

Prediction Regions for Several Predictions from a Single Regression Line*

GERALD J. LIEBERMAN

Stanford University

When a linear relationship has been fitted by least squares, the methods for securing a prediction interval for the response at some fixed value of the independent variable are explained in many statistical text books. This paper describes the somewhat more complex problem of determining the joint prediction interval for the responses at each of K separate settings of the independent variables when all K predictions must be based upon the original fitted model.

I. INTRODUCTION

Determining the relationship between several variables is a problem which often arises in engineering. In particular, a relationship of the form

$$\text{Average value of } y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

is usually assumed and estimates of $\beta_0, \beta_1, \cdots, \beta_p$ are obtained from sample data. More important, however, these results are used to make statements about future values of y for given values of the X 's. For example, consider the following simple situation. The speed of a missile is a critical factor, which is determined only by measurement after firing. On the other hand, the orifice opening of the valve which admits the fuel is easily obtained by bench tests. Suppose that there exists an underlying relationship between these two variables of the form

$$\text{Average value of speed} = \beta_0 + \beta_1 (\text{orifice opening}).$$

Hence, speed corresponds to y and orifice opening corresponds to X_1 ($p = 1$) in the above linear relationship. N missiles are produced and fired so that N pairs of observations, $(X_{11}, y_1), (X_{12}, y_2), \cdots, (X_{1N}, y_N)$ are taken. Estimates of β_0 and β_1 are obtained, usually by the method of least squares. If a new missile is produced a prediction of its speed based on a knowledge of its orifice opening may be desirable. The usual techniques can be used to obtain a prediction interval such that the speed of the missile, having a given orifice opening X^* , will lie in the interval with preassigned probability, say 0.95. Expressions for such an interval can be obtained from any standard text.¹

* Work done under the sponsorship of the Office of Naval Research under Contract N6 onr-25126.

¹ For example, see Chapter 9 of *Engineering Statistics* by Bowker and Lieberman, Prentice Hall, 1959.

A more difficult problem arises when K additional missiles are produced (rather than one), and a prediction region is desired for the K speeds given the K orifice openings and based on the original N pairs of observations. Using the results for the one missile K times is incorrect since the prediction intervals are *not* independent. They are all based upon the same original N observations. It is the purpose of this paper to give three methods for obtaining a prediction region for K future observations $(y_1^*, y_2^*, \dots, y_K^*)$ based upon the same estimated linear regression given the values of the independent variables

$$(X_{11}^*, X_{21}^*, \dots, X_{p1}^*; X_{12}^*, X_{22}^*, \dots, X_{p2}^*; \dots; X_{1K}^*, X_{2K}^*, \dots, X_{pK}^*).$$

II. AN EXACT PREDICTION REGION

(1) Let y_α , $\alpha = 1, \dots, N$ be independent normally distributed random variables with

$$E(y_\alpha) = \beta_0 + \beta_1 X_{1\alpha} + \dots + \beta_p X_{p\alpha}$$

and

$$\sigma^2(y_\alpha) = E[y_\alpha - E(y_\alpha)]^2 = \sigma^2 \quad \text{for all } \alpha.$$

Define

$$a_{ij} = \sum_{\alpha=1}^N X_{i\alpha} X_{j\alpha} - N \bar{X}_i \bar{X}_j; \quad A = (a_{ij}).$$

Let b_0, b_1, \dots, b_p be the least square estimates of $\beta_0, \beta_1, \beta_2, \dots, \beta_p$, respectively. Let

$$X_r^* = (X_{r1}^*, X_{r2}^*, \dots, X_{rp}^*), \quad r = 1, 2, \dots, K$$

be a fixed value of the vector

$$X = (X_1, X_2, \dots, X_p),$$

and

$$\hat{y}_r^* = b_0 + b_1 X_{r1}^* + b_2 X_{r2}^* + \dots + b_p X_{rp}^*.$$

Then the covariance of \hat{y}_r^* and \hat{y}_s^* is given by

$$(2) \quad \sigma_{\hat{y}_r^*, \hat{y}_s^*} = \sigma^2 \left[\frac{1}{N} + \sum_{i=1}^p \sum_{j=1}^p a^{ij} (X_{ri}^* - \bar{X}_i)(X_{sj}^* - \bar{X}_j) \right]$$

where a^{ij} is the element in the i th row and j th column in the inverse matrix of A .

Denote by y_r^* the future r th observation on y corresponding to the vector X_r^* . The quantities $y_r^* - \hat{y}_r^*$ are normally distributed random variables with zero expectation and covariance

$$(3) \quad \begin{aligned} E(y_r^* - \hat{y}_r^*)(y_s^* - \hat{y}_s^*) &= E[\hat{y}_r^* - E(y_r^*)][\hat{y}_s^* - E(y_s^*)], \quad r \neq s \\ E(y_r^* - \hat{y}_r^*)^2 &= E[y_r^* - E(y_r^*)]^2 + E[\hat{y}_r^* - E(y_r^*)]^2, \quad r = s. \end{aligned}$$

Let $\Sigma = (\sigma_{rs})$ be the matrix of covariance and variances for the random vari-

ables $(y^* - \hat{y}^*)$. From (1), (2) and (3) it follows that

$$(4) \quad \begin{aligned} \sigma_{rr} &= \sigma^2 \left[1 + \frac{1}{N} + \sum_{i=1}^p \sum_{j=1}^p a^{ij} (X_{ri}^* - \bar{X}_i)(X_{rj}^* - \bar{X}_j) \right] \\ \sigma_{rs} &= \sigma^2 \left[\frac{1}{N} + \sum_{i=1}^p \sum_{j=1}^p a^{ij} (X_{ri}^* - \bar{X}_i)(X_{sj}^* - \bar{X}_j) \right] \quad \text{for } r \neq s. \end{aligned}$$

Thus $(y^* - \hat{y}^*)$ has a multivariate normal distribution with zero mean and covariance matrix Σ . Let

$$Q = \sum_{r=1}^K \sum_{s=1}^K \sigma^{rs} (y_r^* - \hat{y}_r^*)(y_s^* - \hat{y}_s^*)$$

where σ^{rs} is the element in the r th row and s th column in the inverse matrix of Σ . Then Q has a χ^2 distribution with K degrees of freedom. If we define

$$s_{v|X}^2 = \sum_{\alpha=1}^N [y_\alpha - (b_0 + b_1 X_{1\alpha} + \cdots + b_p X_{p\alpha})]^2 / (N - p - 1)$$

then $(N - p - 1) s_{v|X}^2 / \sigma^2$ is independent of Q and has a χ^2 distribution with $N - p - 1$ degrees of freedom. It follows, therefore, that the quantity

$$U = \frac{\sum_{r=1}^K \sum_{s=1}^K \sigma^{rs} (y_r^* - \hat{y}_r^*)(y_s^* - \hat{y}_s^*)}{K s_{v|X}^2 / \sigma^2}$$

has an F distribution with K and $N - p - 1$ degrees of freedom and is independent of σ^2 . Thus for any level of significance α the equation

$$U \leq F_{\alpha; K, N-p-1}$$

gives a prediction region for $y^* = (y_1^*, \cdots, y_K^*)$, where $F_{\alpha; K, N-p-1}$ is the α percentage point of the F distribution having K and $N - p - 1$ degrees of freedom. Whether a particular set of values $y_1^*, y_2^*, \cdots, y_K^*$ falls within the prediction region is easily ascertained by substituting these values into the expression for U and determining whether the result is less than or equal to $F_{\alpha; K, N-p-1}$.

III. APPROXIMATE PREDICTION REGIONS

The prediction region given in the previous section is clearly an ellipsoid. Although this ellipsoid yields an exact prediction region, its interpretation may be difficult. Often, one is interested in prediction regions which are intervals for each future observation. This section describes two such regions. Both of these regions will be approximate in that the probability will be *at least* $(1 - \alpha)$ that they contain the future observations $y_1^*, y_2^*, \cdots, y_K^*$.

Following the notation given in Section II, an exact prediction interval for $K = 1$ is given by

$$y_1^* \pm (t_{\alpha/2; N-p-1}) s_{v|X} \sqrt{1 + \frac{1}{N} + \sum_{i=1}^p \sum_{j=1}^p a^{ij} (X_{1i}^* - \bar{X}_i)(X_{1j}^* - \bar{X}_j)}$$

where $t_{\alpha/2; N-p-1}$ is the $\alpha/2$ percentage point of the t distribution having $N - p - 1$ degrees of freedom. Suppose there are now K future observations and the above prediction interval is used for each (varying the value of the independent variables each time of course) replacing $t_{\alpha/2; N-p-1}$ by $t_{\alpha/2K; N-p-1}$. The probability that *all* of the K future observations fall into their respective intervals is *at least* $(1 - \alpha)$. This is easily seen by taking the special case of $K = 2$. Let A be the event that the first future observation falls outside of its prediction interval, and let B be the event that the second future observation falls outside of its prediction interval. Then A or B is the failure of the pair of intervals to simultaneously bracket the two future observations. Choose $P(A) = P(B) = \alpha/2$. Hence

$$P(A \text{ or } B) = P(A) + P(B) - P(AB) \leq P(A) + P(B) = \alpha.$$

It then follows that the probability that both future observations simultaneously fall into their prediction intervals is at least $1 - \alpha$.

The second prediction region is obtained by circumscribing a rectangular region about the ellipsoid. Since this rectangular region includes the ellipsoid, the probability will be at least $(1 - \alpha)$ that it contains the future observations $y_1^*, y_2^*, \dots, y_K^*$. The values of the end points of the intervals which describe the rectangular region are easily obtained. Scheffé² has pointed out that an ellipsoid can be generated by tracing out all of the supporting hyperplanes. In particular, the region of interest can be described by a small subset of these hyperplanes; namely, those which occur at the maximum and minimum of the ellipsoid. These supporting hyperplanes lead to the following intervals for the future observations

$$\hat{y}_i^* \pm s_{y_i|x} \sqrt{KF_{\alpha; K, N-p-1} \sigma_{ii} / \sigma^2}; \quad i = 1, 2, \dots, K.$$

This set of intervals generates the rectangular region which circumscribes the ellipsoid.

IV. EXAMPLE

Consider the example mentioned earlier about predicting the speed of a missile from a measurement of its nozzle opening. Suppose 5 pair of observations are taken and are given below

speed (miles/hr)	5,940	5,550	5,070	4,520	4,360
opening (inches)	1.40	1.36	1.34	1.32	1.31.

If two more missiles are produced having nozzle openings of 1.37 and 1.33 inches respectively, 90% prediction regions using the three methods given in this paper are as follows:

Exact Method.

The least squares estimates of β_0 and β_1 are given by

² For example, see *The Analysis of Variance* by H. Scheffé, John Wiley, 1959.

$$b_1 = \frac{\sum_{i=1}^5 (X_i - \bar{X})(y_i - \bar{y})}{\sum_{i=1}^5 (X_i - \bar{X})^2} = 18273.44$$

and

$$b_0 = \bar{y} - b_1\bar{X} = -19508.05$$

so that the equation of the estimated line is

$$\hat{y} = -19508.05 + 18273.44X.$$

Furthermore

$$s_{y|X} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n - 2}} = 165.95.$$

The variance and covariance matrix Σ is given by

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{21} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}$$

where

$$\sigma_{11} = \sigma^2 \left[1 + \frac{1}{N} + \frac{(X_1^* - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] = 1.31\sigma^2$$

$$\sigma_{12} = \sigma_{21} = \sigma^2 \left[\frac{1}{N} + \frac{(X_1^* - \bar{X})(X_2^* - \bar{X})}{\sum (X_i - \bar{X})^2} \right] = .275\sigma^2$$

and

$$\sigma_{22} = \sigma^2 \left[1 + \frac{1}{N} + \frac{(X_2^* - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right] = 1.25\sigma^2.$$

Hence

$$\Sigma^{-1} = \frac{1}{\sigma^2} \begin{bmatrix} 0.799 & -0.176 \\ -0.176 & 0.839 \end{bmatrix}$$

so that

$$\sigma^{11} = \frac{0.799}{\sigma^2}, \quad \sigma^{12} = \sigma^{21} = \frac{-0.176}{\sigma^2}, \quad \sigma^{22} = \frac{0.839}{\sigma^2}.$$

The prediction region is then given by

$$U = \frac{\sigma^{11}(y_1^* - \hat{y}_1^*)^2 + 2\sigma^{12}(y_1^* - \hat{y}_1^*)(y_2^* - \hat{y}_2^*) + \sigma^{22}(y_2^* - \hat{y}_2^*)^2}{2s_{y|X}^2/\sigma^2} \leq F_{0.10;2,3} = 5.46$$

or

$$U = 0.0000145y_1^{*2} + 0.0000152y_2^{*2} - 0.00000638y_1^*y_2^* - 0.130y_1^* - 0.111y_2^* + 623.996 \leq 5.46.$$

This ellipse is plotted in Figure 1.

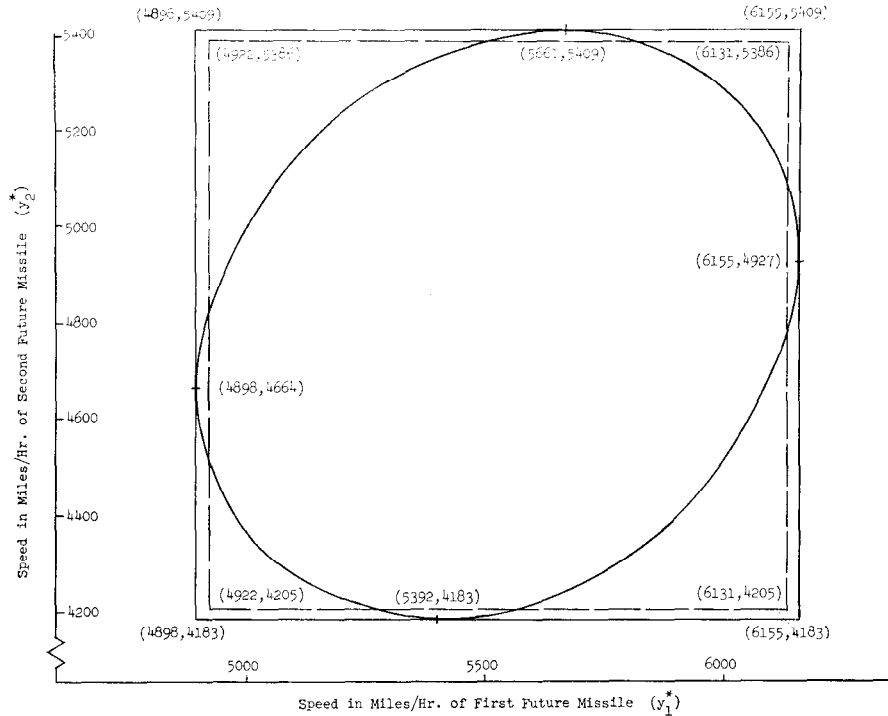


FIGURE 1—Prediction regions for two future missiles having nozzle openings of 1.37 and 1.33 inches respectively.

Approximate Methods.

For the first approximate method the prediction interval for y_1^* is given by

$$\hat{y}_1^* \pm (t_{0.025,3})s_{y_1|x} \sqrt{1 + \frac{1}{5} + \frac{(X_1^* - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

or

$$[4922, 6131].$$

The prediction interval for y_2^* is given by

$$\hat{y}_2^* \pm (t_{0.025,3})s_{y_2|x} \sqrt{1 + \frac{1}{5} + \frac{(X_2^* - \bar{X})^2}{\sum (X_i - \bar{X})^2}}$$

or

$$[4205, 5386].$$

The probability that both future observations fall within their respective intervals is greater than 0.90. This region is shown by the dashed inner rectangle in Figure 1.

For the second approximate method (prediction region which circumscribes the ellipsoid), the prediction interval for y_1^* is given by

$$\hat{y}_1^* \pm s_{y|X} \sqrt{2F_{0.10;2,3} \sigma_{11} / \sigma^2} = 5526.56 \pm 165.95 \sqrt{(2)(5.46)(1.31)}$$

or

$$[4898, 6155].$$

The prediction interval for y_2^* is given by

$$\hat{y}_2^* \pm s_{y|X} \sqrt{2F_{0.10;2,3} \sigma_{22} / \sigma^2} = 4795.63 \pm 165.95 \sqrt{(2)(5.46)(1.25)}$$

or

$$[4183, 5409].$$

The probability that both future observations fall within their respective intervals is greater than 0.90. This region is shown by the solid outer rectangle in Figure 1.

V. REMARKS

It is rather surprising that the exact prediction region given by the ellipse does not lie totally within the rectangle obtained from using the first approximate method. This approximation appears to be crude. N is small (five) in this example so that the prediction for the future observations is strongly dependent upon the estimate of the initial regression line. If N were large, this approximate method would be expected to give good results since the regression line would be "essentially" known. If the regression line is known exactly, this approximate method becomes exact.

Only in the situation where the ellipsoid lies totally within the rectangle obtained from the first approximate method is the second approximate procedure desirable. This procedure circumscribes the ellipsoid and is easy to use. In fact, both approximate procedures have the advantage of ease of computation.

VI. ACKNOWLEDGEMENT

The author is indebted to the late Professor M. A. Girshick and Professor Rupert Miller for their helpful discussions.