Machine Learning and Big data in econometrics: a Machine Learning based specification test

Gilles HACHEME
Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS
Marseille, France
gilles.hacheme@univ-amu.fr

1. Introduction

- Machine Learning (ML) and Big data become ubiquitous in many scientific fields. But their contribution to social sciences is not yet very clear.
- ML methods has shown their relevance in extracting information from massive data.
- Can ML techniques revolutionize Econometrics?
- We suggest some way ML can be used in Econometrics to test model functional specification.
- We suggest a new specification test named BootML: A bootstrapped test using a ML model (Random Forest (RF) in this Paper).
- BootML is computationally lighter and suffer less from the curse of dimensionality compared to kernel based tests as a result of ML models properties such as the ones of RF.
- We showed through simulations that BootML (using RF) is more powerful and has lower type I error than the parametric Regression Error Specification Test (RESET).
- BootML is then suitable for large datasets where it is very difficult to use kernel-based tests.

2. Specification tests

- Model specification in econometrics historically involves statistics methods and economic modelling.
- Historical model specification approaches are based on vanilla parametric statistical methods such as linear regression or structural models. Suitable for small sample data.
- Structural models are subject to misspecification (MS) (Sims, 1980).
- Some types of MS: Serial correlation (Durbin Watson test) , Heteroskedasticity (Breusch, Godfrey test), multicolinearity, endogeneity and incorrect functional form [6, 2].
- We focus on *incorrect functional form* as defined by [6].

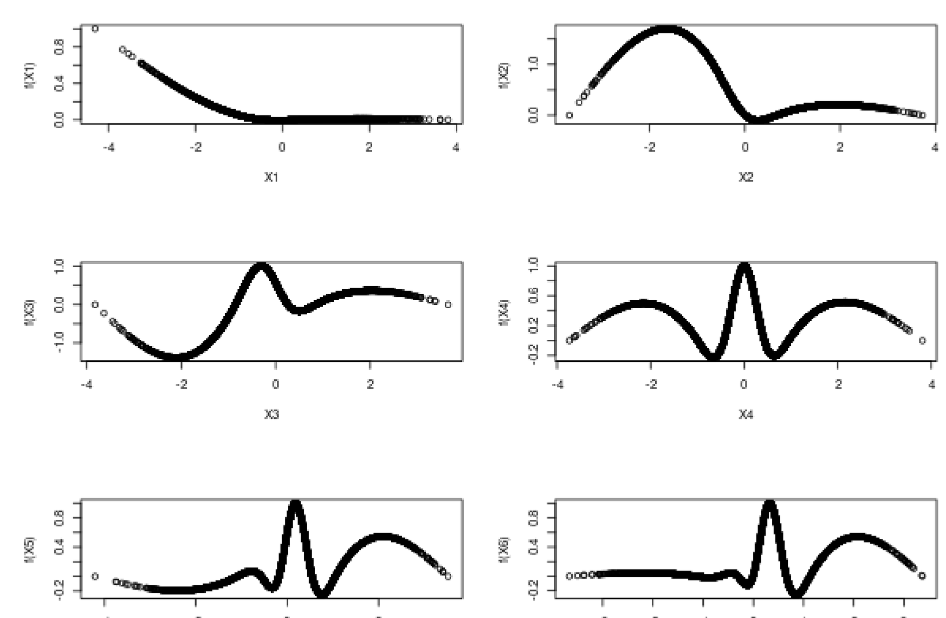
Let y a dependent variable and X a set of k explanatory variables such that: $y = m(X) + \epsilon$, where ϵ is the random error term and m is an unknown function.
The specification test is the following:

$$\begin{cases} H_0 : m(X) = f(X, \beta) \\ H_1 : m(X) \neq f(X, \beta) \end{cases},$$

where f is a known parametric or semi-parametric function and β a vector of parameters. If m is linear under H_0, $m(X) = X\beta$.

- Regression Error Specification Test (RESET): m is a parametric non-linear function (polynomial function of order 3). A Fisher test is done to compare the linear and the non-linear models.
- Problem: if the true non-linearity is far from this function or if there are strong interactions, then it might increases type II error resulting in a less powerful test.
- Solution: Non parametric model specification tests make the promise of more flexible models leading to more consistent tests.
- Robinson [7] highlights that non-parametric methods allow better efficiency than misspecified parametric models.
- Non-parametric functions for m: generally based on kernel regressions.
- But the biggest downside of kernel-based specification tests are by definition the limits of kernel regressions.
  - Curse of dimensionality [3].
  - Another problem comes from hyperparameters estimation such as the bandwidth which is computationally heavy as mentioned by [5]
- Then kernel-based specification tests are in practice limited to very small datasets whereas in almost every single discipline scholars are increasingly working with larger and larger datasets
- To solve issues raised by kernel methods, we introduce a Machine Learning based approach. Here we use RF [1] to replace the kernel regression.
- Indeed, RF is well known to less suffer from the curse of dimensionality while it embeds most of the advantages we can derive from classical non-parametric methods such as kernel regression and even beyond.
- RF is capable to filter out irrelevant variables: acting as a variable selection method. RF also handles well potential non-linearities and interactions.

3. Methodology



Smooth variables from a cubic spline basis

Our method uses a ML model to estimate the m function and uses bootstrapping to test if the ML model's out-of-sample Mean Squared Error (MSE) is significantly different from the one of a candidate parametric model. H_0 is rejected only if the difference is significant and if the ML's MSE is lower than the one of the candidate model (see 4.

$$\begin{cases} H_0 : MSE^{ml}_{test} \geq MSE^{c}_{test} \\ H_1 : MSE^{ml}_{test} < MSE^{c}_{test} \end{cases},$$

where MSE^{ml}_{test} is the MSE on test set by the ML model, and MSE^{c}_{test} is the MSE on test set by the candidate model.

Simulation results for rejection frequencies

	k = 4					
	n=600		n=6.000		n=12.000 "	
	reset	boot.ml	reset	boot.ml	reset	boot.ml
Linear	0.010	0	0.070	0	0.040	0
Non Linear	0.330	0	0.110	1	1	1
Non Linear + Interactions	1	1	1	1	1	1
	k = 6					
	reset	boot.ml	reset	boot.ml	reset	boot.ml
Linear	0.070	0	0.020	0	0.050	0
Non Linear	0.030	0	0.030	0.460	0.960	1
Non Linear + Interactions	1	1	1	1	1	1

4. Simulation results

In our case, we used RF as our ML model and the default candidate is a linear model.
Here are the DGP used in our simulations:

1. Linear DGP: $y = \sum_{j=1}^{k} \alpha_j X_j + \epsilon$, where $X \sim \mathcal{N}(0, \mathbf{1}_k)$ $\epsilon \sim \mathcal{N}(0,1)$ is an iid random error term, and $\alpha_j \sim \mathcal{U}(0,1)$,
2. Non-linear DGP: $y = \sum_{j=1}^{k} \alpha_j X_j + \sum_{j=1}^{k} \beta_j f_j(X_j) + \epsilon$, where and α_j and $\beta_j \sim \mathcal{U}(0,1)$,
   We generate non-linearities using a cubic spline basis where the wiggliness is controlled by the number of knots n_k [8, 4, 9]. The cubic spline basis creates n_k smooth terms for each variable,
   To generate a single smooth term for each variable with different wigglinesses, we proceed as follows:
   - for each X_j we generate cubic spline basis smooth terms R with $j+3$ knots $\forall j = 1, \ldots, k$,
   - but we only keep the j^{th} element: $f_j(x_j) = R(x_j, x_j^*)$, where x_j^* is the j^{th} knot of X_j.
3. Non-linear + interactions DGP:
   $$y = \sum_{j=1}^{k} \alpha_j X_j + \sum_{j=1}^{k} \beta_j f(X_j) + \sum_{j=1}^{p} Z_j + \epsilon,$$
   where Z are linear interactions $X_j X_l$ where $j \neq l$, and we only include the first half interactions in the DGP. So, the number of included interactions $p = k(k-1)/4$.

Algorithm 1 BootML test

1-Estimation using the original training set:
Candidate model : $y = f(X, \beta) + \epsilon$
ML : $y = m(X) + \epsilon$

2-Computing the MSE on the original test set:
Candidate model: $MSE^{c}_{test} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}^{c}_i)^2$
ML: $MSE^{ml}_{test} = \frac{1}{n_{test}} \sum_{i=1}^{n_{test}} (y_i - \hat{y}^{ml}_i)^2$
3-Computing the difference between ML's MSE and Candidate model's MSE: $\hat{\Delta} = MSE^{ml}_{test} - MSE^{c}_{test}$
4-Estimating the mean and variance of $\hat{\Delta}$ using bootstrapping
for $i \leftarrow 1$ **to** $1,000$ **by** 1 **do**
  Resampling with replacement using same size as the test sample to obtain: $\tilde{y}, \hat{\tilde{y}}^{c}, \hat{\tilde{y}}^{ml}$
  Run step 2 using $\tilde{y}, \hat{\tilde{y}}^{c}, \hat{\tilde{y}}^{ml}$ to obtain $\hat{\Delta}_i$
end

Mean: $\Delta = \frac{1}{1,000} \sum_{i=1}^{1,000} \hat{\Delta}_i$

Variance: $V = \frac{1}{1,000} \sum_{i=1}^{1,000} (\hat{\Delta}_i - \Delta)^2$

Test statistic: t = $(\hat{\Delta} - \Delta)/\sqrt{V}$
Asymptotically t should follow a standard normal distribution under the null hypothesis: $t \sim \mathcal{N}(0,1)$
5- Test at threshold error α:
if $|t| > \Phi^{-1}(1 - \alpha/2)$ **then**
  if $t > 0$ **then**
    H_0 is not rejected
  end
  else
    H_0 is rejected
  end
end
else
  H_0 is not rejected
end
NB: Φ^{-1} is the quantile function of the normal distribution

---

We made simulations for different values of $n = 600, 6,000, 12,000$. The RF is trained on two third (2/3) of the data. And BootML specification test is made on the remaining data to avoid overfitting. But the RESET is done as usual on the training set.
Table 1 presents results from simulations:

- $k = 4$: The RESET's power is better for small n (400). But for higher n (4,000 and 8,000), BootML have a perfect power while it does not make any type I error. For $n = 8,000$, RESET also has a perfect power, but has a 4% type II error rate.
- $k = 6$: RESET seems to be less powerful than when $k = 4$ and its type I error is higher except for $n = 4,000$. BootML's power is also lower but higher than RESET's power for $n = 6.000$ and $n = 12,000$.

5. Conclusion

- Non-parametric functional form specification tests are known to be more consistent than parametric ones.
- Kernel based specification tests are the most used non-parametric tests, but they are limited: kernel regressions are computationally heavy and suffer from the curse of dimensionality.
- Solution BootML: using instead a ML model fast enough and less subject to the curse of dimensionality such as RF.
- Simulation results showed that BootML (using RF) is more powerful and less subject to type I error compared to RESET when the number of observations is large enough.
- BootML is then more suitable for Big data: faster and less subject to the curse of dimensionality than kernel based methods.
- Next steps: How do we use a ML model to improve a parametric model to get closer to an unkown DGP?

EcoSta 2021, 4th International Conference on Econometrics and Statistics, 24-26 June 2021, Virtual Conference, HKUST Hong Kong

References

[1] BREIMAN, L. Random forests. *Machine learning 45*, 1 (2001), 5–32.

[2] DAVIDSON, R., AND ZINDE-WALSH, V. Advances in specification testing. *Canadian Journal of Economics/Revue canadienne d'économique 50*, 5 (2017), 1595–1631.

[3] GEENENS, G. Curse of dimensionality and related issues in nonparametric functional regression. *Statistics Surveys 5*, none (2011), 30 – 43.

[4] GU, C. *Smoothing spline ANOVA models*, vol. 297. Springer Science & Business Media, 2013.

[5] HAYFIELD, T., AND RACINE, J. S. Nonparametric econometrics: The np package. *Journal of statistical software 27*, 5 (2008), 1–32.

[6] RAMSEY, J. B. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society: Series B (Methodological) 31*, 2 (1969), 350–371.

[7] ROBINSON, P. M. Hypothesis testing in semiparametric and nonparametric models for econometric time series. *The Review of Economic Studies 56*, 4 (1989), 511–534.

[8] WAHBA, G. *Spline models for observational data*. SIAM, 1990.

[9] WOOD, S. N. *Generalized additive models: an introduction with R*. CRC press, 2017.