Categorical Data Analysis

Chapter 6

Deyuan Li
School of Management
Fudan University

Fall 2015

Outline

6.1 Strategies in model selection

Model selection becomes harder as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions.

Two goals of model selection:

- the model should be complex enough to fit the data well;
- the model should be simple to interpret, smoothing rather than overfitting the data.

6.1 Strategies in model selection

Most models are designed to answer certain questions, which guide the choice of model terms:

- Confirmatory（验证性）analysis use a restrict set of models; e.g., a study hypothesis about an effect may be tested by comparing models with and without that effect.
- exploratory（探索性）studies may provide clues about the dependence structure and raise questions for future research.

In either case, it is helpful first to study the effect on Y of each predictor by itself using graphics for a continuous predictor or a contingency table for a discrete predictor. This gives a "feel" for the marginal effects.

6.1 Strategies in model selection

For unbalanced data with relatively few responses of one type, one guideline to limit the number of explanatory variables (x-terms) is *at least 10 outcomes of each type should occur for every x-term*.

For instance, if $y = 1$ only 30 times out of $n = 1000$, the model should contain no more than about 3 x-terms.

Such guideline is approximate. This does not mean that if you have 500 outcomes of each type you are well served by a model with 50 x-terms.

6.1 Strategies in model selection

A model with several predictors may suffer from multicollinearity（多重共线性）, i.e., correlations among predictors making it seem that no one predictor is important when all the others are in the model.

A predictor may seem to have little effect because it overlaps considerably with other predictors in the model, i.e., itself being predicted well by the other predictors.

Deleting such a redundant predictor can be helpful to reduce standard errors of other estimated effects.

6.1.1 Horseshoe crab example revisited

The horseshoe crab data set in Table 4.3 has four predictors: color (4 categories), spine condition (3 categories), weight and width of the carapace shell.

Outcome: The crab has satellites ($y = 1$) or not ($y = 0$).

Preliminary model: 4 predictors, only main effects

$$\begin{aligned} \text{logit}[P(Y=1)] &= \alpha + \beta_1 \,\text{weight} + \beta_2 \,\text{width} \\ &+ \beta_3 c_1 + \beta_4 c_2 + \beta_5 c_3 + \beta_6 s_1 + \beta_7 s_2 \,, \end{aligned}$$

treating color (c_i) and spine condition (s_j) as qualitative (factors), with dummy variables for the first 3 colors and the first 2 spine conditions. Table 6.1 shows results.

TABLE 6.1 Computer Output from Fitting Model with All Main Effects to Horseshoe Crab Data

Testing Global Null Hypothesis: BETA = 0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	40.5565	7	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	−9.2734	3.8378	5.8386	0.0157
weight	0.8258	0.7038	1.3765	0.2407
width	0.2631	0.1953	1.8152	0.1779
color 1	1.6087	0.9355	2.9567	0.0855
color 2	1.5058	0.5667	7.0607	0.0079
color 3	1.1198	0.5933	3.5624	0.0591
spine 1	−0.4003	0.5027	0.6340	0.4259
spine 2	−0.4963	0.6292	0.6222	0.4302

The likelihood-ratio test of $H_0 : \beta_1 = \cdots = \beta_7 = 0$, i.e., Y is jointly independent of the predictors, gives a test statistic of 40.6, with df$= 7$ and $P < 0.0001$. $\Rightarrow$ Extremely strong evidence that at least one explanatory variable has an effect.

6.1.1 Horseshoe crab example revisited

Although the overall test is highly significant, the individual effects are not very significant (except for color 2).
$\Rightarrow$ A warning sign of multicollinearity.

Weight and width have a strong correlation (0.887)!

For practical purposes they are equally good predictors, but it is nearly redundant to use them both.

$\Rightarrow$ Further analysis uses width (W) with color (C) and spine condition (S) as predictors.

6.1.1 Horseshoe crab example revisited

For simplicity, we symbolize models by their highest-order terms, regarding C and S as factors.

For instance, $(C + S + W)$ denotes a model with main effects, whereas $(C + S * W)$ denotes a model that has those main effects plus an $S \times W$ interaction.

It is not usually sensible to consider a model with interaction but not the main effects that make up that interaction.

6.1.2 Stepwise procedures

Two stepwise procedures for model selection.

- Forward selection
  - Add terms sequentially until further additions do not improve the fit.
  - At each stage it selects the term(s) giving the greatest improvement in fit.
  - The minimum *P*-value for testing the term in the model is a sensible criterion, since reductions in deviance for different terms may have different df values.

6.1.2 Stepwise procedures

- Backward selection
  - Begin with a complex model and sequentially removes terms.
  - At each stage it selects the term(s) for which its removal has the least damaging effect on the model (e.g., largest P-value).
  - The process stops when any further deletion leads to a significantly poorer fit.

- For predictor (factor) with more than two categories, the process should consider the entire factor at any stage, rather than just individual dummy variables.
  $\Rightarrow$ Add or drop the entire factor rather than just one of its dummies. Otherwise, the result depends on the coding.

6.1.3 Backward elimination for horseshoe crab example

TABLE 6.2 Results of Fitting Several Logistic Regression Models to Horseshoe Crab Data

Model	Predictors[a]	Deviance G^2	df	AIC	Models Compared	Deviance Difference	Corr. $r(y,\hat{\mu})$
1	$(C*S*W)$	170.44	149	212.4	—	—	
2	$(C*S + C*W + S*W)$	173.68	155	209.7	(2)−(1)	3.2 (df = 6)	
3a	$(C*S + S*W)$	177.34	158	207.3	(3a)−(2)	3.7 (df = 3)	
3b	$(C*W + S*W)$	181.56	161	205.6	(3b)−(2)	7.9 (df = 6)	
3c	$(C*S + C*W)$	173.69	157	205.7	(3c)−(2)	0.0 (df = 2)	
4a	$(S + C*W)$	181.64	163	201.6	(4a)−(3c)	8.0 (df = 6)	
4b	$(W + C*S)$	177.61	160	203.6	(4b)−(3c)	3.9 (df = 3)	
5	$(C + S + W)$	186.61	166	200.6	(5)−(4b)	9.0 (df = 6)	
6a	$(C + S)$	208.83	167	220.8	(6a)−(5)	22.2 (df = 1)	
6b	$(S + W)$	194.42	169	202.4	(6b)−(5)	7.8 (df = 3)	
6c	$(C + W)$	187.46	168	197.5	(6c)−(5)	0.8 (df = 2)	0.452
7a	(C)	212.06	169	220.1	(7a)−(6c)	24.5 (df = 1)	0.285
7b	(W)	194.45	171	198.5	(7b)−(6c)	7.0 (df = 3)	0.402
8	$(C = \text{dark} + W)$	187.96	170	194.0	(8)−(6c)	0.5 (df = 2)	0.447
9	None	225.76	172	227.8	(9)−(8)	37.8 (df = 2)	0.000

[a] C, color; S, spine condition; W, width.

6.1.3 Backward elimination for horseshoe crab example

Formula for Model (C*S*W):

6.1.3 Backward elimination for horseshoe crab example

- The deviance (G^2) compares the model to the saturated model (Model 1). It is not approximately chi-squared when a predictor is continuous.
- However, if the two models differ by a modest number of parameters, the difference in deviance is the likelihood-ratio statistic $-2(L_0 - L_1)$ comparing the models, and it has an approximate null chi-squared distribution.
- We use backward procedure to select a model. At each step, we test only the highest-order terms for each predictor, since it is inappropriate to remove a main effect term if the model has interactions involving that term.

6.1.3 Backward elimination for horseshoe crab example

Model 1: symbolized by $(C * S * W)$.

The most complex model, with main effects for each predictors (C, S, W), two-factor interactions $(C * S, C * W, S * W)$ and three-factor interaction $(C * S * W)$.

Model 2: symbolized by $(C * S + C * W + S * W)$.

Comparing Model 1 with Model 2, the likelihood-ratio statistic equals 3.2 with df$=3$ and $P = 0.36$.
$\Rightarrow$ The three-factor interaction is not needed.

6.1.3 Backward elimination for horseshoe crab example

Model 3: consider three models: Models 3a, 3b and 3c. Model 3c ($C * S + C * W$) gives essentially the same fit as Model 2 (the deviance difference = 0), so we drop the $S * W$ interaction.

Model 4: comparing Model 4a ($S + C * W$) with Model 3c gives a deviance difference of 8.0 on df= 6 ($P = 0.24$).
$\Rightarrow$ The $C * S$ interaction can be dropped.

Comparing Model 4b ($W + C * S$) with Model 3c gives a deviance difference of 3.9 on df= 3 ($P = 0.27$).
$\Rightarrow$ The $C * W$ interaction can be dropped.

Model 5: proceeding from Model 4b, and dropping the remaining $C * S$ interaction results in Model 5 ($C + S + W$) with a deviance difference of 9.0 on df= 6 ($P = 0.17$).

6.1.3 Backward elimination for horseshoe crab example

Model 6:

a Dropping W results in Model 6a ($C + S$). Compared with Model 5, the effect of W is nonnegligible with a deviance difference of 22.2 on df$= 1$. $\Rightarrow$ Keep W in model.

b Dropping C results in Model 6b ($S + W$). Compared with Model 5, the deviance difference is 7.0 on df$= 3$ ($P = 0.07$). $\Rightarrow$ Keep C in model.

c Dropping S results in Model 6c ($C + W$). Compared with Model 5, S has only little effect. $\Rightarrow$ Drop S.

6.1.3 Backward elimination for horseshoe crab example

Model 8: binary color coding model in Section 5.4.6.

Compared with Model 6c, the deviance difference equals 0.5 on df$=2$. $\Rightarrow$ Model 8 fits essentially as well as Model 6c.

Further simplification (Models 7a, 7b and 9) results in large increases in deviance and is unjustified.

6.1.4 AIC, model selection, and the correct model

Akaike information criterion (AIC)

- Select the model that minimizes

$$\text{AIC} = -2(\text{maximized log likelihood} - \text{number of parameters in model}).$$

- This penalizes a model for having many parameters.

Table 6.2 also lists AIC values:

- Model 6c ($C + W$) has the smallest AIC (197.5).
- The Model 8 has an even smaller AIC (194.0). Need to balance Model 8 and Model 6c.

6.1.5 Using causal hypotheses to guide model building

Selection procedures are helpful tools, but model-building process should utilize theory and common sense.

Often, a time ordering among the variables suggests possible causal relationships.

Analyzing a certain sequence of models helps to investigate those relationships. Table 6.3: $2 \times 2 \times 2 \times 2$ table.

TABLE 6.3 Marital Status by Report of Pre- and Extramarital Sex (PMS and EMS)

		Gender							
		Women				Men			
	PMS:	Yes		No		Yes		No	
Marital Status	EMS:	Yes	No	Yes	No	Yes	No	Yes	No
Divorced		17	54	36	214	28	60	17	68
Still married		4	25	4	322	11	42	4	130

Source: G. N. Gilbert, *Modelling Society* (London: George Allen & Unwin, 1981). Reprinted with permission from Unwin Hyman Ltd.

6.1.5 Using causal hypotheses to guide model building

G = gender (women or men),
P = premarital sex (yes or no),
E = extramarital sex (yes or no),
M = current marital status (divorced or still married).

The time points at which responses on the four variables occur suggests the following ordering of the variables:

Any of these is an explanatory variable when a variable listed to its right is the response.

6.1.5 Using causal hypotheses to guide model building

Figure 6.1 shows one possible causal structure:

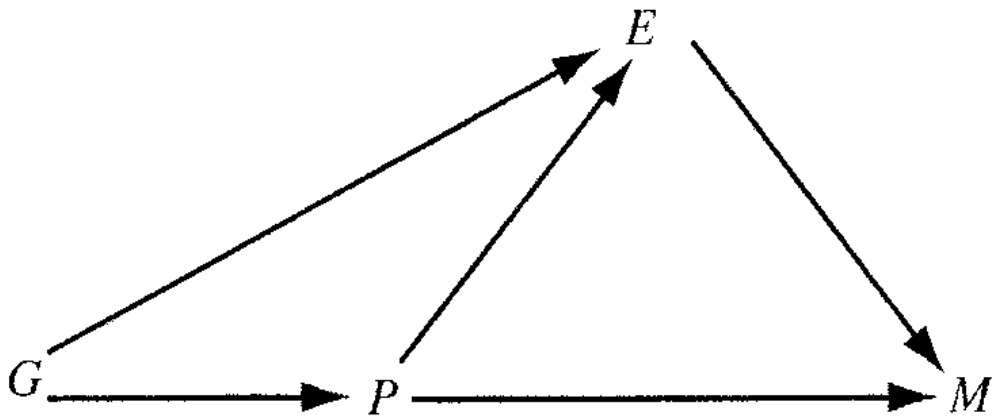

FIGURE 6.1 Causal diagram for Table 6.3.

- A variable at the tip of an arrow is a response for a model at some stage.
- The explanatory variables have arrows pointing to the response, directly or indirectly.

6.1.5 Using causal hypotheses to guide model building

Stage 1: P is the response.

Figure 6.1 predicts that G has a direct effect on P.

Table 6.4 shows results:

TABLE 6.4 Goodness of Fit of Various Models for Table 6.3[a]

Stage	Response Variable	Potential Explanatory	Actual Explanatory	G^2	df
1	P	G	None	75.3	1
			(G)	0.0	0
2	E	G, P	None	48.9	3
			(P)	2.9	2
			$(G + P)$	0.0	1
3	M	G, P, E	$(E + P)$	18.2	5
			$(E*P)$	5.2	4
			$(E*P + G)$	0.7	3

[a] P, premarital sex; E, extramarital sex; M, marital status; G, gender.

6.1.5 Using causal hypotheses to guide model building

Compared with the model without G, the $G^2 = 75.3$ on 1 df.
$\Rightarrow$ Strong evidence of a GP association.
$\Rightarrow$ The model of independence of these variables is inadequate.

The sample odds ratio for their marginal table is

$$\frac{(\text{Women with PMS}) \times (\text{Men without PMS})}{(\text{Women without PMS}) \times (\text{Men with PMS})}$$
$$= \frac{(17+54+4+25) \times (17+68+4+130)}{(36+214+4+322) \times (28+60+11+42)} = 0.27.$$

$\Rightarrow$ The estimated odds of premarital sex for females are 0.27 times that for males.

6.1.5 Using causal hypotheses to guide model building

Stage 2: E is the response. Figure 6.1 predicts that

- P and G have direct effects on E;
- G has an indirect effect on E, through its effect on P.
- If G has only an indirect effect on E, the model with P alone as a predictor is adequate.
  $\Rightarrow$ controlling P, E and G are conditionally independent.

Table 6.4 shows that: after adding G to a model already containing P as a predictor, the G^2 drops by 2.9 on 1 df. $\Rightarrow$ Weak evidence occurs that G had a direct as well as an indirect effect on E.

For this model, the estimated EP conditional odds ratio is 0.40.

6.1.5 Using causal hypotheses to guide model building

Stage 3: M is the response. Figure 6.1 predicts that

- E has a direct effect on M;
- P has a direct effect and an indirect effect on M (through E);
- G has an indirect effect on M (through P and E).

Table 6.4 shows results:

- The model with main effects of E and P, fits poorly.
- The model ($E * P$) with additional $E \times P$ interaction fits much better (G^2 decrease of 13.0, df$=1$).
- The model ($E * P + G$) that also has a main effect for G fits slightly better yet (G^2 decrease of 4.5, df$=1$).

Outline

6.2.1 Pearson, deviance, and standardized residuals

With categorical predictors, it is useful to form residuals to compare observed and fitted counts.

Let

y_i = the binomial variate for n_i trials at setting i ($i = 1, \ldots, N$) of the predictors.

$\hat{\pi}_i$ = model estimate of $P(Y = 1)$.

Then $n_i \hat{\pi}_i$ is the fitted number of successes.

6.2.1 Pearson, deviance, and standardized residuals

Pearson residual: (Section 4.5.5, equation (4.36))

For a GLM with binomial random component, the Pearson residual for this fit is

The Pearson statistic for testing the model fit satisfies

$$X^2 = \sum_{i=1}^{N} e_i^2 \quad \Rightarrow \text{each } e_i^2 \text{ is a component of } X^2.$$

Note that: e_i is not $N(0,1)$. Its variance is smaller than 1.
In fact $E(\sum_i e_i^2)/N \approx \nu/N < 1$ with $\nu =$ df of X^2.

6.2.1 Pearson, deviance, and standardized residuals

Standardized Pearson residual: (Section 4.5.5, equation (4.38))

where $\{\hat{h}_i\}$ is from an estimated hat matrix.

Note that: $|r_i| > |e_i|$ and r_i is approximately $N(0,1)$ when the model holds.

6.2.1 Pearson, deviance, and standardized residuals

Deviance residual: (Section 4.5.5, equation (4.35))

Such residual uses components of the G^2 fit statistic.

For observation i, the deviance residual（偏离度残差）is

$$\sqrt{d_i} \times \text{sign}(y_i - n_i \hat{\pi}_i) \,,$$

where

The deviance residual also tends to be less variable than $N(0,1)$ and can be standardized.
$\Rightarrow$ Need to be standardized!

6.2.2 Heart disease example

Response: Developed coronary heart disease during a six-year follow-up period or not.

Predictor: Blood pressure, grouped into 8 categories.

Model fitting: (see Table 6.5)

TABLE 6.5 Standardized Pearson Residuals for Logit Models Fitted to Data on Blood Pressure and Heart Disease

Blood Pressure	Sample Size	Observed Heart Disease	Fitted		Residual	
			Indep. Model	Linear Logit	Indep. Model	Linear Logit
< 117	156	3	10.8	5.2	−2.62	−1.11
117–126	252	17	17.4	10.6	−0.12	2.37
127–136	284	12	19.7	15.1	−2.02	−0.95
137–146	271	16	18.8	18.1	−0.74	−0.57
147–156	139	12	9.6	11.6	0.84	0.13
157–166	85	8	5.9	8.9	0.93	−0.33
167–186	99	16	6.9	14.2	3.76	0.65
> 186	43	8	3.0	8.4	3.07	−0.18

Source: Data from Cornfield (1962).

6.2.2 Heart disease example

Let π_i be the probability of heart disease for category i.

Model 1: $\text{logit}(\pi_i) = \alpha$, treats the response as independent of blood pressure.

- Some residuals are large.
- The model fits poorly ($G^2 = 30.0$, $X^2 = 33.4$, df $= 7$).
- A plot of the residuals show an increasing trend.

Model 2: $\text{logit}(\pi_i) = \alpha + \beta x_i$, a linear logit model as suggested by the residuals of Model 1.

The $\{x_i\}$ are scores for blood pressure categories. We used scores $(111.5, 121.5, 131.5, 141.5, 151.5, 161.5, 176.5, 191.5)$.

- The trend in residuals disappears for this model.
- Only the second category shows some evidence of lack of fit.

6.2.2 Heart disease example

TABLE 6.6 Residuals Reported in SAS for Heart Disease Data of Table 6.5[a]

			Observation Statistics			
Observ	disease	n	blood	Reschi	Resdev	StReschi
1	3	156	111.5	−0.9794	−1.0617	−1.1058
2	17	252	121.5	2.0057	1.8501	2.3746
3	12	284	131.5	−0.8133	−0.8420	−0.9453
4	16	271	141.5	−0.5067	−0.5162	−0.5727
5	12	139	151.5	0.1176	0.1170	0.1261
6	8	85	161.5	−0.3042	−0.3088	−0.3261
7	16	99	176.5	0.5135	0.5050	0.6520
8	8	43	191.5	−0.1395	−0.1402	−0.1773

[a] Reschi, Pearson residual; StReschi, adjusted residual.

6.2.2 Heart disease example

Comments on Table 6.6:

- The Pearson residuals (Reschi), deviance residuals (Resdev) and standardized Pearson residuals (StReschi) show similar results.
- One relatively large residual is not surprising. With many residuals, some may be large purely by chance.
- The overall fit statistics ($G^2 = 5.9$, $X^2 = 6.3$ with df$= 6$) do not indicate problems.

Plots: To check lack of fit, we can compare observed and fitted proportions by either plotting them against each other (Q-Q plot) or plotting both of them against explanatory variables.

6.2.2 Heart disease example

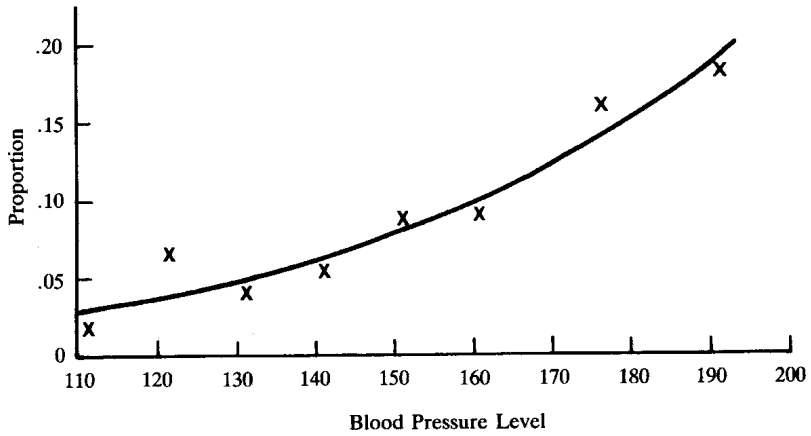

FIGURE 6.2 Observed and predicted proportions of heart disease for linear logit model.

The fit seems decent.

Remarks: Studying residuals helps us understand (1) why a model fits poorly and (2) where there is lack of fit in a generally good-fitting model (see the next example).

6.2.3 Graduate admissions example

TABLE 6.7 Data Relating Admission to Gender and Department for Model with No Gender Effect

	Females		Males		Std. Res		Females		Males		Std. Res
Dept	Yes	No	Yes	No	(Fem,Yes)	Dept	Yes	No	Yes	No	(Fem,Yes)
anth	32	81	21	41	−0.76	ling	21	10	7	8	1.37
astr	6	0	3	8	2.87	math	25	18	31	37	1.29
chem	12	43	34	110	−0.27	phil	3	0	9	6	1.34
clas	3	1	4	0	−1.07	phys	10	11	25	53	1.32
comm	52	149	5	10	−0.63	poli	25	34	39	49	−0.23
comp	8	7	6	12	1.16	psyc	2	123	4	41	−2.27
engl	35	100	30	112	0.94	reli	3	3	0	2	1.26
geog	9	1	11	11	2.17	roma	29	13	6	3	0.14
geol	6	3	15	6	−0.26	soci	16	33	7	17	0.30
germ	17	0	4	1	1.89	stat	23	9	36	14	−0.01
hist	9	9	21	19	−0.18	zool	4	62	10	54	−1.76
lati	26	7	25	16	1.65						

Source: Data courtesy of James Booth.

6.2.3 Graduate admissions example

Response variable: admitted or not (A),

Explanatory variables: department (D, total 23 departments), gender (G).

Let

n_{ik} = number of applications for gender i in department k,
y_{ik} = number of admitted for gender i in department k,
π_{ik} = probability of admission for gender i in department k.

We treat $\{Y_{ik}\}$ as independent Bin(n_{ik}, π_{ik}).

Consider the admission decision is independent of gender, i.e.
Model: logit$(\pi_{ik}) = \alpha + \beta_k^D$.

The model fits rather poorly ($G^2 = 44.7$, $X^2 = 40.9$, df$= 23$).

6.2.3 Graduate admissions example

Departments with large standardized Pearson residuals reveal the reason for the lack of fit.

Significantly more females were admitted than the model predicts in the astronomy (astr) and geography (geog) departments, and fewer in the psychology (psyc) department. Without these 3 departments, the model fits reasonably well ($G^2 = 24.4, X^2 = 22.8, \text{df} = 20$).

For the complete data, adding a gender effect to the model does not provide an improved fit ($G^2 = 42.4, X^2 = 39.0, \text{df} = 22$), because those 3 departments have associations in different directions and of greater magnitude than other departments.

6.2.4 Influence diagnostics for logistic regression

Whenever a residual indicates that a model fits an observation poorly, it can be informative to delete that observation and refit the model to the remaining observations.

Diagnostic tools:

1) *Plot*. Plots of ordered residuals against normal percentiles.

2) *Analyze an observation's influence on parameter estimates and fit statistics*

6.2.4 Influence diagnostics for logistic regression

Influential measures for each observation include:

1. For each model parameter β_i, the change in the parameter estimate when the observation is deleted. This change divided by its standard error, is called Dfbeta, i.e., $(\hat{\beta}_i - \hat{\beta}_{i(j)})/\text{SE}(\hat{\beta}_i - \hat{\beta}_{i(j)})$, where $\hat{\beta}_{i(j)}$ is the parameter estimate when observation j is deleted, $j = 1, \ldots, N$ (total number of observations).
2. A measure of change in a joint confidence interval (CI) for the parameters produced by deleting the observation. Denote this CI displacement diagnostic is denoted by c.
3. The change in X^2 or G^2 goodness-of-fit statistics when the observation is deleted.

For each measure, the larger the value, the greater the influence.

6.2.4 Influence diagnostics for logistic regression

TABLE 6.8 Diagnostic Measures for Logistic Regression Models Fitted to Heart Disease Data

Blood Pressure	*Dfbeta*	*c*	Pearson X^2 Diff.	Likelihood-Ratio G^2 Diff.	Pearson X^2 Diff.[a]	Likelihood-Ratio G^2 Diff.[a]
111.5	0.49	0.34	1.22	1.39	6.86	9.13
121.5	−1.14	2.26	5.64	5.04	0.02	0.02
131.5	0.33	0.31	0.89	0.94	4.08	4.56
141.5	0.08	0.09	0.33	0.34	0.55	0.57
151.5	0.01	0.00	0.02	0.02	0.70	0.66
161.5	−0.07	0.02	0.11	0.11	0.87	0.80
176.5	0.40	0.26	0.42	0.42	14.17	10.83
191.5	−0.12	0.02	0.03	0.03	9.41	6.73

[a] Independence model; other values refer to model with blood pressure predictor.
Source: Data from Cornfield (1962).

Outline

6.3 Inference about conditional associations in 2 × 2 × K tables

Table 6.9: Results of a clinical trial comparing a binary response (success or failure) between two treatment groups (active drug or control) in eight strata (centers).

TABLE 6.9 Clinical Trial Relating Treatment to Response for Eight Centers

		Response				
Center	Treatment	Success	Failure	Odds Ratio	μ_{11k}	var(n_{11k})
1	Drug	11	25	1.19	10.36	3.79
	Control	10	27			
2	Drug	16	4	1.82	14.62	2.47
	Control	22	10			
3	Drug	14	5	4.80	10.50	2.41
	Control	7	12			
4	Drug	2	14	2.29	1.45	0.70
	Control	1	16			
5	Drug	6	11	∞	3.52	1.20
	Control	0	12			
6	Drug	1	10	∞	0.52	0.25
	Control	0	10			
7	Drug	1	4	2.0	0.71	0.42
	Control	1	8			
8	Drug	4	2	0.33	4.62	0.62
	Control	6	1			

Source: Beitler and Landis (1985).

6.3.1 Using logit models to test conditional independence

For a binary response Y, we study the effect of a binary predictor X, controlling for a qualitative covariate Z.

Let $\pi_{ik} = P(Y = 1|X = i, Z = k)$. Consider the model

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z, \quad i = 1, 2, \quad (6.4)$$

$k = 1, ..., K$, where $x_1 = 1$ and $x_2 = 0$.

This model assumes that the XY conditional odds ratio is the same at each category of Z, namely $\exp(\beta)$.

The null hypothesis of XY conditional independence is $H_0 : \beta = 0$. The Wald statistic is $(\hat{\beta}/\text{SE})^2$.

The likelihood-ratio statistic is the difference between G^2 statistics for the full model and the reduce model $\text{logit}(\pi_{ik}) = \alpha + \beta_k^Z$.

6.3.2 Cochran-Mantel-Haenszel test of conditional independence

Mantel and Haeszel (1959)

Consider H_0 : conditional independence in $2 \times 2 \times K$ tables, i.e. $H_0 : \theta_{XY(1)} = ... = \theta_{XY(K)} = 1$.

Focusing on retrospective studies of disease, they treated response (column) marginal totals as fixed.

Thus, in each 2×2 partial table k ($k = 1, \ldots, K$) of cell counts $\{n_{ijk}\}$, their analysis conditions on both the predictor totals (n_{1+k}, n_{2+k}) and the response outcome totals (n_{+1k}, n_{+2k}).

The usual sampling schemes then yield a hypergeometric distribution (3.16) for the first cell count n_{11k} in each partial table.

6.3.2 Cochran-Mantel-Haenszel test of conditional independence

That count determines $(n_{12k}, n_{21k}, n_{22k})$, given the marginal totals. Under H_0, the mean and variance of n_{11k} are

$$\begin{aligned} \mu_{11k} &= E(n_{11k}) = n_{1+k}\, n_{+1k}/n_{++k}\,, \\ \text{var}(n_{11k}) &= n_{1+k}\, n_{2+k}\, n_{+1k}\, n_{+2k}/[n_{++k}^2\,(n_{++k}-1)]\,. \end{aligned}$$

Cell counts from different partial tables are independent.

The test statistic combines information from the K tables by comparing $\sum_k n_{11k}$ to its null expected value. It equals

which converges to chi-squared distribution with df$=1$.

6.3.2 Cochran-Mantel-Haenszel test of conditional independence

Cochran (1954)

He proposed a similar statistic, but treated the rows in each 2×2 table as two independent binomials rather than a hypergeometric.

Cochran's statistic is the same as the CMH above, but $\text{var}(n_{11k})$ is replaced by

$$\text{var}(n_{11k}) = n_{1+k}\, n_{2+k}\, n_{+1k}\, n_{+2k} / n_{++k}^3 .$$

6.3.2 Cochran-Mantel-Haenszel test of conditional independence

Remarks

- Because of the similarity in Mantel-Haenszel and Cochran's approaches, we call the CMH expression above the *Cochran-Mantel-Haenszel* (CMH) *statistic*.
- The Mantel-Haenszel approach is more general in that it also applies to some cases in which the rows are not independent binomial samples from two populations.
- The CMH statistic can be generalized for $I \times J \times K$ tables (Section 7.5.3).

6.3.2 Cochran-Mantel-Haenszel test of conditional independence

TABLE A.11 SAS Code for Cumulative Logit and Probit Models with Mental Impairment Data in Table 7.5

```
data impair;
input mental ses life;
datalines;
1 1 1
...
4 0 9
;
proc genmod ;
  model mental=life ses/dist=multinomial link=clogit lrci type3;
proc logistic;
  model mental=life ses/link=probit;
```

6.3.3 Multicenter clinical trial example

Table 6.9 reports the sample odds ratio for each partial table (i.e., each center) and expected value and variance of the number of successes for the drug treatment (n_{11k}) under H_0 : conditional independence.

In each sub-table except the last, the sample odds ratio shows a positive association (i.e., odds ratio > 1).

Thus, it makes sense to combine results with CMH$=$ 6.38, with df$=$ 1 and $P = 0.012$. $\Rightarrow$ Considerable evidence against H_0.

Similar results occur in test $H_0 : \beta = 0$ in the logit model.

The model fit has $\hat{\beta} = 0.777$ with SE$=$ 0.307. The Wald statistic is $(0.777/0.307)^2 = 6.42$ with $P = 0.011$.

The likelihood-ratio statistic equals 6.67 with $P = 0.010$.

6.3.4 CMH test and sparse data

Contingency tables with relatively few observations are referred to as *sparse*(稀疏).

Usually, when K grows with n, each stratum has few observations; e.g., in case-control studies, a few controls match with each case.

The nonstandard setting in which $K \to \infty$ as $n \to \infty$ is called *sparse-data asymptotics*.

An advantage of CMH is that its chi-squared limit also applies in this situation. The likelihood-ratio and Wald tests requires the number of parameters (and hence K) to be fixed, so it does not apply to this alternative scheme.

6.3.5 Estimation of common odds ratio

When the association seems stable among partial tables, it is helpful to combine the K sample odds ratios into a summary measure of conditional association.

Model-based estimator

The model: $\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z$, implies homogeneous XY association, i.e., $\theta_{XY(1)} = \cdots = \theta_{XY(K)} = \exp(\beta)$.

The ML estimate of the common odds ratio is $\exp(\hat{\beta})$.

Mantel and Haenszel's estimator

6.3.5 Estimation of common odds ratio

where $R_k = n_{11k}n_{22k}/n_{++k}$, $S_k = n_{12k}n_{21k}/n_{++k}$ and $p_{ij|k} = n_{ijk}/n_{++k}$.

When K is large and the data are sparse, $\hat{\theta}_{MH}$ is preferred over the model-based ML estimator because the ML estimator $\hat{\beta}$ tends to be too large in absolute value.

If the true odds ratios are not identical across strata but do not vary drastically, $\hat{\theta}_{MH}$ still is a useful summary of the conditional associations.

6.3.5 Estimation of common odds ratio

Example: for the eight-center clinical trial data in Table 6.9,

$$\hat{\theta}_{MH} = \frac{(11 \times 27)/73 + \cdots + (4 \times 1)/13}{(25 \times 10)/73 + \cdots + (2 \times 6)/13} = 2.13\,.$$

For $\log(\hat{\theta}_{MH}) = \log(2.13) = 0.758$, we obtained $\hat{\sigma}[\log(\hat{\theta}_{MH})] = 0.303$. A 95% CI for the common odds ratio is $\exp(0.758 \pm 1.96 \times 0.303) = (1.18, 3.87)$.

Similar results occur using the logit model (6.4) with $\hat{\beta} = 0.777$ and $\text{SE}(\hat{\beta}) = 0.307$ (see Section 6.3.3):

- A 95% Wald CI for the common odds ratio is $\exp(0.777 \pm 1.96 \times 0.307) = (1.19, 3.97)$.
- A 95% likelihood-ratio CI for the common odds ratio is $(1.20, 4.02)$.

6.3.6 Testing homogeneity of odds ratios

The homogeneous association condition $\theta_{XY(1)} = \cdots = \theta_{XY(K)}$ for $2 \times 2 \times K$ tables is equivalent to the logit model (6.4) without XZ interaction.

The usual G^2 and X^2 test statistics provide this, with df$= K - 1$; i.e., they test that in the saturated model the $K - 1$ coefficients of interaction terms [cross products of the dummy variable for x with $(K - 1)$ dummy variables for categories of Z] all equal 0.

Example: For the eight-center clinical trial data in Table 6.9, $G^2 = 9.7$ and $X^2 = 8.0$ with df$= 7$.

$\Rightarrow$ Do not contradict the hypothesis of equal odds ratios.

$\Rightarrow$ It is reasonable to summarize the conditional association by a single odds ratio (e.g., $\hat{\theta}_{MH} = 2.1$) for all eight partial tables.