# CHAPTER 21

## REGRESSION AND LINEAR MODELS

Fitting a theoretical curve to a set of data points is one of the most common statistical problems faced by scientists, engineers, and economists. This field is very large, because there is no one solution that applies to all cases. Instead, we have a number of quite different problems, depending on just what prior information we had about the phenomenon being observed, the measurement errors, and the unknown parameters.

At the end of Chapter 8 we noted briefly some problems that orthodox theory encounters here because of the difficulty in distinguishing between "random" and "nonrandom" quantities. Another difficulty is even more troublesome in practice.

### Unwanted Parameters

That differences in the prior information can generate qualitatively different mathematical problems has, of course, been well recognized in the voluminous orthodox literature. Some 'sharp and drastic' differences can be expressed adequately by different choices of a model (for example the judgment that a certain parameter should or should not be present at all). But some more 'gentle' differences in the prior information can be expressed precisely only by differences in the corresponding prior probabilities within a model. Orthodox theory, which does not admit the existence of the needed prior probabilities, is helpless to take such information into account, although it may be fully as cogent as the data.

This is not merely a philosophical problem; it leads to a serious technical problem, of "nuisance parameters," *i.e.*, parameters which are physically present in the phenomenon and so cannot be safely disregarded in the model, although we are not interested in estimating them. But once in the model they cannot be eliminated by orthodox principles, and one is obliged to estimate them along with the interesting parameters.

In Bayesian methods, nuisance parameters cause very little trouble – any uninteresting parameters are removed by integratiing out with respect to their prior probabilities. But this gives rise to another technical question whose answer will be important for future extensions of Bayesian theory to more and more complex problems. When parameters are integrated out, what effect does this have on the accuracy of our estimates of the remaining ones?

In many cases, the presence of an unwanted and unknown parameter that has to be integrated out, will cause a deterioration of our ability to estimate another parameter. Thus, consider estimation of the mean $\mu$ of a normal distribution form the sample data $D \equiv \{x_1, x_2, \ldots, x_n\}$. If $\sigma^2$ is known, the posterior distribution $p(du|\sigma, D)$ is still a normal distribution, leading to the 90% interval estimate (i.e., the shortest interval that contains 90% of the posterior probability):

$$(\mu)_{\text{est}} = \bar{x} \pm 1.645 \frac{\sigma}{\sqrt{n}}.$$

But if $\sigma$ is completely unknown and must be integrated out with respect to a Jeffrey's prior $d\sigma/\sigma$, we are, in effect, estimating $\sigma^2$ by the sample variance $s^2 = \overline{x^2} - \bar{x}^2$. But this estimate is uncertain, and the integration over $\sigma$ averages the normal distribution $p(d\mu|\sigma, D)$ over this uncertainty. It then becomes a $t$-distribution, with density function $\propto [s^2 + (\mu - \bar{x})^2]^{-n/2}$; and the 90% interval estimate becomes

$$(\mu)_{\rm est} = \bar{x} \pm t_n \frac{s}{\sqrt{n-1}}$$

where $t_n$ is the upper critical value of the $t$-statistic at the 95% level for $f = (n-1)$ degrees of freedom. From the $t$-tables we find that $\{t_2, t_3, t_4, t_{10}\} = \{6.3, 2.92, 2.35, 1.83\}$ respectively; but as $n \to \infty$, $t_n \to 1.645$. Thus for small samples the penalty for failure to know $\sigma$ is not a change in the actual point estimate, but an appreciable loss of accuracy which we may claim in our estimate of $\mu$. With large samples $\sigma$ is determined by the data more and more accurately, and so we approach the accuracy with $\sigma$ known.

Now suppose there were many parameters $\{\sigma_1, \sigma_2 \ldots \sigma_k\}$ that all had to be integrated out. If each had a comparable effect, then if $k = n$, no useful estimates would be possible at all. There seems to be a general belief – presumably for this reason -that models with large numbers of parameters are, *ipso facto,* intractable, any useful inference requiring that the number of observations be large compared to te number of parameters. Thus various authors [such as Kempthorne and Folks (1971 p. 425)] repeat the folk-theorem that no inference is possible if the number of parameters is greater than the number of different "statistics" that appear in the sampling distribution.

On the other hand, Lindley (1971) notes a problem of the type we study here, which provides a counter-example to the folk-theorem, the presence of many unwanted parameters doing no appreciable harm. It will be important for us to understand the exact conditions for this good behavior.

## Linear Models–A First Look

There are pairs of "true" values $(X_i, Y_i)$ and the corresponding measured values $(x_i, y_i)$,

$$\begin{aligned} x_i &= X_i + e_i, \qquad i = 1, 2, \ldots n \\ y_i &= Y_i + f_i \end{aligned} \tag{21–1}$$

where the errors $e_i$, $f_i$ are supposed independent and $N(0, \sigma_x)$, $N(0, \sigma_y)$ respectively; $\sigma_x$ and $\sigma_y$ may be known, but usually are not. The probability, given $\{\sigma_x, \sigma_y, X_1 \ldots X_n, Y_1 \ldots Y_n\}$ that we shall see the data $D \equiv \{(x_1, y_1), \ldots, (x_n, y_n)\}$, within tolerances $dx \equiv dx_1 \ldots dx_n$, $dy \equiv dy_1 \ldots dy_n$, is

$$p(dx\,dy|\sigma_x \sigma_y XY) = (2\pi\sigma_x\sigma_y)^{-n} \exp\left(-\frac{1}{2}R\right)\, dx\,dy \tag{21–2}$$

where

$$R \equiv \sum_{i=1}^{n} \left[\frac{(x_i - X_i)^2}{\sigma_x^2} + \frac{(y_i - Y_i)^2}{\sigma_y^2}\right] \tag{21–3}$$

and we could integrate out either $dx$ or $dy$ to obtain the marginal distribution; i.e.,

$$p(dy|\sigma_y, Y) = \left(\frac{1}{2\pi\sigma_y^2}\right)^{\frac{n}{2}} \exp\left\{-\sum_i \frac{(y_i - Y_i)^2}{2\sigma_y^2}\right\} dy. \tag{21–4}$$

At this stage, we have two independent problems, of inference about $X_i$, $Y_i$ separately. But now the problems are tied together by a "model"; i.e., a postulated functional relationship between $X$ and $Y$:

$$f(X, Y, \theta_1, \theta_2, \ldots) = 0 \tag{21–5}$$

This model equation contains certain parameters $\theta_k$; and the problem then becomes: to estimate the $\theta_k$. The common experimentalist's problem of fitting a line to a set of data points, corresponds to choosing the model equation

$$Y_i = \alpha + \beta X_i \tag{21–6}$$

and using the data to estimate $(\alpha, \beta)$.

Note that in the older literature the word "Linear" in "Linear Models" is usually taken to mean that the model equation is linear in the parameters and in the errors, but not necessarily in the measured variables. Thus $Y_i^2 = \alpha + \beta \cos X_i$ would be termed a "linear model" if $e_i$, $f_i$ are small enough so that we can write $\cos X_i = cos x_i + e_i \sin x_i$, etc; but $Y_i = \alpha^2 + \beta x_i$ would not [see, for example, Graybill (1961); p. 97]. This terminology was unfortunate, because it is hard to invent any model equation (21—5) that cannot, merely by a redefinition of $\{\theta_i, X, Y\}$, be made linear in the $\theta_i$. Thus the term "linear" was almost meaningless as far as the real content of the theory was concerned – it really meant only "small errors."

The uninitiated were falling constantly into the trap of supposing that "linear" refers to the fact that (23–6) is the equation of a straight line (and the term would be more appropriate and useful if it did!). In 1985, M. DeGroot made a break with this terminology and redefined the term "linear model" to mean straight-line fitting. We shall follow this reform in terminology.

**Case 1. $\sigma_x \equiv 0$; $\sigma_y$ known**

The simplest case is that in which the error is all in the $Y$ measurements (i.e., $x_i \equiv X_i$), and $\sigma_y$ is known. The terms in $\sigma_x$ are then absent, and the sampling distribution (21–4) is appropriate. Using (21–6), it reduces to

$$p(dy|\sigma_y, \alpha, \beta, X) = \left(\frac{1}{2\pi\sigma_y^2}\right)^{\frac{n}{2}} \exp\left[-\frac{n}{2\sigma_y^2} Q(\alpha, \beta)\right] dy \tag{21–7}$$

where

$$Q(\alpha, \beta) \equiv \frac{1}{n} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{21–8}$$

is a positive definite quadratic form in $\alpha$, $\beta$ that proves to be fundamental in several problems below. We digress to consider the many ways of writing this out in detail.

**Forms of $Q(\alpha, \beta)$**

For various purposes, several different forms of $Q(\alpha, \beta)$ are convenient. Writing out (21–8) in full, we get six terms:

$$Q(\alpha, \beta) = \overline{y^2} + \alpha^2 + \beta^2 \overline{x^2} - 2\alpha\overline{y} + 2\alpha\beta\overline{x} - 2\beta\overline{xy} \tag{21–9}$$

where the sample first moments

$$\overline{x} \equiv \frac{1}{n} \sum x_i, \qquad \overline{y} \equiv \frac{1}{n} \sum y_i \tag{21–10}$$

and second moments

$$\overline{x^2} \equiv \frac{1}{n}\sum x_i^2, \quad \overline{y^2} \equiv \frac{1}{n}\sum y_i^2, \quad \overline{xy} \equiv \frac{1}{n}\sum x_i y_i \tag{21–11}$$

are, of course, known from the data. Often, we are interested primarily in $\beta$, not in $\alpha$ [for example, a chemist may want to know how a reaction rate varies with temperature, a meteorologist may wish to determine if there is evidence for a slow warming trend over the past decade; or an economist may want to know how the demand for gasoline or steel varies with its price]. We will then want to integrate $\alpha$ out of the problem. In preparation for this we must complete the square first on $\alpha$:

$$Q(\alpha,\beta) = (\alpha-\alpha')^2 + s_x^2(\beta-\beta^*)^2 + s_y^2(1-r^2). \tag{21–12}$$

Here we have introduced the notation

$$\alpha' \equiv \overline{y} - \beta\overline{x}, \qquad \beta^* \equiv \frac{s_{xy}}{s_x^2} = \frac{s_y}{s_x}r \tag{21–sam}$$

where the sample variances and covariance

$$s_x^2 \equiv \overline{x^2} - \overline{x}^2, \qquad s_y^2 \equiv \overline{y^2} - \overline{y}^2, \qquad s_{xy} \equiv \overline{xy} - \overline{xy} \tag{21–15}$$

and the sample correlation coefficient

$$r \equiv \frac{s_{xy}}{s_x s_y} \tag{21–16}$$

are, of course, also known from the data. As is apparent already from (21–12), $\beta^*$ is going to emerge as a "natural" estimator for $\beta$.

On the other hand, we might be interested primarily in $\alpha$ rather than $\beta$. [For example, a physical chemist measuring ionic conductivity has to make measurements at finite concentrations $(= x)$; but it is the extrapolation to infinite dilution $(x = 0)$ that is the fundamental quantity to be compared with theory. Or, a spectroscopist may wish to determine atomic energy levels by extrapolation the measurable Zeeman levels back to zero magnetic field, as in the famous Lamb shift experiment.] In this case, we will want to integrate $\beta$ out of the problem; completing the square first on $\beta$, we get

$$Q(\alpha,\beta) = \overline{x^2}(\beta-\beta')^2 + \frac{s_x^2}{\overline{x^2}}(\alpha-\alpha^*)2 + s_y^2(1-r^2) \tag{21–17}$$

where

$$\beta' \equiv \frac{\overline{xy} - \alpha\overline{x}}{\overline{x^2}} \tag{21–18}$$

and

$$\alpha^* \equiv \overline{y} - \beta^*\overline{x} \tag{21–19}$$

is a "natural" estimator of $\alpha$. Even at this stage, we can see that to make the estimates $(\alpha^*, \beta^*)$, means that we would take the line passing through the data centroid $(\overline{x}, \overline{y})$ with slope $\beta^*$, as our estimate of the "true" line.

Finally, we may be interested in both $\alpha$ and $beta$, or in some function $f(\alpha,\beta)$ that involves both; and we wish to get their joint posterior *pdf* $p(d\alpha d\beta|D)$ in a form that treats them symmetrically. For this we introduce the coefficients $C_{ij}$ of the quadratic form:

$$C(\alpha,\beta) = c_{11}(\alpha-\alpha^*)^2 + 2C_{12}(\alpha-\alpha^*)(\beta-\beta^*) + C_{22}(\beta-\beta^*)^2. \tag{21–20}$$

Comparison with (21–8) shows that if we choose the matrix elements $C_{ij}$ to be

$$C_{ij} = \begin{pmatrix} 1 & \overline{x} \\ \overline{x} & x^2 \end{pmatrix} \tag{21–21}$$

we have

$$Q(\alpha,\beta) = C(\alpha,\beta) + s_y^2(1-r^2). \tag{21–22}$$

Now, from this and (21–2), (21–7) we see that (since only the dependence on $\alpha$ and $\beta$ matters; i.e., any factors independent of $\alpha$ and $\beta$ are going to be absorbed into normalizing constants anyway) the joint likelihood of $(\alpha,\beta)$ may be taken simply as

$$L(\alpha,\beta) = \exp\left\{-\frac{n}{2\sigma^2}C(\alpha,\beta)\right\} \tag{21–23}$$

and so with uniform priors, their joint posterior distribution is the bivariate normal based on the matrix $C$ and peaked at $(\alpha^*,\beta^*)$. Writing $p(d\alpha d\beta|D) = F(\alpha,\beta)d\alpha d\beta$, this joint posterior density is

$$F(\alpha,\beta) = A\exp\left\{-\frac{n}{2\sigma^2}C(\alpha,\beta)\right\} \tag{21–24}$$

with

$$C(\alpha,\beta) \equiv (\alpha-\alpha^*)^2 + 2\overline{x}(\alpha-\alpha^*)(\beta-\beta^*) + \overline{x^2}(\beta-\beta^*)^2 \tag{21–25}$$

and the normalizing constant is

$$A = \frac{n}{2\pi\sigma_y^2}\,|\det(C)|^{\frac{1}{2}} = \frac{ns_x}{2\pi\sigma_y^2}. \tag{21–26}$$

the second central moment of (21–24) are given by the inverse matrix to $C$:

$$D = C^{-1} = \frac{1}{s_x^2}\begin{pmatrix} \overline{x^2} & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}. \tag{21–27}$$

Thus

$$\langle(\alpha-\alpha^*)^2\rangle = \frac{\sigma_y^2}{n}D_{11} = \frac{\sigma_y^2}{n}\frac{\overline{x^2}}{s_x^2} \tag{21–28}$$

as may also be read off by inspection of (21–17); and

$$\langle(\beta-\beta^*)^2\rangle = \frac{\sigma_y^2}{n}D_{22} = \frac{\sigma_y^2}{n}\frac{1}{s_x^2} \tag{21–29}$$

as is evident from (23–12). The covariance is

$$\langle(\alpha-\alpha^*)(\beta-\beta^*)\rangle = \frac{\sigma_y^2}{n} D_{12} = -\frac{\sigma_y^2\,\overline{x}}{n s_x^2} \tag{21–30}$$

leading to the correlation coefficient

$$\rho = \frac{\langle(\alpha-\alpha^*)(\beta-\beta^*)\rangle}{[\langle(\alpha-\alpha^*)^2\rangle\langle(\beta-\beta^*)^2\rangle]^{\frac{1}{2}}} = -\frac{\overline{x}}{\left(\overline{x^2}\right)^{\frac{1}{2}}}. \tag{21–31}$$

It is interesting that (21–28) – (21–31) involve only the $x$–measurements which are with error. Now if $\sigma_y$ is known and the $x$–measurements are without error, then often one can decide in advance how many measurements to make, and as what values of $x_i$; whereupon we know just what accuracy our $\alpha$ and $\beta$ estimates will have. The entire shape and width of the posterior (21–24) can be known in advance of the experiment, only the location of the peak $(\alpha^*, \beta^*)$ awaiting the actual data.

Of course, this is a rather artificial and oversimplified example; it is not often that one knows $\sigma_y$ in advance. For most measurements it would be more realistic to go to the opposite extreme, and suppose $\sigma_y$ entirely undetermined by the prior information, whereupon it must be estimated from the consistency of the data (i.e., if all the data points lie very accurately on a straight line, our common sense tells us that $\sigma_y$ must have been very small, *etc*).

**Case 2. $\sigma_x \equiv 0$, $\sigma_y$ Unknown**

To express "complete initial ignorance" of $\sigma_y$ we must, as noted in Chapter 12, use the Jeffreys prior

$$p(d\sigma|X) = \frac{d\sigma_y}{\sigma_y} \tag{21–32}$$

and the dependence of the likelihood on $\sigma_y$ must be retained; thus we cannot use (21–23), but must go back to the sampling distribution (21–7) which, in its dependence on $\{\alpha, \beta, \sigma_y\}$ gives their joint likelihood:

$$L(\alpha,\beta,\sigma_x) = \sigma_y^{-n} \exp\left[-\frac{n}{2\sigma_y^2} Q(\alpha,\beta)\right]. \tag{21–33}$$

With uniform priors for $\alpha$ and $\beta$, their joint posterior *pdf* has the form

$$p(d\alpha d\beta d\sigma_y|D) = A' \frac{d\alpha d\beta d\sigma_y}{\sigma_y^{n+1}} \exp\left[-\frac{n}{2\sigma_y^2} Q(\alpha,\beta)\right] \tag{21–34}$$

and if we care only about $\alpha$, $\beta$, we integrate out $\sigma_y$ to obtain

$$p(d\alpha d\beta|D) = A Q^{-n/2} d\alpha d\beta. \tag{21–36}$$

The two normalizing constants being related by

$$A = \left(\frac{2}{n}\right)^{n/2} \Gamma\left(\frac{n}{2}\right) A'. \tag{21–36}$$

We thus have the bivariate $t$-distribution (21—35), instead of the bivariate normal distribution (21–24), as the price we incur for not knowing $\sigma_y$. The distributions are qualitatively similar, the $t$-distribution having wider tails which, for small $n$, represent a significant deterioration in the accuracy of our estimates.

************************** MUCH MORE TO COME! ******************