

## Studying for Exam II

Focus on homeworks 5-9, examples done in the notes, and Sections:

5.1, 5.2, 5.3, 5.4

6.1, 6.2, 6.3, 6.6

7.1, 7.2

10.1, 10.2

11.3, 11.4

12.1, 12.2, 12.3, 12.6.

## *Second half of course*

### **Chapter 5: logistic regression I**

- Logistic regression with one predictor.
- Parameter estimates give odds ratios.
- Case/control (retrospective) studies don't change parameter estimates (only intercept estimate).
- More crab analyses.
- GOF for logistic regression. Grouped data versus ungrouped. Hosmer and Lemeshow.
- Categorical predictors.
- Multiple predictors.
- A bit on fitting.

## Chapter 6: Logistic regression II

- Building models. Type III tests for dropping variables. Hierarchical models. Backwards elimination. AIC. Crab data (yet again).
- Diagnostics: residuals (Pearson and standardized Pearson  $r_i$ ), Cook's distance-type influence statistic  $c_i$ .  $Df\beta_{ij}$ .
- Logistic regression residuals borderline useless without replication. How does replication affect GOF tests?

- $2 \times 2 \times K$  tables: CMH (Cochran-Mantel-Haenszel) versus logistic approach. Estimation of stratum (block) effects (useful for model checking in GLMM!). Testing  $X \perp Y|Z$ : additive versus interaction alternatives. Clinical trial data on infection cream. Additive = homogeneous association: one overall treatment effect.
- Sample size and power in study design.
- Alternate links: probit, complimentary log-log, Cauchy. Left out: nonparametric estimation of link.
- Left out: small sample testing of  $\beta_j$  in logistic regression. Not in book: Bayesian approach works for small samples.

## Chapter 7: extending the logistic regression model to nominal and ordinal multinomial outcomes

- Baseline-category logit models for *nominal* multinomial response. Alligator food!!! Know how to write down model and obtain probabilities.
- Cumulative logit (proportional odds) models for *ordinal* multinomial response.

$$\log \frac{P(Y \leq j | \mathbf{x}_1) / P(Y > j | \mathbf{x}_1)}{P(Y \leq j | \mathbf{x}_2) / P(Y > j | \mathbf{x}_2)} = \boldsymbol{\beta}'(\mathbf{x}_1 - \mathbf{x}_2)$$

is log cumulative odds ratio.

- Latent variable motivation. Mental impairment example.
- Continuation-ratio (hazard regression) models.

## Chapter 10: marginal versus conditional modeling, basic ideas

- Marginal analysis of dependent proportions. Prime minister approval rating data!!! McNemar's test of marginal homogeneity for  $2 \times 2$  table.
- Conditional logistic regression. *Matched* case/control studies: gives different conditional likelihood than unmatched case/control data. Introduces idea of subject-specific effects  $u_i$ . In PROC LOGISTIC add a STRATA statement. 12.1.5 (p. 496): “Conditional ML is also appropriate with retrospective sampling. In that case, bias can occur with a random effects approach because the clusters are not randomly sampled.”

## Chapter 11: Marginal modeling of clustered data: GEE approach

- GEE approach to marginal modeling. Focuses on estimation of population averaged (marginal) effects.
- “Working” correlation structures: exchangeable, AR(1), etc.
- Sandwich estimator  $\widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$  uses estimated working covariance matrix as well as empirical estimate; requires proper specification of the mean  $E(Y_{ij}) = g^{-1}(\mathbf{x}'_{ij}\boldsymbol{\beta})$  to be valid (as most models do).
- Longitudinal mental depression data. Interaction of time and treatment.
- $2 \times 2 \times K$  tables where correlation among observations in a stratum (or *block*) accounted for in estimation (in homework).

## Chapter 12: Conditional modeling of clustered data: GLMMs

- Spent a lot of time here, reflects GLMM widespread use. and widespread use of random effects and latent variables models in general.
- Basic idea: random effect  $\mathbf{u}_i$  induces correlation among repeated measurements in cluster  $i$ :  $(Y_{i1}, \dots, Y_{in_i})$ . Can represent latent, unmeasured covariates or predisposition toward the event being modeled. e.g. level of sleeplessness, tolerance for pain, clinic population effect. Only looked at univariate  $u_i$ .
- Logistic-normal model. Marginal from conditional:  
 $P(Y_{ij} = 1) = E(Y_{ij}) \approx e^{c\mathbf{x}'_{ij}\boldsymbol{\beta}} / (1 + e^{c\mathbf{x}'_{ij}\boldsymbol{\beta}})$  where  
 $c = 1/\sqrt{1 + 0.6\sigma^2}$ .

- Longitudinal mental depression example again. Differences in interpretation between GEE approach and GLMM. Clinical trials example again.
- $2 \times 2 \times K$  tables where stratum effect  $u_i$  modeled explicitly via random effects (homework problem).
- Testing  $H_0 : \sigma = 0$  in logistic-normal model from fitting model with and without random effects. Is Wald test from table of coefficients okay here?
- **Left out:** nonparametric modeling of random effects  $\mathbf{u}_1, \dots, \mathbf{u}_n$ . Diagnostics. Multilevel models with layers of random effects. Other correlation models, e.g. temporal, spatial. PQL approach (fast, easy, and inaccurate – similar to GLIMMIX).

## Omitted or only briefly mentioned...

- Log-linear modeling (Chapters 8 and 9): useful to assess dependence among several categorical outcomes, e.g.  $I \times J \times K \times L$  tables. We only looked at  $I \times J$  in Chapter 3 and tested  $H_0 : X \perp Y$ . Notes from 2007 and 2008 have intro to log-linear models.
- Tests for symmetry, rater agreement, ROC curves.
- Various models: quasi-symmetry, quasi-independence, Bradley-Terry model (Chapter 10); adjacent categories logits, discrete choice (covariates change w/ choice!) (Chapter 7), etc...
- Additive models: various alternative fitting approaches, interaction surfaces, etc...
- Much, much more...scratched surface but covered a lot of ground.

**For exam II, focus on:**

- Models: (a) logistic regression with continuous and categorical predictors, (b) baseline category logit for nominal and ordinal, (c) proportional odds for ordinal, (d) marginal and conditional (i.e. random effects) versions of these. Be able to obtain odds and *probabilities*, relative risks, et cetera from these models for any covariate combination.
- Marginal approaches (Chapters 10 and 11). Course focused more on GEE approach of Chapter 11, but some material in Chapter 10 and two homework problems. Understand what handful of working correlation structures imply about clusters of outcomes. Be able to interpret SAS output. Understand difference and interpretation between marginal and conditional approaches.

- Conditional approaches (Chapters 10 and 12). Course focused more on maximum likelihood approach to fully specified model ( $u_1, \dots, u_n \stackrel{iid}{\sim} N(0, \sigma^2)$ ). Correct test for  $H_0 : \sigma = 0$ . Interpretation and comparison to *fixed effects* analogue. I think of random effects as a sample from some large (theoretically infinite) population; if *iid* they imply exchangeability. Fixed effects are used if you are actually interested in them, or there's only a few of them. How many parameters does each approach add to a basic (independence) model?

Either way, these effects are usually of second interest (they imply blocks) to “treatment” or “population” effects.