This is the accepted manuscript of the article:

Campayo-Sanchez, F., & Nicolau, J. L. (2024). Optimizing control variable selection with algorithms: Parsimony and precision in regression analysis. *Tourism Economics*, 13548166241287953.

OPTIMIZING CONTROL VARIABLE SELECTION WITH ALGORITHMS: PARSIMONY AND PRECISION IN REGRESSION ANALYSIS

Abstract

This research note explores the pivotal role of control variables in any tourism and hospitality research that utilizes regression models in statistical analyses. While theory-driven independent variables offer insight into expected effects, the inclusion of control variables is crucial for mitigating potential confounding factors. In an attempt to strike a balance between model complexity and parsimony, researchers face the challenge of selecting the optimal control variables. To address this issue, the study tests three alternative methods: genetic algorithms, lasso models, and the branch and bound algorithm. Despite their underutilization in tourism research, these methods offer efficient means of selecting control variables, enhancing model precision and interpretation without unnecessarily convoluting the model with irrelevant factors.

Keywords: variable selection; control variables; genetic algorithms; lasso models; branch and bound algorithm.

Introduction

There are multiple studies that include control variables in their models of tourism demand or hotel financial performance (e.g., Li et al., 2020; Wang et al., 2019). Guided by theories that explain a phenomenon in tourism and hospitality, researchers extensively use regression models wherein the main independent variables aligned with those theories are expected to explain the dependent variable. Additionally, control variables[1], while not necessarily informing theory, must be included to account for other potential effects, which may represent alternative sources of variation of the dependent variable that, if not included, could distort the interpretation of the influence of the independent variables of interest (Su et al., 2019). Their inclusion facilitates the isolation of the specific effects of the independent variables, thereby increasing the precision and accuracy of the parameter estimates (Chernozhukov et al., 2015). However, the researcher is always confronted with the dilemma of which control variables to include. Too many control variables could give rise to the issue of overfitting, and too few control variables may lead to the risk of) not capturing external confounding factors (Chernozhukov et al., 2015; Su et al. 2019).

More importantly, should the selection of control variables be guided by economic theories or statistical criteria? Models should balance theory-driven variables and the need for model parsimony. Consequently, researchers must start—and this is indisputable—with economic theories to identify potential control variables, thereby warranting and ensuring that any variables eventually selected and included in the model contribute meaningfully to the understanding of the phenomenon under investigation. However, the final selection may be reached after a process that considers model diagnostics.

This research note presents and compares three alternative methods for variable selection not widely used in tourism research despite their usefulness in selecting control variables efficiently: 1) methods based on genetic algorithms (Holland, 1975) (for a tourism example see Olorunsola et al., 2023); lasso models (Tibshirani, 1996) (for a tourism example see You et al., 2021); and 3) the branch and bound algorithm developed by Hofmann et al. (2020) (see Appendix A for a description of each method).

The two main advantages of these methods are that they allow for the identification of a combination of control variables that enhance the explanatory/predictive power of the model and lead to more parsimonious estimates by eliminating control variables that are statistically irrelevant in explaining the dependent variable. Thus, through the use of these methodologies, the approach proposed in the present study contributes to changing how statistical inferences are made. The main benefit of this method is that researchers can fully focus on the research questions they aim to explore (through the variables of theoretical interest) by allowing mathematical algorithms to handle the selection of control variables that should be included in the regression model.

[1] Control variables in regression analysis are independent variables that researchers include in a regression model to account for potential confounders (Su et al., 2019). These variables are not the main focus of the study, but are controlled to reduce biases and to improve the accuracy of the estimated relationships between the main independent variable(s) and the dependent variable (Chernozhukov et al., 2015).

Empirical Application

Using Compustat data, we compiled an unbalanced panel dataset with 162 annual observations from nine hotel companies (Choice Hotels, Extended Stay America, Hilton, Hyatt, InterContinental Hotels, Marriott, Red Lion, Starwood, and Wyndham) from 1999 to 2022. For illustrative purposes, based on the assumption that firm size is a determinant factor of firm performance, the independent variable will be the Total assets, with Profits and Sales as dependent variables. Possible control variables include Acquisitions, Dividends, Number of employees, Net income, Net cash flow from operating activities, Retained earnings, Book value per share, and dummies related to hotels and years (to control for other potential unobservable factors). Six regressions are estimated with Sales as the dependent variable:

$$Sales_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \mu_i + \phi_t + \varepsilon_{it}\ , (1)$$

$$Sales_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k=2}^{K} \beta_k x_{itk} + \mu_i + \phi_t + \varepsilon_{it,}\ (2)$$

$$Sales_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k^{CVL}=2}^{K^{CVL}} \beta_{k^{CVL}}\, x_{itk^{CVL}} + \mu_i + \phi_t + \varepsilon_{it}\ , (3)$$

$$Sales_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k^{AL}=2}^{K^{AL}} \beta_{k^{AL}}\, x_{itk^{AL}} + \mu_i + \phi_t + \varepsilon_{it}\ , (4)$$

$$Sales_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k^{BBA}=2}^{K^{BBA}} \beta_{k^{BBA}}\, x_{itk^{BBA}} + \mu_i + \phi_t + \varepsilon_{it}\ , (5)$$

$$Sales_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k^{GA}=2}^{K^{GA}} \beta_{k^{GA}}\, x_{itk^{GA}} + \mu_i + \phi_t + \varepsilon_{it}\ , (6)$$

where $\sum_{k=2}^{K} \beta_k x_{itk}$ represents all control variables, and $\sum_{k^{CVL}=2}^{K^{CVL}} \beta_{k^{CVL}}\, x_{itk^{CVL}}$, $\sum_{k^{AL}=2}^{K^{AL}} \beta_{k^{AL}}\, x_{itk^{AL}} + \varepsilon_i$, $\sum_{k^{BBA}=2}^{K^{BBA}} \beta_{k^{BBA}}\, x_{itk^{BBA}}$, and $\sum_{k^{GA}=2}^{K^{GA}} \beta_{k^{GA}}\, x_{itk^{GA}}$ represent the control variables selected through lasso with cross validation, adaptive lasso, the branch and bound algorithm, and the genetic algorithm, respectively, and μ_i and ϕ_t represent the hotel and year fixed effects, respectively. We compare the six models to assess the importance of including control variables in the regression and analyze which methodology contributes most to improving the goodness of fit. We perform the same procedure using Profits as the dependent variable:

$$Profits_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \mu_i + \phi_t + \varepsilon_{it}\ , (7)$$

$$Profits_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k=2}^{K} \beta_k x_{itk} + \mu_i + \phi_t + \varepsilon_{it,}\ (8)$$

$$Profits_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k^{CVL}=2}^{K^{CVL}} \beta_{k^{CVL}}\, x_{itk^{CVL}} + \mu_i + \phi_t + \varepsilon_{it}\ , (9)$$

$$Profits_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k^{AL}=2}^{K^{AL}} \beta_{k^{AL}}\, x_{it\ \ AL} + \mu_i + \phi_t + \varepsilon_{it}\ , (10)$$

$$Profits_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k^{BBA}=2}^{K^{BBA}} \beta_{k^{BBA}} x_{it\ \ BBA} + \mu_i + \phi_t + \varepsilon_{it}\ , (11)$$

$$Profits_{it} = \beta_0 + \beta_1 Total\ assets_{it} + \sum_{k^{GA}=2}^{K^{GA}} \beta_{k^{GA}} x_{itk^{GA}} + \mu_i + \phi_t + \varepsilon_{it}\ , (12)$$

Finally, we compare Models 3–6 with Models 9–12 to assess whether the methods used can capture nuances and distinctive elements that impact each dependent variable by selecting control variables.

Results

Table 1 presents the models using Sales (Models 1–6) and Profits (Models 7–12) as dependent variables. Robust standard errors for heteroscedasticity are used because the Breusch-Pagan tests are significant. According to Hair et al. (2010), collinearity is not an issue because the VIF values do not exceed the threshold of 10 (see Appendix B which includes the correlation matrix and the VIFs of the models).

Focusing on the goodness-of-fit parameters of the regressions in Table 1, we observe that Model 2 (which includes the independent and control variables) is substantially better than Model 1 (containing only the independent variable). Specifically, adding control variables significantly enhances the adjusted R-squared (increasing by 0.321), and the Akaike Information Criterion (decreasing by 277.388). Hence, the need to include control variables in the model is evident. However, is it possible to improve goodness of fit by optimizing variable selection? Comparing Model 2 with Models 4, 5, and 6 (which include only the variables that the adaptive lasso model, the branch and bound algorithm, and the genetic algorithm deems relevant for explaining hotel sales, respectively), we detect improvements in adjusted R-squared (increasing by 0.001), and in the Akaike Information Criterion (decreasing by 3.375). This finding demystifies the tendency to think that controlling more aspects within a regression model makes it more feasible to increase the precision of results and underscores the utility of resorting to algorithms for optimal control variable selection[2]. Similar conclusions are reached when the variable Profits is used as the dependent variable in the regressions estimated in Table 3. However, in this case, only the branch and bound algorithm and the genetic algorithm improve the goodness-of-fit parameters. Consequently, it appears that these two methods are superior to lasso.

[2] The lasso model with cross validation chooses all variables, so it exactly matches Model 2.

Table 1. Results of regression models using Sales (Models 1–6) and Profits (Models 7–12) as the dependent variable (standard errors in parentheses)

	Model 1	Model 2	Model 3 CVL	Model 4 AL	Model 5 BBA	Model 6 GA	Model 7	Model 8	Model 9 CVL	Model 10 AL	Model 11 BBA	Model 12 GA
Intercept	463.968	1146.056	1146.056	1156.666	1156.666	1156.666	350.428[a]	1494.338[a]	1467.649[a]	1472.919[a]	1473.401[a]	1473.401[a]
	(294.732)	(827.904)	(827.904)	(825.096)	(825.096)	(825.096)	(47.565)	(162.641)	(171.000)	(169.949)	(170.379)	(170.379)
Independent variable												
AssetsTotal	0.556[a]	0.147[a]	0.147[a]	0.145[a]	0.145[a]	0.145[a]	0.098[a]	0.068[a]	0.074[a]	0.073[a]	0.073[a]	0.073[a]
	(0.048)	(0.041)	(0.041)	(0.040)	(0.040)	(0.040)	(0.007)	(0.012)	(0.011)	(0.010)	(0.011)	(0.011)
Control variables												
Acquisitions		-0.101	-0.101					-0.047	-0.039			
		(0.253)	(0.253)					(0.036)	(0.035)			
DividendsTotal		-0.208	-0.208					0.079	0.062	0.067		
		(0.420)	(0.420)					(0.145)	(0.141)	(0.142)		
Employees		48.856[a]	48.856[a]	48.779[a]	48.779[a]	48.779[a]		-7.882[a]	-8.047[a]	-8.056[a]	-8.041[a]	-8.041[a]
		(3.564)	(3.564)	(3.606)	(3.606)	(3.606)		(0.997)	(0.993)	(0.998)	(0.997)	(0.997)
NetIncomeLoss		0.392	0.392	0.399	0.399	0.399		0.261	0.281[c]	0.286[c]	0.289[c]	0.289[c]
		(0.452)	(0.452)	(0.445)	(0.445)	(0.445)		(0.162)	(0.164)	(0.163)	(0.162)	(0.162)
OperatingActivitiesNetCash		1.864[a]	1.864[a]	1.882[a]	1.882[a]	1.882[a]		0.494[a]	0.468[b]	0.479[a]	0.482[a]	0.482[a]
		(0.524)	(0.524)	(0.522)	(0.522)	(0.522)		(0.183)	(0.183)	(0.182)	(0.183)	(0.183)
Retained Earnings		0.205[a]	0.205[a]	0.196[a]	0.196[a]	0.196[a]		-0.051[a]	-0.055[a]	-0.058[a]	-0.057[a]	-0.057[a]
		(0.064)	(0.064)	(0.062)	(0.062)	(0.062)		(0.016)	(0.016)	(0.015)	(0.015)	(0.015)
BookValuePerShare		-65.181[a]	-65.181[a]	-65.721[a]	-65.721[a]	-65.721[a]		8.792[c]				
		(22.083)	(22.083)	(21.718)	(21.718)	(21.718)		(4.695)				
Year fixed effects	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Hotel fixed effects	No	Yes	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes
Observations	162	162	162	162	162	162	162	162	162	162	162	162
Adjusted R-squared	0.623	0.944	0.944	0.945	0.945	0.945	0.581	0.886	0.884	0.884	0.885	0.885
F-statistic	266.5[a]	72.27[a]	72.27[a]	77.21[a]	77.21[a]	77.21[a]	224.40[a]	33.82[a]	34.17[a]	35.19[a]	36.40[a]	36.40[a]
Akaike Information Criterion	3036.227	2758.839	2758.839	2755.464	2755.464	2755.464	2501.791	2322.862	2324.464	2323.349	2321.705	2321.705

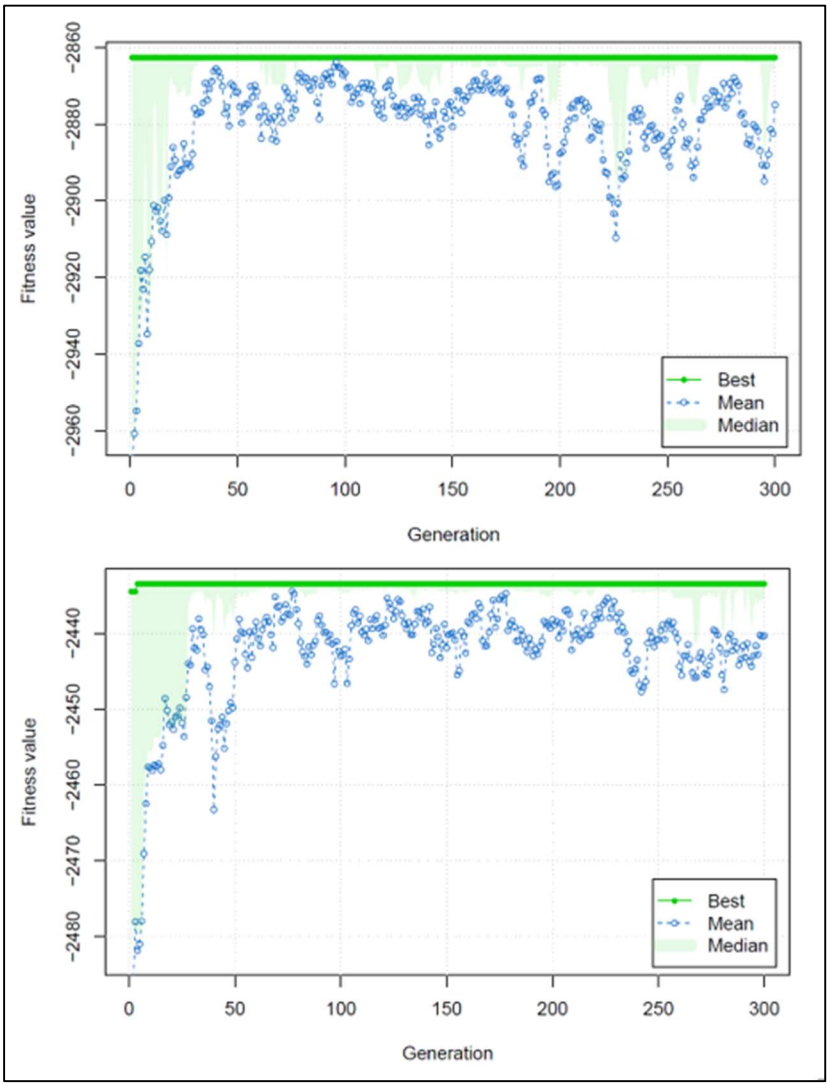
a=p-value<0.01; b=p-value<0.05; c= p-value<0.1

Note: Following the approach proposed by Franco et al. (2020), we use total assets (and not a logarithmic transformation) in the analyses developed in the study.

Lastly, if we compare Models 3–6 with Models 9–12, we observe that the algorithms used have selected different regressors, even though both dependent variables reflect outcome measures. This finding appears to reflect the capacity of these stochastic search techniques to uncover nuances (not always explainable from a theoretical standpoint) underlying the data.

In short, and according to the results, the best methods for the selection of control variables seem to be the branch and bound algorithm and the genetic algorithm. However, since genetic algorithms are general-purpose optimizers (which can be used with other techniques in addition to linear regression), they can be particularly attractive when researchers are not faced with a problem of high complexity. The graphs in Figure 1 summarize the stochastic search process followed by the genetic algorithm to select the variables that maximize the Akaike Information Criterion in Models 6 and 12.

Figure 1. Best model and mean and median Akaike Information Criterion of all chromosomes in each of the 300 generations of the genetic algorithms used to select variables for Models 6 and 12



Finally, the analyses in Appendix C do not reveal that our estimates are affected by endogeneity.

Concluding remarks

In this research note, we underscore the importance of parsimony when selecting control variables by presenting the use of different algorithms in the context of regression analysis.

Explanatory power, measured by R-squared values, tends to increase with the addition of a new independent variable to the model, regardless of the actual significance of the variables. This outcome may create a misleading impression of improved model fit solely due to the inclusion of more variables, even if their relationship with the dependent variable is merely coincidental. In fact, excessive inclusion of independent variables in a model can lead to overfitting, where the model becomes overly tailored to the idiosyncrasies and random noise present in the sample data rather than accurately representing the broader population, thus inflating R-squared values and undermining the model's ability to make accurate predictions. Contrary to the common misconception that a higher R-squared value always indicates a better model (James et al. 2013), it is critical to recall that, while a high R-squared value suggests that the independent variables explain a large portion of the variance in the dependent variable, it does not necessarily mean that the model is accurate. As pointed out, inflated R-squared values derived from overfitting can lead to a false sense of model performance.

Therefore, the selection of control variables is critical in the context of regression analysis because, while adding more independent variables increases the complexity of the model, it is essential to strike a balance between model complexity and parsimony. While including more variables may improve the model's fit to the sample data (as measured by R-squared), it can also lead to overfitting and poor generalization to new data. The general goal of regression analysis is to build a model that can accurately predict the dependent variable in new, unseen data. Accordingly, by selecting independent variables that capture the underlying relationships in the data without overfitting to the idiosyncrasies of the sample data, the model can better generalize to new observations.

While theory should guide the initial choice of control variables, statistical diagnostics ultimately determine the selection of control variables. At this stage, genetic algorithms and branch and bound algorithms offer an efficient selection method. The empirical results obtained, as an exemplificative application, show that controlling for many factors does not necessarily increase the precision of the results. More importantly, our results show that the use of genetic algorithms and branch and bound algorithms proves to be valuable in selecting optimal control variables

References

Chernozhukov, V., Hansen, C., & Spindler, M. (2015). Post-selection and post-regularization inference in linear models with many controls and instruments. *American Economic Review*, *105*(5), 486-490.

Franco, S., Caroli, M. G., Cappa, F., & Del Chiappa, G. (2020). Are you good enough? CSR, quality management and corporate financial performance in the hospitality industry. *International Journal of Hospitality Management*, *88*, 102395.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010) *Multivariate Data Analysis. Boston*. Cengage.

Hofmann, M., Gatu, C., Kontoghiorghes, E.J., Colubi Cervero, A.M., & Zeileis, A. (2020). Lmsubsets: Exact variable-subset selection in linear regression for R. *Journal of Statistical Software*.

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Li, Z., Shu, H., Tan, T., Huang, S., & Zha, J. (2020). Does the demographic structure affect outbound tourism demand? A panel smooth transition regression approach. *Journal of Travel Research*, *59*(5), 893-908.

Olorunsola, V.O., Saydam, M.B., Arici, H. E., & Köseoglu, M.A. (2023). The predictive roles of financial indicators and governance scores on firms' emission performance in the tourism and hospitality industry. Tourism Economics, 13548166231207132.

Su, L., Ura, T., & Zhang, Y. (2019). Non-separable models with high-dimensional data. *Journal of Econometrics*, *212*(2), 646-677.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267-288.

Wang, K., Miao, Y., Chen, M. H., & Hu, D. (2019). Philanthropic giving, sales growth, and tourism firm performance: An empirical test of a theoretical assumption. *Tourism Economics*, *25*(6), 835-855.

You, P.S., Chen, M.H., & Su, C.H. (2021). Travel agent's tour selection and sightseeing bus schedule for group package tour planning. Tourism Economics, 27(1), 220-242.

Appendix A. Methods for variable selection

Genetic Algorithms

Genetic algorithms replicate certain implicit aspects of biological evolution and natural selection in species (e.g., Goldberg, 1989; Holland, 1975). With the analogy of living organisms' adaptation, each of them possessing chromosomes that undergo alterations as members reproduce, resulting in species more likely to endure, let us consider the estimation of multiple linear regression with a sample composed of $i = 1, \dots, n$ observations:

where y_i is the dependent variable, $X = \{x_{i1}, x_{i2}, \dots, x_{iK}\}$ represents the $k = 1, \dots, K$ explanatory variables, $\beta = \{\beta_0, \beta_1, \beta_2, .., \beta_K\}$ are the model coefficients and ε_i is the error term. We turn to a genetic algorithm to select regressors providing genuine information about the dependent variable.

First, we randomly generate a starting population with $m = 1, \dots, S$ vectors (also termed chromosomes) representing different model specifications. Specifically, each vector identifies a numeric sequence (or string) comprising as many elements (genes) as variables that could potentially be included in the regression:

$$\begin{pmatrix} m_1^{(0)} \\ \vdots \\ m_S^{(0)} \end{pmatrix} = \begin{pmatrix} \left(g_{11}^{(0)}, g_{12}^{(0)}, \dots, g_{1K}^{(0)}\right) \\ \vdots \\ \left(g_{S1}^{(0)}, g_{S2}^{(0)}, \dots, g_{SK}^{(0)}\right) \end{pmatrix}$$

Each gene is randomly assigned a binary code based on the presence (1) or absence (0) of the variable they identify within the model[3]. For example, if $m_1^{(0)} = \{1,0,1,1\}$, all variables except the second one will be included when estimating the regression. Once the initial population is created, the performance of potential candidate models represented by different chromosomes is evaluated using a fitness function. In this research note, we use the Akaike Information Criterion, one of the most widely used goodness-of-fit measures in the literature, to select the best regression (e.g., Wan & Song, 2018).

$$\begin{pmatrix} f\left(m_1^{(0)}\right) \\ \vdots \\ f\left(m_S^{(0)}\right) \end{pmatrix} = \begin{pmatrix} AIC\left(m_1^{(0)}\right) \\ \vdots \\ AIC\left(m_S^{(0)}\right) \end{pmatrix}$$

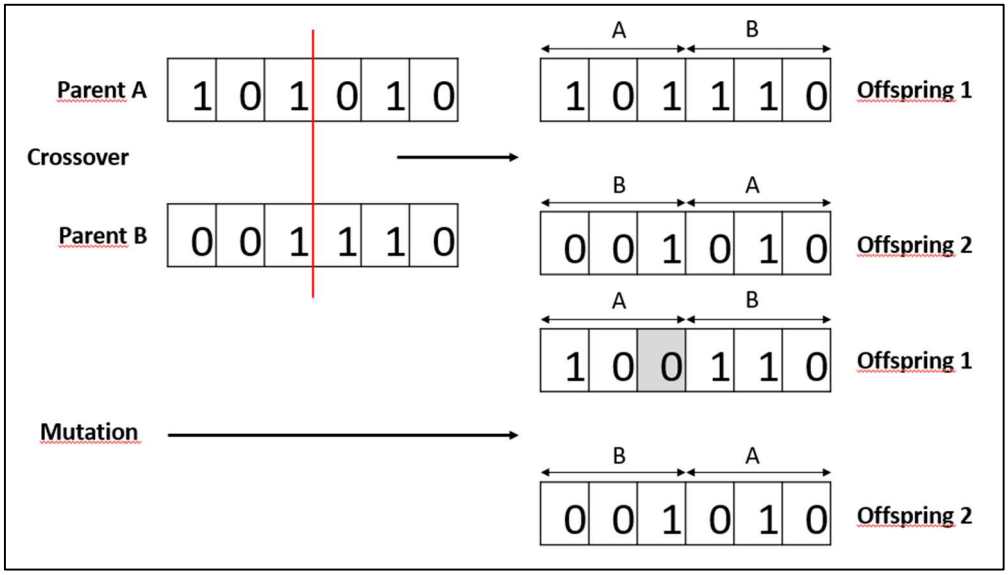
Subsequently, we construct a new population through two processes known as crossover and mutation (see Figure A1)[4]. In the crossover process, new candidate models are generated by combining the genomes of two sequences from the initial population. Through mutation, the value of one of the genes in the new chromosome is randomly altered (changing from 0 to 1 or from 1 to 0). We utilize the fitness function to determine which models will reproduce, thereby

[3] We do not add a gene for the intercept within each chromosome because we will always include it by default in the regression models.

[4] Consistent with prior literature (Scrucca, 2013), we will set the crossover and mutation probabilities at 80% and 10%, respectively.

transmitting their genetic heritage. Specifically, chromosomes with smaller Akaike Information Criterion values will have a higher probability of being chosen for offspring. However, because our aim is to estimate the best model, we also employ an elitist strategy, consistently incorporating the top 5% of models with the best Akaike Information Criterion into the resulting new population. We will iterate through these processes over 300 generations[5], and will ultimately select and estimate the regression with the lowest Akaike Information Criterion identified by the algorithm.

Figure A1. Crossover and mutation



Lasso model with cross-validation and adaptive Lasso model

The lasso models find the optimal number of regressors by minimizing the prediction error under the condition that the model is not too complex. These algorithms measure the complexity through the sum of the absolute values of the $K = \{\beta_1, \beta_2, \ldots, \beta_K\}$ coefficients associated with the variables included in the estimation. The optimal solution is obtained by minimizing the following function:

$$min\left(\frac{1}{2N}(Y - X\beta')'(Y - X\beta') + \lambda\sum_{k=1}^{K}|\beta_k|\right)$$

where $RSS = (Y - X\beta')'(Y - X\beta')$ is the residual sum of squares and $\varphi = \lambda\sum_{k=1}^{K}|\beta_k|$ is the term that penalizes complexity and encourages the omission of econometrically irrelevant regressors, making it possible to estimate more parsimonious and sparse models. The higher the value associated with λ the greater the penalty, leading to the selection of a model that uses a smaller number of explanatory variables.

[5] While 100 generations are usually more than enough (Goldberg 1989), in this research note, we tripled this value to ensure that the genetic algorithm will always stably find the best model.

The stochastic search algorithm identifies λ_{gmax} as the λ associated with the regression that does not include any control variables. Since $\lambda = 0$ corresponds to a model that presents the maximum level of complexity, this method will try to find a λ^* that minimizes the prediction error without greatly increasing the complexity. Therefore, the first step is to estimate the model associated with λ_{gmax} (which does not include control variables). This regression will constitute the starting point to begin the search. The algorithm will gradually reduce the value of λ (so that subsequent models will have as many or more control variables as the models preceding them) until λ^* is reached. However, since the data sets used in the regressions are finite, we will have to fit additional models to be reasonably sure that we are choosing λ^* and the best regression.

One of the methods we will use in this paper for variable selection is called cross validation. The criterion used is based on minimizing an estimate of the out-of-sample prediction error represented by the function $f(\lambda)$. To get to the λ^* that minimizes that function, we will fit several models with λs close to λ^*. Specifically, when we go from model(λ_{k-1}) to model(λ_k), he algorithm performs the following calculation:

$$\frac{deviance\{model(\lambda_{k-1})\} - deviance\{model(\lambda_k)\}}{deviance\{model(\lambda_{k-1})\}}$$

This relative difference is a measure of how much predictive capability is added by model(λ_k) to model(λ_{k-1}). If the difference is meager, the divergence between the cross validation function values (λ_{k-1}) and $f(\lambda_k)$ will be small, so the changes in the function for λs with smaller values will be even smaller[6]. In this scenario we will consider that we have reached the λ^* in model(λ_k) if $f(\lambda_1 = \lambda_{gmax}) > \cdots > f(\lambda_{k-1}) > f(\lambda_k) < f(\lambda_{k+1}) < f(\lambda_{k+2}) < f(\lambda_{k+3})$.

The second method we will resort to for variable selection is the adaptive lasso. Although the procedure is analogous to that used in the lasso with cross validation, in the adaptive lasso the solution is obtained by minimizing the following expression:

$$min\left(\frac{1}{2N}(Y - X\beta')'(Y - X\beta') + \lambda\sum_{k=1}^{K}\omega_k|\beta_k|\right)$$

where $\omega_k = 1/|\hat{\beta}_k|$ is the vector of adaptive weights. The initial estimators, denoted as $\hat{\beta}$, are obtained through cross-validation lasso model estimation[7]. The algorithm eliminates all those control variables for which $\hat{\beta}_k = 0$. The inclusion of ω_k causes the adaptive lasso to penalize coefficients with lower initial estimates more heavily, which means that the specification finally chosen tends to have a smaller number of regressors compared to the models selected through the cross-validation lasso method.

The branch and bound algorithm developed by Hofmann et al. (2020)

[6] Following the STATA (2018) manual, we will stop estimating models associated with smaller λ when the relative difference between model(λ_k) and model(λ_{k-1}) is less than 1e-5.

[7] Other authors recommend a ridge regression when collinearity may become a problem if the control variables are highly correlated (Zou, 2006).

Hofmann et al. (2020) propose an efficient methodology to select control variables. n ordinary least squares $\hat{\beta}_{OLS} = \underset{\beta}{\operatorname{argmin}}\, RSS(\beta)$ is the estimator of β and $RSS(\beta) = \|Y - X\beta\|_2^2$ is the residual sum of squares. Let $P = \{1, \ldots, K\}$ be the set of all control variables that can potentially be included in the regression, a submodel (composed of some of these regressors) is denoted by M, where $M \subseteq P$. Given a criterion function f, the problem of selecting the best subset consists of solving the following equation:

$$M^* = \underset{M \subseteq K}{\operatorname{argmin}}\, f(M)$$

where $f(M) = F(\alpha, \theta)$ is a function that depends on $\alpha = |M|$ and $\theta = RSS(M)$, the number of regressors selected, and the RSS of the OLS estimator of M, respectively. Furthermore, $f(M)$ is assumed to be a monotonic function with respect to $RSS(M)$, i.e.,

$$RSS(M_1) \leq RSS(M_2) \Longrightarrow f(M_1) \leq f(M_1), \text{when } |M_1| = |M_2|$$

In this paper we will use the Akaike information criterion (AIC) because it exhibits precisely this property:

$$AIC_\gamma = M + M\, log\, 2\pi + M\, log(RSS/M) + \gamma(n + 1)$$

where the scalar γ represents a penalty for the inclusion of parameters in the regression which takes the value 2 following the econometric literature (Miller, 2002). Therefore, $M^* = M_p^*$, where $p = \underset{k}{\operatorname{argmin}}\, f(M_k^*)$ and $M_k^* = \underset{|M|=k}{\operatorname{argmin}}\, RSS(M)$ for $k = 1, \cdots, K$. Finding the solution M_k^* is called the "all subset selection problem". Therefore, the "best subset selection problem" can be solved through a two-step procedure. First, for each size k, we must find the subset $M_k^*(|M_k^*| = k)$ with the smallest RSS. Then, we must compute $f(M_k^*)$ for all k and determine p such that $f(M_p^*)$ is minimum.

To reach the optimal solution, Hofmann et al. (2020) resort to a variation of the descending column algorithm (DCA). Given that submodels can be organized in a regression tree in an efficient way, DCA analyzes the $2^{(K-1)}$ nodes that identify all potential combinations of variables. Each node of the regression tree can be represented by a pair of parameters (M, v), where the vector $M = \{m_1, \ldots, m_k\}$ identifies a subset of k variables, with $k = 0, \ldots, K$, and $v = 0, \ldots, K - 1$. The subleader models are defined as $\{m_1, \ldots, m_{v+1}\}, \ldots, \{m_1, \ldots, m_k\}$, whose RSS are computed for each visited node. The root node $(P, 0)$ identifies the complete model (the one that includes all control variables). Nodes descending from this parent arise after eliminating a single regressor:

$$drop(M, s) = (M \setminus \{m_s\}, s - 1)\,, where\ s = v + 1, \ldots, k - 1\,.$$

The DCA algorithm uses Givens rotations to move from one node to another in an efficient way; however, this procedure is computationally demanding. As a consequence, in order to reduce the number of nodes generated, Hofmann et al. (2020) implemented a branch and bound algorithm by eliminating subtrees that serve to find the best solution. A new node is only generated if $RSS(M) < r_s\ (s = v + 1, \ldots, k - 1)$. No regression of the subset extracted from the subtree can have a

smaller RSS. To further reduce the computational cost, these authors have reformulated the regression problem for all subsets as:

$$M_k^* = \underset{|M|=k}{\operatorname{argmin}} RSS(M) \ for\ k = k_{min}, \ldots, k_{max}$$

being $1 \leq k_{min} \leq k_{max} \leq K$. Specifically, the (M, s) nodes are not computed if $|M| < k_{min}$ or if $s \geq k_{max}$. Moreover, to further reduce the computational cost, the algorithm relaxes the cutoff test by employing a set of tolerance parameters $\xi_k \geq 0$ $(k = 1, \ldots, k)$. Therefore, a $drop(M, s)$ node is only generated if there exists at least one $z = v, \ldots, k-1$ such that:

$$(1 + \xi_z) \cdot (RSS(M) - RSS_{full}) < (r_x - RSS_{full})$$

where $RSS_{full} = RSS(P)$ is the residual sum of squares of the model using all control variables. In this paper, we will consider $\xi_z = 0$ to guarantee obtaining the best solution. Finally, to facilitate the computation of the optimal solution, we will generate a node only if $AIC(s, RSS(S)) < r_f$, being r_f the current best solution under AIC.

References

Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison-Wesley: New York.

Scrucca, L. (2013). GA: A package for genetic algorithms in R. *Journal of Statistical Software*, *53*, 1-37.

Wan, S. K., & Song, H. (2018). Forecasting turning points in tourism growth. *Annals of Tourism Research*, *72*, 156-167.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, *101*(476), 1418-1429.

Appendix B. Correlation matrix of the variables used in the study and Variance Inflation Factors (VIFs) of the models estimated in Table 1

Table B.1. Correlation matrix

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
1.GrossProfitLoss	1									
2.AssetsTotal	**0.764**	1								
3.Acquisitions	0.089	**0.253**	1							
4.DividendsTotal	**0.238**	0.113	-0.033	1						
5.Employees	**0.548**	**0.765**	0.138	0.072	1					
6.NetIncomeLoss	**0.551**	**0.490**	0.092	**0.344**	**0.449**	1				
7.OperatingActivitiesNetCash	**0.783**	**0.765**	0.068	**0.236**	**0.714**	**0.595**	1			
8.RetainedEarnings	0.067	**0.178**	0.113	0.127	0.100	**0.319**	**0.236**	1		
9.BookValuePerShare	0.037	**0.233**	**0.199**	**-0.207**	**0.206**	-0.038	-0.028	**0.172**	1	
10.SalesTurnoverNet	**0.663**	**0.790**	0.122	**0.166**	**0.811**	**0.607**	**0.845**	**0.408**	0.067	1

Note: Numbers in bold are statistically significant at 5%.

Table B.2. Variance Inflation Factors (VIFs) of the Estimated Models in Table 1

	Model 1	Model 2	Model 3 CVL	Model 4 AL	Model 5 BBA	Model 6 GA	Model 7	Model 8	Model 9 CVL	Model 10 AL	Model 11 BBA	Model 12 GA
Independent variable												
AssetsTotal		6.625	6.625	6.549	6.549	6.549		6.625	5.401	5.249	5.249	5.249
Control variables												
Acquisitions		1.429	1.429					1.429	1.411			
DividendsTotal		2.180	2.180					2.180	2.164	2.159		
Employees		5.708	5.708	5.701	5.701	5.701		5.708	5.629	5.628	5.621	5.621
NetIncomeLoss		3.091	3.091	3.054	3.054	3.054		3.091	3.026	3.011	3.001	3.001
OperatingActivitiesNetCash		5.241	5.241	5.170	5.170	5.170		5.241	5.118	5.030	5.018	5.018
Retained Earnings		2.749	2.749	2.557	2.557	2.557		2.749	2.640	2.489	2.469	2.469
BookValuePerShare		6.609	6.609	6.475	6.475	6.475		6.609				

Appendix C. Robustness Analysis of the Impact of Endogeneity

Given that the algorithms do not consider the selection of endogenous control variables, we have attempted to explore this issue to assess the extent to which bias may be introduced in our estimates.

Following the approach proposed by Shaver (1998), we can consider that managerial decisions are potentially endogenous because they are based on unobservable perceptions of sales or company performance. Otherwise, we would be assuming that decision-making processes are not strategic (i.e., that managers do not intend to influence firm outcomes with these choices). Considering this circumstance, the two regressors most likely to be potentially endogenous are Retained earnings and Number of employees (even though we are not strictly measuring decisions with either). To assess the extent to which our results might be affected by potential endogeneity, we have calculated Durbin-Wu-Hausman (DWH) tests. To perform this analysis, we used an instrumental variables approach in Models 4, 5, 11, and 12 of Table 1 (since these are the regressions where the goodness-of-fit parameters show the most improvement). As instrumental variables, we used lags of the potentially endogenous variables (Rutz et al., 2012) and a dummy variable that takes the value 1 if the auditor added explanations to their report (despite issuing an unqualified opinion regarding the financial statements of the hotel companies) and 0 otherwise. Semadeni et al. (2014) note that "weak and/or endogenous instruments yield suspect results whereas stronger, exogenous instruments reveal endogeneity" (p. 1077). Therefore, before analyzing the existence of endogeneity, we assessed the suitability of the selected instrumental variables.

In the models recommended by the branch and bound algorithm and the genetic algorithm, the Cragg-Donald Wald F statistic is 50.75 and 49.54 when we use Sales (Models 5 and 6 of Table 1) and Profits (Models 11 and 12 of Table 1) as the dependent variable, respectively. Since both figures are well above 13.43 (Stock & Yogo, 2005), we can conclude that the instruments are strong and have sufficient power to explain the potential endogenous variables. On the other hand, the Sargan statistic is not statistically significant in Models 5 and 6 ($\chi^2 = 1.179, p = 0.278$) and in Models 11 and 12 ($\chi^2 = 0.526, p = 0.468$) of Table 1. This result shows that the instruments are not correlated with the error term. Finally, in the regressions selected by the branch and bound algorithm and the genetic algorithm, the DWH test is not statistically significant when we use Sales ($\chi^2 = 1.557, p = 0.459$) and Profits ($\chi^2 = 2.119, p = 0.347$) as the dependent variable. This result suggests that the variables Retained earnings and Number of employees are exogenous, so it is unnecessary to resort to an instrumental variables approach to mitigate potential biases arising from endogeneity.

References

Rutz, O. J., Bucklin, R. E., & Sonnier, G. P. (2012). A latent instrumental variables approach to modeling keyword conversion in paid search advertising. *Journal of Marketing Research*, *49*(3), 306-319.

Semadeni, M., Withers, M. C., & Trevis Certo, S. (2014). The perils of endogeneity and instrumental variables in strategy research: Understanding through simulations. *Strategic Management Journal*, *35*(7), 1070-1079.

Shaver, J. M. (1998). Accounting for endogeneity when assessing strategy performance: Does entry mode choice affect FDI survival?. *Management science*, *44*(4), 571-585.

Stock, J. H., & Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In D. W. K. Andrews, & J. H. Stock (Eds.), Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg (pp. 80-108). Cambridge University Press.