The Assessment of Fit in the Class of Logistic Regression Models: A Pathway out of the Jungle of Pseudo-R^2s Using Stata

Meeting of the German Stata User Group at GESIS in Cologne, June 10th, 2016

“Models are to be used, but not to be believed.”
Henri Theil

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

Associate Assistant Professor
Université du
Luxembourg

Contents:

- 1. What is the problem?
- 2. Summary of the econometric Monte-Carlo studies for Pseudo-R^2s
- 3. The generalization of the McKelvey & Zavoina Pseudo-R^2 for multinomial logit model
- 4. An application of the generalized M&Z Pseudo-R^2 in an election study of East German students
- 5. Conclusions

1. What is the problem ?

Current situation in applied research:

- An increasing number of people uses logistic models for qualitative dependent variables
- But users often complain about the bad fit of logistic models especially for the multinomial ones
- There is no general agreement on how to assess their fit corresponding to practical significance
- Let me show you the pathway out of the jungle of the pseudo-coefficients of determination

Which solutions does Stata provide?

- Indeed, for binary, ordinal and multinomial logit model Stata calculates only the McFadden Pseudo-R^2

- but J.Scott Long & Jeremy Freese have published their fitstat.ado in 2000. It calculates a set of Pseudo-R^2s for binary, ordinal, multinomial logit or limited dependent variable models discussed by Long in 1997

2. Summary of the econometric Monte-Carlo studies for testing Pseudo-R^2s

- Econometricians made a lot of Monte-Carlo studies in the early 90s:
  - Hagle & Mitchell 1992
  - Veall & Zimmermann 1992, 1993, 1994
  - Windmeijer 1995
  - DeMaris 2002
- They tested systematically the most common Pseudo-R^2s for binary and ordinal probit / logit models

Which Pseudo-R²s were tested in these studies?

- Likelihood-based measures:
  - Maddala / Cox & Snell Pseudo-R² (1983 / 1989)
  - Cragg & Uhler / Nagelkerke Pseudo-R² (1970 / 1992)
- Log-Likelihood-based measures:
  - McFadden Pseudo-R² (1974)
  - Aldrich & Nelson Pseudo-R² (1984)
  - Aldrich & Nelson Pseudo-R² with the Veall & Zimmermann correction (1992)
- Basing on the estimated probabilities:
  - Efron / Lave Pseudo-R² (1970 / 1978)
- Basing on the variance decomposition of the estimated Probits / Logits:
  - McKelvey & Zavoina Pseudo-R² (1975)

Results of the Monte-Carlo-Studies for binary and ordinal logits or probits

- The McKelvey & Zavoina Pseudo-R² is the best estimator for the “true R²” of the OLS regression
- The Aldrich & Nelson Pseudo-R² with the Veall & Zimmermann correction is the best approximation of the McKelvey & Zavoina Pseudo-R²
- Lave / Efron, Aldrich & Nelson, McFadden and Cragg & Uhler Pseudo-R² severely underestimate the “true R²” of the OLS regression
- My personal advice:
  - Use the McKelvey&Zavoina Pseudo-R² to assess the fit of binary and ordinal logit models

Let’s have a detailed look at the winner

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

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

Range: $0 \leq$ M&Z-Pseudo-R² ≤ 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

3. Generalization of McKelvey&Zavoina Pseudo-R^2 to multinomial logit model

- Equations of a multinomial logit model (MNL) for a dependent variable Y with 3 categories
  - Simultaneous estimation of the parameters of two logit equations instead of 2 separate binary logit models

$$(1)\ \log\left[\frac{P_{3i}}{P_{1i}}\right]=\sum_{k=0}^{K}\beta_{31k}X_{ki}\ \left\{+\varepsilon_{31i}\right\}$$

$$(2)\ \log\left[\frac{P_{2i}}{P_{1i}}\right]=\sum_{k=0}^{K}\beta_{21k}X_{ki}\ \left\{+\varepsilon_{21i}\right\}$$

Conditions of getting unbiased estimates

- Independence of Irrelevant Alternatives (IIA)-Axiom:
  - Comparison of two alternatives is independent of the existence of a third one
  - By using the MNL as a nonlinear probability model the IIA-assumption is fulfilled by the discrete and disjunctive categories of the dependent variable Y
- IID-Axiom formulated by Hensher, Rose & Greene (2005: 77):
  - The error terms ε are independently and identically distributed
    - Stochastic independence of ε_{21} and ε_{31}
    - Identical density function of ε_{21} and ε_{31}

Reasons to apply M&Z-Pseudo-R^2 to MNL

- The multinomial logit model (MNL) is ...
  - A multi-equation model
  - It has independent error terms ε_{21} and ε_{31}
  - ε_{21} and ε_{31} follow the logistic density function
- Therefore we can calculate the McKelvey & Zavoina Pseudo-R^2 separately for each comparison of categories
  - Simultaneous estimation by the multinomial logit model
  - Estimation by k-1 separate binary logit models (Begg & Gray 1984)
- Therefore I use the binary McKelvey-Zavoina-Pseudo-R^2s to validate the ones of the MNL

4. Application of the generalized M&Z Pseudo-R^2 in an election study

The Student Election Survey 1998 in Sachsen-Anhalt

- Population
  - 31.000 Students in 150 schools
  - All 5th thru 12th classes in all educational tracks
  - Age 10 thru 18 years

- Sample
  - Representative probability sample of 3.500 students in 22 schools
  - Survey date: 4 days after the general federal election (october 1st,1998)

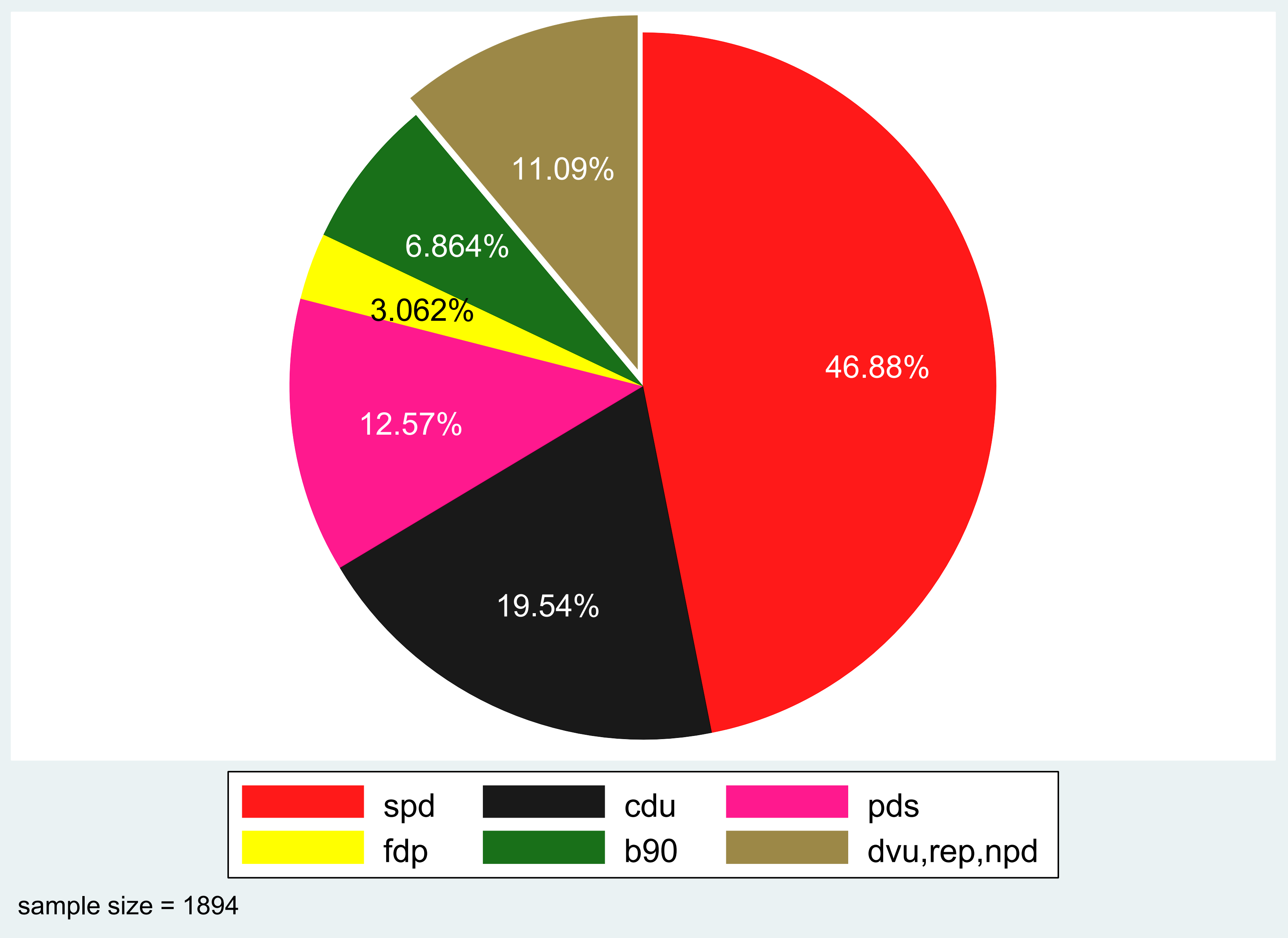
Independent variables

- C_AGE in years (centered)
- GENDER: boys vs. girls
- SCHOOL TYPE: GRAMMAR school, VOCATIONAL school vs. secondary school,
- Internal and external political C_EFFICACY (centered)
- Perceived influence of the peers on the vote (PEERS)
- Perceived influence of the parents (PARENTS)
- Perceived influence of the media (MEDIA)
- Perceived influence of the teachers (TEACHERS)
- Countryside vs. city (LOCATION)

Dependent variable

- VOTING for party
  - Social Democratic Party (SPD) [0]
  - Christian Democratic Union (CDU) [1]
  - Party of Democratic Socialism / Ex-SED communist party (PDS) [2]
  - Free Demokratic Party / Liberals (FDP) [3]
  - Alliance 90 / the Green (B90) [4]
  - Right-wing extremist parties (DVU, REP, NPD) [5]

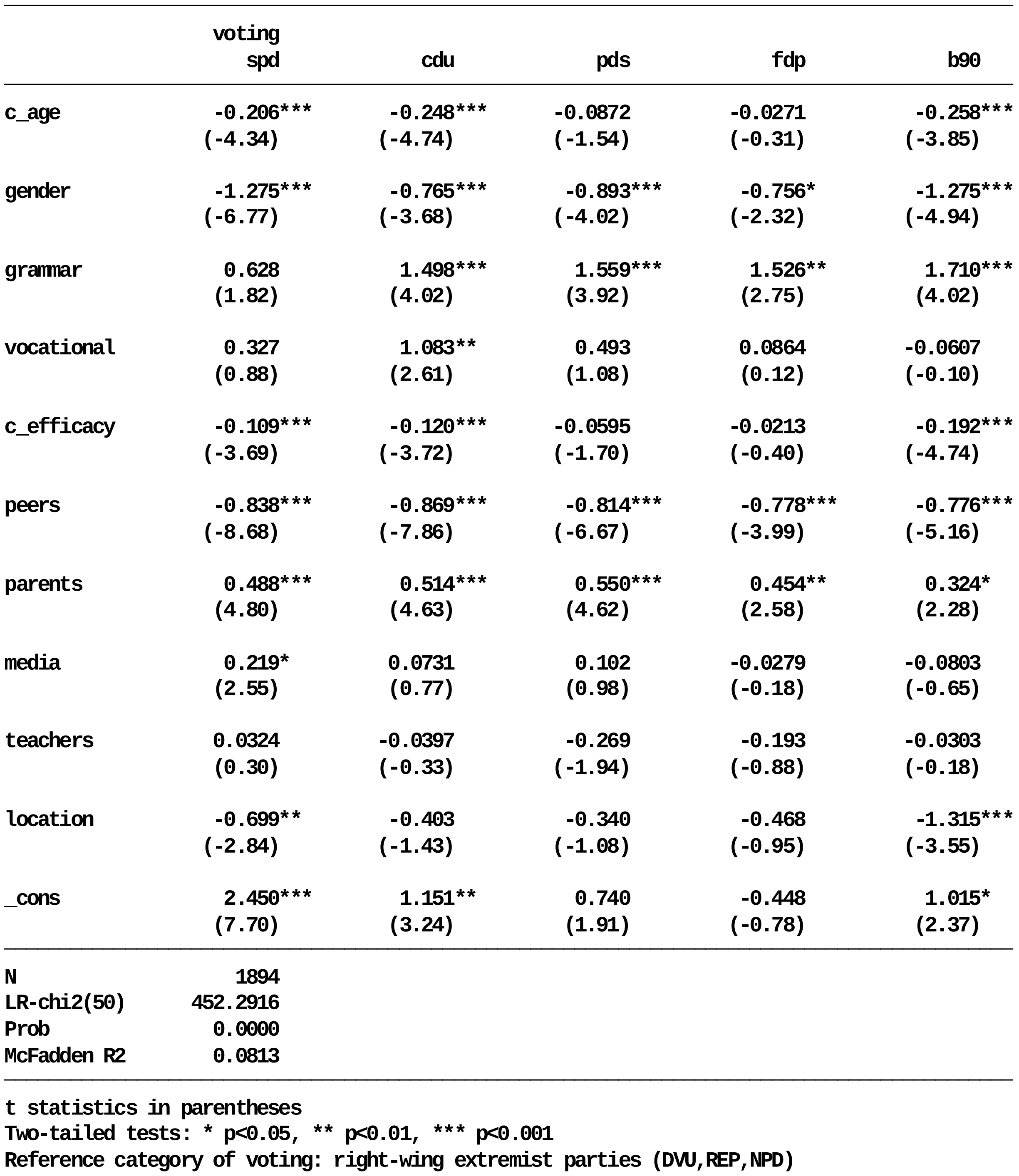
Students' party votes in LSA 1998



sample size = 1894

Estimated multinomial logit model for voting

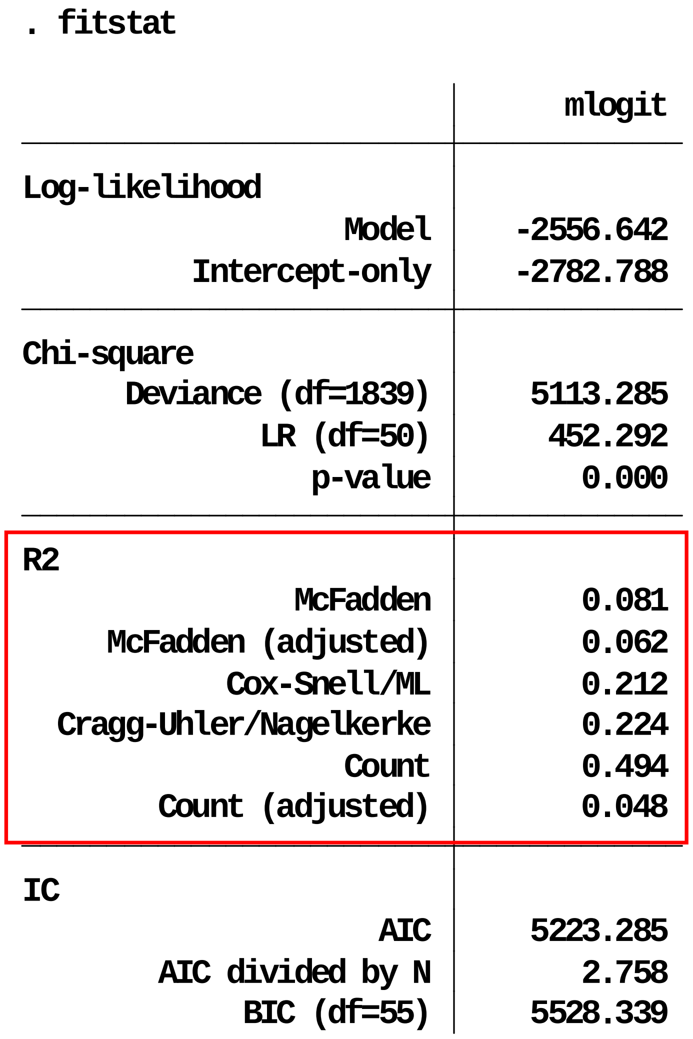
- Choice of the base outcome category
  - The comparison of right wing extremist vs. established parties marks the main political conflict line in East-Germany
- Stata mlogit output formated with Ben Jann esttab.ado



```
--------------------------------------------------------------------------------------------
                        voting
                           spd            cdu            pds            fdp            b90
--------------------------------------------------------------------------------------------
c_age                -0.206***      -0.248***      -0.0872        -0.0271        -0.258***
                    (-4.34)        (-4.74)        (-1.54)        (-0.31)        (-3.85)

gender               -1.275***      -0.765***      -0.893***      -0.756*        -1.275***
                    (-6.77)        (-3.68)        (-4.02)        (-2.32)        (-4.94)

grammar               0.628          1.498***       1.559***       1.526**        1.710***
                     (1.82)         (4.02)         (3.92)         (2.75)         (4.02)

vocational            0.327          1.083**        0.493          0.0864        -0.0607
                     (0.88)         (2.61)         (1.08)         (0.12)        (-0.10)

c_efficacy           -0.109***      -0.120***      -0.0595        -0.0213        -0.192***
                    (-3.69)        (-3.72)        (-1.70)        (-0.40)        (-4.74)

peers                -0.838***      -0.869***      -0.814***      -0.778***      -0.776***
                    (-8.68)        (-7.86)        (-6.67)        (-3.99)        (-5.16)

parents               0.488***       0.514***       0.550***       0.454**        0.324*
                     (4.80)         (4.63)         (4.62)         (2.58)         (2.28)

media                 0.219*         0.0731         0.102         -0.0279        -0.0803
                     (2.55)         (0.77)         (0.98)        (-0.18)        (-0.65)

teachers              0.0324        -0.0397        -0.269         -0.193         -0.0303
                     (0.30)        (-0.33)        (-1.94)        (-0.88)        (-0.18)

location             -0.699**       -0.403         -0.340         -0.468         -1.315***
                    (-2.84)        (-1.43)        (-1.08)        (-0.95)        (-3.55)

_cons                 2.450***       1.151**        0.740         -0.448          1.015*
                     (7.70)         (3.24)         (1.91)        (-0.78)         (2.37)
--------------------------------------------------------------------------------------------
N                      1894
LR-chi2(50)        452.2916
Prob                 0.0000
McFadden R2          0.0813
--------------------------------------------------------------------------------------------
t statistics in parentheses
Two-tailed tests: * p<0.05, ** p<0.01, *** p<0.001
Reference category of voting: right-wing extremist parties (DVU,REP,NPD)
```

Classical fit indices and Pseudo-R^{2}s

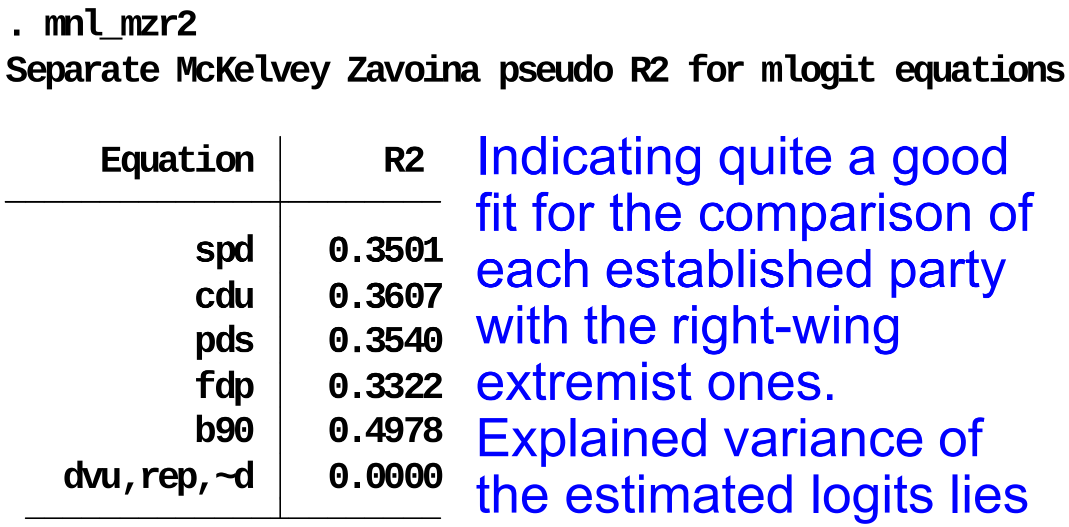
- **Calculated with Long & Freese's fitstat.ado**



```
. fitstat

                                   |        mlogit
-----------------------------------+---------------
Log-likelihood                     |
                             Model |     -2556.642
                    Intercept-only |     -2782.788
-----------------------------------+---------------
Chi-square                         |
                Deviance (df=1839) |      5113.285
                       LR (df=50)  |       452.292
                           p-value |         0.000
-----------------------------------+---------------
R2                                 |
                          McFadden |         0.081
               McFadden (adjusted) |         0.062
                       Cox-Snell/ML|         0.212
           Cragg-Uhler/Nagelkerke  |         0.224
                             Count |         0.494
                  Count (adjusted) |         0.048
-----------------------------------+---------------
IC                                 |
                               AIC |      5223.285
                 AIC divided by N  |         2.758
                     BIC (df=55)   |      5528.339
```

Indicating a bad overall fit of the MNL!

- **McKelvey&Zavoina Pseudo-R^2 for each of k-1 comparisons of Y using my mnl_mrz2.ado**

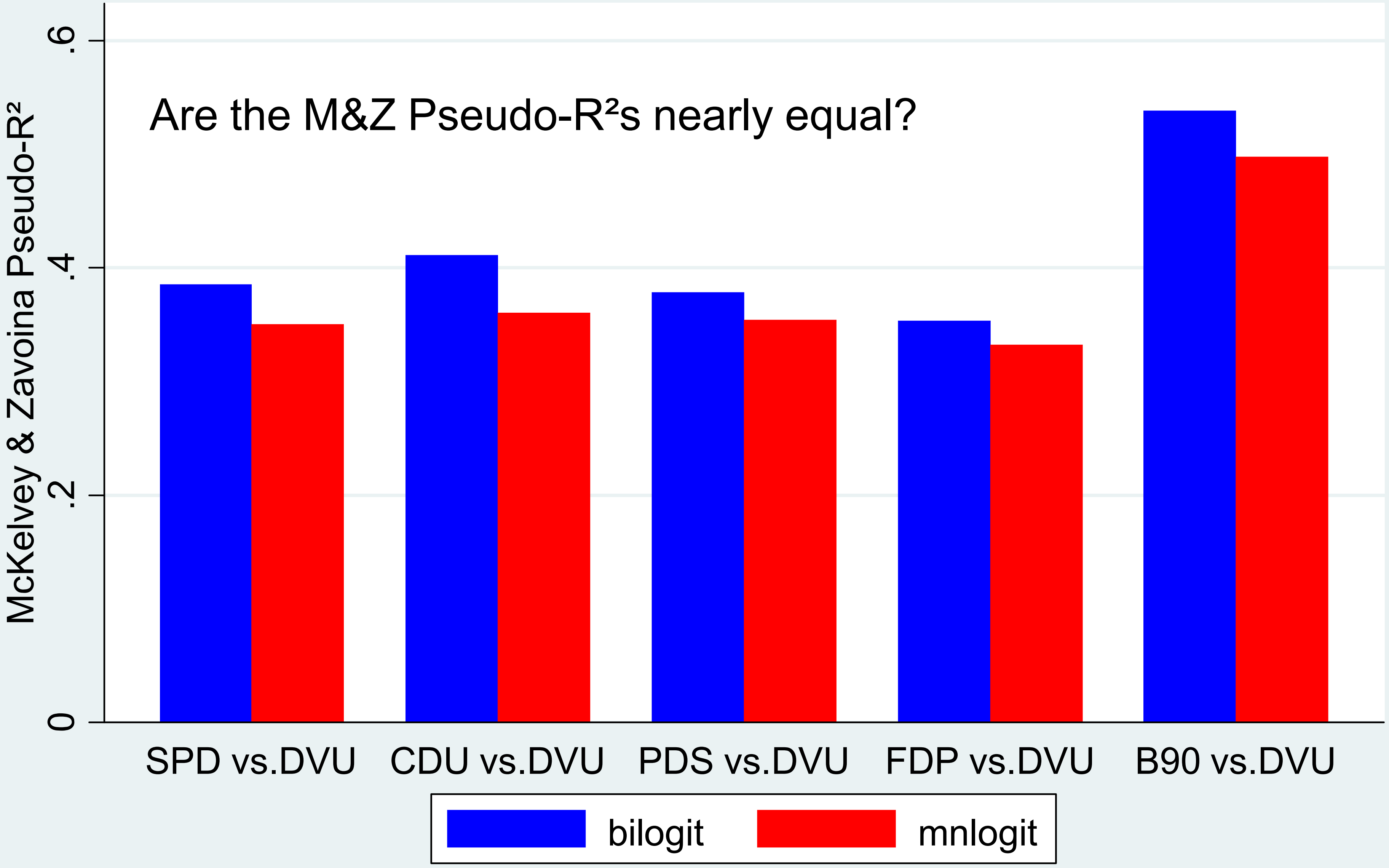


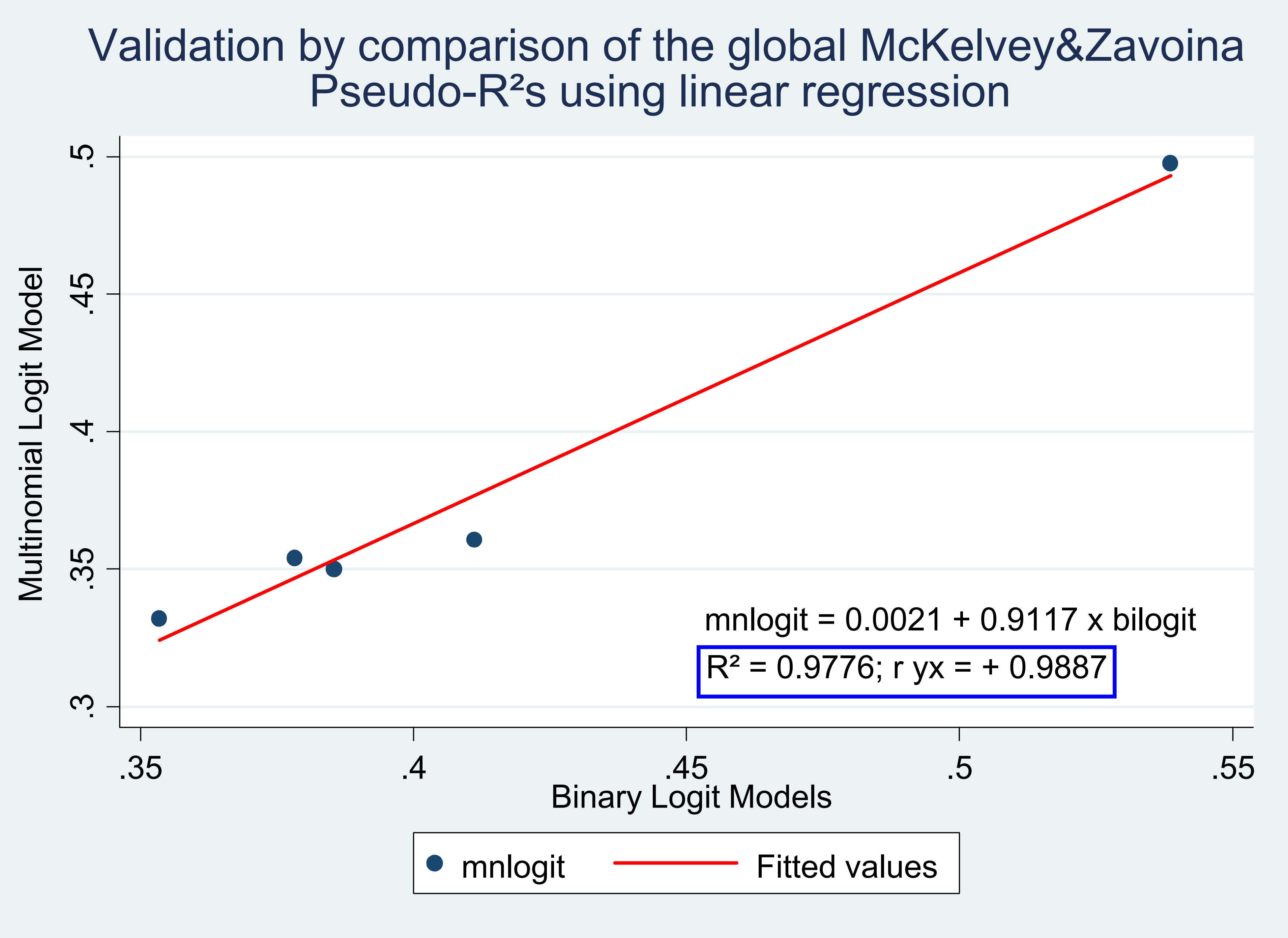
```
. mnl_mzr2
Separate McKelvey Zavoina pseudo R2 for mlogit equations

   Equation |        R2
------------+-----------
        spd |    0.3501
        cdu |    0.3607
        pds |    0.3540
        fdp |    0.3322
        b90 |    0.4978
 dvu,rep,~d |    0.0000
------------------------
```

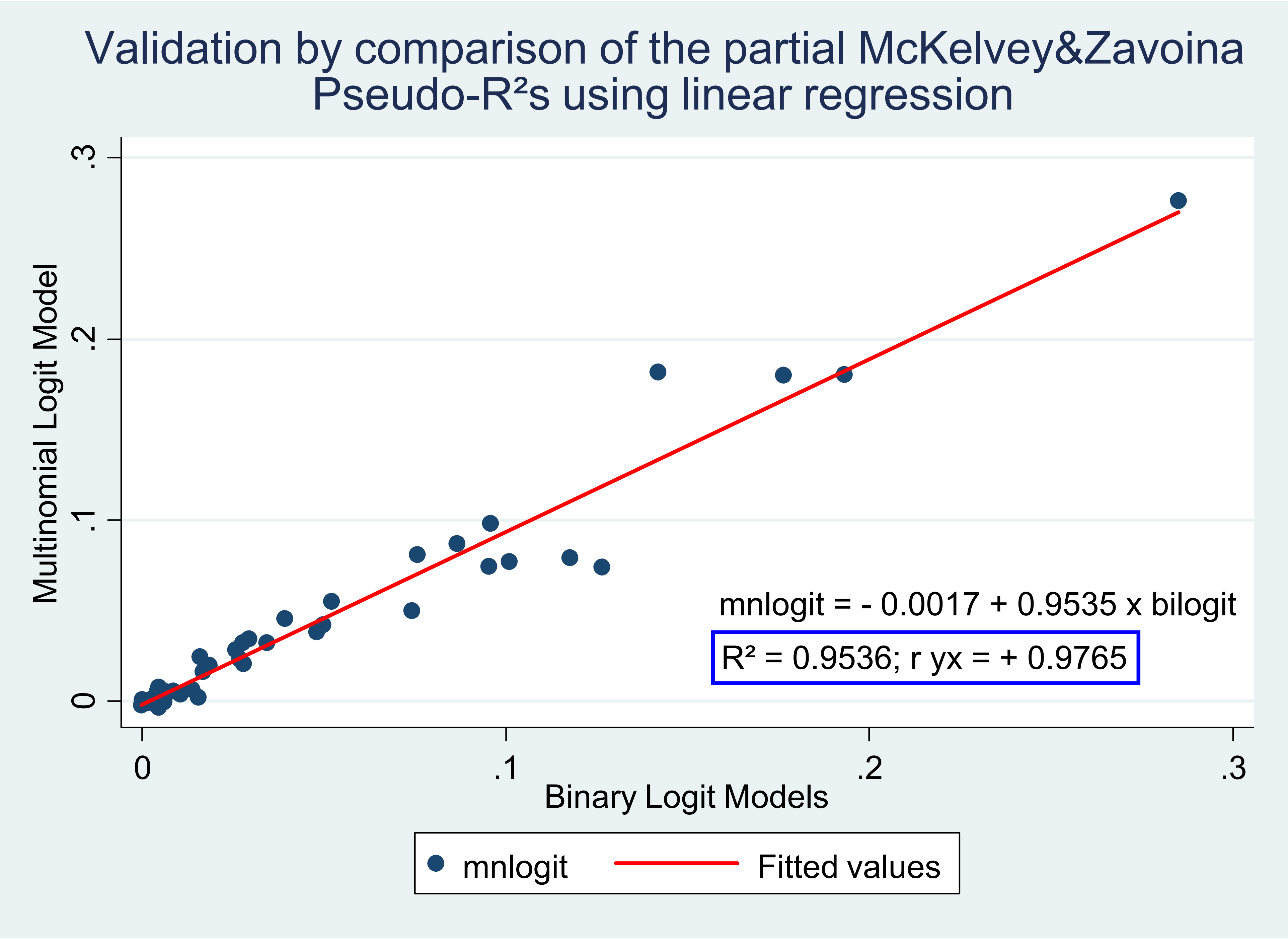
Indicating quite a good fit for the comparison of each established party with the right-wing extremist ones. Explained variance of the estimated logits lies between 33% and 50%.

This table presents the best fit of all possible base outcome categories of voting!

Validation by comparison of the overall fit of the multinominal and binary logit models




Validation by comparison of the global McKelvey&Zavoina Pseudo-R²s using linear regression
Multinomial Logit Model
.3
.35
.4
.45
.5
.35
.4
.45
.5
.55
Binary Logit Models
mnlogit = 0.0021 + 0.9117 x bilogit
R² = 0.9776; r yx = + 0.9887
● mnlogit
Fitted values



Validation by comparison of the partial McKelvey&Zavoina Pseudo-R²s using linear regression

5. Conclusions

- Known
  - The Monte-Carlo-simulation studies show that the McKelvey&Zavoina Pseudo-R² is the best fit measure for binary and ordinal logit models
- New
  - Generalization of the M&Z-Pseudo-R² to the multinomial logit model to identify its differential fit for its k-1 binary comparisons
  - Successful validation of these global and partial M&Z-Pseudo-R²s by those of the corresponding binary logit models
- That’s why
  - I suggest to use my mnl_mzr2.ado file to assess the differential fit of the multinomial logit model

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