

N. H. Bingham  
John M. Fry

SPRINGER UNDERGRADUATE  
MATHEMATICS SERIES

# Regression


Linear Models in Statistics

S

U

M

S

 Springer

# Springer Undergraduate Mathematics Series

## **Advisory Board**

M.A.J. Chaplain *University of Dundee*

K. Erdmann *University of Oxford*

A. MacIntyre *Queen Mary, University of London*

E. Sili *University of Oxford*

J.F. Toland *University of Bath*

For other titles published in this series, go to  
[www.springer.com/series/3423](http://www.springer.com/series/3423)



N.H. Bingham • John M. Fry

# Regression

Linear Models in Statistics

 Springer

N.H. Bingham  
Imperial College, London  
UK  
nick.bingham@btinternet.com

John M. Fry  
University of East London  
UK  
frymaths@googlemail.com

Springer Undergraduate Mathematics Series ISSN 1615-2085  
ISBN 978-1-84882-968-8 e-ISBN 978-1-84882-969-5  
DOI 10.1007/978-1-84882-969-5  
Springer London Dordrecht Heidelberg New York

British Library Cataloguing in Publication Data  
A catalogue record for this book is available from the British Library

Library of Congress Control Number: 2010935297

Mathematics Subject Classification (2010): 62J05, 62J10, 62J12, 97K70

© Springer-Verlag London Limited 2010

Apart from any fair dealing for the purposes of research or private study, or criticism or review, as permitted under the Copyright, Designs and Patents Act 1988, this publication may only be reproduced, stored or transmitted, in any form or by any means, with the prior permission in writing of the publishers, or in the case of reprographic reproduction in accordance with the terms of licenses issued by the Copyright Licensing Agency. Enquiries concerning reproduction outside those terms should be sent to the publishers. The use of registered names, trademarks, etc., in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implied, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

*Cover design:* Deblík

Printed on acid-free paper

Springer is part of Springer Science+Business Media ([www.springer.com](http://www.springer.com))

*To James, Ruth and Tom*

Nick

*To my parents Ingrid Fry and Martyn Fry*

John



# Preface

The subject of regression, or of the linear model, is central to the subject of statistics. It concerns what can be said about some quantity of interest, which we may not be able to measure, starting from information about one or more other quantities, in which we may not be interested but which we can measure. We model our variable of interest as a linear combination of these variables (called covariates), together with some error. It turns out that this simple prescription is very flexible, very powerful and useful.

If only because regression is inherently a subject in two or more dimensions, it is not the first topic one studies in statistics. So this book should not be the first book in statistics that the student uses. That said, the statistical prerequisites we assume are modest, and will be covered by any first course on the subject: ideas of sample, population, variation and randomness; the basics of parameter estimation, hypothesis testing,  $p$ -values, confidence intervals etc.; the standard distributions and their uses (normal, Student  $t$ , Fisher  $F$  and chi-square – though we develop what we need of  $F$  and chi-square for ourselves).

Just as important as a first course in statistics is a first course in probability. Again, we need nothing beyond what is met in any first course on the subject: random variables; probability distribution and densities; standard examples of distributions; means, variances and moments; some prior exposure to moment-generating functions and/or characteristic functions is useful but not essential (we include all we need here). Our needs are well served by John Haigh's book *Probability models* in the SUMS series, Haigh (2002).

Since the terms regression and linear model are largely synonymous in statistics, it is hardly surprising that we make extensive use of linear algebra and matrix theory. Again, our needs are well served within the SUMS series, in the two books by Blyth and Robertson, *Basic linear algebra* and *Further linear algebra*, Blyth and Robertson (2002a), (2002b). We make particular use of the



material developed there on sums of orthogonal projections. It will be a pleasure for those familiar with this very attractive material from pure mathematics to see it being put to good use in statistics.

Practical implementation of much of the material of this book requires computer assistance – that is, access to one of the many specialist statistical packages. Since we assume that the student has already taken a first course in statistics, for which this is also true, it is reasonable for us to assume here too that the student has some prior knowledge of and experience with a statistical package. As with any other modern student text on statistics, one is here faced with various choices. One does not want to tie the exposition too tightly to any one package; one cannot cover all packages, and shouldn't try – but one wants to include some specifics, to give the text focus. We have relied here mainly on S-Plus/R<sup>®</sup>.<sup>1</sup>

Most of the contents are standard undergraduate material. The boundary between higher-level undergraduate courses and Master's level courses is not a sharp one, and this is reflected in our style of treatment. We have generally included complete proofs except in the last two chapters on more advanced material: Chapter 8, on Generalised Linear Models (GLMs), and Chapter 9, on special topics. One subject going well beyond what we cover – Time Series, with its extensive use of autoregressive models – is commonly taught at both undergraduate and Master's level in the UK. We have included in the last chapter some material, on non-parametric regression, which – while no harder – is perhaps as yet more commonly taught at Master's level in the UK.

In accordance with the very sensible SUMS policy, we have included exercises at the end of each chapter (except the last), as well as worked examples. One then has to choose between making the book more student-friendly, by including solutions, or more lecturer-friendly, by not doing so. We have nailed our colours firmly to the mast here by including full solutions to all exercises. We hope that the book will nevertheless be useful to lecturers also (e.g., in inclusion of references and historical background).

Rather than numbering equations, we have labelled important equations acronymically (thus the normal equations are *(NE)*, etc.), and included such equation labels in the index. Within proofs, we have occasionally used local numbering of equations: *(\*)*, *(a)*, *(b)* etc.

In pure mathematics, it is generally agreed that the two most attractive subjects, at least at student level, are complex analysis and linear algebra. In statistics, it is likewise generally agreed that the most attractive part of the subject is

---

<sup>1</sup> S+, S-PLUS, S+FinMetrics, S+EnvironmentalStats, S+SeqTrial, S+SpatialStats, S+Wavelets, S+ArrayAnalyzer, S-PLUS Graphlets, Graphlet, Trellis, and Trellis Graphics are either trademarks or registered trademarks of Insightful Corporation in the United States and/or other countries. Insightful Corporation 1700 Westlake Avenue N, Suite 500 Seattle, Washington 98109 USA.

regression and the linear model. It is also extremely useful. This lovely combination of good mathematics and practical usefulness provides a counter-example, we feel, to the opinion of one of our distinguished colleagues. Mathematical statistics, Professor x opines, combines the worst aspects of mathematics with the worst aspects of statistics. We profoundly disagree, and we hope that the reader will disagree too.

The book has been influenced by our experience of learning this material, and teaching it, at a number of universities over many years, in particular by the first author's thirty years in the University of London and by the time both authors spent at the University of Sheffield. It is a pleasure to thank Charles Goldie and John Haigh for their very careful reading of the manuscript, and Karen Borthwick and her colleagues at Springer for their kind help throughout this project. We thank our families for their support and forbearance.

NHB, JMF

Imperial College, London and the University of East London, March 2010



# Contents

<b>1. Linear Regression</b> . . . . .	1
1.1 Introduction . . . . .	1
1.2 The Method of Least Squares . . . . .	3
1.2.1 Correlation version . . . . .	7
1.2.2 Large-sample limit . . . . .	8
1.3 The origins of regression . . . . .	9
1.4 Applications of regression . . . . .	11
1.5 The Bivariate Normal Distribution . . . . .	14
1.6 Maximum Likelihood and Least Squares . . . . .	21
1.7 Sums of Squares . . . . .	23
1.8 Two regressors . . . . .	26
Exercises . . . . .	28
<b>2. The Analysis of Variance (ANOVA)</b> . . . . .	33
2.1 The Chi-Square Distribution . . . . .	33
2.2 Change of variable formula and Jacobians . . . . .	36
2.3 The Fisher F-distribution . . . . .	37
2.4 Orthogonality . . . . .	38
2.5 Normal sample mean and sample variance . . . . .	39
2.6 One-Way Analysis of Variance . . . . .	42
2.7 Two-Way ANOVA; No Replications . . . . .	49
2.8 Two-Way ANOVA: Replications and Interaction . . . . .	52
Exercises . . . . .	56
<b>3. Multiple Regression</b> . . . . .	61
3.1 The Normal Equations . . . . .	61

3.2	Solution of the Normal Equations	64
3.3	Properties of Least-Squares Estimators	70
3.4	Sum-of-Squares Decompositions	73
3.4.1	Coefficient of determination	79
3.5	Chi-Square Decomposition	80
3.5.1	Idempotence, Trace and Rank	81
3.5.2	Quadratic forms in normal variates	82
3.5.3	Sums of Projections	82
3.6	Orthogonal Projections and Pythagoras's Theorem	85
3.7	Worked examples	89
	Exercises	94
<b>4.</b>	<b>Further Multilinear Regression</b>	<b>99</b>
4.1	Polynomial Regression	99
4.1.1	The Principle of Parsimony	102
4.1.2	Orthogonal polynomials	103
4.1.3	Packages	103
4.2	Analysis of Variance	104
4.3	The Multivariate Normal Distribution	105
4.4	The Multinormal Density	111
4.4.1	Estimation for the multivariate normal	113
4.5	Conditioning and Regression	115
4.6	Mean-square prediction	121
4.7	Generalised least squares and weighted regression	123
	Exercises	125
<b>5.</b>	<b>Adding additional covariates and the Analysis of Covariance</b>	<b>129</b>
5.1	Introducing further explanatory variables	129
5.1.1	Orthogonal parameters	133
5.2	ANCOVA	135
5.2.1	Nested Models	139
5.3	Examples	140
	Exercises	145
<b>6.</b>	<b>Linear Hypotheses</b>	<b>149</b>
6.1	Minimisation Under Constraints	149
6.2	Sum-of-Squares Decomposition and F-Test	152
6.3	Applications: Sequential Methods	157
6.3.1	Forward selection	157
6.3.2	Backward selection	158
6.3.3	Stepwise regression	159
	Exercises	160

<b>7. Model Checking and Transformation of Data</b> .....	163
7.1 Deviations from Standard Assumptions .....	163
7.2 Transformation of Data .....	168
7.3 Variance-Stabilising Transformations .....	171
7.4 Multicollinearity .....	174
Exercises .....	177
<b>8. Generalised Linear Models</b> .....	181
8.1 Introduction .....	181
8.2 Definitions and examples .....	183
8.2.1 Statistical testing and model comparisons .....	185
8.2.2 Analysis of residuals .....	187
8.2.3 Athletics times .....	188
8.3 Binary models .....	190
8.4 Count data, contingency tables and log-linear models .....	193
8.5 Over-dispersion and the Negative Binomial Distribution .....	197
8.5.1 Practical applications: Analysis of over-dispersed models in R® .....	199
Exercises .....	200
<b>9. Other topics</b> .....	203
9.1 Mixed models .....	203
9.1.1 Mixed models and Generalised Least Squares .....	206
9.2 Non-parametric regression .....	211
9.2.1 Kriging .....	213
9.3 Experimental Design .....	215
9.3.1 Optimality criteria .....	215
9.3.2 Incomplete designs .....	216
9.4 Time series .....	219
9.4.1 Cointegration and spurious regression .....	220
9.5 Survival analysis .....	222
9.5.1 Proportional hazards .....	224
9.6 $p \gg n$ .....	225
<b>Solutions</b> .....	227
<b>Dramatis Personae: Who did what when</b> .....	269
<b>Bibliography</b> .....	271
<b>Index</b> .....	279



# 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  $0^\circ$  and the boiling point is  $100^\circ$ , while in the Fahrenheit scale, freezing point is  $32^\circ$  and boiling point is  $212^\circ$  (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,



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

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

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,  $(x_1, y_1), \dots, (x_n, y_n)$  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<sup>®</sup>,<sup>1</sup> for instance, using the command **Regression**, or S-Plus/R<sup>®</sup> 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

<sup>1</sup> Minitab<sup>®</sup>, Quality Companion by Minitab<sup>®</sup>, Quality Trainer by Minitab<sup>®</sup>, Quality Analysis. Results<sup>®</sup> 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

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

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  $i$ th point, the left-hand side above will be 0 if the fit is exact, but  $\epsilon_i$  say (denoting the  $i$ th error) in view of the observational errors. We wish to keep the errors  $\epsilon_i$  small; we wish also to put positive and negative  $\epsilon_i$  on the same footing, which we may do by looking at the squared errors  $\epsilon_i^2$ . 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  $(x_1, y_1), \dots, (x_n, y_n)$ . 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

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

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

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

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

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

and then

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

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

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

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

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

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

$$\begin{aligned} S_{xx} &:= \sum_{i=1}^n (x_i - \bar{x})^2, & S_{xy} &:= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \\ b &= S_{xy}/S_{xx}, & a &= \bar{y} - b\bar{x}. \end{aligned}$$

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

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

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

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

### 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

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

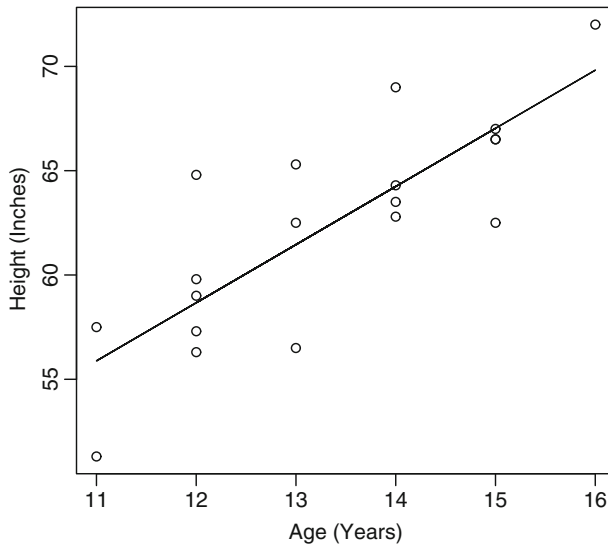
where  $t_{n-2}(1 - \alpha/2)$  denotes the  $1 - \alpha/2$  quantile of the Student  $t$  distribution with  $n - 2$  degrees of freedom and  $s$  is given by

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

### Example 1.2

We fit the line of best fit to model  $y = \text{Height}$  (in inches) based on  $x = \text{Age}$  (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

$$S_{xx} = \sum x_i y_i - n \bar{x} \bar{y},$$

$$S_{xy} = \sum x_i^2 - n \bar{x}^2.$$

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  $(x_1, y_1), \dots, (x_n, y_n)$  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

$$y - \bar{y} = b(x - \bar{x}), \quad b = r_{xy} s_y / s_x. \quad (SRL)$$

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,  $\mu_x$  and  $\mu_y$ , two variances  $\sigma_x^2$  and  $\sigma_y^2$  (or their square roots, the standard deviations  $\sigma_x$  and  $\sigma_y$ ), and one correlation,  $\rho_{xy}$ . The two means are measures of *location*, and serve to identify the point  $(\mu_x, \mu_y)$ , or its sample counterpart,  $(\bar{x}, \bar{y})$  – 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

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

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

## 1.2.2 Large-sample limit

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

$$\bar{x} \rightarrow Ex \quad (n \rightarrow \infty).$$

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

$$s_x^2 = s_{xx} = \overline{x^2} - (\bar{x}^2) \rightarrow E(x^2) - (Ex)^2 = \text{var}(x),$$

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

$$s_{xy} = \overline{xy} - \bar{x}\bar{y} \rightarrow E(xy) - Ex.Ey = \text{cov}(x, y),$$

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

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

tends to the line

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

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

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

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

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

### 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*),  $\sigma_x$  and  $\sigma_y$  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

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

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

$$\rho^n \rightarrow 0 \quad (n \rightarrow \infty):$$

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