GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Dr. Christian Czymara

FORSCHUNGSPRAKTIKUM I UND II: LÄNGSSCHNITTDATENANALYSE IN R

Logistic Fixed and Random Effects models
session ix

AGENDA

- Logistic FE and RE models
- Linear probability FE

RECAP: PROBABILITY MODELS

	Linear Probability Model	Logistic Regression Model
Estimator	Ordinary Least Squares (OLS)	Maximum Likelihood (ML)
Optimization function	Minimize sum of squared residuals	Maximize likelihood function
Solution	Analytical (calculated)	Numerical (iterative "trying")
Effects of explanatory variables	Linear: effect of x is the same for all values/at all levels of x Additive: effect of x is independent from other x	Non-linear: effect of x depends on its value/level Multiplicative: effect of x depends on the values of other x
Predicted values	Can be below 0 or above 1	$[0,1]$
Interpretation of coefficients	Percentage point change in probability of $y = 1$	Absolute change in logged odds (logits) of $y = 1$ *or* relative change in odds ratios of $y = 1$

LOGISTIC POOLED MODELS

LINEAR VS LOGISTIC POOLED

	Linear	Logistic
Properties	Unbiased (if assumptions met)	Consistent (if assumptions met)
Estimator	OLS	ML
u_i	Ignored	Ignored
e_{it}	Distribution assumed $\sim N(0, \sigma_e^2)$	Distribution assumed $\sim logistic(0, \frac{\pi^2}{3})$

- The same as linear & logistic regression for cross-sectional data → idea of pooled models is to ignore the panel structure

LOGISTIC FIXED EFFECTS

FIXED EFFECTS

- In linear case: eliminate everything time-constant (observed z and unobserved u_i)
  - Time-demeaning (Fixed Effects transformation): remove unit-specific means
  - Dummy for each unit
- Based on within-unit variation over time
- FE ≙ conditional estimates (OLS: given unit-specific means)

LOGISTIC REGRESSION FOR PANEL DATA

- $\ln\left(\frac{\mathrm{pr}(y=1|x)}{1-\mathrm{pr}(y=1|x)}\right) = u_i + \beta_1 x_1 + \cdots + \beta_k x_k$
- u_i: unobserved heterogeneity
- Time-constant variables z part of u_i
- Problem: u_i not easily eliminated from likelihood function
- Unit dummies no alternative (*incidental parameters problem*: ML consistent, based on $n \to \infty$)

CONDITIONAL MAXIMUM LIKELIHOOD

- In linear case: FE based on transforming data
- In logistic case: also conditional estimation

→Maximize probability of observing a specific sequence conditional on certain number of 1s

- Control frequency of 1s means to control for average level of y
- Sequences with no change in y do not contribute to conditional likelihood ($pr = 1$)

→Conditional Maximum Likelihood (CML) only based on within variance, similar to linear FE

EXAMPLE: UNION MEMBERSHIP

- Andreß et al. (2014): page 231
- $t = 2$
- Union membership = dummy (1: yes, 0: no)
- Possible sequences
  - No member in both years: 00 ($n = 386$)
  - Became member: 01 ($n = 22$)
  - Left union: 10 ($n = 37$)
  - Member in both years: 11 ($n = 100$)

EXAMPLE: UNION MEMBERSHIP

Member at $t = 1$	Member at $t = 2$		Total
	0	1	
0	386	22	408
1	37	100	137
Total	423	122	545

Sequences	Number of 1s	Frequency	Conditional probability	#outcomes/ #sample
00	0	386	$\Pr(00 \mid 1s = 0)$	$386/386 = 1$
11	2	100	$\Pr(11 \mid 1s = 2)$	$100/100 = 1$
01	1	22	$\Pr(01 \mid 1s = 1)$	$22/(22 + 37) = 0.37$
10	1	37	$\Pr(10 \mid 1s = 1)$	$37/(22 + 37) = 0.63$

EXAMPLE

- $Pr(01 \mid 1s = 1)$

$$= \frac{pr(y_{i1} = 0)\, and\, pr(y_{i2} = 1)}{\left(pr(y_{i1} = 0)\, and\, pr(y_{i2} = 1)\right)\, or\, \left(pr(y_{i1} = 1)\, and\ \, pr(y_{i2} = 0)\right)}$$

$$= \frac{e^{u_i + \beta_1 x_1}}{1 + e^{u_i + \beta_1 x_1}}$$

- I spare you the math (see Andreß et al. (2014): 229 ff.)
- For $t = 2$ CML can use traditional ML
  - Outcome: sequence 01 vs. sequence 10
  - Explanatory variables: First differences of x
- When $t > 2$ traditional ML cannot be used

EXAMPLE FROM LANCEE & PARDOS-PRADO (2013)

	M2 Fixed effects
Female	
Age	
Marital status	
Married	Ref.
Single	0.891 (0.101)
Divorced/separated/widowed	1.097 (0.128)
Educational attainment	
Inadequately/general elementary	0.865 (0.234)
Basic vocational	0.974 (0.176)
Intermediate vocational/general	Ref.
General/vocational maturity	0.946 (0.139)
Tertiary education	0.763 (0.138)
Year and federal state dummies	Yes
Unemployment rate	
Proportion foreigners	
Constant	
Log-likelihood	−15531.6
N observations	40,359
N subjects	6,181

- Interpretation: Exactly like traditional logistic regression
- *Single* ($e^{\beta} = 0.891$)
  - Negative relationship: Singles are less concerned compared to married individuals
  - Odds of being concerned lower by a factor of 0.891 (or $1 - 0.891 = 10.9\%$)
- *Divorced* ($e^{\beta} = 1.097$)
  - Positive relationship: Divorced are more concerned compared to married individuals
  - Odds of being concerned higher by a factor of 1.097 (or 9.7%)

- Logistic FE predicting probability to be very concerned about immigration, odds ratios (standard errors)
- Note: Only part of actual model shown here. For full model see Lancee & Pardos-Prado (2013): 121 f.

LOGISTIC REGRESSION COEFFICIENTS

	Interpretation
β_0	logged odds of $y = 1$ (instead of $y = 0$) when $x = 0$
e^{β_0}	odds of $y = 1$ (instead of $y = 0$) when $x = 0$
β_1	change in logged odds of $y = 1$ (instead of $y = 0$) for an increase in x by one unit
e^{β_1}	change in odds of $y = 1$ (instead of $y = 0$) for an increase in x by one unit → odds ratio

SUMMARY

- Logistic Fixed Effects estimated with CML
- Conditional probability of certain sequence given all sequences that include the same number of 1s
- When $t = 2$ similar to First Difference estimation for continuous y
  - $y = pr(01\ vs\ 10)$
  - $x =$ first differences

LINEAR VS LOGISTIC FIXED EFFECTS

	Linear	Logistic
Conditional estimation	… given unit-specific means	… given certain number of 1s
Units without change	Time-demeaning (Fixed Effects Transformation) → Units without change of y do not contribute to within variance	Sequences without change → Do not contribute to conditional likelihood function

LINEAR VS LOGISTIC FE

	Linear	Logistic
Properties	Unbiased (if (fewer) assumptions met)	Consistent (if (fewer) assumptions met)
Estimator	OLS (demeaned) or LSDV	Conditional ML
u_i	Estimated (LSDV) or controlled (through demeaning)	Controlled
e_{it}	Distribution assumed $\sim N(0, \sigma_e^2)$	Distribution assumed $\sim logistic(0, \frac{\pi^2}{3})$

LOGISTIC RANDOM EFFECTS

ASSUMPTIONS

	Pooled	RE	FE
Omitted Variable Bias			
Not in e_{it} (strict exogeneity): $cov(e_{it}, x) = 0$	✓	✓	✓
Not in u_i (RE assumption): $cov(u_i, x) = 0$	✓	✓	✗
Serial correlation			
Not in e_{it}: $cov(e_{it}, e_{is}) = 0$	✓	✓	✓
Not in ε_{it}: $corr(\varepsilon_{it}, \varepsilon_{is}) = var(u_i) = 0$	✓	✗	✗

RANDOM EFFECTS

- As in linear case: RE need extra assumption compared to FE
- Issue 1: u_i independent of x *and* z variables (model correctly specified)
- Issue 2: Potential serial correlation (biased standard errors)

SOLUTION TO MISSPECIFICATION DUE TO z IN LOGISTIC RE MODELS

- None (like linear case)
- If this is an issue, use FE

SOLUTION TO SERIAL CORRELATION DUE TO u_i IN LOGISTIC RE MODELS

- Make assumption about distribution of u_i
- E. g.: normally distributed, $u_i \sim N(0, \sigma_u^2)$
- Estimate serial correlation: $corr(\varepsilon_{it}, \varepsilon_{is}) = \rho$ (rho)
- ρ: Fraction of total error variance due to u_i (ICC)
- σ_e^2 in logistic models fixed to $\frac{\pi^2}{3}$

$\rightarrow \rho = \frac{\sigma_u^2}{\sigma_u^2 + \frac{\pi^2}{3}}$

TESTING RE ASSUMPTIONS IN LINEAR AND LOGISTIC REGRESSION

	Assumption	Linear	Logistic
FE vs RE	Omitted variable bias of z?	Hausman Test H_0: $FE = RE$	Hausman Test H_0: $FE = RE$
RE vs pooled	Serial correlation?	Breusch-Pagan Test H_0: $\sigma_u^2 = 0$	Likelihood Ratio Test H_0: $\rho = 0$

- Test for omitted variable bias follows same logic in both cases
- Test for serial correlation
  - Breusch Pagan Test
  - Likelihood Ratio Test: Compare RE with pooled model (pooled model nested in RE with restriction $\sigma_u^2 = 0$)

		Idiosyncratic error: e_{it}	Unobserved heterogeneity: u_i	Extra assumptions
Linear models	Pooled	Assumed $\sim N(0, \sigma_u^2)$	-	• $Cov(u_i, z_i) = 0$ / $Cov(u_i, x_{it}) = 0$ • $Cov(e_{it}, e_{is}) = 0$
	RE	Variance estimated (Feasible Generalized Least Squares: FGLS)	Variance estimated (FGLS)	$Cov(u_i, z_i) = 0$/ $Cov(u_i, x_{it}) = 0$
	FE	Assumed $\sim N(0, \sigma_u^2)$	Estimated (LSDV) or controlled (FE)	-
Logistic models	Pooled	Assumed $\sim logistic(0, \frac{\pi^2}{3})$	-	• $Cov(u_i, z_i) = 0$ / $Cov(u_i, x_{it}) = 0$ • $Cov(e_{it}, e_{is}) = 0$
	RE		Assumed $\sim N(0, \sigma_u^2)$	• $Cov(u_i, z_i) =$ 0 / $Cov(u_i, x_{it}) = 0$
	FE		Controlled	-

FE VS RE: EXAMPLE FROM LANCEE & PARDOS-PRADO (2013)

	M1 Random effects		M2 Fixed effects	
Female	0.825***	(0.041)		
Age	1.005*	(0.002)		
Marital status				
Married	Ref.		Ref.	
Single	0.883	(0.089)	0.891	(0.101)
Divorced/separated/widowed	1.076	(0.068)	1.097	(0.128)
Educational attainment				
Inadequately/general elementary	1.541***	(0.181)	0.865	(0.234)
Basic vocational	1.309***	(0.076)	0.974	(0.176)
Intermediate vocational/general	Ref.		Ref.	
General/vocational maturity	0.392***	(0.035)	0.946	(0.139)
Tertiary education	0.219***	(0.016)	0.763	(0.138)
Year and federal state dummies	Yes		Yes	
Unemployment rate				
Proportion foreigners				
Constant	4.658***	(0.142)		
Log-likelihood	−37285.3		−15531.6	
N observations	77,691		40,359	
N subjects	15,694		6,181	

- Interpretation still the same as traditional logistic regression
- RE estimate effects of variables *Female* and *Age*, which are constant across respondents
- Standard errors larger for FE then for RE
- → FE only based on within variation → reduced sample size (like linear case)

- Logistic FE predicting probability to be very concerned about immigration, odds ratios (standard errors)
- Note: Only part of actual model shown here. For full model see Lancee & Pardos-Prado (2013): 121 f.

FE VS RE: EXAMPLE FROM LANCEE & PARDOS-PRADO (2013)

	M1 Random effects	M2 Fixed effects
Class		
High service	0.823** (0.055)	0.863* (0.051)
Low service	Ref.	Ref.
Routine non-manual	1.024 (0.092)	0.958 (0.104)
Routine, service sales	1.092 (0.097)	0.899 (0.068)
Self-employed	0.964 (0.049)	0.929 (0.131)
Self-employed no Employment	1.001 (0.067)	1.085 (0.125)
Skilled manual	1.542*** (0.118)	1.141 (0.117)
Semi/unskilled manual	1.408*** (0.088)	0.969 (0.068)
Farm labor	1.173 (0.243)	1.118 (0.264)
Self-employed Farm	0.683 (0.192)	1.303 (0.431)
Not working/Unemployed	1.502*** (0.147)	1.318*** (0.110)
Not working/Pensioner	1.926 (0.674)	1.384 (0.463)
Difficulty to find a new job if current one is lost	1.284** (0.104)	1.096 (0.057)
Interested in politics	0.974 (0.044)	1.292** (0.119)
Attending church or religious events	0.875** (0.041)	0.931 (0.060)
General life satisfaction	0.378*** (0.022)	0.658** (0.086)
Friends that are not from Germany	0.530*** (0.091)	0.698* (0.110)

- Most effects of class not statistically significant with FE
- Something time-constant seems to confound the effect of class on concerns
- Effect on unemployment stable
- → Unemployment matters for everyone

- Logistic FE predicting probability to be very concerned about immigration, odds ratios (standard errors)
- Note: Only part of actual model shown here. For full model see Lancee & Pardos-Prado (2013): 121 f.

LINEAR AND LOGISTIC PANEL REGRESSION

	Linear	Logistic
Standard errors: pooled < RE < FE	✓	✓
Estimates: RE always between FE and BE	✓	✗
Hybrid models replicate FE and BE exactly	✓ (FE and BE uncorrelated by design)	✗ (only similar due to non-linearity)

SUMMARY: UNOBSERVED HETEROGENEITY

- FE: Fixed value
  - Linear model: Can be estimated (as dummies)
  - Logistic model: Cannot be estimated, but controlled
- RE: Random variable
  - Single value not interesting, but variation
  - ML (linear or logistic): Needs assumption about distribution of u_i
  - GLS (linear): Quasi-demeaning, no assumption about distribution of u_i neccessary
  - Assumption: u_i uncorrelated with x *and* z

LINEAR VS LOGISTIC RE

	Linear	Logistic
Properties	Unbiased (if assumptions met)	Consistent (if assumptions met)
Estimator	Generalized Least Squares (quasi-demeaning) or Maximum Likelihood (random intercepts)	ML
u_i	Distribution assumed $\sim N(0, \sigma_u^2)$	Distribution (often) assumed $\sim logistic(0, \sigma_u^2)$
e_{it}	Distribution assumed $\sim N(0, \sigma_e^2)$	Distribution assumed $\sim logistic(0, \frac{\pi^2}{3})$

LINEAR PROBABILITY MODEL FOR PANEL DATA

LINEAR PROBABILITY MODEL

- Use linear FE for dummy outcome variable to get linear probability model (as in cross-sectional case)
- Circumvents many of the issues related to the use of non-linear models
- But also same problems as in linear case (potentially predicting probabilities below 0 or above 1; functional form misspecification)

EXAMPLE FROM CZYMARA & DOCHOW (2018)

	(1) Main model
Media salience, past 21 days	0.050****
	(0.001)
Party preference (ref.: no preference)	
CDU/CSU (Christian Democrats)	0.027****
	(0.004)
SPD (Social Democrats)	−0.007*
	(0.004)
Die Grünen (The Greens)	−0.012*
	(0.007)
Die Linke (The Left)	−0.005
	(0.008)
FDP (Free Democrats)	0.019**
	(0.009)
Others and mixed	0.015
	(0.010)
Radical right	0.144****
	(0.013)

- Interpretation: Like linear FE but change in probability to be concerned
- Effect of *Media salience* ($\beta = 0.05$)
  - Positive relationship: "*A one unit increase in media salience predicts an increase in the probability of being very concerned by five percentage points.*"

- Note: Only part of actual model shown here. For full model see Czymara & Dochow (2018): 390 f.

HIERARCHICAL LINEAR PROBABILITY MODELS

- Main benefits
  - Easy interpretation of effects
  - Meaningful (changes of) variance components
  - Comparability of coefficients between models
  - Fast estimation
- Main drawbacks
  - Predicted probabilities can be < 0 or > 1
  - Effects do not depend on the level of x (relevant when effects for "extreme" observations are calculated)

EXTENSIONS OF LOGISTIC PANEL REGRESSION

OTHER NON-LINEAR MODELS

- Many interesting outcome variables are not continuous
- E. g. attitudes measured on 3-point scale (agree, neutral, disagree)
- Multinomial Logit Model with Fixed Effects for *nominal* y
- Fixed Effects Ordered Logit Regression for *ordinal* y
- Interpretation analogous to traditional multinominal or ordered logistic regression

SIMPLER WAY TO DEAL WITH CATEGORICAL OUTCOMES

- With few categories
  - Collapse variable into dummy (e. g.: 0: “don‘t agree” & ”partly agree” vs. 1: “totally agree“)
  - Drawback: throwing away information
- With “many” categories
  - Treat as continuous (i. e., run simple linear regression)
  - Drawbacks:
  - Predictions can be out of the outcome’s possible bounds
  - OLS assumptions likely violated → OLS not BLUE

LITERATURE

- Chapter 5.1 (pages 157 ff.) in: Andreß, Golsch & Schmidt (2014). Applied panel data analysis for economic and social surveys. Springer Science & Business Media

- Study applying logistic fixed and random effects models: *Lancee & Pardos-Prado (2013). Group conflict theory in a longitudinal perspective: Analyzing the dynamic side of ethnic competition. International Migration Review, 47(1), 106-131*