Chapter 5 Building and Applying Logistic Regression Models

Outline

Outline

Goals

The selection process becomes more challenging as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions. There are two competing goals: The model should be complex enough to fit the data well, but simpler models are easier to interpret.

To answer those questions, confirmatory analyses use a restricted set of models. A study's theory about an effect may be tested by comparing models with and without that effect. In the absence of underlying theory, some studies are exploratory rather than confirmatory.

5.1.1 How Many Predictors Can You Use?

- Unbalanced data limits the number of predictors for which effects can be estimated precisely. One guideline suggests there should ideally be at least 10 outcomes of each type for every predictor.
- This guideline is approximate. In practice, often the number of variables is large, sometimes even of similar magnitude as the number of observations. However, when the guideline is violated, ML estimates may be quite biased and estimates of standard errors may be poor.
- Cautions that apply to building ordinary regression models hold for any GLM.

5.1.2 Example: Horseshoe Crabs Revisited

Example

Consider a model with all the main effects. Let $\{c_1, c_2, c_3\}$ be indicator variables for the first three (of four) colors and let $\{s_1, s_2\}$ be indicator variables for the first two (of three) spine conditions. The model

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

treats color and spine condition as nominal-scale factors. Table 5.1 shows the results.

5.1.2 Example: Horseshoe Crabs Revisited

Table 5.1 Parameter Estimates for Main Effects Model with Horseshoe Crab Data

Parameter	Estimate	SE
Intercept	−9.273	3.838
Color(1)	1.609	0.936
Color(2)	1.506	0.567
Color(3)	1.120	0.593
Spine(1)	−0.400	0.503
Spine(2)	−0.496	0.629
Weight	0.826	0.704
Width	0.263	0.195

Although this overall test is highly significant, the Table 5.1 results are discouraging. The estimates for weight and width are only slightly larger than their SE values.

5.1.2 Example: Horseshoe Crabs Revisited

For simplicity below, we symbolize models by their highest-order terms, regarding C and S as factors. For instance, $(C + S + W)$ denotes the model with main effects, whereas $(C + S * W)$ denotes the model with those main effects plus an $S \times W$ interaction. It is not sensible to use a model with interaction but not the main effects that make up that interaction. A reason for including lower-order terms is that, otherwise, the statistical significance and practical interpretation of a higher-order term depends on how the variables are coded. This is undesirable. By including all the lower-order effects that make up an interaction, the same results occur no matter how variables are coded.

5.1.3 Stepwise Variable Selection Algorithms

Forward and Backward Selection

Forward selection adds terms sequentially until further additions do not improve the fit. Backward elimination begins with a complex model and sequentially removes terms.

At a given stage, it eliminates the term in the model that has the largest P-value in the test that its parameters equal zero. We test only the highest-order terms for each variable. It is inappropriate, for instance, to remove a main effect term if the model contains higher-order interactions involving that term. The process stops when any further deletion leads to a significantly poorer fit.

5.1.3 Stepwise Variable Selection Algorithms

Remark

- With either approach, for categorical predictors with more than two categories, the process should consider the entire variable at any stage rather than just individual indicator variables. Otherwise, the result depends on how you choose the baseline category for the indicator variables. Add or drop the entire variable rather than just one of its indicators.
- Variable selection methods need not yield a meaningful model. Use them with caution! When you evaluate many terms, one or two that are not truly important may look impressive merely due to chance.

5.1.3 Stepwise Variable Selection Algorithms

Remark

- In any case, statistical significance should not be the sole criterion for whether to include a term in a model. It is sensible to include a variable that is important for the purposes of the study and report its estimated effect even if it is not statistically significant. Keeping it in the model may help reduce bias in estimating effects of other predictors and may make it possible to compare results with other studies where the effect is significant (perhaps because of a larger sample size). Likewise, with a very large sample size sometimes a term might be statistically significant but not practically significant. You might then exclude it from the model because the simpler model is easier to interpret-for example, when the term is a complex interaction.

5.1.4 Example: Backward Elimination for Horseshoe Crabs

Recall that the deviance of a GLM is the likelihood-ratio statistic for comparing the model to the saturated model, which has a separate parameter for each observation (Section 3.4.3). As Section 3.4.4 showed, the likelihood-ratio test statistic $-2(L_0 - L_1)$ for comparing the models is the difference between the deviances for the models.

Table 5.2 summarizes results of fitting and comparing several logistic regression models. To select a model, we use a modified backward elimination procedure.

5.1.4 Example: Backward Elimination for Horseshoe Crabs

Table 5.2 Results of Fitting Several Logistic Regression Models to Horseshoe Crab Data

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)	$37.8(df = 2)$

Note:C = color, S = spine condition, W = width.

5.1.5 AIC, Model Selection, and the "Correct" Model

- In selecting a model, you should not think that you have found the "correct" one. Any model is a simplification of reality.
- Other criteria besides significance tests can help select a good model. The best known is the *Akaike information criterion* (AIC).
- The optimal model is the one that tends to have its fitted values closest to the true outcome probabilities. This is the model that minimizes

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

5.1.6 Summarizing Predictive Power: Classification Tables

1. Classification Table

It cross classifies the binary outcome y with a prediction of whether $y = 0$ or 1. The prediction is $\hat{y} = 1$ when $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_0$, for some cutoff π_0. One possibility is to take $\pi_0 = 0.50$. However, if a low(high) proportion of observations have $y = 1$, the model fit may never(always) have $\hat{\pi}_i > 0.50$, in which case one never(always) predicts $\hat{y} = 1$. Another possibility takes π_0 as the sample proportion of 1 outcomes, which is $\hat{\pi}_i$ for the model containing only an intercept term.

5.1.6 Summarizing Predictive Power: Classification Tables

We illustrate for the model using width and color as predictors of whether a horseshoe crab has a satellite. Of the 173 crabs, 111 had a satellite, for a sample proportion of 0.642. Table 5.3 shows classification tables using $\pi_0 = 0.50$ and $\pi_0 = 0.642$.

Table 5.3. Classification Tables for Horseshoe Crab Data

	Prediction, $\pi_0 = 0.64$		Prediction, $\pi_0 = 0.50$		
Actual	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	Total
$y = 1$	74	37	94	17	111
$y = 0$	20	42	37	25	62

5.1.6 Summarizing Predictive Power: Classification Tables

2. Other Summaries of Predictor Power

$$\text{sensitivity} = P(\hat{y} = 1 | y = 1)$$

$$\text{specificity} = P(\hat{y} = 0 | y = 0)$$

$$\begin{aligned}
&\text{overall proportion of correct classification} \\
&= P(y = 1, \hat{y} = 1) + P(y = 0, \hat{y} = 0) \\
&= P(\hat{y} = 1 | y = 1) P(y = 1) + P(\hat{y} = 0 | y = 0) P(y = 0)
\end{aligned}$$

5.1.6 Summarizing Predictive Power: Classification Tables

A classification table has limitations: It collapses continuous predictive values $\hat{\pi}$ into binary ones. The choice of π_0 is arbitrary. Results are sensitive to the relative numbers of times that $y = 1$ and $y = 0$.

5.1.7 Summarizing Predictive Power: ROC Curves

ROC Curve

A receiver operating characteristic (ROC) curve is a plot of sensitivity as a function of (1-specificity) for the possible cutoffs π_0. An ROC curve is more informative than a classification table, because it summarizes predictive power for all possible π_0.

5.1.7 Summarizing Predictive Power: ROC Curves

When π_0 gets near 0, almost all predictions are $\hat{y} = 1$; then, sensitivity is near 1, specificity is near 0, and the point for $(1 - specificity, sensitivity)$ has coordinates near $(1, 1)$. When π_0 gets near 1, almost all predictions are $\hat{y} = 0$; then, sensitivity is near 0, specificity is near 1, and the point for $(1 - specificity, sensitivity)$ has coordinates near $(0, 0)$. The ROC curve usually has a concave shape connecting the points $(0, 0)$ and $(1, 1)$.

5.1.7 Summarizing Predictive Power: ROC Curves

For a given specificity, better predictive power correspond to higher sensitivity. So, the better the predictive power, the higher the ROC curve.

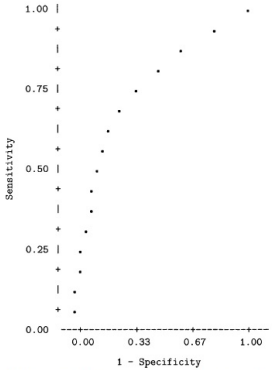

Figure 5.1. ROC curve for logistic regression model with horseshoe crab data.

5.1.7 Summarizing Predictive Power: ROC Curves

Concordance Index

The area under the ROC curve is identical to the value of a measure of predictive power called the *concordance index*. Consider all pairs of observations (i, j) such that $y_i = 1$ and $y_j = 0$. The concordance index c estimates the probability that the predictions and the outcomes are concordant, which means that the observation with the larger y also has the larger $\hat{\pi}$.

A value $c = 0.50$ means predictions were no better than random guessing. This corresponds to a model having only an intercept term. Its ROC curve is a straight line connecting the points $(0, 0)$ and $(1, 1)$.

5.1.8 Summarizing Predictive Power: A Correlation

- For a GLM, a way to summarize prediction power is by the correlation R between the observed responses $\{y_i\}$ and the model's fitted values $\{\hat{\mu}_i\}$.
- For a binary regression model, R is the correlation between the n binary $\{y_i\}$ observations (1 or 0 for each) and the estimated probabilities $\{\hat{\pi}_i\}$.
- For the horseshoe crab data, using color alone does not do nearly as well as using width alone ($R = 0.285$ *vs* $R = 0.402$). Using both predictors together increases R to 0.452. The simpler model that uses color merely to indicate whether a crab is dark does essentially as well, with $R = 0.447$.

Outline

5.2.1 Likelihood-Ratio Model Comparison Tests

One way to detect lack of fit uses a likelihood-ratio test to compare the model with more complex ones. A more complex model might contain a nonlinear effect. Models with multiple predictors would consider interaction terms. If more complex models do not fit better, this provides some assurance that a chosen model is adequate.

$$\text{logit}[\pi(x)] = \alpha + \beta x$$
$$\text{logit}[\hat{\pi}(x)] = \alpha + \beta_1 x + \beta_2 x^2$$

5.2.2 Goodness of Fit and the Deviance

- A more general way to detect lack of fit searches for any way the model fails. A goodness-of-fit test compares the model fit with the data. This approach regards the data as representing the fit of the most complex model possible-the saturated model, which has a separate parameter for each observation.
- Denote the working model by M. In testing the fit of M, we test whether all parameters that are in the saturated model but not in M equal zero. In GLM terminology, the likelihood-ratio statistic for this test is the deviance of the model (Section 3.4.3). In certain cases, this test statistic has a large-sample chi-squared null distribution.

5.2.2 Goodness of Fit and the Deviance

- When the predictors are solely categorical, the data are summarized by counts in a contingency table.
- For the n_i subjects at setting i of the predictors, multiplying the estimated probabilities of the two outcomes by n_i yields estimated expected frequencies for $y = 0$ and $y = 1$. These are the *fitted values* for that setting.

Deviance G^2

$$G^2(M) = 2\sum \text{observed}[\log(\text{observed}/\text{fitted})]$$

For two-way contingency tables $G^2 = 2\sum n_{ij}\log(\frac{n_{ij}}{\mu_{ij}})$

Test of independence $G^2 = 2\sum n_{ij}\log(\frac{n_{ij}}{\hat{\mu}_{ij}})$, $\hat{\mu}_i j = \frac{n_{i+}n_{+j}}{n}$

5.2.2 Goodness of Fit and the Deviance

Pearson X^2

$$X^2(M) = \sum (\text{observed} - \text{fitted})^2/\text{fitted}$$

- For two-way contingency tables $X^2 = \sum \frac{(n_{ij}-\mu_{ij})^2}{\mu_{ij}}$; for test independence $X^2 = \sum \frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$.
- For a fixed number of settings, when the fitted counts are all at least about 5, $X^2(M)$ and $G^2(M)$ have approximate chi-squared null distributions.
- The degrees of freedom, called the residual df for the model, subtract the number of parameters in the model from the number of parameters in the saturated model.

5.2.2 Goodness of Fit and the Deviance

Table 4.4. Development of AIDS Symptoms by AZT Use and Race

		Symptoms	
Race	AZT Use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

$$\text{logit}(\hat{\pi}) = -1.074 - 0.720x + 0.056z$$

$G^2(M) = 1.38, X^2(M) = 1.39,\ df = 4 - 3 = 1,\ P = 0.24.$

5.2.3 Checking Fit: Grouped Data, Ungrouped Data, and Continuous Predictors

Remark

Although the ML estimates of parameters are the same for either form of data, the X^2 and G^2 statistics are not. These goodness-of-fit tests only make sense for the grouped data. The large-sample theory for X^2 and G^2 applies for contingency tables when the fitted counts mostly exceed about 5.

5.2.3 Checking Fit: Grouped Data, Ungrouped Data, and Continuous Predictors

Question

When calculated for logistic regression models fitted with continuous or nearly continuous predictors, the X^2 and G^2 statistics do not have approximate chi-squared distributions. How can we check the adequacy of a model for such data?

5.2.3 Checking Fit: Grouped Data, Ungrouped Data, and Continuous Predictors

- One way creates categories for each predictor, and then applies X^2 or G^2 to observed and fitted counts for the grouped data.
- An alternative way of grouping the data forms observed and fitted values based on a partitioning of the estimated probabilities. With 10 groups of equal size, the first pair of observed counts and corresponding fitted counts refers to the $n/10$ observations having the highest estimated probabilities, the next pair refers to the $n/10$ observations having the second decile of estimated probabilities, and so forth. The fitted value for an outcome is the sum of the estimated probabilities for that outcome for all observations in that group.

5.2.3 Checking Fit: Grouped Data, Ungrouped Data, and Continuous Predictors

- The *Hosmer-Lemeshow test* uses a Pearson test statistic to compare the observed and fitted counts for this partition. The test statistic does not have exactly a limiting chi-squared distribution.
- However, Hosmer and Lemeshow(2000, pp.147-156) noted that, when the number of distinct patterns of covariate values (for the original data) is close to the sample size, the null distribution is approximated by chi-squared with $df =$ number of groups -2.

5.2.4 Residuals for Logit Models

Pearson Residual

$$\text{Pearson residual} = e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{[n_i\hat{\pi}_i(1-\hat{\pi}_i)]}}$$

When n_i is large, e_i has an approximate normal distribution. When the model holds,$\{e_i\}$ has an approximate expected value of zero but a smaller variance than a standard normal variate. For GLM, $e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{Var}(y_i)}}$.

5.2.4 Residuals for Logit Models

Standardized Residual

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

GLM $\frac{y_i-\hat{\mu}_i}{SE}$, $SE = [\hat{Var}(y_i)(1-h_i)]^{1/2}$; contingency table $\frac{n_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-p_{i+})(1-p_{+j})}}$.

5.2.4 Residuals for Logit Models

- The term h_i in this formula is the observation's leverage, its element from the diagonal of the so-called hat matrix.
- The standardized residual equals $e_i/\sqrt{(1-h_i)}$, so it is larger in absolute value than the Pearson residual e_i. It is approximately standard normal when the model holds. We prefer it.
- When fitted values are very small, we have noted that X^2 and G^2 do not have approximate null chi-squared distributions. Similarly, residuals have limited meaning in that case. When data can be grouped into sets of observations having common predictor values, it is better to compute residuals for the grouped data than for individual subjects.

5.2.5 Example: Graduate Admissions at University of Florida

Example

Table 5.5 refers to graduate school applications to the 23 departments in the College of Liberal Arts and Sciences at the University of Florida, during the 1997-98 academic year. It cross-classifies whether the applicant was admitted (Y), the applicant's gender (G), and the applicant's department (D). For the n_{ik} applications by gender i in department k, let y_{ik} denote the number admitted and let π_{ik} denote the probability of admission. We treat $\{Y_{ik}\}$ as independent binomial variates for $\{n_{ik}\}$ trials with success probabilities $\{\pi_{ik}\}$.

5.2.5 Example: Graduate Admissions at University of Florida

Table 5.5. Table Relating Whether Admitted to Graduate School at Florida to Gender and Department, Showing Standardized Residuals 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						

Note: Thanks to Dr. James Booth for showing me these data.

5.2.5 Example: Graduate Admissions at University of Florida

The model with no gender effect, given department, is

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

$G^2 = 44.7$ and $X^2 = 40.9$, $df = 23$, $P-$values $= 0.004, 0.012$.

However, the model may be inadequate, perhaps because a gender effect exists in some departments or because the binomial assumption of an identical probability of admission for all applicants of a given gender to a department is unrealistic.

5.2.5 Example: Graduate Admissions at University of Florida

- Departments with large standardized residuals are responsible for the lack of fit. Significantly more females were admitted than the model predicts in the Astronomy and Geography departments, and fewer were admitted in the Psychology department. Without these three departments, the model fits adequately ($G^2 = 24.4$, $X^2 = 22.8$, $df = 20$).
- For the complete data, next we consider the model that also has a gender effect. It does not provide an improved fit ($G^2 = 42.4$, $X^2 = 39.0$, $df = 22$), because the departments just described have associations in different directions and of greater magnitude than other departments.

5.2.5 Example: Graduate Admissions at University of Florida

- This model has an ML estimate of 1.19 for the GY conditional odds ratio: The estimated odds of admission were 19% higher for females than males, given department.
- By contrast, the marginal table collapsed over department has a GY sample odds ratio of 0.94, the overall odds of admission being 6% lower for females.
- This illustrates Simpson's paradox, because the conditional association has a different direction than the marginal association.

5.2.6 Influence Diagnostics for Logistic Regression

Several diagnostics describe various aspects of influence. Many of them relate to the effect on certain characteristics of removing the observation from the data set. In logistic regression, the observation could be a single binary response or a binomial response for a set of subjects all having the same predictor values (i.e., grouped data). These diagnostics are algebraically related to an observation's leverage.

5.2.6 Influence Diagnostics for Logistic Regression

Influence diagnostics for each observation

1. For each model parameter, the change in the parameter estimate when the observation is deleted. This change, divided by its standard error, is called *Dfbeta*.
2. A measure of the change in a joint confidence interval for the parameters produced by deleting the observation. This confidence interval 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 diagnostic, the larger the value, the greater the influence.

5.2.7 Example: Heart Disease and Blood Pressure

Example

Table 5.6 is from an early analysis of data from the Framingham study, a longitudinal study of male subjects in Framingham, Massachusetts. In this analysis, men aged 40-59 were classified on x = blood pressure and y = whether developed heart disease during a 6 year follow-up period.

5.2.7 Example: Heart Disease and Blood Pressure

Example

Let π_i be the probability of heart disease for blood pressure category i. The table shows the fit for the linear logit model,

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

with scores $\{x_i\}$ for blood pressure level.We used scores (111.5, 121.5, 131.5, 141.5, 151.5, 161.5, 176.5, 191.5). The nonextreme scores are midpoints for the intervals of blood pressure.

5.2.7 Example: Heart Disease and Blood Pressure

Table 5.6. Diagnostic Measures for Logistic Regression Models Fitted to Heart Disease Data

Blood Pressure	Sample Size	Observed Disease	Fitted Disease	Standardized Residual	*Dfbeta*	*c*	Pearson Difference	LR Difference
111.5	156	3	5.2	−1.11	0.49	0.34	1.22	1.39
121.5	252	17	10.6	2.37	−1.14	2.26	5.64	5.04
131.5	284	12	15.1	−0.95	0.33	0.31	0.89	0.94
141.5	271	16	18.1	−0.57	0.08	0.09	0.33	0.34
151.5	139	12	11.6	0.13	0.01	0.00	0.02	0.02
161.5	85	8	8.9	−0.33	−0.07	0.02	0.11	0.11
176.5	99	16	14.2	0.65	0.40	0.26	0.42	0.42
191.5	43	8	8.4	−0.18	−0.12	0.02	0.03	0.03

Source: J. Cornfield, *Fed. Proc.*, **21**(suppl. 11): 58–61, 1962.

5.2.7 Example: Heart Disease and Blood Pressure

Another useful graphical display for showing lack of fit compares observed and fitted proportions by plotting them against each other, or by plotting both of them against explanatory variables.

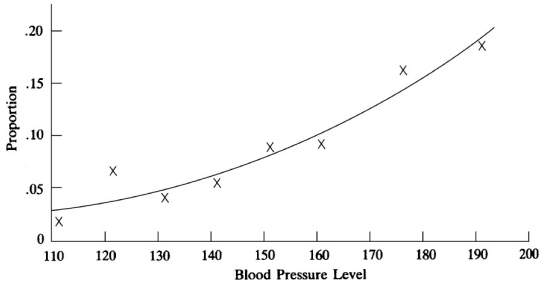

Figure 5.2. Observed proportion (x) and estimated probability of heart disease (curve) for linear logit model.

Outline

Certain data patterns present difficulties, with the ML estimates being infinite or not existing. For quantitative or categorical predictors, this relates to observing only successes or only failures over certain ranges of predictor values.

5.3.1 Infinite Effect Estimate: Quantitative Predictor

Consider first the case of a single quantitative predictor. The ML estimate for its effect is infinite when the predictor values having $y = 0$ are completely below or completely above those having $y = 1$

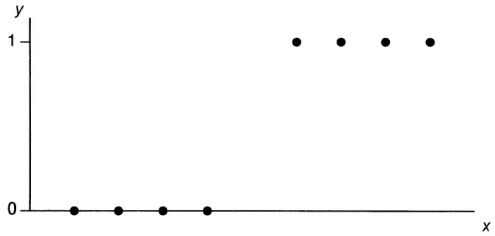

Figure 5.3. Perfect discrimination resulting in infinite logistic regression parameter estimate.

5.3.1 Infinite Effect Estimate: Quantitative Predictor

Remark

In practice, most software fails to recognize when $\hat{\beta} = \infty$. After a few cycles of the iterative fitting process, the log likelihood looks flat at the working estimate, and convergence criteria are satisfied. Because the log likelihood is so flat and because standard errors of parameter estimates become greater when the curvature is less, software then reports huge standard errors. A danger is that you might not realize when reported estimated effects and results of statistical inferences are invalid.

5.3.1 Infinite Effect Estimate: Quantitative Predictor

With several predictors, consider the multidimensional space for displaying the data. Suppose you could pass a plane through the space of predictor values such that on one side of that plane $y = 0$ for all observations, whereas on the other side $y = 1$ always. There is then *perfect discrimination*: You can predict the sample outcomes perfectly by knowing the predictor values (except possibly at boundary points between the two regions). Again, at least one estimate will be infinite. When the spaces overlap where $y = 1$ and where $y = 0$, the ML estimates are finite.

5.3.2 Infinite Effect Estimate: Categorical Predictors

Infinite estimates also can occur with categorical predictors.

Sparse

With two or more categorical predictors, the data are counts in a multiway contingency table. When the table has a large number of cells, most cell counts are usually small and many may equal 0. Contingency tables having many cells with small counts are said to be *sparse*.

Sparseness is common in contingency tables with many variables or with classifications having several categories.

5.3.2 Infinite Effect Estimate: Categorical Predictors

Sampling zero

A cell with a count of 0 is said to be *empty*. Although empty, in the population the cells true probability is almost always positive. That is, it is theoretically possible to have observations in the cell, and a positive count would occur if the sample size were sufficiently large. To emphasize this, such an empty cell is often called a *sampling zero*.

Depending on the model, sampling zeroes can cause ML estimates of model parameters to be infinite. When all cell counts are positive, all parameter estimates are necessarily finite. When any marginal counts corresponding to terms in a model equal zero, infinite estimates occur for that term.

5.3.2 Infinite Effect Estimate: Categorical Predictors

- For instance, consider a three-way table with binary predictors X_1 and X_2 for a binary response Y. When a marginal total equals zero in the 2×2 table relating Y to X_1, then the ML estimate of the effect of X_1 in the logistic regression model is infinite.
- ML estimates are finite when all the marginal totals corresponding to terms in the model are positive.
- Empty cells and sparse tables can also cause bias in estimators of odds ratios.
- When a ML parameter estimate is infinite, this is not fatal to data analysis. When the ML estimate of an odds ratio is $+\infty$, a likelihood-ratio confidence interval has a finite lower bound.

5.3.2 Infinite Effect Estimate: Categorical Predictors

- When your software's fitting processes fail to converge because of infinite estimates, adding a very small constant (such as 10^{-8}) is adequate for ensuring convergence.
- For each possibly influential observation, delete it or move it to another cell to check how much the results vary with small perturbations to the data.
- Often, some associations are not affected by the empty cells and give stable results for the various analyses, whereas others that are affected are highly unstable. Use caution in making conclusions about an association if small changes in the data are influential.
- The Bayesian approach to statistical inference typically provides a finite estimate in cases for which an ML estimate is infinite. See O'Hagan and Forster(2004) for details.

5.3.3 Example: Clinical Trial with Sparse Data

Table 5.7. Clinical Trial Relating Treatment (X) to Response (Y) for Five Centers (Z), with XY and YZ Marginal Tables

		Response (Y)		YZ Marginal	
Center (Z)	Treatment (X)	Success	Failure	Success	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	Active drug	12	36		
Marginal	Placebo	4	42		

Source: Diane Connell, Sandoz Pharmaceuticals Corp.

5.3.3 Example: Clinical Trial with Sparse Data

Example

The purpose was to compare an active drug to placebo for treating fungal infections (1 = success, 0 = failure). For these data, let Y = Response, X = Treatment (Active drug or Placebo), and Z = Center. Centers 1 and 3 had no successes. Thus, the 5×2 marginal table relating center to response, collapsed over treatment, contains zero counts. This marginal table is shown in the last two columns of Table 5.7

$$\text{logit}[P(Y = 1)] = \alpha + \beta x + \beta_k^Z$$

5.3.3 Example: Clinical Trial with Sparse Data

Because centers 1 and 3 had no successes, the ML estimates of the terms β_1^Z and β_3^Z pertaining to their effects equal $-\infty$. The fitted logits for those centers equal $-\infty$, for which the fitted probability of success is 0.

5.3.3 Example: Clinical Trial with Sparse Data

- The empty cells in Table 5.7 affect the center estimates, but not the treatment estimates, for this model. The estimated log odds ratio equals 1.55 for the treatment effect ($SE = 0.70$). The deviance (G^2) goodness-of-fit statistic equals 0.50 ($df = 4, P = 0.97$).
- Centers with no successes or with no failures can be useful for estimating some parameters, such as the difference of proportions, but they do not help us estimate odds ratios for logistic regression models or give us information about whether a treatment effect exists in the population.

5.3.4 Effect of Small Samples on X^2 and G^2 Tests

- When a model for a binary response has only categorical predictors, the true sampling distributions of goodness-of-fit statistics are approximately chi-squared, for large sample size n.
- The X^2 statistic tends to be valid with smaller samples and sparser tables than G^2. The distribution of G^2 is usually poorly approximated by chi-squared when $n/(number of cells)$ is less than 5.
- For fixed values of n and the number of cells, the chi-squared approximation is better for tests with smaller values of df.
- When cell counts are so small that chi-squared approximations may be inadequate, one could combine categories of variables to obtain larger counts. However,...

Outline

For inference about logistic regression parameters, the ordinary sampling distributions are *approximately* normal or chi-squared. The approximation improves as the sample size increases. For small samples, it is better to use the *exact* sampling distributions. Methods that find and use the exact distribution are now feasible due to recent advances in computer power and software.

5.4.1 Conditional Maximum Likelihood Inference

Conditional MLE

The conditional maximum likelihood estimate of a parameter is the value at which the conditional likelihood function achieves it maximum. When the sample size is small, or when there are many parameters relative to the sample size, conditional ML estimates of parameters work better than ordinary ML estimators.

5.4.1 Conditional Maximum Likelihood Inference

The exact inference approach deals with the primary parameters of interest using a conditional likelihood function that eliminates the other parameters. The technique uses a conditional probability distribution defined for potential samples that provide the same information about the other parameters that occurs in the observed sample. The distribution and the related conditional likelihood function depend only on the parameters of interest.

5.4.2 Small-Sample Tests for Contingency Tables

Single explanatory variable

$$\text{logit}[\pi(x)] = \alpha + \beta x$$

Fixing both sets of marginal totals yields a hypergeometric distribution for n_{11}, for which the probabilities do not depend on unknown parameters. The resulting exact test of $H_0 : \beta = 0$ is the same as Fisher's exact test

The unknown parameter α refers to the relative number of outcomes of $y = 1$ and $y = 0$, which are the column totals. Software eliminates α from the likelihood by conditioning also on the column totals, which are the information in the data about α.

5.4.2 Small-Sample Tests for Contingency Tables

Two explanatory factors

$$\text{logit}(\pi) = \alpha + \beta x + \beta_k^Z$$

The exact test eliminates the other parameters by conditioning on the marginal totals in each partial table. This gives an exact test of conditional independence between X and Y , controlling for Z.

For $2 \times 2 \times K$ tables $\{n_{ijk}\}$, conditional on the marginal totals in each partial table, the Cochran-Mantel-Haenszel test of conditional independence is a large-sample approximate method that compares $\sum_k n_{11k}$ to its null expected value.

5.4.3 Example: Promotion Discrimination

Table 5.8 Promotion Decisions by Race and by Month

	July Promotions		August Promotions		September Promotions	
Race	Yes	No	Yes	No	Yes	NO
Black	0	7	0	7	0	8
White	4	16	4	13	2	13

Source:J. Gastwirth, Statistical Reasoning in Law and Public Policy, Academic Press, New York(1988), p.266

5.4.3 Example: Promotion Discrimination

Example

Table 5.8 refers to US Government computer specialists of similar seniority considered for promotion from classification level GS-13 to level GS-14. The table cross classifies promotion decision, considered for three separate months, by employee's race. We test conditional independence of promotion decision and race. The table contains several small counts. The overall sample size is not small ($n = 74$), but one marginal count (collapsing over month of decision) equals zero, so we might be wary of using the CMH test.

5.4.3 Example: Promotion Discrimination

- $H_a : \beta < 0$. This corresponds to potential discrimination against black employees, their probability of promotion being lower than for white employees. The observed $\sum_k n_{11k} = 0$.
- Because the sample result is the most extreme possible, the conditional ML estimator of the effect of race in the logistic regression model is $\hat{\beta} = -\infty$.
- A two-sided P-value, based on summing the probabilities of all tables having probabilities no greater than the observed table, equals 0.056. There is some evidence, but not strong, that promotion is associated with race.

5.4.4 Small-Sample Confidence Intervals for Logistic Parameters and Odds Ratios

The 95% confidence interval for β consists of all values β_0 for which the P- value for $H_0 : \beta = \beta_0$ is larger than 0.05 in the exact test.

Consider again Table 5.8 on promotion decisions and race. When $\hat{\beta}$ is infinite, a confidence interval is still useful because it reports a finite bound in the other direction. StatXact reports an exact 95% confidence interval for β of $(-\infty, 0.01)$. This corresponds to the interval $(e^{-\infty}, e^{0.01}) = (0, 1.01)$ for the true conditional odds ratio in each partial table.

5.4.5 Limitations of Small-Sample Exact Methods

- Although the use of exact distributions is appealing, the conditioning on certain margins can make that distribution highly discrete.
- To alleviate conservativeness, we recommend inference based on the mid P-value.
- When any predictor is continuous, the discreteness can be so extreme that the exact conditional distribution is degenerate-it is completely concentrated at the observed result.
- Generally, small-sample exact conditional inference works with contingency tables but not with continuous predictors.

Outline

5.5.1 Sample Size for Comparing Two Proportions

Power

Consider the hypothesis that the group "success" probabilities π_1 and π_2 are identical. We could conduct a test for the 2×2 table that cross-classifies group by response, rejecting H_0 if the P-value $\leq \alpha$ for some fixed α. To determine sample size, we must specify the probability β of failing to detect a difference between π_1 and π_2 of some fixed size considered to be practically important. For this size of effect, β is the probability of failing to reject H_0 at the α level. Then, $\alpha = P(\text{type I error})$ and $\beta = P(\text{type II error})$. The *power* of the test equals $1 - \beta$.

5.5.1 Sample Size for Comparing Two Proportions

A study using equal group sample sizes requires approximately

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

This formula also provides the sample sizes needed for a comparable confidence interval for $\pi_1 - \pi_2$. Then, α is the error probability for the interval and β equals the probability that the confidence interval indicates a plausible lack of effect, in the sense that it contains the value zero.

5.5.1 Sample Size for Comparing Two Proportions

Hypothesis Testing

The null hypothesis $H_0 : \pi_1 = \pi_2$ in a 2×2 table corresponds to one for a parameter in a logistic regression model having the form

$$\text{logit}(\pi) = \beta_0 + \beta_1 x$$

where $x = 1$ for group 1 and $x = 0$ for group 2. (We use the β_0 and β_1 notation so as not to confuse these with the error probabilities.) H_0 corresponds to a log odds ratio of 0, or $\beta_1 = 0$. Thus, this example relates to sample size determination for a simple logistic regression model.

5.5.2 Sample Size in Logistic Regression

$$n = [z_\alpha + z_\beta \exp(-\lambda^2/4)]^2(1 + 2\bar{\pi}\delta)/(\bar{\pi}\lambda^2)$$

where

$$\delta = [1 + (1 + \lambda^2)\exp(5\lambda^2/4)]/[1 + \exp(-\lambda^2/4)],$$

$\lambda = \log(\theta)$, $\bar{\pi}$ is the probability of success at the mean of x.

5.5.2 Sample Size in Logistic Regression

Example

The dependence of the probability of severe heart disease on $x =$ cholesterol level for a middle-aged population. $H_0 : \beta_1 = 0$ v.s. $H_a : \beta_1 > 0$. Suppose previous studies have suggested that $\bar{\pi}$ is about 0.08, and we want the test to be sensitive to a 50% increase (i.e., to 0.12), for a standard deviation increase in cholesterol. $\alpha = 0.05$, $\beta = 0.10$.

5.5.3 Sample Size in Multiple Logistic Regression

Let R denote the multiple correlation between the predictor X of interest and the others in the model. One divides the above formula for n by $(1 - R^2)$. In that formula, $\bar{\pi}$ denotes the probability at the mean value of all the explanatory variables, and the odds ratio refers to the effect of the predictor of interest at the mean level of the others.

Outline

Homework 5

1. Analysis the GDS5037 data. Suppose the samples are randomly chosen. Let Y denote patient's status, $Y = 1$ for mild asthma (MMA) and severe asthma (SA), $Y = 0$ for control.
   (1) Use one and only one identifier each time as the independent variable to built a logistic regression. Try all identifiers. Report top 5 best models based on AIC.
   (2) In (1), Calculate the 5 identifiers' sample correlation matrix.
   (3) Choose 10 identifiers randomly as independent variables and use stepwise method to choose the best model with just main effect. Is your model good enough? If yes, why? If no, do you have any suggestions to improve it?

2. Problems in textbook 5.3, 5.4, 5.10, 5.17, 5.20, 5.22.