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	$115	1-year subscription	$145
2-year subscription	$210	2-year subscription	$270
3-year subscription	$285	3-year subscription	$375
1-year student subscription	$ 85	1-year student subscription	$115
1-year institutional subscription	$345	1-year institutional subscription	$375
2-year institutional subscription	$625	2-year institutional subscription	$685
3-year institutional subscription	$875	3-year institutional subscription	$965
Electronic only		**Electronic only**	
1-year subscription	$ 85	1-year subscription	$ 85
2-year subscription	$155	2-year subscription	$155
3-year subscription	$215	3-year subscription	$215
1-year student subscription	$ 55	1-year student subscription	$ 55

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 (2014)
14, Number 4, pp. 778–797

dhreg, xtdhreg, and bootdhreg: Commands to implement double-hurdle regression

Christoph Engel
Max Planck Institute for Research
on Collective Goods
Bonn, Germany
engel@coll.mpg.de

Peter G. Moffatt
School of Economics
University of East Anglia
Norwich, UK
P.Moffatt@uea.ac.uk

Abstract. The `dhreg` command implements maximum likelihood estimation of the double-hurdle model for continuously distributed outcomes. The command includes the option to fit a p-tobit model, that is, a model that estimates only an intercept for the hurdle equation. The `bootdhreg` command (the bootstrap version of `dhreg`) may be convenient if the data-generating process is more complicated or if heteroskedasticity is suspected. The `xtdhreg` command is a random-effects version of `dhreg` applicable to panel data. However, this estimator differs from standard random-effects estimators in the sense that the outcome of the first hurdle applies to the complete set of observations for a given subject instead of applying at the level of individual observations. Command options include estimation of a correlation parameter capturing dependence between the two hurdles.

Keywords: st0359, dhreg, xtdhreg, bootdhreg, hurdle model, double-hurdle model, random-effects double-hurdle model, tobit, p-tobit, inverse Mills ratio, bootstrapping

1 Introduction

The double-hurdle model, introduced by Cragg (1971), embodies the idea that an individual's decision on the extent of participation in an activity is the result of two processes: the first hurdle, determining whether the individual is a zero type, and the second hurdle, determining the extent of participation given that the individual is not a zero type. A key feature of the model is that there are two types of zero observations: an individual can be a zero type, and the outcome will always be zero whatever his or her circumstances at the time of the decision; alternatively, the individual might not be a zero type, but his or her current circumstances might dictate that the outcome is zero—this sort of zero is usually classified as a censored zero after (Tobin 1958).

In addition to naturally incorporating the zero type, the double-hurdle model allows estimation of the proportion of the population that is of the zero type. Better still, it allows the probability of a subject's being zero type to depend on the subject's characteristics.

The double-hurdle model has previously been applied in a variety of contexts. Jones (1989) applied it to cigarette consumption by individuals, with the justification that a proportion of the population would never smoke whatever circumstances they found themselves in. Burton, Tomlinson, and Young (1994) applied the model to meat consumption by single-adult households, with the justification that a proportion of the population of single adults must be vegetarian and therefore destined to record zero consumption of meat. The model has also been applied in models of loan default (Dionne, Artís, and Guillén 1996; Moffatt 2005), where it is assumed that a proportion of borrowers would, out of principle, never default on a loan.

The practice of fitting hurdle models is well developed in the context of count-data outcomes. See, for example, the user-written Stata commands `ztpnm` and `hnbclg`. McDowell (2003) has provided advice from the Stata help desk on the required programming. The practice is less developed in the context of continuous outcomes, the case of interest here. After our paper was accepted for publication, the `dblhurdle` command and the accompanying article by Garcia (2013) became available. Unlike our commands, the `dblhurdle` command offers weighted estimation. By contrast, our `dhreg` command allows the capture of possible correlation of the error terms between the hurdle equation and the equation for choices that pass the hurdle by using the inverse Mills's ratio.

We also extend the hurdle framework to panel data (again going beyond `dblhurdle`). The panel-hurdle model has been applied to household milk consumption by Dong and Kaiser (2008). From that starting point, we extend the panel-hurdle model in an important way: we assume a nonzero correlation between the individual-specific error terms in the two hurdles. Hence, we achieve superior efficiency by fitting the panel-hurdle model with dependence. In a double-hurdle model with just one observation per subject, there are problems identifying this correlation (Smith 2003). In contrast, with panel data, we shall see that the parameter can be estimated with reasonably high precision.

The panel-hurdle model is potentially important in experimental economics, in which there are several natural applications, including public goods experiments and dictator games. Sometimes, each subject engages in only one task; however, it is more common for each subject to engage in a sequence of tasks throughout an experiment. The extension of the model to panel data requires care because the outcome of the first hurdle—that is, the determination of whether a respondent is of the zero type—must apply to that respondent for every period. Switching in and out of the zero type is ruled out. In contrast, the outcome of the second hurdle—that is, the amount consumed or contributed in any period—is determined at the level of individual observations. In principle, respondents classified as nonparticipants must necessarily consume or contribute zero in every period. We also offer a bootstrap version of the estimator, again going beyond the `dblhurdle` command. Bootstrapping can, for instance, be helpful if choices are nested in individuals who are, in turn, nested in higher-level units.

In section 2, we cover theoretical aspects of the double-hurdle model, specifying the likelihood function for each model. In section 3, we do the same for the panel-hurdle model. We then introduce the user-written Stata commands and syntax in section 4. We demonstrate the commands using a simulated dataset in section 5.

2 Double-hurdle and related models

When referring to examples in the description of models, we will use the term "contribution" to represent the outcome variable; the term "consumption" could also be used.

2.1 Tobit

A natural starting point is the tobit model (Tobin 1958) because the hurdle model is an extension of it. Tobit-type models, or censored regression models, are required when the dependent variable is censored, that is, when there is an accumulation of observations at the limits of the range of the variable. The lower limit of the range is usually zero, and censoring is usually zero censoring, although sometimes, we are required to deal with upper censoring, where there is an accumulation of observations at the maximum. In other contexts, there is censoring from below but at a cutoff point different from zero. The software handles all of these. Yet when introducing the theory, we shall focus on lower censoring at zero.

We start with a linear equation in which the dependent variable is a latent (unobserved) variable, y_i^*, representing the desired contribution of subject i:

$$y_i^* = \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_i$$
$$\varepsilon_i \sim N(0, \sigma^2)$$

The desired contribution is assumed to be a linear function of the observed subject characteristics and treatment variables contained in the vector $\mathbf{x}_i$, plus a normally distributed random error. The important feature of y_i^* is that it can be negative: subjects are permitted to desire to contribute a negative amount. Of course, if a subject does desire to contribute a negative amount, most experimental designs would constrain the subject to contribute zero;[1] if the subject desires to contribute any positive amount, this amount will be his or her actual observed contribution. This amounts to what is known as a censoring rule:

$$y_i = y_i^* \text{ if } \; y_i^* > 0$$
$$y_i = 0 \text{ if } \; y_i^* \leqslant 0 \qquad (1)$$

y_i is the observed contribution of subject i. Therefore, the censoring rule shows the relationship between desired and actual contributions.

1. An interesting new development is the emergence of take games, dictator games in which some treatments allow dictators to take money away from the recipient, that is, to give less than zero (see List [2007]; Bardsley [2008]).

In the situation where we have lower censoring at zero, there are two regimes of behavior: zero observations and positive observations. The sample log-likelihood function is constructed by combining contributions for each regime as follows:

$$\text{Log}L = \sum_{i=1}^{n} \left[I_{y_i=0} \ln \left\{ \Phi \left(-\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma} \right) \right\} + I_{y_i>0} \ln \left\{ \frac{1}{\sigma} \phi \left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma} \right) \right\} \right] \tag{2}$$

I is the indicator function, taking the value one if the subscripted expression is true, and zero otherwise. Log$L$ is maximized with respect to the parameters contained in the vector $\boldsymbol{\beta}$ and the standard deviation parameter σ.

2.2 p-tobit

The overrestrictive feature of the tobit model described in section 2.1 is that it allows only one type of zero observation, and the implicit assumption is that zeros arise because of subject circumstances and treatments. The obvious way to relax this is to assume the existence of an additional class of subjects who would never contribute under any circumstances.

In the first instance, let us assume that the proportion of the population who are potential contributors is p, so that the proportion of the population who would never contribute is $1 - p$. For the former group, the tobit model applies, while for the latter group, the contribution is automatically zero.

This assumption leads to the p-tobit model, originally proposed by Deaton and Irish (1984) in the context of household consumption decisions, where they were essentially allowing for a class of abstinent consumers for each good modeled. The log-likelihood function for the p-tobit model is[2]

$$\text{Log}L = \sum_{0} \ln \left\{ 1 - p\Phi \left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma} \right) \right\} + \sum_{+} \ln \left\{ p \frac{1}{\sigma} \phi \left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma} \right) \right\} \tag{3}$$

Maximizing (3) returns an estimate of the parameter p, in addition to those of $\boldsymbol{\beta}$ and σ obtained under tobit.

2.3 Double hurdle

Because the class of subjects who would never contribute may be the focus of the analysis, it is desirable to investigate which types of subjects are most likely to appear in this class. With this in mind, we assume that the probability of a subject's being in the said class depends on a set of subject characteristics. In other words, we shall generalize the p-tobit model of section 2.2 by allowing the parameter p to vary according to subject characteristics. This generalization leads us to the double-hurdle model.

2. For readers unfamiliar with the structure of log likelihoods such as (2) and (3), a useful basic principle is that because of the symmetry of the normal distribution, $\Phi(-z) = 1 - \Phi(z)$.

As the model name suggests, subjects must cross two hurdles to contribute. The first hurdle needs to be crossed to be a potential contributor. Given that the subject is a potential contributor, his or her current circumstances and treatment in the experiment dictate whether he or she contributes—this is the second hurdle.

The double-hurdle model contains two equations and can be given the interpretation of a combined probit and tobit estimator. We write

$$\begin{aligned} d_i^* &= \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{1,i} \\ y_i^{**} &= \mathbf{x}_i'\boldsymbol{\beta} + \varepsilon_{2,i} \\ \begin{pmatrix}\varepsilon_{1,i}\\ \varepsilon_{2,i}\end{pmatrix} &\sim N\left[\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 0\\ 0 & \sigma^2\end{pmatrix}\right] \end{aligned} \tag{4}$$

The variance of $\varepsilon_{1,i}$ is normalized to 1, as required for identification, because the outcome of the first hurdle is binary. The diagonality of the covariance matrix implies that the two error terms are assumed to be independently distributed.

The first hurdle is represented by

$$\begin{aligned} d_i &= 1 \text{ if } d_i^* > 0 \\ d_i &= 0 \text{ if } d_i^* \leqslant 0 \end{aligned} \tag{5}$$

The first hurdle is thus assumed to be defined by the latent variable d_i^*. The second hurdle closely resembles the tobit model (1):

$$y_i^* = \max(y_i^{**}, 0) \tag{6}$$

Finally, the observed variable, y_i, is determined as

$$y_i = d_i y_i^* \tag{7}$$

The log-likelihood function for the double-hurdle model is

$$\text{Log}L = \sum_0 \ln\left\{1 - \Phi(\mathbf{z}_i'\boldsymbol{\alpha})\Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right\} + \sum_+ \ln\left\{\Phi(\mathbf{z}_i'\boldsymbol{\alpha})\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right\} \tag{8}$$

When the lower hurdle is $y_{\min} \neq 0$, the first term in (8) changes to $\sum_{\min} \ln[1 - \Phi(\mathbf{z}_i'\boldsymbol{\alpha})\Phi\{(\mathbf{x}_i'\boldsymbol{\beta} - y_{\min})/\sigma\}]$. When the hurdle is an upper hurdle, $y_{\max}$, the first term becomes $\sum_{\max} \ln[1 - \Phi(\mathbf{z}_i'\boldsymbol{\alpha})\Phi\{(y_{\max} - \mathbf{x}_i'\boldsymbol{\beta})/\sigma\}]$.

Figure 1 is useful for understanding the model defined in (4)–(7). The concentric ellipses are contours of the joint distribution of the latent variables d^* and y^{**}. These ellipses are centered on the point $(\mathbf{z}_i'\boldsymbol{\alpha}, \mathbf{x}_i'\boldsymbol{\beta})$ so that the whole distribution moves around with changes in the values taken by the explanatory variables. The likelihood component associated with noncontribution [that is, the first term in curly braces in (8)] is represented by the probability mass under the L-shaped region comprising the

northwest, southwest, and southeast quadrants of the graph; the likelihood component associated with a contribution [the second term in curly braces in (8)] is represented by a thin strip of the probability mass within the northeast quadrant at the value of the observed contribution (one such value is depicted in the diagram).

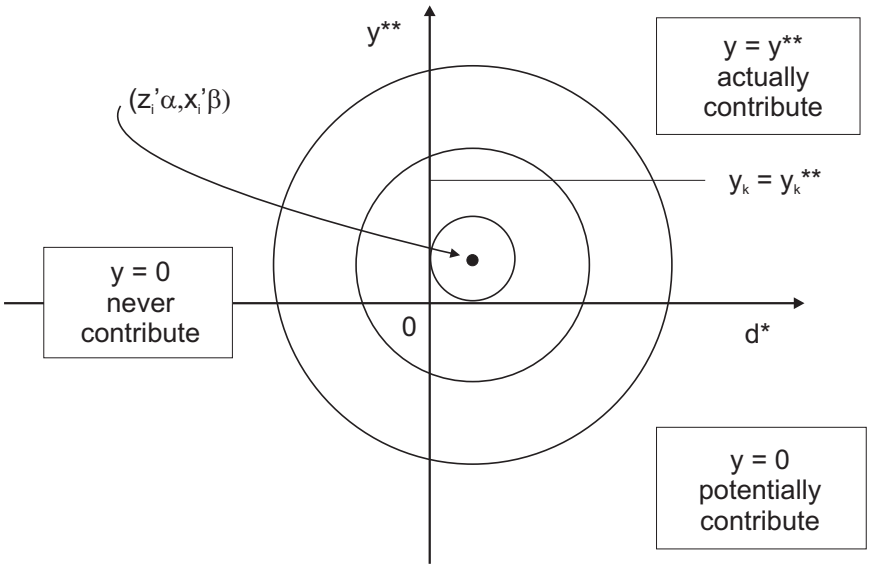

Figure 1. The relationship between latent (d* and y**) and observed (y) variables in the double-hurdle model

Finally, consider a double-hurdle model in which there are no explanatory variables in the first-hurdle equation. There is only an intercept, α_0. The likelihood function becomes

$$\text{Log}L = \sum_0 \ln\left\{1 - \Phi(\alpha_0)\Phi\left(\frac{\mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right\} + \sum_+ \ln\left\{\Phi(\alpha_0)\frac{1}{\sigma}\phi\left(\frac{y_i - \mathbf{x}_i'\boldsymbol{\beta}}{\sigma}\right)\right\}$$

$\Phi(\alpha_0)$ is now a scalar. If we rename this scalar as p, we have the p-tobit model defined in (3).

This gives us a way of fitting the p-tobit model. We fit the double-hurdle model with no explanatory variables in the first hurdle. We then transform the estimate of the intercept parameter in the first hurdle, α_0, using

$$p = \Phi(\alpha_0)$$

This gives the estimate of the parameter p in the p-tobit model. The delta method is required to obtain a standard error for this estimate.

2.4 The single-hurdle model

The single-hurdle model is a model that has the property of first-hurdle dominance (Jones 1989). This essentially requires that any individual who passes the first hurdle necessarily has a positive outcome. Hence, there is only one source of zeros, the zero type; censored zeros are ruled out.

The formal definition of the single-hurdle model is similar to that of the double-hurdle model given in section 2.3; the only difference is that (6) changes from a rule embodying zero censoring to one embodying zero truncation:

$$\begin{aligned} y_i^* &= y_i^{**} \quad \text{if} \quad y_i^{**} > 0 \\ y_i^* &\quad \text{unobserved if} \quad y_i^{**} \leqslant 0 \end{aligned}$$

As will be mentioned in the next section, logical problems arise when we try to extend the single-hurdle model to the panel-data setting. For this reason, we do not pay close attention to this model.

3 Extension to panel data

3.1 The basic panel-hurdle model

Until this point in the article, we have been concerned with estimation with one cross-section of data. We now progress to panel data. The panel-hurdle model was developed by Dong and Kaiser (2008), who applied the model to household milk consumption. Here we assume that we have n subjects, each of whom participated in T tasks. We denote y_{it} as the decision (that is, the contribution) of subject i in task t. The two hurdles are defined as follows:

First hurdle

$$\begin{aligned} d_i^* &= \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{1,i} \\ d_i &= 1 \text{ if } d_i^* > 0; d_i = 0 \text{ otherwise} \\ \varepsilon_{1,i} &\sim N(0,1) \end{aligned}$$

Second hurdle

$$\begin{aligned} y_{it}^{**} &= \mathbf{x}_{it}'\boldsymbol{\beta} + u_i + \varepsilon_{2,it} \\ y_{it}^{*} &= \max(y_{it}^{**}, 0) \\ \begin{pmatrix} \varepsilon_{1,i} \\ u_i \\ \varepsilon_{2,it} \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho\sigma_u & 0 \\ \rho\sigma_u & \sigma_u^2 & 0 \\ 0 & 0 & \sigma^2 \end{pmatrix} \right] \end{aligned}$$

Observed

$$y_{it} = d_i y_{it}^*$$

The central feature of the panel-hurdle model is that the first hurdle has only one outcome per subject, and that outcome applies to all observations for that subject. For example, if subject i falls at the first hurdle ($d_i = 0$), then all observations on y for subject i must be 0 ($y_{it} = 0$, $t = 1, \ldots, T$). This feature is essential to capture the concept of the zero type. If a subject is a zero type, then the subject will necessarily contribute zero on every occasion on which he or she is observed.

Note that the second hurdle contains a subject-specific random-effects term (u_i) that allows between-subject heterogeneity and thereby within-subject dependence. In the specification of the joint distribution of the three stochastic terms, we have assumed that the correlation between $\varepsilon_{1,i}$ and u_i is ρ. In this section, we are considering the model with independence between the two hurdles, so we assume that $\rho = 0$. In section 3.4 (panel-hurdle model with dependence), we will allow ρ to be nonzero and consider the strategy for estimating this parameter.

3.2 The panel single-hurdle model

We introduced the single-hurdle model in section 2.4. This is a model satisfying first-hurdle dominance: passing the first hurdle necessarily implies a positive outcome. In the panel setting, first-hurdle dominance gives rise to a logical problem. If an individual passes the first hurdle, his or her outcomes would need to be positive in every period. We already know that if the individual falls at the first hurdle, the outcome is zero every time. Hence, first-hurdle dominance rules out a mixture of zero and positive outcomes for a given individual. This is clearly a serious problem because most panel-data sets would be expected to contain such mixtures. For this reason, we shall restrict attention to the framework of the panel double-hurdle model introduced in section 3.1, in which the zero censoring assumed in the second hurdle allows mixtures of zeros and positive observations for a given individual.

3.3 Construction of likelihood function

Conditional on $d_i = 1$ (and also on the heterogeneity term u_i), we obtain something very similar to the random-effects tobit likelihood:

$$(L_i|d_i = 1, u_i) = \prod_{t=1}^{T} \left\{1 - \Phi\left(\frac{\mathbf{x}_{it}'\boldsymbol{\beta} + u_i}{\sigma}\right)\right\}^{I(y_{it}=0)} \left\{\frac{1}{\sigma}\phi\left(\frac{y_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - u_i}{\sigma}\right)\right\}^{I(y_{it}>0)} \tag{9}$$

Conditional on $d_i = 0$, the likelihood is trivial and depends on whether all observations are zero for subject i:

$$\begin{aligned}(L_i|d_i = 0) &= 0 \text{ if } \sum_{t=1}^{T} y_{it} > 0 \\ &= 1 \text{ if } \sum_{t=1}^{T} y_{it} = 0 \end{aligned} \tag{10}$$

The likelihood (conditional on u_i) for subject i is then obtained as a weighted average of (9) and (10), with weights given by the probabilities $P(d_i = 1)$ and $P(d_i = 0)$, which are obtained from the first-hurdle equation.

$$(L_i|u_i) = \Phi(\mathbf{z}_i'\boldsymbol{\alpha})(L_i|d_i = 1, u_i) + \{1 - \Phi(\mathbf{z}_i'\boldsymbol{\alpha})\}(L_i|d_i = 0) \tag{11}$$

Finally, the marginal likelihood for subject i is obtained by integrating (11) over u,

$$L_i = \int_{-\infty}^{\infty} (L_i|u) f(u) du$$

where $f(u)$ is the normal $(0, \sigma_u^2)$ density function for u.

The sample log-likelihood function is then given by

$$\text{Log}L = \sum_{i=1}^{n} \ln L_i$$

3.4 Panel-hurdle model with dependence

The model developed above assumes that there is no correlation between the error terms in the two hurdles. In this section, this assumption is relaxed.

Subject i's idiosyncratic propensity to pass the first hurdle is represented by the error term $\varepsilon_{1,i}$; i's idiosyncratic propensity to contribute, conditional on passing the first hurdle, is represented by u_i. It is between these two terms that we introduce a correlation:

$$\text{corr}(\varepsilon_1, u) = \rho$$

How is the correlation parameter ρ incorporated in estimation? Let us return to the first hurdle:

$$\begin{aligned} d_i^* &= \mathbf{z}_i'\boldsymbol{\alpha} + \varepsilon_{1,i} \\ d_i &= 1 \text{ if } d_i^* > 0;\ d_i = 0 \text{ otherwise} \\ \varepsilon_{1,i} &\sim N(0, 1) \end{aligned}$$

Because $\text{corr}(\varepsilon_1, u) = \rho$, we can represent ε_1 as

$$\varepsilon_1 = \rho\frac{u}{\sigma_u} + \sqrt{1-\rho^2}\xi$$

where $\xi \sim N(0,1)$ and $\xi \perp u$. The requirement for passing the first hurdle becomes

$$d_i = 1 \text{ if } \xi > -\frac{\mathbf{z}_i'\boldsymbol{\alpha} + \rho\dfrac{u}{\sigma_u}}{\sqrt{1-\rho^2}}$$

from which the probability of passing the first hurdle becomes

$$\Phi\left(\frac{\mathbf{z}_i'\boldsymbol{\alpha} + \rho\dfrac{u}{\sigma_u}}{\sqrt{1-\rho^2}}\right) \tag{12}$$

In estimation, the Halton draws used to represent realizations of u in the second hurdle also appear in the probability of passing the first hurdle in accordance with (12).

In the standard double-hurdle model with dependence (one observation per subject), there can be problems identifying the correlation coefficient ρ (Smith 2003). However, with panel data and use of the estimation approach outlined in this section, the parameter can be estimated precisely.

3.5 Two-step estimation of the dependence model

Heckman (1979) developed a procedure that treats correlation as an omitted variable problem. If the error terms are indeed correlated, the inverse Mills ratio from the first component must have explanatory power for the second component. Specifically, the coefficient of this additional regressor is precisely the covariance of the two error terms. Using this approach, we have a tractable way of fitting a double-hurdle model if the error terms for the d_i^* and the y_i^{**} are possibly correlated:

- Estimate the double-hurdle model assuming covariance to be zero.
- Generate the inverse Mills ratio from the first component. If the first component is estimated with probit, it is given by

$$\frac{\phi(\mathbf{z}_i'\boldsymbol{\alpha})}{\Phi(\mathbf{z}_i'\boldsymbol{\alpha})}$$

- Refit the double-hurdle model with the first component as before, but with the second component (the y_i^{**} equation) estimated with the inverse Mills ratio as an additional explanatory variable. If the additional regressor turns out significant, this suggests that the two processes are indeed correlated.

If the problem at hand invites exclusion restrictions for the first hurdle, these restrictions identify the effect of the outer hurdle. Otherwise, as in a standard Heckman model, identification is through functional form only.

3.6 Bootstrap version of the model

The panel version of the double-hurdle model assumes a specific, randomly distributed error, uncorrelated with observables. Because there is no acknowledged fixed-effects version of the panel tobit model, a fortiori there is no reliable fixed-effects estimator for double-hurdle regression, which builds on tobit. As a safeguard, we offer a nonparametric version using bootstrapping, the `bootdhreg` estimator, which is the bootstrap version of the `dhreg` estimator. Users can match the panel structure of their data-generating process by sampling from participants as clusters. This command may also be convenient if there are reasons to doubt the normality assumption on which the maximum likelihood procedure of `dhreg` is based or if there are higher-level clusters. An illustration in the framework of the smoking example would be two samples taken from different subpopulations, say, the pupils of two different schools. Of course, the bootstrap clusters are less efficient than the random-effects model because they do not make any assumptions about covariance terms.

3.7 Estimation

Estimation of the panel-hurdle model is performed using the method of maximum simulated likelihood (Train 2009). This requires the use of Halton draws, which, when converted to normality, represent simulated realizations of the random-effects term u. In the model with dependence, in accordance with (12), the simulated values also appear in the probability of passing the first hurdle. Maximization of the simulated likelihood function is performed using the ml routine in Stata.

4 The dhreg, xtdhreg, and bootdhreg commands

4.1 Syntax

`dhreg` *depvar indepvars* [*if*] [*in*] [`, up ptobit hd(`*varlist*`) millr`]

`xtdhreg` *depvar indepvars* [*if*] [*in*] [`, up ptobit hd(`*varlist*`) uncorr trace difficult constraints(`*numlist*`)`]

`bootdhreg` *depvar indepvars* [*if*] [*in*] [`, up ptobit hd(`*varlist*`) millr margins(`*string*`) seed(`*integer*`) reps(`*integer*`) strata(`*varlist*`) cluster(`*varlist*`) capt maxiter(`*integer*`)`]

4.2 Options for dhreg

`up` specifies that the upper, not the lower, limit of the support of the dependent variable be treated as the hurdle.

`ptobit` estimates the equation for the outer hurdle with just the intercept.

`hd(`*varlist*`)` allows a set of explanatory variables for the outer hurdle that differs from the explanatory variables for the inner hurdle and those realizations of the dependent variable that surmount the hurdle.

`millr` estimates a second version of the model with the inverse Mills ratio controlling for potential correlation of the error terms.

4.3 Options for xtdhreg

`up` specifies that the upper, not the lower, limit of the support of the dependent variable be treated as the hurdle.

`ptobit` estimates the equation for the outer hurdle with just the intercept.

`hd(`*varlist*`)` allows a set of explanatory variables for the outer hurdle that differs from the explanatory variables for the inner hurdle and those realizations of the dependent variable that surmount the hurdle.

`uncorr` assumes that the error terms of the hurdle equation and of the main equation are uncorrelated.

`trace` displays coefficients from each iteration.

`difficult` uses an alternative, more calculation-intense algorithm for approximation (which may help if the model does not converge).

`constraints(`*numlist*`)` makes it possible for users to constrain the model.

4.4 Options for bootdhreg

`up` specifies that the upper, not the lower, limit of the support of the dependent variable be treated as the hurdle.

`ptobit` estimates the equation for the outer hurdle with just the intercept.

`hd(`*varlist*`)` allows a set of explanatory variables for the outer hurdle that differs from the explanatory variables for the inner hurdle and those realizations of the dependent variable that surmount the hurdle.

`millr` estimates a second version of the model with the inverse Mills ratio controlling for potential correlation of the error terms.

`margins(`*string*`)` calls for bootstrap estimates of marginal effects.

`seed(`*integer*`)` fixes a seed for the randomization (as a default, one seed is implemented so that results can be replicated).

`reps(`*integer*`)` defines the number of bootstrap repetitions. The default is `reps(50)`.

`strata(`*varlist*`)` orders the bootstrap to be stratified.

`cluster(`*varlist*`)` defines higher-order aggregates from which samples are drawn.

`capt` ignores repetitions that do not converge. It may be useful if, for instance, the program does not converge for some samples because of clustering. The bootstrap results are then taken from the remaining draws.

`maxiter(`*integer*`)` limits the number of iterations. `maxiter()` is also useful if the maximum likelihood routine has a hard time converging. `maxiter()` should be combined with `capt`. The default is `maxiter(50)`.

5 Examples

Here is a stylized example for using the `dhreg` command. It works with simulated data with the following data-generating process:

$$
\begin{aligned}
d_i^* &= \begin{cases} 1 & \text{if } -2 + 4 \times z_i + \varepsilon_{i1} > 0 \\ 0 & \text{otherwise} \end{cases} \\
y_i^{**} &= 0.5 + 0.3 \times x_i + \varepsilon_{i2} \\
y_i^* &= \begin{cases} y_i^{**} & \text{if } y_i^{**} > 0 \\ 0 & \text{otherwise} \end{cases} \\
y_i &= d_i^* \times y_i^* \\
\varepsilon_{i1} &= 0.5 \times \varepsilon_{i2} + \sqrt{(1 - 0.5^2)} \times \eta_i \\
\varepsilon_{i2}, \eta_i &\sim N(0, 1) \\
\text{corr}(\varepsilon_{i1}, \varepsilon_{i2}) &= 0.5 \\
z_i, x_i &\sim U(0, 1)
\end{aligned}
$$

In these data, the latent process defined by the first equation generates the first hurdle, which is determined by a constant, the (uniformly distributed) exogenous variable `z`, and the error term for this process, ε_{i1}. This error term has a correlation of 0.5 with the error term of the second process, ε_{i2}; these correlated errors are simulated by a separate normal variate, η_i. The latent process defined by the second equation generates the magnitude of those observations that pass both hurdles. This second process is determined by (uniformly distributed) exogenous variable `x` and the (normally distributed) error term, ε_{i2}. Through the final equation ($y_i = d_i^* \times y_i^*$), the observed dependent variable, y_i, is zero if either the first or the second hurdle is not passed and otherwise has the magnitude of the second latent variable, y_i^{**}. Figure 2 shows the resulting data. Nearly half fall at the first hurdle. A sizable portion falls at the second hurdle.

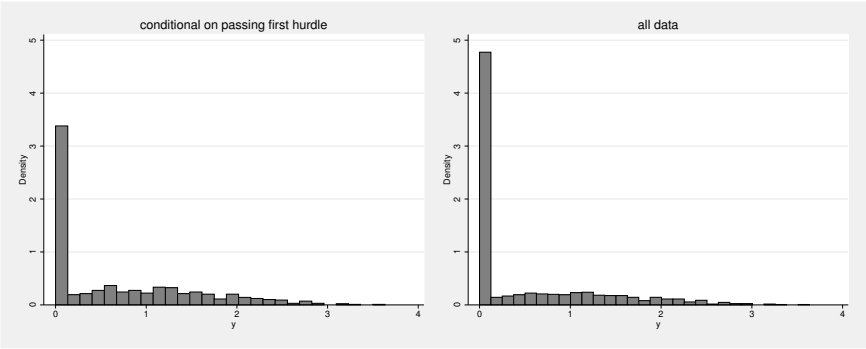

Figure 2. Data-generating process cross-section example

In the first step, we show that tobit is not appropriate for this data-generating process. The model does not pick up the effect of `x`. The coefficient is insignificant and about half as large as the actual effect.

```
. tobit y x, ll(0)
Tobit regression                                  Number of obs   =       1000
                                                  LR chi2(1)      =       0.70
                                                  Prob > chi2     =     0.4015
Log likelihood = -1127.5531                       Pseudo R2       =     0.0003

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   .1549814   .1846556     0.84   0.402    -.2073759    .5173388
       _cons |  -.3249669   .1145513    -2.84   0.005    -.5497557    -.100178
-------------+----------------------------------------------------------------
      /sigma |   1.490576   .0589589                      1.374879    1.606274
------------------------------------------------------------------------------

  Obs. summary:        580  left-censored observations at y<=0
                       420     uncensored observations
                         0 right-censored observations
```

In the next step, we present output from the double-hurdle model but still (wrongly, given our simulation) assume that error terms are uncorrelated. For expositional reasons, we had the effect of `x` designed to be small. It is now properly estimated, as is the effect of `z` and the estimate of the standard error. However, all coefficients are still slightly biased.

```
. dhreg y x, hd(z)
  (output omitted)
maximum likelihood estimates of double hurdle model

N = 1000
log likelihood = -947.24225
chi square hurdle equation = 81.817097
p hurdle equation = 1.493e-19
chi square above equation = 4.7952012
p above equation = .02853912
chi square overall = 90.440744
p overall = 2.296e-20
```

	coef	se	z	p	lower CI	upper CI
hurdle						
z	3.877918	.4287228	9.04528	0	3.037636	4.718199
_cons	-1.852446	.152426	-12.15308	5.52e-34	-2.151195	-1.553696
above						
x	.3673263	.1677446	2.189795	.0285391	.0385529	.6960997
_cons	.6856148	.113751	6.027331	1.67e-09	.462667	.9085626
sigma						
_cons	.9423028	.0564864	16.68193	0	.8315915	1.053014

In the final step, we estimate the correlation using the inverse Mills ratio. The model rightly suggests that error terms are correlated (the coefficient of the inverse Mills ratio is highly significant). Coefficients are now almost in line with the data-generating process. Comparing the Wald tests, we see that the model has a considerably better fit.

```
. dhreg y x, hd(z) millr
  (output omitted)
 second stage results

N = 1000
log likelihood = -937.13325
chi square hurdle equation = 121.026
p hurdle equation = 3.772e-28
chi square above equation = 29.057293
p above equation = 4.901e-07
chi square overall = 127.67061
p overall = 1.891e-28
```

	coef	se	z	p	lower CI	upper CI
hurdle						
z	4.049705	.3681154	11.00118	0	3.328212	4.771198
_cons	-2.002341	.1479133	-13.53727	9.42e-42	-2.292246	-1.712437
above						
x	.3243295	.1560948	2.077773	.0377303	.0183894	.6302696
_mill	.5337018	.1108932	4.812754	1.49e-06	.3163551	.7510485
_cons	.5127246	.1099401	4.663672	3.11e-06	.2972459	.7282032
sigma						
_cons	.8811395	.0446539	19.73263	0	.7936194	.9686596

Our second example is also from simulated data, meant to illustrate the random effects and bootstrap estimators. The data-generating process is as follows:

$$
\begin{aligned}
d_i^* &= \begin{cases} 1 & \text{if } -2 + 4 \times z_i + \varepsilon_{i1} > 0 \\ 0 & \text{otherwise} \end{cases} \\
y_{it}^{**} &= 0.5 + 0.3 \times x_{it} + u_i + \varepsilon_{it2} \\
y_{it}^* &= \begin{cases} y_{it}^{**} & \text{if } y_{it}^{**} > 0 \\ 0 & \text{otherwise} \end{cases} \\
y_{it} &= d_i^* \times y_{it}^* \\
\varepsilon_{i1} &= 0.9 \times u_i + \sqrt{(1 - 0.9^2)} \times \eta_i \\
\begin{pmatrix} \varepsilon_{it2} \\ u_i \end{pmatrix} &\sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & \sigma^2 \end{pmatrix} \right] \\
\eta_i &\sim N(0, \sigma^2)
\end{aligned}
$$

This dataset is a panel. Individuals i are observed at multiple points in time t. Their choices are determined by the first hurdle, d_i^*, which is assumed to be time invariant, and by the second hurdle, $y_{it}^* = 0$, which is allowed to differ by individual and period. In the latent-variable defining choices, provided the first hurdle is passed, there is a random effect u_i. The correlation of error terms involves this random effect and not the residual error ε_{it2}.

The output shows that the estimator finds the effects of both explanatory variables, including the very small effect of x_{it}, and the correlation of the error terms.

```
. xtdhreg y x, hd(z)
  (output omitted)
                                                Number of obs   =      10000
                                                Wald chi2(1)    =      43.39
Log likelihood = -6197.9509                     Prob > chi2     =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hurdle						
z	4.736641	.719114	6.59	0.000	3.327203	6.146079
_cons	-2.443042	.45771	-5.34	0.000	-3.340137	-1.545946
above						
x	.2585995	.0530704	4.87	0.000	.1545835	.3626156
_cons	.986435	.0538315	18.32	0.000	.8809273	1.091943
sigma_u						
_cons	1.114397	.0493999	22.56	0.000	1.017575	1.211219
sigma_e						
_cons	1.00247	.0121348	82.61	0.000	.9786863	1.026254
transformed~o						
_cons	.7980986	.5459896	1.46	0.144	-.2720214	1.868219

```
 generate estimate of correlation in error terms, with confidence interval

        rho:  tanh([transformed_rho]_cons)
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
rho	.6629725	.3060095	2.17	0.030	.0632049	1.26274

```
 separate Wald tests for joint significance of all explanatory variables
 note
 if you use factor variables, i.e. the i., c., # and ## notation, you must run
 the Wald test by hand. For detail see help file
 estimates of joint significance
chi square hurdle equation = 43.385582
p hurdle equation = 4.495e-11
chi square above equation = 23.743844
p above equation = 1.100e-06
chi square overall = 67.117399
p overall = 2.665e-15
```

In this case, we know this estimator to be appropriate. All normality assumptions are met, and the data are not nested. Therefore, there is no reason to resort to bootstrapping. Yet in real applications, it may be less obvious that the assumptions underlying a random-effects model are justified. If users instead or additionally run a bootstrap, the output looks as follows:

```
. bootdhreg y x, hd(z) cluster(i) capt
  (output omitted)
 maximum likelihood estimates of double hurdle model

N = 10000
log likelihood = -9860.9387
chi square hurdle equation = 9.3724107
p hurdle equation = .00220276
chi square main equation = 9.3724107
p main equation = .00220276
chi square overall = 570.78004
p overall = 1.14e-124
bootstrap results
```

	coef	se	p	lowciz	upciz	lowcip
hurdle						
z	4.218375	3.43218	.1095236	-2.508698	10.94545	2.856766
_cons	-2.180051	1.124344	.0262535	-4.383766	.0236639	-6.814596
main						
x	.243663	.0644732	.0000786	.1172956	.3700305	.1066285
_cons	.851831	.3136393	.0033042	.2370979	1.466564	.0889678
sigma						
_cons	1.370452	.1429195	0	1.09033	1.650574	1.147232

	upcip
hurdle	
z	16.95187
_cons	-1.541749
main	
x	.39398
_cons	1.338123
sigma	
_cons	1.699634

Bootstrap standard errors are larger than standard errors from the random-effects model. This is as expected. The bootstrap assumes less structure (there is no random effect), and it coarsens the data by the sampling process. The bootstrap routine resamples with replacement. Because the `cluster(i)` option is used, entire sets of observations per individual are sampled. With our seed (which users are free to change with the `seed()` option), some bootstrap samples do not converge, which is why the `capt` option is used. It confines the calculation of the bootstrap standard errors to those (multiple) instances that converge. As is standard in bootstrapping, coefficients are taken from using `dhreg` on the original data, but standard errors are taken from the bootstrap. The reported bootstrap standard error is the standard deviation of coefficients from all bootstrap instances. The resulting p-value and the `lowciz` and `upciz` confidence intervals assume that these results are normally distributed. The two final columns report a distribution-free estimate of the confidence interval. It results from the lower and the upper 2.5% of the empirical distribution of coefficients. If users want to rely on these estimates, they should check whether zero lies outside this interval. If

users wish to allow for correlation of the error terms, they can combine `bootdhreg` with the `millr` option.

Additionally, coefficients from all bootstrap repetitions of the statistical model are stored in variables `_res`. These data are useful if the user wants to run additional tests, such as a Wald-like test for a net effect. This can be done by generating a new variable that sums up the main effect and the interaction effect. With the `summarize` command, one generates the mean and the standard error of this new variable. `min((1-normal(r(mean)/r(sd))), (normal(r(mean)/r(sd))))` generates the p-value.

We have enabled all estimators to take factor variables. Users can therefore rely on the `i.a##c.b` or on the `c.b##c.b` notation to generate interaction terms and other multiplicative terms, and they can use the `margins` command to derive model predictions, as we show in the example below.[3] With the `margins` command, users can also calculate average marginal effects with the `dydx()` option. Yet for recovering the average marginal change in probability, which tends to be more interesting, the following, somewhat unintuitive, command must be used. It transforms the (average of the) linear marginal effect into a probability (because we use a probit specification for the second equation). In the example, a one-unit change in variable `z` (which would be a change from one extreme to the other, given the variable has range 0–1) increases the probability of the first hurdle being passed by almost 97%.

```
. margins, dydx(z) expression(normal(xb(hurdle)))
Average marginal effects                        Number of obs   =       1000
Model VCE    : OIM
Expression   : normal(xb(hurdle))
dy/dx w.r.t. : z
```

	Delta-method dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
z	.9675177	.0294781	32.82	0.000	.9097416	1.025294

6 References

Bardsley, N. 2008. Dictator game giving: Altruism or artefact? *Experimental Economics* 11: 122–133.

Burton, M., M. Tomlinson, and T. Young. 1994. Consumers' decisions whether or not to purchase meat: A double hurdle analysis of single adult households. *Journal of Agricultural Economics* 45: 202–212.

3. Because there is more than one equation in our model, through `predict(equation(eqname))`, users must specify the equation to which the coefficient in question refers. The hurdle equation is always denoted hurdle, while the equation estimating the dependent variable conditional on the first hurdle being passed is denoted above if a lower hurdle is estimated and below if an upper hurdle is estimated.

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

Deaton, A., and M. Irish. 1984. Statistical models for zero expenditures in household budgets. *Journal of Public Economics* 23: 59–80.

Dionne, G., M. Artís, and M. Guillén. 1996. Count data models for a credit scoring system. *Journal of Empirical Finance* 3: 303–325.

Dong, D., and H. M. Kaiser. 2008. Studying household purchasing and nonpurchasing behaviour for a frequently consumed commodity: Two models. *Applied Economics* 40: 1941–1951.

Garcia, B. 2013. Implementation of a double-hurdle model. *Stata Journal* 13: 776–794.

Heckman, J. J. 1979. Sample selection bias as a specification error. *Econometrica* 47: 153–161.

Jones, A. M. 1989. A double-hurdle model of cigarette consumption. *Journal of Applied Econometrics* 4: 23–39.

List, J. A. 2007. On the interpretation of giving in dictator games. *Journal of Political Economy* 115: 482–493.

McDowell, A. 2003. From the help desk: Hurdle models. *Stata Journal* 3: 178–184.

Moffatt, P. G. 2005. Hurdle models of loan default. *Journal of the Operational Research Society* 56: 1063–1071.

Smith, M. D. 2003. On dependency in double-hurdle models. *Statistical Papers* 44: 581–595.

Tobin, J. 1958. Estimation of relationships for limited dependent variables. *Econometrica* 26: 24–36.

Train, K. E. 2009. *Discrete Choice Methods with Simulation*. 2nd ed. Cambridge: Cambridge University Press.

About the authors

Christoph Engel is a director of the Max Planck Institute for Research on Collective Goods, Bonn, Germany.

Peter Moffatt is a professor of econometrics in the School of Economics at the University of East Anglia, Norwich, UK.