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 (2015)
15, Number 1, pp. 3–20

twopm: Two-part models

Federico Belotti
Centre for Economic and International Studies
University of Rome Tor Vergata
Rome, Italy
federico.belotti@uniroma2.it

Partha Deb
Hunter College and Graduate Center, CUNY
New York, NY
and National Bureau of Economic Research
Cambridge, MA
partha.deb@hunter.cuny.edu

Willard G. Manning[1]
University of Chicago
Chicago, IL

Edward C. Norton
University of Michigan
Ann Arbor, MI
and National Bureau of Economic Research
Cambridge, MA
ecnorton@umich.edu

Abstract. In this article, we describe `twopm`, a command for fitting two-part models for mixed discrete-continuous outcomes. In the two-part model, a binary choice model is fit for the probability of observing a positive-versus-zero outcome. Then, conditional on a positive outcome, an appropriate regression model is fit for the positive outcome. The `twopm` command allows the user to leverage the capabilities of `predict` and `margins` to calculate predictions and marginal effects and their standard errors from the combined first- and second-part models.

Keywords: st0368, twopm, two-part models, cross-sectional data, predictions, marginal effects

1 Introduction

Many outcomes (y_i) in empirical analyses are mixed discrete-continuous random variables. They have two basic statistical features: 1) $y_i \geq 0$, and 2) $y_i = 0$ is observed often enough that there are compelling substantive and statistical reasons for special treatment. In other words, because of the mass point at zero, a single index model for such data may not be desirable. The two-part model provides one approach to account for the mass of zeros. In the two-part model, a binary choice model is fit for the

[1] Willard G. Manning passed away in November 2014.

probability of observing a positive-versus-zero outcome. Then, conditional on a positive outcome, an appropriate regression model is fit for the positive outcome. In this article, we describe the command `twopm`, which can be used to conveniently fit two-part models and calculate predictions and marginal effects.

The two-part model has a long history. Since the 1970s, meteorologists have used versions of a two-part model for rainfall (Cole and Sherriff 1972; Todorovic and Woolhiser 1975; Katz 1977). Economists also used two-part models in the 1970s. Cragg (1971) developed the two-part model as an extension of the tobit model. The two-part model became widely used in health economics and health services research after a team at RAND Corporation used it to model health care expenditures in the context of the Health Insurance Experiment (Duan et al. 1984) (see Mihaylova et al. [2011] for more on the widespread use of the two-part model for health care cost data). Two-part models are also appropriate for other mixed discrete-continuous outcomes such as household-level consumption of food items and other consumables.

The two-part model has a commonly used counterpart for count data called the "hurdle" model (see Cameron and Trivedi [2013]; Jones [1989]; and Hilbe [2005]). We use the term "two-part" model to distinguish models for continuous outcomes from models for count data. Hilbe (2005) provides a command for hurdle models for count data.

The Heckman selection model (Heckman 1979), also referred to as the adjusted or generalized tobit (Amemiya 1985; Maddala 1983), is a multiple-index model that can also be fit as an alternative to the two-part model for mixed discrete-continuous outcomes. However, there are conceptual and statistical differences between the two models, and these have been debated extensively in the literature (see Poirier and Ruud [1981]; Duan et al. [1984]; Hay and Olsen [1984]; Manning, Duan, and Rogers [1987]; Hay, Leu, and Rohrer [1987]; Leung and Yu [1996]; and Dow and Norton [2003]).

A few points are important to reiterate here. First, despite their superficial similarity, the two-part model should not be viewed as being nested within the Heckman selection model and equivalent when there is no selection on unobservables. The two-part model does not make any assumption about the correlation between the errors of the binary and continuous equations. Second, from a conceptual standpoint, the zeros in the Heckman selection model denote censored values of the positive outcome, while zeros in the two-part model are true zeros. Third, Monte Carlo evidence shows that when the data are generated from the generalized tobit model without exclusion restrictions to identify the "zeros" equation, the two-part model generally produces better estimates of the conditional mean and of marginal effects than the correctly specified generalized tobit model: the reason is that the correlation parameter is very poorly identified. When data are generated from a generalized tobit with an exclusion restriction, the two-part model estimates of the conditional mean and marginal effects are not much worse than those obtained from the generalized tobit model. Because there are usually few situations in which exclusion restrictions distinguish the "zeros" equation from the "positives" equation, assuming that the analyst is interested in estimates of $E(y|\mathbf{x})$ and of $\partial E(y|\mathbf{x})/\partial \mathbf{x}$, the two-part model is almost always an adequate (if not superior on precision grounds) way to model mixed discrete-continuous outcomes if there are no exclusion restrictions.

The `twopm` package has several advantages compared with estimating the parameters of each part separately. First, it incorporates `svy:`, so it can adjust for complex survey design in the parameter estimates and the standard errors of those estimates. Complex survey design is common in large surveys; ignoring the survey structure can lead to biased estimates of population parameters. Second, it is easy to conduct joint statistical tests of parameters from both parts of the two-part model. Sometimes, it is appropriate to conduct a test of the joint significance of a variable that appears in both parts of the model. Third, it is easy to recover overall predicted values of the dependent variable and marginal effects for the combined model using the postestimation commands `predict` and `margins`. Note that these predicted values will be for the entire sample, as opposed to predictions based on the second (conditional) part of the model, which would typically be for the conditional sample of those with positive values. Fourth, our program produces estimates of predictions on the y scale (the raw scale), incorporating appropriate retransformation from the estimation scales when $\ln(y)$ is regressed using ordinary least squares (OLS) in the second part. Fifth, it automatically computes standard errors of predicted values and marginal effects and accounts for both parts of the model, any complex survey design, and robust standard errors based on the delta method. In terms of the amount of effort saved by the user, this is perhaps the most important feature of the `twopm` command. However, standard errors for margins and marginal effects in the model that require retransformation must be obtained via bootstrap methods.

2 Two-part models

A two-part model is a flexible statistical model specifically designed to deal with limited dependent variables. The distinguishing feature of these variables is that the range of values they may assume has a lower bound occurring in a fair number of observations. The basic framework is as follows. Suppose that there is an event that may or may not occur. When it does occur, one observes a positive random variable. When it does not, the observed outcome takes a zero value, thus becoming a zero-censored variable. For instance, in explaining individual annual health expenditure, the event is represented by a specific disease. If the illness occurs, then some not-for-free treatment will be needed, and a positive expense will be observed. In these situations, a two-part model allows the censoring mechanism and the outcome to be modeled to use separate processes. In other words, it permits the zeros and nonzeros to be generated by different densities as a special type of mixture model. The zeros are typically handled using a model for the probability of a positive outcome,

$$\phi(y>0) = \Pr(y>0|\mathbf{x}) = F(\mathbf{x}\boldsymbol{\delta})$$

where $\mathbf{x}$ is a vector of explanatory variables, $\boldsymbol{\delta}$ is the corresponding vector of parameters to be estimated, and F is the cumulative distribution function of an independent and identically distributed error term, typically chosen to be from extreme value (logit) or normal (probit) distributions. For the positives, the model is usually represented as

$$\phi(y|y>0,\mathbf{x}) = g(\mathbf{x}\boldsymbol{\gamma})$$

where $\mathbf{x}$ is a vector of explanatory variables, $\boldsymbol{\gamma}$ is the corresponding vector of parameters to be estimated, and g is an appropriate density function for $y|y>0$. The likelihood contribution for an observation can be written as

$$\phi(y) = \{1 - F(\mathbf{x}\boldsymbol{\delta})\}^{i(i=0)} \times \{F(\mathbf{x}\boldsymbol{\delta})g(\mathbf{x}\boldsymbol{\gamma})\}^{i(y>0)}$$

where $i(.)$ denotes the indicator function. Then, the log-likelihood contribution is

$$\ln\{\phi(y)\} = i(i=0)\ln\{1 - F(\mathbf{x}\boldsymbol{\delta})\} + i(i>0)[\ln\{F(\mathbf{x}\boldsymbol{\delta})\} + \ln\{g(\mathbf{x}\boldsymbol{\gamma})\}]$$

Because the $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$ parameters are additively separable in the log-likelihood contribution for each observation, the models for the zeros and the positives can be estimated separately.

Note that the overall mean can be written as the product of expectations from the first and second parts of the model, as follows:

$$E(y|\mathbf{x}) = \Pr(y>0|\mathbf{x}) \times E(y|y>0,\mathbf{x})$$

This is derived from the first principles of statistics decomposition of a joint distribution into marginal and conditional distributions. It is always true, with or without separability or specific F and $g(\cdot)$.

Estimating the parameters of the two-part model is straightforward. The threshold, $\Pr(y>0|\mathbf{x})$, is modeled using a regression model for binary outcomes such as the probit or logit. The positives, $E(y|y>0,\mathbf{x})$ or $g(y|y>0,\mathbf{x})$, where $g(\cdot)$ denotes a density function, are modeled using a regression framework for a continuous outcome; for example, they can be modeled using OLS regression or a generalized linear model (GLM). The second part is commonly modeled by OLS regression, with or without a transformation applied to $y|y>0$. It is straightforward to use OLS regression specified as $y = \mathbf{x}\boldsymbol{\gamma}+\varepsilon$ to estimate the second part. But, in many applications, and ubiquitous in the health economics and health services literature, the second part is specified as OLS regression of $\ln(y|y>0,\mathbf{x})$ written as $\ln(y) = \mathbf{x}\boldsymbol{\gamma}+\varepsilon$. In that case, if ε is independent and identically normally distributed, then

$$E(y|y>0,\mathbf{x}) = e^{\mathbf{x}\boldsymbol{\gamma}} \times e^{0.5\sigma^2} \tag{1}$$

where σ^2 is the variance of the distribution of ε; that is, it is the variance of the error on the log scale. If ε is not normally distributed but it is homoskedastic, then Duan (1983) showed that

$$E(y|y>0,\mathbf{x}) = e^{\mathbf{x}\boldsymbol{\gamma}} \times E\left(e^{\varepsilon}\right) \tag{2}$$

More recently, researchers have used the GLM framework (McCullagh and Nelder 1989) to model $(y|y>0,\mathbf{x})$ using a nonlinear transformation of a linear index function directly. Then

$$E(y|y>0,\mathbf{x}) = g^{-1}(\mathbf{x}\boldsymbol{\gamma})$$

where g is the link function in the GLM. Other approaches such as regressions with Box–Cox transformations and quantile regressions may also be used (not available in `twopm`).

The error terms in the two equations do not need to be independent to get consistent estimates of the parameters $\boldsymbol{\delta}$ and $\boldsymbol{\gamma}$. There is a misconception, especially in the early literature, that the two-part model assumes independence of binary outcomes and is conditional on positive, continuous outcomes. Also note that in the description above, the vector of covariates **x** is the same in both parts of the model. Although this is likely in most applications, sometimes, there may be legitimate theoretical (conceptual) or statistical reasons for using different independent variables in the two equations. For completeness, `twopm` has a syntax that allows for different covariates in each equation, but we do not generally recommend its use without appropriate justification.

Predictions of y_i, $(\widehat{y}_i|\mathbf{x}_i)$ can be constructed by multiplying predictions from each part of the model, observation by observation; that is,

$$\widehat{y}_i|\mathbf{x}_i = (\widehat{p}_i|\mathbf{x}_i) \times (\widehat{y}_i|y_i > 0, \mathbf{x}_i) \tag{3}$$

where $\widehat{p}_i|\mathbf{x}_i$ is the predicted probability that $y_i > 0$. Predictions for each part, confidence intervals for those predictions, and marginal effects of covariates on the outcomes in each part can be computed with existing commands. While one can construct overall predictions and marginal effects with a few lines of code, `twopm` makes it very easy to calculate them with the standard postestimation commands `predict` and `margins`. Unless retransformation is required, `predict` and `margins` produce standard errors of these predictions or marginal effects by using the delta method. When postestimation retransformation is required, `bootstrap` can be used with `predict` and `margins` to obtain standard errors.

Note that `margins` calls the prediction programs associated with the estimation command; that is, using `margins` following `twopm` calls `predict`, which in turn calls our program to calculate predictions of y based on (3).

3 The twopm command

`twopm` fits two-part models with logit and probit specifications for the first part and OLS [on y and on $\ln(y)$] and GLM regression for the second part. `twopm` can be specified using one of two syntaxes. The first syntax automatically specifies the same regressors (and functional forms in the index) in the first and second parts and is generally recommended. The second syntax allows the user to specify different regressors in the first and second parts. Although not generally recommended, there may be theoretically or statistically motivated situations where such a model may be applicable.

3.1 Syntax

The syntax for using `twopm` with specification of the same regressors in the first and second parts is

`twopm` *depvar* [*indepvars*] [*if*] [*in*] [*weight*] `,` `firstpart(`*f_options*`)`
`secondpart(`*s_options*`)` [`vce(`*vcetype*`)` `robust` `cluster(`*clustvar*`)` `suest`
`level(`*#*`)` `nocnsreport` *display_options*]

Syntax for using `twopm` with specification of different regressors in the first and second parts is

`twopm` *equation1 equation2* [*if*] [*in*] [*weight*] `,` `firstpart(`*f_options*`)`
`secondpart(`*s_options*`)` [`vce(`*vcetype*`)` `robust` `cluster(`*clustvar*`)` `suest`
`level(`*#*`)` `nocnsreport` *display_options*]

where *equation1* and *equation2* are specified as

`(`*depvar* [`=`] [*indepvars*]`)`

Note that *indepvars* may contain factor variables, and *depvar* and *indepvars* may contain time-series operators. `iweight`s, `aweight`s, and `pweight`s are allowed. `twopm` may be used with the `svy:` and `bootstrap` prefixes.

3.2 Options

`firstpart(`*f_options*`)` specifies the first part of the model for a binary outcome. It should be `logit` or `probit`. Each can be specified with its options except `vce()`, which should be specified as a `twopm` option. See the manual entries for [R] **logit** and [R] **probit**. `firstpart()` is required.

`secondpart(`*s_options*`)` specifies the second part of the model for a positive outcome. It should be `regress` or `glm`. Each can be specified with its options except `vce()`, which should be specified as a `twopm` option. See the manual entries for [R] **regress** and [R] **glm**. `secondpart()` is required.

`vce(`*vcetype*`)` specifies the type of standard error reported, including types that are derived from asymptotic theory, that are robust to some kinds of misspecification, that allow for intragroup correlation, and that use bootstrap or jackknife methods; see [R] ***vce_option***.

`vce(conventional)`, the default, uses the conventionally derived variance estimators for the first and second part of the model.

Note that options related to the variance estimators for both parts must be specified using `vce(`*vcetype*`)` in the `twopm` syntax. Specifying `vce(robust)` is equivalent to specifying `vce(cluster` *clustvar*`)`.

`robust` is the synonym for `vce(robust)`.

`cluster(`*clustvar*`)` is the synonym for `vce(cluster` *clustvar*`)`.

`suest` combines the estimation results of the first and second parts of the model to derive a simultaneous (co)variance matrix of the sandwich or robust type. Typical applications of `suest` are tests for cross-part hypotheses using `test` or `testnl`.

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

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

display_options: `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`; see [R] **estimation options**.

3.3 Postestimation

`predict` [*type*] *newvar* [*if*] [*in*] `,` [{`normal`|`duan`} `scores` `nooffset`]

and

`predict` [*type*] {*stub*`*`|*newvar1* ... *newvarq*} [*if*] [*in*] `,` `scores`

calculate predicted values or estimates of $E(y|x)$ and equation-level scores, respectively. While the first syntax is available both in and out of sample, type `predict` ... `if e(sample)` if predictions are wanted only for the estimation sample and if the second syntax for equation-level scores is restricted to the estimation sample. For predicted values estimated after the second-part regression of $\ln(y|y > 0)$, the following options are available:

`normal` uses normal theory retransformation to obtain fitted values. Either `normal` or `duan` must be specified when a linear regression of the log of the second-part outcome is estimated.

`duan` uses Duan's (1983) smearing retransformation to obtain fitted values. Either `normal` or `duan` must be specified when a linear regression of the log of the second-part outcome is estimated.

`scores` creates a score variable for each part in the model. Because the score for the second part of the model makes sense only for the estimation subsample (where $Y > 0$), the calculation is automatically restricted to the estimation subsample.

`nooffset` specifies that the calculation should be made ignoring any offset or exposure variable specified when fitting the model. This may be used with most statistics.

If neither the `offset(`*varname*`)` option nor the `exposure(`*varname*`)` option is specified when fitting the model, specifying `nooffset` does nothing.

4 Examples

We show two examples of two-part models for total annual health care expenditures using the medical expenditure panel survey 2004 data. We use two common versions of the two-part model to estimate predicted values of total expenditures and to calculate marginal or incremental effects of age and gender. In the first example, we fit a probit model in the first part and a GLM with the log link and gamma distribution for the second part. In the second example, we fit a logit model in the first part and an OLS regression with a logged dependent variable for the second part. We limit the covariates to just age and gender. The `twopm` command is compatible with complex survey commands, so after reading in the data, we set up the data for survey commands using `svyset`.

```
. * Use MEPS data on health care expenditures
. use http://www.econometrics.it/stata/data/meps_ashe_subset5
(MEPS04 date with edits)

. svyset [pweight=wtdper], strata(varstr) psu(varpsu)

      pweight: wtdper
          VCE: linearized
  Single unit: missing
     Strata 1: varstr
         SU 1: varpsu
        FPC 1: <zero>
```

After adjusting for the complex survey design, we see that the mean of health care expenditures is $3,839, with nearly 18% having a value of 0. The mean age is about 46 (range from 18 to 85) and just over half of participants are women.

```
. * Summarize data
. svy: mean exp_tot age female
(running mean on estimation sample)

Survey: Mean estimation

Number of strata =     203       Number of obs    =      19386
Number of PSUs   =     448       Population size  =  187973715
                                 Design df        =        245

─────────────┬──────────────────────────────────────────────────
             │             Linearized
             │       Mean   Std. Err.     [95% Conf. Interval]
─────────────┼──────────────────────────────────────────────────
     exp_tot │   3838.939   99.94525      3642.078    4035.801
         age │   45.79115   .2293769      45.33935    46.24295
      female │   .5201957   .0031165      .5140571    .5263343
─────────────┴──────────────────────────────────────────────────
```

4.1 Probit with GLM with log link and gamma distribution

Here we provide the command to estimate the parameters of the two-part model with a probit in the first part and a GLM with the log link and gamma distribution in the second part, taking into account the complex survey design.

```
. * Two-part model, with probit first part and GLM second part
. svy: twopm exp_tot c.age i.female, firstpart(probit)
> secondpart(glm, family(gamma) link(log))
(running twopm on estimation sample)
Survey data analysis
Number of strata   =       203                  Number of obs      =      19386
Number of PSUs     =       448                  Population size    =  187973715
                                                Design df          =        245
                                                F(   2,    244)    =     671.26
                                                Prob > F           =     0.0000

------------------------------------------------------------------------------
             |             Linearized
     exp_tot |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
probit       |
         age |   .0250999    .000793    31.65   0.000     .0235379    .0266618
    1.female |    .564196   .0271783    20.76   0.000     .5106631    .6177289
       _cons |  -.2386055   .0389997    -6.12   0.000    -.3154229   -.1617881
-------------+----------------------------------------------------------------
glm          |
         age |   .0287867   .0012973    22.19   0.000     .0262314    .0313421
    1.female |   .1995253   .0538871     3.70   0.000     .0933842    .3056665
       _cons |    6.80357    .086506    78.65   0.000      6.63318     6.97396
------------------------------------------------------------------------------
```

The estimated coefficients for `age` and `female` are positive in both parts and statistically significant at the 1% level. Both the probability of spending and the amount of spending conditional on any spending increase with age. Women are more likely than men to spend at least $1, and, conditional on spending any amount, they are more likely to spend more than men. In this simple example, we have not controlled for or tested for heteroskedasticity.

We can use the `margins` command as a postestimation command to predict the total spending. The predicted total spending is about $3,870 per person per year, which is relatively close to the actual average of $3,839.

```
. * Overall conditional mean
. margins
Predictive margins                                Number of obs      =      19386
Model VCE    : Linearized
Expression   : twopm combined expected values, predict()

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   3870.714   94.98674    40.75   0.000     3684.544    4056.885
------------------------------------------------------------------------------
```

Next, we show the marginal (or incremental) effects for the combined probit and GLM version of the two-part model. The marginal effect of age averages $128 per year of age, and women spend more than men by about $1,140. Note that if a covariate had opposite signs in each part of the model, then it would be possible for the joint test of significance of the coefficients to be statistically significant, along with the overall marginal effect being insignificant (although that is not the case here).

```
. * Marginal effects, averaged over the sample
. margins, dydx(*)

Average marginal effects                        Number of obs     =      19386
Model VCE    : Linearized

Expression   : twopm combined expected values, predict()
dy/dx w.r.t. : age 1.female

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   127.8325   6.372966    20.06   0.000     115.3417    140.3232
    1.female |   1139.541   186.7794     6.10   0.000     773.4597    1505.621
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

Because the marginal effects vary over the life course, we computed marginal effects conditional at four ages (20, 40, 60, and 80). When we calculate the marginal effects over the life course, we see that the marginal effects of both age and gender increase with age. For example, although women spend more than men at all ages, this difference is much greater for elderly women than for young women. This is due to the assumed log link in GLM, even with a simple linear specification of age.

```
. * Marginal effects at different ages
. margins, dydx(*) at(age=(20(20)80))

Conditional marginal effects                    Number of obs     =      19386
Model VCE    : Linearized

Expression   : twopm combined expected values, predict()
dy/dx w.r.t. : age 1.female

1._at        : age             =          20

2._at        : age             =          40

3._at        : age             =          60

4._at        : age             =          80

------------------------------------------------------------------------------
             |            Delta-method
             |      dy/dx   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
age          |
         _at |
          1  |   51.35857   1.357531    37.83   0.000     48.69786    54.01929
          2  |   95.64313   3.140771    30.45   0.000     89.48733    101.7989
          3  |    169.311    9.60291    17.63   0.000     150.4896    188.1324
          4  |   295.7016   24.66708    11.99   0.000      247.355    344.0482
-------------+----------------------------------------------------------------
1.female     |
         _at |
          1  |   589.6436   60.45588     9.75   0.000     471.1522    708.1349
          2  |   942.5697    127.344     7.40   0.000     692.9801    1192.159
          3  |   1431.437   260.9784     5.48   0.000     919.9285    1942.945
          4  |   2228.771   505.6082     4.41   0.000     1237.797    3219.745
------------------------------------------------------------------------------
Note: dy/dx for factor levels is the discrete change from the base level.
```

It is often of interest to know whether a covariate is jointly significant in both parts of the two-part model. In this example, age and gender are statistically significant in each part, so it is no surprise that they are each jointly significant in both parts.

```
. * Test whether coefficients on interaction terms are jointly zero
. test age

Adjusted Wald test

 ( 1)  [probit]age = 0
 ( 2)  [glm]age = 0

       F(  2,   244) =  803.99
            Prob > F =    0.0000

. test 1.female

Adjusted Wald test

 ( 1)  [probit]1.female = 0
 ( 2)  [glm]1.female = 0

       F(  2,   244) =  226.39
            Prob > F =    0.0000
```

When `twopm` is used together with the `svy:` prefix (and the default option for the (co)variance matrix `vce(linearized)`), a simultaneous "linearized" (co)variance matrix of the sandwich or robust type is automatically estimated. This ensures that hypotheses involving parameters across both parts can be correctly tested with `test` or `testnl`. When estimation is performed without the `svy:` prefix and cross-part hypotheses are of interest, we suggest using the `suest` option within `twopm`. This option produces a simultaneous (co)variance matrix of the sandwich or robust type; thus `test` (or `testnl`) will use the correct formula to perform the Wald test (see [R] **suest**).

4.2 Logit with OLS with logged dependent variable

Next, we provide an example using another common model, the two-part model with logit in the first part and OLS with log-transformed y in the second part. For the retransformation to the raw scale, we do not impose the restrictive assumption that the log-scale errors have a normal distribution. This assumption is often wrong and can lead to widely biased estimates of the conditional mean and marginal effects. Instead, we use Duan's (1983) smearing estimator. The `twopm` command automatically calculates the smearing estimate for use in postestimation commands.

In this example, we do not control for complex survey design. When one uses bootstrapping (which is necessary in this model with retransformation), the simple way of bootstrapping is incorrect. Here we focus on the importance of bootstrapping to account for the uncertainty in the estimated retransformation parameter.

```
. * Two-part model, with logit first part and OLS second part
. twopm exp_tot c.age i.female, firstpart(logit) secondpart(regress, log)

Fitting logit regression for first part:

Iteration 0:   log likelihood = -9062.9759
Iteration 1:   log likelihood = -8139.4972
Iteration 2:   log likelihood = -8062.7898
Iteration 3:   log likelihood = -8062.5899
Iteration 4:   log likelihood = -8062.5899

Fitting OLS regression for second part:

Two-part model
------------------------------------------------------------------------------
Log pseudolikelihood =  -37216.38                 Number of obs   =      19386

Part 1: logit
------------------------------------------------------------------------------
                                                  Number of obs   =      19386
                                                  LR chi2(2)      =    2000.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -8062.5899                       Pseudo R2       =     0.1104

Part 2: regress_log
------------------------------------------------------------------------------
                                                  Number of obs   =      15946
                                                  F(   2,  15943) =    1490.33
                                                  Prob > F        =     0.0000
                                                  R-squared       =     0.1575
                                                  Adj R-squared   =     0.1574
Log likelihood =  -29153.79                       Root MSE        =     1.5060
------------------------------------------------------------------------------
     exp_tot |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
logit        |
         age |    .047287   .0013987    33.81   0.000     .0445456    .0500284
             |
      female |
           1 |   .9684718   .0404988    23.91   0.000     .8890957    1.047848
             |
       _cons |  -.8706272   .0597288   -14.58   0.000    -.9876934   -.7535609
-------------+----------------------------------------------------------------
regress_log  |
         age |   .0358123    .000678    52.82   0.000     .0344835    .0371412
             |
      female |
           1 |   .3511679   .0242542    14.48   0.000     .3036305    .3987054
             |
       _cons |   5.329011    .037319   142.80   0.000     5.255867    5.402155
------------------------------------------------------------------------------
```

As before, the estimated coefficients for `age` and `female` are positive in both parts and statistically significant at the 1% level. The z statistics are similar in the logit and probit models, as expected. Again both the probability of spending and the amount of spending conditional on any spending increase with age. Women are more likely than men to spend at least $1, and (conditional on spending any amount) they spend more. Again we have not controlled for or tested for heteroskedasticity.

The predicted total expenditures from this model are considerably higher than in the model with probit and GLM. The predicted total expenditures are about $4,090 per person per year, which is far higher than the actual average. This calculation uses Duan (1983) smearing as part of the retransformation of the second part.

```
. * Overall conditional mean
. margins, predict(duan) post
Warning: cannot perform check for estimable functions.

Predictive margins                              Number of obs   =      19386

Expression   : twopm combined expected values, predict(duan)

------------------------------------------------------------------------------
             |            Delta-method
             |     Margin   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   4090.519    59.4288    68.83   0.000     3974.041    4206.998
------------------------------------------------------------------------------
```

Alternatively, we could have created a variable for the conditional mean for each observation using `predict yhat_duan, duan`.

Note that `margins` does not produce the correct standard errors for estimates when using retransformation. More specifically, while `margins` takes the uncertainty of parameter estimates into account in the index function for each part of the model, it does not account for estimation of σ^2 in (1) or $E(e^{\varepsilon})$ in (2). Although the `margins` command automatically computes the unconditional marginal effects after running `twopm`, the default delta-method standard errors are incorrect and will generally be too small. Therefore, after fitting a log OLS model in the second part, one must calculate standard errors and confidence intervals for `margins` using a nonparametric bootstrap.

The following is a simple program to bootstrap the standard errors for `margins`:

```
. * Overall conditional mean
. capture program drop Ey_boot

. program define Ey_boot, eclass
  1.     twopm exp_tot c.age i.female, firstpart(logit) secondpart(regress, log)
  2.     margins, predict(duan) nose post
  3. end

. bootstrap _b, seed(14) reps(1000): Ey_boot
(running Ey_boot on estimation sample)

Bootstrap replications (1000)
───┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
..................................................   850
..................................................   900
..................................................   950
..................................................  1000

Predictive margins                              Number of obs      =      19386
                                                Replications       =       1000

─────────────┬────────────────────────────────────────────────────────────────
             │   Observed   Bootstrap                         Normal-based
             │      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
─────────────┼────────────────────────────────────────────────────────────────
       _cons │   4090.519    97.54505    41.93   0.000     3899.335    4281.704
─────────────┴────────────────────────────────────────────────────────────────
```

The bootstrapped standard errors are roughly twice as large as the delta-method standard errors. In our experience, ignoring the uncertainty in the retransformation factor will bias the standard errors downward by a large amount, as in this example.

For the marginal effects, we again need to bootstrap the standard errors when using `margins`. In the two-part model with the logit and OLS with $\ln(y)$, age has a marginal effect of about $165 per year, while female has an incremental effect of almost $1,800.

```
. * Marginal effects, averaged over the sample
. capture program drop dydx_boot

. program define dydx_boot, eclass
  1.     twopm exp_tot c.age i.female, firstpart(logit) secondpart(regress, log)
  2.     margins, dydx(*) predict(duan) nose post
  3. end

. bootstrap _b, seed(14) reps(1000): dydx_boot
(running dydx_boot on estimation sample)

Bootstrap replications (1000)
────┼─── 1 ───┼─── 2 ───┼─── 3 ───┼─── 4 ───┼─── 5
..................................................    50
..................................................   100
..................................................   150
..................................................   200
..................................................   250
..................................................   300
..................................................   350
..................................................   400
..................................................   450
..................................................   500
..................................................   550
..................................................   600
..................................................   650
..................................................   700
..................................................   750
..................................................   800
..................................................   850
..................................................   900
..................................................   950
..................................................  1000

Average marginal effects                        Number of obs      =     19386
                                                Replications       =      1000

──────────────────────────────────────────────────────────────────────────────
             │   Observed   Bootstrap                         Normal-based
             │      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
─────────────┼────────────────────────────────────────────────────────────────
         age │   165.6376   5.193646    31.89   0.000     155.4583     175.817
    1.female │   1784.333   96.14553    18.56   0.000     1595.892    1972.775
──────────────────────────────────────────────────────────────────────────────
```

The different results demonstrate that the model used does matter. However, without further testing, it is unclear which model performs better in a statistical sense. We believe that using the two-part model can make a substantial difference, as can the retransformation approach for $\ln(y)$ models, as Duan (1983) showed. Both are likely sources of the differences between estimates in our examples.

5 Discussion

This version of twopm considers only a subset of two-part models where the positive outcomes are continuous. It does not deal with discrete or count outcomes. twopm allows for modeling of the second part using OLS (regress) or GLM (glm) but not numerous other plausible models for continuous outcomes, such as regressions with Box–Cox transformations (boxcox), quantile regressions (qreg), and other approaches available in user-written packages.

The two-part model is typically specified using the same set of covariates in both parts, and this is how we have specified our examples. However, this restriction is generally not required for all two-part model applications. The issue is not just about the same variables appearing in each part; model selection (with suitable safeguards against overfitting) may suggest different functional forms for variables in the index functions. For example, income may be either income or income and income^2, or ln(income). Alternatively, in our example, we used `age` and `female`, but a more adequate function may involve interactions and polynomials, which could vary by model part. One can still obtain marginal effects of `age` and `female` without restricting the functional form to be the same.

When the second part of the two-part model is modeled using OLS regression of $\ln(y)$, a retransformation is required to go from $\widehat{\ln(y)}$ to $\widehat{y}$. `twopm` provides retransformations based on homoskedastic, normally distributed errors and a nonparametric approach by Duan (1983) that also assumes homoskedastic errors. But heteroskedasticity is common in this context, and the retransformations based on homoskedastic errors are not consistent. Because of the complexity of dealing with heteroskedastic retransformations, we have not allowed for this possibility. We suggest users consider the gamma GLM with log link as an alternative for consistent estimation of coefficients, predictions, and marginal effects.

As with all estimation approaches, we suggest checking the specification of the two-part model to see whether the specification is appropriate for the given data. The fit for each of two equations for the probability of any use or expenditure and the level of use or expenditures can be assessed with conventional tests and approaches in the literature as well as with link (Pregibon 1980) and regression-equation specification error tests (Ramsey 1969). But the overall fit of the two parts combined has a more limited set of checks available. The `twopm` postestimation commands provide predictions that can be used to calculate various tests, including the modified Hosmer–Lemeshow test (Hosmer and Lemeshow 1980) and the Pearson correlation test as implemented in Manning, Basu, and Mullahy (2005).

6 Acknowledgments

We thank Vincenzo Atella for stimulating discussions and thank Tom Weichle, Andrea Piano Mortari, Joanna Kopinska, and Valentina Conti for giving `twopm` a thorough workout and for offering insightful suggestions. We also thank an anonymous referee for comments that improved the article.

7 References

Amemiya, T. 1985. *Advanced Econometrics*. Cambridge: Harvard University Press.

Cameron, A. C., and P. K. Trivedi. 2013. *Regression Analysis of Count Data*. 2nd ed. Cambridge: Cambridge University Press.

Cole, J. A., and J. D. F. Sherriff. 1972. Some single- and multi-site models of rainfall within discrete time increments. *Journal of Hydrology* 17: 97–113.

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

Dow, W. H., and E. C. Norton. 2003. Choosing between and interpreting the Heckit and two-part models for corner solutions. *Health Services and Outcomes Research Methodology* 4: 5–18.

Duan, N. 1983. Smearing estimate: A nonparametric retransformation method. *Journal of the American Statistical Association* 78: 605–610.

Duan, N., W. G. Manning, Jr., C. N. Morris, and J. P. Newhouse. 1984. Choosing between the sample-selection model and the multi-part model. *Journal of Business and Economic Statistics* 2: 283–289.

Hay, J. W., R. Leu, and P. Rohrer. 1987. Ordinary least squares and sample-selection models of health-care demand: Monte Carlo comparison. *Journal of Business and Economic Statistics* 5: 499–506.

Hay, J. W., and R. J. Olsen. 1984. Let them eat cake: A note on comparing alternative models of the demand for medical care. *Journal of Business and Economic Statistics* 2: 279–282.

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

Hilbe, J. 2005. hplogit: Stata module to estimate Poisson-logit hurdle regression. Statistical Software Components S456405, Department of Economics, Boston College. http://ideas.repec.org/c/boc/bocode/s456405.html.

Hosmer, D. W., Jr., and S. Lemeshow. 1980. Goodness of fit tests for the multiple logistic regression model. *Communications in Statistics—Theory and Methods* 9: 1043–1069.

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

Katz, R. W. 1977. Precipitation as a chain-dependent process. *Journal of Applied Meteorology* 16: 671–676.

Leung, S. F., and S. Yu. 1996. On the choice between sample selection and two-part models. *Journal of Econometrics* 72: 197–229.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.

Manning, W. G., A. Basu, and J. Mullahy. 2005. Generalized modeling approaches to risk adjustment of skewed outcomes data. *Journal of Health Economics* 24: 465–488.

Manning, W. G., N. Duan, and W. H. Rogers. 1987. Monte Carlo evidence on the choice between sample selection and two-part models. *Journal of Econometrics* 35: 59–82.

McCullagh, P., and J. A. Nelder. 1989. *Generalized Linear Models*. 2nd ed. London: Chapman & Hall/CRC.

Mihaylova, B., A. Briggs, A. O'Hagan, and S. G. Thompson. 2011. Review of statistical methods for analysing healthcare resources and costs. *Health Economics* 20: 897–916.

Poirier, D. J., and P. A. Ruud. 1981. On the appropriateness of endogenous switching. *Journal of Econometrics* 16: 249–256.

Pregibon, D. 1980. Goodness of link tests for generalized linear models. *Journal of the Royal Statistical Society, Series C* 29: 15–23.

Ramsey, J. B. 1969. Tests for specification errors in classical linear least-squares regression analysis. *Journal of the Royal Statistical Society, Series B* 31: 350–371.

Todorovic, P., and D. A. Woolhiser. 1975. A stochastic model of n-day precipitation. *Journal of Applied Meteorology* 14: 17–24.

About the authors

Federico Belotti is a researcher at the Centre for Economics and International Studies of the University of Rome Tor Vergata.

Partha Deb is a professor of economics at Hunter College and the Graduate Center, CUNY, and a research associate at the National Bureau of Economic Research.

Will Manning was an influential health economist, starting with his pioneering work on the RAND Health Insurance Experiment in the 1970s. With others at RAND, he advocated moving away from tobit and sample-selection models to deal with distributions of dependent variables that had a large mass at zero. The two-part model, in all its forms, is now the dominant model for health care costs and expenditures. He continued to push the field of health econometrics by helping to develop new methods and advocating the work of others who have found better ways of modeling health care and its costs. His passing is a great loss to the profession and to so many of us personally.

Edward C. Norton is a professor in the Department of Health Management and Policy and in the Department of Economics at the University of Michigan, and he is a research associate at the National Bureau of Economic Research.