What to do today (01/25)?

1. Introdution and Preparation

2. Analysis with Binary Variables (Chp 1-2)

2.1 Analysis with binary variables I (Chp 1)

2.1.1 On one binary variable (Chp1.1)

2.1.2 On two binary variables (Chp1.2)

2.1.2A Introduction

2.1.2B Inference with two binary variables

2.1.2C Further topics

2.2 Analysis with binary response II (Chp 2)

2.2.1 Regression models (Chp 2.1, Chp2.2.1)

2.2.2 Simple logistic regression analysis (Chp2.2.2-7)

2.1.2C Further topics: Matched pairs data

Example. Crossover Study to Compare Drug with Placebo.
86 subjects randomly assigned to receive drug then placebo or else placebo then drug. Binary response (success, failure) for each treatment.

Treatment	success	failure	Total
drug	61	25	86
placebo	22	64	86

- *Methods so far (e.g., X^2, G^2 test of indep., CI for the log-OR θ from logistic regression) assume independent samples: they are inappropriate for dependent samples (e.g., same subjects in each sample, which yield matched pairs).*

Example. Crossover Study to Compare Drug with Placebo. (cont'd)
To reflect the dependence, display the data as 86 (pair) observations rather than 2×86 observations:

		Placebo		
		success	failure	Total
Drug	success	12	49	61
	failure	10	15	25
		22	64	86

To compare the drug vs placebo with the data $\Longrightarrow$ to check if $\pi_{1+} - \pi_{+1}$, the difference in success rate between the drug and placebo, is zero.

2.1.2C Further topics: Matched pairs data

In general, with a 2×2 table, there is *marginal homogeneity* if $\pi_{1+} - \pi_{+1} = 0$

- Since $\pi_{1+} - \pi_{+1} = [\pi_{11} + \pi_{12}] - [\pi_{11} + \pi_{21}] = \pi_{12} - \pi_{21}$, *marginal homogeneity* $\Leftrightarrow \pi_{12} = \pi_{21}$ (*symmetry*).

- With H_0: *marginal homogeneity*, $\frac{\pi_{12}}{\pi_{12}+\pi_{21}} = \frac{1}{2}$. Thus, under H_0 and conditioanl on $n^* = N_{12} + N_{21}$, $N_{12} \sim Bin(n^*, \frac{1}{2})$ with mean $E(N_{12}) = \frac{n^*}{2}$ and std.dev$=\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}$

McNemar's Test on H_0: $\pi_{1+} = \pi_{+1}$
By normal approximation to binomial, for large n^* and under H_0,

$$Z = \frac{N_{12} - \frac{n^*}{2}}{\sqrt{n^*(\frac{1}{2})(\frac{1}{2})}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \sim N(0, 1)$$

or, equivalently

$$Z^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi^2(1)$$

CI for $\pi_{1+} - \pi_{+1}$:

$$[\hat{\pi}_{1+} - \hat{\pi}_{+1}] \pm z_{1-\alpha/2} SE_{(\hat{\pi}_{1+} - \hat{\pi}_{+1})}$$

with $SE = \frac{1}{n}\sqrt{(n_{12} + n_{21}) - \frac{(n_{12} - n_{21})^2}{n}}$.

Example. Crossover Study to Compare Drug with Placebo. (cont'd)

Test on $H_0 : \pi_{1+} = \pi_{+1}$ vs $H_1 : \pi_{1+} > \pi_{+1}$

$$z_{obs} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{49 - 10}{\sqrt{49 + 10}} = 5.1$$

$\Longrightarrow p < 0.001$: extremely strong evidence for the probability of success is higher with drug compared to placebo.

95% CI for $\pi_{1+} - \pi_{+1}$

$$[\frac{n_{11} + n_{12}}{n} - \frac{n_{11} + n_{21}}{n}] \pm 1.96SE = 0.453 \pm 1.96(.075) = (0.31, 0.60)$$

Conclude that with 95% confidence the probability of success with drug is between .31 and .60 higher than it with placebo.

2.1.2C Further topics: Larger contingency tables

Recall ... ...

- **Contingency Table**

  a table with cells contain *frequency counts* of outcome according to categorical variables

- **2-Way Contingency Table**

  a table with cells contain *frequency counts* of outcome according to 2 categorical variables

- **I$\times$ J Contingency Table**

  a table with cells contain *frequency counts* of outcome according to 2 categorical variables, one with I levels and one with J levels

Larger contingency tables ...

- **I$\times$ J Contingency Table** with $I > 2$ and/or $J > 2$
- **K-Way Contingency Table** with $K > 2$

2.1.2C Further topics: Larger contingency tables

Where do young people live?

TABLE 23.1 Young adults by age and living arrangement

LIVING ARRANGEMENT	AGE (YEARS)				TOTAL
	19	20	21	22	
Parents' home	324	378	337	318	1357
Another person's home	37	47	40	38	162
Your own place	116	279	372	487	1254
Group quarters	58	60	49	25	192
Other	5	2	3	9	19
Total	540	766	801	877	2984

TABLE 23.2 Percents of each age group who have each living arrangement (read down columns)

LIVING ARRANGEMENT	AGE (YEARS)			
	19	20	21	22
Parents' home	60.0	49.3	42.1	36.3
Another person's home	6.9	6.1	5.0	4.3
Your own place	21.5	36.4	46.4	55.5
Group quarters	10.7	7.8	6.1	2.9
Other	0.9	0.3	0.4	1.0
Total	100.0	99.9	100.0	100.0

Test on independence between age and living place:

$$\mathcal{X}^2 = \sum_{\text{all } 5\times 4 \text{ cells}} \frac{(\text{observed count - expected count})^2}{\text{expected count}} = \sum_{i,j} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \sim \chi^2(df)$$

$df = (5-1) * (4-1)$ and $\mathcal{X}^2_{obs} = 193.56$

$\Longrightarrow$ p-value$=P(\mathcal{X}^2 > \mathcal{X}^2_{obs}) = P(\mathcal{X}^2 > 193.56) < 0.001$

Conclusion. There is very strong evidence that living arrangements of young people are not the same across the groups of age 19,20,21, and 22.

Further, the percentage table shows how young people become more independent as they grow older.

To conduct the test by LRT?

2.2 Analysis with binary response (Chp 2)

2.2.1 Regression models: Overview

Recall how we have studied so far about the association of variables X, Y ... ...

- One type of approaches ...
  Don't assign the different roles to them, and consider

- Another type of approaches ...
  Choose one as the dependent (response) variable and the other as the independent (explanatory) variable, and conduct a regression analysis

2.2.1 Regression models: Logistic regression model

Example. Female Horseshoe Crabs and their Satellites ("An Introduction to Categorical Data Analysis" A. Agresti, 2007)

- **Study Goal:** To investigate factors that affect whether a female crab had any males (satellites) residing nearby her.
- **Study Variables:**
  - Response: number of satellites
  - Explanatory Variables: female crab's color; spine condition; weight; carapace width.
- **Study Data ... ...**

Data of Female Horseshoe Crab Study

C	S	W	Wt	Sa	C	S	W	Wt	Sa
2	3	28.3	3.05	8	3	3	22.5	1.55	0
1	1	26.0	2.30	9	3	3	24.8	2.10	0
3	3	26.0	2.60	4	2	3	23.8	2.10	0
		...	...				...	...	

Source: *Ethology*, 102: 1-21, 1996

Note: C=color (1=light medium, 2=medium, 3=dark medium, 4=dark); S=spine condition (1=both good, 2=one worn/broken, 3=both worn/broken); W=carapace width (cm); Wt=weight(kg); Sa=number of satellites.

- How does "the number of satellites" depend on "color" / "width"?

How does Y depend on X?

How about to consider a simplified problem:

- Consider the response variable Y as $Y = 0$ or 1 if "the number of satellites" is 0 or > 0
- Consider only one explanatory variable X="color"

Thinking ... ... To conduct a regression analysis? What regression model to consider?

- How about $Y = \alpha + \beta X + \epsilon$?
  not working well ... $\Longrightarrow$
- How about $\pi(x) = \alpha + \beta x$ with
  $\pi(x) = E(Y|X = x) = P(Y = 1|X = x)$?
  still not working well ... $\Longrightarrow$
- To consider $\log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x$?

2.2.1 Regression models: Logistic regression model

General Setting:

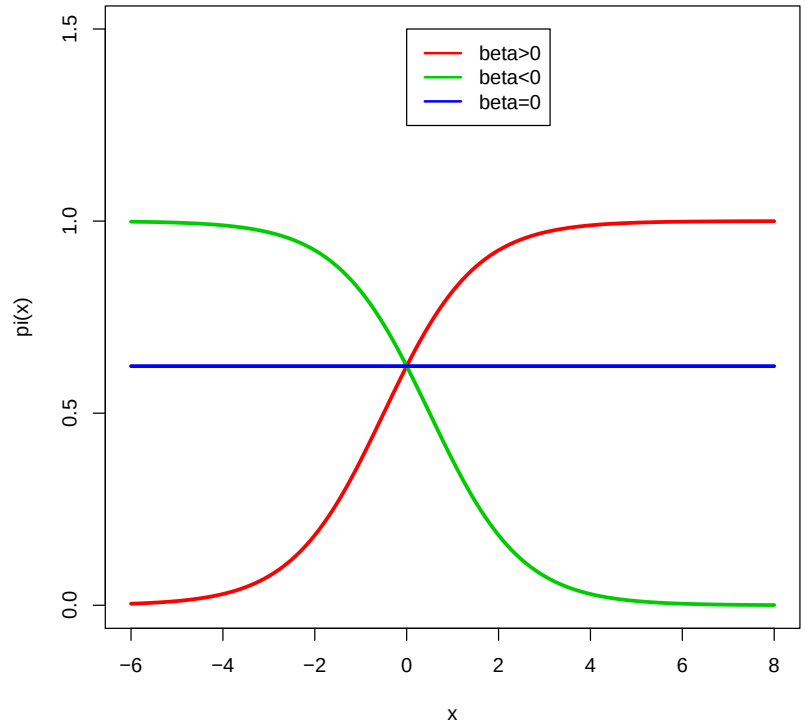
- a binary response Y (e.g. success (1)/failure (0)); one explanatory variable X
- to study how the prob of success depends on X
  That is, to find out about the function
  $\pi(x) = E(Y|X = x) = P(Y = 1|X = x)$
- Here $Y|X = x \sim Bernoulli\big(\pi(x)\big)$

Simple Logistic Regression Model:

$$logit\big[\pi(x)\big] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x$$

equivalent to $\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$.

2.2.2A Simple logistic regression analysis: Modeling and Interpretation



Properties:

- Always in between 0 and 1 regardless of α and β's values
- If $\beta = 0$, $\pi(x) = \frac{e^{\alpha}}{1+e^{\alpha}}$; if $\beta > 0 (< 0)$, $\pi(x) \uparrow (\downarrow)$ as $x \uparrow$
- S-shaped – often desirable and meeting the common sense

2.2.2A Simple logistic regression analysis: Modeling and Interpretation

When $\mathit{logit}\left[\pi(x)\right] = \alpha + \beta x$**,** $\pi(x)$ **is not a linear function.**

- **Linear Approximation:** $\pi(x) \approx a + \left[\frac{d\pi(x)}{dx}\right]x$,

$$\frac{d\pi(x)}{dx} = \beta\pi(x)\left[1 - \pi(x)\right]$$

β determines the increase/decrease of $\pi(x)$ as $x \uparrow$:

- with $\beta = 0$, $\frac{d\pi(x)}{dx} = 0$ and thus $\pi(x) = \mathit{constant} \Longrightarrow Y \perp\!\!\!\perp X$
- with a fixed β, the largest (steepest) linear slope is when $\pi(x) = 1/2$ and thus $x = -\alpha/\beta$ (the median effective level)
- with a fixed β, if $\pi(x) \approx 0$ or $1 \Longrightarrow$ the flattest slope

2.2.2A Simple logistic regression analysis: Modeling and Interpretation

When $\mathit{logit}\left[\pi(x)\right] = \alpha + \beta x$**,** $\pi(x)$ **is not a linear function.**

- **Odds Ratio Interpretation:** If X is binary, e.g male (1)/female (0), the odds ratio of success with male vs female:

$$\theta = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} = \frac{\exp(\alpha+\beta*1)}{\exp(\alpha+\beta*0)} = \exp(\beta) \Longrightarrow \log(\theta) = \beta$$

2.2.2B Simple logistic regression analysis: Statistical Inference

Modeling: With the simple logistic regression model,
$logit\left[\pi(x)\right] = \alpha + \beta x$,
$\Longrightarrow Y|X = x \sim Bernoulli(\pi(x))$

Available data: data from a study with n independent individuals:
$\{(X_i, Y_i) : i = 1, \ldots, n\}$.

What to do?

- **estimate** α, β
- **test on hypothese about** α, β
- **estimate** $\pi(x)$
- **model checking: is "**$logit\left[\pi(x)\right] = \alpha + \beta x$**" a good model?**

2.2.2B Simple logistic regression analysis: Estimation of α and β

the likelihood function:

$$L(\alpha, \beta) = \prod_{i=1}^{n} \pi(x_i; \alpha, \beta)^{Y_i} \left[1 - \pi(x_i; \alpha, \beta)\right]^{1-Y_i}$$

the log-likelihood function:

$$\log L(\alpha, \beta) = \sum_{i=1}^{n} \left[Y_i(\alpha + \beta X_i) - \log(1 + e^{\alpha+\beta X_i}) \right]$$

Solving the equations to obtain the MLE $\hat{\alpha}$ and $\hat{\beta}$:

$$\frac{\partial \log L(\alpha, \beta)}{\partial \alpha} = \sum_{i=1}^{n} \left[Y_i - \pi(X_i; \alpha, \beta) \right] = 0; \quad \frac{\partial \log L(\alpha, \beta)}{\partial \beta} = \sum_{i=1}^{n} \left[Y_i - \pi(X_i; \alpha, \beta) \right] X_i = 0$$

To implement it by *R*.

2.2.2B Simple logistic regression analysis: Estimation of α and β

Properties of the MLE $\hat{\alpha}$ and $\hat{\beta}$:

- consistent estm
- asymptotic normality: $\hat{\beta} \sim N(\beta, AV_{\hat{\beta}})$ and $\hat{\alpha} \sim N(\alpha, AV_{\hat{\alpha}})$, as $n >> 1$

More on the MLE ...

- MLE can be obtained with iterative numerical methods ... ...
- Estimation of the asymptotic variance $AV_{\hat{\alpha}}$ and $AV_{\hat{\beta}}$: $\left[AV_{\hat{\theta}}(\theta)\right]^{-1} = nI(\theta) \approx -\partial^2 \log L(\theta)/\partial\theta^2$.
- Confidence intervals for α, β: the Wald type e.g. $\hat{\beta} \pm 1.96 ASE_{\hat{\beta}}$ with $ASE_{\hat{\beta}} = \sqrt{AV_{\hat{\beta}}}$

2.2.2B Simple logistic regression analysis: Testing

To test on $H_0 : \beta = \beta_0$:

- the Wald-type test statistic: if $n >> 1$, under H_0, approximately

$$Z = \frac{\hat{\beta} - \beta_0}{ASE_{\hat{\beta}}} \sim N(0, 1)$$

  equivalently $Z^2 \sim \chi^2(1)$

- the LRT statistic: if $n >> 1$, under H_0, approximately

$$-2 \log \left[\frac{L(\hat{\alpha}_0, \beta_0)}{L(\hat{\alpha}, \hat{\beta})} \right] \sim \chi^2(1)$$

- the score test statistic: if $n >> 1$, under H_0, approximately

$$\frac{\partial \log L(\alpha, \beta)}{\partial \beta}\Big|_{\beta=\beta_0} \sim N(0, I(\alpha, \beta_0))$$

2.2.2B Simple logistic regression analysis: Estimation of $\pi(x;\alpha,\beta)$

With the MLE of α,β, the MLE of $\pi(x;\alpha,\beta)$ is

$$\hat{\pi}(x) = \pi(x;\hat{\alpha},\hat{\beta}) = \frac{\exp(\hat{\alpha}+\hat{\beta}x)}{1+\exp(\hat{\alpha}+\hat{\beta}x)}$$

How about an approximate 95% CI of $\pi(x;\alpha,\beta)$?

- $\hat{\pi}(x) \pm (1.96)(SE_{\hat{\pi}(x)})$, or
- $logit^{-1}\left[(\hat{\alpha}+\hat{\beta}x) \pm (1.96)(SE_{\hat{\alpha}+\hat{\beta}x})\right]$

$var(\hat{\alpha}+\hat{\beta}x) = var(\hat{\alpha}) + var(\hat{\beta})x^2 + 2cov(\hat{\alpha},\hat{\beta})x$ and

$$var(\hat{\alpha},\hat{\beta}) = \begin{bmatrix} var(\hat{\alpha}) & cov(\hat{\alpha},\hat{\beta}) \\ cov(\hat{\alpha},\hat{\beta}) & var(\hat{\beta}) \end{bmatrix} \approx \left[-\partial^2 \log L(\alpha,\beta)/\partial(\alpha,\beta)^2\right]^{-1}.$$

What will we do next week?

A. New material to study:

- *1. Introduction and Preparation*
- **2. Analysis with Binary Variables (Chp 1-2)**
  - *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)**
    - *2.2.4 Generalized linear models (Chp2.3)*

B. Midterm 1:

- **A brief review**
- **Exam: AQ3005; 10:30-11:20am**