Stata Conference

*Portland 2024*

# Estimating a Probit Model with a continuous endogenous covariate and using complex survey data: an application to socioeconomic mobility analysis in Mexico.

August 1-2, 2024


Sylvia Beatriz Guillermo Peón

Benemérita Universidad Autónoma de Puebla

Alejandro Miguel Castañeda Valencia

Benemérita Universidad Autónoma de Puebla

Juan Enrique Huerta Wong

Vocería de la Presidencia de la República


BUAP | Facultad de Economía

# Research Objectives

- To **estimate the probability** of an individual's destination being at a high socioeconomic level as a function of a set of explanatory variables.
  - A high socioeconomic level is defined as being located at the top tertile of the economic resources index distribution.
  - Given the endogeneity of the education variable, the probability function is estimated using **a Probit model with an instrumental variable under the context of a complex survey data set.**

# Research Objectives

- We analyze the influence of higher education and parental socioeconomic status on the offspring's probability of a high socioeconomic destination in three residence areas of Mexico: South Region, Mexico City and Nuevo Leon. These are Mexico's three most referenced and contrasting geographical areas regarding inequality of opportunities, poverty, and development.

  - Nuevo Leon and Mexico City (CEEY, 2019c; CEEY, 2023) are the two federal entities reported with the highest opportunities for social mobility, more extensive possibilities of social ascension, and hence larger opportunities of poverty overcoming.
  - Southern region states are the ones reported with the lowest degree of upward social mobility

# Data

We use data from the **two latest surveys conducted by the Center of Studies Espinosa Yglesias** (CEEY)

- The 2017 ESRU **Survey on Social Mobility in Mexico** (ESRU-EMOVI-2017)

  This national survey provides current and retrospective information on the interviewees' characteristics and their parents; it has statistical representation for women and men at the regional level, including five regions in Mexico: North, Northwest, Center-North, Center, and South. Additionally, within the Center region, the sampling design includes a Mexico City representative sample.

- The 2021 **ESRU Survey on Social Mobility in Nuevo Leon** (ESRU-EMOVI Nuevo Leon 2021)

- The data for the South region and Mexico City are merged with the EMOVI-Nuevo Leon to construct a database considering the complex sampling design characteristics of the two surveys (primary sampling units, strata and expansion factors).

# Procedure

## 1) Measuring Socioeconomic Level

- We estimate two indexes of total economic resources to measure parental and informants' socioeconomic levels.
- The indexes are divided into tertiles so that parental and offspring socioeconomic levels are defined by their corresponding tertile of the economic resources indexes distribution.
- Indexes are estimated using multiple correspondence analysis on a matrix of categorical variables expressing the individual's asset holdings.

# Procedure

## 2) The Structural Model

The dependent variable:

$hd_i^*$ = high destination; is a continuous and unobserved (latent) variable representing the individual's propensity to be located in the top socioeconomic stratum.

$hd_i$ = tertile of the socioeconomic (total economic resources) index distribution in which each interviewee (offspring) is located and takes on two values:

$hd_i = 1$ if the interviewee's current hierarchical position in the socioeconomic structure is in the third (top) tertile and
$hd_i = 0$ otherwise.

And the relationship between the observed (binary) and unobserved (continuous) variables is:

$$hd_i = \begin{cases} 1 \;\; if \;\; hd_i^* > 0 \;\; propensity \; of \; destination \; at \; the \; high \; socioeconomic \; strata \\ 0 \;\; if \;\; hd_i^* \leq 0 \;\; propensity \; of \; destination \; at \; the \; non - high \; socioeconomic \; strata \end{cases} \quad \textbf{(1)}$$

# Procedure

Under the previous definition, the model can be formally expressed as:

$$hd_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \gamma educ_i + e_i \qquad \textbf{(2)}$$

$$educ_i = \boldsymbol{x}_i\boldsymbol{\alpha} + \boldsymbol{z}_i\boldsymbol{\theta} + u_i \qquad \textbf{(3)}$$

Where:

$\boldsymbol{x}_i$ = raw vector of $K$ exogenous explanatory variables for the interviewed individual $i$

$educ_i$ = individual $i$'s years of schooling (endogenous variable)

$\boldsymbol{\beta}$ = column vector of $K$ structural parameters associated with the exogenous explanatory variables

$\gamma$ = the structural parameter associated with years of schooling

$\boldsymbol{z}_i$ is a raw vector of L=3 external instruments (instrumental variables)

$\boldsymbol{\alpha}$ and $\boldsymbol{\theta}$ are the $K \times 1$ and $L \times 1$ vectors of the reduced form parameters

$e_i$ and $u_i$ are the standard normal distributed structural error and reduced form error terms, respectively.

# Procedure

The likelihood function is derived considering that the joint density $f(hd_i, educ_i|\boldsymbol{x}_i, \boldsymbol{z}_i)$ can be written as (Wooldridge, 2010: p. 476; Stata 17: p. 1142):

$$f(hd_i, educ_i|\boldsymbol{x}_i, \boldsymbol{z}_i) = f(hd_i|educ_i, \boldsymbol{x}_i, \boldsymbol{z}_i) \times f(educ_i|\boldsymbol{x}_i, \boldsymbol{z}_i) \quad \textbf{(4)}$$

Therefore, the log likelihood function is expressed as:

$$\ln L = \sum_{i=1}^{N} w_i \left\{ hd_i \ln \boldsymbol{\Phi}(m_i) + (1 - hd_i) \ln[1 - \Phi(m_i)] + \ln \phi \left( \frac{educ_i - \boldsymbol{x}_i\boldsymbol{\alpha} - \boldsymbol{z}_i\boldsymbol{\theta}}{\sigma} \right) - ln\sigma \right\} \quad (\mathbf{5})$$

where

$$m_i = \frac{\boldsymbol{x}_i\boldsymbol{\beta} + \gamma educ_i + \rho(educ_i - \boldsymbol{x}_i\boldsymbol{\alpha} - \boldsymbol{z}_i\boldsymbol{\theta})/\sigma}{(1 - \rho^2)^{1/2}}$$

# Procedure

The probability of a destination at a high socioeconomic level for an individual as a function of a set of explanatory variables can be expressed as (Wooldridge, 2010: p. 476):

$$P(hd_i = 1|\boldsymbol{x}_i, educ_i) = \boldsymbol{\Phi}\left[\frac{\boldsymbol{x}_i\boldsymbol{\beta} + \gamma educ_i + \rho(educ_i - \boldsymbol{x}_i\boldsymbol{\alpha} - \boldsymbol{z}_i\boldsymbol{\theta})/\sigma}{(1-\rho^2)^{1/2}}\right] \qquad (7)$$

# Procedure

## 3) IV Probit or Standard Probit?

In order to choose the appropriate estimation method, we must test if the variable ***educ*** is endogenous in the model. That is, we need to test:

$$\rho = \text{Corr}\,(educ_i, e_i) = 0$$ ➔ ***educ*** is exogenous ➔ Standard Probit

This is a Wald exogeneity test.

The STATA ***ivprobit*** command **used with Survey Data Analysis** does not provide the Wald exogeneity test (it does so but only for the *ivprobit* without considering the survey design).

To obtain the Wald's test statistic, we estimate the model using STATA ***ivprobit*** command (also using the conditional maximum-likelihood estimator) and **using the expansion factors as sampling weights as well as clustered robust standard errors**, where the cluster variable is the primary sampling unit (Long and Freese, 2014). Point estimates and their standard errors have exactly the same values as those obtained with ***ivprobit*** under the survey data analysis setup.

# Procedure

| | ivprobit | | ivprobit_svy | |
|---|---|---|---|---|
| educ_y | .2076758 | *** | .2076758 | *** |
| | (.017296) | | (.0173042) | |
| Sex | | | | |
| female | -.0295701 | | -.0295701 | |
| | (.0566901) | | (.0563435) | |
| age | .0435869 | ** | .0435869 | ** |
| | (.0175668) | | (.0174567) | |
| age # age | -.0002026 | | -.0002026 | |
| | (.0002037) | | (.0002027) | |
| sec_origin | | | | |
| medium | .2532842 | *** | .2532842 | *** |
| | (.0969488) | | (.0962878) | |
| high | .6978133 | *** | .6978133 | *** |
| | (.1385345) | | (.1385158) | |
| skin_tone | | | | |
| dark | -.1478777 | ** | -.1478777 | ** |
| | (.0673965) | | (.0673212) | |
| area | | | | |
| rural | -.082247 | | -.082247 | |
| | (.0881862) | | (.0883836) | |

Based on Long & Freese (2014) estimation procedure we get the Walt Test of exogeneity with STATA ivprobit command

| | | |
|---|---|---|
| corr(e.educ_y,e.hd) | -0.3462 | 0.0648 |
| sd(e.educ_y) | 3.4510 | 0.0519 |

```
Wald test of exogeneity (corr = 0): chi2(1) = 24.04
Instrumented: educ_y
 Instruments: 1.sex age c.age#c.age 2.sec_origin 3.sec
              2.region 3.region educ_yho overcrowding_
```

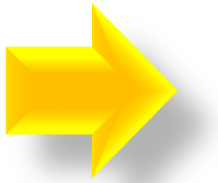
The appropriate procedure is
**IVPROBIT**

# Procedure

## 4) Testing Instruments' strength

Given that the model has only one endogenous covariate, we only need one strong instrument. Tthe test for instruments' strength is simply a joint significance test after the ***ivprobit*** command with survey analysis:

```
. test  educ_yho overcrowding_ho 1.floor_ho

Adjusted Wald test

 ( 1)  [educ_y]educ_yho = 0
 ( 2)  [educ_y]overcrowding_ho = 0
 ( 3)  [educ_y]1.floor_ho = 0

       F(  3,  1047) =  106.48
            Prob > F =    0.0000
```



At least one instrument is strong

# Procedure

## 5) Overidentification test for the exogeneity of instruments

- No STATA command is available to test the exogeneity of instruments (instruments' validity test).
- When estimating discrete choice models, testing instruments exogeneity is somehow more difficult because the error term of the model is latent (not observed); so we performed **the Refutability Test**, developed by Angelo Guevara (2018).
- This test states that:

  Under Ho, all instruments are exogenous (valid)

  Under H1, at least one instrument is endogenous

# Procedure

**Stage 1:**
Estimate the reduced form equation for the endogenous variable (*educ*) by OLS and obtain the residuals.

$$educ_i = \boldsymbol{x}_i\boldsymbol{\alpha} + \boldsymbol{z}_i\boldsymbol{\theta} + u_i \quad \rightarrow \hat{u}_i$$

**Stage 2:**
Estimate the structural equation, including the residuals from Stage 1 as an auxiliary variable to control for the endogeneity, and retrieve the log-likelihood of this restricted **Control Function** $L_R^{CF}$.

$$hd_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \gamma educ_i + \delta\hat{u}_i + v_i \quad \rightarrow \quad L_R^{CF}$$

**Stage 3:**
Estimate the structural model again, including now not only $\hat{u}_i$, but also $L-1$ (two) of the instruments as additional variables, and retrieve the log-likelihood of this unrestricted Control Function $L_U^{CF}$.

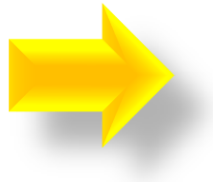
$$hd_i^* = \boldsymbol{x}_i\boldsymbol{\beta} + \gamma educ_i + \delta\hat{u}_i + \alpha_1 z_{1i} + \alpha_2 z_{2i} + v_i \quad \rightarrow L_U^{CF}$$

# Procedure

**The test statistic of the Refutability tes**t is calculated as a **likelihood ratio test** in which the model estimated in Stage 2 is the restricted version of the model estimated in Stage 3:

$$LR = -2\left(L_U^{CF} - L_R^{CF}\right) \sim \chi^2_{(L-1)}$$

```
. lrtest probitu1 probitr, force

Likelihood-ratio test
Assumption: probitr nested within probitu1

 LR chi2(2) =   2.03
Prob > chi2 = 0.3626
```



Cannot reject Ho:
All instruments are exogenous

# Estimation Results and Analysis

```
. svy linearized : ivprobit hd i.sex  age c.age#c.age  i.sec_origin i.skin_tone i.area b1.region (educ_y=
> educ_yho overcrowding_ho i.floor_ho), cformat(%5.4f)  pformat(%5.3f) sformat(%5.3f)
(running ivprobit on estimation sample)

Survey: Probit model with endogenous regressors

Number of strata =    19                        Number of obs    =      8,465
Number of PSUs   = 1,068                        Population size = 22,279.071
                                                 Design df        =      1,049
                                                 F(10, 1040)      =     100.03
                                                 Prob > F         =     0.0000
```

| | Linearized Coefficient | std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| hd | | | | | | |
| educ_y | 0.2077 | 0.0173 | 12.001 | 0.000 | 0.1737 | 0.2416 |
| sex | | | | | | |
| female | -0.0296 | 0.0563 | -0.525 | 0.600 | -0.1401 | 0.0810 |
| age | 0.0436 | 0.0175 | 2.497 | 0.013 | 0.0093 | 0.0778 |
| c.age#c.age | -0.0002 | 0.0002 | -1.000 | 0.318 | -0.0006 | 0.0002 |
| sec_origin | | | | | | |
| medium | 0.2533 | 0.0963 | 2.630 | 0.009 | 0.0643 | 0.4422 |
| high | 0.6978 | 0.1385 | 5.038 | 0.000 | 0.4260 | 0.9696 |
| skin_tone | | | | | | |
| dark | -0.1479 | 0.0673 | -2.197 | 0.028 | -0.2800 | -0.0158 |
| area | | | | | | |
| rural | -0.0822 | 0.0884 | -0.931 | 0.352 | -0.2557 | 0.0912 |
| region | | | | | | |
| Mexico City | 0.7176 | 0.0742 | 9.675 | 0.000 | 0.5720 | 0.8631 |
| Nuevo Leon | 0.9807 | 0.0866 | 11.321 | 0.000 | 0.8107 | 1.1506 |
| _cons | -4.5574 | 0.4129 | -11.037 | 0.000 | -5.3676 | -3.7471 |

| | | | | | | |
|---|---|---|---|---|---|---|
| educ_y | | | | | | |
| sex | | | | | | |
| female | -0.7516 | 0.1259 | -5.968 | 0.000 | -0.9987 | -0.5045 |
| age | 0.0364 | 0.0461 | 0.788 | 0.431 | -0.0542 | 0.1269 |
| c.age#c.age | -0.0012 | 0.0005 | -2.250 | 0.025 | -0.0022 | -0.0002 |
| sec_origin | | | | | | |
| medium | 0.8222 | 0.2089 | 3.935 | 0.000 | 0.4122 | 1.2321 |
| high | 2.1481 | 0.3012 | 7.131 | 0.000 | 1.5570 | 2.7392 |
| skin_tone | | | | | | |
| dark | -0.6066 | 0.1376 | -4.410 | 0.000 | -0.8766 | -0.3367 |
| area | | | | | | |
| rural | -0.5673 | 0.1957 | -2.899 | 0.004 | -0.9514 | -0.1833 |
| region | | | | | | |
| Mexico City | -0.3199 | 0.1962 | -1.631 | 0.103 | -0.7049 | 0.0651 |
| Nuevo Leon | -0.8060 | 0.2177 | -3.702 | 0.000 | -1.2333 | -0.3788 |
| educ_yho | 0.2909 | 0.0173 | 16.829 | 0.000 | 0.2570 | 0.3248 |
| overcrowding_ho | -0.0894 | 0.0282 | -3.170 | 0.002 | -0.1447 | -0.0341 |
| 1.floor_ho | -1.3862 | 0.2225 | -6.229 | 0.000 | -1.8229 | -0.9496 |
| _cons | 10.0287 | 0.9597 | 10.450 | 0.000 | 8.1455 | 11.9118 |
| /athrho2_1 | -0.3611 | 0.0737 | -4.897 | 0.000 | -0.5058 | -0.2164 |
| /lnsigma2 | 1.2387 | 0.0149 | 83.038 | 0.000 | 1.2094 | 1.2679 |
| corr(e.educ_y,e.hd) | -0.3462 | 0.0649 | | | -0.4666 | -0.2131 |
| sd(e.educ_y) | 3.4510 | 0.0515 | | | 3.3515 | 3.5535 |

Instrumented: educ_y
Instruments: 1.sex age c.age#c.age 2.sec_origin 3.sec_origin 1.skin_tone 1.area
2.region 3.region educ_yho overcrowding_ho 1.floor_ho

# Estimation Results and Analysis

## 6) Estimating Average Marginal Effects

Average marginal effects

Number of strata = 19 — Number of obs = 8,465
Number of PSUs = 1,068 — Population size = 22,279.071
Model VCE: Linearized — Design df = 1,049

Expression: Average structural function probabilities, predict(pr)
dy/dx wrt: educ_y 1.sex age 2.sec_origin 3.sec_origin 1.skin_tone 1.area 2.region 3.region

| | Delta-method dy/dx | std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| educ_y | .0470281 | .0045422 | 10.35 | 0.000 | .0381153 | .0559408 |
| sex | | | | | | |
| female | -.0067013 | .0127602 | -0.53 | 0.600 | -.0317397 | .0183371 |
| age | .0061183 | .0008608 | 7.11 | 0.000 | .0044292 | .0078075 |
| sec_origin | | | | | | |
| medium | .0612959 | .0226895 | 2.70 | 0.007 | .0167738 | .1058179 |
| high | .1825759 | .0363959 | 5.02 | 0.000 | .1111589 | .253993 |
| skin_tone | | | | | | |
| dark | -.0338293 | .015444 | -2.19 | 0.029 | -.064134 | -.0035245 |
| area | | | | | | |
| rural | -.0187855 | .0201318 | -0.93 | 0.351 | -.0582887 | .0207177 |
| region | | | | | | |
| Mexico City | .1882577 | .0206496 | 9.12 | 0.000 | .1477384 | .2287769 |
| Nuevo Leon | .2649556 | .0255709 | 10.36 | 0.000 | .2147796 | .3151316 |

Note: dy/dx for factor levels is the discrete change from the base level.

# Estimation Results and Analysis

## 7) Computing the percentage of Correctly Classified outcomes

STATA does not provide a function to obtain the percentage of correctly classified outcomes under the Survey Data Analysis framework, and if we use the ***pweights*** option (as we do with the Long & Freese procedure) the ***estat classification*** command is not allowed.

**Our proposal**

**Step 1**

Obtain the estimated probabilities after estimating the ***ivprobit*** model with **SVY**,

```
predict prhat, pr
```

**Step 2**

Define the estimated 0, 1 outcomes based on the estimated probabilities

```
gen hd_hat=(prhat>=0.5)
```

# Estimation Results and Analysis

**Step 3**

Generate a variable (correct) taking on value 1 iff predicted outcome = observed outcome and 0 otherwise

```
gen correct=(hd_hat==hd)
```

**Step 4**

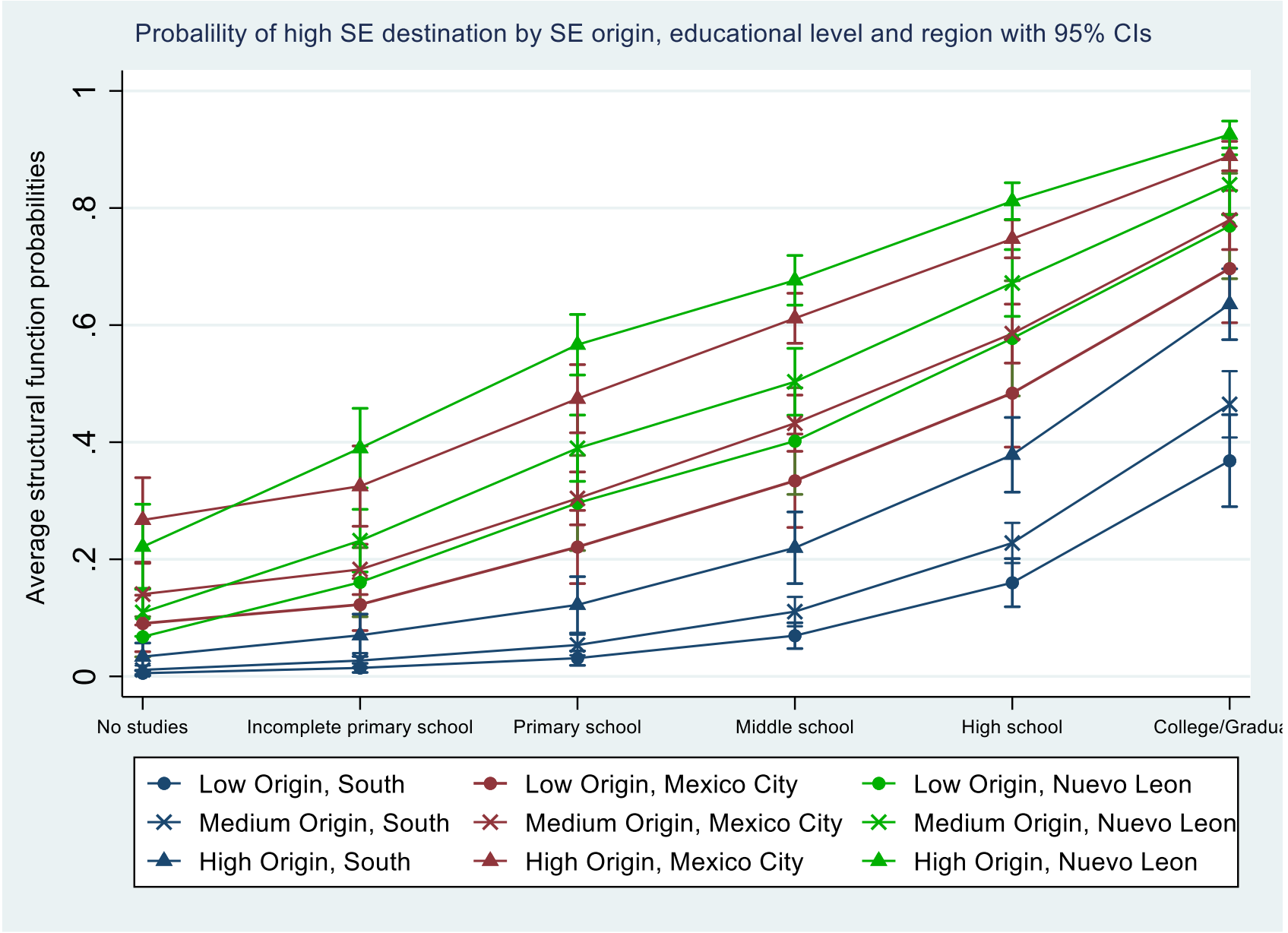
Tabulate

```
tab correct
```

| correct | Freq. | Percent | Cum. |
|---|---|---|---|
| 0 | 2,059 | 24.32 | 24.32 |
| 1 | 6,406 | 75.68 | 100.00 |
| Total | 8,465 | 100.00 | |

# Estimation Results and Analysis

**8) Comparing probabilities of high Socioeconomic destination by SE origin, educational level and region**



# Estimation Results and Analysis

## 9) Estimating Odds Ratios

### Social Reproduction matters: The probability premium of higher education by socioeconomic origin

We computed some odds ratios to analyze how the probability premium of higher education changes by socioeconomic origin

$$\frac{\overline{\boldsymbol{P}}(hd_i = 1|\boldsymbol{x}_i, educ_i, sec\_origin_i, region_i)}{\overline{\boldsymbol{P}}(hd_i = 1|\boldsymbol{x}_i, educ_i = high\ shool, sec\_origin_i, region_i)}$$

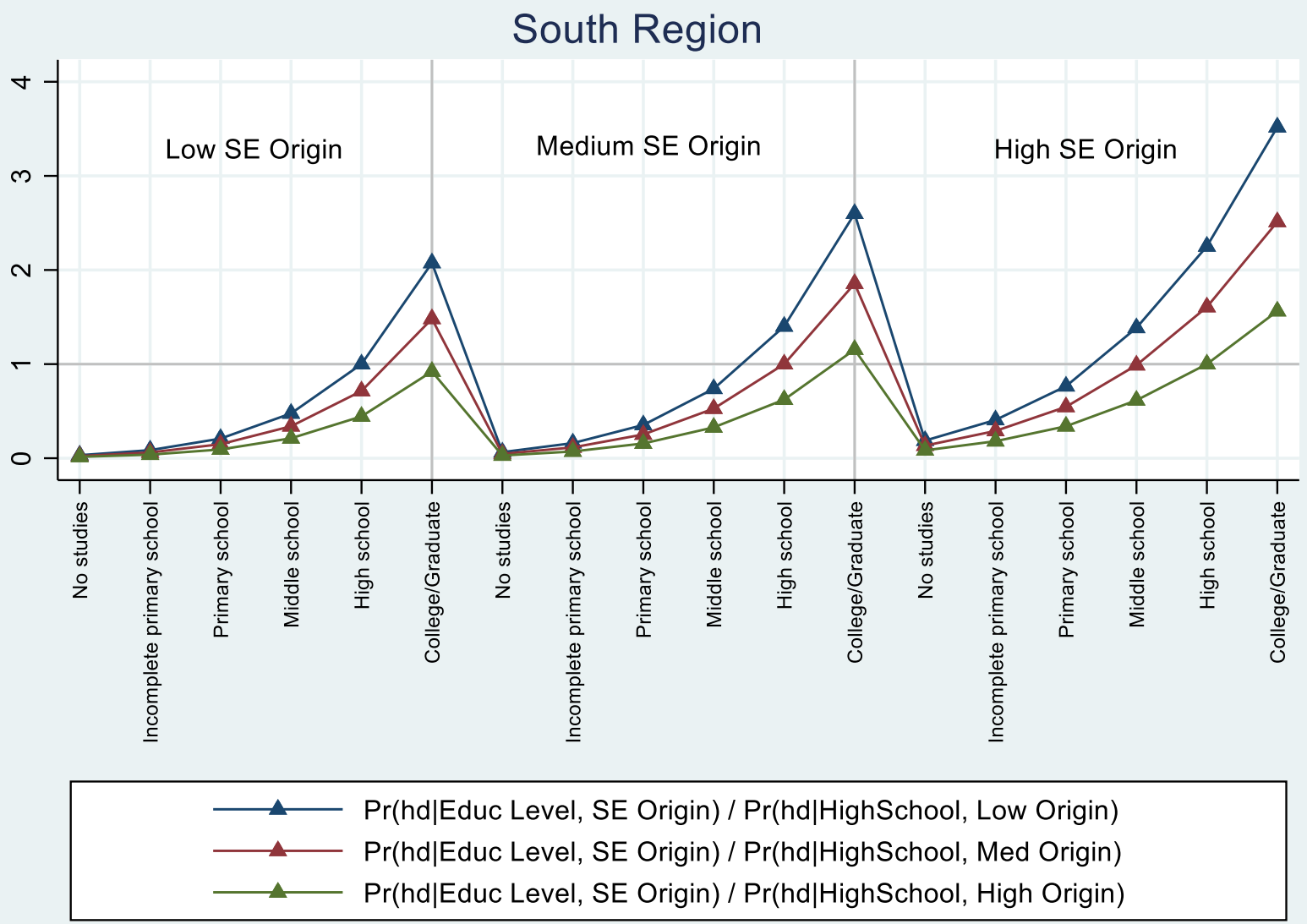
We did these calculations using the margins postestimation functions

# Estimation Results and Analysis

```
. margins sec_origin#region, by(educ) predict(pr) cformat(%5.3f)  pformat(%5.2f) sformat(%5.2f)  post

Predictive margins

Number of strata =    19                        Number of obs    =      8,465
Number of PSUs   = 1,068                        Population size = 22,279.071
Model VCE: Linearized                           Design df        =      1,049

Expression: Average structural function probabilities, predict(pr)
Over:       educ
```

| | Delta-method Margin | std. err. | t | P>\|t\| | [95% conf. | interval] |
|---|---|---|---|---|---|---|
| educ#sec_origin#region | | | | | | |
| No studies#low#South | 0.006 | 0.002 | 2.70 | 0.01 | 0.002 | 0.010 |
| No studies#low#Mexico City | 0.036 | 0.011 | 3.32 | 0.00 | 0.015 | 0.058 |
| No studies#low#Nuevo Leon | 0.064 | 0.017 | 3.66 | 0.00 | 0.030 | 0.098 |

# Estimation Results and Analysis

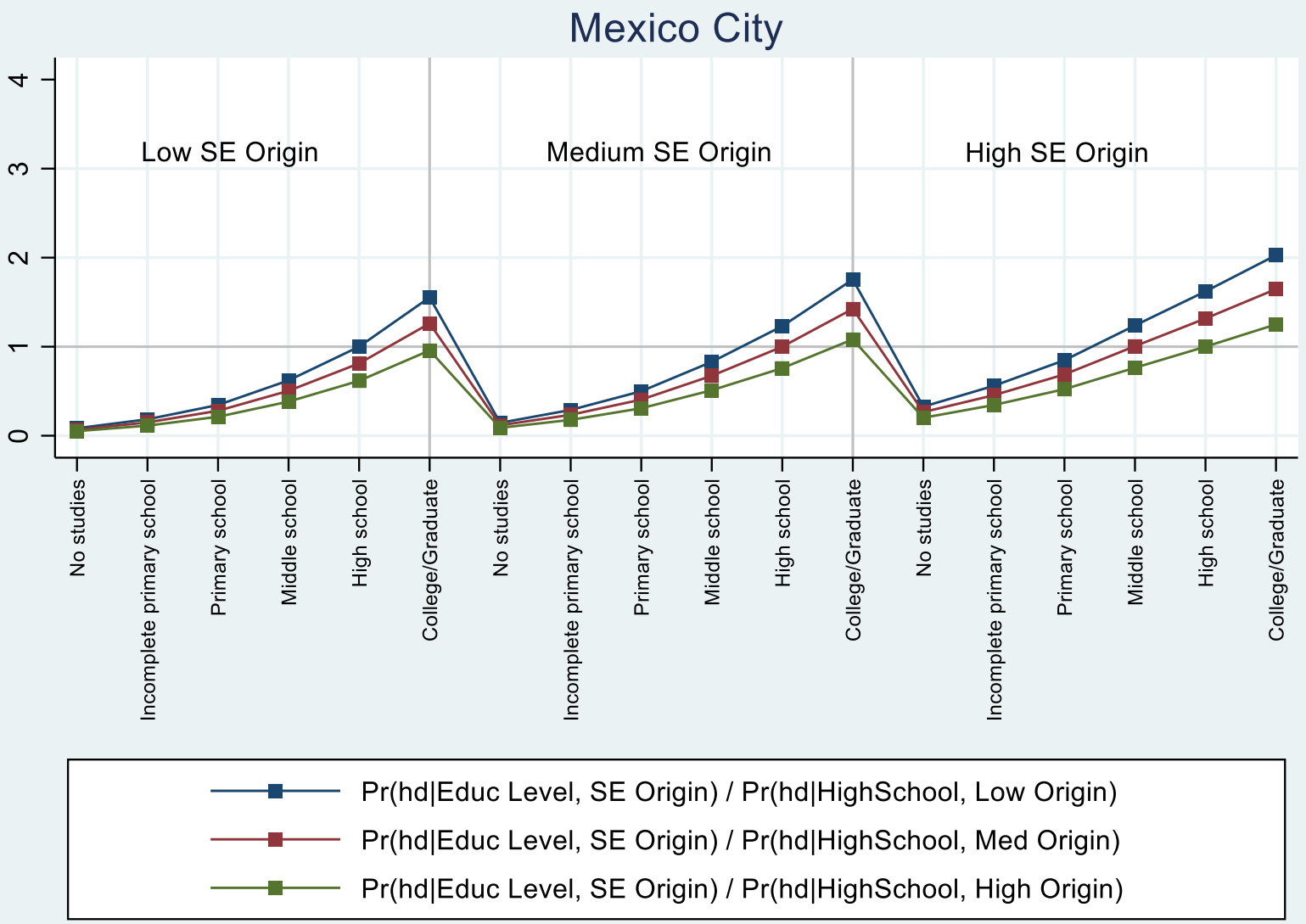



## 9) Odds Ratios

**probability premium of higher education by socioeconomic origin**

$$\frac{\bar{\boldsymbol{P}}(hd_i = 1|\boldsymbol{x}_i, educ_i, sec\_origin_i, region_i)}{\bar{\boldsymbol{P}}(hd_i = 1|\boldsymbol{x}_i, educ_i = high\ shool, sec\_origin_i, region_i)}$$

# Estimation Results and Analysis

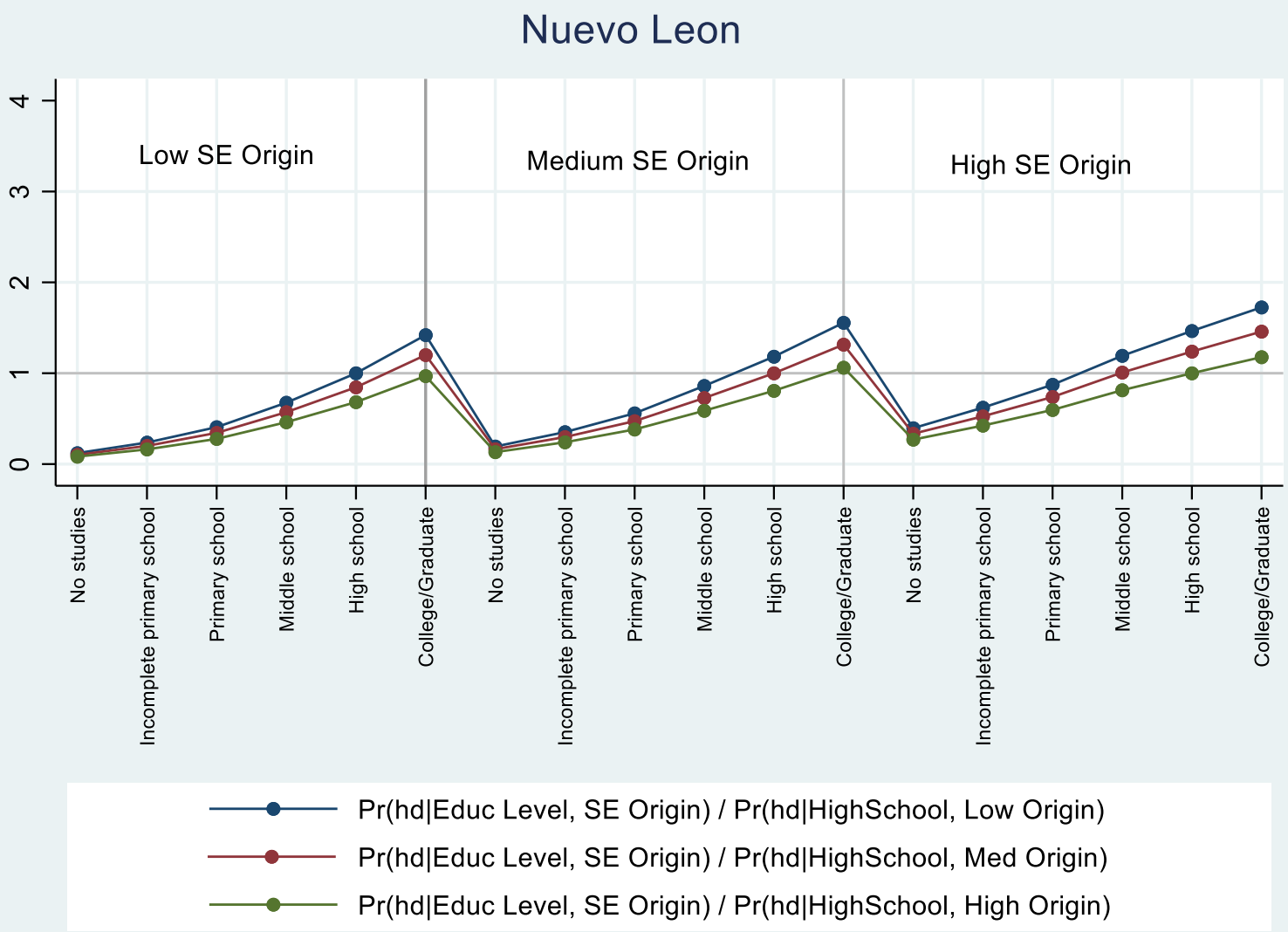



## 9) Odds Ratios

**probability premium of higher education by socioeconomic origin**

$$\frac{\bar{\boldsymbol{P}}(hd_i = 1|\boldsymbol{x}_i, educ_i, sec\_origin_i, region_i)}{\bar{\boldsymbol{P}}(hd_i = 1|\boldsymbol{x}_i, educ_i = high\ shool, sec\_origin_i, region_i)}$$

# Estimation Results and Analysis




## 9) Odds Ratios

**probability premium of higher education by socioeconomic origin**

$$\frac{\overline{\boldsymbol{P}}(hd_i = 1|\boldsymbol{x}_i, educ_i, sec\_origin_i, region_i)}{\overline{\boldsymbol{P}}(hd_i = 1|\boldsymbol{x}_i, educ_i = high\ shool, sec\_origin_i, region_i)}$$

# Estimation Results and Analysis

| Socioeconomic Stratum of origin | Educational level | High school, Low origin | High school, Medium origin | High school, High origin |
|---|---|---|---|---|
| Low | | | | |
| | No studies | 0.03 *** | 0.02 *** | 0.01 *** |
| | Incomplete primary school | 0.09 *** | 0.06 *** | 0.04 *** |
| | Primary school | 0.21 *** | 0.15 *** | 0.09 *** |
| | Middle school | 0.48 *** | 0.34 *** | 0.21 *** |
| | High school | 1.00 | 0.71 *** | 0.44 *** |
| | College/Graduate | 2.07 *** | 1.48 *** | 0.92 |
| Medium | | | | |
| | No studies | 0.06 *** | 0.05 *** | 0.03 ** |
| | Incomplete primary school | 0.16 *** | 0.11 *** | 0.07 ** |
| | Primary school | 0.35 *** | 0.25 *** | 0.16 ** |
| | Middle school | 0.74 ** | 0.53 *** | 0.33 ** |
| | High school | 1.40 ** | 1.00 | 0.62 *** |
| | College/Graduate | 2.60 *** | 1.85 *** | 1.15 |
| High | | | | |
| | No studies | 0.19 *** | 0.13 *** | 0.08 *** |
| | Incomplete primary school | 0.41 *** | 0.29 *** | 0.18 *** |
| | Primary school | 0.76 | 0.55 *** | 0.34 *** |
| | Middle school | 1.39 | 0.99 | 0.62 *** |
| | High school | 2.25 *** | 1.61 *** | 1.00 |
| | College/Graduate | 3.52 *** | 2.51 *** | 1.56 *** |

## South Region

## Odds Ratio Test

## Testing Odds=1

```
. testnl _b[6.educ#1.sec_origin#1.region]/_b[5.educ#1.sec_origin#1.region] = 1

  (1)  _b[6.educ#1.sec_origin#1.region]/_b[5.educ#1.sec_origin#1.region] = 1

           chi2(1) =      105.61
       Prob > chi2 =      0.0000
```

# Estimation Results and Analysis

## 10) Testing the Lucky High Schooler Hypothesis

"Individuals with no more than high school education (the *lucky high schooler*) have the same probability of a high destination in the socioeconomic distribution compared to those who have attained a university educational level."

Using estimated average probabilities (STATA margins functions), we were able to test if

$$H_0: \; P(hd_i = 1|\boldsymbol{x}_i, educ_i = college/graduate\,) - \; P(hd_i = 1|\boldsymbol{x}_i, educ_i = high\ school\,) = 0$$

```
. test 6.educ=5.educ

Adjusted Wald test

 ( 1)  - 5bn.educ + 6.educ = 0

       F(  1,  1049) =  496.12
            Prob > F =    0.0000
```

# Challenges found

- Wald exogeneity test for Survey Data Analysis
- Test of Instruments' strength (particularly if there is more than one endogenous covariate)
- Overidentification test for the exogeneity of instruments (instruments validity test)
- Percentage of Correctly Classified outcomes for the estimated model
- Pseudo-$R^2$

# References


CEEY, Centro de Estudios Espinosa Yglesias (2019c). Movilidad Social en la Ciudad de México. 2019. Available at: https://ceey.org.mx/informe-de-movilidad-social-en-la-ciudad-de-mexico-2019/#:~:text=Los%20resultados%20del%20estudio%20muestran,lo%20largo%20de%20su%20vida.

CEEY, Centro de Estudios Espinosa Yglesias (2023). Documento Metodológico Encuesta ESRU de Movilidad Social en Nuevo Leon 2021. Availible at: https://ceey.org.mx/contenido/que-hacemos/emovi/

Guevara A. (2018). Overidentification tests for the exogeneity of instruments in discrete choice models. Transportation Research Part B 114 (2018) 241–253. https://www.sciencedirect.com/science/article/abs/pii/S0191261518302303?via%3Dihub

Long J. Scott and Freese Jeremy (2014). Regression Models for Categorical Dependent Variables Using Stata, Third Edition. Stata Press

STATA 17. Base Reference Manual, Release 2017. Ivprobit-Probit model with continuous endogenous covariates. Available at: https://www.stata.com/manuals/rivprobit.pdf

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

# Stata Conference

*Portland 2024*

BUAP | Facultad de Economía