What to do today (02/01)?

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

2.1 Analysis with binary variables I (Chp 1)

2.2 Analysis with binary response II (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)

- **Midterm 1: 10:30 - 11:20**

2.2.2 Simple logistic regression analysis

- a binary response Y (e.g. success (1)/failure (0)); one explanatory variable X
- to find out about the function $\pi(x) = P(Y = 1|X = x)$

Simple Logistic Regression Model:

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

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

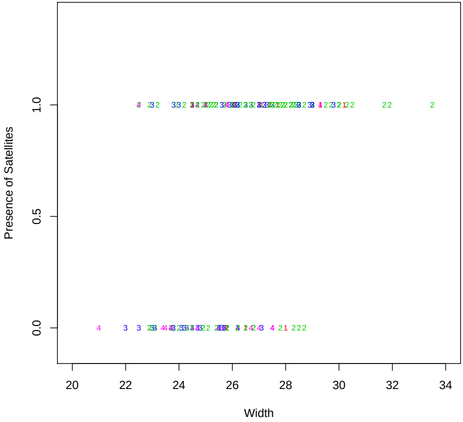
Properties:

- 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 commen sense

Example. Female Horseshoe Crabs and their Satellites: Revisit I

To consider a simplified problem: the response variable $Y = 1$ or 0 for if presence of satellite; one predictor X="width"

How does Y depend on X? What is $\pi(x) = P(Y = 1|X = x)$?

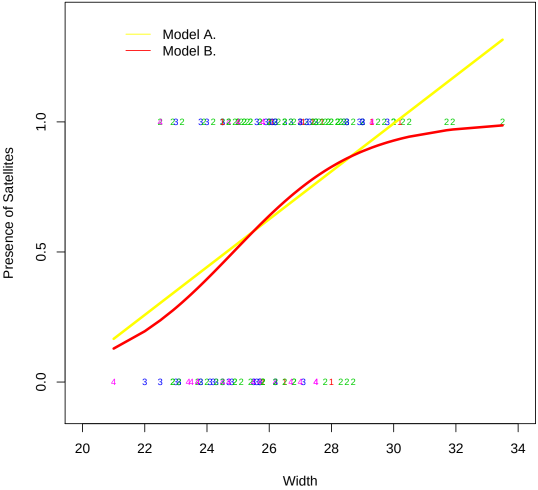


Example. Female Horseshoe Crabs and their Satellites: Revisit I

with **Model A.** $\hat{\pi}(x) = -1.766 + 0.092x$

with **Model B.** the simple logistic regression model

$logit\left[\pi(x;\alpha,\beta)\right] = \alpha + \beta x$: $\hat{\alpha} = -12.351$, $\hat{\beta} = 0.497$

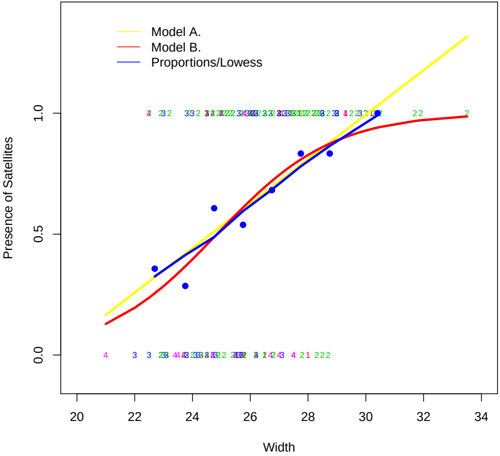


Example. Female Horseshoe Crabs and their Satellites: Revisit I

With **Model B.**, $\hat{\pi}(x) = \frac{\exp(-12.35+.497x)}{1+\exp(-12.35+.497x)}$

- $\hat{\beta} > 0$: $\hat{\pi}(x_{min}) = 0.129$ and $\hat{\pi}(x_{max}) = 0.987$
- the median effective level (the steepest slope of the curve, $\pi(x; \alpha, \beta) = 1/2$): $EL_{50} = -\hat{\alpha}/\hat{\beta} = 24.8$
- at the sample mean width of $\bar{x} = 26.5$cm, $\hat{\pi} = .674$ and the slope $\hat{\beta}\hat{\pi}(1-\hat{\pi}) = 0.11$
- the odds ratio for each cm increases in width: $\exp(\hat{\beta}) = 1.64$ e.g., $x = 26.3$ vs $x = 27.3 \Longrightarrow$ odds=2.07 vs 3.40.
- 95% CI for the size of the width's effect: $\hat{\beta} \pm Z_{.025}ASE = (0.298, 0.697)$
- Testing (significant effect?): $H_0 : \beta = 0$, $Z_{obs} = \hat{\beta}/ASE = 4.9$
- 95% CI for $\pi(26.5)$: $(.61, .77)$, obtained by getting CI for $\alpha + \beta x$ and then logit^{-1}-transfering.

Example. Female Horseshoe Crabs and their Satellites: Revisit I



Is the simple logistic regression model provides a good fit to the data? Are there any other predictors?

2.2.3 Multiple logistic regression analysis

- a binary response Y (e.g. success (1)/failure (0)); several explanatory variables $X_1, \ldots, X_K$
- 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:

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

equivalently to $\pi(x_1, \ldots, x_K; \alpha, \beta_1, \ldots, \beta_K) = \frac{\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}{1+\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}$.

- always in between 0 and 1 regardless of α and β's values
- if $\beta_1, \ldots, \beta_K = 0$, $\pi(x_1, \ldots, x_K) = \frac{e^\alpha}{1+e^\alpha}$; if $\beta_1 > 0 (< 0)$, $\pi(x_1, \ldots, x_K) \uparrow (\downarrow)$ as $x_1 \uparrow$ and fixed other predictors
- S-shaped [with a predictor] – often desirable and meeting the commen sense

2.2.3A Modeling and interpretation

Multiple Logistic Regression Model:

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

equivalently to
$\pi(x_1,\ldots,x_K;\alpha,\beta_1,\ldots,\beta_K) = \frac{\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}{1+\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}$.

For example, when $X_1 = 1$ or 0, β_1 is the effect of X_1 on the log odds of $Y = 1$, controlling the other explanatory variables

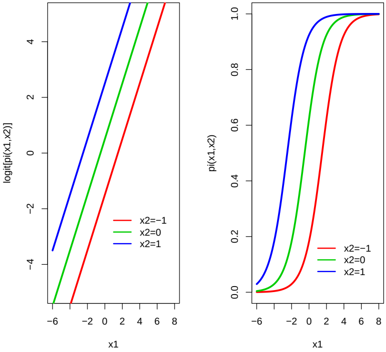
- Testing on $H_0 : \beta_1 = 0 \Rightarrow$ whether X_1 is a significant predictor in the presence of the other ones

2.2.3A Modeling and interpretation

Multiple Logistic Regression Model:

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

equivalently to $\pi(x_1,\ldots,x_K;\alpha,\beta_1,\ldots,\beta_K) = \frac{\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}{1+\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}$.
For example, when $K = 2$, X_2 is fixed at diff values, the curves $\pi(x_1, x_2)$ of x_1 are "parallel"



2.2.3B Statistical inference

Suppose the data from a study are

$$\{(y_i, x_{i1}, \ldots, x_{iK}) : i = 1, \ldots, n\}$$

all the individuals follow the same multiple logistic regression model:

$$Y_i \big| X_{i1}, \ldots, X_{iK} \sim Bernoulli(\pi_i)$$

$$\pi_i = \pi(x_{i1}, \ldots, x_{iK}; \alpha, \beta_1, \ldots, \beta_K) = \frac{\exp(\alpha+\beta_1 x_{i1}+\ldots+\beta_K x_{iK})}{1+\exp(\alpha+\beta_1 x_{i1}+\ldots+\beta_K x_{iK})}$$

What to do with the model next?

- estimation of $\alpha, \beta_1, \ldots, \beta_K$
- testing for hypothese about $\alpha, \beta_1, \ldots, \beta_K$
- estimation of $\pi(x_1, \ldots, x_K; \alpha, \beta_1, \ldots, \beta_K)$
- model checking and variable selection (* to study in Chp 5)

Estimation of $\alpha, \beta_1, \ldots, \beta_K$
the likelihood function:

$$L(\alpha, \beta_1, \ldots, \beta_K) = \prod_{i=1}^{n} \pi_i^{Y_i} \big(1 - \pi_i\big)^{1-Y_i}$$

with $\pi_i = \pi(x_{i1}, \ldots, x_{iK}; \alpha, \beta_1, \ldots, \beta_K) = \frac{\exp(\alpha+\beta_1 x_{i1}+\ldots+\beta_K x_{iK})}{1+\exp(\alpha+\beta_1 x_{i1}+\ldots+\beta_K x_{iK})}$

$\Longrightarrow$ the MLE $(\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_K)$: consistent and asymptotically normal

$$N\left(\begin{pmatrix} \alpha \\ \beta_1 \\ \ldots \\ \beta_K \end{pmatrix}, \Sigma_{(K+1)\times(K+1)}\big(\alpha, \ldots, \beta_k\big) \right)$$

$\Longrightarrow$ confidence interval/region: for example,

- for each of the parameters: $\hat{\beta}_1 \pm z_{0.975} SE_{\hat{\beta}_1}$
- joint (simutaneous) CI/CR: e.g.

$$\left\{ \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} : \left(\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \right)' \Sigma^{-1} \left(\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \right) \leq c \right\}$$

Estimation of $\pi(x_1, \ldots, x_K)$
Recall $logit\big(\pi(x_1, \ldots, x_K)\big) = \alpha + \beta_1 x_1 + \ldots + \beta_K x_K$ is equivalent to $\pi(x_1, \ldots, x_K; \alpha, \beta_1, \ldots, \beta_K) = \frac{\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}{1+\exp(\alpha+\beta_1 x_1+\ldots+\beta_K x_K)}$.

- Point Estimator (MLE).

$$\hat{\pi}(x_1, \ldots, x_K) = \frac{\exp(\hat{\alpha} + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_K x_K)}{1 + \exp(\hat{\alpha} + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_K x_K)}.$$

- CI. $\hat{\pi} \pm z_{0.975} SE_{\hat{\pi}}$;
  an alternative method:
    - first to obtain a CI for $\alpha + \beta_1 x_1 + \ldots + \beta_K x_K$ using the estms of α, β's and the estm of $\Sigma_{(K+1)\times(K+1)}$
    - then to take the $logit^{-1}$-transformation to attain a CI for $\pi(x_1, \ldots, x_K)$

Hypothesis Testing

For example, when $K = 2$, $H_0 : \beta_2 = 0$ vs H_1 : otherwise (regarding the specified multiple logistic regression model for $\pi(x_1, x_2)$)

- Approach 1: using the MLE of β_2 and

$$Z = \frac{\hat{\beta}_2 - \beta_{20}}{SE_{\hat{\beta}_2}} \sim N(0, 1)$$

  approximately under H_0 when $n >> 1$

- Approach 2: using the LRT

$$\mathcal{G}^2 = -2 \log \left[\frac{\max L(\alpha, \beta_1, 0)}{\max L(\alpha, \beta_1, \beta_2)} \right] \sim \chi^2(1)$$

  approximately under H_0 when $n >> 1$

Hypothesis Testing (cont'd)

For another example, when $K = 2$, $H_0 : \beta_1 = 0, \beta_2 = 0$ vs H_1 : otherwise (regarding the specified multiple logistic regression model for $\pi(x_1, x_2)$)

- Approach 1: using the MLE of β_1, β_2 and the Wald type test with

$$\left(\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right)' \Sigma^{-1} \left(\begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} - \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right) \sim \chi^2(2)$$

  approximately under H_0 when $n >> 1$

- Approach 2: using the LRT

$$\mathcal{G}^2 = -2 \log \left[\frac{\max L(\alpha, 0, 0)}{\max L(\alpha, \beta_1, \beta_2)} \right] \sim \chi^2(2)$$

approximately under H_0 when $n >> 1$

What to do next?

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)**

Marked Midterm 1 papers will be returned on Feb 6