Model Ambiguity in `R`: The `sValues` Package

Carlos Cinelli *

October 25, 2015

1 The problem: *ad-hoc* specification searches

A researcher is studying economic growth and is specifically interested in the role of Government (Nominal) GDP Share. After trying some preliminary models, he comes up with a "good", "parsimonious" specification with 10 control variables. The coefficient is negative, "significant" and it even resists some "robustness" checks. How reliable is this finding? Actually, not much. But this practice is quite common.

Researchers usually engage in *ad-hoc* specification searches but present only their favorite models. This, however, can easily underestimate the uncertainty caused by model selection and lead to overconfident inferences. Since we are dealing with nonexperimental data, the set of controls can be virtually unlimited and the theory ambiguous about which ones do matter. In this example, it turns out that one can come up with a different set of 10 controls in which the coefficient for Government GDP Share is positive and "significant". In fact, there are 67 possible control variables, which could generate 148 *quintillion* different models!

So how can we tackle that problem? This presentation will introduce the `R` package `sValues`, which implements a measure of sturdiness of coefficients proposed by Leamer[4] and discussed in Leamer[3]. The S-values try to provide a sensible framework to assess the sensitivity of coefficient estimates to model ambiguity. But before going to the `R` implementation, let's see a *brief* description of the method.

*This vignette is a draft based on a poster presented on useR! 2015. I've learned a great deal from discussions with Ed Leamer! I also thank Rasmus Baath, Danilo Freire and Douglas Araujo for their comments. Of course, all remaining errors are my own. And all opinions expressed in this material are mine and do not necessarily reflect the views of the CBB. Contact: carloscinelli@hotmail.com

2 S-Values: measures of the sturdiness of regression coefficients

2.1 Extreme bounds for the coefficients

The different estimates for Government GDP Share can be interpreted as the result of *different strong prior beliefs*: the coefficients of the *ommited* variables are *exactly zero*, whereas, for the ones *included* in the model, *we believe whatever the data says*. This suggests that a Bayesian approach could be useful to model this problem. Consider the linear model $y = X\beta + \epsilon$ where $\epsilon \sim N(0, \sigma^2 I)$ and $\beta \sim N(0, V)$. The OLS estimate of β is $b = (X'X)^{-1}X'y$ with precision matrix $H = (X'X)/\sigma^2$. Then, the posterior mean of β is:

$$\hat{\beta}(V) = (H + V^{-1})^{-1}Hb \tag{1}$$

Notice that within this framework we can express model specifications in terms of *beliefs about the prior variance* V. For example, regressions with subsets of explanatory variables are akin to saying that the diagonal of V is really really large (infinite) for some of them and really really small (zero) for others. If we knew V exactly (or had a distribution for V), then we would just have an estimation problem. But, if V is ambiguous or disputable, then we might want to know how sensitive $\hat{\beta}$ is to "sensible" variations in the prior variance.

To come up with a set for possible Vs, we might want to bound it from below, excluding dogmatic priors of zero variances which would lead to inferences unaffected by data. We might also want to bound it from above, preventing the data to speak freely and limiting the influence of unimportant variables. So, given that V is bounded from above and from below, $V_* \leq V \leq V^*$, Leamer[2] shows that $\hat{\beta}$ lies in the ellipsoid:

$$(\hat{\beta} - f)G(\hat{\beta} - f) \leq c \tag{2}$$

Where:

$$G = (H + V^{*-1})(V_*^{-1} - V^{*-1})^{-1}(H + V^{*-1}) + (H + V^{*-1})$$

$$f = (H + V_*^{-1})^{-1}[Hb + (V_*^{-1} - V^{*-1})(H + V^{*-1})^{-1}Hb/2]$$

$$c = b'H(H + V^{*-1})^{-1}(V_*^{-1} - V^{*-1})(H + V^{*-1})^{-1}Hb/4$$

Therefore the extreme bounds for a linear combination $\psi'\hat{\beta}(V)$ are given by:

$$\psi f \pm (\psi' G^{-1}\psi)^{\frac{1}{2}} c^{\frac{1}{2}} \tag{3}$$

And our measure of the sturdiness of a coefficient, the S-value, can be defined as:

$$s = \frac{\psi f}{(\psi' G^{-1}\psi)^{\frac{1}{2}} c^{\frac{1}{2}}} \tag{4}$$

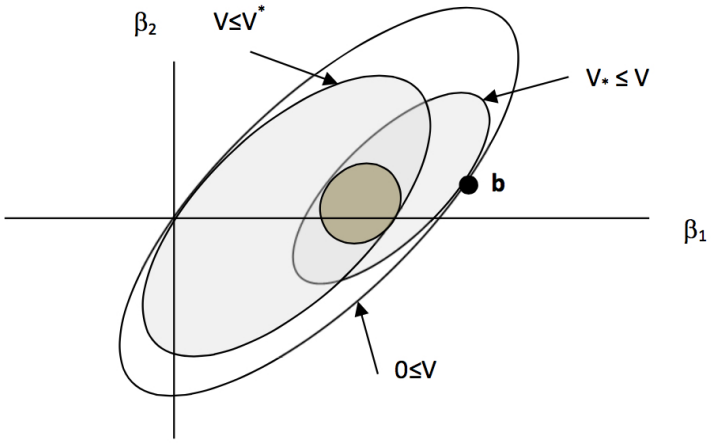

Figure 1: Ellipses of estimates. Source: Leamer[4]

When the S-value is less than 1 in absolute value this means that the coefficient is not *sturdy* - that is, it changes sign when V varies within the upper and lower bounds. Figure 1 illustrates these ideas.

2.2 Conventional bounds for the prior variances

The problem now is how to find specific numerical bounds for V. Choosing bounds for the variances of the coefficients can be a challenging task. So, instead, Leamer[4] suggests that we focus on the expected R^2 of the model (which, probably, most people would find easier). After standardizing the variables and considering bounds proportional to the identity matrix, the prior variance v^2 of each beta-coefficient equals to the expected R^2 divided by the number of parameters k of the model, that is, $v^2 = E(R^2)/k$.

This would give us the bounds:

$$v_{low}^2 = \frac{E(R^2)_{low}}{k} \leq v^2 \leq \frac{E(R^2)_{up}}{k} = v_{up}^2 \tag{5}$$

As for the ranges of expected R^2, Leamer[4] proposes three choices: (i) a context-minimal range $[0.1, 1.0]$; (ii) a pessimistic range $[0.1, 0.5]$; and, (iii) an optimistic range $[0.5, 1.0]$. We can generalize this to allow sets of "favorite" variables. Note that whereas the bounds of V are diagonal, we are allowing non-diagonal priors.

3 The sValues R package: A Growth Regressions Example

The sValues package comes with an example dataset on economic growth used by various papers (FSL[1], SDM[5] and Leamer[4]). This dataset comprises the growth of real GDP per capita from 1960 to 1996 and other 67 explanatory variables from 87 countries.

The main function of the package is the sValues function. The standard approach is to provide a formula specifying the model, a data.frame with the data and a numerical vector with the R^2 bounds (default values are 0.1, 0.5, and 1). As a shortcut, you can omit the formula and the function will automatically consider the first column as the dependent variable and the rest as the independent variables. Let's run the analysis for the economic growth data.

```
> library(sValues) # loads package
> data("economic_growth") # loads data
> eg <- sValues(economic_growth) # runs analysis
> eg # prints basic results

Data: economic_growth,    Formula: GR6096 ~ .
R2 bounds: 0.1 - 0.5 - 1

abs(S-value) > 1:
  R2 (0.1, 1): None
  R2 (0.1, 0.5): None
  R2 (0.5, 1): BUDDHA CONFUC EAST IPRICE1 P60 RERD

abs(t-value) > 2:
  Bayesian (R2 = 0.1): EAST
  Bayesian (R2 = 0.5): IPRICE1
  Bayesian (R2 = 1): IPRICE1
  Unconstrained OLS: IPRICE1
```

As we can see from the results, only in the "optimistic" scenario some variables are robust to model ambiguity. Moreover, if we look at the sample uncertainty (t-values), there is only one variable (IPRICE1) which has both $|s| > 1$ and $|t| > 2$. This means that any precise inferences about the sign of almost all the coefficients *require stronger prior information about preference for some variables*. It is worth mentioning that these results are in contrast with those obtained by using the BMA methodologies proposed by FSL[1] and SDM[5], which can also be implemented in R using the BMS package (see Zeugner[6]).

You can access specific coefficient values with the coef function setting the argument type to the desired statistic (for example, type = "s_values", type = "t_values" or type = "extreme_bounds"). For each type, there is also a wrapper function with the same name, so the command coef(x, type = "s_values") is equivalent to s_values(x).

```
> # gets a complete table like in Leamer[3]
> full_table <- coef(eg)
> full_table[1:5, 1:5] # showing only first five columns and rows

        ols_simple b_bayes_0.1 b_bayes_0.5  b_bayes_1    ols_all
IPRICE1 -0.4443075 -0.06260783  -0.1277645 -0.1596315 -0.3827615
CONFUC   0.4735595  0.06838785   0.1149072  0.1311292  0.2761033
EAST     0.5303796  0.07882938   0.1308950  0.1485540  0.1279553
BUDDHA   0.4447188  0.06251667   0.1004677  0.1106633  0.3369219
P60      0.5742210  0.05701799   0.1155436  0.1488612  0.4574343

> # gets just the s_values
> just_svalues <- coef(eg, type = "s_values")
> just_svalues[1:5, ] # showing only first five rows

           s_R2_0.1_1 s_R2_0.1_0.5  s_R2_0.5_1
ABSLATIT  0.036236556  0.073865639  0.10136981
AIRDIST   0.002881237 -0.004704409  0.07160259
AVELF    -0.135043178 -0.196852073 -0.50902679
BRIT      0.140973114  0.175892738  0.59795954
BUDDHA    0.309053323  0.440686400  1.21091252
```

Let's print the extreme bounds in the optimistic case of two specific coefficients: `GOVNOM1` and `IPRICE1`.

```
> extreme_bounds(eg)[c("GOVNOM1", "IPRICE1"),
+                    c("R2_0.5_1.low", "R2_0.5_1.up")]

        R2_0.5_1.low R2_0.5_1.up
GOVNOM1   -0.1566860  0.03367756
IPRICE1   -0.2167692 -0.07062673
```

As we had seen, `IPRICE1` is robust to different prior variances, with extreme bounds of -0.22 and -0.07. On the other hand, `GOVNOM1` may change its sign with different specifications varying from -0.16 to 0.03.

The package comes with some plot methods to explore the results. Let's plot the t-statistics versus the s-values per coefficient, highlighting the uncertain and fragile estimates, as shown in Figure 2.

```
> plot(eg, type = "t_s_plot", R2_bounds = c(0.5, 1))
```

Also, let's investigate how the coefficient for Government GDP Share varies with different prior R^2 as shown in Figure 3 (the Bayesian estimates consider a diagonal V with the corresponding v^2 specified before and can be thought of weighted averages of the 2^k regressions [4]).

```
> plot(eg, type = "beta_plot", variables = "GOVNOM1",
+      error_bar = TRUE, ext_bounds_shades = TRUE)
```

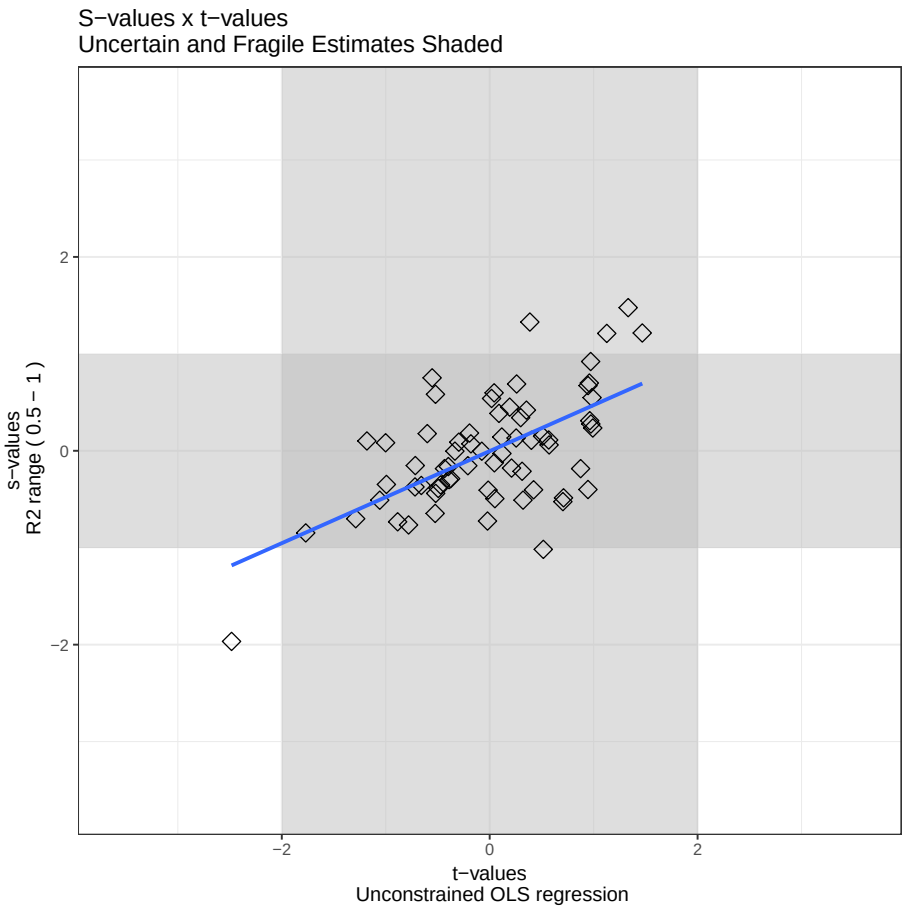

Figure 2: t-statistics vs s-values

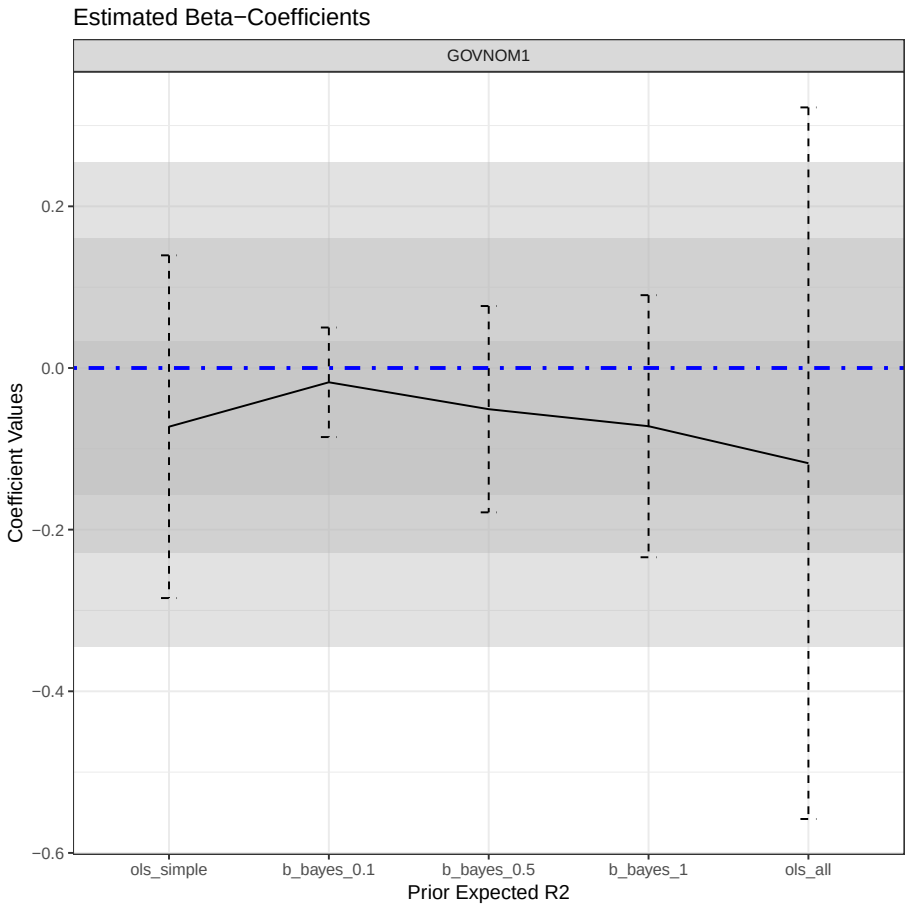

Figure 3: Bayesian estimates for GOVNOM1, with error bars and extreme bounds (shaded areas).

The `sValues` function allows you to define some of the variables as "favorites" with larger prior variances. In that case, you need to specify a `favorites` paramater - with the names of the favorite variables - and a `R2_favorites` parameter with the R^2 bounds for the favorite variables. For example, the code below reproduces the favorite variables chosen in Leamer[4]:

```
> favorites <- c("GDPCH60L", "OTHFRAC", "ABSLATIT",
+                "LT100CR", "BRIT", "GOVNOM1",
+                "WARTIME", "SCOUT","P60", "PRIEXP70",
+                "OIL", "H60", "POP1560", "POP6560")
> eg_fav <- sValues(economic_growth, R2_bounds = c(0.5, 1),
+                   favorites = favorites, R2_favorites = c(0.4, 0.8))
> eg_fav

Data: economic_growth,    Formula: GR6096 ~ .
R2 bounds: 0.5 - 1
Favorites: GDPCH60L OTHFRAC ABSLATIT LT100CR BRIT GOVNOM1 and 8 more.
R2 favorites: 0.4 - 0.8

abs(S-value) > 1:
  R2 (0.5, 1): EAST GDPCH60L IPRICE1 OTHFRAC P60 PRIEXP70

abs(t-value) > 2:
  Bayesian (R2 = 0.5): P60
  Bayesian (R2 = 1): P60
  Unconstrained OLS: IPRICE1
```

Further developments

We need more tools that help us study the sensitivity of our inferences and help us communicate it effectively. The idea of the `sValues` package is to bring one of these tools to the `R` community, with functions that (hopefully) make some of these tasks easier. This is still a work in progress though, and there is a lot that can be improved: what kind of tables, summaries or visualizations do you think would be most helpful both for exploring and for reporting the results? In what directions should the method be extended? For comments or suggestions, feel free to contact me or to make pull requests on github.

References

[1] Fernandez, C., Ley, E., and Steel, M. F. J. Model uncertainty in cross-country growth regressions. *Journal of Applied Econometrics 16*, 5 (2001).

[2] Leamer, E. E. Sets of posterior means with bounded variance priors. *Econometrica 50*, 3 (1982).

[3] Leamer, E. E. S-values and bayesian weighted all-subsets regressions. *European Economic Review 81* (2016). Model Uncertainty in Economics.

[4] Leamer, E. E. S-values: Conventional context-minimal measures of the sturdiness of regression coefficients. *Journal of Econometrics 193*, 1 (2016).

[5] Sala-i-Martin, X., Doppelhofer, G., and Miller, R. I. Determinants of long-term growth: A Bayesian Averaging of Classical Estimates (BACE) approach. *The American Economic Review 94*, 4 (2004).

[6] Zeugner, S. Bayesian model averaging with BMS. *R Package Vignette* (2011).