Estimating effects from extended regression models

David M. Drukker

Executive Director of Econometrics
Stata

2017 UK Stata Users Group meeting
8 September 2017

- Fictional data on wellness program from large company

```
. use wprogram2

. describe wchange age over phealth prog wtprog wtsamp

              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
wchange         float   %9.0g      changel    Weight change level
age             float   %9.0g                 Years over 50
over            float   %9.0g                 Overweight (tens of pounds)
phealth         float   %9.0g                 Prior health score
prog            float   %9.0g      yesno      Participate in wellness program
wtprog          float   %9.0g      yesno      Offered work time to participate
                                                in program
wtsamp          float   %9.0g                 Offered work time to participate
                                                in sample
```

- Three levels of `wchange`

```
. tabulate wchange prog
    Weight |    Participate in
    change |   wellness program
     level |        No        Yes |     Total
-----------+----------------------+----------
      Loss |       194        962 |     1,156
 No change |       306        188 |       494
      Gain |       152         14 |       166
-----------+----------------------+----------
     Total |       652      1,164 |     1,816
```

- Data are observational
- Table does not account for how observed covariates and/or unobserved errors that affect program participation also affect the outcome variable

I use an ordered probit model to control for observable covariates that could affect both `wchange` and `prog`

```
. eoprobit wchange i.prog age over phealth, vsquish nolog
Extended ordered probit regression              Number of obs     =      1,816
                                                Wald chi2(4)      =     548.00
Log likelihood = -1267.3173                     Prob > chi2       =     0.0000
```

wchange	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
prog						
Yes	-1.486537	.0687325	-21.63	0.000	-1.621251	-1.351824
age	.0371479	.0969554	0.38	0.702	-.1528811	.2271769
over	-.1682472	.0626191	-2.69	0.007	-.2909785	-.0455159
phealth	-.1378776	.0528111	-2.61	0.009	-.2413854	-.0343699
cut1	-.7693622	.076155			-.9186233	-.6201011
cut2	.5106948	.0763306			.3610895	.6603

```
. eoprobit wchange i.prog age over phealth, vsquish nolog
Extended ordered probit regression              Number of obs     =      1,816
                                                Wald chi2(4)      =     548.00
Log likelihood = -1267.3173                     Prob > chi2       =     0.0000
```

wchange	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
prog						
Yes	-1.486537	.0687325	-21.63	0.000	-1.621251	-1.351824
age	.0371479	.0969554	0.38	0.702	-.1528811	.2271769
over	-.1682472	.0626191	-2.69	0.007	-.2909785	-.0455159
phealth	-.1378776	.0528111	-2.61	0.009	-.2413854	-.0343699
cut1	-.7693622	.076155			-.9186233	-.6201011
cut2	.5106948	.0763306			.3610895	.6603

$$wchange = \begin{cases} \text{``}\textit{Loss}\text{''} & \text{if } \qquad \beta_1 \texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut1 \\ \text{``}\textit{No change}\text{''} & \text{if } cut1 < \beta_1 \texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut2 \\ \text{``}\textit{Gain}\text{''} & \text{if } cut2 < \beta_1 \texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \end{cases}$$

$$\mathbf{x}\boldsymbol{\beta} = \beta_2 \texttt{age} + \beta_3 \texttt{over} + \beta_4 \texttt{phealth}$$

```
. margins r.prog, contrast(nowald) post
Contrasts of predictive margins
Model VCE     : OIM
1._predict    : Pr(wchange==Loss), predict(outlevel(0))
2._predict    : Pr(wchange==No change), predict(outlevel(1))
3._predict    : Pr(wchange==Gain), predict(outlevel(2))
```

	Contrast	Delta-method Std. Err.	[95% Conf.	Interval]
prog@_predict				
(Yes vs No) 1	.5293751	.0213456	.4875385	.5712116
(Yes vs No) 2	-.313256	.0170586	-.3466903	-.2798217
(Yes vs No) 3	-.2161191	.0156092	-.2467126	-.1855256

- When everyone joins the program instead of when no one participants in the program,
  - On average, the probability of “Loss” goes up by .52
  - On average, the probability of “No change” goes down by .31
  - On average, the probability of “Gain” goes down .22

- I suspect that unobservables that increase program participation are negatively correlated with unobservables that affect weight gain

  Those most likely to participate are most likely to lose weight, after controlling for observable covariates
- I want a model that
  - allows observed covariates to affect both `wchange` and assignment to `prog`
  - allows the errors that affect `prog` to be correlated with the errors that affect `wchange`
- In other words, I want to model `prog` as endogenous

A model when prog is endogenous

$$
wchange = \begin{cases} \text{"}\mathit{Loss}\text{"} & \text{if } \qquad \beta_1 \texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut1 \\ \text{"}\mathit{No\ change}\text{"} & \text{if } cut1 < \beta_1 \texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut2 \\ \text{"}\mathit{Gain}\text{"} & \text{if } cut2 < \beta_1 \texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \end{cases}
$$

$$
\begin{aligned}
prog &= (\mathbf{x}\boldsymbol{\gamma} + \gamma_1 \texttt{wtime} + \eta > 0) \\
& \quad \epsilon \text{ and } \eta \text{ are correlated and joint normal} \\
\mathbf{x}\boldsymbol{\beta} &= \beta_2 \texttt{age} + \beta_3 \texttt{over} + \beta_4 \texttt{phealth} \\
\mathbf{x}\boldsymbol{\gamma} &= \gamma_2 \texttt{age} + \gamma_3 \texttt{over} + \gamma_4 \texttt{phealth}
\end{aligned}
$$

- `wtime` is an instrumental variable
  - It is included in the model for treatment
  - It is excluded from the model for the potential outcomes of `wchange`

$$
\begin{aligned}
wchange &= \begin{cases} \text{``}\mathit{Loss}\text{''} & \text{if } \qquad \beta_1\texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut1 \\ \text{``}\mathit{No\ change}\text{''} & \text{if } cut1 < \beta_1\texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \leq cut2 \\ \text{``}\mathit{Gain}\text{''} & \text{if } cut2 < \beta_1\texttt{prog} + \mathbf{x}\boldsymbol{\beta} + \epsilon \end{cases} \\
prog &= (\mathbf{x}\boldsymbol{\gamma} + \gamma_1\texttt{wtime} + \eta > 0) \\
&\quad \epsilon \text{ and } \eta \text{ are correlated and joint normal} \\
\mathbf{x}\boldsymbol{\beta} &= \beta_2\texttt{age} + \beta_3\texttt{over} + \beta_4\texttt{phealth} \\
\mathbf{x}\boldsymbol{\gamma} &= \gamma_2\texttt{age} + \gamma_3\texttt{over} + \gamma_4\texttt{phealth}
\end{aligned}
$$

Fit by: `eoprobit wchange age over phealth ,`
`endog(prog = age over phealth wtime, probit)`

```
. eoprobit wchange age over phealth ,                     ///
>        endog(prog = age over phealth wtprog, probit) ///
>        vsquish nolog
Extended ordered probit regression                Number of obs     =      1,816
                                                  Wald chi2(4)      =      98.47
Log likelihood = -2177.6691                       Prob > chi2       =     0.0000
```

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
wchange						
age	.204564	.0980909	2.09	0.037	.0123094	.3968186
over	.0278124	.0687223	0.40	0.686	-.1068808	.1625055
phealth	-.3028088	.0575207	-5.26	0.000	-.4155473	-.1900703
prog						
Yes	-.628258	.1582358	-3.97	0.000	-.9383945	-.3181215
prog						
age	-.8484251	.1076217	-7.88	0.000	-1.05936	-.6374904
over	-1.071231	.0757757	-14.14	0.000	-1.219748	-.9227131
phealth	.873563	.0623242	14.02	0.000	.7514097	.9957163
wtprog	1.618161	.113306	14.28	0.000	1.396086	1.840237
_cons	.0856418	.0687773	1.25	0.213	-.0491592	.2204428
/wchange						
cut1	-.2589072	.1119722			-.4783686	-.0394458
cut2	.927279	.0900163			.7508504	1.103708
corr(e.prog, e.wchange)	-.5305974	.0772131	-6.87	0.000	-.6649372	-.3630029

Log likelihood = -2177.6691 Prob > chi2 = 0.0000

	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
wchange						
age	.204564	.0980909	2.09	0.037	.0123094	.3968186
over	.0278124	.0687223	0.40	0.686	-.1068808	.1625055
phealth	-.3028088	.0575207	-5.26	0.000	-.4155473	-.1900703
prog						
Yes	-.628258	.1582358	-3.97	0.000	-.9383945	-.3181215
prog						
age	-.8484251	.1076217	-7.88	0.000	-1.05936	-.6374904
over	-1.071231	.0757757	-14.14	0.000	-1.219748	-.9227131
phealth	.873563	.0623242	14.02	0.000	.7514097	.9957163
wtprog	1.618161	.113306	14.28	0.000	1.396086	1.840237
_cons	.0856418	.0687773	1.25	0.213	-.0491592	.2204428
/wchange						
cut1	-.2589072	.1119722			-.4783686	-.0394458
cut2	.927279	.0900163			.7508504	1.103708
corr(e.prog, e.wchange)	-.5305974	.0772131	-6.87	0.000	-.6649372	-.3630029

- The coefficient on `wtprog` and its standard error give the impression that the instrument is relevant

cut2	.927279	.0900163			.7508504	1.103708
corr(e.prog, e.wchange)	-.5305974	.0772131	-6.87	0.000	-.6649372	-.3630029

- The nonzero correlation between `e.prog` and `e.wchange` indicates that `prog` is endogenous
- Those who are more likely to participate are more likely to lose weight

```
. margins r.prog,                                        ///
>         predict(fix(prog) outlevel("Loss"))            ///
>         predict(fix(prog) outlevel("No change"))       ///
>         predict(fix(prog) outlevel("Gain"))            ///
>         contrast(nowald)

Contrasts of predictive margins
Model VCE     : OIM

1._predict    : Pr(wchange==Loss), predict(fix(prog) outlevel("Loss"))
2._predict    : Pr(wchange==No change), predict(fix(prog) outlevel("No
                change"))
3._predict    : Pr(wchange==Gain), predict(fix(prog) outlevel("Gain"))

--------------------------------------------------------------------
                |             Delta-method
                |   Contrast   Std. Err.     [95% Conf. Interval]
----------------+---------------------------------------------------
 prog@_predict  |
(Yes vs No) 1   |    .231068    .0583617      .1166812    .3454547
(Yes vs No) 2   |   -.146159    .0392355     -.2230591   -.0692589
(Yes vs No) 3   |   -.084909    .0201163     -.1243361   -.0454818
--------------------------------------------------------------------
```

- When everyone joins the program instead of when no one participants in the program,
    - On average, the probability of “Loss” goes up by .23
    - On average, the probability of “No change” goes down by .15
    - On average, the probability of “Gain” goes down by .08

- `fix(prog)` gets us the effect of the program that is not contaminated by the correlation between ϵ and η that increases the participation among people more likely to lose weight

- If you specify `fix(prog)`, `predict` ignores the correlation between `prog` and ϵ in estimating the prediction
  - Specifying `fix(prog)` gets the prediction you want to estimate the effect of the progam that is not contaminated by the endogenous selection into the program

- If you do not specify `fix(prog)`, `predict` includes the correlation between `prog` and ϵ in estimating the prediction
  - Not specifying `fix(prog)` gets the prediction you want if you are betting on whether someone with specific covariates and program status will lose weight

- `fix(prog)` predictions are sometimes called the structural prediction or an average structural function; see Blundell and Powell (2003), Blundell and Powell (2004), Wooldridge (2010), and Wooldridge (2014),
- The difference between the mean of the average of the structural predictions when `prog=1` and the mean of the average of the structural predictions when `prog=0` is an average treatment effect (Blundell and Powell (2003) and Wooldridge (2014))

Standard errors for population versus sample

- The delta-method standard errors reported by `margins` hold the covariates fixed at their sample values
  - The delta-method standard errors are for a sample-average treatment effect instead of a population-averaged treatment effect
  - The sample-averaged treatment effect is for those individuals that showed up in that run of the treatment
  - The population-averaged treatment effect is for a random draw of individuals from the population
- To get standard errors for the population-average treatment effect, specify `vce(robust)` to the estimation command and specify `vce(unconditional)` to `margins`

```
. quietly eoprobit wchange age over phealth ,                ///
>         endog(prog = age over phealth wtprog, probit) ///
>         vce(robust)
. margins r.prog,                                     ///
>         predict(fix(prog) outlevel("Loss"))         ///
>         predict(fix(prog) outlevel("No change")) ///
>         predict(fix(prog) outlevel("Gain"))         ///
>         contrast(nowald) vce(unconditional) post
Contrasts of predictive margins
1._predict   : Pr(wchange==Loss), predict(fix(prog) outlevel("Loss"))
2._predict   : Pr(wchange==No change), predict(fix(prog) outlevel("No
               change"))
3._predict   : Pr(wchange==Gain), predict(fix(prog) outlevel("Gain"))
```

	Contrast	Unconditional Std. Err.	[95% Conf.	Interval]
prog@_predict				
(Yes vs No) 1	.231068	.0583663	.1166721	.3454639
(Yes vs No) 2	-.146159	.0391262	-.222845	-.069473
(Yes vs No) 3	-.084909	.0202105	-.1245208	-.0452971

Interacting an endogenous variable with other covariates

```
. eoprobit wchange i.prog i.prog#c.(age over phealth) ,   ///
>         endog(prog = age over phealth wtprog, nomain probit)    ///
>         vce(robust) vsquish nolog
Extended ordered probit regression                 Number of obs     =      1,816
                                                   Wald chi2(7)      =     111.63
Log pseudolikelihood = -2158.8165                  Prob > chi2       =     0.0000
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
wchange						
prog						
Yes	.0018457	.1781571	0.01	0.992	-.3473357	.3510272
prog#c.age						
No	.3123571	.1331677	2.35	0.019	.0513531	.573361
Yes	.0730845	.1298635	0.56	0.574	-.1814432	.3276122
prog#c.over						
No	.17194	.0854484	2.01	0.044	.0044641	.3394158
Yes	-.2479575	.1063778	-2.33	0.020	-.456454	-.0394609
prog#c.phealth						
No	-.0730391	.0899687	-0.81	0.417	-.2493744	.1032963
Yes	-.5054434	.0741897	-6.81	0.000	-.6508525	-.3600342
prog						
age	-.8543462	.106038	-8.06	0.000	-1.062177	-.6465156
over	-1.069359	.0736758	-14.51	0.000	-1.213761	-.9249569
phealth	.8570916	.0608459	14.09	0.000	.7378359	.9763473
wtprog	1.627213	.1077598	15.10	0.000	1.416007	1.838418

Log pseudolikelihood = -2158.8165 Prob > chi2 = 0.0000

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
wchange						
prog						
Yes	.0018457	.1781571	0.01	0.992	-.3473357	.3510272
prog#c.age						
No	.3123571	.1331677	2.35	0.019	.0513531	.573361
Yes	.0730845	.1298635	0.56	0.574	-.1814432	.3276122
prog#c.over						
No	.17194	.0854484	2.01	0.044	.0044641	.3394158
Yes	-.2479575	.1063778	-2.33	0.020	-.456454	-.0394609
prog# c.phealth						
No	-.0730391	.0899687	-0.81	0.417	-.2493744	.1032963
Yes	-.5054434	.0741897	-6.81	0.000	-.6508525	-.3600342
prog						
age	-.8543462	.106038	-8.06	0.000	-1.062177	-.6465156
over	-1.069359	.0736758	-14.51	0.000	-1.213761	-.9249569
phealth	.8570916	.0608459	14.09	0.000	.7378359	.9763473
wtprog	1.627213	.1077598	15.10	0.000	1.416007	1.838418
_cons	.0965657	.0688104	1.40	0.161	-.0383003	.2314316
/wchange						
cut1	.0358062	.115777			-.1911124	.2627249
cut2	1.227726	.097207			1.037204	1.418248
corr(e.prog, e.wchange)	-.5476024	.076449	-7.16	0.000	-.6799189	-.3807508

```
. margins r.prog,                                  ///
>         predict(fix(prog) outlevel("Loss"))      ///
>         predict(fix(prog) outlevel("No change")) ///
>         predict(fix(prog) outlevel("Gain"))      ///
>         contrast(nowald) vce(unconditional) post
Contrasts of predictive margins
1._predict   : Pr(wchange==Loss), predict(fix(prog) outlevel("Loss"))
2._predict   : Pr(wchange==No change), predict(fix(prog) outlevel("No
               change"))
3._predict   : Pr(wchange==Gain), predict(fix(prog) outlevel("Gain"))
```

	Contrast	Unconditional Std. Err.	[95% Conf.	Interval]
prog@_predict				
(Yes vs No) 1	.2357078	.0600875	.1179385	.3534772
(Yes vs No) 2	-.1546622	.0401367	-.2333286	-.0759957
(Yes vs No) 3	-.0810456	.0209271	-.1220621	-.0400292

- When everyone joins the program instead of when no one participants in the program,
  - On average, the probability of “Loss” goes up by .24
  - On average, the probability of “No change” goes down by .15
  - On average, the probability of “Gain” goes down by .08

Endogenous treatment model

```
. eoprobit wchange (age over phealth) ,            ///
>         entreat(prog = age over phealth wtprog ) ///
>         vce(robust) vsquish nolog
Extended ordered probit regression              Number of obs     =      1,816
                                                Wald chi2(6)      =      61.42
Log pseudolikelihood = -2158.1656               Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wchange      |
  prog#c.age |
          No |   .3122714   .1314859     2.37   0.018     .0545638    .5699791
         Yes |    .071914   .1308732     0.55   0.583    -.1845927    .3284208
 prog#c.over |
          No |   .1742641   .0843392     2.07   0.039     .0089624    .3395659
         Yes |  -.2519632    .107001    -2.35   0.019    -.4616814    -.042245
       prog# |
   c.phealth |
          No |  -.0765452   .0887458    -0.86   0.388    -.2504837    .0973933
         Yes |  -.5094441   .0751039    -6.78   0.000     -.656645   -.3622432
-------------+----------------------------------------------------------------
prog         |
         age |  -.8545688   .1060258    -8.06   0.000    -1.062375   -.6467621
        over |  -1.069774   .0736061   -14.53   0.000     -1.21404   -.9255089
     phealth |   .8569976   .0608534    14.08   0.000     .7377271    .9762682
      wtprog |   1.627411    .107371    15.16   0.000     1.416967    1.837854
```

	Coef.	Robust Std. Err.	z	P>|z|	[95% Conf.	Interval]
wchange						
prog#c.age						
No	.3122714	.1314859	2.37	0.018	.0545638	.5699791
Yes	.071914	.1308732	0.55	0.583	-.1845927	.3284208
prog#c.over						
No	.1742641	.0843392	2.07	0.039	.0089624	.3395659
Yes	-.2519632	.107001	-2.35	0.019	-.4616814	-.042245
prog# c.phealth						
No	-.0765452	.0887458	-0.86	0.388	-.2504837	.0973933
Yes	-.5094441	.0751039	-6.78	0.000	-.656645	-.3622432
prog						
age	-.8545688	.1060258	-8.06	0.000	-1.062375	-.6467621
over	-1.069774	.0736061	-14.53	0.000	-1.21404	-.9255089
phealth	.8569976	.0608534	14.08	0.000	.7377271	.9762682
wtprog	1.627411	.107371	15.16	0.000	1.416967	1.837854
_cons	.0978712	.0689011	1.42	0.155	-.0371724	.2329148
/wchange						
prog#c.cut1						
No	.0527717	.1152427			-.1730998	.2786432
Yes	.0186214	.107202			-.1914907	.2287335
prog#c.cut2						
No	1.210627	.0970201			1.020471	1.400783
Yes	1.301471	.151592			1.004356	1.598586
corr(e.prog, e.wchange)	-.5501941	.0753943	-7.30	0.000	-.680788	-.3856995

prog#c.age						
No	.3122714	.1314859	2.37	0.018	.0545638	.5699791
Yes	.071914	.1308732	0.55	0.583	-.1845927	.3284208
prog#c.over						
No	.1742641	.0843392	2.07	0.039	.0089624	.3395659
Yes	-.2519632	.107001	-2.35	0.019	-.4616814	-.042245
prog# c.phealth						
No	-.0765452	.0887458	-0.86	0.388	-.2504837	.0973933
Yes	-.5094441	.0751039	-6.78	0.000	-.656645	-.3622432
prog						
age	-.8545688	.1060258	-8.06	0.000	-1.062375	-.6467621
over	-1.069774	.0736061	-14.53	0.000	-1.21404	-.9255089
phealth	.8569976	.0608534	14.08	0.000	.7377271	.9762682
wtprog	1.627411	.107371	15.16	0.000	1.416967	1.837854
_cons	.0978712	.0689011	1.42	0.155	-.0371724	.2329148
/wchange						
prog#c.cut1						
No	.0527717	.1152427			-.1730998	.2786432
Yes	.0186214	.107202			-.1914907	.2287335
prog#c.cut2						
No	1.210627	.0970201			1.020471	1.400783
Yes	1.301471	.151592			1.004356	1.598586
corr(e.prog, e.wchange)	-.5501941	.0753943	-7.30	0.000	-.680788	-.3856995

```
. estat teffects
Predictive margins                              Number of obs     =      1,816
ATE_Pr0      : Pr(wchange=0=Loss)
ATE_Pr1      : Pr(wchange=1=No change)
ATE_Pr2      : Pr(wchange=2=Gain)
```

	Margin	Unconditional Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE_Pr0 prog (Yes vs No)	.2252647	.0600534	3.75	0.000	.1075623	.3429671
ATE_Pr1 prog (Yes vs No)	-.1349272	.0438049	-3.08	0.002	-.2207833	-.0490711
ATE_Pr2 prog (Yes vs No)	-.0903375	.0216817	-4.17	0.000	-.1328328	-.0478422

- When everyone joins the program instead of when no one participants in the program,
  - On average, the probability of “Loss” goes up by .23
  - On average, the probability of “No change” goes down by .14
  - On average, the probability of “Gain” goes down .09

```
. margins r.prog,                                  ///
>         predict(fix(prog) outlevel("Loss"))      ///
>         predict(fix(prog) outlevel("No change")) ///
>         predict(fix(prog) outlevel("Gain"))      ///
>         contrast(nowald) vce(unconditional) post

Contrasts of predictive margins

1._predict    : Pr(wchange==Loss), predict(fix(prog) outlevel("Loss"))
2._predict    : Pr(wchange==No change), predict(fix(prog) outlevel("No
               change"))
3._predict    : Pr(wchange==Gain), predict(fix(prog) outlevel("Gain"))

-----------------------------------------------------------------
               |               Unconditional
               |   Contrast   Std. Err.     [95% Conf. Interval]
---------------+-------------------------------------------------
 prog@_predict |
(Yes vs No) 1  |   .2252647   .0600534      .1075623    .3429671
(Yes vs No) 2  |  -.1349272   .0438049     -.2207833   -.0490711
(Yes vs No) 3  |  -.0903375   .0216817     -.1328328   -.0478422
-----------------------------------------------------------------
```

Endogenous sample selection

- Reconsider our fictional weight-loss program
    - Some program participants and some nonparticipants did not show up for the final weigh in
      This is commonly known as lost to follow up
    - If unobservables that affect whether someone is lost to follow up
        - are independent of the unobservables that affect program participantion
        - and they are independent of the unobservables that affect the outcomes with and without the program,
    - the previously discussed estimator consistently estimates the effects
- Any dependence among the unobservables must be modeled

Data

```
. describe
Contains data from wprogram2.dta
  obs:         3,000
 vars:             8                          6 Sep 2017 12:10
 size:        96,000
-------------------------------------------------------------------------------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------------
wchange         float   %9.0g      changel    Weight change level
age             float   %9.0g                 Years over 50
over            float   %9.0g                 Overweight (tens of pounds)
phealth         float   %9.0g                 Prior health score
prog            float   %9.0g      yesno      Participate in wellness program
wtprog          float   %9.0g      yesno      Offered work time to participate
                                                in program
wtsamp          float   %9.0g                 Offered work time to participate
                                                in sample
insamp          float   %9.0g                 In sample: attended initial and
                                                final weigh in
-------------------------------------------------------------------------------
Sorted by:
     Note: Dataset has changed since last saved.
```

$$
\begin{aligned}
\textit{insamp} &= (\mathbf{x}\boldsymbol{\alpha} + \alpha_1 \texttt{wtsamp} + \xi > 0) \\
\textit{prog} &= (\mathbf{x}\boldsymbol{\gamma} + \gamma_1 \texttt{wtprog} + \eta > 0) \\
\textit{wchange} &= \begin{cases} \text{"}\textit{Loss}\text{"} & \text{if } \quad\quad \mathbf{x}\boldsymbol{\beta}_0 + \epsilon \leq \textit{cut}1_0 \\ \text{"}\textit{No change}\text{"} & \text{if } \textit{cut}1_0 < \mathbf{x}\boldsymbol{\beta}_0 + \epsilon \leq \textit{cut}2_0 \\ \text{"}\textit{Gain}\text{"} & \text{if } \textit{cut}2_0 < \mathbf{x}\boldsymbol{\beta}_0 + \epsilon \end{cases} \\
\mathbf{x}\boldsymbol{\beta}_0 &= \beta_{1,0}\texttt{age} + \beta_{2,0}\texttt{over} + \beta_{3,0}\texttt{phealth} \\
&\quad \text{for the observations at which } \texttt{prog}{=}0\text{, and} \\
\textit{wchange} &= \begin{cases} \text{"}\textit{Loss}\text{"} & \text{if } \quad\quad \mathbf{x}\boldsymbol{\beta}_1 + \epsilon \leq \textit{cut}1_1 \\ \text{"}\textit{No change}\text{"} & \text{if } \textit{cut}1_1 < \mathbf{x}\boldsymbol{\beta}_1 + \epsilon \leq \textit{cut}2_1 \\ \text{"}\textit{Gain}\text{"} & \text{if } \textit{cut}2_1 < \mathbf{x}\boldsymbol{\beta}_1 + \epsilon \end{cases} \\
\mathbf{x}\boldsymbol{\beta}_1 &= \beta_{1,1}\texttt{age} + \beta_{2,1}\texttt{over} + \beta_{3,1}\texttt{phealth} \\
&\quad \text{for the observations at which } \texttt{prog}{=}1 \\
&\quad \xi, \epsilon \text{ and } \eta \text{ are correlated and joint normal}
\end{aligned}
$$

Fit by:
```
eoprobit wchange (age over phealth) ,
        entreat(prog = age over phealth wtprog )
        select(samp = age over phealth wtsamp )
        vce(robust)
```

```
. eoprobit wchange (age over phealth) ,        ///
>         entreat(prog = age over phealth wtprog ) ///
>         select(insamp = age over phealth wtsamp )  ///
>         vce(robust) vsquish nolog
Extended ordered probit regression              Number of obs     =      3,000
                                                      Selected    =      1,816
                                                      Nonselected =      1,184
                                                Wald chi2(6)      =     200.16
Log pseudolikelihood = -4402.4852               Prob > chi2       =     0.0000
------------------------------------------------------------------------------
             |               Robust
             |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
wchange      |
  prog#c.age |
         No  |   .2977275   .1074884     2.77   0.006     .0870541    .5084009
        Yes  |   .1653358   .1026185     1.61   0.107    -.0357928    .3664644
 prog#c.over |
         No  |   .5347524   .0680259     7.86   0.000      .401424    .6680808
        Yes  |     .23094   .0953423     2.42   0.015     .0440725    .4178075
       prog# |
   c.phealth |
         No  |  -.4092577   .0708874    -5.77   0.000    -.5481944   -.2703209
        Yes  |    -.75997   .0626118   -12.14   0.000    -.8826868   -.6372532
-------------+----------------------------------------------------------------
insamp       |
         age |   .0511126   .0789532     0.65   0.517    -.1036328    .2058579
        over |  -.7893401   .0445127   -17.73   0.000    -.8765835   -.7020968
     phealth |   .7739903   .0461381    16.78   0.000     .6835613    .8644193
      wtsamp |    2.20639   .4215291     5.23   0.000     1.380208    3.032571
       _cons |   .3026734   .0507938     5.96   0.000     .2031193    .4022275
```

c.phealth						
No	-.4092577	.0708874	-5.77	0.000	-.5481944	-.2703209
Yes	-.75997	.0626118	-12.14	0.000	-.8826868	-.6372532
insamp						
age	.0511126	.0789532	0.65	0.517	-.1036328	.2058579
over	-.7893401	.0445127	-17.73	0.000	-.8765835	-.7020968
phealth	.7739903	.0461381	16.78	0.000	.6835613	.8644193
wtsamp	2.20639	.4215291	5.23	0.000	1.380208	3.032571
_cons	.3026734	.0507938	5.96	0.000	.2031193	.4022275
prog						
age	-.9408839	.0823665	-11.42	0.000	-1.102319	-.7794485
over	-1.061503	.050071	-21.20	0.000	-1.15964	-.9633653
phealth	.8896701	.0494006	18.01	0.000	.7928467	.9864935
wtprog	1.629244	.0764087	21.32	0.000	1.479486	1.779002
_cons	.0199176	.0530267	0.38	0.707	-.0840128	.1238481
/wchange						
prog#c.cut1						
No	-.3821007	.0926799			-.5637499	-.2004514
Yes	-.4393841	.0802464			-.5966641	-.2821041
prog#c.cut2						
No	.5051071	.1022236			.3047525	.7054618
Yes	.5437111	.1399479			.2694182	.818004
corr(e.insamp, e.wchange)	-.8266016	.0514301	-16.07	0.000	-.9043439	-.6957701
corr(e.prog, e.wchange)	-.4910402	.0594322	-8.26	0.000	-.5985767	-.366119
corr(e.prog, e.insamp)	.0835352	.0350767	2.38	0.017	.0144972	.1517805

Yes	.5437111	.1399479			.2694182	.818004
corr(e.insamp, e.wchange)	-.8266016	.0514301	-16.07	0.000	-.9043439	-.6957701
corr(e.prog, e.wchange)	-.4910402	.0594322	-8.26	0.000	-.5985767	-.366119
corr(e.prog, e.insamp)	.0835352	.0350767	2.38	0.017	.0144972	.1517805

- Nonzero correlation between `e.insamp` and `e.wchange` implies endogenous sample selection for outcomes
    - Those more likely to show up for final weigh in are more likely to lose weight
- Nonzero correlation between `e.prog` and `e.wchange` implies endogenous treatment assignment
    - Those more likely to participate in program are more likely to lose weight
- Nonzero correlation between `e.prog` and `e.insamp` implies endogenous sample selection for program
    - Those more likely to participate in program are more likely to show up for the final weigh in

```
. estat teffects
Predictive margins                              Number of obs     =      3,000
ATE_Pr0         : Pr(wchange=0=Loss)
ATE_Pr1         : Pr(wchange=1=No change)
ATE_Pr2         : Pr(wchange=2=Gain)
```

	Margin	Unconditional Std. Err.	z	P>|z|	[95% Conf.	Interval]
ATE_Pr0 prog (Yes vs No)	.1616204	.0403782	4.00	0.000	.0824805	.2407603
ATE_Pr1 prog (Yes vs No)	.0021599	.0256098	0.08	0.933	-.0480345	.0523542
ATE_Pr2 prog (Yes vs No)	-.1637803	.0372978	-4.39	0.000	-.2368826	-.0906779

- When everyone joins the program instead of when no one participants in the program,
  - On average, the probability of "Loss" goes up by .16
  - On average, the probability of "No change" does not change
  - On average, the probability of "Gain" goes down by .16

Modeling endogeneity and sample selection can make a difference

```
. estimates table exog linear interact full essample
```

Variable	exog	linear	interact	full	essample
prog@ _predict					
(Yes vs No) 1	.52937505	.23106797	.23570783	.22526472	.16162041
(Yes vs No) 2	-.31325598	-.14615901	-.15466218	-.13492724	.00215985
(Yes vs No) 3	-.21611907	-.08490896	-.08104565	-.09033748	-.16378026

More about ERM commands

- Extended regression model (ERM) is a Stata term for a class of regression models
- The commands `eregress`, `eprobit`, and `eintreg` fit ERMs handle continuous-and-unbounded, binary, and censored/corner outcomes
- Look at

  http://www.stata.com/manuals/erm.pdf

  for more examples and a wealth of details

Blundell, R. W., and J. L. Powell. 2003. Endogeity in nonparametric and semiparametric regression models. In *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, ed. M. Dewatripont, L. P. Hansen, and S. J. Turnovsky, vol. 2, 312–357. Cambridge: Cambridge University Press.

———. 2004. Endogeneity in semiparametric binary response models. *Review of Economic Studies* 71: 655–679.

Wooldridge, J. M. 2010. *Econometric Analysis of Cross Section and Panel Data*. 2nd ed. Cambridge, Massachusetts: MIT Press.

———. 2014. Quasi-maximum likelihood estimation and testing for nonlinear models with endogenous explanatory variables. *Journal of Econometrics* 182: 226–234.