The Economic and Social Review, Vol. 26, No. 1, October, 1994, pp. 69-74

Artificial Regression Based Mis-Specification Tests for Discrete Choice Models

ANTHONY MURPHY
University College, Dublin

Abstract: Lagrange Multiplier (LM) tests for omitted variables, neglected heteroscedasticity and other mis-specifications in general discrete choice models may be simply and conveniently calculated using an artificial regression. This artificial regression approach is likely to have better small sample properties than the more common outer product gradient (OPG) form of LM test.

I INTRODUCTION

Davidson and MacKinnon (1984b) derive convenient Lagrange Multiplier (LM) tests for omitted variables and neglected heteroscedasticity in logit and probit models.[1] Their LM tests are convenient since they are based on artificial regressions. They are also likely to have good small sample properties since they do not use the outer product gradient (OPG) form of the LM test.[2] Instead the information matrix is calculated as the expectation of the outer product of the score and is not just approximated by the outer product of the score. The Davidson and MacKinnon approach may be used to derive many other mis-specification tests in both binary choice models (e.g. tests of normality in probit models and asymmetry in logit models) and some more general discrete choice models (e.g. normality in censored bivariate probit models).

Paper presented at the Eighth Annual Conference of the Irish Economic Association.

1. Engle (1984) also suggests using this approach.
2. See Engle (1984) and Godfrey (1988) for an extensive survey of LM tests and Davidson and MacKinnon (1983; 1984a; 1993), Godfrey, (1988) and MacKinnon (1992) for a discussion of limitations of the OPG form of LM test.

In this paper artificial regression based LM tests for general discrete choice models are derived. The tests are likely to have good small sample properties since the OPG form of the LM test is not used. The proposed LM tests may be used to detect a range of mis-specifications such as omitted variables, neglected heterogeneity, incorrect functional form and non-normality/asymmetry in ordered probit/logit models.

II GENERAL DISCRETE CHOICE MODEL

Consider a discrete choice model with a random sample of N individuals, denoted by subscript i, and J+1 alternatives numbered from 0 to J. Let y_{ij} be an indicator variable for individual i and alternative j. Thus, y_{ij} equals one if individual i selects alternative j; otherwise y_{ij} equals zero. Let p_{ij} be the probability that i selects alternative j. p_{ij} depends on the parameter vector θ. The true parameter is assumed to be in the interior of the parameter space. For any individual both the sum of the y_{ij}'s and the p_{ij}'s across the J+1 alternatives equal one.

With a random sample the log likelihood is:

$$l = \sum_{i=1}^{N} \sum_{j=0}^{J} y_{ij} \ln p_{ij} \tag{1}$$

and the score equals:

$$\frac{\delta l}{\delta \theta} = \sum_i \sum_j \frac{y_{ij}}{p_{ij}} \frac{\delta p_{ij}}{\delta \theta} \tag{2}$$

which may be recast as:

$$\begin{aligned} \frac{\delta l}{\delta \theta} &= \sum_i \sum_j \left(\frac{y_{ij} - p_{ij}}{\sqrt{p_{ij}}} \right) \left(\frac{1}{\sqrt{p_{ij}}} \frac{\delta p_{ij}}{\delta \theta} \right) \\ &= \sum_i \sum_j u_{ij} z_{ij} \end{aligned} \tag{3}$$

since $\sum_j \delta p_{ij} / \delta \theta = 0$. The u_{ij}'s may be thought of as standardised residuals. They have zero means, variances equal to $1 - p_{ij}$ and covariances equal to $-(p_{ij}\, p_{ik})^{\frac{1}{2}}$ when $j \neq k$. The information matrix is:

$$
\begin{aligned}
I_{\theta\theta} &= \lim_{N\to\infty} E\frac{1}{N}\frac{\delta l}{\delta\theta}\frac{\delta l}{\delta\theta'} \\
&= \lim_{N\to\infty} E\frac{1}{N}\sum_i\sum_j \frac{1}{p_{ij}}\frac{\delta p_{ij}}{\delta\theta}\frac{\delta p_{ij}}{\delta\theta'} \\
&= \lim_{N\to\infty} E\frac{1}{N}\sum_i\sum_j z_{ij}z_{ij}'
\end{aligned}
\tag{4}
$$

since the sample is random, $Ey_{ij}^2 = Ey_{ij} = p_{ij}$ and $Ey_{ij}y_{ik} = 0$ when $j \neq k$. The information matrix is assumed to be non-singular in the neighbourhood of the true parameter value.

III LM TEST STATISTIC

The LM test statistic of the null $\theta = \theta_0$ is:

$$
LM = \frac{1}{N}\frac{\delta l}{\delta\hat{\theta}'} I_{\hat{\theta}\hat{\theta}}^{o-1} \frac{\delta l}{\delta\hat{\theta}} \tag{5}
$$

where both the score and the observed information matrix $I_{\hat{\theta}\hat{\theta}}^{o}$ are evaluated using the restricted parameter estimates $\hat{\theta}$.

The observed information matrix is:

$$
\begin{aligned}
I_{\hat{\theta}\hat{\theta}}^{o} &= E\frac{1}{N}\sum_i\sum_j \frac{\delta l}{\delta\hat{\theta}}\frac{\delta l}{\delta\hat{\theta}'} \\
&= E\frac{1}{N}\sum_i\sum_j \frac{1}{\hat{p}_{ij}}\frac{\delta p_{ij}}{\delta\hat{\theta}'}\frac{\delta p_{ij}}{\delta\hat{\theta}'}.
\end{aligned}
\tag{6}
$$

Under standard weak regularity conditions, the LM test statistic has a chi-squared distribution with degrees of freedom equal to the number of restrictions under the null.

IV ARTIFICIAL REGRESSION BASED LM TEST STATISTIC

The LM test statistic may be calculated as:

$$
LM = \left(\sum_i\sum_j \hat{u}_{ij}\hat{z}_{ij}\right)' \left(\sum_i\sum_j \hat{z}_{ij}\hat{z}_{ij}'\right)^{-1} \left(\sum \hat{u}_{ij}\hat{z}_{ij}\right) \tag{7}
$$

where $\hat{u}_{ij}$ and $\hat{z}_{ij}$ are the estimates of u_{ij} and z_{ij} at the restricted parameter estimates:

$$\hat{u}_{ij} = \frac{y_{ij} - \hat{p}_{ij}}{\sqrt{\hat{p}_{ij}}}$$

$$\hat{z}_{ij} = \frac{1}{\sqrt{\hat{p}_{ij}}} \frac{\delta p_{ij}}{\delta \hat{\theta}_{ij}}$$

In (7) the LM test statistic is simply the explained sum of squares from the uncentred auxiliary regression of the $\hat{u}_{ij}$'s on the $\hat{z}_{ij}$'s across all J+1 alternatives and N individuals.

The LM test statistic is also asymptotically equal to NJ times the R^2 from this auxiliary regression. The proof is the same as in McFadden, 1987. Note that:

$$NJR^2 = \frac{LM}{\frac{1}{NJ}\sum_i \sum_j \hat{u}_{ij}}$$

and

$$\frac{1}{NJ}\sum_i \sum_j \hat{u}_{ij}^2 = \frac{1}{NJ}\sum_i \sum_j \left(\hat{u}_{ij}^2 - u_{ij}^2\right) + \frac{1}{NJ}\sum_i \sum_j u_{ij}^2 .$$

Use the mean value theorem to expand the first term as the product of a stochastically bounded expression and $\hat{\theta} - \theta$ which has a plim of zero. Thus the first term has a plim of zero. Finally, note that the expectation of the second term is one, since $Eu_{ij}^2 = 1 - p_{ij}$ and $\sum_i \sum_j Eu_{ij}^2 / NJ = 1$, and the variance tends to zero as N tends to infinity. Thus the plim of NJ times the R^2 from the auxiliary regression equals the LM test statistic.

V BINARY CHOICE MODEL

In the special case of two alternatives 0 and 1, the log likelihood equals:

$$l = \sum_{i-1}^{N} \left\{ (1 - y_i) \ln(1 - p_i) + y_i \ln p_i \right\} \tag{1'}$$

and the score equals:

$$\begin{aligned} \frac{\delta l}{\delta \theta} &= \sum_i \left[-\left(\frac{1 - y_i}{1 - p_i} \right) + \left(\frac{y_i}{p_i} \right) \right] \frac{\delta p_i}{\delta \theta} \\ &= \sum_i \left(\frac{y_i - p_i}{\sqrt{p_i(1 - p_i)}} \right) \left(\frac{1}{\sqrt{p_i(1 - p_i)}} \frac{\delta p_i}{\delta \theta} \right) \\ &= \sum_i r_i x_i \end{aligned} \tag{3'}$$

where y_i is an indicator variable (i.e. y_i is one if alternative one is chosen and zero otherwise) and p_i is the probability that alternative one is chosen. The r_i are just the scaled residuals — they have a zero mean and a unit variance. Note that, using the notation of the previous section:

$$y_{i0} = 1 - y_i, y_{i1} = y_i, p_{i0} = 1 - p_i, p_{i1} = p_i, r_i = \sum_{j=0}^{1} u_{ij}, x_i = \sum_{j=0}^{1} z_{ij}$$

The information matrix is:

$$I_{\theta\theta} = \lim_{N\to\infty} E\frac{1}{N}\sum_i x_i x_i' \tag{5'}$$

and the LM test statistic is:

$$LM = \left(\sum_i \hat{r}_i \hat{x}_i'\right)\left(\sum_i \hat{x}_i \hat{x}_i'\right)^{-1}\left(\sum_i \hat{r}_i \hat{x}_i\right) \tag{7'}$$

where the observed scaled residuals and regressors are:

$$\hat{r}_i = \frac{y_i - \hat{p}_i}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}$$

$$\hat{x}_i = \frac{1}{\sqrt{\hat{p}_i(1-\hat{p}_i)}}\frac{\delta p_i}{\delta \hat{\theta}}$$

As Davidson and MacKinnon (1984) point out, the LM test statistic is just the explained sum of squares from the uncentred auxiliary regression of $\hat{u}_i$ on $\hat{x}_i$. Asymptotically N times the R^2 from this regression equals the LM test statistic.

VI CONCLUSION

LM test statistics for omitted variables, neglected heteroscedasticity and other mis-specifications in general discrete choice models may be readily calculated using an artificial regression, the same as in binary choice models. However the form of the artificial regression is different in the general case. This artificial regression approach is both convenient and likely to have better small sample properties than the more common outer product gradient form of LM statistic.

REFERENCES

DAVIDSON, R., and J.G. MacKINNON, 1983. "Small Sample Properties of Alternative Forms of the Lagrange Multiplier Test", *Economic Letters*, Vol. 12, pp. 269-275.

DAVIDSON, R., and J.G. MacKINNON, 1984a. "Model Specification Tests Based on Artificial Linear Regressions", *International Economic Review*, Vol. 25, pp. 485-502.

DAVIDSON, R., and J.G. MacKINNON, 1984b. "Convenient Specification Tests for the Logit and Probit Models", *Journal of Econometrics,* Vol. 25, pp. 241-262.

DAVIDSON, R., and J.G. MacKINNON, 1993. *Estimation and Inference in Econometrics*, New York: Oxford University Press.

ENGLE, R.F., 1984. "Wald, Likelihood Ratio and Lagrange Multiplier Tests in Econometrics", Ch. 13, in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics,* Vol. II, pp. 775-826. Amsterdam: North-Holland.

GODFREY, L.G., 1988. *Mis-Specification Tests in Econometrics,* Cambridge: Cambridge University Press.

MacKINNON, J.G., 1992. "Model Specification Tests and Artificial Regressions", *Journal of Economic Literature,* Vol. XXX, pp. 102-146.

McFADDEN, D., 1987. "Regression Based Specification Tests for the Multinomial Logit Model", *Journal of Econometrics*, Vol. 34, pp. 63-82.