Fractional Response Models with Endogenous Explanatory Variables and Heterogeneity

Jeffrey M. Wooldridge
Michigan State University

1. Introduction
2. Fractional Probit with "Heteroskedasticity"
3. Fractional Probit with an Endogenous Explanatory Variable
4. Linear Unobserved Effects Models with Unbalanced Panels
5. Nonlinear UE Models with Unbalanced Panels

1. Introduction

- A fractional response y satisfies $0 \leq y \leq 1$, possibly with $P(y = 0) > 1$ or $P(y = 1) > 0$ (or both).
- Assume y is the variable we would like to explain in terms of covariates, $\mathbf{x} = (x_1, \ldots, x_K)$. (No data censoring, but y may be a "corner solution.)
- Focus here is on mean response. If $\mathbf{x}$ is exogenous, goal is to estimate $E(y|\mathbf{x})$.

- Can always use a linear model for $E(y|\mathbf{x})$, but it is at best an approximation.
- Papke and Wooldridge (1996, *Journal of Applied Econometrics*): Model $E(y|\mathbf{x})$ using models of the form $G(\mathbf{x}\boldsymbol{\beta})$ for $0 < G(\cdot) < 1$ (or nonindex forms).

- So-called “fractional response” models (fractional probit, fractional logit) easily estimated using `glm`, and robust inference is trivial (and very important: MLE standard errors are too *large*).
- For panel data, can use `xtgee`. Papke and Wooldridge (2008, *Journal of Econometrics*) show how to use correlated random effects approaches to estimate fractional response models for panel data. But for balanced panels.
- Wooldridge (2005, Rothenberg Festschrift; 2010, MIT Press) considers models with continuous endogenous explanatory variables (EEVs). Proposes two-step control function approach.

- Papke and Wooldridge (2008): heterogeneity and continuous EEV. Combination of CRE and control function methods for fractional probit. But balanced panel, and only two-step estimators.
- What if we want a one-step quasi-MLE (which simplifies inference and may have better finite-sample properties)? So y_1 is a fractional response and y_2 a continuous EEV. Wooldridge (2011, unpublished) shows that the `ivprobit` log-likelihood identifies the (scaled) parameters under correct specification of $E(y_1|y_2, \mathbf{z}_1, a_1)$ where a_1 is the omitted variable. (and y_2 follows classical linear model).

- What if y_1 is a fractional response and y_2 a binary EEV? Two-step "forbidden regression" is not valid. Wooldridge (2011) shows the `biprobit` log likelihood identifies the (scaled) parameters if $E(y_1|y_2, \mathbf{z}_1, a_1)$ is correctly specified (and y_2 follows a probit).
- Neither `ivprobit` nor `biprobit` allow y_1 to be a fractional response. Neither does `cmp` (Roodman, 2009).
- Bottom line: Many existing Stata commands could be used to estimate flexible fractional response models allowing for endogeneity and unbalanced panel by removing the "data checks" on the response variable.

2. Fractional Probit with "Heteroskedasticity"

- Let $\mathbf{x} = (x_1, x_2, \ldots, x_K)$. Fractional probit model is

$$E(y|\mathbf{x}) = \Phi(\beta_0 + \mathbf{x}\boldsymbol{\beta}) = \Phi(\beta_0 + \beta_1 x_1 + \ldots + \beta_K x_K)$$

- Might want more flexibility. If $P(y = 0) > 0$, could use a two-part model.
- But can directly make model for $E(y|\mathbf{x})$ more flexible, for example,

$$E(y|\mathbf{x}) = \Phi[(\beta_0 + \mathbf{x}\boldsymbol{\beta})\exp(-\mathbf{z}\boldsymbol{\delta}/2)]$$

where $\mathbf{z}$ $(1 \times M)$ is a function of $(x_1, x_2, \ldots, x_K)$ that does not include a constant.

- The β_j and δ_h are consistently estimated using the Bernoulli quasi-MLE if $E(y|\mathbf{x})$ is correctly specified. As usual, need to use robust inference because y is not binary. (The conditional mean may be misspecified, anyway.)
- Ideally, just type

```
hetprobit y x1 ... x2, het(z1 z2 ... zM), robust
```

- But y is turned into a binary response.
- Can easily test $H_0 : \boldsymbol{\delta} = \mathbf{0}$ with robust Wald statistic.

```
clear

capture program drop frac_het

program frac_het
version 11
args llf xb zg
quietly replace ‘llf’ = $ML_y1*log(normal(‘xb’*exp(-‘zg’))) ///
     + (1 - $ML_y1)*log(1 - normal(‘xb’*exp(-‘zg’)))
end

ml model lf frac_het (prate = mrate ltotemp age sole) ///
     (mrate ltotemp age sole, nocons), vce(robust)

ml max
```

Number of obs = 4075
Wald chi2(4) = 152.
Log pseudolikelihood = -1674.5212 Prob > chi2 = 0.0000

prate	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval
eq1						
mrate	1.384717	.2238623	6.19	0.000	.9459552	1.823479
ltotemp	-.1495098	.0139662	-10.71	0.000	-.1768831	-.1221365
age	.0670733	.0100639	6.66	0.000	.0473484	.0867981
sole	-.1182667	.0932352	-1.27	0.205	-.3010042	.0644708
_cons	1.679383	.1059	15.86	0.000	1.471823	1.886944
eq2						
mrate	.2403586	.053781	4.47	0.000	.1349497	.3457674
ltotemp	.0375202	.0144216	2.60	0.009	.0092543	.0657861
age	.0171714	.0027289	6.29	0.000	.011823	.0225199
sole	-.1627509	.0631067	-2.58	0.010	-.2864378	-.039064

```
. test [eq2]

 ( 1)  [eq2]mrate = 0
 ( 2)  [eq2]ltotemp = 0
 ( 3)  [eq2]age = 0
 ( 4)  [eq2]sole = 0

           chi2(  4) =  109.26
         Prob > chi2 =    0.0000

. * Usual fractional probit (could use glm):

capture program drop frac_probit

program frac_probit
version 11
args llf xb
quietly replace ‘llf’ = $ML_y1*log(normal(‘xb’)) ///
    + (1 - $ML_y1)*log(1 - normal(‘xb’))
end

ml model lf frac_probit (prate = mrate ltotemp age sole), vce(robust)

ml max
```

```
                                                  Number of obs   =       4075
                                                  Wald chi2(4)    =     695.
Log pseudolikelihood = -1681.9607                 Prob > chi2     =     0.0000

------------------------------------------------------------------------------
             |               Robust
       prate |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval
-------------+----------------------------------------------------------------
       mrate |   .5955726    .038756    15.37   0.000     .5196123    .6715329
    ltotemp |  -.1172851   .0080003   -14.66   0.000    -.1329655   -.1016048
         age |   .0180259   .0014218    12.68   0.000     .0152392    .0208126
        sole |   .0944158   .0271696     3.48   0.001     .0411645    .1476672
       _cons |   1.428854   .0593694    24.07   0.000     1.312493    1.545216
------------------------------------------------------------------------------
```

- Should do a comparison of average partial effects between ordinary fractional probit and heteroskedastic fractional probit.
- The “hetprobit” quasi-MLE is needed for nonlinear CRE panel models with unbalanced panels.

3. Fractional Probit with an Endogenous Explanatory Variable

- Adapted from Wooldridge (2011, unpublished). Set up endogeneity as an omitted variable problem, and start by assuming y_2 is continuous:

$$E(y_1|\mathbf{z}, y_2, a_1) = \Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + a_1).$$

$$y_2 = \mathbf{z}\boldsymbol{\delta}_2 + v_2,$$

where $\mathbf{x}_1$ is a general nonlinear function of $(\mathbf{z}_1, y_2)$, a_1 is an omitted factor thought to be correlated with y_2 but independent of the exogenous variables $\mathbf{z}$.

• The average partial effects in this model are obtained from the "average structural function" (ASF):

$$ASF(\mathbf{x}_1) = E_{a_1}[\Phi(\mathbf{x}_1\boldsymbol{\beta}_1 + a_1)] = \Phi(\mathbf{x}_1\boldsymbol{\beta}_{a1})$$

where

$$\boldsymbol{\beta}_{a1} = \boldsymbol{\beta}_1/(1 + \sigma_{a_1}^2)^{1/2}.$$

• Happily, these are precisely the parameters that are identified.

• If (a_1, v_2) is jointly normal, a two-step control function method is valid (Wooldridge, 2005). Note that the distribution of y_1 is not further restricted.

(i) Regress y_{i2} on $\mathbf{z}_i$ and obtain the residuals, $\hat{v}_{i2}$.

(ii) Use "probit" of y_{i1} on $\mathbf{x}_{i1}, \hat{v}_{i2}$ to estimate parameters with different scales, say $\hat{\boldsymbol{\beta}}_{e1}$ and $\hat{\gamma}_{e1}$. (Can implement as a "generalized linear model.")

• The "average structural function" (ASF) is consistently estimated as

$$\widehat{ASF}(y_2, \mathbf{z}_1) = N^{-1} \sum_{i=1}^{N} \Phi(\mathbf{x}_1 \hat{\boldsymbol{\beta}}_{e1} + \hat{\gamma}_{e1} \hat{v}_{i2}),$$

and this can be used to obtain APEs with respect to y_2 or $\mathbf{z}_1$ (Wooldridge, 2005).

• What about a quasi-LIML approach? Can show that

$$E(y_1|y_2,\mathbf{z}) = \Phi\left[\frac{\mathbf{x}_1\boldsymbol{\beta}_{r1} + (\rho_1/\tau_2)(y_2 - \mathbf{z}\boldsymbol{\delta}_2)}{(1-\rho_1^2)^{1/2}}\right]$$

and so we can plug this mean function into the Bernoulli quasi-log likelihood. This gives $q_1(y_1, y_2, \mathbf{z}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$. Identify $\boldsymbol{\delta}_2$ and τ_2 using the Gaussian QLL, which gives $q_2(y_2, \mathbf{z}, \boldsymbol{\theta}_2)$.

• The same objective function we get for MLE with y_1 binary can be used when y_1 is fractional – continuous or otherwise.

• In other words, `ivprobit` could be easily modified – and use robust inference.

• A similar argument holds when y_2 is binary and follows a probit model:

$$
\begin{aligned}
y_2 &= 1[\mathbf{z}\boldsymbol{\delta}_2 + v_2 \geq 0] \\
v_2|\mathbf{z} &\sim Normal(0,1)
\end{aligned}
$$

• Can show that $E(y_1|y_2,\mathbf{z})$ has the same form as the response probability in the so-called "bivariate probit" model.

- For example,

$$E(y_1|y_2 = 1, \mathbf{z}) = \int_{-\mathbf{z}\boldsymbol{\delta}_2}^{\infty} \Phi\left[\frac{\mathbf{x}_1\boldsymbol{\beta}_{r1} + \rho_1 \mathfrak{v}_2}{(1-\rho_1^2)^{1/2}} \right] d\mathfrak{v}_2$$

- So for $q_2(y_2, \mathbf{z}, \boldsymbol{\theta}_2)$ we use the usual probit log-likelihood. For $q_1(y_1, y_2, \mathbf{z}, \boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ we use the Bernoulli QLL associated with bivariate probit.

- So if y_1 were allowed to be fractional, `biprobit` with a “robust” option could be used.

4. Linear Unobserved Effects Models with Unbalanced Panels

- Model for a random draw i has T *potential* time periods:

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + c_i + u_{it}, t = 1, \ldots, T$$

$$E(u_{it}|\mathbf{x}_{i1}, \ldots, \mathbf{x}_{iT}, c_i) = 0.$$

- Given access to a balanced random sample, the zero conditional mean assumption is sufficient for FE to be consistent (as $N \to \infty$, T fixed) and $\sqrt{N}$-asymptotically normal, provided all elements of $\mathbf{x}_{it}$ have some time variation.

- Let $\{s_{it} : t = 1, \ldots, T\}$ be a sequence of "selection indicators": $s_{it} = 1$ if and only if observation (i, t) is used. These are generally outcomes of random variables.
- The number of time periods available for unit i is $T_i = \sum_{r=1}^{T} s_{ir}$; this is properly viewed as random.

Fixed Effects on the Unbalanced Panel

• The time-demeaned data uses a different number of time periods for different i. Let

$$\ddot{y}_{it} = y_{it} - T_i^{-1} \sum_{r=1}^{T} s_{ir} y_{ir}$$

$$\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - T_i^{-1} \sum_{r=1}^{T} s_{ir} \mathbf{x}_{ir}$$

- The FE estimator is then

$$\hat{\boldsymbol{\beta}}_{FE} = \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{it}' \ddot{\mathbf{x}}_{it} \right)^{-1} \left(N^{-1} \sum_{i=1}^{N} \sum_{t=1}^{T} s_{it} \ddot{\mathbf{x}}_{it}' \ddot{y}_{it} \right),$$

- A sufficient condition for consistency of FE on the unbalanced panel is an extension of the usual strict exogeneity assumption:

$$E(u_{it} | \mathbf{x}_i, \mathbf{s}_i, c_i) = 0, \; t = 1, \ldots, T$$

$$\mathbf{s}_i = (s_{i1}, \ldots, s_{iT})$$

- Both the covariates and selection are strictly exogenous conditional on c_i. Rules out selection in any time period depending on the shocks in any time period. That is, the condition is generally violated if $Cov(s_{ir}, u_{it}) \neq 0$ for any (r, t) pair.
- Importantly, it allows s_{it} to depend on c_i in an unrestricted way.
- `xtreg` allows unbalanced panels and properly computes standard errors and test statistics.

Random Effects on the Unbalanced Panel

- The quasi-time-demeaning value for unit i is

$$\hat{\theta}_i = 1 - \left\{ \frac{1}{[1 + T_i(\hat{\sigma}_c^2/\hat{\sigma}_u^2)]} \right\}^{1/2}.$$

Now define

$$\breve{y}_{it} = y_{it} - \hat{\theta}_i \bar{y}_i$$

where $\bar{y}_i = T_i^{-1} \sum_{r=1}^{T} s_{ir} y_{ir}$, and similarly for $\breve{\mathbf{x}}_{it}$. Then, RE is POLS of $\breve{y}_{it}$ on $\breve{\mathbf{x}}_{it}$ using the $s_{it} = 1$ data points.

- Useful equivalence result (Wooldridge, 2010, unpublished). Define

$$\bar{\mathbf{x}}_i = T_i^{-1} \sum_{r=1}^{T} s_{ir} \mathbf{x}_{ir}$$

and consider either POLS or RE estimation of the following equation on the unbalanced panel:

$$y_{it} = \alpha + \mathbf{x}_{it} \boldsymbol{\beta} + \bar{\mathbf{x}}_i \boldsymbol{\xi} + v_{it}$$

Then $\hat{\boldsymbol{\beta}}_{POLS} = \hat{\boldsymbol{\beta}}_{RE} = \hat{\boldsymbol{\beta}}_{FE}$. Generally, $\hat{\boldsymbol{\xi}}_{POLS} \neq \hat{\boldsymbol{\xi}}_{RE}$

- Must be careful in constructing $\bar{\mathbf{x}}_i$; only use periods where all variables are observed ($s_{it} = 1$).
- Must now include the time averages of year dummies because these are no longer constants in an unbalanced panel.
- Same result holds when add any other time-constant covariates. Implies that the CRE is robust even with unbalanced panels.
- Basis for robust Hausman test. $H_0 : \boldsymbol{\xi} = 0$. Use RE with all time-constant controls included.

Heterogeneous Slopes

- Suppose the population model is

$$E(y_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i) = a_i + \mathbf{x}_{it}\mathbf{b}_i,$$

so, in the population, $\{\mathbf{x}_{it} : t = 1, \ldots, T\}$ is strictly exogenous conditional on $(a_i, \mathbf{b}_i)$.

- Define $a_i = \alpha + c_i$, $\mathbf{b}_i = \boldsymbol{\beta} + \mathbf{d}_i$ and write

$$y_{it} = \alpha + \mathbf{x}_{it}\boldsymbol{\beta} + c_i + \mathbf{x}_{it}\mathbf{d}_i + u_{it}$$

where $E(u_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i) = E(u_{it}|\mathbf{x}_i, c_i, \mathbf{d}_i) = 0$ for all t.

- Assume that selection may be related to $(\mathbf{x}_i, a_i, \mathbf{b}_i)$ but not the idiosyncratic shocks:

$$E(u_{it}|\mathbf{x}_i, a_i, \mathbf{b}_i, \mathbf{s}_i) = 0,\ t = 1, \ldots, T.$$

- Multiply population equation by the selection indicator:

$$s_i y_{it} = s_{it}\alpha + s_{it}\mathbf{x}_{it}\boldsymbol{\beta} + s_{it}c_i + s_{it}\mathbf{x}_{it}\mathbf{d}_i + s_{it}u_{it}$$

- Find an estimating equation by conditioning on $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1, \ldots, T\}$.

• Let $\mathbf{h}_i \equiv \{\mathbf{h}_{it} : t = 1,\ldots,T\} \equiv \{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1,\ldots,T\}$ and consider

$$E(s_i y_{it}|\mathbf{h}_i) = s_{it}\alpha + s_{it}\mathbf{x}_{it}\boldsymbol{\beta} + s_{it}E(c_i|\mathbf{h}_i) + s_{it}\mathbf{x}_{it}E(\mathbf{d}_i|\mathbf{h}_i)$$

and then make assumptions concerning $E(c_i|\mathbf{h}_i)$ and $E(\mathbf{d}_i|\mathbf{h}_i)$.

• We might choose

$$\mathbf{w}_i \equiv (T_i, \bar{\mathbf{x}}_i)$$

as the exchangeable functions satisfying

$$E(c_i|\mathbf{h}_i) = E(c_i|\mathbf{w}_i),\ E(\mathbf{d}_i|\mathbf{h}_i) = E(\mathbf{d}_i|\mathbf{w}_i).$$

- A flexible specification with $g_{ir} \equiv 1[T_i = r]$:

$$E(c_i|T_i,\bar{\mathbf{x}}_i) = \sum_{r=1}^{T} \psi_r(g_{ir} - \rho_r) + \sum_{r=1}^{T} g_{ir} \cdot (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_r)\boldsymbol{\xi}_r$$

$$E(\mathbf{d}_i|T_i,\bar{\mathbf{x}}_i) = \sum_{r=1}^{T}(g_{ir} - \rho_r)\boldsymbol{\kappa}_r + \sum_{r=1}^{T} g_{ir} \cdot (\bar{\mathbf{x}}_i - \boldsymbol{\mu}_r) \otimes \mathbf{I}_K]\boldsymbol{\eta}_r,$$

where the $\boldsymbol{\mu}_r$ are the expected values of $\bar{\mathbf{x}}_i$ given r time periods observed and ρ_r is the fraction of observations with r time periods:

$$\boldsymbol{\mu}_r = E(\bar{\mathbf{x}}_i|T_i = r),\ \rho_r = E\{1[T_i = r]\}$$

• This formulation is identical to running separate regressions for each T_i:

$$y_{it} \text{ on } 1,\ \mathbf{x}_{it},\ \bar{\mathbf{x}}_i,\ (\bar{\mathbf{x}}_i - \hat{\boldsymbol{\mu}}_r) \otimes \mathbf{x}_{it},\ \text{for } s_{it} = 1$$

where $\hat{\boldsymbol{\mu}}_r = N_r^{-1} \sum_{i=1}^{N} 1[T_i = r]\bar{\mathbf{x}}_i$ and N_r is the number of observations with $T_i = r$.

• The coefficient on $\mathbf{x}_{it}$, $\hat{\boldsymbol{\beta}}_r$, is the APE given $T_i = r$. Average these across r to obtain the overall APE. Cannot identify the APE for $T_i = 1$.

• A simple test of the null that the $\boldsymbol{\beta}_r$ do not change. Augmented equation is

$$y_{it} = \mathbf{x}_{it}\boldsymbol{\beta} + 1[T_i = 2] \cdot \mathbf{x}_{it}\boldsymbol{\gamma}_2 + \ldots + 1[T_i = T - 1] \cdot \mathbf{x}_{it}\boldsymbol{\gamma}_{T-1} + c_i + u_{it}$$

where the base group is $T_i = T$. Use FE on the unbalanced panel and obtain a fully robust test of

$$H_0 : \boldsymbol{\gamma}_2 = \mathbf{0}, \ldots, \boldsymbol{\gamma}_{T-1} = \mathbf{0}$$

This is like a Chow test where the slopes are allowed to differ by the number of available time periods for each unit.

```
. use meap94_98

. xtset schid year

. egen tobs = sum(1), by(schid)

. tab tobs

       tobs |      Freq.     Percent        Cum.
------------+-----------------------------------
          3 |      1,512       21.15       21.15
          4 |      1,028       14.38       35.52
          5 |      4,610       64.48      100.00
------------+-----------------------------------
      Total |      7,150      100.00

. gen tobs4 = tobs == 4

. gen tobs3 = tobs == 3

. gen tobs3_lavgrexp = tobs3*lavgrexp

. gen tobs4_lavgrexp = tobs4*lavgrexp
```

```
. xtreg math4 lavgrexp lunch lenrol y95 y96 y97 y98, fe cluster(distid)

.
.
.
                              (Std. Err. adjusted for 467 clusters in distid
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval
-------------+----------------------------------------------------------------
    lavgrexp |   6.288376   3.132334     2.01   0.045     .1331271    12.44363
       lunch |  -.0215072   .0399206    -0.54   0.590    -.0999539    .0569395
      lenrol |  -2.038461   2.098607    -0.97   0.332    -6.162365    2.085443
         y95 |    11.6192   .7210398    16.11   0.000     10.20231     13.0361
         y96 |   13.05561   .9326851    14.00   0.000     11.22282     14.8884
         y97 |   10.14771   .9576417    10.60   0.000      8.26588    12.02954
         y98 |   23.41404   1.027313    22.79   0.000      21.3953    25.43278
       _cons |   11.84422   32.68429     0.36   0.717    -52.38262    76.07107
-------------+----------------------------------------------------------------
     sigma_u |   15.84958
     sigma_e |  11.325028
         rho |  .66200804   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. xtreg math4 lavgrexp tobs3_lavgrexp tobs4_lavgrexp lunch lenrol y95 y96
        y97 y98, fe cluster(distid)

.
.
.
                               (Std. Err. adjusted for 467 clusters in distid
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval
-------------+----------------------------------------------------------------
    lavgrexp |   3.501465   3.547611     0.99   0.324    -3.469832    10.47276
tobs3_lavg~p |   8.048717   4.190867     1.92   0.055    -.1866205    16.28405
tobs4_lavg~p |   9.103049   6.809195     1.34   0.182    -4.277481    22.48358
       lunch |  -.0292364   .0380268    -0.77   0.442    -.1039616    .0454889
      lenrol |  -2.169307   2.074624    -1.05   0.296    -6.246084     1.90747
         y95 |   12.01813     .69288    17.35   0.000     10.65657    13.37968
         y96 |   13.56065   .9018155    15.04   0.000     11.78852    15.33278
         y97 |   10.60934   .9648135    11.00   0.000     8.713416    12.50526
         y98 |   23.84989   1.061322    22.47   0.000     21.76432    25.93546
       _cons |    10.6043   31.12293     0.34   0.733    -50.55438    71.76297
-------------+----------------------------------------------------------------
     sigma_u |  41.080099
     sigma_e |  11.319318
         rho |  .92943391   (fraction of variance due to u_i)
------------------------------------------------------------------------------
```

```
. test  tobs3_lavgrexp tobs4_lavgrexp

 ( 1)  tobs3_lavgrexp = 0
 ( 2)  tobs4_lavgrexp = 0

       F(  2,   466) =    2.37
            Prob > F =    0.0942

. * Might get away with using the pooled equations.
```

5. Nonlinear UE Models with Unbalanced Panels

- Adapted from Wooldridge (2010, unpublished). Interested in

$$E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i),$$

where $0 \leq y_{it} \leq 1$ and $\mathbf{c}_i$ is unobserved heterogeneity. (Binary response as special case.)

- Again, unbalanced panel. Assume strictly exogenous covariates conditional on $\mathbf{c}_i$ and ignorable selection:

$$E(y_{it}|\mathbf{x}_i, \mathbf{c}_i, \mathbf{s}_i) = E(y_{it}|\mathbf{x}_{it}, \mathbf{c}_i), t = 1, \ldots, T.$$

- Do not model serial correlation. Make inference robust.
- Specify models for

$$D(\mathbf{c}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1,\ldots,T\}).$$

- Let $\mathbf{w}_i$ be a vector of known functions of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1,\ldots,T\}$ that act as sufficient statistics, so that

$$D(\mathbf{c}_i|\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1,\ldots,T\}) \;\; = D(\mathbf{c}_i|\mathbf{w}_i)$$

- For simplicity, take

$$E(y_{it}|\mathbf{x}_i, c_i) = E(y_{it}|\mathbf{x}_{it}, c_i) = \Phi(\mathbf{x}_{it}\boldsymbol{\beta} + c_i),\ t = 1, \ldots, T$$

where $\mathbf{x}_{it}$ can include time dummies or other aggregate time variables.

- Assume that selection is conditionally ignorable for all t, that is,

$$E(y_{it}|\mathbf{x}_i, c_i, \mathbf{s}_i) = E(y_{it}|\mathbf{x}_i, c_i).$$

• All that is left is to specify a model for $D(c_i|\mathbf{w}_i)$ for suitably chosen functions $\mathbf{w}_i$ of $\{(s_{it}, s_{it}\mathbf{x}_{it}) : t = 1,\ldots,T\}$. Simplest is the time average on the selected periods, $\bar{\mathbf{x}}_i$, and the number of time periods, T_i.

• A specification linear in $\bar{\mathbf{x}}_i$ but with intercept and slopes different for each T_i is

$$E(c_i|\mathbf{w}_i) = \sum_{r=1}^{T} \psi_r 1[T_i = r] + \sum_{r=1}^{T} 1[T_i = r] \cdot \bar{\mathbf{x}}_i \boldsymbol{\xi}_r$$

• At a minimum, should let the variance of c_i change with T_i:

$$Var(c_i|\mathbf{w}_i) = \exp\left(\tau + \sum_{r=1}^{T-1} 1[T_i = r]\boldsymbol{\omega}_r\right)$$

• If we also maintain that $D(c_i|\mathbf{w}_i)$ is normal, then we obtain the following:

$$E(y_{it}|\mathbf{x}_{it}, \mathbf{w}_i, s_{it} = 1) = \Phi\left[\frac{\mathbf{x}_{it}\boldsymbol{\beta} + \sum_{r=1}^{T}\psi_r g_{ir} + \sum_{r=1}^{T} g_{ir} \cdot \bar{\mathbf{x}}_i\boldsymbol{\xi}_r}{\exp\left(\sum_{r=2}^{T} g_{ir}\omega_r\right)^{1/2}}\right]$$

where $g_{ir} = 1[T_i = r]$.

• No difficulty in adding $g_{ir} \cdot \bar{\mathbf{x}}_i$ for $r = 1,\ldots,T$ to the variance function.

• Can use "heteroskedastic probit" software provided the response variable can be fractional.

• The explanatory variables at time t are $(1, \mathbf{x}_{it}, g_{i1}, \ldots, g_{iT}, g_{i1} \cdot \bar{\mathbf{x}}_i, \ldots, g_{iT} \cdot \bar{\mathbf{x}}_i)$ and the explanatory variables in the variance are simply the dummy variables $(g_{i2}, \ldots, g_{iT})$, or also add $g_{i1} \cdot \bar{\mathbf{x}}_i, \ldots, g_{iT} \cdot \bar{\mathbf{x}}_i$.

• Might want to impose restrictions, such as constant slopes on $\bar{\mathbf{x}}_i$.

• The average partial effects are easy to obtain from the estimated "average structural function":

$$\widehat{ASF}(\mathbf{x}_t) = N^{-1}\sum_{i=1}^{N}\Phi\left[\frac{\mathbf{x}_t\hat{\boldsymbol{\beta}} + \sum_{r=1}^{T}\hat{\psi}_r g_{ir} + \sum_{r=1}^{T} g_{ir}\cdot\bar{\mathbf{x}}_i\hat{\boldsymbol{\xi}}_r}{\exp\left(\sum_{r=2}^{T} g_{ir}\hat{\omega}_r\right)^{1/2}}\right],$$

where the coefficients with "^" are from the pooled heteroskedastic fractional probit estimation.

• The functions of $(T_i, \bar{\mathbf{x}}_i)$ are averaged out, leaving the result a function of $\mathbf{x}_t$. Take derivatives or changes with respect to x_{tj}.

```
. use meap94_98

xtset schid year
       panel variable:  schid (unbalanced)
        time variable:  year, 1994 to 1998, but with gaps
                delta:  1 unit

. tab tobs

  number of |
       time |
    periods |      Freq.     Percent        Cum.
------------+-----------------------------------
          3 |      1,512       21.15       21.15
          4 |      1,028       14.38       35.52
          5 |      4,610       64.48      100.00
------------+-----------------------------------
      Total |      7,150      100.00

. gen tobs3 = tobs == 3

. gen tobs4 = tobs == 4

. replace math4 = math4/100
(7150 real changes made)
```

```
. capture program drop frac_het
.
. program frac_het
  1. version 11
  2. args llf xb zg
  3. quietly replace ‘llf’ = $ML_y1*log(normal(‘xb’*exp(-‘zg’))) ///
     + (1 - $ML_y1)*log(1 - normal(‘xb’*exp(-‘zg’)))
  4. end
.
.
end of do-file

. ml model lf frac_het (math4 = lavgrexp lunch lenrol y95 y96 y97 y98 lavgrexpb
  lunchb lenrolb y95b y96b y97b y98b tobs3 tobs4) (tobs3 tobs4, nocons),
  vce(cluster schid)

. ml max
```

Log pseudolikelihood = -4414.8409 Prob > chi2 = 0.0000

(Std. Err. adjusted for 1683 clusters in schid

math4	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval
eq1						
lavgrexp	.1142198	.0735598	1.55	0.120	-.0299547	.2583943
lunch	-.0013961	.001221	-1.14	0.253	-.0037891	.0009969
lenrol	-.067624	.0561521	-1.20	0.228	-.1776801	.0424321
y95	.3241894	.0150181	21.59	0.000	.2947545	.3536243
y96	.3724917	.0203004	18.35	0.000	.3327036	.4122797
y97	.2830853	.0217498	13.02	0.000	.2404565	.325714
y98	.7162732	.0239386	29.92	0.000	.6693544	.7631921
lavgrexpb	.1622914	.0957332	1.70	0.090	-.0253422	.349925
lunchb	-.0126246	.0012652	-9.98	0.000	-.0151044	-.0101448
lenrolb	-.0029272	.0610953	-0.05	0.962	-.1226718	.1168175
y95b	.8794288	.5371528	1.64	0.102	-.1733713	1.932229
y96b	.7270724	.2073897	3.51	0.000	.320596	1.133549
y97b	.6338092	.4187642	1.51	0.130	-.1869536	1.454572
y98b	.2733774	.4579278	0.60	0.551	-.6241446	1.170899
tobs3	.022217	.056255	0.39	0.693	-.0880406	.1324747
tobs4	.088465	.0891877	0.99	0.321	-.0863396	.2632697
_cons	-1.856404	.6052342	-3.07	0.002	-3.042641	-.6701668
eq2						
tobs3	.2007713	.0566528	3.54	0.000	.0897339	.3118087
tobs4	.5504922	.1162983	4.73	0.000	.3225517	.7784327

```
. ml model lf frac_probit (math4 = lavgrexp lunch lenrol y95 y96 y97 y98
  lavgrexpb lunchb lenrolb y95b y96b y97b y98b tobs3 tobs4), vce(cluster schid
. ml max

Log pseudolikelihood = -4420.8672                    Prob > chi2     =     0.0000

                              (Std. Err. adjusted for 1683 clusters in schid
------------------------------------------------------------------------------
             |               Robust
       math4 |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval
-------------+----------------------------------------------------------------
    lavgrexp |   .1227898   .0669842     1.83   0.067    -.0084967    .2540764
       lunch |  -.0008316   .0010475    -0.79   0.427    -.0028847    .0012215
      lenrol |  -.0556512   .0490405    -1.13   0.256    -.1517689    .0404665
         y95 |   .3186249   .0143788    22.16   0.000     .2904429    .3468069
         y96 |   .3647386   .0189796    19.22   0.000     .3275393    .4019379
         y97 |   .2860664   .0201033    14.23   0.000     .2466647    .3254682
         y98 |   .6760248   .0217182    31.13   0.000     .6334579    .7185917
   lavgrexpb |   .1658169     .08903     1.86   0.063    -.0086786    .3403125
      lunchb |  -.0113902   .0010958   -10.39   0.000    -.0135381   -.0092424
     lenrolb |   .0202697   .0531842     0.38   0.703    -.0839694    .1245088
        y95b |   .9325259   .3529265     2.64   0.008     .2408026    1.624249
        y96b |   .5439736   .1438847     3.78   0.000     .2619647    .8259826
        y97b |   .6807815   .2587424     2.63   0.009     .1736557    1.187907
        y98b |   .2624711    .338214     0.78   0.438    -.4004161    .9253584
       tobs3 |  -.0431248    .044767    -0.96   0.335    -.1308666     .044617
       tobs4 |  -.0771368   .0413601    -1.87   0.062     -.158201    .0039274
       _cons |  -2.194584   .5328879    -4.12   0.000    -3.239025   -1.150142
------------------------------------------------------------------------------
```