Econometric Genetic Programming Outperforms Traditional Econometric Algorithms for Regression Tasks

André Luiz Farias Novaes
Electrical Engineering Department
PUC-Rio, Brazil
andrelfnovaes@gmail.com

Ricardo Tanscheit
Electrical Engineering Department
PUC-Rio, Brazil
ricardo@ele.puc-rio.br

Douglas Mota Dias
Department of Electronics and Telecommunications Engineering
UERJ, Brazil
douglasmota.uerj@gmail.com

ABSTRACT

Econometric Genetic Programming (EGP) evolves multiple linear regressions through Genetic Programming (GP), which is responsible for model selection, aiming to generate high accuracy regressions with potential interpretability of parameters. It uses statistical significance as a feature selection tool, directly and efficiently identifying introns and controlling bloat. In this paper, EGP is tested against traditional feature-selection econometric algorithms in regression tasks – namely Partial Least Squares Regression, Ridge Regression and Stepwise Forward Regression – outperforming them in all three datasets. The way EGP explores search space of possible regressors and models is crucial for its results. EGP is carefully constructed considering econometric theory on cross-sectional datasets, giving rigorous treatment on topics like homoscedasticity and heteroscedasticity, statistical inference for estimated parameters and sampling criteria. It also benefits by the mathematical proof on accuracy and statistical significance: accuracy will only increase if the regressor presents a test's statistics module in a two-sided hypothesis testing higher than a predefined value.

KEYWORDS

Genetic Programming, Multiple Regression, Model Selection, Feature Selection

GECCO '17 Companion, July 15-19, 2017, Berlin, Germany

ACM ISBN 978-1-4503-4939-0/17/07 $15.00
http://dx.doi.org/10.1145/3067695.30825060

1 INTRODUCTION

The very first form of regression was the method of Least Squares (LS), published by [1] and [2]. The term "regression" was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. As stated in [3], the method for Symbolic Regression (SR) proposed by [4] is an alternative approach to curve fitting. The technique creates mathematical expressions to fit a set of data points using the evolutionary process of GP.

Past literature relates some pioneer works that hybridize linear regression with GP. For a generous list on it, see [5]. In the following, just papers that most influenced EGP will be described.

Works [3] and [6,7] create polynomial regression models for SR tasks. The Weierstrass approximation theorem (1885) states that every continuous function defined on a closed interval [a, b] can be uniformly approximated as closely as desired by a polynomial function. By itself, the theorem would be sufficient for a great effort on approximating the dependent variable in a regression task by polynomials. EGP follows this direction.

EGP, which was first and partially introduced in [8] for regression tasks and tested against exhaustive search algorithms, is carefully constructed considering econometric theory on cross-sectional datasets. Rigorous treatment on topics like homoscedasticity [9], statistical inference for estimated parameters and sampling criteria are made. These considerations represent a significant difference with its predecessors, which relax some hypothesis or even do not test them on datasets and models, although each of them has its own contribution.

Kaizen Programming (KP) [13] is an interesting evolutionary tool based on concepts of continuous improvement from Kaizen methodology, which was successfully tested against traditional SR benchmark functions. EGP is similar to KP in the sense that both are concerned about bloat, introns and use $\overline{R}^2$ (see section 2) as a comparison metric between models. But they also differ in several aspects, in particular the fact that EGP exclusively evolves polynomials (enhancing potential interpretability of parameters), while KP evolves linear-in-parameters individuals in a more general shape.

In this paper, EGP is tested against traditional feature-selection econometric algorithms in regression tasks outperforming them in all three datasets. The main reason EGP outperforms traditional econometric algorithms is its capability to explore the regressors' and models' search space. While the number of regressors for each of the feature-selection algorithms are just a few for each dataset, EGP generates models with 50 statistically significant regressors or more. EGP explores non-linearity of features, by multiplying different features, maintaining models with linear structure.

This paper is organized as follows: Section 2 describes the elements of econometrics used by EGP: there is no intention to fully exhaust the theme; justification on these elements is presented when necessary. Section 3 succinctly describes EGP. Sections 4 proposes experiments and discusses results. Conclusion is done in Section 5.

2 ECONOMETRICS

2.1 Linear Regression Model, Least Squares, QR Decomposition

As in [9], the multiple linear regression model with k parameters and n observations is:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u} \tag{1}$$

where $\boldsymbol{y}_{n \times 1}$ is the dependent variable vector, $\boldsymbol{X}_{n \times (k+1)}$ is the regressor's matrix, $\boldsymbol{\beta}_{(k+1) \times 1}$ is the vector representing the terms that adjust $\boldsymbol{X}$ to $\boldsymbol{y}$ and $\boldsymbol{u}_{n \times 1}$ is an error vector.

Vector $\boldsymbol{\beta}$ is an unknown statistical population parameter usually estimated by LS, which generates $\widehat{\boldsymbol{\beta}}$, the ideal multiplier for $\boldsymbol{X}$ on $\delta(\boldsymbol{X})$ (the column space of $\boldsymbol{X}$) that makes $\boldsymbol{X}\widehat{\boldsymbol{\beta}}$ the most closely projection to $\boldsymbol{y}$ on $\delta(\boldsymbol{X})$. As $\delta(\boldsymbol{X})$ is orthogonal to $\boldsymbol{u}$, some matrix manipulation leads to:

$$\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{y} \tag{2}$$

In [10], it is stated: "The most commonly used, and in many ways the most important, estimation technique in econometrics is LS". In general, calculations on (2) via matrix inversion are numerically unstable and QR Decomposition is recommended by [11,12] in such cases.

2.3 Hypothesis Test

Under conditions stated in [14], $\widehat{\boldsymbol{\beta}}$ is a BLUE estimator for $\boldsymbol{\beta}$ [9], which does not guarantee statistical significance for $\boldsymbol{\beta} = [\beta_1 \cdots \beta_i \cdots \beta_k]^{\mathrm{T}}$. I.e., it is possible that some β_i in (1), or even all $\boldsymbol{\beta}$, is a pure random effect on $\boldsymbol{y}$ and does not present any causal relationship with it. To check statistical significance, it is natural to perform HT on β_j, individually, or on $\boldsymbol{\beta}$. Just fully satisfiability of conditions stated in [14] allows to perform HT as described in the following and that is the case of models generated by EGP and datasets used in this article.

HT is constructed with a null and alternative hypothesis ($\boldsymbol{H}_0$ and $\boldsymbol{H}_1$, respectively), a test statistics and a decision criteria. For a two-sided HT for β_i, hypothesis $\boldsymbol{H}_0$ and $\boldsymbol{H}_1$ are frequently:

$$\begin{aligned} \boldsymbol{H}_0&: \beta_i = 0 \\ \boldsymbol{H}_1&: \beta_i \neq 0 \end{aligned} \tag{3}$$

and $\widehat{\boldsymbol{\beta}}|\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1})$, while T is a test statistic with a known probability distribution:

$$T = \frac{\hat{\beta}_i - \beta_i}{SE(\hat{\beta}_i)/\sqrt{n}} \sim t\,(n-k-1) \tag{4}$$

with $SE(\hat{\beta}_i)$ the standard error of $\hat{\beta}_i$. As n increases, $T_n \xrightarrow{d} N(0,1)$. Decision criteria is defined following Fig. 1.

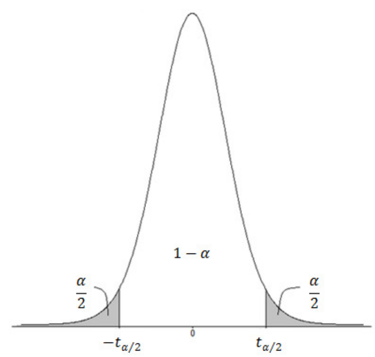

Figure 1: Distribution of T under $\boldsymbol{H}_0$.

The quantity T_{obs} is the observed value of the random variable T when all variables in (4) are substituted by their respective values. If $|T_{obs}| > t_{\alpha/2}$, T_{obs} is far away from the average of the curve in Fig. 1 and thus is less likely that T_{obs} is indeed generated by the distribution of T under $\boldsymbol{H}_0$. In this case, $\boldsymbol{H}_0$ is rejected and β_i remains in (1). Otherwise, if $|T_{obs}| < t_{\alpha/2}$, T_{obs} is probably generated by the distribution of T under $\boldsymbol{H}_0$. In this case, $\boldsymbol{H}_0$ is not rejected and β_i quits (1), because it is not statistical significant. Typically, $\alpha = 5\%$, $t_{\alpha/2} = 1.96$ and $-t_{\alpha/2} = -1.96$.

2.4 Statistical Significance and Accuracy

The Root Mean Square Error (RMSE) is a typical accuracy measure used in SR experiments. Its relation with $\overline{\mathrm{R}}^2$, a typical metric of fitness in cross-sectional econometrics that linearly penalizes by addition of non-statistically significant regressors with $|T_{obs}| < 1.00$, is given by:

$$\overline{\mathrm{R}}^2 = 1 - \left(\frac{n(\mathrm{RMSE})^2}{(\boldsymbol{y}-\overline{\boldsymbol{y}})^{\mathrm{T}}(\boldsymbol{y}-\overline{\boldsymbol{y}})}\right)\left(1 + \frac{k}{n-k-1}\right) \tag{5}$$

with $\bar{\boldsymbol{y}}$ the average of $\boldsymbol{y}$. By (5), it is concluded that the RMSE minimization implies $\bar{\mathrm{R}}^2$ maximization, maintaining others factors constant.

It is stated in [9] that $\bar{\mathrm{R}}^2$ will increase if, and only if, $|T_{obs}| > 1.00$ in a two-sided HT for β_i, with null and alternative hypothesis, test statistics and decision criteria stated as before. EGP uses this mathematical proof to increase accuracy of its individuals by statistical significance, while simultaneously controlling bloat. To achieve fitness improvement with statistical significance, EGP uses 1.96 as threshold in HT instead of 1.00.

3 ECONOMETRIC GENETIC PROGRAMMING

EGP evolves models in format of (1) through GP, which is responsible for model selection. GP is mainly based in configuration presented in [15], as well as EGP parameters.

3.1 Representation

Individuals / programs / regressions / models are multigenic. Any constant in any program comes from LS in (2), i.e. there are no ephemeral constants. The terminal set, namely Ω, is purely composed by variables. The primitive set, namely ϑ, is composed just by variables and operations of sum and multiplication, due (1) format.

3.2 Initial Population

EGP uses a probabilistic version of ramped half-and-half method. Fig. 2 shows a possible individual generated by EGP.

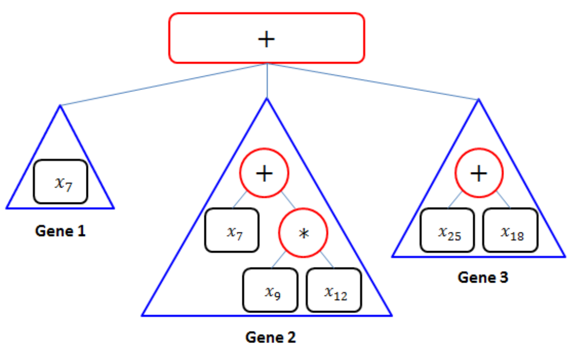

Figure 2: A possible individual generated by EGP.

Set Ω is composed by K features (independent variables). Every individual has its own set of regressors, forming its own $\boldsymbol{X}$, composed by simple or combined elements of Ω. As an example, it is possible that x_1, $x_3x_{11}^2$ and $x_3x_4x_6$ are regressors of a particular individual, formed by features x_1, x_3, x_4, x_6 and x_{11}.

3.3 Accuracy

RMSE is the objective function. The $\bar{\mathrm{R}}^2$ is just used to compare models. Ideally, $\bar{\mathrm{R}}^2$ would be the objective function, but some issues, principally when $k > n$, makes it little suitable in practice.

To calculate accuracy in an EGP individual, the one showed in Fig. 2 needs to be transformed into a model like $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$. Then, EGP will solve $\hat{\boldsymbol{\beta}}$ for $X\hat{\boldsymbol{\beta}} = \beta_1x_7 + \beta_2x_9x_{12} + \beta_3x_{18}x_{25}$. If any of the regressors are not statistically significant, they will be removed from $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{u}$. In sequence, $\hat{\boldsymbol{\beta}}$ is recalculated just with statistically significant regressors. RMSE is finally calculated using $\hat{\boldsymbol{\beta}}$ after these steps. This routine is traditional in econometric studies, ensuring statistical significance over a determined significance level α, and that is the way EGP performs feature selection. Modifications described are necessary just for accuracy calculation, therefore individuals will keep their multigene structure to mutation, crossover and elitism.

EGP is not a kind of stepwise regression, because it does not build a model sequentially, variable by variable, as described in [20]: "(Forward) Stepwise regression builds a model sequentially, adding one variable at a time. At each step, it identifies the best variable to include in the active set, and then updates the least squares fit to include all the active variables."

EGP does not estimate on genes, just on regressors, by two main reasons: possible multicollinearity problem, interfering on HT for β_i, and lack of interpretation for $\hat{\beta}_i$ when it is related to a gene.

3.4 Selection

Tournament selection with $n_{tourn} = 7$ and repetitions allowed, with a variation on lexicographic parsimony pressure of [16], is used. Individuals with a large number of statistically significant regressors will be preferred over others with a few number, in the same range of fitness. Therefore, EGP is parsimonious in its nature, because it avoids the individuals with a large amount of i*ntrons* (in this case, non statistically significant regressors).

3.5 Mutation, Crossover and Elitism

Types of mutation used: traditional mutation proposed by [4] and mutation by regressors' substitution. Types of crossover used: intergenic and intragenic crossovers. Mutation and crossover rates vary through evolution following automatic adaptation of operators as described in [17]. Elitism rate is settled to 5% of individuals by generation.

3.6 Tools and Parameters

EGP is implemented through a modification on GPTIPS, a *Matlab* toolbox, presented in [18]. Information on EGP parameters are shown in Table 1.

Table 1: **EGP Parameters**

Parameters	
- Population size	150.
- Generations	50.
- Maximum gene Depth	5.
- Maximum number of genes by individual	5.
- Probability of traditional mutation [4]	95%.
- Probability of intragenic crossover	50%.

4 EXPERIMENTS AND RESULTS

Table 2 presents the results. The training set contains 70% of the data samples. Results for EGP ("EGP-Regression" in the table) are the average of 50 runs for best individuals. EGP is carefully constructed considering econometric theory on cross-sectional datasets and thus needs cross-sectional datasets to be tested. Time series datasets are largely available, while the same is not true for cross-sectional datasets. Considering availability and the quality of datasets, the following ones have been chosen to compare EGP with traditional feature-selection econometric algorithms: "Concrete Compressive Strength"; "Housing" and "Airfoil Self-Noise (Nasa)". All information on datasets can be found in UCI Machine Learning Repository [19].

Table 2: **EGP and traditional feature-selection econometric algorithms**

Dataset : Concrete Compressive Strength

Position	Algorithm	Adjusted R^2 Training Set	Adjusted R^2 Test Set
1	**EGP-Regression**	**0,799**	**0,822**
2	Partial Least Squares Regression	0,597	0,658
3	Ridge Regression	0,604	0,643
4	Stepwise Forward Regression	0,604	0,640

(**to details on the dataset**: https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength)

Dataset : Housing

Position	Algorithm	Adjusted R^2 Training Set	Adjusted R^2 Test Set
1	**EGP-Regression**	**0,769**	**0,746**
2	Stepwise Forward Regression	0,739	0,710
3	Ridge Regression	0,699	0,673
4	Partial Least Squares Regression	0,624	0,572

(**to details on the dataset**: https://archive.ics.uci.edu/ml/datasets/Housing)

Dataset : Airfoil Self-Noise (Nasa)

Position	Algorithm	Adjusted R^2 Training Set	Adjusted R^2 Test Set
1	**EGP-Regression**	**0,629**	**0,673**
2	Stepwise Forward Regression	0,516	0,451
3	Partial Least Squares Regression	0,480	0,418
4	Ridge Regression	... *1	... *1

(**to details on the dataset**: https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise)
(***1**: Adjusted R^2 for Training and Test Sets were inconclusive in sinal and were not reported)

The main reason EGP outperforms traditional econometric feature-selection methods is its capability to explore the regressors' and models' search space. The number of regressors for each of the algorithms shows it. Partial Least Squares Regression, Ridge Regression and Stepwise Forward Regression have at most 8, 13, and 5 regressors, respectively, in its generated regressions for each dataset. For Housing Dataset, as an example, EGP generates models with 50 statistically significant regressors or more. EGP explores non-linearity of features, by multiplying different features, maintaining models with linear structure.

5 CONCLUSION

EGP was successful in achieving its objective of generating high accuracy regressions with potential interpretability of parameters. Feature and model selection performed well, when comparing with traditional methods, as previously shown in the results. Statistical significance proved to be a powerful feature selection tool, directly and efficiently identifying introns and controlling bloat.

REFERENCES

[1] A.M. Legendre, 1805. Nouvelles méthodes pour la détermination des orbites des comètes, Firmin Didot *Commun.* Firmin Didot, Paris.

[2] C.F. Gauss, 1809. Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientum.

[3] J. W. Davidson, D. Savic, and G. A.Walters. 1999. Method for the identification of explicit polynomial formulae for the friction in turbulent pipe flow. *Comm. Journal of Hydroinformatics* 1, 2 (1999), 115–126.

[4] J. R. Koza. 1992. *Genetic Programming: On the Programming of Computers by Means of Natural Selection (Complex Adaptive Systems)* (1st. ed.). The MIT Press.

[5] I. Arnaldo, K. Krawiec, and U.-M. O'Reilly. 2014. Multiple regression genetic programming In *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation* (GECCO'14). ACM, New York, NY, USA, 879–886.

[6] J. W. Davidson, D. Savic, and G. A.Walters. 2003. Symbolic and numerical regression: experiments and applications. *Comm. Information Sciences* 150, 1-2 (2003), 95–117.

[7] O. Giustolisi, and D. Savic. 2006. A symbolic data-driven technique based on evolutionary polynomial regression. *Comm. Journal of Hydroinformatics* 8, 3 (2006), 207-222.

[8] A. L. F. Novaes, R. Tanscheit, and D. M. Dias. 2016. Programação Genética Econométrica Aplicada a Problemas de Regressão em Conjuntos de Dados Seccionais. In *Proceedings of XIII Encontro Nacional de Inteligência Artificial* (ENIAC'16). Recife, PE.

[9] J. Wooldridge. 2009. *Introductory Econometrics: A Modern Approach* (4 ed.). Cengage Learning.

[10] R. Davidson, and J. MacKinnon. 1993. *Estimation and Inference in Econometrics* (1 ed.). Oxford University Press.

[11] J. M. Chambers. 1977. *Computational Methods for Data Analysis (Probability & Mathematical Statistics)* (1 ed.). John Wiley & Sons, New York.

[12] J. H. Maindonald. 1984. *Statistical Computation* (1 ed.). Wiley, New York.

[13] V. V. De Melo. 2014. Kaizen programming. In *Proceedings of the 2014 Conference on Genetic and Evolutionary Computation* (GECCO'14). ACM, New York, NY, USA, 895–902.

[14] A. L. F. Novaes. 2015. *Programação Genética Econométrica: uma Nova Abordagem para Problemas de Regressão e Classificação em Conjuntos de Dados Seccionais.* Master's thesis. Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Brazil.

[15] R. Poli, W. B. Langdon, and N. F. McPhee. 2008. *A Field Guide to Genetic Programming* (1 ed.). Lulu Enterprises, United Kingdom.

[16] S. Luke, and L. Panait. 2002. Lexicographic parsimony pressure. In *Proceedings of the 2002 Conference on Genetic and Evolutionary Computation* (GECCO'02). ACM, San Francisco, CA, 829–836.

[17] S. Silva, and J. Almeida. 2003. Gplab – a genetic programming toolbox for matlab. In *Proceedings of the Nordic MATLAB conference*. 273–278.

[18] D. P. Searson, D.E. Leahy, and M. J. Willis. 2010. GPTIPS: an open source genetic programming toolbox for multigene symbolic regression. In *Proceedings of The International Multiconference of Engineers and Computer Scientists 2010* (IMECS'10). Hong Kong,, 77–80.

[19] UCI Machine Learning Repository, http://archive.ics.uci.edu/ml, last accessed 2015/02/24, University of California, School of Information and Computer Science, Irvine, CA.

[20] T. Hastie, R. Tibshirani, and J. Friedman. 2011. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2 ed.). Springer.