Published by: South Asian Academic Research Journals

ACADEMICIA:

An International Multidisciplinary Research Journal

CRITERIA FOR SELECTION OF REGRESSORS IN ECONOMETRICS

PROF. M. VENKATARAMANAIAH*; M. SUDARSANA RAO**

*Department of Statistics,
SVUCS,
Tirupati, India.
**Research Scholar,
Department of Statistics,
SVUCS,
Tirupati, India.

ABSTRACT

A common problem in applied regression analysis is to select the variables that enter a linear regression. Examples are selection among capital stock series constructed with different depreciation assumptions, or use of variables that depend on unknown parameters, such as Box-Cox transformations, linear splines with parametric knots, and exponential functions with parametric decay rates. It is often computationally convenient to estimate such models by least squares, with variables selected from possible candidates by enumeration, grid search, or Gauss-Newton iteration to maximize their conventional least squares significance level; term this method Prescreened Least Squares (PLS). This note shows that PLS is equivalent to direct estimation by non-linear least squares, and thus statistically consistent under mild regularity conditions. However, standard errors and test statistics provided by least squares are biased. When explanatory variables are smooth in the parameters that index the selection alternatives, Gauss-Newton auxiliary regression is a convenient procedure for obtaining consistent covariance matrix estimates. In cases where smoothness is absent or the true index parameter is isolated, covariance matrix estimates obtained by kernel-smoothing or bootstrap methods appear from examples to be reasonably accurate for samples of moderate size.

INTRODUCTION

Often in applied linear regression analysis one must select an explanatory variable from a set of candidates. For example, in estimating production functions one must select among alternative measures of capital stock constructed using different depreciation assumptions. Or, in hedonic analysis of housing prices, one may use indicator or ramp variables that measure distance from spatial features such as parks or industrial plants, with cutoffs at distances that are determined as parameters. In the second example, the problem can be cast as one of nonlinear regression. However, when there are many linear parameters in the regression, direct nonlinear regression canbe computationally inefficient, with convergence problematic. It is often more practical to approach this as a linear regression problem with variable selection. This paper shows that selecting variables in a linear regression to maximize their conventional least squares significance level is equivalent to direct application of non-linear least squares. Thus, this method provides a practical computational shortcut that shares the statistical properties of the nonlinear least squares solution. However, standard errors and test statistics produced by least squares are biased by variable selection, and are often inconsistent. I give practical consistent estimators for covariances and test statistics, and show in examples that kernel-smoothing or bootstrap methods appear to give adequate approximations in samples of moderate size.

Stated formally, the problem is to estimate the parameters of the linear model

(1) y = X + Z( ) + u, _ ,

where y is n×1, X is an n×p array of observations on fixed explanatory variables,

Z = Z( ) is an

n×q

array of observations on selected explanatory variables, where indexes candidates from a set of alternatives , and u is an n×1 vector of disturbances with a scalar covariance matrix. Let k = p + q, and assume _ _h. The set is finite in the traditional problem of variable selection, but will be a continuum for parametric data transformations. Assume the data in (1) are generated by independent random sampling from a model y = x_ o + z(o,w)_ o + u, where (y,x,w) _ _×_p×_m is an observed data vector, z: ×_m __ _q is a "well-behaved" parametric transformation of the data w, (o, o, o) denote the true parameter values, and the distribution of u is independent of x and w, and has mean zero and variance o 2. There may be overlapping variables in w and x. Examples of parametric data transformations are (a) a Box-Cox transformation z(,w) = w -1/(-1) for _ 0 and z(0,w) = log(w); (b) a ramp (or linear spline) function z(,w) = Max(-w,0) with a knot at ; (c) a structural break z(,w) = **1**(w<) with a break at ; and (d) an exponential decay z(,w) = e- w. 2

One criterion for variable selection is to pick a c _ that maximizes the conventional least squares test statistic for the significance of the resulting Z(c) variables, using enumeration, a grid search, or Gauss-Newton iteration, and then pick the least-squares estimates (a,b,s2) of (o, o, o 2) in (1) using the selected Z(c). As a shorthand, term this the Prescreened Least Squares (PLS) criterion

for estimating (1). I will show that PLS is equivalent to selecting Z( ) to maximize R2, and is also

equivalent to estimating (, , , 2) in (1) jointly by nonlinear least squares. Hence, PLS shares the large-sample statistical properties of nonlinear least squares. However, standard errors and test statistics for a and b that are provided by least squares at the selected Z(c) fail to account for the

impact of variable selection, and will usually be biased downward. When is a continuum, a Gauss-Newton auxiliary regression associated with the nonlinear least squares formulation of the problem can be used in many cases to obtain consistent estimates of standard errors and test statistics. When is finite, the effects of variable selection will be asymptotically negligible, but least squares estimates of standard errors will be biased downward in finite samples.

Let M = I - X(X_X)-1X_, and rewrite the model (1) in the form

(2) y = X[+ (X_X)-1X_Z( )] + MZ( ) + u .

The explanatory variables X and MZ( ) in (2) are orthogonal by construction, so that the sum of squared residuals satisfies

(3) SSR( ) = y_My - y_MZ( )[Z( )_MZ( )]-1Z( )_My .

Then, the estimate c that minimizes SSR( ) for _ also maximizes the expression

(4) S( ) _ n-1_y_MZ( )[Z( )_MZ( )]-1Z( )_My .

The nonlinear least squares estimators for , , and 2 can be obtained by applying least squares to (1) using Z = Z(c); in particular, the estimator of 2 is s2 = SSR(c)/(n-k) and the estimator of is b = [Z(c)_MZ(c)]-1Z(c)_My. Since R2 is monotone decreasing in SSR( ), and therefore monotone increasing in S( ), the estimator c also maximizes R2. Least squares estimation of (1) also yields an estimator Ve(b) = s2[Z(c)_MZ(c)]-1 of the covariance matrix of the estimator b; however, this estimator does not take account of the impact of estimation of the embedded parameter on the distribution of the least squares estimates. The conventional least-squares F-statistic for the null hypothesis that the coefficients in (1) are zero, treating Z as if it were predetermined rather than a function of the embedded estimator c, is

(5) F = b_Ve(b)-1b/q = y_MZ(c)[Z(c)_MZ(c)]-1Z(c)_My/s2q = (n_k)_S(c) . q_(y_My/n _ S(c))

But the nonlinear least squares estimator selects _ to maximize S( ), and (5) is an increasing function of S( ). Then estimation of (, , , 2) in (1) by nonlinear least squares, with _ , is equivalent to estimation of this equation by least squares with c _ selected to maximize the F-statistic (5) for a least squares test of significance for the hypothesis that = 0. When there is a single variable that depends on the embedded parameter , the F-statistic equals the square of the 3\T-statistic for the significance of the coefficient , and the PLS procedure is equivalent to selecting c to maximize the "significance" of the T-statistic.I have not found this result stated explicitly in the literature, but it is an easy special case of selection of regressors in nonlinear least squares using Unconditional Mean Square Prediction Error, which in this application where all candidate vectors are of the same dimension coincides with the Mallows Criterion and the Akaike Information Criterion; see Amemiya (1980). Many studies have noted the impact of variable selection or embedded parameters on covariance matrix estimates, and given examples showing that least squares estimates that ignore these impacts can be substantially biased; see

Amemiya (1978), Freedman (1983), Freedman-Navidi-Peters (1988), Lovell (1983), Newey-McFadden (1994), and Peters-Freedman (1984).

MODEL SELECTION

- Data generation process (DGP):
- joint distribution of all variables in economy
- Economic mechanism plus measurement system
- Huge dimensionality; highly non-stationary
- Impossible to model precisely
- Need to reduce to manageable size:
- local DGP (LDGP) is DGP in space of variables

MODEL SELECTION IN ECONOMETRICS

- Many features not derivable from economic theory
- institutional knowledge, or previous evidence:
- lag reactions; structural breaks; non-linear functions
- All have to be data-based on available sample–
- major problems of model specification and selection
- Former mainly up to investigator; latter is daunting
- May have several hundred candidate variables
- Computer-automated econometric

MODEL SELECTION

- seek to locate LDGP
- General-to-specific approach embodied in Autometrics
- In Monte Carlo, Gets recovers LDGP accurately
- Clarifies ‘data mining’ in economics

HOW TO SELECT AN EMPIRICAL MODEL?

Many grounds on which to select empirical models:

- theoretical
- empirical
- aesthetic
- philosophical
- Within each category, many criteria:
- theory: generality; internal consistency; invariance
- empirical: goodness-of-fit; congruence; parsimony;
- consistency with theory; constancy; encompassing;
- forecast accuracy
- aesthetic:
- elegance; relevance; ‘tell a story’
- philosophical:
- novelty; excess content; making money....

IMPLICATIONS

- Any test + decision = selection, so ubiquitous
- Most decisions undocumented
- often not recognized as selection
- Unfortunately, model selection theory is difficult:
- all statistics have interdependent distributions
- altered by every modelling decision
- Fortunately, computer selection algorithms allow
- operational studies of alternative strategies

MAINLY CONSIDER GETS: General-to-specific modeling

- Explain approach and review progress
- How costly to search many alternatives?
- If 1000 candidate variables, 21000 $\simeq$ 10300 possible models
- Makes task sound impossible
- Tests have non-zero rejection frequencies under null,
- but type-I errors do not accumulate
- Selection really only involves one decision:
- which variables to retain (equivalently, eliminate)
- Repeated-testing claims too pessimistic
- Fix by small null-rejection frequency:
- at some cost in lower power
- For 1000 candidate variables and 0.1% significance
- would retain just 1 variable by chance
- and on average eliminate 999–vast increase in knowledge
- Yet t(0.1%)$\simeq$ 3.4, so only small power loss

AUTOMATIC MODEL

SELECTION

- Hoover and Perez (1999) evaluate Gets:
- follow many search paths from congruent GUM;
- terminate if no reductions; or significant diagnostics
- Much better than Lovell’s (1983) ‘data mining’ critique
- Lower ‘size’ and raise power by improved algorithm
- Other experiments demonstrate:
- no major loss of power;

- correct ‘size’;
- accurate ‘goodness-of-fit’ estimates;
- standard errors accurate
- ‘Pre-testing’ implies biased coefficients:
- so literature suggests search has high costs
- But can bias correct selected models

GETS-BASED SELECTION

- Based on general-to-specific modeling.
- Start from general dynamic statistical model (GUM):
- check GUM captures essential characteristics of data
- Then eliminate statistically-insignificant variables,
- to reduce its complexity;
- check validity of reductions by diagnostic tests,
- to ensure congruence of final model
- Test final selection encompasses rival contenders
- Progressive research strategy (PRS) key concept

CONCLUSIONS

ELECTION

Major recent developments in theory and practice of automatic model selection: multi-path searches, encompassing choices impulse saturation non-linearity Autometrics provides powerful model-selection procedure: null rejection frequency close to nominal; power close to starting with LDGP; near unbiased estimates of fit and standard errors; can bias-correct estimated parameters; can handle more variables than observations Turn to the origins of empirical models

REFERENCES

1. Jushan Bai, 1997. "Estimation Of A Change Point In Multiple Regression Models," The Review of Economics and Statistics, MIT Press, vol. 79(4), pages 551-563, November.

2. Perron, P. & Bai, J., 1995. "Estimating and Testing Linear Models with Multiple Structural Changes," Cahiers de recherche 9552, Centre interuniversitaire de recherche en économie quantitative, CIREQ.

3. Jushan Bai & Pierre Perron, 1998. "Estimating and Testing Linear Models with Multiple Structural Changes," Econometrica, Econometric Society, vol. 66(1), pages 47-78, January.

4. Perron, P. & Bai, J., 1995. "Estimating and Testing Linear Models with Multiple Structural Changes," Cahiers de recherche 9552, Universite de Montreal, Departement de sciences economiques.

5. Ninomiya, Yoshiyuki, 2005. "Information criterion for Gaussian change-point model," Statistics & Probability Letters, Elsevier, vol. 72(3), pages 237-247, May.

6. Hansen, Bruce E., 2009. "Averaging Estimators For Regressions With A Possible Structural Break," Econometric Theory, Cambridge University Press, vol. 25(06), pages 1498-1514, December.

7. Andrews, Donald W. K. & Lee, Inpyo & Ploberger, Werner, 1996. "Optimal changepoint tests for normal linear regression," Journal of Econometrics, Elsevier, vol. 70(1), pages 9-38, January.

8. Donald W.K. Andrews & Inpyo Lee & Werner Ploberger, 1992. "Optimal Changepoint Tests for Normal Linear Regression," Cowles Foundation Discussion Papers 1016, Cowles Foundation for Research in Economics, Yale University.

9. Andrews, Donald W K, 1993. "Tests for Parameter Instability and Structural Change with Unknown Change Point," Econometrica, Econometric Society, vol. 61(4), pages 821-56, July.

10. Donald W.K. Andrews, 1990. "Tests for Parameter Instability and Structural Change with Unknown Change Point," Cowles Foundation Discussion Papers 943, Cowles Foundation for Research in Economics, Yale University.

11. Zhongjun Qu & Pierre Perron, 2007. "Estimating and Testing Structural Changes in Multivariate Regressions," Econometrica, Econometric Society, vol. 75(2), pages 459-502, 03.

12. Zhongjun Qu & Pierre Perron, 2005. "Estimating and testing structural changes in multivariate regressions," Boston University - Department of Economics - Working Papers Series WP2005-012, Boston University - Department of Economics.

13. Bai, Jushan, 1997. "Estimating Multiple Breaks One at a Time," Econometric Theory, Cambridge University Press, vol. 13(03), pages 315-352, June.

14. Jushan Bai, 1995. "Estimating Multiple Breaks One at a Time," Working papers 95-18, Massachusetts Institute of Technology (MIT), Department of Economics.

15. Davis, Richard A. & Lee, Thomas C.M. & Rodriguez-Yam, Gabriel A., 2006. "Structural Break Estimation for Nonstationary Time Series Models," Journal of the American Statistical Association, American Statistical Association, vol. 101, pages 223-239, March.

16. Jushan Bai, 1999. "Vector Autoregressive Models with Structural Changes in Regression Coefficients and in Variance-Covariance Matrices," CEMA Working Papers 24, China Economics and Management Academy, Central University of Finance and Economics, revised Oct 2000.

17. Jushan Bai, 2000. "Vector Autoregressive Models with Structural Changes in Regression Coefficients and in Variance-Covariance Matrices," Annals of Economics and Finance, Society for AEF, vol. 1(2), pages 303-339, November.

18. Jushan Bai & Pierre Perron, 2003. "Computation and analysis of multiple structural change models," Journal of Applied Econometrics, John Wiley & Sons, Ltd., vol. 18(1), pages 1-22.

19. Tom Doan, . "RATS programs to replicate examples of Bai-Perron procedure," Statistical Software Components RTZ00008, Boston College Department of Economics.

20. Tom Doan, . "BAIPERRON: RATS procedure to perform Bai-Perron Test for Multiple Structural Changes," Statistical Software Components RTS00013, Boston College Department of Economics.

21. Tom Doan, . "MULTIPLEBREAKS: RATS procedure to perform multiple structural change analysis," Statistical Software Components RTS00138, Boston College Department of Economics.

22. BAI, Jushan & PERRON, Pierre, 1998. "Computation and Analysis of Multiple Structural-Change Models," Cahiers de recherche 9807, Universite de Montreal, Departement de sciences economiques.

23. Bai, Jushan, 1999. "Likelihood ratio tests for multiple structural changes," Journal of Econometrics, Elsevier, vol. 91(2), pages 299-323, August.