Environment Canada Imaging Cover Page

Report N.: 

SKP Box Number: 672672428

ATMOSPHERIC ENVIRONMENT SERVICE
DEPARTMENT OF THE ENVIRONMENT – CANADA

Technical Memoranda

EVENT PROBABILITIES ASSOCIATED WITH REGRESSION ESTIMATES OF PREDICTAND VALUES

by

R. ROBINSON

U.D.C. 551.509.313
CN ISSN 0068-7804

TEC 788
7 JUNE 1973

18/10/77

ENVIRONMENT CANADA - ATMOSPHERIC ENVIRONMENT SERVICE
4905 Dufferin Street
Downsview, Ontario

EVENT PROBABILITIES ASSOCIATED WITH REGRESSION ESTIMATES OF PREDICTAND VALUES

by

R. Robinson

ABSTRACT

Forecasts of predictand values based on linear regression equations do not necessarily meet requirements of individual consumers. But such forecasts carry implications which permit their modification and/or interpretation in terms of probability, and which consequently may enhance their value.

These implications, and the possible use of an inflation factor (or constant multiplier) for modification, are explored.

PROBABILITÉ D'ÉVENEMENTS ASSOCIÉE AVEC LA PRÉDICTION D'UNE VARIABLE AU MOYEN DE RÉGRESSION LINÉAIRE

par

R. Robinson

RÉSUMÉ

Les prévisions des valeurs de prédiction basées sur les équations de régression linéaire ne rencontrent pas nécessairement les exigences des consommateurs individuels. Mais de telles prévisions comportent des implications qui permettent leur modification et/ou interprétation en termes de probabilité et qui, conséquemment, peuvent rehausser leur valuer.

Ces implications et l'utilisation possible d'un facteur d'inflation (ou multiplicateur constant) à fin de modification sont explorées.

EVENT PROBABILITIES ASSOCIATED WITH REGRESSION ESTIMATES OF PREDICTAND VALUES

by

R. Robinson

(Manuscript received December 6, 1973)

1. Introduction

Regression equations linking a variable Y with one or more other variables X_i ($i = 1, 2, \ldots$) are often used to define a predictand-predictor relationship with which observed values of X_i are used to provide forecast values of Y. As a consequence of this procedure, the consumer receives forecasts which will have the property that the expected value of the squared error is minimized. If this suits his purpose, the consumer will obtain optimum results from the forecasts by using predicted values of Y without modification, and may be given an estimate of average forecast worth in terms of the mean square error computed as a function of the correlation coefficient between Y and X_i.

For many consumers, however, losses stemming from incorrect forecasts are not at all proportional to the square of the error, and the maximization of the benefits obtained from regression-based predictions may require either systematic modification of the forecasts and/or interpretation of forecast skill which goes beyond the estimate of mean square error.

One such consumer, for example, may be subject to a loss L whenever a certain threshold value of the predictand Y is exceeded, unless prior action at cost C has been taken. If action is based on forecast values of Y, even large errors in the forecast carry no penalty provided that forecast and observed values are both above or both below the threshold. On the other hand, even small errors will be costly if forecast and observed values are on opposite sides of the threshold.

If, for a given forecast, P represents the probability of the threshold being subsequently exceeded, the action should be taken if $P > C/L$, and to achieve optimum results from the forecast service, this consumer must be given either:

i) an amended version of the predictions in which the threshold value will be exceeded when and only when $P > C/L$, or

ii) an estimate of P.

Advance estimates of the worth of the forecast service for this application, whether or not the regression predictions are modified, will involve the expected frequency of occurrence of values above the threshold, and the conditional probability that observation of above the threshold will be "correctly" forecast.

In the following sections, conditional probabilities of the occurrence of forecast and observed values in specified ranges are developed as functions of the correlation coefficient linking the predictand Y with the set of predictors X_i. Multiplication of regression-based forecasts by a constant, or "inflation factor" is examined as one possible method of providing forecasts to better serve specific consumers, and its impact on mean square error and probability distributions noted. Other procedures, such as application of a discriminant function for categorical prediction, are not considered in this paper.

2. Linear Regression

For any variables Y, X_i ($i = 1, 2, \ldots$), a linear correlation coefficient R may be defined between Y and X_i. For the sake of simplicity of the following derivation, only one independent variable X is considered, and both it and the predictand Y are assumed to be normalized and standardized, i.e. with means $\overline{X} = \overline{Y} = 0$ and standard deviations $\sigma X = \sigma Y = 1$. The generality of the results is not affected.

If Y*(X) is the value of Y obtained from the regression of Y on X, then:

$$Y^*(X) = R_{XY}.X, \text{ where } R_{XY} = \overline{XY} \qquad (1)$$

By substitution, $\overline{Y}^*(X) = 0$, and $\sigma^2 Y^*(X) = R^2_{XY}$. The inclusion of an inflation factor 'm' as an arbitrary method of altering the characteristics of Y*(X) may be done by redefining the predicted value as

$$Y^* = m.R_{XY}.X \qquad (2)$$

If error is defined as $E = Y^* - Y$, then

$$\overline{E} = 0$$

and $$\sigma^2 E = E^{\overline{2}} = 1 - R^2_{XY}(2m - m^2) \qquad \text{(Fig. 10)}$$

Regression of E on Y or Y* now provides a predicted error E* as a function of either the observed predictand Y or the forecast predictand Y*.

$$E^*(Y) = (mR^2_{XY} - 1)Y \tag{3}$$

$$E^*(Y^*) = \frac{m - 1}{m} \cdot Y^* \tag{4}$$

And as the difference between predicted and observed values of the predictand defined E, so the differences between E and E*(Y), and between E and E*(Y*), define variables $F_Y = E^*(Y) - E$ and $F_{Y^*} = E^*(Y^*) - E$. By using previous definitions, and averaging over all values of Y and Y*,

$$\bar{F}_Y = \bar{F}_{Y^*} = 0$$

and

$$\sigma^2_{F_Y} = \overline{F^2}_Y = m^2R^2_{XY}(1 - R^2_{XY})$$

$$\sigma^2_{F_{Y^*}} = \overline{F^2}_{Y^*} = 1 - R^2_{XY}$$

3. Event Probabilities

F_Y is uncorrelated to Y and F_{Y^*} uncorrelated to Y*. Provided that the means and variances given above apply not only to the total sample of observations but to the hypothetical subsets of F_Y defined by Y = constant, and to the subsets of F_{Y^*} with Y* = constant, and assuming that the probability density distribution of such subsets follows the normal curve,

$$PD(F) = \frac{1}{\sqrt{2\pi}\ \sigma_F} \cdot e^{\frac{-(F - \bar{F})^2}{2\ \sigma_F{}^2}}$$

the variables E* and F may be used to assess event probabilities, where "event" is defined as the co-occurrence of forecast and observation within specified ranges.

If the transformation $Z = F_Y / \sigma F_Y$ is made, the variable Z will have a probability density distribution given by the standard normal curve, $PD(Z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{Z^2}{2}}$. The probability that F_Y lies between limits L_1 and L_2 is identical to the probability that Z lies between $L_1 / \sigma F_Y$ and $L_2 / \sigma F_Y$. For any event which can be translated into a restriction on F_Y (or F_{Y^*}), tables for the standard normal curve will then yield the associated probability. Three such events are considered below.

Event I

For a large class of consumers, the most useful forecast on any given occasion is not the simple quantity Y* (even if this is the most probable value) but the probability of the subsequent observation lying below (or above) a particular threshold. This probability may be assessed by considering the event: "With a given prediction Y*, the observed value is less than a constant, K". (K, like Y and Y*, is measured in standard deviations of Y.)

The probability of $Y < K$ is equivalent to that of

$$Y^* - E < K,$$

or

$$F_{Y^*} - E^*(Y^*) + Y^* < K,$$

or

$$F_{Y^*} < K - Y^*/m,$$

which in turn is equivalent to the probability that the standard normal distribution has values less than $(K - Y^*/m)/\sqrt{(1 - R^2_{XY})}$. The dependence of this probability on K, Y*, m, and R_{XY} is shown in Figures 1-3, for selected values of Y*.

When m = 1, the probability of a given error is linearly uncorrelated with Y*. In the case of m = 1, and for fixed correlation coefficient, comparison of Figures 1 and 2 thus shows equal probabilities for pairs of Y*, K having identical differences Y* - K. When Y* - K is near zero in absolute value, the probability of observations less (or greater) than K is relatively unaffected by changing correlation coefficient, except as R approaches unity and probabilities change rapidly from near 50% to either zero or one. Even for values of Y* - K as large as one standard deviation, the probability of observation less than K decreases from a value of about 16% with R = 0 by only a few percentage points as R increases to .5, and a value of 1% is not attained until R is over .9.

If m is greater than unity, the systematic error of $\frac{m-1}{m} \cdot Y^*$ increases the probability of errors having the same sign as Y*. In a comparison of Figure 3 with Figure 2, this shows as a shift of the family of probability curves toward higher probability values.

Event II:

A simple measure of forecast success on any one occasion is the error, or difference between forecast and observation. The economic consequence of a given error is not, however, necessarily (or even usually) a constant over the whole range of possible

forecasts and/or observations. Errors which may be tolerated when the observed value turns out to be at one level may carry large penalties when they occur with observations at a second level. (c.f. consequences to the public of an unforecasted occurrence of "bad" weather compared to the consequences of an unforecasted occurrence of "good" weather).

Under these conditions, a more pertinent measure of success on a single occasion will be a function not only of error size but also of the observed value Y, and the prior evaluation of the success to be expected from a series of forecasts will require the conditional probabilities of given errors occurring with given values of the observation. Examples of such probabilities may be developed by considering the event: "For a given observation Y, a prediction Y* will have been made such that the absolute error $|E|$ is less than a positive constant K".

The probability of $|E| < K$ is equivalent to that of

$$|E^*(Y) - F_Y| < K, \text{ or}$$

i) if $E^*(Y)$ and F_Y have opposite sign,

$$|F_Y| < K - |E^*(Y)|$$

ii) if $E^*(Y)$ and F_Y have same sign,

$$|F_Y| > |E^*(Y)| - K \quad \text{for} \quad |F_Y| < |E^*(Y)|$$

$$|F_Y| < K + |E^*(Y)| \quad \text{for} \quad |F_Y| > |E^*(Y)|$$

Figures 4-8 indicate the dependence of this event probability on Y, K, m and R_{XY}.

The variation of probability with correlation coefficient reflects both the effect of changing the expected mean error and of altering the scatter about that mean. When R_{XY} is near zero, the forecasts are concentrated in a narrow range about zero, and the error is essentially equivalent to the observed value. Errors will thus always be less, or always greater, than a given K, depending only on wether K is greater or less than the observation Y. As R_{XY} increases, forecasts will be made with higher absolute values, increasing the chance that some forecasts will be wrong for small observed values of the predictand and that some forecasts will be correct for large observed values of the predictand. Reversal in the downward trend for probability of error within a specified limit,

when the observed value is small, occurs as R_{XY} reaches approximately 0.7. As R_{XY} approaches unity (and if m = 1) forecast and observed values become identical, and errors are always less than an arbitrarily small constant.

The effect of a magnification factor less than unity is, broadly speaking, to reduce the probability of any absolute-value error when the observed value is low (in terms of distance from the mean), and to increase it when the observed value is high. For m greater than one, this trend is reversed.

In both cases, the amount of change is dependent on the correlation coefficient. With R_{XY} near unity, error is essentially equal to (m-1) Y and the probability of absolute error less than K is either zero or one depending on whether K is less or greater than (m-1)Y. For $m = 1/R^2_{XY}$, the probability of any specified error is independent of the observed value Y.

Event III

Another possible measurement of success comes from the comparison of forecast error with the error made by a prediction always equal to the mean value of the predictand. The latter error will of course be equal in magnitude to the observed value of the predictand: the "prediction" yielding it may be considered to be no prediction at all, but merely the best value to assume in the absence of a forecast.

For some purposes, it may be sufficient to classify a single forecast as successful if it yields an error smaller in magnitude than that given by this non-forecast, without regard to the size of the difference. For a series of forecasts, success will be measured by the percentage of occasions on which the forecast produces at least a little better result than the assumption of mean value. A prior estimate of this measure may be made by defining an event not limited to specific values of forecast or observation. Thus we consider Event III: "The absolute value of the error of the forecast is less than the absolute value of the observation".

If, for a specified Y, the probability of error less than Y is U(Y), then the probability over the whole range of Y is equal to

$\int_{-\infty}^{\infty} U(Y) \cdot P(Y) \cdot dY$, where P(Y) is the probability density function of Y.

Using values of U(Y) determined similarly to the probabilities discussed earlier, and assuming that Y is normally distributed, a finite difference approximation to the above integral yields the probabilities given in Figure 9 as functions of R_{XY} and m.

The percentage of forecasts better than no forecasts increases slowly from the 50% expected with R_{XY} near zero until correlation coefficients of about 0.7 are reached. Without inflation, a correlation of 0.9 provides improvement about 75% of the time. Deflation (m less than one) increases the percentage of successful forecasts, except at the limiting values $R_{XY} = 0$ and $R_{XY} = 1$.

4. Conclusion

The families of probabilities derived in the preceding sections are not rigourously true for predictand/predictors which have non-normal distributions or which have a non-linear interdependence, or when the assumption of a joint probability distribution of random variables is inappropriate. Nevertheless, these probabilities can be used for the translation of regression-based forecasts into probabilistic terms which may be more useful to some consumers, and a means for estimating the success of the forecasts in meeting objectives other than the minimization of mean square error. They show morevoer that the success of regression-based forecasts in meeting such other objectives can be enhanced by the use of an inflation factor, and that for some consumers the modified forecasts can be used as an alternative to explicit statements of probability.

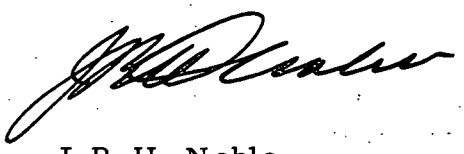
If, for instance, it is essential that very few occurrences of extreme values be unpredicted (i.e. that the ratio of protection cost to potential loss is small) then an inflation factor greater than unity can in general raise the incidence of forecasts at and over the threshold level, and by so doing provide a criterion for consumer action equivalent to the statement that the probability of occurrence is greater than C/L. Conversely, if cost/loss ratios are large, or if importance is attached to the frequency rather than the amount of improvement that the forecasts give relative to a prediction of the mean value, then inflation factors less than unity (deflation, or hedging) will produce appropriate results.

In sum, regression-based forecasts, basically designed to minimize mean square error, can be modified and/or interpreted to suit the requirements of specific consumers. Properly exploited, this ability can lead to an increase in the value of a "customized" forecast service.

Acknowledgements

The diagrams used to illustrate this article were prepared by Mr. Andre Jacques, Extended Forecast Unit, Canadian Meteorological Centre, Montreal.

APPROVED,



J.R.H. Noble,
Assistant Deputy Minister,
Atmospheric Environment Service.

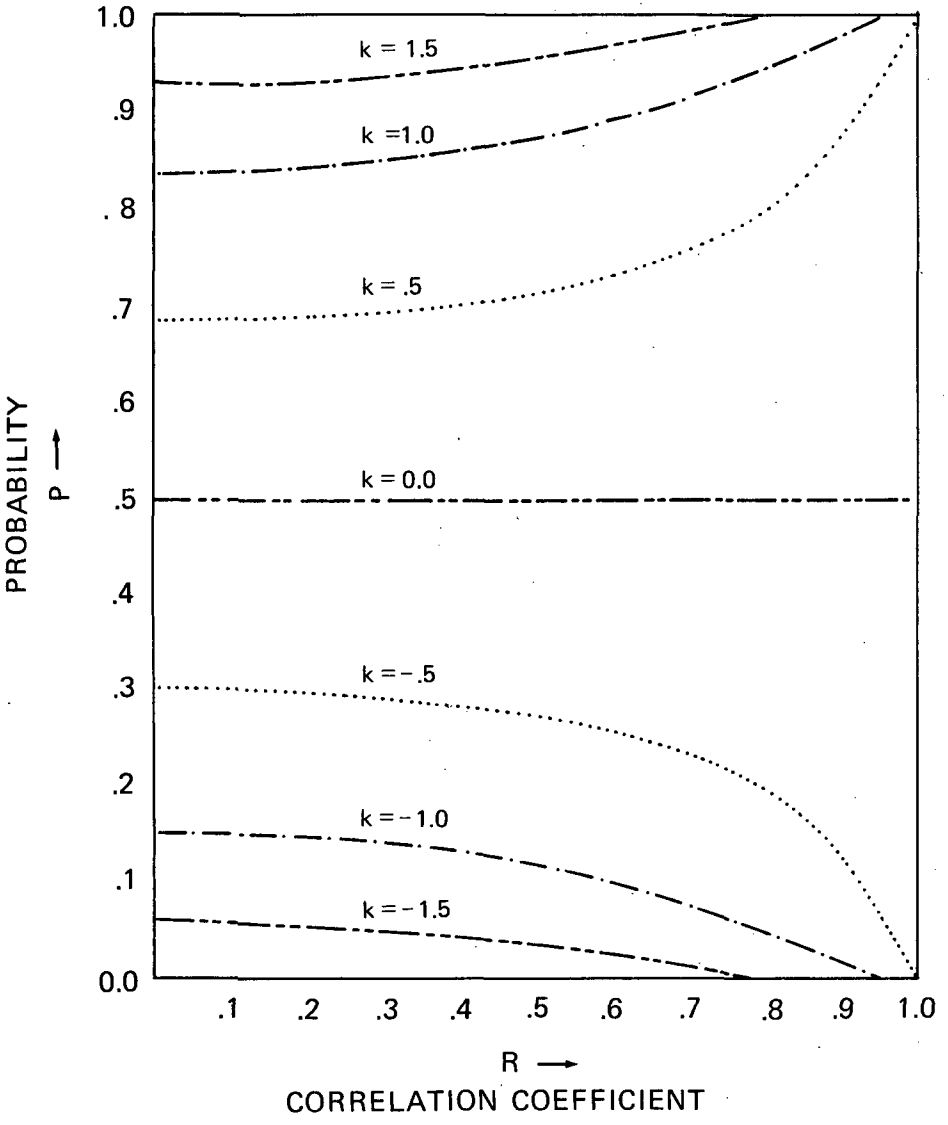


Figure 1.
Probability of Observation (Y) Less than Constant (K)
when Y* = 0 (m = 1)

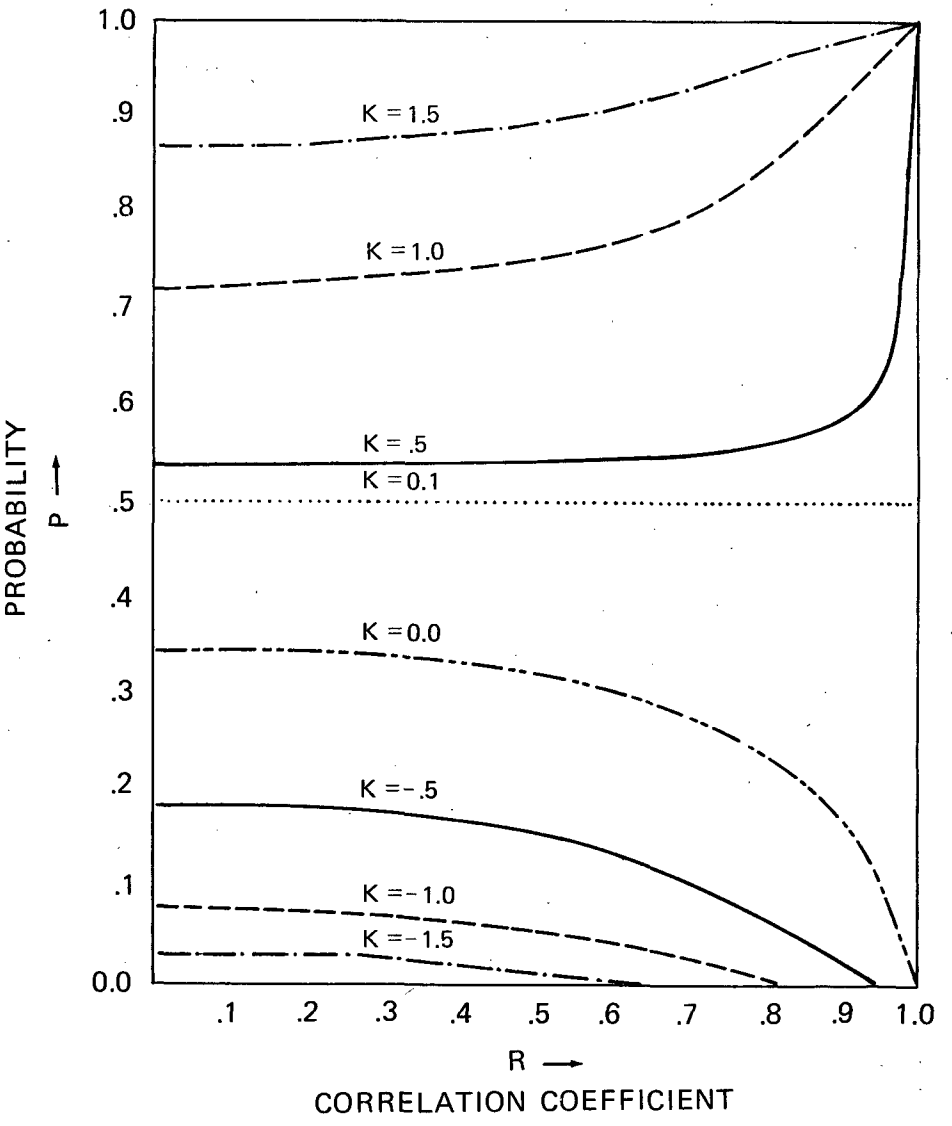


Figure 2.
Probability of Observation (Y) Less than Constant (K)
when Y* = .4 (m = 1)

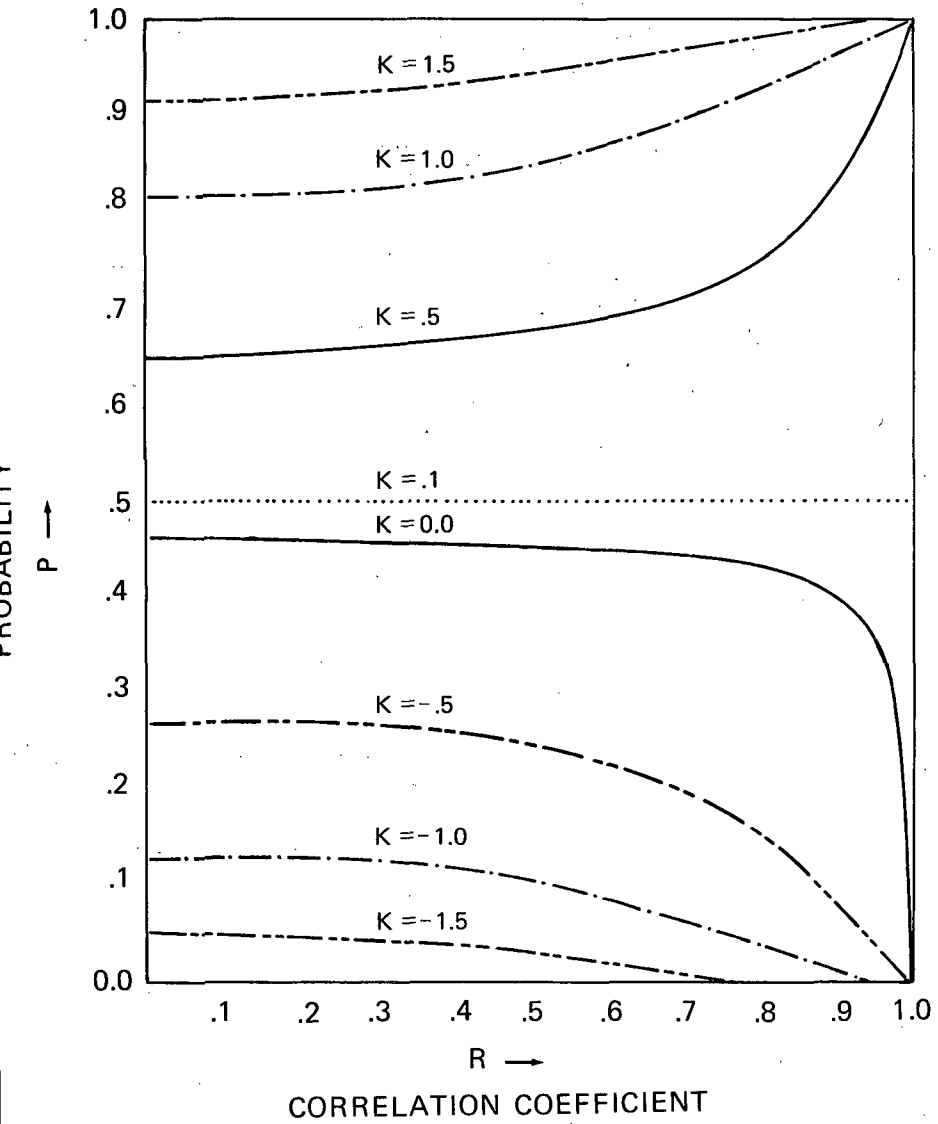


Figure 3.
Probability of Observation (Y) Less than Constant (K) when Y* = .4 (m = 4.0)

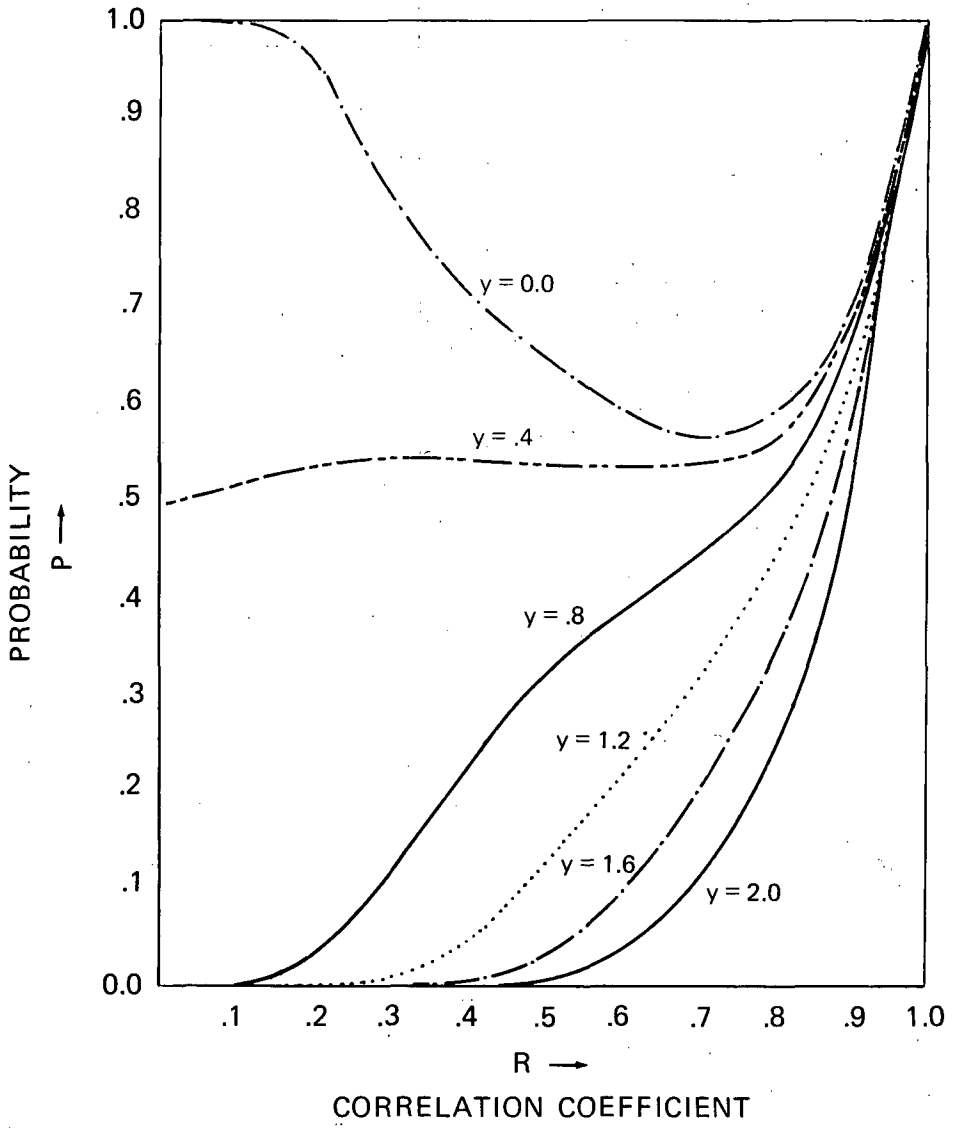


Figure 4.
Probability of Absolute Error Less than .4 for Observation (Y) and Forecast Y* = mRX (m = 1)

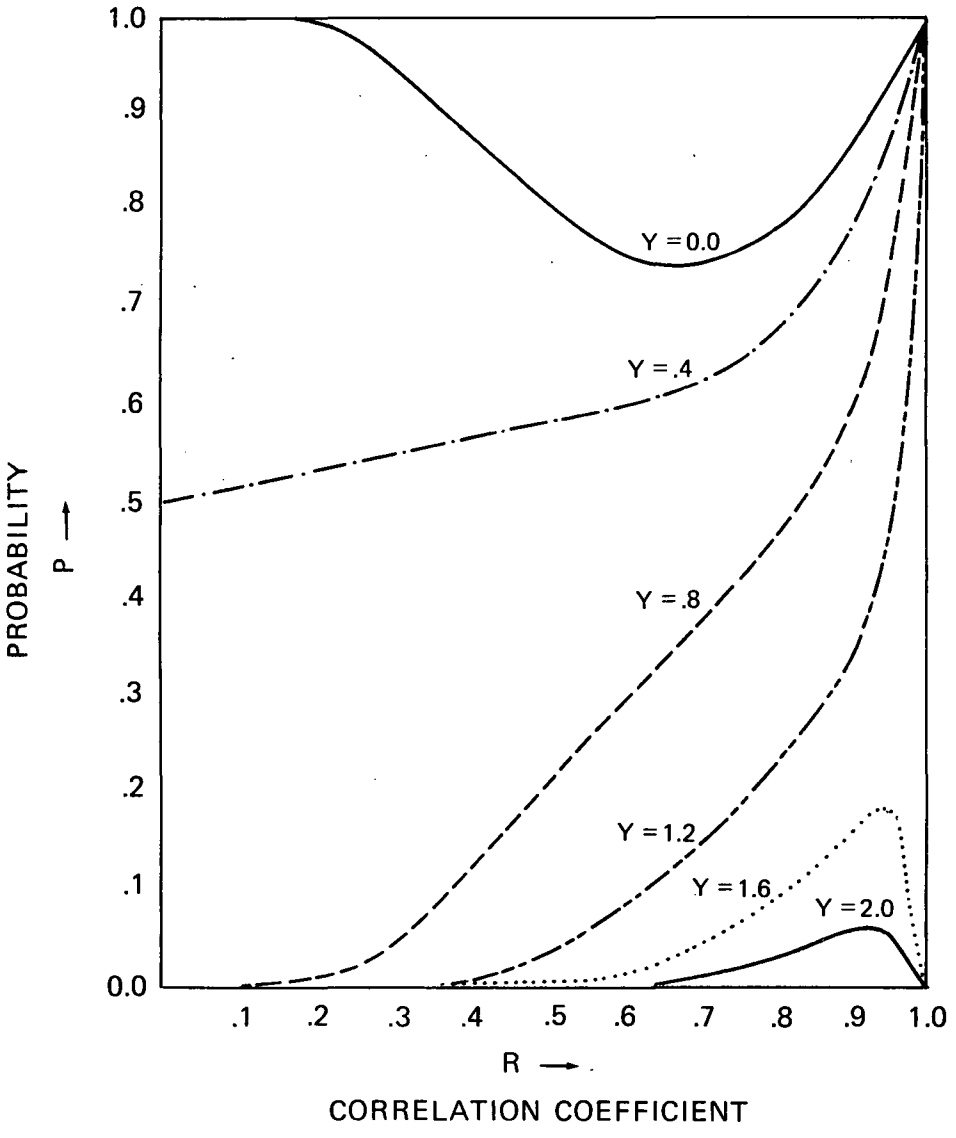


Figure 5.
Probability of Absolute Error Less than .4 for Observation(Y) and Forecast Y* = mRX(m = .7)

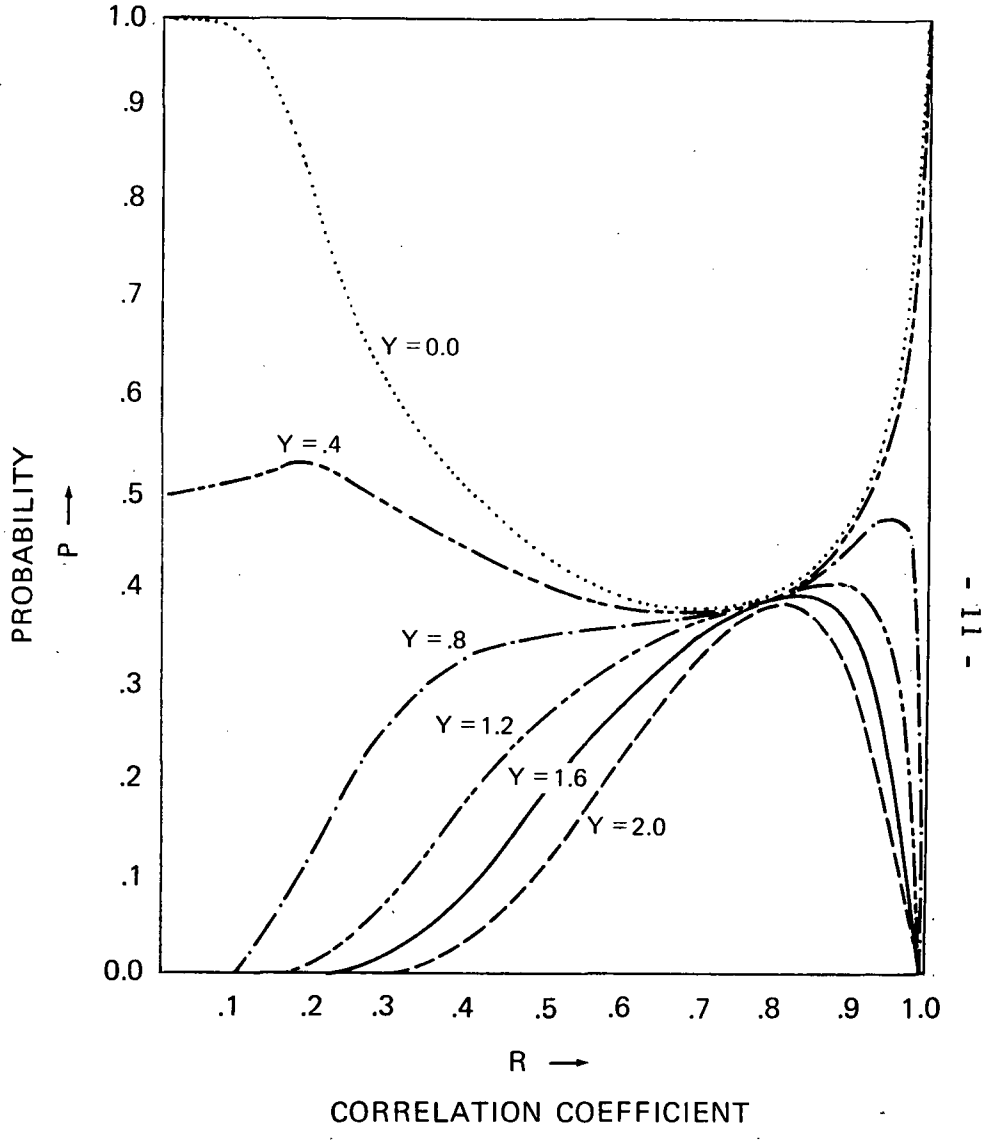


Figure 6.
Probability of Absolute Error Less than .4 for Observation(Y) and Forecast Y* = mRX (m = 1.6)

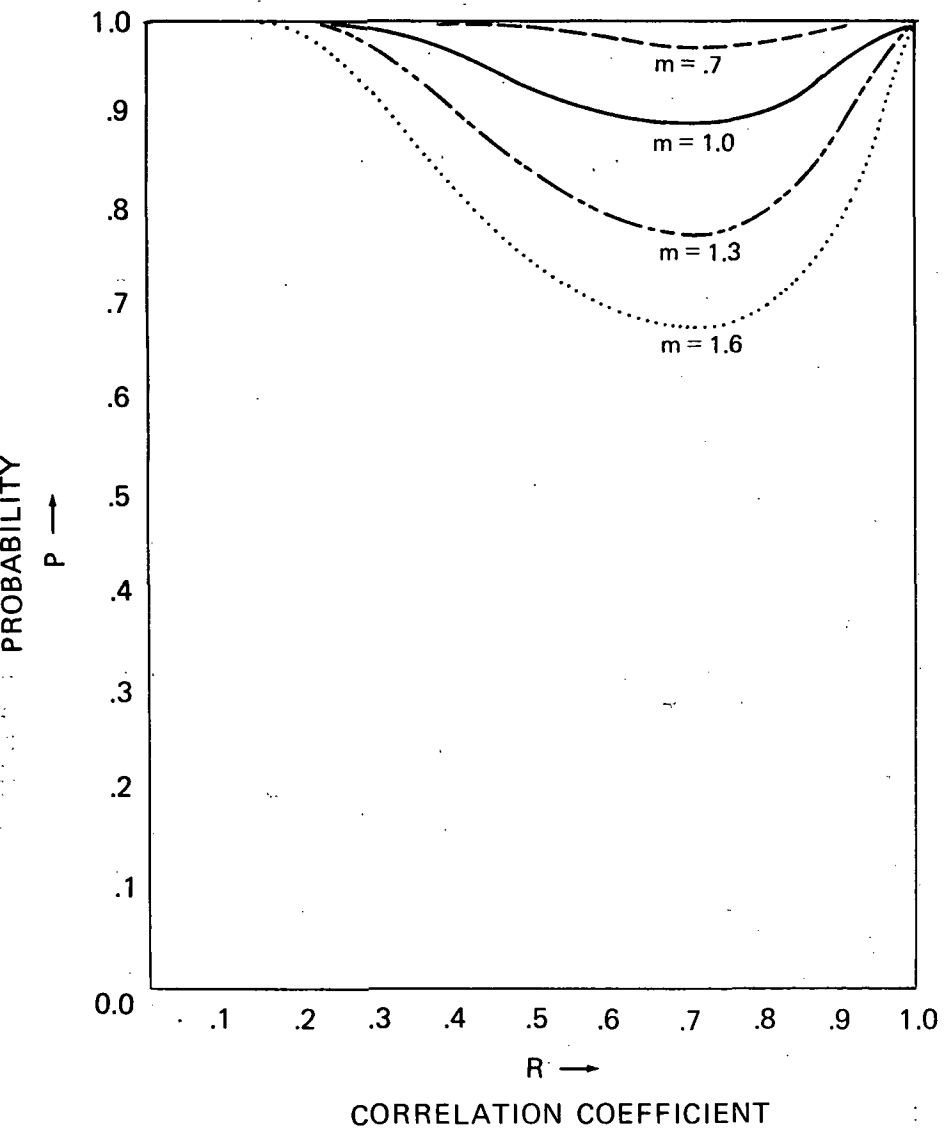


Figure 7.
Probability of Absolute Error Less than .8 for Observation Y = 0 and Forecast Y* = mRX

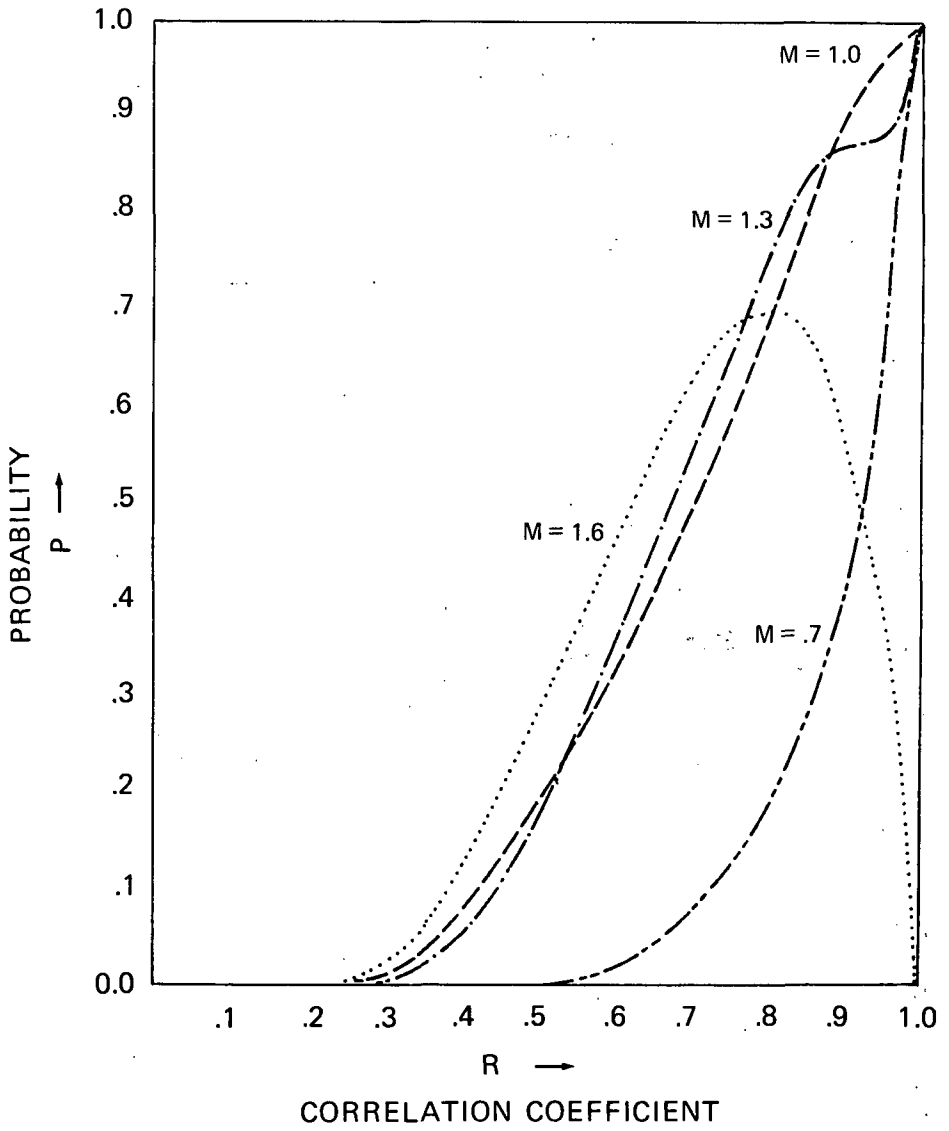


Figure 8.
Probability of Absolute Error Less than .8 for Observation Y = 2.0 and Forecast Y* = mRX

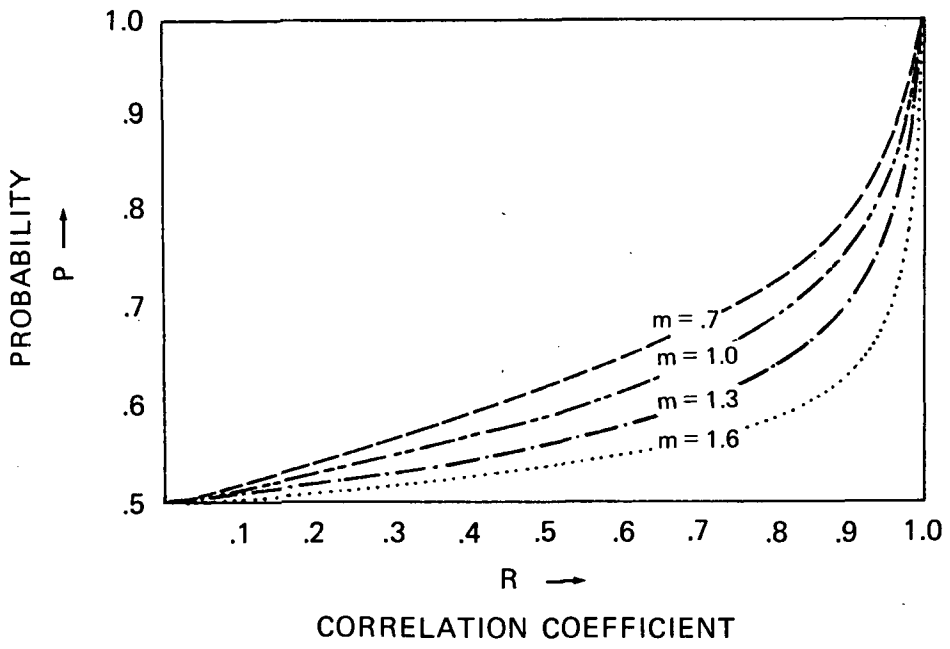


Figure 9.
Probability of Absolute Error Less than the Absolute Value of the Observation (Y) for Predictions Y*=mRX

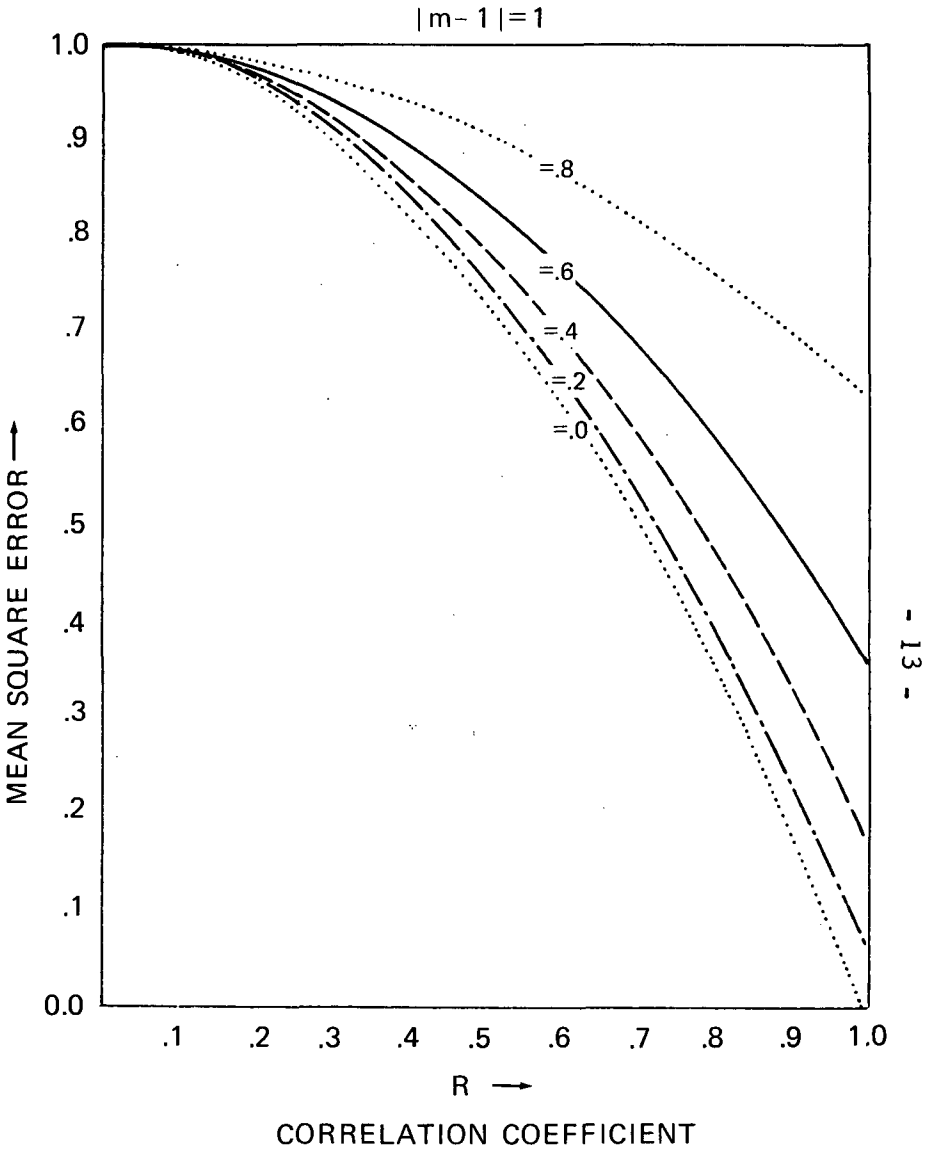


Figure 10.
Mean Square Error as Function of Correlation Coefficient (R) and Inflation Factor (m)

TEC-788 UDC: 551.509.313
7 June 1973

CANADA

Environment - Atmospheric Environment Service
4905 Dufferin Street, Downsview, Ontario

Event Probabilities Associated with Regression Estimates of Predictand Values
by R. Robinson

8 pps. 10 figs.

Subject Reference: 1. Probability Forecasting

TEC-788 UDC: 551.509.313
7 June 1973

CANADA

Environment - Atmospheric Environment Service
4905 Dufferin Street, Downsview, Ontario

Event Probabilities Associated with Regression Estimates of Predictand Values
by R. Robinson

8 pps. 10 figs.

Subject Reference: 1. Probability Forecasting

TEC-788 UDC: 551.509.313
7 June 1973

CANADA

Environment - Atmospheric Environment Service
4905 Dufferin Street, Downsview, Ontario

Event Probabilities Associated with Regression Estimates of Predictand Values
by R. Robinson

8 pps. 10 figs.

Subject Reference: 1. Probability Forecasting

TEC-788 UDC: 551.509.313
7 June 1973

CANADA

Environment - Atmospheric Environment Service
4905 Dufferin Street, Downsview, Ontario

Event Probabilities Associated with Regression Estimates of Predictand Values
by R. Robinson

8 pps. 10 figs.

Subject Reference: 1. Probability Forecasting

ABSTRACT: Forecasts of predictand values based on linear regression equations do not necessarily meet requirements of individual consumers. But such forecasts carry implications which permit their modification and/ or interpretation in terms of probability, and which consequently may enhance their value.
These implications, and the possible use of an inflation factor (or constant multiplier) for modification, are explored.

ABSTRACT: Forecasts of predictand values based on linear regression equations do not necessarily meet requirements of individual consumers. But such forecasts carry implications which permit their modification and/ or interpretation in terms of probability, and which consequently may enhance their value.
These implications, and the possible use of an inflation factor (or constant multiplier) for modification, are explored.

ABSTRACT: Forecasts of predictand values based on linear regression equations do not necessarily meet requirements of individual consumers. But such forecasts carry implications which permit their modification and/ or interpretation in terms of probability, and which consequently may enhance their value.
These implications, and the possible use of an inflation factor (or constant multiplier) for modification, are explored.

ABSTRACT: Forecasts of predictand values based on linear regression equations do not necessarily meet requirements of individual consumers. But such forecasts carry implications which permit their modification and/ or interpretation in terms of probability, and which consequently may enhance their value.
These implications, and the possible use of an inflation factor (or constant multiplier) for modification, are explored.