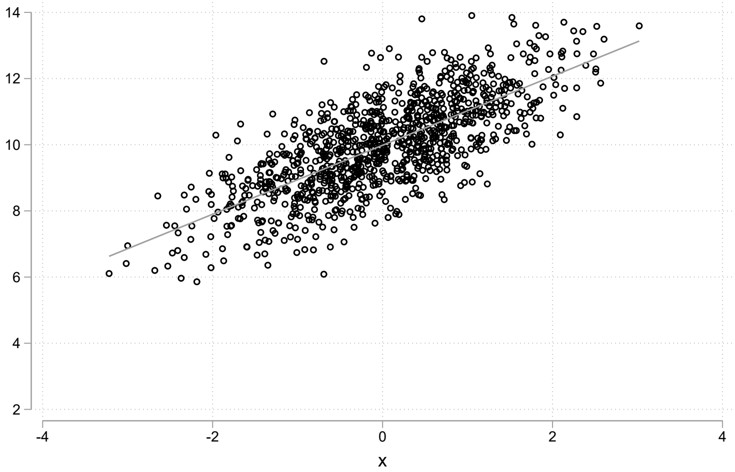
# `gtsheckman`: Generalized two-step Heckman Estimator

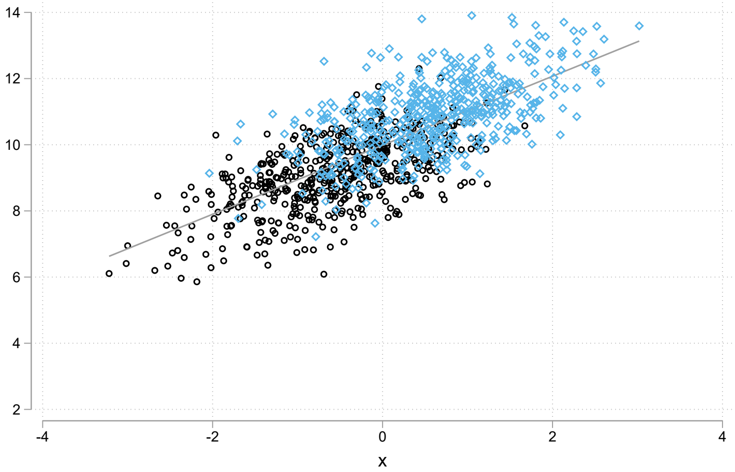
Alyssa H. Carlson
University of Missouri

August, 2022

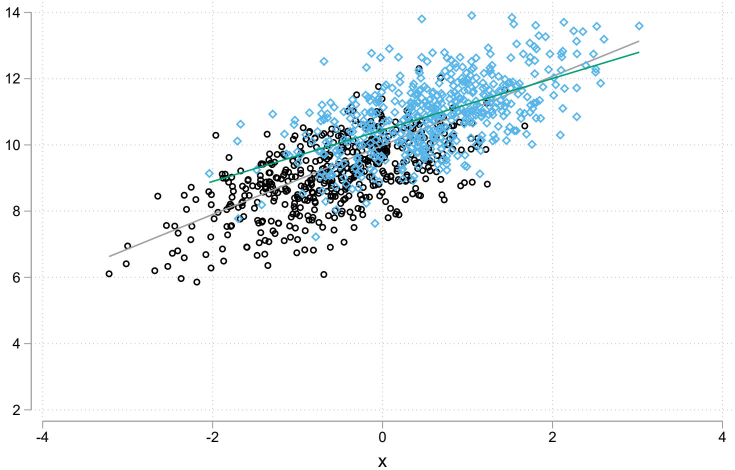
# Intro



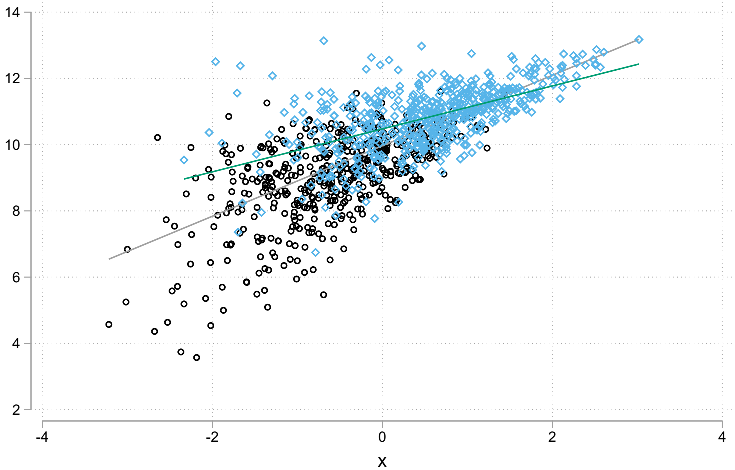
# Intro



# Intro



# Intro



# Today's talk

New command for a two-step Heckman sample selection estimator under heteroskedasticity.

Outline of talk

1. Background
   - endogenous sample selection model
   - two-step Heckman estimator
2. Introduce heteroskedasticity - generalized two-step Heckman Estimator
   - `gtsheckman`
3. Example
   - Mroz (1987)
   - `use http://fmwww.bc.edu/ec-p/data/wooldridge/mroz, clear`

# Sample Selection

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 $s_i = 1$,

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

- 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

$$\begin{aligned} \ln(wage_i) =& \beta_0 + educ_i\beta_1 + u_{1i} \\ inlf_i =& 1(\gamma_0 + educ_i\gamma_1 + nwifinc_i\gamma_2 + u_{2i} > 0) \end{aligned}$$

# Sample Selection

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

Heckman (1979) famous paper assume

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix}\right)$$

Which suggests two possible estimators:

1. Full information ML: maximum likelihood over the joint distribution of $y_i$ and $s_i$.
2. Limit information ML: two-step estimator based on the conditional distribution of $y_i|s_i = 1$

# two-step Heckman Estimator

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

Heckman (1979) famous paper assume

$$
\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim \mathrm{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma^2 & \rho\sigma \\ \rho\sigma & 1 \end{pmatrix} \right)
$$

The two-step estimator builds follows from

$$
E(u_{i1}|s_i = 1, \mathbf{x}_{1i}, \mathbf{x}_{2i}) = \rho\sigma \frac{\phi(\mathbf{x}_{2i}\boldsymbol{\gamma}/1)}{\Phi(\mathbf{x}_{2i}\boldsymbol{\gamma}/1) \times 1}
$$

and therefore

$$
E(y_i|s_{it} = 1, \mathbf{x}_{1i}, \mathbf{x}_{2i}) = \mathbf{x}_{1i}\boldsymbol{\beta} + \rho\sigma \frac{\phi(\mathbf{x}_{2i}\boldsymbol{\gamma}/1)}{\Phi(\mathbf{x}_{2i}\boldsymbol{\gamma}/1) \times 1}
$$

# two-step Heckman Estimator

two-step Heckman Estimator

1. Estimate the binary choice in equation (2) using **`probit`**, calculate the estimated inverse mills ratio: $\widehat{\lambda}_i = \phi(\mathbf{x}_{2i}\widehat{\boldsymbol{\gamma}}/1)/(\Phi(\mathbf{x}_{2i}\widehat{\boldsymbol{\gamma}}/1) \times 1)$.
2. Estimate the following augmented regression:

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

Stata command:

`heckman` *depvar* [ *indepvars* ] , `select(`$depvar_s$ `=` $varlist_s$`) twostep`

# two-step Heckman Estimator

```
. 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 |      Coef.   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
------------------------------------------------------------------------------
```

# two-step Heckman Estimator

```
. 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
```

| | Coef. | 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

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

Now allowing for heteroskedasticity

$$\begin{pmatrix} u_{i1} \\ u_{i2} \end{pmatrix} \sim \mathrm{N}\left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_{1i}^2 & \sigma_{12i} \\ \sigma_{12i} & \sigma_{2i}^2 \end{pmatrix} \right)$$

Consider parametric models for the heteroskedasticity:

$$\sigma_{2i}^2 = \{\exp(\mathbf{z}_{2i}\boldsymbol{\delta})\}^2 \tag{3}$$
$$\sigma_{12i} = \mathbf{z}_{12i}\boldsymbol{\pi} \tag{4}$$

then

$$E(y_i \mid s_i = 1, \mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{z}_{2i}, \mathbf{z}_{12i}) = \mathbf{x}_{1i}\boldsymbol{\beta} + \mathbf{z}_{12i}\boldsymbol{\pi} \frac{\phi(\mathbf{x}_{2i}\boldsymbol{\gamma}/\exp(\mathbf{z}_{2i}\boldsymbol{\delta}))}{\Phi(\mathbf{x}_{2i}\boldsymbol{\gamma}/\exp(\mathbf{z}_{2i}\boldsymbol{\delta}))\exp(\mathbf{z}_{2i}\boldsymbol{\delta})}$$

# generalized two-step Heckman Estimator

generalized two-step Heckman Estimator

1. Estimate the binary choice in equation (2) with exponential heteroskedasticity in equation (3) via a pooled MLE approach using **`hetprobit`**, calculate the scaled estimated inverse mills ratio:

$$\widehat{\lambda}_i = \frac{\phi(\mathbf{x}_{2i}\widehat{\boldsymbol{\gamma}}/\exp(\mathbf{z}_{2i}\widehat{\boldsymbol{\delta}}_2))}{\Phi(\mathbf{x}_{2i}\widehat{\boldsymbol{\gamma}}/\exp(\mathbf{z}_{2i}\widehat{\boldsymbol{\delta}}_2))\exp(\mathbf{z}_{2i}\widehat{\boldsymbol{\delta}}_2)}.$$

2. Estimate the following augmented regression

$$y_i = \mathbf{x}_{1i}\boldsymbol{\beta} + \widehat{\lambda}_i \mathbf{z}_{12i}\boldsymbol{\pi} + \varepsilon_i. \tag{5}$$

Stata command:

**`gtsheckman`** *depvar* [ *indepvars* ] , <u>`sel`</u>`ect(`*depvar*$_s$ `=` *varlist*$_s$`)`
[ `het(`*varlist*$_1$`)` `clp(`*varlist*$_2$`)` `vce(`*vcetype*`)` ]

# 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

- variables that determine the heterogeneity in variance of the latent sample selection
- variables with a heterogeneous effect on sample selection
- all of $\mathbf{x}_{2i}$ to allow for flexibility in the distributional assumption (probit)

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

- it always includes a constant
- variables that determine the heterogeneity in the endogeneity of sample selection

# generalized two-step Heckman Estimator

```
. gtsheckman lwage educ, select(inlf = educ nwifeinc) het(educ nwifeinc) clp(e
> duc) vce(robust)
```

Generalized Two Step Heckman Estimator

| | |
|---|---|
| Number of obs = | 753 |
| Selected = | 428 |
| Nonselected = | 325 |

First-stage heteroskedastic probit estimates

| inlf | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| **inlf** | | | | | | |
| educ | .1070307 | .0722545 | 1.48 | 0.139 | -.0345855 | .2486469 |
| nwifeinc | -.0197132 | .0128005 | -1.54 | 0.124 | -.0448017 | .0053753 |
| _cons | -.8254513 | .5930636 | -1.39 | 0.164 | -1.987835 | .3369321 |
| **lnsigma** | | | | | | |
| educ | -.0539838 | .0461509 | -1.17 | 0.242 | -.144438 | .0364704 |
| nwifeinc | .021201 | .0130363 | 1.63 | 0.104 | -.0043496 | .0467516 |

Second-stage augmented regression estimates

| | Coef. | Robust Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| **lwage** | | | | | | |
| educ | .1878758 | .0846033 | 2.22 | 0.026 | .0220564 | .3536953 |
| lambda | 1.28414 | 1.064629 | 1.21 | 0.228 | -.8024936 | 3.370774 |
| c.lambda#c.educ | -.0914331 | .0605547 | -1.51 | 0.131 | -.2101182 | .027252 |
| _cons | -1.326688 | 1.402169 | -0.95 | 0.344 | -4.074889 | 1.421514 |

# Conclusion

`gtsheckman`: generalized two-step Heckman sample selection estimator

- available at `https://carlsonah.mufaculty.umsystem.edu/research`
- Carlson and Joshi (2021) utilizes the `gtsheckman` estimator for panel data with heterogeneous coefficients and sample selection

# References I

Carlson, A., and R. Joshi (2021): "Sample Selection in Linear Panel Data Models with Heterogeneous Coefficients," Working Papers 2103, Department of Economics, University of Missouri.

Heckman, J. J. (1979): "Sample selection bias as a specification error," *Econometrica: Journal of the econometric society*, pp. 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.