STATISTICAL HORIZONS

Longitudinal Data Analysis Using R

Stephen Vaisey, Ph.D.

Upcoming Seminar:

April 8-10, 2021, Remote Seminar

Outline

1. Opportunities and challenges of panel data
2. Linear models
3. Logistic regression models
4. Count data models
5. Linear structural equation models

3

3

Section 1

OPPORTUNITIES AND CHALLENGES

4

4

Panel data

Data in which variables are measured at multiple points in time for the same individuals.

Response variable y_{it} with t = 1, 2,…, T

Vector of predictor variables x_{it} .

Some of these may vary with time, others may not.

Assume, for now, that time points are the same for everyone in the sample. (For some methods that assumption is not essential).

5

5

Why are panel data desirable?

In *Econometric Analysis of Panel Data* (2008), Baltagi lists six potential benefits of panel data:

1. Ability to control for individual heterogeneity
2. More informative data
3. Better ability to study the dynamics of adjustment
4. Ability to identify and measure effects not detectable in cross-sections
5. Ability to test more complicated behavioral models
6. Avoidance of aggregation bias

6

6

My list

1. Ability to control for unobservables
2. Ability to investigate causal ordering (sometimes!)
3. Ability to study the effect of a treatment on the trajectory of an outcome

Problems with panel data

- Attrition and missing data
- Dependence among multiple observations from same individual
  – observations are correlated; individuals have tendencies
  – conventional methods assume independent observations
  – for this reason, estimated SEs tend to be too low

Estimating mixed models in R

- For mixed models that assume exchangeability, `lme4::lmer` is probably best and easiest to use (and has nice plotting tools)
- For models that relax the exchangeability assumption, `nlme::lme` is required
- Both can do random slopes/intercepts
- Both can do maximum likelihood
- Can get robust SEs much more easily after `nlme::lme`

49

Example `lme4` syntax

```
mix.ri <- lmer(anti ~ self + pov + black + hispanic + childage + married +
                  gender + momage + momwork + wave + (1 | id),
                data = nlsy_long,
                REML = FALSE)
```

specify the random intercept as a term in the formula, grouped by the cluster variable

this means use ML instead of "restricted maximum likelihood." REML can be better in small samples, but doesn't allow model comparisons based on different specifications of the predictors

50

```
> summary(mix.ri)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: anti ~ self + pov + black + hispanic + childage + married +
 gender +
    momage + momwork + wave + (1 | id)
   Data: nlsy_long

     AIC      BIC   logLik deviance df.resid
  5882.4   5958.9  -2927.2   5854.4     1729

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.5962 -0.5637 -0.1141  0.5087  3.3422

Random effects:
 Groups   Name        Variance Std.Dev.
 id       (Intercept) 1.2827   1.1326
 Residual             0.9929   0.9964
Number of obs: 1743, groups:  id, 581

Fixed effects:
             Estimate Std. Error t value
(Intercept)  2.531431   1.089759   2.323
self        -0.062076   0.009487  -6.543
pov          0.247138   0.080136   3.084
black        0.226754   0.125000   1.814
hispanic    -0.218209   0.137456  -1.587
childage     0.088456   0.090583   0.977
married     -0.049565   0.125717  -0.394
gender      -0.483449   0.105925  -4.564
momage      -0.021920   0.025147  -0.872
momwork      0.261132   0.114058   2.289
wave2        0.047340   0.058530   0.809
wave3        0.216381   0.058702   3.686
```

ICC = 1.2827/(.9929 + 1.2827) = .564

the same as the value of ρ in the GLS-ML with exchangeable residuals

51

51

	generalized least squares (1)	linear mixed-effects (2)
self	-.062***	-.062***
	(.010)	(.009)
pov	.247***	.247***
	(.080)	(.080)
black	.227*	.227*
	(.125)	(.125)
hispanic	-.218	-.218
	(.138)	(.137)
childage	.088	.088
	(.091)	(.091)
married	-.050	-.050
	(.126)	(.126)
gender	-.483***	-.483***
	(.106)	(.106)
momage	-.022	-.022
	(.025)	(.025)
momwork	.261**	.261**
	(.114)	(.114)
wave2	.047	.047
	(.059)	(.059)
wave3	.216***	.216***
	(.059)	(.059)
Constant	2.531**	2.531**
	(1.094)	(1.090)
Observations	1,743	1,743
Log Likelihood	-2,927.199	-2,927.199
Akaike Inf. Crit.	5,882.398	5,882.398
Bayesian Inf. Crit.	5,958.885	5,958.885

The two models are the same!

(other than a few tiny rounding differences in the reported SEs)

52

52

Software note

If you are looking for a robust approach to estimating these models using `lme4`, check out:

Koller, M. (2016). robustlmm: An R Package for Robust Estimation of Linear Mixed-Effects Models. *Journal of Statistical Software*, *75*(6). https://doi.org/10.18637/jss.v075.i06

Wang, T., & Merkle, E. C. (2016). Derivative Computations and Robust Standard Errors for Linear Mixed Effects Models in lme4. *ArXiv:1612.04911 [Stat]*. Retrieved from http://arxiv.org/abs/1612.04911

You can use `clubSandwich` after `nlme::lme`.

53

53

Random Coefficients

We can allow for random coefficients for the time-varying predictors by writing

$$y_{it} = \mu_t + \beta_i x_{it} + \gamma z_i + \alpha_i + \varepsilon_{it}$$

where the β coefficients now have a subscript *i*. In practice, we might only do this for a subset (or just one of) the variables. We assume that this β is a normally distributed random variable (with a mean and a variance that we'll estimate) that is uncorrelated with everything else.

54

54

Recap: GEE vs. RE-ML for logistic models

Why prefer GEE?

- GEE is much faster
- GEE can allow for departures from exchangeability

Why prefer RE-ML?

- Subject-specific coefficients
- Weaker assumptions for missing data (MAR vs. MCAR)
- Random coefficients
- More than two levels

For both methods, robust SEs are preferable, but this is hard in `lme4`.

117

117

Solution 4: Fixed effects models

The fixed effects model has the same general form as the random effects model

$$\log\left(\frac{p_{it}}{1-p_{it}}\right)=\mu_t+\beta x_{it}+\gamma z_i+\alpha_i$$

However, α_i is now regarded as a set of fixed constants rather than as random variables. This allows for any correlation between α_i, z_i and x_{it}.

118

118

For the linear FE model, we could handle α_i by including dummy variables for each person. For logistic models, that produces estimates biased away from 0, sometimes severely (unless T is large).

The preferred method is conditional likelihood, which conditions on the number of 1's and 0's for each person. In effect, we are asking: "Given that a girl is in poverty for two out of five years, **why did poverty occur in years 2 and 4, rather than in years 1, 3, or 5**?"

Clearly, if a girl is in (or out of) poverty all five years, there is nothing to explain. So girls with these response patterns are effectively eliminated from the analysis.

119

119

Conditional logit with `survival::clogit`

```
clogit(pov ~ mother + spouse + inschool + hours + wave + strata(id),
       data = teenpov_long)
```

I omitted time-constant variables before specifying the model

`strata()` is where you specify the index for individuals

120

120

```
Call:
coxph(formula = Surv(rep(1, 5755L), pov) ~ mother + spouse +
    inschool + hours + wave + strata(id), data = teenpov_long,
    method = "exact")

  n= 5755, number of events= 2169

             coef exp(coef) se(coef)      z Pr(>|z|)
mother    0.58243   1.79039  0.15958  3.650 0.000263 ***
spouse   -0.74776   0.47343  0.17535 -4.264 2.00e-05 ***
inschool  0.27187   1.31241  0.11273  2.412 0.015883 *
hours    -0.01965   0.98055  0.00315 -6.236 4.49e-10 ***
wave2     0.33178   1.39345  0.10156  3.267 0.001088 **
wave3     0.33498   1.39791  0.10825  3.094 0.001971 **
wave4     0.43277   1.54151  0.11651  3.714 0.000204 ***
wave5     0.40250   1.49556  0.12753  3.156 0.001598 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

         exp(coef) exp(-coef) lower .95 upper .95
mother      1.7904     0.5585    1.3095    2.4478
spouse      0.4734     2.1123    0.3357    0.6676
inschool    1.3124     0.7620    1.0522    1.6369
hours       0.9805     1.0198    0.9745    0.9866
wave2       1.3934     0.7176    1.1419    1.7004
wave3       1.3979     0.7154    1.1307    1.7283
wave4       1.5415     0.6487    1.2268    1.9370
wave5       1.4956     0.6686    1.1648    1.9202

Rsquare= 0.017   (max possible= 0.42 )
Likelihood ratio test= 97.28  on 8 df,   p=<2e-16
Wald test            = 90.56  on 8 df,   p=4e-16
Score (logrank) test = 94.58  on 8 df,   p=<2e-16
```

121

121

	RE Logit	Conditional Logit
age	-.063	
	(.047)	
black	.609***	
	(.098)	
mother	1.010***	.582***
	(.118)	(.160)
spouse	-1.172***	-.748***
	(.151)	(.175)
inschool	-.115	.272*
	(.099)	(.113)
hours	-.026***	-.020***
	(.003)	(.003)
wave2	.283**	.332**
	(.100)	(.102)
wave3	.213*	.335**
	(.104)	(.108)
wave4	.242*	.433***
	(.109)	(.117)
wave5	.145	.403**
	(.116)	(.128)
Constant	-.005	
	(.762)	
Observations	5,755	5,755

No results are reported for AGE and BLACK. These time-invariant predictors cannot explain why poverty occurred in some years but not in others.

Compared with RE estimates, the coefficients for MOTHER, SPOUSE and HOURS are much smaller in magnitude, with higher standard errors. The INSCHOOL coefficient has actually changed sign and is now statistically significant.

It's common for FE results to be substantially different from RE results, because FE controls for all stable characteristics of the individuals.

122

122

Adding TV × TC interactions

```
> teenpov_long$age0 <- teenpov_long$age-14 # recode so youngest age (14) is 0
> clogit2 <- clogit(pov ~ mother + spouse + inschool + hours + wave +
+                   mother:age0 + strata(id),
+                 data = teenpov_long)
> summary(clogit2)
Call:
coxph(formula = Surv(rep(1, 5755L), pov) ~ mother + spouse +
    inschool + hours + wave + mother:age0 + strata(id), data = teenpov_long,
    method = "exact")

  n= 5755, number of events= 2169

                 coef exp(coef)  se(coef)      z Pr(>|z|)
mother       1.084243  2.957200  0.281584  3.851 0.000118 ***
spouse      -0.744261  0.475085  0.175755 -4.235 2.29e-05 ***
inschool     0.284232  1.328741  0.113072  2.514 0.011946 *
hours       -0.019745  0.980449  0.003153 -6.262 3.80e-10 ***
wave2        0.338248  1.402489  0.101524  3.332 0.000863 ***
wave3        0.345848  1.413188  0.108348  3.192 0.001413 **
wave4        0.441904  1.555666  0.116678  3.787 0.000152 ***
wave5        0.409744  1.506432  0.127652  3.210 0.001328 **
mother:age0 -0.293634  0.745549  0.133896 -2.193 0.028307 *
```

The poverty risk of motherhood declines with age

123

123

Considerations for conditional logit

Like the RE model, the FE model assumes exchangeability: correlations among the response variables at different times don't depend on how close or far apart in time they are. Because that assumption is not realistic, it would be preferable to use robust standard errors, which *don't* assume exchangeability. Unfortunately, `survival::clogit` doesn't offer that option.

124

124