b(x - \overline{x}), \qquad b = s_{xy}/s_{xx} = S_{xy}/S_{xx}. \qquad (SRL)$$

It is called the *sample regression line*, for reasons which will emerge later.

Notice that the line goes through the point $(\overline{x}, \overline{y})$ – the *centroid*, or centre of mass, of the scatter diagram $(x_1, y_1), \ldots, (x_n, y_n)$.

Note 1.1

We will see later that if we assume that the errors are *independent* and identically distributed (which we abbreviate to iid) and normal, $N(0, \sigma^2)$ say, then these formulas for a and b also give the maximum likelihood estimates. Further, $100(1-\alpha)\%$ confidence intervals in this case can be calculated from points $\hat{a}$ and $\hat{b}$ as

$$\begin{aligned} a &= \hat{a} \pm t_{n-2}(1-\alpha/2)s\sqrt{\frac{\sum x_i^2}{nS_{xx}}}, \\ b &= \hat{b} \pm \frac{t_{n-2}(1-\alpha/2)s}{\sqrt{S_{xx}}}, \end{aligned}$$

where $t_{n-2}(1-\alpha/2)$ denotes the $1-\alpha/2$ quantile of the Student t distribution with $n-2$ degrees of freedom and s is given by

$$s = \sqrt{\frac{1}{n-2}\left(S_{yy} - \frac{S_{xy}^2}{S_{xx}}\right)}.$$

Example 1.2

We fit the line of best fit to model y = Height (in inches) based on x = Age (in years) for the following data:
x=(14, 13, 13, 14, 14, 12, 12, 15, 13, 12, 11, 14, 12, 15, 16, 12, 15, 11, 15),
y=(69, 56.5, 65.3, 62.8, 63.5, 57.3, 59.8, 62.5, 62.5, 59.0, 51.3, 64.3, 56.3, 66.5, 72.0, 64.8, 67.0, 57.5, 66.5).

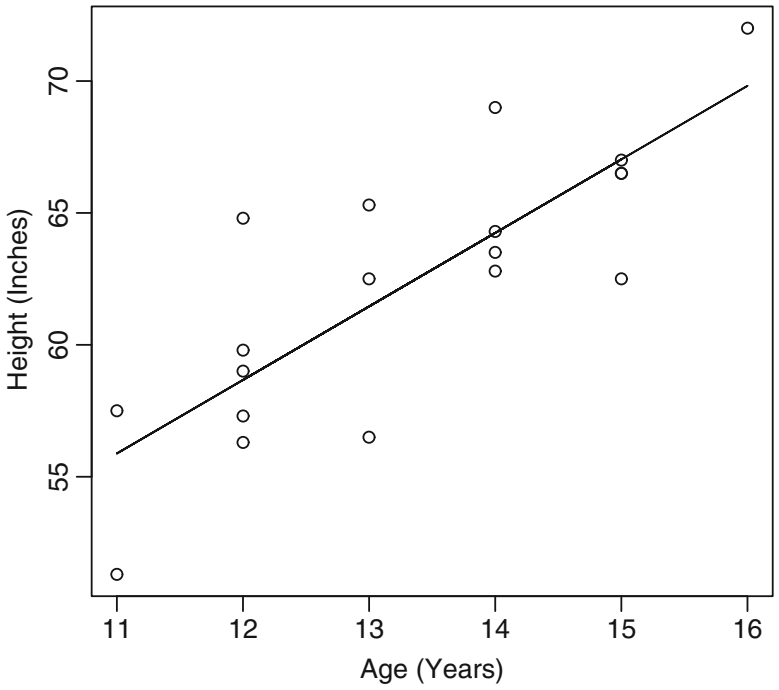

Figure 1.1 Scatter plot of the data in Example 1.2 plus fitted straight line

One may also calculate S_{xx} and S_{xy} as

$$\begin{aligned} S_{xx} &= \sum x_i y_i - n\overline{xy}, \\ S_{xy} &= \sum x_i^2 - n\overline{x}^2. \end{aligned}$$

Since $\sum x_i y_i = 15883, \bar{x} = 13.316, \bar{y} = 62.337, \sum x_i^2 = 3409$, $n = 19$, we have that

$$b = \frac{15883 - 19(13.316)(62.337)}{3409 - 19(13.316^2)} = 2.787 \text{ (3 d.p.)}.$$

Rearranging, we see that a becomes $62.33684 - 2.787156(13.31579) = 25.224$. This model suggests that the children are growing by just under three inches

per year. A plot of the observed data and the fitted straight line is shown in Figure 1.1 and appears reasonable, although some deviation from the fitted straight line is observed.

1.2.1 Correlation version

The *sample correlation coefficient* $r = r_{xy}$ is defined as

$$r = r_{xy} := \frac{s_{xy}}{s_x s_y},$$

the quotient of the sample covariance and the product of the sample standard deviations. Thus r is dimensionless, unlike the other quantities encountered so far. One has (see Exercise 1.1)

$$-1 \leq r \leq 1,$$

with equality if and only if (iff) all the points $(x_1, y_1), \ldots, (x_n, y_n)$ lie on a straight line. Using $s_{xy} = r_{xy} s_x s_y$ and $s_{xx} = s_x^2$, we may alternatively write the sample regression line as

$$y - \overline{y} = b(x - \overline{x}), \qquad b = r_{xy} s_y / s_x. \tag{SRL}$$

Note also that the slope b has the same sign as the sample covariance and sample correlation coefficient. These will be approximately the population covariance and correlation coefficient for large n (see below), so will have slope near zero when y and x are uncorrelated – in particular, when they are independent, and will have positive (negative) slope when x, y are positively (negatively) correlated.

We now have *five* parameters in play: two means, μ_x and μ_y, two variances σ_x^2 and σ_y^2 (or their square roots, the standard deviations σ_x and σ_y), and one correlation, ρ_{xy}. The two means are measures of *location*, and serve to identify the point – (μ_x, μ_y), or its sample counterpart, $(\overline{x}, \overline{y})$ – which serves as a natural choice of *origin*. The two variances (or standard deviations) are measures of *scale*, and serve as natural units of length along coordinate axes centred at this choice of origin. The correlation, which is dimensionless, serves as a measure of *dependence*, or *linkage*, or *association*, and indicates how closely y depends on x – that is, how informative x is about y. Note how differently these behave under affine transformations, $x \mapsto ax + b$. The mean transforms linearly:

$$E(ax + b) = aEx + b;$$

the variance transforms by

$$\mathrm{var}(ax + b) = a^2 \mathrm{var}(x);$$

the correlation is unchanged – it is *invariant* under affine transformations.

1.2.2 Large-sample limit

When $x_1, \ldots, x_n$ are independent copies of a random variable x, and x has mean Ex, the Law of Large Numbers says that

$$\overline{x} \to Ex \qquad (n \to \infty).$$

See e.g. Haigh (2002), §6.3. There are in fact several versions of the Law of Large Numbers (LLN). The Weak LLN (or WLLN) gives convergence in probability (for which see e.g. Haigh (2002). The Strong LLN (or SLLN) gives convergence with probability one (or 'almost surely', or 'a.s.'); see Haigh (2002) for a short proof under stronger moment assumptions (fourth moment finite), or Grimmett and Stirzaker (2001), §7.5 for a proof under the minimal condition – existence of the mean. While one should bear in mind that the SLLN holds only off some exceptional set of probability zero, we shall feel free to state the result as above, with this restriction understood. Note the content of the SLLN: thinking of a random variable as its mean plus an error, *independent errors tend to cancel when one averages*. This is essentially what makes Statistics work: the basic technique in Statistics is *averaging*.

All this applies similarly with x replaced by y, x^2, y^2, xy, when all these have means. Then

$$s_x^2 = s_{xx} = \overline{x^2} - \left(\overline{x}^2\right) \to E\left(x^2\right) - (Ex)^2 = var(x),$$

the population variance – also written $\sigma_x^2 = \sigma_{xx}$ – and

$$s_{xy} = \overline{xy} - \overline{x}.\overline{y} \to E(xy) - Ex.Ey = \mathrm{cov}(x, y),$$

the population covariance – also written σ_{xy}. Thus as the sample size n increases, the sample regression line

$$y - \overline{y} = b(x - \overline{x}), \qquad b = s_{xy}/s_{xx}$$

tends to the line

$$y - Ey = \beta(x - Ex), \qquad \beta = \sigma_{xy}/\sigma_{xx}. \qquad (PRL)$$

This – its population counterpart – is accordingly called the *population regression line*.

Again, there is a version involving correlation, this time the *population correlation coefficient*

$$\rho = \rho_{xy} := \frac{\sigma_{xy}}{\sigma_x \sigma_y} :$$

$$y - Ey = \beta(x - Ex), \qquad \beta = \rho_{xy}\sigma_y/\sigma_x. \qquad (PRL)$$

Note 1.3

The following illustration is worth bearing in mind here. Imagine a school Physics teacher, with a class of twenty pupils; they are under time pressure revising for an exam, he is under time pressure marking. He divides the class into ten pairs, gives them an experiment to do over a double period, and withdraws to do his marking. Eighteen pupils gang up on the remaining two, the best two in the class, and threaten them into agreeing to do the experiment for them. This pair's results are then stolen by the others, who to disguise what has happened change the last two significant figures, say. Unknown to all, the best pair's instrument was dropped the previous day, and was reading way too high – so the *first* significant figures in their results, and hence all the others, were wrong. In this example, the insignificant 'rounding errors' in the last significant figures *are* independent and *do* cancel – but no significant figures are correct for any of the ten pairs, because of the strong dependence between the ten readings. Here the tenfold replication is only apparent rather than real, and is valueless. We shall see more serious examples of correlated errors in Time Series in §9.4, where high values tend to be succeeded by high values, and low values tend to be succeeded by low values.

1.3 The origins of regression

The modern era in this area was inaugurated by Sir Francis Galton (1822–1911), in his book *Hereditary genius – An enquiry into its laws and consequences* of 1869, and his paper 'Regression towards mediocrity in hereditary stature' of 1886. Galton's real interest was in intelligence, and how it is inherited. But intelligence, though vitally important and easily recognisable, is an elusive concept – human ability is infinitely variable (and certainly multi–dimensional!), and although numerical measurements of general ability exist (intelligence quotient, or IQ) and can be measured, they can serve only as a proxy for intelligence itself. Galton had a passion for measurement, and resolved to study something that *could* be easily measured; he chose human height. In a classic study, he measured the heights of 928 adults, born to 205 sets of parents. He took the average of the father's and mother's height ('mid-parental height') as the predictor variable x, and height of offspring as response variable y. (Because men are statistically taller than women, one needs to take the gender of the offspring into account. It is conceptually simpler to treat the sexes separately – and focus on sons, say – though Galton actually used an adjustment factor to compensate for women being shorter.) When he displayed his data in tabular form, Galton noticed that it showed *elliptical contours* – that is, that squares in the

(x, y)-plane containing equal numbers of points seemed to lie approximately on ellipses. The explanation for this lies in the *bivariate normal distribution*; see §1.5 below. What is most relevant here is Galton's interpretation of the sample and population regression lines (SRL) and (PRL). In (PRL), σ_x and σ_y are measures of *variability* in the parental and offspring generations. There is no reason to think that variability of height is changing (though *mean* height has visibly increased from the first author's generation to his children). So (at least to a first approximation) we may take these as equal, when (PRL) simplifies to

$$y - Ey = \rho_{xy}(x - Ex). \qquad (PRL)$$

Hence Galton's celebrated interpretation: for every inch of height above (or below) the average, the parents transmit to their children *on average* ρ inches, where ρ is the population correlation coefficient between parental height and offspring height. A further generation will introduce a further factor ρ, so the parents will transmit – again, *on average* – ρ^2 inches to their grandchildren. This will become ρ^3 inches for the great-grandchildren, and so on. Thus for every inch of height above (or below) the average, the parents transmit to their descendants after n generations *on average* ρ^n inches of height. Now

$$0 < \rho < 1$$

($\rho > 0$ as the genes for tallness or shortness are transmitted, and parental and offspring height are positively correlated; $\rho < 1$ as $\rho = 1$ would imply that parental height is *completely* informative about offspring height, which is patently not the case). So

$$\rho^n \to 0 \qquad (n \to \infty):$$

the effect of each inch of height above or below the mean is damped out with succeeding generations, and disappears in the limit. Galton summarised this as 'Regression towards mediocrity in hereditary stature', or more briefly, *regression towards the mean* (Galton originally used the term *reversion* instead, and indeed the term *mean reversion* still survives). This explains the name of the whole subject.

Note 1.4

1. We are more interested in intelligence than in height, and are more likely to take note of the corresponding conclusion for intelligence.

2. Galton found the conclusion above depressing – as may be seen from his use of the term mediocrity (to call someone average may be factual, to call

them mediocre is disparaging). Galton had a typically Victorian enthusiasm for *eugenics* – the improvement of the race. Indeed, the senior chair in Statistics in the UK (or the world), at University College London, was originally called the Galton Chair of Eugenics. This was long before the term eugenics became discredited as a result of its use by the Nazis.

3. The above assumes *random mating*. This is a reasonable assumption to make for height: height is not particularly important, while choice of mate is very important, and so few people choose their life partner with height as a prime consideration. Intelligence is quite another matter: intelligence *is* important. Furthermore, we can all observe the tendency of intelligent people to prefer and seek out each others' company, and as a natural consequence, to mate with them preferentially. This is an example of *assortative mating*. It is, of course, the best defence for intelligent people who wish to transmit their intelligence to posterity against regression to the mean. What this in fact does is to stratify the population: intelligent assortative maters are still subject to regression to the mean, but it is to a different mean – not the general population mean, but the mean among the social group in question – graduates, the learned professions or whatever.

1.4 Applications of regression

Before turning to the underlying theory, we pause to mention a variety of contexts in which regression is of great practical use, to illustrate why the subject is worth study in some detail.

1. *Examination scores.*

   This example may be of particular interest to undergraduates! The context is that of an elite institution of higher education. The proof of elite status is an excess of well-qualified applicants. These have to be ranked in merit order in some way. Procedures differ in detail, but in broad outline all relevant pieces of information – A Level scores, UCAS forms, performance in interview, admissions officer's assessment of potential etc. – are used, coded in numerical form and then combined according to some formula to give a numerical score. This is used as the predictor variable x, which measures the quality of *incoming students*; candidates are ranked by score, and places filled on merit, top down, until the quota is reached. At the end of the course, students graduate, with a classified degree. The task of the Examiners' Meeting is to award classes of degree. While at the margin

this involves detailed discussion of individual cases, it is usual to table among the papers for the meeting a numerical score for each candidate, obtained by combining the relevant pieces of information – performance on the examinations taken throughout the course, assessed course-work etc. – into a numerical score, again according to some formula. This score is y, the response variable, which measures the quality of *graduating students*. The question is how well the institution picks students – that is, how good a *predictor* of eventual performance y the incoming score x is. Of course, the most important single factor here is the innate ability and personality of the individual student, plus the quality of their school education. These will be powerfully influential on *both* x *and* y. But they are not directly measurable, while x is, so x serves here as a proxy for them. These underlying factors remain unchanged during the student's study, and are the most important determinant of y. However, other factors intervene. Some students come to university if anything under-prepared, grow up and find their feet, and get steadily better. By contrast, some students arrive if anything over-prepared (usually as a result of expensively purchased 'cramming') and revert to their natural level of performance, while some others arrive studious and succumb to the temptations of wine, women (or men) and song, etc. The upshot is that, while x serves as a good proxy for the ability and intelligence which really matter, there is a considerable amount of unpredictability, or *noise*, here.

The question of how well institutions pick students is of great interest, to several kinds of people:

a) admissions tutors to elite institutions of higher education,

b) potential students and their parents,

c) the state, which largely finances higher education (note that in the UK in recent years, a monitoring body, OFFA – the Office for Fair Access, popularly referred to as Oftoff – has been set up to monitor such issues).

2. *Height.*

Although height is of limited importance, proud parents are consumed with a desire to foresee the future for their offspring. There are various rules of thumb for predicting the eventual future height as an adult of a small child (roughly speaking: measure at age two and double – the details vary according to sex). This is of limited practical importance nowadays, but we note in passing that some institutions or professions (the Brigade of Guards etc.) have upper and lower limits on heights of entrants.

3. *Athletic Performance*

   a) *Distance.*

      Often an athlete competes at two different distances. These may be half-marathon and marathon (or ten miles and half-marathon) for the longer distances, ten kilometres and ten miles – or 5k and 10k – for the middle distances; for track, there are numerous possible pairs: 100m and 200m, 200m and 400m, 400m and 800m, 800m and 1500m, 1500m and 5,000m, 5,000m and 10,000m. In each case, what is needed – by the athlete, coach, commentator or follower of the sport – is an indication of how informative a time x over one distance is on time y over the other.

   b) *Age.*

      An athlete's career has three broad phases. In the first, one completes growth and muscle development, and develops cardio-vascular fitness as the body reacts to the stresses of a training regime of running. In the second, the plateau stage, one attains one's best performances. In the third, the body is past its best, and deteriorates gradually with age. Within this third phase, age is actually a good predictor: the Rule of Thumb for ageing marathon runners (such as the first author) is that every extra year costs about an extra minute on one's marathon time.

4. *House Prices and Earnings.*

   Under normal market conditions, the most important single predictor variable for house prices is earnings. The second most important predictor variable is interest rates: earnings affect the purchaser's ability to raise finance, by way of mortgage, interest rates affect ability to pay for it by servicing the mortgage. This example, incidentally, points towards the use of two predictor variables rather than one, to which we shall return below. (Under the abnormal market conditions that prevail following the Crash of 2008, or Credit Crunch, the two most relevant factors are availability of mortgage finance (which involves liquidity, credit, etc.), and confidence (which involves economic confidence, job security, unemployment, etc.).)

1.5 The Bivariate Normal Distribution

Recall two of the key ingredients of statistics:

(a) *The normal distribution*, $N(\mu, \sigma^2)$:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\},$$

which has mean $EX = \mu$ and variance $varX = \sigma^2$.

(b) *Linear regression by the method of least squares* – above.

This is for *two-dimensional* (or bivariate) data $(X_1, Y_1), \ldots, (X_n, Y_n)$. Two questions arise:

(i) Why linear?

(ii) What (if any) is the two-dimensional analogue of the normal law?

Writing

$$\phi(x) := \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}$$

for the standard normal density, $\int$ for $\int_{-\infty}^{\infty}$, we shall need

(i) *recognising normal integrals:*

a) $\int \phi(x)dx = 1$ (*'normal density'*),

b) $\int x\phi(x)dx = 0$ (*'normal mean'* - or, 'symmetry'),

c) $\int x^2\phi(x)dx = 1$ (*'normal variance'*),

(ii) *completing the square*: as for solving quadratic equations!

In view of the work above, we need an analogue in *two* dimensions of the normal distribution $N(\mu, \sigma^2)$ in one dimension. Just as in one dimension we need *two* parameters, μ and σ, in two dimensions we must expect to need *five*, by the above.

Consider the following bivariate density:

$$f(x, y) = c \exp\left\{-\frac{1}{2}Q(x, y)\right\},$$

where c is a constant, Q a positive definite quadratic form in x and y. Specifically:

$$\begin{aligned} c &= \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}, \\ Q &= \frac{1}{1-\rho^2}\left[\Big(\frac{x-\mu_1}{\sigma_1}\Big)^2 - 2\rho\Big(\frac{x-\mu_1}{\sigma_1}\Big)\Big(\frac{y-\mu_2}{\sigma_2}\Big) + \Big(\frac{y-\mu_2}{\sigma_2}\Big)^2\right]. \end{aligned}$$

Here $\sigma_i > 0$, μ_i are real, $-1 < \rho < 1$. Since f is clearly non-negative, to show that f is a (probability density) function (in two dimensions), it suffices to show that f integrates to 1:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x,y)\,dx\,dy = 1, \qquad \text{or} \qquad \int\int f = 1.$$

Write

$$f_1(x) := \int_{-\infty}^{\infty} f(x,y)\,dy, \qquad f_2(y) := \int_{-\infty}^{\infty} f(x,y)\,dx.$$

Then to show $\int\int f = 1$, we need to show $\int_{-\infty}^{\infty} f_1(x)\,dx = 1$ (or $\int_{-\infty}^{\infty} f_2(y)\,dy = 1$). Then f_1, f_2 are densities, in *one* dimension. If $f(x,y) = f_{X,Y}(x,y)$ is the *joint* density of *two* random variables X, Y, then $f_1(x)$ is the density $f_X(x)$ of X, $f_2(y)$ the density $f_Y(y)$ of Y (f_1, f_2, or f_X, f_Y, are called the *marginal* densities of the *joint* density f, or $f_{X,Y}$).

To perform the integrations, we have to *complete the square*. We have the algebraic identity

$$(1-\rho^2)Q \equiv \Big[\Big(\frac{y-\mu_2}{\sigma_2}\Big) - \rho\Big(\frac{x-\mu_1}{\sigma_1}\Big)\Big]^2 + \big(1-\rho^2\big)\Big(\frac{x-\mu_1}{\sigma_1}\Big)^2$$

(reducing the number of occurrences of y to 1, as we intend to integrate out y first). Then (taking the terms free of y out through the y-integral)

$$f_1(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2\right)}{\sigma_1\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{1}{\sigma_2\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left(\frac{-\frac{1}{2}(y-c_x)^2}{\sigma_2^2\,(1-\rho^2)}\right) dy, \tag{$*$}$$

where

$$c_x := \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1).$$

The integral is 1 ('normal density'). So

$$f_1(x) = \frac{\exp\left(-\frac{1}{2}(x-\mu_1)^2/\sigma_1^2\right)}{\sigma_1\sqrt{2\pi}},$$

which integrates to 1 ('normal density'), proving

Fact 1. $f(x, y)$ is a joint density function (two-dimensional), with marginal density functions $f_1(x), f_2(y)$ (one-dimensional).

So we can write

$$f(x, y) = f_{X,Y}(x, y), \qquad f_1(x) = f_X(x), \qquad f_2(y) = f_Y(y).$$

Fact 2. X, Y are normal: X is $N(\mu_1, \sigma_1^2)$, Y is $N(\mu_2, \sigma_2^2)$. For, we showed $f_1 = f_X$ to be the $N(\mu_1, \sigma_1^2)$ density above, and similarly for Y by symmetry.
Fact 3. $EX = \mu_1, EY = \mu_2, \text{var } X = \sigma_1^2, \text{var } Y = \sigma_2^2$.

This identifies four out of the five parameters: two means μ_i, two variances σ_i^2.

Next, recall the definition of conditional probability:

$$P(A|B) := P(A \cap B)/P(B).$$

In the *discrete* case, if X, Y take possible values x_i, y_j with probabilities $f_X(x_i), f_Y(y_j)$, (X, Y) takes possible values (x_i, y_j) with corresponding probabilities $f_{X,Y}(x_i, y_j)$:

$$f_X(x_i) = P(X = x_i) = \Sigma_j P(X = x_i, Y = y_j) = \Sigma_j f_{X,Y}(x_i, y_j).$$

Then the *conditional* distribution of Y *given* $X = x_i$ is

$$f_{Y|X}(y_j|x_i) = \frac{P(Y = y_j, X = x_i)}{P(X = x_i)} = \frac{f_{X,Y}(x_i, y_j)}{\sum_j f_{X,Y}(x_i, y_j)},$$

and similarly with X, Y interchanged.

In the *density* case, we have to replace *sums* by *integrals*. Thus the conditional *density* of Y *given* $X = x$ is (see e.g. Haigh (2002), Def. 4.19, p. 80)

$$f_{Y|X}(y|x) := \frac{f_{X,Y}(x, y)}{f_X(x)} = \frac{f_{X,Y}(x, y)}{\int_{-\infty}^{\infty} f_{X,Y}(x, y)\ dy}.$$

Returning to the bivariate normal:
Fact 4. The conditional distribution of y given $X = x$ is

$$N\left(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \quad \sigma_2^2\left(1 - \rho^2\right)\right).$$

Proof

Go back to completing the square (or, return to $(*)$ with $\int$ and dy deleted):

$$f(x, y) = \frac{\exp\left\{-\frac{1}{2}\left(x - \mu_1\right)^2/\sigma_1^2\right\}}{\sigma_1\sqrt{2\pi}} \cdot \frac{\exp\left\{-\frac{1}{2}\left(y - c_x\right)^2/\left(\sigma_2^2\left(1 - \rho^2\right)\right)\right\}}{\sigma_2\sqrt{2\pi}\sqrt{1 - \rho^2}}.$$

The first factor is $f_1(x)$, by Fact 1. So, $f_{Y|X}(y|x) = f(x,y)/f_1(x)$ is the second factor:

$$f_{Y|X}(y|x) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}} \exp\left\{\frac{-(y-c_x)^2}{2\sigma_2^2(1-\rho^2)}\right\},$$

where c_x is the linear function of x given below $(*)$. □

This not only completes the proof of Fact 4 but gives
Fact 5. The conditional mean $E(Y|X=x)$ is *linear* in x:

$$E(Y|X=x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1).$$

Note 1.5

1. This simplifies when X and Y are equally variable, $\sigma_1 = \sigma_2$:

   $$E(Y|X=x) = \mu_2 + \rho(x-\mu_1)$$

   (recall $EX = \mu_1, EY = \mu_2$). Recall that in Galton's height example, this says: for every inch of mid-parental height above/below the average, $x-\mu_1$, the parents pass on to their child, *on average*, ρ inches, and continuing in this way: *on average*, after n generations, each inch above/below average becomes *on average* ρ^n inches, and $\rho^n \to 0$ as $n \to \infty$, giving *regression towards the mean.*

2. This line is the population regression line (PRL), the population version of the sample regression line (SRL).

3. The relationship in Fact 5 can be generalised (§4.5): a population regression function – more briefly, a regression – is a *conditional mean.*
   This also gives

Fact 6. The conditional variance of Y given $X = x$ is

$$var(Y|X=x) = \sigma_2^2\left(1-\rho^2\right).$$

Recall (Fact 3) that the variability (= variance) of Y is $\mathrm{var}Y = \sigma_2^2$. By Fact 5, the variability remaining in Y when X is given (i.e., not accounted for by knowledge of X) is $\sigma_2^2(1-\rho^2)$. Subtracting, the variability of Y which *is* accounted for by knowledge of X is $\sigma_2^2\rho^2$. That is, ρ^2 is the *proportion of the*

variability of Y accounted for by knowledge of X. So ρ is a measure of the *strength of association* between Y and X.

Recall that the *covariance* is defined by

$$\begin{aligned}\mathrm{cov}(X,Y) &:= E[(X-EX)(Y-EY)] = E[(X-\mu_1)(Y-\mu_2)],\\ &= E(XY)-(EX)(EY),\end{aligned}$$

and the *correlation coefficient* ρ, or $\rho(X,Y)$, defined by

$$\rho = \rho(X,Y) := \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}X}\sqrt{\mathrm{var}Y}} = \frac{E[(X-\mu_1)(Y-\mu_2)]}{\sigma_1\sigma_2}$$

is the usual measure of the strength of association between X and Y ($-1 \le \rho \le 1$; $\rho = \pm 1$ iff one of X, Y is a function of the other). That this is consistent with the use of the symbol ρ for a parameter in the density $f(x,y)$ is shown by the fact below.

Fact 7. If $(X,Y)^T$ is bivariate normal, the correlation coefficient of X, Y is ρ.

Proof

$$\rho(X,Y) := E\left[\left(\frac{X-\mu_1}{\sigma_1}\right)\left(\frac{Y-\mu_2}{\sigma_2}\right)\right] = \int\int \left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) f(x,y)\,dx\,dy.$$

Substitute for $f(x,y) = c\exp(-\frac{1}{2}Q)$, and make the change of variables $u := (x-\mu_1)/\sigma_1$, $v := (y-\mu_2)/\sigma_2$:

$$\rho(X,Y) = \frac{1}{2\pi\sqrt{1-\rho^2}} \int\int uv \exp\left(\frac{-\left[u^2-2\rho uv+v^2\right]}{2(1-\rho^2)}\right)\, du\; dv.$$

Completing the square as before, $[u^2-2\rho uv+v^2] = (v-\rho u)^2 + (1-\rho^2)u^2$. So

$$\rho(X,Y) = \frac{1}{\sqrt{2\pi}} \int u \exp\left(-\frac{u^2}{2}\right) du . \frac{1}{\sqrt{2\pi}\sqrt{1-\rho^2}} \int v \exp\left(-\frac{(v-\rho u)^2}{2(1-\rho^2)}\right) dv.$$

Replace v in the inner integral by $(v-\rho u)+\rho u$, and calculate the two resulting integrals separately. The first is zero ('normal mean', or symmetry), the second is ρu ('normal density'). So

$$\rho(X,Y) = \frac{1}{\sqrt{2\pi}} . \rho \int u^2 \exp\left(-\frac{u^2}{2}\right) du = \rho$$

('normal variance'), as required. □

This completes the identification of all five parameters in the bivariate normal distribution: two means μ_i, two variances σ_i^2, one correlation ρ.

Note 1.6

1. The above holds for $-1 < \rho < 1$; always, $-1 \leq \rho \leq 1$, by the Cauchy-Schwarz inequality (see e.g. Garling (2007) p.15, Haigh (2002) Ex 3.20 p.86, or Howie (2001) p.22 and Exercises 1.1-1.2). In the limiting cases $\rho = \pm 1$, one of X, Y is then a linear function of the other: $Y = aX + b$, say, as in the temperature example (Fahrenheit and Centigrade). The situation is not really two-dimensional: we can (and should) use only *one* of X and Y, reducing to a one-dimensional problem.

2. The slope of the regression line $y = c_x$ is $\rho\sigma_2/\sigma_1 = (\rho\sigma_1\sigma_2)/(\sigma_1^2)$, which can be written as $\mathrm{cov}(X,Y)/\mathrm{var}X = \sigma_{12}/\sigma_{11}$, or σ_{12}/σ_1^2: the line is

$$y - EY = \frac{\sigma_{12}}{\sigma_{11}}(x - EX).$$

   This is the *population* version (what else?!) of the *sample regression line*

$$y - \overline{y} = \frac{s_{XY}}{s_{XX}}(x - \overline{x}),$$

   familiar from linear regression.

   The case $\rho = \pm 1$ – apparently two-dimensional, but really one-dimensional – is *singular*; the case $-1 < \rho < 1$ (genuinely two-dimensional) is *non-singular*, or (see below) *full rank*.

We note in passing

Fact 8. The bivariate normal law has *elliptical contours*.

For, the contours are $Q(x, y) = const$, which are ellipses (as Galton found).

Moment Generating Function (MGF). Recall (see e.g. Haigh (2002), §5.2) the definition of the moment generating function (MGF) of a random variable X. This is the function

$$M(t), \quad \text{or} \quad M_X(t) := E\exp\{tX\}$$

for t real, and such that the expectation (typically a summation or integration, which may be infinite) converges (absolutely). For X normal $N(\mu, \sigma^2)$,

$$M(t) = \frac{1}{\sigma\sqrt{2\pi}} \int e^{tx} \exp\left(-\frac{1}{2}(x-\mu)^2/\sigma^2\right)\, dx.$$

Change variable to $u := (x - \mu)/\sigma$:

$$M(t) = \frac{1}{\sqrt{2\pi}} \int \exp\left(\mu t + \sigma u t - \frac{1}{2}u^2\right)\, du.$$

Completing the square,

$$M(t) = e^{\mu t}\frac{1}{\sqrt{2\pi}}\int \exp\left(-\frac{1}{2}(u-\sigma t)^2\right)\ du.e^{\frac{1}{2}\sigma^2 t^2},$$

or $M_X(t) = \exp(\mu t + \frac{1}{2}\sigma^2 t^2)$ (recognising that the central term on the right is 1 – 'normal density') . So $M_{X-\mu}(t) = \exp(\frac{1}{2}\sigma^2 t^2)$. Then (check)

$$\mu = EX = M_X'(0), \text{var } X = E[(X-\mu)^2] = M_{X-\mu}''(0).$$

Similarly in the bivariate case: the MGF is

$$M_{X,Y}(t_1,t_2) := E\exp(t_1X + t_2Y).$$

In the bivariate normal case:

$$\begin{aligned} M(t_1,t_2) &= E(\exp(t_1X+t_2Y)) = \int\int \exp(t_1x+t_2y)f(x,y)\ dx\ dy \\ &= \int \exp(t_1x)f_1(x)\ dx \int \exp(t_2y)f(y|x)\ dy. \end{aligned}$$

The inner integral is the MGF of $Y|X=x$, which is $N(c_x, \sigma_2^2, (1-\rho^2))$, so is $\exp(c_xt_2 + \frac{1}{2}\sigma_2^2(1-\rho^2)t_2^2)$. By Fact 5

$$c_xt_2 = [\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x-\mu_1)]t_2,$$

so $M(t_1,t_2)$ is equal to

$$\exp\left(t_2\mu_2 - t_2\frac{\sigma_2}{\sigma_1}\mu_1 + \frac{1}{2}\sigma_2^2\left(1-\rho^2\right)t_2^2\right)\int \exp\left(\left[t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}\right]x\right)f_1(x)\ dx.$$

Since $f_1(x)$ is $N(\mu_1,\sigma_1^2)$, the inner integral is a normal MGF, which is thus

$$\exp(\mu_1[t_1 + t_2\rho\frac{\sigma_2}{\sigma_1}] + \frac{1}{2}\sigma_1^2[\ldots]^2).$$

Combining the two terms and simplifying, we obtain
Fact 9. The joint MGF is

$$M_{X,Y}(t_1,t_2) = M(t_1,t_2) = \exp\left(\mu_1t_1 + \mu_2t_2 + \frac{1}{2}\left[\sigma_1^2t_1^2 + 2\rho\sigma_1\sigma_2t_1t_2 + \sigma_2^2t_2^2\right]\right).$$

Fact 10. X, Y are independent iff $\rho = 0$.

Proof

For densities: X, Y are independent iff the joint density $f_{X,Y}(x,y)$ *factorises* as the *product* of the marginal densities $f_X(x).f_Y(y)$ (see e.g. Haigh (2002), Cor. 4.17).

For MGFs: X, Y are independent iff the joint MGF $M_{X,Y}(t_1,t_2)$ *factorises* as the *product* of the marginal MGFs $M_X(t_1).M_Y(t_2)$. From Fact 9, this occurs iff $\rho = 0$. □

Note 1.7

1. X, Y independent implies X, Y uncorrelated ($\rho = 0$) in general (when the correlation exists). The converse is *false* in general, but *true*, by Fact 10, in the bivariate normal case.

2. *Characteristic functions (CFs).* The *characteristic function*, or CF, of X is

$$\phi_X(t) := E(e^{itX}).$$

Compared to the MGF, this has the drawback of involving complex numbers, but the great advantage of always existing for t real. Indeed,

$$|\phi_X(t)| = \left|E(e^{itX})\right| \leq E\left|\left(e^{itX}\right)\right| = E1 = 1.$$

By contrast, the expectation defining the MGF $M_X(t)$ may diverge for some real t (as we shall see in §2.1 with the chi-square distribution.) For background on CFs, see e.g. Grimmett and Stirzaker (2001) §5.7. For our purposes one may pass from MGF to CF by formally replacing t by it (though one actually needs analytic continuation – see e.g. Copson (1935), §4.6 – or Cauchy's Theorem – see e.g. Copson (1935), §6.7, or Howie (2003), Example 9.19). Thus for the univariate normal distribution $N(\mu, \sigma^2)$ the CF is

$$\phi_X(t) = \exp\left\{i\mu t - \frac{1}{2}\sigma^2 t^2\right\}$$

and for the bivariate normal distribution the CF of X, Y is

$$\phi_{X,Y}(t_1, t_2) = \exp\left\{i\mu_1 t_1 + i\mu_2 t_2 - \frac{1}{2}\left[\sigma_1^2 t_1^2 + 2\rho\sigma_1\sigma_2 t_1 t_2 + \sigma_2 t_2^2\right]\right\}.$$

1.6 Maximum Likelihood and Least Squares

By Fact 4, the conditional distribution of y given $X = x$ is

$$N(\mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1), \quad \sigma_2^2(1 - \rho^2)).$$

Thus y is decomposed into two components, a *linear trend* in x – the systematic part – and a normal error, with mean zero and constant variance – the random part. Changing the notation, we can write this as

$$y = a + bx + \epsilon, \qquad \epsilon \sim \ N(0, \sigma^2).$$

With n values of the predictor variable x, we can similarly write

$$y_i = a + bx_i + \epsilon_i, \qquad \epsilon_i \sim \ N(0, \sigma^2).$$

To complete the specification of the model, we need to specify the dependence or correlation structure of the errors $\epsilon_1, \ldots, \epsilon_n$. This can be done in various ways (see Chapter 4 for more on this). Here we restrict attention to the simplest and most important case, where the errors ϵ_i are iid:

$$y_i = a + bx_i + \epsilon_i, \qquad \epsilon_i \quad \text{iid} \quad N(0, \sigma^2). \tag{$*$}$$

This is the basic model for simple linear regression.

Since each y_i is now normally distributed, we can write down its density. Since the y_i are *independent*, the joint density of $y_1, \ldots, y_n$ *factorises* as the product of the marginal (separate) densities. This joint density, regarded as a function of the *parameters*, a, b and σ, is called the *likelihood*, L (one of many contributions by the great English statistician R. A. Fisher (1890-1962), later Sir Ronald Fisher, in 1912). Thus

$$\begin{aligned} L &= \frac{1}{\sigma^n (2\pi)^{\frac{1}{2}n}} \prod_{i=1}^n \exp\{-\frac{1}{2}(y_i - a - bx_i)^2/\sigma^2\} \\ &= \frac{1}{\sigma^n (2\pi)^{\frac{1}{2}n}} \exp\{-\frac{1}{2}\sum_{i=1}^n (y_i - a - bx_i)^2/\sigma^2\}. \end{aligned}$$

Fisher suggested choosing as our estimates of the parameters the values that maximise the likelihood. This is the Method of Maximum Likelihood; the resulting estimators are the maximum likelihood estimators or MLEs. Now maximising the likelihood L and maximising its logarithm $\ell := \log L$ are the same, since the function log is increasing. Since

$$\ell := \log L = -\frac{1}{2} n \log 2\pi - n \log \sigma - \frac{1}{2}\sum_{i=1}^n (y_i - a - bx_i)^2/\sigma^2,$$

so far as maximising with respect to a and b are concerned (leaving σ to one side for the moment), this is the same as minimising the sum of squares $SS := \sum_{i=1}^n (y_i - a - bx_i)^2$ – just as in the Method of Least Squares. Summarising:

Theorem 1.8

For the normal model $(*)$, the Method of Least Squares and the Method of Maximum Likelihood are equivalent ways of estimating the parameters a and b.

It is interesting to note here that the Method of Least Squares of Legendre and Gauss belongs to the early nineteenth century, whereas Fisher's Method of Maximum Likelihood belongs to the early twentieth century. For background on the history of statistics in that period, and an explanation of the 'long pause' between least squares and maximum likelihood, see Stigler (1986).

There remains the estimation of the parameter σ, equivalently the variance σ^2. Using maximum likelihood as above gives

$$\partial\ell/\partial\sigma = \frac{-n}{\sigma} + \frac{1}{\sigma^3}\sum_{i=1}^{n}(y_i - a - bx_i)^2 = 0,$$

or

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2.$$

At the maximum, a and b have their maximising values $\hat{a}$, $\hat{b}$ as above, and then the maximising value $\hat{\sigma}$ is given by

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 = \frac{1}{n}\sum_{1}^{n}(y_i - \hat{y}_i)^2.$$

Note that the sum of squares SS above involves unknown parameters, a and b. Because these are unknown, one cannot calculate this sum of squares numerically from the data. In the next section, we will meet other sums of squares, which can be calculated from the data – that is, which are functions of the data, or *statistics*. Rather than proliferate notation, we will again denote the largest of these sums of squares by SS; we will then break this down into a sum of smaller sums of squares (giving a *sum of squares decomposition*). In Chapters 3 and 4, we will meet multidimensional analogues of all this, which we will handle by matrix algebra. It turns out that all sums of squares will be expressible as quadratic forms in normal variates (since the parameters, while unknown, are constant, the distribution theory of sums of squares with and without unknown parameters is the same).

1.7 Sums of Squares

Recall the sample regression line in the form

$$y = \overline{y} + b(x - \overline{x}), \qquad b = s_{xy}/s_{xx} = S_{xy}/S_{xx}. \qquad (SRL)$$

We now ask how much of the variation in y is accounted for by knowledge of x – or, as one says, by regression. The data are y_i. The fitted values are $\hat{y}_i$, the left-hand sides above with x on the right replaced by x_i. Write

$$y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}),$$

square both sides and add. On the left, we get

$$SS := \sum_{i=1}^{n} (y_i - \overline{y})^2,$$

the *total sum of squares* or *sum of squares* for short. On the right, we get three terms:

$$SSR := \sum_i (\hat{y}_i - \overline{y})^2,$$

which we call the *sum of squares for regression*,

$$SSE := \sum_i (y_i - \hat{y}_i)^2,$$

the *sum of squares for error* (since this sum of squares measures the errors between the fitted values on the regression line and the data), and a cross term

$$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = n \frac{1}{n} \sum_i (y_i - \hat{y}_i)(\hat{y}_i - \overline{y}) = n.\overline{(y - \hat{y})(y - \overline{y})}.$$

By (SRL), $\hat{y}_i - \overline{y} = b(x_i - \overline{x})$ with $b = S_{xy}/S_{xx} = S_{xy}/S_x^2$, and

$$y_i - \hat{y} = (y_i - \overline{y}) - b(x_i - \overline{x}).$$

So the right above is n times

$$\frac{1}{n} \sum_i b(x_i - \overline{x})[(y_i - \overline{y}) - b(x_i - \overline{x})] = bS_{xy} - b^2 S_x^2 = b\left(S_{xy} - bS_x^2\right) = 0,$$

as $b = S_{xy}/S_x^2$. Combining, we have

Theorem 1.9

$$SS = SSR + SSE.$$

In terms of the sample correlation coefficient r^2, this yields as a corollary

Theorem 1.10

$$r^2 = SSR/SS, \qquad 1 - r^2 = SSE/SS.$$

Proof

It suffices to prove the first.

$$\frac{SSR}{SS} = \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{\sum b^2(x_i - \overline{x})^2}{\sum(y_i - \overline{y})^2} = \frac{b^2 S_x^2}{S_y^2} = \frac{S_{xy}^2}{S_x^4} . \frac{S_x^2}{S_y^2} = \frac{S_{xy}^2}{S_x^2 S_y^2} = r^2,$$

as $b = S_{xy}/S_x^2$. □

The interpretation is that $r^2 = SSR/SS$ is the proportion of variability in y accounted for by knowledge of x, that is, by regression (and $1 - r^2 = SSE/SS$ is that unaccounted for by knowledge of x, that is, by error). This is just the sample version of what we encountered in §1.5 on the bivariate normal distribution, where (see below Fact 6 in §1.5) ρ^2 has the interpretation of the proportion of variability in y accounted for by knowledge of x. Recall that r^2 tends to ρ^2 in the large-sample limit, by the Law of Large Numbers, so the population theory of §1.5 is the large-sample limit of the sample theory here.

Example 1.11

We wish to predict y, winning speeds (mph) in a car race, given the year x, by a linear regression. The data for years one to ten are y=(140.3, 143.1, 147.4, 151.4, 144.3, 151.2, 152.9, 156.9, 155.7, 157.7). The estimates for a and b now become $\hat{a} = 139.967$ and $\hat{b} = 1.841$. Assuming normally distributed errors in our regression model means that we can now calculate confidence intervals for the parameters and express a level of uncertainty around these estimates. In this case the formulae for 95% confidence intervals give (135.928, 144.005) for a and (1.190, 2.491) for b.

Distribution theory. Consider first the case $b = 0$, when the slope is zero, there is no linear trend, and the y_i are identically distributed, $N(a, \sigma^2)$. Then $\overline{y}$ and $y_i - \overline{y}$ are also normally distributed, with zero mean. It is perhaps surprising, but true, that $\sum(y_i - \overline{y})^2$ and $\overline{y}$ are independent; we prove this in §2.5 below. The distribution of the quadratic form $\sum(y_i - \overline{y})^2$ involves the *chi-square distribution*; see §2.1 below. In this case, SSR and SSE are independent chi-square variates, and $SS = SSR + SSE$ is an instance of *chi-square decompositions*, which we meet in §3.5.

In the general case with the slope b non-zero, there is a linear trend, and a sloping regression line is more successful in explaining the data than a flat one. One quantifies this by using a ratio of sums of squares (ratio of independent chi-squares) that *increases* when the slope b is non-zero, so large values are evidence *against* zero slope. This statistic is an *F-statistic* (§2.3: F for Fisher). Such F-tests may be used to test a large variety of such *linear hypotheses* (Chapter 6).

When b is non-zero, the $y_i - \overline{y}$ are normally distributed as before, but with *non-zero* mean. Their sum of squares $\sum(y_i - \overline{y})^2$ then has a *non-central chi-square distribution*. The theory of such distributions is omitted here, but can be found in, e.g., Kendall and Stuart (1979), Ch. 24.

1.8 Two regressors

Suppose now that we have *two* regressor variables, u and v say, for the response variable y. Several possible settings have been prefigured in the discussion above:

1. *Height.*

   Galton measured the father's height u and the mother's height v in each case, before averaging to form the mid-parental height $x := (u+v)/2$. What happens if we use u and v in place of x?

2. *Predicting grain yields.*

   Here y is the grain yield after the summer harvest. Because the *price* that the grain will fetch is determined by the balance of supply and demand, and demand is fairly inflexible while supply is unpredictable, being determined largely by the weather, it is of great economic and financial importance to be able to *predict* grain yields in advance. The two most important predictors are the amount of rainfall (in cm, u say) and sunshine (in hours, v say) during the spring growing season. Given this information at the end of spring, how can we use it to best predict yield in the summer harvest? Of course, the actual harvest is still subject to events in the future, most notably the possibility of torrential rain in the harvest season flattening the crops. Note that for the sizeable market in grain futures, such predictions are highly price-sensitive information.

3. *House prices.*

   In the example above, house prices y depended on earnings u and interest rates v. We would expect to be able to get better predictions using both these as predictors than using either on its own.

4. *Athletics times.*

   We saw that both age and distance can be used separately; one ought to be able to do better by using them together.

5. *Timber.*

   The economic value of a tree grown for timber depends on the volume of usable timber when the tree has been felled and taken to the sawmill. When choosing which trees to fell, it is important to be able to estimate this volume without needing to fell the tree. The usual predictor variables here are girth (in cm, say – measured by running a tape-measure round the trunk at some standard height – one metre, say – above the ground) and height (measured by use of a surveyor's instrument and trigonometry).

With two regressors u and v and response variable y, given a sample of size n of points $(u_1, v_1, y_1), \ldots, (u_n, v_n, y_n)$ we have to fit a least-squares *plane* – that is, we have to choose parameters a, b, c to minimise the sum of squares

$$SS := \sum_{i=1}^{n} (y_i - c - au_i - bv_i)^2.$$

Taking $\partial SS/\partial c = 0$ gives

$$\sum_{i=1}^{n} (y_i - c - au_i - bv_i) = 0 : \qquad c = \overline{y} - a\overline{u} - b\overline{v}.$$

We rewrite SS as

$$SS = \sum_{i=1}^{n} [(y_i - \overline{y}) - a(u_i - \overline{u}) - b(v_i - \overline{v})]^2.$$

Then $\partial SS/\partial a = 0$ and $\partial SS/\partial b = 0$ give

$$\sum_{i=1}^{n} (u_i - \overline{u})[(y_i - \overline{y}) - a(u_i - \overline{u}) - b(v_i - \overline{v})] = 0,$$

$$\sum_{i=1}^{n} (v_i - \overline{v})[(y_i - \overline{y}) - a(u_i - \overline{u}) - b(v_i - \overline{v})] = 0.$$

Multiply out, divide by n to turn the sums into averages, and re-arrange using our earlier notation of sample variances and sample covariance: the above equations become

$$as_{uu} + bs_{uv} = s_{yu},$$

$$as_{uv} + bs_{vv} = s_{yv}.$$

These are the *normal equations* for a and b. The determinant is

$$s_{uu}s_{vv} - s_{uv}^2 = s_{uu}s_{vv}(1 - r_{uv}^2)$$

(since $r_{uv} := s_{uv}/(s_u s_v)$). This is non-zero iff $r_{uv} \neq \pm 1$ – that is, iff the points $(u_1, v_1), \ldots, (u_n, v_n)$ are not collinear – and this is the condition for the normal equations to have a unique solution.

The extension to three or more regressors may be handled in just the same way: with p regressors we obtain p normal equations. The general case is best handled by the matrix methods of Chapter 3.

Note 1.12

As with the linear regression case, under the assumption of iid $N(0, \sigma^2)$ errors these formulas for a and b also give the maximum likelihood estimates. Further,

$100(1-\alpha)\%$ confidence intervals can be returned routinely using standard software packages, and in this case can be calculated as

$$c = \hat{c} \pm t_{n-3}(1-\alpha/2)s\sqrt{\frac{\sum u_i^2 \sum v_i^2 - \left(\sum u_i v_i\right)^2}{n\sum u_i^2 S_{vv} + n\sum u_i v_i \left[2n\overline{uv} - \sum u_i v_i\right] - n^2\overline{u}^2\sum v_i^2}},$$

$$a = \hat{a} \pm t_{n-3}(1-\alpha/2)s\sqrt{\frac{S_{vv}}{\sum u_i^2 S_{vv} + \sum u_i v_i \left[2n\overline{uv} - \sum u_i v_i\right] - n\overline{u}^2\sum v_i^2}},$$

$$b = \hat{b} \pm t_{n-3}(1-\alpha/2)s\sqrt{\frac{S_{uu}}{\sum u_i^2 S_{vv} + \sum u_i v_i \left[2n\overline{uv} - \sum u_i v_i\right] - n\overline{u}^2\sum v_i^2}},$$

where

$$s = \sqrt{\frac{1}{n-3}\left(S_{yy} - \hat{a}S_{uy} - \hat{b}S_{vy}\right)};$$

see Exercise 3.10.

Note 1.13 (Joint confidence regions)

In the above, we restrict ourselves to confidence intervals for individual parameters, as is done in e.g. S-Plus/R®. One can give confidence regions for two or more parameters together, we refer for detail to Draper and Smith (1998), Ch. 5.

EXERCISES

1.1. By considering the quadratic

$$Q(\lambda) := \frac{1}{n}\sum_{i=1}^{n}(\lambda(x_i - \overline{x}) + (y_i - \overline{y}))^2,$$

show that the sample correlation coefficient r satisfies
(i) $-1 \le r \le 1$;
(ii) $r = \pm 1$ iff there is a linear relationship between x_i and y_i,

$$ax_i + by_i = c \quad (i = 1, \ldots, n).$$

1.2. By considering the quadratic

$$Q(\lambda) := E[(\lambda(x - \overline{x}) + (y - \overline{y}))^2],$$

show that the population correlation coefficient ρ satisfies
(i) $-1 \le \rho \le 1$;

(ii) $\rho = \pm 1$ iff there is a linear relationship between x and y, $ax + by = c$ with probability 1.

(These results are both instances of the Cauchy–Schwarz inequality for sums and integrals respectively.)

1.3. *The effect of ageing on athletic performance.* The data in Table 1.1 gives the first author's times for the marathon and half-marathon (in minutes).
(i) Fit the model $\log(time) = a + b\log(age)$ and give estimates and

Age	Half-marathon	Age	Marathon
46	85.62	46.5	166.87
48	84.90	47.0	173.25
49	87.88	47.5	175.17
50	87.88	49.5	178.97
51	87.57	50.5	176.63
57	90.25	54.5	175.03
59	88.40	56.0	180.32
60	89.45	58.5	183.02
61	96.38	59.5	192.33
62	94.62	60.0	191.73

Table 1.1 Data for Exercise 1.3

95% confidence intervals for a and b.
(ii) Compare your results with the runners' Rule of Thumb that, for ageing athletes, every year of age adds roughly half a minute to the half-marathon time and a full minute to the marathon time.

1.4. Look at the data for Example 1.11 on car speeds. Plot the data along with the fitted regression line. Fit the model $y = a + bx + cx^2$ and test for the significance of a quadratic term. Predict the speeds for x=(-3, 13) and compare with the actual observations of 135.9 and 158.6 respectively. Which model seems to predict best out of sample? Do your results change much when you add these two observations to your sample?

1.5. Give the solution to the normal equations for the regression model with two regressors in §1.8

1.6. Consider the data in Table 1.2 giving the first author's half-marathon times:

Age (x)	Time (y)	Age (x)	Time (y)
42	92.00	51	87.57
43	92.00	57	90.25
44	91.25	59	88.40
46	85.62	60	89.45
48	84.90	61	96.38
49	87.88	62	94.62
50	87.88	63	91.23

Table 1.2 Data for Exercise 1.6

(i) Fit the models $y = a + bx$ and $y = a + bx + cx^2$. Does the extra quadratic term appear necessary?
(ii) *Effect of club membership upon performance.* Use the following proxy $v = (0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$ to gauge the effect of club membership. ($v = 1$ corresponds to being a member of a club). Consider the model $y = a + bx + cv$. How does membership of a club appear to affect athletic performance?

1.7. The following data, $y = (9.8, 11.0, 13.2, 15.1, 16.0)$ give the price index y in years one to five.
(i) Which of the models $y = a + bt$, $y = Ae^{bt}$ fits the data best?
(ii) Does the quadratic model, $y = a + bt + ct^2$ offer a meaningful improvement over the simple linear regression model?

1.8. The following data in Table 1.3 give the US population in millions. Fit a suitable model and interpret your findings.

Year	Population	Year	Population
1790	3.93	1890	62.90
1800	5.31	1900	76.00
1810	7.24	1910	92.00
1820	9.64	1920	105.70
1830	12.90	1930	122.80
1840	17.10	1940	131.70
1850	23.20	1950	151.30
1860	31.40	1960	179.30
1870	39.80	1970	203.20
1880	50.20		

Table 1.3 Data for Exercise 1.8.

1.9. *One-dimensional change-of-variable formula.* Let X be a continuous random variable with density $f_X(x)$. Let $Y = g(X)$ for some monotonic function $g(\cdot)$.
(i) Show that

$$f_Y(x) = f_X\left(g^{-1}(x)\right)\left|\frac{dg^{-1}(x)}{dx}\right|.$$

(ii) Suppose $X \sim N(\mu, \sigma^2)$. Show that $Y = e^X$ has probability density function

$$f_Y(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(\log x - \mu)^2}{2\sigma^2}\right\}.$$

[Note that this gives the log-normal distribution, important in the Black–Scholes model of mathematical finance.]

1.10. The following exercise motivates a discussion of Student's t distribution as a normal variance mixture (see Exercise 1.11). Let $U \sim \chi^2_r$ be a chi-squared distribution with r degrees of freedom (for which see §2.1), with density

$$f_U(x) = \frac{x^{\frac{1}{2}r-1}e^{-\frac{1}{2}x}}{2^{\frac{1}{2}r}\Gamma(\frac{r}{2})}.$$

(i) Show, using Exercise 1.9 or differentiation under the integral sign that $Y = r/U$ has density

$$f_Y(x) = \frac{r^{\frac{1}{2}r}x^{-1-\frac{1}{2}r}e^{-\frac{1}{2}rx^{-1}}}{2^{\frac{1}{2}r}\Gamma(\frac{r}{2})}.$$

(ii) Show that if $X \sim \Gamma(a, b)$ with density

$$f_X(x) = \frac{x^{a-1}b^a e^{-bx}}{\Gamma(a)},$$

then $Y = X^{-1}$ has density

$$f_Y(x) = \frac{b^a x^{-1-a}e^{-b/x}}{\Gamma(a)}.$$

Deduce the value of

$$\int_0^\infty x^{-1-a}e^{-b/x}dx.$$

1.11. *Student's t distribution.* A Student t distribution $t(r)$ with r degrees of freedom can be constructed as follows:
1. Generate u from $f_Y(\cdot)$.
2. Generate x from $N(0, u)$,
where $f_Y(\cdot)$ is the probability density in Exercise 1.10 (ii). Show that

$$f_{t(r)}(x) = \frac{\Gamma\left(\frac{r}{2} + \frac{1}{2}\right)}{\sqrt{\pi r}\Gamma(\frac{r}{2})}\left(1 + \frac{x^2}{r}\right)^{-\frac{1}{2}(r+1)}.$$

The Student t distribution often arises in connection with the chi-square distribution (see Chapter 2). If $X \sim N(0,1)$ and $Y \sim \chi^2_r$ with X and Y independent then

$$\frac{X}{\sqrt{Y/r}} \sim t(r).$$