What to do today (Apr 3)?

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 Ect Models for Correlated Data*

Plan for the rest of this term

6.3.2A Marginal Modeling: Quasi-Score

Recall *inference with GLM* ...

A. Modelling:

- Assume a GLM model,
  - **Random Component.** response r.v. $Y|X = x, Z = z \sim f(y|x,z)$ with $\mu(x,z) = E(Y|X = x, Z = z)$
  - **Systematic Component.** $h(x,z) = \beta_0 + \beta_1 x + \beta_2 z$
  - **Link Function.** $g(\mu) = h(x,z)$
- That is, assume $Y|X = x, Z = z \sim f(y|\mu) = f(y|x,z;\beta_0,\beta_1,\beta_2)$

B. Data: $\{(y_i, x_i, z_i) : i = 1, \ldots, n\}$ from n indpt units

C. Statistical Inference with GLM: the likelihood-based methods

What if we can't confidently specify response r.v. $Y|X = x, Z = z \sim f(y|x,z)$?

If, instead, we'd like to assume $Var(Y|X = x, Z = z) = I(\mu(x,z))$, such as $\phi\mu(x,z)$ in the Quasi-Poisson case. $\Longrightarrow$ **Moment (Marginal) Modeling**

What if the observations are not indpt? Examples?

6.3.2B Marginal Modeling: GEE Approach

A. Modelling: Assume r.v. Y with $\mu(x, z) = E(Y|X = x, Z = z)$

B. Data: $\{(y_{ij}, x_{ij}, z_{ij}) : j = 1, \ldots, J_i; i = 1, \ldots, n\}$ from n indpt units: n indpt clusters of observations

C. GEE approach:

- $R : gee(formula, id, data, family, corstr)$:
  - id: identifies the clusters
  - family=gaussian, binomial, poisson, Gamma, and quasi
  - corstr: the covariance structure of the response observations within a cluster, such"independence", "fixed", "stat_M_dep", "non_stat_M_dep", "exchangeable", "AR-M" and "unstructured"
- An alternative function $R : geeglm()$

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

$R: tmp.out1 < -glm(counts \sim (AUse * CUse * Race + AUse * Gender + AUse * MUse$
$+ CUse * MUse + MUse * Gender), data = Table713, family = poisson)$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	5.97802	0.04847	123.323	< 2e-16	***
AUseno	-5.87657	0.46542	-12.626	< 2e-16	***
CUseno	-3.03133	0.15235	-19.898	< 2e-16	***
Raceother	-2.65694	0.10614	-25.033	< 2e-16	***
Gendermale	0.14457	0.06473	2.233	0.025522	*
MUseno	-0.38955	0.07089	-5.495	3.9e-08	***
AUseno:CUseno	2.20630	0.19227	11.475	< 2e-16	***
AUseno:Raceother	1.37601	0.37288	3.690	0.000224	***
CUseno:Raceother	0.21459	0.19606	1.095	0.273716	
AUseno:Gendermale	0.29852	0.12743	2.343	0.019147	*
AUseno:MUseno	3.00592	0.46484	6.467	1.0e-10	***
CUseno:MUseno	2.84789	0.16384	17.382	< 2e-16	***
Gendermale:MUseno	-0.26929	0.09039	-2.979	0.002891	**
AUseno:CUseno:Raceother	-1.09579	0.45240	-2.422	0.015428	*

Null deviance: 4818.051 on 31 degrees of freedom
Residual deviance: 15.154 on 18 degrees of freedom
AIC: 179.39

$$R : tmp.out1b <- glm(counts \sim (AUse * CUse * Race + AUse * Gender + AUse * MUse + CUse * MUse + MUse * Gender), data = Table713, family = quasipoisson)$$

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	5.97802	0.04266	140.130	< 2e-16	***
AUseno	-5.87657	0.40960	-14.347	2.71e-11	***
CUseno	-3.03133	0.13407	-22.609	1.15e-14	***
Raceother	-2.65694	0.09341	-28.445	< 2e-16	***
Gendermale	0.14457	0.05697	2.538	0.020619	*
MUseno	-0.38955	0.06238	-6.244	6.86e-06	***
AUseno:CUseno	2.20630	0.16921	13.039	1.31e-10	***
AUseno:Raceother	1.37601	0.32816	4.193	0.000547	***
CUseno:Raceother	0.21459	0.17254	1.244	0.229558	
AUseno:Gendermale	0.29852	0.11214	2.662	0.015883	*
AUseno:MUseno	3.00592	0.40909	7.348	8.05e-07	***
CUseno:MUseno	2.84789	0.14419	19.751	1.20e-13	***
Gendermale:MUseno	-0.26929	0.07955	-3.385	0.003298	**
AUseno:CUseno:Raceother	-1.09579	0.39814	-2.752	0.013108	*

(Dispersion parameter for quasipoisson family taken to be 0.7745045)

AIC: NA

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

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

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-5.8248	0.4659	-12.501	$<$ 2e-16	***
CUseno	2.1937	0.1928	11.377	$<$ 2e-16	***
Raceother	1.2046	0.3884	3.102	0.00192	**
Gendermale	0.2677	0.1364	1.963	0.04967	*
MUseno	2.9831	0.4651	6.414	1.42e-10	***
CUseno:Raceother	-0.9500	0.4675	-2.032	0.04217	*

the estimated log(OR) of using A comparing using M vs not:

- from tmp.out12: $\hat{\beta}_2^M - \hat{\beta}_1^M = 2.98$
- from tmp.out1: $\hat{\lambda}_{22}^{AM} + \hat{\lambda}_{11}^{AM} - \hat{\lambda}_{21}^{AM} - \hat{\lambda}_{21}^{AM} = 3.01$

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

    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)
- 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

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

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

	Estimate	Std.err	Wald	$Pr(>\|W\|)$	
(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			

What are we going to do next?

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 Ect Models for Correlated Data** (**Thu Apr 5**)

... ... and

- **10:30-11:20 Tuesday Apr 10: A review by Zhiyang Zhou**
- **Final Exam: 15:30-18:30 Mon Apr 23; BLU 9660**