Chapter 2
Point Fitting Problems in One and Two Dimensions

Abstract We begin our analysis by considering the fitting of a single point to a number of point observations in one-dimensional space. Using the L_t-norm as optimality criterion with $t = 1$, $t = 2$ or $t = \infty$, we obtain the median, mean and midrange of a set of observations respectively. Similarly, applying the same three optimality criteria in the two-dimensional case, we obtain the mediancentre or centre of population, the centroid or centre of gravity and the unnamed centre of the circle of smallest radius respectively. Moreover, if we omit some of the more extreme observations then we obtain truncated variants of these procedures. As noted in Chap. 7, the midrange and its generalisations may be associated with a set of more or less familiar geometrical instruments: The univariate midrange with a pair of callipers, the bivariate midrange with a pair of compasses and the minimax fitted line of Chap. 3 with a pair of parallel rules.

Keywords Centre of gravity · Centre of population · Centroid · L_1-norm · L_2-norm · L_∞-norm · Linear programming · Mean · Median · Midrange · Mediancentre · Oja's bivariate median · Truncated midrange

2.1 Point Fitting Problems in One Dimension

Let $y_1, y_2, \ldots, y_n$ represent a set of n observations on a single variable Y, then these n observations may be represented by the n points at $y = y_1, y = y_2, \ldots, y = y_n$ on the y-axis of a Cartesian diagram. Moreover, we may identify a point of best fit to these n points by choosing a value for a in such a way that the sum of the squared distances

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

is minimised.

R. W. Farebrother, *L_1-Norm and L_∞-Norm Estimation*,
SpringerBriefs in Statistics, DOI: 10.1007/978-3-642-36300-9_2,

Taking square roots in this optimality function, we find that we may alternatively choose a value for a to minimise the root mean squared deviation function

$$\left[\sum_{i=1}^{n} |y_i - a|^2\right]^{\frac{1}{2}}.$$

and, replacing 2 by p with $0 < p \leq \infty$, in this expression, we have the optimality criterion employed in the more general L_t-norm point fitting problem

$$\left[\sum_{i=1}^{n} |y_i - a|^t\right]^{\frac{1}{t}}.$$

In this brief, we shall be largely concerned with two special cases of the L_t-norm problem: the first is identified by setting $t = 1$, when we have the sum of absolute deviations optimality criterion employed in the L_1-norm point fitting problem:

$$\sum_{i=1}^{n} |y_i - a|$$

and, in the limit as p tends to ∞, we have the minimax absolute residual optimality criterion employed in the corresponding L_∞-norm point fitting problem:

$$max_{i=1}^{n} |y_i - a|.$$

Both of these special cases of the general point fitting problem are readily solved: the L_t-norm optimality criterion defines the median (or middlemost) observation when $t = 1$, the arithmetic mean of the observations $\bar{y} = \sum_{i=1}^{n} y_i/n$ when $t = 2$, and the midrange (or midpoint of the shortest line segment containing all n observations) in the limit as t tends to ∞.

To define the median and the conventional midrange of the n observations, $y_1, y_2, \ldots, y_n$, we arranged these observations in increasing order as $y_{[1]} \leq y_{[2]} \leq \ldots \leq y_{[n]}$, then the median value of these n observations is given by $(y_{[m]} + y_{[m]})/2 = y_{[m]}$ when $n = 2m - 1$ is odd and by $(y_{[m]} + y_{[m+1]})/2$ when $n = 2m$ is even. Similarly, the conventional midrange is given by $(y_{[1]} + y_{[n]})/2$.

Now, all three of these expressions take the form $(y_{[r+1]} + y_{[n-r]})/2$ where $0 \leq r \leq n/2$, which we shall call the $(r,\ r)$-level symmetrically truncated midrange as its computation ignores the r smallest values and the r largest values of y_i.

As a more general variant of this expression, we may define the $(r,\ s)$-level nonsymmetrically truncated midrange $(y_{[r+1]} + y_{[n-s]})/2$ whose computation ignores the r smallest values and the s largest values of y_i. In particular, if we wish to retain m observations in the nonsymmetric case, then we have to choose values for r and s in such a way as to exclude $r + s = n - m$ observations from the computation so

that the remaining $n - r - s = m$ observations define the midpoint of the shortest line segment covering m of the observations. Moreover, if $n = 2m$ or $n = 2m - 1$ then this expression excludes one-half (or almost one-half) of the observations from the computation, and the optimal nonsymmetric truncated midrange is known as the 'shortest half'.

2.2 Point Fitting Problems in Two Dimensions

Generalising the representation of Sect. 2.1 to the 2-dimensional case, we find that we have n observations on the two variables X and Y. Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ represent a set of n matched pairs of observations on the two variables X and Y, then, for $i = 1, 2, \ldots, n$, the ith observation may be represented by a point at $(x, y) = (x_i, y_i)$ in the xy-plane of a two-dimensional Cartesian diagram.

In this context, and for each choice of $t > 0$, two definitions of the point of best fit become available: we may either separately minimise the L_t-norm of the distances measured perpendicular to the x-axis (and thus parallel to the y-axis)

$$\left[\sum_{i=1}^{n} |x_i - c|^t\right]^{\frac{1}{t}}$$

to obtain an optimal value for c at the same time as minimising the L_t-norm of the absolute distances measured perpendicular to the y-axis (and thus parallel to the x-axis)

$$\left[\sum_{i=1}^{n} |y_i - a|^t\right]^{\frac{1}{t}}$$

to obtain an optimal value for a.

Alternatively, we may simultaneously choose values for a and c to minimise the L_t-norm of the n Euclidean distances

$$\left\{\sum_{i=1}^{n} \left[(x_i - c)^2 + (y_i - a)^2\right]\right\}^{\frac{1}{t}}.$$

In the special case when $t = 2$ the square of this last expression may be written as

$$\left[\sum_{i=1}^{n} (x_i - c)^2\right] + \left[\sum_{i=1}^{n} (y_i - a)^2\right]$$

so that we obtain the same values $c = \bar{x}$ and $a = \bar{y}$ in this context as in the componentwise case mentioned above, where $\bar{y} = \sum_{i=1}^{n} y_i/n$ and $\bar{x} = \sum_{i=1}^{n} x_i/n$.

Thus, this alternative expression defines the mediancentre or centre of population of the n observations when $t = 1$, the centroid or centre of gravity when $t = 2$, and the unnamed centre of the circle of smallest area (and hence smallest radius) which just covers all n points in the limit as t tends to ∞.

As in Sect. 2.1, and for all values of $m \leq n$, we may readily generalise our definition of the one-dimensional nonsymmetrically truncated midrange to the two-dimensional case by replacing the midpoint of the shortest line segment which just covers $m \leq n$ points by the centre of the circle with smallest area which just covers $m \leq n$ points. Indeed, and for all values of $p \geq 1$, this last definition may be further generalised to the centre of the p-dimensional sphere with minimal p-dimensional volume which just covers $m \leq n$ points.

2.3 Truncated Point Fitting Problems in Two Dimensions

The mediancentre of a set of two-dimensional observations is a point (x_0, y_0) or (c, a) chosen in such a way that the sum of the lengths (or Euclidean distances) of the line segments joining the n given points to this arbitrary point takes its minimum value. [Gower (1974) has supplied an algorithm for performing the necessary calculations.]

Now, the line segments in this definition of the mediancentre may be replaced by triangles, triangular pyramids, ..., that is, by p-dimensional simplices where $p = 2$, 3, ... Thus an alternative definition of a central point of a set of two-dimensional observations is a point (x_0, y_0) chosen in such a way as to minimise the sum of the areas of the ${}^nC_2 = n(n-1)/2$ triangles defined by any two of the n given points and the arbitrary point. The centre defined in this way is known as Oja's (1983) bivariate median.

Now, for all $i < j$, the area of the triangle with vertices (x_0, y_0), (x_i, y_i) and (x_j, y_j) is given by one-half of the absolute value of the determinant of the 3×3 matrix

$$\begin{bmatrix} 1 & x_0 & y_0 \\ 1 & x_i & y_i \\ 1 & x_j & y_j \end{bmatrix}$$

that is, by one-half of the absolute value of

$$(x_i y_j - x_j y_i) - (y_j - y_i)x_0 + (x_j - x_i)y_0$$

or one-half of $w_{ij}|e_{ij}|$ where

$$\begin{aligned} w_{ij} &= |x_j - x_i| \\ e_{ij} &= y_0 - a_{ij} - x_0 b_{ij} \\ a_{ij} &= (x_j y_i - x_i y_j)/(x_j - x_i) \\ b_{ij} &= (y_j - y_i)/(x_j - x_i) \end{aligned}$$

and our problem takes the form of a weighted least absolute residuals fitting problem of the type discussed in Sect. 3.4 below provided that the coefficient of y_0 is nonzero, that is, provided that the x_i are distinct.

In other words, we have to choose x_0 and y_0 in such a way as to minimise a weighted sum of absolute values. Thus, the procedure for determining the value of Oja's bivariate median may be implemented in the form of a linear programming problem. Niinimaa et al. (1992) have provided such an algorithm.

The concept underlying Oja's bivariate median may readily be generalised to higher dimensions if we choose to minimise the sum of the p-dimensional volumes of the ${}^nC_p$ distinct p-dimensional simplices defined by the arbitrary point and any set of p of the n given points. In this case, we have to minimise the sum of the absolute values of the determinants of a set of $(p+1) \times (P+1)$ matrices divided by $p!$, see Farebrother (1992) for details.

References

Dodge, Y. (Ed.). (1992). L_1-*Statistical analysis and related methods*. Amsterdam: North-Holland Publishing Company.

Farebrother, R. W. (1992). The geometrical foundations of a class of estimation procedures which minimise sums of Euclidean distances and related quantities, in Dodge (pp. 337–349).

Gower, J. C. (1974). Mediancentre. *Applied Statistics*, *23*, 466–470.

Niinimaa, A., Oja, H., & Nyblom, J. (1992). The Oja bivariate median. *Applied Statistics*, *41*, 611–617.

Oja, H. (1983). Descriptive statistics for multivariate distributions. *Statistics and Probability Letters*, *1*, 327–332.

http://www.springer.com/978-3-642-36299-6

L1-Norm and L∞-Norm Estimation
An Introduction to the Least Absolute Residuals, the Minimax Absolute Residual and Related Fitting Procedures
Farebrother, R.
2013, VI, 58 p., Softcover
ISBN: 978-3-642-36299-6