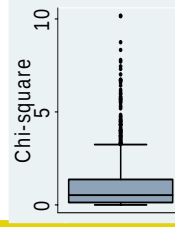
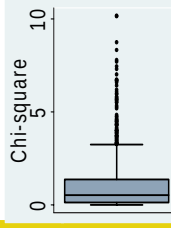


Longitudinal and Hierarchical Analytic Strategies for OAI Data

Charles E. McCulloch,
Division of Biostatistics,
Dept of Epidemiology and Biostatistics,
UCSF

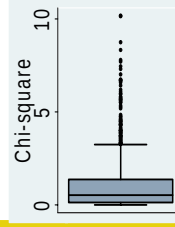
OARSI – Montreal September 10, 2009

Outline



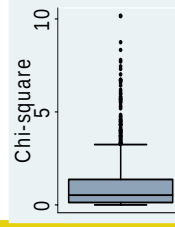
1. Introduction and examples.
2. Data layouts for hierarchical data.
3. Accommodating clustered or repeated measures data.
4. Analysis considerations.
5. Examples.
6. Tips and Tricks.
7. Analyzing change scores.
8. Summary.

Introduction



- Analysis technique depends on nature of the outcome variable and research question.
  - Binary: logistic regression (e.g., presence of osteophytes)
    - Odds ratios, area under ROC curve
  - Numeric: linear regression (e.g., WOMAC disability)
  - Also – time to event (Cox model or pooled logistic regression), count outcomes (Poisson regression)
- Methods need to be modified if there are clustered data or repeated measures for the *outcomes*.

Prototypical examples



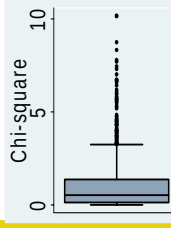
Example 1: (cross sectional) Is KOOS quality of life (QoL) at baseline related to joint space narrowing (by Xray) in either knee at baseline?

(not repeated measures)

Example 2: (clustered by knee) Is difference between men and women in the WOMAC pain score the same for those with and without symptomatic knee OA at baseline?

Example 3: (longitudinal/change) What is the rate of accumulation of bone marrow lesions or meniscal damage? Which tends to occur first?

Prototypical examples

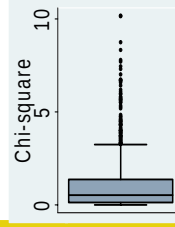


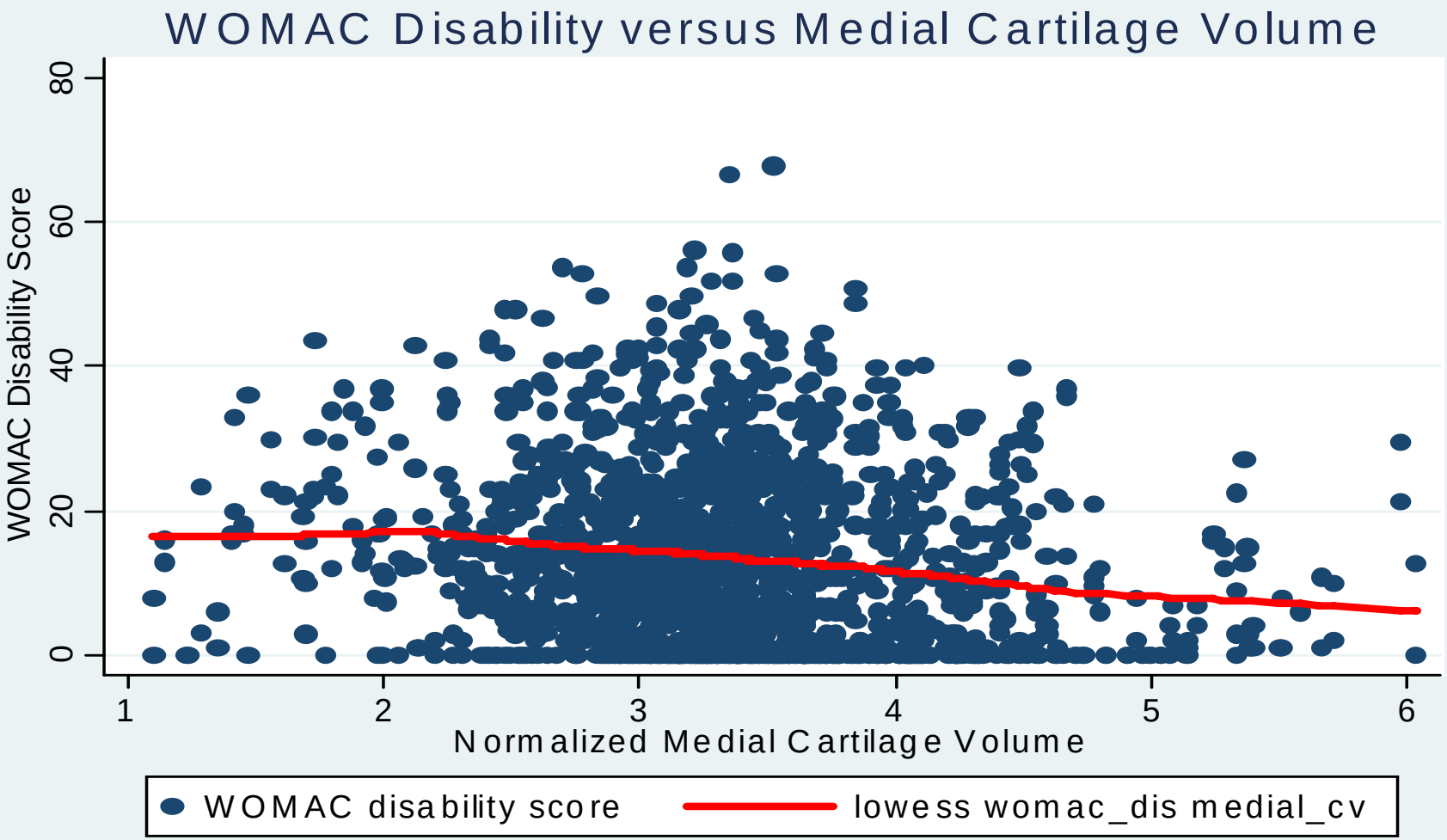
Example 4: (longitudinal/binary outcome) Does medial mean thickness at baseline predict change in activity limitation due to knee pain?

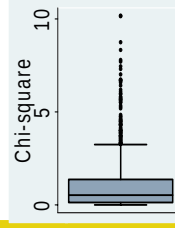
Example 5: (longitudinal/change/clustered) Is the change in WOMAC disability (knee specific) over 24 months related to the change in medial cartilage volume?

Example 6: (clustered/mediational) Does the change in cartilage volume between 0 and 12 months predict change in WOMAC disability between 12 and 24 months?

Ex 4: Is WOMAC disability related to Medial Cartilage Volume?





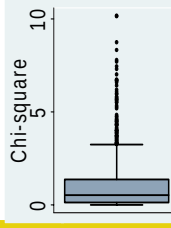


Data layouts for longitudinal/clustered data

For change score analyses: “wide format”

id	p01bmi	v01bmi	v02bmi
9000296	29.8	29.4	29.1
9000798	32.4	32.3	32.5

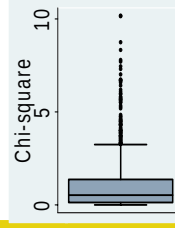
Data layouts for longitudinal/clustered data



For longitudinal analyses: “long format”

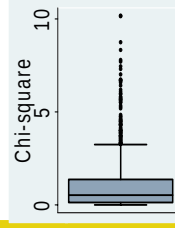
id	visit	knee	bmi	sxkoa
9000296	0	L	29.8	0
9000296	0	R	29.8	1
9000296	12	L	29.4	0
9000296	12	R	29.4	1
…				
9000798	0	R	32.4	0
9000798	12	L	32.3	0
9000798	12	R	32.3	0
9000798	18	L	32.5	0
9000798	18	R	32.5	1

Accommodating clustered or repeated measures data



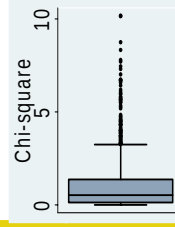
- Important to accommodate clustering and repeated measures.
- Otherwise SEs, p-values and confidence intervals can be incorrect, sometimes grossly so.
- Not possible to predict how the results will change when the proper analysis is used.

Efficiency of analyses of clustered data



- For between person predictors (e.g. BMI), the proper, clustered-data (e.g.,outcome measured on two knees) analysis will usually have larger SEs.
  - Intuition: for between person predictors an analysis that assumes all knees are independent over-represents the information content.
- For within person predictors (e.g., knee-specific), the proper, clustered-data analysis will usually have smaller SEs.
  - Intuition: Using each person as their own control increases efficiency.

Ex 2: Is there a sex by SX OA interaction for the WOMAC pain score at baseline?

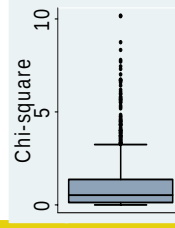


Mean WOMAC pain score	Males	Females	Difference
No Sx Knee OA	1.52	1.57	0.05
Sx Knee OA	3.58	4.49	0.91

When analyzing knees, effect of failing to allow correlation between a person's knees

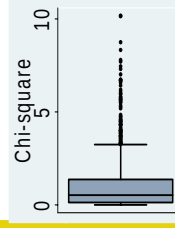
Analysis	Coeff	SE	p-value
Assume indep	-0.87	0.27	0.001
Allow correlation	-0.87	0.37	0.02

Accommodating clustered or repeated measures data



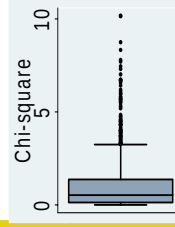
- Many methods exist to accommodate
  - Mixed models (e.g., SAS Proc MIXED, NLMIXED)
  - Repeated measures ANOVA (e.g., SAS Proc GLM)
  - Alternating logistic regression (in SAS Proc GENMOD)
  - Generalized Estimating Equations (GEEs). Invoked in SAS using Proc GENMOD using the REPEATED statement.

Accommodating clustered or repeated measures data



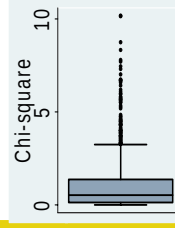
- Repeated measures/clustering is an issue for the *outcome* variable, not the predictor.
- Example: Are days missed from work predicted by knee pain (separate values for left and right knee). Does not have repeated measures on the outcome.
- Can accommodate by including both left and right knee values as predictors or by calculating summary measure(s) (e.g., average knee pain).

Desirable features for an analysis method



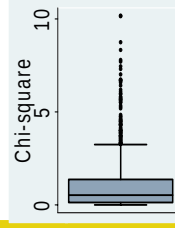
- Can accommodate a variety of outcome types (e.g., binary and numeric).
- Can accommodate clustering by knee, person (over time) and perhaps even different regions of interest (ROI) within a knee.
- Does not require extensive modeling of the correlation over time or between knees or between ROI in the knee.

Recommended analysis strategy – Generalized Estimating Equations (GEEs)



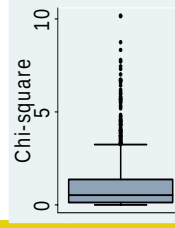
- Works with many types of outcomes.
- Robust variance estimate – obviates need to model correlation structure.
- Works well with not too many repeated measures per subject and a large number of subjects.
- So ideal for analyses incorporating multiple knees and time points. Somewhat less good if there are also multiple ROI per knee treated as outcomes (e.g., tibial and femoral cartilage loss).

Recommended analysis strategy - GEEs



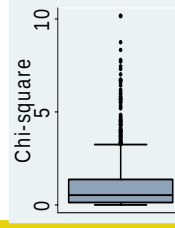
- Accommodates unbalanced data, e.g., some subjects contribute one knee while others contribute two.
- Accommodates unequally spaced data, e.g., missed visits or image data unavailable.
- BUT – always be wary of the pattern of missing data. If the fact that the data are missing is informative (e.g., those with missed visits are in extreme pain), virtually no standard statistical method will get the right answer.

Analyzing change with longitudinal data

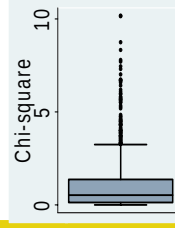


- Including a variable for *time* (or *visit*) describes the change over time, e.g., progression.
- Inclusion of *time* (or *visit*) interactions with baseline predictors allows analysis of whether baseline predictors are associated with change over time.
- Inclusion of a time-varying predictor (e.g., MRI findings at sequential visits) allows analysis of whether change in that predictor is associated with change in the outcome.

Analyzing change with longitudinal data



- Can use lagged variables to ask if prior values of risk factors predict later onset of disease (Is it prognostic?)
- Helps to strengthen inference of causation.
- With longitudinal data be wary of adjusting for baseline values of the outcome. Doing so will usually bias estimates of change.



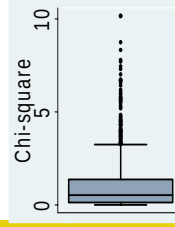
Ex 4: Medial thickness (at BL)/Activity

Binary outcome, so logistic regression, GEE analysis.

Medial mean cartilage thickness at baseline is a time-invariant variable, so include an interaction with time (visit).

Being 1 SD above normal in medial cartilage thickness at baseline is associated with an 8% increase in the per year odds ratio for activity limitation (1.08, 95% CI [0.8, 1.6], p=.68).

Ex 5: Is WOMAC Disability related to Medial Cartilage Volume?



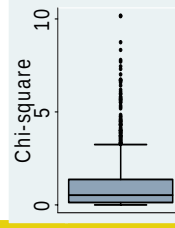
Treat WOMAC disability as approximately normally distributed.

Both WOMAC disability and cartilage volume are time varying variables. So include medial cartilage volume as a predictor.

Coefficient of cartilage volume is -2.04 (95% CI [-2.86, -1.23], $p<0.0005$).

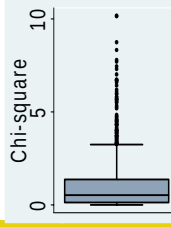
So a 1 unit increase in cartilage volume is associated with about a 2 unit decrease in WOMAC disability score.

Tips and Tricks



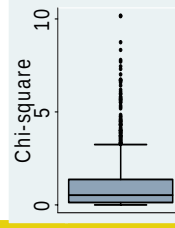
- If you use data from multiple imaging projects from the same vendor, include project as a categorical predictor (to account for any systematic differences) and remove duplicated subjects (or average their results).
- Variables that vary both within and between subjects may be susceptible to between person confounding. A solution is to partition *predictors* into within and between subject components as follows.
- Instead of just using, say, cartilage volume at time *t* as a predictor, include two predictors
  - Average cartilage volume for that subject across all times
  - The difference of the volume measurement and the average
  - The second predictor (difference) isolates the within component.

Analyzing change over time: What about analyzing change scores?



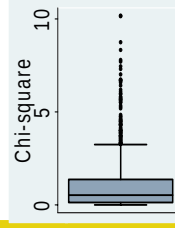
- An excellent and simple method when there are only two time points of interest and most subjects have complete data.
- Not as attractive with multiple time points or unbalanced data. Some loss of efficiency.

Analyzing change over time: What about analyzing change scores?

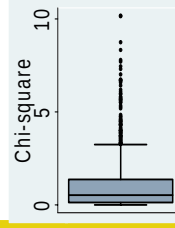


Mean WOMAC knee pain (N)		
	SX KOA	
Visit	No	Yes
BL	1.55	4.13
	1,954	730
12 month	1.42	3.72
	1,860	676

Analyzing change over time: What about analyzing change scores?



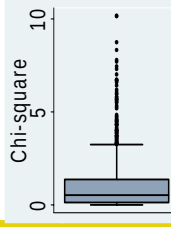
- Using a longitudinal analysis the difference in change (BL to 12 month visit) between the OA and non-OA groups is 0.268 with a SE of 0.1333 and a p-value of 0.045.
- Using the change score analysis the difference is 0.264 with a SE of 0.1339 and a p-value of 0.049.
- Adjusting for the baseline value gives a difference of 0.42 with a p-value approximately 0. So the adjusted analysis addresses a different question.



Summary

- Use Generalized Estimating Equations to accommodate longitudinal and/or clustered data.
- Include time-varying predictors directly to model whether change in outcome is associated with change in predictor.
- Include a time (e.g., visit) variable and interactions with time-invariant predictors (e.g., baseline characteristics) to model association of a predictor with change in outcome.

Questions?



Contact information:

chuck@biostat.ucsf.edu