

Missing the data or missing the point?

Stas Kolenikov

March 14, 2002

In this short paper, I propose a simulation study of a simple case of non-ignorable / systematically missing data. I show that with as few as 5% of the observations missing, the rate of false rejections can be as high as 90% at a nominal 5% significance level.

The setup for simulation is as follows. I generate a sample of 1000 i.i.d. observations from a bivariate standard normal vector:

$$\begin{pmatrix} x \\ y \end{pmatrix} \sim N(\mathbf{0}, I_2) \quad (1)$$

so x and y are independent. The y variable is missing according to the following scenario:

$$P = I\{y \text{ is observed}\} = I\{x - y < \Phi^{-1}(1 - \tau)\sqrt{2}\} \quad (2)$$

where P is the indicator of y being observed, Φ is the standard normal c.d.f., and the condition in the RHS of (2) removes $\tau\%$ of the data. I am taking $\tau = 5\%$ in this study, so y is missing if it is below the line $y = x - 2.326$. The sketch is shown on Fig. 1; clearly, the lower right part of the circular scatterplot has been cut (can you guess that without knowing that it is bound to be missing? The response that I was receiving when presenting this to a group of economists was, “Now that you said what to look at, it is apparently missing”). What is the effect of missing data?

Table 1 reports the comparison results for three different estimation procedures for the above data. The second column marked “Full sample” is an OLS regression that uses all 1000 values of x and y . Suppose however that the researcher only observes the 940 observations censored according to (2). Then the results of the regression of y on x are as reported in column 3, “Censored sample”. Clearly, the results are not satisfactory. The regressor comes out to be statistically significant with p -value of 0.27%.

The fourth column of Table 1 reports the results of estimation procedure that accounts for the missing data. In this case, this is an interval regression, which is a (sort of) generalization of tobit model where censoring is allowed to be dependent on some covariates. This is a full information

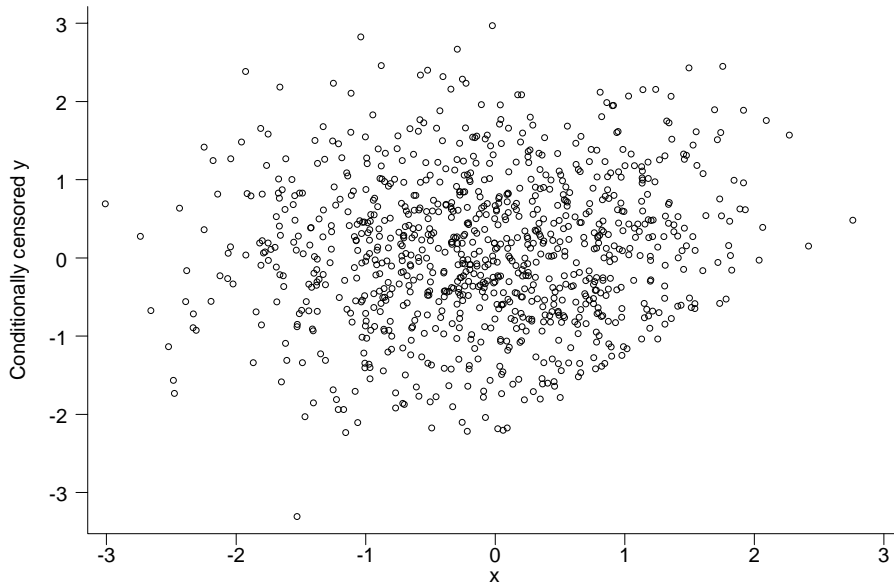


Figure 1: The scatterplot with 5% data missing in a systematic way.

maximum likelihood procedure. The likelihood function for an individual observation is

$$L(\theta|y, P, x) = \begin{cases} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{(y-\beta_0-\beta x)^2}{2}\right], & P = 1 \\ \Phi(x - 2.326), & P = 0 \end{cases} \quad (3)$$

In other words, the likelihood involves the observed y in the standard OLS manner whenever y is available, and it uses the information on **why** y is missing, otherwise. Note that in this particular simulation case we are happy enough to know this “why” issue. In real research, this information would hardly be available.

The interval regression produced quite sensible results. Most importantly, it did not flag significance of the regressor. It also produced the point estimates that are quite close to those of the OLS regression based on the complete sample. Note higher standard errors; this is the price that we always pay when some data is missing. The ratio of squared standard errors of the slope as reported for the interval regression and the full sample regression is 1.070, which roughly corresponds to 6% of missing data (recall that we were aiming at 5% of data to be missing in our missing data mechanism). This is what might have been expected, as the variance of the coefficient estimates is the information matrix, the expected value of which is proportional to the number of observations in simple enough i.i.d. cases. Thus employing the likelihood procedure that accounts for the missing

	Full sample	Censored sample	Interval regression	Heckman model
Slope	-0.0257 (0.0315)	0.0982 (0.0327)**	-0.0295 (0.0326)	-0.0289 (0.0329)
Constant	-0.0072 (0.0318)	0.0916 (0.0311)**	-0.0094 (0.0321)	-0.0087 (0.0325)
Observations	1000	940	1000	1000
R-squared	0.00	0.01	N/A	N/A
LR	0.6667	9.0142	0.8202	0.7693
p-value	0.4142	0.0027	0.3651	0.3804

Standard errors in parentheses. *, significant at 5%; **, significant at 1%.

Table 1: Comparison of different estimation methods.

data does combat the bias, but does not necessarily provide much efficiency gains.

Yet an alternative method which uses even less information on the censoring mechanism is Heckman model so popular in economics that its author received Nobel prize in 2000 for this contribution. This is a two equation model where the first equation is a probit regression with the indicator of censoring of y as a dependent variable, and a bunch of regressors, and the second equation is the usual regression for the observed cases of y on (presumably, a different) bunch of regressors with the additional term derived from the first equation (called inverse Mills ratio) that corrects for the bias caused by missing data. Then the joint likelihood of the two equations that accounts for probable correlation of the disturbance terms in those equations is written down and supplied into a numerical maximum likelihood procedure. The results of the Heckman model are reported in the last column of Table 1. The censoring equation is reported to be $-1.002x + 2.260$ which is, up to a sign, is quite close to the censoring mechanism (2).

It might be noted in passing that this particular data problem might turn out to be numerically challenging for Heckman model. The selection equation is deterministic, in the sense that there is a strict threshold for censoring, so the variance of the disturbance term is actually zero, and the predicted probabilities are either zeroes and ones. This is usually a very tough case for the probit model, as estimates tend to diverge to infinity. Also, the disturbances of the two equations were reported to be perfectly correlated, which reflects the fact that y was used in the selection process. In fact, achieving convergence took only 3 Newton-Raphson steps for the interval regression, and 21 such steps for Heckman model. It should also be taken into account that Heckman model has 6 parameters, two regression slopes, two intercepts, the variance of the regression disturbance, and the correlation of the disturbances, while the interval regression model has only three, the slope and the

	Mean	S.d.	Min	Max
<i>Full sample OLS</i>				
$\hat{\beta}_x$	0.000273	0.0319	-0.0982	0.1026
S.e. $\hat{\beta}_x$	0.0316	0.00102	0.0282	0.0346
R^2	0.00101	0.00139	$2.06 \cdot 10^{-10}$	0.0101
% H_0 rejected	5.1%			
<i>Censored sample OLS</i>				
$\hat{\beta}_x$	0.104	0.0308	0.0155	0.197
S.e. $\hat{\beta}_x$	0.0323	0.00105	0.0290	0.0353
p -value	0.0197	0.0629	$1.11 \cdot 10^{-9}$	0.627
R^2	0.0118	0.00649	0.000252	0.0384
% H_0 rejected	91.2%			
<i>Interval regression</i>				
$\hat{\beta}_x$	$-0.147 \cdot 10^{-3}$	0.0326	-0.103	0.103
S.e. $\hat{\beta}_x$	0.0324	0.00104	0.0288	0.0352
Efficiency				
loss	4.86%	1.86	-1.57%	11.9%
p -value	0.500	0.292	0.00146	0.9992
% H_0 rejected	4.5%			

Percentage H_0 rejected is for 5% significance level. Efficiency loss is the squared ratio of standard errors of the slope coefficients in the full sample OLS and interval regression models.

Table 2: Comparison of different estimation methods: 1000 simulations.

intercept, along with the nuisance variance parameter. Still the difference in the number of iterations is too large to be accounted for by the different dimensionality of the parameter space solely.

The above results were obtained in just one simulation. When the process is repeated a number of times, a better picture of the statistical properties of the estimation procedures can be obtained. 1000 samples were drawn, and the three regressions (except more time-consuming Heckman model) were estimated for each of them. The results are reported in Table 2¹.

Clearly, the OLS based on the censored sample produces biased coefficient estimates. The bias is strong enough to break down the inference: instead of 5% rejections, a whopping 91.2% is observed. Both full sample OLS and the interval regression are producing unbiased estimates of the slope with

¹All simulations have been performed in Stata 7. The code is available from the author upon request. Suggestions as to what else can be studied without making the setup too complex are welcome.

the correct size of the two-sided tests. The distribution of the p -values in the case of interval regression is almost perfectly close to the uniform distribution. Distribution free Kolmogorov-Smirnov test gives the distance between the two distribution functions of 0.0005 with p -value reported as 1.000.

The reported standard errors are OK, although the regression for the censored sample seems to overstate the standard errors a little bit (compare the variability in $\hat{\beta}_x$ as evidenced by the (empirical) standard deviation of 0.308, and the average reported standard error of 0.323). Also worth noting is “improvement” in fit (R^2) for the censored sample OLS. The efficiency loss measured as the ratio of the estimated variances of $\hat{\beta}_x$ for the interval regression and the full sample OLS is quite close to 5% which is the average amount of missing data.

Thus the simulations confirm that even small amount of missing data — and 5% would certainly be judged as rather lucky case of almost all data available — can lead to serious biases in both point estimates and hypothesis testing when the data are systematically missing. The likelihood based procedures do improve the situation when the mechanism that produces the missing data is known, or can be guessed with reasonable accuracy.