How to assess the fit of choice models with Stata?

2024 German Stata Conference
at GESIS in Mannheim,
June 7, 2024

"There is no safety in numbers." Howard S. Wainer

Dr. Wolfgang Langer
Martin-Luther-Universität
Halle-Wittenberg
Institut für Soziologie

Associate Assistant Professor
University of
Luxembourg

Outline

- **What is the problem?**
- **What is the solution in Stata?**
- **Example of application**
- **Conclusions**

What is the problem?

- **In 1992 Stata V3 introduced the clogit-command to estimate Conditional (fixed-effects) logistic regression model which calculates the McFadden Pseudo R^2**
- **In 2007 Stata V10 introduced the asclogit-command to estimate the alternative-specific conditional logit model**
- **In 2019 Stata V16 introduced the Choice Models (cm) commands**
- **But none of them calculates the Likelihood-Ratio-chi² test statistic and any Pseudo R^2 to assess the fit of the model !**

What is the solution in Stata?

- **My fit_cmclogit.ado calculates for McFadden's conditional logit choice model the following test statistic and Pseudo R^2s tested by Monte Carlo simulation studies in the 1990s / 2000s**
  - **Likelihood-Ratio-χ^2 test statistic using a zero model with alternative-specific constants**
  - **McFadden Pseudo R^2 (likelihood-ratio-index) (1974)**
  - **Adjusted McFadden Pseudo R^2 (1985)**
  - **Maddala Pseudo R^2 (1983)**
  - **Cragg & Uhler Pseudo R^2 (1970)**
  - **Aldrich & Nelson Pseudo R^2 (1984)**
  - **Aldrich & Nelson Pseudo R^2 with Veall & Zimmermann correction (1994)**

Example of application

- **North Rhine-Westphalia Election Study of 1995**
  - **CATI Survey with 504 respondents (eligible voters)**
  - **Endogenous variable: voting intention for the German parties SPD, FDP or CDU: 1) yes 0) no**
  - **Exogenous variables**
    - **Generic / alternative specific: long term preference for one of the three parties (gprefall): 1) yes 0) no**
    - **Case-specific variables:**
      - **Religious denomination (confession): 1) yes 0) no**
      - **Educational degree (education): 1) secondary modern 2) secondary modern+ 3) grammar school 4) college/university**
  - **Balanced hierarchical data structure**
    - **Party alternatives are nested within respondents**

Stata 18 Output

```
Conditional logit choice model                   Number of obs      =      1,512
Case ID variable: probnr                         Number of cases    =        504

Alternatives variable: party                     Alts per case: min =          3
                                                                avg =        3.0
                                                                max =          3

                                                    Wald chi2(9)    =     263.97
Log likelihood = -259.67913                         Prob > chi2     =     0.0000

------------------------------------------------------------------------------------
              vote | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------------+----------------------------------------------------------------
party              |
          gprefall |
               yes |   2.193726   .1401447    15.65   0.000     1.919048    2.468405
-------------------+----------------------------------------------------------------
SPD                |
        confession |
               yes |  -.9000949   .3084457    -2.92   0.004    -1.504637   -.2955524
                   |
         education |
       sec.modern+ |  -.1846034   .3412324    -0.54   0.589    -.8534067    .4841998
    grammar school |   -.645902   .5506053    -1.17   0.241    -1.725069    .4332646
college/university |   -1.03819   .6887728    -1.51   0.132     -2.38816    .3117801
                   |
             _cons |   .4353825   .2737489     1.59   0.112    -.1011554    .9719205
-------------------+----------------------------------------------------------------
FDP                |
        confession |
               yes |  -.6455168   .3947333    -1.64   0.102     -1.41918    .1281462
                   |
         education |
       sec.modern+ |   1.393966   .4604399     3.03   0.002     .4915205    2.296412
    grammar school |   2.076665   .6434303     3.23   0.001     .8155643    3.337765
college/university |   3.160799   .5990928     5.28   0.000     1.986598    4.334999
                   |
             _cons |  -1.956077   .3772011    -5.19   0.000    -2.695377   -1.216776
-------------------+----------------------------------------------------------------
CDU                |  (base alternative)
------------------------------------------------------------------------------------
```

Output of my fit_cmclogit.ado

```
. fit_cmclogit

Likelihood-Ratio-chi2 test against zero model with ASCs

H0: all alternative-/case-specific-effects are zero in the population

LR chi2(  9) =    463.32    Prob > chi2 = 0.0000

Fit-Indices for the Alternative-Specific-Conditional-Logit model:

McFadden Pseudo R2 (compared with zero model with ASCs) = 0.4715

McFadden Pseudo R2 with Ben-Akiva & Lerman correction = 0.4532

Maddala ML Pseudo R2 = 0.6012

Cragg & Uhler Pseudo R2 = 0.7009

Aldrich & Nelson Pseudo R2 = 0.4790

Aldrich & Nelson Pseudo R2 with Veall & Zimmermann correction = 0.7246
```

Excellent fit!

My ado returns the following r-containers

```
. return list

scalars:
                 r(logl_m0) =  -491.3376899127339
                 r(logl_ma) =  -259.679126768394 8
               r(an_pr2_vz) =  .724628733565413 9
                  r(an_pr2) =  .478971284284291 3
                  r(cu_pr2) =  .700944868912334 4
                  r(ml_pr2) =  .601193926861091 2
                r(rho2_bar) =  .453168091346473 5
                    r(rho2) =  .471485432321472 8
                    r(lr_p) =  0
                   r(lr_df) =  9
                 r(lr_chi2) =  463.317126288678 2
```

Conclusions

- **What have I shown?**
  - **My fit_cmclogit.ado allows to assess the fit of McFadden's choice model in a user-friendly way.**
  - **It provides all information we need to evaluate the model fit.**
- **What's in progress?**
  - **Following extensions are in the pipeline**
    - **Construction of McFadden's prediction-success table**
    - **Calculation of a separate McKelvey & Zavoina Pseudo R^2 for each logit equation**

Closing words

- **Thank you for your attention**
- **Do you have some questions?**

Contact

- **Affiliation**

  - **Dr. Wolfgang Langer**
    University of Halle
    Institute of Sociology
    D 06099 Halle (Saale)

  - **Email:**
    - **wolfgang.langer@soziologie.uni-halle.de**

  - **Url:**
    - **https://langer.soziologie.uni-halle.de**

References

- Aldrich, J.H. & Nelson, F.D. (1984):
  Linear probability, logit, and probit models. Newbury Park: SAGE (Quantitative Applications in the Social Sciences, 45)
- Amemiya, T. (1981):
  Qualitative response models: a survey. *Journal of Economic Literature*, 21, pp.1483-1536
- Ben-Akiva,M. & S.R.Lerman 1991[4] (1985):
  Discrete choice analysis. Theory and application to travel demand. Cambridge, Mass: MIT-Press
- Cox, D.R.& Snell, E.J. (1989):
  The analysis of binary data. London: Chapman&Hill
- Cragg, S.G.& Uhler, R. (1970):
  The demand for automobiles. *Canadian Journal of Economics*, 3, pp. 386-406
- DeMaris, A.(2002):
  Explained variances in logistic regression. A Monte Carlo study of proposed measures.*Sociological Methods&Research*, 11, 1, pp. 27-74
- Domencich,T.A. & McFadden, D.L. (1975): Urban travel demand. A behavioral analysis. Amsterdam u. Oxford: North Holland Publishing Company
- Efron, B. (1978):
  Regression and Anova with zero-one data. Measures of residual variation. *Journal of American Statistical Association*, 73, pp. 113-121
- Hagle, T.M. & Mitchell II,G.E. (1992):
  Goodness of fit measures for probit and Logit. *American Journal of Political Science*, 36, 3, pp. 762-784

References 2

- **Hensher, D.A., Rose, J.M. & Greene (2005):**
  Applied choice analysis. A primer. Cambridge: Cambridge University Press
- **Long, J.S. (1997):**
  Regression models for categorical and limited dependent variables. Thousand Oaks, Ca : Sage
- **Long, J.S. & Freese, J. (2000):**
  Scalar measures of fit for regression models. Bloomington, : Indiana University
- **Long, J.S. & Freese, J. (2003[2]):**
  Regression models for categorical dependent variables using Stata. College Station, Tx: Stata
- **Maddala, G.S. (1983):**
  Limited-dependent and qualitative variables in econometrics. Cambridge: Cambridge University Press
- **McFadden, D. (1974): Conditional logit analysis of qualitative choice behavior. In: P.Zarembka (ed.), Frontiers in Econometrics. New York: Academice Press, pp. 105-142**
- **McFadden, D. (1978):**
  Quantitative methods for analysing travel behaviour of individuals: some recent developments. In: D.A. Hensher & P.R. Stopher: (eds): Behavioural travel modelling. London: Croom Helm, pp. 279-318
- **McKelvey, R. & Zavoina, W. (1975):**
  A statistical model for the analysis of ordinal level dependent variables. *Journal of Mathematical Sociology*, 4, pp. 103-20
- **Nagelkerke, N.J.D. (1991):**
  A note on a general definition of the coefficient of determination. *Biometrika*, 78, 3, pp.691-693

References 3

- Veall, M.R. & Zimmermann, K.F. (1992): Pseudo-R2 in the ordinal probit model. *Journal of Mathematical Sociology*, 16, 4, pp. 333-342
- Veall, M.R. & Zimmermann, K.F. (1994): Evaluating Pseudo-R2's for binary probit models. *Quality&Quantity*, 28, pp. 151 - 164
- Windmeijer, F.A.G. (1995): Goodness-of-fit measures in binary choice models. *Econometric Reviews*, 14, 1, pp. 101-116
- Zimmermann, K.F. (1993): Goodness of fit in qualitative choice models: review and evaluation. In: H. Schneeweiß & K. Zimmermann (eds): Studies in applied econometrics. Heidelberg: Physika, pp. 25-74

Appendix

What is the solution?

- **Short review of the Monte-Carlo studies made by econometricians to test systematically the most common Pseudo R^2s for binary and ordinal probit / logit models**
  - **Hagle & Mitchell 1992**
  - **Veall & Zimmermann 1992, 1993, 1994**
  - **Windmeijer 1995**
  - **DeMaris 2002**

- **My fit_cmclogit.ado to calculate the most important Pseudo-R^2s**

Which Pseudo-R^2s were tested in the MC studies?

- **Likelihood-based measures:**
  - **Maddala / Cox & Snell Pseudo R^2 (1983/1989)**
  - **Cragg & Uhler / Nagelkerke Pseudo R^2 (1970/1992)**
- **Log-Likelihood-based measures:**
  - **McFadden Pseudo R^2 (1974)**
  - **Aldrich & Nelson Pseudo R^2 (1984)**
  - **Aldrich & Nelson Pseudo R^2 with the Veall & Zimmermann correction (1992)**
- **Basing on the estimated probabilities:**
  - **Efron / Lave Pseudo R^2 (1970 / 1978)**
- **Basing on the variance decomposition of the estimated Probits / Logits:**
  - **McKelvey & Zavoina Pseudo R^2 (1975)**

Results of the Monte-Carlo-studies for binary / ordinal logits or probits

- **The McKelvey & Zavoina Pseudo R^2 is the best estimator for the "true R^2" of the OLS regression**
- **The Aldrich & Nelson Pseudo R^2 with the Veall & Zimmermann correction is the best approximation of the McKelvey & Zavoina Pseudo R^2**
- **Lave / Efron, Aldrich & Nelson, McFadden and Cragg & Uhler Pseudo R^2 underestimate the "true R^2" of the OLS regression**
- **My personal advice: Use the McKelvey & Zavoina Pseudo R^2 or the Aldrich & Nelson Pseudo R^2 with Veall & Zimmermann correction to assess the fit of binary and ordinal logit models**

Log-Likelihood-based measures 1

- **McFadden-Pseudo-R^2 (1974) provided by Stata**

$$McFadden\ Pseudo\ R^2\left(\rho^2\right) = 1 - \left[\frac{\log L_A}{\log L_0}\right]$$

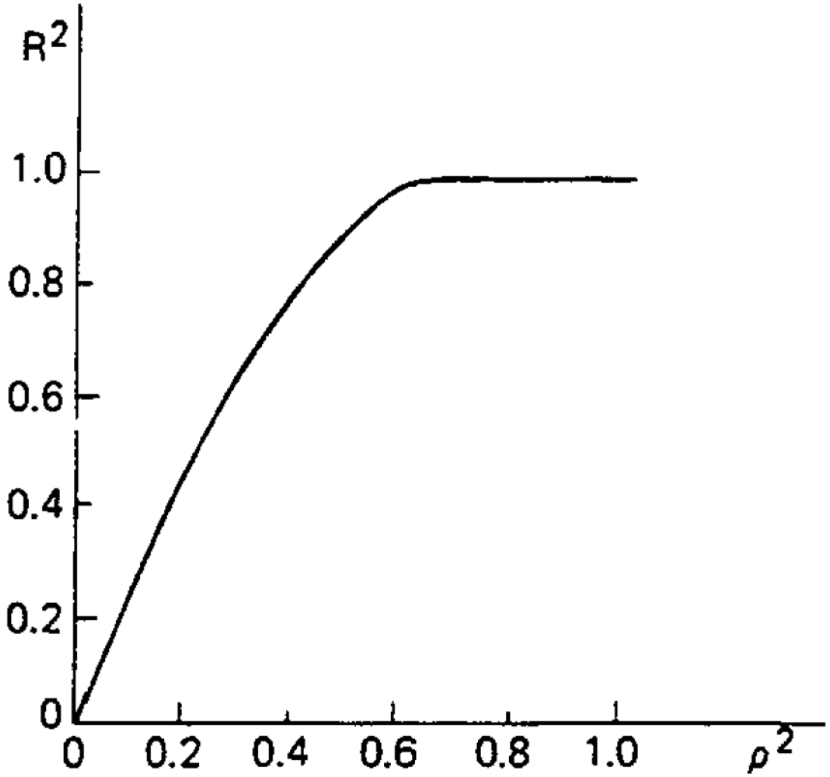
Theoretical range: $0 \leq$ McFadden Pseudo $R^2 \leq 1$

but ρ^2 does not reach its maximum of one!

Rule of thumb: $0.20 \leq$ McFadden Pseudo $R^2 \leq 0.40$ marks an excellent fit. It is equivalent to $0.40 \leq R^2 \leq 0.80$ of a linear regression model (McFadden 1978: 307)

Legend: $\log L_A$: Log-Likelihood of alternative model
$\log L_0$: Log-Likelihood of zero model

Relationship between McFadden's ρ^2 and R^2 of the regression model



(Figure 5.5 in Domencich & McFadden 1975: 124)

Interpretation

„Those unfamiliar with the *p2* index should be forewarned that its values tend to be considerably lower than those of the R2 index and should not be judged by the standards for a 'good fit' in ordinary regression analysis. For example, values of 0.2 to 0.4 for *p2* represent an excellent fit."

(McFadden 1978: 307)

Log-Likelihood-based measures 2

- **Adjusted McFadden Pseudo R^2 (1985)**

$$McFadden\ Pseudo\ R^2_{adjusted}\left(\bar{\rho}^2\right) = 1 - \left[\frac{\log L_A - K}{\log L_0}\right]$$

Correction of McFadden Pseudo R² by the total number of estimated logistic slopes (K) proposed by Ben-Akiva & Lerman (1985: 167)

Range: $0 \leq$ McFadden Pseudo $R^2_{adjusted} \leq 1$,
but it does not reach its maximum of one!

Likelihood-based measures 1

- **Maddala Pseudo-R^2 (1983) or Cox & Snell Pseudo R^2 (1989):**

$$Maddala\ Pseudo\ R^2 (R^2_{ML}) = 1 - \left[\frac{L_0}{L_A} \right]^{\frac{2}{n}}$$

$$= 1 - \exp\left(\frac{-L.R.\chi^2}{n} \right) = 1 - \exp\left(\frac{-2 \times \left(\log L_A - \log L_0 \right)}{n} \right)$$

$$Range: \quad 0 \leq Maddala\ Pseudo\ R^2 \leq 1 - L_0^{\frac{2}{n}}$$

Legend:
L_0 : Likelihood of zero model (constant only)
L_A : Likelihood of alternative model
n : number of cases

Likelihood-based measures 2

- **Cragg & Uhler Pseudo R² (1970) or Nagelkerke Pseudo R² (1991)**

$$Cragg\ \&\ Uhler\ Pseudo\ R^2 = \frac{R^2_{ML}}{\max .R^2_{ML}}$$

$$= \frac{1 - \left[\frac{L_0}{L_A}\right]^{\frac{2}{n}}}{1 - L_0^{\frac{2}{n}}} = \frac{1 - \exp\left(\frac{-L.R.\chi^2}{n}\right)}{1 - \exp\left(\frac{2}{n} \times \log L_0\right)}$$

Correction of the Maddala Pseudo R^2 by its own theoretical maximum➜ Range: $0 \leq$ C&U Pseudo $R^2 \leq 1$

Legend: log: Logarithmus naturalis
exp: Exponential function

Log-Likelihood-based measures 3

- **Aldrich & Nelson Pseudo R^2 (1984)**

$$Aldrich\ \&\ Nelson\ Pseudo\ R^2 = \frac{L.R.\chi^2}{L.R.\chi^2 + n}$$

$$= \frac{2 \times (\log L_A - \log L_0)}{2 \times (\log L_A - \log L_0) + n}$$

Veall & Zimmermann Correction

- **Veall & Zimmermann (1994) propose a correction of the Aldrich & Nelson Pseudo R^2 by its upper limit**
  - **Range of the A&N Pseudo R^2**

$$0 \leq A\,\&\,N\ Pseudo\ R^2 \leq \frac{-2 \times \log L_0}{n - 2 \times \log L_0}$$

  - **Correction formula**

$$A\,\&\,N\ Pseudo\ R^2_{V\,\&\,Z} = \frac{\dfrac{2 \times \left(\log L_A - \log L_o\right)}{2 \times \left(\log L_A - \log L_o\right) + n}}{\dfrac{-2 \times \log L_0}{n - 2 \times \log L_0}}$$

Basing on the estimated probabilities

- **Lave / Efron Pseudo R^2 (1979)**

$$Lave\,/\,Efron\ Pseudo\ R^2 = 1 - \frac{\sum_{i=1}^{n}\left(Y_i - \hat{p}_i\right)^2}{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}$$

with $\overline{Y} = \frac{1}{n} \times \sum_{i=1}^{n} Y_i$

Legend:

Y_i: Value of the dependent variable for case i (1 or 0)

$\hat{p}_i$: Estimated probability Y=1 for case i

$\overline{Y}$: Relative frequency of Y=1

Variance decomposition of estimated Y*

- **McKelvey & Zavoina Pseudo R^2 (M&Z Pseudo R^2)**

$$M\ \&\ Z\ Pseudo\ R^2 = \frac{Var\left(\hat{y}^*\right)}{Var\left(\hat{y}^*\right)+Var\left(\varepsilon\right)} = \frac{\dfrac{\sum_{i=1}^{n}\left(\hat{y}_i^* - \overline{\hat{y}^*}\right)^2}{n}}{\dfrac{\sum_{i=1}^{n}\left(\hat{y}_i^* - \overline{\hat{y}^*}\right)^2}{n} + n \times \frac{\pi^2}{3}}$$

Range: $0 \leq$ M&Z Pseudo $R^2 \leq 1$

Legend:

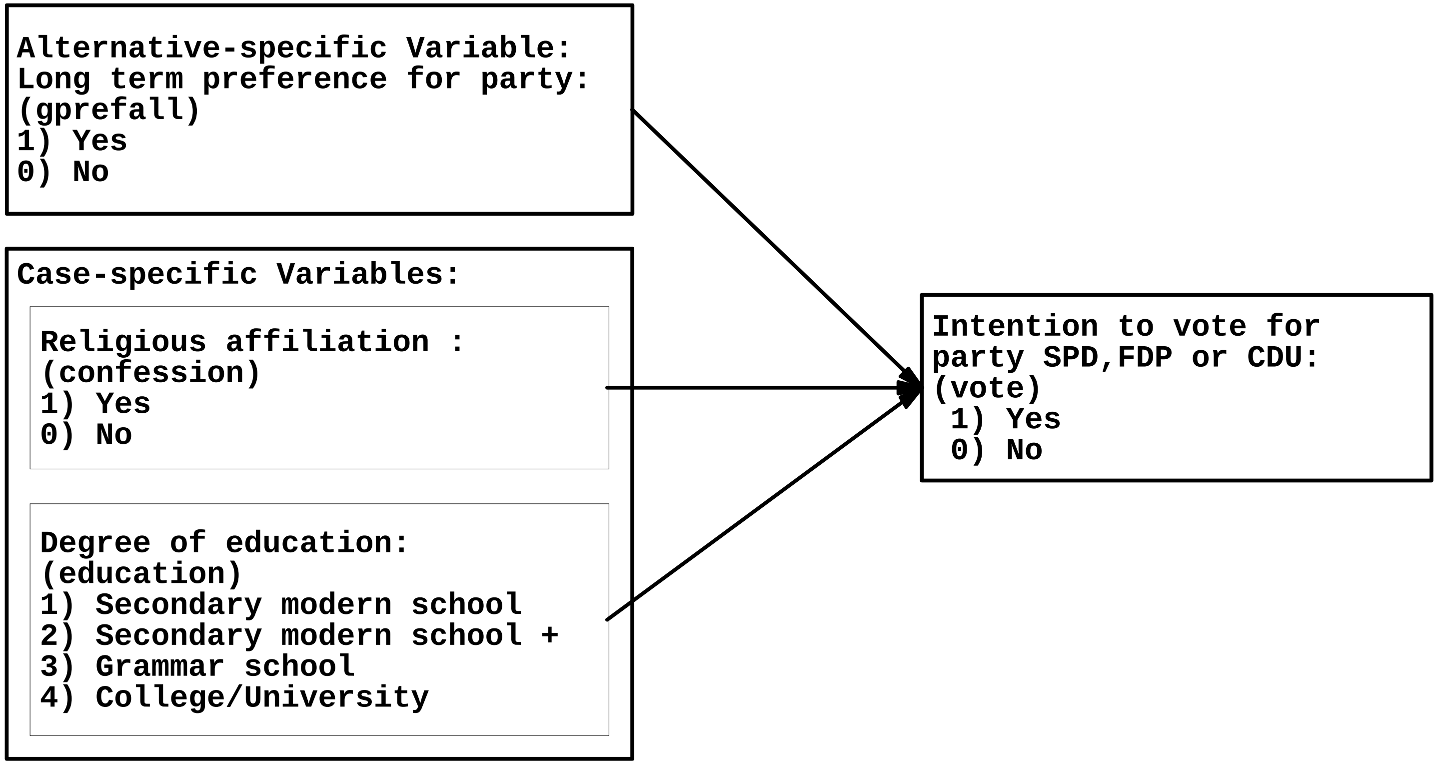
$Var\left(\hat{y}^*\right)$: Variance of the estimated logits (latent variable Y^*)

$\hat{y}_i^*$: Estimated logit of case i

$\overline{\hat{y}^*}$: Mean of the estimated logits

$\frac{\pi^2}{3}$: Variance of logistic density function

Theoretical Model



- **Reference group: respondents with secondary modern degree, without religious affiliation and no party preference**

McFadden's choice model (cmclogit)

- **Estimation equation**

$$\ln\left[\frac{P_i(Y=j)}{P_i(Y=J)}\right] = \sum_{k=1}^{K} \gamma_k \times (z_{ijk} - z_{iJk}) + \sum_{j}^{J-1} \alpha_j + \sum_{l=1}^{L} ß_l \times X_{il}$$

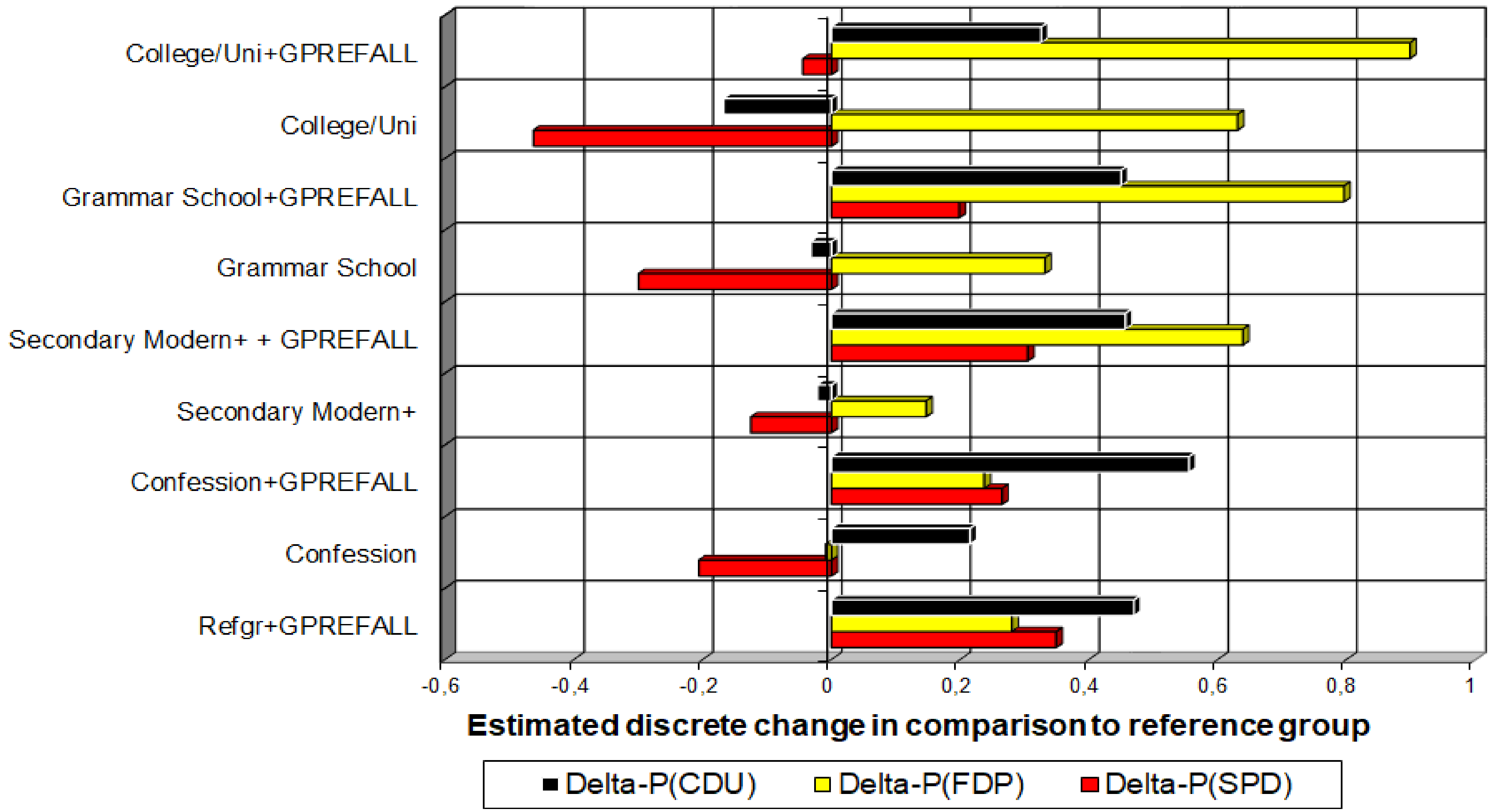
Rational Choice-part with alternative-specific γ-logit slopes for the difference of Z_k

Multinomial logit model to estimate the effects case-specific exogenous variables

β-logistic slope of the effect of X_l on comparison j vs. J

Logistic constant for the comparison j vs. J

Estimated effects of exogenous variables



Reference group: **P(CDU)=0.3722** **P(FDP) = 0.0526** **P(SPD)= 0.5751**