19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011
http://mssanz.org.au/modsim2011

Testing for structural breaks in discrete choice models

Johnathan Wongsosaputro[a], Laurent L. Pauwels[a], and Felix Chan[b]

[a]*The University of Sydney Business School, New South Wales 2006, Australia*
[b]*School of Economics and Finance, Curtin University, GPO Box U1987, Perth, WA 6845, Australia*
Email: laurent.pauwels@sydney.edu.au

Abstract:

A structural break refers to a shift in the parameters of the model of interest. When the conditional relationship between the dependent and explanatory variables contains a structural break, estimates of model coefficients will be inaccurate across different regimes. As such, estimations that do not account for structural breaks will be biased and inconsistent.

Ever since the seminal work of Chow [1960], there have been numerous other tests proposed for detecting various forms of structural breaks in different contexts. Chow [1960] proposed an F-statistic to detect a single structural break with known location in the context of a linear model. One of the most commonly used tests is the Andrews [1993], which generalised Chow [1960] to the Sup Wald, LR, and LM tests for linear models when the position of the breakpoint is unknown. Other important contributions to the literature include Bai and Perron [1998] and Bai [1999], both of which constructed tests that detect multiple structural breaks in linear models.

Studying the properties of such tests is particularly important because the theoretical distribution of most of the test statistics have only been identified asymptotically, but the same critical values are also used for smaller sample sizes in practice. Furthermore, the theoretical properties of the test statistics are usually established only under certain restrictions such as $i.i.d.$ assumptions that may not hold in practice for various reasons. While present literature does include studies of structural break tests where the changepoint is unknown, such as Diebold and Chen [1996], and Bai and Perron [2004], these have mostly been restricted to linear regression models. To our knowledge, no study has been carried out thus far to evaluate the properties of any structural break test in the context of binary choice models, such as probit models, which will be the main contribution of this paper.

This paper considers the size and power of the three Andrews [1993] Sup-type tests when applied to probit models with different levels of autocorrelation and varying sample sizes using a simulation-based approach similar to Diebold and Chen [1996], which tested the sizes of the tests in the linear regression model . We carry out the same procedure with a different data generating process, but also further the study by comparing the results in the linear model with that of a probit model. In addition, we also consider the power of the tests in both models, as well as a few different levels of data trimming.

The main findings of this paper are that the shortcomings of the Andrews [1993] Sup-type tests in linear models are magnified in probit models. In particular, the tests exhibit greater size distortion, lower power, and become more imprecise in identifying the position of the structural break when the samples are small or when the errors are autocorrelated.

Keywords: Structural break, unknown breakpoint, binary choice, probit model, autocorrelation, simulation

1 BINARY CHOICE MODELS AND THE ANDREWS [1993]

This paper considers two models, one linear and one probit, both of which consist of a single explanatory variable. Under the null hypothesis of no structural change, the restricted models are as follows:

$$\textbf{Linear model:} \quad y_t^* = \beta x_t + \epsilon_t \,, \quad \textbf{Probit model:} \quad y_t = \begin{cases} 1 & \text{if } y_t^* < 0 \\ 0 & \text{otherwise} \end{cases} \,, \tag{1}$$

The explanatory variable, x_t, and error terms, ϵ_t, are the same across both models, but they influence the dependent variable, y_t, in different ways. In the linear model, y_t^* has a linear relationship with x_t and ϵ_t, with the relative influence of both inputs depending on the magnitude of the coefficient, β. This implies that β has a direct impact on y_t. In the probit model, y_t^* is an unobserved latent variable that influences the observed binary variable y_t. As a result, β no longer impacts y_t directly and instead influences the probability of y_t taking on one of two values.

In the linear case, the unrestricted model containing a single break is defined as

$$\begin{aligned} y_t^* &= \beta_1 x_t + \epsilon_t, \qquad t = 1, ..., \pi T, \\ y_t^* &= \beta_2 x_t + \epsilon_t, \qquad t = (\pi T + 1), ..., T.\text{'} \end{aligned} \tag{2}$$

while the unrestricted probit model is

$$\begin{aligned} y_t &= \begin{cases} 1 & \text{if } y_t^* < 0 \\ 0 & \text{otherwise} \end{cases} \,, \quad t = 1, ..., \pi T, \\ y_t &= \begin{cases} 1 & \text{if } y_t^* < 0 \\ 0 & \text{otherwise} \end{cases} \,, \quad t = (\pi T + 1), ..., T, \end{aligned} \tag{3}$$

where π denotes the proportion of observations before the breakpoint in the unrestricted model being estimated, and πT denotes the position of the breakpoint. Under the Andrews [1993] framework, the null hypothesis assuming no structural break in the data is tested against the alternative of a single structural break:

$$H_0 : \beta_t = \beta \text{ for all } t \geq 1 \text{ for some } \beta_0 \in B \subset R^p.$$

$$H_{1T} : \beta_t = \begin{cases} \beta_1(\pi) & \text{for } t = 1, ..., \pi T \\ \beta_2(\pi) & \text{for } t = \pi T + 1, ... \end{cases} \,,$$

In the linear model, W, LM, and LR will be

$$W(\pi) = T\left[\frac{\hat{\epsilon}^T\hat{\epsilon} - \hat{\epsilon}_1^T\hat{\epsilon}_1 - \hat{\epsilon}_2^T\hat{\epsilon}_2}{\hat{\epsilon}_1^T\hat{\epsilon}_1 + \hat{\epsilon}_2^T\hat{\epsilon}_2}\right], LM(\pi) = T\left[\frac{\hat{\epsilon}^T\hat{\epsilon} - \hat{\epsilon}_1^T\hat{\epsilon}_1 - \hat{\epsilon}_2^T\hat{\epsilon}_2}{\hat{\epsilon}^T\hat{\epsilon}}\right], LR(\pi) = T\ln\left[\frac{\hat{\epsilon}^T\hat{\epsilon}}{\hat{\epsilon}_1^T\hat{\epsilon}_1 + \hat{\epsilon}_2^T\hat{\epsilon}_2}\right],$$

with W being asymptotically equivalent to the Chow [1960] test. In the probit model, we define $l_1(.)$ and $l_2(.)$ as the log-likelihood before and after the break respectively. The log-likelihood of the unrestricted model will then be $l_1(.) + l_2(.)$. When the model is correctly specified with spherical errors, the test statistics are

$$\boldsymbol{W}_T(\pi) = T\left(\hat{\beta}_1(\pi) - \hat{\beta}_2(\pi)\right)^T \left(\hat{V}_1(\pi)/\pi + \hat{V}_2(\pi)/(1-\pi)\right)^{-1} \left(\hat{\beta}_1(\pi) - \hat{\beta}_2(\pi)\right),$$

$$\boldsymbol{LM}_T(\pi) = \frac{1}{\pi(1-\pi)} \left(\frac{\partial l_1(\hat{\beta})}{\partial \hat{\beta}}\right)^T \left[\hat{V}(\pi)^{-1}\right] \left(\frac{\partial l_1(\hat{\beta})}{\partial \hat{\beta}}\right), \boldsymbol{LR}_T(\pi) = 2\left[l_1\left(\hat{\beta}_1\right) + l_2\left(\hat{\beta}_2\right) - l\left(\hat{\beta}\right)\right],$$

where $\hat{V}(\pi)$ is estimated using the negative of the hessian matrix in accordance with the information matrix equality. $\left(\frac{\partial l_1(\hat{\beta})}{\partial \hat{\beta}}\right)$ is the first derivative of the log-likelihood of the subsample before the breakpoint in the unrestricted model evaluated at the maximum likelihood estimate for the restricted model.

When testing the null of no structural break against the alternative of a single structural break, the statistic that is chosen from the three above will then be computed for every observation within a pre-defined region, Π. The corresponding test statistics are then $\sup_{\pi\in\Pi} W_T(\pi), \sup_{\pi\in\Pi} LM_T(\pi),$ and $\sup_{\pi\in\Pi} LR_T(\pi)$. Π must be bounded away from zero and one in order for the the test statistics to converge in distribution. Andrews [1993] provided critical values for various choices of Π, but also arbitrarily suggested $\Pi = [0.15, 0.85]$, with 15% trimming on both ends of the sample.

2 CRITICAL VALUES

2.1 Asymptotic critical values

The critical values from Andrews [2003] that are relevant to our simulations are as follows, with π denoting the proportion of trimming on each end:

	Significance level		
π	10%	5%	1%
.15	7.12	8.68	12.16
.10	7.58	9.11	12.59
.05	8.13	9.71	13.17

The critical values shown above are only valid asymptotically, and the test statistics are known to exhibit varying levels of size distortion when those critical values are used in small samples. An alternative method of obtaining critical values is to use bootstrapping.

2.2 Bootstrapping

Bootstrapping, which involves generating bootstrap pseudo-samples through resampling of the data available, has frequently been shown to improve the finite-sample performances of various hypothesis tests. In particular, Diebold and Chen [1996] affirmed a size improvement for the Andrews [1993] tests in small samples even when the data exhibits high persistence.

We carry out parametric bootstrapping using the fast bootstrap procedure introduced in Davidson and MacKinnon [1999] and evaluated in Lamarche [2004]. Unlike the usual procedure where B bootstrap samples are generated at each iteration, the fast bootstrap generates a single bootstrap sample using the using a new set of $N(0,1)$ errors at each iteration. The test statistics generated from these bootstrap samples are then used as an approximation of the bootstrap distribution, with the critical values obtained from corresponding percentiles. Davidson and MacKinnon [1999] showed that the fast bootstrap is valid when the test statistic and bootstrap DGP are independent or asymptotically independent, a condition that is satisfied in parametric bootstrapping when the parameters are estimated under the null distribution, which is the case in this paper.

3 SIMULATION DESIGN

The linear and probit models described in equations $(1)-(3)$ are simulated with x_t and ϵ_t defined as

$$x_t = e_t, \quad e_t \sim N(0,4), \quad \text{and}$$
$$\epsilon_t = \rho\epsilon_{t-1} + u_t, \quad u_t \sim N(0,1), \quad t = 1, ..., T.$$

Such a specification ensures that the performances of the tests are fully comparable across both models, since the inputs, x_t and ϵ_t, are the same in both models. At the same time, varying ρ allows us to investigate the effect of autocorrelation on the test statistics. For this reason, the models are estimated without accounting for persistence in the errors.

The simulation procedure is as follows:

1. Simulate the linear and probit data as specified in (1) when testing for size and $(2)-(3)$ when for power with x_t and ϵ_t defined above.

2. Estimate $\hat{\beta}$ in the linear model using OLS and in the probit model using maximum likelihood and Compute $\sup_{\pi\in\Pi} W_T(\pi), \sup_{\pi\in\Pi} LM_T(\pi),$ and $\sup_{\pi\in\Pi} LR_T(\pi)$.

3. Generate a T-vector of bootstrap residuals, ϵ_t^b, where $\epsilon_t^b \sim N(0,1)$.

4. Use ϵ_t^b to generate a bootstrap sample under the null for both models by replacing ϵ_t with ϵ_t^b in (1).

5. Estimate $\hat{\beta}^b$ in both bootstrap samples and compute the bootstrap test statistics.

6. Repeat steps 1 to 6 for N iterations.

7. Determine the asymptotic size/power for each test based on the proportion of iterations in which the test statistic exceeds the asymptotic critical values in section 2.2.

8. Determine the bootstrap size/power for each test based on the proportion of iterations in which the test statistic exceeds the bootstrap critical values, defined as the value corresponding to the $(1-\alpha)\times 100^{th}$ percentile of the N bootstrap test statistics, where α is the selected nominal size.

When testing for size, we formulate the models as defined, with $\beta = 1$ throughout. When assessing power against a single structural break, we set a structural break at $0.5T$, and simulate two specifications with $\hat{\beta}_1 = 1, \hat{\beta}_2 = 1.5$, and $\hat{\beta}_1 = 1, \hat{\beta}_2 = 2$. When it comes to power against two breaks, we set structural breaks at $0.3T$ and $0.7T$ and three specifications with $\hat{\beta}_1 = 1$, $\hat{\beta}_2 = 1.5$, and $\hat{\beta}_3$ taking on values of 2, 1, and 0.5. We run simulations for values of ρ between 0 and 0.99 in increments of 0.10, and sample sizes $T = 10$, 30, 50, 100, 250, 500, 1000. All simulations are run over $N = 1000$ iterations.

4 RESULTS

We present our results compactly surface plots, constructed via OLS regressions against third-degree expansions and cross-products of the simulation parameters. For brevity, only selected results for 15% trimming and 10% significance level are shown. Asymptotic and bootstrap results are denoted by the prefix Asy and $Boot$ respectively, followed by the name of the test and the suffixes $-L$ for linear models and $-P$ for probit models.

4.1 Size

The surface plots of the sizes in linear models mirror the results shown in Diebold and Chen [1996]. The empirical sizes of the asymptotic test statistics converge to their nominal sizes when the sample size is large, as expected. In smaller samples, $AsySupLM - L$ is undersized, while $AsySupLR - L$ and $AsySupWald - L$ are oversized. Bootstrapping reduces the size distortion caused by the smaller sample size, as evidenced by the flatter surface of $BootSupLM - P$. Diebold and Chen [1996], however, found that the the bootstrap test statistics remained close to perfect in the presence of high autocorrelation, while we observed a slight positive distortion in $BootSup - L$. This can be attributed to a difference in model specifications, as well as our use of the fast parametric bootstrap as opposed to Diebold and Chen [1996]'s empirical bootstrap.

Figure 1. Size of the Andrews [1993] test

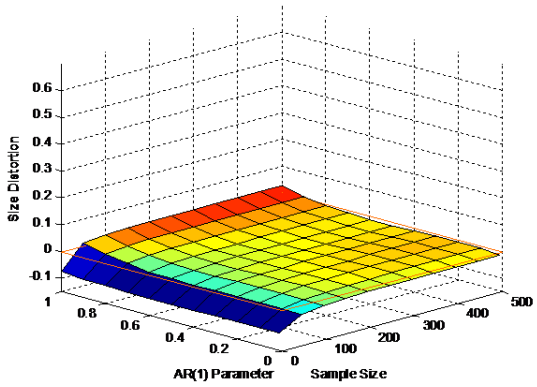

Figure 1a: Size distortion of $AsySupLM - L$

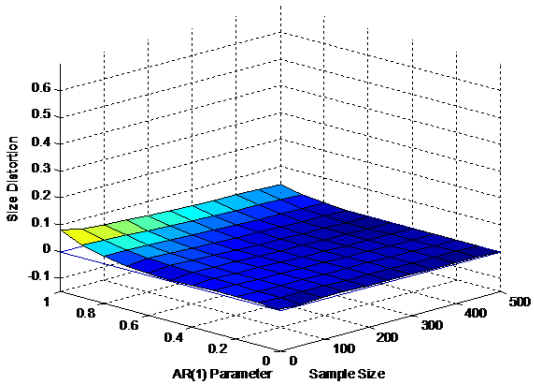

Figure 1b: Size distortion of linear $BootSupLM - L$

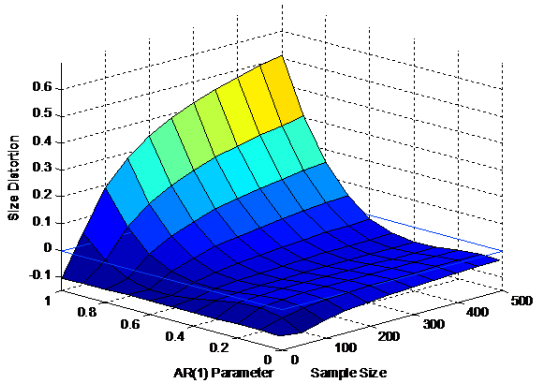

Figure 1c: Size distortion of $AsySupW - P$

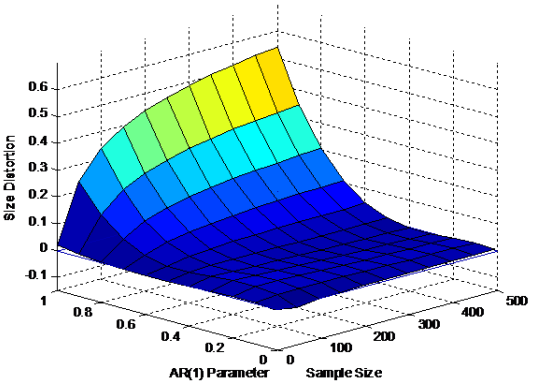

Figure 1d: Size distortion of $BootSupW - P$

The size results for $AsySup - P$, however, stand in stark contrast to their linear counterparts. While the asymptotic test statistics do converge to their nominal sizes, the convergence is much slower than in the linear models. $AsySupW - P$, in particular, remains undersized even when the sample size is 500. In addition, the effect of high autocorrelation is especially harsh on the test statistics when applied to Probit models, with the empirical sizes becoming oversized to unmanageable levels. Somewhat peculiarly, this size distortion worsens as the sample size increases, perhaps because the effects of autocorrelation build up fairly slowly and require a larger sample size to have an impact on the test statistic. Similar to the linear case, bootstrapping greatly reduces size distortion in smaller sample sizes, but this only occurs in the probit model when the autocorrelation is low. Under higher levels of persistence, bootstrapping appears to exacerbate the size distortion, although this most likely occurs because the undersizing of the asymptotic test statistics cancels some of the oversizing caused by the autocorrelation. Overall, these results suggest that correcting for the presence of autocorrelation is particularly important when testing for structural breaks in Probit models, more so than in linear models.

4.2 Power

The surface plots show that the power of the tests increase with the sample size and eventually converge to 1. It was observed in the linear case that $AsySupW-L$ had the highest power, followed by $AsySupLR-L$ and then $AsySupLM - L$. It has to be noted, however, that this difference is most likely due to size

Figure 2. Single break: $\beta_1 = 1, \beta_2 = 1.5$

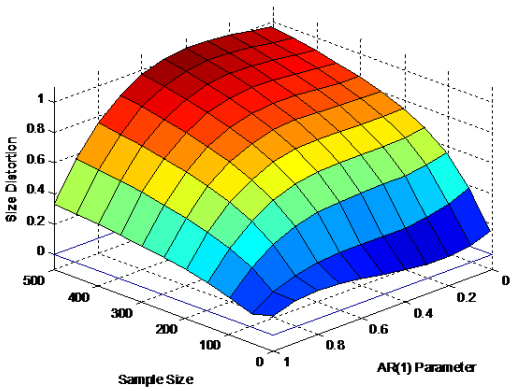

Figure 2a: Power of $AsySupLR - L$

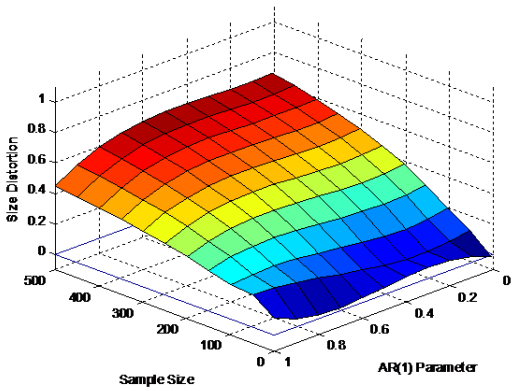

Figure 2b: Power of linear $AsySupLR - P$

distortions, since $AsySupW - L$ has the highest size distortion of the three. Autocorrelation has a substantial negative impact on power, with the loss in power becoming more noticeable as the sample size increases.

Comparing the probit results with the linear results, we observe the same patterns in $AsySup - P$ as described above, except that $AsySup - P$ generally has considerably less power than its linear counterparts. While the effect of autocorrelation on power appears to be diminished in the probit model, this is again most likely due to the test statistics being grossly oversized in the probit model. Indeed, the rejection rates of the tests is barely higher than their empirical sizes when persistence is high. The other simulations also yield a few other interesting observations. When we set $\beta_2 = 2$, the power of all the tests increased, which shows that the magnitude of the break has a positive effect on Sup-type tests. In addition, we also observed slight power gains under high autocorrelation when the level of trimming increases, which concurs with the results of Bai and Perron [2004]. Finally, we also found that the tests became less precise in identifying the position of the identified breakpoint as the level of autocorrelation increases. This was once again exacerbated in the Probit models as compared to the linear ones.

The simulations containing multiple breaks show that the power of the test is greatly reduced when the breakpoints are in opposite directions, a result that was also seen in Bai and Perron [2004]. The power of the test statistics when $\beta_3 = 2$ was greater than a single change from $\beta_1 = 1$ to $\beta_2 = 1.5$, but less than a single change to $\beta_2 = 2$. In contrast, the power of the tests when $\beta_2 = 1.5$ and $\beta_3 = 1$ was the lowest out of our simulation specifications. We observed, once again, that this result was more noticeable in the probit models. This has implications on data sets containing temporary structural breaks and models involving regime-switching, since the tests will find it difficult to pick up these breaks. The specification in which $\beta_1 = 1$, $\beta_2 = 1.5$, and $\beta_3 = 0.5$ involves a second structural break that is greater in magnitude and in the opposite direction from the first. In this case, the tests exhibited greater power than a single change to $\beta_2 = 1.5$, but less than a staggered change with $\beta_2 = 1.5$ and $\beta_3 = 2$. This seems to suggest that the loss in power occurring due to the breaks of opposite sign cancelling out can be alleviated provided that one of the breaks has a larger magnitude than the other.

5 Conclusion

This paper presented a simulation analysis comparing the size and power of the Andrews [1993] Sup-type test statistics under finite-samples and varying degrees of autocorrelation in Probit and linear models. Our results for linear models match with previous studies carried out in literature, but also showed that many of the shortcomings of the test statistics in linear models are magnified in probit models. In particular, the test statistics had larger size distortions, lower power, and were more inaccurate in identifying the

location of the breakpoint. While some of these problems could be solved in the linear models through bootstrapping, it was found that bootstrapping could only solve the problems caused by smaller samples but not the ones associated with autocorrelation.

REFERENCES

Andrews, D. W. K. [1993], 'Tests for parameter instability and structural change with unknown change point', *Econometrica* **61**(4), 821–56.

Andrews, D. W. K. [2003], 'Tests for parameter instability and structural change with unknown change point: A corrigendum', *Econometrica* **71**(1), 395–397.

Bai, J. [1999], 'Likelihood ratio tests for multiple structural changes', *Journal of Econometrics* **91**(2), 299–323.

Bai, J. and Perron, P. [1998], 'Estimating and testing linear models with multiple structural changes', *Econometrica* **66**(1), 47–78.

Bai, J. and Perron, P. [2004], Multiple structural change models: a simulation analysis. Unpublished manuscript.

Chow, G. [1960], 'Tests of equality between sets of coefficients in two linear regressions', *Econometrica* **28**, 591–605.

Davidson, R. and MacKinnon, J. G. [1999], 'The size distortion of bootstrap tests', *Econometric Theory* **15**(3), 361–376.

Diebold, F. X. and Chen, C. [1996], 'Testing structural stability with endogenous breakpoint a size comparison of analytic and bootstrap procedures', *Journal of Econometrics* **70**(1), 221–241.

Lamarche, J.-F. [2004], 'The numerical performance of fast bootstrap procedures', *Computational Economics* **23**, 379–389.