Brief overview of Longitudinal data analysis

ehsan.karim@ubc.ca

Nov 6th, 2020

SPPH 504/007

Ref

- Reference for reading

[BOOK] **Extending** the **linear model** with **R**: **generalized linear**, **mixed effects** and **nonparametric regression models**

JJ Faraway - 2016 - content.taylorfrancis.com

Using the VitalSource® ebook Access to the VitalBookTM ebook accompanying this book is via VitalSource® Bookshelf – an ebook reader which allows you to make and share notes and highlights on your ebooks and search across all of the ebooks that you hold on your VitalSource …

☆ 99 Cited by 2603 Related articles All 12 versions Import into BibTeX

[BOOK] **A handbook** of **statistical analyses using R**

T Hothorn, BS Everitt - 2014 - taylorfrancis.com

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the …

☆ 99 Cited by 957 Related articles All 18 versions Import into BibTeX »

Content

- Longitudinal Data example
- Wide form vs long form data
- Review of linear model
- Mixed effect
  - Random intercept
  - Random slope
- Compare models
- Model diagnostics
- Mixed effect vs. marginal models (GEE)

Longitudinal Data

- In longitudinal study,
  - the variable of interests (i.e., outcome variable) are continuously measured repeatedly over a period of time.
  - Also known as panel data.
- We will cover very basic ideas without math.
- Vast topic. Would recommend taking courses if interested in this topic:
  - SPPH 501 Analysis of Longitudinal Data from Epidemiological Studies

Example study: 'Beat the Blues'

Clinical trial of an interactive multimedia program.

Data (100 subjects) from

Computerized, interactive, multimedia cognitive-behavioural program for anxiety and depression in general practice

J Proudfoot, D Goldberg, A Mann, B Everitt... - ... **medicine**, 2003 - cambridge.org

Background. **Cognitive-behavioural** therapy (CBT) brings about significant clinical improvement in **anxiety** and **depression**, but therapists are in short supply. We report the first phase of a randomized controlled trial of an interactive **multimedia program** of **cognitive** ...

☆ 99 Cited by 433 Related articles All 10 versions Import into BibTeX

Example study: 'Beat the Blues'

- **Drug (D)**: Anti-depressant drugs (No/Yes).
- **Length (L)**: length of the current episode of depression (<6m / >6m)
- **Treatment (A)**: Exposure group
  - ***TAU*** (treatment as usual) and ***BtheB*** (Beat the Blues)
- **Bdi.pre (B)**: Beck Depression Inventory II before treatment.
- Outcome (Y): **Bdi.2m, bdi.3m, bdi.5m, bdi.8m**: Beck Depression Inventory II after 2/3/5/8 months follow-up.

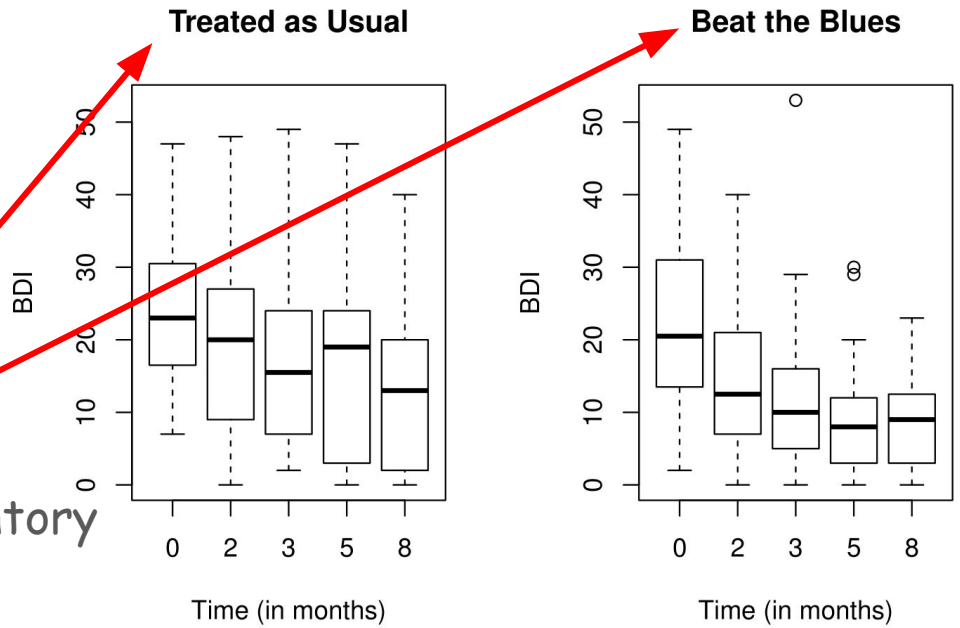
Wide-form data

drug	length	treatment	bdi.pre	bdi.2m	bdi.3m	bdi.5m	bdi.8m
No	>6m	TAU	29	2	2	NA	NA
Yes	>6m	BtheB	32	16	24	17	20
Yes	<6m	TAU	25	20	NA	NA	NA
No	>6m	BtheB	21	17	16	10	9
Yes	>6m	BtheB	26	23	NA	NA	NA
Yes	<6m	BtheB	7	0	0	0	0

- **Times (T)**: month of follow-up (2/3/5/8).

Example study: 'Beat the Blues'

- **Drug (D)**: Anti-depressant drugs (No/Yes).
- **Length (L)**: length of the current episode of depression (<6m / >6m)
- **Treatment (A)**: Exposure group
  - ***TAU*** (treatment as usual) and ***BtheB*** (Beat the Blues)
- **Bdi.pre (B)**: Beck Depression Inventory II before treatment.
- Outcome (Y): **Bdi.2m, bdi.3m, bdi.5m, bdi.8m**: Beck Depression Inventory II after 2/3/5/8 months follow-up.



- **Times (T)**: month of follow-up (2/3/5/8).

Step 1 of analysis: Wide to Long

Reshape data to convert outcome into 1 column

for feeding into software

	drug	length	treatment	bdi.pre	subject	time	bdi
1.2m	No	>6m	TAU	29	1	2	2
2.2m	Yes	>6m	BtheB	32	2	2	16
3.2m	Yes	<6m	TAU	25	3	2	20
4.2m	No	>6m	BtheB	21	4	2	17
5.2m	Yes	>6m	BtheB	26	5	2	23
6.2m	Yes	<6m	BtheB	7	6	2	0

drug	length	treatment	bdi.pre	bdi.2m	bdi.3m	bdi.5m	bdi.8m
No	>6m	TAU	29	2	2	NA	NA
Yes	>6m	BtheB	32	16	24	17	20
Yes	<6m	TAU	25	20	NA	NA	NA
No	>6m	BtheB	21	17	16	10	9
Yes	>6m	BtheB	26	23	NA	NA	NA
Yes	<6m	BtheB	7	0	0	0	0

Total 100 subjects

Long-form data

Example study: 'Beat the Blues'

Closer inspection of data:

- Data from subject 99
- Data from subject 2

4 rows per person!

	drug	length	treatment	bdi.pre	subject	time	bdi
99.2m	No	<6m	TAU	13	99	2	5
99.3m	No	<6m	TAU	13	99	3	5
99.5m	No	<6m	TAU	13	99	5	0
99.8m	No	<6m	TAU	13	99	8	6

	drug	length	treatment	bdi.pre	subject	time	bdi
2.2m	Yes	>6m	BtheB	32	2	2	16
2.3m	Yes	>6m	BtheB	32	2	3	24
2.5m	Yes	>6m	BtheB	32	2	5	17
2.8m	Yes	>6m	BtheB	32	2	8	20

Data analysis - 1

Regular model fixed intercept and slope

$$Y \sim (\text{common alpha}) + (\text{common beta}) \cdot X + \text{error term}$$

Where Y = outcome, X = covariate matrix (includes exposure)

Every subject has a common slope (beta) and common intercept (alpha).

Data analysis - 1

bdi ~ bdi.pre + time + treatment + drug +length + error term

Use lm function

	Est.	S.E.	t val.	p	
(Intercept)	7.32	1.73	4.24	0.00	***
bdi.pre	0.57	0.05	10.44	0.00	***
time	-0.94	0.24	-3.97	0.00	***
treatmentBtheB	-3.32	1.10	-3.02	0.00	**
drugYes	-3.57	1.15	-3.11	0.00	**
length>6m	1.71	1.11	1.54	0.12	

Standard errors: OLS

AIC	1887.49
BIC	1916.57
Pseudo-R^2 (fixed effects)	0.39
Pseudo-R^2 (total)	0.79

Data analysis - 1

In repeated measurement design, outcome measurements are taken repeatedly from each individual. Therefore, for each individual, the measurements are correlated.

Impact on beta or SE or none?

Data analysis - II

Random intercept but fixed slope

Y ~ (common alpha) + (**alpha for subject i**) +

(common beta) . X + error term

= Y ~ ('common alpha' + 'alpha for subject i') +

('common beta') . X + error term

Each subject (source of repeated measurements) has an individual intercept (alpha for i) and common slope (beta).

Data analysis - II

bdi ~ bdi.pre + time + treatment + drug + length

+ (1 | subject) + error term

The interpretation of estimated coefficients will be similar to regular linear model.

Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	5.59	2.24	2.49	108.98	0.01	*
bdi.pre	0.64	0.08	8.21	104.08	0.00	***
time	-0.70	0.15	-4.81	199.32	0.00	***
treatmentBtheB	-2.33	1.67	-1.39	97.12	0.17	
drugYes	-2.82	1.73	-1.64	98.20	0.11	
length>6m	0.20	1.64	0.12	100.26	0.90	

p values calculated using Satterthwaite d.f.

Data analysis - II

bdi ~ bdi.pre + time + treatment + drug + length

+ (1 | subject) + error term

AIC	1887.49
BIC	1916.57
Pseudo-R^2 (fixed effects)	0.39
Pseudo-R^2 (total)	0.79

AIC, BIC, and -loglik etc are goodness-of-fit statistics, which tells you how will the model fits your data.

Data analysis - II

bdi ~ bdi.pre + time + treatment + drug + length

+ (1 | subject) + error term

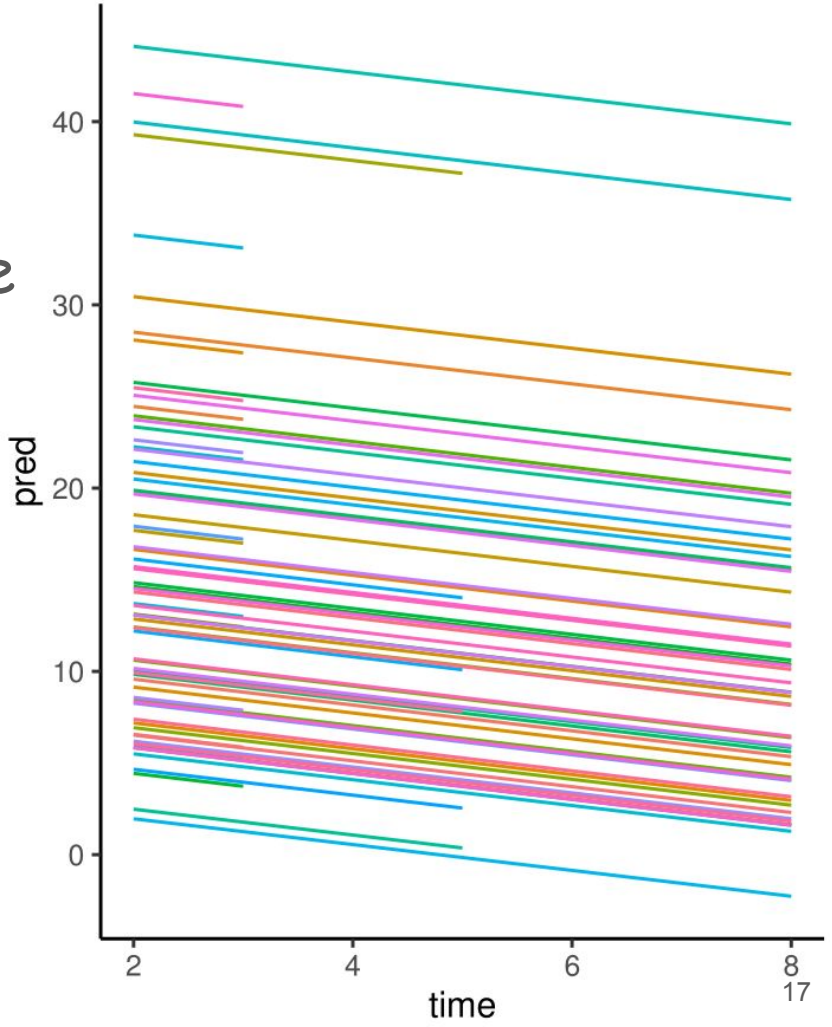
Random Effects		
Group	Parameter	Std. Dev.
subject	(Intercept)	6.98
Residual		5.01

Std.Dev is SD not SE, thus it is not to estimated uncertainty of the estimate. It provides SD across the subjects' mean BDI.

No coefficients for random effects.

Data analysis - II

Random intercept but fixed slope



Data analysis - III

Random slope and random intercept

Y ~ (common alpha) + (alpha for subject i) +

(common beta) . X + (**beta for time j**) . X + error term

= Y ~ ('common alpha' + 'alpha for subject i') +

('common beta' + 'beta for time j') . X + error term

Each subject has an individual intercept and different effects of time on outcome [slope].

Data analysis - III

bdi ~ bdi.pre + time + treatment + drug + length

+ (time | subject)

The interpretation of estimated coefficients will be similar to regular linear model.

Fixed Effects

	Est.	S.E.	t val.	d.f.	p	
(Intercept)	5.61	2.25	2.50	106.79	0.01	*
bdi.pre	0.64	0.08	8.25	102.78	0.00	***
time	-0.70	0.15	-4.56	57.70	0.00	***
treatmentBtheB	-2.38	1.67	-1.42	97.12	0.16	
drugYes	-2.87	1.73	-1.66	98.18	0.10	.
length>6m	0.14	1.64	0.09	100.05	0.93	

p values calculated using Satterthwaite d.f.

Data analysis - III

bdi ~ bdi.pre + time + treatment + drug + length

+ (time | subject)

AIC	1891.04
BIC	1927.39
Pseudo-R^2 (fixed effects)	0.39
Pseudo-R^2 (total)	0.80

AIC, BIC, and -loglik etc are goodness-of-fit statistics, which tells you how will the model fits your data.

Data analysis - III

bdi ~ bdi.pre + time + treatment + drug + length

+ (time | subject)

Random Effects		
Group	Parameter	Std. Dev.
subject	(Intercept)	7.12
subject	time	0.43
Residual		4.90

SD across the subjects' mean BDI.
SD across the times' mean BDI.

No coefficients for random effects.

Data analysis - III

bdi ~ bdi.pre + time + treatment + drug + length

+ (time | subject)

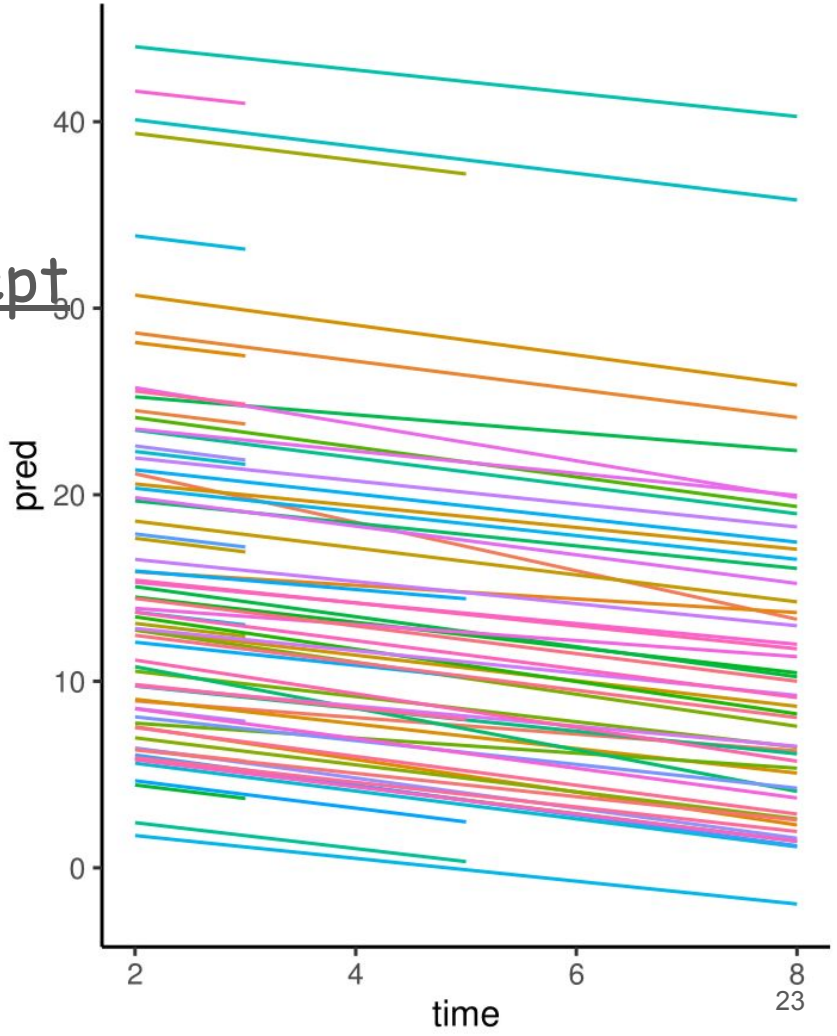
Grouping Variables		
Group	# groups	ICC
subject	97	0.68

ICC: "It describes how strongly units in the same group resemble each other".

3 groups had all missing Y and got excluded from analysis.

Data analysis - II

Random slope and random intercept



Which is better model?

ANOVA: likelihood ratio test as well as AIC/BIC

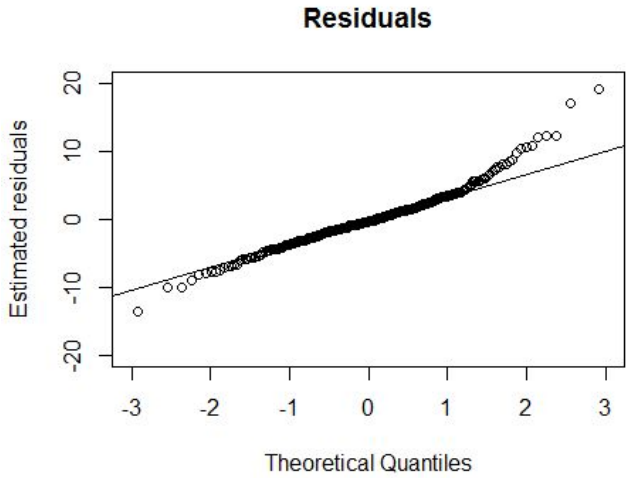
```
## Data: BtheB_long
## Models:
## BtheB_lmer1: bdi ~ bdi.pre + time + treatment + drug + length + (1 | subject)
## BtheB_lmer2: bdi ~ bdi.pre + time + treatment + drug + length + (time | subject)
##              Df    AIC    BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## BtheB_lmer1   8 1887.5 1916.6 -935.75   1871.5
## BtheB_lmer2  10 1891.0 1927.4 -935.52   1871.0 0.4542      2     0.7969
```

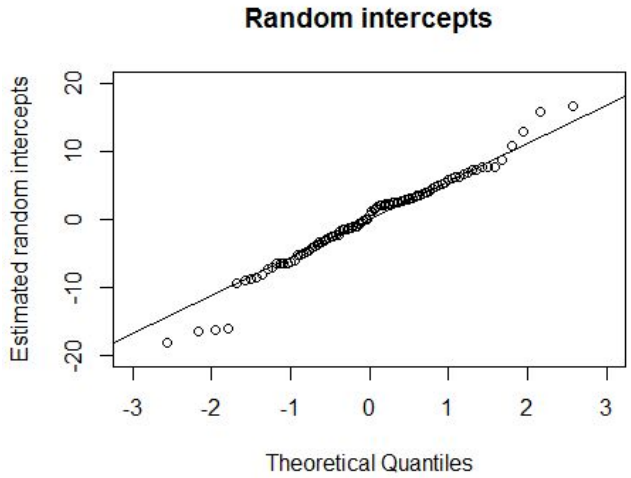
The non-significant p-value shows that the second model (random slope + random intercept) is not statistically different from the first model (random intercept).

Which model is preferable via Anova?

Assumptions

- Normality for error term + beta for subject i
- Predicted values can be used to examine the assumptions we have for linear mixed effect model.





Alternative model: Marginal model

- Mixed effect model is conditional on random effects.
- Alternative is marginal model (population-averaged).
  - Basic idea is treating data as cross-sectional, but incorporate correlation structure.

Alternative model: Marginal model

- Repeated measurement and longitudinal data has responses taken at different time points.
- We could simply review them as many series of cross-sectional data.
- Each sectional data can be analyzed using GLM.
- Only change will be that a correlation structure will be assumed to connect different "cross-sections".

Correlation structure

An identity matrix

$$\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

An exchangeable correlation matrix

$$\begin{pmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{pmatrix}$$

Autoregressive correlation matrix

$$\begin{pmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{pmatrix}$$

Unstructured correlation matrix

$$\begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}$$

GEE

- Generalized estimating equation or GEE is a marginal model.
- The estimated regression coefficients are marginal (or population-averaged) effects,
- The interpretation are at population-level.
- Inference on any specific individual or cluster is not feasible from a GEE.

GEE

bdi ~ bdi.pre + treatment + length + drug

- Cluster = subject
- Outcome follows Gaussian/Normal
- Correlation structure = "exchangeable"

GEE

Independence (as if independent)

exchangeable

```
##                  Estimate Naive S.E.   Naive z Robust S.E.   Robust z
## (Intercept)     3.5686314  1.4833349  2.405816  2.26947617  1.5724472
## bdi.pre         0.5818494  0.0563904 10.318235  0.09156455  6.3545274
## treatmentBtheB -3.2372285  1.1295569 -2.865928  1.77459534 -1.8242066
## length>6m       1.4577182  1.1380277  1.280916  1.48255866  0.9832449
## drugYes        -3.7412982  1.1766321 -3.179667  1.78271179 -2.0986557
##
## Estimated Scale Parameter:  79.25813
## Number of Iterations:  1
##
## Working Correlation
##      [,1] [,2] [,3] [,4]
## [1,]    1    0    0    0
## [2,]    0    1    0    0
## [3,]    0    0    1    0
## [4,]    0    0    0    1
```

```
## Coefficients:
##                  Estimate Naive S.E.     Naive z Robust S.E.   Robust z
## (Intercept)     3.0231602 2.30390185  1.31219140  2.23204410  1.3544357
## bdi.pre         0.6479276 0.08228567  7.87412417  0.08351405  7.7583066
## treatmentBtheB -2.1692863 1.76642861 -1.22806339  1.73614385 -1.2494854
## length>6m      -0.1112910 1.73091679 -0.06429596  1.55092705 -0.0717577
## drugYes        -2.9995608 1.82569913 -1.64296559  1.73155411 -1.7322940
##
## Estimated Scale Parameter:  81.7349
## Number of Iterations:  5
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.6757951 0.6757951 0.6757951
## [2,] 0.6757951 1.0000000 0.6757951 0.6757951
## [3,] 0.6757951 0.6757951 1.0000000 0.6757951
## [4,] 0.6757951 0.6757951 0.6757951 1.0000000
```

GEE

Fixed Effects						
	Est.	S.E.	t val.	d.f.	p	
(Intercept)	5.61	2.25	2.50	106.79	0.01	*
bdi.pre	0.64	0.08	8.25	102.78	0.00	***
time	-0.70	0.15	-4.56	57.70	0.00	***
treatmentBtheB	-2.38	1.67	-1.42	97.12	0.16	
drugYes	-2.87	1.73	-1.66	98.18	0.10	
length>6m	0.14	1.64	0.09	100.05	0.93	

p values calculated using Satterthwaite d.f.

exchangeable

```
## Coefficients:
##                 Estimate Naive S.E.     Naive z Robust S.E.   Robust z
## (Intercept)    3.0231602 2.30390185  1.31219140  2.23204410  1.3544357
## bdi.pre        0.6479276 0.08228567  7.87412417  0.08351405  7.7583066
## treatmentBtheB -2.1692863 1.76642861 -1.22806339  1.73614385 -1.2494854
## length>6m      -0.1112910 1.73091679 -0.06429596  1.55092705 -0.0717577
## drugYes        -2.9995608 1.82569913 -1.64296559  1.73155411 -1.7322940
##
## Estimated Scale Parameter:  81.7349
## Number of Iterations:  5
##
## Working Correlation
##           [,1]      [,2]      [,3]      [,4]
## [1,] 1.0000000 0.6757951 0.6757951 0.6757951
## [2,] 0.6757951 1.0000000 0.6757951 0.6757951
## [3,] 0.6757951 0.6757951 1.0000000 0.6757951
## [4,] 0.6757951 0.6757951 0.6757951 1.0000000
```

Mixed model results different?

Interpretation is different (cond. vs marg.)

Thanks!

ehsan.karim@ubc.ca
www.ehsankarim.com