Replication Exercise

Luke Duggan

10th April 2022

1 Introduction and Data Source

This exercise replicates and extends the results from the economist Ray C. Fair's 1978 paper "A Theory of Extramarital Affairs". The data was obtained from professor Fair's website (accessed 10/04/2022): https://fairmodel.econ.yale.edu/rayfair/pdf/2011b.htm)

The data was preprocessed in Excel, and the attached csv file "fair1978" was used for the analysis. Preprocessing was minimal, being essentially limited to adding column names and dropping some columns containing only 0's.

2 Description of the Dataset

Fair's paper contains two distinct datasets: both are surveys conducted by magazines, one by *Psychology Today* (*PT*) and one by Redbook. I used only the former data in this exercise. The *PT* survey was conducted in July, 1969, and there were 601 recorded respondents. I will now briefly discuss the variables in the dataset and some general features of this type of data.

The outcome variable of interest is a sort of count variable measuring how often a respondent had engaged in extramarital affairs over the previous year. I say "a sort of count variable" because, as is often the case with survey data, several distinct counts have been collapsed into one category. In particular, the permissible values for the outcome are:

0 = none, 1 = once, 2 =
twice, 3 = 3 times, 7 = 4–10
times, 12 = monthly, 12 =
weekly, 12 = daily

The explanatory variables are: sex, age, how long the respondent has been married, how religious they are, their level of education, their occupation, and how they would subjectively rate their satisfaction with their marriage. Most of these are self-explanatory, and only three call for clarification. The occupation variable is categorical, and corresponds to the Hollingshead "four-factor" scale of varieties of job. The religiosity and satisfaction with marriage variables range from 1 to 4 and 1 to 5 respectively, with an increase signifying greater religiosity or marital satisfaction.

Perhaps the most obvious and immediate problem with this sort of data is that it cannot be plausibly interpreted as a random sample of married Americans. The format of the *PT* survey was that the survey was published in the magazine and readers were invited to mail in their answers. Thus, our outcome variable will be correlated with whatever characteristics tend to be possessed people who mail in surveys to *Psychology Today.* Since the survey-answering readership of *Psychology Today* is not a random sample of married Americans, the use of respondents to the survey in place of a random sample is likely to induce bias. This phenomenon has been called "responder bias" and is particularly relevant for these sorts of questions about intimate aspects of people's lives. Our extension of Fair's work will focus on this shortcoming in the data collection process.

3 Results and Replication

Fair writes that "... because many values of y [...] are zero, it would clearly be incorrect to use ordinary least squares to estimate the equations. The obvious technique to use in this case is the Tobit estimator" (pg.57). More generally: one would like to distinguish two conceptually distinct issues. The first is how our explanatory variables affect people's decisions on whether to have an affair *at all*; the second is how the variables affect the *extent* of a person's extramarital affairs, given that the outcome is nonzero.

This problem is structurally similar to one encountered by the great economist James Tobin. In examining determinants of consumer spending on luxury goods, he found that many of the consumers in his dataset spent exactly $0 on luxury goods. The so-called "Tobit estimator" aimed to estimate the effect of a marginal increase in an explanatory variable on the "latent" or "underlying" propensity to consume luxury goods, bearing in mind that for values of the latent variable below some threshold, some consumers would still choose to consume no luxury goods at all.

Fair's original parameter estimates were:

Fair's Results

* α = 0.1
** α = 0.05
*** α = 0.01

Variable		Coeff. Est.	t-Stat.	Coeff. Est.	t-Stat.
Constant	...	7.60	1.92 **	8.17	2.96 **
Occupation	z_7	.213	.67	.326	1.29 *
Education	z_6	.0252	.11	...	...
Husband's occupation	...	...	...	...	...
Marital happiness	z_8	−2.27	−5.48 ***	−2.28	−5.61 ***
Age	z_2	−.193	−2.37 ***	−.179	−2.26 **
No. years married	z_3	.533	3.63 ***	.554	4.13 ***
Children	z_4	1.02	.79	...	...
Degree of religiosity	z_5	−1.70	−4.15 ***	−1.69	−4.14 ***
Sex	z_1	.945	.88	...	...
	$\hat{\sigma}^2_{PT}$	8.26	...	8.25	...
No. observations		601‡		601‡	

Note that there are two sets of coefficient estimates: the latter column is for a reduced model that drops several variables from the first, to little effect overall. Fair's analysis is mainly qualitative, focusing on the sign of the estimates and only secondarily on their magnitude. The results are, in any case, quite commonsensical: excepting the intercept, the largest and most significant effects are associated with marital satisfaction and religiosity: as the levels of these categorical variables increase, there is a large drop-off in the estimated propensity to engage in extramarital affairs. There are also smaller but significant estimates on two other variables: the propensity to engage in extramarital affairs is estimated to decline with age and increase with the length of time for which one has been married.

Replicating these results in R is not too difficult thanks to an "off the shelf" function for estimating a Tobit model from the package AER:

```
library(AER)
model <- tobit(y ~ sex + age + years_married + children + how_religious +
                   education + occupation + rate_marriage, data = data,
                 left = 0)
summary(model)
```

The results are:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	7.60849	3.90599	1.948	0.051426	.
sex	0.94579	1.06287	0.890	0.373548	
age	-0.19270	0.08097	-2.380	0.017316	*
years_married	0.53319	0.14661	3.637	0.000276	***
children	1.01918	1.27957	0.797	0.425741	
how_religious	-1.69900	0.40548	-4.190	2.79e-05	***
education	0.02536	0.22767	0.111	0.911304	
occupation	0.21298	0.32116	0.663	0.507220	
rate_marriage	-2.27328	0.41541	-5.472	4.44e-08	***

It will be seen that, up to differences in rounding, the point estimates are identical to Fair's. There is a slight difference in the computation of standard errors: Fair uses an exact t-distribution with 601 degrees of freedom, whereas R takes it that 601 is a more than sufficient degrees of freedom to justify using the asymptotically valid normal distribution. This doesn't affect the significance of any parameter estimate.

4 Extension

Fair seems to have been unaware of Heckman selection models when writing this paper (Heckman's first paper on the subject was written about two years prior). Heckman selection models are perhaps more appropriate to the issues raised at the outset of this exercise, since they involve the estimation of distinct *selection* and *outcome* equations: that is, they estimate both the effect of an explanatory variable on the *probability* of the outcome variable's being nonzero and, separately and conditional upon its being nonzero, on the *magnitude* of the outcome variable.

Estimating a Heckman selection model in R is also not too difficult, owing to the existence of an "off the shelf" function. The only subtlety is that, for the selection equation, we have to recode our outcome variable to be binary: zero or nonzero.

```
### Estimate a Heckman model.
library(sampleSelection)

# We have to define a binary outcome variable for the selection equation.

thing <- numeric(601)
for (i in 1:length(data$y)) {
  if (data$y[i] > 0) {
    thing[i] <- 1
  }
  if (data$y[i] == 0) {
    thing[i] <- 0
  }
}

# Then estimate the model.

model2 <- heckit(selection = thing ~ sex + age + years_married + children + how_religious +
                   education + occupation + rate_marriage,
                 outcome = y ~ sex + age + years_married + children + how_religious +
                   education + occupation + rate_marriage,
                 data = data)
summary(model2)
```

The results are:

```
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.77940    0.51255   1.521 0.128895
sex             0.17346    0.13799   1.257 0.209253
age            -0.02458    0.01042  -2.360 0.018612 *
years_married   0.05434    0.01881   2.889 0.004005 **
children        0.21664    0.16517   1.312 0.190154
how_religious  -0.18547    0.05163  -3.593 0.000355 ***
education       0.01126    0.02952   0.382 0.702930
occupation      0.01367    0.04140   0.330 0.741421
rate_marriage  -0.27179    0.05347  -5.083 5.03e-07 ***
Outcome equation:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     4.57504   11.15810   0.410    0.682
sex             2.65669    4.55357   0.583    0.560
age            -0.51448    0.57499  -0.895    0.371
years_married   1.35905    1.21799   1.116    0.265
children        3.49256    6.06633   0.576    0.565
how_religious  -4.42642    4.13525  -1.070    0.285
education       0.04795    0.63916   0.075    0.940
occupation      0.56871    0.87044   0.653    0.514
rate_marriage  -5.90339    5.63577  -1.047    0.295
Multiple R-Squared:0.1876,      Adjusted R-Squared:0.1353
   Error terms:
               Estimate Std. Error t value Pr(>|t|)
invMillsRatio    27.089     27.318   0.992    0.322
sigma            23.261         NA      NA       NA
rho               1.165         NA      NA       NA
```

These results vindicate, in part, Fair's original estimates. Exactly the same variables have turned out to be significant in the selection equation as were estimated to be significant in Fair's original paper: furthermore, the estimated signs on the coefficients agree, as do their relative magnitudes.

However, the results also qualify Fair's conclusion. As with our analysis of the Mroz dataset in class, *none* of the estimates is statistically significant in the outcome equation. This suggests that religiosity and marital satisfaction, and to a lesser extent age and duration of marriage, are significant in affecting a person's *decision* on whether or not to have an affair, but not on the extent of such behaviour given that it occurs.

Finally, we note that the inverse Mills ratio is not statistically significantly different from zero. Thus, we cannot reject the null that there is no selection bias effect. It would be worth looking at the power of this test for, as discussed above, it is *a priori* quite likely that there is indeed a selection bias here.