Chapter 8

Correlation and Regression

8.1 Introduction

This chapter concerns itself entirely with the relationship between *two* continuous variables, in contrast with Section 4.2, which considers two categorical variables.

We have already seen in Chapter 1 how to present data of this form (using MINITAB) and make a suitable comment. This descriptive a is now supplemented with more formal techniques, in particular estimation and hypothesis testing for the parameters of a linear relationship which are defined below.
The data consist of n pairs of measurements on two variables *X* and *Y*:

$(x_1, y_1), (x_2, y_2), \ldots, (x_i, y_i), \ldots (x_n, y_n)$

These data can have arisen from a random sample of n individual from a population, or from an experiment in which one variable, conventionally X, is held fixed or controlled at certain chosen levels and independent measurements of the *response* variable, conventionally *Y*, are taken for each of these level.

Examp1e 15 Ten males were randomly selected from a well-specified population and their weight and height were measured:

Height(X)	63	71	72	68	75	66	68	76	71	70
Weight(Y)	145	158	156	148	163	155	153	158	150	154

These data are plotted in Figure 5.1 There inn approximate positive linear

Figure 5.1: Weight and Height of 10 randomly selected males

relationship between the two variables. One way of characterising this is to say that the *mean* weight for a given height seems to be an increasing *linear* function .f height, within the range of heights examined here.

Example 16 (data again taken from Chatfield's book). An experiment was conducted to investigate the variation in specific heat of a certain chemical with temperature. Two measurements of specific heat were taken at a series of six equally spaced temperatures in the

range 50°—100°C.

The important feature of this experiment is that one variable, temperature, is controlled and only the other, specific heat, is subject to random variation (mainly due to measurement error). It is not necessary to have equally spaced values nor is it necessary to have equal *replications* at those values.

Temperature °C(X)	50	60	70	80	90	100
Specific Heat *(Y)*	1.60	1.63	1.67	1.70	1.71	1.71
(cal/gm/°C)	1.64	1.65	1.67	1.72	1.72	1.74

The data are plotted in Figure 5.2, where the '2' means that there are two observations at that point. As in Example 15, there is a positive relationship but now the function giving mean specific heat for a given temperature, although increasing, may be *curvilinear*, possibly quadratic. (We will nevertheless analyse these data for a linear relationship *only.)*

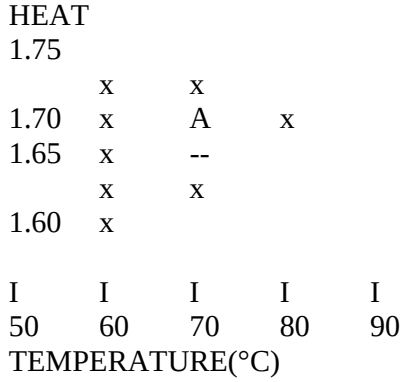

Figure 5.2: Experiment to determine variation of Specific Heat with Temperature

You should read the suggested references for a discussion of other possible scatter plots and their interpretation. The methods in this chapter are applicable only if the data show a *linear* relationship, whether strong or weak, negative or positive. Either or both variables may need to be transformed to achieve an approximate linear relationship.

8.2 Correlation

The *only* measure of association we will be covering here is Pearson's *product moment linear correlation coefficient r* (when referring to this statistic, all the words here can be omitted with the exception of linear *correlation).* The formula for *r* is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

If we define the *corrected sum of products of X and* Y to be

$$CS(x,y)=\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$$

which is calculated by , and the *corrected* sum *of* squares *of* X to be

$$CS(x,x)=\sum_{i=1}^{n}(x_i-\bar{x})^2$$

and similarly for *Y*, calculated using the *fundamental identity* (on page ())

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}y_i^2-n\bar{y}^2$$

then *r* can be written more compactly as

$$r=\frac{CS(x,y)}{\sqrt{CS(x,x)CS(y,y)}}$$

Alternatively we can use the sample *covariance*

$$S_{XY}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}=CS(x,y)/n-1$$

to write

$$r=\frac{S_{XY}}{S_X S_Y}$$

in terms of sample variances and covariance.

Theorem 2: For any set of data (where neither the x-values nor the y-values are all the same)

$$-1\le r\le 1$$

with equality if and only if X and *Y* are linear combinations of each other for the data set.

Proof

For any k

$$\sum_{i=1}^{n}[(y_i-\bar{y})-k(x_i-\bar{x})]^2\ge 0$$

as it the sum of non-negative quantities. Expanding the quadratic and using the laws assumption,

$$k^2\sum_{i=1}^{n}(x_i-\bar{x})^2-2k\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})+\sum_{i=1}^{n}(y_i-\bar{y})^2\ge 0\qquad\forall k$$

Consider the left hand side of this inequality as a. function of *k.* By the inequality, it must have no zeroes or two identical zeroes, so the quadratic discriminant B^2 – 4AC must be non positive and hence

$$4[C(S,S)]^2\le 4CS(X,X)CS(y,y).$$

Thus $r^2\le 1$ which proves the result, but note that equality occurs if and only if there are two zeroes,i.e when there exits a k such that

$$\sum_{i=1}^{n}[(y_i - \bar{y}) - k(x_i - \bar{x})]^2 = 0$$

which can only occur if

$$(y_i - \bar{y}) - k(x_i - \bar{x}) = 0 \quad i = 1,2,\ldots,n$$

This is equivalent to the existence of constants *k* and *c* such that (1 = 1,2,...,n), which is what is meant by the two variables being linear combinations of each other.

What does the observed value of *r* tell us about the relationship between the two variables *in the population?*

Assume that *both* samples are random and from Normal populations (easily checked with Normal probability plots), and that the joint sample of pairs is random from a Bivariate Normal distribution. (The first part of this assumption, which we *can* check, follows from the second part, which we *cannot* check, nor do we need to how the full meaning of the second part.) We can.use *r* as a test statistic for the null hypothesis H_0 : $\rho = 0$ (no *linear* association) versus H_1 : $\rho \neq 0$, where ρ is the population parameter measuring the correlation of X and Y. The sampling distribution of *r* is complicated but it can be shown that

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim t_{n-2} \text{ under } H_0$$

We. can test H_0 using Table which gives critical values of r for $v = n - 2$ degrees of freedom, or by referring the observed value of *t* above to Table 7.

Example 15 revisited You should always examine the scatter plot first before undertaking any calculations or tests. Here it clearly shows that there is weak *linear* association.

Using hand calculations,

$\sum_{i=1}^{n} y_i = 700, \sum_{i=1}^{n} x_i = 1540, \ \sum_{i=1}^{n} y_i^2 .. = 49140.., \sum_{i=1}^{n} x.^2_i = 237412$

$\sum_{i=1}^{n} x_i y_i$ so $CS(y,y) = 140$, $CS(x,y) = 146$, $CS(x,x) = 252$ hence

$$r = .\frac{146}{\sqrt{140 \times 252}} = 0.777$$

which we refer to Table 15 with v = 8. critical values are 0.7646 *(P* = 0.01) and 0.8721 *(P* =0.001) for a two-sided test so we conclude that there is strong evidence for linear correlation as 0.001 <*P* <0.01.

The same conclusion is obtained calculating $= 0.777 \times \sqrt{8} / \sqrt{1 - 0.777^2} = 3.49$ which gives 0.01 > *P* > 0.005.

Note that the data of Example 16 are not Suitable for correlation analysis as one variable (*X*) is controlled and not random. Other data sets which show non-linear association can either be analysed by the technique above after suitable transformation(s) or by using *Rank correlation methods* which are outside the scope of this course. Also excluded from this course is Fisher's z-transformation which can be used (among other things) to construct a 95% confidence interval for the unknown population parameter ρ.

Regression

Correlation analysis may establish a linear relationship but does not allow us to *use* it to say, predict the value of one variable given the value of another. *Regression Analysis* allows us to do this and more. It is also applicable when one of the variables (X) is controlled.

We will assume that the scatter plot of *Y* versus X shows a roughly linear relationship and in addition that the *spread* in the Y-direction is roughly

constant with X. It may be necessary to transform one or both variables to achieve this.

We postulate a linear *model* for the data (and the population from which the data has been drawn):

$$Y = \alpha + \beta X + \in$$

where *Y* is the response variable, X is the regressor or explanatory variable and $\in$ is a random error with zero mean and constant variance a2 (unknown) for each value of X. The unknown parameters α and β represent the intercept and slope of the (unknown) *population regression line* $a + \beta X$.

The estimate of this line is $a + bX$ where

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

$$b = \frac{CS(x,y)}{CS(x,x)}$$

are unbiased estimators of α and β respectively. (The derivation of these estimates by the Principle of Least Squares or by Maximum Likelihood is outside the scope of this course)

Example 16 revisited $n = 12$, $\sum_{i=1}^{n} x_i = 900$, $\sum y_i = 20.16$, $\bar{x} = 75$, $\bar{y} = 1.68$. The raw sums of squares and products are $\sum_{i=1}^{n} x_i y_i = 1519.9$, $\sum_{i=1}^{n} x_i^2 = 71000$, $\sum_{i=1}^{n} y_i^2 = 33.8894$

$$b = \frac{\sum_{i=1}^{n} x_i y_i - n\overline{xy}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} = \frac{1519.9 - 12 \times 75 \times 1.68}{71000 - 12 \times 75^2} = \frac{7.9}{3500} = 0.00226$$

This can be interpreted as the (estimated) average increase in specific heat per increase temperature of 1^0C.

$$a = 1.68 - 0.00228 \times 75 = 1.511$$

This has *no* interpretation as say, the specific heat at 0°C, as the data tell us nothing about the relationship outside the range of temperatures considered.

The fitted line *Y* = 1.511 + 000226X should now be plotted on the original scatter diagram, as a check to calculations, and prior to checking the assumptions of the model.

To make inferences on the unknown parameters α, β, σ^2 and to make predictions (with appropriate confidence intervals) of the response variable we need to make further distributional assumptions on the random errors in the model. We assume that the random errors $\in$ are independent and Normally distributed (zero mean and constant variance already assumed).

The assumptions of the regression model can be checked by calculating and examining (by

suitable plots) estimates of the random errors, called *residuals.* The i^{th} residual is given by

$e_i = y_i - \alpha - bx_i = y_{.i} - \hat{y}_i$

where $\hat{y}_i$ is called the i^{th} fitted value. Graphically the residual is just the vertical distance of the observed i^{th} point from the fitted line, If the assumptions are correct then these residuals should look approximately like a random sample from a Normal distribution with zero mean and variance $\sigma_{.}^{2}$.

They should be plotted (on scatter diagrams) against

- the values of the regressor variable X - to indicate a departure from the model assumptions in the form of a non-linear term, and also to check whether the variance changes with X.

- the fitted values $\hat{y}_i$ to indicate a possibly non-constant variance. (In many data sets the variance or spread is found to increase with the estimated mean or fitted value and a transformation of the response variable is often the appropriate remedy.)

- expected Normal order statistics — to indicate possible non-Normality in the error distribution (MINITAB: NSCORES)

- the values of any other variable observed on the sample units — a systematic pattern here indicates the need for a *multiple regression,* outside the scope of this course.

Example 16 revisited

		observed	fitted	residual
.	i	x_i	y_I $\hat{y}_i$	e_i
1	50	1.60	1.624	- 0.024
2	50	1.44	1.624	+ 0.016
	3	60	1.63	1.647	- 0.017
	4	60	1.65	1.647	+ 0.003
	5	70	1.67	1.669	+0.001
	6	70	1.67	1.669	+ 0.001
	7	80	1.70	1.692	+ 0.008
	8	80	1.72	1.692	+ 0.028
	9	90	1.71	1.714	- 0.004
	10	90	1.72	1.714	+ 0.006
	11	100	1.71	1.737	- 0.027
	12	100	1.74	1.737	+ 0.003

Another purpose of calculating the residuals and/or plotting them is to detect possible *outliers* i.e points which do not seem to belong with the remaining data on account of their large residual. To assess the size of residuals we need to estimate a'^{2}, the variance about the regression line. This variation is measured by the *residual* or *error sum of squares,* denoted *RSS* and given by

$$RSS = \sum_{i=1}^{n} e_i^2 = CS(y,y) - [CS(x,y)]^2 / CS(x,x)$$

(proof not required). It can be shown that the *residual mean* square, given by $s^2 = RSS/(n - 2)$ (the same notation as for the sample variance), is an unbiased estimator of σ^2 and $(n — 2)s^2/\sigma^2$ has a chi-square sampling distribution with $n-2$ degrees of freedom.

Example 16 revisited *CS(s, y)* = 7.9, *CS(z, z)* = 3500, and *CS(y, y)* = 33.8894—12 x $(1.68)^2$ = 0.0206. Thus *RSS* = 0.0206 – $(7.9)^2/3500$ = 0.00277 and so s^2 = *RSS / 10* = 2.77 x 10^{-4} and *s* = 0.0166.

In the table of residuals above our largest residual in absolute value is 0.028 which could well have been generated from a Normal distribution with zero mean and standard deviation 0.0166 so this value is not 'surprising' and we have no reason to believe the point is an outlier.

Our first task having checked the assumptions is to establish whether there is any real evidence of a linear relationship. If there is not, we may as well treat the response values as a random sample We test

$$H_o : \beta = 0 \text{ versus } H_1 : \beta \neq 0$$

It can be shown that

$$t \sim N(\beta, \sigma^2 / CS(x,x))$$

independently of s^2 so that

$$T = \frac{b}{s/\sqrt{CS(x,x)}} \sim t_{n-2} \text{ under } H_o$$

is a suitable test statistic.

Example 16 revisited The fitted slope *b* is 0.00226 so the observed
Value of T is

$$t = \frac{0.00226}{0.0166/\sqrt{3500}} = 8.05$$

Referring this to Table? with 10 degrees of freedom, we obtain *P* <0.001, which is very strong evidence against *Ho.* So we have established that there is a linear relationship, which we can now use to make predictions etc.

A $100(1-\alpha)\%$ confidence interval for β is

Example 16 revisited A 95% confidence interval for the average in-crease in specific heat per °C is

$$0.00226 \pm \frac{0.0166 \times 2.228}{\sqrt{3500}} = 0.00226 \pm 0.000625$$

which gives (0.001635, 0.002885).

Suppose we wish to estimate the (true) mean response $\alpha + \beta x$ at a specified value of the regressor x_o in the range of the data. (We have no evidence that the model holds outside this range.) Then it can be shown that

$$\alpha + bx_o \sim N\left(\alpha + \beta x_o, \sigma^2\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{CS(x,x)}\right]\right)$$

so that a $100(1-\alpha)\%$ confidence interval for $\alpha + \beta x_o$ is

$$\alpha + bx_o \pm t_{n-2}(\alpha/2)\, s\left[\frac{1}{n} + \frac{(x_o - \bar{x})^2}{CS(x,x)}\right]^{1/3}$$

For the *prediction* of a single measurement of response at x_0, we need an interval which contains it with probability $1-\alpha$ so we must take into account the extra random variation of a single measurement about the mean. This variance is just, σ^2 so the $100(1-\alpha)\%$ *prediction interval* is

$$\alpha + bx_o \pm t_{n-2}(\alpha/2)\, s\left[1 + \frac{1}{n} + \frac{(x_o - \bar{x})^2}{CS(x,x)}\right]^{1/3}$$

Example 16 revisited We require a 95% confidence interval for the true specific heat at 60°C.

We *could* use the two observations at this temperature and disregard the rest of the observations. Using the methods of Section 3.2, this would give an interval 1.64 ± 0.1271. As it is based on a sample of size 2 we would not expect this interval to be very precise i.e. narrow. However, the interval does not depend on the validity of the linear regression model so *robustness* has been obtained at the expense of precision. Moreover this method can clearly *not* be used where there is no data, but still within the range, for example at x_o = 67°C.

Using the fitted regression model our estimate is a + b x 60 = 1.647 (same as the fitted value because we are estimating at a data point). A 95% confidence interval is

$$1.67 \pm 2.228 \times 0.0166 \times \sqrt{\frac{1}{12} + \frac{(60-75)^2}{3500}} = 1.647 \pm 0.0142$$

which gives an interval (1.63?8, 1.6612), far narrower than the one considered above.

For predicting a single measurement at x =60 the reader can easily verify that the interval is 1.647 ± 0.0396 or (1.607, 1.687).

In general, confidence intervals based on a fitted regression are narrower if

s^2 is smaller i.e. there is less variability about the line

the sample size is larger

. the significance level α is larger i.e. less ‘confidence’ is required

x_0 is nearer to 2 i.e. predictions are more precise near the centre of the data

CS(x, x) is larger i.e. data are more spread in the s-range

Outside the scope of this course is the *analysis of variance* for the regression which gives an alternative but equivalent F-test for H_0 and a simple tabular representation about how fitting the model 'explains' the variation in response values. This topic is of great importance in multiple regression.

Assignment 5

1 (a) Recall your descriptive analysis of the PULSE data in Assignment 1.
Now use MINITAB to produce a *separate* plot of weight versus height for both males and females.

Fit the linear regression of weight (as a response)on height (as a regressor) for both groups separately and predict the weight of an *individual* (i) male (ii) female student whose height is 68 inches, giving a 95% interval in each case.

Check your MINITAB intervals by hand, obtaining the value for *CS(x, x)* by using the fact that the standard error of!, is $s/\sqrt{CS(x,x)}$.

A study was conducted to determine the effects of sleep deprivation on subjects' ability to solve simple problems. The amount of sleep deprivation varied from 8 to 24 hours and was carefully controlled. A total of 10 subjects IT participated in the study, 2 at each sleep deprivation level. After the specified period, each subject was given a set of simple addition problems and the number of errors was recorded. The data were

Errors ' y'	8 6	6 10	8 14	14 12	16 12
hours 'x'	8	12	16	20	24

Regarding y as a continuous variable, perform a linear regression analysis on these data:

(a) Plot the data on a scatter diagram.

(b) Postulate a regression model for these data mentioning carefully what assumptions are made on the random errors.

(c) Fit the model by estimating the parameters and plot the fitted line on the diagram.

(d) Give a 90% confidence interval for the 'slope' parameter after first testing the hypothesis of no linear association.

(e) Check the assumptions of the model by calculating the fitted values and (crude) residuals, giving suitable plots and comments. Check in particular whether account should be taken (using multiple regression) of the variable *z* defined to be 'normal hours of sleep' taking respective values for the subjects in the study of 7, 8, 7,8, 8,9, 6,6,9,7.

Predict the number of errors made by an individual who was deprived of sleep for 14 hours, giving a 95% interval.

(g) Estimate the mean number of errors made for 24 hours deprivation giving a 90% interval.