REGRESSION ANALYSIS: AN ASSESSMENT OF CONFIDENCE

1. INTRODUCTION

Evidence in court cases on the potential earnings of individuals is usually in the form of calculated *sample averages*. These averages may be distinguished by province, five-year age group, sex and educational attainment or occupation[1]. Additional information such as the work and earnings history of the individual is often incorporated in the analysis.

Occasionally, estimates of an individual's potential earnings are based on a more complicated but also more powerful approach, *multiple regression analysis*. In an economic research setting, regression analysis is used predominantly to explore the relationships between variables to be predicted (dependent variables, e.g. earnings) and those variables which are thought to influence them (explanatory variables, e.g. age and education). Regression results can also be used to generate estimates of earnings for given values of the explanatory variables. The issue then arises of how confident the user of such estimates can be with respect to the accuracy of the forecasts.

Where an estimate is produced in the form of a single number, say, a sample average, a confidence interval can be calculated to convey information on the reliability of the estimate[2]. In the context of regression analysis, the calculation of confidence intervals is not a standard procedure. Furthermore, whenever a series of values is estimated, confidence intervals may no longer be as useful. To illustrate this point, suppose estimates of potential future earnings were presented for a relatively long time horizon, say, 40 years of working life remaining. This would require the production of 40 confidence intervals for 40 predicted values and, of course, 40 separate decisions on whether to accept or reject all, none or some of the 40 estimates.

Rather than examine the accuracy of each individual estimate, it may be more useful to satisfy oneself as to the reliability of the methodology itself, and to examine the quality of the results obtained in the regression analysis. It is suggested that as long as a sound methodology is applied correctly to an appropriate set of data, and as long as the resulting estimate of the relationship between earnings and its determinants is a "high-quality estimate", then there is no reason to believe that the predictions which flow from those results are bad predictions. Section 2 below provides a brief description of

[1] These are custom data calculated by Statistics Canada on the basis of information provided by the 20% of Canada's population who completed the "long form" of the Census questionnaire.

[2] In plain language, confidence intervals are expressed as "This value is considered accurate plus or minus x% so many times out of a hundred". As a rule, the lower the plus/minus and the more the "times out of a hundred", the higher the confidence in the estimated value. One ought to bear in mind, however, that acceptance or rejection of the estimate is ultimately a subjective choice on the part of the user, because it is the user who determines whether the plus/minus is low enough and the "times out of a hundred" sufficiently high.

the fundamentals of regression analysis. Section 3 provides a perspective on how one might go about determining the quality of regression results.

2. FUNDAMENTALS OF REGRESSION ANALYSIS

Regression analysis is a statistical technique which is used to estimate the relationship between a dependent variable and an independent or explanatory variable. One of the questions which can be answered by regression analysis is: how can I predict Y, given my belief that X determines its value? The belief (or hypothesis) that X somehow determines the value of Y is usually expressed as:

$$Y = bX + \varepsilon$$

This expression shows "Y" as the dependent variable, "X" as the explanatory variable, "b" as the link between the two variables (referred to as a "coefficient"), and an error term, "ε". The error term captures any fluctuations in Y which cannot be attributed uniquely to X. An important feature of many regression techniques is that the average of the estimated errors is zero[3]. The measure of interest is therefore the coefficient estimate: once an estimate of the coefficient has been obtained, projections of Y can be generated by choosing X and multiplying it by the estimate of b.

"Multiple" regression analysis works in the same way, except that a number of Xs are included in the analysis instead of a single X. The question answered is: how can I predict Y, given my belief that a number of Xs somehow determine the value of Y? In the multiple regression extension, there will be several coefficient estimates, each representing an estimate of the link between Y and one of the Xs included. As in the case of simple regression analysis, "ε" captures variability in Y which cannot be attributed uniquely to any of the Xs, and again usually averages to zero.

Many different regression techniques are available that may be used to estimate the relationship between a dependent variable Y and an X or a number of Xs. Each procedure has its advantages and disadvantages, and its strengths and weaknesses. Some may not be appropriate in all circumstances, others may be appropriate only in specific situations. Whatever procedure is chosen, however, the goal is always the same: to estimate the relationship between Y and X .

[3] Note that this is a mechanical property of some techniques because the technique itself may rest on the assumption that the errors have an average of zero. This is a reflection of the belief that the positive and negative effects of chance events, unobserved and unmeasurable characteristics tend to offset each other. Where the assumption is found to be incorrect, a different estimation technique might be used that rests on a more appropriate assumption.

3. MEASURING THE QUALITY OF REGRESSION OUTPUT

Users of estimates based on regression analysis may derive a certain amount of confidence from the knowledge that regression analysis is a "tried, tested and true" methodology. For many years, a great deal of empirical research in economics has used regression techniques to explore the relationships between dependent variables and their determinants. In addition, the properties of the different estimation techniques have been extensively researched and are well understood.

Another way to gauge the trustworthiness of regression results is to compare them to the results other researchers have obtained in the past in similar contexts, and how they conform to expectations formed on the basis of economic theory. For example, human capital theory would predict that all else equal in general more education will result in higher predicted earnings. This particular theoretical expectation has been confirmed by countless empirical studies. If regression output were to indicate that education has no effect or perhaps even a negative effect on earnings, then the researcher should either be able to reconcile the result with other aspects of economic theory or should investigate the possibility of estimation problems.

Confidence may also be derived from the fact that regression analysis can be considered an extension (albeit a very sophisticated and potentially rather complicated one) to simple averaging. It can be shown that in the simplest possible application, a regression procedure will produce a predicted value equal to the sample average.

Several technical measures are produced as part of the regression output which usually serve as indicators of confidence. One obvious candidate is "R^2", a measure familiar to many entry-level statistics students[4]. This measure indicates the percentage of the total variation observed in the dependent variable which is explained uniquely by the explanatory variables. The draw-back of the R^2 measure is that there does not seem to be a consensus among researchers as to what constitutes a "good" R^2. Some researchers draw conclusions from regression results despite the fact that only 10% or 20% of the total variation observed is explained by their model, others will be thrilled to

[4] Two alternative R^2 measures are usually produced. "Unadjusted" R^2 has the mechanical property of increasing every time a variable is added, no matter how inconsequential or irrelevant the additional variable might be. Quite literally, one could add the price of widgets to any regression analysis and R^2 would increase. Therefore, a more appropriate measure would be "adjusted R^2", which imposes a penalty every time a variable is added. If the additional variable does not explain a sufficient amount of variation the adjusted R^2 will tend to decrease even though unadjusted R^2 will increase.

report an R^2 greater than 30%, and yet others will hardly comment on the fact that R^2 was 70%. Thus, the R^2 measure is very much subject to personal evaluation.

One can also examine the statistical significance of the estimated coefficients (the links between the explanatory variables and the dependent variable) by examining their so-called "t-ratios". T-ratios provide a handy measure of statistical significance. Essentially, a t-ratio close to or greater than 1.96 tells the researcher that the variable to which it relates is one of the determinants of the dependent variable[5]. Furthermore, the higher the t-ratio the narrower the confidence interval around the estimated value of the coefficient.

- -

To summarize, regression analysis has a long history in economic research, and the properties of many different techniques are well understood. When used appropriately, these techniques can provide a great deal of insight into the determinants of an individual's earnings, and can be used to generate high-quality estimates of earning capacity. Although confidence intervals for such estimates are not easy to calculate, some of the information such intervals would convey can be obtained from the regression output.

[5] It should be noted that when the objective is to generate predictions of the dependent variable, some researchers choose not to eliminate variables with low t-ratios. The reasoning for this is that as long as the confidence interval around a coefficient estimate is centred on the true value, and as long as there is a good theoretical reason to believe that the variable to which the coefficient estimate relates is a predictor of the dependent variable, the width of the confidence interval around the estimate of the coefficient (as implied by the t-ratio) is irrelevant.

REGRESSION ANALYSIS:

AN ASSESSMENT OF CONFIDENCE

Jutta G. Heinrichs, B.A., M.A.
Associated Economic Consultants Ltd.
Vancouver, B.C.

Submitted to
The Verdict and *The Barrister*
for publication May 14, 1998