STAT 7030: Categorical Data Analysis

6. More Discussions on Logistic Regression

Peng Zeng

Department of Mathematics and Statistics
Auburn University

Fall 2012

Outline

1. Logistic regression for contingency tables
   - Conditional association
   - AZT and AIDS symptoms
   - A clinic study
2. More topics on logistic regression
   - Model selection
   - Model checking
3. Case study: UIS

Logistic Regression for Contingency Tables

When all the variables are categorical, the data are usually presented in terms of a contingency table.

We can analyze a contingency table using logistic regression if one variable is response and the remaining ones are predictors.

When there is only one predictor, the table is $I \times 2$. The advantage of logistic regression is not clear.

When there are more than one predictor, it is better to analyze the contingency table using a model approach.

We first look at some concepts in terms a three-way contingency table.

Confounding

When studying the effect of X on Y, one should control confounding variables that can influence that relationship because they are associated with both X and Y.

- Kid's age (Z) is associated with both reading ability (Y) and height (X).

In order to eliminate the influence of confounding variables,

- In experimental studies, randomly assigning subjects to different levels of X.
- In observational studies, analyze data at fixed levels of covariate.

Partial Tables

We control for Z by studying X–Y relationship at fixed levels of Z.

Partial table splits the original two-way contingency table according to levels of Z. Original table is called XY marginal table.

The associations in partial tables are called conditional associations, because they refer to the effect of X on Y conditional on fixing Z at some level.

It may be misleading to study only the marginal table.

Death Penalty Example

The 674 subjects were the defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987.

Victims's Race	Defendant's Race	Death Penalty		Percent Yes
		Yes	No	
white	white	53	414	11.3
	black	11	37	22.9
black	white	0	16	0.0
	black	4	139	2.8
total	white	53	430	11.0
	black	15	176	7.9

The three variables are

Y death penalty verdict (yes, no)
X race of defendant (white, black)
Z race of victims (white, black)

Simpson's paradox

From the marginal table, the white receive death penalty more likely than the black.

From the partial table, the black receive death penalty more likely than the white in both categories.

Why?? Because of strong association between victim's race and defendant's race.

	defendant	
victim	white	black
white	467	48
black	16	143

The result that a marginal association can have a difference direction from each conditional association is called Simpson's paradox.

Conditional Odds Ratio

For $2 \times 2 \times K$ tables, where K is the number of categories of Z. Let $\{n_{ijk}\}$ denote cell frequencies.

Within a fixed category k of Z, the conditional odds ratio is

$$\hat{\theta}_{XY(k)} = n_{11k} n_{22k} / n_{12k} n_{21k}$$

These can be quite different from the marginal odds ratio,

$$\hat{\theta}_{XY} = n_{11+} n_{22+} / n_{12+} n_{21+}$$

Example. In the example, $\hat{\theta}_{XY(i)} < 1$ but $\hat{\theta}_{XY} > 1$.

$$\hat{\theta}_{XY(1)} = 0.43, \quad \hat{\theta}_{XY(2)} = 0, \quad \hat{\theta}_{XY} = 1.45$$

Conditional Independence

If X and Y are independent in partial table k, then X and Y are conditionally independent at level k of Z.

$$P(Y = j \mid X = i, Z = k) = P(Y = j \mid Z = k), \quad \text{for all } i, j$$

X and Y are conditionally independent given Z when they are independent at every level of Z.

Assume a single multinomial distribution to the three way table, and if X and Y are conditionally independent given Z, then

$$\begin{aligned}\pi_{ijk} &= P(X = i, Z = k)\, P(Y = j \mid X = i, Z = k) \\ &= \pi_{i+k} P(Y = j \mid Z = k) = \pi_{i+k} P(Y = j, Z = k)/P(Z = k) \\ &= \pi_{i+k}\pi_{+jk}/\pi_{++k}, \quad \text{for all } i, j, \text{ and } k\end{aligned}$$

Conditional vs Marginal Independence

Conditional independence does not imply marginal independence.

		Y =Response	
Z = Clinic	X = Treatment	Yes	No
1	A	18	12
	B	12	8
2	A	2	8
	B	8	32
total	A	20	20
	B	20	40

$$\theta_{XY(1)} = \theta_{XY(2)} = 1.0, \quad \text{but } \theta_{XY} = 2.0$$

Homogeneous Association

A $2 \times 2 \times K$ table has homogeneous XY association when

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$$

Conditional independence of X and Y is a special case when $\theta_{XY(k)} = 1.0$.

When it occurs, it is said there is no interaction between two variables in their effects on the other variable.

The Breslow-Day test in SAS can be used to test homogeneous association. Under H_0, the statistic approximately follows χ^2 with $K - 1$ degrees of freedom.

SAS Code

```
proc freq data = SAS-Data-set order = data;
   weight count;
   table zvar * xvar * yvar / measures cmh;
run;
```

where xvar and yvar are the two variables of interest and zvar is the covariate.

The option cmh requires Breslow-Day test for homogeneity of the conditional odds ratios.

Example: AZT

The following data are from a study on the effects of AZT in slowing the development of AIDS symptoms. In the study, 338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait until their T-cells showed severe immune weakness.

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

The response is binary and there are two categorical predictors.

Logistic Regression Model

Consider the following logistic regression

$$\text{logit}[P(Y=1)] = \beta_0 + \beta_1 x + \beta_2 z$$

where Y denotes whether AIDS symptoms develop, X and Z are dummy variables for AZT treatment and race, respectively.

$$x = \begin{cases} 1, & \text{for immediate AZT use} \\ 0, & \text{otherwise} \end{cases} \qquad z = \begin{cases} 1, & \text{for whites} \\ 0, & \text{for blacks} \end{cases}$$

In this model, we assume there is no interaction between X and Z. The effect of one factor is the same at each level of the other factor.

Interpret Parameters

The following table shows the logit values at the four combinations of values of the two predictors.

x	z	logit
1	1	$\beta_0 + \beta_1 + \beta_2$
0	1	$\beta_0 + \beta_2$
1	0	$\beta_0 + \beta_1$
0	0	β_0

Therefore, the three parameters can be interpreted as

- β_0 is the log odds of developing AIDS symptoms for black subjects without immediate AZT use.
- β_1 is the increment to the log odds for those with immediate AZT use.
- β_2 is the increment to the log odds for white subjects.

More Interpretation

At a fixed level z of Z, the effect on the logit of changing categories of X is

$$\begin{aligned}\beta_1 &= [\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] \\ &= \log \frac{\text{odds at } x = 1 \text{ and } z}{\text{odds at } x = 0 \text{ and } z}\end{aligned}$$

Thus e^{β_1} is the conditional odds ratio between X and Y.

This conditional odds ratio is the same at each level of Z, that is, there is homogeneous XY association controlling for Z.

When $\beta_1 = 0$, the common odds ratio equals 1. In this case, X and Y are independent in the partial table, or conditional independent given Z.

Alternative Representation

An alternative representation of such factors resembles the way that ANOVA models often express them.

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

where β_i^X denotes the effect on the logit of classification in category i of X.

Conditional independence between X and Y, given Z, corresponds to $\beta_1^X = \cdots = \beta_I^X$, which means $P(Y=1)$ does not change as i changes.

SAS Code and Output

```
proc logistic data = SAS-Dataset order = data;
   class var1 (ref = first) var2 / param = reference;
   model yvar = var1 var2;
run;
```

- param = reference specifies reference cell coding. The default is effect coding.
- ref = first designates the first ordered level as reference. Similarly, we can use ref = last to designates the last ordered level as reference, or just write ref = 'name-of-level'.

Different Coding Schemes

The following table shows parameter estimates for three ways of defining factor parameters.

	definition of parameters		
parameter	last = zero	first = zero	sum = zero
Intercept	−1.074	−1.738	−1.406
AZT yes	−0.720	0.000	−0.360
AZT no	0.000	0.720	0.360
Race white	0.056	0.000	0.028
Race black	0.000	−0.056	−0.028

Notice that the estimated logit/probability of developing AIDS symptoms ($\hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 z$) is the same. For example, logit of immediate AZT use and white race is

$$-1.074 - 0.720 + 0.056 = -1.738 + 0 + 0 = -1.406 - 0.360 + 0.028$$

Results: Estimation

Let us focus on the result when $x = 1$ means immediate AZT use and $z = 1$ means whites. The SAS output shows

Parameter	Estimate	Std Error	Chi-Square	Pr > ChiSq
Intercept	−1.0736	0.2629	16.6705	< .0001
azt yes	−0.7195	0.2790	6.6507	0.0099
race white	0.0555	0.2886	0.0370	0.8476

The estimated (conditional) odds ratio between immediate AZT use and development of AIDS symptoms equals $e^{-0.7195} = 0.487$. For each race, the estimated odds of symptoms are half as high for those who took AZT immediately.

The Wald 95% confidence interval for this effect is

$$\exp\{-0.7195 \pm (1.96)(0.279)\} = (0.28, 0.84)$$

Testing Significance of Model

Test whether the prob of developing AIDS symptoms depends on immediate AZT use and race.

$$H_0 : \beta_1 = \beta_2 = 0, \quad H_a : \text{not both } \beta_1 \text{ and } \beta_2 \text{ are zero}$$

In this case,

$$\begin{aligned} \text{reduced model:} \quad & \text{logit}(\pi_{ik}) = \beta_0 \\ \text{full model:} \quad & \text{logit}(\pi_{ik}) = \beta_0 + \beta_1 x + \beta_2 z \end{aligned}$$

The likelihood ratio statistic is $342.118 - 335.151 = 6.967$ with 2 degrees of freedom. The p-value is 0.0307, and we should reject H_0.

Results: Testing Single Parameter

The hypothesis of conditional independence of AZT treatment and development of AIDS symptoms, controlling for race, is

$$H_0 : \beta_1 = 0, \quad H_a : \beta_1 \neq 0.$$

The Wald chi-squared statistic is $(-0.7195/0.2790)^2 = 6.6507$ with p-value 0.0099. We should reject H_0, which means AZT treatment and development of AIDS symptoms are not conditional independent.

The hypothesis of conditional independence of race and development of AIDS symptoms, controlling for AZT treatment is

$$H_0 : \beta_2 = 0, \quad H_a : \beta_2 \neq 0.$$

The test statistic is $(0.0555/0.2886)^2 = 0.0370$ with p-value 0.8476. Thus H_0 is accepted, which means the probability of developing AIDS symptoms does not depend on race, controlling for AZT treatment.

Goodness-of-fit Test

Recall that likelihood ratio test compares the maximized log-likelihood (or equivalently deviance) of a reduced model to that of a full model.

Goodness-of-fit test compare the model of interests

$$\text{logit}(\pi_{ik}) = \beta_0 + \beta_1 x + \beta_2 z$$

to the saturated model

$$\text{logit}(\pi_{ik}) = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$$

The likelihood ratio test statistic is $G^2 = 1.38$ with one degree of freedom. The Pearson chi-squared statistic is $X^2 = 1.39$. We should accept H_0, which means the model without interaction adequately model the variability in data.

Conditional Associations

The table (in the next slide) shows results of a clinic trail with eight centers. The study compared two cream preparations, an active drug and a control, on their success in curing an infection.

This table illustrate a common pharmaceutical application, comparing two treatments on a binary response with observations from several strata (for example, clinics, age groups, etc.).

Use logistic regression to investigate whether an association exists between a treatment variable and a disease outcome after controlling for a possibly confounding variable that might influence that association.

Example: Clinic Trails

		response		
center	treatment	success	failure	odds-ratio
1	drug	11	25	1.19
	control	10	27	
2	drug	16	4	1.82
	control	22	10	
3	drug	14	5	4.80
	control	7	12	
4	drug	2	14	2.29
	control	1	16	
5	drug	6	11	inf
	control	0	12	
6	drug	1	10	inf
	control	0	10	
7	drug	1	4	2.00
	control	1	8	
8	drug	4	2	0.33
	control	6	1	

Logistic Regression and its Application

Let $\pi_{ik} = P(Y = 1|x = i, Z = k)$ and consider the model

$$\text{logit}(\pi_{ik}) = \alpha + \beta x + \beta_k^Z, \quad i = 0, 1;\ k = 1, \ldots, 8$$

where x is a dummy variable for treatment

$$x = 1 \text{ for drug and } x = 0 \text{ for control.}$$

The unknown parameters are α, β, $\{\beta_1^Z, \ldots, \beta_8^Z\}$. We need a constraint on β_k^Z, for example $\beta_8^Z = 0$.

This model implies homogeneous XY association because there is no interaction between X and Z.

Homogeneous Association

Recall that a $2 \times 2 \times K$ table has homogeneous XY association when

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}$$

The logistic regression provides a way for testing homogeneous XY association. A test of homogeneous association is essentially a goodness-of-fit test of model $\text{logit}(\pi_{ik}) = \alpha + \beta x + \beta_k^Z$. The full model or the saturated model is the one with interactions. The degrees of freedom of G^2 or X^2 are $K - 1$.

In this example, $G^2 = 277.020 - 267.274 = 9.746$ with df $= 7$. The critical value is $\chi^2_{7,0.05} = 14.07$. Therefore, we should accept H_0, which means the table has homogeneous XY association.

Another test is the Breslow-Day test given in proc freq.

Common Odds Ratio from Model

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.

From the logistic regression $\text{logit}(\pi_{ik}) = \alpha + \beta x + \beta_k^Z$, the common (conditional) odds ratio is estimated by $e^{\hat{\beta}}$.

For this example, the SAS output shows

$$\hat{\beta} = 0.7769, \quad \text{SE}(\hat{\beta}) = 0.3067.$$

Therefore, an estimate of the common conditional odds ratio is

$$e^{0.7769} = 2.1747$$

An approximate 95% confidence interval is

$$e^{0.7769 \pm (1.96)(0.3067)} = (e^{0.1758}, e^{1.3780}) = (1.1922, 3.9670)$$

Common Odds Ratio

An alternative estimate of the common conditional odds ratio is

$$\hat{\theta}_{MH} = \frac{\sum_k (n_{11k} n_{22k} / n_{++k})}{\sum_k (n_{12k} n_{21k} / n_{++k})}$$

It is preferred over the MLE when K is large and the data are sparse (each stratum has few observations).

The common conditional odds ratio in this example is

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

Testing Conditional Independence

The conditional independence is a special case of homogeneous association when $\theta_{XY(k)} = 1$.

Consider the logistic regression

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

The null hypothesis of XY conditional independence is $H_0 : \beta = 0$. We can either use a Wald test or a likelihood ratio test.

Notice that $\hat{\beta} = 0.7769$ with SE $= 0.3067$. The p-value is 0.0113. We should reject H_0, which means X and Y are not conditionally independent.

An alternative approach is the Cochran-Mantel-Haenszel test of conditional independence in proc freq.

Model Selection

The model selection process determines which predictors and/or interactions should be included in the final model.

Two competing goals for model selection

- The model should be complicated enough to fit the data well.
- Simpler models are easier to interpret.

It is recommended that at least 10 outcomes of each type should occur for every predictor (only an approximate rule).

For example. If $y = 1$ only 30 times out of $n = 3000$, the model should contain no more than about three predictors.

More Comments

It is helpful first to study the effect of each predictor on Y by itself using graphics for a continuous predictor or a contingency table for a discrete predictor.

Cautions that apply to ordinary linear regression hold for any generalized linear model.

For example. Models with several predictors often suffer from multi-collinearity. Correlations among predictors make it seem that no one variable is important when all the others are in the model.

Example: Horseshoe Crab Data

Consider the horseshoe crab data again. Fit a logistic regression model containing all main effects.

$$\begin{aligned}\text{logit}(\pi) = \beta_0 + \beta_1\,\text{weight} + \beta_2\,\text{width} \\ + \beta_3 c_1 + \beta_4 c_2 + \beta_5 c_3 + \beta_6 s_1 + \beta_7 s_2\end{aligned}$$

where π is the probability of a female crab having satellite, s_1 and s_2 are two dummy variables for spine condition (three categories)

$s_1 = 1$ means both spines are good and $s_1 = 0$ otherwise,

$s_2 = 1$ means one worn or broken and $s_2 = 0$ otherwise.

and c_1, c_2, c_3 are three dummy variables for color (four categories).

SAS Output

The test $H_0 : \beta_1 = \cdots = \beta_7 = 0$ is highly significant. The likelihood ratio statistic is 40.5565 with 7 degrees of freedom. The p-value is $< .0001$. At least one predictor has an effect.

Parameter	Estimate	Standard Error	Wald 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
c1	1.6087	0.9355	2.9567	0.0855
c2	1.5058	0.5667	7.0607	0.0079
c3	1.1198	0.5933	3.5624	0.0591
s1	−0.4003	0.5027	0.6340	0.4259
s2	−0.4963	0.6292	0.6222	0.4302

Many variables are not significant.

Results

The small p-value for the overall test yet the lack of significance for individual effects is a warning sign of multi-collinearity.

Weight and width have a strong correlation (0.887). For practical purpose they are equally good predictors, but it is nearly redundant to use them both in the model.

In the further analysis, we use width (W), color (C), and spine condition (S) as predictors.

Recall: how to do model selection in linear regression?

Stepwise Procedure

Forward selection adds terms sequentially until further additions do not improve the fit.

- Begin with a model containing only the intercept.
- At each stage, add the term giving the greatest improvement in fit. (minimum p-value for testing the selected term)

Backward elimination removes terms until further deletion leads to a significantly poorer fit.

- Begin with a complicated model (always contains all available predictors).
- At each stage, remove the term for which its removal has the least damaging effect on the model. (largest p-value)

Some Comments

For categorical predictors with more than two categories, add or delete the entire variable at any stage rather than just individual dummy variables.

Each step corresponds to a testing that compares two nested models. (add if the test is rejected. delete if the test is accepted.)

Forward selection and backward elimination do not necessarily lead to the same model. Many statisticians prefer backward elimination.

Stepwise selection means at each stage of forward selection we retests terms added at previous stages to see if they are still significant.

Results for Horseshoe Crab Example

For simplicity, we symbolize models by their highest-order terms.

- $(C + S + W)$ denotes a model with main effects
- $(C + S * W)$ denotes a model that has main effects plus an $S \times W$ interaction
- .....

Apply backward elimination algorithm to the horseshoe crab data

- We start with a model containing W, C, S and their interactions $(W * C * S)$, and sequentially remove insignificant predictors.

(see the next slide)

Backward Elimination

					Models	Deviance		
	Predictor	-2*logL	AIC	df	Compared	Diff	df	P-val
1	(C*S*W)	170.446	212.446	152	---	---		
2	(C*S+C*W+S*W)	173.674	209.674	155	(2)-(1)	3.228	3	0.36 X
3a	(C*S+S*W)	177.336	207.336	158	(3a)-(2)	3.662	3	0.30
3b	(C*W+S*W)	181.559	205.559	161	(3b)-(2)	7.885	6	0.25
3c	(C*S+C*W)	173.677	205.677	157	(3c)-(2)	0.003	2	1.00 X
4a	(S+C*W)	181.637	201.637	163	(4a)-(3c)	7.96	6	0.24
4b	(W+C*S)	177.597	203.597	160	(4b)-(3c)	3.92	3	0.27 X
5	(C+S+W)	186.612	200.612	166	(5)-(4b)	9.015	6	0.17 X
6a	(C+S)	208.834	220.834	167	(6a)-(5)	22.222	1	0.00
6b	(S+W)	194.425	202.425	169	(6b)-(5)	7.813	3	0.05
6c	(C+W)	187.457	197.457	168	(6c)-(5)	0.845	2	0.66 X
7a	(C)	212.061	220.061	169	(7a)-(6c)	24.604	1	0.00
7b	(W)	194.453	198.453	171	(7b)-(6c)	6.996	3	0.07
8	(C=dark + W)	187.958	193.958	170	(8)-(6c)	0.501	2	0.78 X
9	None	225.759	227.759	172	(9)-(8)	37.801	2	0.00

Comments

How many parameters in model $C * S * W$?

intercept	1
main effects	C: 3, S: 2, W: 1
two-factor interactions	$C * S$: 6, $C * W$: 3, $S * W$: 2
three-factor interactions	$C * S * W$: 6

The total parameters should be

$$1 + (3 + 2 + 1) + (6 + 3 + 2) + 6 = 24$$

and the df should be $n - 24 = 173 - 24 = 149$?

Why df = 152? SAS output shows that three parameters of the three-factor interactions are set to be 0 because the corresponding variables are exactly linear combinations of some other variables.

More Explanation

The fact that there is only one observation in certain cells makes some parameters not estimable.

	spine			
color	1 (s_1)	2 (s_2)	3	total
2 (c_1)	9	2	1	12
3 (c_2)	24	8	63	95
4 (c_3)	3	4	37	44
5	1	1	20	22
total	37	15	121	173

$c_1 - c_1 s_1 - c_1 s_2 = 1$ for one observation and 0 otherwise.
$s_1 - c_1 s_1 - c_2 s_1 - c_3 s_1 = 1$ for one observation and 0 otherwise.
$s_2 - c_1 s_2 - c_2 s_2 - c_3 s_2 = 1$ for one observation and 0 otherwise.

If all frequencies are > 1, then all the interactions are estimable.

SAS Code

```
proc logistic data = SAS-Dataset;
   model response = predictor-list / selection = stepwise
                slentry = p1 slstay = p2;
run;
```

The options are

selection	specifies effect selection method (backward, forward, stepwise)
slentry	specifies significance level for entering effects (add if p-value is smaller than "slentry")
slstay	specifies significance level for removing effects (remove if p-value is larger than "slstay")

More options can be found in the SAS online documents.

Akaike Information Criterion

Other criteria besides significant tests can help select a good model.

Akaike information criterion (AIC) judges a model by how close its fitted values tend to be to the true values. The smaller, the better.

$$\begin{aligned} \text{AIC} &= -2(\text{max log likelihood} - \#\text{ of parameter}) \\ &= -2(\text{max log likelihood}) + 2(\#\text{ of parameter}) \end{aligned}$$

The parameters include the intercept.

In the previous table, we should choose model ($C = dark + W$) according to AIC.

Classification Tables

A classification table can be used to summarize the predictive power of a binary regression model.

The prediction is $\hat{y} = 1$ when $\hat{\pi}_i > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi}_i \leq \pi_0$.

- One possibility is to take $\pi_0 = 0.5$.
- Or choose π_0 as the sample proportion of successes ($y = 1$).

A classification table cross classifies the binary outcome y with a prediction of whether $y = 0$ or 1.

	pred $\pi_0 = 0.64$		pred $\pi_0 = 0.50$		
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	34	28	62

Sensitivity and Specificity

Two useful summaries of predictive power are

$$\text{sensitivity} = P(\hat{y} = 1 | y = 1). \quad \text{specificity} = P(\hat{y} = 0 | y = 0).$$

When $\pi_0 = 0.64$, the estimated sensitivity is $74/111 = 0.667$ and the specificity is $42/62 = 0.677$.

A classification table has limitations:

- It collapses continuous predictive value $\hat{\pi}$ into binary ones.
- The choice of π_0 is arbitrary.

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.

ROC curve usually has a concave shape connecting the points $(0, 0)$ and $(1, 1)$.

The area under a ROC curve is the concordance index c, which estimates the probability that the predictions and the outcomes are concordant, which means that the observations with the larger y also has the larger $\hat{\pi}$.

- The larger the concordance index is, the better.
- $c = 0.5$ corresponds random guessing.

SAS Code for ROC Curve

```
ods listing gpath = 'your-directory';
ods graphics on;
proc logistic data = SAS-Dataset plots = roc;
   model resp = list-of-predictors;
run;
```

Assess Model Adequacy

We can assess model adequacy by

- comparing the model to a more complicated one (with more interactions or higher-order terms).
- applying goodness-of-fit test.

When there is only one predictor and it is continuous, we have discussed how to conduct a goodness-of-fit test by grouping the data for the crab example.

When there are more than one continuous variables, we can conduct Hosmer-Lemeshow test.

Hosmer-Lemeshow Test

The Hosmer and Lemeshow test for a logistic model is requested by specifying the lackfit option in the model statement (after /).

- The subjects are divided into approximately ten groups of roughly the same size based on the percentiles of the estimated probabilities.
- The discrepancies between the observed and expected number of observations in these groups are summarized by a statistic similar to the Pearson chi-squared statistic.
- Compared the statistic to χ^2_d distribution where d is the number of groups minus 2. A large p-value (> 0.05) indicates the model fits the data.

Notice that we used a similar strategy for crab data, and construct the groups according to the width of crabs.

Pearson Residuals

The Hosmer-Lemeshow test (or other test for assessing model adequacy) checks model fit in a global sense. Diagnostic analysis may suggest a reason for the lack of fit.

Let y_i denote the number of "successes" for n_i trails at setting i of the explanatory variables. Let $\hat{\pi}_i$ denote the estimated probability of of success. Thus the raw residual is $(y_i - n_i\hat{\pi}_i)$.

For a GLM with binomial random component, the Pearson residual

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}}$$

When the model holds, the mean of e_i is approximately zero, but its variance is smaller than one.

More Residuals

The standardized Pearson residual divides $(y_i - n_i\hat{\pi}_i)$ by its SE.

$$\frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)(1-h_i)}}$$

where h_i called the observation's leverage.

- The greater an observation's leverage, the greater its potential influence on the model fit.
- The standardized Pearson residual approximately follows $N(0,1)$.
- An absolute value larger than roughly 2 or 3 provides evidence of lack of fit.

We can also use deviance residuals and/or standardized deviance residuals. They can be interpreted similarly.

Example: Admission to Graduate School

The table (next slide) refers to graduate school applications to the 23 departments in a university. It cross-classifies whether the applicant was admitted, the applicant's gender, and the applicant's department.

Other things being equal, one would hope that admissions decision is independent of gender. Consider the following model

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

where π_{ik} is the probability of admission for gender i in department k.

This model fits rather poorly. (The df are 23 for both statistics.)
The Pearson chi-squared statistic is $X^2 = 40.9$ with p-value 0.012.
The likelihood ratio statistic is $G^2 = 44.7$ with p-value 0.004.

The standardized Pearson residuals for the number of females who were admitted for this model.

	Female		Male		StResChi			Female		Male		StResChi
DEPT	YES	NO	YES	NO	(Fem.Yes)		DEPT	YES	NO	YES	NO	(Fem.Yes)
anth	32	81	21	41	-0.76457		ling	21	10	7	8	1.37298
astr	6	0	3	8	2.87096	*	math	25	18	31	37	1.28844
chem	12	43	34	110	-0.26830		phil	3	0	9	6	1.34164
clas	3	1	4	0	-1.06904		phys	10	11	25	53	1.32458
comm	52	149	5	10	-0.63260		poli	25	34	39	49	-0.23318
comp	8	7	6	12	1.15752		psyc	2	123	4	41	-2.27222 *
engl	35	100	30	112	0.94209		reli	3	3	0	2	1.26491
geog	9	1	11	11	2.16641	*	roma	29	13	6	3	0.13970
geol	6	3	15	6	-0.26082		soci	16	33	7	17	0.30123
germ	17	0	4	1	1.88730		stat	23	9	36	14	-0.01229
hist	9	9	21	19	-0.17627		zool	4	62	10	54	-1.75873
lati	26	7	25	16	1.64564							

In some department, the probability of admitting females is much larger/smaller than expected.

SAS Code

```
proc genmod data = SAS-dataset;
   model y = x / dist = bin link = logit r;
   output out = residuals stdreschi = streschi;
run;
```

Create a new SAS dataset containing residuals. The names can be reschi, resdev, stdreschi, stdreschi, ...

Results

Significantly more females were admitted than the model predicts in the astronomy and geography departments, and fewer than in the psychology department.

Without these three departments, the model fits reasonably well. The Pearson chi-squared statistic is $X^2 = 22.8$ with p-value 0.30. The likelihood ratio statistic is $G^2 = 24.4$ with p-value 0.225. The df are 20 for both statistics.

For the complete data, adding a gender effect to the model does not provide an improved fit.

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

The Pearson chi-squared statistic is $X^2 = 39.0$ with p-value 0.014. The likelihood ratio statistic is $G^2 = 42.4$ with p-value 0.006. The df are 22 for both statistics.

Influence Diagnostics

Some observations may have too much influence in determining the parameter estimates. The fit could be quite different if they were deleted.

Influence diagnostics usually relate to the effect on certain characteristics of removing the observation from the data set. Influence measures for each observation include:

- 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.
- 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.
- The change in X^2 or G^2 goodness-of-fit statistics when the observation is deleted.

Example: Coronary Heart Disease

A sample of male residents aged 40 through 59 were classified on blood pressure. Consider the following model

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

where π is the probability of developing coronary heart disease during a six-year follow-up period and x is blood pressure.

The overall fit statistics do not indicate lack of fit. The Pearson chi-squares statistic is $X^2 = 6.3$ with p-value 0.39. The likelihood ratio statistic is $G^2 = 5.9$ with p-value 0.43. The df are 6 for both statistics.

Add influence in the model statement of proc logistic for influence diagnostics. Plots are available with ods graphics on;.

Results for Coronary Heart Disease

The following table list the dfbeta measure for the coefficient of blood pressure, the confidence interval diagnostic c, the change in X^2 and the change in G^2.

blood pressure	total	$y = 1$	scores	dfbeta	c	Pearson difference	LR difference
< 117	156	3	111.5	0.49	0.34	1.22	1.39
$117 - 126$	252	17	121.5	-1.14	2.27	5.64	5.04
$127 - 136$	284	12	131.5	0.33	0.31	0.89	0.94
$137 - 146$	271	16	141.5	0.08	0.09	0.33	0.34
$147 - 156$	139	12	151.5	0.01	0.00	0.02	0.02
$157 - 166$	85	8	161.5	-0.07	0.02	0.11	0.11
$167 - 186$	99	16	176.5	0.40	0.26	0.42	0.42
> 186	43	8	191.5	-0.12	0.02	0.03	0.03

All the influence diagnostics show that deleting the second observation has the greatest effect.

Case Study: UIS

Consider a subset of data from the University of Massachusetts Aids Research Unit (UMARU) IMPACT Study (UIS). This was a 5-year (1989-1994) collaborative research project comprised of two concurrent randomized trials of residential treatment for drug abuse.

The purpose of the study was to compare treatment programs (A or B) of different planned durations (short or long) designed to reduce drug abuse and to prevent high-risk HIV behavior.

The UIS sought to determine whether alternative residential treatment approaches are variable in effectiveness and whether efficacy depends on planned program duration.

How Data Are Collected

The trial at site A randomized 444 participants and was a comparison of 3- and 6-month modified therapeutic communities which incorporated elements of health education and relapse prevention.

Clients in the relapse prevention/health education program (site A) were taught to recognize "high-risk" situations that are triggers to relapse and were taught the skills to enable them to cope with these situations without using drugs.

In the trial at site B, 184 clients were randomized to receive either a 6- or 12-month therapeutic community program involving a highly structured life-style in a communal living setting.

List of Variables

name	description	codes/values
id	identification code	1-575
age	age at enrollment	years
beck	Beck depression score at admission	0.000-54.000
ivhx	IV drug use history at admission	1 = never
		2 = previous
		3 = recent
ndrugtx	number of prior drug treatments	0-40
race	subject's race	0 = white
		1 = other
treat	treatment randomization assignment	0 = short
		1 = long
site	treatment site	0 = A
		1 = B
dfree	returned to drug use prior to the	1 = remained drug free
	scheduled end of the treatment program	0 = otherwise

Summary of Data

The number of observations is $n = 575$.

The response is dfree. In this study, 147 out of the 575 subjects (25.57%) remained drug free for at least one year.

The seven explanatory variables are

age, beck, ivhx, ndrugtx, race, treat, site.

The goal is to determine whether there is a difference between two treatment programs after adjusting for potential confounding and interaction variables.

Fit Univariate Logistic Regression

First, we fit the univariate logistic regression models to these data to assess how dfree depends on the individual variables.

	coeff	Std.Err.	odds ratio	95% CI	G2	P-value
age	0.018	0.0153	1.20	(0.89, 1.62)	1.40	0.237
beck	-0.008	0.0103	0.96	(0.87, 1.06)	0.63	0.425
ndrugtx	-0.075	0.0247	0.93	(0.88, 0.97)	11.84	<0.001
ivhx_2	-0.481	0.2657	0.62	(0.37, 1.04)		
ivhx_3	-0.775	0.2166	0.46	(0.30, 0.70)	13.35	0.001
race	0.459	0.2110	1.58	(1.04, 2.39)	4.62	0.032
treat	0.437	0.1931	1.55	(1.06, 2.26)	5.18	0.023
site	0.264	0.2034	1.30	(0.87, 1.94)	1.67	0.197

- The odds ratio for age is for a 10-year increase and the odds ratio for beck is for a 5-point increase. It is because a change of 1 year or 1 point would not be clinically meaningful.
- The 95% confidence intervals are for odds ratio.
- Some variables are not significant (age, beck, and site).
- The degrees of freedom are 2 for variable ivhx.

A Tentative Model

Fit a tentative multiple logistic regression with all variables but beck.

parameter	estimate	std err	chi-square	Pr > ChiSq
intercept	−2.4054	0.5548	18.7975	< .0001
age	0.0504	0.0173	8.4550	0.0036
ivhx 2	−0.6033	0.2872	4.4118	0.0357
ivhx 3	−0.7327	0.2523	8.4328	0.0037
ndrugtx	−0.0615	0.0256	5.7559	0.0164
race	0.2261	0.2233	1.0251	0.3113
treat	0.4425	0.1993	4.9302	0.0264
site	0.1486	0.2172	0.4681	0.4939

Some Comments

The table in the previous slide indicates weaker associations for some covariates when controlling for other variables. In particular, the significance level for the Wald test for the coefficient for site is $p = 0.4939$ and for race is $p = 0.3113$.

Strict adherence to conventional levels of statistical significance would dictate that we consider a smaller model deleting these two covariates.

However, due to the fact that subjects were randomized to treatment within site we keep site in the model. Race is an important control variable, and we also keep it.

For two Continuous Variables

We want to check the scale of the continuous covariates in the model.

For age, a linear term is enough. ($\chi^2_{1,0.05} = 3.84$)

	-2 LogL	vs null	vs linear
Not in model	627.801		
linear	619.248	8.553	
quadratic	618.931	8.870	0.317

For ndrgutx, a linear term is enough. ($\chi^2_{1,0.05} = 3.84$)

	-2 LogL	vs null	vs linear
Not in model	626.176		
linear	619.248	6.928	
quadratic	617.345	8.831	1.903

Add Interactions

Add one interaction to the main effect model at each time. Only two interactions are significant.

Effect	DF	Chi-Square	Pr > ChiSq
age*ndrugtx	1	4.3497	0.0370
age*ivhx	2	0.8274	0.6612
age*race	1	0.2495	0.6174
age*treat	1	2.5147	0.1128
age*site	1	1.2934	0.2554
ndrugtx*ivhx	2	0.3082	0.8572
ndrugtx*race	1	2.8832	0.0895
ndrugtx*treat	1	1.0151	0.3137
ndrugtx*site	1	0.0005	0.9826
race*ivhx	2	1.6843	0.4308
treat*ivhx	2	0.0870	0.9575
site*ivhx	2	0.6578	0.7197
race*treat	1	1.1631	0.2808
race*site	1	8.1020	0.0044
treat*site	1	0.1274	0.7212

Automatic Model Selection

We can also use variable selection procedure to identify a best subset of variables to build the model.

In this example, stepwise selection and backward elimination yield different results.

Goodness-of-fit tests are accepted for both models.

		Hosmer and Lemeshow Test		
Model selected by	# terms	Chi-Square	DF	Pr > ChiSq
Stepwise	5	7.4584	8	0.4881
Backward	8	3.7026	8	0.8829

We prefer the second model.

Result of Model Fitting

Now consider the second model.

parameter	estimate	std err	Chi-Square	Pr > ChiSq
intercept	−1.6046	0.7457	4.6303	0.0314
age	0.0224	0.0228	0.9678	0.3252
ndrugtx	−0.3846	0.1564	6.0513	0.0139
ivhx 2	−0.6364	0.2945	4.6703	0.0307
ivhx 3	−0.6959	0.2568	7.3435	0.0067
race	0.6662	0.2618	6.4745	0.0109
treat	0.4636	0.2023	5.2533	0.0219
site	0.5161	0.2514	4.2158	0.0400
age*ndrugtx	0.00888	0.00415	4.5747	0.0324
race*site	−1.5193	0.5266	8.3248	0.0039

Some Interpretation

The SAS output contains the following results for odds ratio.

effect	estimate	95% CI	
ivhx 2 vs 1	0.529	0.297	0.942
ivhx 3 vs 1	0.499	0.301	0.825
treat	1.590	1.069	2.363

The odds ratio for treat is $1.590 = e^{0.4636}$. The odds of remaining drug free increases 59% if we change treatment from short to long.

A 95% confidence interval for the odds ratio for treat is $\exp(0.4636 \pm 1.96 \times 0.2023) = (1.069, 2.363)$.

The odds of remaining drug free increases 2.3% if age increases 1 year. ($e^{0.0224} = 1.023$), and a 95% confidence interval is $\exp(0.0224 \pm 1.96 \times 0.0228) = (0.978, 1.069)$.

Estimated Probability

Suppose a 30-year-old white man recently use drugs and the number of previous drug treatment is 10. His Beck depression score is 20. He received the long treatment at site A. The logit of remaining drug free is

$$\begin{aligned}\text{logit}(\hat{\pi}) = &-1.6046 + (30)(0.0224) - (10)(0.3846) \\ &- 0.6959 + 0.4636 + (30)(10)(0.00888) = -2.3469\end{aligned}$$

and the estimated probability is

$$\hat{\pi} = \frac{\exp(-2.3469)}{1 + \exp(-2.3469)} = 0.087$$

A 95% confidence interval for this probability is $(0.046, 0.160)$.