SECTION V FITTING CURVES TO OBSERVATIONS

5.1 General

Let $x = g(y)$ be the "structural" relation between two random variables, X denoting the dependent, and Y the independent variable. The general fitting problem can be said to consist in a determination of the parameters of the function g, in such a way, that the deviations Δ_i of the data points (x_i, y_i) from the curve representing the function g be as small as possible. Since the deviation Δ is a random variable itself, the "smallness" of the deviations is subject to some arbitrarily chosen criterion defining the goodness of fit.

The general problem will here be limited by the assumption that Y is a non-random variable, that is, it can be measured with any degree of accuracy. To each value y_i there exists a corresponding random variable X defined by its cdf., as usually, put in the form $F[(x-\mu)/\beta]$.

Consider now the case that the function g contains two parameters a,b, and that observed values of X belong to two groups only, one x_{1i} corresponding to y_1, the other x_{2i} corresponding to y_2.

If the two numbers of observations are sufficiently large, it would be possible to determine the distribution parameters μ,β, and possibly α for each group. The curve g can then be fitted by a proper choice of a and b to pass through an arbitrary point of the two distributions, be it their modes, means, medians or any other percentage points for example, the lower bound of the distributions, if existent.

It appears to be a sound requirement that the two points chosen should correspond to the same percentage. This condition is satisfied by anyone of the mentioned points provided identical cdf. at y_1 and y_2, otherwise the mean and the mode may correspond to different percentages.

Consider now the other extreme, that there are only one single observation x_i corresponding to each of a number of levels y_i, the number being larger than the number of parameters in the function g.

Clearly, it is impossible to determine the cdf. of X corresponding to each y_i, and the distribution has to be known or assumed, for example, by assuming it to be normal or, if containing a parameter α, knowing the relation between α and y. In addition, the relation between β and y is also needed. If known, the parameter β may be eliminated by substituting the "normalized" variable $x' = x/\beta$ for x.

In this way, there only remains as unknown the location parameter μ and the fitting problem could be formulated as the problem of determining the function $\mu = g(y)$ subject to some criterion of best fit, where μ may

represent an arbitrary location statistic of the distributions.

For the further discussions the simplifying assumption will be made that the distribution function is independent of y and the parameter β is known.

A general fitting procedure, including as particular cases, the methods of least squares and linear estimators, can be visualized by means of the mechanical model demonstrated in (4.2.4).

Let $\hat{x}$ denote a point on the fitted curve

$$\hat{x} = g(y; a,b) \tag{173}$$

(the notation indicates that any point of this curve is an estimated point).

This curve is assumed to be attracted by forces H_i emanating from the points x_i and defined by

$$H_i = -\frac{1}{\beta} f'\left(\frac{\Delta_i}{\beta} + \tilde{z}\right) \Big/ f\left(\frac{\Delta_i}{\beta} + \tilde{z}\right) \tag{174}$$

where

$$\Delta_i = x_i - \hat{x} = x_i - g(y_i; a,b) \tag{175}$$

Under the influence of these forces the curve g takes a stable equilibrium which implies that the resultant force and the resultant moment are equal to zero, that is,

$$\Sigma H_i = 0 \qquad \text{and} \qquad \Sigma y_i \cdot H_i = 0 \tag{176}$$

The only unknown quantities in (176) are the parameters a and b which can be determined by a cut-and-try method without prior rectification of the curve g.

Since for normal distributions $H_i = (x_i - \hat{x})/\sigma^2$, it follows that the proposed procedure in this particular case reduces to the least-squares method.

Some applications will be demonstrated for illustrative purpose.

5.2 Maximum Likelihood Conditions

Let the fr.f. be a known function and β an unknown constant, both of them independent of y, while

$$\mu = g(y; a,b) \tag{177}$$

The likelihood L is then according to (118)

$$L = -n \log\beta + \Sigma \log f\left[(x_i - g(y_i; a,b))/\beta\right] \tag{178}$$

The parameters β, a, and b are determined by the conditions, $\partial L/\partial\beta = \partial L/\partial a = \partial L/\partial b = 0$.

These general formulas will be applied to the fr.f.

$$\frac{m}{\beta} \cdot z^{m-1} \cdot e^{-z^m}$$

for the particular case that

$$g(y) = a + by \tag{179}$$

In analogy with (123) we then have

$$\beta^m = \frac{\Sigma(x_i - a - b y_i)^m}{n} = \frac{m}{m-1} \cdot \frac{\Sigma(x_i - a - b y_i)^{m-1}}{\Sigma(x_i - a - b y_i)^{-1}} = \frac{m}{m-1} \frac{\Sigma y_i (x_i - a - b y_i)^{m-1}}{\Sigma y_i (x_i - a - b y_i)^{-1}} \tag{180}$$

From this system a and b can be computed by a cut-and-try method.

A modified method appears to be practicable on the condition that a number of observations are available for discrete y-values. For each group the mode can be estimated by (123) and a regression line can be fitted to the modes by means of the least-squares method. This is a quite legitimate procedure, because small disturbances from the equilibrium points are resisted by forces proportional to the deviation and this is the requisite for the applicability of the least-squares method. Each point should be weighted in proportion to the number of points it represents.

5.3 Linear Regression Formulas

In the case of normal distributions, the force H is proportional to the deviation from the equilibrium point, and the condition $\Sigma H_i = 0$ is identical with minimizing the sum of the squared deviations which is exactly the condition of the least-squares method.

It can thus be concluded that this method is efficient on the condition that the deviations from the fitted curve are normally distributed.

In other cases loss of efficiency will result (Cf.4.4).

Let

$$\hat{X} - X_e = b(Y - Y_e) \tag{181}$$

be the straight line, called the regression line, which is to be fitted to the observations. (In this case, capital letters are used because small letters will be used for denoting the deviations from the mean, e.g., $x = X - \bar{X}$).

The criterion of best fit is now

$$2M = \sum_{i=1}^{n} (X_i - \hat{X})^2 / \sigma_i^2 = \text{Minimum} \tag{182}$$

where σ_i^2 is the variance of X_i.

Let, for simplicity, σ_i^2 be equal for all values and introduce (181) into (182). The condition may then be formulated

$$\Sigma \left[(X_i - X_e) - b(Y_i - Y_e) \right]^2 = \text{Minimum} \tag{183}$$

As a first alternative let (X_e, Y_e) be a given point and b the only parameter to dispose of.

The condition $\partial M / \partial b = 0$ is satisfied by

$$\hat{b} = \Sigma (X_i - X_e)(Y_i - Y_e) / \Sigma (Y_i - Y_e)^2 \tag{184}$$

For the particular case $X_e = Y_e = 0$
we have

$$\hat{b} = \Sigma X_i Y_i / \Sigma Y_i^2 \tag{185}$$

The second alternative arises, if we can dispose freely also of (X_e, Y_e). The condition $\partial M / \partial X_e = 0$ gives

$$\Sigma \left[(X_i - X_e) - b(Y_i - Y_e) \right] = 0$$

This condition is satisfied for any value of b if but only if

$$X_e = \bar{X} = \Sigma X_i / n \quad \text{and} \quad Y_e = \bar{Y} = \Sigma Y / n \tag{186}$$

With the notations

$$x_i = X_i - \bar{X} \quad \text{and} \quad y_i = Y_i - \bar{Y} \tag{187}$$

it follows that

$$\hat{b} = \Sigma x_i y_i / \Sigma y_i^2 \tag{188}$$

which is a linear estimate in the observed values x_i.

The variance of the estimate $\hat{b}$ is, as will be deduced in the next paragraph,

$$\mathrm{var}(\hat{b}) = \sigma^2/\Sigma y_i^2 \tag{189}$$

where σ^2 is the variance of X_i.

The sum of the squared deviations is by (183), (186), and (187)

$$M = \Sigma(x - by)^2$$

Introducing (188), the minimal value

$$\mathrm{Min}(M) = \Sigma x^2 - (\Sigma xy)^2/\Sigma y^2 \tag{190}$$

An unbiased estimate of $\mathrm{var}(\hat{X}_i)$ is obtained by

$$\mathrm{var}(\hat{X}_i) = \sigma^2 = \left[\Sigma x^2 - (\Sigma xy)^2/\Sigma y^2\right]/(n-2) \tag{191}$$

It is required that this estimate shall be correct for any value of n. Since $M = 0$ for $r = 2$, no other denominator than $(n-2)$ is possible.

5.4 Linear Estimators for Independent Values

This method applied to order statistics is as demonstrated in (4.2.5), complicated due to the fact that order statistics are dependent of each other. In the present case the observations are independent. The procedure is for this reason considerably simplified, because then all covariances are equal to zero.

Let the regression line be

$$\hat{X} = a + bY \tag{192}$$

and

$$U = \Sigma v_i X_i \tag{193}$$

a linear estimator of the parameter

$$k_1 a + k_2 b \tag{194}$$

which may be reduced to a or to b by proper choice of the constants k_1 and k_2.

The system of equations (144) now take the simple form

$$\left.\begin{aligned} \sigma_i^2 \cdot w_i + \lambda + \mu Y_i &= 0 \\ \Sigma w_i &= k_1 \\ \Sigma w_i Y_i &= k_2 \end{aligned}\right\} \tag{195}$$

Introducing the notation

$$n_i = 1/\sigma_i^2 \tag{196}$$

which is called the weight of X_i, we obtain after some calculations the estimates by

$$\hat{b} = \Sigma n_i x_i y_i / \Sigma n_i y_i^2 \tag{197}$$

and

$$\hat{a} = \bar{X} - \hat{b}\bar{Y} \tag{198}$$

where $\bar{X}$ and $\bar{Y}$ now are the weighted means

$$\bar{X} = \Sigma n_i X_i / \Sigma n_i; \quad \bar{Y} = \Sigma n_i Y_i / \Sigma n_i \tag{199}$$

For $n_i = 1$, that is, for $\sigma_i = \sigma$, the estimates are identical with those deduced by means of the least-squares method, (186) and (188).

Since $\Sigma y_i = 0$ the term $\Sigma \bar{X} y_i = 0$
and consequently

$$\hat{b} = \Sigma x_i y_i / \Sigma y_i^2 = \Sigma X_i y_i / \Sigma y_i^2$$

The coefficients w_i of the estimator are

$$w_i = y_i / \Sigma y_i^2 \tag{200}$$

The variance of the estimator U is, considering (143),

$$\text{var}(U) = \Sigma w_i^2 \cdot \sigma_1^2 \tag{201}$$

Introducing (200) into (201) and putting in (194) $k_1 = 0$, $k_2 = 1$ the variance of $\hat{b}$ is

$$\text{var}(\hat{b}) = \sigma^2 \cdot \Sigma y_i^2 / (\Sigma y_i^2)^2 = \sigma^2 / \Sigma y_i^2 \tag{202}$$

as already stated by (189).

BIBLIOGRAPHY

Blom, G. (1958), "Statistical estimates and transformed beta-variables". Stockholm.

Cramér, H. (1945), "Mathematical methods of statistics". Uppsala.

Cummings, H.N., Stulen, F.B. & Schulte, W.C. (1955), "Investigation of materials fatigue problems applicable to propeller design". WADC TR 54-531.

Computation Laboratory of Harvard University, (1955), "Tables of the cumulative binomial probability distribution". Cambridge, Mass.

Dixon, W.J. & Massey, F.J. (1957), "Introduction to Statistical analysis", 2nd ed., New York.

Fisher, R.A. (1912), "On an absolute criterion for fitting frequency curves". Mess.of Math., Vol. 14, p. 155.

- - - - - (1921), "On the mathematical foundation of theoretical statistics". Trans.Roy.Soc.A. Vol. 222, p. 309.

- - - - - (1925), "Theory of statistical estimation", Proc.Cambr. Phil.Soc., Vol. 22, p. 700.

Fisher, R.A. & Tippett, L.H.C. (1928), "Limiting forms of the frequency distribution of the largest or smallest member of a sample". Proc.Cambr.Phil.Soc., Vol. 24, p. 180.

Fisher, R.A. & Yates, F. (1943), "Statistical tables". Edinburgh & London

Gumbel, E.J. (1954), "Statistical theory of extreme values and some practical applications". Nat.Bur.Stand.,Appl.Math.Ser.33, Washington.

- - - - - (1958), "Statistics of extremes". New York.

Kendall, M.G. & Babington Smith, B. (1938), "Randomness and random sampling numbers". J.Roy.Soc., Vol. 101, p. 592

Lieblein, J. (1954), "A new method of analyzing extreme-value data". NACA TN 3053.

- - - - - (1955), "On moments of order statistics from the Weibull distribution". Ann.Math.Stat. Vol. 26 (2), 330-333.

Lloyd, E.H. (1952), "Least squares estimation of location and scale parameters using order statistics". Biometrika, Vol. 39, p. 88.

Rao, C.R. (1945), "Information and accuracy attainable in the estimation of statistical parameters". Bull.Calcutta Math. Soc., Vol. 37, p. 81.

Tippett, L.H.C. (1927), "Random sampling numbers". Tracts for computers No. 15.

Weibull, W. (1939), "The phenomenon of rupture in solids". Ing.Vetensk. Akad.Handl. No. 153.