What to do today (Mar 13)?

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**

6. *Additional Topics (Chp 6)*

5A. Model Selection and Evaluation: Overview

Model selection in regression.

- to identify an appropriate probability model
- to identify an appropriate set of explanatory variables in the appropriate model: *variable selection*

Model evalaution in regression.

- residuals: graphical assessment of residuals
- goodness-of-fit
- influence: influence measures such as *leverage*

5A. Model Selection and Evaluation: Overview

Model comparison criteria.

$$IC(k) = -2\log\left(L(\hat{\beta}|data)\right) + kr$$

with sample size n and r (non-redundant) parameters.

- **Akaike's Information Criterion** (AIC):

$$AIC = IC(2) = -2\log\left(L(\hat{\beta}|data)\right) + 2r$$

- **Corrected AIC**:

$$AIC_c = IC(\frac{2n}{n-r-1}) = -2\log\left(L(\hat{\beta}|data)\right) + \frac{2n}{n-r-1}r$$

- **Bayesian Information Criterion** (BIC; Schwarz criterion):

$$BIC = IC(\log(n)) = -2\log\left(L(\hat{\beta}|data)\right) + \log(n)r$$

5A. Model Selection and Evaluation: Overview

Variable selection.

Applying a method for model checking "dynamically" to achieve the "best" model of a class of models, with a specified criterion at each step

- **forward selection** starting from a model without any predictor, and adding predictor to the regression model one by one
- **backward elimination** starting from a regression model with all potential predictors, and removing not important predictor from the model one by one
- **forward-backward or backward-forward selection** combinations of forward and backward selection

5B. Model Selection and Evaluation: in the simple logistic regression

Statistical inference in the simple logistic regression.

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:
$\{(X_i, Y_i) : i = 1, \ldots, n\}$.

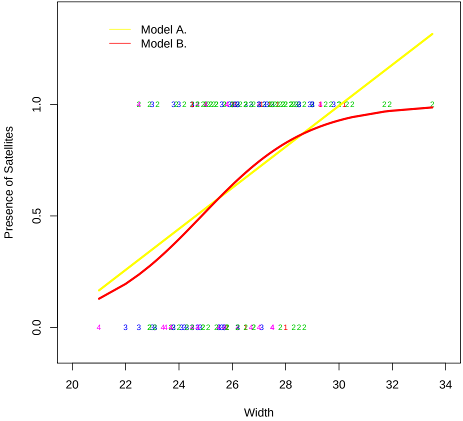
What to do?

- *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?**

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 $\pi(x) = P(Y = 1|X = x)$?



Fitted model $logit\left[\pi(x)\right] = -12.35 + 0.497x$

5B. Model Selection and Evaluation: in the simple logistic regression

Case (i) If X is categorical with I levels

- The study data can be summarized by an $I \times 2$ table, as Y is binary.
- To diagnose the simple logistic regression model:
  to test on $H_0 : logit\big[\pi(x)\big] = \alpha + \beta x$ vs H_1: otherwise
- If the cell counts in the table ≥ 5 and the overall total $n >> 1$,
  $\Longrightarrow$ applications of the Pearson's χ^2-test and LRT-test with the two way contingency table:

5B. Model Selection and Evaluation: in the simple logistic regression

Case (i) If X is categorical with I levels

Under H_0 and $df = I - 2$,

$$\mathcal{K}^2 = \sum \frac{(observed - fitted)^2}{fitted} \sim \chi^2(df);$$

$$\mathcal{G}^2 = 2 \sum (observed) \log\left(\frac{observed}{fitted}\right) \sim \chi^2(df)$$

$fitted = \hat{\pi}(x) * \left(\#subjects\ in\ x\ group\right)$ or
$fitted = \left(1 - \hat{\pi}(x)\right) * \left(\#subjects\ in\ x\ group\right)$

5B. Model Selection and Evaluation: in the simple logistic regression

Case (ii) If X is continuous or discrete but with large I^* levels

- Group the values of X into a finite number of I such that $n/I \geq 5$
  - the larger I is, the less coarsening but the $I \times 2$ table's cell counts are smaller
  - the smaller I is, the more coarsening and thus more away from the really value
- Form the $I \times 2$ table and then use the approaches in Case (i)

different grouping/partitioning $\Rightarrow$ *different conclusion?*

5B. Model Selection and Evaluation: in the simple logistic regression

Likelihood-Ratio Model Comparison Test.
In general, to compare a "smaller" model to a "larger" model in good fit: H_0 : *model* M_0 vs H_1 : *model* M_1 with $M_0 \subset M_1$

For example, $M_1 : \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$ and $M_0 : \pi(x) = \frac{e^{\alpha}}{1+e^{\alpha}}$

The LRT-test statistic

$$\mathcal{G}^2(M_0|M_1) = -2 \log \left(\frac{\max L_{M_0}}{\max L_{M_1}} \right) \sim \chi^2(df)$$

approximately under H_0, with $df = df_{M_1} - df_{M_0}$.

5B. Model Selection and Evaluation: in the simple logistic regression

Likelihood-Ratio Model Comparison Test.
Often, to obtain $\mathcal{G}^2(M_0|M_1) = \mathcal{G}^2(M_0|M_s) - \mathcal{G}^2(M_1|M_s)$

- M_s is the "saturated" model: the model gives the perfect fit – its number of parameters is the same as the df of the data
- $\mathcal{G}^2(M_0|M_s)$ and $\mathcal{G}^2(M_1|M_s)$ are referred to as the deviances of M_0 and M_1 (to M_s), denoted by $\mathcal{G}^2(M_0)$ and $\mathcal{G}^2(M_1)$ sometime

more about this later ... ...

5B. Model Selection and Evaluation: in the simple logistic regression

Residuals for the Logit Model.

The Pearson's $\chi^2 - test\ \mathcal{K}^2 = \sum \frac{(observed - fitted)^2}{fitted}$ is the same as

$$\sum e_k^2 : \quad e_k = \frac{s_k - n_k \hat{\pi}_k}{\sqrt{n_k \hat{\pi}_k (1 - \hat{\pi}_k)}}$$

$n_k = \#x_i = k$ with s_k successes and $\pi_k = P(success|X = k)$.
e_k's: the Pearson's residuals

- If $n_k \uparrow$, $e_k \sim N(0, var(e_k))$ under H_0 approximately: $var(e_k) < 1$.
- If $e_k \geq 2 \rightarrow$ possible lack of fit.
- Graphical displays of e_k's: residual plots

5B. Model Selection and Evaluation: in the simple logistic regression

Residuals for the Logit Model.
Often the adjusted residuals are used:

$$e_k^* = \frac{e_k}{\sqrt{1 - h_k}} = \frac{s_k - n_k \hat{\pi}_k}{\sqrt{var(s_k - n_k \hat{\pi}_k)}} \sim N(0, 1)$$

approximately uner H_0

Diagnostic measures of influence:

- values of e_k's or e_k^*
- outliers: extrem values?
  - deleting outliers to obtain a better fit?
  - taking them as important signals?

Example. Female Horseshoe Crabs and their Satellites Revisit I (cont'd)

- Model checking A: Is the logistic model appropriate?
  - Classifying width values into 8 groups: $(0, 23.25], (23.25, 24.25], ..., (28.25, 29.25], (29.25, \infty)$
  - Form 8×2 table
  - Obtain $\mathcal{K}^2_{obs} = 5.3$ and $\mathcal{G}^2_{obs} = 6.2$ (df=6)
  - Conclusion: *no evidence of lack of fit*

Q: original 66×2? Can be $\mathcal{K}^2$-test or $\mathcal{G}^2$-test directly applied?

Example. Female Horseshoe Crabs and their Satellites Revisit I (cont'd)

- Model checking B: Can the term of X in the logistic model be omitted?
  With the 8×2 table:
  $H_0 : logit(\pi(x)) = \alpha$ vs $H_1 : logit(\pi(x)) = \alpha + \beta x$
  $G^2(M_1 \big| M_0) = 34.0 - 6 = 28 (df = 1)$

strong evidence against H_0

Example. Female Horseshoe Crabs and their Satellites Revisit I (cont'd)

- Model checking B: Can the term of X in the logistic model be omitted?

With the original data (66×2 table):
$G^2(M_1|M_0) = 225.76 - 194.45 = 31.3(df = 1)$ *strong evidence against* H_0

```
tmpy<-ifelse(ex.crab[,5]>0,1,0)
tmpout<-glm(tmpy~ex.crab[,3], family=binomial)
summary(tmpout)

==================================
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0281  -1.0458   0.5480   0.9066   1.6942

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 194.45  on 171  degrees of freedom
AIC: 198.45
===============================
```

Example. Female Horseshoe Crabs and their Satellites Revisit I (cont'd)

Compared to testing $H_0 : \beta = 0$?

- 8×2 table: $\frac{\hat{\beta}}{SE_{\hat{\beta}}} = 0.46316/0.09787 = 4.732 \Rightarrow p < 0.001$
- 66×2 table: $\frac{\hat{\beta}}{SE_{\hat{\beta}}} = 0.4972/0.1017 = 4.887 \Rightarrow p < 0.001$

Example. Female Horseshoe Crabs and their Satellites: More Revisits ...

About What Aspects?

- More than one factors, including qualitative predicators
- The original response: not binary but count of satellites

$\Longrightarrow$ their model evaluation and selection?

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.

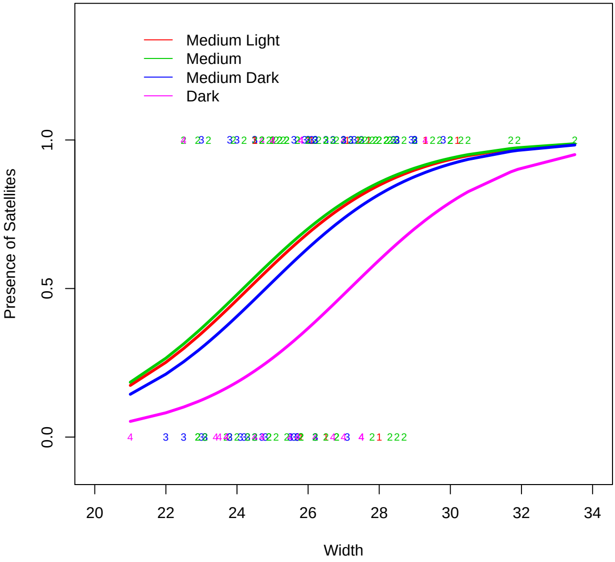
- **Criteria for Variable Selection:**
  - **classical criterion** selecting/keeping only predictors according to a pre-specified significance level
  - **Information criteria**: e.g. to achieve the min AIC, or corrected AIC or BIC

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.

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

```
--------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
```

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**

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

6. *Additional Topics (Chp 6)*