*ISSN 1471-0498*


# DEPARTMENT OF ECONOMICS

# DISCUSSION PAPER SERIES

## AUTOMATIC SELECTION FOR NON-LINEAR MODELS

**Jennifer L. Castle and David F. Hendry**

Number 473
January 2010

Manor Road Building, Oxford OX1 3UQ

# Automatic Selection for Non-linear Models

Jennifer L. Castle and David F. Hendry*
Department of Economics, Oxford University.

December 11, 2009

**Abstract**

Our strategy for automatic selection in potentially non-linear processes is: test for non-linearity in the unrestricted linear formulation; if that test rejects, specify a general model using polynomials, to be simplified to a minimal congruent representation; finally select by encompassing tests of specific non-linear forms against the selected model. Non-linearity poses many problems: extreme observations leading to non-normal (fat-tailed) distributions; collinearity between non-linear functions; usually more variables than observations when approximating the non-linearity; and excess retention of irrelevant variables; but solutions are proposed. A returns-to-education empirical application demonstrates the feasibility of the non-linear automatic model selection algorithm *Autometrics*.

*JEL classification:* C51; C22; C87

*Keywords:* Econometric methodology; model selection; *Autometrics*; non-linearity; outliers; returns to education.

## Contents



*Prepared for the *Festschrift* in Honor of Peter Young. We thank participants of the Royal Economic Society Conference 2006, Econometric Society European and Australasian Meetings, 2006, *Journal of Econometrics* Conference, 2007, and the *Arne Ryde* Lectures 2007 for helpful comments and suggestions on an earlier version. Financial support from the ESRC under grant RES 051 270035 is gratefully acknowledged.

# 1 Introduction

It is a pleasure to contribute a chapter on non-linear model selection to a volume in honor of Peter C. Young, who has himself contributed so much to modeling, to understanding and capturing key aspects of non-linearity, and to data basing the choice of which models work in a wide range of important areas in statistics, environmental studies and economics. While we do not also address his interests in forecasting, we share them strongly and have tried to advance that subject in other publications–and as a further objective, trying to establish the general approach adopted here for dynamic, non-stationary processes. We congratulate Peter on his successes to date and look forward to many more.

Economic processes are complicated entities, which are often modeled by linear approximations, leading to possible mis-specification when non-linearity matters. This chapter develops a strategy for selecting non-linear in variables models for cross-section data, following the automatic general-to-specific (*Gets*) multi-path search algorithms of *PcGets* (see Hendry and Krolzig, 2001, which built on Hoover and Perez, 1999), and *Autometrics* within *PcGive* (see Doornik, 2009, and Hendry and Doornik, 2009). The general properties of *Autometrics* model selection are established in Castle, Doornik and Hendry (2009a), multiple breaks are investigated by Castle, Doornik and Hendry (2009b), and an empirical application is provided in Hendry and Mizon (2009). These properties of *Autometrics* can be summarized as follows for a linear static model. When there are $K$ candidate variables, and $k$ of these are relevant, then $\alpha\,(K-k)$ irrelevant variables will be retained on average, where $\alpha$ is the chosen significance level. Because it selects variables ($K$), rather than models ($2^K$), that result continues to hold even when $K$ is greater than the sample size, $N$, provided $N > k$. Also, the $k$ relevant variables will be retained with a probability close to the theoretical t-test powers determined by the non-centralities of their parameters. For example, if $K - k = 100$ and $\alpha = 0.01$, then one irrelevant variable will be retained on average by chance sampling, despite the plethora of candidate variables. Moreover, coefficients with $|\mathsf{t}|$-values greater than about $c_\alpha = 2.6$ will be retained on average. Next, although selection only retains variables whose estimated coefficients have $|\mathsf{t}| \geq c_\alpha$, the resulting selection bias is easily corrected, which greatly reduces the mean-square errors (MSEs) of retained irrelevant variables: see Hendry and Krolzig (2005). Finally, the terminal models found by *Autometrics* will be congruent (well specified), undominated reductions of the initial general unrestricted model (GUM). We will not discuss the details of the multi-path search algorithms that have made such developments feasible, as these are well covered elsewhere (see e.g., Hendry and Krolzig, 2001, Hendry and Doornik, 2009, Doornik, 2009, and Doornik, Hendry and Nielsen, 2009): the reader is referred to those publications for bibliographic perspective on this exciting

and burgeoning new field. The latest version of the model selection algorithm *Autometrics* is likelihood based, so can accommodate discrete variable models such as logit and probit, along with many other econometric specifications, but we focus on non-linear regression analysis here.

Thus, we investigate non-linear modelling as part of a general strategy of empirical model discovery. Commencing with a low-dimensional portmanteau test for non-linearity (see Castle and Hendry, 2009), non-rejection entails remaining with a linear specification, whereas rejection leads to specifying a general non-linear, identified and congruent approximation. Next, the multi-path search procedure seeks a parsimonious, still congruent, non-linear model, and that in turn can be tested against specific non-linear functional forms using encompassing tests (see, e.g., Mizon and Richard, 1986, and Hendry and Richard, 1989), and simplified to them if appropriate.

Since the class is one of non-linear in variables, but linear in parameters, the most obvious approach is to redefine non-linear functions as new variables (e.g., $x_i^2 = z_i$ say), so the model becomes linear but larger, and standard selection theory applies. However, non-linearity *per se* introduces five specific additional problems even in cross sections, solutions to which need to be implemented as follows.

First, determining whether there is non-linearity. The low-dimensional portmanteau test for non-linearity in Castle and Hendry (2009) is applied to the unrestricted linear regression to check whether any non-linear extension is needed. Their test is related to the test for heteroskedasticity proposed by White (1980), but by using squares and cubics of the principal components of the linear variables, the test circumvents problems of high-dimensionality and collinearity, and is not restricted to quadratic departures. Providing there are fewer linear variables, $K$, than about a quarter of the sample size, $N$, the test can accommodate large numbers, $M_K$, of potential non-linear terms, including more than $N$, where for a cubic polynomial:

$$M_K = K\left(K+1\right)\left(K+5\right)/6.$$

If the test does not reject, the usual *Gets* approach is applied to the linear model. Otherwise, a non-linear, or indeed non-constant, model is needed to characterize the evidence, so these possibilities must be handled jointly, as we do below.

Second, including both the linear and non-linear transformations of a variable can generate substantial collinearity, similar to slowly-varying regressors (as in Phillips, 2007). Such collinearity can be problematic for estimation and selection procedures, as the information content of the extra collinear variables is small, yet disrupts existing information attribution. When the additional transformed variables are in fact irrelevant, model selection algorithms may select poorly between the relevant and irrelevant variables, depending on chance sampling. In a sense, automatic algorithms still perform adequately, as they usually keep a 'representative' of the relevant effect. Nevertheless, orthogonality is beneficial for model selection in general, both for that reason, and because deleting small, insignificant coefficients leaves the retained estimates almost unaltered. We use a simple operational de-meaning rule to eliminate one important non-orthogonality prior to undertaking model selection.

Third, non-linear functions can generate extreme outcomes, and the resulting 'fat tails' are problematic for inference and model selection, as the assumption of normality is in-built into most procedures' critical values. Non-linear functions can also 'align' with outliers, causing the functions to be retained spuriously, which can be detrimental for forecasting and policy. Thus, data contamination, outliers and non-linearity interact, so need to be treated together. To do so, we use impulse-indicator saturation (denoted IIS), which adds an indicator for every observation to the candidate regressor set (see Hendry, Johansen and Santos, 2008, and Johansen and Nielsen, 2009) to remove the impact of breaks and extreme

observations in both regressors and regressand, and ensure near normality. Johansen and Nielsen (2009) show that IIS is a robust estimation method, and that despite adding greatly to the number of variables in the search, there is little efficiency loss under the null of no contamination. In the present context, there is also a potentially large gain by avoiding non-linear terms that chance to capture unmodeled outliers, but there are always bound to be more candidate variables for selection than the sample size.

General non-linear functional approximations alone can create more variables than observations. However, building on Hendry and Krolzig (2005), *Autometrics* already handles such situations by a combination of expanding and contracting searches (see Doornik, 2007). Nevertheless, the number of potential regressors, $M_K$, grows rapidly as $K$ increases:

$$\begin{array}{cccccccccccc} K & 1 & 2 & 3 & 4 & 5 & 10 & 15 & 20 & 30 & 40 \\ M_K & 3 & 9 & 19 & 30 & 55 & 285 & 679 & 1539 & 5455 & 12300 \end{array} \tag{1}$$

An additional exponential component adds $K$ more to $M_K$, and impulse-indicator saturation (IIS) adds $N$ more dummies for a sample of size $N$ (below, we use more than $5000$ observations). Selections of such a magnitude are now feasible but lead to the next problem.

The fourth is the related problem of excess retention of linear and non-linear functions and indicators due to a highly over-parameterized GUM. This is controlled by implementing a 'super-conservative' strategy for the non-linear functions, where selection is undertaken at stringent significance levels to control the null rejection frequency. For example, when $M_K + K + N = 8000$ and no variables actually matter, a significance level of $\alpha = 0.001$ would lead on average to $8$ irrelevant retentions, of which $5$ would simply be indicators, which just dummy out their respective observations (so is 99.9% efficient). As discussed in Hendry and Krolzig (2005) and Castle *et al.* (2009b), post-selection bias correction will drive the estimated coefficients of adventitiously retained variables towards the origin, leading to small mean square errors, so is not a problematic outcome from learning that $7992$ of the candidate variables do not in fact matter. Thus the distribution under the null is established as retaining $\alpha\,(M_K + K + N - k)$ chance significant effects when $k$ variables matter.

Finally, non-linearity comprises everything other than the linear terms, so some functional form class needs to be assigned to search across, and that is almost bound to be an approximation in practice. In a cross-section context, polynomials often make sense, so we use that as the basis class. To then implement any economic-theory based information, encompassing tests of the entailed non-linear form against the selected model can be undertaken, and this order of proceeding avoids the potential identification problems that can arise when starting with non-linear-in-parameters models (see Granger and Teräsvirta, 1993). However, we do not focus on that aspect here.

We undertake an empirical study of returns to education for US males, using 1980 census data, applying the proposed non-linear algorithm after finding strong evidence for non-linearity using the Castle and Hendry (2009) test. The log-wage data are non-normal, but we use IIS to obtain an approximation to normality, adding the indicators to a general non-linear GUM, which controls for a wide range of covariates such as education, experience, ability, usual hours worked, marital status, race, etc. The non-linear selection algorithm finds a congruent model in which non-linear functions play a key role in explaining the data.

The structure of the chapter is as follows. Section 2 outlines the non-linear specification procedure to which a model selection algorithm such as *Autometrics* is applied, and details the non-linear functions used, related to the RETINA algorithm in Perez-Amaral, Gallo and White (2003). Section 3 addresses the five intrinsic problems of selecting models that are non-linear in the regressors. First, §3.1 sketches

the non-linearity test, then §3.2 demonstrates the collinearity between linear and non-linear functions, and proposes a solution by simply de-meaning all functions of variables. Third, §3.3 outlines the issue of non-normality, with a Monte Carlo study that highlights the problem of extreme observations for model selection, and explains the application of IIS jointly with selecting variables. Finally, §3.5 discusses the super-conservative strategy to ensure non-linear functions are retained only when there is definite evidence of non-linearity in the data. Section 4 applies the non-linear selection algorithm to a cross section of log wages, modeling the returns to education: there is strong evidence both for non-linearity and outliers that are captured by the algorithm. Finally, Section 5 concludes.

## 2 The non-linear algorithm

Finding a unique non-linear representation of an economic process can be formidable given the complexity of possible local data generating processes (LDGPs, namely the DGP in the space of the variables under analysis). As there are an infinite number of potential functional forms that the LDGP may take, specifying a GUM that nests the unknown LDGP is problematic. Here, we assume the LDGP is given by:

$$y_i = f\left(x_{1,i}, \ldots, x_{k,i}; \boldsymbol{\theta}\right) + \epsilon_i \text{ where } \epsilon_i \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right] \tag{2}$$

for $i = 1, \ldots, N$, with $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Three key concerns for the econometrician are the specification of the functional form, $f(\cdot)$, the identification of $\boldsymbol{\theta}$, and the selection of the potentially relevant variables, $\mathbf{x}_i' = (x_{1,i}, \ldots x_{k,i})$ from an available set of candidates $(x_{1,i}, \ldots x_{K,i})$ where $K \geq k$.

The initial GUM includes all $K$ candidates, in some non-linear form $g(\cdot)$:

$$y_i = g\left(x_{1,i}, \ldots, x_{K,i}; \boldsymbol{\phi}\right) + v_i \text{ where } v_i \sim \mathsf{IN}\left[0, \sigma_v^2\right] \tag{3}$$

Economic theory, past empirical and historical evidence, and institutional knowledge all inform the specification of the variables in the GUM and their functional form. If the initial specification is too parsimonious, relevant variables may be omitted leading to a mis-specified final model. Theory often has little to say regarding the functional-form specification, so an approximating class is required from the infinite possibilities of non-linear functions. Many non-linear models–including smooth-transition regressions, regime-switching models, neural networks and non-linear equations–can be approximated by Taylor expansions, so polynomials form a flexible approximating class for a range of possible LDGPs.

A Taylor-series expansion of (3) around zero results in (see e.g., Priestley, 1981):

$$g\left(x_{1,i}, \ldots x_{K,i}; \boldsymbol{\phi}\right) = \phi_0 + \sum_{j=1}^{K} \phi_{1,j} x_{j,i} + \sum_{j=1}^{K} \sum_{l=1}^{j} \phi_{2,j,l} x_{j,i} x_{l,i} + \sum_{j=1}^{K} \sum_{l=1}^{j} \sum_{m=1}^{l} \phi_{3,j,l,m} x_{j,i} x_{l,i} x_{m,i} + \cdots \tag{4}$$

While motivating the use of polynomial functions, (4) demonstrates how quickly the number of parameters increases as (1), shows, exacerbated when $N$ impulse indicators are added. Polynomial functions are often used in economics because of Weierstrass's approximation theorem whereby any continuous function on a closed and bounded interval can be approximated as closely as one wishes by a polynomial, so if $x \in [a, b]$, for any $\eta > 0$ there exists a polynomial $p(x) \in [a, b]$ such that $|f(x) - p(x)| < \eta \; \forall x \in [a, b]$. However, the goodness of the approximation is unknown *a priori* in any given application, although it can be evaluated by testing against a higher-order formulation and by mis-specification tests.

A wide range of non-linear functions has been considered to approximate (2), including various orthogonal polynomials, such as Hermite, Fourier series, asymptotic series (see e.g., Copson, 1965), squashing functions (see White, 1992), and confluent hypergeometric functions (see Abadir, 1999). Here, we include cubic functions, as these are sign-preserving (so could represent, say, non-linear demand or price responses), and add to the flexibility of the transformations, potentially approximating ogives. We do not include exponential components, although the most general test in Castle and Hendry (2009) does. If the LDGP contains an inverse polynomial function, the polynomial will detect this form of non-linearity due to the high correlation between the variable and its inverse. Although the selected model might then be prone to misinterpretation, we consider the polynomial approximation to be an intermediate stage before testing parsimonious encompassing of by a specific functional form.

Many other functional forms have been proposed in the literature: for example, RETINA (see Perez-Amaral *et al.*, 2003) uses the transformations (see Castle, 2005):

$$\sum_{j=1}^{K}\sum_{l=1}^{K}\beta_{j,l}x_{j,i}^{\lambda_1}x_{l,i}^{\lambda_2} \quad \text{for } \lambda_1,\lambda_2=-1,0,1 \tag{5}$$

Although we exclude inverses, squared inverses, and ratios due to their unstable behavior potentially creating outliers, and adequate correlations with levels (4) includes the remaining terms. Also, for example, logistic smooth transition models (LSTAR: see e.g., Teräsvirta, 1994) will be approximated by the third-order Taylor expansion given by (4). Thus, (4) approximates or nests many non-linear specifications.

While (4) already looks almost intractable, the inclusion of more variables than observations does not in fact make it infeasible for an automatic algorithm, enabling considerable flexibility when examining non-linear models despite the number of potential regressors being large. When $N > K$, the *Gets* approach is to specify a GUM that nests the LDGP in (2), to ensure the initial formulation is congruent. As $K > N$, both expanding and contracting searches are required, and congruence can only be established after some initial simplification to make it feasible to estimate the remaining model. Here, we propose using the general formulation:

$$y_i=\phi_0+\sum_{j=1}^{K}\phi_{1,j}x_{j,i}+\sum_{j=1}^{K}\sum_{l=1}^{j}\phi_{2,j,l}x_{j,i}x_{l,i}+\sum_{j=1}^{K}\sum_{l=1}^{j}\sum_{m=1}^{l}\phi_{3,j,l,m}x_{j,i}x_{l,i}x_{m,i}+\sum_{j=1}^{N}\delta_j 1_{\{j=i\}}+u_i \tag{6}$$

with $K$ potential linear regressors, $\mathbf{x}_i$, where $1_{\{j=i\}}$ is an indicator for the $i$th observation.

## 3 Problems when selecting non-linear models

There are five problems that arise when selecting from a GUM that consists of a large set of polynomial regressors as in (6). These problems include first detecting non-linearity (§3.1), reducing collinearity (§3.2), handling non-normality (§3.3) leading to more variables than observations (§3.4), and avoiding potential excess retention of irrelevant regressors (§3.5). Solutions to all of these problems are now proposed, confirming the feasibility of our non-linear model selection strategy.

### 3.1 Testing for non-linearity

The LDGP in (2) has $k$ relevant and $K-k$ irrelevant variables when $f(\cdot)$ is linear. The first stage is to apply the test for non-linearity in Castle and Hendry (2009) to see if it is viable to reduce (6) directly to:

$$y_i = \sum_{j=1}^{K} \beta_j x_{j,i} + \sum_{j=1}^{N} \delta_j 1_{\{j=i\}} + e_i \tag{7}$$

If outliers are likely to be problematic, IIS could first be applied to (7) to ascertain any major discrepancies, leading to say $r$ indicators being retained (see §3.4):

$$y_i = \sum_{j=1}^{K} \beta_j x_{j,i} + \sum_{j=1}^{r} \delta_j 1_{\{j=i\}} + e_i \tag{8}$$

When $\mathbf{x}_i$ denotes the set of linear candidate regressor variables, to calculate their principal components, denoted $\mathbf{z}_i$, define $\mathbf{H}$ and $\mathbf{\Lambda}$ as the eigenvectors and eigenvalues of $N^{-1}\mathbf{X}'\mathbf{X}$, such that:

$$\mathbf{z}_i = \mathbf{\Lambda}^{-\frac{1}{2}} \left[ \left( \mathbf{H}'\mathbf{x}_i \right) - \overline{\left( \mathbf{H}'\mathbf{x}_i \right)} \right] \tag{9}$$

Let $z^2_{j,i} = w_{j,i}$ and $z^3_{j,i} = s_{j,i}$, then the test for non-linearity is the $\mathsf{F}$-test of $\mathsf{H}_0$: $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 = 0$ in:

$$y_i = \beta_0 + \boldsymbol{\beta}_1'\mathbf{x}_i + \boldsymbol{\beta}_2'\mathbf{w}_i + \boldsymbol{\beta}_3'\mathbf{s}_i + \sum_{j=1}^{r} \delta_j 1_{\{j=i\}} + \epsilon_i \tag{10}$$

where $r = 0$ if IIS is not first applied. If the $\mathsf{F}$-test does not reject, the GUM is taken to be linear, and the usual selection algorithm is applied to select the relevant regressors. Conversely, if the test rejects, non-linearity is established at the selected significance level, so the remaining four problems need resolving for a viable approach. If IIS was not applied, non-linearity is only contingently established, as it may be proxying outliers as §3.3 shows.

### 3.2 Collinearity

Multicollinearity was first outlined by Frisch (1934) within the context of static general-equilibrium linear relations. Confluence analysis was developed to address the problem, although that method is not in common practice now (see Hendry and Morgan, 1989). The definition of collinearity has shifted over the years, but for an $N \times K$ regressor matrix $\mathbf{X}$, we can define perfect collinearity as $|\mathbf{X}'\mathbf{X}| = 0$, and perfect orthogonality as a diagonal $(\mathbf{X}'\mathbf{X})$ matrix. Since collinearity is not invariant under linear transformations, it is difficult to define a 'degree of collinearity', as a linear model is equivariant under linear transformations, and so the same model could be defined by various isomorphic representations, which nevertheless deliver very different inter-correlations. Hence, collinearity is a property of the parametrization of the model, and not the variables *per se*. Moreover, $|\mathbf{X}'\mathbf{X}| = 0$ whenever $N > K$ anyway.

Nevertheless non-linear transformations can generate substantial collinearity between the linear and non-linear functions. We consider a simple case in which we add the irrelevant transformation $f(w_i) = w_i^2$ to a linear model in $w_i$. This polynomial transform is common in economics: see section 4 for an

empirical application. The degree of collinearity varies as the statistical properties of the process vary: collinearity between $w_i$ and $w_i^2$ is zero when $\mathsf{E}[w_i] = 0$, but dramatically increases to almost perfect collinearity as $\mathsf{E}[w_i] = \mu$ increases. To see that, consider the DGP given by the linear conditional relation:

$$y_i = \beta w_i + e_i = 0 + \beta w_i + 0 w_i^2 + \epsilon_i \tag{11}$$

where $\epsilon_i \sim \mathsf{IN}\left[0, \sigma_\epsilon^2\right]$ with $i = 1, \ldots, N$, and:

$$w_i \sim \mathsf{IN}\left[0, \sigma_w^2\right] \tag{12}$$

Since (11) is equivariant under linear transformations, in that both the dependent variable and the error process are unaffected, it can also be written for $z_i = w_i + \mu$ as:

$$\begin{aligned} y_i &= -\beta\mu + \beta\left(w_i + \mu\right) + 0\left(w_i + \mu\right)^2 + \epsilon_i \\ &= -\beta\overline{z} + \beta z_i + 0 z_i^2 + \epsilon_i \\ &= 0 + \beta\left(z_i - \overline{z}\right) + 0\left(z_i - \overline{z}\right)^2 + \epsilon_i. \end{aligned} \tag{13}$$

Correspondingly, there are three models, namely, the original zero-mean case:

$$y_i = \beta_0 + \beta_1 w_i + \beta_2 w_i^2 + u_i \tag{14}$$

the non-zero-mean case:

$$y_i = \gamma_0 + \gamma_1 z_i + \gamma_2 z_i^2 + u_i \tag{15}$$

and the transformed zero-mean case:

$$y_i = \lambda_0 + \lambda_1 z_i + \lambda_2\left(z_i^2 - \overline{z^2}\right) + u_i \tag{16}$$

where $\overline{z^2}$ is the sample mean of $z_i^2$.

First, letting $\mathbf{X}$ denote the general regressor matrix, for (15) with a non-zero mean:

$$\begin{aligned} \mathsf{E}\left[N^{-1}\mathbf{X}'\mathbf{X}_{(\mu)}\right] &= \mathsf{E}\left[\begin{pmatrix} 1.0 & \overline{z} & \overline{z^2} \\ \overline{z} & N^{-1}\sum z_i^2 & N^{-1}\sum z_i^3 \\ \overline{z^2} & N^{-1}\sum z_i^3 & N^{-1}\sum z_i^4 \end{pmatrix}\right] \\ &= \begin{pmatrix} 1.0 & \mu & \mu^2 + \sigma_w^2 \\ \mu & \mu^2 + \sigma_w^2 & \mu^3 + 3\mu\sigma_w^2 \\ \mu^2 + \sigma_w^2 & \mu^3 + 3\mu\sigma_w^2 & 3\sigma_w^4 + \mu^4 + 6\mu^2\sigma_w^2 \end{pmatrix} \end{aligned} \tag{17}$$

with the inverse:

$$\left(\mathsf{E}\left[N^{-1}\mathbf{X}'\mathbf{X}_{(\mu)}\right]\right)^{-1} = \frac{1}{2\sigma_w^6}\begin{pmatrix} \mu^4\sigma_w^2 + 3\sigma_w^6 & -2\mu^3\sigma_w^2 & \mu^2\sigma_w^2 - \sigma_w^4 \\ -2\mu^3\sigma_w^2 & 2\sigma_w^4 + 4\mu^2\sigma_w^2 & -2\mu\sigma_w^2 \\ \mu^2\sigma_w^2 - \sigma_w^4 & -2\mu\sigma_w^2 & \sigma_w^2 \end{pmatrix} \tag{18}$$

There is substantial collinearity between the variables, except for the squared term, which is irrelevant in the DGP. As $\mu$–an incidental parameter here–increases, $\mathsf{E}[N^{-1}\mathbf{X}'\mathbf{X}_{(\mu)}]$ tends towards singularity, and for $\sigma_w^2 = 1$, the ratio $R$ of the largest to the smallest eigenvalues in (18) grows dramatically from $R = 5.83$

when $\mu = 0$ through $R = 60223$ for $\mu = 4$ to $R = 5.6 \times 10^7$ when $\mu = 10$. Note that age enters some regressions below, often with a mean above 20.

Next, in the zero-mean model in (14):

$$\mathsf{E}\left[N^{-1}\mathbf{X}'\mathbf{X}_{(0)}\right] = \mathsf{E}\left[\begin{pmatrix} 1.0 & \overline{w} & \overline{w^2} \\ \overline{w} & N^{-1}\sum w_i^2 & N^{-1}\sum w_i^3 \\ \overline{w^2} & N^{-1}\sum w_i^3 & N^{-1}\sum w_i^4 \end{pmatrix}\right] = \begin{pmatrix} 1.0 & 0.0 & \sigma_w^2 \\ 0.0 & \sigma_w^2 & 0.0 \\ \sigma_w^2 & 0.0 & 3\sigma_w^4 \end{pmatrix} \tag{19}$$

so the inverse is:

$$\left(\mathsf{E}\left[N^{-1}\mathbf{X}'\mathbf{X}_{(0)}\right]\right)^{-1} = \frac{1}{2\sigma_w^6}\begin{pmatrix} 3\sigma_w^6 & 0 & -\sigma_w^4 \\ 0 & 2\sigma_w^4 & 0 \\ -\sigma_w^4 & 0 & \sigma_w^2 \end{pmatrix} \tag{20}$$

There is no collinearity between $w_i$ and $w_i^2$ although there is an effect on the intercept, but this does not cause a problem for either estimation or a selection algorithm.

Finally, in the transformed zero-mean model in equation (16):

$$\left(\mathsf{E}\left[N^{-1}\mathbf{X}'\mathbf{X}_{(0,0)}\right]\right)^{-1} = \frac{1}{3\sigma_w^6}\begin{pmatrix} 3\sigma_w^6 & 0.0 & 0.0 \\ 0.0 & 3\sigma_w^4 & 0.0 \\ 0.0 & 0.0 & \sigma_w^2 \end{pmatrix} \tag{21}$$

Thus, a near orthogonal representation can be achieved simply by taking deviations from means, which re-creates the specification in terms of the original variables $w_i$ and $w_i^2$ as $z_i = w_i + \mu$ where $\mathsf{E}[\overline{z}] = \mu$ and $\mathsf{E}[\overline{z^2}] = \mu^2 + \sigma_w^2$:

$$\mathsf{E}\left[N^{-1}\mathbf{X}'\mathbf{X}_{(\overline{\mu})}\right] = \begin{pmatrix} 1.0 & 0.0 & 0.0 \\ 0.0 & \sigma_w^2 & 2\mu\sigma_w^2 \\ 0.0 & 2\mu\sigma_w^2 & 3\sigma_w^6 - \sigma_w^4 + 4\mu^2\sigma_w^2 \end{pmatrix} \tag{22}$$

with the inverse:

$$\left(\mathsf{E}\left[N^{-1}\mathbf{X}'\mathbf{X}_{(\overline{\mu})}\right]\right)^{-1} = \frac{1}{\sigma_w^6\left(3\sigma_w^2 - 1\right)}\begin{pmatrix} 3\sigma_w^8 - \sigma_w^6 & 0.0 & 0.0 \\ 0.0 & 3\sigma_w^6 - \sigma_w^4 + 4\mu^2\sigma_w^2 & -2\mu\sigma_w^2 \\ 0.0 & -2\mu\sigma_w^2 & \sigma_w^2 \end{pmatrix} \tag{23}$$

Taking deviations from sample means delivers a reduction in collinearity, which is particularly marked for the intercept, but worse for the linear term $(z_i - \overline{z})$. Again the irrelevant squared term 'benefits'. To remove the collinearity, first de-mean $z_i$, then also de-mean $z_i^2$. The linear term remains $(z_i - \overline{z})$, but the squared term becomes $(z_i - \overline{z})^2 - [\mathsf{E}(z_i - \overline{z})]^2$ which will result in a model that is identical to equation (16). Double de-meaning thus removes the collinearity generated by the non-zero mean, and Monte Carlo evidence confirms this is an effective solution to mean-induced collinearity.

A non-linear selection strategy should automatically double de-mean the generated polynomial functions prior to formulating the GUM. Two caveats apply. First, the orthogonalizing rules will not remove all collinearity between higher-order polynomials. We considered orthogonalizing using the Choleski method (see Rushton, 1951), but double de-meaning removed enough collinearity to ensure the *Autometrics* selection had the appropriate properties. Second, any information contained in the intercepts of the explanatory variables will be removed, although there is rarely a theory of the intercept when developing econometric models, especially for cross-section data.

### 3.3 Non-normality

Normality is a central assumption for inference, as conventional critical values tend to be used, so null rejection frequencies would be incorrect for non-normality. Normality tends to be even more vital for selection, when many decisions are made. In non-linear models, normality is essential, as problems arise when fat-tailed distributions result in extreme observations, as there is an increased probability that non-linear functions will align with extreme observations, effectively acting as indicators and therefore being retained too often (see e.g., Castle, Fawcett and Hendry, 2010).

We now show by a Monte Carlo example that non-normal variables pose similar problems. Consider these DGPs for four variables:

$$x_{i,t} = \epsilon_{i,t} \quad \epsilon_{i,t} \sim \mathsf{IN}\left[0,1\right] \quad \text{for } i = 1,\ldots,4. \tag{24}$$

We generate non-linear functions given by the inverses of these normal distributions (as in RETINA):

$$x_{i,t}^{-1} = \frac{1}{x_{i,t}}. \tag{25}$$

The GUM contains twenty irrelevant variables given by:

$$x_{1,t}^{-1} = \rho_0 + \sum_{i=1}^{4} \rho_i x_{1,t-i}^{-1} + \sum_{j=2}^{4}\sum_{m=0}^{4} \rho_{j,m} x_{j,t-m}^{-1} + \epsilon_t. \tag{26}$$

Then selecting from (26) leads to |t|-values as large as 19 for variables with zero non-centralities. Such a variable would unequivocally, but incorrectly, be retained as a DGP variable. On average, two of the twenty irrelevant regressors are retained at the 1% significance level. This implies that a fat-tailed distribution would have a null rejection frequency of 10% at the 1% significance level. If the dependent variable is $x_{i,t}$ rather than $x_{i,t}^{-1}$, the retention probabilities are correct as normality results. Non-normal errors can also pose a similar problem (see Castle *et al.*, 2009b). Hence, the problem of model selection is exacerbated by the inclusion of non-linear functions, such as inverses, which generate extreme observations.

### 3.4 Impulse-indicator saturation

Hendry *et al.* (2008) propose the use of impulse-indicator saturation to detect and remove outliers and breaks, utilizing the fact that *Autometrics* can handle more variables than observations. Here the aim is to ensure that the selection process will not overly favor non-linear functions that chance to capture outliers. The modeling procedure generates impulse indicators for every observation, $1_{\{i=s\}} \forall s$. The indicators are divided into $J$ subsets, which form the initial GUMs (including an intercept) and *Autometrics* selects the significant indicators from each subset, which are then stored as terminal models. The joint model is formulated as the union of the terminal models and *Autometrics* re-selects the indicators. Under the null that there are no outliers, $\alpha N$ indicators will be retained on average for a significance level $\alpha$. Johansen and Nielsen (2009) show that the cost of testing for the significance of $N$ indicators under the null is low for small $\alpha$: for example, when $\alpha = 1/N$, only one observation is 'removed' on average. Also, Castle *et al.* (2009b) show that IIS alleviates fat-tailed draws, and allows near-normal inference, important both during search and for the post-selection bias correction which assume normality.

Impulse-indicator saturation also overcomes the problem of 'undetectable' outliers. One concern with non-linearity is that it is difficult to distinguish between extreme observations that are outliers or data contamination and extreme observations that are due to the non-linearity in the data. Non-linear functions can 'hide' outliers by fitting to the extreme values, or conversely, methods that remove extreme observations could be in danger of removing the underlying non-linearity that should be modeled. IIS avoids this problem by including all potentially relevant variables as well as indicators for all observations in the initial GUM, effectively applying IIS to the residuals of the model as opposed to the dependent variable itself. Removing the extreme observations in conjunction with selecting the non-linear functions avoids both problems of removing observations that generate the non-linearity and finding spurious non-linearity that merely captures outliers.

In fact the empirical example does not carry out the strategy precisely as proposed here because the distributions transpired to be so highly non-normal, specifically very badly skewed. Since there were more variables (including indicators) than observations, initial selection inferences based on subsets of variables could be distorted by that skewness. Thus, we added a stage of pre-selecting indicators to 'normalize' the dependent variable. Johansen and Nielsen (2009) show the close relationship of IIS to robust statistics: both can handle data contamination and outliers, and IIS appears to be a low cost way of doing so. Thus, in the spirit of robust statistics, we sought the sub-sample that would be near normal, representing the most discrepant observations by indicators rather than dropping them, so this was only a transient stage. Those indicators are then retained as if they were additional regressors. If the indicators are essential, then better initial selection inferences will ensue, and if they really are not needed, as there were no outliers after the non-linear terms were included, then they should drop out during selection.

### 3.5 Super-conservative strategy

Irrelevant non-linear functions that are adventitiously retained are likely to be detrimental to modeling and forecasting, making such models less robust than linear models, by 'amplifying' changes in collinearity between regressors (see e.g., Clements and Hendry, 1998), and location shifts within the equation or in any retained irrelevant variables. Hence, non-linear functions should only be retained if there is strong evidence. Given the possible excess retention of irrelevant functions due to the large number of potential non-linear functions in the candidate set, much more stringent critical values must be used for the non-linear, than linear, functions during multi-path searches. Critical values should also increase with the number of functions included in the model, and with the sample size, although as with all significance levels, the choice can also depend on the preferences of the econometrician and the likely uses of the resulting model. Parsimonious encompassing of the feasible GUM by the final selected model helps control the performance of the selection algorithm: see Doornik (2008).

A potential problem could arise if the selection procedure eliminated all non-linear functions, contradicting the results of the non-linearity test: it is feasible that the ellipsoid for a joint test at a looser significance level does not include the origin, whereas the $p$-value hyper-square from individual tests at a tighter significance level does. This can be avoided by then repeating the multi-stage strategy with tests undertaken at consecutively looser significance levels. Rules for the super-conservative strategy could be similar to those implemented for the Schwarz information criterion (see Campos, Hendry and Krolzig, 2003), so the selection strategy should deliver an undominated, congruent, specific non-linear model that parsimoniously encompasses the feasible GUM.

We have now resolved the main problems likely to distort selection for a non-linear model, relative

to what is known about its performance in linear settings, so now apply the approach in *Autometrics* to empirically modeling the returns to education.

# 4 Empirical Application: Returns to Education

A natural application of the non-linear algorithm is returns to education. The literature is replete with empirical studies: see, *inter alia*, Garen (1984), Harmon and Walker (1995) and Altonji and Dunn (1996). We focus on a one-factor model, where education is summarized as a single measure defined by years of schooling, in keeping with the homogeneous returns literature of Griliches (1977) and Card (1999). We do not allow for unobserved heterogeneity, capturing heterogeneity through the conditioning variables, following Dearden (1999). There are a range of estimation procedures commonly used, including instrumental variables, control functions and matching methods (see Blundell, Dearden and Sianesi, 2005, for an overview), all of which have been developed to mitigate the biases induced by least-squares estimation. There are 3 sources of biases in a least-squares regression of wage on schooling:
(i) the ability bias, where there is a correlation between the length of schooling and an individual's inherent, but unobserved, ability;
(ii) the returns bias, where the marginal return is correlated with the length of schooling; and
(iii) measurement-error bias due to incorrect measurement of the schooling variable.
In our simple one-factor model, these biases are likely to be small, and Card (1999) argues that there is some evidence that the biases balance out, resulting in near consistent OLS estimates of the returns' coefficient. In order to reduce the biases it is important to include many control variables that can capture omitted factors. Since the functional forms cannot be deduced from theory in this context, a non-linear model must be postulated and so an automatic selection algorithm is a natural tool to use.

We use data from the 1980 US census, based on a random draw of 0.01% of the population of US males in employment, resulting in $5173$ observations. Wage income has been top coded at $\$75,000$, resulting in $204$ observations that are truncated. Figure 1 records the density and distribution of log wages ($w_i$) with their Gaussian reference counterparts. Normality is strongly rejected for $w$ as $\chi^2(2) = 1018.0^{**}$, with substantial skewness in the left tail. Many studies have considered alternative distributions to the log-normal including the Pareto, Champernowne and inverse Gaussian: see Staehle (1943), Lehergott (1959), Harrison (1981) and Ahmed (2007). Instead, we apply IIS as outlined in section 3.4. Table 1 records summary statistics for wages and the covariates.

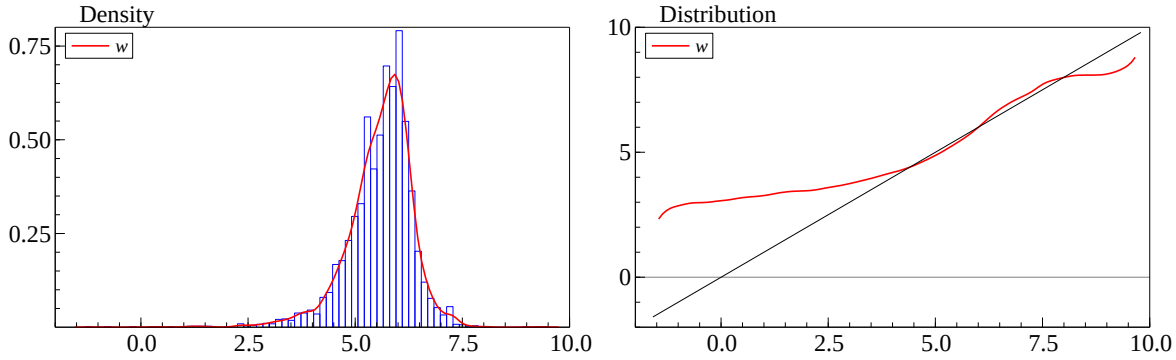


Figure 1: Distribution of log wages

| Variable | Label | Definition | Mean | Variance | Min | Max |
|---|---|---|---|---|---|---|
| Wage | $w$ | logs | 5.58 | 0.59 | -1.24 | 9.47 |
| Experience | $exp$ | Age$-$years education$-6$ | 18.24 | 181.25 | -3 | 57 |
| Education | $edu$ | Grade completed (21 categories) | 12.64 | 9.67 | 0 | 20 |
| Usual hours worked | $hrs$ | Log ave. hours worked in 1979 | 3.70 | 0.11 | 0 | 4.60 |
| Metropolitan status | $met$ | City/rural (5 categories) | 2.27 | 1.56 | 0 | 4 |
| Race | $race$ | (9 categories) | 1.19 | 0.50 | 1 | 7 |
| State | $state$ | FIPS code (62 categories) | 28.52 | 242.77 | 1 | 56 |
| No. of own children | $child$ | in household | 1.01 | 1.69 | 0 | 9 |
| Marital status | $mar$ | (6 categories) | 2.42 | 4.52 | 1 | 6 |
| Educational attainment | $attain$ | (9 categories) | 6.97 | 3.12 | 1 | 9 |

Table 1: Potential explanatory variables

### 4.1 Fitting the theory model

The standard reduced-form model of returns to education is the Mincer regression (Mincer, 1958, 1974):

$$w_i = \beta_0 + \beta_1 edu_i + \beta_2 exp_i + \beta_3 exp_i^2 + u_i \tag{27}$$

where $\beta_1$ measures the 'rate of return to education' which is assumed to be the same for all education levels, and $\mathsf{E}[u_i|edu_i, exp_i] = 0$. In practice, conditioning on additional covariates reduces the impact of omitted variable bias. Here, the results for the augmented Mincer regression are:

$$\begin{aligned} w_i &= \underset{(0.12)}{2.85} + \underset{(0.008)}{0.067} edu_i + \underset{(0.003)}{0.045} exp_i - \underset{(0.00006)}{0.0007}\, exp_i^2 + \underset{(0.014)}{0.003} attain_i + \underset{(0.029)}{0.408} hrs_i \\ &\quad + \underset{(0.007)}{0.043} met_i - \underset{(0.013)}{0.050} race_i - \underset{(0.0006)}{0.002}\, state_i + \underset{(0.009)}{0.019} child_i - \underset{(0.006)}{0.047} mar_i \\ \mathsf{R}^2 &= 0.288 \quad \widehat{\sigma} = 0.651 \quad \chi^2(2) = 1947.7^{**} \quad SC = 1.997 \quad N = 5173 \\ &\quad \mathsf{F}_{\mathsf{het}}(19, 5142) = 4.59^{**} \quad \mathsf{F}_{\mathsf{reset}}(1, 5161) = 4.64^{*} \end{aligned} \tag{28}$$

In (28), $\mathsf{R}^2$ is the squared multiple correlation, $\widehat{\sigma}$ is the residual standard deviation, *SC* is the Schwarz criterion (see Schwarz, 1978), and coefficient standard errors are shown in parentheses. The diagnostic tests are of the form $\mathsf{F}_{\mathsf{j}}(k, T-l)$ which denotes an approximate F-test against the alternative hypothesis j for: heteroskedasticity ($\mathsf{F}_{\mathsf{het}}$: see White, 1980) and the RESET test ($\mathsf{F}_{\mathsf{reset}}$: see Ramsey, 1969); and a chi-square test for normality ($\chi^2_{\mathsf{nd}}(2)$: see Doornik and Hansen, 2008). $^{*}$ and $^{**}$ denote rejection at 5% and 1% respectively.

The model shows a positive *ex post* average rate of return to education of 7% which is broadly in line with the Mincer regression results in Heckman, Lochner and Todd (2006, Table 2) although these are slightly higher at 10-13% as they consider separate regressions for blacks and whites, whereas we take a random sample of the population and condition on a race variable that includes 9 separate categories. We also condition on a further 6 additional explanatory variables to control for omitted variable bias. The economic theory leads to a relatively poor fit ($R^2 = 29\%$), and does not capture well the behavior of the observed data as the model fails mis-specification tests for normality, heteroskedasticity and the RESET test for functional form. Despite poor model specification, the elasticity signs are 'correct', with positive returns to education and experience and an earnings profile that is concave with a significant negative estimated coefficient for experience squared ($\mathsf{t}_{exp^2} = -11$).

## 4.2 Theory equation with IIS

Given the poor performance of the theory model, and the highly significant non-normality test statistic, we next introduce IIS into the specification, using a 0.001 significance level. The resulting model is:

$$\begin{aligned} w_i &= \underset{(0.097)}{3.58} + \underset{(0.006)}{0.063} edu_i + \underset{(0.002)}{0.039} exp_i - \underset{(0.00004)}{0.0006}\, exp_i^2 - \underset{(0.010)}{0.007} attain_i + \underset{(0.023)}{0.268} hrs_i \\ &\quad + \underset{(0.005)}{0.039} met_i - \underset{(0.009)}{0.048} race_i - \underset{(0.0004)}{0.001}\, state_i + \underset{(0.006)}{0.013} child_i - \underset{(0.004)}{0.039} mar_i + 301 \text{ indicators} \\ \mathsf{R}^2 &= 0.670 \quad \widehat{\sigma} = 0.457 \quad \chi^2(2) = 194.2^{**} \quad SC = 1.725 \quad N = 5173 \\ &\quad \mathsf{F}_{\mathsf{het}}(19, 4852) = 5.439^{**} \quad \mathsf{F}_{\mathsf{reset}}(1, 4860) = 0.945 \end{aligned} \tag{29}$$

IIS does not remove the heteroskedasticity found in (28) (note that the test for heteroskedasticity excludes the indicators from the variable set), which suggests that an alternative functional form should be sought. The RESET test indicates that there is no functional form mis-specification, although the RESET test including squares and cubics rejects at the 5% significance level ($\mathsf{F}_{\mathsf{reset23}}(2, 4859) = 4.29[0.014]^{*}$); we will see if we can improve on the functional-form specification in section 4.3.[1] The normality test still fails, but the statistic value is vastly reduced. At a significance level of 0.1%, with 5173 observations, 5 variables will be retained on average under the null, and t-statistics of approximately 3.3 or greater would be retained under normality. Autometrics finds 301 indicators (less than 6% of observations) and this greatly reduces non-normality (excluding the covariates, the test for normality after IIS is $\chi^2(2) = 77.17^{**}$). The test is only an indication, as there is a mass at zero due to the indicators, although Hendry and Santos (2005) show that forming indexes of the indicators can avoid this problem. Figure 2 records the density and QQ plot of log wages once the indicators have been included: there is some deviation from the normal distribution in the tails with the distribution falling outside the pointwise asymptotic 95% standard error bands.

We also applied IIS at $\alpha = 0.05\%$ and $\alpha = 0.01\%$, which would imply that under the null of no outliers we would retain 2.5 and 0.5 of an indicator on average. The resulting Mincer regressions are similar to (29) with 58 and 17 indicators retained.

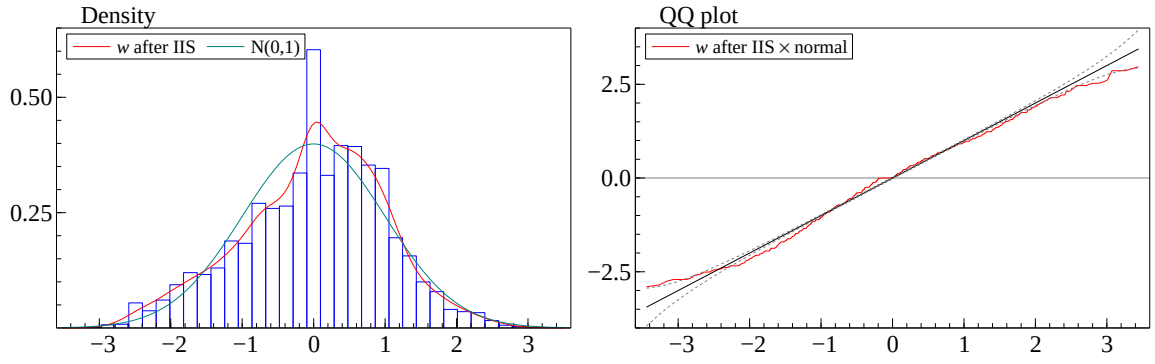


Figure 2: Log wages adjusted for extreme observations

[1] p-values shown in brackets.

### 4.3 Non-linear models

In this section, we extend the Mincer regression in (27) to allow for non-linearities that may enter other than through the experience squared term. We apply the non-linear algorithm presented in section 2, first without IIS and then with IIS to assess the importance of removing outliers.

#### 4.3.1 Testing non-linearity

The first stage of the algorithm is to test for non-linearity using the test proposed by Castle and Hendry (2009). Here $\mathbf{x}_i' = \{exp_i, edu_i, hrs_i, met_i, race_i, state_i, child_i, mar_i, attain_i\}$, so the regressors are a combination of discrete and continuous variables with very different ranges, but principal components standardize the linear combinations. We apply IIS to the linear model in which we fix the linear regressors in the model, i.e. do not select over them, and apply model selection to the impulse-indicators, which is equivalent to applying IIS to the residuals after conditioning on the linear regressors. We retain $r = 316$ indicators ($\mathsf{F}(316, 4847) = 19.09[0.00]^{**}$). We then compute the non-linearity test (10) based on (9). The test statistic, $\mathsf{F}(18, 4829) = 20.15[0.00]^{**}$, strongly rejects the null hypothesis of linearity. Given the strong evidence for a squared experience term in (28) and (29), the test may seem redundant, but we wish to illustrate the general approach in action. In many applications, theory does not provide such a direct non-linear functional-form specification, so there is value in confirming the need for a non-linear specification in advance of model selection to avoid over-parameterizing the GUM with non-linear functions when they are not required.

#### 4.3.2 Modeling non-linearity without IIS

We form the non-linear GUM given by (6), but excluding the impulse indicators, which results in 220 regressors (we also exclude ratios and inverses as highly collinear, and as some variables are discrete with realizations of 0, resulting in numerical problems: RETINA naturally excludes such ratios and inverses). The resulting model nests the Mincer regression (28). All functions are double de-meaned as in section 3.2. The GUM equation standard error is $\widehat{\sigma}_{GUM} = 0.631$. Selection is undertaken using *Autometrics* at the 0.1% significance level, and equation (30) reports the selected model.

$$\begin{aligned} w_i &= \underset{(0.219)}{2.48} + \underset{(0.004)}{0.076} edu_i + \underset{(0.001)}{0.018} exp_i - \underset{(0.00007)}{0.0008}\, exp_i^2 + \underset{(0.056)}{0.382} hrs_i \\ &\quad + \underset{(0.018)}{0.134} met_i + \underset{(0.013)}{0.083} child_i + 48 \text{ non-linear variables} \\ \mathsf{R}^2 &= 0.334 \quad \widehat{\sigma} = 0.632 \quad \chi^2(2) = 1820.6^{**} \quad SC = 2.003 \quad N = 5173 \\ &\quad \mathsf{F}_{\mathsf{het}}(103, 5014) = 1.356^{*} \quad \mathsf{F}_{\mathsf{reset}}(1, 5117) = 10.09^{**} \end{aligned} \tag{30}$$

Education, experience and experience squared are retained with the correct signs and are highly significant, although the coefficient on experience is smaller due to additional non-linear functions of experience that are retained. There is a small improvement in fit compared to (28) from an $\mathsf{R}^2$ of 29% to 33%, but again the model fails the diagnostic tests and selection using critical values based on the normal distribution is clearly violated. Further, 48 additional non-linear variables are retained, possibly representing the problem of over-fitting when outliers are not accounted for, which could lead to poor predictions. We next consider a model that includes both the non-linear functions and IIS.

### 4.3.3 Modeling non-linearity with IIS

The previous regressions demonstrate that both augmenting the Mincer regression with additional non-linear functions and applying IIS to account for outliers are necessary but insufficient steps on their own in developing a theory-consistent model that also captures the key characteristics of the data. Instead of applying both jointly, we add a preliminary step in which IIS is first applied by itself to the linear model (7) to eliminate the most extreme observations: from section 4.3.1 we find $r = 316$ indicators. Johansen and Nielsen (2009) show that under the null, impulse-indicator saturation can be applied to any asymmetric distribution as long as the first four moments exist, and the distribution satisfies some smoothness properties. The reason for this preliminary stage, as opposed to the simultaneous application of IIS and selection of non-linear functions (as recommended above to overcome the problem of extreme observations), is that by obtaining a reasonable first approximation to normality, conventional critical values are then applicable throughout the selection process, which perforce includes both expanding as well as the usual contracting searches as all variables cannot be included in the GUM from the outset. By selecting over the indicators again in the non-linear GUM, the problem of extreme observations is overcome, and this second stage can be undertaken at looser significance levels as the procedure will involve fewer variables than observations.

Augmenting the GUM in section 4.3.2 with the 316 impulse indicators results in 536 regressors in the initial GUM. The GUM equation standard error is $\widehat{\sigma}_{GUM} = 0.431$, which is only slightly smaller than (29), although an F-test of the reduction to (29) (excluding indicators) is rejected ($\mathsf{F}(209, 4637) = 2.601[0.00]^{**}$). Selection is undertaken using *Autometrics* at the 0.1% significance level, and equation (31) reports the selected model, with figure 3 recording the residual density and residual QQ plot.

$$\begin{aligned}
\widehat{w}_i \;=\;& \underset{(0.10)}{3.19} + \underset{(0.002)}{0.059}\mathit{edu}_i + \underset{(0.001)}{0.015}\mathit{exp}_i - \underset{(0.00006)}{0.0008}\,\mathit{exp}_i^2 + \underset{(0.0000)}{0.000014}\mathit{exp}_i^3 + \underset{(0.023)}{0.342}\mathit{hrs}_i \\
&+\underset{(0.012)}{0.127}\mathit{met}_i - \underset{(0.005)}{0.032}\mathit{met}_i^2 - \underset{(0.004)}{0.032}\mathit{met}_i^3 - \underset{(0.023)}{0.142}\mathit{race}_i + \underset{(0.005)}{0.025}\mathit{race}_i^2 \\
&+\underset{(0.008)}{0.041}\mathit{child}_i - \underset{(0.003)}{0.013}\mathit{child}_i^2 - \underset{(0.0004)}{0.003}\,\mathit{mar}_i^3 - \underset{(0.0009)}{0.006}\left(\mathit{hrs}^2 \times \mathit{state}\right)_i \\
&+\underset{(0.013)}{0.069}\left(\mathit{hrs}^2 \times \mathit{child}\right)_i + 242 \text{ indicators.} \qquad (31) \\
\mathsf{R}^2 \;=\;& 0.667 \;\; \widehat{\sigma} = 0.457 \;\; \chi^2_{\mathsf{nd}}(2) = 193.63^{**} \;\; \mathit{SC} = 1.646 \quad N = 5173. \\
&\mathsf{F}_{\mathsf{het}}(25, 4905) = 0.913 \;\; \mathsf{F}_{\mathsf{reset}}(1, 4914) = 3.349
\end{aligned}$$

15 explanatory variables are retained from the 220 candidates, all with t-values greater than 4.6. Also 242 indicators are retained, picking up most of the left-tail skewness. The model passes all diagnostics except for normality, partly due to the large number of indicators putting a mass at the origin, and partly due to some residual skewness in the tails: Fig. 3b records the QQ plot with 95% pointwise standard error bands around the normal and there are significant deviations in the tails. Experience enters as a level, quadratic and cubic, indicating strong non-linearity, as many authors have found when including age and age-squared terms. Characteristics such as usual hours worked, race, metropolitan area, and the number of children also help explain wages, with some strong interactions and non-linear terms. Some effects enter with opposite signs on the level and quadratic term suggesting concave functions. The equation standard error is similar to the GUM: the parsimonious encompassing test of the specific model against the GUM is $\mathsf{F}(278, 4637) = 0.998$, so a valid reduction has been undertaken.

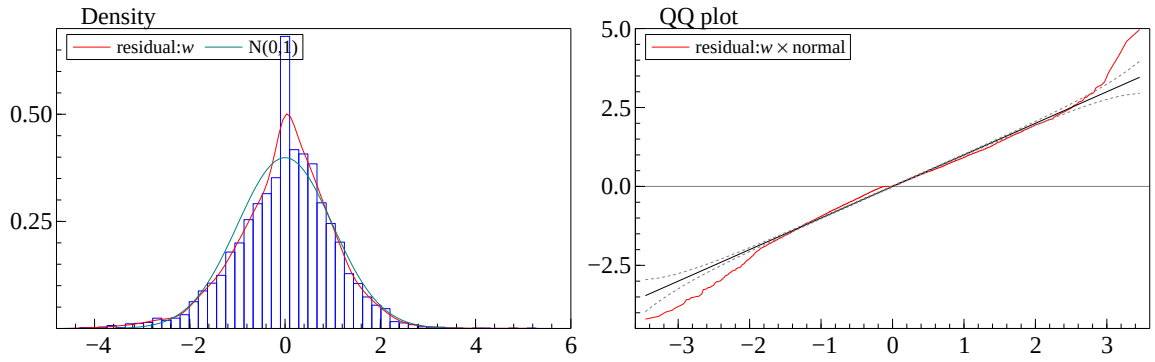


Figure 3: Non-linear wage model with IIS: residual density and residual QQ plot.

Double de-meaning was important: the correlation between *exp* and $exp^2$ was 0.974, but after double de-meaning the correlation was reduced to -0.327. Impulse-indicator saturation was also needed to obtain near-normality for selection and inference. Finally, tight significance levels were vital to prevent excess retention of irrelevant variables.

Although we do not have a substantive functional form specification deduced from a prior theory to test as an encompassing reduction here, the logic thereof is fairly clear. Adding such a functional form to (31) should eliminate many of the selected non-linear terms in favor of the theory-based form, thereby delivering a more robust, identified, interpretable and parsimonious form that does not impugn the congruence of the model or its parsimonious encompassing of the initial GUM, and indeed could even improve the fit while reducing the number of parameters. Equally, such a theory-based function might not remove all the non-linearity, so simply imposing it from the outset would have led to a poorer final model.

# 5 Conclusion

This chapter develops a strategy for the selection of non-linear models, designed to be embedded within the automatic model selection algorithm of *Autometrics*. First, a GUM is formulated in which all potential variables that are thought to explain the phenomenon of interest are included and a test of linearity is applied to that approximation. If the null is accepted, standard selection procedures are applied to the linear GUM. If the null is rejected, a non-linear functional form is generated using polynomial transformations of the regressors in which all functions are double de-meaned prior to inclusion in the GUM to remove one potential collinearity. A set of $N$ impulse indicators is also generated for a sample of size $N$, and included in the GUM to remove outliers and data contamination concurrently with selection of the specific model. Above, because normality was so strongly rejected, a preliminary stage was applied with impulse-indicator saturation alone, to ensure more appropriate initial inferences. Selection is then performed using the techniques developed to handle more variables than observations.

The chapter has shown that in order to achieve a successful algorithm, it is important to jointly implement all the developments discussed above, namely:
testing for the need to select a non-linear model when there are many candidates;
transformations to a near-orthogonal representation;
impulse-indicator saturation to remove extreme observations;

tight significance levels to avoid excess retention of irrelevant non-linear functions;
handling more variables than observations.
Removing any one of these ingredients would be deleterious to selection, and hence to the quality of the resulting model.

An empirical study of returns to education demonstrated the applicability of the approach. Fitting theory-based models such as the Mincer equation without paying attention to the data characteristics by addressing evidence of mis-specification and outliers, can result in poor models. Further, many previous empirical studies did not address the implications of induced collinearity by including age and age squared (or experience) without prior de-meaning. The empirical application is large in dimension, with over $5000$ observations and many linear covariates, leading to a large number of candidate non-linear functions as well as indicators. Fortunately, advances in automatic model selection mean that problems of this scale are now tractable; and the analyses and simulations in recent research demonstrate the high success rates of such an approach.

## References


Abadir, K. M. (1999). An introduction to hypergeometric functions for economists. *Econometric Reviews*, **18**, 287–330.

Ahmed, S. (2007). Econometric issues on the return to education. M.Phil Thesis, University of Oxford.

Altonji, J., and Dunn, T. (1996). Using siblings to estimate the effect of schooling quality on wages. *Review of Economics and Statistics*, **78**, 665–671.

Blundell, R., Dearden, L., and Sianesi, B. (2005). Evaluating the effect of education on earnings: Models, methods and results from the National Child Development Survey. *Journal of Royal Statistical Society Series A*, **168**, 473–512.

Campos, J., Hendry, D. F., and Krolzig, H.-M. (2003). Consistent model selection by an automatic *Gets* approach. *Oxford Bulletin of Economics and Statistics*, **65**, 803–819.

Card, D. (1999). The causal effect of education on earnings. In Ashenfelter, O., and Card, D. (eds.), *Handbook of Labor Economics*, Vol. 3A, pp. 1801–1863. Amsterdam: North-Holland.

Castle, J. L. (2005). Evaluating PcGets and RETINA as automatic model selection algorithms. *Oxford Bulletin of Economics and Statistics*, **67**, 837–880.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2009a). Evaluating automatic model selection. Working paper, Economics Department, University of Oxford.

Castle, J. L., Doornik, J. A., and Hendry, D. F. (2009b). Model selection when there are multiple breaks. Working paper, Economics Department, University of Oxford.

Castle, J. L., Fawcett, N. W. P., and Hendry, D. F. (2010). Forecasting Breaks and During Breaks. In Clements, M. P., and Hendry, D. F. (eds.), *Oxford Handbook of Economic Forecasting*, Ch. 11. Forthcoming, Oxford: OUP.

Castle, J. L., and Hendry, D. F. (2009). A low-dimension, portmanteau test for non-linearity. *Journal of Econometrics*, forthcoming.

Castle, J. L., and Shephard, N. (eds.)(2009). *The Methodology and Practice of Econometrics*. Oxford: Oxford University Press.

Clements, M. P., and Hendry, D. F. (1998). *Forecasting Economic Time Series*. Cambridge: Cambridge University Press.

Copson, E. T. (1965). *Asymptotic Expansions*. Cambridge: Cambridge University Press.

Dearden, L. (1999). The effects of families and ability on men's education and earnings in Britain. *Labour Economics*, **6**, 551–567.

Doornik, J. A. (2007). Econometric model selection with more variables than observations. Working paper, Economics Department, University of Oxford.

Doornik, J. A. (2008). Encompassing and automatic model selection. *Oxford Bulletin of Economics and Statistics*, **70**, 915–925.

Doornik, J. A. (2009). Autometrics. In Castle, and Shephard (2009), pp. 88–121.

Doornik, J. A., and Hansen, H. (2008). An omnibus test for univariate and multivariate normality. *Oxford Bulletin of Economics and Statistics*, **70**, 927–939.

Doornik, J. A., Hendry, D. F., and Nielsen, B. (2009). *Empirical Model Discovery*. Economics Department: Oxford University.

Frisch, R. (1934). *Statistical Confluence Analysis by means of Complete Regression Systems*. Oslo: University Institute of Economics.

Garen, J. (1984). The returns to schooling: A selectivity bias approach with a continuous choice variable. *Econometrica*, **52(5)**, 1199–1218.

Granger, C. W. J., and Teräsvirta, T. (1993). *Modelling Nonlinear Economic Relationships*. Oxford: Oxford University Press.

Griliches, Z. (1977). Estimating the returns to schooling: Some econometric problems. *Econometrica*, **45**, 1–22.

Harmon, C., and Walker, I. (1995). Estimates of the economic return to schooling for the UK. *American Economic Review*, **85**, 1278–1286.

Harrison, A. (1981). Earnings by size: A tale of two distributions. *Review of Economic Studies*, **48**, 621–631.

Heckman, J. J., Lochner, L. J., and Todd, P. E. (2006). Earnings functions, rates of return and treatment effects: The Mincer equation and beyond. In Hanushek, E., and Welch, F. (eds.), *Handbook of the Economics of Education*, Volume 1, Ch. 7. Amsterdam: North Holland.

Hendry, D. F., and Doornik, J. A. (2009). *Empirical Econometric Modelling using PcGive: Volume I*. London: Timberlake Consultants Press.

Hendry, D. F., Johansen, S., and Santos, C. (2008). Automatic selection of indicators in a fully saturated regression. *Computational Statistics*, **33**, 317–335. Erratum, 337–339.

Hendry, D. F., and Krolzig, H.-M. (2001). *Automatic Econometric Model Selection*. London: Timberlake Consultants Press.

Hendry, D. F., and Krolzig, H.-M. (2005). The properties of automatic Gets modelling. *Economic Journal*, **115**, C32–C61.

Hendry, D. F., and Mizon, G. E. (2009). Econometric modelling of changing time series. Unpublished paper, Economics Department, Oxford University.

Hendry, D. F., and Morgan, M. S. (1989). A re-analysis of confluence analysis. *Oxford Economic Papers*,

**41**, 35–52.

Hendry, D. F., and Richard, J.-F. (1989). Recent developments in the theory of encompassing. In Cornet, B., and Tulkens, H. (eds.), *Contributions to Operations Research and Economics. The XXth Anniversary of CORE*, pp. 393–440. Cambridge, MA: MIT Press. Reprinted in Campos, J., Ericsson, N.R. and Hendry, D.F. (eds.), *General to Specific Modelling*. Edward Elgar, 2005.

Hendry, D. F., and Santos, C. (2005). Regression models with data-based indicator variables. *Oxford Bulletin of Economics and Statistics*, **67**, 571–595.

Hoover, K. D., and Perez, S. J. (1999). Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal*, **2**, 167–191.

Mincer, J. (1958). Investment in human capital and personal income distribution. *Journal of Political Economy*, **66(4)**, 281–302.

Mincer, J. (1974). *Schooling, Experience and Earnings*. New York: Natonal Bureau of Economic Research.

Johansen, S., and Nielsen, B. (2009). An analysis of the indicator saturation estimator as a robust regression estimator. In Castle, and Shephard (2009), pp. 1–36.

Lehergott, S. (1959). The shape of the income distribution. *American Economic Review*, **49**, 328–347.

Mizon, G. E., and Richard, J.-F. (1986). The encompassing principle and its application to non-nested hypothesis tests. *Econometrica*, **54**, 657–678.

Perez-Amaral, T., Gallo, G. M., and White, H. (2003). A flexible tool for model building: the relevant transformation of the inputs network approach (RETINA). *Oxford Bulletin of Economics and Statistics*, **65**, 821–838.

Phillips, P. C. B. (2007). Regression with slowly varying regressors and nonlinear trends. *Econometric Theory*, **23**, 557–614.

Priestley, M. B. (1981). *Spectral Analysis and Time Series*. New York: Academic Press.

Ramsey, J. B. (1969). Tests for specification errors in classical linear least squares regression analysis. *Journal of the Royal Statistical Society B*, **31**, 350–371.

Rushton, S. (1951). On least squares fitting by orthogonal polynomials using the Choleski method. *Journal of the Royal Statistical Society, B*, **13**, 92–99.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, **6**, 461–464.

Staehle, H. (1943). Ability, wages and income. *Review of Economics and Statistics*, **25**, 77–87.

Teräsvirta, T. (1994). Specification, estimation and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association*, **89**, 208–218.

White, H. (1980). A heteroskedastic-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, **48**, 817–838.

White, H. (1992). *Artificial Neural Networks: Approximation and Learning Theory*. Oxford: Oxford University Press.