PREDICTIONS BASED ON REGRESSION ANALYSIS

By: Nooreha Husain

The use of statistical methods is becoming increasingly important in all fields of everday task. Infact it is quite impossible to name an activity which does not employ its own particular statistics as an aid to influencing human behaviour. But the usefulness of any statistical enquiry depends entirely on the competence of those who attempt to interpret the results they obtain.

Tomorrow's high temperature will be between 75°F and 78°F;

The results of ITM students based on their GPA is forecasted to increase between 10% and 20% during the next semester;

About 4 million motorists will take to the Karak Highway during this holiday weekend.

The three sets of statements above refer to something that might happen in the future. it is not an **estimate,** because it has not yet happened. These three statements are **forecasts** or **PREDICTIONS.** Here, we will be concerned about the two questions : How do you prepare predictions? What do they mean?

One of the statistical tools that can be use to make **prediction** is the linear regression analysis. The prediction can be done graphically by reading off from the graph the y-value corresponding to the given x value or by substituting the given x-value into the regression equation of y on x, that reveals the relationship between x and y; and calculating the value of y.

(NOTE: This equation should not be used to predict x for a given value of y. If that kind of prediction is required the regression of x on y must be calculated and used).

If in the case of the y-value being predicted for an x-value within the range of x values in the original data, this is called **interpolation.** The particular x-value used in this procedure is almost equal to the mean of the x-values in the original data and thus interpolation can be said to be a very respectable procedure which will lead to sensible predictions. If appropriate distributional assumptions hold concerning the data and we can proceed to find confidence limits on the regression parameter estimates β and β_1 and on the predictions from the regression, the limits on an interpolated prédition will be tight.

In linear regression analysis, two types of estimates or predictions of values of the dependent variable are made. **The first** type of estimate involves predicting an individual value of the dependent variable Y, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$. For example suppose we want to estimate the Grade-Point Average (GPA) of a particular student, based on a linear regression equation between the dependent variable y, the GPA and the independent variable x, the first exam score.

The second is an estimate of a conditional mean, estimating the mean of the Y population for a specified X. Again the single estimated value is simply $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ but an internal of reasonable values is different than if we are interested in a single Y value. For example, we might be interested in the final mean score of the population of students scoring 70% on the first examination. this is obviously different than being interested in the final score of a single student scoring 70% on the first examination, and the intervals are different. Another problem is to give intervals for the mean of all Y population simultaneously. This is equivalent to specifying a region in which the entire true regression line lies with reasonable assurance.

I. PREDICTION INTERVAL FOR AN INDIVIDUAL VALUE OF Y.

Looking at the example below

Two sets of examination scores for the students were recorded for the month of Feb 1990 and April 1990.

First exam (X) Feb 1990	**Last exam (Y) April 1990**
63	68
65	75
72	70
73	76
80	81
85	78
86	89
93	84

From the data above, we obtain the following calculations:

$\bar{X} = 77.125$ $\quad \hat{\sigma}^2 = 18.3078$

$\bar{Y} = 77.625$

$$\Sigma(x-\bar{x})^2 = \Sigma x^2 - (\Sigma x)^2/n$$
$$= 48{,}377 - 47{,}586.125$$
$$= 790.875$$

$$\Sigma(Y-\bar{Y})^2 = \Sigma Y^2 - (\Sigma Y)^2/n$$
$$= 48{,}547 - 48{,}205.125$$
$$= 341.875$$

$$\Sigma(x-\bar{x})(Y-\bar{Y}) = \Sigma xY - (\Sigma x)(\Sigma Y)/n$$
$$= 48{,}323 - 47{,}894{,}625$$
$$= 428.375$$

Since the least squares estimates of β_0 and β_1 are given by the formulae

$$\hat{\beta}_1 = \frac{\Sigma(x-\bar{x})(Y-\bar{Y})}{\Sigma(x-\bar{x})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{x}$$

The quantities obtained are:

$$\hat{\beta}_1 = \frac{425.375}{790.875} = 0.5416$$

$$\hat{\beta}_0 = 77.625 - (0.5416)(77.125)$$
$$= 35.8541$$

So the regression line is given by the equation

$$Y = 35.8541 + 0.5416X$$

Suppose a lecturer has been teaching large lecture sections for many years, and using standardized examinations, has obtained a regression line relating course averages (Y) to the first exam scores, for example a student scoring 70 on the first exam and is thinking of dropping the course. The lecturer can use the regression line and the first score of the student to obtain a prediction interval for the students' course average score.

The end point of a 95% confidence interval are given by:

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{tab}\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma (X - \bar{X})^2}} \qquad (1)$$

Where X_0 = X value of interest (70 in this case)

t_{tab} has n-2 degrees of freedom, and other quantities are calculated from the regression data.

$$\sqrt{1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma (X - \bar{X})^2}} = 1.0905$$

Thus

$$35.8540 + (70)(0.5416) \pm (2.4469)\,(4.2788)\,(1.0905)$$

$$73.766 \pm 11.4173$$

63.3488 and 85.1834

Thus the 95% prediction interval for the student is from 62.3488 to 85.1834. In a real situation such a prediction interval could be of value in counseling a student.

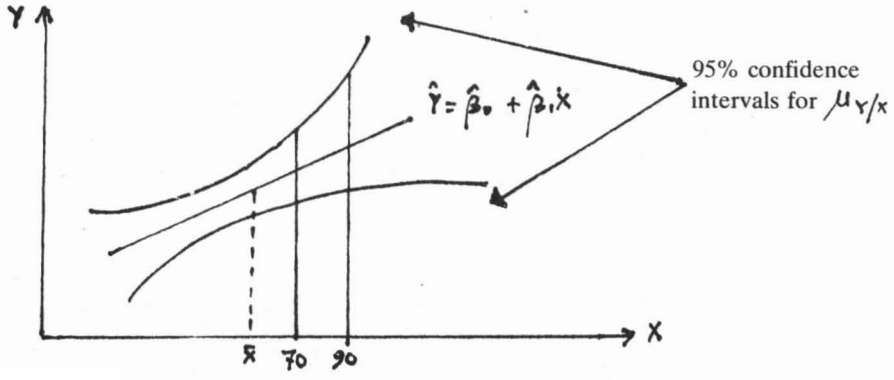

FIG. 1. The 95% confidence-interval band

II. CONFIDENCE INTERVAL ESTIMATE OF A CONDITIONAL MEAN

It seems quite intuitive that we should be able to locate the mean score more precisely than that of a single student. The confidence interval for the Y population mean at a given X value is

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm t_{tab}\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X - \bar{X})^2}} \qquad (2)$$

Carrying the calculations through for 95% confidence, we find

$$\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\Sigma(X - \bar{X})^2}} = 0.4350$$

So the end points of our intervals are

$73.7661 \pm (2.4469)\,(4.2788)\,(0.4350)$

73.7661 ± 4.5544

69.2117 and 78.3205

Thus we would be 95% confident that the course average for the population of students scoring 70 on the first exam lies between 69.2117 and 78.3205. The interval for the Y population mean is considerably shorter than the interval for a single predicted observation.

III. CONFIDENCE INTERVAL ESTIMATES FOR THE Y POPULATION MEANS FOR ALL X SIMULTANEOUSLY (the entire population regresion line)

A formula does exist for setting joint confidence intervals on the mean of the Y populations for all X_0 i.e. $\beta_0 + \beta_1 X_0$ for all X_0.

The formula is,

$$\hat{\beta}_0 + \hat{\beta}_1 X_0 \pm \sqrt{2F_{tab}}\;\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(X_0 - \bar{X})}{\Sigma(X - \bar{X})^2}} \qquad (3)$$

Where F_{tab} is the tabulated F value for the desired level of confidence with 2 degrees of freedom for the numinator and n-2 degrees of freedom for the denominator. It is apparent that a wider intervals is required if we wish to set confidence limits simultaneously on all values of $\beta_0 + \beta_1 X$ than if we are satisfied with one X only. In this equation we multiply by $\sqrt{2F_{tab}}$ instead of t_{tab} because $\sqrt{F_{tab}}$ is the same as t_{tab}, the only difference is equation (3) is determined by $\sqrt{2}$. Thus equation (3) gives an interval at least 40% wider than that given by equation (2).

Carrying the calculation, $X_0 = 70$ with the current example, we find $\sqrt{2F_{tab}} = 3.2073$ for 95% confidence and the end points of the interval are given by

$$73.7661 \pm (3.2073)\ (4.2788)\ (0.4350)$$
$$73.7661 \pm 5.9691$$
67.7964 and 79.7358

Thus, the average score for all students population mean lies between 67.7964 and 79.7358.

REFERENCES

1. *Draper, N.R. and H. Smith (1966) Applied Regression Analysis, New York : Wiley.*

2. *Fisher, R.A. (1933). Statistical Methods for Research Workers, Edinburgh; Oliver & Boyd.*

3. *Snedecor, George W. (1934). Calculation and Interpretation of Analysis of Variance and Covariance, Ames, Ia : Collegiate Press.*

4. *Morris Hamburg (1970) Statistical Analysis for Decision Making; HBJ International.*

5. *Howard B. Christensen; Statistics Stop by step by step.*

6. *Gordon Bancrft/George O'Shulhran; Maths and Statistics for Accounting and Business Studies.*

7. *Richard L. Mills; Statistics for Applied Economics and Business.*