Estimating Logistic Regressions with Two-Stage Least Squares

Zach Flynn*

May 25, 2018

Abstract

I develop an algorithm to estimate a flexible binary regression model with endogeneity by repeatedly solving a two-stage least squares problem; the algorithm is numerically stable and guaranteed to converge regardless of starting value. The method is numerically stable even when a successful outcome is rare because it has a uniformly small condition number, unlike Newton methods with maximum likelihood estimation whose condition number is unbounded across potential parameter values. The instrumental variable method does not require choosing a special regressor or making assumptions on the first stage relationship between covariates and instruments other than a rank restriction to ensure the instruments are relevant enough.

1 Introduction

Linear probability models are a bad idea[1] — especially when the probability of success is low or high for some subpopulations — but they are easy to compute and we can handle endogeneity with instrumental variables using standard two-stage least squares. I show how to use two-stage least squares to estimate more realistic probability models with endogeneity and how the method can improve the numerical stability of logistic regression, especially when modeling "rare" events.

I use a flexible parametric model to estimate the causal effect of some vector x on the probability that $y = 1$ using instruments z,

$$\text{pr}\left\{y = 1|z\right\} = \mathbb{E}\left[\frac{1}{1 + \exp\left(-r\left(x\right)^{\top}\beta\right)}\Big|z\right], \tag{1}$$

*E-mail: zflynn@wisc.edu or zlflynn@gmail.com.

[1] See Lewbel, Dong, and Yang (2012), among other papers and textbooks, for many reasons why. The issues with the model stem from its failure to require predicted probabilities be between $[0, 1]$. In fact, if the covariates have unbounded support, the linear probability model cannot be the data generating process.

where r is a vector of known transformations of a vector x. The endogeneity problem is that x not only causally changes the probability of success ($y = 1$) but is also correlated with unobserved determinants of success. We use the instruments z to deal with the endogeneity problem. Economically, the instrumental variable assumption is that the only reason the probability of success changes with z is because of z's relationship to x; z does not otherwise determine success[2].

The advantage of the specification is that the predicted response function can be flexibly specified and will always lie between $[0, 1]$ unlike in a linear probability model, but there are two challenges relative to the linear probability model, both of which I address in the paper:

1. There is not a ready closed form estimate of β, which can make computing the estimator difficult, especially when $r(x)$ is high-dimensional.

2. It is a nonlinear instrumental variable model and these require stronger assumptions on the first stage (the relationship between instruments and covariates) than linear instrumental variable models.

The advantage of this particular instrumental variable assumption over other methods of estimating binary regression models with endogeneity (see Lewbel 2000, Dong and Lewbel 2015, Rothe 2009, Blundell and Smith 1989, Rivers and Vuong 1988, and Heckman 1976) is that the model needs few assumptions beyond the instrumental variable assumption itself to achieve identification. In particular, I do not need to make parametric assumptions about the relationship between x and z (as in maximum likelihood estimation) or to assume a "special regressor", see Lewbel (2000), which imposes support restrictions and restricts how flexible the approximation can be because transformations of the special regressor cannot be included in the specification: if x_1 is the special regressor, x_1^2 cannot also be in $r(x)$. Identification relies only on a relevance condition which I will call a "tilting rank condition" which is analogous to the standard rank condition in linear regression.

The most similar method computationally is the *iterative least squares* approach used to estimate semiparametric binary choice models (without endogeneity) in Wang and Zhou (1995) and Dominitz and Sherman (2005) in the sense that both my method and iterative least squares use linear regression to compute step size in an iterative algorithm. The methods are based on different principles and match different "moments" of the data. My approach extends more naturally to models with endogeneity while iterative least squares extends more naturally to semiparametric models.

I describe my proposed algorithm in Section 2 and, in Section 3, I show it is also more numerically stable than using the Newton method in maximum likelihood estimation, a standard method for estimating logistic regressions (the method used in Stata), because it has a uniformly small condition number across potential parameter vectors while the Newton method applied to maximum likelihood estimation has an unbounded condition number. The

[2]There are other models of endogeneity and kinds of instrumental variable assumptions in binary regression models, see Lewbel (2000).

method I propose is particularly more stable when "success" (the event $y = 1$) is rare (or very common).

Section 4 is a short block of Stata code to implement the estimator.

2 Algorithm

Let $w(z)$ be a vector of transformations of z such that $w(z)$ is equal in dimension to $r(x)$.

I develop an iterative approach to computing β that is guaranteed to converge regardless of starting value.

Algorithm: Repeated Two-Stage Least Squares.

1. Use two-stage least squares to estimate γ *as if* the model were the linear probability model,

$$y = r(x)^\top \gamma + e, \quad \mathbb{E}[w(z)\, e] = 0. \tag{2}$$

2. Set $k = 0$ and let β_0 be some initial value for β.

3. Use two-stage least squares to estimate $\gamma_k = \gamma(\beta_k)$ as if the true model were,

$$\frac{1}{1 + \exp\left(-r(x)^\top \beta_k\right)} = r(x)^\top \gamma_k + e_k, \quad \mathbb{E}[w(z)\, e_k] = 0. \tag{3}$$

4. Set $\beta_{k+1} = \beta_k + \gamma - \gamma_k$.

5. Set $k = k + 1$.

6. Repeat steps 3-5 until $||\gamma - \gamma_k||$ is sufficiently small. Then $\beta = \beta_k$.

The main assumption that both identifies the model and justifies the above algorithm is the tilting rank condition.

Assumption 1 (Tilting Rank Condition)**.** Let $u(x)$ be any scalar function of x such that $u(x) > 0$ for all x. Then, $\mathbb{E}\left[u(x)\, w(z)\, r(x)^\top\right]$ is invertible. This condition is always true without endogeneity, where $w(z) = r(x)$, because $\mathbb{E}\left[u(x)\, r(x)\, r(x)^\top\right]$ is strictly positive definite (assuming no colinearity in r), and so, invertible.

Theorem 1. Define $T(\beta) = \beta + \gamma - \gamma(\beta)$. Given the *Tilting Rank Condition*, T is a contraction mapping so it has a unique fixed point. Any fixed point of T is a value of β such that equation (1) is satisfied. So β is identified, and it is the unique fixed point of T.

Proof. Define $\pi(x,\beta) = \left[1+\exp\left(-r(x)^\top \beta\right)\right]^{-1}$. Write DT (the derivative of T) as (suppressing dependence of w on z, of r on x, and of π on x and β),

$$DT = I - \mathbb{E}\left[wr^\top\right]^{-1}\mathbb{E}\left[wr^\top \pi(1-\pi)\right] \tag{4}$$
$$= \mathbb{E}\left[wr^\top\right]^{-1}\mathbb{E}\left[wr^\top \{1-\pi(1-\pi)\}\right] \tag{5}$$

If λ is an eigenvalue of DT, then, for a nonzero vector v,

$$\mathbb{E}\left[wr^\top\right]^{-1}\mathbb{E}\left[wr^\top \{1-\pi(1-\pi)\}\right] v = \lambda v \tag{6}$$
$$\implies \mathbb{E}\left[wr^\top\right]^{-1}\mathbb{E}\left[wr^\top \{1-\pi(1-\pi)-\lambda\}\right] v = 0 \tag{7}$$
$$\implies \mathbb{E}\left[wr^\top \{1-\pi(1-\pi)-\lambda\}\right] v = 0 \tag{8}$$

If $\lambda \geq 1$, then $\{1-\pi(1-\pi)-\lambda\}$ is strictly negative. So, by the *Tilting Rank Condition*, $\mathbb{E}\left[wr^\top \{1-\pi(1-\pi)-\lambda\}\right]$ is invertible, implying $v = 0$, a contradiction.

If $\lambda \leq 0$, then, similarly, $\{1-\pi(1-\pi)-\lambda\}$ is strictly positive. So, by the *Tilting Rank Condition*, $\mathbb{E}\left[wr^\top \{1-\pi(1-\pi)-\lambda\}\right]$ is invertible, implying $v = 0$, a contradiction.

So $\lambda \in (0,1)$ and the spectral norm of DT is less than 1; T is a contraction mapping with a unique fixed point.

Because, by equation (1),

$$\mathbb{E}[w(z)\,y] = \mathbb{E}\left[w(z)\frac{1}{1+\exp\left(-r(x)^\top \beta\right)}\right], \tag{9}$$

we know $\gamma = \gamma(\beta)$ for any true value of β. So any true value of β is a fixed point of T. Because there is only one fixed point of T, β is identified.

□

Inference on β is identical to making inference on the generalized method of moments estimator based on the moment conditions,

$$\mathbb{E}\left[w(z)\left\{y - \frac{1}{1+\exp\left(-r(x)^\top \beta\right)}\right\}\right] = 0; \tag{10}$$

because repeated two-stage least squares is an algorithm for solving the above system of moment equalities.

3 Numerical stability compared to the Newton method applied to maximum likelihood estimation

Even when there is no endogeneity problem to deal with, the repeated linear regression algorithm is useful because it is numerically more stable than Newton methods used in maximum likelihood estimation, a standard method for estimating logistic regressions (the default method in Stata). Recall that, without endogeneity, the *Tilting Rank Condition* is always satisfied.

Let $L(\beta)$ be the log likelihood of a logistic regression model. The Newton method finds β by iteration, following the rule,

$$\beta_{k+1} = \beta_k - D^2 L(\beta_k)^{-1} DL(\beta_k), \tag{11}$$

where D^2 gives the Hessian and D the gradient.

The repeated linear regression estimator uses the iteration rule,

$$\beta_{k+1} = \beta_k + \gamma - \gamma(\beta_k). \tag{12}$$

The difference between the two methods is how the step size, $\Delta_k = \beta_{k+1} - \beta_k$, is computed.

The Newton method solves the following linear system to determine step size,

$$D^2 L(\beta_k) \Delta_k = -DL(\beta_k). \tag{13}$$

Condition numbers are a measure of how accurate a numerical solution to a problem is; in this setting, condition numbers measure how much the solution to the linear equations would change if the right hand side vector were slightly perturbed. Problems with large condition numbers are numerically unstable. Let $\lambda_{min}(A)$ give the minimum eigenvalue of the matrix A and similarly for $\lambda_{max}(A)$. The condition number of the solution to this linear system is,

$$\kappa_{MLE}(\beta_k) = \frac{\lambda_{max}(D^2 L(\beta_k))}{\lambda_{min}(D^2 L(\beta_k))}, \tag{14}$$

The second derivative of the log likelihood is,

$$D^2 L(\beta) = -\mathbb{E}\left[r(x) r(x)^\top \pi(x,\beta) \times (1 - \pi(x,\beta))\right]. \tag{15}$$

So, for a given parameter vector β, the minimum eigenvalue of $D^2 L(\beta)$ can be arbitrarily close to zero because β can be such that $\pi(x,\beta)$ is arbitrarily close to zero or one. The condition number is large across β,

$$\sup_{\beta} \kappa_{MLE}(\beta) = \infty.$$

Repeated linear regression also determines step size by solving a linear system of equations,

$$\mathbb{E}\left[r(x)\, r(x)^{\top}\right]\Delta_k = \mathbb{E}\left[r(x)\left(y - \frac{1}{1+\exp\left(-r(x)^{\top}\beta_k\right)}\right)\right]. \tag{16}$$

The condition number of the solution to this linear system is,

$$\kappa_{RLR}(\beta_k) = \frac{\lambda_{max}\left(\mathbb{E}\left[r(x)\, r(x)^{\top}\right]\right)}{\lambda_{min}\left(\mathbb{E}\left[r(x)\, r(x)^{\top}\right]\right)}. \tag{17}$$

Two key points:

1. $\kappa_{RLR}(\beta)$ is not a function of β. The starting parameter values of the algorithm do not influence the condition number.
2. $\kappa_{RLR}(\beta)$ depends only on the eigenvalues of $\mathbb{E}\left[r(x)\, r(x)^{\top}\right]$. So long as the eigenvalues of the matrix are bounded away from zero (the x's are not colinear), the condition number will be bounded away from infinity, for all β.

So,

$$\sup_{\beta} \kappa_{RLR}(\beta) = \kappa_{RLR} < \infty.$$

This property of repeated linear regression makes computation more stable, especially when success is rare or common.

4 Stata Implementation

To aid the use of the method in practice, I include an implementation of the estimator in Stata (it assumes the dataset has variables $y, r1, r2, \ldots, w1, w2, \ldots$):

```
ivregress 2sls y (r* = w*)
mat gamma = e(b)
mat beta = J(1,colsof(gamma),1)

local fit = ""
local nr = colsof(gamma)-1
forvalues i=1/‘nr’ {
  local fit = "‘fit’ r‘i’*beta[1,‘i’] + "
}
local fit = "‘fit’ beta[1,‘nr’+1]"’
```

```
gen ybeta = .
mat diff = J(1,1,1)

while (diff[1,1] > 1e-8) {
  qui replace ybeta = 1/(1+exp(-1*(`fit')))
  qui ivregress 2sls ybeta (r* = w*)
  mat beta = beta + gamma - e(b)
  mat diff = (gamma-e(b))*((gamma-e(b))')
}
mat list beta
```

5 Conclusion

I present an algorithm for estimating binary regression models with or without endogeneity. The algorithm is numerically more stable than standard Newton maximum likelihood estimation — especially when success is rare — and it allows for instrumental variables while putting little restrictions on the first stage, on how instruments and covariates are related and without requiring any special regressor or other restriction on the flexibility of the model.

References

Blundell, R. and R. Smith (1989). Estimation in a class of simultaneous equation limited dependent variable models. *The Review of Economic Studies 56*(1), 37–57.

Dominitz, J. and R. Sherman (2005). Some convergence theory for iterative estimation procedures with an application to semiparametric estimation. *Econometric Theory 21*(4), 838–863.

Dong, Y. and A. Lewbel (2015). A simple estimator for binary choice models with endogenous regressors. *Econometric Reviews 34*, 82–105.

Heckman, J. (1976). Simultaneous equation models with both continuous and discrete endogenous variables with and without structural shift in the equations. *Studies in Nonlinear Estimation*.

Lewbel, A. (2000). Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables. *Journal of Econometrics 97*, 145–177.

Lewbel, A., Y. Dong, and T. T. Yang (2012). Comparing features of convenient estimators for binary choice models with endogenous regressors. *The Canadian Journal of Economics 45*(3), 809–829.

Rivers, D. and Q. Vuong (1988). Limited information estimators and exogeneity tests for simultaneous probit models. *Journal of Econometrics 39*(3), 347–366.

Rothe, C. (2009). Semiparametric estimation of binary response models with endogenous regressors. *Journal of econometrics 153*(1), 51–64.

Wang, W. and M. Zhou (1995). Iterative least squares estimator of binary choice models: a semi-parametric approach. *University of Kentucky Working Papers*.