Building, Checking, and Appying Logistic Regression Models

Jinhwan Suk

Department of Mathematical Science, KAIST

July 21, 2020

Contents

- Model selection : Complexity vs. Simplicity
- (Peduzzi et al. 1996) For each type of predictors, it recommends to exist more than 10 outcomes.
- Cautions that apply to ordinary regression hold for any GLM.
- *Correlations* among several explanatory variables make it seem that no one variable is important.
- Deleting such a redundant variable can be helpful to reduce SE of other estimates.

C	S	W	Wt	**Sa**
2	3	28.3	3.05	Yes
3	3	26.0	2.60	Yes
3	3	25.6	2.15	No
4	2	21.0	1.85	No
3	3	22.5	1.55	No

- $\text{logit}[P(Y=1)] =$
  $\alpha + \beta_1 W + \beta_2 Wt + \beta_3 c_1 + \beta_4 c_2 + \beta_5 c_3 + \beta_6 s_1 + \beta_7 s_2$
- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0$ Likelihood-ratio test
- Test statistic $= 40.56$, df $= 7 \cdots (P < 0.0001)$
- At least one predictor has an effect

Strategies in Model Selection

How Many Explanatory Variables Can Be in the Model?

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

- In previous chapter, we showed strong evidence of a width effect.
- Weight and width are equally good predictors, but they have a strong correlation(0.887)

Contents

1. **Forward selection**
    1. Adds terms sequentially
    2. Select the term giving the greatest improvement in fit

       Improvement in fit?? *p-value* or *reduction in deviance*??
    3. Stop when they do not improve the fit significantly

2. **Backward Selection**
    1. Begin with a complex model
    2. Sequentially remove terms

       Cautions :
       Qualitative predictors (more than 2 categories), Interaction effect terms

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	152	212.4	—	—	
2	$(C*S + C*W + S*W)$	173.68	155	209.7	(2)–(1)	3.2 (df = 3)	
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.

Contents

- We are **not** selecting a "correct" model from a set of candidates
- Model is a simplification of reality
- Simple model has the advantages of model parsimony
- Consider a criteria that can help select a good model in terms of estimating quantities of interest

- *Akaike information criterion* (AIC)

$$AIC(\mathcal{M}) = -2(\sup_{M \in \mathcal{M}} \mathcal{L}(\theta_M; y) - p_{\mathcal{M}})$$

- *Bayesian information criterion* (BIC)

$$BIC(\mathcal{M}) = -2 \sup_{M \in \mathcal{M}} \mathcal{L}(\theta_M; y) + p_{\mathcal{M}} \log(n)$$

Determine a set of models that has highest posterior probability.

Unclear when applied with frequentist method.

Contents

Strategies in Model Selection

Example: Using Causal Hypotheses to Guide Model Building

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.

- There is a time ordering of the variables:

$$G \to P \to E \to M$$

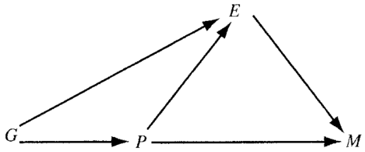


FIGURE 6.1 Causal diagram for Table 6.3.

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.

Contents

- Standard residuals :

$$r_i = y_i - \hat{\mu}_i = y_i - n_i\hat{\pi}_i$$

- Pearson residuals :

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{\widehat{Var}(Y_i)}} = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}, \quad X^2 = \sum_{i=1}^{N} e_i^2$$

- Deviance residuals :

$$\sqrt{d_i} \times \text{sign}(y_i - \hat{\mu}_i), \quad G^2 = \sum_{i=1}^{N} d_i$$

- We prefer to use standardized residuals :

$$r_i = e_i / \sqrt{1 - \hat{h}_i}$$

- Asymptotic covariance of $\hat{\beta}$ is given by

$$\begin{aligned} asymp.\ Cov(\hat{\beta}) &= \mathcal{I}(\hat{\beta})^{-1} \quad \textit{(property of MLE)} \\ &= \mathbb{E}\left[\left(\frac{\partial L}{\partial \beta} \right)^T \left(\frac{\partial L}{\partial \beta} \right) \right]^{-1} \\ &= (X^T W X)^{-1} \qquad W = diag\left(\frac{(\partial \mu_i / \partial \eta_i)^2}{Var(Y_i)} \right) \end{aligned}$$

- Let D denote the diagonal matrix with $\partial \mu_i / \partial \eta_i$

$$W = DV^{-1}D \quad V = DW^{-1}D$$

- (Grouped Data) Asymptotic covariance of $\hat{\eta} = X\hat{\beta}$ is

$$asymp.\ Cov(\hat{\eta}) = X(X^T WX)^{-1}X^T$$

- Asymptotic covariance of $\hat{\mu} = g^{-1}(\hat{\eta})$ is

$$asymp.\ Cov(\hat{\mu}) = DX(X^T WX)^{-1}X^T D$$

- Since $y - \hat{\mu}$ and $\hat{\mu} - \mu$ is asymptotically uncorrelated[1],

$$\begin{aligned} asymp.\ Cov(y - \hat{\mu}) &= asymp.\ Cov(y - \mu) - asymp.\ Cov(\hat{\mu} - \mu) \\ &= V - Cov(\hat{\mu}) \\ &= DW^{-1}D - DX(X^T WX)^{-1}X^T D \\ &= V^{1/2}[I - H_{at}]V^{1/2} \end{aligned}$$

[1] P 142, 4.5.7

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: Indep. Model	Fitted: Linear Logit	Residual: Indep. Model	Residual: 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).

- Independent model : $\text{logit}(\pi_i) = \alpha$
- The (Standardized) residual of independent model shows an increasing trend $\rightarrow$ linear logit model
- The trend in standardized residuals disappears!
- $G^2 = 5.91, X^2 = 6.29, df = 6$

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.

- The admissions decision is independent of gender :

$$\text{logit}(\pi_{ik}) = \alpha + \beta_k^D$$

- $G^2 = 44.74, X^2 = 40.85, df = 23 \rightarrow$ poor fit

Contents

- $R = Corr(y, \hat{\mu})$
- $R^2 = 1 - \frac{\sum_i (y_i - \hat{\pi}_i)^2}{\sum_i (y_i - \bar{y})^2}$

Contents

Summarizing the Predictive Power of a Model

Likelihood and Deviance Measures

- $L_{(M/S/0)}$: the maximized log likelihood for a (given model / saturated model / model only with an intercept)
- $L_0 \leq L_M \leq L_S \leq 0$
- Measure of predictive power(D) :

$$\frac{L_M - L_0}{L_S - L_0} \in [0, 1]$$

- For N independent Bernoulli observations,

$$L_0 = N[\bar{y} \log \bar{y} + (1 - \bar{y}) \log(1 - \bar{y})] \quad \leftarrow \hat{\pi}_i = \bar{y}$$

$$L_S = 0 \quad \leftarrow \hat{\pi}_i = y_i, \; n_i = 1$$

$$D = 1 - \frac{L_M}{L_0}$$

Contents

Summarizing the Predictive Power of a Model

Classification Tables

Test Indicator

Outcome		No	Yes
	No	a True Negative	b False Positive
	Yes	c False Negative	d True Positive

- $\hat{y} = 1$ when $\hat{\pi} > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi} < \pi_0$
- The proportion of correct classifications is

$$\begin{aligned} P_{cor} &= P(y = 1 \; and \; \hat{y} = 1) + P(y = 0 \; and \; \hat{y} = 0) \\ &= P(\hat{y} = 1 \, | y = 1) P(y = 1) + P(\hat{y} = 0 \, | y = 0) P(y = 0) \\ &= Sensitivity * P(y = 1) + Specificity * P(y = 0) \end{aligned}$$

- Sensitive to relative numbers of $y = 1$ and $y = 0$ ...?

Contents

- The classification table depends on the cutoff π_0
- ROC curve is plot of Sensitivity as a function of (1-Specificity)
- When $\pi_0 \approx 0$, Sensitivity ≈ 1 and Specificity $\approx 0 \rightarrow (1, 1)$
- When $\pi_0 \approx 1$, Sensitivity ≈ 0 and Specificity $\approx 1 \rightarrow (0, 0)$
- The area under a ROC curve : *concordance index*

Contents

TABLE 6.9 Clinical Trial Relating Treatment to Response for Eight Centers

Center	Treatment	Response: Success	Response: Failure	Odds Ratio	μ_{11k}	$\text{var}(n_{11k})$
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).

- *Simpson's Paradox*
- For a binary response Y, we analyze the effect of binary predictor X, conditional on the covariate Z

$$\pi_{ik} = P(Y = 1 | X = i, Z = k)$$

- Our model is

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z$$

- This model (implicitly) assumes that the XY conditional odds ratio is same

$$\theta_1 = \theta_2 = \cdots = \theta_K = \exp(\beta)$$

- **Method 1 :** $H_0 : \beta = 0$ ($\Leftrightarrow XY$ conditional independence)
  1. Wald statistic : $(\hat{\beta}/SE)^2, \quad df = 1$
  2. LR statistic : $G^2(M|M_0), \quad df = 1$ where the model M_0 is

$$\text{logit}(\pi_{ik}) = \alpha + \beta_k^D$$

- **Method 2 :** Goodness-of-fit test($df = K$) of the model

$$\text{logit}(\pi_{ik}) = \alpha + \beta_k^D$$

where saturated model is given by

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^D + \beta_{ik} x_i \beta_k^D$$

- When we can assume that $\beta_{ik} \approx 0$, Method 2 is less powerful, especially when K is large
- When the direction of the conditional XY association varies among categories of Z, Method 1 can be less powerful

Contents

- Non model based test of H_0 : conditional independence in $2 \times 2 \times K$
- Hypergeometric sampling for n_{11k}
- Under H_0, each n_{11k} follows hypergeometric mean and variace are

$$\mu_{11k} = \mathbb{E} n_{11k} = n_{1+k} n_{+1k} / n_{++k}$$
$$var(n_{11k}) = n_{1+k} n_{2+k} n_{+1k} n_{+2k} / [n_{++k}^2 (n_{++k} - 1)]$$

- Each partial tables are independent :

$$CMH = \frac{(\sum_k n_{11k} - \sum_k \mu_{11k})^2}{var(\sum_k n_{11k})} = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k var(n_{11k})}$$

- This statistic has a large-sample χ^2 null distribution with df=1

- Cochran (1954) treated the rows in each 2×2 table as two independent binomials

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

- Mantel-Haenszel approach:

  1. Retrospective study
  2. Randomized clinical trials with volunteers randomly allocated to two treatments

- The multicenter clinical trial reports the sample odds ratio
- Except last, the sample odds ratio shows a positive association
- Combine results using CMH = 6.38 with df=1(P=0.012)
- Testing $H_0 : \beta = 0$ in logistic model
  - Model fit : $\hat{\beta} = 0.777$ with $SE = 0.307$
  - Wald statistic $= 6.42$ $(P = 0.011)$
  - LR statistic $= 6.67$ $(P = 0.010)$
  - *Score statistic = CMH*

- As $n \to \infty$ with fixed K,

$$\textit{Wald, LR, CMH tests } \to \chi^2_1 \quad \textit{under } H_0$$

- Advantage of CMH statistic is that when $K \to \infty$ as $n \to \infty$

$$\textit{CMH test } \to \chi^2_1 \textit{ under } H_0$$

- $n = 2K$, so $K \to \infty$ as $n \to \infty$(Sparse-data asymptotics)
- The first case in Table 6.10
  - $\mu_{11k} - n_{11k} = 0$
  - $var(n_{11k}) = 0$
- The second case in Table 6.10
  - $\mu_{11k} = 0.50$
  - $var(n_{11k}) = 0.25$
- By CLT, the CMH is approximately chi-squared

TABLE 6.10 Stratum Containing a Matched Pair

Element of Pair	Response		Response	
	Success	Failure	Success	Failure
First	1	0	1	0
Second	1	0	0	1

Contents

- The logistic model implies homogeneous association

$$\theta_{XY(1)} = \cdots = \theta_{XY(1)} = \exp(\beta)$$

- The ML estimate of the common odds ratio is $\exp(\beta)$
- Mantel and Haenszel (1959) proposed

$$\hat{\theta}_{MH} = \frac{\sum_k n_{++k} p_{11|k} p_{22|k}}{\sum_k n_{++k} p_{12|k} p_{21|k}}$$

  It is preferred over the ML estimator when K is large and the data are very sparse
- Robins et al. (1986) derived an estimated variance for $\log(\hat{\theta}_{MH})$

Contents

Meta-analyses for Summarizing Multiple 2×2 Tables

- *Meta-analysis* is a statistical analysis that combines information from several studies.
- Assume that the population values of the particular effect measure(Odds ratio or difference of proportion) are identical in each study
- Significance test of no effect(conditional independence)
  1. Test $H_0 : \beta = 0$ using Wald or LR
  2. CMH test (Advantageous for highly sparse data)
  3. Small sample $\rightarrow$ Generalization of Fisher's exact test (7.3.5)
- Common odds ratio
  1. Logistic model : $\exp(\hat{\beta})$
  2. Highly sparse data : $\hat{\theta}_{MH}$

- ML estimate : Common difference of proportion(δ) in a model

$$\pi_{ik} = \alpha + \delta x_i + \beta_k^Z$$

- Greenland and Robins (1985) proposed Mantel-Haenszel-type estimates

$$\hat{\delta}_{MH} = \sum_k w_k \hat{\delta}_k / \sum_k w_k$$
$$w_k = n_{1+k} n_{2+k} / (n_{1+k} + n_{2+k})$$

- ML estimator is more efficient but

  But, π_{ik} must be constrained to fall between 0 and 1

- Alternative approach :
  1. Score or Profile likelihood confidence interval

$$d_k \pm z_{\alpha/2} s_k$$

  2. Then taking weight,

$$\hat{\delta} = \sum_k w_k d_k, \quad SE = \sum_k [1/n_{11k}^2]^{-1/2}$$

- The multicenter clinical trial
  1. $\hat{\delta}_{MH} = 0.130$, $SE = 0.050$
  2. $\hat{\delta} = 0.128$, $SE = 0.049$

- *Collapsibility condition for Odds Ratio* : When $\theta_{XY(k)}$ is identical at every level k of Z, that value equals to θ_{XY} if either Z and X are conditionally independent or if Z and Y are conditionally independent.
- Consider the logistic model

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z$$

- The estimated odds ratio $\exp(\hat{\beta})$ differs from the sample odds ratio in marginal 2×2 table.
- When center effects are negligible and the model

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i$$

fits well, then the collapsibility holds

- The homogeneous association condition for $2 \times 2 \times K$

$$\theta_{XY(1)} = \cdots = \theta_{XY(K)}$$

  is equivalent to logistic model

$$\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z$$

- G^2 and X^2 with $df = K - 1$
- Saturated model : $\text{logit}(\pi_{ik}) = \alpha + \beta x_i + \beta_k^Z + \beta_{ik} x_i$
- The multicenter clinical trial
  - $G^2 = 9.75$, $X^2 = 8.03$, $df = 7$
  - Do not contradict the hypothesis of equal odds ratio
  - $\hat{\theta}_{MH} = 2.13$ or $e^{\hat{\beta}} = 2.17$

Contents

Definition (Complete separation)

The space of explanatory variable values is said to have **complete separation** when a hyperplane can pass through that space, i.e., there exists a vector **b** such that

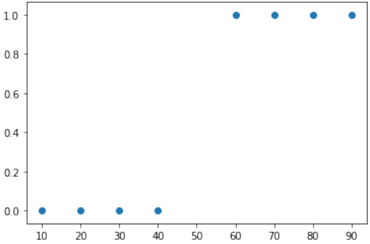
$$\mathbf{b}^T x_i > 0 \text{ whenever } y_i = 1$$
$$\mathbf{b}^T x_i < 0 \text{ whenever } y_i = 0$$

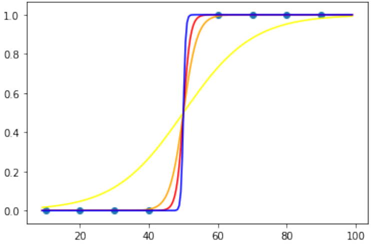
Definition (Quasi-complete separation)

The space of explanatory variable values is said to have **quasi-complete separation** when a hyperplane separates explanatory variables with $y = 1$ and with $y = 0$, but cases exist with both outcomes on that hyperplane

Detecting and Dealing with Infinite Estimates

Complete or Quasi-complete Separation





- Set $\alpha = -50\beta$ and $\beta \to \infty$
- As β increases, it becomes closer to a perfect fit
- Wald inferecne is useless
- We can still get confidence interval by **inverting** LR test or Score test

  95% confidence interval for β : $(0.06, \infty)$

		Response(Y)		YZ Marginal	
Center(Z)	Treatment(X)	Success	Failure	Sucess	Failure
1	Active drug	0	5	0	14
	Placebo	0	9		
2	Active drug	1	12	1	22
	Placebo	0	10		
3	Active drug	0	7	0	12
	Placebo	0	5		
4	Active drug	6	3	8	9
	Placebo	2	6		
5	Active drug	5	9	7	21
	Placebo	2	12		
XY marginal	Active drug	12	36		
	Placebo	4	42		

Contents

- Inverting LR or Score test
- Smoothing data
- **The Bayesian approach**(Sec 7.2)
- Instead, maximize Penalized likelihood function(Sec 7.4.5)

Contents

- We want to determine whether a particular variable has an effective on a response
- Strong effects are likely to be detected even when n is small
- Detection of weak effects requires large n
- Test $H_0 : \pi_1 = \pi_2$. The study using equal sample size requires approximately

$$n_1 = n_2 = (z_{\alpha/2} + z_{\beta})^2 [\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)] / (\pi_1 - \pi_2)^2$$