The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
13, Number 4, pp. 776–794

Implementation of a double-hurdle model

Bruno García
The College of William and Mary
Williamsburg, VA
bsgarcia@email.wm.edu

Abstract. Corner solution responses are frequently observed in the social sciences. One common approach to model phenomena that give rise to corner solution responses is to use the tobit model. If the decision to participate in the market is decoupled from the consumption amount decision, then the tobit model is inappropriate. In these cases, the double-hurdle model presented in Cragg (1971, *Econometrica* 39: 829–844) is an appropriate alternative to the tobit model. In this article, I introduce a command, `dblhurdle`, that fits the double-hurdle model. The implementation allows the errors of the participation decision and the amount decision to be correlated. The capabilities of `predict` after `dblhurdle` are also discussed.

Keywords: st0317, dblhurdle, tobit, Heckman, Cragg, double hurdle, hurdle

1 Introduction

Double-hurdle models are used with dependent variables that take on the endpoints of an interval with positive probability and that are continuously distributed over the interior of the interval. For example, you observe the amount of alcohol individuals consume over a fixed period of time. The distribution of the amounts will be roughly continuous over positive values, but there will be a "pile up" at zero, which is the corner solution to the consumption problem the individuals face; no individual can consume a negative amount of alcohol.

One common approach to modeling such situations is to use the tobit model. Suppose the dependent variable y is continuous over positive values, but $\Pr(y = 0) > 0$ and $\Pr(y < 0) = 0$. Letting $\Phi(\,)$ denote a standard normal cumulative distribution function (CDF) and $\phi(\,)$ denote a standard normal density function, recall that the log-likelihood function for the tobit model is

$$\log(L) = \sum_{y_i=0} \left[\log \left\{ 1 - \Phi \left(\frac{x_i\beta}{\sigma} \right) \right\} \right] + \sum_{y_i>0} \left[\log \left\{ \phi \left(\frac{y_i - x_i\beta}{\sigma} \right) \right\} - \log(\sigma) \right]$$

The functional form of the tobit model imposes a restriction on the underlying stochastic process: $x_i\beta$ parameterizes both the conditional probability that $y_i = 0$ and the conditional density associated with the magnitude of y_i whenever $y_i > 0$. Thus the tobit model cannot properly handle the situation where the effect of a covariate on the probability of participation $\Pr(y_i > 0)$ and the effect of the same covariate on the amount of participation have different signs. For example, it might be the case that

attending AA meetings lowers the probability of engaging in the consumption of alcohol, but if alcohol is consumed, a high quantity of consumption is likely because of binge drinking. A similar situation can be seen in the work of Martínez-Espiñeira (2006), who examined a survey asking respondents to state a reasonable tax amount to protect coyotes by compensating farmers for livestock losses. Martínez-Espiñeira (2006) finds that "respondents who hunt stated support for significantly lower levels of tax than nonhunters. However, hunters are less likely to state a zero amount of tax".

2 The double-hurdle model

The consumer-choice example described below provides intuition about the structure in the double-hurdle model. The model is not limited to problems in this context and can also be applied in epidemiology and other applied biostatistical fields.

Suppose individuals make their consumption decisions in two steps. First, the individual determines whether he or she wants to participate in the market. This is called the participation decision. Then the individual determines an optimal consumption amount (which may be 0) given his or her circumstances. This is called the quantity decision. If y_i represents the observed consumption amount of the individual, we can model it as

$$y_i = \begin{cases} x_i\beta + \epsilon_i & \text{if } \min(x_i\beta + \epsilon_i, z_i\gamma + u_i) > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{pmatrix} \epsilon_i \\ u_i \end{pmatrix} \sim N\left(0, \Sigma\right), \Sigma = \begin{pmatrix} 1 & \sigma_{12} \\ \sigma_{12} & \sigma \end{pmatrix}$$

Letting $\Psi\left(x, y, \rho\right)$ denote the CDF of a bivariate normal with correlation ρ, the log-likelihood function for the double-hurdle model is

$$\begin{aligned} \log(L) = &\sum_{y_i=0} \left[\log\left\{ 1 - \Phi\left(z_i\gamma, \frac{x_i\beta}{\sigma}, \rho \right) \right\} \right] \\ &+ \sum_{y_i>0} \left(\log\left[\Phi\left\{ \frac{z_i\gamma + \frac{\rho}{\sigma}\left(y_i - x_i\beta\right)}{\sqrt{1-\rho^2}} \right\} \right] - \log\left[\sigma\right] + \log\left\{ \phi\left(\frac{y_i - x_i\beta}{\sigma} \right) \right\} \right) \end{aligned}$$

The double-hurdle model can be reduced to the tobit model by setting $\rho = 0$ and taking the limit $z_i\gamma \to +\infty$.

3 The dblhurdle command

The `dblhurdle` command implements the double-hurdle model, where the error terms of the participation equation and the quantity equation are jointly normal and may be correlated. Letting $x_i\beta + \epsilon_i$ model the quantity equation and $z_i\gamma + u_i$ model the participation equation, the command estimates β, γ, ρ, and σ, where $\sigma = \text{Var}\left(\epsilon\right)$. We restrict $\text{Var}\left(u\right)$ to equal 1; otherwise, the model is not identified.

3.1 Syntax

`dblhurdle` *depvar* [*indepvars*] [*if*] [*in*] [*weight*], {`ll(`#`)`|`ul(`#`)`}
[`peq(`*varlist*`,` [`noconstant`]`)` `ptobit` `noconstant` `constraints(`*numlist*`)`
`vce(`*vcetype*`)` `level(`#`)` `correlation` *display_options* *maximize_options*]

indepvars and `peq()` may contain factor variables; see [U] **11.4.3 Factor variables**.

3.2 Options

`ll(`#`)` indicates a lower corner. Observations with *depvar* $\leq$ # are considered at the corner. One of `ul(`#`)` or `ll(`#`)` must be specified.

`ul(`#`)` indicates an upper corner. Observations with *depvar* $\geq$ # are considered at the corner. One of `ul(`#`)` or `ll(`#`)` must be specified.

`peq(`*varlist*`,` [`noconstant`]`)` specifies the set of regressors for the participation equation if these are different from those of the quantity equation.

`ptobit` specifies that the participation equation should consist of a constant only. This option cannot be specified with the `peq()` option.

`noconstant`; see [R] **estimation options**.

`constraints(`*numlist*`)` is used to specify any constraints the researcher may want to impose on the model.

`vce(`*vcetype*`)` specifies the type of standard error reported. *vcetype* may be `oim` (default), `robust`, or `cluster` *clustvar*.

`level(`#`)`; see [R] **estimation options**.

`correlation` displays the correlation between the error terms of the quantity equation and the participation equation. The covariance is not shown when this option is specified.

display_options; see *Reporting* under [R] **estimation options**.

maximize_options: `technique(`*algorithm_spec*`)`, `iterate(`#`)`, [`no`]`log`, `tolerance(`#`)`, `ltolerance(`#`)`, `nrtolerance(`#`)`, and `from(`*init_specs*`)`; see [R] **maximize**. These options are seldom used.

3.3 Stored results

`dblhurdle` stores the following in `e()`:

Scalars

`e(N)`	number of observations	`e(ulopt)`	contents of `ul()`
`e(ll)`	log likelihood	`e(llopt)`	contents of `ll()`
`e(converged)`	`1` if converged, `0` otherwise		

Macros

`e(cmd)`	`dblhurdle`	`e(predict)`	program used to implement `predict`
`e(cmdline)`	command as typed		
`e(depvar)`	name of dependent variable	`e(marginsok)`	predictions allowed by `margins`
`e(title)`	title in estimation output	`e(qvars)`	variables in quantity equation
`e(vce)`	*vcetype* specified in `vce()`	`e(pvars)`	variables in participation equation
`e(properties)`	`b V`		

Matrices

`e(b)`	coefficient vector	`e(V)`	variance–covariance matrix of the estimators
`e(Cns)`	constraints matrix		

Functions

`e(sample)`	marks estimation sample

4 Postestimation: predict

4.1 Syntax

`predict` [*type*] *newvarname* [*if*] [*in*] [`, xb zb xbstdp zbstdp ppar ycond yexpected stepnum(#)`]

4.2 Options

`xb` calculates the linear prediction for the quantity equation. This is the default option when no options are specified in addition to `stepnum()`.

`zb` calculates the linear prediction for the participation equation.

`xbstdp` calculates the standard error of the linear prediction of the quantity equation, `xb`.

`zbstdp` calculates the standard error of the linear prediction of the participation equation, `zb`.

`ppar` is the probability of being away from the corner conditional on the covariates.

`ycond` is the expectation of the dependent variable conditional on the covariates and on the dependent variable being away from the corner.

`yexpected` is the expectation of the dependent variable conditional on the covariates.

`stepnum(#)` controls the number of steps to be taken for predictions that require integration (`yexpected` and `ycond`). More specifically, # will be the number of steps taken per unit of the smallest standard deviations of the normal distributions used

in the prediction. The default is `stepnum(10)`. You can fine-tune the value of this parameter by trial and error until increasing the parameter results in no or little change in the predicted value.

5 Example

We illustrate the use of the `dblhurdle` command using `smoke.dta` from Wooldridge (2010).[1]

We begin our example by describing the dataset:

```
. use smoke
. describe
Contains data from smoke.dta
  obs:           807
 vars:            10                          15 Aug 2012 19:00
 size:        19,368
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
educ            float   %9.0g                 years of schooling
cigpric         float   %9.0g                 state cig. price, cents/pack
white           byte    %8.0g                 =1 if white
age             byte    %8.0g                 in years
income          int     %8.0g                 annual income, $
cigs            byte    %8.0g                 cigs. smoked per day
restaurn        byte    %8.0g                 =1 if rest. smk. restrictions
lincome         float   %9.0g                 log(income)
agesq           float   %9.0g                 age^2
lcigpric        float   %9.0g                 log(cigprice)
-------------------------------------------------------------------------------
Sorted by:
. misstable summarize
(variables nonmissing or string)
```

We will model the number of cigarettes smoked per day, so the dependent variable will be `cigs`. The explanatory variables we use are `educ` (number of years of schooling); the log of income; the log of the price of cigarettes in the individual's state; `restaurn`, which takes the value 1 if the individual's state has restrictions against smoking in restaurants and 0 otherwise; and we include the individual's age and the age squared. Not all variables will be included in both equations.

The fact that `cigs` (the dependent variable) is a byte should remind us that we are implicitly relaxing an assumption of the double-hurdle model. The hypothesized data-generating process generates values over a continuous range of values, but all the observed number of cigarettes are integers.

1. The data were downloaded from http://fmwww.bc.edu/ec-p/data/wooldridge/smoke.dta, and the variables were labeled according to http://fmwww.bc.edu/ec-p/data/wooldridge/smoke.des.

It is always good to check for any missing values; because we have no string variables, the output of `misstable summarize` ensures that there are no missing values.

The dependent variable should have a "corner" at zero because all nonsmokers will report smoking zero cigarettes per day. We verify this point by tabulating the dependent variable. This simple check is important because it might be the case that our data contain only smokers with positive entries in the variable `cigs`, in which case a truncated regression model would be more appropriate. We perform the simple check:

```
. tabulate cigs
```

cigs. smoked per day	Freq.	Percent	Cum.
0	497	61.59	61.59
1	7	0.87	62.45
2	5	0.62	63.07
3	5	0.62	63.69
4	2	0.25	63.94
5	7	0.87	64.81
6	3	0.37	65.18
7	2	0.25	65.43
8	3	0.37	65.80
9	2	0.25	66.05
10	28	3.47	69.52
11	2	0.25	69.76
12	4	0.50	70.26
13	2	0.25	70.51
14	1	0.12	70.63
15	23	2.85	73.48
16	1	0.12	73.61
18	3	0.37	73.98
19	1	0.12	74.10
20	101	12.52	86.62
25	7	0.87	87.48
28	3	0.37	87.86
30	42	5.20	93.06
33	1	0.12	93.18
35	2	0.25	93.43
40	37	4.58	98.02
50	6	0.74	98.76
55	1	0.12	98.88
60	8	0.99	99.88
80	1	0.12	100.00
Total	807	100.00	

The tabulation of the dependent variable reveals that about 60% of the individuals in the sample smoked 0 cigarettes. Strangely, we also see that individuals seem to smoke cigarettes in multiples of five—in part, this may be due to a reporting heuristic used by individuals.

We estimate the parameters of a double-hurdle model by typing

```
. dblhurdle cigs educ restaurn lincome lcigpric, peq(educ c.age##c.age) ll(0)
> nolog
Double-Hurdle regression                        Number of obs   =        807
```

cigs	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
cigs						
educ	4.373058	.8969167	4.88	0.000	2.615134	6.130983
restaurn	-6.629484	2.630784	-2.52	0.012	-11.78573	-1.473241
lincome	3.236915	1.534674	2.11	0.035	.2290102	6.24482
lcigpric	-2.376598	12.02945	-0.20	0.843	-25.95388	21.20068
_cons	-44.41139	50.5775	-0.88	0.380	-143.5415	54.71869
peq						
educ	-.2053851	.0324439	-6.33	0.000	-.2689739	-.1417963
age	.0867284	.015593	5.56	0.000	.0561666	.1172901
c.age#c.age	-.0010174	.0001755	-5.80	0.000	-.0013615	-.0006734
_cons	1.093345	.4821582	2.27	0.023	.1483324	2.038358
/sigma	24.58939	2.904478			18.89671	30.28206
/covariance	-20.70667	3.881986	-5.33	0.000	-28.31523	-13.09812

The command showcases some of the features implemented. We used factor variables to include both age and age squared.

The command displays the number of observations in the sample. It lacks a test against a benchmark model. Most estimation commands implement a test against a benchmark constant-only model. For the double-hurdle model, the choice of model to test against has been left to the user. This test can be carried out with standard postestimation tools.

The estimation table shows results for four equations. In the econometric sense, we estimated the parameters from two equations and two dependence parameters. The first equation displays the coefficients of the quantity equation, which is titled `cigs` after the dependent variable. The second equation, titled `peq`, which is short for participation equation, displays the coefficients of the participation equation. The third equation, titled `/sigma`, displays the estimated value of the standard deviation of the error term of the quantity equation. As mentioned, the analogous parameter of the participation equation is set to 1; otherwise, the model is not identified. The fourth equation, titled `/covariance`, displays the estimated value of the covariance between the error terms of the quantity equation and the participation equation. If the `correlation` option is specified, the correlation is displayed instead, and the equation title changes to `/rho`.

The results allow us to appreciate the strengths of the double-hurdle model. For example, the coefficient of `educ` has a positive value on the quantity equation, while the analogous coefficient in the participation equation has a negative value. This implies that more educated individuals will be less likely to smoke, but if they smoke, they will tend to smoke more than less educated individuals.

So a small increment in the number of years of schooling will positively affect the number of daily cigarettes smoked given that an individual is a smoker but negatively affect the probability that the individual is a smoker. Naturally, we may want to know which effect, if any, dominates. For nonlinear problems like this one, which effect dominates depends on the other characteristics of the individual. In these situations, researchers often calculate marginal effects. In our example, we illustrate how to compute the average marginal effect of the number of years of schooling (`educ`) on three different quantities of interest:

- The probability of smoking
- The expected number of cigarettes smoked given that you smoke
- The expected number of cigarettes smoked

Given the signs of the coefficients, we know that the average marginal effect of `educ` on the probability of smoking will be negative. We also expect the average marginal effect of `educ` on the number of cigarettes smoked given that you are a smoker will be positive. The final quantity, the marginal effect of `educ` on the number of cigarettes smoked regardless of smoker status, is ambiguous.

To estimate these quantities, we use the `predict()` option in conjunction with the `margins` command. First, we calculate the average marginal effect of `educ` on the probability that the individual is a smoker by using the `ppar` option:

```
. margins, dydx(educ) predict(ppar)
Average marginal effects                        Number of obs   =        807
Model VCE    : OIM
Expression   : predict(ppar)
dy/dx w.r.t. : educ
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	-.0348973	.0052745	-6.62	0.000	-.0452352	-.0245595

Note that the effect is negative, as expected, and significant.

Next we compute the average marginal effect of education on the number of cigarettes smoked given that the individual is a smoker by using the `ycond` option. We will carry on this computation twice to illustrate the use of the `stepnum()` option.

```
. set r on
r; t=0.00 11:43:17
. margins, dydx(educ) predict(ycond)
Average marginal effects                          Number of obs   =        807
Model VCE    : OIM
Expression   : predict(ycond)
dy/dx w.r.t. : educ
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.691684	.2795245	2.47	0.013	.143826	1.239542

```
r; t=120.64 11:45:17
. margins, dydx(educ) predict(ycond stepnum(100))
Average marginal effects                          Number of obs   =        807
Model VCE    : OIM
Expression   : predict(ycond stepnum(100))
dy/dx w.r.t. : educ
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	.6916836	.2795244	2.47	0.013	.1438259	1.239541

```
r; t=1153.86 12:04:31
. set r off
```

First, we calculate the average marginal effect with the default value of `stepnum()`, which is 10. We note that the effect is positive, as expected, and significant. We also note that when the calculation is repeated with a `stepnum()` of 100, we observe a change in the sixth decimal point, which in this context is meaningless, but it comes at the expense of a tenfold increase in run time. Hence, `stepnum()` should be used with caution. My advice is to tune it by using `predict` with the `ycond` option until the predicted values show little or no sensitivity to positive changes in `stepnum()`.

Finally, we use the `yexpected` option of `predict` in `margins` to calculate the average marginal effect `educ` has on the number of cigarettes smoked per day regardless of the individual's smoker status:

```
. margins, dydx(educ) predict(yexpected)
Average marginal effects                          Number of obs   =        807
Model VCE    : OIM
Expression   : predict(yexpected)
dy/dx w.r.t. : educ
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
educ	-.5487611	.1473763	-3.72	0.000	-.8376132	-.2599089

We note that the effect is negative and that it is statistically significant. Hence, on average, a higher education will lower the expected number of cigarettes an individual smokes.

6 Monte Carlo simulation

This section describes some Monte Carlo simulations used to investigate the finite-sample properties of the estimator. Point estimates of the parameters should be close to their true values, and the rejection rate of the true null hypothesis should be close to the nominal size of the test.

To this end, we perform a Monte Carlo simulation, and we look at three measures of performance:

- The mean of the estimated parameters should be close to their true values.
- The mean standard error of the estimated parameters over the repetitions should be close to the standard deviation of the point estimates.
- The rejection rate of hypothesis tests should be close to the nominal size of the test.

The first step consists of choosing the parameters of the model. The quantity equation was chosen to have one continuous covariate, one indicator variable, and an intercept. The variance of the error associated with this equation is equal to 1. The participation equation consists of a different continuous variable, indicator variable, and intercept. The error terms will be drawn so that they are independent. Thus the correlation between the error terms will be 0. We set an upper corner at 0. The data-generating process can be summarized as follows:

$$y = \begin{cases} \min(0, 2x_1 - d_1 + 0.5 + \epsilon) & \text{if } x_2 - 2d_2 + 1 + u < 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{pmatrix} \epsilon \\ u \end{pmatrix} \sim N\left(0, \Sigma\right), \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

A dataset of 2,000 observations was created containing the covariates. The x's were drawn from a standard normal distribution, and the d's were drawn from a Bernoulli with $p = 1/2$. In the pseudocode below, we refer to this dataset as "base".

Now we describe an iteration of the simulation:

1. Use "base".
2. For each observation, draw (`gen`) ϵ from a standard normal.
3. For each observation, draw (`gen`) u from a standard normal.

4. For each observation, compute y according to the data-generating process presented above.

5. Fit the model, and save the values of interest with `post`.

The values of interest during each iteration are the point estimates of the parameters; the standard errors of the parameters; and, for each parameter, whether the 95% confidence interval around the estimated parameter excluded the true value of the parameter. At the conclusion of the simulation, we have a dataset of 10,000 observations, where each observation is a realization of the values of interest.

The following table summarizes the results. It shows the mean estimated coefficient, or "mean"; the standard deviation of the sample of estimated coefficients, or "std. dev."; the mean estimated standard error, or "std. err."; and the proportion of the time a test of size 0.05 rejected the true null hypothesis, denoted by "rej. rate".

Table 1. Results of the simulation

parameter	true value	mean	std. dev	std. err	rej. rate
β_{x_1}	2	2.0007	0.0563	0.0561	0.0524
β_{d_1}	−1	−1.0001	0.0860	0.0856	0.0507
β_{cons_1}	0.5	0.5007	0.0881	0.0885	0.0497
γ_{x_2}	1	1.0095	0.0823	0.0811	0.0520
γ_{d_2}	−2	−2.0156	0.1424	0.1426	0.0486
γ_{cons_2}	1	1.0068	0.0862	0.0863	0.0507
sigma	1	0.9979	0.0364	0.0364	0.0542
covariance	0	0.0016	0.1046	0.1036	0.0532

The results show that the statistical properties of the estimates are as desired. Other simulations were done to see how these results would change under extreme circumstances, such as correlations close to the extremes of −1 or 1. The results were qualitatively similar to those above for correlations as high as 0.95 and as low as −0.95. There were instances where the tests did not achieve their nominal size. Rather than being driven by the extreme values of the input parameters, these issues seem to be driven primarily by the proportion of observations at the corner. As this proportion gets close to either extreme (0 or 1), the nominal size of a test of the covariance deviates from the true size. This becomes an issue once the proportion of observations at the corner is above 95% or below 5%.

The other parameters can also be affected by this, but for those parameters, this is more intuitive because it can be viewed through the lens of a small-sample problem. For example, if most of your observations are at the corner, you will have very little data to estimate the parameters associated with the quantity equation. Because the confidence intervals produced by maximum likelihood are normal only asymptotically, we cannot expect them to achieve their nominal size on small samples.

Figure 1 summarizes this information. Each scatterplot contains the observed rejection rate of a test of nominal size 0.05 on the vertical axis, and the proportion of observations at the corner on the horizontal axis. Each point on the scatterplot represents a variation on the parameterization of the data-generating process presented above. I held the coefficients of the quantity and participation equations fixed, and I tried every combination of upper or lower corner; corner at -2, 0, 2; $\sigma \in \{0.2, 1, 10\}$; and $\rho \in \{-0.95, 0, 0.95\}$.

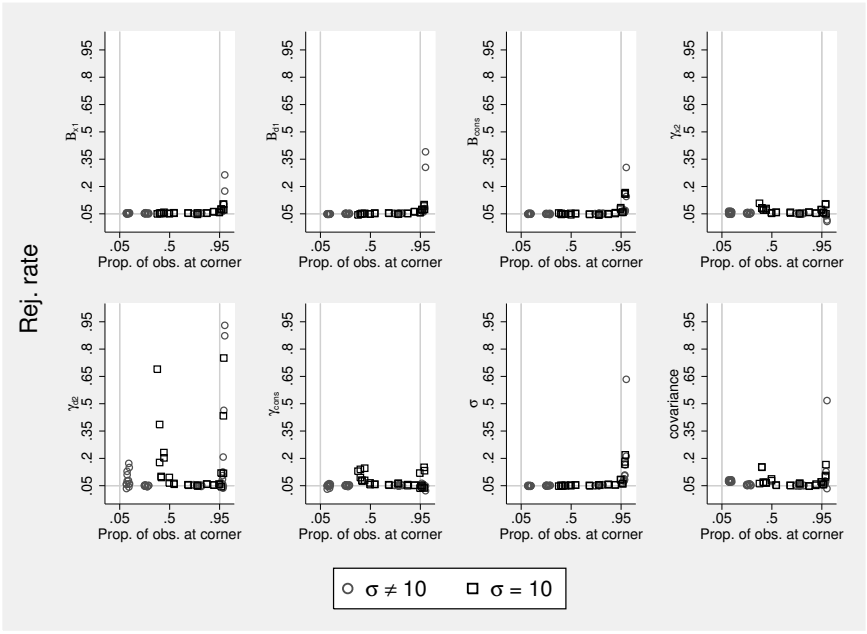

Figure 1. Scatterplots showing rejection rate of a test of nominal size 0.05 and proportion of observations at the corner

Of these, only the parameterization where $\sigma = 10$ seems to induce a discrepancy between the nominal size of the test and the attained size of the test, particularly for γ_{d_2}. Hence, I decided to mark those points with a square instead of a circle.

Notice that the nominal size is almost never achieved once you cross the 0.95 proportion (marked with a vertical line). Also notice that tests involving the gamma coefficients (those of the participation equation) also deviate from their nominal size (albeit less markedly) when the proportion of censored observations is low. This is most obvious for the γ_{d_2} coefficient.

A less intuitive issue occurs when the set of regressors in the participation equation is equal to the set of regressors of the quantity equation. In this case, the model is weakly identified, and the nominal sizes will differ from the true size of the test. To illustrate, we attempt to recover the parameters of the following data-generating process:

$$y = \begin{cases} \min(0, 2x_1 - d_1 + 0.5 + \epsilon) & \text{if } 2x_1 - d_1 + 0.5 + u < 0 \\ 0 & \text{otherwise} \end{cases}$$
$$\begin{pmatrix} \epsilon \\ u \end{pmatrix} \sim N(0, \Sigma), \Sigma = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

The results, summarized in the following table, suggest that the point estimates can be trusted but that the size of the tests may deviate from the advertised values.

Table 2. Results of the data-generating process

parameter	true value	mean	std. dev	std. err	rej. rate
β_{x_1}	2	2.0043	0.0907	0.0877	0.0711
β_{d_1}	−1	−1.0029	0.0925	0.0925	0.0535
β_{cons_1}	0.5	0.5077	0.1618	0.1569	0.0762
γ_{x_1}	2	2.0625	0.2898	0.2699	0.0846
γ_{d_1}	−1	−1.0270	0.2216	0.2114	0.0560
γ_{cons_1}	0.5	0.5417	0.2548	0.2447	0.0777
sigma	1	1.0009	0.0331	0.0328	0.0534
covariance	0	0.0374	0.2754	0.2541	0.1118

Figure 2 is analogous to figure 1. We note that tests on the covariance are particularly unreliable, that the distinction between the cases where $\sigma \neq 10$ and $\sigma = 10$ seems not to matter, and that the rejection rate exceeds the nominal size of the test when the proportion of observations at the corner is around 0.9. However, when the proportion of observations at the corner is between 0.3 and 0.8, the sizes are mostly reliable with the notable exception of tests of the covariance.

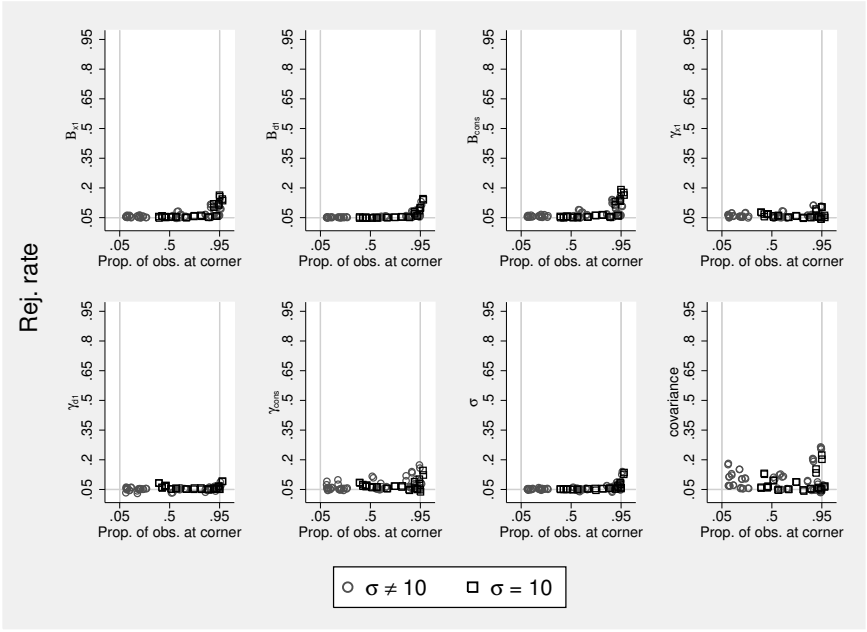

Figure 2. Scatterplots showing rejection rate of a test of nominal size 0.05 and proportion of observations at the corner

7 Methods and formulas

7.1 Log likelihood

In Cragg (1971), a variety of double-hurdle models were first proposed. Jones (1992) applies the double-hurdle model with correlation in the error terms to data on tobacco expenditures. Letting $\Phi(\bullet)$ denote the standard normal CDF, $\phi(\bullet)$ denote the standard normal density function, and $\Psi(x, y, \rho)$ denote the CDF of the bivariate normal with correlation ρ, the log-likelihood function for the double-hurdle model with a lower corner at c is

$$a = \frac{\left\{z\gamma - c + \frac{\rho}{\sigma}(y - x\beta)\right\}}{\sqrt{1-\rho^2}}, b = \frac{x\beta - c}{\sigma}$$

$$\log(L) = \sum_{y_i = c} \left[\log\left\{1 - \Psi(z\gamma - c, b, \rho)\right\}\right] + \sum_{y_i > c} \left[\log\left\{\Phi(a)\right\} - \log(\sigma) + \log\left\{\phi\left(\frac{y - x\beta}{\sigma}\right)\right\}\right]$$

If the upper corner is at c, then

$$\log(L) = \sum_{y_i=c} \left[\log\left\{1 - \Psi\left(c - z\gamma, -b, \rho\right)\right\}\right] + \sum_{y_i<c} \left[\log\left\{\Phi\left(-a\right)\right\} - \log\left(\sigma\right) + \log\left\{\phi\left(\frac{x\beta - y}{\sigma}\right)\right\}\right]$$

7.2 Choosing the initial point

The optimization routine `optimize()` requires an initial point from which to initialize the optimization algorithm. My choice of starting point is $[x\beta, z\gamma, 0, 5]'$, where β are the ordinary least-squares estimates of a regression of the dependent variable of the model on x, the variables in the quantity equation; γ are the ordinary least-squares estimates of a regression of the dependent variable of the model on z, the variables in the quantity equation; and ρ and σ are chosen to be 0 and 5, respectively.

There is no guarantee that the initial point will be feasible. If the initial point is infeasible, the use of the `from()` option is recommended.

7.3 First derivatives

The first derivatives of the log likelihood (if β_j is the constant, simply let $x_j = 1$ and likewise for γ_j) are given below. These were adapted from Jones and Yen (2000). Letting $\psi\left(x, y, \rho\right)$ be the density of a bivariate normal with correlation ρ,

$$\Psi = \Psi\left(z\gamma - c, b, \rho\right)$$

$$\psi = \Psi\left(z\gamma - c, b, \rho\right)$$

$$\Phi_{12} = \Phi\left(\frac{z\gamma - c - b\rho}{\sqrt{1-\rho^2}}\right)$$

$$\phi_{12} = \Phi\left(\frac{z\gamma - c - b\rho}{\sqrt{1-\rho^2}}\right)$$

$$\Phi_{21} = \Phi\left(\frac{b - \rho\left(z\gamma - c\right)}{\sqrt{1-\rho^2}}\right)$$

$$\frac{d\beta_j}{d\log\left(L\right)} = \frac{x_j}{\sigma}\left[\sum_{y=c}\left\{\frac{\phi\left(b\right)\Phi_{12}}{\Psi - 1}\right\} + \sum_{y>c}\left\{\frac{-\rho\phi\left(a\right)}{\sqrt{1-\rho^2}\Phi\left(a\right)} + \frac{y - x\beta}{\sigma}\right\}\right]$$

$$\frac{d\gamma_j}{d\log\left(L\right)} = z_j\left[\sum_{y=c}\left\{\frac{\phi\left(z-c\right)\Phi_{21}}{\Psi - 1}\right\} + \sum_{y>c}\left\{\frac{\phi\left(a\right)}{\sqrt{1-\rho^2}\Phi\left(a\right)}\right\}\right]$$

$$\frac{d\sigma_{12}}{d\log\left(L\right)} = \frac{1}{\sigma}\left(\sum_{y=c}\left(\frac{\psi}{\Psi - 1}\right) + \sum_{y>x}\left[\frac{y - x\beta}{\sigma}\left\{\frac{\phi\left(a\right)}{\Phi\left(a\right)\sqrt{1-\rho^2}}\right\} + \frac{a\rho}{\sqrt{1-\rho^2}}\right]\right)$$

$$\begin{aligned}\frac{d\sigma}{d\log\left(L\right)} = {} & \frac{1}{\sigma}\sum_{y=c}\left[b\left\{\frac{\Phi_{12}\phi\left(b\right)}{1-\Psi}\right\} + \frac{\rho\psi}{1-\Psi}\right] \\ & + \frac{1}{\sigma}\sum_{y>c}\left[\left(\frac{y - x\beta}{\sigma}\right)^2 - 1 + \left\{\frac{-\rho\phi\left(a\right)}{\Phi\left(a\right)\sqrt{1-\rho^2}}\right\}\left(2\frac{y - x\beta}{\sigma} + \frac{a\rho}{\sqrt{1-\rho^2}}\right)\right]\end{aligned}$$

The implementation of the derivatives for an upper corner at c requires a few minor changes. First, the derivatives with respect to β_j and γ_j should be multiplied by -1. Finally, multiply a, b, $z\gamma - c$, and $(y - x\beta)/\sigma$ by -1.

7.4 Weights

The weighting schemes implemented for `dblhurdle` are frequency weights (`fw`), sampling weights (`pw`), and importance weights (`iw`). Recall that the likelihood function is summed over observations. To implement the weights, you need to multiply the ith term of the summation over observations by the weight of the ith observation. The frequency weights are only allowed to be positive integers.

When frequency weights are specified, the sample size is adjusted so that it is equal to the sum of the weights. The importance weights are allowed to be any real number. No sample-size adjustments are made when importance weights are specified. The sampling weights are like the importance weights, but a `robust` estimator of the variance is computed instead of the default `oim` estimator. No sample-size adjustment is made when sampling weights are specified, and the weights are not allowed to be negative.

Finally, analytic weights (`aw`) are not allowed. This command was written with the `tobit` command in mind. In that case, the `aweights` (normalized) divide the variance of the error term. In the case of `dblhurdle`, the rationale for dividing the variance by the normalized weights does not carry over well because we also have to estimate the covariance between the error terms.

7.5 Prediction

There are three options in the prediction program that require some explanation. The `ppar` option computes the probability of being away from the corner conditional on the covariates. Thus this option computes

$$\Pr\left(y > c|x,z\right) = \Phi\left(z\gamma - c, \frac{x\beta - c}{\sigma}, \rho\right)$$

The option `ycond` computes the following expectation:

$$E\left(y|x,z,y>c\right) = \int_{c}^{+\infty} y f(y|u > c - z\gamma, \epsilon > c - x\beta)dy$$

$$f(y|u > c - z\gamma, \epsilon > c - x\beta) = \frac{\phi\left(\frac{y-x\beta}{\sigma}\right)\Phi\left\{\frac{z\gamma - c + \frac{\rho}{\sigma}(y-x\beta)}{\sqrt{1-\rho^2}}\right\}}{\sigma\Phi\left(z\gamma - c, \frac{x\beta - c}{\sigma}, \rho\right)}$$

Finally, the option `yexpected` computes the expected value of y conditional on x and z:

$$E\left(y|x,z\right) = c\left\{1 - \Pr\left(y > c|x,z\right)\right\} + \Pr\left(y > c|x,z\right) E\left(y|x,z,y>c\right)$$

Note that the options that involve integration are time consuming. Thus the option `stepnum()` was added to the prediction program to allow the user some control of the execution time for the integration. Letting n_s denote the `stepnum()`, the step size is chosen to be

$$\frac{\min\left(\sigma, \sqrt{1-\rho^2}\right)}{n_s}$$

Execution is faster when the `stepnum()` is smaller, but the improved run time comes at a cost to accuracy. The default is `stepnum(10)`.

When the corner is above, the expressions that change become

$$\Pr(y < c|x,z) = \Phi\left(c - z\gamma, \frac{c - x\beta}{\sigma}, \rho\right)$$

$$E(y|x,z,y<c) = \int_{-\infty}^{c} y f(y|u < z\gamma - c, \epsilon < x\beta - c)dy$$

$$f(y|u < z\gamma - c, \epsilon < x\beta - c) = \frac{\phi\left(\frac{x\beta - y}{\sigma}\right)\Phi\left\{\frac{c - z\gamma - \frac{\rho}{\sigma}(y - x\beta)}{\sqrt{1-\rho^2}}\right\}}{\sigma\Phi\left(c - z\gamma, \frac{c - x\beta}{\sigma}, \rho\right)}$$

8 Conclusion

The double-hurdle model was an important contribution to the econometric toolkit used by researchers. I hope that readers will consider this model and, in particular, the `dblhurdle` command when their first instinct is to use the tobit model. The example presented in section 5 illustrates the flexibility of the model. It allows the researcher to break down the modeled quantity along two useful dimensions, the "quantity" dimension and the "participation" dimension.

The command presented in this article only allows for a single corner in the data. One desirable feature to add is the capability to handle dependent variables with two corners. Such variables are common (for example, 401k contributions), so this feature would certainly provide higher value to users.

9 Acknowledgments

I wrote this article and the command described therein during a summer internship at StataCorp. It was exciting to meet the individuals behind Stata. I thank David Drukker for his support and for the time he spent going over the intricate details of the models. I also thank Rafal Raciborski for all of his comments, suggestions, and tips. Any errors in my work are my own.

10 References

Cragg, J. G. 1971. Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica* 39: 829–844.

Jones, A. M. 1992. A note on computation of the double-hurdle model with dependence with an application to tobacco expenditure. *Bulletin of Economic Research* 44: 67–74.

Jones, A. M., and S. T. Yen. 2000. A Box-Cox double-hurdle model. *Manchester School* 68: 203–221.

Martínez-Espiñeira, R. 2006. A Box-Cox Double-Hurdle model of wildlife valuation: The citizen's perspective. *Ecological Economics* 58: 192–208.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, MA: MIT Press.

About the author

Bruno García is working toward a master's degree in computational operations research at the College of William and Mary. He received his bachelor's degree in applied mathematics and economics from Brown University.