Outline

- Strategies for model selection
  - [illegible]
    - Model checking

How many predictors can you use?

Data are unbalanced on Y if y=1 occurs relatively few times or if y=0 occurs relatively few times.

- One guideline: suggest at least 10 outcomes of each type should occur for every X term. e.g, if y=1 only 30 times out of n=1000 the model should contain no more than 3 predictors.

Example: Horseshoe crabs revisited

4 predictors: color,spine,weight(wt),and width(w).

response:

$$y = \begin{cases} 1 & \text{if at least one satellites} \\ 0 & \text{o.w} \end{cases}$$

model : $logit[P(Y = 1)] = a + b_1 \quad wt + b_2 \quad w + b_3 \quad c_1 + b_4 \quad c_2 + b_5 \quad c_3 + b_6 \quad s_1 + b_7 \quad s_2$

$H_0 : b_1 = b_2 = b_3 = b_4 = b_5 = b_6 = b_7 = 0$

$G^2(M) = 40.6$ with df=7(p-value<0.0001).

Example: Horseshoe crabs revisited

4 predictors: color,spine,weight(wt),and width(w).

response:

$$y = \begin{matrix} 1 & \text{if at least one satellites} \\ 0 & \text{o.w} \end{matrix}$$

model : $logit[P(Y = 1)] = a + b_1 \quad wt + b_2 \quad w + b_3 \quad c_1 + b_4 \quad c_2 + b_5 \quad c_3 + b_6 \quad s_1 + b_7 \quad s_2$

$H_0 : b_1 = b_2 = b_3 = b_4 = b_5 = b_6 = b_7 = 0$

$G^2(M) = 40.6$ with df=7(p-value

Example: Horseshoe crabs revisited

Coefficients:

	Estimate	Std. Error	z value	$Pr(>\|z\|)$	
(Intercept)	-9.2728	3.8275	-2.423	0.01541	*
color1	1.6087	0.9344	1.722	0.08512	.
color2	1.5057	0.5659	2.661	0.00779	**
color3	1.1198	0.5924	1.890	0.05874	.
spine1	-0.4003	0.5016	-0.798	0.42479	
spine2	-0.4963	0.6287	-0.790	0.42973	
weight	0.8257	0.7025	1.175	0.23984	
width	0.2631	0.1949	1.350	0.17703	

Example: Backward elimination for horseshoe crabs

				Models	Deviances	
Model	Deviance	df	AIC	Compared	Difference	p-value
1	173.7	155	209.7	-	-	-
2	186.6	166	200.6	(2) (1)	12.9(df=11)	0.2999
3a	208.8	167	220.8	(3a)		

AIC,model seG04seandco,model

2278.d
ModelDeviance AIC

Classification Table of horseshoe crabs (P.142)

1. Choosing cutoff points:

Classification Table of horseshoe crabs (P.142)

1. Choosing cutoff points:
   - For this data, the proportion of Y=5 is $\frac{111}{111+62} = 0.642$
   - Cutoff points p_0 may be 0.5 or 0.642
2. For width and color as predictors, fit a logistic regression. Y ~ W+C

Classification Table of horseshoe crabs (P.142)

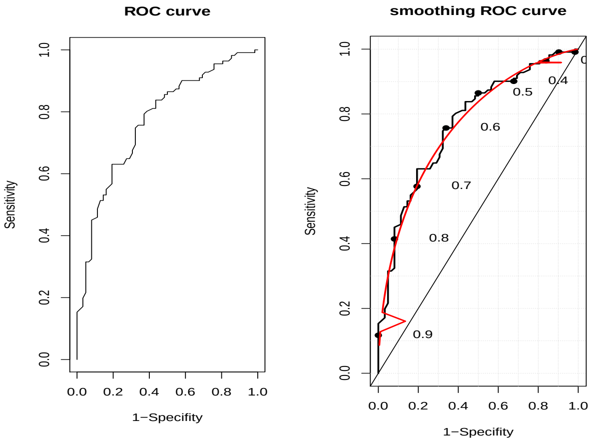
ROC curves

- Plot of (1-specificity,sensitivity)
- More informative than classification table,for use all possible p_0 to summarize predictive power.
- The ROC curves pass (0,0) and (1,1)
- Concordance index = Area of the ROC curve

ROC curves

For p_0=0.642, ROC curves pass (0.5,0.865) (*)

ROC curves



And the concordance index = 0.7714

Models concordance index

Model	Concordance index
CS+CW+SW	0.8059
C+S+W	0.7749
C+S	0.6694
S+W	0.7427
C+W	0.7714
C	0.6386
W	0.7424
C=dark+W	0.7720
None	0.5000

Models concordance index

Model	Concordance index
CS+CW+SW	0.8059
C+S+W	0.7749
C+S	0.6694
S+W	0.7427
C+W	0.7714
C	0.6386
W	0.7424
C=dark+W	0.7720
None	0.5000

$\Rightarrow$ Choose a model simple, but concordance index higher.

L-R Model Comparison Test

Example: AZT on AIDS symptoms table

AZT	Symptoms	Freq	Pi

Example: AZT on AIDS symptoms table

AZT	Symptoms	Race	Freq	Pi_hat	Fitted
Yes	Yes	White	14	0.1496	16
No	Yes	White	32	0.2654	30
Yes	Yes	Black	11	0.1427	9
No	Yes	Black	12	0.2547	14
Yes	No	White	93	0.8504	91
No	No	White	81	0.7346	83
Yes	No	Black	52	0.8573	54
No	No	Black	43	0.7453	41

- $G^2(M) = 1.3835$ with df=1, and P-value = 0.2395059

⇒ Don't reject H_0, model fit well.

Grouped Data and Ungrouped Data

- The group data are the totals of successes and failures at

Standardized residual

$$\text{standardized residual} = \frac{y_i - n_i\hat{p}_i}{SE} = \frac{y_i - n_i\hat{p}_i}{\sqrt{(n_i\hat{p}_i(1-\hat{p})(1-h_i))}}$$

- h_i : the observation's leverage, it's element from the

- We would hope the admissions decision is independent of gender, and we treat Y_{ik} as independent binomial variates for n_{ik} trials with success probabilities p_{ik}.
- The model with no gender effect, given department

- We would hope the admissions decision is independent of gender, and we treat Y_{ik} as independent binomial variates for n_{ik} trials with success probabilities p_{ik}.
- The model with no gender effect, given department

$$\Longrightarrow \ logit(p_{ik}) = a + b_k^D$$

- Simpson's Paradox ?

Table 5.5. Table Rel

Influence Diagnostics for Logistic Regression

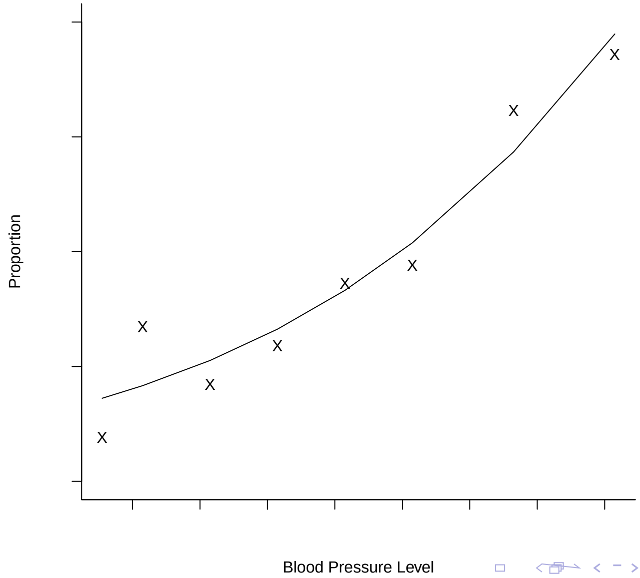
The observation could be a single binary response or a binomial response for a set of subjects(grouped data).

When the certain observation is deleted

- the change divided by it's standard error :*Dfbeta*
- The change in a joint *CI* displacement diagnostic : c.
- The change X^2 or G^2 in G-O-F statistics.

Example:Heart Disease and Blood Pressure

- p^{i}_i


Proportion
X
X
X
X
X
X
X
X
Blood Pressure Level