CONFIDENCE REGIONS FOR LINEAR REGRESSION

PAUL G. HOEL
UNIVERSITY OF CALIFORNIA, LOS ANGELES

1. Introduction

It is well known that Student's t-distribution yields a confidence interval for the ordinate, y, of a regression line corresponding to any fixed valued of x, under the assumption that the sample x's are fixed variates and the corresponding sample ordinates are independently normally distributed about the regression line with a common variance. Less well known is the result of Hotelling and Working [1] in which a confidence band is obtained for the entire regression line, although with the additional assumption that the common variance of the sample ordinates is known.

Confidence bands are a particularly useful tool in those sampling problems that produce an estimate of a curve, such as a growth curve. Very often such curves can be treated as special cases of linear regression in several variables. It is not difficult to extend the methods employed in [1] to linear regression in several variables and thus obtain confidence bands for such curves.

In attempting to obtain a confidence band, it is desirable to seek for one that is as narrow as possible, in some sense, over the range of interest. The confidence band obtained in [1] was derived with mathematical convenience in mind, rather than with optimum properties dominant. The purpose of this paper is to derive confidence bands from an optimum point of view and to study the extent to which the confidence band of [1] is optimum. For simplicity of explanation, the discussion will be limited mostly to the regression line; however a generalization to linear regression in several variables is straightforward.

2. General confidence bands

This section will be concerned with deriving the equations that define a fairly general confidence band for a regression line. Consider a fixed set of x's: $x_1, x_2, \ldots, x_n$. Let y_i corresponding to x_i be normally distributed with mean $\alpha + \beta(x_i - \bar{x})$ and variance σ^2, and let the y_i be independently distributed. It will be assumed that σ^2 is known; however in a later section this restriction will be removed. Let

$$(1) \qquad \alpha^* = \bar{y}\,, \qquad \beta^* = \frac{\sum (x_i - \bar{x})\, y_i}{\sum (x_i - \bar{x})^2},$$

$$u = \frac{\sqrt{n}\,(\alpha - \alpha^*)}{\sigma}, \qquad v = \frac{\sqrt{n}\, s_1 (\beta - \beta^*)}{\sigma},$$

This research was done under the sponsorship of the Office of Naval Research.

where $n s_1^2 = \sum (x_i - \bar{x})^2$. It is easily shown [2, p. 550] that u and v are independently normally distributed with zero means and unit variances.

One method for finding a confidence band for the line $y = \alpha + \beta(x - \bar{x})$ depends upon finding the envelope of a single parameter family of lines obtained by restricting α and β to satisfy a certain functional relationship. The following derivation proceeds along such lines. For this purpose let $g = g(u^2, v^2)$ be a single valued function of the random variables u and v that possesses a probability density function, $f(g)$, which is positive, except possibly at its extremities. Let g_c be the value of g such that

$$\int_{-\infty}^{g_c} f(g)\, dg = c .$$

It will also be assumed that the function g is such that the equation

$$g(u^2, v^2) = g_c \tag{2}$$

defines a closed curve in the u, v plane and that this equation may be written in the explicit form

$$u^2 = t(v^2) ,$$

and hence in the form

$$u = \pm h(v) . \tag{3}$$

For the equivalence of (2) and (3), it suffices that g possess continuous derivatives with respect to u^2 and v^2 and that the derivative with respect to u^2 does not vanish.

For a given sample, u and v are functions of α and β; consequently the equation of the regression line $y = \alpha + \beta(x - \bar{x})$ may be written in the form

$$y = \alpha^* + \frac{\sigma}{\sqrt{n}} u + \left(\beta^* + \frac{\sigma}{s_1\sqrt{n}} v\right)(x - \bar{x}) ,$$

where u and v are now treated as the unknown parameters. If u and v are restricted to satisfy (2), a single parameter family of possible regression lines will be obtained. There will be two such families, corresponding to the two signs in (3). The equations of these two families, with v as the parameter, will be

$$y = \alpha^* \pm \frac{\sigma}{\sqrt{n}} h(v) + \left(\beta^* + \frac{\sigma}{s_1\sqrt{n}} v\right)(x - \bar{x}) . \tag{4}$$

If these families of lines possess envelopes, the equation of either envelope may be obtained by eliminating v between equation (4) and the equation obtained by differentiating (4) with respect to v. The derivative of (4) reduces to

$$0 = \pm h'(v) + \frac{x - \bar{x}}{s_1} . \tag{5}$$

Treating v as the parameter, equations (4) and (5) yield the following equations of the two envelopes, expressed in parametric form:

$$x = \bar{x} - s_1 h'(v) \tag{6}$$

$$y = \alpha^* + \frac{\sigma}{\sqrt{n}} h(v) - \left(\beta^* + \frac{\sigma}{s_1\sqrt{n}} v\right) s_1 h'(v)$$

and

$$x = \bar{x} + s_1 h'(v) \tag{7}$$

$$y = a^* - \frac{\sigma}{\sqrt{n}} h(v) + \left(\beta^* + \frac{\sigma}{s_1\sqrt{n}} v\right) s_1 h'(v).$$

Now consider the restrictions that must be placed on $h(v)$ in order that (6) and (7) will yield envelopes that will determine acceptable confidence bands for the regression line. It will suffice to inspect (6). It is essential that the envelope (6) be single valued, that it exist for all values of x, and that all members of the family lie on one side only of the envelope. It is clear from (6) that the first two of these properties will be satisfied if, and only if, $h'(v)$ is a monotonic function of v which takes on all real values. For convenience, $h(v)$ as given by (3) will be defined so that $h'(v)$ is a decreasing function. It is readily verified that the third condition will be satisfied with the preceding restriction on $h(v)$ and that the two envelopes defined by (6) and (7) always exist, with the curve (6) lying above the curve (7). Furthermore, all lines of the families (4) will lie between these two curves. Since the two families of (4) correspond to a single family for restriction (2), curves (6) and (7) will be called the upper and lower branches of the envelope of the regression lines obtained by restricting u and v to satisfy (2).

Now suppose that $g(u^2, v^2)$ of (2) is such that if g_c is decreased the new curve will lie inside the old curve and that every interior point will lie on one such curve. From (4) it is clear that every pair of values of u and v that satisfies the inequality

$$g(u^2, v^2) < g_c \tag{8}$$

will correspond to a regression line that lies between the curves (6) and (7) because the value of $h(v)$ for any fixed v will be decreased, whereas the slope of the line will be unchanged. Thus, if u and v satisfy (8), the true regression line will lie between the two branches of the envelope given by (6) and (7). Since g_c was selected so that the probability is c that (8) will be satisfied, the region between the two curves whose equations are given by (6) and (7) will constitute a confidence band for the true regression line with confidence coefficient c.

3. General optimum confidence bands

It is desirable to find a confidence band for the regression line that will be as narrow as possible. The natural geometrical property to seek is minimum area in some sense. Since the area of such a band is infinite, whereas extreme values of the line are of little interest, it is necessary to introduce a weight function that will make the weighted area finite and that will weight y relative to the frequency with which its corresponding x is likely to occur. Although the x's are treated as fixed variates, their values are often obtained from sampling a bivariate population.

If the weight function selected is denoted by $w(x)$, a confidence band with confidence coefficient c will be said to be optimum if it minimizes the weighted area of the band, that is, if it minimizes the integral

$$I = \int_{-\infty}^{\infty} (y_2 - y_1) w(x)\, dx,$$

subject to restriction (2), where y_2 and y_1 are the upper and lower envelope curves given by (6) and (7). If $w(x)$ has been normalized, I may be treated as the expected value of $y_2 - y_1$; hence

$$I = E(y_2) - E(y_1).$$

The weight function that will be chosen here is the normal function

$$w(x) = \frac{e^{-(x-\bar{x}/s_1)^2/2}}{s_1\sqrt{2\pi}}. \tag{9}$$

Since the x's are treated as fixed variates, $\bar{x}$ and s_1 are known constants. Because of the symmetry of $w(x)$, it follows from (6) and (7) that

$$E(y_2) = a^* + \frac{\sigma}{\sqrt{n}}E(h) - \frac{\sigma}{\sqrt{n}}E(vh'),$$

$$E(y_1) = a^* - \frac{\sigma}{\sqrt{n}}E(h) + \frac{\sigma}{\sqrt{n}}E(vh'),$$

and hence that

$$I = \frac{2\sigma}{\sqrt{n}}[E(h) - E(vh')].$$

Changing variables from x to v by means of (6), it follows that

$$E(h) = -\frac{1}{\sqrt{2\pi}}\int_{v_1}^{v_2} hh''e^{-(h')^2/2}dv,$$

where v_1 and v_2 correspond to $-\infty$ and $+\infty$. Similarly,

$$E(vh') = -\frac{1}{\sqrt{2\pi}}\int_{v_1}^{v_2} vh'h''e^{-(h')^2/2}dv.$$

Since, by (2) and (3), $h(v)$ is an even function, it suffices to integrate over positive values of v only; hence these two results give

$$I = -\frac{4\sigma}{\sqrt{2\pi n}}\int_0^{v_2} h''(h - vh')e^{-(h')^2/2}dv.$$

Now from (1), together with symmetry and the equivalence of (2) and (3), it follows that condition (8) can be written in the form

$$\int_0^{v_2}\int_0^{h} e^{-(u^2+v^2)/2}du\,dv = \frac{c\pi}{2}. \tag{10}$$

Shifting to more customary notation, the problem of finding an optimum confidence band for a regression line has been reduced to finding a function $y(x)$, satisfying certain conditions enumerated earlier, that minimizes the integral

$$J = -\int_0^a y''(y - xy')e^{-(y')^2/2}dx \tag{11}$$

subject to the restriction

$$k = \int_0^a e^{-x^2/2}\int_0^y e^{-t^2/2}dt\,dx. \tag{12}$$

At first glance, this problem appears to be a standard problem in the calculus of

variations; however unless the class of admissible arcs is restricted properly, the solution will be the trivial one in which $y(x)$ is constant. But, as was pointed out in the discussion following (7), $y'(x)$ must be a monotonic function taking on all values if an acceptable confidence band is to be obtained.

4. Special optimum confidence bands

The confidence band derived in [1] is easily seen to correspond to choosing the curve (2) to be a circle. The simplest class of arcs that will satisfy the essential conditions on $y'(x)$ and contain a circle as a special case is the family of ellipses with axes coinciding with the coordinate axes; hence $y(x)$ will be so restricted. The equation of this family will be written in the form $y = b\sqrt{a^2 - x^2}$.

For the purpose of minimizing J, consider its derivative. J may be treated as a function of a, since b is a function of a through (12). Then, differentiating (11),

$$\frac{dJ}{da} = -\int_0^a \left\{ (y - xy') \frac{\partial y''}{\partial a} - (xy'' + yy'y'' - xy'^2y'') \frac{\partial y'}{\partial a} + y'' \frac{\partial y}{\partial a} \right\} e^{-(y')^2/2} dx,$$

because the integrand of J vanishes at the upper limit. By integrating by parts and applying boundary conditions, it follows readily that

$$\int_0^a (y - xy') \frac{\partial y''}{\partial a} e^{-(y')^2/2} dx = \int_0^a (xy'' + yy'y'' - xy'^2y'') \frac{\partial y'}{\partial a} e^{-(y')^2/2} dx.$$

Consequently,

$$\frac{dJ}{da} = -\int_0^a y'' \frac{\partial y}{\partial a} e^{-(y')^2/2} dx.$$

Since $y = b\sqrt{a^2 - x^2}$ here, calculations yield

$$(13) \quad \frac{dJ}{da} = a^2 b \int_0^a (a^2 - x^2)^{-3/2} \left[ab(a^2 - x^2)^{-1/2} + (a^2 - x^2)^{1/2} \frac{db}{da} \right] \times e^{-b^2x^2/2(a^2 - x^2)} dx.$$

An expression for $\frac{db}{da}$ may be obtained by differentiating (12), thus

$$(14) \quad 0 = \int_0^a \left[ab(a^2 - x^2)^{-1/2} + (a^2 - x^2)^{1/2} \frac{db}{da} \right] e^{-(x^2 + y^2)/2} dx.$$

If the change of variable $x = at$ is made in both (13) and (14), and then (14) is solved for $\frac{db}{da}$ and substituted in (13), it will be found that (13) reduces to

$$(15) \quad \frac{dJ}{da} = b^2 e^{b^2/2} \left\{ \int_0^1 (1 - t^2)^{-2} e^{-b^2/2(1 - t^2)} dt - \int_0^1 (1 - t^2)^{-1} e^{-b^2/2(1 - t^2)} dt \times \frac{\int_0^1 (1 - t^2)^{-1/2} e^{-a^2(1 - b^2)t^2/2} dt}{\int_0^1 (1 - t^2)^{1/2} e^{-a^2(1 - b^2)t^2/2} dt} \right\}.$$

5. Classical confidence band

The procedure employed in [1] consists in recognizing that $u^2 + v^2$ possesses a χ^2-distribution and then constructing a confidence band with given confidence

coefficient by choosing the curve (2) to be the proper circle. This case is obtained by letting $b = 1$, in which case (15) reduces to

$$\frac{dJ}{da} = e^{1/2}\left\{\int_0^1 (1-t^2)^{-2} e^{-1/2(1-t^2)}dt - 2\int_0^1 (1-t^2)^{-1} e^{-1/2(1-t^2)}dt\right\}.$$

Calculations show that the value of this expression is not zero; therefore the classical confidence band for linear regression is not optimum as defined in section 3.

For a given confidence coefficient c, it is possible to solve the equation $\frac{dJ}{da} = 0$ by using (12) to obtain the numerical relation between a and b and then making successive approximations to the root of the equation. If c is chosen as .95, it turns out that the optimum ellipse has semiaxes of 2.62 and 2.32 as compared to the classical case with 2.45 and 2.45. Since the optimum ellipse differs so little from the classical circle, it is illuminating to compare the values of the integral J. It will be found that the classical case yields a value of J which is less than 1% larger than that for the optimum ellipse. Thus, the classical approach based largely on mathematical convenience turns out to be surprisingly efficient as judged by a more critical approach. Although the optimum curve satisfying the essential restrictions may not be an ellipse, the restrictions are such that it appears that the optimum curve is likely to differ but little from an ellipse, and consequently the optimum value of J is likely to be only slightly smaller than that for the circle.

If the normal weight function (9) is replaced by the weight function

$$w(x) = \left[1 + \left(\frac{x - \bar{x}}{s_1}\right)^2\right]^{-3/2},$$

and if families of ellipses only are considered, it will be found that the classical confidence band corresponding to a circle will now be optimum.

6. Unknown variance

The preceding sections have assumed that σ was known. When σ is unknown, it suffices to consider the additional variable $\xi^2 = n\sigma^{*2}/\sigma^2$, where $n\sigma^{*2} = \sum [y_i - \alpha^* - \beta^*(x_i - \bar{x})]^2$. Then the variables $y = u\sqrt{n-2}/\xi$ and $z = v\sqrt{n-2}/\xi$ will possess Student's t-distributions, and hence any probability density function $g = g(y^2, z^2)$ will possess a distribution that is independent of α, β, and σ. The methods of the preceding sections now apply to the variables y and z instead of the variables u and v. Restriction (8) gives rise to a more complicated expression than (10); consequently it is formally more difficult to study optimum properties when σ is unknown. Since the results of the preceding sections are independent of n, a confidence band with σ unknown will not be optimum for general n if it is not optimum when σ is known.

7. Multivariate regression

Since confidence regions for linear regression in several variables do not seem to have been considered in the literature, for completeness their derivation will be outlined here. If the regression equation is

$$y = \alpha + \beta_1(x_1 - \bar{x}_1) + \ldots + \beta_k(x_k - \bar{x}_k), \tag{16}$$

and σ is known, one employs the variables

$$(17)\quad u = \frac{(a - a^*)\sqrt{n}}{\sigma}, \qquad v_1 = \frac{(\beta_1 - \beta_1^*)\sqrt{n}\, s_1}{\sigma}, \quad \ldots,$$

$$v_k = \frac{(\beta_k - \beta_k^*)\sqrt{n}\, s_k}{\sigma},$$

which will be normally distributed with zero means and unit variances. If one considers a function g such that

$$(18)\qquad g(u^2, v_1^2, \ldots, v_k^2) = g_c$$

defines a closed surface with properties analogous to those for g of (2), one proceeds exactly as before to derive an envelope surface for the family of planes given by (16), (17), and (18). Its equation in parametric form with parameters $v_1, \ldots, v_k$ will be

$$x_1 = \bar{x}_1 \mp s_1 h_1, \ldots, x_k = \bar{x}_k \mp s_k h_k,$$

$$y = a^* \pm \frac{\sigma}{\sqrt{n}} h \mp \left(\beta_1^* + \frac{\sigma}{s_1\sqrt{n}} v_1\right) s_1 h_1 \mp \ldots \mp \left(\beta_k^* + \frac{\sigma}{s_k\sqrt{n}} v_k\right) s_k h_k,$$

where h_i denotes the derivative of h with respect to v_i and where the two sets of signs correspond to the upper and lower surfaces that bound the confidence region.

If σ is unknown, one introduces a generalization of ξ given in section 6 and employs the variables

$$y = \frac{u}{\xi}, \qquad z_1 = \frac{v_1}{\xi}, \quad \ldots, \quad z_k = \frac{v_k}{\xi},$$

which, when multiplied by the proper constant, will possess Student's t-distributions. The methods employed here are the same as those for the case when σ is known.

REFERENCES

[1] H. HOTELLING and H. WORKING, "Applications of the theory of error to the interpretation of trends," *Jour. Amer. Stat. Assoc., Suppl.*, Vol. 24 (1929), pp. 73–85.

[2] H. CRAMÉR, *Mathematical Methods of Statistics*, Princeton University Press, Princeton, 1946.