A Bayesian Heckman Selection Model: An Application in R with the Stan Programming Language [*]

Steven V. Miller *Clemson University*

Heckman corrections/models are "two step" procedures, though the reliance on convenience functions like `selection` in the `sampleSelection` package or `heckit` in Stata mask that the modeling process is two-fold. I show that insight from Bayesian approaches show how simple and straightforward this "two step" procedure is. Simulating fake data approximating a real-world example, I show that estimating a Bayesian probit model, calculating the inverse Mills ratio for selected observations, and plugging the inverse Mills ratio into a linear model produces parameters that do well to capture the known population parameters. The results effectively equal `heckit` in Stata and the output that `sampleSelection` produces. I close with a conclusion that I admittedly have not written yet and I probably should.

Keywords: Heckman models, sample selection, Bayesian methods

Introduction

Researchers in social science settings often encounter sample selection problems (or so called "selection effects") in their data. These are generally situations where an outcome of interest is continuous/interval, but only for the non-random subset of the data that "selected" into the data-generating process that produced the outcome. Applications here are multiple in the social/political sciences. The most famous example is the determinants of a women's wage, contingent on her selection into the labor force (e.g. Heckman, 1978, 1979). Heckman's defining work on methodology for sample selection problems and his application to women's wages in the labor force are why sample selection correction techniques bear his name (i.e. "Heckman models" or "Heckman corrections"). Other salient examples include the severity of a conflict, contingent on dyadic partners selecting into conflict initiation (e.g. Reed, 2000; Senese and Vasquez, 2003), the volume of economic assistance, given selection to receive economic assistance (e.g. Meernik, Krueger and Poe, 1998; Gibler, 2008), and the volume of donations, given a decision to donate (e.g. Hart, 2001), to name just a few. Though Heckman introduced these techniques in the late 1970s, their rise in social/political science research owes largely to the rise of cheap computing power in the late 1990s as well as the statistical software program Stata, which introduced a Heckman model in its software around the same time. Since 2000, about 186 articles have appeared in the *American Political Science Review*, *American Journal of Political Science*, and *Journal of Politics* that have advertised Heckman corrections for sample selection.

[*] This is a work in progress. Some care should be exercised here. I think I know what I'm doing, but this is new territory for me. Feedback welcome: svmille@clemson.edu.

Researchers who either cannot afford Stata or prefer to use the R programming language have not had the same opportunity to pursue research questions for which Heckman models are appropriate tools. `sampleSelection` (Toomet and Henningsen, 2008) came around 2008 for these models, years after Stata had introduced it as a feature in its software. To be clear, `sampleSelection` offers multiple tools and is great for simple selection problems. However, it will struggle with more complex selection models, often resulting in an inability to calculate standard errors.[1] Users either migrating to R from Stata or who are trying to compare results from `sampleSelection` (in R) to `heckman` (in Stata) will find comparison difficult. The results will never perfectly align, which is a natural (and to be clear: minor) byproduct of different programming approaches. More importantly, Stata's Heckman approach calculates standard errors differently (and communicates them for all parameters). Stata also reports different interpretations of the important parameters of the Heckman selection model (i.e. σ, ρ, and λ), which communciate the scope of selection. In the case of `sampleSelection`, only the inverse Mills ratio (λ) of the Heckman selection model comes with an estimate of uncertainty. Researchers in R may explore other similar models (e.g. a lognormal hurdle model) that are better supported and are more flexible, but econometric research has generally found the Heckman model to be preferable to these alternatives (e.g. Wooldridge, 2010).

In this paper, I introduce a Bayesian approach to a Heckman selection model that breaks down the two-stage estimation procedure into its constituent components. Researchers are accustomed to having just one command in their chosen program do this modeling for them, but the use of one command masks the two relatively simple procedures that are happening. The first procedure is a simple probit model predicting selection into the outcome. From this, the transformed fitted values of the probit model become the inverse Mills ratio (λ) to be included in the outcome model, which is a linear model. Whereas ρ is defined as the inverse Mills ratio (λ) divided over the residual standard error of the regression equations (σ), users can leverage recent programming advances to easily do this themselves. A Bayesian approach is novel because the posterior distribution of regression results creates more intuitive measures of uncertainty (and, importantly, intuitive measures of uncertainty for the selection parameters ρ, λ, and σ).

The remainder of this paper proceeds as follows. First, I describe the statistical procedure and underlying set of assumptions that go into the Heckman model, emphasizing the importance of corrections for sample selection. Thereafter, I introduce a simulation of fake data approximating real-world data that are intuitive to researchers of all walks who have read about Heckman models. The next section describes the coding procedure I recommend for approaching Heckman models from a Bayesian perspective. Whereas this is a dynamic document (Xie, 2013) and all simulations/estimations have reproducible seeds, users in R can install the libraries mentioned in this document and execute the code they see to get identical results. I close with an implication of the value of approaching Heckman models from a Bayesian perspective.

[1] See these examples: https://stat.ethz.ch/pipermail/r-help/2011-November/296781.html, https://stackoverflow.com/questions/36701318/errors-in-heckman-replication-from-stata-to-r

The Problem of Sample Selection, and How to Model It

Limited dependent variables appear in many forms in social/political science, typically clustering on issues of "truncation" and "censoring."[2] In this context, "truncation" occurs when a researcher attempts to make inferences about a larger population based on an available sample drawn from a distinct subpopulation. Truncation can happen, for example, when a researcher gathers data on income based on incomes above or below a certain threshold (e.g. the poverty line), making this kind of problem more about the sampling frame. "Censoring" occurs when values of the dependent variable are condensed into a single value (e.g. a researcher recodes all incomes below the poverty line to the same value). Statisticians have developed models over time to deal with both. Tobin (1958) famously introduced his "Tobit" estimator for modeling household purchases on durable goods when data are censored. Cragg (1971) extended Tobin's (1958) model to make inferences on from a truncated normal distribution for limited dependent variables.

The Heckman model concerns a closely related concept of "sample selection", alternatively known as "incidental truncation", that combines both issues of limited dependent variables. This is a common problem in labor economics for researchers interested in the determinants of wages. A researcher wanting to explain the determinants of a worker's wages can only receive data on people in the work force. Researchers out of the work force have no wage, making the truncation of wage "incidental" because it depends on another variable (namely, labor force participation). However, a researcher who does not appropriately model the incidental truncation of their dependent variable in their OLS model will receive biased estimates from the model's output. There would be omitted variables that determine both the probability of an outcome being observed and the variation observed in the outcome. More formally, a naive OLS model would produce biased estimates the extent to which the errors determining the probability of the outcome being observed and the variation in observed outcome are correlated.

The usual approach to dealing with sample selection is to add an explicit selection equation to the modeling procedure, resulting in a "two step" estimation procedure.[3] First, define the determinants of sample selection to be

$$z^* = w'\gamma + u$$

where $\boldsymbol{w}$ is exogenous an uncorrelated with the error terms at either the selection-level (u) or outcome-level (ϵ). When $z_i* > 0$, the outcome y_i is observed. An outcome is not observed otherwise. Formally:

$$z_i = \begin{cases} 1 & \text{if } z_i* > 0 \\ 0 & \text{if } z_i* \leq 0 \end{cases}$$

There is also a basic outcome equation that exemplifies the incidental truncation of the

[2] The language I use here purposely hews close to Wooldridge's (2010) terminology.

[3] This "two step" procedure should not be confused with a "two-part" procedure like that advocated by Manning, Duan and Rogers (1987).

data.

$$y_i = \begin{cases} x_i\beta + \epsilon_i & \text{if } z_i* > 0 \\ - & \text{if } z_i* \leq 0 \end{cases}$$

There are additional assumptions about the error terms in the selection equation (u_i) and the outcome equation (ϵ_i). The error term in the selection equation is assumed to follow a standard normal distribution while the errors in the outcome equation are assumed to have a mean of 0 and a variance equal to σ^2. The correlation between u_i and ϵ_i is ρ, an important parameter in the Heckman two-step procedure the extent to which it communicates the scope and direction of the incidental truncation. Formally:

$$\begin{aligned} u_i &\sim N(0,1) \\ \epsilon_i &\sim N(0,\sigma^2) \\ \text{corr}(u_i,\epsilon_i) &= \rho \end{aligned}$$

Put differently, the errors are assumed to follow a bivariate normal distribution with means of 0, a correlation of ρ, and a covariance matrix defined as

$$\begin{bmatrix} \sigma_\epsilon{}^2 & \varrho \\ \varrho & 1 \end{bmatrix}$$

where ϱ is the covariance between ϵ and u. The determinants of sample selection can be estimated with the following probit model.[4]

$$P(z = 1|w) = \mathbf{\Phi}(w'\gamma)$$

This probit model produces a $\hat{\gamma}$, the estimate of γ in the selection equation. From there, the individual estimates of λ (i.e. $\hat{\lambda^i}$) are obtained with the following formula:

$$\hat{\lambda} = \frac{\boldsymbol{\phi}(w_i\hat{\gamma})}{\mathbf{\Phi}(w_i\hat{\gamma})}$$

where $\boldsymbol{\phi}$ is the standard normal probability density function and $\mathbf{\Phi}$ is the standard normal cumulative distribution function. The second step of the "two step" procedure estimates the variation in the observed outcome to be

$$y = x\beta + \beta_\lambda\hat{\lambda} + \epsilon$$

where x is a vector of exogenous variables, $\hat{\lambda}$ is the inverse Mills ratio derived from the selection model and added to the outcome model, and β_λ is defined as $\rho\sigma_\epsilon$. When these assumptions are met, the model gives a β that is unbiased and consistent. Any case where $\rho \neq 0$ produces biased OLS estimates that are fixed with this procedure.

[4] Though regression modelers may use logistic regressions more than probit regressions when modeling binary dependent variables, the standard normal cumulative distribution function $\mathbf{\Phi}(.)$ in the probit model is conjugate to the normal distribution and is required for this procedure.

Researchers in the social and political science are particularly drawn to Heckman models or Heckman error corrections for data that have a sample selection component and an outcome that is continuous. Their use is common in labor economics, whose applied researchers devised this method for sample selection problems of observed wages contingent on labor force participation (Vella, 1998, for a review). Empirical applications of this approach also include conflict escalation (Reed, 2000; Senese and Vasquez, 2003), the allocation of economic aid (Meernik, Krueger and Poe, 1998; Gibler, 2008), campaign donations (e.g. Hart, 2001), among other diverse research topics. Though the pioneering work on sample selection had been around since the 1970s, their increased use in social scientific settings (especially in political science) owes to the rise of cheap computing power in the late 1990s and the widespread adoption of the statistical software program Stata. Stata has long included Heckman models as part of its software bundle.

How researchers estimate Heckman corrections with available software packages could stand to be improved. For one, Stata's `heckman` procedure does not appear to directly calculate ρ (the correlation of the two error terms in the modeling process) or σ (the standard error of the residual in the outcome equation). Instead, Stata's `heckman` command estimates the inverse hyperbolic tangent of ρ (i.e. $athro = \frac{1}{2}(\frac{1+\rho}{1-\rho}))$ and extracts an estimate of ρ from that for background information at the end of the output. Likewise, σ is not directly estimated either. Stata instead estimates the natural log of σ for numerical stability and exponentiating it for additional background information at the end of the output. Stata's `heckman` procedure does calculate standard errors for these estimates using the delta method, even for the important parameters of ρ and σ that are only indirectly estimated.

The `sampleSelection` package (Toomet and Henningsen, 2008) offers the most convenient way of estimating Heckman models in R. The package offers support for multiple types of sample selection and in a framework that would be intuitive to users migrating from Stata to R. However, some limitations emerge. For one, only the inverse Mills ratio (λ) has an estimate of uncertainty. ρ and σ are calculated and the estimates for them will almost always match what comes from Stata output, but there is no uncertainty provided with those estimates.

The next section proposes a means of estimating Heckman corrections with a Bayesian approach made possible by the Stan probabilistic programming language and made tractable and intuitive with the `brms` package. The following packages will be required for this exercise.

```r
library(tidyverse) # for most things workflow
library(brms) # for a convenient wrapper for Stan
library(mvtnorm) # for simulating random values from a covariance matrix
library(sampleSelection) # for comparison
library(stevemisc) # my toy package, for graph formatting
# ^ devtools::install_github("svmiller/stevemisc")
```

A Simulation

The basis for this simulation will be familiar to long-time Stata users, and indeed most people who learned about Heckman corrections by reference to women participation in the labor force. Stata has long provided a sample file named `womenwk.dta` (StataCorp, 2013). The data set has 2,000 observations total with 1,343 uncensored (i.e. participating in the labor force and receiving a wage) and 657 censored (i.e. women who do not participate in the labor force and thus have no wage). The data are fake, but approximate real-life examples that motivated Heckman's pioneering research on the estimation procedure that bears his name in Stata. A simple two-step Heckman model on the fake data produced the following results in the user guide (Table 1).

Table 1: The Coefficients and Standard Errors from a Simple Pedagogical Heckman Model in Stata

Level	Term	Coef.	Std. Err.
Outcome	Education	0.9899537	0.0532565
	Age	0.2131294	0.0206031
	Intercept	0.4857752	1.0770370
Selection	Married	0.4451721	0.0673954
	Children	0.4387068	0.0277828
	Education	0.0557318	0.0107349
	Age	0.0365098	0.0041533
	Intercept	-2.4910150	0.1893402
Heckman Parameter	athrho	0.8742086	0.1014225
	log(sigma)	1.7925590	0.0275980
	rho	0.7035061	0.0512264
	sigma	6.0047970	0.1657202
	lambda	4.2244120	0.3992265

Stata gives no clues to the data-generating process for this fake data, but this simulation will assume the model's point estimates are perfectly correct (even to the seven decimal points that come default in Stata output) and try to recapture them from a manual Bayesian approach. Toward that end, let us first define the number of observations (n = 2,000) and, importantly, define the covariance matrix. We will assume that the error term of the outcome equation and selection equation are bivariate normal. The variance of ϵ is defined as σ^2 and the variance of u is normalized to 1. Pursuant to the Heckman estimation procedure, the correlation between ϵ and u is ρ and the covariance of ϵ and u is equal to $\rho\sigma$. We will assume the Stata model was perfectly correct and that the true population ρ is .7035061 and the true population σ is 6.004797.

```r
N <- 2000

# covariance matrix where top-left is sigma_e^2,
# top-right and bottom-left is varrho, and bottom-right is 1.
covmat <- matrix(c(6.004797^2,
                  .7035061*6.004797,
                  .7035061*6.004797, 1), ncol=2)
```

Thereafter, we will set a reproducible seed (8675309) and simulate the bivariate normal errors. Do note the default output from these simulations is an object of class matrix, but we will eventually this matrix to a data frame for an eventual merge into the simulated data. Therein, the first "column" of the simulated errors is the simulated ϵ while the second column is the simulated u.

```r
set.seed(8675309) # Jenny, I got your number...
errors <- rmvnorm(N, sigma=covmat)

errors %>%
  as.data.frame %>% tbl_df() %>%
  rename(e = V1, u = V2) -> errorsdf
```

Next, we'll simulate the data to match what is provided in the womenwk.dta file in Stata. educ will have a range from 1 to 16 and will represent higher education through a simulated proxy of a woman's reported years in school. age will represent the women's age and range from 18 to 64. married will be a dummy variable indicating whether a woman is married or not. children will represent how many children a woman has, ranging from 0 to 5.

```r
set.seed(8675309) # Jenny, I got your number...
tibble(educ = sample(1:16, N, replace = T),
       age = sample(18:64, N, replace = T),
       married = sample(0:1, N, replace = T),
       children = sample(0:5, N, replace = T)) -> Data
```

Afterward, we are going to add the simulated errors to the fake data we generated and create additional data that mimic the Stata output in Table 1. We will define an outcome lnwage that is the natural log of a woman's wage, provided she is in the work force. The natural log of a woman's wage will be .4857752 + .9899537*educ + .2131294*age + e.[5] z_star will represent the selection component of this procedure and will be defined as

[5] The y-intercept is not significant in the Stata output provided in its user guide and has a z-value that is about .45. This implies the intercept in the outcome equation was probably simulated at 0 and the intercept that emerged from the Stata procedure comes from random chance. No matter, we will proceed with this y-intercept for this simulation.

`-2.491015 + .4451721*married + .4387068*children + .0557318*educ + .0365098*age + u`. `observed` will equal 1 (i.e. the woman received a wage) if `z_star` is greater than 0. For clarity, we will also hard-code missingness in `lnwage` if `observed` is 0 (i.e. `z_star` $\leq$ 0).

```r
Data %>%
  # add error terms
  bind_cols(., errorsdf) %>%
  mutate(lnwage = .4857752 + .9899537*educ + .2131294*age  + e,
         z_star = -2.491015 + .4451721*married + .4387068*children +
           .0557318*educ + .0365098*age + u,
         observed = ifelse(z_star > 0, 1, 0),
         lnwage = ifelse(observed == 0, NA, lnwage)) -> Data
```

The next part of the procedure will run a probit model for the selection component of the "two-step" procedure. First, however, we will set some weakly informative priors for the probit model. Setting priors is a sine qua non feature of Bayesian analysis, though researchers are free to invest more or less time and energy into this as they see fit. However, knowing that the first step of this procedure is a probit model, we can set some reasonably informative, but still ignorant, prior distributions on the regression parameters and intercept to be 0 with a standard deviation equal to 2.5.

```r
probit_priors <- c(set_prior("normal(0,2.5)", class="b"),
                   set_prior("normal(0,2.5)", class="Intercept"))
```

Thereafter, we will run the first step of the "two step" procedure. This is a minimal probit model modeling whether a wage is observed or not. Researchers at this stage should look carefully at the model output and weigh the importance of potential diagnostic errors that Stan communicates. The researcher is also free to add more or fewer chains or cores as they see fit. For convenience sake, though the researcher should have the same chains and iterations at the first step as they have at the second step. Whereas this is a probit model, the researcher should also set `inits = 0` in the `brms` wrapper for Stan. Otherwise, the sampling will start over non-sensical values in the probit context, reject them, and slow down the convergence procedure.

```r
P1 <- brm(observed ~ age + educ + married + children,
          data=Data, family=bernoulli(link="probit"),
          seed = 8675309, # Jenny, I got your number.
          inits = 0, chains = 4, cores = 4,
          prior=probit_priors)
```

The researcher can stop here and look at these results as they see fit to see what is discernibly associated with the outcome being observed. This guide will push forward, though. First, extract the residuals from the probit model and save them for later.

```
residP1 <- resid(P1, summary= F) %>% # give us *all the residuals*
  as.data.frame %>% tbl_df()

residP1 %>%
  # create a rowid
  rowid_to_column() %>%
  # group_by it.
  group_by(rowid) %>%
  # turn wide to even longer
  gather(var, val, 2:ncol(.), -rowid) %>%
  # Give us standard deviation of those residuals
  summarise(proberrorsds = sd(val)) -> residP1sds
```

The final part of this first step will extract the inverse Mills ratio (i.e. $\hat{\lambda}$) from the fitted values of this probit model. Recall, the inverse Mills ratio is computed as $\frac{\phi(w_i\hat{\gamma})}{\Phi(w_i\hat{\gamma})}$. This is a simple computation in R, given the output from a brms wrapper for Stan.

```
# extract fitted values
fitP1 <- fitted(P1, scale="linear")
# below is the inverse Mills ratio
Mills <- dnorm(fitP1)/pnorm(fitP1)

# Let's add it to our simulated data.
Mills %>% tbl_df() %>%
  select(Estimate) %>%
  bind_cols(Data, .) %>% tbl_df() %>%
  rename(imr = Estimate) -> Data
```

This will allow us to proceed with the second step of the "two step" procedure. First, let's set some weakly informative priors on the outcome model. Of note, we will retain the normal(0, 2.5) prior for the betas (including the beta for the inverse Mills ratio). The researcher should give more thought to the priors for the intercept and σ in the linear model absent an intuitive transformation of the data (i.e. scaling the inputs by two standard deviations, a la Gelman (2008)). The intercept may be all over the place and we will not know ahead of time just how much is left unexplained by the regression inputs. Therefore, we will set more diffuse ignorance priors of normal(0, 5) for the intercept and cauchy(0,10) for σ.[6]

[6] It should be obvious that σ has a hard left bound at 0 and can never be negative. Thus, Bayesian modelers prefer to refer to these kinds of priors as "half-Cauchy" because the Stan probabilistic programming language will not sample left of 0 for a parameter like σ.

```
lm_priors <- c(set_prior("normal(0,2.5)", class="b"),
               set_prior("normal(0,10)", class="Intercept"),
               set_prior("cauchy(0,10)", class="sigma"))
```

The final part of the "two step" procedure is the estimation of variation in the outcome, provided the outcome is observed. One thing that may not be obvious to the researcher who relies on one command to do a "two step" procedure is the second step includes the invese Mills ratio (i.e. the transformed fitted values from the first step) as another regression input. The coefficient that emerges from that is typically communicated outside the outcome equation as a parameter of the Heckman model, masking what it truly is in the procedure.

```
L1 <- brm(lnwage ~ age + educ + imr,
          data = subset(Data, observed == 1),
          family="gaussian",
          seed = 8675309, # Jenny, I got your number...
          chains = 4, cores = 4,
          prior = lm_priors)
```

Before evaluating the model output, we should create a data frame of important parameters from the Heckman model. First, we will create a data frame that includes all the simulated σs from the outcome equation.

```
rseL1 <- VarCorr(L1, summary=F) %>%
  as.data.frame %>% tbl_df() %>% rename(sigma = 1)
```

Finally, let's create a data frame that includes all the simulated values for the intercept and the betas for the `age` variable, the `educ` variable, and the inverse Mills ratio we plugged into the outcome equation. We will add the simulated σs from the outcome equation and the simulated standard deviations of the residuals from the probit model in the first "step". The extent to which $\rho\sigma = \lambda$, we can calculate the simulated values of ρ to be λ/σ but with one important caveat. The σ that Stata calculates as part of its output includes the errors from the first step as well, which means we will need to add those errors to the σs from the outcome equation or risk underestimating σ and overestimating ρ.

```
fixef(L1, summary=F) %>% tbl_df() %>%
    bind_cols(., rseL1) %>%
    bind_cols(., residP1ds) %>%
    mutate(rho = imr/(sigma+proberrorsds)) -> heck_params
```

Finally, we can run the same procedure with the `selection()` function in `sampleSelection`. This will allow us to compare our results.

```
H1 <- selection(observed ~ age + educ + married + children,
                lnwage ~ age + educ, data=Data, method = '2step')
```

We will start with a comparison of the results of a manual Bayesian estimation approach for the selection equation and the outcome equation with the results from `sampleSelection`. Whereas the data are simulated assuming the Stata results are perfectly correct, we can add the information from Table 1 to our summary data as well.

```
H1coefnames <- rownames(summary(H1)$estimate) # rownames, helpfully.
summary(H1)$estimate %>%
  as.data.frame %>%
  mutate(term = H1coefnames) %>%
  select(term, everything()) %>%
  rename(estimate = 2,
         std.error = 3) %>%
  select(1:3) %>%
  # by default, brms tidiers include 90% intervals
  mutate(lower = estimate - 1.645*std.error,
         upper = estimate + 1.645*std.error,
         term = ifelse(term == "(Intercept)", "Intercept", term)) %>%
  mutate(level = c(rep("Selection", 5),
                   rep("Outcome", 3),
                   rep("Heckman Parameter", 3)),
         model = "sampleSelection") -> tidyH1

broom::tidy(P1) %>%
  mutate(term = str_replace(term, "b_","")) %>%
  filter(term != "lp__") %>%
  mutate(level = "Selection",
         model = "Manual Bayesian") -> tidyP1

broom::tidy(L1) %>%
  mutate(term = str_replace(term, "b_","")) %>%
  slice(1:3) %>% mutate(level = "Outcome",
                        model = "Manual Bayesian") -> tidyL1

bind_rows(tidyP1, tidyL1) %>%
  bind_rows(tidyH1) -> tidyAll

heck_examp %>%
  fill(level) %>% mutate(model = "Stata's heckit (Assumed Parameters)") %>%
  na.omit %>%
  # create "uppers" and "lowers"
  mutate(lower = estimate - 1.645*std.error,
         upper = estimate + 1.645*std.error) %>%
  mutate(term = ifelse(term == "(Intercept)", "Intercept", term),
         term = ifelse(term == "Education", "educ", term),
         term = ifelse(term %in% c("Married", "Children", "Age"),
                       str_to_lower(term), term)) %>%
  bind_rows(tidyAll, .) %>% tbl_df() -> tidyAll
```

The results in Figure 1 show that a manual Bayesian approach to a Heckman model matches well with what sampleSelection will produce. The results are functionally identical at both the outcome and the selection equation, with the only caveat that the intercept of the outcome equation is considerably diffuse. In almost every application, the 90% intervals around the mean coefficient for the manual Bayesian approach include the assumed population parameter produced from Stata's heckit estimation. The only clear instance where it did not was the coefficient for married women at the selection phase. The true population effect is .4451721, but the mean coefficient from the Bayesian probit model was 0.323 with a standard deviation of 0.069. Yet, this might be a function of the random simulation of the data because the results are identical to the coefficient estimate (0.323) and standard error (0.069) that sampleSelection produces.

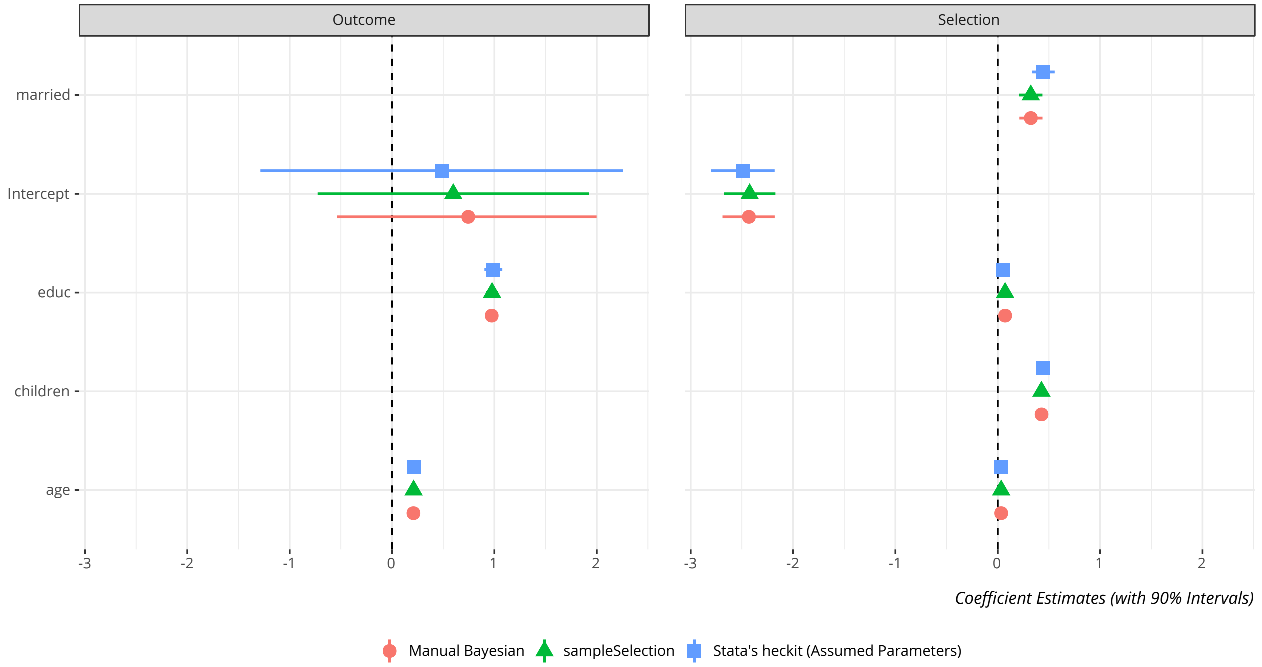
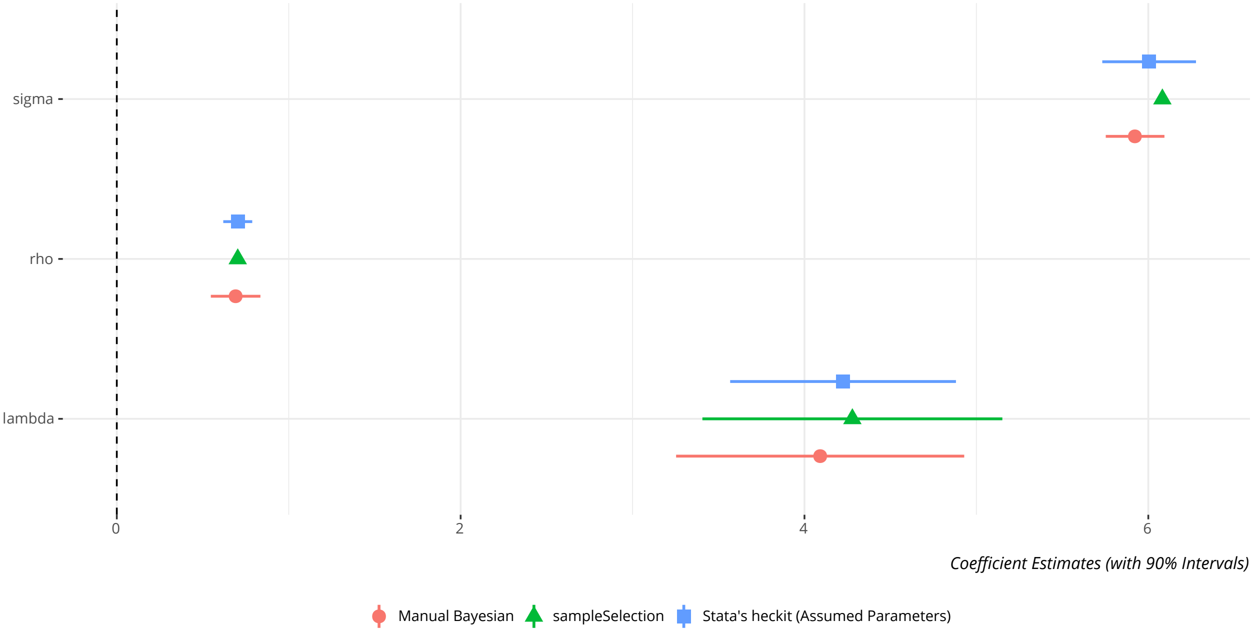


Figure 1: A Comparison of a Manual Bayesian Approach to a Heckman Two-Step Procedure (Coefficients)

Yet the real value of a manual Bayesian approach to a two-step Heckman procedure is the simulations of the important parameters of the Heckman model: ρ, σ, and λ. These are the important parameters of the Heckman model that communicate the scope of selection, but it is foggy how we can interpret the uncertainty of those estimated parameters. Stata does not directly calculate ρ or σ, but instead the inverse hyperbolic tangent of ρ and the natural log of σ. Thus, ρ and σ are only indirectly estimated, but come with standard errors calculated by the delta method. `sampleSelection` provides no estimate of uncertainty for ρ and σ, only for λ. From a Bayesian perspective, this is a simple solution. The posterior distribution that emerges from the modeling procedure gives the researcher an estimate of central tendency (mean or median, per the researcher's discretion) and a host of ways of calculating uncertainty around it (i.e. standard deviation, quantiles).

Figure 2 shows that a manual Bayesian approach approximates known/assumed population parameters for λ, ρ, and σ reasonably well. The ρ from the manual Bayesian approach is functionally identical to the ρ calculated by `sampleSelection` and Stata's `heckit` command (which is treated as the assumed population ρ). The λ (the coefficient for the inverse Mills ratio in the outcome equation) effectively captures the known population λ. An adjustment to σ that also includes the standard deviation of the residuals from the probit model produces an overall σ that captures the known population σ as well.



Note: sampleSelection provides no estimate of uncertainty for sigma or rho.

Figure 2: A Comparison of a Manual Bayesian Approach to a Heckman Two-Step Procedure (Heckman Parameters)

Overall, the results of this simulation show how to faithfully perform a Heckman "two step" model in the absence of Stata or independent of `sampleSelection`. Researchers accustomed to having one command perform a "two step" procedure miss that the function is performing two regressions, not one. Recent advances in Bayesian approaches and the clarity of these functions allow the researcher to not only do this themselves, but give the researcher greater flexibility over the parameters of Heckman approach. The important parameters communicating the scope of selection are not only easily estimated and extracted, but come with intuitive estimates of uncertainty often opaque or unavailable in alternative functions and programs.

Conclusion

I'll need to write this. Basically, I think the major value of this is random effects are super simple from Bayesian perspective and that's the real value in doing this.

References

Cragg, John G. 1971. "Some Statistical Models for Limited Dependent Variables with Application to the Demand for Durable Goods." *Econometrica* 39(5):829–844.

Gelman, Andrew. 2008. "Scaling Regression Inputs by Dividing by Two Standard Deviations." *Statistics in Medicine* 27(15):2865–2873.

Gibler, Douglas M. 2008. "United States Economic Aid and Repression: The Opportunity Cost Argument." *Journal of Politics* 70(2):513–526.

Hart, David M. 2001. "Why Do Some Firms Give? Why Do Some Give A Lot?: High-Tech PACs, 1977-1996." *Journal of Politics* 63(4):1230–1249.

Heckman, James J. 1978. "Dummy Endogenous Variable in a Simultaneous Equation System." *Econometrica* 46:931–939.

Heckman, James J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47(1):153–161.

Manning, W.G., N. Duan and W.H. Rogers. 1987. "Monte Carlo Evidence on the Choice between Sample Selection and Two-Part Models." *Journal of Econometrics* 35(1):59–82.

Meernik, James, Eric L. Krueger and Steven C. Poe. 1998. *Testing Models of U.S. Foreign Policy: Foreign Aid during and after the Cold War*. Vol. 60.

Reed, William. 2000. "A Unified Statistical Model of Conflict Onset and Escalation." *American Journal of Political Science* 44(1):84–93.

Senese, Paul D. and John A. Vasquez. 2003. "A Unified Explanation of Territorial Conflict: Testing the Impact of Sampling Bias, 1919-1992." *International Studies Quarterly* 47(2):275–298.

StataCorp. 2013. *Stata User's Guide Release 13*. College Station, TX: Stata Press.

Tobin, James. 1958. "Estimation of Relationships for Limited Dependent Variables." *Econometrics* 26(1):24–36.

Toomet, Ott and Arne Henningsen. 2008. "Sample Selection Models in R: Package sampleSelection." *Journal of Statistical Software* 27(7):1–23.

Vella, Francis. 1998. "Estimating Models with Sample Selection Bias: A Survey." *Journal of Human Resources* 33(1):127–169.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Xie, Yihui. 2013. *Dynamic Documents with R and knitr*. Boca Raton, FL: CRC Press.