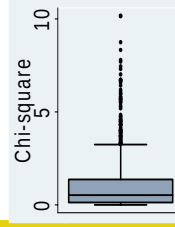
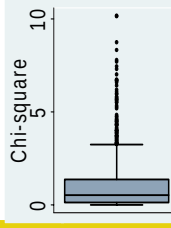


Analytic Strategies for the OAI Data

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

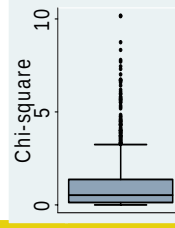
ACR October 2008

Outline



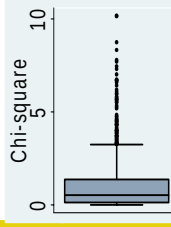
1. Introduction and examples.
2. General analysis considerations.
3. Accommodating correlations between knees within a person
4. Accommodating correlations over time
5. Analyzing change.
6. Questions from the participants.

Introduction



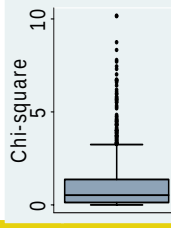
- Analysis technique depends on nature of the outcome variable and research question.
  - Binary: logistic regression (e.g., presence of osteophytes)
    - Odds ratios, area under ROC curve
  - Numeric: linear regression (e.g., WOMAC pain)
  - Also – time to event (Cox model or pooled logistic regression),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

Prototypical examples



Example 1: (cross sectional) Is KOOS quality of life related to BMI at baseline?

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?

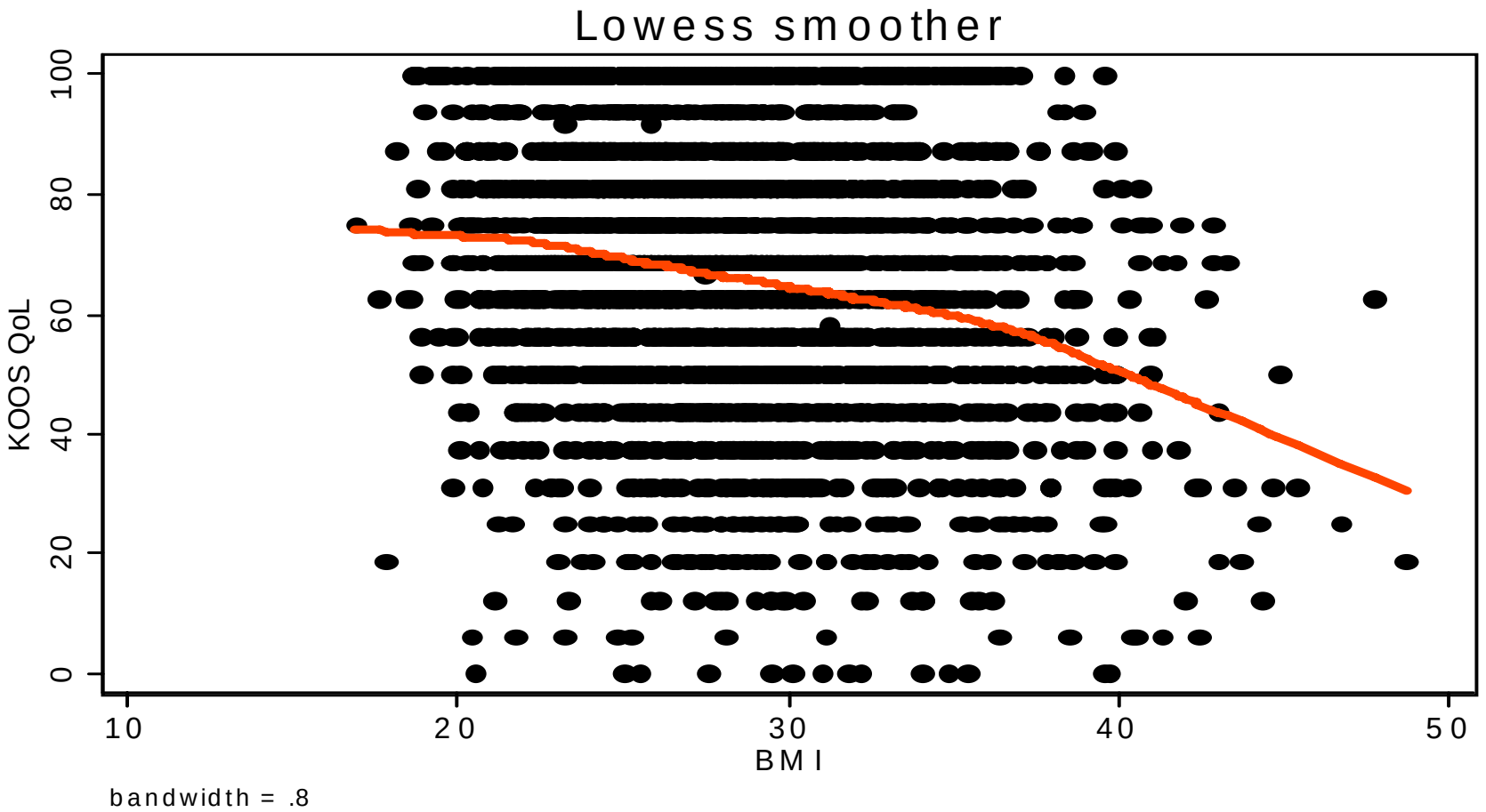


Prototypical examples

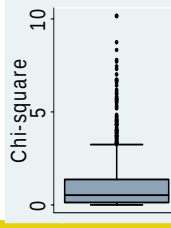
Example 3: (clustered data) Is the presence of osteophytes at baseline predicted by knee pain?

Example 4: (longitudinal/change) Is the 18 month change in WOMAC pain score the same or different for those with symptomatic knee OA at baseline?

Ex 1: Is KOOS QoL related to baseline BMI?



Ex 1: Is KOOS QoL related to baseline BMI?

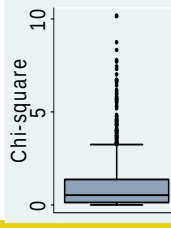


Analysis: linear regression.

Regression coefficient is -1.01 with a SE of 0.09 and a p-value of <0.0001.

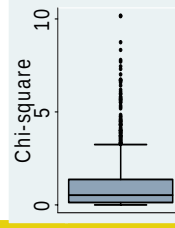
Not clustered data.

Accommodating clustered or repeated measures data



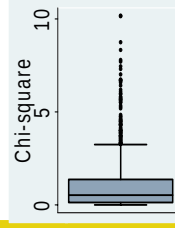
- Important to accommodate clustering and repeated measures.
- Otherwise SEs, p-values and confidence intervals can be incorrect, sometimes grossly so.
- Not possible to predict how the results will change when the proper analysis is used.

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 baseline SX OA interaction for the WOMAC pain score?

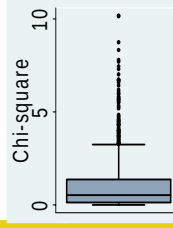


When analyzing knees, effect of failing to allow correlation between a person's knees

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

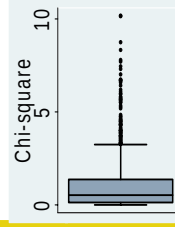
Mean WOMAC pain score	Males	Females	Difference
No Knee OA	1.52	1.57	0.05
Knee OA	3.58	4.49	0.91

Accommodating clustered or repeated measures data



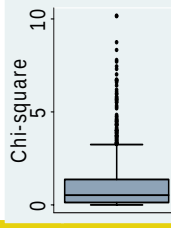
- Many methods exist to accommodate
  - Mixed models (e.g., SAS Proc MIXED, NLMIXED)
  - Repeated measures ANOVA (e.g., SAS Proc GLM)
  - Alternating logistic regression (in SAS Proc GENMOD)
  - Generalized Estimating Equations (GEEs). Invoked in SAS Proc GENMOD using the REPEATED statement.

Accommodating clustered or repeated measures data



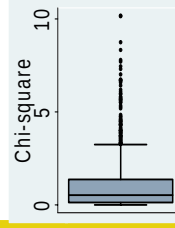
- Repeated measures/clustering is an issue for the *outcome* variable, not the predictor.
- Example: Are days missed from work predicted by knee pain (separate values for left and right knee). Does not have repeated measures on the outcome.
- Can accommodate by including both left and right knee values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

Desirable features for an analysis method



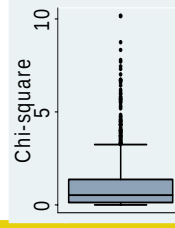
- Can accommodate a variety of outcome types (e.g., binary and numeric).
- Can accommodate clustering by knee, person (over time) and perhaps even different regions of interest (ROI) within a knee.
- Does not require extensive modeling of the correlation over time or between knees or between ROI in the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

Recommended analysis strategy - GEEs



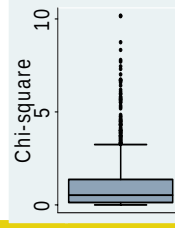
- Works with many types of outcomes.
- Robust variance estimate – obviates need to model correlation structure.
- Works well with not too many repeated measures per subject and a large number of subjects.
- So ideal for analyses incorporating multiple knees and time points. Somewhat less good if there 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.
- 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.

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



Effect of different analysis methods:

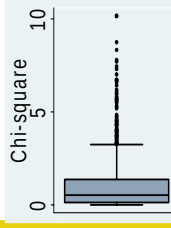
Analysis	Coeff	SE	p-value
Assume indep	-0.87	0.27	0.0013
GEE - robust	-0.87	0.37	0.02
Mixed	-0.87	0.32	0.01
Mixed - empirical	-0.87	0.37	0.02

SAS – GENMOD “REPEATED” option

SAS – MIXED “EMPIRICAL” option

Stata – “cluster()” or “vce(robust)” or “robust”

Ex 2: Does pain predict presence of osteophytes at baseline?

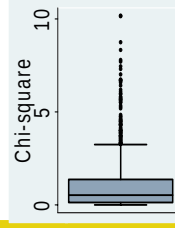


The odds of an osteophyte increase by 12.5% with each increase in pain score of 1 (0-10 scale).

Odds ratio of 1.12 (95% CI 1.09, 1.15).

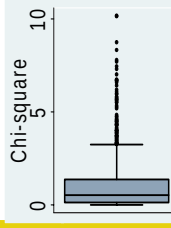
Accounts for clustering by subject.

Analyzing change with longitudinal data



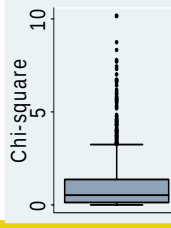
- Including a variable for *time* (or *visit*) describes the change over time, e.g., progression.
- Inclusion of *time* (or *visit*) interactions with baseline predictors allows analysis of whether baseline predictors are associated with change over time.
- Inclusion of a time-varying predictor (e.g., 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

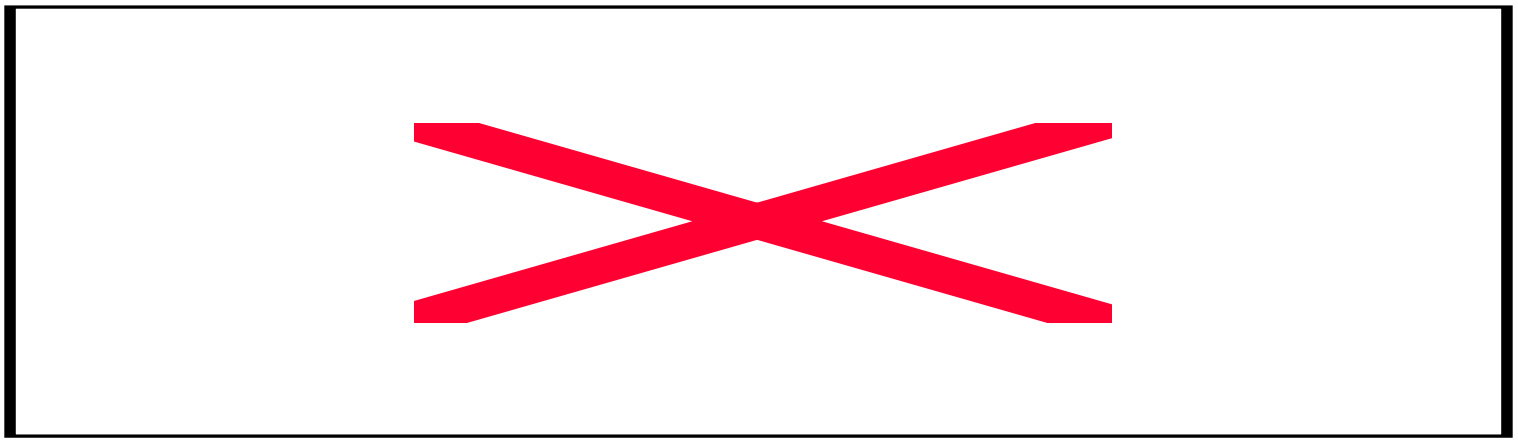


- Can use lagged variables to ask if prior values of risk factors predict later onset of disease (Is it prognostic?)
- Helps to strengthen inference of causation.

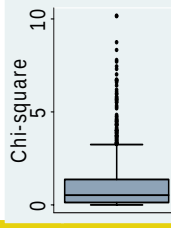
Ex 3: Does 18 month change in WOMAC pain depend on baseline SX K OA?



Include a SX K OA by visit interaction in the model.

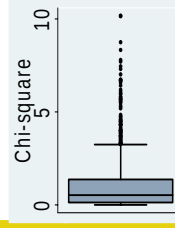


Analyzing change over time: What about analyzing change scores?



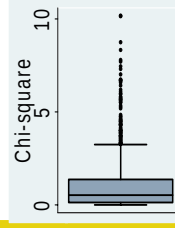
- An excellent and simple method when there are only two time points of interest and most subjects have complete data.
- Not as attractive with multiple time points or unbalanced data. Some loss of efficiency.
- If you do analyze change scores, be very wary of adjusting for baseline values of the change scores. Doing so will usually bias estimates of change.

Analyzing change over time: What about analyzing change scores?

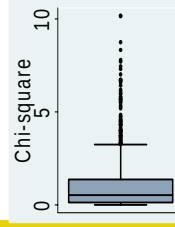


Mean WOMAC knee pain (N)		
	SX KOA	
Visit	No	Yes
BL	1.55	4.13
	1,954	730
12 month	1.42	3.72
	1,860	676

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.

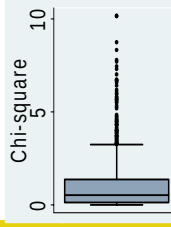


Would more detailed workshops be useful?

- Hands-on in a computer lab?
- What would you like to see?
  - How to explore the data to find variables
  - How to generate simple descriptive statistics from the web site
  - How to work with the image data
  - More specific guidance on statistical analysis methods

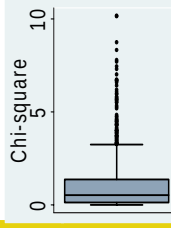
- Sign up sheet in back to give us feedback

Questions and Answers?



Contact information:

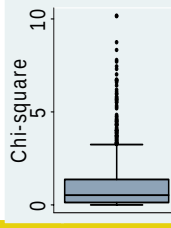
chuck@biostat.ucsf.edu



Data layouts for longitudinal/clustered data

For longitudinal analyses: "long format" one row of data per outcome (multiple rows per subject)

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



Data layouts for longitudinal/clustered data

For change score analyses: “wide format”. Append all the variables as multiple columns, one row per subject.

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

Below is sample SAS code that was used to generate the examples in the talk. These are intended to indicate the variety of analyses possible with the OAI data with a focus on accommodating the repeated measures/clustered data nature of the dataset. They do not represent thorough data analyses.

```
data all;
set cemwork.oai_example;
/* Pick out half the subjects so examples run faster*/
if id<9547618;
/* Symptomatic knee OA */
sxkoa=P01SXKOA>0;
/* Define definite presence of osteophytes */
defosteo=(svgkost=2);
run;

/* For repeated measures analyses, data should be sorted by id */
proc sort;
by id visit;
run;

/* Extract baseline data only for first three analyses */
data bl;
set work.all;
if visit=0;
run;

title "Cross-sectional analysis of KOOS QoL";
/* Example 1: Association of KOOS QoL as a function of baseline BMI */
proc glm data=work.bl;
model koosqol=bmi;
run;

title "Clustered baseline of WOMAC pain score";
/* Example 2: Comparing association of WOMAC pain score with
symptomatic knee OA
between men and women at baseline only */
/* Using GEE methods assuming independence working correlation */
proc genmod data=work.bl;
class p02sex id;
model womkpg=SXKOA p02sex SXKOA*p02sex;
repeated subject=id;
run;

/* Example 2: GEE specifying an exchangeable working correlation
structure */
proc genmod data=work.bl;
class p02sex id;
model womkpg=sxkoa p02sex sxkoa*p02sex;
repeated subject=id / type=exch;
run;

/* Example 2: using mixed model */
proc mixed data=work.bl noclprint;
class p02sex id;
model womkpg=sxkoa p02sex sxkoa*p02sex/solution;
random intercept/ subject=id;
run;
```

```
/* Example 2: using mixed model with EMPIRICAL option */
proc mixed data=work.bl empirical noclprint;
class p02sex id;
model womkpg=sxkoa p02sex sxkoa*p02sex/solution;
random intercept/subject=id;
run;

title "Binary outcome clustered";
/* Example 3:  Does pain scale predict presence of osteophytes at
baseline */
proc genmod data=work.bl descending;
class  id;
model defosteo=P7GKRCV / dist=bin;
estimate "pain effect"  p7gkrcv 1 / exp;
repeated subject=id / type=exch;
run;

title "Longitudinal";
/* Example 4: Does change in WOMAC pain depend on baseline SX KOA?*/
/* Using GEEs */
proc genmod data=work.all;
class p02sex id visit;
model womkpg=sxkoa p02sex visit sxkoa*p02sex sxkoa*visit;
repeated subject=id;
run;

/* Example 4: GEE specifying an exchangeable working correlation
structure */
proc genmod data=work.all;
class p02sex id visit;
model womkpg=sxkoa p02sex visit sxkoa*p02sex sxkoa*visit;
repeated subject=id/type=exch;
run;

/* Example 4: mixed model specifying an exchangeable working
correlation structure */
proc mixed data=work.all noclprint;
class p02sex id visit;
model womkpg=sxkoa p02sex visit sxkoa*p02sex sxkoa*visit/solution;
random id;
run;

/* Example 4: mixed model using EMPIRICAL option */
proc mixed data=work.all noclprint empirical;
class p02sex id visit;
model womkpg=sxkoa p02sex visit sxkoa*p02sex sxkoa*visit/solution;
random intercept/subject=id;
run;

/* Example 4: mixed model specifying a nested error structure */
proc mixed data=work.all noclprint;
class p02sex id visit;
model womkpg=sxkoa p02sex visit sxkoa*p02sex sxkoa*visit/solution;
random id visit(id);
run;
```