Economics Bulletin

Volume 36, Issue 4

I just ran two trillion regressions

Christoph Hanck
Universität Duisburg-Essen

Abstract

The computational effort required to conduct a full model search to identify the most useful specification in problems that feature a large set of potential explanatory variables is widely perceived to be large. To circumvent or mitigate this challenge, the literature has proposed a host of techniques, many of which are not easy to implement. Using the example of a standard cross-country growth regression data set, we demonstrate that the computational effort in conducting a full model search will often be negligible. We provide an assessment of how this finding generalizes to model spaces of different sizes.

I gratefully acknowledge the support of DFG through project HA 6766/2-2.
Citation: Christoph Hanck, (2016) "I just ran two trillion regressions", *Economics Bulletin*, Volume 36, Issue 4, pages 2037-2042
Contact: Christoph Hanck - christoph.hanck@vwl.uni-due.de.
Submitted: April 05, 2016. **Published:** November 09, 2016.

1. Introduction

Finding determinants of economic growth of a cross-section of countries is, for obvious reasons, among the most relevant issues in economic research. Unsurprisingly, the topic has therefore attracted substantial attention. However, economic theories explaining growth are typically "open-ended" (Brock and Durlauf, 2001), implying that one set of determinants predicted by some theory does not rule out others. Hence, given that the list of potential determinants of growth is quite long, it may not come as a surprise that the literature does not naturally converge to a consensus.

Ultimately, the question therefore needs to be answered empirically. Now, with K potential determinants, there are 2^K possible linear models of the type

$$y = c + X_\ell \beta_\ell + \text{error}, \qquad \ell = 1, \ldots, 2^K$$

A natural and seemingly straightforward approach is to fit all possible 2^K models and pick the "best" one according to some suitable criterion. (Clearly, such a criterion needs to penalize complexity to avoid picking the largest possible, and presumably overfitted, model. Examples include the Bayesian Information Criterion or Mallows' C_p.) For instance, Fernandez, Ley and Steel (2001) construct, building on earlier work of Sala-i-Martin (1997), a famous dataset with $K = 41$ that has been widely used in many subsequent empirical and methodological studies (e.g., Hendry and Krolzig, 2004; Ley and Steel, 2009; Eicher, Papageorgiou and Raftery, 2011; Schneider and Wagner, 2012; Deckers and Hanck, 2014). With $K = 41$, there are thus roughly $2^{41} \approx 2 \cdot 10^{12}$, i.e., two trillion potential models. A full model search is then however widely perceived to be "unwieldy" (Moral-Benito, 2015) or even "prohibitive," see e.g. Hendry and Krolzig (2004, p. 803).

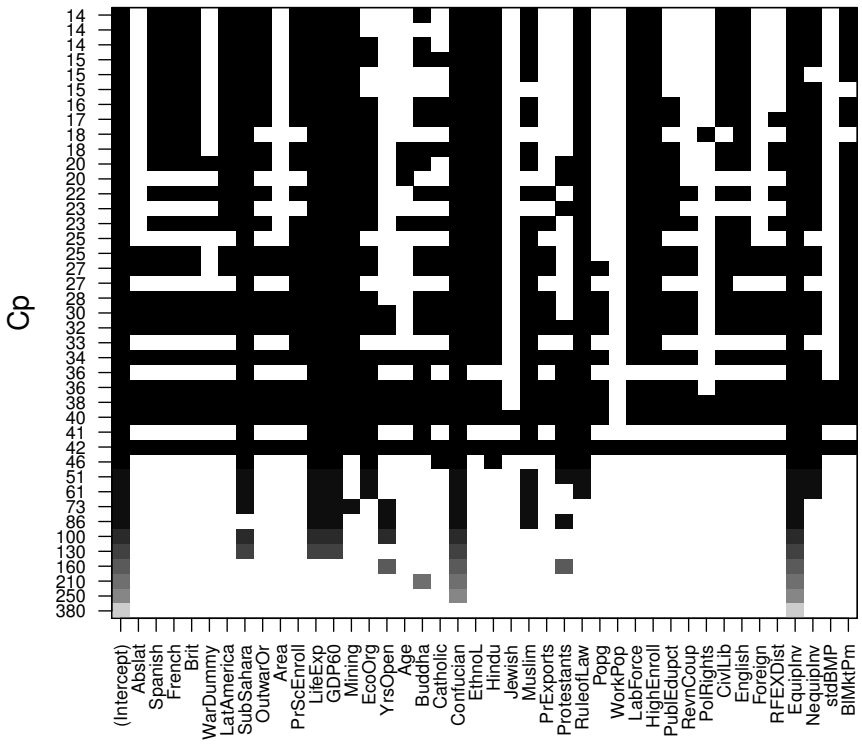
Econometricians have devoted much effort to circumvent the challenge of a full model search. Sala-i-Martin (1997) restricts himself to regressions including six predictors. An approach that has recently received increasing attention relies on Bayesian Model Averaging (BMA), which uses cleverly designed Markov Chain Monte Carlo Model Composition (MC^3) algorithms to at least explore the most compelling parts of the model space (e.g., Fernandez *et al.*, 2001). The "sheer number of models" (Schneider and Wagner, 2012) then however also implies that computing exact BMA estimates is infeasible, see also Fernandez *et al.* (2001).[1] Other examples include the PcGets approach of Hoover and Perez (1999) and Hendry and Krolzig (2004), the Lasso (Schneider and Wagner, 2012) or multiple testing techniques applied in Deckers and Hanck (2014). The statistical literature has proposed devices such as forward or backwise stepwise selection (e.g., Hastie, Tibshirani and Friedman, 2009).

All these approaches have in common that not visiting the entire model space implies the risk of not considering a model with better performance than the ones that are considered.[2]

This note is to demonstrate that for the type of problems considered in Fernandez *et al.* (2001), the arguably natural approach of a full model search—known as full subset selection in the statistical literature (e.g., Hastie *et al.*, 2009)—is a negligible task both computationally and in terms of user input.

[1] Recent advances in the model averaging literature include Magnus, Powell and Prüfer (2010) or Magnus and Wang (2014). See Moral-Benito (2015) for a survey.

[2] That said, there are fairly tight bounds on this risk in terms of, e.g., parsimony and prediction of for example forward regression at least when predictors do not exhibit strong correlation (see, e.g., Das and Kempe, 2008). A full discussion of this issue however is outside the scope of this note.



The figure shows, for each model size k, the included variables of the best model of that size along with its C_p criterion value. The models are sorted from best (above) to worst (below).

Figure 1: Best models for different model sizes k

2. Full subset regression

Full subset regression is conceptually straightforward. For a given model size $k \in \{1, \ldots, K\}$, fit all possible $\binom{K}{k}$ models. Of these, choose the one with the lowest sum of squared residuals. As all these models have k parameters, none has an unfair advantage over the others using this criterion. Of the resulting set of optimal models of a given dimension, $\{\mathcal{M}_k^\star, k = 1, \ldots, K\}$, choose the one with the smallest value of some information criterion such as Mallows' C_p. Notice that no hypothesis testing is involved, so that misspecification of smaller models is not an issue.

Using the `R` (R Core Team, 2014) package `leaps` (Lumley, 2009), we perform best subset selection on the Fernandez *et al.* (2001) data ($n = 72$), provided in for instance the BMS package (see Feldkircher and Zeugner, 2009).[3] More specifically, we invoke `{regsubsets(y~., data=datafls,nvmax=41)`. This task took a little more than three minutes on a standard desktop PC (Intel i7-3770 3.40GHz CPU) without any attempts to speed up computation by, say, parallelization. Calling the resulting object `regfit.full` yields Figure 1 on calling

[3]While `leaps` is guaranteed to find the best model, it does not literally run 2^K OLS regressions. The underlying Fortran code (by Allan Miller) employs a branch-and-bound algorithm (Furnival and Wilson, 1974). The algorithm avoids visiting parts of the model space which cannot contain the optimum given the results of models fitted earlier in the search, effectively exploiting that dropping variables from a specification will never increase the R^2. To give an indication of the efficiency of the branch and bound method, fitting the full model with $K = 41$ using `lm` requires 0.02 seconds. Hence, fewer than $10.000 = 10^4$ ($\approx 187/0.02$) such regressions could be fit until the algorithm has optimized over the entire model space of approximate dimension $2 \cdot 10^{12}$. For details, see e.g. Hand (1981).

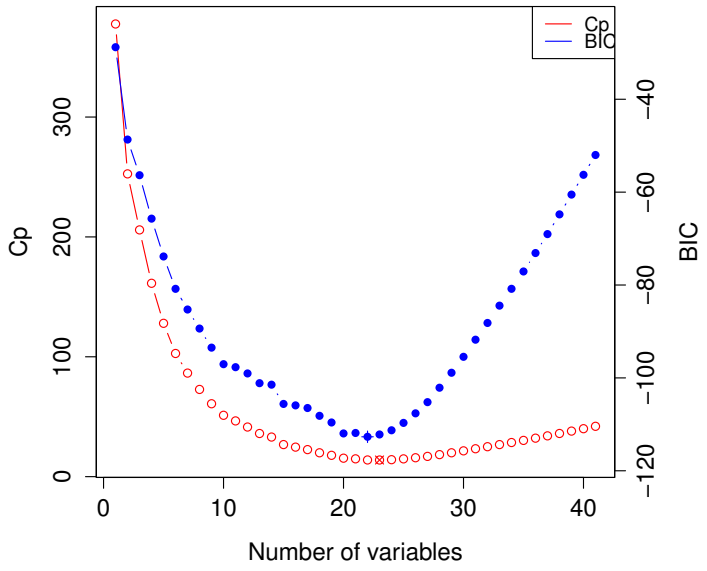


Figure 2: Information criteria as a function of model size k

`plot(regfit.full)`.

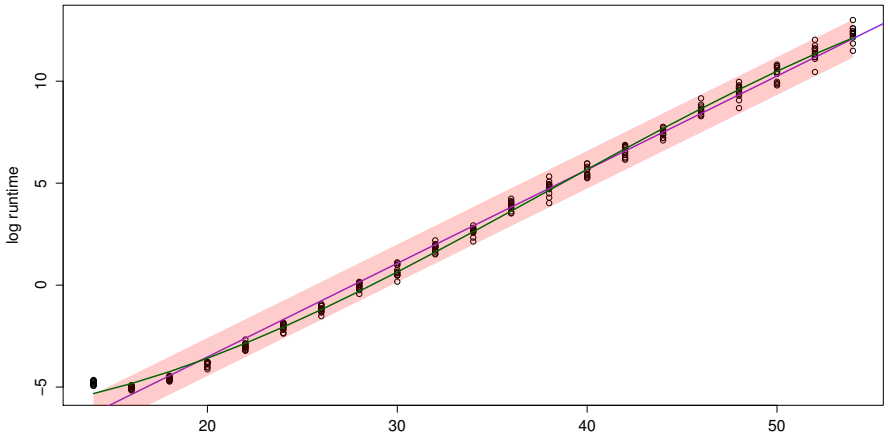
The best-performing model then is the one given in the first row, i.e., the one with the lowest information criterion C_p. It includes 23 predictors, with fitted model

$$\begin{aligned}\hat{y} \;=\; & 0.0733 + 0.0134 Spanish + 0.0104 French + 0.0074 Brit - 0.0116 LatAmerica \\ & - 0.02 SubSahara - 0.0039 OutwarOr + 0.0242 PrScEnroll + 0.0009 LifeExp \\ & - 0.0178 GDP60 + 0.0344 Mining + 0.0064 Buddha + 0.0767 Confucian \\ & + 0.0160 EthnoL - 0.1065 Hindu + 0.0089 Muslim + 0.0123 RuleofLaw \\ & + 3.68\text{e-}7 LabForce - 0.1177 HighEnroll - 0.0028 CivlLib \\ & - 0.0072 English + 0.1475 EquipInv + 0.0294 NequipInv - 0.0057 BlMktPm\end{aligned}$$

We do however notice that there are several models whose C_p only differs slightly from that of the best one and hence perform fairly similarly, see also Figure 2. Figure 2 also highlights the familiar bias-variance tradeoff, in that highly parameterized models fit better, but are more variable. Reassuringly, these top models also only differ moderately in terms of the variables selected, indicating a certain robustness.

Figure 2 shows that the BIC favors a model of very similar but, as expected, slightly smaller size with 22 explanatory variables.

These results are broadly in line with those of other model selection approaches on the Fernandez *et al.* (2001) data. In particular, variables with high posterior inclusion probability in Fernandez *et al.* (2001) or small adjusted p-values in Deckers and Hanck (2014), such as initial GDP, life expectancy or the fraction of Confucians, are also included by best subset selection. See Deckers and Hanck (2014, Table 6) for a more complete comparison. With 23 or 22 included variables, best subset selection is within the range of the number of selected variables by other model selection procedures, but at the higher end (cf. e.g. Eicher *et al.*, 2011, Table II).



Magenta: linear fit (with prediction interval). Green: Cubic fit.

Figure 3: Log-runtimes as a function of k_j

3. Some simulations

Of course, the conclusions that may be drawn from the above example are specific to data sets with comparable k. As the size of the model space doubles with each additional regressor, one needs to expect computation time to grow exponentially in k.[4] To shed some light on how the above findings generalize to models with other k, we record computation times for samples drawn from a linear model $y = X_{k_j}\beta_{k_j} + u$ with a k_j-dimensional zero-mean multivariate normal regressor matrix X_{k_j} with covariance matrix $\Sigma'\Sigma$ where the entries of Σ are $N(0,1)$, just as those of u. The first $k_j/2$ elements of β_{k_j} are zero, representing irrelevant regressors, and the remaining entries are drawn from a uniform distribution on $[5,10]$. We use a sample size of $n = 100$, take $k_j \in \{14, 16, \ldots, 52, 54\}$ and draw $M = 12$ samples for each k_j.[5]

Figure 3 summarizes the results, plotting log-runtimes against k_j. The excellent linear fit of the regression ($R^2 = 0.993$) confirms that computation time generally grows exponentially in k_j. The slope of the regression line (0.46) however reveals that the branch-and-bound algorithm is capable of chopping off increasingly large parts of the model space with growing k_j, as computation time does not double when k_j increases by one, but only by $100 \cdot [\exp(0.46) - 1] \approx 58\%$. The cubic fit provides tentative evidence that the increase in computation time might even flatten out for very large k_j, but the necessary computation time for such k_j makes this assessment speculative at today's computational speeds.

4. Concluding remarks

This note shows that conducting a full model search is computationally feasible for a much larger class of models than what appears to be commonly thought. Of course, we do not wish to argue that clever approaches to model selection are not worth considering. For instance,

[4] For example, had we felt the necessity to include a full set of interaction terms of the $k = 41$ regressors in Section 2, we would have been left with $41 \cdot 40/2$ additional variables, prohibitively increasing the computational burden. There is hence a tradeoff between the desired flexibility of the model and the possibility to perform a full model search.

[5] Experimenting with other conventional choices of n and β_j had a minor impact on runtime. Moderately larger sample sizes do not substantially increase computation time because the computationally costly operation is to compute $(X'X)^{-1}$, which becomes more burdensome as k increases.

full subset regression increases the likelihood of selecting a spurious model with poor out-of-sample explanatory performance, while forward selection (Hastie *et al.*, 2009) only requires comparing models along the selection path.[6] Similarly, genome-association studies routinely (and occasionally also some variable selection problems in economics) face situations in which $K \gg n$, the number of observations, calling for, e.g., sparse approaches such as the Lasso (Tibshirani, 1996). Also, in many cases, the analyst may not only be interested in a single specification best describing the data (i.e., "model search"), but rather in, say, prediction. (Indeed, many popular "ensemble" machine learning algorithms like random forests, boosting or bagging (e.g., Hastie *et al.*, 2009) trade the neat, possibly structural, interpretability of a single final specification with instead a predictive performance that is often superior.) In such a situation, the averaged predictions of (frequentist or Bayesian) model averaging may offer distinct advantages, as it is widely documented that the average forecast of different models often outperforms that of a single model.[7]

References

Brock WA, Durlauf SN. 2001. Growth empirics and reality. *The World Bank Economic Review* **15**: 229–272.

Das A, Kempe D. 2008. Algorithms for subset selection in linear regression. In *Proceedings of the Fortieth Annual ACM Symposium on Theory of Computing*, STOC '08, New York, NY, USA: ACM, pages 45–54.

Deckers T, Hanck C. 2014. Variable selection in cross-section regressions: Comparisons and extensions. *Oxford Bulletin of Economics and Statistics* **76**: 841–873.

Eicher TS, Papageorgiou C, Raftery AE. 2011. Default priors and predictive performance in Bayesian model averaging, with application to growth determinants. *Journal of Applied Econometrics* **26**: 30–55.

Feldkircher M, Zeugner S. 2009. Benchmark priors revisited: On adaptive shrinkage and the supermodel effect in Bayesian model averaging. *IMF Working Paper* **09**: 1–39.

Fernandez C, Ley E, Steel MF. 2001. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics* **16**: 563–576.

Furnival GM, Wilson RW. 1974. Regressions by leaps and bounds. *Technometrics* **16**: 499–511.

Hand DJ. 1981. Branch and bound in statistical data analysis. *The Statistician* **30**: 1–13.

Hastie T, Tibshirani R, Friedman J. 2009. *The Elements of Statistical Learning*. Springer, 2nd edn.

Hendry DF, Krolzig HM. 2004. We ran one regression. *Oxford Bulletin of Economics and Statistics* **66**: 799–810.

Hoover KD, Perez SJ. 1999. Data mining reconsidered: Encompassing and the general-to-specific approach to specification search. *Econometrics Journal* **2**: 167–191.

Leeb H, Pötscher B. 2005. Model selection and inference: Facts and fiction. *Econometric Theory* **21**: 21–59.

Ley E, Steel MF. 2009. On the effect of prior assumptions in Bayesian model averaging with applications to growth regression. *Journal of Applied Econometrics* **24**: 651–674.

[6] Recent research begins to more systematically assess the statistical tradeoffs implied by using approximate solutions to computationally difficult estimation problems. This literature demonstrates that local optima identified by approximate solutions are generally statistically "well-behaved" (see, e.g., Loh and Wainwright, 2015, for details).

[7] That said, if either the best model from a full model search of any of the above averaging methods are to be used for inferential purposes, issues such as post model selection distortions (e.g., Leeb and Pötscher, 2005) arise, requiring careful consideration and adjustments. A detailed discussion of this aspect is, however, beyond the scope of this note.

Loh PL, Wainwright MJ. 2015. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research* **16**: 559–616.

Lumley T. 2009. *Leaps: Regression Subset Selection*. R package version 2.9.

Magnus JR, Powell O, Prüfer P. 2010. A comparison of two model averaging techniques with an application to growth empirics. *Journal of Econometrics* **154**: 139–153.

Magnus JR, Wang W. 2014. Concept-based bayesian model averaging and growth empirics. *Oxford Bulletin of Economics and Statistics* **76**: 874–897.

Moral-Benito E. 2015. Model averaging in economics: An overview. *Journal of Economic Surveys* **29**: 46–75.

R Core Team. 2014. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sala-i-Martin XX. 1997. I just ran two million regressions. *American Economic Review* **87**: 178–183.

Schneider U, Wagner M. 2012. Catching growth determinants with the adaptive lasso. *German Economic Review* **13**: 71–85.

Tibshirani R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B* **58**: 267–288.