What to do today (Feb 6)?

1. *Introdution and Preparation*

2. **Analysis with Binary Variables (Chp1-2)**
   - *2.1 Analysis with binary variables I (Chp 1)*
   - **2.2 Analysis with binary response (Chp 2)**
     - *2.2.1 Regression models (Chp2.1, Chp2.2.1)*
     - *2.2.2 Simple logistic regression analysis (Chp2.2.2-7)*
     - **2.2.3 Multiple logistic regression analysis (Chp2.2.2-7)**
   - **2.3 Generalized linear models (Chp2.3)**

3. *Analysis with Multicategory Variables (Chp3)*

2.2.3C Interactions and transformations of predictors

When to consider two predictors X_1, X_2
the logistic regression model I:

$$logit\left[\pi(x_1, x_2)\right] = \log\left[\frac{\pi(x_1, x_2)}{1 - \pi(x_1, x_2)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2$$

- **What if the interaction of X_1, X_2 is of interest?** $\Longrightarrow$ to add the term for $X_3 = X_1 X_2$ to the model:

$$logit\left[\pi(x_1, x_2)\right] = \log\left[\frac{\pi(x_1, x_2)}{1 - \pi(x_1, x_2)}\right] = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

- **What if the effect of $h(X_2)$ is of interest, instead of X_2?** $\Longrightarrow$ to replace X_2 with $X_2^* = h(X_2)$ in to the model:

$$logit\left[\pi(x_1, x_2)\right] = \log\left[\frac{\pi(x_1, x_2)}{1 - \pi(x_1, x_2)}\right] = \alpha + \beta_1 x_1 + \beta_2 h(x_2)$$

2.2.3D Qualitative explanatory variables

Example. AZT Use and AIDS (NY Times, 1991): a clinical trial with n=338 HIV infected subjects

Development of AIDS by AZT Use and Race

		AIDS Symptoms	
Race	AZT Use	yes	no
white	yes	14	93
	no	32	81
black	yes	11	52
	no	12	43

- binary response Y: AIDS developed or not
- two factors $X=$ AZT Use: received immediately or not, and Z=Race: white or black
- multiple logistic model: $logit\left[\pi(x,z)\right] = \alpha + \beta_1 X + \beta_2 Z$

Multiple logistic regression model:

$$logit\left[\pi(x,z)\right] = \alpha + \beta_1 X + \beta_2 Z$$

$\Leftrightarrow \pi(x,z) = \pi(i,k)$, $i = 1, 2$ and $k = 1, 2$:
ANOVA Representation $logit[\pi(i,k)] = \alpha + \beta_i^X + \beta_k^Z$

- Coding Scheme I (eg. SAS): $\beta_1^X = \beta_1$, the coef to X; $\beta_2^X = 0$;
  The log OR of AIDS with AZT use vs not is $\beta_1^X - \beta_2^X = \beta_1$.
- Coding Scheme II (eg. R): $\beta_1^X = 0$ and $\beta_2^X = \beta_1$, the coef to X;
  The log OR of AIDS with AZT use vs not is $\beta_1^X - \beta_2^X = -\beta_1$.
- Coding Scheme III (eg. ANOVA-type): $\beta_1^X = -\beta_2^X \Leftrightarrow$
  $\beta_1^X + \beta_2^X = 0$, and $\beta_1^X = \beta_1$, the coef to X;
  The log OR of AIDS with AZT use vs not is $\beta_1^X - \beta_2^X = 2\beta_1$.

If consider two factor interactions ... ...

Multiple logistic regression model:

$$logit\left[\pi(x,z)\right] = \alpha + \beta_1 X + \beta_2 Z + \beta_{12} XZ$$

$\pi(x,z) = P(Y=1|X=x, Z=z) \Leftrightarrow \pi(x,z) = \pi(i,k),\ i = 1,2$ and $k = 1,2$:

ANOVA Representation

$$logit[\pi(i,k)] = \alpha + \beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$$

2.2.3E Example of the crab study (cont'd)

Revisit II: A 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 $logit(\pi) = \alpha + \beta_1 x_1 + \beta_{21} x_{21} + \beta_{22} x_{22} + \beta_{23} x_{23}$

- For $x_1 = 26.3$, a medium-light crab's predicted probability is $\hat{\pi}(26.3, 0, 0, 0) = .715$ and 95% CI $(.392, .908)$:
  - calculate 95% CI for $logit(\pi) = \alpha + \beta_1 26.3$:

$$(\hat{\alpha}+\hat{\beta}_1 26.3) \pm 1.96 * \sqrt{\hat{var}(\hat{\alpha}) + \hat{var}(\hat{\beta}_1) * 26.3^2 + 2\hat{cov}(\hat{\alpha}, \hat{\beta}_1) * 26.3}$$

$$\Longrightarrow (-.44, 2.28)$$

  - calculate 95% CI for π:

$$\left(\frac{e^{-.44}}{1 + e^{-.44}}, \frac{e^{2.28}}{1 + e^{2.28}}\right) = (.392, .908)$$

Revisit II: A multiple logistic regression analysis – using the regression outcome

- For $x_1 = 26.3$ (average width) and a medium-light crab, its odds is $.715/.285 = 2.51$
- For $x_1 = 26.3$ and a dark crab, its prob of having satellites is .399 and odds is $.399/(1 - .399) = 0.66$
- The odds ratio of having satellites for medium-light vs dark crabs with average width is $2.51/.66 = 3.8$

  $\Longrightarrow$ a dark crab of average width is less likely than a medium-light crab to have satellites.

Revisit II: An alternative multiple logistic regression analysis
(Quantitative Treatment of the Ordinal Predictor, color)

$color = x_2 = 1, 2, 3, 4$ for the color categories and fit
$logit(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2$

- using the regression outcome
  The estm for β_1 and β_2 along with their ASE values show strong evidence of an effect for each.
- goodness-of-it?

to add in more predictors? how about two predictors' interactions

Revisit III: 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.

To be studied in Chp 5 systematically.

2.3A Generalized linear models: Introduction

- Ordinary Linear Regression Models (OLM)

  To study $Y \leftarrow X, Z$? with continuous response Y and two explanatory variables: $Y = \alpha + \beta X + \gamma Z + \eta XZ + \epsilon$ with $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$

  - *R: glmout< −glm(Y∼X*Z, family=gaussian)*

- Logistic Regression Models (Logit)

  To study $Y \leftarrow X, Z$? with binary response $Y = 1$, or 0:
  $logit\left[\pi(x,z)\right] = \alpha + \beta X + \gamma Z + \eta XZ$ with
  $P(Y = 1|X = x, Z = z) = \pi(x,z)$ and
  $Y \sim Bernoulli\left(\pi(x,z)\right)$

  - *R: glmout< −glm(Y∼X*Z, family=binomial)*

2.3B Generalized linear models: Components

What common features in the examples of regression models, OLM, Logit? Recall how to conduct the analysis with *R: glm(formula, family)* $\Longrightarrow$
GOAL: to study how $Y \leftarrow X_1, \ldots, X_K$

Generalized Linear Models:

- **Random Component.** response r.v. Y with $\mu(x_1, \ldots, x_k) = E(Y|x_1, \ldots, x_k)$ to be examined

- **Systematic Component.** $\alpha + \beta_1 x_1 + \ldots + \beta_K x_K$
  Some x_k can be based on others: e.g. $x_3 = x_1 x_2$.

- **Link Function.** $g(\mu) = \alpha + \beta_1 x_1 + \ldots + \beta_K x_K$
  The link function $g(\cdot)$ links the *random componet* through its mean and the *systematic component*. **More on GLM later**

What will we study next?

1. Introdution and Preparation

2. Analysis with Binary Variables (Chp1-2)

3. Analysis with Multicategory Variables (Chp 3)

- **3.1 Revisit to Analysis with Contingency Table (Chp 3.1-2)**
  - **3.1.1 Review of two-way contingency tables**
  - **3.1.2 Analysis of $I \times J$ contingency table**
  - **3.1.3 Multi-way contingency tables (supplementary)**
- **3.2 Analysis with Multicategory Response (Chp 3.3-5)**