Local Predictive Influence

by

Michael Lavine
University of Minnesota
Technical Report No. 503
November 1987

Research supported by National Institute of Health Grant No. GM 25271.

LOCAL PREDICTIVE INFLUENCE

by

Michael Lavine

University of Minnesota

0. Introduction

This paper gives a specific application of a general paradigm that was described by Cook (1986), and McCulloch (1985). Let M represent the ingredients of a statistical problem, M = (model, data) where the model consists of a set of sampling distributions and, for Bayesians, a set of prior distributions on the sampling distributions. An analysis technique T maps each M into an answer: T(M) = a where a might be a parameter estimate, a confidence interval, a probability or any other type of inference.

Let M be a function of a vector ω where ω_0 is a standard and other values of ω represent perturbations of the standard. For example, in a regression setting, ω may be an n-vector of case weights, an n-vector of perturbations in the observations, or an $n \times p$ matrix of perturbations in the covariates. For these examples, ω_0 would be the vector of all 1's, the 0 vector, and the 0 matrix.

Let D be a discrepancy function between pairs of answers, where $D(a_1, a_2) \in R$. The function D measures the influence that a perturbation scheme has on the outcome of the analysis. Cook (1986) suggests that we often want to examine the function

$$h(\omega) = D\left[T(M(\omega_0)),\ T(M(\omega))\right] \text{ for } \omega\text{'s}$$

in a neighborhood around ω_0.

Many useful choices for D will satisfy $D(a_1, a_2) \geq 0$ and $D(a, a) = 0$. Assume, from now on, that these conditions are met and therefore that h has a local minimum at $\omega = \omega_0$. The shape of h at ω_0 is an indicator of how drastically the inference changes as a function of ω, at least locally.

When h is twice differentiable the shape of h at ω_0 can be studied through the curvature, which in turn can be studied through the curvature in one direction at a time. Any vector ω can be written as $\omega = r \cdot d$ where r is a scalar and d is a unit vector. The curvature C_d in the direction d is defined to be

$$C_d = \left.\frac{\partial^2 h(\omega)}{\partial r^2}\right|_{r=0}.$$

If the maximum curvature, $\sup_d C_d$, is large then small changes in ω can make large changes in the inference. On the other hand, a small maximum curvature is evidence that the analysis is robust to small changes in M.

The remaining sections of this paper show to to compute c_d and $\sup C_d$ for one particular type of analysis, perturbation scheme and discrepancy function.

1.2 Framework

Let the data consist of independent random variables $Y_1, \ldots, Y_n$ and p-dimensional covariates $X_1, \ldots, X_n$. Assume that the normal linear model with different case weights applies, i.e.,

$$Y_i \sim N(X_i^t \beta, \sigma_i^2)$$

Let X be the matrix $(X_1, X_2, \ldots, X_n)^t$ so the model can be written

$$Y \sim N(X^t \beta, \sigma^2 S)$$

where β is the p×1 vector of regression coefficients, σ^2 is a positive scalar, and S is a positive-definite diagonal matrix. A standard assumption is that all the case weights are equal. Let $\omega = (\omega_1, \ldots, \omega_n)^t$ be a vector representing changes from identical case weights, so that the diagonal of S is $(1/(1+\omega_1), \ldots, 1/(1+\omega_n))$. The 0 vector is ω_0.

Let the prior be the usual improper, non-informative prior proportional to $\sigma^{-2} d\beta d\sigma^2$, and suppose that the goal of the analysis is to compute a predictive density for a future random variable Z at known covariate w that satisfies

$$Z \sim N(w^t \beta, \sigma^2).$$

The Kullback-Leibler directed divergence betwen two densities f and g is defined to be $I(f,g) = \int ln(f(x)/g(x))f(x)dx$. Let the discrepancy function D be the Kullback-Leibler divergence, so that $h(\omega) = I(f, f_\omega)$ where f is the predictive density computed with equal weights and f_ω is the predictive density computed with weights $(1+\omega_i)$.

By the linear transformation $X^*=S^{1/2}X$ and $Y^*=S^{1/2}Y$ we get the new model $Y^* \sim N(X^{*t}\beta, \sigma^2 I)$ that has the same weight for every case.

he distribution of Z given w, X and Y is the Student distribution $St(n-p, w^t\hat{\beta}, (1+v)s^2)$ where p is the dimension of β, $\hat{\beta}=(X^{*t}X^*)^{-1}X^{*t}Y^*$, $v=w^t(X^{*t}X^*)^{-1}w$, $s^2=Y^{*t}QY^*/(n-p)$, $Q=I-X^*(X^{*t}X^*)^{-1}X^{*t}$ is the orthogonal projection operator parallel to the column space of X^* and the distribution St(a,b,c) has density proportional to $dz[1+(z-b)^2/ac]^{-(b+1)/2}$ (Geisser (1965), Johnson and Geisser (1982)).

By interchanging integration and differentiation and after some tedious calculus we see that

$$C_d \text{ is } d^t(M1 + M2 + M3 + M4)d$$

where M1, M2, M3, and M4 are each rank one matrices. They are defined in terms of $z^t=(z_1,\ldots,z_n)=w^t(X^tX)^{-1}X^t$ and the vector of residuals $QY=r=(r_1,\ldots,r_n)^t$. The four matrices are

$M1 = (n-p)/(2(n-p+3)(1+v)^2) \cdot [z^\circ z]\ [z^\circ z]^t$

$M2 = -(n-p)/((n-p+3)(1+v)Y^tQY) \cdot [z^\circ z]\ [r^\circ r]^t$

$M3 = (n-p)/(2(n-p+3)(Y^tQY)^2) \cdot [r^\circ r]\ [r^\circ r]^t$

$M4 = (n-p)(n-p+1)/((n-p+3)(1+v)(Y^tQY)) \cdot [r^\circ z]\ [r^\circ z]^t$

where ° denotes elementwise multiplication. Section 3 sketches a proof of this result.

The direction that maximizes the second derivative is the eigenvector corresponding to the largest eigenvalue of M1 + M2 + M3 + M4. Since each summand has rank 1 the sum has at most rank 4. Thus there is only a four dimensional space of weight changes that effect the Kullback-Leibler divergence of the predictive density, at least locally.

2. Example

For a numerical example consider, as does Cook (1986), the Snow Geese data for observer 1 from Weisberg (1985). The data are X=flock size estimated by the observer and Y=flock size determined from a photograph. We believe Y to be the true flock size. We are interested in true flock size Z for flocks which have not been photographed but whose sizes have been estimated as w by the same observer. Figure 1 is a scatterplot of the data.

This is a calibration problem. Aitchison and Dunsmore (1975) show that if

1) the conditional distribution of X_i given Y_i, β and σ^2 is $N(\beta_0+\beta_1 Y_i, \sigma^2)$,
2) the conditional distribution of w given Z, β and σ^2 is $N(\beta_0+\beta_1 Z, \sigma^2)$,
3) the conditional distribution of Z given Y is $St(n-3, \bar{Y}, (1+1/n)\Sigma(Y_i-\bar{Y})^2/(n-3))$ and
4) the prior for β and σ^2 is proportional to $\sigma^{-2}d\beta d\sigma^2$

then the predictive distribution for Z given X, Y and w is St(n-2,a,b) where

$$a = \frac{\bar{Y} + (Z-\bar{X})\cdot\Sigma(X_i-\bar{X})(Y_i-\bar{Y})}{\Sigma(X_i-\bar{X})^2} \quad \text{and}$$

$$b = \frac{RSS\cdot\Sigma(X_i-\bar{X})^2}{(n-2)\cdot\Sigma(Y_i-\bar{Y})^2}\left(1 + \frac{1}{n} + \frac{(Z-\bar{Y})^2}{\Sigma(X_i-\bar{X})^2} \right) \quad \text{and}$$

RSS is the residual sum of squares from the regression of Y on X. Geisser (1985) points out that the Aitchison and Dunsmore result is identical to the predictive distribution for Z given X, Y and w if

1′) the conditional distribution of Y_i given X_i, β and σ^2 is $N(\beta_0+\beta_1X_i,\sigma^2)$,

2′) the conditional distribution of Z given w, β and σ^2 is $N(\beta_0+\beta_1w,\sigma^2)$ and

4′) (=4) the prior for β and σ^2 is proportional to $\sigma^{-2}d\beta d\sigma^2$.

Therefore we can solve the calibration problem as a straightforward linear regression prediction problem by reversing the roles of X and Y.

Let's consider predicting true flock size for three values of estimated flock size, say $w\in\{30,100,300\}$. For each value of w we can find d_{max}, the direction that maximizes C_d. Figure 8.2 is a plot of the coordinates of d_{max} for each value of w as a function of observer count. Each coordinate of d_{max} corresponds to one data case. A large coordinate indicates a case that would cause a large change in the predictive distribution if its weight were changed slightly.

These plots are similar to a plot by Cook of the coordinates of d_{max}

as a function of observer count. Cook treated σ^2 as known and used a discrepancy function that depends only on point estimates of β. The main difference between his plot and our plots is in the value for the point where X=500. In Cook's analysis that point corresponded to the largest coordinate of d_{max} and would have been the most influential under a set of small weight changes. In our analysis the influence of that point depends on the value of the covariate.

Another interesting feature is that for w=30 the biggest change in the discrepancy function comes when the points at X=500 and X=250 get weight changes of the same sign. For w=300 the biggest change comes when those points get weight changes of opposite signs. This effect may arise because for w=300 changing the weights with opposite signs will make a large change in the location of the predictive distribution. For w=30 changing the weights with the same signs will make a large change in the variance of the predictive distribution.

APPENDIX B

3. Computation of Curvature

This appendix gives a rough outline and a few intermediate calculations for proving the result in Section 1. Let r be a scalar and $d=(d_1,\ldots,d_n)^t$ be a unit vector. Define

$$S = \begin{pmatrix} 1 + r\cdot d_1 & & & 0 \\ & \cdot & & \\ & & \cdot & \\ 0 & & & 1 + r\cdot d_n \end{pmatrix}.$$

Under the linear model $Y \sim N(X^t\beta, \sigma^2(S^{-1}))$ with prior $\sigma^{-2} d\beta d\sigma^2$ the predictive distribution for a future observable Z with known covariate w is $St(n-p,\ w^t\hat{\beta},\ (1+v)s^2)$ where

X is $n \times p$

$X^* = S^{1/2}X$

$Y^* = S^{1/2}Y$

$\hat{\beta} = (X^{*t}X^*)^{-1}X^{*t}Y^*$

$v = w^t(X^{*t}X^*)^{-1}w$

$s^2 = Y^{*t}QY^*/(n-p)$

and $Q = I - X^*(X^{*t}X^*)^{-1}X^{*t}$

Let f_w be the predictive distribution of Z given above. We want to compute

$$C_d = \left.\frac{\partial\, I^2(f_0, f_w)}{\partial r^2}\right|_{r=0}$$

where I is defined in Section 1.

Let $A = (1+v)s^2, \quad A_0 = A|_{r=0}$

$B = (z-w^t\hat{\beta})^2, \quad B_0 = B|_{r=0}.$

The first step in computing C_d is to differentiate and evaluate at r=0 inside the integral. The derivatives of terms involving only A_0 and B_0 are 0 because A_0 and B_0 do not depend on r. Terms involving only A can come outside of the integral. Letting ′ denote differentiation with respect to r we get

$$C_d = -\frac{n-p}{2}\frac{AA'' - (A')^2}{A^2}\bigg|_{r=0} + \frac{n-p+1}{2}\int \frac{((n-p)A+B)\,((n-p)A''+B) - ((n-p)A'+B)^2}{((n-p)A+B)^2}\bigg|_{r=0} f_0(z)dz$$

Note that

$$\frac{f_0(z)dz}{((n-p)A+B)^2} = \frac{(n-p+2)(n-p)\ g(z)dz}{(n-p+3)(n-p+1)(1+v_0)^2(Y^tQ_0Y)}$$

where g is the Student $(n-p+4,\ w^t\hat{\beta}_0,\ (n-p)A_0/(n-p+4))$ density and a subscript 0 indicates evaluation at r=0. Multiplying out the numerator of the integrand gives

$$C_d = -\frac{n-p}{2}\frac{AA''-(A')^2}{A^2}\bigg|_{r=0}$$

$$+ \frac{(n-p+2)(n-p)}{2(n-p+3)(1+v_0)^2(Y^tQ_0Y)}\Big[(n-p)^2AA'' + (n-p)\int B''g(z)dz$$

$$+ (n-p)A''\int Bg(z)dz + \int BB''g(z)dz$$

$$- (n-p)^2(A')^2 - 2(n-p)A'\int B'g(z)dz$$

$$- \int (B')^2g(z)dz \Big]\Big|_{r=0}.$$

Next evaluate B and its derivatives.

$\int Bg(z)dz|_{r=0} = \text{var}(g) = (1+v_0)Y^tQ_0Y/(n-p+2)$.

$\int B'g(z)dz = 0$ because the integral is an odd central moment of a symmetric density.

B'' does not involve z and comes outside of the integral. Using $((X^tSX)^{-1})' = -(X^tSX)^{-1}(X^tSX)'(X^tSX)^{-1}$ (Rogers (1980)) and $(X^tSX)' = X^tDX$ where $D = \text{diag}(d_1,\dots,d_n)$ yields

$B''|_{r=0} = 2(w^t(X^tX)^{-1}X^tDQ_0Y)^2$.

$$\int (B')^2g(z)dz|_{r=0} = 4(w^t\hat{\beta}')^2|_{r=0} \cdot \text{var}(g)$$

$$= 4(w^t(X^tX)^{-1}X^tDQ_0Y)^2\ (1+v_0)Y^tQ_0Y/(n-p+2)$$

and hence

$$C_d = (A')^2|_{r=0} \cdot (n-p)^3/(2(n-p+3)(1+v_0)^2(Y^tQ_0Y)^2)$$

$$+ (w^t(X^tX)^{-1}X^tDQ_0Y)^2 \cdot (n-p+1)(n-p)/((n-p+3)(1+v_0)(Y^tQ_0Y)).$$

Evaluating A' at r=0 and substituting back into C_d yields C_d as the sum of four terms.

$$C_d = \frac{n-p}{2(n-p+3)(1+v_0)^2} \cdot (w^t(X^tX)^{-1}X^tDX(X^tX)^{-1}w)^2$$
$$- \frac{n-p}{(n-p+3)(1+v_0)Y^tQ_0Y} \cdot (w^t(X^tX)^{-1}X^tDX(X^tX)^{-1}w)\ (Y^tQ_0DQ_0Y)$$
$$+ \frac{n-p}{2(n-p+3)(Y^tQ_0Y)^2} \cdot (Y^tQ_0DQ_0Y)^2$$
$$+ \frac{(n-p+1)\ (n-p)}{(n-p+3)(1+v_0)(Y^tQ_0Y)} \cdot (w^t(X^tX)^{-1}X^tDQ_0Y)^2$$

Let $e = Q_0Y$, the vector of residuals.

Let $m = X(X^tX)^{-1}w$.

Let $\circ$ denote elementwise multiplication. Then

$$C_d = d^t\ (\ M1 + M2 + M3 + M4\ )\ d \text{ where}$$

$$M1 = \frac{n-p}{2(n-p+3)(1+v_0)^2} \cdot (\ m \circ m\ )(\ m \circ m\ )^t$$

$$M2 = \frac{-(\ n-p\ )}{(n-p+3)(1+v_0)Y^tQ_0Y} \cdot (\ m \circ m\ )(\ e \circ e\ )^t$$

$$M3 = \frac{n-p}{2(n-p+3)(Y^tQ_0Y)^2} \cdot (\ e \circ e\ )(\ e \circ e\ )^t$$

$$M4 = \frac{(n-p+1)\ (n-p)}{(n-p+3)(1+v_0)(Y^tQ_0Y)} \cdot (\ e \circ m\ )(\ e \circ m\ )^t$$

References

Aitchison, J. and Dunsmore, I.R. (1975). Statistical Prediction Analysis, Cambridge University Press, Cambridge.

Cook, R.D. (1986). Assessment of local influence (with discussion). JRSS B 48, 133-169.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. Ann. Math. Statist. 36, 150-159.

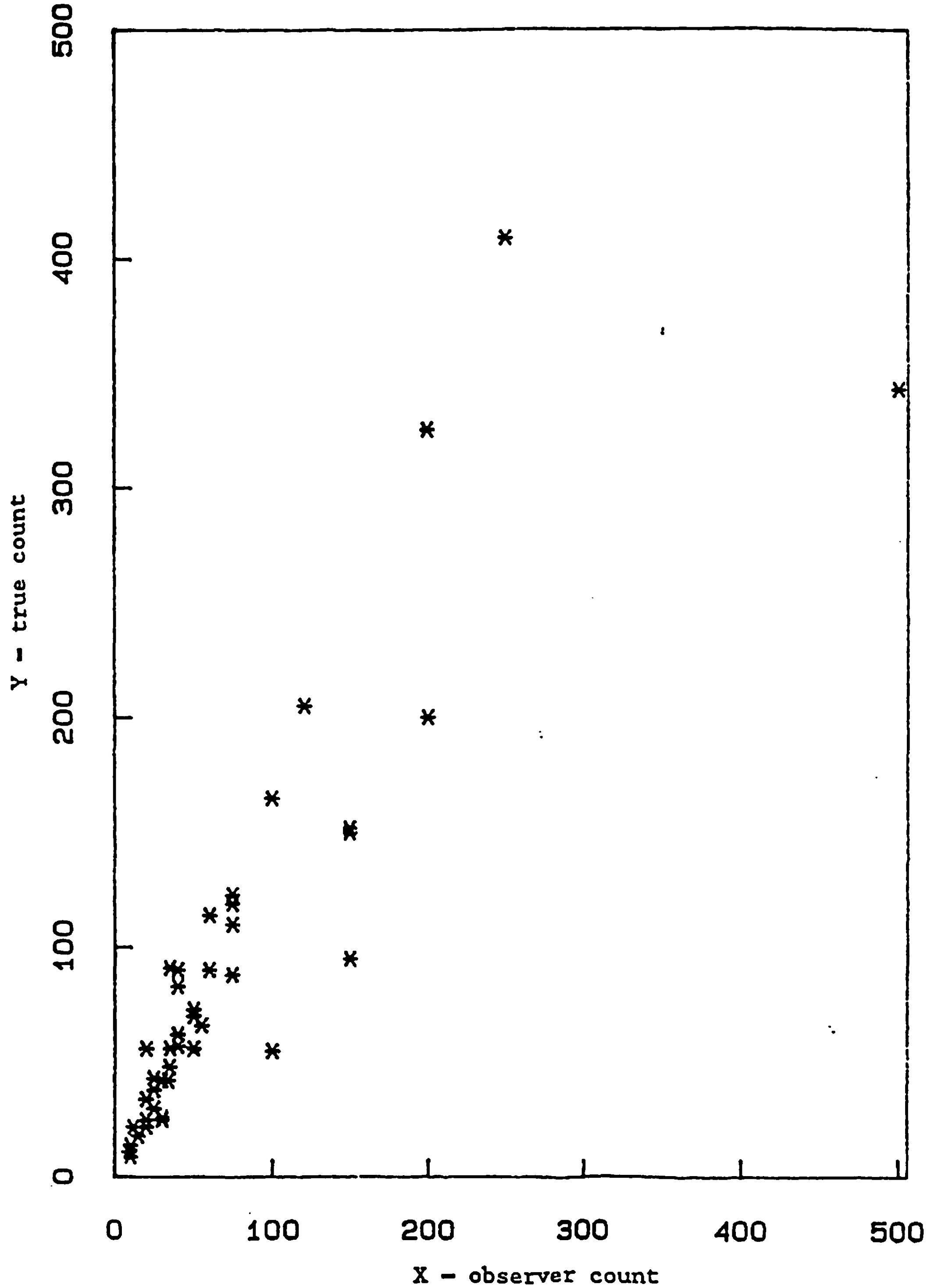
Geisser, S. (1985). Reply to the discussion on On the prediction of observables: a selective update (with disucssion) in Bayesian Statistics 2, Bernardo et al eds., North-Holland, Amsterdam.

Johnson, W. and Geisser, S. (1982). Assessing the predictive influence of observations. In Statistics and Probability Essays in Honor of C.R. Rao, Kallianpur et al eds., North-Holland, Amsterdam.

McCulloch, R. (1986). Local prior influence. University of Minnesota Technical Report No. 477.

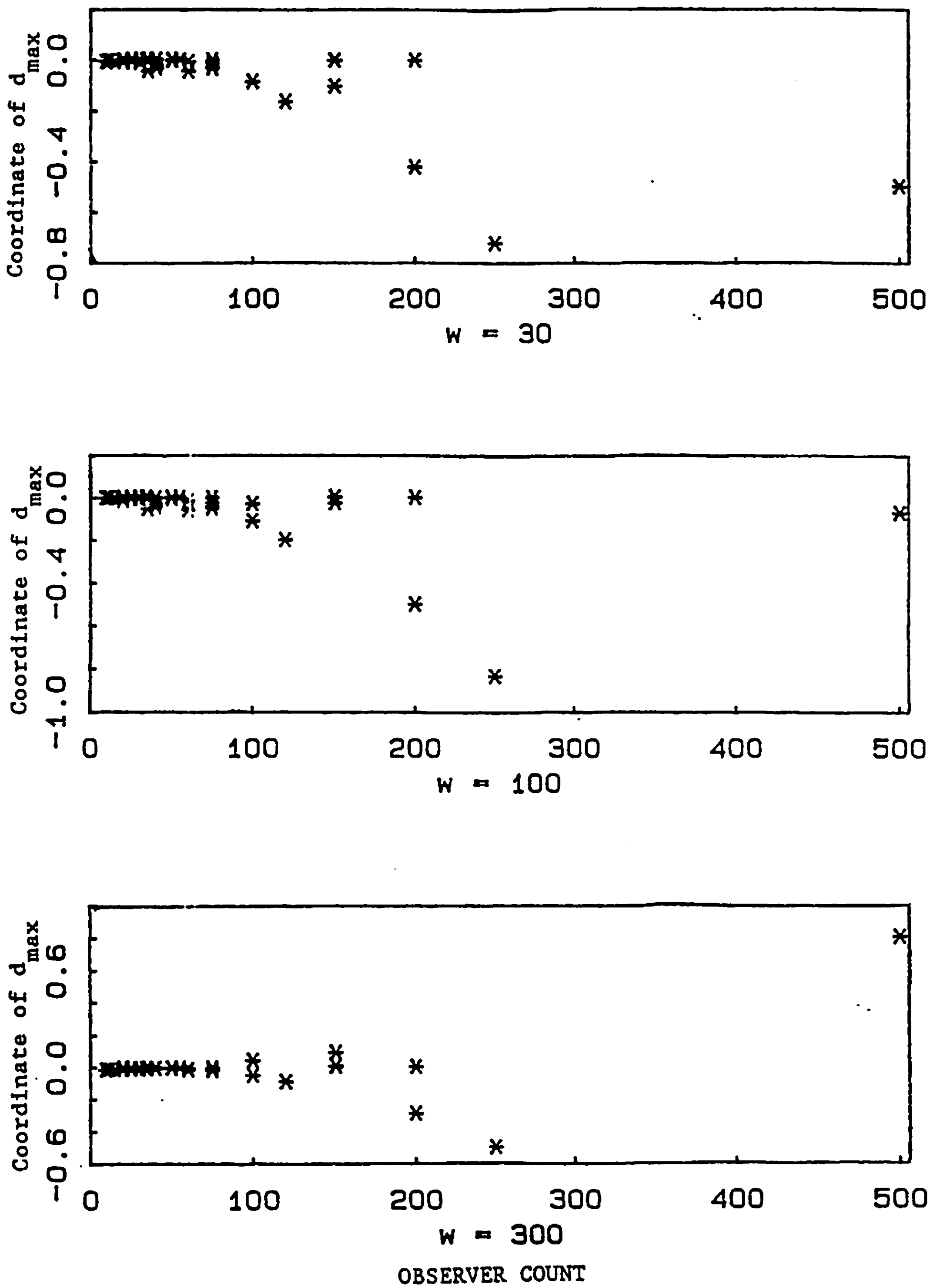
Rogers, G.S. (1980). Matrix Derivatives, Marcel Dekker, New York.

Weisberg, S. (1985). Applied Linear Regression, Wiley, New York.



Snow Geese Data

FIGURE 1



Coordinate of d_{max}
-0.8
-0.4
0.0
0
100
200
300
400
500
W = 30
Coordinate of d_{max}
-1.0
-0.4
0.0
0
100
200
300
400
500
W = 100
Coordinate of d_{max}
-0.6
0.0
0.6
0
100
200
300
400
500
W = 300
OBSERVER COUNT

FIGURE 2