What to do today (Mar 15)?

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)

- **5.1 Variable selection (Chp 5.1.1-4)**
- **5.2 Tools to assess model fit (Chp 5.2)**
- **5.3 Examples**

Midterm 2: AQ 3005; 10:30-11:20

6. *Additional Topics (Chp 6)*

5C. Model Selection and Evaluation: in multiple logistic regression

General Setting:
A binary response Y (e.g. success (1)/failure (0)); several explanatory variables $X_1, \ldots, X_K$ (e.g. width, weight, color): to find out about the function $\pi(x_1, \ldots, x_K) = P(Y = 1|X_1 = x_1, \ldots, X_K = x_K)$
Multiple Logistic Regression Model:

$$logit\big[\pi(x_1, \ldots, x_K)\big] = \log\left[\frac{\pi(x_1, \ldots, x_K)}{1 - \pi(x_1, \ldots, x_K)}\right] = \alpha + \beta_1 x_1 + \ldots + \beta_K x_K$$

equivalently to $\pi(x_1, \ldots, x_K) = \frac{\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}{1+\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}$.
Available Data: $\big\{(y_i, x_{i1}, \ldots, x_{iK}) : i = 1, \ldots, n\big\}$ from indpt units.
Statistical inference under the model with the data:

- estimation of $\alpha, \beta_1, \ldots, \beta_K$: *MLE; CI/CR*; testing hypothese about $\alpha, \beta_1, \ldots, \beta_K$; estimation of $\pi(x_1, \ldots, x_K)$: *MLE; CI*
- model checking and variable selection: *compare the analysis with the nonparametric one; residuals analysis; model comparison;* model/variable selection

5C. Model Selection and Evaluation: in multiple logistic regression

Model Checking:

- **inferential methods**
  - after grouping data according to $X_1, \ldots, X_K$, applications of the Pearson's χ^2-test and the LRT
  - applying the LRT for comparing M_0 vs M_1,
    $\mathcal{G}^2(M_0|M_1) \sim \chi^2(df)$

- **graphical methods**: various residual plots
  - Pearson's residual: $e_k = \frac{y_k - n_k\hat{\pi}_k}{\sqrt{n_k\hat{\pi}_k(1-\hat{\pi}_k)}}$
    y_k=num of successes with n_k trials
  - the standardized (adjusted) Pearson's residual: $e_k^* = \frac{e_k}{\sqrt{1-h_k}}$
    h_k is the observation's leverage: the diagonal elements of estimated $\Sigma_{(K+1)\times(K+1)}$

5C. Model Selection and Evaluation: in multiple logistic regression

- **Variable Selection.**

  Caution in using multiple regression model about "multi-collinearity":
  If there are strong correlations in $X_1, \ldots, X_K$, none of them could seem important in the presence of the others in the model.

- **Criteria for Variable Selection:**
  - **classical criterion** selecting/keeping only predictors according to a pre-specified significance level
  - **Information criteria**: e.g. to achieve the min AIC, or corrected AIC or BIC

Example. Female Horseshoe Crabs and their Satellites: Revisit II. multiple logistic regression analysis

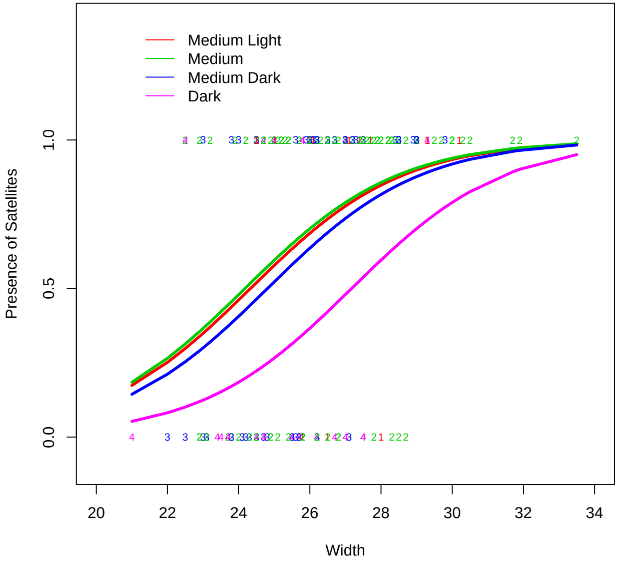
- Using Color and Width Predictors – $X_1 = width$, $X_2 = color$: (a surrogate for age) light (not sampled), medium light, medium, medium dark, dark:
  - $X_{21} = 1$ for medium, $= 0$ otherwise
  - $X_{22} = 1$ for medium dark, $= 0$ otherwise
  - $X_{23} = 1$ for dark, $= 0$ otherwise
- Consider $logit(\pi) = \alpha + \beta_1 x_1 + \beta_{21} x_{21} + \beta_{22} x_{22} + \beta_{23} x_{23}$

```
-------------------R Codes--------------------------------------------
tmpy<-ifelse(ex.crab[,5]>0,1,0)
tmpx1<-ex.crab[,3]
tmpx2<-ex.crab[,1]
tmpout<-glm(tmpy~tmpx1+as.factor(tmpx2), family=binomial)
summary(tmpout)
```

---------------R Output --

```
Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.1124  -0.9848   0.5243   0.8513   2.1413

Coefficients:                Estimate Std. Error z value Pr(>|z|)
(Intercept)       -11.38519    2.87346  -3.962 7.43e-05 ***
tmpx1               0.46796    0.10554   4.434 9.26e-06 ***
as.factor(tmpx2)2   0.07242    0.73989   0.098    0.922
as.factor(tmpx2)3  -0.22380    0.77708  -0.288    0.773
as.factor(tmpx2)4  -1.32992    0.85252  -1.560    0.119
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
 Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.46
```



Medium Light
Medium
Medium Dark
Dark
Presence of Satellites
0.0
0.5
1.0
20
22
24
26
28
30
32
34
Width

Revisit II.1: a multiple logistic regression analysis – goodness-of-fit? Inferential Procedures

- Compared to other models
  - to the null model ($M_0 : \pi = \frac{e^{\alpha}}{1+e^{\alpha}}$)
    $\mathcal{G}^2(M_0|M_1)_{obs} = 225.76 - 187.46$ with
    $df = 5 - 1 = [173 - 1] - [173 - 5]$
    $\Longrightarrow p - value < .001$, a significant improvement
  - to the simple logistic model with width only
    ($M_0 : \pi = \frac{e^{\alpha+\beta_1 x_1}}{1+e^{\alpha+\beta_1 x_1}}$)
    $\mathcal{G}^2(M_0|M_1)_{obs} = 194.45 - 187.46$ with
    $df = 5 - 2 = [173 - 2] - [173 - 5]$
    $\Longrightarrow p - value = .072$, a marginal improvement
    (the reduced model has the advantage of simpler interpretations)

Revisit II.2: multiple logistic regression analysis – To add in more predictors? How about two predictors' interactions?

Model selection (Backward Elimination)

Consider the multiple logistic regression with different sets of predictors:

Model	predictors	Deviance	df	AIC	Models Compared	Deviance Difference
1	C S + C W + S W	173.7	155	209.7	-	-
2	C + S + W	186.6	166	200.6	(2)-(1)	12.9 (df = 11)
3a	C + S	208.8	167	220.8	(3a)-(2)	22.2 (df = 1)
3b	S + W	194.4	169	202.4	(3b)-(2)	7.8 (df = 3)
3c	C + W	187.5	168	197.5	(3c)-(2)	0.9 (df = 2)
4a	C	212.1	169	220.1	(4a)-(3c)	24.6 (df = 1)
4b	W	194.5	171	198.5	(4b)-(3c)	7.0 (df = 3)
5	(C = dark) + W	188.0	170	194.0	(5)-(3c)	0.5 (df = 2)
6	None	225.8	172	227.8	(6)-(5)	37.8 (df = 2)

C=color; S=spine condition; W=width.

Note: A strong linear correlation between width and weight: sample corr=0.887. So weight is not included.

Revisit II.2: Model selection (Backward Elimination)

My variable selection by R ... ...

Using R function *step*(): a stepwise algorithm.

$step(object,\ direction = c(\text{"both"}, \text{"backward"}, \text{"forward"}))$

```
--------R Codes ----------------------------------
tmpy<-ifelse(ex.crab[,5]>0,1,0)
tmpx1<-ex.crab[,3]
tmpx2<-as.factor(ex.crab[,1])
tmpx3<-as.factor(ex.crab[,2])
tmpout3<-glm(tmpy~tmpx1*tmpx2*tmpx3, family=binomial)
step(tmpout3)
```

```
----------------R Output ----------------------
Step:  AIC=199.08
tmpy ~ tmpx1 + tmpx2 + tmpx1:tmpx2

              Df Deviance    AIC
- tmpx1:tmpx2  3   187.46 197.46
<none>             183.08 199.08

Step:  AIC=197.46
tmpy ~ tmpx1 + tmpx2

        Df Deviance    AIC
<none>       187.46 197.46
- tmpx2  3   194.45 198.45
- tmpx1  1   212.06 220.06

Call:  glm(formula = tmpy ~ tmpx1 + tmpx2, family = binomial)

Coefficients:
(Intercept)        tmpx1       tmpx22       tmpx23       tmpx24
  -11.38519      0.46796      0.07242     -0.22380     -1.32992

Degrees of Freedom: 172 Total (i.e. Null);  168 Residual
Null Deviance:       225.8
Residual Deviance: 187.5  AIC: 197.5
```

5D. Model Selection and Evaluation: in loglinear regression

e.g. loglinear model for three-way contingency tables:

Recall that

- how to establish the association of the cell counts, $N_{ijk} \sim Poisson(\mu_{ijk})$, with X, Y, and Z, three categorical variables?

Saturated Loglinear Model (XYZ) (including all main effects, two factor interactions, three factor interactions: df=IJK)

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ} + \lambda_{ijk}^{XYZ}$$

Loglinear Model of Multual Independence (X,Y,Z) (including only main effects: df =I+J+K-2)

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$$

Loglinear Model of Homogeneous Association (XY,YZ,XZ) (including all main effects, two factor interactions: assuming $\lambda_{ijk}^{XYZ} = 0$)

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{jk}^{YZ} + \lambda_{ik}^{XZ}$$

Parameter Interpretation for Model (XY,YZ,XZ): when I=J=2, X-Y conditional odds ratio at $Z = k$ for any k is

$$\log \theta_{XY(k)} = \log \left(\frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} \right) = [\lambda_{11}^{XY} + \lambda_{22}^{XY}] - [\lambda_{12}^{XY} + \lambda_{21}^{XY}]$$

$\Longrightarrow$ *Homogeneous Conditional Association of X-Y*
Further, if $\lambda_{ij}^{XY} = 0$,

- $\Longrightarrow$ Model (YZ,XZ)
- $\log \theta_{XY(k)} = 0$, for all $k \Longrightarrow X \perp Y | Z$

Statistical Inference with (the loglinear (Poisson) regression with 3 categorical predictors):

- Be careful with coding X, Y, Z
- Choice of models: e.g. (X,Y,Z), (X,YZ), (YZ,XZ), (XY,YZ,XZ), (XYZ)
- Variouse inference procedures:
  - Estm model parameters; estm μ_{ijk}; estm OR
  - Model checking/comparison: Pearson's χ^2-test, LRT-test

Example. Alcohol, Cigarette and Marijuana Use

Alcohol Use (A)	Cigarette Use (C)	Marijuana Use (M)	
		Yes	No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

Source: a survey conducted in 1992 by the Wright State Univ. School of Medicine and the United Health Services in Dayton.

Using *read.table* to read in data and *as.data.frame* to form it into R's data format, or

- $counts < -c(911, 44, 3, 2, 538, 456, 43, 279)$
- $A < -gl(2, 2, 8);\ C < -gl(2, 1, 8);\ M < -gl(2, 4, 8);$
  $\#\#1 = yes, 2 = no$
- $ACM.data < -cbind(A, C, M, counts)$

Run R to fit different models with the data: for example,

- For Model (ACM)

$$tmp.out < -glm(counts \sim A * C * M, family = poisson);$$

- For Model (AC,CM,AM)

$$tmp2.out < -glm(counts \sim A * C + C * M + A * M, family = poisson);$$

- For Model (CM,AM)

$$tmp3.out < -glm(counts \sim C * M + A * M, family = poisson);$$

- For Model (AC,M)

$$tmp4.out < -glm(counts \sim A * C + M, family = poisson);$$

- For Model (A,C,M)

$$tmp5.out < -glm(counts \sim A + C + M, family = poisson);$$

Step 1. Fitted Values for Loglinear Models:

- Plug in the estm for the parameters in the models to attain the fitted values, or
- Use "*tmp.out$fitted*", for example

The fit for (AC,AM,CM) is close to the observed data, the same as the fitted values for (ACM).

Fitted Values for Loglinear Models:

			Loglinear Model				
A	C	M	(A,C,M)	(AC,M)	(AM,CM)	(AC,AM,CM)	(ACM)
Yes	Yes	Yes	540.0	611.2	909.24	910.4	911
		No	740.2	837.8	438.84	538.6	538
	No	Yes	282.1	210.9	45.76	44.6	44
		No	386.7	289.1	555.16	455.4	456
No	Yes	Yes	90.6	19.4	4.76	3.6	3
		No	124.2	26.6	142.16	42.4	43
	No	Yes	47.3	118.5	0.24	1.4	2
		No	64.9	162.5	179.84	279.6	279

Step 2. To obtain estimates for what needed based on the analyses

- using the analysis outputs: the estms for the model parameters and their estimated standard errors

- using the fitted counts when applicable

e.g. the OR of alcohol use (A yes vs not) between cigarette use or not (C yes vs not)

- conditional on marijuana use (M=yes or not)
- marginal (regardless of M)

Step 3. Chi-Squared Goodness-of-Fit Tests: Loglinear Residuals

- $G^2[(AC, AM, CM)] = 2\sum n_{ijk}\log(\frac{n_{ijk}}{\hat{\mu}_{ijk}})$
- $X^2[(AM, CM)] = \sum \frac{(n_{ijk}-\hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$.
  - e.g. the Pearson's residuals: $e_{ijk} = \frac{n_{ijk}-\hat{\mu}_{ijk}}{\sqrt{\hat{\mu}_{ijk}}}$
  - residual plots: e.g. scatter plot of e_{ijk} vs A

Step 4. Model Selection: (backward elimination)

Start: AIC=65.04 $counts \sim ACM$			
	Df	Deviance	AIC
- A:C:M	1	0.37399	63.417
$< none >$		0.00000	65.043

Step: AIC=63.42 $counts \sim A + C + M + AC + AM + CM$			
	Df	Deviance	AIC
$< none >$		0.37	63.42
- A:M	1	92.02	153.06
- A:C	1	187.75	248.80
- C:M	1	497.37	558.41

Part V.2.2D for Three-Way Contingency Tables

Step 5. Tests about Partial Associations:

- The test statistic for testing $\lambda^{AC} = 0$ in (AC,AM,CM) is

$$\begin{aligned} G^2[(AM, CM)\big|(AC, AM, CM)] &= G^2(AM, CM) - G^2(AC, AM, CM) \\ &= 187.8 - 0.04, \end{aligned}$$

df=2-1
$\Longrightarrow p < 0.001$: strong evidence against the null hypothesis and in favor of an A-C partial association.

What will we study next?

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)**
   - **5.1 Variable selection (Chp 5.1.1-4)**
   - **5.2 Tools to assess model fit (Chp 5.2)**
   - **5.3 Examples**

6. **Additional Topics (Chp 6)**