HOUGH SPACE AND CONFIDENCE INTERVALS FOR REGRESSION LINES

by

Louis J. Cote
Purdue University

Technical Report #87-30

Department of Statistics
Purdue University

October 1987

HOUGH SPACE AND CONFIDENCE INTERVALS FOR REGRESSION LINES SAMPLE

by

Louis Cote
Purdue University

ABSTRACT

A technique used in pattern recognition called Hough Space is used to derive confidence regions for simple linear regression lines. These confidence regions are based on convex confidence regions for the regression coefficients. Examples are worked out when the latter are ellipses, rectangles, and parallelograms.

HOUGH SPACE AND CONFIDENCE INTERVALS FOR REGRESSION LINES

by

Louis J. Cote
Purdue University

Textbooks usually show diagrams of the loci of confidence intervals for the predictor and of prediction intervals in simple linear regression. These are, in the case of normal errors, hyperbolas whose minor axis is the regression line. The books usually explain that the area between the hyperbolas is not a confidence region for the regression line and at least two, (2, p. 460; 4, p. 154–6), present the confidence region for the regression line as the region between two branches of a hyperbola when the joint confidence region for the regression coefficients is elliptical (6). The same texts also discuss rectangular Bonferroni confidence regions for the regression coefficients. Is there a confidence region for the regression line corresponding to a rectangular confidence region for its coefficients? If there is what does it look like?

It is easy to show that such regions exist. The methods are elementary and can be explained in terms of a technique used in pattern recognition called Hough Space. This originated not in a research paper, but in a U. S. patent (3). In its simplest form a Hough space is a Cartesian space with axes β_1, β_0 in which a point maps into a line, $y = \beta_0 + \beta_1 x$, in the x, y plane. In engineering use, however, the line is expressed in normal form (1, Ch. 9) and the Hough space coordinates are the length of the perpendicular from the origin and the angle it makes with the x-axis. For us it is more convenient to work with the slope intercept-form of simple linear regression. Thus a line in the x, y plane maps into a point in the β_1, β_0 space. A point, (x_0, y_0) determines a bundle of straight lines $y = mx_0 + (y_0 - mx_0)$ which map into the locus,

$$\beta_0 = y_0 - mx_0$$
$$\beta_1 = m,$$

in the Hough space. This is a straight line with slope $-x_0$ and intercept y_0. In this way a point on the x, y plane maps into a line in Hough space. In pattern recognition the x, y plane is monitored by a television camera and points in the raster with sudden changes in brightness are identified. If several of these lie on a line, the lines in Hough space corresponding to them will intersect in the point that maps into the x, y-line. In practice, of course, nothing is done to ultimate accuracy, but the mapping is continuous so all the intersection points of the lines from the x, y discontinuities made by a line will group around the parameters of that line. In this way (x, y) lines can be identified from frequency data in Hough space (5).

Suppose R is a convex set in Hough space. Each point in R maps into a line in the x,y plane. Suppose, for some value of x, $U(x) = \{y : \exists(\beta_1, \beta_0) \epsilon R : \; y = \beta_0 + \beta_1 x\}$. Let y_1 and y_2 be points in $U(x)$ and $y_3 = ay_1 + (1-a)y_2$ with $0 < a < 1$. There are points in Hough space such that $y_1 = \beta_{01} + \beta_{11} x$, and $y_2 = \beta_{02} + \beta_{12} x$. Then

$y_3 = a\beta_{01} + (1-a)\beta_{02} + (a\beta_{11} + (1-a)\beta_{12})x$. Since R is convex, y_3 is in $U(x)$. Therefore $U(x)$ is an interval in the vertical line at x in the x, y plane. The endpoints, $Y_U(x)$ and $Y_L(x)$ are the max and min of $y = \beta_0 + \beta_1 x$. Writing this as the line, $\beta_0 = -x\beta_1 + y$, in Hough space, we see that the max and min of the y's are the intercepts of the lines of slope $-x$ tangent to R.

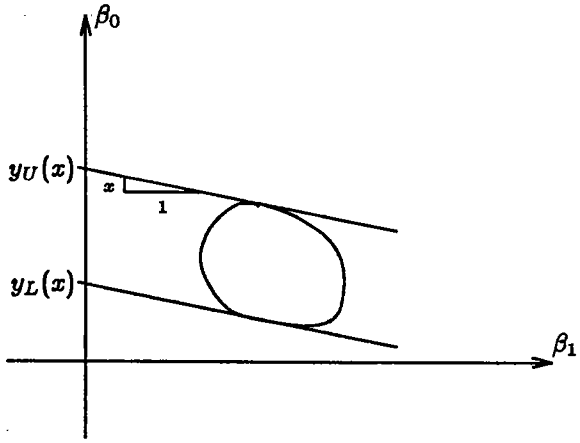

Fig. 1

Example 1. Rectangular R whose sides are parallel to the coordinate axes in Hough space. Let the lower left vertex be (β_{11}, β_{01}), and the upper right vertex, (β_{12}, β_{02}). When $x \neq 0$, the tangent lines to R pass through two opposite vertices.

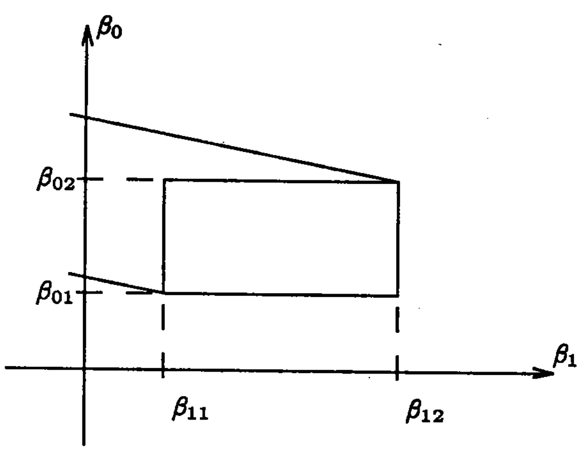

Fig. 2

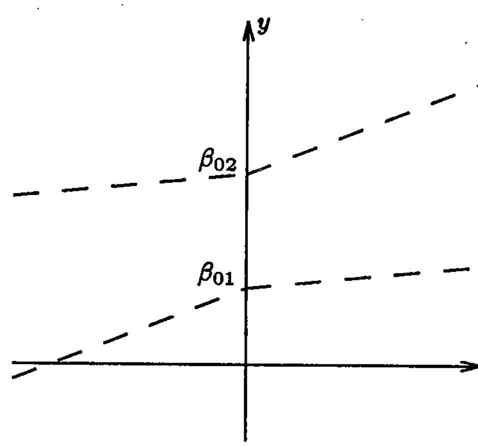

Fig. 3

When $x > 0$, these are (β_{11}, β_{01}) and (β_{12}, β_{02}), the latter giving the maximum. The tangent line through (β_{11}, β_{01}) in Hough space is

$$\beta_0 - \beta_{01} = -x(\beta_1 - \beta_{11}).$$

The intercept of this line, which is the lower end of $U(x)$, is $\beta_{01} + \beta_{11}x$. Thus the locus of the lower ends in the x, y plane is

$$y = \beta_{01} + \beta_{11}x \text{ for } x > 0.$$

The upper end locus, by similar reasoning is

$$y = \beta_{02} + \beta_{12}x \text{ for } x > 0.$$

These are diverging half-lines from the y-intercepts β_{01} and β_{02}. For $x < 0$, the parallel tangents pass through (β_{12}, β_{01}) and (β_{11}, β_{02}), the latter giving the maximum. The points in x, y corresponding to these tangency points are half-lines diverging to the left from the same y-intercepts. If the rectangle in Hough space is a Bonferroni confidence region for β_1, β_0, the regression estimates (b_1, b_0), are usually in its center. The estimated regression line in the x, y plane passes through the midpoints of all the vertical intervals between these boundary lines. The area between these diverging pairs of rays is the confidence region for the regression line with the confidence of the Bonferroni region in Hough space. The narrow neck of this figure would seem to be more appropriately placed near $(\overline{x}, \overline{y})$ rather than on the y-axis. This can easily be arranged by adjusting the Bonferroni region. If we translate the x, y origin to $(\overline{x}, \overline{y})$; i.e.

$$y' = y - \overline{y}, \quad x' = x - \overline{x}.$$

Then a regression line, $y = \beta_0 + \beta_1 x$, referred to the new coordinates, is

$$y' = (\beta_0 - \overline{y} + \beta_1 \overline{x}) + \beta_1 x'$$

The Bonferroni region for the parameters of this line has the form

$$|b_1 - \beta_1| \leq d_1$$

$$|b_0 - \beta_0 + (b_1 - \beta_1)\overline{x}| \leq d_0.$$

This is a parallelogram in the Hough space with two vertical sides, the other sides having the slope $-\overline{x}$. The parallel tangents again touch at vertices, but the change of vertex takes place when the slope passes through $-\overline{x}$ or through infinity. Thus on the x, y plane the break in the diverging lines is at $\overline{x}$ rather than at 0.

Suppose the convex set, R, is bounded by curves so the points of tangency change continuously as the slope, $-x$ changes. If $(\beta_1(x), \beta_0(x))$ is a point of contact on the curve for a tangent of slope $-x$, the upper or lower point on the vertical at x in the x, y plane is given by $y = \beta_0(x) + \beta_1(x)x$.

Example 2. Suppose the confidence set for (β_1, β_0) is the ellipse,

$$A(\beta_1 - b_1)^2 + 2B(\beta_1 - b_1)(\beta_0 - b_0) + C(\beta_0 - b_0)^2 \leq 1.$$

Since this is an ellipse, the left side is positive definite, i.e. $A > 0, C > 0, AC - B^2 > 0$. The condition on r for a line $\beta_0 - b_0 = -x(\beta_1 - b_1) + r$ to be tangent to the ellipse is easily derived by ensuring that the quadratic equation for the β_1-coordinate of the intersection has a double root. It is

$$r^2 = \left(\frac{(B - Cx)^2}{AC - B^2} + 1\right)\frac{1}{C}.$$

Because of the condition on the coefficients of the ellipse, the right side of this expression is positive for all x and the values of r are real. The value of β_1 at a point of contact of a tangent line is given by the double root of the quadratic. The value of β_0 comes from the equation for the tangent line;

$$\beta_1 - b_1 = -r\frac{(B - Cx)}{A - 2Bx + Cx^2}$$
$$\beta_0 - b_0 = r\frac{(A - Bx)}{A - 2Bx + Cx^2}$$

Then the boundary of the x, y confidence interval for the regression lines is given by the equation,

$$y = b_0 + \frac{r(A - \beta x)}{A - 2Bx + Cx^2} + (b_1 - \frac{r(B - Cx)}{A - 2Bx + Cx^2})x$$

which works out to be

$$\frac{(y - b_0 - b_1x)^2}{1/C} - \frac{(B - Cx)^2}{AC - B^2} = 1,$$

a hyperbola.

The equations of the asymptotes of this curve are gotten by setting the left side of the above equation to zero.

$$y - b_0 - b_1x = \pm(B - Cx)/\sqrt{C(AC - B^2)}.$$

The two asymptotes are lines which are equally distant above and below the estimated regression line, $y = b_0 + b_1x$ at any x. The center of the hyperbola is at the point where the asymptotes meet. The x coordinate of this point is given by $B - Cx = 0$.

The elliptical confidence region for (β_1, β_0) given by normal theory has coefficients (4, p. 148),

$$A = \frac{\Sigma X^2}{2\hat{\sigma}^2 F},$$
$$B = \frac{\Sigma X}{2\hat{\sigma}^2 F},$$
$$C = \frac{n}{2\hat{\sigma}^2 F},$$

where $\hat{\sigma}^2$ is the mean squared error from the regression $ANOVA$ and $F = F_{2,n-2}(1 - \alpha)$. The intersection of the asymptotes is, therefore, at $x = \beta/C = \overline{x}$. At $\overline{x}$ the hyperbola intersects the vertical at distances above and below the midpoint,

$$\sqrt{1/C} = \sqrt{2\hat{\sigma}^2 F/n.}$$

The equation for the boundary of this $1 - \alpha$ confidence region is

$$y = b_0 + b_1x \pm \sqrt{\frac{1}{C}}\sqrt{\frac{(B - Cx)^2}{AC - B_2} + 1}$$

or

$$y = b_0 + b_1 x \pm \hat{\sigma}\sqrt{2F}\sqrt{\frac{1}{n} + \frac{(\overline{x} - x)^2}{SXX}}$$

where $SXX = \Sigma(X - \overline{x})^2$. This formula is the same as the one for the locus of endpoints of the confidence limits for the predictor except for the multiplier $\sqrt{2F_{2,n-2}(1-\alpha)}$ instead of $t_{n-2}(1-\alpha/2)$.

A useful comparison of the three examples of confidence regions for regression lines cannot be made without having a loss function unless one region is inside another. It is easy to show that there are no inclusions among the three regions in Hough space and thus none also in the x, y plane. A comparison of the vertical widths of the three regions at their narrowest parts can be made:

$$\text{Bonferroni rectangle } 2t\hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{x}^2}{SXX}}$$

$$\text{Bonferroni parallelogram } 2t\hat{\sigma}\sqrt{\frac{1}{n}}$$

$$\text{Ellipse } 2\sqrt{2F}\hat{\sigma}\sqrt{\frac{1}{n}}.$$

In these formulas $t = t_{n-2}(1-\alpha/4)$. For any reasonable α they are ordered largest to smallest from top to bottom.

BIBLIOGRAPHY

(1) R. O. Duda and P. E. Hart, "Pattern Classification and Scene Analysis." (Wiley, 1973).

(2) S. Ehrenfeld and S. Littauer, "Introduction to Statistical Method." (McGraw Hill, 1964).

(3) P. V. C. Hough, "Method and Means for Recognizing Complex Patterns." U. S. Patent 3069654, (December 18, 1962).

(4) J. Neter, W. Wasserman, and M. H. Kutner, "Applied Linear Statistical Models." (2nd Ed. 1986, Irwin).

(5) S. Tsuji and F. Matsumoto, "Detection of Elliptic and Linear Edges by Searching Two Parameter Spaces." *Proc. 5th Int'l. Conf. on Artif. Intell.* (M. I. T., 1977, Vol. 2).

(6) Working H. Hotelling, "Application of the Theory of Error to the Interpretation of Trends." *JASA* v. **24**, pp. 73–85 (1929).