What to do today (Apr 5)?

1. *Introduction and Preparation*
2. *Analysis with Binary Variables (Chp 1-2)*
3. *Analysis with Multicategory Variables (Chp 3)*
4. *Analysis with Count Variables (Chp 4)*
5. *Model Selection and Evaluation (Chp 5)*

6. **Additional Topics (Chp 6)**
   - *6.1 Exact inference (Chp 6.2)*
   - *6.2 Revisit to Loglinear and Logistic Models for Contingency Tables: the Loglinear-Logit Connection* (Supplementary)
   - **6.3 Revisit III to GLM and Some Advanced Topics (Chp 5.3, Chp 6.5)**
     - *6.3.1 Revisit III to GLM*
     - **6.3.2 Marginal Modeling**
     - **6.3.3 Mixed Effects Models for Correlated Data**

Example. Alcohol, Cigarette, and Marijuana Use for High School Seniors, by Gender (G) and Race (R)

		Marijuana Use (M)							
		White				Other			
		Female		Male		Female		Male	
Alcohol Use (A)	Cigarette Use (C)	Yes	No	Yes	No	Yes	No	Yes	No
Yes	Yes	405	268	453	228	23	23	30	19
	No	13	218	28	201	2	19	1	18
No	Yes	1	17	1	17	0	1	1	8
	No	1	117	1	133	0	12	0	17

the total number of subjects: n=2276

- How are A, C, M associated?
  previous example with a partial table
- How are A,C,M associated, adjusting for R (race) and G (gender)? See the following ...

Step 1. Preliminary Analysis

- 1.1. Loglinear analysis:

- variable selection
  - starting with $(ACGMR)$; variable selection using $R : step()$
    $\Longrightarrow$ (ACGR, AM,CM,GM,MR)
  - further variable selection with (ACG,ACR, AGR, CGR, AM,CM,GM,RM)?
    $\Longrightarrow$ (ACR, AG, AM,CM,GM,MR)

- analysis outcome with the selected model

$$\frac{\begin{array}{r} R : tmp.out1 < -glm(counts \sim (AUse * CUse * Race + AUse * Gender + AUse * MUse \\ +CUse * MUse + MUse * Gender), data = Table713, family = poisson) \end{array}}{\begin{array}{r} R : tmp.out1b < -glm(counts \sim (AUse * CUse * Race + AUse * Gender + AUse * MUse \\ +CUse * MUse + MUse * Gender), data = Table713, family = quasipoisson) \end{array}}$$

- 1.2. Logistic analysis: using $A \sim logit(CR, G, M)$

$$\begin{array}{r} R : tmp.out12 < -glm(AUse \sim CUse * Race + Gender + MUse, \\ weight = counts, data = Table713, family = binomial) \end{array}$$

Step 2. Marginal analysis with a newly defined response

- Defintion.
  - "Response"=using substance ###yes=1; no=0
  - "Type"=the type of substance ###1,2,3 for A,C,M

- Logistic Regression: viewing all observations indpt
  - variable selection from $Response \sim logit(G * R * Type)$ to $Response \sim logit(G * Type, R)$
  - analysis outcome ... ...

$R : tmp.out2 <- glm(Response \sim Gender * Type + Race,$
$data = Table713dataC, family = binomial)$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	1.90766	0.08854	21.545	< 2e-16	***
Gender2	-0.16643	0.12004	-1.386	0.1656	
Type2	-1.21857	0.10835	-11.247	< 2e-16	***
Type3	-2.29661	0.10724	-21.416	< 2e-16	***
Race2	-0.40701	0.10010	-4.066	4.78e-05	***
Gender2:Type2	0.15247	0.14910	1.023	0.3065	
Gender2:Type3	0.36862	0.14716	2.505	0.0123	*

Null deviance: 8883.1 on 6827 degrees of freedom
Residual deviance: 7876.4 on 6821 degrees of freedom
AIC: 7890.4

Alternatively, using two dummy variables S1=1,0 for using A or not, and S2=1,0 for using C or not (as in Agresti, 1996)

$$R : tmp.out2b <- glm(Response \sim Gender * S1 + Gender * S2 + Race, data = Table713dataC, family = binomial)$$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-0.38895	0.06147	-6.327	2.49e-10	***
Gender2	0.20219	0.08515	2.374	0.0176	*
S1	2.29661	0.10724	21.416	$<$ 2e-16	***
S2	1.07804	0.08788	12.267	$<$ 2e-16	***
Race2	-0.40701	0.10010	-4.066	4.78e-05	***
Gender2:S1	-0.36862	0.14716	-2.505	0.0123	*
Gender2:S2	-0.21614	0.12277	-1.761	0.0783	.

Null deviance: 8883.1 on 6827 degrees of freedom
Residual deviance: 7876.4 on 6821 degrees of freedom
AIC: 7890.4

Anything not quite right?

Step 3. GEE analysis with the newly defined response

- Logistic Regression, adjusting for the possible correlation among observations from the same student
  - $n = 2276$ students (clusters): ID used for diff students
  - each student has 3 response obstns:
    working correlation: "exchangable"
    $cor(Y_{iA}, Y_{iC}) = cor(Y_{iA}, Y_{iM}) = cor(Y_{iC}, Y_{iM}) = \rho$
- R : library(gee) $\rightarrow$ gee; library(geepack) $\rightarrow$ geeglm

$R : tmp.out3 <- gee(Response \sim Race + Gender * Type, id = ID,$
$data = Table713dataC, family = binomial, corstr = \text{"exchangeable"})$

	Estimate	Naive S.E	Naive z	Robust S.E.	Robust z
(Intercept)	1.9059457	0.08876452	21.471931	0.08892841	21.432360
Race2	-0.3826952	0.13561541	-2.821915	0.13545120	-2.825336
Gender2	-0.1686674	0.11996805	-1.405936	0.11988703	-1.406886
Type2	-1.2181782	0.08290443	-14.693765	0.08289060	-14.696216
Type3	-2.2956989	0.08237034	-27.870457	0.09056542	-25.348515
Gender2:Type2	0.1523329	0.11372451	1.339490	0.11309395	1.346958
Gender2:Type3	0.3679203	0.11273372	3.263622	0.12163124	3.024883
Working Correlation					
1.0000000	0.4376341	0.4376341			
0.4376341	1.0000000	0.4376341			
0.4376341	0.4376341	1.0000000			

R : $tmp.out3b < -geeglm(Response \sim Race + Gender * Type, id = ID,$
$data = Table713dataC, family = binomial, corstr = \text{"}exchangeable\text{"})$

	Estimate	Std.err	Wald	$Pr(> \lvert W \rvert)$	
(Intercept)	1.90594	0.08893	459.346	$<$ 2e-16	***
Race2	-0.38269	0.13545	7.982	0.00472	**
Gender2	-0.16867	0.11989	1.979	0.15947	
Type2	-1.21818	0.08289	215.979	$<$ 2e-16	***
Type3	-2.29570	0.09057	642.548	$<$ 2e-16	***
Gender2:Type2	0.15233	0.11309	1.814	0.17799	
Gender2:Type3	0.36792	0.12163	9.150	0.00249	**
Estimated Scale Parameters:					
	Estimate	Std.err			
(Intercept)	0.9988	0.02859			
Correlation: Structure = exchangeable Link = identity					
Estimated Correlation Parameters:					
	Estimate	Std.err			
alpha	0.4376	0.02072			
Number of clusters: 2276 Maximum cluster size: 3					

$R: tmp.out32 < -gee(Response \sim Race + Gender * S1 + Gender * S2, id = ID,$
$data = Table713dataC, family = binomial, corstr = \text{"exchangeable"})$

	Estimate	Naive S.E	Naive z	Robust S.E.	Robust z
(Intercept)	-0.3898	0.06179	-6.308	0.06186	-6.300
Race2	-0.3827	0.13562	-2.822	0.13545	-2.825
Gender2	0.1993	0.08512	2.341	0.08511	2.341
S1	2.2957	0.08237	27.870	0.09057	25.349
S2	1.0775	0.06591	16.349	0.06080	17.723
Gender2:S1	-0.3679	0.11273	-3.264	0.12163	-3.025
Gender2:S2	-0.2156	0.09207	-2.342	0.08416	-2.562
Working Correlation					
1.0000000	0.4376341	0.4376341			
0.4376341	1.0000000	0.4376341			
0.4376341	0.4376341	1.0000000			

6.3.3 Mixed Effects Models for Correlated Data: GLMM

Generalized Linear Mixed Models:

- **Random Component.** response r.v. $Y_{ij}|x_{ij}, z_{ij}, b_i \sim f(\cdot|x_{ij}, z_{ij}; b_i)$ with $\mu(x_{ij}, z_{ij}; b_i) = E(Y_{ij}|x_{ij}, z_{ij}; b_i)$: e.g. $Y_{ij} \sim B(1, \pi(x_{ij}, z_{ij}; b_i))$.

- **Systematic Component.** $[\beta_0 + b_{0i}] + [\beta_1 + b_{1i}]x + [\beta_2 + b_{2i}]z$

- **Link Function.** $g(\mu_i) = [\beta_0 + b_{0i}] + [\beta_1 + b_{1i}]x + [\beta_2 + b_{2i}]z$ The link function $g(\cdot)$ links the *random componet* through its mean and the *systematic component*

6.3.3 Mixed Effects Models for Correlated Data: GLMM

Generalized Linear Mixed Models:

- **Random Component.** response r.v. $Y_{ij}|x_{ij}, z_{ij}, b_i \sim f(\cdot|x_{ij}, z_{ij}; b_i)$ with $\mu(x_{ij}, z_{ij}; b_i) = E(Y_{ij}|x_{ij}, z_{ij}; b_i)$: e.g. $Y_{ij} \sim B(1, \pi(x_{ij}, z_{ij}; b_i))$.
- **Systematic Component.** $[\beta_0 + b_{0i}] + [\beta_1 + b_{1i}]x + [\beta_2 + b_{2i}]z$
- **Link Function.** $g(\mu) = [\beta_0 + b_{0i}] + [\beta_1 + b_{1i}]x + [\beta_2 + b_{2i}]z$

Examples of GLMM: regarding the random effects b_i

- $b_{0i} \sim N(0, \sigma_0^2)$ and $b_{1i} = b_{2i} = 0 \Rightarrow$ **GLMM** with random intercept
- $b_{0i} \sim N(0, \sigma_0^2)$, $b_{1i} \sim N(0, \sigma_1^2)$ and $b_{2i} = 0 \Rightarrow$ **GLMM** with random intercept, random slope to (b_{0i}, b_{1i}) have correlation $\rho_{01} \neq 0$.

- Model Fitting: estimating $\beta_0, \beta_1, \beta_2$ and σ_0^2, σ_1^2 etc. Likelihood based estimation procedures
- with R: in the *lme4* package
  e.g. $glmer(formula = response \sim x + (1|b), nAGQ = a, data, family)$
- Inference
  - for fixed-effect parameters
  - *for variance components*: e.g. $\sigma_0^2 = 0$?

Example. Falls with Head Impact (page 423, Schonnop et al, 2013) Fall is a serious problem among elderly, resulting in injuries, medical expenses and sometimes death.
Data: 227 falls among 133 residents at two long-term care facilities in BC with variables resident (id), initial (backward,down,forward,sideways), head (yes,not)

```
> head(fall.head)
  resident  initial head
1       56 Sideways    0
2        9 Backward    0
3       30  Forward    0
4        9     Down    0
5       70 Sideways    0
6       21 Sideways    1
```

```
> mod.glmm.1 <- glmer(formula = head ~ initial + (1|resident
    ), nAGQ = 1, data
> = fall.head, family = binomial)
> summary(mod.glmm.1)$varcor
 Groups   Name        Std.Dev.
 resident (Intercept) 0.25192
> summary(mod.glmm.1)$varcor[[1]][1,1]
[1] 0.06346608
> mod.glmm.5 <- glmer(formula = head ~ initial + (1|resident
    ), nAGQ = 5, data
> = fall.head, family = binomial)
> summary(mod.glmm.5)$varcor
 Groups   Name        Std.Dev.
 resident (Intercept) 0.30342
> mod.glmm.10 <- glmer(formula = head ~ initial + (1|
    resident), nAGQ = 10,
> data = fall.head, family = binomial)
> summary(mod.glmm.10)$varcor
 Groups   Name        Std.Dev.
 resident (Intercept) 0.30342
```

```
> summary(mod.glmm.5)
     AIC      BIC   logLik deviance df.resid
   279.5    296.4   -134.8    269.5      210

Random effects:
 Groups   Name        Variance Std.Dev.
 resident (Intercept) 0.09206  0.3034
Number of obs: 215, groups:  resident, 131

Fixed effects:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.6447     0.2469  -2.611  0.00901 **
initialDown       -1.1705     0.6783  -1.726  0.08440 .
initialForward     0.9581     0.3689   2.597  0.00940 **
initialSideways   -0.1208     0.3768  -0.321  0.74855
---
Correlation of Fixed Effects:
            (Intr) intlDw intlFr
initialDown -0.340
initilFrwrd -0.660  0.230
initilSdwys -0.620  0.240  0.423
>
```

What have we studied?

in STAT-475/675: **Analysis of Categorical Data**

- *1. Introduction and Preparation*
- *2. Analysis with Binary Variables (Chp 1-2)*
- *3. Analysis with Multicategory Variables (Chp 3)*
- *4. Analsyis with Count Response (Chp 4)*
- *5. Model Selection and Evaluation (Chp 5)*
- *6. Additional Topics (Chp 6: Chp6.2 and 6.5)*

All will be covered in the final exam.

Please be reminded ...

- On Tuesday Apr 10 10:30-11:20, Zhiyang will provide a review.
- There will be no tutorial from next week.
- Schedule for our office hours during the final exam period is posted in the webpage.

Thanks for your participation & good luck on the final exam!