CHAPTER 3

Fitting Straight Lines: Special Topics

3.0. SUMMARY AND PRELIMINARIES

This chapter, which can be omitted in a first reading, deals with some special topics relating to straight line fits. These are:

(3.1) Predictions, and confidence statements, for the true mean value of Y at a given X.

(3.2) Inverse regression. We want to predict an X, given a Y-value, after the fit has been made.

(3.3) How we can pick a good design for fitting a straight line using results in Chapters 1 and 2.

(3.4) A brief discussion of, and a suggested method for dealing with, the situation where the errors in X cannot be ignored, as is usually done.

For Section 3.1 we need a preliminary result about the covariance of two linear combinations of $Y_1, Y_2, \ldots, Y_n$.

Covariance of Two Linear Functions

Let a_i and c_i be constants, let Y_i be random, and suppose that

$$a = a_1Y_1 + a_2Y_2 + \cdots + a_nY_n,$$
$$c = c_1Y_1 + c_2Y_2 + \cdots + c_nY_n.$$

Suppose that all the Y_i have the same variance $\sigma^2 = V(Y_i)$ and that the Y's are pairwise uncorrelated, that is, cov $(Y_i, Y_j) = 0$, $i \neq j$. It then follows that

$$\text{cov}(a, c) = (a_1c_1 + a_2c_2 + \cdots + a_nc_n)\sigma^2. \tag{3.0.1}$$

[*Note:* If $\text{cov}(Y_i, Y_j) = \rho_{ij}\sigma^2$ for $i \neq j$, then all possible terms of form $(a_ic_j + a_jc_i)\rho_{ij}$ would be added inside the parentheses on the right of (3.0.1). However, we do not need this for our purposes here.]

Example. Let $a = \overline{Y}$, and $c = b_1$, the slope from a straight line fit. Then

$$a_i = 1/n \tag{3.0.2}$$

and

$$c_i = \frac{X_i - \overline{X}}{\Sigma(X_i - \overline{X})^2} = \frac{X_i - \overline{X}}{S_{XX}} \tag{3.0.3}$$

from the algebra just above (1.4.2). It follows that

$$a_1c_1 + a_2c_2 + \cdots + a_nc_n = (nS_{XX})^{-1}\Sigma(X_i - \overline{X}) = 0,$$

and so

$$\text{cov}(\overline{Y}, b_1) = 0, \tag{3.0.4}$$

that is, $\overline{Y}$ and b_1 are uncorrelated random variables.

3.1. STANDARD ERROR OF $\hat{Y}$

The fitted equation of a straight line, least squares fit can be written $\hat{Y} = b_0 + b_1X$ or, with $b_0 = \overline{Y} - b_1\overline{X}$, as

$$\hat{Y} = \overline{Y} + b_1(X - \overline{X}) \tag{3.1.1}$$

where both $\overline{Y}$ and b_1 are subject to error, which will affect $\hat{Y}$. At a specified value of X, say, X_0, we predict

$$\hat{Y}_0 - \overline{Y} + b_1(X_0 - \overline{X}) \tag{3.1.2}$$

for the mean value of Y at X_0. Because X_0 and $\overline{X}$ are fixed, and because cov $(\overline{Y}, b_1) = 0$, see above, we obtain

$$\begin{aligned} V(\hat{Y}_0) &= V(\overline{Y}) + (X_0 - \overline{X})^2V(b_1) \\ &= \frac{\sigma^2}{n} + \frac{(X_0 - \overline{X})^2\sigma^2}{\Sigma(X_i - \overline{X})^2}. \end{aligned} \tag{3.1.3}$$

Hence

$$\text{se}(\hat{Y}_0) = \text{est. sd}(\hat{Y}_0) = s\left\{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\Sigma(X_i - \overline{X})^2}\right\}^{1/2}. \tag{3.1.4}$$

This is a minimum when $X_0 = \overline{X}$ and increases as we move X_0 away from $\overline{X}$ in either direction. In other words, the greater distance an X_0 is (in either direction) from $\overline{X}$, the larger is the error we may expect to make when predicting, from the regression line, the mean value of Y at X_0. This is intuitively very reasonable. To state the matter loosely, we might expect to make our "best" predictions in the "middle" of our observed range of X and would expect our predictions to be less good away from the "middle." For values of X outside our experience—that is, outside the range observed—we should expect our predictions to be less good, becoming worse as we moved away from the range of observed X-values.

We now apply these results to the steam data of Chapter 1.

Example

$$n = 25, \qquad \Sigma(X_i - \overline{X})^2 = 7154.42;$$

$$s^2 = 0.7923, \qquad \overline{X} = 52.60;$$

$$\text{est. } V(\hat{Y}_0) = s^2\left\{\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\Sigma(X_i - \overline{X})^2}\right\}$$

$$= 0.7923\left\{\frac{1}{25} + \frac{(X_0 - 52.60)^2}{7154.42}\right\}.$$

For example, if $X_0 = \overline{X}$, then $\hat{Y}_0 = \overline{Y}$ and

$$\text{est. } V(\hat{Y}_0) = 0.7923\left\{\frac{1}{25}\right\} = 0.031692,$$

that is, $\text{se}(\hat{Y}_0) = \sqrt{0.031692} = 0.1780$.

If $X_0 = 28.6$,

$$\text{est. } V(\hat{Y}_0) = 0.7923\left\{\frac{1}{25} + \frac{(28.60 - 52.60)^2}{7154.42}\right\} = 0.095480,$$

that is, $\text{se}(\hat{Y}_0) = \sqrt{0.095480} = 0.3090$.

Correspondingly, $\text{se}(\hat{Y}_0)$ is also 0.3090 when $X_0 = 76.60$.

The 95% confidence limits for the true mean value of Y for a given X_0 are then given by $\hat{Y}_0 \pm (2.069)\,\text{se}(\hat{Y}_0)$. The t-value for $\nu = 23$ df is used because the $s^2 = 0.7923$ estimate is based on 23 df. The limits are thus:

$$\text{At } X_0 = \overline{X} = 52.60, \quad \overline{Y} \pm 2.069(0.1780) = 9.424 \pm 0.368$$

$$= (9.056, 9.792).$$

$$\text{At } X_0 = 28.6, \quad (13.623 - 0.079829(28.6)) \pm 2.069(0.3090)$$

$$= 11.340 \pm 0.639 = (10.701, 11.979).$$

$$\text{At } X_0 = 76.6, \quad (13.623 - 0.079829(76.6)) \pm 0.639$$

$$= 7.508 \pm 0.639 = (6.869, 8.147).$$

Note that because the X_0 values 28.6 and 76.6 are the same distance from $\overline{X} = 52.6$, the *widths* of the respective intervals are the same but their positions differ vertically.

If we joined up all the lower end points and all the upper end points of such intervals as X_0 changes, we would get Figure 3.1. The two curves are hyperbolas.

These limits can be interpreted as follows. Suppose repeated samples of Y_i are taken of the same size and at the same fixed values of X as were used to determine the fitted line above. Then of all the 95% confidence intervals constructed for the mean value of Y for a given value of X, say X_0, 95% of these intervals will contain the true mean value of Y at X_0.

If only one prediction $\hat{Y}_0$ is made, say, for $X = X_0$, then we act as though the probability that the calculated interval at this point will contain the true mean is 0.95.

Intervals for Individual Observations and Means of *q* Observations

The variance and standard deviation formulas above apply to the predicted *mean value* of Y for a given X_0. Since the actual observed value of Y varies about the true mean value with variance σ^2 [independent of the $V(\hat{Y})$], a predicted value of an *individual* observation will still be given by $\hat{Y}$ but will have variance

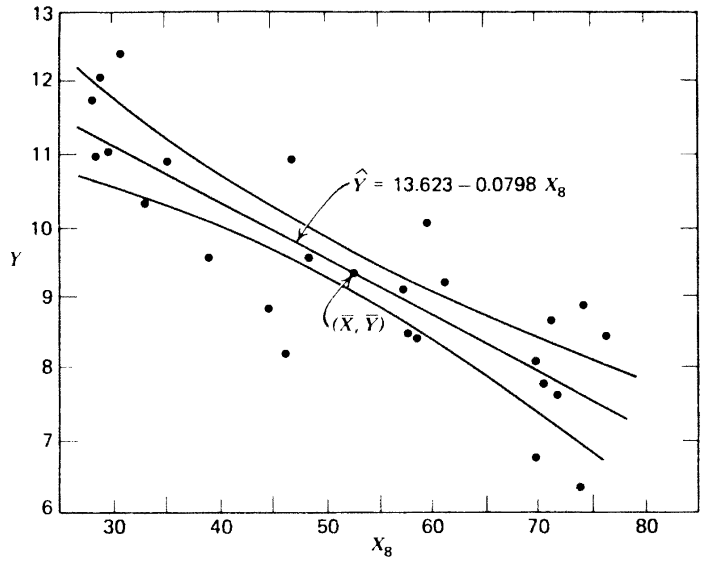

Figure 3.1. The loci of 95% confidence bands for the true mean value of Y.

$$\sigma^2\left\{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right\} \tag{3.1.5}$$

with corresponding estimated value obtained by inserting s^2 for σ^2. Confidence values can be obtained in the same way as before; that is, we can calculate a 95% confidence interval for a new observation, which will be centered on $\hat{Y}_0$ and whose length will depend on an estimate of this new variance from

$$\hat{Y}_0 \pm t(\nu, 0.975)\left\{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right\}^{1/2} s, \tag{3.1.6}$$

where ν is the number of degrees of freedom on which s^2 is based (and equals $n - 2$ here). A confidence interval for the average of q new observations about $\hat{Y}_0$ is obtained similarly as follows:

Let $\bar{Y}_0$ be the mean of q future observations at X_0 (where q could equal 1 to give the case above). Then

$$\bar{Y}_0 \sim N(\beta_0 + \beta_1 X_0, \sigma_0^2/q),$$
$$\hat{Y}_0 \sim N(\beta_0 + \beta_1 X_0, V(\hat{Y}_0)),$$

so that

$$\bar{Y}_0 - \hat{Y}_0 \sim N(0, \sigma_0^2/q + V(\hat{Y}_0)),$$

and $[(\bar{Y}_0 - \hat{Y}_0)/\text{se}(\bar{Y}_0 - \hat{Y}_0)]$ is distributed as a $t(\nu)$ variable, where ν is the number of degrees of freedom on which s^2, the estimate of σ^2, is based. Thus

$$\text{Prob}\left\{|\bar{Y}_0 - \hat{Y}_0| \le t(\nu, 0.975)\left[s^2\left(\frac{1}{q} + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X_i - \bar{X})^2}\right)\right]^{1/2}\right\} = 0.95$$

from which we can obtain 95% confidence limits for $\overline{Y}_0$ about $\hat{Y}_0$ of

$$\hat{Y}_0 \pm t(\nu, 0.975)\left[\frac{1}{q}+\frac{1}{n}+\frac{(X_0-\overline{X})^2}{\Sigma(X_i-\overline{X})^2}\right]^{1/2} s. \tag{3.1.7}$$

These limits are of course wider than those for the mean value of Y for given X_0, since these limits are the ones within which 95% of future observations at X_0 (for $q = 1$) or future means of q observations at X_0 ($q > 1$) are expected to lie.

Note: To obtain simultaneous confidence curves appropriate for the whole regression function over its entire range, it would be necessary to replace $t(\nu, 1 - \frac{1}{2}\alpha)$ by $\{2F(2, n - 2, 1 - \alpha)\}^{1/2}$. See, for example, pp. 110–116 of *Simultaneous Statistical Inference*, 2nd ed., by R. G. Miller, published by Springer-Verlag, New York, in 1981.

Confidence bands are rarely drawn in practice. However, the idea is important to understand and a suitable confidence interval around any $\hat{Y}$ value can always be evaluated numerically by applying a general algebraic formula, no matter how many X's there are. This latter aspect is a valuable one.

3.2. INVERSE REGRESSION (STRAIGHT LINE CASE)

Suppose we have fitted a straight line $\hat{Y} = b_0 + b_1 X$ to a set of data (X_i, Y_i), $i = 1, 2, \ldots, n$, and now, for a specified value of Y, say Y_0, we wish to obtain a predicted value $\hat{X}_0$, the corresponding value of X, as well as some sort of interval confidence statement for X surrounding $\hat{X}_0$. A practical example of such a problem is the following: X_i is an age estimate obtained from counting tree rings, while Y_i is a corresponding age estimate obtained from a carbon-dating process. The fitted straight line provides a "calibration curve" for the carbon-dating method, related to the more accurate tree ring data. Application of the carbon-dating method to an object now gives a reading Y_0. What statements can we make about the object's true age? This problem is called the inverse regression problem. (In other examples, Y_0 might be a true mean value or the average of q observations.)

There are several alternative ways of obtaining the (same) solution to this type of problem. First, let us assume Y_0 is a true mean value, not a single observation or average of q observations. Intuitively reasonable is the following. Draw the fitted straight line and curves that give the end points of the $100(1 - \alpha)\%$ confidence intervals for the true mean value of Y given X. (See Figure 3.2.) Draw a horizontal line parallel to the X axis at a height Y_0. Where this straight line cuts the confidence interval curves, drop perpendiculars onto the X-axis to give lower and upper $100(1 - \alpha)\%$ "fiducial limits," labeled X_L and X_U in Figure 3.2. The perpendicular from the point of intersection of the two straight lines onto the X-axis gives the inverse estimate of X, defined by solving $Y_0 = b_0 + b_1 \hat{X}_0$ for $\hat{X}_0$, namely,

$$\hat{X}_0 = (Y_0 - b_0)/b_1.$$

To obtain the values of X_L and X_U we can proceed as follows. In the figure, X_L is the X-coordinate of the point of intersection of the line

$$Y = Y_0 \qquad (\text{i.e., } Y = b_0 + b_1 \hat{X}_0) \tag{3.2.1}$$

and the curve

$$Y = Y_{X_L} - ts\left\{\frac{1}{n}+\frac{(X_L-\overline{X})^2}{S_{XX}}\right\}^{1/2}, \tag{3.2.2}$$

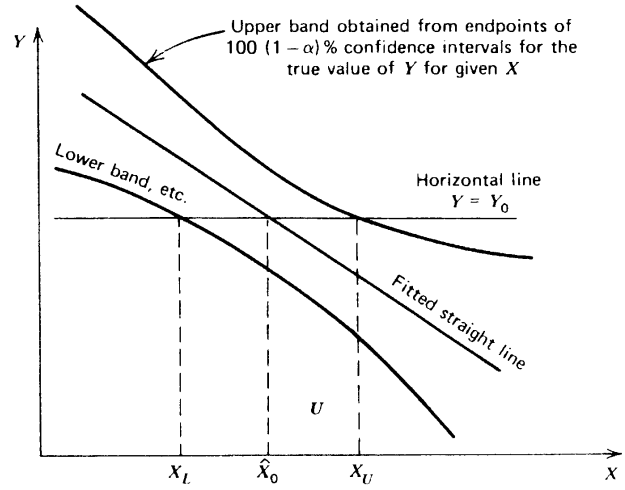

Figure 3.2. Inverse regression: estimating X by $\hat{X}_0$ from a given Y_0 value, and obtaining a $100(1-\alpha)\%$ "fiducial interval" for X.

where $S_{XX} = \Sigma(X_i - \overline{X})^2$, $Y_{X_L} = b_0 + b_1 X_L$, $t = t(\nu, 1 - \alpha/2)$ is the usual t percentage point, and ν is the number of degrees of freedom of s^2. Setting Eqs. (3.2.1) and (3.2.2) equal, canceling a b_0, rearranging to leave the square root alone on one side of the equation, and squaring both sides to get rid of the square root leads to a quadratic equation

$$PX_L^2 + 2QX_L + R = 0, \tag{3.2.3}$$

in X_L, where

$$\begin{aligned} P &= b_1^2 - t^2 s^2 / S_{XX}, \\ Q &= t^2 s^2 \overline{X} / S_{XX} - b_1^2 \hat{X}_0, \\ R &= b_1^2 \hat{X}_0^2 - t^2 s^2 / n - t^2 s^2 \overline{X}^2 / S_{XX}. \end{aligned} \tag{3.2.4}$$

We get exactly the same equation for X_U, so that X_L, X_U are the roots of (3.2.3). These, after some manipulation, are found to be

$$\left.\begin{matrix} X_U \\ X_L \end{matrix}\right\} = \overline{X} + \frac{b_1(Y_0 - \overline{Y}) \pm ts\{[(Y_0 - \overline{Y})^2/S_{XX}] + (b_1^2/n) - (t^2 s^2/nS_{XX})\}^{1/2}}{b_1^2 - (t^2 s^2/S_{XX})} \tag{3.2.5}$$

or, in an alternative form,

$$= \hat{X}_0 + \frac{(\hat{X}_0 - \overline{X})g \pm (ts/b_1)\{[\hat{X}_0 - \overline{X})^2 S_{XX}] + (1-g)/n\}^{1/2}}{1-g}, \tag{3.2.6}$$

where $g = t^2 s^2/(b_1^2/S_{XX})$. When g is "small," say 0.05 or smaller, it is convenient to set $g = 0$ for an approximate answer. Note that we can write

$$\begin{aligned} g &= t^2/\{b_1/(s^2 S_{XX})^{1/2})^2 \\ &= \left\{\frac{t(\nu, 1 - \alpha/2) \text{ percentage point}}{t\text{-statistic derived from } b_1 \text{ divided by its standard error}}\right\}^2. \end{aligned} \tag{3.2.7}$$

Thus the "more significant" b_1 is, the larger will be the denominator of g and the

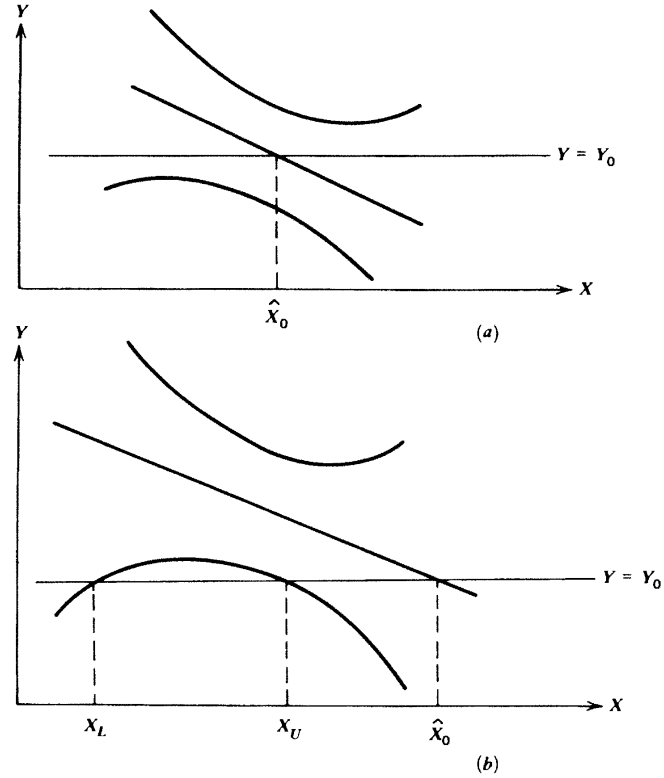

Figure 3.3. Inverse regression peculiarities: (*a*) complex roots and (*b*) real roots but both on the same side of the regression line. Inverse regression would not be of much practical value in these sorts of circumstances. The respective intervals are (*a*) $(-\infty, \infty)$ and (*b*) (X_U, ∞).

smaller will be g. Clearly g will tend to be large if $|b_1|$ is small or badly determined, in which case s^2 will tend to be large or S_{XX} will tend to be small or both. Inverse estimation is, typically, not of much practical value unless the regression is well determined, that is, b_1 is significant, which implies that g should be smaller than (say) about 0.20. (A t-statistic value of 2.236 would achieve this, for example.)

When the regression line is not well determined, peculiarities can arise. For example, the roots X_L and X_U may be complex; or they may both be real but both on one side of the regression line. Drawing a figure will usually make it quite obvious why the peculiarity has arisen. When the line is not well estimated, the hyperbolas defined by the end points of the confidence intervals usually flare out badly, or turn back down (or up as the case may be) sharply. Figure 3.3 shows two examples.

An alternative way of writing the quadratic equation (3.2.3), which enables the generalization to the multiple predictor case to be made more easily, is

$$\{-Y_0 + b_0 + b_1 X\}^2 = t^2 s^2 \left\{\frac{1}{n} + \frac{(X - \overline{X})^2}{S_{XX}}\right\}, \tag{3.2.8}$$

where the X represents the X_L in (3.2.2), or X_U.

Note: The calculations above are true mean value calculations. To obtain a more general formula in which Y_0 is regarded not as a true mean value but as the mean of q observations, replace each $1/n$ by $1/q + 1/n$ in Eqs. (3.2.2), (3.2.4), (3.2.5), (3.2.6), and (3.2.8), as described above. Then $q = 1$ provides an individual observation formula,

applicable to a single new reading, as in the carbon-dating example that introduced this section. When $q = \infty$ we obtain the true mean value formulas shown above.

We have adopted the term "fiducial limits" for (X_L, X_U) from Williams (1959) whose account of this topic in his Chapter 6 is excellent. Rather than becoming involved in the theoretical arguments behind the use of this term, we ask the reader simply to regard such intervals as inverse confidence limits for X given Y_0.

3.3. SOME PRACTICAL DESIGN OF EXPERIMENT IMPLICATIONS OF REGRESSION

We have dealt with the fitting of a straight line model $Y = \beta_0 + \beta_1 X + \epsilon$ to a set of data (X_i, Y_i), $i = 1, 2, \ldots, n$. We have also carefully considered a detailed analysis of how well the line fits, and whether or not repeated observations indicated whether or not there was any evidence in the data to indicate that an alternative model should be used. When we have only one predictor variable X and when the postulated model is a straight line, the alternatives we would consider would often be higher-order polynomials in X; for example, the quadratic $Y = \beta_0 + \beta_1 X + \beta_{11} X^2 + \epsilon$, or the cubic, and so on. We now put all this information into a practical perspective by considering the problem of choosing an experimental strategy for the one-predictor-variable case.

Experimental Strategy Decisions

Suppose an experimenter wants to collect data on a response variable Y at n selected values of a controllable predictor variable in order to determine an empirical relationship between Y and the predictor variable. We assume that the latter is (at least to a satisfactory approximation) not subject to random error, but that Y is, and that the n values of the predictor are not necessarily all distinct, that is, repeat runs are permitted. The experimenter first has to ask, and answer, a number of questions.

1. What range of values of the predictor variable is she currently interested in? This is often difficult to decide. The range must be wide enough to permit useful inference, yet narrow enough to permit representation by as simple a model function as possible. Once the decision is made, the interval can be coded to $(-1, 1)$ without loss of generality. For example, if a temperature range of $140°\text{F} \leq T \leq 200°\text{F}$ is selected, the coding

$$X = (T - 170)/30$$

will convert it to the interval $-1 \leq X \leq 1$. In general, the transformation is given by

$$X = \frac{\text{Original variable} - \text{Midpoint of original interval}}{\text{Half the original range}}.$$

2. What kind of relationship does the experimenter anticipate will hold over the selected range? Is it first order (i.e., straight line), second order (i.e., quadratic), or what? To decide this, she will not only bring to bear her own knowledge but will usually seek the expertise of others as well. To fix ideas, let us assume the experimenter believes the relationship is probably first order but is not absolutely sure.

3. If the relationship tentatively decided upon in (2) is wrong, what alternative does the experimenter expect? For example, if she believes the true model is a straight line, she is most likely to expect that, if she is wrong, the model will be somewhat

curved in a quadratic manner. A more remote possibility is that the true model may be cubic. Typically, she will decide that she *may be* one order too low, if anything. Otherwise, she would probably postulate a higher-order model to begin with.

4. What is the inherent variation in the response? That is, what is $V(Y) = \sigma^2$? The experimenter may have a great deal of experience with similar data and may "know" what σ^2 is. More typically, she may wish to incorporate repeat runs into the experiments so that σ^2 can be estimated at the same time as the relationship between Y and X, and also so that the usual assumptions about the constancy of σ^2 throughout the chosen range of values of the predictor can be checked.

5. How many experimental runs are possible? The experimenter has only limited dollars, staff, facilities, and time. How many runs are justified by the importance of the problem and the costs involved?

6. How many sites (i.e., different X-values) should be chosen? How many repeat runs should be performed at each of the chosen sites?

Let us now continue our discussion in terms of a specific example.

An Example

Suppose our experimenter decides that a straight line relationship is most likely over the range $-1 \leq X \leq 1$ of her coded predictor, that she most fears a quadratic alternative, that she does not know σ^2, and that 14 runs are possible. At what values of X (i.e., what sites) should she perform experimental runs, how many where, and with what justification?

Figure 3.4 shows some of the possibilities. (Each dot represents a run; a pile of dots represents repeated runs.) Let us see how each matches the requirements we have set out.

Each design has 14 degrees of freedom to begin with. Two of these are taken up by the parameter estimates b_0 and b_1 leaving 12 residual degrees of freedom to be allocated between lack of fit and pure error. Lines (1) and (2) of Table 3.1 show how these residual degrees of freedom split up for the various designs. Line (3) gives the value of

$$S_{\bar{X}X}^{1/2} = \{\Sigma(X_i - \bar{X})^2\}^{-1/2},$$

which, by Eq. (1.4.6), is proportional to the standard deviation of b_1 when a straight line is fitted. Line (4) shows the number of parameters that it is *possible* to fit to the design data. We can fit a polynomial of order $p - 1$ (with p parameters including β_0) to a design with p sites. A second reason that this is shown is that p is proportional (when n and σ^2 are fixed) to $p\sigma^2/n$, and the latter is the average size of $V(\hat{Y}(X))$ averaged over all the points of the design when the polynomial of order $p - 1$ is fitted. In other words

$$\sum_{i=1}^{n} V\{\hat{Y}(X_i)\}/n = p\sigma^2/n.$$

This result is true in general for any linear model. For the straight line case, when $p = 2$, it can be deduced by replacing the subscript 0 by i in Eq. (3.1.3), summing from $i = 1, 2, \ldots, n$, and dividing by n. For a general proof using matrices see Exercise R in "Exercises for Chapters 5 and 6."

Note that the lack of fit degrees of freedom is equal to the number of distinct X-sites in the data minus the number of parameters in the postulated model. In fact,

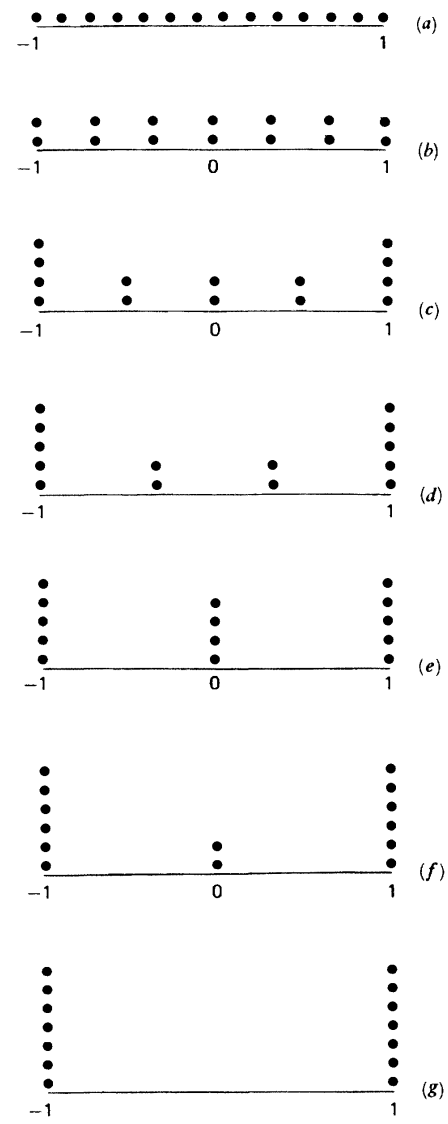

Figure 3.4. Some possible experimental arrangements for obtaining data for fitting a straight line: (*a*) 14 sites, (*b*) 7 sites, (*c*) 5 sites, (*d*) 4 sites, (*e*) 3 sites, (*f*) 3 sites, and (*g*) 2 sites. Which are good, which are bad, in the circumstances described in the text? The sites are equally spaced in cases (*a*)–(*f*).

TABLE 3.1. Characteristics of Various Strategies Depicted in Figure 3.4

	(*a*)	(*b*)	(*c*)	(*d*)	(*e*)	(*f*)	(*g*)
(1) Lack of fit df:	12	5	3	2	1	1	0
(2) Pure error df:	0	7	9	10	11	11	12
(3) sd $(b_1)/\sigma$:	0.43	0.40	0.33	0.31	0.32	0.29	0.27
(4) p sites:	14	7	5	4	3	3	2

because our example has two parameters to be estimated, β_0 and β_1, line (4) minus line (1) equals 2 throughout Table 3.1.

Comments on Table 3.1

The requirement in our example that σ^2 needs to be estimated via pure error makes (*a*) a poor strategy here. If we are to be able to check lack of fit, (*g*) is automatically eliminated, too.

Next consider (*b*). Is it really sensible to use seven different levels if our major alternative is a quadratic model? Not really, because we do not need that many levels to check our alternative. Moreover, of designs remaining, it has the highest $\text{sd}(b_1)/\sigma$. So we drop (*b*) from consideration.

Our best choice clearly lies with one of (*c*), (*d*), (*e*), and (*f*); exactly which would be selected depends on the experimenter's preferences. Only three levels (sites) are strictly necessary to achieve a lack of fit test versus a quadratic alternative but there is only one degree of freedom for lack of fit in (*e*) and (*f*). The latter is a better choice on the basis of the $\text{sd}(b_1)/\sigma$ values. Design (*d*) allows two df for lack of fit; design (*c*) perhaps goes too far in number of sites. So the final choice lies between (*f*) and (*d*) with (*f*) slightly preferred perhaps if the quadratic alternative is all that is anticipated.

Perhaps the most important aspect of this discussion is not so much a specific choice of design, but the immediate elimination of designs that might in some other contexts be regarded as reasonable. For example, design (*a*) would be a very poor choice—who needs 14 levels to estimate a straight line? Again, design (*g*) provides the smallest variance for the slope b_1 but is of no use at all if we want to be able to check possible lack of fit against a quadratic (or indeed any) alternative. When a design must be chosen from a list of alternatives, we recommend consideration of details like those in Table 3.1; such a display can be both helpful and revealing.

3.4. STRAIGHT LINE REGRESSION WHEN BOTH VARIABLES ARE SUBJECT TO ERROR[1]

Whenever we fit the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \qquad i = 1, 2, \ldots, n, \tag{3.4.1}$$

by least squares to a set of n data values (X_i, Y_i), we usually take it for granted that Y_i is subject to the error ϵ_i and X_i is not subject to error. If this is true, and if the errors $\epsilon_1, \epsilon_2, \ldots, \epsilon_n$ are independently and normally distributed, $N(0, \sigma^2)$, maximum likelihood estimation, and least squares estimation, namely,

$$\underset{\beta_0, \beta_1}{\text{Minimize}} \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

provide the same estimates (b_0, b_1) of (β_0, β_1).

What if both X and Y are subject to error? We can write

$$Y_i = \eta_i + \epsilon_i, \tag{3.4.2}$$

$$X_i = \xi_i + \delta_i. \tag{3.4.3}$$

[1]This section is a condensed version of Draper (1992).

We assume that a straight line relationship

$$\eta_i = \beta_0 + \beta_1 \xi_i \tag{3.4.4}$$

holds between the true but unobserved values η_i and the n unknown parameters ξ_i. Substituting (3.4.4) into (3.4.2) and then substituting for ξ_i from (3.4.3) gives

$$Y_i = \beta_0 + \beta_1 X_i + (\epsilon_i - \beta_1 \delta_i). \tag{3.4.5}$$

Let us assume that $\epsilon_i \sim N(0, \sigma^2)$, with the ϵ_i uncorrelated, and $\delta_i \sim N(0, \sigma_\delta^2)$, with the δ_i uncorrelated, with ϵ_i and δ_i uncorrelated, and define

$$\sigma_\xi^2 = \sum_{i=1}^{n} (\xi_i - \bar{\xi})^2/n, \tag{3.4.6}$$

$$\sigma_{\xi\delta} = \text{Covariance } (\xi, \delta), \tag{3.4.7}$$

$$\rho = \sigma_{\xi\delta}/(\sigma_\xi \sigma_\delta), \tag{3.4.8}$$

$$r = \sigma_\delta/\sigma_\xi. \tag{3.4.9}$$

In (3.4.7), $\sigma_{\xi\delta}$ would typically be zero; however, see case (2) below. If, mistakenly, we fit (3.4.1) by least squares, b_1 will be biased. In fact,

$$E(b_1) = \beta_1 - \frac{\beta_1 r(\rho + r)}{1 + 2\rho r + r^2}. \tag{3.4.10}$$

The bias is negative if $\sigma_\xi^2 + \sigma_{\xi\delta} > 0$, that is, if $\rho + r > 0$. The bias arises from the fact that X_i is not independent of the error in (3.4.5), in general. In fact,

$$\text{Covariance } [X_i, (\epsilon_i - \beta_1 \delta_i)] = -\beta_1(\rho + r)\sigma_\xi \sigma_\delta. \tag{3.4.11}$$

We thus see that there are cases where fitting (3.4.1) by least squares will provide little or no bias. These are:

1. If σ_δ^2 is small compared with σ_ξ^2, the errors in the X's are small compared with the spread in the ξ_i's (and so in the X's) and r will be small. The bias in (3.4.10) is then small. This is what is often assumed in practice, when least squares is used.

2. If the X's are fixed and determined by the experimenter (see Berkson, 1950), then $\sigma_{\xi\delta} = \text{Covariance}(X_i - \delta, \delta) = -\sigma_\delta^2$, which means that $\sigma_{\xi\delta} + \sigma_\delta^2 = 0$, or $\rho + r = 0$, implying zero bias in (3.4.10).

3. We wish to fit $Y_i = \eta_i + \epsilon_i$, where $\eta_i = \beta_0 + \beta_1 X_i$ (the observed X_i, note), and not as in (3.4.4).

These formats will not fit all practical cases. One case that occurred at the University of Wisconsin, in connection with a study on wild birds, required the observation of X_i = "the distance the bird was from a path." The student doing the study pointed out that, as she approached a bird, it flew away before she got close enough to see precisely where it had perched. Thus error in recording X was unavoidable.

If we attempt to obtain the maximum likelihood estimates of β_0 and β_1 under the distributional assumptions made in connection with (3.4.5), we find that there is an identifiability problem. The estimation cannot be carried through without some additional information being added, for example, knowledge of the ratio $\lambda = \sigma^2/\sigma_\delta^2$ (Barnett, 1967; Wong, 1989). This is Case III of Sprent and Dolby (1980), discussed below. Various authors have suggested alternative analyses. The literature is too vast to discuss in full detail here, and we provide a selective, perhaps biased, discussion.

Sprent and Dolby (1980) distinguish four cases:

I. (X, Y) are bivariate normal variables and $E(Y|X) = \beta_0 + \beta_1 X$.

II. $Y \sim N(\beta_0 + \beta_1 X, \sigma^2)$. The observed X-values are fixed on realizations of a random variable with any (reasonable) distribution.

In both I and II, estimates via maximum likelihood are the usual least squares estimates $b_1 = S_{XY}/S_{XX}$ and $b_0 = \overline{Y} - b_1\overline{X}$, where $S_{XY} = \Sigma_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})$ and $S_{XX} = \Sigma_{i=1}^n (X_i - \overline{X})^2$.

III. The case of Eqs. (3.4.1) through (3.4.5). If λ were known, maximum likelihood leads to estimates

$$\begin{aligned} \hat{\beta}_1 &= [S_{YY} - \lambda S_{XX} + \{(S_{YY} - \lambda S_{XX})^2 + 4\lambda S_{XY}^2\}^{1/2}]/(2S_{XY}) \\ \hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1\overline{X}, \end{aligned} \tag{3.4.12}$$

where $S_{YY} = \Sigma_{i=1}^n (Y_i - \overline{Y})^2$. Note that, if $\lambda = S_{YY}/S_{XX}$, $\hat{\beta}_1 = (S_{YY}/S_{XX})^{1/2}$, which is the geometric mean functional relationship, after attachment of the sign of S_{XY}. This case is often called the functional relationship model. Note also that, when $\lambda = 1$, the solution (3.4.12) defines the line that minimizes the sum of squares of *perpendicular* deviations from the line. Many people find such a solution intuitively satisfying, but it is appropriate only when $\sigma^2 = \sigma_\delta^2$, that is, when $\lambda = 1$. This solution was first given by Adcock (1878).

IV. Similar to III but with ξ_i a normal random variable, independent of δ, so that because of (3.4.4), (ξ_i, η_i) follow a joint degenerate bivariate normal distribution. This is the so-called structural relationship model and again the case III solution applies if λ is known.

Sprent and Dolby "do not recommend *ad hoc* use of the geometric mean functional relationship when there are errors in both variables," arguing that other *ad hoc* estimates could equally be used. Nevertheless, we shall recommend it below for reasons to be stated.

Riggs, Guarnieri, and Addelman (1978) study, partially through simulations, a variety of 34 different methods of fitting (X, Y) data. While they favor (3.4.12), they warn that a reasonably accurate estimate of λ is desirable. They also point out that the geometric mean functional relationship occupies a "central position" in compromises between the two least squares solutions, Y on X and X on Y, an appealing characteristic (see their Figure 8, p. 1338).

Practical Advice

Although many of us try to avoid the issue of errors in both X and Y by advising "take data where the X-range is large compared with the X-error," this cannot always be done, and one must often suggest something specific. If λ is known (or can reasonably be estimated) use of the maximum likelihood solution (3.4.12) is probably best.

A simple alternative initially suggested by Wald (1940), using two groups, and amended by Bartlett (1949) to three groups is the following: Divide the data into three equal (or as equal as possible) groups with: (1) the smaller, or most negative, X-values; let $P_1 \equiv (\overline{X}_1, \overline{Y}_1)$ be the center of gravity of these. (2) The larger, or least negative, X-values; let $P_3 \equiv (\overline{X}_3, \overline{Y}_3)$ be their center of gravity. (3) The remainder, which are used only in estimating the overall center of gravity, $(\overline{X}, \overline{Y})$. Use the line passing through $(\overline{X}, \overline{Y})$ with slope $(\overline{Y}_3 - \overline{Y}_1)/(\overline{X}_3 - \overline{X}_1)$, that is, parallel to P_1P_3. For

reasoning, see Wald (1940) and Barlett (1949). Later studies by Gibson and Jowett (1957) indicate that maximum efficiency is achieved by a division of observations closer to the ratio 1:2:1, but the exact split is not crucial.

Geometric Mean Functional Relationship

Our own preference is to suggest the geometric mean functional relationship for which the estimators are

$$\hat{\beta}_1 = (S_{YY}/S_{XX})^{1/2}, \qquad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{3.4.13}$$

This does assume, it is true, that we are using (3.4.12) with $\lambda = S_{YY}/S_{XX}$. However, it is appealing otherwise, if this assumption is not unreasonable. The estimator $\hat{\beta}_1$ is the geometric mean of the quantities

$$b_1 = S_{XY}/S_{XX}, \qquad a_1^{-1} = (S_{XY}/S_{YY})^{-1},$$

where b_1 and a_1 are, respectively, the slopes in least squares fits of Y versus X ($\hat{Y} = b_0 + b_1 X$) and of X versus Y ($\hat{X} = a_0 + a_1 Y$). Inverting the latter relationship leads to

$$Y = -a_0/a_1 + a_1^{-1}\hat{X}$$

so the geometric mean $\hat{\beta}_1 = (b_1 a_1^{-1})^{1/2}$ is a compromise value lying in between the two "Y on X equation" slopes. Note that, if the roles of X and Y are reversed, exactly the same line emerges, that is, the fitted line

$$Y = \hat{\beta}_0 + \hat{\beta}_1 X \tag{3.4.14}$$

with coefficients from (3.4.13) is uniquely defined. This natural symmetry is most appealing. The attractiveness of the geometrical mean functional relationship has greatly been enhanced by the independent discoveries of Teissier (1948) and Barker, Soh, and Evans (1988) that this solution is an optimum solution to a specific problem. (See also Harvey and Mace, 1982.) That is, the geometric mean functional relationship minimizes the sum of the areas obtained by drawing horizontal (parallel to the X-axis) and vertical (parallel to the Y-axis) lines from each data point (see Figure 3.5). The symmetry of the solution is again obvious; interchange of the X and Y axes leaves the areas unchanged. One disadvantage of the geometric mean functional relationship is that no easy calculations are available for conducting tests on the parameters or constructing confidence intervals for them. (For the complications involved, see, for example, Creasy, 1956.) It is possible that applying nonlinear estimation and defining the loss function as the sum of the areas might offer some help here. The geometric

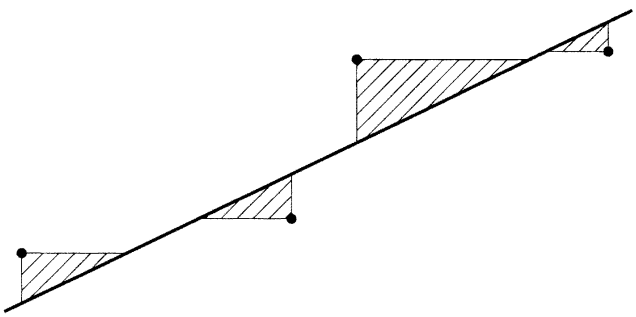

Figure 3.5. The geometrical mean functional relationship line minimizes the sum of the shaded areas. (The dots are data points.)

TABLE 3.2. Star Data from Dressler (1984) and Jefferys (1990)[a]

Coma Sample ($z = 1$)		Virgo Sample ($z = 0$)	
V_{26}	$\log \sigma$	V_{26}	$\log \sigma$
12.60	2.449	11.39	2.242
13.12	2.394	12.53*	1.716*
14.23	2.285	9.98	2.412
14.86	2.166	9.37	2.480
15.88*	1.863*	12.24	2.059
13.92	2.286	9.20	2.355
15.45*	1.761*	12.17	2.009
14.36	2.209	12.01	1.949
15.07	2.113	12.50	2.079
14.07	2.301	8.56	2.474
14.53	2.243	10.28	2.268
15.60	2.169	11.28	2.170
12.27	2.383	8.79	2.528
14.36	2.311	12.02*	1.778*
14.50	2.339	11.92	2.021
15.52	2.251	9.95	2.391
13.46	2.361	11.30	2.185
11.85	2.584	9.88	2.338
15.31	2.007	9.82	2.303
13.98	2.180	8.90	2.514
15.28	2.099	10.97	2.262
14.26	2.275	9.28	2.276
14.11	2.320	11.37	2.207
14.87	2.191		
14.82	2.247		
15.37	2.059		
13.49	2.394		
15.04	2.154		
13.67	2.274		
12.88	2.383		

[a]In the example, the four observations marked with an asterisk will be ignored.

mean functional relationship has been condemned as being inconsistent, that is, the estimates do not tend to their true values as n tends to infinity. However, other estimators are biased, and what happens for large n is often not of concern to those with practical problems and small data sets. While it is true that maximum likelihood methods can make appeal to asymptotic results at this point, such results do not seem to apply too well when n is small, judging by the comments of various authors.

Examples

Example 1. The data in Table 3.2 were used by Jefferys (1990) and taken from Dressler (1984). "They consist of the integrated V magnitudes V_{26}, and log of the central velocity dispersion, log σ, of a sample of 53 galaxies from two galaxy clusters, the Coma and Virgo cluster" (Jefferys, 1990, p. 602). The model

$$\log \sigma = \beta_0 + \beta_1 V_{26}$$

is deemed appropriate with a common β_1 and a different β_0 for each cluster. Four outliers are present (marked by asterisks in Table 3.2), which we ignore. (This bypasses some of the points made by Jefferys, which are not our concern here.) We adopt a dummy or indicator variable z; $z = 1$ for the Coma sample and $z = 0$ for the Virgo sample. Two least squares fits using the models

$$\log \sigma = \beta_0 + \beta_1 V_{26} + \beta_2 z + \epsilon$$

and

$$V_{26} = \alpha_0 + \alpha_1 \log \sigma + \sigma_2 z + \epsilon$$

provide, respectively, fitted equations

$$\hat{\log} \sigma = 3.4795 - 0.116334 V_{26} + 0.44097 z$$

and

$$\hat{V}_{26} = 25.329 - 6.5641 \log \sigma + 3.7685 z.$$

The slope of the geometric mean functional relationship is thus

$$-\{-0.116334/(-6.5641)\}^{1/2} = -0.133127.$$

Putting parallel straight lines with this slope through the individual centers of gravity of the two sets of data provides fitted equations

$$\hat{\log} \sigma = 4.159 - 0.133 V_{26} \quad \text{(Coma sample)}$$

and

$$\hat{\log} \sigma = 3.656 - 0.133 V_{26} \quad \text{(Virgo sample)}.$$

These are very close to the reference solution of Jefferys (1990), which was "an errors-in-variables least squares fit" to the same data. (The method is not further explained.) They are virtually identical to Dressler's (1984) values obtained via a sensible ad hoc procedure (Jefferys's: 4.14, 3.65, −0.132; Dressler's: 4.156, 3.656, −0.1333).

Example 2. The data in Table 3.3, from Kelly (1984), were taken from Miller (1980). Kelly uses the data to illustrate points she is making about (i) estimating the variance of the classical estimators of (3.4.12) and (ii) detecting influential observations. We analyze them using the geometric mean functional relationship estimator.

The two fits to all the data (X = heelstick, Y = catheter) are $\hat{Y} = 2.786 + 0.8805X$, and $\hat{X} = 4.210 + 0.7870Y$, which we can invert to the form $Y = -5.349 + 1.2706\hat{X}$. The geometric mean functional relationship is thus $Y = -0.91 + 1.058X$. Both individual regressions indicate that the second observation is influential, however, and a plot of the data indicates we might consider dropping it. The two fits to the remaining 19 observations give

$$\hat{Y} = -1.628 + 1.1147X$$

and

$$\hat{X} = 5.482 + 0.70462Y,$$

which we can invert to the form

$$Y = -7.780 + 1.4192\hat{X}.$$

TABLE 3.3. Serum Kanamycin Levels in Blood Samples Drawn Simultaneously from an Umbilical Catheter and a Heel Venipuncture in 20 Babies

Baby	Heelstick (X)	Catheter (Y)
1	23.0	25.2
2	33.2	26.0
3	16.6	16.3
4	26.3	27.2
5	20.0	23.2
6	20.0	18.1
7	20.6	22.2
8	18.9	17.2
9	17.8	18.8
10	20.0	16.4
11	26.4	24.8
12	21.8	26.8
13	14.9	15.4
14	17.4	14.9
15	20.0	18.1
16	13.2	16.3
17	28.4	31.3
18	25.9	31.2
19	18.9	18.0
20	13.8	15.6

Then $\hat{\beta} = (1.1147/0.70462)^{1/2} = 1.258$ and the geometric mean functional relationship is

$$\hat{Y} = -4.52 + 1.258X.$$

(If the second observation is not deleted, the parallel result would be $\hat{Y} = -0.91 + 1.058X$, where the 1.058 is the geometric mean of the slopes 0.8805 and 1.2706.) Kelly (1984) obtains two 95% confidence intervals for the slope using all the data, getting (0.76, 1.38) via a bootstrap method, and (0.76, 1.52) via a method based on normal assumptions, given by Kendall and Stuart (1961, pp. 388–390). She concludes that these support the hypothesis that $\beta_0 = 0$, $\beta_1 = 1$, which implies that the methods of measurement that gave rise to Table 3.3 are equivalent. She then points out that removal of the second observation takes the estimated point for (β_0, β_1) "to approximately the edge of a 60% confidence region around" her original estimates based on a maximum likelihood analysis assuming $\lambda = 1$. We interpret that to mean that the hypothesis $\beta_0 = 0$, $\beta_1 = 1$ is no longer supported.

The geometric mean functional relationship does not provide confidence intervals, but we can get a rough feel for the situation by looking at the estimates when all equations are written in Y on X form. When observation 2 is included, the two slopes are 0.8805 and 1.2706 and their geometric mean is 1.0577; the two intercept values are 2.786 and −5.349 and the intercept of the geometric mean functional relationship is −0.91. One feels that the hypothesis—intercept = 0, slope = 1—is not unreasonable. Now remove the second observation. The slopes are now 1.1147 and 1.4192 with a geometric mean of 1.258 (all > 1) and the two intercepts are −1.628 and −7.780 (both < 0) with an intercept of −4.52 from the geometric mean functional relationship. The impression we get is that the hypothesis is not valid. Thus the situation turns on the one influential data point. Can we regard the two lines that lead to the geometric

mean functional relationship as confidence limits of some sort? No properties of them are known, it seems, but using them appears to be common sense.

References

Adcock (1878); Barker, Soh, and Evans (1988); Barnett (1967); Barlett (1949); Creasy (1956); Draper (1992); Dressler (1984); Gibson and Jowett (1957); Harvey and Mace (1982); Jefferys (1990); Kelly (1984); Kendall and Stuart (1961); Miller (1980); Riggs, Guarnieri and Addelman (1978); Sprent and Dolby (1980); Teissier (1948); Wald (1940); Wong (1989). (We are grateful to Dr. Penny Reynolds for supplying some of these references.)

EXERCISES FOR CHAPTERS 1–3

Readers who have not yet read all the material needed in Chapters 1–3 should temporarily skip over parts of questions that refer to unread portions. The residuals from all fits should be calculated and examined, whether or not it is specified in the question.

A. A study was made on the effect of temperature on the yield of a chemical process. The following data (in coded form) were collected:

X	Y
−5	1
−4	5
−3	4
−2	7
−1	10
0	8
1	9
2	13
3	14
4	13
5	18

1. Assuming a model, $Y = \beta_0 + \beta_1 X + \epsilon$, what are the least squares estimates of β_0 and β_1? What is the prediction equation?
2. Construct the analysis of variance table and test the hypothesis $H_0: \beta_1 = 0$ with an α risk of 0.05.
3. What are the confidence limits ($\alpha = 0.05$) for β_1?
4. What are the confidence limits ($\alpha = 0.05$) for the true mean value of Y when $X = 3$?
5. What are the confidence limits ($\alpha = 0.05$) for the difference between the true mean value of Y when $X_1 = 3$ and the true mean value of Y when $X_2 = -2$?
6. Are there any indications that a better model should be tried?
7. Comment on the number of levels of temperature investigated with respect to the estimate of β_1 in the assumed model.

B. A test is to be run on a given process for the purpose of determining the effect of an independent variable X (such as process temperature) on a certain characteristic property of the finished product Y (such as density). Four observations are to be taken at each of five settings of the independent variable X.

1. In what order would you take the 20 observations required in this test?
2. When the test was actually run, the following results were obtained:

$$\overline{X} = 5.0 \qquad \Sigma(X_i - \overline{X})^2 = 160.0 \qquad \Sigma(X_i - \overline{X})(Y_i - \overline{Y}) = 80.0$$
$$\overline{Y} = 3.0 \qquad \Sigma(Y_i - \overline{Y})^2 = 83.2.$$

Assume a model of the type $Y = \beta_0 + \beta_1 X + \epsilon$.

a. Calculate the fitted regression equation.

b. Prepare the analysis of variance table.

c. Determine 95% confidence limits for the true mean value of Y when

(1) $X = 5.0$;

(2) $X = 9.0$.

3. Suppose that the actual data plot is as shown in Figure B1 and that the sum of squares due to replication (pure error) is 42. Answer the following questions on the basis of this additional information.

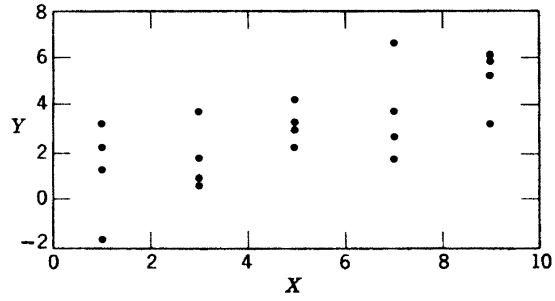

Figure B1

a. Does the regression equation you derived in 2a adequately represent the data? Give reasons and include the result of a test for lack of fit.

b. Are the confidence limits you calculated in 2c applicable? If not, state your reasons.

c. If the model used in part 2 does not seem appropriate, suggest a possible alternate.

4. Suppose that the actual data plot is as shown in Figure B2 and that the sum of squares due to replication is 42.0. Answer the questions 3a, 3b, and 3c on the basis of this information.

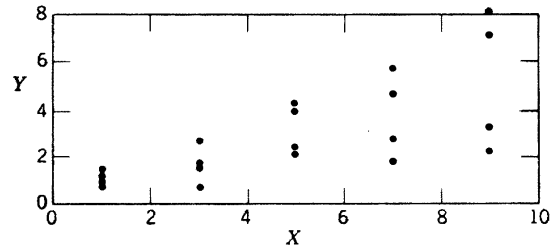

Figure B2

5. Suppose that the actual data plot is as shown in Figure B3, and the sum of squares due to replication is 23.2. Answer the questions 3a, 3b, and 3c on the basis of this information.

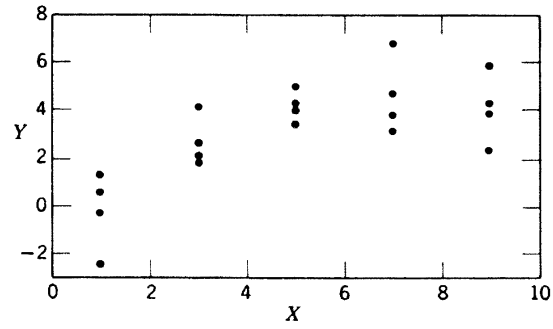

Figure B3

C. Thirteen specimens of 90/10 Cu–Ni alloys, each with a specific iron content, were tested in a corrosion-wheel setup. The wheel was rotated in salt seawater at 30 ft/s for 60 days. The corrosion was measured in weight loss in milligrams/square decimeter/day, MDD. The following data were collected:

X (Fe)	Y (loss in MDD)
0.01	127.6
0.48	124.0
0.71	110.8
0.95	103.9
1.19	101.5
0.01	130.1
0.48	122.0
1.44	92.3
0.71	113.1
1.96	83.7
0.01	128.0
1.44	91.4
1.96	86.2

Requirement. Determine if the effect of iron content on the corrosion resistance of 90/10 Cu–Ni alloys in seawater can justifiably be represented by a straight line model. Assume $\alpha = 0.05$.

D. (*Source:* "Leverage and regression through the origin," by G. Casella, *The American Statistician,* **37,** 1983, 147–152.) There are very few occasions where it makes sense to fit a model without an intercept β_0. If there *were* occasion to fit the model $Y = \beta X + \epsilon$ to a set of data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$, the least squares estimate of β would be

$$b = \Sigma X_i Y_i / \Sigma X_i^2.$$

Suppose you have a programmed calculator that will fit only the intercept model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

but you want to fit the no-intercept model. By adding one more fake data point $(m\bar{X}, m\bar{Y})$ to the data above, where

$$m = n/\{(n + 1)^{1/2} - 1\} = n/a,$$

say, and letting the calculator fit $Y = \beta_0 + \beta_1 X + \epsilon$, you can estimate β by using the b_1 estimate but ignoring b_0, which is not zero in general. Try to show this algebraically.

Check this out with the very small data set $(X, Y) = (1, 0.5), (3, 1.0), (5, 0.9)$ for which $(m\bar{X}, m\bar{Y}) = (9, 2.4)$.

(In MINITAB, the subcommand "noconstant" will get you the fit through the origin.)

E. The data below consist of seven pairs of values of

X = price of alcohol relative to take-home pay,

Y_1 = consumption of absolute alcohol in liters per head of population per year,

Y_2 = cirrhosis deaths per 100,000 of populations,

for seven European countries. Plot Y_1 versus X and Y_2 versus X (Y as ordinate, X as abscissa) and judge whether, in your opinion, a straight line would fit each set of data reasonably well.

u	Country	X	Y_1	Y_2
1	France	0.016	24.66	51.7
2	Italy	0.027	18.00	30.5
3	West Germany	0.026	13.63	29.0

u	Country	X	Y_1	Y_2
4	Belgium	0.022	8.42	14.2
5	United Kingdom	0.057	7.66	4.1
6	Ireland	0.092	7.64	5.0
7	Denmark	0.096	7.50	11.6

1. For the (X, Y_2) only, evaluate the fitted least squares line and draw it on your data plot.
2. Evaluate the residuals to two decimal places. Check that $\Sigma e_u = 0$, within rounding error.
3. Obtain the analysis of variance table.
4. Find the standard errors of b_0 and b_1.
5. Find the formula for the standard error of $\hat{Y}$, and construct 95% confidence bands for the true mean value of Y. (Plot about half a dozen points over the X-range and join them up smoothly.)
6. Test the overall regression via an F-test, and find how much of the variation about the mean $\overline{Y}$ is explained by the fitted line.
7. The data have no actual repeats. One practical trick that is often useful, however, is to consider points that are "sufficiently close together" as being, approximately, repeat runs, and to use this as a basis for working out an approximate pure error sum of squares. After this has been evaluated, the usual steps are followed. The major difficulty in doing such a thing is in deciding exactly what the words "sufficiently close together" mean.

   For our data, it might be sensible to regard the following sets of runs as approximate repeats.

   a. $X = 0.027, 0.026, 0.022$.

   b. $X = 0.092, 0.096$.

   Evaluate an approximate pure error sum of squares using these sets, and proceed with the appropriate approximate analysis. State your conclusions.
8. Reviewing the whole problem, comment on the apparent value and usefulness of the straight line you have fitted. It has been suggested by some that, if the price of alcohol were raised, fewer cirrhosis deaths would result. On the basis of these data, what do *you* think? What possible flaws might exist in your conclusion? (Discuss the situation, briefly.)

F. The moisture of the wet mix of a product is considered to have an effect on the finished product density. The moisture of the mix was controlled and finished product densities were measured as shown in the following data:

Mix Moisture (Coded) X	Density (Coded) Y
4.7	3
5.0	3
5.2	4
5.2	5
5.9	10
4.7	2
5.9	9
5.2	3
5.3	7
5.9	6
5.6	6
5.0	4
$\Sigma X = 63.6$	$\Sigma Y = 62$
$\Sigma X^2 = 339.18$	$\Sigma Y^2 = 390$
$\Sigma x^2 = 2.10$	$\Sigma y^2 = 69.67$
$\overline{X} = 5.3$	$\overline{Y} = 5.17$

$\Sigma XY = 339.1$

$\Sigma xy = 10.5$

Requirements

1. Fit the model $Y = \beta_0 + \beta_1 X + \epsilon$ to the data.
2. Place 95% confidence limits on β_1.
3. Is there any evidence in the data that a more complex model should be tried? (Use $\alpha = 0.05$.)

G. The cost of the maintenance of shipping tractors seems to increase with the age of the tractor. The following data were collected.

Age (yr) X	6 Months Cost ($) Y
4.5	619
4.5	1049
4.5	1033
4.0	495
4.0	723
4.0	681
5.0	890
5.0	1522
5.5	987
5.0	1194
0.5	163
0.5	182
6.0	764
6.0	1373
1.0	978
1.0	466
1.0	549

Requirements

1. Determine if a straight line relationship is sensible. (Use $\alpha = 0.10$.)
2. Can a better model be selected?

H. It has been proposed in a manufacturing organization that a cup loss figure performed on the line supersede a bottle loss analysis, which is a costly, time-consuming laboratory procedure. The cup loss analysis would yield better control of the process because of the gain in time. If it can be shown that cup loss is a function of bottle loss, it would be a reasonable decision to make. Given the following data, what is your conclusion? (Use $\alpha = 0.05$.)

Bottle Loss (%) X	Cup Loss (%) Y
3.0	3.1
3.1	3.9
3.0	3.4
3.6	4.0
3.8	3.6
2.7	3.6
3.1	3.1
2.7	3.6
2.7	2.9
3.3	3.6
3.2	4.1
2.1	2.6
3.0	3.1
2.6	2.8

I. It is thought that the number of cans damaged in a boxcar shipment of cans is a function of the speed of the boxcar at impact. Thirteen boxcars selected at random were used to examine whether this appeared to be true. The data collected were as follows:

Speed of Car at Impact X	Number of Cans Damaged Y
4	27
3	54
5	86
8	136
4	65
3	109
3	28
4	75
3	53
5	33
7	168
3	47
8	52

What are your conclusions? (Use $\alpha = 0.05$.)

J. The effect of the temperature of the deodorizing process on the color of the finished product was determined experimentally. The data collected were as follows:

Temperature X	Color Y
460	0.3
450	0.3
440	0.4
430	0.4
420	0.6
410	0.5
450	0.5
440	0.6
430	0.6
420	0.6
410	0.7
400	0.6
420	0.6
410	0.6
400	0.6

1. Fit the model $Y = \beta_0 + \beta_1 X + \epsilon$.
2. Is this model sensible? (Use $\alpha = 0.05$.)
3. Obtain a 95% confidence interval for the true mean value of Y at any given value of X, say, X_0.

K. The data below (provided by Tom B. Whitaker) show 34 pairs of values of

X = average level of aflatoxin in a mini-lot sample of 120 pounds of peanuts, ppb,

Y = percentage of noncontaminated peanuts in the batch − 99.

Y	X	Y	X	Y	X
0.971	3.0	0.942	18.8	0.863	46.8
0.979	4.7	0.932	18.9	0.811	46.8
0.982	8.3	0.908	21.7	0.877	58.1
0.971	9.3	0.970	21.9	0.798	62.3
0.957	9.9	0.985	22.8	0.855	70.6
0.961	11.0	0.933	24.2	0.788	71.1
0.956	12.3	0.858	25.8	0.821	71.3
0.972	12.5	0.987	30.6	0.830	83.2
0.889	12.6	0.958	36.2	0.718	83.6
0.961	15.9	0.909	39.8	0.642	99.5
0.982	16.7	0.859	44.3	0.658	111.2
0.975	18.8				

1. Plot the data (Y as ordinate, X as abscissa) and, "by eye" only, draw what appears to you to be a "well-fitting straight line 'through' the points." Keep this figure; you will need it for comparison purposes later. Would you say your line is a "good fit"?
2. Also, evaluate ΣX_i, ΣY_i, ΣX_i^2, ΣY_i^2, X_iY_i, all summations being from $i = 1$ through 34.
3. Fit the model $Y = \beta_0 + \beta_1 X + \epsilon$ by least squares. Draw the fitted line on your plot and check how good your "eye-fit" was.
4. Evaluate the residuals to three decimal places. Check that $\Sigma e_i = 0$, within rounding error. Examine plots of the residuals. Give conclusions.
5. Obtain the analysis of variance table in the form given in Table 1.4.
6. Find the standard errors of b_0 and b_1.
7. Find the formula for the standard error of $\hat{Y}$, and construct 95% confidence bands for the true mean value of Y. (Plot about half a dozen points over the X-range and join them up smoothly.)
8. Test the overall regression via an F-test, and find how much of the variation about the mean $\overline{Y}$ is explained by the fitted line.

L. The data given in Exercise K have only two pairs of actual repeats, in fact, at $X = 18.8$ and at $X = 46.8$. One practical trick that is often useful, however, is to consider points that are "sufficiently close together" as being, approximately, repeat runs, and to use this as a basis for working out an approximate pure error sum of squares. After this has been evaluated, the usual steps are followed. The major difficulty in doing such a thing is in deciding exactly what the words "sufficiently close together" mean.

For our data here, it would be sensible to regard the following seven sets of runs as approximate repeats:

$X = 9.3, 9.9$
$X = 12.3, 12.5, 12.6$
$X = 18.8, 18.8, 18.9$
$X = 21.7, 21.9$
$X = 46.8, 46.8$ (these are exact repeats, of course!)
$X = 70.6, 71.1, 71.3$
$X = 83.2, 83.6$

Evaluate an approximate pure error sum of squares using these sets, and proceed with the appropriate approximate analysis, following the details in Section 2.1. State your conclusions.

Reviewing the whole problem in Exercises K and L, comment on the apparent value and usefulness of the straight line you have fitted.

M. For fitting a straight line we have defined

$$R^2 = \text{SS}(b_1|b_0)/\Sigma(Y_i - \overline{Y})^2.$$

What is the maximum possible value of R^2 if:

1. There are no repeat runs at all in the data?
2. There *are* proper repeat runs in the data?

Are your conclusions extendable to general regression situations, do you think?

N. The straight line model $\hat{Y} = 1.692 - 0.0546X$ is fitted to the data below, and the analysis of variance table is as shown. Continue the analysis.

X	Y	$\hat{Y}$	e	X	Y	$\hat{Y}$	e
10	−2	1.1	−3.1	40	0	−0.5	0.5
10	−4	1.1	−5.1	40	1	−0.5	1.5
20	1	0.6	0.4	40	2	−0.5	2.5
20	3	0.6	2.4	50	−2	−1.0	−1.0
30	2	0.1	1.9	50	−3	−1.0	−2.0
30	5	0.1	4.9	50	−4	−1.0	−3.0

ANOVA

Source	df	SS	MS	F
b_0	1	0.083		
$b_1 \mid b_0$	1	7.240	7.240	0.845
Residual	10	85.677	8.568	
Total	12	93.000		

O. In a certain Federal Power Commission hearing some years ago, witnesses presented the following data (here rounded):

X	Y
13.3	3.5
16.9	5.1
19.9	4.8
23.2	6.7
26.3	6.0
30.1	9.5
42.6	8.1

The variables are

Predictor X = Percentage of liquids in gas production output;

Response Y = Unit cost in cents of processing output.

Fit a straight line to these data and find the residual variation via an analysis of variance table.

The witnesses extrapolated the straight line to the points $X = 0$ and $X = 100$ to provide unit costs for situations in which production consisted of no liquids and all liquids. Find the values of $\hat{Y}(0)$ and $\hat{Y}(100)$ and determine 95% confidence limits for the true mean value of Y at $X = 0$ and $X = 100$. What do you conclude?

P. Show that, for a straight line fit, $r^2_{XY} = R^2 = r^2_{Y\hat{Y}}$. (The second equality is true in general.)

Q. Consider the hypothesis that the titles of papers in a journal tend to be longer for short papers and shorter for long papers. Check this hypothesis in the following way. Look at recent issues of any journal in your own field and make a list, for a number of papers (the more the better, but let's say at least 25), of the following data:

X = Number of pages in a paper or note;

Y = Number of words in the paper or note title (hyphenated words count as one word).

Plot these data and fit a straight line $Y = \beta_0 + \beta_1 X + \epsilon$ to them via least squares.

Check for lack of fit, if repeat points (or near repeats) exist. Then, provided that there is no significant lack of fit or, if there is no pure error, provided that the residuals do not exhibit any obvious irregularity that would indicate lack of fit, test H_0:$\beta_1 = 0$ versus H_1:$\beta_1 < 0$ via a one-sided t test. What is your conclusion?

R. (*Source:* Ice crystal growth data from B. F. Ryan, E. R. Wishart, and D. E. Shaw, Commonwealth Scientific and Industrial Research Organisation (C.S.I.R.O.), Australia. See "The growth rates and densities of ice crystals between −3°C and −21°C," *Journal of the Atmospheric Sciences,* **33,** 1976, 842–850.)

Ice crystals are introduced into a chamber, the interior of which is maintained at a fixed temperature (−5°C) and a fixed level of saturation of air with water. The growth of the crystals with time is observed. The 43 sets of measurements presented here are of axial length of the crystals (A) in micrometers for times (T) of 50 seconds to 180 seconds from the introduction of the crystals. Each measurement represents a single complete experiment; the experiments were conducted over a number of days, and were randomized as to observation time. (The actual order in which they were conducted is not available.) It was desired to learn whether or not a straight line model $A = \beta_0 + \beta_1 T + \epsilon$ provided an adequate representation of the growth with time of the axial length of the ice crystal. Perform a full analysis and state your conclusions.

T	A	T	A
50	19	125	28
60	20, 21	130	31, 32
70	17, 22	135	34, 25
80	25, 28	140	26, 33
90	21, 25, 31	145	31
95	25	150	36, 33
100	30, 29, 33	155	41, 33
105	35, 32	160	40, 30, 37
110	30, 28, 30	165	32
115	31, 36, 30	170	35
120	36, 25, 28	180	38

S. Refer to the steam data in Table 1.1, and its subsequent analysis.

1. Suppose we specify a true mean value of $Y_0 = 10$. Use the methods of inverse estimation to provide an estimate $\hat{X}_0$ of the corresponding X value and 95% fiducial limits (X_L, X_U).
2. State the value of g you found in (1). Obtain the approximate "$g = 0$" (X_L, X_U) values and compare them with those in (1).
3. If Y_0 were not a true mean value but the result of a single new observation, what would $\hat{X}_0$ and (X_L, X_U) be?
4. State the value of g you found in (3). Obtain the approximate "$g = 0$" (X_L, X_U) values and compare them with those in (3).

Use accurate figures and carry enough places to minimize round-off errorrs.

T. Recall that

$$V(\hat{Y}_0) = \sigma^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right].$$

An experimenter tells you that he has drawn the loci (loci = plural of locus) of end points of 95% confidence bands for the true mean value of Y given X and that "for all practical purposes over the range I am interested in, $150 \leq X \leq 170$, these loci are parallel straight lines." Would you anticipate that the X-range of his data is

1. Much smaller than 20 X-units?
2. Roughly equal to 20 X-units? or
3. Much greater than 20 X-units?

Why?

U. 1. You are given some data (X_i, Y_i), $i = 1, 2, \ldots, n$ and asked to fit a straight line $\hat{Y} = b_0 + b_1 X$ to it by least squares, and to perform a "full analysis." As you finish, the experimenter comes in and says "I've just found out all my X's are biased. Each X value I gave you is only 90% of the true X. Would you please redo the analysis." Using $1 - q$ instead of 90% (so you can handle the problem more generally) explain what parts of the "full analysis" are changed by this new information, and in what manner. (*Hint:* New $S_{XX} \equiv S^*_{XX} = S_{XX}/(1-q)^2$, etc., so $b_1^* = S^*_{XY}/S^*_{XX} = b_1(1-q)$, and so on.)

2. What would happen if the proportion differed for each observation, for example, was $(1 - q_i)$ for $i = 1, 2, \ldots, n$?

3. If in (1) or (2) the q or q_i were small, so that q^2 or q_i^2 could be ignored, how would the usual full analysis be changed?

V. (*Source:* "Life expectancy" by M. E. Wilson and L. E. Mather, Letter to the Editor, *Journal of the American Medical Association,* **229,** No. 11, 1974, 1421–1422.) The 50 pairs of observations below arose during a study carried out by Dr. L. E. Mather and Dr. M. E. Wilson. The variables are

X = Age of person at death (to nearest year);

Y = Length of lifeline on left hand in centimeters (to nearest 0.15 cm).

Many people believe that the length of one's life is linearly related to the length of one's lifeline. What light do these data throw on such a belief? You may assume that:

$$\Sigma X = 3333, \quad \Sigma X^2 = 231{,}933, \quad \Sigma XY = 30{,}549.75,$$
$$\Sigma Y = 459.9, \quad \Sigma Y^2 = 4308.57.$$

X = Age (yr)	Y = Length (cm)	X = Age (yr)	Y = Length (cm)
19	9.75	68	9.00
40	9.00	69	7.80
42	9.60	69	10.05
42	9.75	70	10.50
47	11.25	71	9.15
49	9.45	71	9.45
50	11.25	71	9.45
54	9.00	72	9.45
56	7.95	73	8.10
56	12.00	74	8.85
57	8.10	74	9.60
57	10.20	75	6.45
58	8.55	75	9.75
61	7.20	75	10.20
62	7.95	76	6.00
62	8.85	77	8.85
65	8.25	80	9.00
65	8.85	82	9.75
65	9.75	82	10.65
66	8.85	82	13.20
66	9.15	83	7.95
66	10.20	86	7.95
67	9.15	88	9.15
68	7.95	88	9.75
68	8.85	94	9.00

W. In *The Chicago Maroon* for Friday, November 10, 1972, The Party Mart advertised per bottle prices for vintage port as given in the accompanying table.

1. Plot the data and examine them. Would it be sensible to fit a regression of the response "price" on the predictor "year"? What disadvantages can you see?
2. What transformation of the predictor "year" would be sensible? (*Hint:* Pretend it is 1972, and think, for example, how you describe your own "year," typically.) Plot price versus your new predictor, and examine the plot. What type of transformation on price would seem sensible here to make the data look "more straight-line-ish"?
3. Plot the data Y = ln(price) versus Z = age of bottle. Fit a straight line through the data by least squares, evaluate the residuals, and produce the analysis of variance table.
4. What do you conclude about the price of vintage port as exhibited by this set of data and your analysis? To the nearest cent, at what per-year rate would you expect the price of a bottle of vintage port to rise *if* a similar price pattern continued into the future?
5. A subsequent advertisement three years later on Tuesday, November 25, 1975, offered 1937 vintage port at $20.00 per bottle. If it can be assumed that a straight line relationship is preserved, and applies also to this new data point, how much per bottle per year does it appear prices have risen in the intervening three years? Are your answers here and in (4) consistent, or does it appear that per year price rises have accelerated?

Year	Price($)	Year	Price ($)
1890	50.00	1941	10.00
1900	35.00	1944	5.99
1920	25.00	1948	8.98
1931	11.98	1950	6.98
1934	15.00	1952	4.99
1935	13.00	1955	5.98
1940	6.98	1960	4.98

Source: Data via Steve Stigler, 1976.

X. (*Source:* "Graphs in statistical analysis," by F. J. Anscombe, The *American Statistician,* **27,** 1973, 17–21.) Fit a straight line model $Y = \beta_0 + \beta_1 X + \epsilon$ to each of the four sets of data below and show that *for each set,* $n = 11$, $\overline{X} = 9$, $\overline{Y} = 7.5$, $\hat{Y} = 3 + 0.5X$, $S_{XX} = 110$, Regression SS $= S_{XY}^2/S_{XX} = 27.5$ (1 df), Residual SS $= S_{YY} - S_{XY}^2/S_{XX} = 13.75$ (9 df), se$(b_1) = 0.118$, $R^2 = 0.667$.

Plot all four sets of data and explain how the sets of data differ and what their main characteristics are.

(Note that data sets 1–3 all have the same X values but different Y's.)

Data Set No.:	1–3	1	2	3	4	4
Variable:	X	Y	Y	Y	X	Y
Obs. No.: 1	10	8.04	9.14	7.46	8	6.58
2	8	6.95	8.14	6.77	8	5.76
3	13	7.58	8.74	12.74	8	7.71
4	9	8.81	8.77	7.11	8	8.84
5	11	8.33	9.26	7.81	8	8.47
6	14	9.96	8.10	8.84	8	7.04
7	6	7.24	6.13	6.08	8	5.25
8	4	4.26	3.10	5.39	8	5.56
9	12	10.84	9.13	8.15	8	7.91
10	7	4.82	7.26	6.42	8	6.89
11	5	5.68	4.74	5.73	19	12.50

Y. Suppose you were asked to do a simulation as follows:

1. Choose a "true" straight line $\eta = \beta_0 + \beta_1 X$.
2. Select n values of X, $X_1, X_2, \ldots, X_n$.
3. Generate n random errors ϵ_i, $i = 1, 2, \ldots, n$, from an $N(0, \sigma^2)$ and so obtain n "observations," $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$.
4. Fit a straight line $\hat{Y} = b_0 + b_1 X$ to the "observations."
5. Obtain a set of residuals $e_i = Y_i - \hat{Y}_i$.
6. Repeat steps 3–5 a total of N times, using the following (for example) parameter values: $n = 11$, $X_i = -1 + (i - 1)0.2$, $\sigma^2 = 1$, $N = 1000$.

Question: Why have no values been specified for β_0 and β_1?
(*Hint:* Show that the residuals e_i do not depend on β_0 and β_1. Thus any values can be used; the simplest choice is $\beta_0 = \beta_1 = 0$.)

Z. The data in Table Z, published by the *Chicago Tribune* on December 7, 1993, show consumption of candy (Y) and population (X) for 17 countries in 1991.

1. Plot Y versus X.
2. Fit $Y = \beta_0 + \beta_1 X + \epsilon$ by least squares and plot the fitted line on your diagram.
3. Evaluate the residuals to two decimal places. Check that $\Sigma e_u = 0$, within rounding error.
4. Obtain an analysis of variance table.
5. Find out how much of the variation about the mean $\overline{Y}$ is explained by the fitted line.
6. There are no exact repeat runs. As an approximation, however, treat the Y values at $X = 3.5$ and 4.3 as repeats; also those at $X = 5.0$ and 5.1; and finally those at $X = 56.9$, 57.7, and 57.8. (One could argue about the first and second of these groupings, of course, and you might wish to do a second, alternative calculation of your choice, to see what difference it makes.) Use the pseudo repeat runs to test (approximately) for lack of fit and state your conclusion. If you find lack of fit, omit parts (7), (8), (9), and (12).
7. Is it appropriate to test for overall regression via the F-test? If so, do it.
8. Find standard errors for b_0 and b_1.
9. Find the formula for the standard error of $\hat{Y}$ and, assuming that s^2 is an appropriate estimate of σ^2, construct 95% confidence bands for the true mean value of Y. (Plot about half a dozen points over the X-range and join them up smoothly.)

TABLE Z. Candy Consumption for Selected Countries in 1991

Country	Consumption (in millions of pounds)	Population (in millions)
1 Australia	327.4	17.3
2 Austria	179.5	7.7
3 Belgium/Luxembourg	279.4	10.4
4 Denmark	139.1	5.1
5 Finland	92.5	5.0
6 France	926.7	56.9
7 Germany	2186.3	79.7
8 Ireland	96.8	3.5
9 Italy	523.9	57.8
10 Japan	935.9	124.0
11 Netherlands	444.2	15.1
12 Norway	119.7	4.3
13 Spain	300.7	39.0
14 Sweden	201.9	8.7
15 Switzerland[a]	194.7	6.9
16 United Kingdom	1592.9	57.7
17 United States[a]	5142.2	252.7

[a]Both Switzerland and the United States include sugar-free candy in their statistics. The U.S. consumption excludes chewing gum.

Source: International Statistics Committee; *Chicago Tribune.*

10. There is a hint in the table footnote that the Swiss and U.S. data might not be consistent with the rest. If they were not, how much effect would the inconsistencies have on the fitted regression, do you think? A lot? Not much?
11. Plot the residuals versus $\hat{Y}$. Does this plot seem to confirm, or deny, the basic assumption of constant error variance. If it does deny it, suggest what could be done.
12. Fit the model $X = \alpha_0 + \alpha_1 Y + \epsilon$ by least squares. Find the fitted line that passes through $(\overline{X}, \overline{Y})$ and has slope

$$c_1 = (b_1/a_1)^{1/2},$$

where b_1 is the least squares estimate of β_1 in the $Y = \beta_0 + \beta_1 X + \epsilon$ fit (2) and a_1 is the least squares estimate of α_1 in the model of this part. (This fitted line is the "geometric mean functional relationship"; see Section 3.4.) State exactly what this third line is in the form $\hat{Y} = c_0 + c_1 X$, and plot all three lines on a new diagram.

AA. The height of soap suds in the dishpan is of importance to soap manufacturers. An experiment was performed by varying the amount of soap and measuring the height of the suds in a standard dishpan after a given amount of agitation. The data are as follows:

Grams of Product, X	Suds Height, Y
4.0	33
4.5	42
5.0	45
5.5	51
6.0	53
6.5	61
7.0	62

Assume that a model of the form $Y = \beta_0 + \beta_1 X + \epsilon$ is reasonable.

Requirements

1. Determine the best-fitting equation.
2. Test it for statistical significance.
3. Calculate the residuals and see if there is any evidence suggesting that a more complicated model would be more suitable.

BB. In an experiment similar to that given in Exercise AA, the experimenter stated that the model $Y = \beta_0 + \beta_1 X + \epsilon$ was "a ridiculous model unless $\beta_0 = 0$, for anyone knows that if you don't put any soap in the dishpan there will be no suds." Thus he insists on using the model $Y = \beta_1 X + \epsilon$. His data are shown as follows:

Grams of Product, X	Suds Height, Y
3.5	24.4
4.0	32.1
4.5	37.1
5.0	40.4
5.5	43.3
6.0	51.4
6.5	61.9
7.0	66.1
7.5	77.2
8.0	79.2

Requirements

1. Accepting the experimenter's model, determine the best estimate of β_1.
2. Using this equation, estimate $\hat{Y}$ for each X.

3. Examine the residuals.
4. Draw conclusions and make recommendations to the experimenter.

CC. The following data indicate the relationship between the amount of β-erythroidine in an aqueous solution and the colorimeter reading of the turbidity:

X Concentration (mg/mL)	Y Colorimeter Reading
40	69
50	175
60	272
70	335
80	490
90	415
40	72
60	265
80	492
50	180

Requirements. Fit the equation $Y = \beta_0 + \beta_1 X + \epsilon$, obtain the residuals, examine them, and comment on the adequacy of the model.

DD. A synthetic fiber, which because of its hairlike appearance has been found suitable in the manufacture of wigs, must necessarily be preshrunk prior to manufacture. This is accomplished in two steps:

Step 1. The fiber is soaked in a dilute solution of chemical A, which is necessary to preserve the luster of the fiber during step 2.

Step 2. The fiber is baked in large ovens at a very high temperature for 1 hour.

It is suggested that the temperature at which the fiber is baked may influence the effectiveness of the preshrinking process. An experiment is performed in which the baking temperature T is varied for various batches of fiber. The finished fiber is then soaked in rainwater for a suitable length of time and put out in the sun to dry. The amount of further shrinkage Y (in percent) resulting from the rainwater test is recorded along with the value of T for each batch:

Batch No.	T	Y
1	280	2.1
2	250	3.0
3	300	3.2
4	320	1.4
5	310	2.6
6	280	3.9
7	320	1.3
8	300	3.4
9	320	2.8

1. Fit a regression line $\hat{Y} = b_0 + b_1 T$ to the data by least squares.
   (*Note:* Coding the variable T may simplify the calculations, but remember in the end to express the fitted equation in terms of the original variable T.)
2. Perform an analysis of variance and test:
   a. The lack of fit.
   b. The significance of the regression.
   What is the percentage variation explained by the regression equation?
3. What is the standard error of b_1? Give a 95% confidence interval for the true regression coefficient β_1.
4. Give the fitted value $\hat{Y}_i$ and the residual $Y_i - \hat{Y}_i$ corresponding to each run (batch).

5. For $T_0 = 315$, find an interval about the predicted value $\hat{Y}_0$ within which a single future observation Y will fall with probability 0.95.
6. Could we use the fitted equation to predict a value of Y at $T = 360$? Give reasons for your answer.

(See Exercise L in "Exercises for Chapters 5 and 6" for a continuation.)

EE. The tree data below, obtained by Tamra J. Burcar and used with her permission, consist of 23 observations of X = tree diameter in inches measured at breast height (DBH) and Y = height in feet.

X	Y	X	Y	X	Y
5.5	58	8.3	68	10.6	80
5.7	60	8.6	65	10.8	82
6.5	64	9.5	70	11.3	70
6.6	60	10.0	63	11.3	74
6.7	65	10.1	75	11.6	68
6.9	56	10.2	72	11.6	68
7.0	57	10.4	78	13.0	82
7.3	70	10.6	65		

1. Plot Y versus X.
2. Fit $Y = \beta_0 + \beta_1 X + \epsilon$ by least squares and plot the fitted line on your diagram.
3. Evaluate the residuals to two decimal places. Check that $\Sigma e_i = 0$, within rounding error.
4. Obtain the analysis of variance table.
5. Find out how much of the variation about the mean $\overline{Y}$ is explained by the fitted line.
6. Use the repeat runs to test for lack of fit and state your conclusions.
7. Is it appropriate to test for overall regression via the F-test? If so, do it.
8. Find standard errors for b_0 and b_1, using s_e^2 or s^2 as seems appropriate from part 6.
9. Find the formula for the standard error of $\hat{Y}$ and, assuming that s^2 is an appropriate estimate of σ^2 for this part, construct 95% confidence bands for the true mean value of Y. (Plot about half a dozen points over the X-range and join them up smoothly.)

In fact two more data points were obtained by Ms. Burcar but were deleted above. They were:

X	Y
5.8	42
18.0	88

The small tree was clearly stunted; the large tree may have been damaged by winds because it protruded above the tree line.

10. Show the new trees on your plot and comment briefly on what you see.

FF. Reconsider the data of Exercise EE. Fit the model $X = \alpha_0 + \alpha_1 Y + \epsilon$ by least squares. Find the fitted line that passes through $(\overline{X}, \overline{Y})$ and has slope

$$c_1 = (b_1/a_1)^{1/2},$$

where b_1 is the least squares estimate of β_1 in the $Y = \beta_0 + \beta_1 X + \epsilon$ fit of the previous exercise and a_1 is the least squares estimate of α_1. (This fitted line is the geometric mean functional relationship, as described in Section 3.4.) State exactly what this third line is in the form $\hat{Y} = c_0 + c_1 X$, and plot all three lines on a new diagram.

(The following comment is not connected specifically to the data of Exercises EE and FF in any way, but since it concerns data of the same type, we reproduce it here for the benefit of those taking similar data.

> Experience suggests certain cautions. First, there is a tendency to select vigorously growing trees of good form for the dimensional analysis, unless this tendency is consciously counteracted. The preference for "good" sample trees implies overestimation of productivity when regressions from these trees are applied to field quadrat data. Second, the largest errors result from applying regressions to the largest

trees in the samples.... If, from the population of large trees in the stand, many of them senescent or with partly broken crowns, a particularly "good" individual has been chosen, the slope of the regression as it extends to larger tree sizes is biased by this individual. The production estimates for the few large trees in the sample quadrat will be overestimates for most of these trees. It is therefore important that errors of estimation for large trees be controlled by some means...."

This excellent advice, which can be summarized by saying "Make sure your data are representative of your population," was written by R. H. Whittaker and P. L. Marks in "Methods of assessing terrestial productivity," Chapter 4, p. 85 of *Primary Productivity of the Biosphere*, edited by H. Lieth and R. H. Whittaker, and published by Springer-Verlag, New York, in 1975.)

GG. The *New York Times* of Thursday, May 24, 1990, showed, on page B5, a chart containing data obtained in the National Child Care Staffing Study released in Fall 1989 by the Child Care Employee Project in Oakland, California. These data are displayed below. Plot Y = turnover versus X = hourly wage. The article stated that one city had a high unemployment rate. Which data point do you think represents that city? Omit the Detroit data point and fit a straight line $Y = \beta_0 + \beta_1 X + \epsilon$ to the remaining four data points. Estimate the true mean turnover if the wage were \$6.00, and put 95% confidence bands around the estimate.

Place	Hourly Wage (in \$) of Staff Who Care for Children	Average Annual Staff Turnover (in %)
Atlanta	4.96	57
Boston	7.28	29
Detroit	4.88	27
Phoenix	4.45	64
Seattle	5.21	41

HH. (*Source:* R. Peter Hypher, Ottawa, Ontario, Canada.) Below are 29 observations on

X = Monthly Protestant receipts in Canadian dollars,

Y = Monthly attendance at Catholic mass,

for a certain Canadian town some years ago. (The figures, read off from a plot, vary slightly from the originals, which are not available.)

Plot the data, and fit the model $Y = \beta_0 + \beta_1 X + \epsilon$ by least squares. Assuming conditions remain the same, can we validly predict attendance at Catholic mass from the Protestant monthly receipts? Estimate the Catholic mass attendance when $X = 0$ and provide 95% confidence limits on the true mean value of Y at $X = 0$.

X	Y	X	Y	X	Y
100	2000	290	2875	390	2500
100	2475	290	2875	400	2875
100	2825	300	2900	400	3800
155	2000	300	3000	440	3825
160	1650	330	3100	440	3900
205	2475	350	2875	450	3100
230	2725	350	3225	460	3050
250	2125	350	3400	500	3475
250	3000	350	3450	550	3200
		360	2950	600	3675

II. The data below consist of

X = Course number,

Y = Questionnaire score on "question 11,"

for 32 teaching assistants in the Statistics Department, University of Wisconsin, in the Spring Semester of 1978. Plot the data, fit a straight line model $Y = \beta_0 + \beta_1 X + \epsilon$, perform all the usual analyses, and state the practical conclusions that arise from your analysis.

Row	Y	X	Row	Y	X	Row	Y	X
1	5.00	349	12	4.41	224	23	4.00	301
2	4.86	333	13	4.38	333	24	3.92	201
3	4.75	301	14	4.29	314	25	3.88	201
4	4.64	301	15	4.25	201	26	3.86	312
5	4.66	302	16	4.22	424	27	3.74	224
6	4.62	824	17	4.21	424	28	3.69	224
7	4.55	710	18	4.15	301	29	3.57	224
8	4.47	301	19	4.07	201	30	3.27	301
9	4.43	301	20	4.08	301	31	2.96	224
10	4.44	310	21	4.00	224	32	2.67	224
11	4.43	702	22	4.00	301			

JJ. (*Source: USA Today* for Monday, December 6, 1993, page 5B.) The data in the accompanying Table JJ consist of average Sunday circulation (Y) and average daily circulation (X) for 48 U.S. newspapers. All values are in thousands, rounded to the nearest thousand. Two of the "top 50" were omitted; the second and 20th were removed because they had no Sunday edition. The figures are for the 6 months ended September 30, 1993 and include bulk sales, defined as lower price sales to, for example, hotels and airlines, who give them free to customers.

1. Plot Y versus X.
2. Fit $Y = \beta_0 + \beta_1 X + \epsilon$ by least squares and plot the fitted line on your diagram.

TABLE JJ. Average Sunday (Y) and Daily (X) Circulations in Thousands for 48 of the Top 50 Newspapers in the United States for the Period March–September 1993

Newspaper	Sunday (Y)	Daily (X)	Newspaper	Sunday (Y)	Daily (X)
1	2313	1886	25	559	342
2	1766	1155	26	483	338
3	1497	1097	27	441	333
4	1140	816	28	428	324
5	928	764	29	214	322
6	826	748	30	378	302
7	1107	694	31	431	291
8	1187	558	32	440	285
9	708	549	33	321	284
10	524	536	34	342	282
11	814	527	35	338	270
12	814	508	36	401	264
13	944	487	37	320	264
14	704	474	38	375	263
15	709	442	39	346	255
16	607	413	40	347	252
17	696	411	41	456	250
18	543	396	42	340	246
19	509	387	43	326	236
20	455	384	44	301	233
21	1186	367	45	496	232
22	548	348	46	411	232
23	408	344	47	321	230
24	453	343	48	504	229

Source: USA Today, Monday, December 6, 1993, p. 5B.

3. Evaluate the residuals to one decimal place. Check that $\Sigma e_u = 0$, within rounding error.
4. Obtain an analysis of variance table.
5. Find out how much of the variation about the mean $\overline{Y}$ is explained by the fitted line.
6. There are only two pairs of exact repeat runs in the rounded data quoted. As an approximation, however, treat the Y values at X = 229, 230, 232, 232, 233, and 236 as repeats; also, those at X = 246, 250, 252, and 255; those at X = 263, 264, and 264; those at X = 282, 284, and 285; those at X = 333 and 338; those at X = 342, 343, 344, and 348; and finally those at X = 411 and 413. (One could argue about these groupings, of course, and you might wish to do a second, alternative calculation using only the two pairs, to see what difference it makes.) Use the pseudo repeat runs to test (approximately) for lack of fit and state your conclusion. If lack of fit is shown, omit parts (7), (8), and (9).
7. Is it appropriate to test for overall regression via the F-test? If so, do it.
8. Find standard errors for b_0 and b_1.
9. Find the formula for the standard error of $\hat{Y}$ and, assuming that s^2 is an appropriate estimate of σ^2 for this part, construct 95% confidence bands for the true mean value of Y. (Plot about half a dozen points over the X-range and join them up smoothly.)
10. Plot the residuals versus $\hat{Y}$. Does this plot seem to confirm, or deny, the basic assumption of constant error variance. If it does deny it, suggest what could be done.
11. Below are the average daily bulk sales for the first nine newspapers in the data list. Adjust the data for these numbers (subtract them) and repeat the analysis. State your conclusions overall.

Newspaper	Sunday (Y)	Daily X
1	408	391
2	10	14
3	9	7
4	1	2
5	1	0
6	1	1
7	5	3
8	1	2
9	6	5

KK. (*Source:* Audit Bureau of Circulations, as given in the *New York Times,* October 31, 1995, page C7.) The average daily circulations of the largest 12 U.S. newspapers (apart from the *Detroit Free Press,* whose union workers were on strike in 1995) for the 6 months ending September 30 are listed in the table for 1994 (X) and 1995 (Y). Circulation figures for papers 2 and 10 do not include Fridays. Fit a straight line model $Y = \beta_0 + \beta_1 X + \epsilon$, evaluate the residuals, check the fit, and amend it as appropriate. State conclusions.

Paper	1994	1995
1. *Wall Street Journal*	1,780,422	1,763,140
2. *USA Today*	1,465,936	1,523,610
3. *New York Times*	1,114,168	1,081,541
4. *Los Angeles Times*	1,062,199	1,012,189
5. *Washington Post*	810,675	793,660
6. *Daily News*	753,024	738,091
7. *Chicago Tribune*	678,081	684,366
8. *Newsday*	693,556	634,627
9. *Houston Chronicle*	409,007	541,478
10. *Dallas Morning News*	491,480	500,358
11. *Boston Globe*	506,543	498,853
12. *San Francisco Chronicle*	509,548	489,238

LL. The manager of a small mail order house hires additional personnel every time there is a peak demand that exceeds the work load of his normal three employees. To check the effectiveness of this idea, he records the daily output of his total crew on various days during various periods, both peak and otherwise. These data are given below. Fit a straight line model $Y = \beta_0 + \beta_1 X + \epsilon$ to the data via least squares, check for lack of fit, and (if there is no lack of fit) test for overall regression. Examine the residuals, in any case, and state the conclusions you draw from the study.

Number of Parcels Dispatched Y	Number of Employees X	Number of men Z^a
50	1[b]	
110	2[b]	
90	2[b]	
150	3	
140	3	
180	3	
190	4	
310	6	
330	6	
340	7	
360	8	
380	10	
360	10	

[a]Needed for Exercise H in "Exercises for Chapters 5 and 6." Ignore for now.
[b]Regular employee(s) sick or on vacation.

Useful Facts. $n = 13$

$$\Sigma X_i = 65, \quad \Sigma Y_i = 2990, \quad \Sigma X_i Y_i = 19{,}120,$$

$$\Sigma X_i^2 = 437, \quad \Sigma Y_i^2 = 857{,}500, \quad \Sigma Y_i^2 - (\Sigma Y_i)^2/n = 169{,}800.$$

MM. (Source: Steven LeMire.) An important factor in the performance of an air filter is its resistance. Measuring this accurately on a large section of filter material is expensive. An inexpensive alternative device was suggested and was used to obtain the data X below. The Y values were obtained from the more expensive method. (All values have been multiplied by 1000.) Fit straight line models $Y = \beta_0 + \beta_1 X + \epsilon$ and $X = \alpha_0 + \alpha_1 Y + \epsilon$ to these data, and plot each set of residuals versus its corresponding set of fitted values, $\hat{Y}$ or $\hat{X}$, respectively. Also fit the geometric mean functional relationship line which has slope $(b/a)^{1/2}$ and which passes through $(\overline{X}, \overline{Y}) = (96.25, 89.1875)$, where b and a are the estimates of β_1 and α_1. Plot the two sets of residuals, horizontal and vertical, from this line versus their corresponding fitted values. Does each set of residuals sum to zero? Now review all the results and comment.

X	Y	X	Y
0	0	79	72
22	22	93	87
26	25	133	118
41	40	122	120
62	59	142	125
67	64	165	150
70	67	190	182
78	70	250	226