Explaining Unemployment Rates with Symbolic Regression

Philip Truscott & Michael F. Korns

Philip Truscott, Singapore University of Technology and Design, 20 Dover Drive, Singapore, 138682
philiptruscott@sutd.edu.sg

Michael F. Korns, AIS Foundation, 98 Perea Street, Makati 1229, Philippines
mkorns@korns.com

Abstract

Much of the research on the accuracy of symbolic regression has focused on artificially constructed search problems where there is zero noise in the data. Such problems admit of exact solutions but cannot tell us how accurate the search process is in a noisy real world domain.

To explore this question symbolic regression is applied here to an area of research which has been well-travelled by regression modelers: the prediction of unemployment rates. A respected dataset was selected, the CEP-OECD Labour Market Institutions Database, to provide a testing environment for a variety of searches.

Metrics of success for this paper went beyond the normal yardsticks of statistical significance to demand 'plausibility'. Here it is assumed that a plausible model must be able to predict unemployment rates out of the sample period for six future years: this metric is referred to as the 'out of sample R-Square'. Moreover successful models must never produce implausible outlier values such as *negative unemployment rates* (which can easily be produced by seemingly accurate regression models).

We conclude that the two packages tested, Eureqa and ARC, can produce models that go beyond the power of tradition stepwise regression. ARC, in particular, is able to replicate the format of published economic research because ARC contains a high level Regression Query Language **RQL**, inspired by the database search language SQL. RQL consists of *one or more search clauses* which together make up a symbolic regression request. Each search clause represents an independent evolutionary island in which a separate symbolic regression search is performed. The best champion from all search islands is the answer to the regression query. ARC's regression query language allows excellent control of the search process and grammar depth.

ARC also allows easy user intervention in the fitness computation. During the evolution, plausibility was enhanced by penalizing the fitness score of models with negative predict unemployment rates and which had coefficients with P-Values over 0.15.

This research produced a number of models that are consistent with published economic research, have in sample R-Squared values over 0.80, no negative unemployment values, and out of sample R-Square values above 0.45. It is argued that SR offers significant new advantages to social science researchers.

Key words: *Abstract Expression Grammars, Genetic Algorithms, Symbolic Regression, Non-Linear Regression.*

"... it would be dangerous to attempt this comparison; for when statistics are not based upon computations which are strictly accurate, they mislead instead of guiding aright. The mind is easily imposed upon by the false affectation of exactness, which prevails even in the misstatements of science, and it adopts with confidence errors which are dressed in the forms of mathematical truth." Democracy in America

(Tocqueville, Reeve, & Commager, 1952)

Symbolic regression researchers are engaged in the quest for an infallible search agent but can they really guarantee to find the correct model in a search space containing trillions of potential candidates? Much of the recent research in this area has used artificially generated test problems with "zero noise." When the search finds the correct formula, the dependent variable can be predicted exactly (or at least within an infinitesimally small range of error). In the real world the noise is not only significant. At times it can be deafening.

This paper asks some fundamental questions about Symbolic Regression. Has the technique reached a point where it can make a real contribution to social science research? Can symbolic regression search languages focus the hunt to achieve better results that brute force universal searches? How far can the goal search be constrained to stay within the academic culture of specific disciplines? Can the plausibility of models be tested to ensure that they can make predictions in the real world, or are they misleading abstractions that, as de Tocqueville put it, are "merely dressed in the forms of mathematical truth?"

These issues will be explored by applying Symbolic Regression to a specific real world problem: the explanation of national unemployment rates. The discussion starts with an examination of one particularly famous unemployment formula and proceeds to show how far it can be extended or improved by evolutionary methods.

Predicting Unemployment with Institutional Variables

Possibly the most important post-war book on unemployment appeared in 1991. It was a landmark study on the relationship between jobless rates and labour market institutions. *Unemployment: macroeconomic performance and the labour market* (Layard, Nickell, & Jackman, 1991) included a fascinating regression model that predicted unemployment rates for 20 OECD countries from 1983-1988 (see Table 1). For the sake of brevity we will call this the LNJ Model (derived from the names of the authors).

The notes in the last column of Table 1 attempt to give a plain language explanation of the LNJ formula. It contains only one predictor variable from traditional macroeconomics: the change in a country's inflation rate was used to capture the effect of consumer demand. For example, sharply falling prices might indicate deflation, falling demand and rising jobless rates. It also included some effects of a country's system of wage bargaining. High rates of labour union membership tend to increase unemployment. However this can be mitigated by high levels of wage bargaining coordination. Countries that negotiate across the whole labour force or a whole industry tend to have fewer attempts to "leapfrog" over other workers and so less need to control inflation by cutting consumer demand.

The formula also captures the amount spent on "active labour market" policies such as training and labour exchanges.

The most politically controversial variables capture the effects of cash benefits paid to the unemployed. According to the LNJ formula the longer the duration of unemployment benefits the higher the jobless rate. The "Benefit Replacement Ratio" attempts to measure the generosity of the payments. The LNJ formula compares the average cash value of unemployment benefits with the average income of the poorest 25% of wage earners. A higher replacement ratio implies more generous benefits and a higher unemployment rate.

Table 1: Original Layard, Nickell and Jackman Unemployment Regression Formula

Explanatory Variable	Coefficient	Sig	Notes
Constant	0.24		
Change in Inflation	-0.35	**	The change in inflation tries to capture changes in demand. A falling inflation rate implies falling demand and thus higher unemployment rates.
Benefit Duration	0.92	**	This measures the number of years a person can receive unemployment benefits. Countries with shorter time limits appeared to have lower unemployment rates.
Replacement Ratio	0.17	**	This compared the average incomes of those on unemployment benefit with the average income of the poorest 25% of workers. Higher ratios were said to make it less attractive to be in work so leading to higher jobless rates.
Active Labour Market Policy spending	-0.13	**	This included the spending, per unemployed person, on training and job matching services. The higher the active spending the lower the jobless rate.
Union Coverage	2.45	**	Union coverage measures the proportion of the national labour force covered by collective bargaining agreements. This posits that more powerful trade union movements may exert upward pressure on wages, increase the cost of labour and so increase unemployment.
Union Coordination	-1.42	*	While union coverage appeared to increase unemployment this effect could be mitigated by careful coordination of the wage bargaining process: where all unions negotiate at the same time inter-union competition is eliminated. Countries were categorized on a three point scale. Those with a wage bargaining covering the whole work force (like Sweden) scored a 3. Those with industry-wide coordination (like Germany) scored a 2. Countries without coordination (e.g. the UK and USA) scored a 1.
Employer Coordination	-4.28	**	This is the employer version of the Union Coordination described above. The same points system was used (3=national coordination, 2=industry coordination, 1= no coordination).
R^2	0.91	**	In sample R-Square
Sample Size	20		

** Significant at 0.05; * Significant at 0.1

The Quest for an "Operational" Model

The LNJ formula achieved an impressive 0.91 in sample R-Square with a small sample size of only 20 countries. Rather than data for individual years, five year averages for each country were used. The formula looks tantalizingly close to an "operational" model to predict unemployment rates in future years. However, constructing a generally respected model of this type has proved elusive. If regressions are

done over a long time period with individual years, it is very easy to produce a model with a high R-Square but where some of the predicted unemployment rates are below zero. Negative unemployment rates cast serious doubt upon the validity of the predictive model.

One common strategy economists use to ensure that the model fits the data realistically is to use a Boolean flag variable for each country in the dataset. For example, the records in the dataset describing the country, France, would have a "French_Flag" variable coded to 1 while all the other countries would be coded to zero. These Boolean flag variables mean adding one extra variable to the list of predictors for each country in the analysis. Those who use this technique have given it an unprepossessing title: "country dummies." One of the authors of the LNJ formula, Stephen Nickell, helped to devise a regression model (S. Nickell, Nunziata, Ochel, & Quintini, 2001) using both country dummies and time dummies (separate Boolean flags for each year in the series). None of the predicted unemployment rates were negative but policy-makers should ask an important question about such research: If so much of the goodness of fit is coming from the dummies are the policy-related variables being reduced to triviality? Ideally such research should be able to show how much jobless rates will fall in relation to a change in the benefit system, wage bargaining or economic policy, but the influence of the policy-related variables declines as the impact of the dummies rises.

In a traditional OLS regression it would be a simple matter to quantify the impact of the dummies. One could compare the R-Square figure for models with and without them. However Nickel et al. (2001) use Generalized Least Squares (GLS) to allow for heteroscedasticity in the data. While most labour market economists would probably agree with their choice of GLS it means their readers are deprived of a readily understandable goodness-of-fit measure like the R-Square. Blanchard and Wolfers (1999) use similar country data and use non-linear least squares that produces an R-Square. While acknowledging that not all economists support least squares regression for this data, we will use it here to search for an "operational" unemployment prediction formula that satisfies these goals:

- It should achieve a high "goodness of fit" (R-Square)
- It should use no "dummy" variables
- It should not produce any negative unemployment rates
- It should appear plausible in relation to out of sample predictions

The Test Data: The CEP OECD Data Set

A fascinating dataset on labour market institutions was published by Nickell (2006) covering most OECD countries for the years 1960-2004: the CEP-OECD Labour Market Institutions Database. This lacked an inflation change variable (required by the LNJ formula) and so was combined with information on inflation and real interest rates from the World Bank's DataBank (2013).

The full dataset contains 617 rows, but some strategic predictor variables were missing for many years (mainly for the earlier years in the series). To strike a reasonable balance between predictive power and time coverage some records had to be deleted from the historical dataset. After some experimentation with fitness measures, it was decided to remove from the analysis records where there was missing data for inflation, active labour market spending or labour market centralization (entitled CEW in the original dataset). This resulted in a database of 282 rows. This data was further subdivided into two separate databases for testing and training. A training database covering the years up to and including 1993

contained 143 records. The testing database covering the years 1994-2000 contained 137 records. In order to assess how well the formulas predicted unemployment *after* 1993 the regression formulas were applied (with the same coefficients) to the post 1993 data. The mean difference between the predicted values and the actual values was calculated in the same manner as the traditional R-Square. We will refer to this statistic below as the "Out-of-Sample R-Square".

All processing was done on a 64-bit Intel Core i5-2520M CPU at 2.5 GHz. The system had 4 GB of installed RAM. All searches in this paper were limited to 30 minutes.

In order to explore the researcher's options with Symbolic regression grammars two very different packages were used. Eureqa (Dubčáková, 2011) is a freely downloadable product with a highly developed user interface. ARC (for Abstract Regression Classification) (Korns, 2007, 2010a, 2010b) is still in the private domain and requires the user to implement commands with its built-in language RQL (for Regression Query Language) and also allows the package to be modified at the LISP programming level.

Testing with Eureqa

Eureqa has probably gone much further than any other software package in making Symbolic Regression readily accessible. It can be downloaded as a freeware version and has a low price option for academic researchers. It has an extremely user-friendly interface that made importing the CEP-OECD database a two-minute project.

The interface requires the user to set a "target" which is Eureqa's term for the symbolic regression search goal. Figure 1 below shows a goal similar to the initial search goal Eureqa suggested after loading the CEP-OECD database. The variable names have been conveniently taken from the column headers of the imported data. With minor editing UNEM was specified as the dependent variable. Some other variables had to be removed from the pool of predictor variables such as "strunem" (the structural unemployment rate) because they were too close in concept to unemployment itself. In general, where there were groups of variables with similar characteristics they were all left in the pool so that Eureqa could choose those that produced the best fit. However, it proved impossible to include both of the active labour market variables. The variables *almp* (active labour market policy spending as a proportion of GDP) and *almp_unem* (the active labour spending divided by the unemployment rate) contained too much information. Eureqa was able to predict the unemployment rate exactly by discovering the relationship between these two variables alone. For this reason, *almp* was dropped from the search.

Figure 1 A Universal Search Goal in Eureqa

```
UNEM = f(Year, Inflation, Inf_Change, RIR, ep, epl, epl_Allard, udnet, udnet_vis, uc_Ochel,
uc_oecd, uc, co, co_oecd, co_oecd_int, cow, cow_int, ce_oecd, ce_oecd_int, cew, cew_int, brr1,
brr23, brr45, bd, brrl, brr_oecd, nrw, minw_med, ed90_50, ed50_10, educ, educ_int, ho, ho_oecd,
ho_comb, almp_unem, regref, pmr, AdminR, EconR, T1, T2, T3, TW, TW_Nicol, sing1a, sing1b, sing2a,
sing2b, sing3a, sing3b, sing4a, sing4b, mp1a, mp1b, mp2a, mp2b, mp3a, mp3b, mp4a, mp4b, ms1a,
ms1b, ms2a, ms2b, ms3a, ms3b, ms4a, ms4b)
```

Running the Eureqa's universal goal specified in Figure 1 produced twelve champion formulas arranged in order of complexity in its "view results" panel.

Two of these twelve champion formulas are shown in Table 2: the simplest, the most complex and a formula of median complexity.

In the context of labour market economics, the output from Eureqa's universal goal (illustrated by Table 2) would be difficult to "sell". The economists' culture requires regression models of a specific form. Generally, the grammar depth is only one with the independent variables modified by unary operators. Another problem is the explicability of the formula. In the simplest formula, for example, a benefit replacement rate is divided by active labour market policy spending per unemployed person. Culturally economists would find it unacceptable to have two such unrelated things on either side of the operator.

Table 2 Three Representative Eureqa Champions

Complexity	In Sample R-Squared	Coefficients	Formula
18	0.85	2	UNEM = (home_owner_rate_comb + ms3a + 4.636*benefit_replacement_rate_oecd - 53.73 – benefit_replacement_rate_years_2&3)/(bargaining_coordination + education_years_mean + almp_unem)
6	0.61	1	UNEM = 4.407 + benefit_replacement_rate_oecd/ active_labour_div_unemployed

In order to try to produce a model more palatable to economists, a search goal was specified to replicate the form of the regression model in Table 1. This target is shown in Figure 2 below (in the interests of readability the short form of the variable names has been used). The target expression groups variables into pools any one of which might be inserted into the model. Thus on line 2 the target does not name a single variable but rather five possible variables any one of which might be selected as the best candidate to represent the benefit replacement ratio variable in the LNJ formula.

Figure 2 A Complex Search Goal expressed in Eureqa

```
1  UNEM=   f0() *f1(bd)
2        + f2() *f3(brr1, brr23, brr45, brrl, brr_oecd)
3        + f4() *f5(almp_unem)
4        + f6() +f7(udnet, udnet_vis)
5        + f8() +f9(co, co_oecd, co_oecd_int, cow, cow_int, ce_oecd, ce_oecd_int, cew, cew_int)
6        + f10()*f11(co, co_oecd, co_oecd_int, cow, cow_int, ce_oecd, ce_oecd_int, cew, cew_int)
7        + f12()*f13(Inflation, Inf_Change, RIR)
8        + f14()+f15(Year, ho, ho_oecd, ho_comb, urban)
```

Two of the champion formulas from the Eureqa search specified in Figure 2 are shown in Table 3 below.

Table 3 Two Complex Regression Champion from Eureqa

Complexity	In Sample R-Squared	Coefficients	Formula
26	0.82	14	UNEM = 2.354 + 0.22*benefit_replacement_rate_oecd + factorial(-0.01192*home_owner_rate_comb) - 0.05175*almp_unem - 1.299*bargaining_coordination_oecd - 2.213*benefit_duration
35	0.84	15	UNEM = 0.2265*brr_oecd + 70.58/(6.68 + almp_unem) + factorial(-0.01179*ho_comb) - 3.698 - 1.132*co_oecd - 1.029*tan(bd^2)

For each level of complexity Eureqa shows only one regression champion. At the time of writing it does not appear there is a method to constrain Eureqa to produce a variety of models at a single complexity level. For this reason most of the searches and plausibility analysis will center on ARC.

Testing with ARC

Next, an attempt was made to replicate the LNJ formula in Table 1 using ARC and RQL. Symbolic regression seems particularly well suited to exploring relationships in our Labour Market Institutions database because often there are many slightly different versions of a variable designed to represent the same characteristic. Whereas in Table 1, the authors used a single variable to capture the benefit replacement rate in Nickel's (2004) database there were no less than six benefit replacement variables. In this situation, we should specify a symbol for all six possible variables rather than mention one specific one. In RQL this symbol is specified by building up a list of variables in the form x[n] where n is the ordinal number representing the position of the predictor variable in the dataset. Thus to allow ARC to choose among all five benefit replacement variables for the 2nd term one writes a goal expression in the form shown in Figure 3 below. Mnemonics have been used to specify the variable names in Figure 3. In reality ARC requires the variables to be defined in terms of their ordinal position in the array x[] (x1, x2, etc.).

This figure replicates an entire RQL specification of an Island in the search space that attempts to replicate the LNJ formula in Table 1. However, wherever several variables might substitute for the original variable all of them have been allowed in to the search space for that term. The variable names in the following regression tables are slightly more verbose in the interests of readability (for example Brr45 has been lengthened to Benefit_Replacement_in_Years_4&5); Figure 3 uses only the abbreviated form.

Figure 3: Specifying the LNJ Formula in RQL

```
1   search universal(1,7,v)
2   where {fitness(nmae)
3   island(smart,smart,100,100,200)
4   op (noop,abs,square,cube,quart,exp,ln)
5   v0 (bd)
6   v1 (bd)
7   v2 (brr1,brr23,brr45,brrl,brr_oecd)
8   v3 (brr1,brr23,brr45,brrl,brr_oecd)
9   v4 (almp, almp_unem)
10  v5 (almp, almp_unem)
11  v6 (udnet,udnet_vis,uc_Ochel,uc_oecd,uc)
12  v7 (udnet,udnet_vis,uc_Ochel,uc_oecd,uc)
13  v8 (co,co_oecd,co_oecd_int,cow,cow_int,ce_oecd,ce_oecd_int,cew,cew_int)
13  v9 (co,co_oecd,co_oecd_int,cow,cow_int,ce_oecd,ce_oecd_int,cew,cew_int)
14  V10(co,co_oecd,co_oecd_int,cow,cow_int,ce_oecd,ce_oecd_int,cew,cew_int)
15  V11(co,co_oecd,co_oecd_int,cow,cow_int,ce_oecd,ce_oecd_int,cew,cew_int)
16  v12(inflation,inflation_change,rer)
17  v13(inflation,inflation_change,rer)
18  v14(year,ho,ho_oecd,ho_comb,urban)}
19  v15(year,ho,ho_oecd,ho_comb,urban)}
```

Figure 3 above shows the entire search goal in RQL. Line 1 specifies that the search is universal but in the following lines, pools of variables have been specified to ensure the final champion formula sticks to the general form of the regression model in Table 1. Line 2 declares the fitness measure shall the Normalized Mean Absolute Error. Korns (2013) explains the island goal on line 3 elsewhere in this volume. The list of operators in line 4 is extremely parsimonious in part to ensure that champion formulas conform to the

culture of economic journals. Binary operators would produce champions that appear to be bloated. Some of the geometric functions (sin, cos, tan) are rarely seen as modifiers in labour market economics and have been excluded.

In general, the islands have been specified to conform as closely as possible to the original LNJ formula save that one extra term has been added which is a pool of socio-demographic variables (describing home ownership and urbanization). Thus, the search goal in Figure 3 will produce a model with 8 variables rather than the 7 shown in Table 1. This is justified here as a means to give the models a reasonably good fit without resorting to the expedient of dummy variables. In general, the searches below select one of the home ownerships variables. It has been argued that a country with a large proportion of owner-occupied housing will have higher unemployment because job-seekers find it difficult to re-locate to find employment (Oswald, 1996).

It should also be noted that the search process was allowed to choose among three possible macroeconomic variables that might affect consumer demand. The original LNJ model used the change in inflation (the current year's inflation rate minus that of the previous year). This search allows ARC to choose the simple inflation rate or the real interest rate.

When the search in Figure 3 was first run, it failed to produce any champions that satisfied two important plausibility goals. The champions had some negative unemployment rates or the predicted unemployment rates in the out of sample data were so different from the actual values that the out of sample R-Square was negative.

Customized Scoring

However, the same search produced significantly better results when the scoring process penalized formulas with large outliers (including negative unemployment rates anywhere in the set of predicted values). The customized penalties can be expressed fairly simply in plain language:

a) Where any unemployment rate is higher than 40% the error was multiplied by 1.015

b) Where any unemployment rate is lower than 0% the error was multiplied by 1.015

A similar penalty was applied to models with at least one insignificant coefficient: where any P-Value exceeded 0.15, the error was also multiplied by 1.015.

It is important to stress that such models were not excluded from the evolution. There survivability was merely impaired.

Table 4 shows the result of the ARC search in Figure 3 where the scoring penalties for minimum and maximum values were applied during the evolution. The R-Square of 0.79 is lower than that in Table 1. However, it was based on 143 records compared to the 20 records of the LNJ regression (where there was only one record per country and one time period based on a five year average). Moreover the same formula produced at Out of sample R-Square of 0.47 when applied to the years 1994-2000.

Many labour market economists would object to the regression model shown in Table 4 because the signs of the benefit variables do not match the conventional theory. It is common in economic journals to read a smug analysis of a regression model which declares proudly "all the signs are in the expected direction." In Table 4 the benefit duration appears to lower unemployment but the benefit replacement variable

increases it, whereas in Table 1 as both the benefit variables increase the jobless rate also increases. This type of dilemma illustrates a philosophical difference between genetic programmers and many economists. Enthusiasts for evolutionary computation claim that the world is a complex non-linear place and that genetic algorithms are about to make a major contribution to the understanding of it.

Table 4 A Modified LNJ Formula

RSQ=[0.89]				
Sign	Coefficient	Variable	T-Statistic	P-Value
-	4.289049	cube(benefit_duration)	-10.262717	0.0005
+	7.409974	ln(benefit_replacement_rate_oecd)	21.012474	0.0005
-	2.532656	ln(active_labour_per_unemployed)	-11.665543	0.0005
+	8.533892E-38	exp(union_density)	3.620270	0.0050
-	0.007280	quart(bargaining_coordination_oecd)	-7.589414	0.0005
-	0.013467	quart(bargaining_coordination_w)	-1.995091	0.0350
-	0.786579	ln(Inflation)	-4.160817	0.0005
+	2.658106E-07	quart(home_owner_rate)	18.380145	0.0005
Plausibility Measures			Predicted	Actual
Highest Unemployment Rate			20.2	22.5
Lowest Unemployment Rate			0.4	0.3
Out of Sample R-Square			0.02	

At their best economists demand some clear theory that underlies a regression model. At their worst they search for the highest R-Square that will confirm a pre-conceived ideological prejudice. Plausibility measures such as the Out of Sample R-Square should afford an objective means to resolve this dilemma for they show how closely a given model helps to map out the real world.

The Table 4 should be a solemn warning to many social scientists who publish regression models; it shows how easy it is to strike fool's gold. Table 4 shows that seven of the eight terms have a P-Value below 0.01. All of them are significant at 0.05. It has a respectable R-Square of 0.89. With the exception of the benefit duration term, the signs of all the coefficients are in the correct direction (i.e. the direction implied by the prevailing economic theory). However as soon as we require this champion to do work in the real world it collapses. As can be seen from the bottom line of Table 4 the out of sample R-Square was only 0.02 implying that it could only explain a tiny proportion of unemployment rates from 1994-2000.

Genetic programmers are often accused of producing over-fitted models that mean nothing in the real world. Researchers who use non-evolutionary methods should answer the same charge. The R-Square, T-Statistics and P-values are often the ones published alongside conventional regression models. How many of the models published in journal articles are equally weak predictors of future events? Of their hundred champion formulas produced by this search none had an out of sample R-Square higher than 0.04. Let us see if some variations on the LNJ formula can produce more convincing results.

Virtual Economists

One of ARC's most interesting features is the ability to specify multiple 'islands' in the search space that will evolve simultaneously. To demonstrate the power of this feature a set of ideological islands were created with the same general form as the island shown in Figure 3 but with pools of variables that represent different ideological preferences. We will run simulations with the existing modified LNJ formula shown in Figure 3. We will call this island "**Pragmatic Centrist**" because some of its ideological underpinnings seem to lean towards the political right (the effect of the benefit system) and some to the political left (support for bargaining coordination).

The second island we will call "**Hayek Libertarian**" after the champion of the free market Freidrich von Hayek. Hayek would be able to agree that the effect of the benefit system will push up unemployment. However, he was a firm opponent of attempts to plan economic behavior so any attempt to promote wage bargaining coordination or centralization would have been anathema to him (Hayek, 1994). On this island, inhabitants can hunt for relationships linked to certain libertarian issues such as the taxation and the minimum wage. However, they are not allowed to include any variables describing bargaining centralization or coordination. The island also includes a large number of variables that describe the extent of market regulation and possible disincentives caused by the tax system.

The last island we will call simply "**Progressive**." Our progressive islanders accept most of the arguments underpinning the LNJ model but are unwilling to consider reducing the duration of unemployment benefits. However, they are willing to try to change the benefit replacement ratio. It should be stressed that this can be done in different ways. A conservative might wish to cut the real value of the benefits. A progressive might try to reduce the replacement ratio by raising the incomes of the poorest quarter of wage earners. This could be done by increasing benefits paid to poorer workers such as the UK's Family Credit or the USA's Earned Income Tax Credit (and ending the 'unemployment trap' is often claimed as a significant reason for doing so).

Results of Hayek Libertarian Island

The Hayek Libertarians managed to find several models within their ideological research constraints. Table 5 achieves an in sample R-Square of 0.88 without using any wage bargaining variables. The sign of both benefit system variables was positive which is direction implied by the LNJ formula (Layard et al., 1991), however many other plausible models developed for the Hayekian island had a positive sign for the benefit replacement rate and a negative sign for benefit duration.

The prevailing argument is that the shorter the time when the jobless receive cash benefits the lower will be the unemployment rate. However, where the search process can select multiple benefit variables the varying signs may indicate that the effect of benefit system generosity does not increase unemployment monotonically or exponentially. The varying positive and negative signs for models with high out-of-sample R-Squares imply the true relationship may be polynomial. This conflicts with much received wisdom in economics where any regression model putting a negative sign before a benefit duration coefficient would probably be un-publishable.

Another area of potential controversy in Table 5 is the impact of the minimum wage. If Hayek were alive to comment on Table 5 he could only smile with a lop-sided grin because the minimum wage variable was selected but with a negative sign. According to the traditional libertarian argument, minimum wage laws price workers out of jobs by making them more expensive to employ. Should one reject the model

on this account? Layard and Nickell (2011) reviewed the evidence on this and wrote "Our reading of the evidence is that minimum wages are set low enough not to have a significant impact on adult male unemployment" (though they thought it might increase jobless rates for younger workers). The negative sign for the minimum wage variable could well capture some undocumented effect such as a tendency for high minimum wage countries to spend more on training (or some other unpredicted interaction). The direct tax rate also has a negative sign, which is also contrary to the Hayekian argument. The model shows a negative sign for active labour market spending per unemployed person and a negative sign for inflation which both in line with most of the published research.

Table 5 A Hayek Libertarian Formula

RSQ=[0.88]				
Sign	Coefficient	Variable	T-Statistic	P-Values
+	0.896966	quart(benefit_duration)	1.885060	0.0400
+	10.182363	ln(benefit_replacement_rate_oecd)	25.070745	0.0005
-	-4.727687	ln(active_labour_per_unemployed)	-16.822152	0.0005
+	1.08E-07	quart(union_density)	6.230725	0.0005
-	-4.080167E-07	quart(minimum_wage_median)	-10.693994	0.0005
-	-0.000352	cube(direct_tax_rate)	-14.012383	0.0005
-	-0.099023	abs(inflation)	-2.051690	0.0300
+	0.087017	home_owner_rate_oecd	12.874764	0.0005
Plausibility Measures			Predicted	Actual
Highest Unemployment Rate			22.5	23.9
Lowest Unemployment Rate			0.3	1.6
Out of Sample R-Square			0.46	

Results of the 'Progressive' Island

As mentioned above the 'Progressive' islanders were allowed to analyze the benefit system but did not want to contemplate cutting the time for paying unemployment benefit. They were allowed two use two benefit system variables (but not benefit duration). Similarly, they were allowed to use two wage bargaining system variables.

The results look encouraging in terms of the P-Values. Seven out of eight are under 0.001 and the other is under 0.05. The out of sample R-Square of 0.51 also looks respectable. The chief difficulty is that that where there are two variables that cover the same general topic the signs of the coefficients do not match. In the LNJ Formula, both benefit duration and benefit generosity (the replacement rate) increase unemployment (both had a positive sign). Similarly, both employee bargaining coordination and employer coordination decrease it (both have a negative sign).

How can we explain the fact that one coefficient for the benefit replacement rate is positive while the other is negative. Is this just the result of an implausibly over-fitted model? If so, why does the formula in Table 6 explain so much of the variation in unemployment rates from 1994 to 2000? Philosophically one can argue that whatever model achieves the highest out of sample R-Square is the model that should be used by policy makers in the real world. The same arguments apply to the coefficients for the two wage bargaining variables in Table 6. Many economists would reject the model on this account. A

genetic programmer might argue this is example of the subtle non-linear relationships that evolutionary algorithms are intended to uncover. Unfortunately, the coefficients in Table 6 do not take the form of an acceptable polynomial. For that to be the case ***the same variable*** would need to appear twice with both positive and negative coefficients.

Table 6 A 'Progressive' Formula

RSQ=[0.87]				
Sign	Coefficient	Variable	T-Statistic	P-Value
-	0.132523	benefit_replacement_years2&3	-6.678575	0.0005
+	11.500410	ln(benefit_replacement_rate_oecd)	14.671151	0.0005
-	3.071967	ln(active_labour_per_unemployed)	-11.074161	0.0005
+	5.895417E-38	exp(union_density)	2.313943	0.0200
+	2.842712	bargaining_coordination_w	6.540536	0.0005
-	1.566878	bargaining_coordination_oecd	-8.270395	0.0005
-	0.230288	abs(real_interest_rate)	-4.481130	0.0005
+	2.191587E-07	quart(home_owner_rate_oecd)	12.904609	0.0005
Plausibility Measures			Predicted	Actual
Highest Unemployment Rate			19.9	23.9
Lowest Unemployment Rate			1.6	1.6
Out of Sample R-Square			0.51	

In order to explore the concept of such a non-linear relationship between replacement rates and unemployment, several islands were defined with the same format as the progressive search goal except that the same benefit variable was repeated in the first and second positions. The different benefit variables tested in this way were brr1 (the replacement rate during the 1st year of unemployment), brrl (the long-term replacement rate), brr_oecd (a summary measure intended to represent the replacement rate across all time periods).

Table 7 Formula with a Polynomial Benefit Replacement Effect

RSQ=[0.85]				
Sign	Coefficient	Variable	T-Statistic	P-Value
-	0.296301	benefit_replacement_rate_oecd	-4.935111	0.0005
+	13.831793	ln(benefit_replacement_rate_oecd)	9.284424	0.0005
-	2.934026E+00	ln(active_labour_per_unemployed)	-8.139578	0.0005
+	4.16E-08	quart(union_density_vis)	2.650339	0.0100
-	1.812492	abs(bargaining_coordination_oecd)	-8.442577	0.0005
+	3.154983	abs(bargaining_coordination_w)	6.070994	0.0005
-	0.239395	real_interest_rate	-4.282545	0.0005
+	2.204215E-07	quart(ho_oecd)	11.569312	0.0005
Plausibility Measures			Predicted	Actual
Highest Unemployment Rate			19.3	23.9
Lowest Unemployment Rate			0.6	1.6
Out of Sample R-Square			0.47	

A final island was defined with bd (the benefit duration variable) included since the format of this variable in the CEP-OECD database captures both the duration and generosity of the payments. ARC's RQL language is possibly the only SR package that would allow this type of highly customized multi-island search to be defined. Of the four hundred regression champions produced by this search, only the brr_oecd variable produced any plausible models. The model with the highest out of sample R-Square is shown in Table 7. Here we see a large positive coefficient applied to the log of benefit replacement combined with a small negative coefficient on brr_oecd in an unmodified form.

A crude plain language explanation of this might be as follows: "unemployment benefits that are extremely ungenerous have a powerful effect in lowering unemployment but this effect tapers away as benefit payments approach the average incomes of the lower paid." This has important political implications because a government could spend billions of dollars on in-work benefits to change the replacement ratio from say 80% down to 70% without much impact on unemployment. The really strong unemployment-reduction effects would be seen in countries with tiny replacement rates (such as Japan where there no benefit at all after the first year). If this non-linearity of benefit levels is true then it lends further weight to Layard and Nickell's call for a job guarantee (2011) rather than benefit reform. Under this scheme the state would have a duty to offer work or training in the first year of unemployment and the jobless would have a corresponding duty to accept the offer.

The polynomial form represented in Table 7 might satisfy some economists but many would still object to any coefficients that are positive when economic theory implies they should be negative (or vice versa). Another way to deal with this type of objection would be to simplify the search goal by only allowing only one benefit variable and one wage bargaining variable (the LNJ formula allowed two of each). The resulting search goal allows for six variables.

Those predisposed to shave their results with Occam's razor will note with satisfaction that their methodological barbering has created a champion that is far more ideologically attractive (if one can call labour market economics an ideology). Table 8 shows a champion model with the highest out of sample R-Square value. All the coefficient signs are consistent with the LNJ formula.

Table 8 A Simplified LNJ Formula with six terms

RSQ=[0.82]				
Sign	Coefficient	Variable	T-Statistic	P-Value
+	6.609785	ln(benefit_replacement_rate_oecd)	15.866457	0.0005
-	2.096486	ln(active_labour_div_unemployed)	-7.220649	0.0005
+	8.957248E-38	exp(union_density)	3.038739	0.0050
-	0.168167	square(bargaining_coordination_oecd)	-6.477663	0.0005
-	0.237584	Inflation_Change	-2.924540	0.0100
+	2.520564E-07	quart(home_owner_rate_oecd)	13.535545	0.0005
Plausibility Measures			Predicted	Actual
Highest Unemployment Rate			19.1	23.9
Lowest Unemployment Rate			0.5	1.6
Out of Sample R-Square			0.48	

"Guided" vs. "Unguided" Searches

Since much effort has been invested in producing automatic regression model searches, it is worth stopping to ponder if equally plausible models could have been produced without using an existing published formula as a starting point. Theoretically, a symbolic regression expert with no knowledge of economics could produce a more plausible model by using generic search heuristics.

Table 9 A Formula Resulting from a Universal Search

RSQ=[0.83]				
Sign	Coefficient	Variable	T-Statistic	P-Value
-	6.177979E-05	cube(home_owner_rate_oecd)	-6.400420	0.0005
+	0.256957	abs(inflation_change)	2.442444	0.0150
-	0.105022	active_labour_per_unemployed	-6.539635	0.0005
+	0.000396	square(union_density)	2.999033	0.0050
+	6.094955	ln(benefit_replacement_rate_oecd)	11.574101	0.0005
+	1.109308E-06	quart(home_owner_rate_oecd)	8.737376	0.0005
-	1.582346	abs(bargaining_centralization_oecd_int)	-7.275434	0.0005
+	0.000150	cube(employment_tax_rate)	5.940433	0.0005
Plausibility Measures			Predicted	Actual
Highest Unemployment Rate			22.2	23.9
Lowest Unemployment Rate			0.7	1.6
Out of Sample R-Square			0.45	

To explore this, a universal search goal was specified without any named variables. The champion model is shown in Table 9. This is the most plausible model to emerge from the search in the sense that it had the highest out of sample R-Square among those models that had no negative unemployment rates. Given that this has an out of sample R-Square of 0.45 it certainly appears to have a respectable fit.

Table 10 Number of Plausible Models by Grammar Depth and Number of Variables

	(1) Number of Minimally Plausible Models (Out of sample RSQ > 0)	(2) Number of Highly Plausible Models (Out of Sample RSQ > 0.3)	(3) Most Plausible Model (Highest Out of Sample RSQ in group)
9 Variables 0 Depth	1	1	0.42
8 Variables 0 Depth	2	1	0.30
7 Variables 0 Depth	19	5	0.42
9 Variables 1 Depth	14	14	0.66
8 Variables 1 Depth	12	9	0.57
7 Variables 1 Depth	5	4	0.39
9 Variables 2 Depth	2	2	0.39
8 Variables 2 Depth	0	0	0.00
7 Variables 2 Depth	18	16	0.56

Interestingly the universal search detected taxes on employment as a predictor of increased unemployment. In theory such taxes should always increase the unemployment rate of a country that relies on them heavily because they increase the cost of home produced goods but not imports. The also increase the cost of goods that are unusually labor-intensive and comparatively cheapen capital intensive goods. The British economist Nicholas Kaldor raised concerns about the employment effects of such taxes when the UK's postwar welfare state was set up (King, 2009). Layard and Nickell reviewed multiple studies (2011) and were not convinced that employment taxes were important. Their appearance in Table 9 might suggest that they are significant in combination with other variables.

In Table 10 we address the question of whether one part of the search space yields higher proportions of plausible models. Universal searches were specified at three 'grammar depths'. At a grammar depth of zero only linear models can be formed. At a grammar depth of one the variables may be modified by a single operator (e.g. *cube(benefit_duration*)). At a depth two the functions may be nested which in the case of unary operators creates terms such as *cube(exp(benefit_duration))*. It is difficult to detect a consistent pattern but intuitively the models at a single grammar depth produced the largest number of plausible models (those where the out of sample R-Square was over 0.3). It would be tempting to look at Table 10 and assume that the model that achieved the highest ORSQ was the "overall winner" but one of its features would deny it acceptability among economists: it uses no less than three home ownership variables out of its total of seven. With universal goals there is no way to constrain ARC from picking the same or similar variables multiple times.

Conclusion

Both Eureqa and ARC offer tools to regression modelers that go significantly beyond the power offered by stepwise regression in packages like Stata, SPSS and SAS. ARC, in particular, allows the statistical explorer to fine-tune the target model in a way that adds real value for the social scientist.

Consider the process of using a package like Stata. Stepwise regression can be used to select independent variables and the best fitting model predicts that some countries have negative unemployment rates. There is no automated process to find plausible models where all the predicted rates are positive and which can predict out of sample data with any accuracy.

Its most telling advantages of SR and ARC in particular are the ability to specify multiple islands on the same run and to control the grammar depth to be searched.

The Greek goddess of hunting, Artemis, had a dog called Lealaps with a singularly useful quality: it always caught its prey. Those developing symbolic regression systems are engaged in a similar pursuit. This paper demonstrates the ability to find a powerful unemployment prediction model using an automatic universal search. This universal search produced an out of sample R-Square of 0.45 which is creditable, but this figure was exceeded by the models shown in tables 5-8 which were inspired by the research of famous economists. Our cybernetic Lealaps is useful, but perhaps he is even more effective when leashed to a human expert

Notes

Blanchard, O., & Wolfers, J. (1999). *The Role of Shocks and Institutions in the Rise of European Unemployment: The Aggregate Evidence* (Working Paper No. 7282). National Bureau of Economic Research.

Dubčáková, R. (2011). Eureqa: software review. *Genetic Programming and Evolvable Machines*, *12*(2), 173–178.

Hayek, F. A. von. (1994). *The road to serfdom*. Chicago: University of Chicago Press.

King, J. E. (2009). *Nicholas Kaldor*. Basingstoke [England]; New York: Palgrave Macmillan.

Korns, M. (2007). Large-Scale, Time-Constrained Symbolic Regression. In R. Riolo, T. Soule, & B. Worzel (Eds.), *Genetic Programming Theory and Practice IV*, Genetic and Evolutionary Computation (pp. 299–314). Springer US.

Korns, M. (2010a). Symbolic Regression of Conditional Target Expressions. In R. Riolo, U.-M. O'Reilly, & T. McConaghy (Eds.), *Genetic Programming Theory and Practice VII*, Genetic and Evolutionary Computation (pp. 211–228). Springer US.

Korns, M. (2010b). Abstract Expression Grammar Symbolic Regression. *Genetic Programming Theory and Practice VIII (ed. Rick Riolo , Una-May O'Reilly and Trent McConaghy)*. Springer.

Layard, R., Nickell, S. J., Eichhorst, W., & Zimmermann, K. F. (2011). *Combatting unemployment*. Oxford; New York: Oxford University Press.

Layard, R., Nickell, S. J., & Jackman, R. (1991). *Unemployment : macroeconomic performance and the labour market*. Oxford [England]; New York: Oxford University Press.

Nickell, S., Nunziata, L., Ochel, W., & Quintini, G. (2001). *The Beveridge curve, unemployment and wages in the OECD from the 1960s to the 1990s - preliminary version*.

Nickell, W. (2006). *The CEP-OECD Institutions Data Set (1960-2004)* (CEP Discussion Paper No. dp0759). Centre for Economic Performance, LSE.

Oswald, A. J. (1996, December). A conjecture on the explanation for high unemployment in the industrialized nations: part 1.

Tocqueville, A. de, Reeve, H., & Commager, H. S. (1952). *Democracy in America,*. London: Oxford University Press.

World Bank. (2013). DataBank. World Bank.