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 3, pp. 627–644

Prediction in linear index models with endogenous regressors

Christopher L. Skeels
Department of Economics
University of Melbourne
Melbourne, Australia
Chris.Skeels@unimelb.edu.au

Larry W. Taylor
Department of Economics
Lehigh University
Bethlehem, PA
larry.taylor@lehigh.edu

Abstract. In this article, we examine prediction in the context of linear index models when one or more of the regressors are endogenous. To facilitate both within-sample and out-of-sample predictions, Stata offers the postestimation command `predict` (see [R] **predict**). We believe that the usefulness of the predictions provided by this command is limited, especially if one is interested in out-of-sample predictions. We demonstrate our point using a probit model with continuous endogenous regressors, although it clearly generalizes readily to other linear index models. We subsequently provide a program that offers one possible implementation of a new command, `ivpredict`, that can be used to address this shortcoming of `predict`, and we then illustrate its use with an empirical example.

Keywords: st0397, predict, probit, logit, ivprobit, prediction, linear index, endogenous regressors, ivpredict, out-of-sample prediction

1 Introduction

In a recent article, Skeels and Taylor (2014) explore current practice for prediction in linear simultaneous-equations models. They demonstrate that a predictor based on replacing unknown coefficients with consistent estimators thereof was itself inconsistent for any population quantity of interest. Here we extend this result to a broader class of linear index models when one or more of the regressors are endogenous. We focus on the case of a probit model with a continuous endogenous regressor, although the intuition clearly extends further (for example, to count-data models). Of particular concern is the implication for current econometric practice in terms of Stata.

To facilitate both within-sample and out-of-sample predictions, Stata offers the intelligent postestimation command `predict` (see [R] **predict**) ("intelligent" in that exactly what `predict` returns is context dependent). For example, following least-squares estimation of the linear regression model

$$y = X\beta + \varepsilon$$

using the `regress` command (see [R] **regress**), the predictions generated by the `predict` command will take the form $X\widehat{\beta}$, where $\widehat{\beta}$ denotes the estimated value of β. Conversely, following `probit` (see [R] **probit**) or `logit` (see [R] **logit**) estimation of a binary dependent variable model of the general form $P(y = 1) = F(X\beta)$, `predict` can return predictions of either $X\widehat{\beta}$ or $F(X\widehat{\beta})$, where $F(\cdot)$ denotes either the standard normal or the logistic distribution function.[1] The common feature of these two examples is that, under standard assumptions, the predictions returned by `predict` are, by design and by default, realizations of consistent predictors for $E(y \mid X)$.[2] Such predictors are of interest because the conditional expectations are mean squared prediction-error minimizers. Moreover, because $E(y \mid X)$ is a regression function, the predictions are readily interpretable.

Unfortunately, when there are endogenous regressors, the linear index $X\widehat{\beta}$ is not a consistent predictor for $E(y \mid X)$ or for any other quantity related to the relevant regression function, calling into question the usefulness of the predictions returned by `predict`.[3] In the next section, we establish notation and derive population quantities of interest for a probit model with continuous endogenous regressors. In section 3, we explore different predictors that may arise with this model. In section 4, we explore various prediction options available in Stata and how they relate to the quantities defined in section 3. In section 5, we present a program that illustrates how to generate predictions consistent for population quantities of interest by augmenting the prediction powers of Stata. In section 6, we use this program in an empirical example. In section 7, we conclude with some discussion of the lessons to be learned.

2 A simultaneous probit model

Although the fundamental tenet of this article is applicable to any linear index model, a concrete example helps to illustrate the idea. Here we consider a particular probit model with continuous endogenous regressors.[4]

To begin, let $y_i^\star$ denote the ith observation on a latent endogenous variable, observed only up to sign; let Y_i' denote the ith observation on a set of n continuous endogenous regressors; and let X_i' and Z_i' denote the ith observations on sets of K_X included and K_Z excluded exogenous regressors, respectively, with $K = K_X + K_Z$ and $i = 1, \ldots, N$. The model is then characterized as being composed of

(i) one structural equation of interest,

$$y_i^\star = Y_i'\beta + X_i'\gamma + \epsilon_i \tag{1}$$

coupled with

1. In general, the latter behavior is probably the more useful and is the default.
2. More generally, the `predict` command aspires to generate predictions that are "numbers related to the $E(y_j \mid x_j)$" (StataCorp 2015, 1889).
3. Here we are concerned with the usefulness of the predictions as forecasts. It is understood that the quantities $X\widehat{\beta}$ may be calculated purely as intermediate steps in some other calculation. For instance, the generalized residuals that are used to construct various test statistics require the calculation of $X\widehat{\beta}$; see, for example, Pagan and Vella (1989) and Skeels and Vella (1999).
4. See Greene (2012, sec. 17.3.5) for an excellent discussion of this model and various approaches to it. See also Cameron and Trivedi (2005, sec. 16.8.2) and, especially, Cameron and Trivedi (2009, sec. 14.8).

(ii) that part of the reduced form required to complete the system

$$Y_i' = [X_i', Z_i']\Pi + V_i' \tag{2}$$

together with

(iii) an endogeneity assumption among the jointly normally distributed disturbances, which are otherwise uncorrelated across observations $i = 1, \dots, N$,

$$[\epsilon_i, V_i']' \sim N(0, \Sigma)$$

where

$$\Sigma = \begin{bmatrix} 1 & \rho'\Omega_{22}^{1/2} \\ \Omega_{22}^{1/2}\rho & \Omega_{22} \end{bmatrix} \tag{3}$$

is positive definite (written $\Sigma > 0$), with ρ denoting the vector of correlations between ϵ_i and V_i, and

(iv) an observation rule of the form

$$y_i = \begin{cases} 1, & \text{if } y_i^\star > 0, \\ 0, & \text{otherwise,} \end{cases} \qquad i = 1, \dots, N \tag{4}$$

Given the joint normality of ϵ_i and V_i, it follows that[5,6,7]

$$\text{vec}[y_i^\star, Y_i'] \sim N\left(\text{vec}\left[[X_i', Z_i']\Pi\beta + X_i'\gamma, [X_i', Z_i']\Pi\right], \begin{bmatrix} 1 & \beta' \\ 0 & I_n \end{bmatrix} \Sigma \begin{bmatrix} 1 & 0 \\ \beta & I_n \end{bmatrix}\right) \tag{5}$$

and the conditional distribution of $y_i^\star$ given Y_i can be shown to be (see, for example, Mardia, Kent, and Bibby [1979, theorem 3.2.4])

$$y_i^\star \mid Y_i \sim N\left(Y_i'\beta + X_i'\gamma + (Y_i' - [X_i', Z_i']\Pi)\Omega_{22}^{-1/2}\rho, 1 - \rho'\rho\right) \tag{6}$$

5. Here vec[·] denotes the usual vec operator of matrix algebra; see, for example, Searle (1982, 332).
6. Here, as throughout the article, we have not made the conditioning on the exogenous variables $[X_i', Z_i']$ notationally explicit because it is ubiquitous and should be taken as read.
7. We now have two mathematically equivalent statements of the model, namely, (1)–(4) and the joint distribution (5) coupled with the observation rule (4). The latter statement of the model is more convenient for our purposes here. For a discussion of the pros and cons of the two statements with the classical linear simultaneous-equations model, see Poskitt and Skeels (2008).

It follows immediately that[8]

$$E(y_i \mid Y_i) = \text{Prob}(y_i = 1 \mid Y_i) = \text{Prob}(y_i^\star > 0 \mid Y_i)$$
$$= \Phi\left(\frac{Y_i'\beta + X_i'\gamma + (Y_i' - [X_i', Z_i']\Pi)\Omega_{22}^{-1/2}\rho}{\sqrt{1-\rho'\rho}}\right) \tag{7}$$

where $\Phi(\cdot)$ denotes the standard normal distribution function.

Note that all the parameters of this model can be consistently estimated using maximum likelihood techniques. If we let the vector y and the matrix Y have rows y_i and Y_i', respectively, with $i = 1, \ldots, n$, and we observe that the joint density of the sample is

$$f(y, Y; \beta, \gamma, \rho, \Omega_{22}) = \prod_{i=1}^{n} \{\text{Prob}(y_i = 0 \mid Y_i)\}^{(1-y_i)}\{\text{Prob}(y_i = 1 \mid Y_i)\}^{y_i} g(Y_i)$$
$$= \prod_{i=1}^{n} \{1 - \text{Prob}(y_i = 1 \mid Y_i)\}^{(1-y_i)}\{\text{Prob}(y_i = 1 \mid Y_i)\}^{y_i} g(Y_i)$$

where $g(Y_i)$ denotes the marginal density of Y_i, then the log likelihood is[9]

$$\sum_{i=1}^{n} [(1-y_i)\ln\{1 - E(y_i \mid Y_i)\} + y_i \ln E(y_i \mid Y_i) + \ln g(Y_i)]$$

In section 3, we will explore some quantities of potential interest for prediction.

8. Equation (7) extends Greene (2012, eq. 17–32) to allow for multiple endogenous regressors.

9. This differs from the expression given by Greene (2012, 748), which exploits the symmetry of the normal distribution about its mean to write

$$1 - E(y_i \mid Y_i) = \Phi\left(-\frac{Y_i'\beta + X_i'\gamma + (Y_i' - [X_i', Z_i']\Pi)\Omega_{22}^{-1/2}\rho}{\sqrt{1-\rho'\rho}}\right)$$

and characterizes sign information by

$$2y_i - 1 = \begin{cases} 1, & \text{if } y_1 = 1, \text{ and} \\ -1, & \text{otherwise} \end{cases}$$

Combining these two results with (7) yields

$$\sum_{i=1}^{n} \{(1-y_i)\ln(1 - E(y_i \mid Y_i)) + y_i \ln E(y_i \mid Y_i)\}$$
$$= \sum_{i=1}^{n} \Phi\left((2y_i - 1)\left(\frac{Y_i'\beta + X_i'\gamma + (Y_i' - [X_i', Z_i']\Pi)\Omega_{22}^{-1/2}\rho}{\sqrt{1-\rho'\rho}}\right)\right)$$

which establishes the equivalence of the two results.

3 Prediction in the simultaneous probit model

In the standard probit model, composed of a latent model of the form

$$y_i^\star = X_i'\gamma^\star + \varepsilon_i \qquad \varepsilon_i \sim N(0, \sigma_\varepsilon^2) \tag{8}$$

together with the observation rule (4), it is standard practice to normalize the latent model to have unit variance. This reflects that it is impossible to measure the scale of a dependent variable that is unobservable, a problem that manifests itself in an inability to separately identify both $\gamma^\star$ and σ_ε^2. As is well known, the usual coefficient estimates are actually estimates of the ratio $\gamma^\star/\sigma_\varepsilon$.[10] Fortunately, if one is interested in $E(y_i \mid X_i) = \Phi(X_i'\gamma^\star/\sigma_\varepsilon)$, then this ratio is the required quantity. Letting $\widehat{\gamma}$ denote a consistent estimator for $\gamma^\star/\sigma_\varepsilon$, we conclude that $\widehat{y}_i = X_i'\widehat{\gamma}$ is probably not a very interesting predictor on its own (because of the scale issues) but that $\Phi(\widehat{y}_i)$ is.

Turning to the simultaneous probit model, our primary concern, we find the discussion is complicated by (1) not being a regression function, in contrast to (8). Consequently, consistent estimation of the coefficients of (8) does not yield a consistent predictor of the conditional mean of $y_i^\star$ given Y_i. Specifically, if $\widehat{\beta}$ and $\widehat{\gamma}$ are consistent estimators for β and γ, respectively, then

$$\widehat{y}_i = Y_i'\widehat{\beta} + X_i'\widehat{\gamma}$$

is not consistent for

$$E(y_i^\star \mid Y_i) = Y_i'\beta + X_i'\gamma + (Y_i' - [X_i', Z_i']\Pi)\Omega_{22}^{-1/2}\rho \tag{9}$$

In terms of (9), it is useful to distinguish between two cases: one where Y_i is known and the other where it is unknown and must be predicted. When Y_i is known, maximum likelihood can then be used to consistently estimate Π, Ω_{22}, and ρ using the two-step method discussed in Greene (2012, 747–750). Letting $\widehat{\Pi}$, $\widehat{\Omega}_{22}$, and $\widehat{\rho}$ denote these estimators, we see that

$$Y_i' - [X_i', Z_i']\Pi = Y_i' - [X_i', Z_i']\widehat{\Pi} + [X_i', Z_i'](\widehat{\Pi} - \Pi) = \widehat{V} + o_p(1)$$

and so a consistent predictor for the conditional mean of $y_i^\star$ given Y_i is

$$\widetilde{y}_i = Y_i'\widehat{\beta} + X_i'\widehat{\gamma} + \widehat{V}\widehat{\Omega}_{22}^{-1/2}\widehat{\rho} \tag{10}$$

Of course, if Y_i is unknown, as would typically be the case when forecasting, then (10) is not operational. Moreover, conditioning on Y_i when it is unknown makes no sense. Consequently, rather than working with the conditional distribution (6), we might reasonably base our prediction on the relevant marginal distribution, which can be obtained from (5) as

$$y_i^\star \sim N\left([X_i', Z_i']\Pi\beta + X_i'\gamma, \sigma_{y^\star}^2\right) \tag{11}$$

10. This is why, in the probit model, it is meaningful to interpret the signs and statistical significance of estimated coefficients but not their magnitudes.

where $\sigma^2_{y^\star} = [1, \beta']\Sigma[1, \beta']' = 1 + 2\rho'\Omega_{22}^{1/2}\beta + \beta'\Omega_{22}\beta$. That is, the relevant conditioning set is composed of the exogenous variables $[X_i', Z_i']$ alone. Defining

$$\widehat{Y}_i = [X_i', Z_i']\widehat{\Pi}$$

we see that $E\left(y_i^\star \mid X_i, Z_i\right)$ can be consistently predicted by

$$\breve{y}_i = \widehat{Y}_i'\widehat{\beta} + X_i'\widehat{\gamma} \tag{12}$$

One advantage of this formulation is that it is also available when Y_i is known, so $\breve{y}_i$ is also useful for within-sample predictions (c.f. Skeels and Taylor 2014).

Now suppose that the quantity of interest is the expected value of y_i, that is, the probability that $y_i^\star$ is positive. In this case, neither $\widetilde{y}_i$ nor $\breve{y}_i$ are adequate. Instead, for known Y_i, we have the predictor

$$\widehat{E}\left(y_i \mid Y_i\right) = \Phi\left(\widetilde{y}_i/\sqrt{1-\widehat{\rho}'\widehat{\rho}}\right)$$

whereas, conditioning on $[X_i', Z_i']$ alone, (11) yields

$$\widehat{E}\left(y_i\right) = \Phi(\breve{y}_i/\widehat{\sigma}_{y^\star}) \tag{13}$$

All the required parameters can be estimated using the two-step maximum-likelihood estimator mentioned above. However, for the purposes of prediction, all that is required are the various ratios of parameters. For example, the parameters required for $\widehat{E}\left(y_i\right)$ could be consistently estimated using a standard probit model with regressors $\widehat{Y}_i$ and X_i. Again, $\widehat{E}\left(y_i \mid Y_i\right)$ is operational only when Y_i is known, whereas $\widehat{E}\left(y_i\right)$ is always available. Note that $\Phi(\widehat{y}_i)$ is not consistent for any population quantity of interest.

4 Stata and prediction

Here we explore the behavior of the `predict` command when used in conjunction with various other commands for fitting probit models with endogenous regressors. As described in section 1, this exploration is complicated by the behavior of `predict` varying with circumstance. For example, to fit the model in section 3, Stata offers the `ivprobit` command (see [R] **ivprobit**), which can fit the model either by using maximum likelihood (the default) or by using Newey's (1987) two-step minimum chi-squared estimator with the `twostep` option. Both procedures allow the `xb` option (linear prediction, the default), while the maximum likelihood estimator also allows the `pr` option, which purportedly generates the probability of a positive outcome—that is, $\text{Prob}(y_i = 1 \mid Y_i) = E\left(y_i \mid Y_i\right)$—an option not available with the two-step estimator.

For both estimators, `predict` with the `xb` option works as advertised. That is, the predictions generated are of the form

$$\widehat{y}_i = Y_i'\widehat{\beta} + X_i'\widehat{\gamma}$$

where $\widehat{\beta}$ and $\widehat{\gamma}$ denote the estimates of β and γ, respectively, generated by `ivprobit`. However, as demonstrated in the previous section, it is not entirely clear why anyone should be interested in $\widehat{y}_i$ because it does not correspond to any population quantity of interest. Moreover, it is not clear how to interpret functions such as $\Phi(\widehat{y}_i)$, because this is not a consistent estimator of the regression function specified in (7) except, of course, in the special case $\rho = 0$. If $\rho = 0$, then simultaneity is obviously not a problem, and the use of any estimator designed to cater for it will be, at best, inefficient relative to the ordinary probit estimator.

One might argue that prediction of the simple-minded form discussed here is not of particular interest in most microeconometric applications and that the primary use of such predictions is to provide intermediate quantities for use in subsequent calculations. Although that may be true in certain cases, in others it is not. For example, the conditional-moment tests described in Pagan and Vella (1989) are based on the generalized residuals of Cox and Snell (1968), which are based on deviations from the relevant conditional expectations. Consequently, practitioners must be aware of exactly when they cannot rely on predictions of the form $\widehat{y}_i$.

Perhaps most surprising is the behavior of the `pr` option for `predict`, which is available with maximum likelihood estimation only. This option generates $\Phi(\widehat{y}_i)$, which does not correspond to any population quantity of interest. This is surprising because, at least for within-sample predictions, the correct quantities are necessarily calculated in construction of the log likelihood but are not returned as predictions.

For out-of-sample predictions, matters are complicated because Y_i is not observed and must be forecast. As previously explained, this suggests the use of $\breve{y}_i$ (or functions of it) to forecast population quantities of interest.

`ivprobit` is not the only available estimation command; several user-written commands are available.[11] Closest in functionality to `ivprobit` are the `cmp` (Roodman 2011) and `cdsimeq` (Keshk 2003) commands.[12] Several commands implement Amemiya's (1978) generalized least-squares estimator for probit models with endogenous regressors, including the `ivprob` command (Harkness 2001) and the `probitiv` command (Gelbach 1999).[13] These latter commands are for a slightly different model than that of section 2. The specific difference is that they are constructed for systems of structural equations characterized by the presence of the dichotomous endogenous variable as a regressor in the equations for the continuous endogenous variable, making this latter equation structural as well. Contrast this with (2), which is a reduced-form equation rather than a structural equation. In light of these differences, our treatment of these commands will be cursory at best.

11. The list of commands considered here should be considered indicative rather than exhaustive.
12. These are "closest" in that they handle the same model as `ivprobit`. `cdsimeq` is written specifically for the model discussed in this article, whereas `ivprobit` provides only a tiny subset of the functionality of `cmp`.
13. The authors of the various packages also refer to Newey (1987) and the discussion of Maddala (1983, 242–252) for further descriptions of what they are implementing. The discussion in this article corresponds to Maddala's (1983, 244–245) Model 3.

Neither `probitiv` nor `cdsimeq` produced predictions of any sort, with the `predict` command resulting in errors when used. With respect to prediction, the `ivprob` command behaves exactly as `ivprobit` with the `twostep` option, which is reasonable because it also is a two-step estimator. Hence, it produces linear predictions of the form $\widehat{y}_i$ but not predictions of the probability of a positive outcome. Finally, in our experience, `cmp` also generates the same prediction behavior as `ivprobit`. That said, `cmp` offers an option, `reducedform`, for use with the `predict` command, which reads as though it should use predictions $\widehat{Y}_i'$ in place of the actual Y_i' in constructing predictions, as suggested here. Unfortunately, that was not our experience, and we could observe no change in prediction behavior when this option was used. Consequently, we were unable to conclude that `cmp` could produce the predictions that we were seeking.[14]

5 A program for consistent predictions: ivpredict

At the suggestion of a referee, we offer a program, `ivpredict`, that complements the functionality provided by `predict` in linear index models with endogenous regressors. We offer the routine primarily as a proof of concept.[15] Notionally, however, the routine has applicability to a wide variety of linear index models, although we have not explored this beyond the probit model considered in this article. An example using this program in the context of an empirical application appears in section 6.

5.1 Implementation

The program is presented in table 1.

14. We acknowledge the possibility that we were using the option incorrectly, but we did invest considerable time in trying to make it work because it would have absolved us from needing to write the next section.
15. Obvious improvements to our humble offering would include the ability to pass options to some of the commands used within. We do not pursue this for two reasons. First, we would much rather Stata improve the interaction between `predict` and those estimation routines specifically designed to deal with endogenous regressors, thereby rendering `ivpredict` redundant. Second, if this is not to happen, at the very least, an `ivpredict`-like command should be developed by better programmers than us.

Table 1. `ivpredict.do`

```
capture program drop ivpredict
program ivpredict
  gettoken stub 0 : 0
  gettoken subcmd 0 : 0
  gettoken dependent 0 : 0
  gettoken endo rest : 0 , parse(" :")
  gettoken paren exog : rest
  gettoken inexog exexog : exog , match(check)
  matrix myb=e(b)
  capture matrix coef = myb[1,"`dependent´:"]
  if _rc !=0 {
     matrix coleq myb = `dependent´
     matrix coef = myb
  }
  *  Initialize variables returning predictions
  capture drop `stub´pr
  capture drop `stub´xb
  generate `stub´xb=0
  *  Create `stub´xb (predictions of the index)
  foreach var in `endo´ {
     capture drop p`var´
     regress `var´ `inexog´ `exexog´
     predict p`var´
     local pendo "`pendo´ p`var´"
     matrix coeff=coef[1,"`dependent´:`var´"]
     scalar coefff=det(coeff)
     capture replace `stub´xb = `stub´xb + p`var´ * coefff
     matrix drop coeff
     scalar drop coefff
  }
  foreach var in `inexog´ {
     matrix coeff=coef[1,"`dependent´:`var´"]
     scalar coefff=det(coeff)
     capture replace `stub´xb = `stub´xb + `var´ * coefff
     matrix drop coeff
     scalar drop coefff
  }
  *  Be careful with the constant
  capture matrix coeff=coef[1,"`dependent´:_cons"]
  if _rc == 0 {
     scalar coefff=det(coeff)
     capture replace `stub´xb = `stub´xb + coefff
     matrix drop coeff
     scalar drop coefff
     `subcmd´ `dependent´ `pendo´ `inexog´
     capture predict `stub´pr, pr
  }
  else {
     `subcmd´ `dependent´ `pendo´ `inexog´, noconstant
     capture predict `stub´pr, pr
  }
end
```

Like the `predict` command, the `ivpredict` program is used after an estimation command. Its first task is to gather coefficient estimates. It then attempts to produce predictions of the expected values of $y_i^\star$ and, if applicable, y_i using (12) and (13), respectively. That is, the predictions produced are conditioned on X_i' and Z_i' alone but not on Y_i'. Additionally, various intermediate results are also produced, as explained in section 5.2 below.

Two aspects of the code are worth mentioning. First, the construction of $\breve{y}_i$ would be done more directly using matrix computations. However, in the example in section 6, we exceeded `matsize`; so, for better or worse, we have explicitly rolled out the matrix calculation as a set of loops mimicking the sums

$$\breve{y} = \widehat{Y}\widehat{\beta} + X\widehat{\gamma} = \sum_{j=1}^{n} \widehat{Y}_j\widehat{\beta}_j + \sum_{k=1}^{K_X} X_k\widehat{\gamma}_k$$

where the subscripts refer to either columns of data matrices or rows of coefficient vectors.

Second, Stata displays inconsistent behavior in how different estimation commands present results when working with systems of equations. Suppose that rather than using `ivprobit`, the model in section 6 was fit using two-stage least squares with `ivregress` (see [R] **ivregress**), such as

```
ivregress 2sls ins $xlist2 (linc=$ivlist2)
```

Comparing the contents of `e(b)` immediately following each of these estimation commands reveals two related differences. `ivprobit` provides results for both the structural equation of interest and the reduced-form equations used to generate the instruments. In contrast, `ivregress` provides coefficient estimates only for the structural equation of interest.[16] One consequence of these differing quantities of results is that in the case of `ivprobit`, the column names of `e(b)` are decorated with equation names, whereas those for `ivregress` are not. This created some complications, which is why we have a somewhat-convoluted definition of `coef` in `ivpredict.do` (see table 1).[17]

5.2 Results

`ivpredict` returns the results of fitting each reduced equation (one per endogenous regressor) together with estimation results for (13) or its equivalent.

One new variable is always generated, with a name of the form `‘`*stub*`’xb`, where `‘`*stub*`’` is replaced with whatever text was supplied with `ivpredict`. This variable contains the

16. The behavior of `ivregress` is also shared by `ivreg2` (Baum, Schaffer, and Stillman 2002).
17. Our treatment of these different column-naming approaches is somewhat ad hoc and will break if `ivpredict` is used following an estimation command that adopts yet another approach. We have not encountered such a command, and so we believe that `ivpredict` stands as a proof of concept, although we certainly do not claim that our experimentation with different estimation commands has been exhaustive.

predictions $\breve{y}_i$ [equation (12)], which correspond to those predictions of the conditional expectation of the linear index that would be obtained using `predict` following the `ivprobit` command, except that Y_i' has been replaced by $\widehat{Y}_i'$ in construction of the index.

A second variable may also be created, containing either probabilities that the latent variable is positive or, equivalently, $\widehat{E}(y_i)$ [equation (13)], the probability of a "success". This variable will be called `‘stub’pr`. Note that this latter variable is produced only when modeling variables that are binary in nature by using commands such as `logit`, `probit`, and `ivprobit`. In particular, `ivpredict` will not produce probabilities of the form Prob($a < y_i < b$), such as can be obtained from `predict` following the use of `regress`.[18]

The important aspect of these various predictions is that they are predictions of the conditional expectations of the variables $y_i^\star$ and y_i, given all the instruments in the model (X_i' and Z_i'). In particular, we do not condition on Y_i'.

5.3 Use

We illustrate the use in the following two examples. For $i = 1, \ldots, n$, consider the model

$$y_i^\star = Y_{1i}\beta_1 + Y_{2i}\beta_2 + X_{1i}\gamma_1 + X_{2i}\gamma_2 + \epsilon_i \tag{14}$$

$$[Y_{1i}, Y_{2i}] = [X_{1i}, X_{2i}, Z_{1i}, Z_{2i}]\Pi + [V_{1i}, V_{2i}]$$

coupled with the observation rule (4). The coefficients in the structural equation are scalars, but Π is of dimension (4×2).

Example 1

The code

```
ivpredict mypredict probit y Y1 Y2 : (X1 X2) Z1 Z2
```

creates the two sets of predictions described in section 5.2, stored in the variables `mypredictxb` and `mypredictpr`.

Example 2

Now suppose that (14) is replaced in the model by

$$y_i^\star = Y_{1i}\beta_1 + Y_{2i}\beta_2 + \epsilon_i \tag{15}$$

18. As mentioned above, there is no mechanism for passing the bounds (a, b) to `ivpredict`.

Here the appropriate code to generate the predictions of interest is

`ivpredict mypredict probit` y Y_1 Y_2 `: ( )` Z_1 Z_2

Observe that although there are no exogenous variables included in (15), it is still necessary to include an empty list surrounded by parentheses. Again, the predictions will be stored in the newly created variables, `mypredictxb` and `mypredictpr`. More generally, the syntax of `ivpredict` is

`ivpredict` *stub cmd depvar enreg* `: (`*inexog*`)` *exexog*

where

stub is the stub of the names of any predictions produced. Specifically, predictions of the linear index will be saved in *stub*`xb`, and predictions of probabilities will be saved in *stub*`pr`;

cmd is the estimation command used to fit (13) or its equivalent;

depvar is the left-hand-side endogenous regressor (y in the notation of section 2);

enreg is the list of right-hand-side endogenous regressors (Y in the notation of section 2);

`:` is a required separator of the endogenous variables from the list of instruments; and

`(`*inexog*`)` is the list of instruments (or predetermined variables) included in the structural equation of interest (excluding the intercept), and *exexog* are those instruments excluded from this equation. In the notation of section 2, `(`*inexog*`)` and *exexog* are denoted X and Z, respectively. Note that the parentheses around *inexog* are required, and in the event that *inexog* is empty, it is still necessary to include an empty pair of parentheses: `( )`. It is assumed that an intercept appears somewhere in the system of equations, although it need not appear in the structural equation of interest.

All arguments of `ivpredict`, as specified above, are required.

We believe that because *cmd* is user specified, `ivpredict` will work for other linear index models. For example, if one was fitting a structural model using `ivpoisson`, then setting *cmd* to `poisson` should work. However, this has not been tested extensively.

6 An application using ivpredict

In this section, we provide a concrete example of the use of `ivpredict`, and we illustrate one of the vagaries of its use.

Our application is drawn from Cameron and Trivedi (2009, sec. 14.8).[19] It uses data from the Health and Retirement Study (HRS) to model the probability of the elderly purchasing private supplementary insurance (`ins`), with the logarithm of household income (`linc`) as the sole endogenous right-hand-side variable. Exogenous variables in the structural equation are measures of individual characteristics, including age, gender, race, ethnicity, marital status, years of education, a health-status dummy variable, the total number of chronic conditions, and the number of limitations on activities of daily living (respectively, `age`, `female`, `white`, `hisp`, `married`, `educyear`, `hstatusg`, `chronic`, and `adl`). The structural equation also includes the square of age (`age2`) as an explanatory variable. The names of all of these exogenous variables compose the elements of the global macro `$xlist2` in the output discussed below. The exogenous variables excluded from the structural equation that serve as additional instruments are the individual's retirement status and his or her spouse's retirement status (`retire` and `sretire`, respectively). In the output discussed below, the global macro `$ivlist2` is the concatenation of `$xlist2`, `retire`, and `sretire`. The original dataset contains observations on 3,206 individuals, although 9 data points are lost during estimation because, for these individuals, household income is allegedly 0, and so it is impossible to construct the endogenous regressor `linc`.

The commands used to generate the output are contained in the file `example.do`, which is presented in table 2.

Table 2. `example.do`

```
do ivpredict.do
use http://www.stata-press.com/data/mus/mus14data, clear
generate linc = ln(hhincome)
global xlist2 female age age2 educyear married hisp white chronic adl hstatusg
global ivlist2 $xlist2 retire sretire
ivprobit ins $xlist2 (linc = $ivlist2), vce(robust)
predict ivins, xb
predict ivinspr, pr
ivpredict mypredict probit ins linc : ($xlist2) retire sretire
correlate ivins mypredictxb
correlate ivinspr mypredictpr
```

After some initial setup, the first task of commands in `example.do` is to fit the model using `ivprobit` and, on the basis of these results, obtain predictions of the form $\widehat{y}_i$ and $\Phi(\widehat{y}_i)$ using the `predict` command. These predictions are stored in the variables `ivins` and `ivinspr`, respectively. The `ivprobit` estimation results are presented in table 3.

19. Cameron and Trivedi (2009, 449) provide a more complete description of the data. The data are contained in the file `mus14data.dta`, which is available from http://www.stata-press.com/data/mus.html.

Table 3. Output from `example.do`: `ivprobit` estimation results

```
. ivprobit ins $xlist2 (linc = $ivlist2), vce(robust)
  (output omitted)
Probit model with endogenous regressors         Number of obs     =      3,197
                                                Wald chi2(11)     =     382.34
Log pseudolikelihood = -5407.7151               Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        linc |  -.5338185   .3852357    -1.39   0.166    -1.288867    .2212296
      female |  -.1394069   .0494475    -2.82   0.005    -.2363223   -.0424915
         age |   .2862283   .1280838     2.23   0.025     .0351886    .5372679
        age2 |  -.0021472   .0009318    -2.30   0.021    -.0039736   -.0003209
    educyear |   .1136877   .0237927     4.78   0.000     .0670548    .1603205
     married |   .7058269   .2377731     2.97   0.003     .2398002    1.171854
        hisp |  -.5094513   .1049488    -4.85   0.000    -.7151473   -.3037554
       white |    .156344   .1035713     1.51   0.131    -.0466521    .3593401
     chronic |   .0061943   .0275259     0.23   0.822    -.0477556    .0601441
         adl |  -.1347663     .03498    -3.85   0.000    -.2033259   -.0662067
    hstatusg |   .2341782   .0709769     3.30   0.001      .095066    .3732904
       _cons |  -10.00785   4.065795    -2.46   0.014    -17.97666   -2.039039
-------------+----------------------------------------------------------------
     /athrho |     .67453   .3599915     1.87   0.061    -.0310404      1.3801
    /lnsigma |   -.331594   .0233799   -14.18   0.000    -.3774178   -.2857703
-------------+----------------------------------------------------------------
         rho |   .5879518    .235547                     -.0310305    .8809738
       sigma |   .7177787   .0167816                      .6856296    .7514352
------------------------------------------------------------------------------
Instrumented:  linc
Instruments:   female age age2 educyear married hisp white chronic adl
               hstatusg retire sretire
------------------------------------------------------------------------------
Wald test of exogeneity (/athrho = 0): chi2(1) = 3.51     Prob > chi2 = 0.0610
```

Note that in this example, the null hypothesis of exogeneity of `linc` is barely accepted if testing at the 5% level of significance but clearly rejected if testing at the 10% level.

As explained in section 5.2, `ivpredict` presents estimation results for each reduced-form equation that is fit. In this example, the only endogenous regressor is `linc`. This is modeled by least squares, and predictions for `linc` are stored in the new variable `plinc`. Here the naming convention for each new variable created at this stage is to prepend the name of the endogenous regressor with the letter '`p`'.

The next step is to replace each endogenous regressor (`linc`, in this example) by its linear predictor (`plinc`) and then obtain values for $\breve{y}_i$ and $\Phi(\breve{y}_i/\widehat{\sigma}_{y^\star})$, which are stored in `mypredictxb` and `mypredictpr`, respectively. Note that the `ivprobit` coefficient estimates are the ones used by `ivpredict` to construct $\breve{y}_i$ (or `mypredictxb`).

When constructing `mypredictpr`, `ivpredict` refits the model by (in this case, probit) maximum likelihood but now with `plinc` replacing `linc` in the list of regressors. This procedure automatically adjusts the coefficient estimates for scaling by $\widehat{\sigma}_{y^\star}$. These probit results are the second set appearing in table 4.

Table 4. Output from `example.do`: `ivpredict` estimation results

```
. ivpredict mypredict probit ins linc : ($xlist2) retire sretire

      Source |       SS           df       MS      Number of obs   =     3,197
-------------+----------------------------------   F(12, 3184)     =    188.99
       Model |  1173.12053        12  97.7600445   Prob > F        =    0.0000
    Residual |  1647.03826     3,184  .517285885   R-squared       =    0.4160
-------------+----------------------------------   Adj R-squared   =    0.4138
       Total |  2820.15879     3,196  .882402626   Root MSE        =    .71923

------------------------------------------------------------------------------
        linc |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -.0936494   .0297304    -3.15   0.002     -.151942   -.0353569
         age |   .2669284   .0627794     4.25   0.000     .1438361    .3900206
        age2 |  -.0019065   .0004648    -4.10   0.000    -.0028178   -.0009952
    educyear |    .094801   .0043535    21.78   0.000     .0862651    .1033369
     married |   .7918411   .0367275    21.56   0.000     .7198291    .8638531
        hisp |  -.2372014   .0523874    -4.53   0.000    -.3399179    -.134485
       white |   .2324672   .0347744     6.69   0.000     .1642847    .3006496
     chronic |  -.0388345   .0100852    -3.85   0.000    -.0586086   -.0190604
         adl |  -.0739895   .0173458    -4.27   0.000    -.1079995   -.0399795
    hstatusg |   .1748137   .0338519     5.16   0.000       .10844    .2411875
      retire |  -.0909581   .0288119    -3.16   0.002    -.1474499   -.0344663
     sretire |  -.0443106   .0317252    -1.40   0.163    -.1065145    .0178932
       _cons |  -7.702456   2.118657    -3.64   0.000    -11.85653   -3.548385
------------------------------------------------------------------------------
(option xb assumed; fitted values)
Iteration 0:   log likelihood = -2139.7712
Iteration 1:   log likelihood =  -1989.887
Iteration 2:   log likelihood = -1988.8213
Iteration 3:   log likelihood = -1988.8206
Iteration 4:   log likelihood = -1988.8206
Probit regression                               Number of obs     =      3,206
                                                LR chi2(11)       =     301.90
                                                Prob > chi2       =     0.0000
Log likelihood = -1988.8206                     Pseudo R2         =     0.0705

------------------------------------------------------------------------------
         ins |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       plinc |  -.5086222   .5010422    -1.02   0.310    -1.490647    .4734024
      female |  -.1540794   .0678164    -2.27   0.023     -.286997   -.0211618
         age |   .3250768   .1683096     1.93   0.053    -.0048039    .6549576
        age2 |  -.0024523   .0012331    -1.99   0.047    -.0048692   -.0000354
    educyear |   .1255632   .0475677     2.64   0.008     .0323322    .2187942
     married |   .7501984   .3865768     1.94   0.052    -.0074781    1.507875
        hisp |  -.5781927   .1604455    -3.60   0.000    -.8926601   -.2637253
       white |   .1547357   .1336158     1.16   0.247    -.1071465    .4166179
     chronic |    .013837   .0270804     0.51   0.609    -.0392396    .0669136
         adl |  -.1552079   .0501382    -3.10   0.002    -.2534769   -.0569388
    hstatusg |   .2522909   .1073601     2.35   0.019     .0418689    .4627129
       _cons |  -11.55901   5.208077    -2.22   0.026    -21.76666   -1.351369
------------------------------------------------------------------------------
```

Note the aforementioned vagary of `ivpredict`. Observe that the sample size used to fit the probit model is the full 3,206 observations rather than the 3,197 used to fit the reduced-form model. This is because `linc` has been replaced by predictions of it (namely, `plinc`). Therefore, losing nine observations because `linc` is undefined in those cases is immaterial to `plinc`. That is, in the notation of section 2, provided that we have observations on X_i' and Z_i', we can produce forecast $\widehat{Y}_i'$ regardless of whether or not Y_i' is well defined. This slight inconsistency is something that the user may wish to ponder. If deemed a problem, then it can be addressed by simply dropping the offending observations from the overall analysis before invoking `ivpredict`.

To further explore the practical implications of the different prediction methods, we look at correlations between the pairs (`ivins` and `mypredictxb`) and (`ivinspr` and `mypredictpr`). These are 0.6885 and 0.6704, respectively, which are not especially large. This provides some prima facie evidence that getting the prediction model correct can substantially impact the predictions obtained.

In closing, we emphasize that `ivpredict` is potentially useful for many linear index models with one or more endogenous right-hand-side variables and not just for endogenous probit estimation with one endogenous regressor. For example, `ivpredict` can be used for linear, logit, tobit, or Poisson regression. When fitting a linear model using two-stage least squares, one simply uses `regress` for *cmd*.

7 Concluding remarks

In linear index models with endogenous regressors, simply replacing parameters with consistent estimators thereof does not necessarily yield consistent estimators of the expectations of interest. In section 3, we illustrate this phenomenon. Of course, the developments we discuss in sections 2 and 3 are not new. For instance, equation (6) [or (7)] provides a foundation for one form of the Hausman (1978) test for exogeneity, which involves augmenting the structural equation of interest with reduced-form residuals and then testing that the coefficients of these artificial regressors are jointly zero. Similarly, Terza, Basu, and Rathouz (2008), and the references cited therein, explore issues of estimation in probit models with endogenous regressors, among other things. Another example of this phenomenon, that of the linear simultaneous-equations model, is explored by Skeels and Taylor (2014).

In section 4, we explore how Stata, and some user-written extensions to Stata, deal with prediction in probit models with endogenous regressors. The overwhelming impression is that none of them address the problem particularly well. In light of this assessment, we recommend that `predict` offer an additional option that provides $\breve{y}_i$ rather than $\widehat{y}_i$. Indeed, if this new option were available, then it makes sense for it to impact the behavior of the `pr` option as well so that probability predictions are based on (13) rather than $\Phi(\widehat{y}_i)$. In section 5, we discuss a program that provides a slightly less-elegant solution than that offered by the more-intelligent `predict` command in Stata. As noted throughout the article, although the various types of predictions we discuss here all have important uses, those that correspond to conditional expectations are typically the easiest to interpret.

8 Acknowledgments

We thank Steve Pudney for helpful discussions early in this article's genesis, while concurrently absolving him from any shortcomings contained herein. We also thank the editor, Joe Newton, and an anonymous referee for offering positive and constructive comments.

9 References

Amemiya, T. 1978. The estimation of a simultaneous equation generalized probit model. *Econometrica* 46: 1193–1205.

Baum, C. F., M. E. Schaffer, and S. Stillman. 2002. ivreg2: Stata module for extended instrumental variables/2SLS and GMM estimation. Statistical Software Components S425401, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s425401.html.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications*. New York: Cambridge University Press.

———. 2009. *Microeconometrics Using Stata*. College Station, TX: Stata Press.

Cox, D. R., and E. J. Snell. 1968. A general definition of residuals. *Journal of the Royal Statistical Society, Series B* 30: 248–275.

Gelbach, J. B. 1999. probitiv: Stata module to perform instrumental variables probit. Statistical Software Components S371401, Department of Economics, Boston College. https://ideas.repec.org/c/boc/bocode/s371401.html.

Greene, W. H. 2012. *Econometric Analysis*. 7th ed. Upper Saddle River, NJ: Prentice Hall.

Harkness, J. 2001. ivprob-ivtobit.pkg. Technical report, online publication. http://fmwww.bc.edu/repec/bocode/i.

Hausman, J. A. 1978. Specification tests in econometrics. *Econometrica* 46: 1251–1271.

Keshk, O. M. G. 2003. CDSIMEQ: A program to implement two-stage probit least squares. *Stata Journal* 3: 157–167.

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

Mardia, K. V., J. T. Kent, and J. M. Bibby. 1979. *Multivariate Analysis*. London: Academic Press.

Newey, W. K. 1987. Efficient estimation of limited dependent variable models with endogenous explanatory variables. *Journal of Econometrics* 36: 231–250.

Pagan, A., and F. Vella. 1989. Diagnostic tests for models based on individual data: A survey. *Journal of Applied Econometrics* 4: S29–S59.

Poskitt, D. S., and C. L. Skeels. 2008. Conceptual frameworks and experimental design in simultaneous equations. *Economics Letters* 100: 138–142.

Roodman, D. 2011. Fitting fully observed recursive mixed-process models with cmp. *Stata Journal* 11: 159–206.

Searle, S. R. 1982. *Matrix Algebra Useful for Statistics*. New York: Wiley.

Skeels, C. L., and L. W. Taylor. 2014. Prediction after IV estimation. *Economics Letters* 122: 420–422.

Skeels, C. L., and F. Vella. 1999. A Monte Carlo investigation of the sampling behavior of conditional moment tests in tobit and probit models. *Journal of Econometrics* 92: 275–294.

StataCorp. 2015. *Stata 14 Base Reference Manual*. College Station, TX: Stata Press.

Terza, J. V., A. Basu, and P. J. Rathouz. 2008. Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal of Health Economics* 27: 531–543.

About the authors

Chris Skeels is an associate professor in the Department of Economics at the University of Melbourne.

Larry Taylor is a professor of economics in the Department of Economics at Lehigh University.