# Applied Quantitative Methods II

## Lecture 9: Limited dependent variables

Klára Kalíšková

# Motivation

- Many topics when *dependent* variable is a dummy variable
- For any discrete choice, dependent variable is typically a dummy variable:
    - Will a person get a loan?
    - Will a customer buy a product?
    - Will a person study college?
    - Will a woman work if she has 2+ kids?
    - Will there be re-offense in cases of domestic violence if the offender is arrested on the spot?

# Outline

- Today: models with outcome variable

$$Y_i = \begin{cases} 1 & \\ 0 & \end{cases},$$

  depending on qualitative choice (*binary models*)
- These will be:
  - Linear Probability Model (LPM)
  - Logit Model
  - Probit Model

# Outline

# Probability distribution of $Y_i$

- $Y_i$ is a discrete random variable with Bernoulli distribution:

$$Y_i = \begin{cases} 1 & \text{with probability } p_i \\ 0 & \text{with probability } 1 - p_i \end{cases}$$

  - We can find the expected value:

$$E\left[Y_i\right] = 1 \cdot p_i + 0 \cdot (1 - p_i) = p_i$$

  - and the variance:

$$\begin{aligned} Var\left[Y_i\right] &= E\left[Y_i^2\right] - \left(E\left[Y_i\right]\right)^2 = 1^2 \cdot p_i + 0^2 \cdot (1 - p_i) - p_i^2 \\ &= p_i - p_i^2 = p_i(1 - p_i) \end{aligned}$$

# Linear Probability Model

- Running the usual OLS on dummy dependent variable:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

  - Why we call it the "linear probability" model?
  - Let us take the expected value:

$$\begin{aligned} E\left[Y_i\right] &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + E\left[\varepsilon_i\right] \\ p_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} \end{aligned}$$

  - Hence, $p_i = Prob\,(Y_i = 1)$ is a linear function of explanatory variables

# Example

- Angrist, J. (2006) *Instrumental Variables Methods in Experimental Criminological Research: What, Why, and How?*
- Estimate determinants of re-offense status $y$ for cases of domestic violence ($y$ is dummy indicating cases when re-offense occurred)
- Main explanatory variable:

$$d\_coddled = \begin{cases} 1 & \text{if the offender was } \textbf{not} \text{ arrested} \\ 0 & \text{if the offender was arrested} \end{cases}$$

- Other controls:
    - race dummies
    - dummies indicating the presence of weapons and drugs

# Example

- OLS (robust SE)

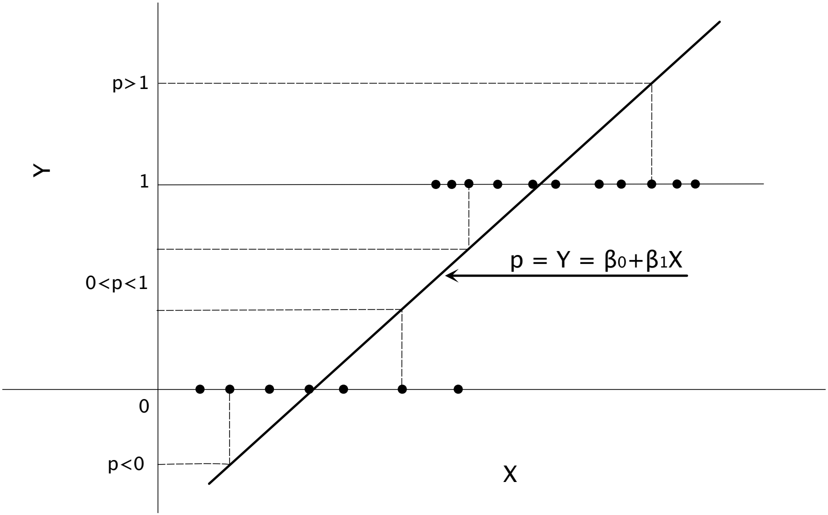
```
Linear regression                               Number of obs =        330
                                                F(  5,    324) =       1.54
                                                Prob > F       =     0.1763
                                                R-squared      =     0.0239
                                                Root MSE       =     .38457
```

| y | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| d_coddled | .0873254 | .0410044 | 2.13 | 0.034 | .0066569 | .1679938 |
| drugs | .0479707 | .0437274 | 1.10 | 0.273 | -.0380548 | .1339962 |
| weapon | .0113562 | .0480876 | 0.24 | 0.813 | -.0832472 | .1059597 |
| nonwhite | -.0274346 | .0425991 | -0.64 | 0.520 | -.1112405 | .0563712 |
| mixed | .07402 | .051851 | 1.43 | 0.154 | -.0279871 | .1760271 |
| _cons | .0901995 | .0511667 | 1.76 | 0.079 | -.0104615 | .1908604 |

# Problems with LPM

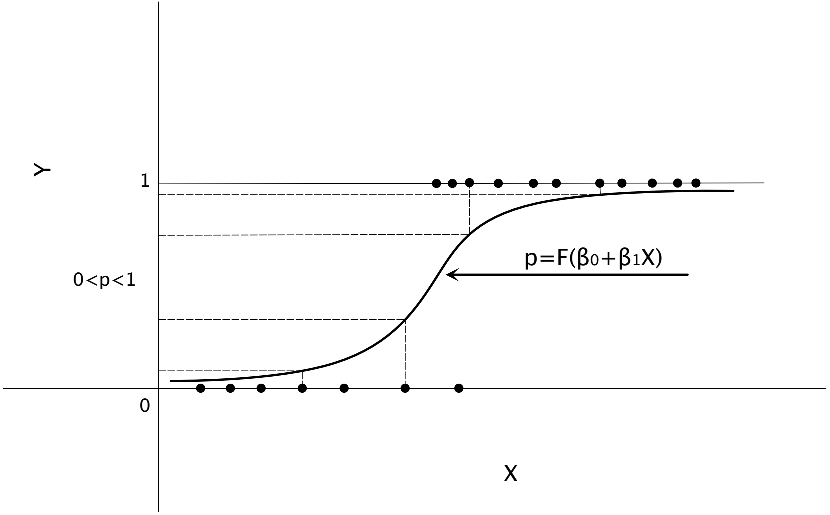
1. Error term not normally distributed:
   - because $Y_i$ has only two values, error term
   $$\varepsilon_i = Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})$$
   also binomial
2. Error term is inherently heteroskedastic:
   - we have
   $$Var\left[\varepsilon_i\right] = Var[Y_i - (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik})] = ... = Var\left[Y_i\right] = p_i(1-$$
   where $p_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik}$ so variance is a function of x's, not constant
   - we can find estimator with higher efficiency (e.g. WLS)
3. The probability is not bounded by 0 and 1:
$$\widehat{p}_i = \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \widehat{\beta}_2 x_{i2} + \ldots + \widehat{\beta}_k x_{ik}$$

# Problems with LPM



# We would like something like this:

Figure:



# We would like something like this:

- We would like to transform LPM

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i$$

- to a function

$$y_i = F(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i)$$

- such that

$$F = \begin{cases} 0 & \text{for } -\infty \\ 1 & \text{for } +\infty \end{cases}$$

# Possible F's

- Standard normal:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} \mathrm{d}t$$

- Logistic:

$$f(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

$$F(x) = \frac{1}{1+\exp(-x)}$$

# Latent variable approach

- Suppose we have a continuous variable $y_i^*$ (called *latent variable*), following:

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \tag{1}$$

and the relationship

$$Y_i = \begin{cases} 1 & \text{for } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

- Equations (1) and (2) together define the binary model
- Underlying heuristic: the value of the qualitative dependent variable depends on a choice based on a latent (unobserved) continuous utility and a simple decision rule
- Leads to derivation of Logit and Probit models

# Latent variable approach

- Let us express the probability that $Y_i = 1$ under this approach:

$$
\begin{aligned}
p_i &= Prob(Y_i = 1) = Prob(y_i^* > 0) \\
&= Prob(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \varepsilon_i > 0) \\
&= Prob(\varepsilon_i > -\beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) \\
&= 1 - Prob(\varepsilon_i \leq -\beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) \\
&= 1 - F(-\beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) \quad ,
\end{aligned}
$$

where $F(.)$ denotes the cumulative distribution function (cdf) of the error term $\varepsilon_i$

# Possible distributions of the error term

- Standard normal:

$$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{x^2}{2}\right\}$$

$$F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{t^2}{2}\right\} \mathrm{d}t$$

- Logistic:

$$f(x) = \frac{\exp(-x)}{(1+\exp(-x))^2}$$

$$F(x) = \frac{1}{1+\exp(-x)}$$

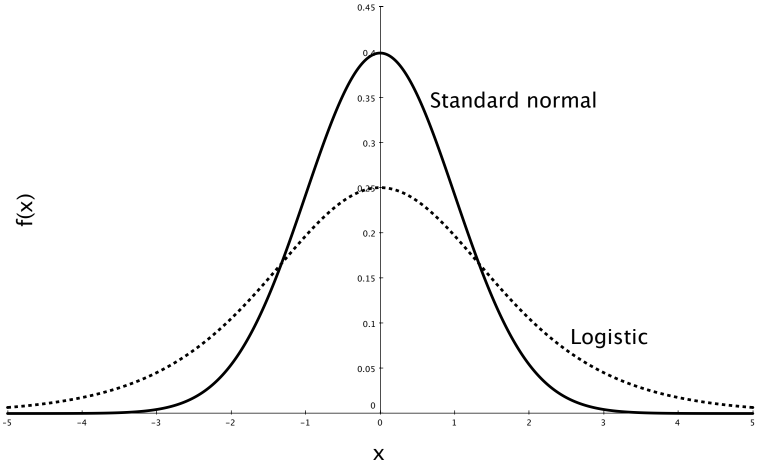
- Both distributions satisfy:

$$1 - F(-x) = F(x)$$

- This allows us to write:

$$\begin{aligned} p_i = Prob(Y_i = 1) &= 1 - F(-\beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}) \\ &= F(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}) \end{aligned}$$

# Possible distributions: pdf's



# Outline

# Probit and Logit Models

- Both models define probability of $Y_i = 1$ as a function of explanatory variables:

$$p_i = Prob(Y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik}) \quad ,$$

where $F(.)$ denotes he cdf of error term $\varepsilon_i$

- Probit model - uses standard normal cdf
- Logit model - uses the logistic cdf
- Parameters $\beta_0$, $\beta_2$, ..., $\beta_k$ are estimated by the Maximum Likelihood method

# Maximum Likelihood Estimator

- The principle of the MLE is to maximize the likelihood function $L$ as a function of the parameter which is to be estimated
- The likelihood function represents the probability of the sample as we observe it
- For binary models with $n$ observations, it looks as

$$L = \prod_{i=1}^{n} p_i^{Y_i} (1 - p_i)^{(1-Y_i)}$$

with

$$p_i = Prob(Y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})$$

# Maximum Likelihood Estimator

- The MLE estimates of $\beta_0$, $\beta_2$, ..., $\beta_k$ are such that they maximize the logarithm of the likelihood function

$$\ln L = \sum_{i=1}^{n} Y_i \ln p_i + (1 - Y_i) \ln(1 - p_i)$$

with

$$p_i = Prob(Y_i = 1) = F(\beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik})$$

- The choice of $F(.)$ depends on whether we use Probit or Logit model
- Testing multiple hypothesis - Wald or LR test
- Both models are **consistent** and **efficient** under the **condition** that the **choice of** $F(x)$ **is correct** (very limiting!)

# Comparison of the models

- In the LPM model, we had

$$\widehat{p}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik} \quad ,$$

  which was not bounded by 0 and 1

- In the Logit an Probit models, we have

$$\widehat{p}_i = F(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik}) \quad ,$$

  which is bounded by 0 and 1 thanks to the properties of a cumulative distribution function

# Interpretation

- In the LPM model, we had

$$\widehat{p}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik} \quad ,$$

which gave a simple interpretation of the coefficients:

$$\frac{\partial \widehat{p}_i}{\partial x_{ij}} = \widehat{\beta}_j$$

- In the Logit an Probit models, we have

$$\widehat{p}_i = F(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik}) \quad ,$$

which gives:

$$\frac{\partial \widehat{p}_i}{\partial x_{ij}} = f(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik}) \cdot \widehat{\beta}_j$$

# Logit and probit: marginal effects

- Logit and probit: more than in coefficients $\widehat{\beta}_j$, we are interested in marginal effects of the explanatory variables on the probability of $Y_i = 1$ :
$$\frac{\partial \widehat{p}_i}{\partial x_{ij}} = f(\widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \ldots + \widehat{\beta}_k x_{ik}) \cdot \widehat{\beta}_j$$
- In order to obtain an average marginal effect (impact of $xj$ on the probability of $Y_i = 1$), the function $f(.)$ in this expression is usually evaluated at the mean of observations:
$$\frac{\partial \widehat{p}}{\partial x_j} = f(\widehat{\beta}_0 + \widehat{\beta}_1 \overline{x}_1 + \ldots + \widehat{\beta}_k \overline{x}_k) \cdot \widehat{\beta}_j$$

# Back to the example

## Logit

```
Logistic regression                               Number of obs   =        330
                                                  LR chi2(5)      =       7.97
                                                  Prob > chi2     =     0.1580
Log likelihood = -152.48188                       Pseudo R2       =     0.0255
```

| y | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| d_coddled | .6235318 | .3151227 | 1.98 | 0.048 | .0059026 | 1.241161 |
| drugs | .3339199 | .3070374 | 1.09 | 0.277 | -.2678624 | .9357022 |
| weapon | .0745484 | .3323755 | 0.22 | 0.823 | -.5768956 | .7259925 |
| nonwhite | -.194676 | .3013182 | -0.65 | 0.518 | -.7852489 | .3958969 |
| mixed | .4732988 | .3159317 | 1.50 | 0.134 | -.1459159 | 1.092513 |
| _cons | -2.189955 | .3998198 | -5.48 | 0.000 | -2.973588 | -1.406323 |

# Back to the example

## Marginal effects after Logit:

```
Marginal effects after logit
      y  = Pr(y) (predict)
         =  .17410803
```

| variable | dy/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. | ] | X |
|---|---|---|---|---|---|---|---|
| d_codd~d* | .0867072 | .04168 | 2.08 | 0.037 | .005016 | .168399 | .587879 |
| drugs* | .0468831 | .0419 | 1.12 | 0.263 | -.035245 | .129011 | .612121 |
| weapon* | .0108449 | .0489 | 0.22 | 0.825 | -.085007 | .106697 | .260606 |
| nonwhite* | -.0277203 | .04241 | -0.65 | 0.513 | -.110851 | .05541 | .421212 |
| mixed* | .0731195 | .05185 | 1.41 | 0.158 | -.028497 | .174736 | .263636 |

(*) dy/dx is for discrete change of dummy variable from 0 to 1

```
Probit regression                               Number of obs   =        330
                                                LR chi2(5)      =       7.84
                                                Prob > chi2     =     0.1653
Log likelihood = -152.54647                     Pseudo R2       =     0.0251
```

| y | Coef. | Std. Err. | z | P>\|z\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| d_coddled | .3494254 | .1749548 | 2.00 | 0.046 | .0065202 | .6923306 |
| drugs | .1839146 | .1719692 | 1.07 | 0.285 | -.1531388 | .520968 |
| weapon | .044509 | .1885126 | 0.24 | 0.813 | -.324969 | .4139869 |
| nonwhite | -.1106221 | .1687968 | -0.66 | 0.512 | -.4414577 | .2202135 |
| mixed | .2563258 | .1826897 | 1.40 | 0.161 | -.1017394 | .614391 |
| _cons | -1.281978 | .2182593 | -5.87 | 0.000 | -1.709759 | -.8541981 |

# Back to the example

## Marginal effects after probit:

```
Marginal effects after probit
      y  = Pr(y) (predict)
         =  .17582338
```

| variable | dy/dx | Std. Err. | z | P>\|z\| | [ 95% C.I. ] | | X |
|---|---|---|---|---|---|---|---|
| d_codd~d* | .0877187 | .04232 | 2.07 | 0.038 | .004782 | .170655 | .587879 |
| drugs* | .0466293 | .04268 | 1.09 | 0.275 | -.037014 | .130272 | .612121 |
| weapon* | .0116215 | .0497 | 0.23 | 0.815 | -.085786 | .109029 | .260606 |
| nonwhite* | -.0283666 | .04289 | -0.66 | 0.508 | -.112433 | .0557 | .421212 |
| mixed* | .0699435 | .05233 | 1.34 | 0.181 | -.03263 | .172517 | .263636 |

(*) dy/dx is for discrete change of dummy variable from 0 to 1

- LPM:

| y | Coef. | Robust Std. Err. | t | P>\|t\| | [95% Conf. | Interval] |
|---|---|---|---|---|---|---|
| d_coddled | .0873254 | .0410044 | 2.13 | 0.034 | .0066569 | .1679938 |

- Logit (marginal effect):

| variable | dy/dx | Std. Err. | z | P>\|z\| | [ 95% | C.I. ] | X |
|---|---|---|---|---|---|---|---|
| d_codd~d* | .0867072 | .04168 | 2.08 | 0.037 | .005016 | .168399 | .587879 |

- Probit (marginal effect) :

| variable | dy/dx | Std. Err. | z | P>\|z\| | [ 95% | C.I. ] | X |
|---|---|---|---|---|---|---|---|
| d_codd~d* | .0877187 | .04232 | 2.07 | 0.038 | .004782 | .170655 | .587879 |

# Outline

# Tobit estimation

- When Y is roughly continuous in positive values, but a lot of observations zero
  - corner solutions

**Example**

Charity donations - many people give $> 0$, but many give $= 0$

- Problem: with OLS we would obtain below-zero fitted values
- Can be modelled with latent-variable approach as well:

$$y_i^* = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \varepsilon_i \tag{3}$$

and the observed variable is:

$$Y_i = \begin{cases} y & \text{for } y_i^* > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

- latent var. $y$ is homoskedastic and normally distributed
- Model is estimated with MLE method

# Tobit estimation

- Interpretation of coefficient **different** than OLS
  - Often similar values as OLS - tempting
  - adjustment factors can be calculated
  - Stata: postestimation *margins*
- Limitation: Relies on normality and homoskedasticity of latent variable
- Generally, Tobit one of censored regression models
  - Censored data - due to some contraints, some Y could not be realized
    - corner solutions - no negative hours worked
  - Truncated data - due to some contraints, some Y was realized but not observed
    - we have no data on subset of population

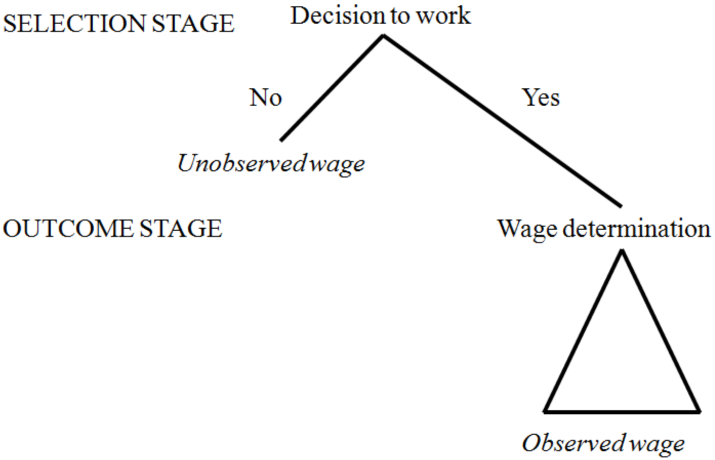
# Other estimation techniques

1. Poisson estimation
    - Count data - non-negative integers $\{0, 1, 2, ...\}$
        - E.g. Number of children born to a woman
    - OLS may again not produce a good fit
2. Multinomial logit/probit
    - when more than two categories
3. Ordered probit
    - We have increasing discrete values of dep. variable - ranking
        - Ordinal variable, e.g. survey answers on 10-point scale
4. Interval regression
    - Data not continuous, but elicited in intervals
        - e.g. income bracket
5. ... many more

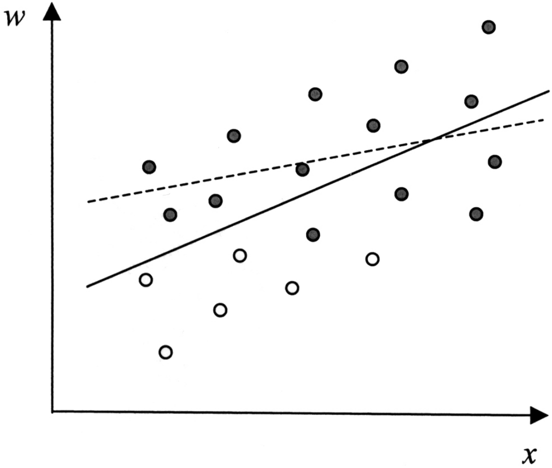
# Outline

# Intuition

- **Goal**: we want to estimate wages of women
- We observe only wages of working women (truncation)
- OK if selection into working and not working random: is it?
- Working women probably smarter, more career–oriented, more ambitious
- Bias: non-random sample selection
- Can lead to wrong conclusions and bad policies
- *Crucial: do we know, how the selection is made?*


SELECTION STAGE
Decision to work
No
Yes
Unobserved wage
OUTCOME STAGE
Wage determination
Observed wage

# Intuition



**Two-equation behavioral model**

selection equation

$$z_i = w_i'\gamma + e_i$$

outcome equation

$$y_i = x_i'\beta + u_i$$

- where $y$ is observed only when $z > 0$ (or some other threshold)
- we observe wages ($y$) only for people who work ($z > 0$)

$$E[y_i|x_i, z_i > 0] = x_i'\beta + E[u_i|z_i > 0] = x_i'\beta + E[u_i|e_i > -w_i'\gamma]$$

# Heckman's sample selection model

$$E[y_i|x_i, z_i > 0] = x_i'\beta + E[u_i|z_i > 0] = x_i'\beta + E[u_i|e_i > -w_i'\gamma]$$

- If $u_i$ and $e_i$ are independent, $E[u_i|e_i > -w_i'\beta] = 0$.
  - but unobservables in the two equations are likely to be correlated
  - e.g. ability driving both the participation decision and wages
- Instead assume that $u_i$ and $e_i$ are *jointly normal*,
  - with covariance $\sigma_{12}$ and variances $\sigma_1^2$ and $\sigma_2^2$, respectively.

$$E[y_i|x_i, z_i > 0] = x_i'\beta + \frac{\sigma_{12}}{\sigma_2}\frac{\phi(w_i'\gamma/\sigma_2)}{\Phi(w_i'\gamma/\sigma_2)} = x_i'\beta + \sigma_\lambda\lambda(w_i'\gamma)$$

# Heckman's sample selection model

$$E[y_i|x_i, z_i > 0] = x_i'\beta + \frac{\sigma_{12}}{\sigma_2}\frac{\phi(w_i'\gamma/\sigma_2)}{\Phi(w_i'\gamma/\sigma_2)} = x_i'\beta + \sigma_\lambda\lambda(w_i'\gamma)$$

, where $\frac{\phi(w_i'\gamma/\sigma_2)}{\Phi(w_i'\gamma/\sigma_2)}$ is the inverse Mills ratio (Heckman's lambda).

- **We can consistently estimate $\beta$ on the selected sample** if we include $\lambda(w_i'\gamma)$ as an additional regressor into the outcome equation.
- Source: Heckman, J. (1979). Sample selection bias as a specification error. *Econometrica*, 47, pp. 153-61.
  - *Note: Heckman got the Nobel prize for this paper.*

# Practical quidelines

1. Estimate **selection** equation using **all observations**.
   - $z_i = w_i'\gamma + e_i$
   - obtain estimates of parameters $\hat{\gamma}$
   - compute the inverse Mills ratio: $\frac{\phi(w_i'\hat{\gamma})}{\Phi(w_i'\hat{\gamma})} = \hat{\lambda}(w_i'\gamma)$
2. Estimate the **outcome** equation using **only the selected** observations.
   - $y_i = x_i'\beta + \sigma_\lambda \hat{\lambda}(w_i'\gamma) + u_i$
   - we can test selection bias by testing significance of the lambda term (standard t-test)

- Note: standard errors have to be adjusted
  - we use $\hat{\lambda}(w_i'\gamma)$ instead of $\lambda(w_i'\gamma)$ in the estimation

# Identification issues

- selection equation: $z_i = w_i'\gamma + e_i$
- outcome equation: $y_i = x_i'\beta + \sigma_\lambda \hat{\lambda}(w_i'\gamma) + u_i$
- **Can we estimate $\beta$ and $\sigma_\lambda$ if $x_i = w_i$?**
  - i.e., can we use Heckman's two–step model if the determinants of participation are the same as determinants of wages?
  - Yes, we can estimate it even if $x_i = w_i$ because $\lambda$ is a nonlinear function.
  - However, we should not rely on nonlinearity of $\lambda$ function!
    - Lambda can be very close to a linear function.
    - Thus, $\lambda(w_i'\gamma)$ might be highly correlated with $x_i$ if $x_i = w_i$.
    - Multicollinearity problem!
  - We should try to find *exclusion restriction*.

# Identification issues

- selection equation: $z_i = w_i'\gamma + e_i$
- outcome equation: $y_i = x_i'\beta + \sigma_\lambda \hat{\lambda}(w_i'\gamma) + u_i$
- Identification should be based on **exclusion restriction**.
  - Exclusion restriction is a variable that explains selection (participation), but not the outcome variable.
  - There is at least one variable which is in $w_i$, which is not in $x_i$.
  - $x_i$ should be a strict subset of $w_i$.
  - E.g.: presence of small children affects participation on the labor market, but not wages of women.

# Outline

# Eissa and Hoynes (2004)

*Taxes and the labor market participation of married couples: The earned income tax credit*

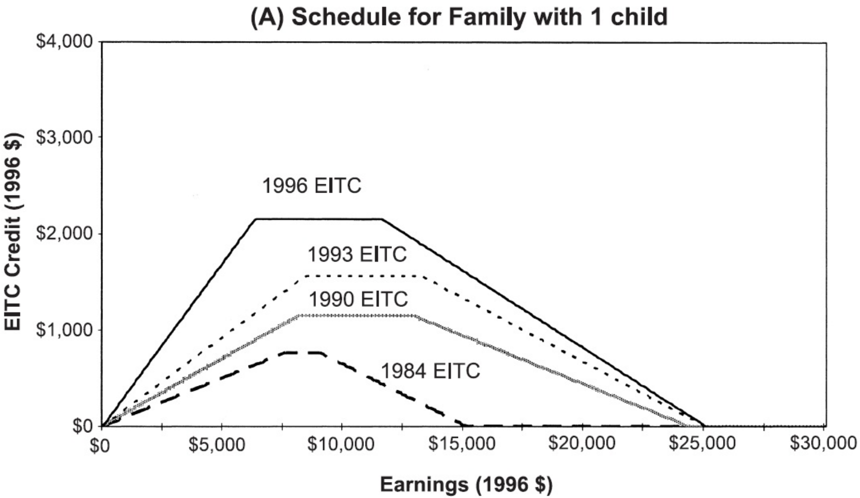
- **Goal**: Estimate the impact of EITC on female labor supply.
- Earned Income Tax Credit (EITC):
    - largest cash-transfer program (negative income tax) for working poor (low-income) families with children (20m families)
    - conditions for eligibility: some positive earnings (*work*) and total family income below certain threshold
    - Why: "promote both the values of family and work"
    - Traditional welfare programs - adverse incentives to work
    - EITC should not distort labor supply
    - Does it really work?
- Potential side-effects
    - based on family income => disincentives for the secondary earner
        - men increase but women decrease labor supply
    - EITC may thus reduce overall family labor supply of married couples

# Eissa and Hoynes (2004)

*Taxes and the labor market participation of married couples: The earned income tax credit*

- Data for 1984 to 1996
- 6.4m to 19.5m recipient families
- EITC from \$755 to \$3556
- Authors restrict sample to low-educated couples.
  - endogenous sample selection?
  - no, because education is explanatory variable
  - why not restricting the sample to low-income instead?
  - income driven by unobserved characteristics that drive participation!

# Eissa and Hoynes (2004)



# Labor supply effects of EITC

- EITC encourages work among single women.
    - Meyer and Rosenbaum (2011)
- Effect on primary earners (men or single women) is also positive.
    - Those who already work are either better off or not affected
    - Those who do not work are not affected
- BUT: the effect on secondary earners (married women) might be negative.
    - Example: Husband's income qualifies family for EITC. If wife starts working, family might not be eligible anymore (her income will shift the family income above the threshold for eligibility).

Comparison of before/after treated/control:

Table 3
EITC maximum credit and mean labor force participation rates of married couples

| | Before expansion (1989 – 1993) | After expansion (1994 – 1996) | Change | Relative (to no kids) change |
|---|---|---|---|---|
| *Panel A: maximum EITC (1989 to 1996, in 1996 dollars)* | | | | |
| 2+ Children | \$1151 | \$3556 | \$2405 | \$2082 |
| One child | \$1151 | \$2152 | \$1001 | \$678 |
| No children | \$0 | \$323 | \$323 | |
| *Panel B: married women* | | | | |
| 2+ kids ($N$=7095) | 0.533 (0.007) | 0.504 (0.010) | – 0.029 (0.012) | – 0.051 (0.022) |
| One kid ($N$=2648) | 0.642 (0.011) | 0.642 (0.017) | +0.001 (0.020) | – 0.021 (0.027) |
| No kids ($N$=3120) | 0.653 (0.010) | 0.676 (0.015) | +0.023 (0.018) | |
| *Panel C: married men* | | | | |
| 2+ kids ($N$=7095) | 0.955 (0.003) | 0.958 (0.004) | +0.003 (0.005) | +0.014 (0.010) |
| One kid ($N$=2648) | 0.968 (0.004) | 0.962 (0.007) | – 0.006 (0.008) | +0.005 (0.012) |
| No kids ($N$=3120) | 0.954 (0.005) | 0.943 (0.008) | – 0.011 (0.009) | |

Source: Authors' tabulations of March CPS for years 1990 – 1997. EITC figures are in nominal dollars. Sample includes married couples where the wife has less than 12 years of education. See text for further sample selection.

# Eissa and Hoynes (2004): Estimation approach

1. "Natural experiment" approach:
   - Using policy reforms of EITC expansion
   - Difference-in-differences method
   - Treatment group: low-educated married women with children
   - Control group: low-educated married women without children
2. They estimate participation equation as a function of net wages (after EITC):
   - Use two–step Heckman's method to predict wages for both working and non–working
   - Exclusion restriction: family characteristics (number of children, presence of young children)

# Eissa and Hoynes (2004): Estimation approach

1. Participation equation for the Heckman wage equation:

$$P_i = w_i'\gamma + v_i = z_i'\gamma_z + \gamma_1 children_i + \gamma_2 young\_child + v_i$$

2. Wage equation with Heckman's selection term:

$$wage_i = z_i'\beta + \sigma_\lambda \hat{\lambda}(w_i'\gamma) + u_i$$

3. Participation equation of interest (impact of EITC captured through changes in tax rates):

$$P_{it} = \alpha_1 other\_inc_{it} + \alpha_2 \hat{wage}_{it}(1 - ATR)_{it} + x_{it}'\rho + e_{it}$$

# Eissa and Hoynes (2004): Results

Results from diff–in–diffs estimation:

Table 4
Difference in difference estimates of labor force participation rates for married couples with and without children

| | Married women ($dp/dx$) | Married men ($dp/dx$) |
|---|---|---|
| *Panel A: unconditional means (any kids)* | | |
| Any children | – 0.047 (0.021) | 0.011 (0.010) |
| *Panel B: basic estimates (any kids)* | | |
| $\gamma$ (any children) | – 0.039 (0.021) | 0.008 (0.008) |
| Log likelihood / ($R^2$) | – 8106 | – 1967 |
| *Panel C: kids, 2+ kids unconditional means* | | |
| EITC1 (one child) | – 0.024 (0.027) | 0.005 (0.010) |
| EITC2 (2+ children) | – 0.052 (0.022) | 0.014 (0.012) |
| *Panel D: kids, 2+ kids, basic estimates* | | |
| $\gamma_g$ (any kids) | – 0.014 (0.027) | 0.003 (0.010) |
| $\gamma_{g2}$ (2+ children) | – 0.034 (0.024) | 0.006 (0.009) |
| Log likelihood / ($R^2$) | – 8105 | – 1967 |
| Mean of the dependent variable | 0.58 | 0.96 |
| Other controls (all specifications) | Demographics, state unemployment rate, state dummies, time dummies | |
| Observations | | 12,863 |

# Eissa and Hoynes (2004): Results

Results from reduced form participation equation:

Table 6
Parameter estimates for labor force participation equation for married couples with children, 1984–1996

| Variable | Married women | Married men |
|---|---|---|
| *Specification: average tax rate evaluated at full-time (40 h)* | | |
| # of children | – 0.045 (0.0065) | – 0.003 (0.001) |
| # preschool children | – 0.109 (0.006) | – 0.005 (0.001) |
| Black | 0.076 (0.017) | – 0.025 (0.007) |
| Other race | 0.014 (0.017) | – 0.048 (0.008) |
| Age | 0.045 (0.006) | 0.001 (0.002) |
| Age squared ( per 100) | – 0.067 (0.008) | – 0.001 (0.002) |
| State unemployment rate | – 0.004 (0.004) | – 0.004 (0.001) |
| Net wage, $w(1-\tau^a)$ | 0.027 (0.005) | 0.003 (0.001) |
| Net unearned income, $y^n$ | – 0.001 (0.0003) | – 0.005 (0.0003) |
| Other controls | State, time dummies, 2+ children × time interactions | |
| Pseudo $R^2$ | 0.07 | 0.18 |
| Mean of dep. variable | 0.556 | 0.960 |
| Observations | | 17,178 |
| *"Elasticity" of participation* | | |
| Wage | 0.267 | 0.032 |
| Income | – 0.039 | – 0.007 |

# Eissa and Hoynes (2004): Results

Results from reduced form participation equation:

Table 8
Simulated changes in labor force participation responses for EITC expansion 1984–1996

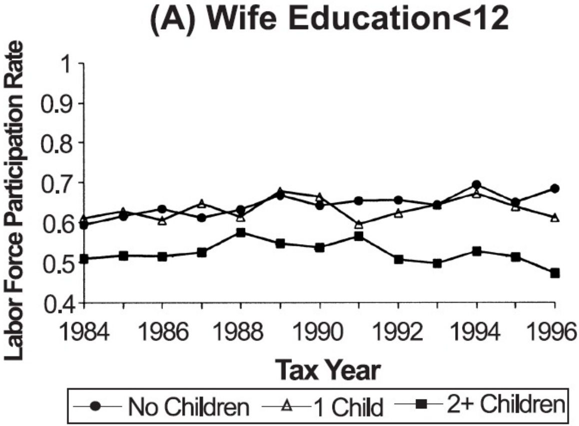
| | Percent of sample | Married women | | Married men | | Family | |
|---|---|---|---|---|---|---|---|
| | | Change in employment probability | | Change in employment probability | | EITC | |
| | | Level | Percent | Level | Percent | Gross | Net |
| Overall | 100 | – 0.011 | – 2.4 | 0.002 | 0.2 | 927 | 858 |
| *Grouping by husband's predicted wage* | | | | | | | |
| Decile 1 | | – 0.017 | – 4.2 | 0.006 | 0.6 | 1379 | 1315 |
| Decile 2 | | – 0.016 | – 3.8 | 0.004 | 0.4 | 1349 | 1279 |
| Decile 3 | | – 0.015 | – 3.6 | 0.003 | 0.3 | 1218 | 1132 |
| Decile 4 | | – 0.013 | – 3.0 | 0.003 | 0.3 | 1087 | 1022 |
| Decile 5 | | – 0.013 | – 2.3 | 0.002 | 0.2 | 1019 | 939 |
| Decile 6 | | – 0.011 | – 1.8 | 0.002 | 0.2 | 778 | 718 |
| Decile 7 | | – 0.007 | – 1.5 | 0.002 | 0.2 | 736 | 704 |
| Decile 8 | | – 0.010 | – 1.8 | 0.000 | 0.0 | 650 | 539 |
| Decile 9 | | – 0.009 | – 1.7 | 0.000 | 0.0 | 642 | 546 |
| Decile 10 | | – 0.005 | – 0.9 | 0.000 | 0.0 | 415 | 356 |
| *Grouping by location in 1996 EITC segment* | | | | | | | |
| Phase-in | 8.8 | 0.011 | 10.0 | 0.004 | 0.6 | 1144 | 1289 |
| Flat | 6.0 | – 0.015 | – 6.5 | 0.002 | 0.2 | 2424 | 2355 |
| Phase-out | 42.9 | – 0.021 | – 5.0 | 0.002 | 0.2 | 1591 | 1455 |
| >Phase-out | 42.3 | – 0.006 | – 0.8 | 0.001 | 0.1 | 0 | – 41 |

# Eissa and Hoynes (2004): Downsides of the paper (1)

- Assumptions of the diff-in-diffs approach:

1. Common trend assumption of the same trend
    - families with and without children can be different!!!
    - the two groups need to face the same trend in labor supply
    - problem would be if work preferences of mothers *changed* differently that those of non-mothers
2. assumption of **no composition changes**
    - composition of groups stays the same over time
    - no effect of EITC on decision to get married and have children

**Assumption of the common trend in LFP**



# Eissa and Hoynes (2004): Downsides of the paper (3)

- **Unitary household labor supply model:**
  - Wife's participation decision has no effect on husband's.
  - Do you think that there are many families in which husband decides to stay at home if his wife is working, while he would go to work if his wife is at home?
- **Participation in the shadow economy:**
  - Can the results be invalidated because authors did not consider shadow economy?
  - Diff-in-diffs approach: assumption of the same trend.
  - It would be invalidated only if treated women were more likely to start working in the shadow economy after the EITC expansion than the control group women.

# Bibliography


- Eissa, N., & Hoynes, H. W. (2004). Taxes and the labor market participation of married couples: The earned income tax credit. *Journal of Public Economics*, 88(9-10), 1931–1958. doi:10.1016/j.jpubeco.2003.09.005
- Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, 47(1), 153–161. doi:10.2307/1912352

EVROPSKÁ UNIE
Evropské strukturální a investiční fondy
Operační program Výzkum, vývoj a vzdělávání

MINISTERSTVO ŠKOLSTVÍ,
MLÁDEŽE A TĚLOVÝCHOVY