Econometrics, Exercise 3

Demirol, Engelen & Kuschnig

26 Mai 2018

Part 1, Task 1

We have written a function as per requirements and will demonstrate its capabilities and usefulness over the course of this document. The definition is displayed here, for the full code please consult the appendix.

- **Parameters**: In its most basic form the function would only take the endogenous (Y) and exogenous (X) variables as parameters. Considering the importance of finding the right values for $\tau_{0,1}$ we should also be able to modify these parameters. According to most of the literature we have set the default value of τ_1 as $\tau_0 * 100$. For τ_0 we will utilise an automatic approach (see George, Sun & Ni 2006) as the default value, which we will discuss later. We also allow for varying the number of iterations - both in terms of values stored and burned. We adopt the non-informative priors from class but allow for fiddling by including them as parameters. Last but not least we provide an option for centering and scaling the endogenous and exogenous variables, which we will also discuss later.
- **Output**: For handing over all the results that we might need we use a list. The central element of the output should generally be the posterior inclusion probabilities, but mean values and variances of β and the mean σ^2 of the saved models should also prove helpful. We also include an element with metadata, i.e. parameters of the function call, with $\tau_{0,1}$ being the most interesting ones in the case of automatic estimation.

```
ssvs = function(y,
                X,
                tau0 = NULL,
                tau1 = tau0 * 100,
                save = 4000,
                burn = 1000,
                s_prior = 0.01,
                S_prior = 0.01,
                standardise = TRUE) {...}
```

Before testing the water we load the required dataset and decide on values for τ_0 and τ_1 (which we handle as a scalar for the former).

```
data(datafls)

y = datafls[, 1]
X = datafls[, 2:ncol(datafls)]
rm(datafls)

tau0s = c(1, 1e-2, 1e-5, 1e-15)
tau1_scales = c(100, 1000)
```

Then we run our function with the specified parameters. We do so three (ignoring the repetitions due to differing $\tau_{0,1}$) times:
1. with the plain data, i.e. non-standardised
2. with standardised data
3. with standardised data and our automatic approach to setting $\tau_{0,1}$

```
i = 1
for(scale in tau1_scales) {
```

```
  for(tau in tau0s) {
    pips[[i]] = ssvs(y, X, tau0 = tau, tau1 = tau * scale, standardise = FALSE)[[1]]
    names(pips)[i] = paste0(tau, " & *", scale)
    i = i + 1
  }
}

i = 1
for(scale in tau1_scales) {
  for(tau in tau0s) {
    pips_z[[i]] = ssvs(y, X, tau0 = tau, tau1 = tau * scale, standardise = TRUE)[[1]]
    names(pips_z)[i] = paste0(tau, " & *", scale)
    i = i + 1
  }
}

pips_auto = ssvs(y, X)[[1]]
```

We immediately take note of the fact that setting τ_0 to very small values (such as 1e-15) is probably best suited for simulating the outcome of a coinflip. Pictured are the posterior inclusion probabilities of non-standardised and standardised run-throughs with $\tau_0 = 1\text{e-}15$ and $\tau_1 = 100$.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4570  0.4905  0.4980  0.5009  0.5050  0.6480

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4615  0.4850  0.4958  0.4956  0.5072  0.5300
```

Furthermore we notice that choosing a bigger scalar for τ_1 leads to lower posterior inclusion probabilities, but comparable results overall. This is in line with the literature, where it is used to influence the amount of variables that should make it into the final model. Pictured are the posterior inclusion probabilities of standardised run-throughs with $\tau_0 = 0.01$ and $\tau_1 = 100 \ / \ 1000$.
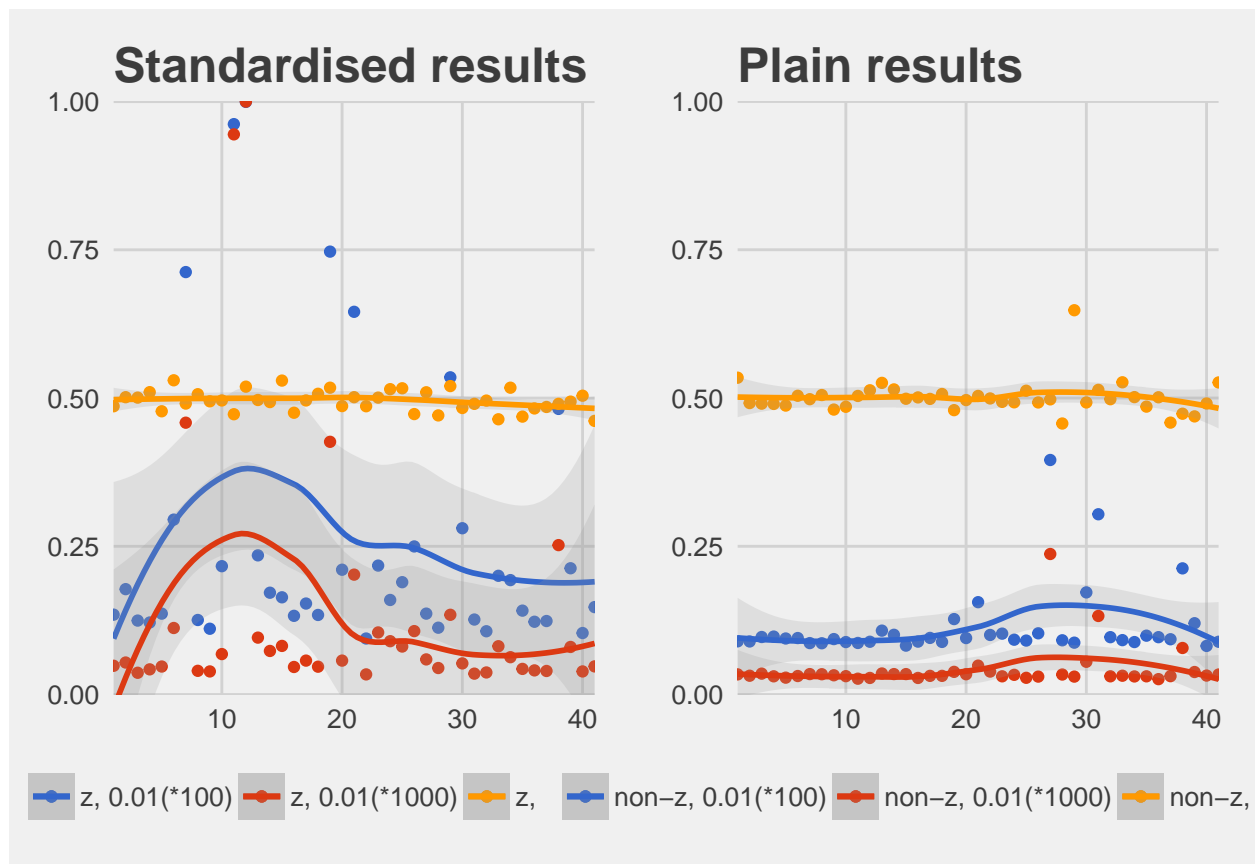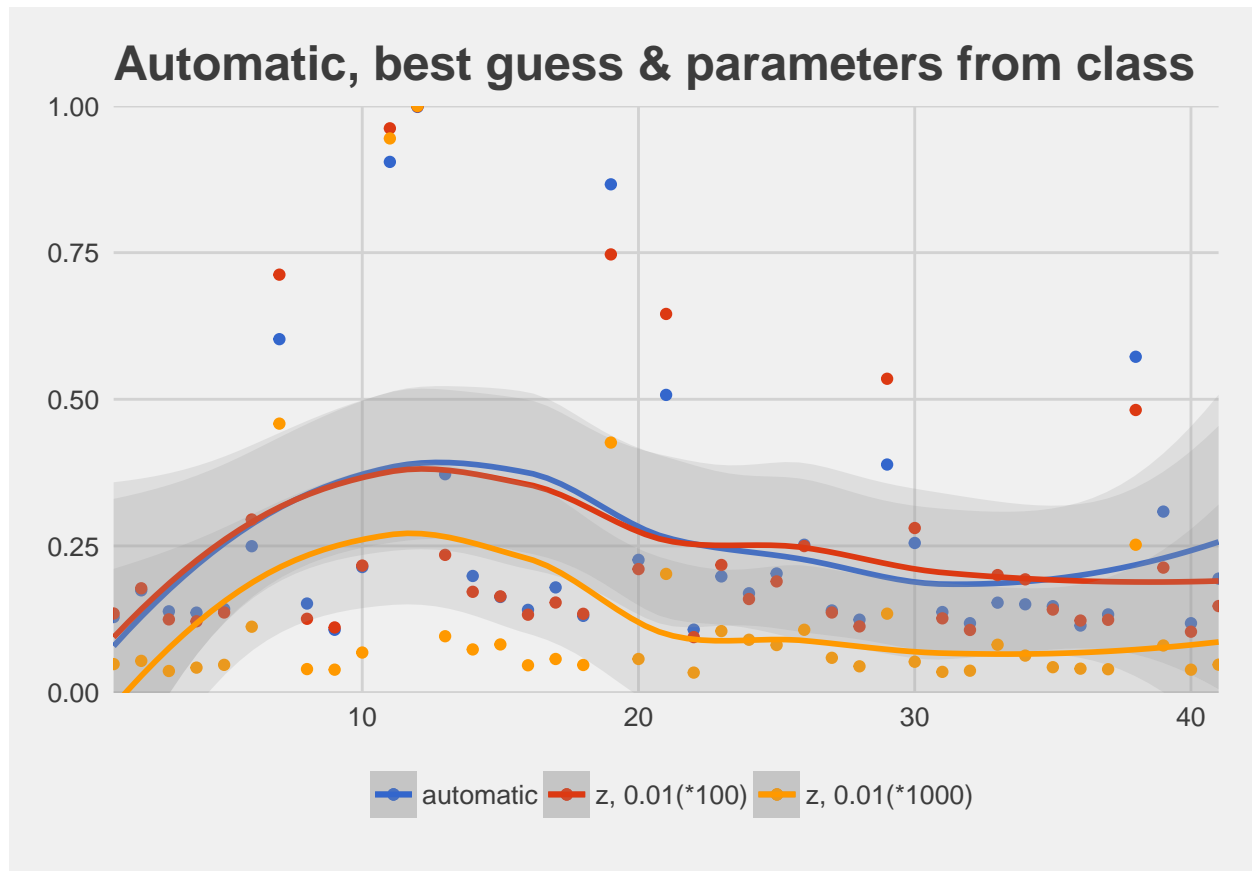
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0945  0.1265  0.1640  0.2597  0.2347  0.9995

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03400 0.04325 0.05725 0.13285 0.09600 1.00000
```

Processing the results graphically carries some further information: 1. our coinflip-hypothesis is holding up 2. standardising the data leads to more diverse and "nicer" posterior inclusion probabilities 3. the automatic approach of using OLS estimates of the variance for $\tau_{0,1}$ has very similar results to our best guess of $\tau_0 = 0.01$ and $\tau_1 = 1$.

Automatic, best guess & parameters from class

Legend: automatic · z, 0.01(*100) · z, 0.01(*1000)

Note: While fitting a line to the observations doesn't seem to make a lot of sense considering the data, we think it provides an easily noticeable and overall decent visual representation.