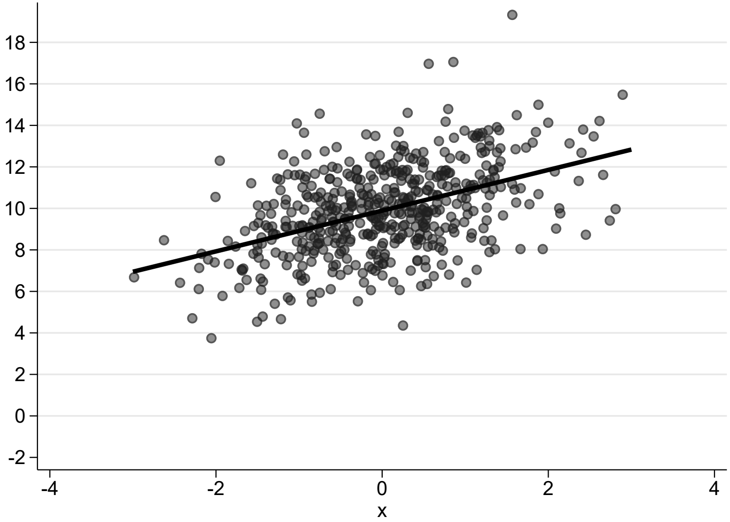
Heteroskedasticity and Heterogeneity in Sample Selection Models

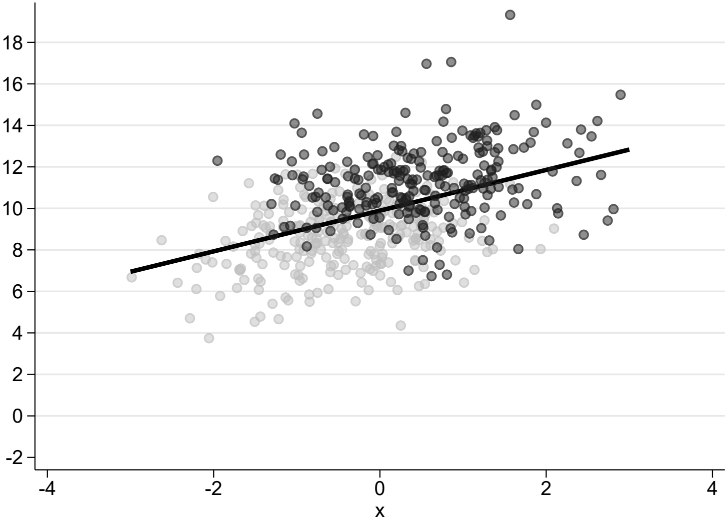
Alyssa H. Carlson
University of Missouri

November, 2024
Virtual Stata Symposium

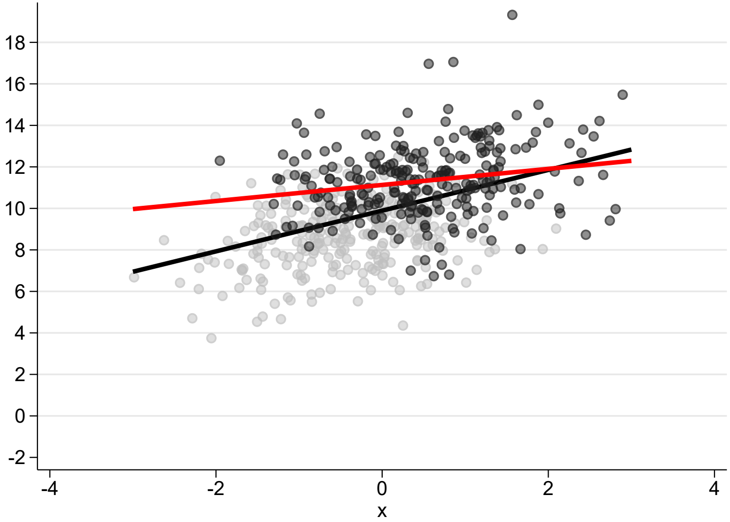
Intro



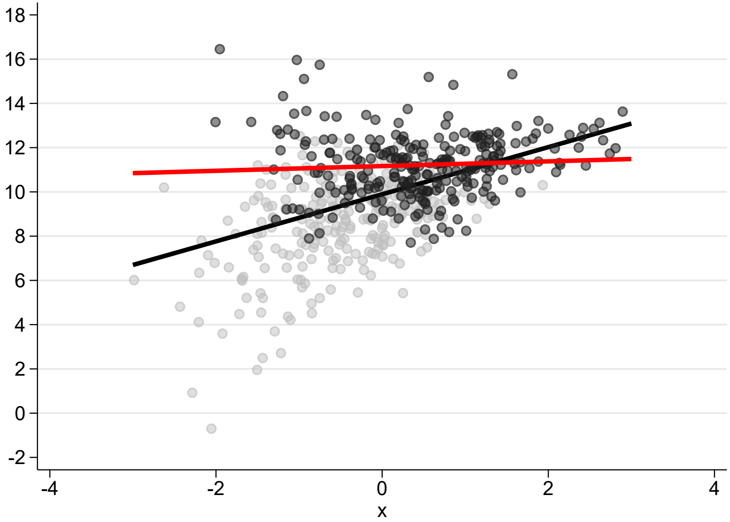
Intro



Intro



Intro



Today's talk

1. What causes heteroskedasticity in Sample Selection models?
   - heterogeneity!
2. What are the consequences of heteroskedasticity in Sample Selection models?
   - LIML vs FIML estimators
   - heteroskedasticity in outcome vs selection equation
3. Can we test for heteroskedasticity?
   - LIML over FIML – (demeaned) Breusch and Pagan (1979) test and Hausman (1978) test
   - Validity of LIML – MCC test
4. Is there an alternative estimator for sample selection models with general forms of heteroskedasticity.
   - `gtsheckman`

Today's talk

Carlson and Joshi (2024) " Sample Selection in linear panel data models with heterogenous coefficients," *Journal of Applied Econometrics*, 39(2), 237-255.

Carlson (2022) "GTSHECKMAN: Stata module to compute generalised two-step Heckman selection model," Statistical Software Components, Boston College Department of Economics.

- Forthcoming Stata Journal article (Carlson, forthcoming)

Carlson and Zhao (2023) "Heckman sample selection estimators under heteroskedasticity," Working Paper 2303, Department of Economics, University of Missouri.

- Forthcoming Stata Journal article

Sample Selection Model

The outcome is modeled as

$$y_i = \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} \tag{1}$$

but the outcome is not always observed.

y_i is only observed when $s_i = 1$,

$$s_i = 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) \tag{2}$$

- both $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ include a constant
- often $\mathbf{x}_{2i} = (\mathbf{x}_{1i}, \mathbf{w}_i)$
- Ex: Estimating married woman wages

$$\begin{aligned} \ln(wage_i) =& \beta_0 + educ_i\beta_1 + u_{1i} \\ inlf_i =& 1(\delta_0 + educ_i\delta_1 + nwifinc_i\delta_2 + u_{2i} > 0) \end{aligned}$$

Sample Selection Model

$$y_i = \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} \tag{1}$$
$$s_i = 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) \tag{2}$$

Problem: want to estimate

$$\mathrm{E}(y_i|\mathbf{x}_{1i}) = \mathbf{x}_{1i}\boldsymbol{\beta}$$

but you can only use the observed sample,

$$\mathrm{E}(y_i|\mathbf{x}_{1i}, s_i = 1) \neq \mathbf{x}_{1i}\boldsymbol{\beta}$$

if u_{1i} is correlated with u_{2i}

Sample Selection Estimators

$$
\begin{aligned}
y_i =& \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} && (1)\\
s_i =& 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) && (2)
\end{aligned}
$$

Heckman (1979) assumes

$$
\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \Bigg| \mathbf{x}_{1i}, \mathbf{x}_{2i} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right) \tag{3}
$$

Which suggests two possible estimators:

Sample Selection Estimators

$$y_i = \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} \tag{1}$$
$$s_i = 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) \tag{2}$$

Heckman (1979) assumes

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \Bigg| \mathbf{x}_{1i}, \mathbf{x}_{2i} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right) \tag{3}$$

Which suggests two possible estimators:

1. Full information ML (FIML): maximum likelihood over the joint distribution of y_i and s_i.
   - Requires joint distribution to be correctly specified

FIML

Stata command:

`heckman` *depvar* [*indepvars*] , `select(`*depvar*$_s$ `=` *varlist*$_s$`)`

1. Maximize the joint log likelihood

$$\begin{aligned}\ell_i =& (1-s_i)\ln[1-\Phi(\mathbf{x}_{2i}\boldsymbol{\delta})] + s_i \ln\left[\Phi\left(\frac{\mathbf{x}_{2i}\boldsymbol{\delta}+\rho(y_i-\mathbf{x}_{1i}\boldsymbol{\beta})/\sigma_1}{\sqrt{1-\rho^2}}\right)\right] \\ &- s_i\left[\frac{(y_i-\mathbf{x}_{1i}\boldsymbol{\beta})^2}{2\sigma_1^2}+\ln(\sigma_1)\right]\end{aligned}$$

   with respect to $\boldsymbol{\delta}, \boldsymbol{\beta}, \rho, \sigma_1$.

Sample Selection Estimators

$$
\begin{aligned}
y_i =& \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} && (1) \\
s_i =& 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) && (2)
\end{aligned}
$$

Heckman (1979) assumes

$$
\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \Bigg| \mathbf{x}_{1i}, \mathbf{x}_{2i} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right) \tag{3}
$$

Which suggests two possible estimators:

2. Limit information ML (LIML): two-step estimator based on the conditional distribution of $y_i \mid s_i = 1$
    - Requires **Minimial Consistency Condition (MCC)** (Wooldridge, 2010, Assumption 19.1):

$$
\begin{aligned}
u_{2i} \mid \mathbf{x}_{1i}, \mathbf{x}_{2i} &\sim \mathrm{N}(0, 1) \\
\mathrm{E}(u_{1i} \mid u_{2i}, \mathbf{x}_{1i}, \mathbf{x}_{2i}) &= \gamma u_{2i}
\end{aligned} \tag{4}
$$

LIML

Under MCC

$$\mathrm{E}(y_i \mid s_i = 1, \mathbf{x}_{1i}, \mathbf{x}_{2i}) = \mathbf{x}_{1i}\boldsymbol{\beta} + \gamma\lambda(\mathbf{x}_{2i}\boldsymbol{\delta}) \tag{5}$$

where

$$\lambda(\mathbf{x}_{2i}\boldsymbol{\delta}) \equiv \frac{\phi(\mathbf{x}_{2i}\boldsymbol{\delta})}{\Phi(\mathbf{x}_{2i}\boldsymbol{\delta})} = \mathrm{E}(u_{2i} \mid s_i = 1, \mathbf{x}_{1i}, \mathbf{x}_{2i})$$

Stata command:

`heckman` *depvar* [*indepvars*] , `select(`*depvar*$_s$ `=` *varlist*$_s$`) twostep`

1. Estimate the binary choice in equation (2) using `probit`, calculate the estimated inverse mills ratio:

$$\widehat{\lambda}_i = \frac{\phi(\mathbf{x}_{2i}\widehat{\boldsymbol{\delta}})}{\Phi(\mathbf{x}_{2i}\widehat{\boldsymbol{\delta}})}$$

2. Estimate the following augmented regression:

$$y_i = \mathbf{x}_{1i}\boldsymbol{\beta} + \gamma\widehat{\lambda}_i + \varepsilon_i.$$

FIML and LIML in Stata

```
. use http://fmwww.bc.edu/ec-p/data/wooldridge/mroz, clear

. reg lwage educ

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(1, 426)       =     56.93
       Model |  26.3264237         1  26.3264237   Prob > F        =    0.0000
    Residual |  197.001028       426  .462443727   R-squared       =    0.1179
-------------+----------------------------------   Adj R-squared   =    0.1158
       Total |  223.327451       427  .523015108   Root MSE        =    .68003

------------------------------------------------------------------------------
       lwage | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .1086487   .0143998     7.55   0.000     .0803451    .1369523
       _cons |  -.1851969   .1852259    -1.00   0.318    -.5492674    .1788735
------------------------------------------------------------------------------
```

Mroz (1987) PSID data on the wages of 428 working, married women

FIML and LIML in Stata

```
. heckman lwage educ, select(inlf = educ nwifeinc) nolog

Heckman selection model                         Number of obs     =        753
(regression model with sample selection)              Selected    =        428
                                                      Nonselected =        325

                                                Wald chi2(1)      =      49.21
Log likelihood = -929.6295                      Prob > chi2       =     0.0000

------------------------------------------------------------------------------
             | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
lwage        |
        educ |   .1176578   .0167722     7.02   0.000     .0847848    .1505307
       _cons |  -.3920955   .2700523    -1.45   0.147    -.9213882    .1371972
-------------+----------------------------------------------------------------
inlf         |
        educ |   .1433212   .0226377     6.33   0.000     .0989522    .1876902
    nwifeinc |  -.0216104   .0043381    -4.98   0.000    -.0301129   -.0131079
       _cons |  -1.144077   .2656869    -4.31   0.000    -1.664813   -.6233399
-------------+----------------------------------------------------------------
     /athrho |    .209972   .2003914     1.05   0.295    -.1827879    .6027318
    /lnsigma |  -.3759879   .0415655    -9.05   0.000    -.4574549   -.2945209
-------------+----------------------------------------------------------------
         rho |   .2069397   .1918098                      -.180779    .5389906
       sigma |   .6866106   .0285393                      .6328924    .7448884
      lambda |    .142087   .1351505                     -.1228031    .4069771
------------------------------------------------------------------------------
LR test of indep. eqns. (rho = 0): chi2(1) = 0.85           Prob > chi2 = 0.3575
```

FIML and LIML in Stata

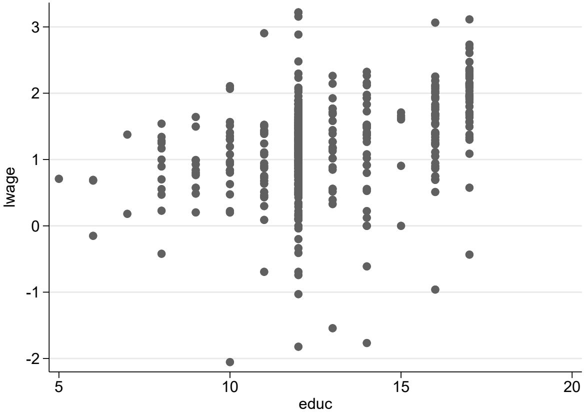
```
. heckman lwage educ, select(inlf = educ nwifeinc) twostep

Heckman selection model -- two-step estimates   Number of obs     =        753
(regression model with sample selection)              Selected    =        428
                                                      Nonselected =        325

                                                Wald chi2(1)      =      34.07
                                                Prob > chi2       =     0.0000
```

	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
lwage						
educ	.1282506	.021972	5.84	0.000	.0851862	.171315
_cons	-.6339939	.4179628	-1.52	0.129	-1.453186	.1851981
inlf						
educ	.1418686	.0225342	6.30	0.000	.0977025	.1860348
nwifeinc	-.0213744	.0043692	-4.89	0.000	-.0299378	-.0128109
_cons	-1.130936	.2644248	-4.28	0.000	-1.649199	-.6126727
/mills						
lambda	.306887	.2544542	1.21	0.228	-.1918341	.8056081
rho	0.42874					
sigma	.71578623					

Introducing Heteroskedasticity

What causes heteroskedasticity in Sample Selection models?

- Variation in wages is changing for different education levels



Introducing Heteroskedasticity

What causes heteroskedasticity in Sample Selection models?

- Heterogeneous effects

$$\ln(wage_i) = \beta_0 + educ_i b_{1i} + u_{1i} \tag{6}$$

$$inlf_i = 1(\delta_0 + educ_i d_{1i} + nwifinc_i d_{2i} + u_{2i} > 0) \tag{7}$$

Introducing Heteroskedasticity

What causes heteroskedasticity in Sample Selection models?

- Heterogeneous effects

$$\ln(wage_i) = \beta_0 + educ_i b_{1i} + u_{1i} \tag{6}$$

$$inlf_i = 1(\delta_0 + educ_i d_{1i} + nwifinc_i d_{2i} + u_{2i} > 0) \tag{7}$$

let $\beta_1 = \mathrm{E}(b_{1i})$, $\delta_1 = \mathrm{E}(d_{1i})$, and $\delta_2 = \mathrm{E}(d_{2i})$, then

$$\ln(wage_i) = \beta_0 + educ_i \beta_1 + \tilde{u}_{1i} \tag{8}$$

$$inlf_i = 1(\delta_0 + educ_i \delta_1 + nwifinc_i \delta_2 + \tilde{u}_{2i} > 0) \tag{9}$$

where

$$\tilde{u}_{1i} = u_{1i} + (b_{1i} - \beta_1) educ_i \tag{10}$$

$$\tilde{u}_{2i} = u_{2i} + (d_{1i} - \delta_1) educ_i + (d_{2i} - \delta_2) nwifinc_i \tag{11}$$

Introducing Heteroskedasticity

What causes heteroskedasticity in Sample Selection models?

- Heterogeneous effects

$$\begin{aligned} \ln(wage_i) =& \beta_0 + educ_i b_{1i} + u_{1i} & (6) \\ inlf_i =& 1(\delta_0 + educ_i d_{1i} + nwifinc_i d_{2i} + u_{2i} > 0) & (7) \end{aligned}$$

let $\beta_1 = \mathrm{E}(b_{1i})$, $\delta_1 = \mathrm{E}(d_{1i})$, and $\delta_2 = \mathrm{E}(d_{2i})$, then

$$\begin{aligned} \ln(wage_i) =& \beta_0 + educ_i \beta_1 + \tilde{u}_{1i} & (8) \\ inlf_i =& 1(\delta_0 + educ_i \delta_1 + nwifinc_i \delta_2 + \tilde{u}_{2i} > 0) & (9) \end{aligned}$$

then

$$\begin{aligned} \mathrm{Var}(\tilde{u}_{1i} \mid educ_i, nwifinc_i) =& \sigma^2 + \sigma^2_{b1} educ_i^2 \\ \mathrm{Var}(\tilde{u}_{2i} \mid educ_i, nwifinc_i) =& 1 + \sigma^2_{d1} educ_i^2 + \sigma^2_{d2} nwifinc_i^2 \\ & + \sigma^2_{d1d2} educ_i \times nwifinc_i \end{aligned}$$

$$\mathrm{Cov}(\tilde{u}_{1i}, \tilde{u}_{2i} \mid educ_i, nwifinc_i) = \rho\sigma + \sigma_{b1,d1} educ_i^2 + \sigma_{b1,d2} educ_i \times nwifinc_i$$

(assuming $(u_{1i}, u_{2i}) \perp (b_{1i}, d_{1i}, d_{2i})$)

Introducing Heteroskedasticity

$$
\begin{aligned}
y_i =& \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} && (1)\\
s_i =& 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) && (2)
\end{aligned}
$$

Suppose we have heteroskedasticity

$$
\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \rho_i\sigma_{1i}\sigma_{2i} \\ \rho_i\sigma_{1i}\sigma_{2i} & \sigma_{2i}^2 \end{pmatrix} \right) \tag{12}
$$

What are the consequences?

Introducing Heteroskedasticity

$$
\begin{aligned}
y_i =& \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} && (1)\\
s_i =& 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) && (2)
\end{aligned}
$$

Suppose we have heteroskedasticity

$$
\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \rho_i\sigma_{1i}\sigma_{2i} \\ \rho_i\sigma_{1i}\sigma_{2i} & \sigma_{2i}^2 \end{pmatrix} \right) \tag{12}
$$

What are the consequences?

1. FIML
    - joint distribution is misspecified $\rightarrow$ **inconsistent!**
    - robust standard errors does not fix this!

Introducing Heteroskedasticity

$$
\begin{aligned}
y_i =& \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} && (1)\\
s_i =& 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) && (2)
\end{aligned}
$$

Suppose we have heteroskedasticity

$$
\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \rho_i\sigma_{1i}\sigma_{2i} \\ \rho_i\sigma_{1i}\sigma_{2i} & \sigma_{2i}^2 \end{pmatrix} \right) \tag{12}
$$

What are the consequences?

1. FIML
   - joint distribution is misspecified → **inconsistent!**
   - robust standard errors does not fix this!
2. LIML
   - if MCC holds → **consistent**

     $$
     \begin{aligned}
     \sigma_{2i} &= 1\\
     \rho_i\sigma_{1i} &= \gamma
     \end{aligned}
     $$

     but need robust standard errors!
   - if MCC does not hold → **inconsistent!**

Introducing Heteroskedasticity

Stata LIML estimator does not produce heteroskedastic robust standard errors,

```
. heckman lwage educ, select(inlf = educ nwifeinc) twostep vce(robust)
vcetype robust not allowed
r(198);
```

Introducing Heteroskedasticity

DGP 1: Homoskedastic

Estimator	N	Bias	StdDev	SE	RSE	CR	RCR
FIML	500	-0.004	0.108	0.103	0.104	0.94	0.94
LIML	500	-0.004	0.112	0.111	0.110	0.95	0.94
FIML	1000	-0.002	0.075	0.073	0.073	0.94	0.94
LIML	1000	-0.003	0.079	0.078	0.078	0.95	0.94
FIML	2000	0.000	0.052	0.051	0.051	0.95	0.95
LIML	2000	0.002	0.056	0.055	0.055	0.95	0.94

Introducing Heteroskedasticity

DGP 2: Heteroskedastic (MCC)

Estimator	N	Bias	StdDev	SE	RSE	CR	RCR
FIML	**500**	**0.240**	**0.264**	**0.115**	**0.180**	**0.38**	**0.64**
LIML	**500**	**-0.010**	**0.249**	**0.144**	**0.236**	**0.76**	**0.95**
FIML	**1000**	**0.253**	**0.182**	**0.081**	**0.134**	**0.23**	**0.48**
LIML	**1000**	**-0.008**	**0.188**	**0.103**	**0.176**	**0.72**	**0.94**
FIML	**2000**	**0.263**	**0.111**	**0.057**	**0.099**	**0.07**	**0.21**
LIML	**2000**	**0.001**	**0.130**	**0.073**	**0.128**	**0.74**	**0.96**

Introducing Heteroskedasticity

DGP 3: Heteroskedastic (no MCC)

Estimator	N	Bias	StdDev	SE	RSE	CR	RCR
FIML	500	0.898	0.163	0.171	0.158	0.00	0.00
LIML	500	0.807	0.146	0.171	0.138	0.00	0.00
FIML	1000	0.896	0.110	0.121	0.111	0.00	0.00
LIML	1000	0.800	0.098	0.121	0.097	0.00	0.00
FIML	2000	0.894	0.078	0.085	0.078	0.00	0.00
LIML	2000	0.799	0.070	0.085	0.068	0.00	0.00

Introducing Heteroskedasticity

What are the consequences of heteroskedasticity?

- If there is heteroskedasiticty in outcome equation – FIML is inconsistent, LIML can be consisent
- If MCC does not hold – both FIML and LIML are inconsistent

Can we test for this?

Introducing Heteroskedasticity

What are the consequences of heteroskedasticity?

- If there is heteroskedasiticty in outcome equation – FIML is inconsistent, LIML can be consisent
- If MCC does not hold – both FIML and LIML are inconsistent

Can we test for this? **Yes!**

- Testing for heteroskedasiticty in outcome equation – (demeaned) Breusch and Pagan (1979) test and Hausman (1978) test
- Testing for MCC – MCC test (using `gtsheckman` command)

Testing for Heteroskedasticity

Testing for heteroskedasticity in outcome equation

- Without sample selection – Breusch and Pagan (1979) test

$$y_i = \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} \tag{1}$$

homoskedasticity implies $\mathrm{E}(u_{1i}^2 \mid \mathbf{x}_{1i}) = \sigma_1^2$

1. Regress y_i on $\mathbf{x}_{1i}$, obtain residuals squared, $\widehat{u}_{1i}^2$.
2. Regress $\widehat{u}_{1i}^2$ on $\mathbf{x}_{1i}$ evaluate overall test of significance

Testing for Heteroskedasticity

Testing for heteroskedasticity in outcome equation

- With sample selection – Naive Breusch and Pagan (1979) test

$$
\begin{aligned}
y_i &= \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} && (1)\\
s_i &= 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) && (2)
\end{aligned}
$$

1. Estimate the sample selection model using LIML, obtain residuals squared for the observed sample,
$$\widehat{u}_{1i}^2 = (y_i - \mathbf{x}_{1i}\widehat{\boldsymbol{\beta}})^2$$
2. Regress $\widehat{u}_{1i}^2$ on $\mathbf{x}_{1i}$ on the observed sample and evaluate overall test of significance

Testing for Heteroskedasticity

Naive Breusch Pagan test

```
. quietly heckman lwage educ, select(inlf = educ nwifeinc) twostep

. gen uhatsq = (lwage - (_b[lwage:_cons]+_b[lwage:educ]*educ))^2 if inlf==1
(325 missing values generated)

. reg uhatsq educ
```

Source	SS	df	MS
Model	.006770959	1	.006770959
Residual	447.077377	426	1.04947741
Total	447.084148	427	1.04703548

Number of obs	=	428
F(1, 426)	=	0.01
Prob > F	=	0.9360
R-squared	=	0.0000
Adj R-squared	=	-0.0023
Root MSE	=	1.0244

uhatsq	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
educ	.0017424	.0216928	0.08	0.936	-.0408958	.0443806
_cons	.4804915	.2790351	1.72	0.086	-.0679654	1.028948

Testing for Heteroskedasticity

Testing for heteroskedasticity in outcome equation

- With sample selection – Naive Breusch and Pagan (1979) test

$$
\begin{aligned}
y_i =& \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} & (1)\\
s_i =& 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) & (2)
\end{aligned}
$$

1. Estimate the sample selection model using LIML, obtain residuals squared for the observed sample,

$$\widehat{u}_{1i}^2 = (y_i - \mathbf{x}_{1i}\widehat{\boldsymbol{\beta}})^2$$

2. Regress $\widehat{u}_{1i}^2$ on $\mathbf{x}_{1i}$ on the observed sample and evaluate overall test of significance

- **But this is not a valid test!**

Testing for Heteroskedasticity

Testing for heteroskedasticity in outcome equation

- With sample selection – Naive Breusch and Pagan (1979) test

$$
\begin{aligned}
y_i &= \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} && (1)\\
s_i &= 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) && (2)
\end{aligned}
$$

1. Estimate the sample selection model using LIML, obtain residuals squared for the observed sample,

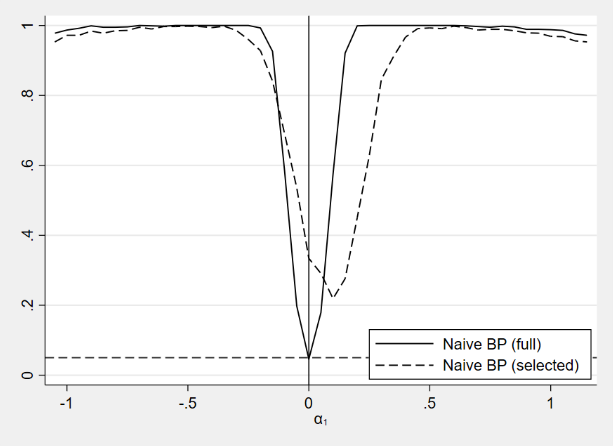
$$\widehat{u}_{1i}^2 = (y_i - \mathbf{x}_{1i}\widehat{\boldsymbol{\beta}})^2$$

2. Regress $\widehat{u}_{1i}^2$ on $\mathbf{x}_{1i}$ on the observed sample and evaluate overall test of significance

- **But this is not a valid test!**
- Even with homoskedasticity in u_{1i}, conditioning on the selected sample looks like heteroskedasticity

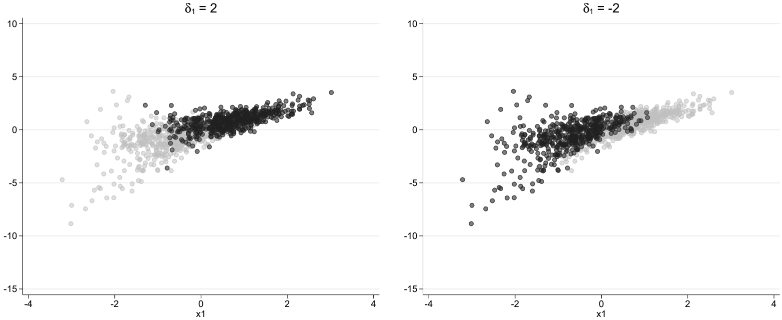
$$\mathrm{E}(u_{1i}^2 \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}, s_i = 1) = \sigma_1^2 - \gamma^2 \lambda(\mathbf{x}_{2i}\boldsymbol{\delta})\mathbf{x}_{2i}\boldsymbol{\delta}$$

Testing for Heteroskedasticity



Testing for Heteroskedasticity

Why was it asymmetric?



Depends on how the selection relates to the heteroskedasticity

Testing for Heteroskedasticity

How should we test for heteroskedasticity in outcome equation with sample selection?

Testing for Heteroskedasticity

How should we test for heteroskedasticity in outcome equation with sample selection?

- Demeaned Breusch and Pagan (1979) test
  - With homoskedasticity in u_{1i},

$$\mathrm{E}(u_{1i}^2 \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}, s_i = 1) = \sigma_1^2 - \gamma^2 \lambda(\mathbf{x}_{2i}\boldsymbol{\delta})\mathbf{x}_{2i}\boldsymbol{\delta}$$

instead we can demean it!

$$\mathrm{E}(\underbrace{u_{1i}^2 + \gamma^2 \lambda(\mathbf{x}_{2i}\boldsymbol{\delta})\mathbf{x}_{2i}\boldsymbol{\delta}}_{\tilde{u}_{1i}^2} \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}, s_i = 1) = \sigma_1^2$$

Testing for Heteroskedasticity

How should we test for heteroskedasticity in outcome equation with sample selection?

- Demeaned Breusch and Pagan (1979) test
  - With homoskedasticity in u_{1i},

$$\mathrm{E}(u_{1i}^2 \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}, s_i = 1) = \sigma_1^2 - \gamma^2 \lambda(\mathbf{x}_{2i}\boldsymbol{\delta})\mathbf{x}_{2i}\boldsymbol{\delta}$$

instead we can demean it!

$$\mathrm{E}(\underbrace{u_{1i}^2 + \gamma^2 \lambda(\mathbf{x}_{2i}\boldsymbol{\delta})\mathbf{x}_{2i}\boldsymbol{\delta}}_{\tilde{u}_{1i}^2} \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}, s_i = 1) = \sigma_1^2$$

Execute in the following steps

1. Estimate the sample selection model using LIML, obtain demeaned residuals squared for the observed sample,

$$\tilde{u}_{1i}^2 = (y_i - \mathbf{x}_{1i}\widehat{\boldsymbol{\beta}})^2 + \widehat{\gamma}^2 \widehat{\lambda}_i \mathbf{x}_{2i}\widehat{\boldsymbol{\delta}}$$

2. Regress $\tilde{u}_{1i}^2$ on $\mathbf{x}_{1i}$ on the observed sample and evaluate overall test of significance

Testing for Heteroskedasticity

Demeaned Breusch Pagan test

```
. quietly heckman lwage educ, select(inlf = educ nwifeinc) twostep mills(lambdah
> at)

. gen uhatsq_dm = uhatsq +_b[/mills:lambda]^2*lambdahat*(_b[inlf:_cons]+_b[inlf:
> educ]*educ +_b[inlf:nwifeinc]*nwifeinc)
(325 missing values generated)

. reg uhatsq_dm educ

      Source |       SS           df       MS      Number of obs   =       428
-------------+----------------------------------   F(1, 426)       =      0.11
       Model |  .113992521         1  .113992521   Prob > F        =    0.7419
    Residual |  447.307608       426  1.05001786   R-squared       =    0.0003
-------------+----------------------------------   Adj R-squared   =   -0.0021
       Total |  447.421601       427  1.04782576   Root MSE        =    1.0247

------------------------------------------------------------------------------
   uhatsq_dm | Coefficient  Std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
        educ |   .0071494   .0216984     0.33   0.742    -.0354998    .0497985
       _cons |   .4218472   .2791069     1.51   0.131     -.126751    .9704453
------------------------------------------------------------------------------
```

Testing for Heteroskedasticity

How should we test for heteroskedasticity in outcome equation with sample selection?

- Hausman (1978) test
  - With homoskedasticity both FIML and LIML are consistent, FIML is efficient
  - Without homoskedasticity (with MCC), only LIML is consistent
  - In Stata,

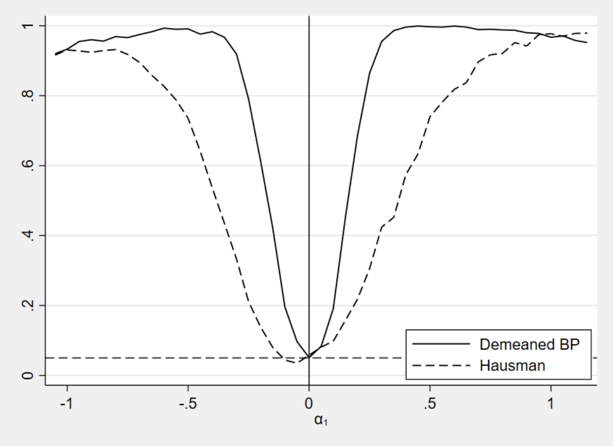
    ```
    hausman FIML LIML
    ```

Testing for Heteroskedasticity

Hausman test

```
. quietly heckman lwage educ, select(inlf = educ nwifeinc) twostep

. estimates store LIML

. quietly heckman lwage educ, select(inlf = educ nwifeinc)

. estimates store FIML

. hausman LIML FIML

                 ---- Coefficients ----
             |      (b)          (B)            (b-B)     sqrt(diag(V_b-V_B))
             |      LIML         FIML        Difference       Std. err.
-------------+----------------------------------------------------------------
        educ |    .1282506     .1176578        .0105929        .0141937
------------------------------------------------------------------------------
                          b = Consistent under H0 and Ha; obtained from heckman.
           B = Inconsistent under Ha, efficient under H0; obtained from heckman.

Test of H0: Difference in coefficients not systematic

    chi2(1) = (b-B)'[(V_b-V_B)^(-1)](b-B)
            =   0.56
Prob > chi2 = 0.4555
```

Testing for Heteroskedasticity



Estimation with Heteroskedasticity

Consistency of LIML is fundamentally reliant on MCC

- Can we test for this?
- Can we get a consistent estimator without MCC?

Estimation with Heteroskedasticity

$$y_i = \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} \tag{1}$$
$$s_i = 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) \tag{2}$$

Suppose we have heteroskedasticity

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \rho_i\sigma_{1i}\sigma_{2i} \\ \rho_i\sigma_{1i}\sigma_{2i} & \sigma_{2i}^2 \end{pmatrix} \right) \tag{12}$$

Can we still derive a correction?

Estimation with Heteroskedasticity

$$\begin{aligned} y_i =& \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} \quad (1) \\ s_i =& 1(\mathbf{x}_{2i}\boldsymbol{\delta} + u_{2i} > 0) \quad (2) \end{aligned}$$

Suppose we have heteroskedasticity

$$\begin{pmatrix} u_{1i} \\ u_{2i} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \rho_i\sigma_{1i}\sigma_{2i} \\ \rho_i\sigma_{1i}\sigma_{2i} & \sigma_{2i}^2 \end{pmatrix}\right) \quad (12)$$

Can we still derive a correction? **Yes!**

$$\lambda_i \equiv \frac{\phi(\mathbf{x}_{2i}\boldsymbol{\delta}/\sigma_{2i})}{\sigma_{2i}\Phi(\mathbf{x}_{2i}\boldsymbol{\delta}/\sigma_{2i})}$$

$$\gamma_i \equiv \rho_i\sigma_{1i}\sigma_{2i}$$

then

$$\mathrm{E}(y_i \mid s_i = 1, \mathbf{x}_{1i}, \mathbf{x}_{2i}) = \mathbf{x}_{1i}\boldsymbol{\beta} + \gamma_i\lambda_i \quad (13)$$

Estimation with Heteroskedasticity

$$\begin{aligned} y_i =& \mathbf{x}_{1i}\boldsymbol{\beta} + u_{1i} && (1) \\ s_i =& 1(\mathbf{x}_{2i}\boldsymbol{\gamma} + u_{2i} > 0) && (2) \end{aligned}$$

Estimation depends on modeling σ_{2i} and γ_i

$$\lambda_i = \frac{\phi(\mathbf{x}_{2i}\boldsymbol{\delta}/\sigma_{2i})}{\sigma_{2i}\Phi(\mathbf{x}_{2i}\boldsymbol{\delta}/\sigma_{2i})}$$

$$\mathrm{E}(y_i \mid s_i = 1, \mathbf{x}_{1i}, \mathbf{x}_{2i}) = \mathbf{x}_{1i}\boldsymbol{\beta} + \gamma_i\lambda_i \tag{13}$$

Consider parametric models for the heteroskedasticity:

$$\sigma_{2i}^2 = \mathrm{Var}(u_{2i} \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}) = \{\exp(\mathbf{z}_{2i}\boldsymbol{\pi})\}^2 \tag{14}$$

$$\gamma_i = \mathrm{Cov}(u_{1i}, u_{2i} \mid \mathbf{x}_{1i}, \mathbf{x}_{2i}) = \mathbf{z}_{12i}\boldsymbol{\rho} \tag{15}$$

generalized two-step Heckman Estimator

What to include in $\mathbf{z}_{2i}$ and $\mathbf{z}_{12i}$?

$\mathbf{z}_{2i}$ are the covariates in the conditional variance of the binary sample selection equation

- never includes a constant (binary response only identified to scale)
- variables with a heterogeneous effect on sample selection

$$\begin{aligned}\mathrm{Var}(\tilde{u}_{2i} \mid educ_i, nwifinc_i) =& 1 + \sigma^2_{d1} educ_i^2 + \sigma^2_{d2} nwifinc_i^2 \\ & + \sigma^2_{d1d2} educ_i \times nwifinc_i\end{aligned}$$

generalized two-step Heckman Estimator

What to include in $\mathbf{z}_{2i}$ and $\mathbf{z}_{12i}$?

$\mathbf{z}_{12i}$ are the covariates in the conditional covariance across the outcome and sample selection equations

- it always includes a constant (first element)
- variables whose heterogeneous effects could be correlated across equations

$$\mathrm{Cov}(\tilde{u}_{1i}, \tilde{u}_{2i} \mid educ_i, nwifinc_i) = \rho\sigma + \sigma_{b1,d1} educ_i^2 + \sigma_{b1,d2} educ_i \times nwifinc_i$$

generalized two-step Heckman Estimator

generalized two-step Heckman Estimator

1. Estimate the binary choice in equation (2) with exponential heteroskedasticity in equation (14) via a MLE approach using **`hetprobit`**, calculate the scaled estimated inverse mills ratio:

$$\widehat{\lambda}_i = \frac{\phi(\mathbf{x}_{2i}\widehat{\boldsymbol{\delta}}/\exp(\mathbf{z}_{2i}\widehat{\boldsymbol{\pi}}))}{\Phi(\mathbf{x}_{2i}\widehat{\boldsymbol{\delta}}/\exp(\mathbf{z}_{2i}\widehat{\boldsymbol{\pi}}))\exp(\mathbf{z}_{2i}\widehat{\boldsymbol{\pi}})}.$$

2. Estimate the following augmented regression

$$y_i = \mathbf{x}_{1i}\boldsymbol{\beta} + \widehat{\lambda}_i\mathbf{z}_{12i}\boldsymbol{\rho} + \varepsilon_i. \tag{16}$$

Stata command:

`gtsheckman` *depvar* [*indepvars*] , `select(`*depvar*$_s$ `=` *varlist*$_s$`)`
[`het(`*varlist*$_1$`)` `clp(`*varlist*$_2$`)` `vce(`*vcetype*`)`]

generalized two-step Heckman Estimator

```
. gtsheckman lwage educ, select(inlf = educ nwifeinc) het(educ nwifeinc c.educ#c
> .nwifeinc) clp(c.educ#c.(educ nwifeinc)) vce(robust) nolog

Generalized Two Step Heckman Estimator          Number of obs =        753
                                                    Selected =        428
                                                 Nonselected =        325
First-stage heteroskedastic probit estimates
```

inlf	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
inlf						
educ	.1064374	.1287153	0.83	0.408	-.14584	.3587147
nwifeinc	-.0196065	.0230412	-0.85	0.395	-.0647664	.0255534
_cons	-.820798	1.024283	-0.80	0.423	-2.828355	1.186759
lnsigma						
educ	-.0543857	.0859826	-0.63	0.527	-.2229085	.1141371
nwifeinc	.0208967	.056284	0.37	0.710	-.089418	.1312114
c.educ#c.nwifeinc	.0000222	.0039891	0.01	0.996	-.0077964	.0078407

generalized two-step Heckman Estimator

Second-stage augmented regression estimates

	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
lwage						
educ	.2219466	.142621	1.56	0.120	-.0575855	.5014787
lambda	1.019362	1.11739	0.91	0.362	-1.170681	3.209405
c.lambda#c.educ#c.educ	-.0063103	.0054442	-1.16	0.246	-.0169808	.0043603
c.lambda#c.educ#c.nwifeinc	.0004444	.0007381	0.60	0.547	-.0010023	.001891
_cons	-1.7458	2.280336	-0.77	0.444	-6.215176	2.723576

Testing for MCC

Consistency of LIML is fundamentally reliant on MCC

$$\begin{aligned} u_{2i} \mid \mathbf{x}_{1i}, \mathbf{x}_{2i} &\sim \mathrm{N}(0,1) \\ \mathrm{E}(u_{1i} \mid u_{2i}, \mathbf{x}_{1i}, \mathbf{x}_{2i}) &= \gamma u_{2i} \end{aligned} \tag{4}$$

Can we test for this?

Testing for MCC

Consistency of LIML is fundamentally reliant on MCC

$$\begin{aligned} u_{2i} \mid \mathbf{x}_{1i}, \mathbf{x}_{2i} &\sim \mathrm{N}(0,1) \\ \mathrm{E}(u_{1i} \mid u_{2i}, \mathbf{x}_{1i}, \mathbf{x}_{2i}) &= \gamma u_{2i} \end{aligned} \tag{4}$$

Can we test for this?

the `gtsheckman` command does not rely on MCC

$$\begin{aligned} u_{2i} \mid \mathbf{x}_{1i}, \mathbf{x}_{2i} &\sim \mathrm{N}(0, \{\exp(\mathbf{z}_{2i}\boldsymbol{\pi})\}^2) \\ \mathrm{E}(u_{1i} \mid u_{2i}, \mathbf{x}_{1i}, \mathbf{x}_{2i}) &= \frac{\mathbf{z}_{12i}\boldsymbol{\rho}}{\{\exp(\mathbf{z}_{2i}\boldsymbol{\pi})\}^2} u_{2i} \end{aligned} \tag{17}$$

but MCC holds if

$$\boldsymbol{\pi} = 0 \text{ and } \mathbf{z}_{12i}\boldsymbol{\rho} = \gamma$$

in other words, test of all the heteroskedasticity terms and all covariance terms (not including the constant)

Testing for MCC

```
. hetprobit inlf educ nwifeinc, het(educ nwifeinc c.educ#c.nwifeinc) nolog

Heteroskedastic probit model                    Number of obs     =        753
                                                Zero outcomes     =        325
                                                Nonzero outcomes  =        428

                                                Wald chi2(2)      =       0.73
Log likelihood = -486.2947                      Prob > chi2       =     0.6959
```

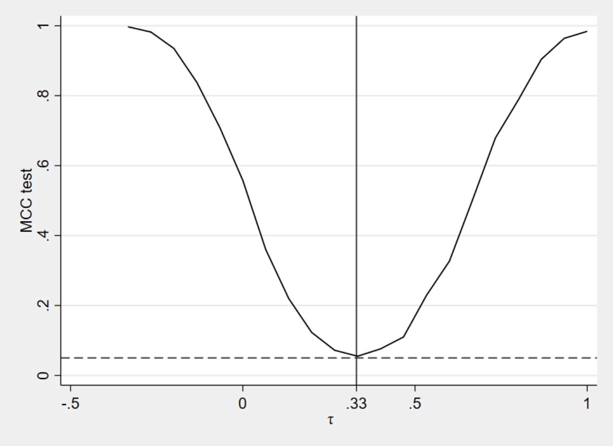
inlf	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
inlf						
educ	.1064374	.1287153	0.83	0.408	-.14584	.3587147
nwifeinc	-.0196065	.0230412	-0.85	0.395	-.0647664	.0255534
_cons	-.820798	1.024283	-0.80	0.423	-2.828355	1.186759
lnsigma						
educ	-.0543857	.0859826	-0.63	0.527	-.2229085	.1141371
nwifeinc	.0208967	.056284	0.37	0.710	-.089418	.1312114
c.educ#c.nwifeinc	.0000222	.0039891	0.01	0.996	-.0077964	.0078407

```
LR test of lnsigma=0: chi2(3) = 5.00                      Prob > chi2 = 0.1721
```

Testing for MCC

```
. quietly gtsheckman lwage educ, select(inlf = educ nwifeinc) het(educ nwifeinc
> c.educ#c.nwifeinc) clp(c.educ#c.(educ nwifeinc)) vce(robust) nolog

. test c.lambda#c.educ#c.educ c.lambda#c.educ#c.nwifeinc

 ( 1)  [lwage]c.lambda#c.educ#c.educ = 0
 ( 2)  [lwage]c.lambda#c.educ#c.nwifeinc = 0

           chi2(  2) =    1.34
         Prob > chi2 =    0.5106
```

Testing for MCC



Conclusion

1. What causes heteroskedasticity in Sample Selection models?
   - heterogeneity!
2. What are the consequences of heteroskedasticity in Sample Selection models?
   - LIML vs FIML estimators
   - heteroskedasticity in outcome vs selection equation
3. Can we test for heteroskedasticity?
   - LIML over FIML – (demeaned) Breusch and Pagan (1979) test and Hausman (1978) test
   - Validity of LIML – MCC test
4. Is there an alternative estimator for sample selection models with general forms of heteroskedasticity.
   - `gtsheckman`

References I

Breusch, T. S., and A. R. Pagan (1979): "A simple test for heteroscedasticity and random coefficient variation," *Econometrica*, 47(5), 1287–1294.

Carlson, A. (2022): "GTSHECKMAN: Stata module to compute a generalized two-step Heckman selection model," Statistical Software Components, Boston College Department of Economics, revised 06 Aug 2024.

——— (forthcoming): "gtsheckman: Generalized two-step Heckman estimator," *The Stata Journal*.

Carlson, A., and R. Joshi (2024): "Sample selection in linear panel data models with heterogeneous coefficients," *Journal of Applied Econometrics*, 39(2), 237–255.

Carlson, A., and W. Zhao (2023): "Heckman sample selection estimators under heteroskedasticity," Working Papers 2303, Department of Economics, University of Missouri.

Hausman, J. A. (1978): "Specification tests in econometrics," *Econometrica*, 46(6), 1251–1271.

Heckman, J. J. (1979): "Sample selection bias as a specification error," *Econometrica*, 47(1), 153–161.

Mroz, T. A. (1987): "The sensitivity of an empirical model of married women's hours of work to economic and statistical assumptions," *Econometrica*, 55(4), 765–799.

Wooldridge, J. M. (2010): *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge, MA, 2nd edn.