STAT 426

Author: Wenxiao Yang

Institute: Department of Mathematics, University of Illinois at Urbana-Champaign

All models are wrong, but some are useful.

Contents

Chapter 1 Basic of Categorical Data

1.1 Variable Measurement

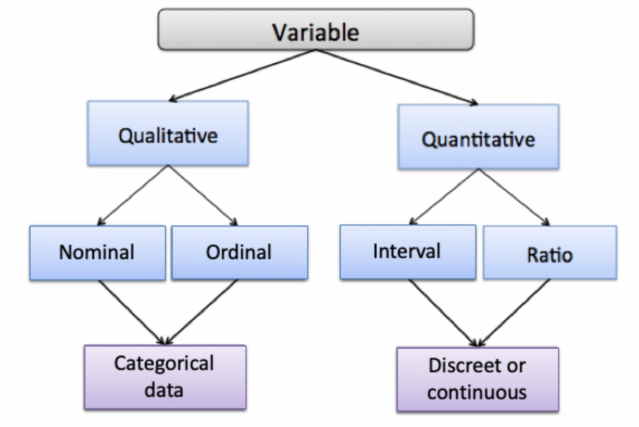

Figure 1.1: Variable Type

a) Nominal: Categories do not have a natural order. Ex. blood type, gender.

b) Ordinal: Categories have a natural order. Ex. low/middle/high education level

c) Interval: There is a numerical distance (difference between two different values is meaningful) between any two values. Ex. blood pressure level, 100 blood pressure doesn't mean the double degree of 50 pressure.

d) Ratio: An interval variable where ratios are valid (presence of absolute zero, i.e. zero is meaningful). Ex. weight, 4g is double degree of 2g, distance run by an athlete.

Levels of measurements

A variable's level of measurement determines the statistical methods to be used for its analysis.

Variables hierarchy: Ratio $>$ Interval $>$ Ordinal $>$ Nominal

Statistical methods applied to variables at a lower level can be used with variables at a higher level, but the contrary is not true.

1.2 Statistical Inference for Categorical Data

There is a distribution $F(\beta)$ with p.d.f. (p.m.f.) $f(x \mid \beta)$, where β a generic unknown parameter and $\hat{\beta}$ the parameter estimate.

1.2.1 Maximum likelihood Estimation (MLE)

Given a set of observations $\vec{x} = (x_1, ..., x_n)$, the likelihood function of these observations with parameter β is $l(\vec{x} \mid \beta)$. We want to find parameter $\hat{\beta}$ that maximizes the likelihood function,

$$\hat{\beta} = \arg\max_{\beta} l(\vec{x} \mid \beta)$$

which is also equivalent to maximizing the logarithm of the likelihood function $L(\vec{x} \mid \beta) = \log(l(\vec{x} \mid \beta))$,

$$\hat{\beta} = \arg\max_{\beta} L(\vec{x} \mid \beta)$$

Definition 1.1 (score function)

The score function is

$$u(\beta, \vec{x}) = \nabla_\beta L(\vec{x} \mid \beta) = \frac{\nabla_\beta l(\vec{x} \mid \beta)}{l(\vec{x} \mid \beta)}$$

♣

Lemma 1.1 (mean of score function)

The mean of score function is 0,

$$\mathbb{E}_{\vec{x}} u(\beta, \vec{x}) = 0$$

♡

Proof 1.1

$$\begin{aligned}
\mathbb{E}_{\vec{x}} u(\beta, \vec{x}) &= \int_{\vec{x}} l(\vec{x} \mid \beta) \frac{\nabla_\beta l(\vec{x} \mid \beta)}{l(\vec{x} \mid \beta)} d\vec{x} \\
&= \int_{\vec{x}} \nabla_\beta l(\vec{x} \mid \beta) d\vec{x} \\
&= \nabla_\beta \left(\int_{\vec{x}} l(\vec{x} \mid \beta) d\vec{x} \right) \\
&= \nabla_\beta 1 = 0
\end{aligned}$$

Lemma 1.2 (variance of score function)

The variance of the score function is

$$\mathrm{Var}_{\vec{x}}(u(\beta, \vec{x})) = \mathbb{E}_{\vec{x}} \left(u(\beta, \vec{x}) u(\beta, \vec{x})^T \right)$$

♡

Proof 1.2

Prove by the zero mean.

Definition 1.2 (Fisher information)

The (Fisher) information is

$$\iota(\beta) = -\mathbb{E}_{\vec{x}}\left[\nabla^2_\beta L(\vec{x} \mid \beta)\right]$$

♣

Lemma 1.3

The Fisher information is equal to the variance of score function.

$$\text{Var}_{\vec{x}}(u(\beta, \vec{x})) = \mathbb{E}_{\vec{x}}\left(u(\beta, \vec{x})u(\beta, \vec{x})^T\right) = -\mathbb{E}_{\vec{x}}\left[\nabla^2_\beta L(\vec{x} \mid \beta)\right] = \iota(\beta)$$

♡

Proof 1.3

$$\mathbb{E}_{\vec{x}}\left[\nabla^2_\beta L(\vec{x} \mid \beta)\right] = \mathbb{E}_{\vec{x}}\left(\frac{\partial \frac{\nabla_\beta l(\vec{x}|\beta)}{l(\vec{x}|\beta)}}{\partial \beta}\right) = \mathbb{E}_{\vec{x}}\left(\frac{\nabla^2_\beta l(\vec{x} \mid \beta)}{l(\vec{x} \mid \beta)} - \frac{\nabla_\beta l(\vec{x} \mid \beta)\nabla_\beta l(\vec{x} \mid \beta)^T}{(l(\vec{x} \mid \beta))^2}\right)$$

where $\mathbb{E}_{\vec{x}}\left(\frac{\nabla^2_\beta l(\vec{x}|\beta)}{l(\vec{x}|\beta)}\right) = \int_{\vec{x}} l(\vec{x} \mid \beta)\frac{\nabla^2_\beta l(\vec{x}|\beta)}{l(\vec{x}|\beta)}d\vec{x} = \int_{\vec{x}} \nabla^2_\beta l(\vec{x} \mid \beta)d\vec{x} = \nabla^2_\beta \int_{\vec{x}} l(\vec{x} \mid \beta)d\vec{x} = \nabla^2_\beta 1 = 0$

Hence,

$$\mathbb{E}_{\vec{x}}\left[\nabla^2_\beta L(\vec{x} \mid \beta)\right] = -\mathbb{E}_{\vec{x}}\left(\frac{\nabla_\beta l(\vec{x} \mid \beta)\nabla_\beta l(\vec{x} \mid \beta)^T}{(l(\vec{x} \mid \beta))^2}\right) = -\mathbb{E}_{\vec{x}}\left(u(\beta, \vec{x})u(\beta, \vec{x})^T\right)$$

Proposition 1.1

When the sample x is made up of i.i.d. observations, the covariance matrix of the maximum likelihood estimator $\hat{\beta}$ is approximately equal to the inverse of the information matrix.

$$\text{Cov}(\hat{\beta}) \approx (\iota(\beta))^{-1}$$

♠

Hence, the covariance matrix can be estimated as $(\iota(\hat{\beta}))^{-1}$. Similarly, *SE* is estimated by $\sqrt{(\iota(\hat{\beta}))^{-1}}$.

1.2.2 Likelihood Inference (Wald, Likelihood-Ratio, Score)

We want to test

$$H_0 : \beta = \beta_0 \qquad H_\alpha : \beta \neq \beta_0$$

or form a confidence interval (CI) for β.

Definition 1.3 (Wald Test)

The Wald statistic:

$$z_W = \frac{\hat{\beta} - \beta_0}{SE} = \frac{\hat{\beta} - \beta_0}{\sqrt{(\iota(\hat{\beta}))^{-1}}}$$

where $SE = \sqrt{(\iota(\hat{\beta}))^{-1}}$.

Usually, as $n \to \infty$, $z_W \xrightarrow{d} N(0, 1)$ under $H_0 : \beta = \beta_0$.

(1) We reject the H_0 if $|z_W| \geq z_{\frac{\alpha}{2}}$ for a two-sided level α test.

(2) The $(1-\alpha)100\%$ Wald (confidence) interval is

$$\{\beta_0 : |z_W| = \frac{|\hat{\beta} - \beta_0|}{SE} < z_{\frac{\alpha}{2}}\} = (\hat{\beta} - z_{\frac{\alpha}{2}} SE, \hat{\beta} + z_{\frac{\alpha}{2}} SE)$$

(3) The Wald test also has a chi-squared form, using

$$z_W^2 = \frac{(\hat{\beta} - \beta_0)^2}{(\iota(\hat{\beta}))^{-1}} \sim \chi_1^2 \quad \text{(under } H_0\text{)}$$

♣

Definition 1.4 (Likelihood Ratio Test)

Let

$$\Lambda = \frac{l(\vec{x} \mid \beta_0)}{l(\vec{x} \mid \hat{\beta})}$$

where $l(\vec{x} \mid \hat{\beta}) = \max_\beta l(\vec{x} \mid \beta)$, so the ratio $\Lambda \in [0, 1]$.

The **likelihood-ratio test (LRT) chi-squared statistic**:

$$-2 \ln \Lambda = -2 \left(L(\beta_0) - L(\hat{\beta}) \right)$$

It has an approximate χ_1^2 distribution under $H_0 : \beta = \beta_0$, and otherwise tends to be larger.

(1) Thus, reject H_0 if

$$-2 \ln \Lambda \geq \chi_1^2(\alpha)$$

(2) The $(1-\alpha)100\%$ likelihood-ratio (confidence) interval is

$$\{\beta_0 : -2 \ln \Lambda = -2 \left(L(\beta_0) - L(\hat{\beta}) \right) < \chi_1^2(\alpha)\}$$

Unlike Wald, this interval is not degenerate. (i.e., For general case, the interval does not have an explicit form.)

♣

Definition 1.5 (Score Test)

The **score statistic**:

$$z_S = \frac{u(\beta_0)}{\sqrt{\iota(\beta_0)}}$$

As $n \to \infty$, $z_S \xrightarrow{d} N(0, 1)$ under $H_0 : \beta = \beta_0$. Otherwise, it tends to be further from zero.

(1) Thus, reject H_0 if $|z_S| \geq z_{\frac{\alpha}{2}}$ for a two-sided level α test.

(2) The $(1-\alpha)100\%$ score (confidence) interval is

$$\{\beta_0 : |z_S| = \frac{|u(\beta_0)|}{\sqrt{\iota(\beta_0)}} < z_{\frac{\alpha}{2}}\}$$

Unlike Wald, it is not degenerate for some distributions.

(3) There is also a chi-squared form:

$$z_S^2 = \frac{u(\beta_0)^2}{\iota(\beta_0)} \sim \chi_1^2 \quad \text{(under } H_0\text{)}$$

♣

We can also use P-value to measure the probability of the statistic is more extreme under the H_0. We can reject H_0 if the P-value is $\leq \alpha$.

All three kinds tend to be "asymptotically equivalent" as $n \to \infty$. For smaller n, the likelihood-ratio and score methods are preferred.

Chapter 2 Association in Contingency Tables

2.1 Association in Two-Way Contingency Tables

Consider joint observations of two categorical variables: X with I categories, Y with J categories.

We can summarize data in an $I \times J$ **contingency table**:

		Y		
		1	$\cdots$	J
X	1			
	$\vdots$			
	I			

Each **cell** contains a count n_{ij}.

2.1.1 Distribution

If both X and Y are random, let

$$\pi_{ij} = P(X \text{ in row } i, Y \text{ in col } j)$$

be the **joint** distribution of X and Y.

The **marginal** distribution of X is defined by

$$\pi_{i+} = P(X \text{ in row } i)$$

and similarly for Y:

$$\pi_{+j} = P(Y \text{ in col } j)$$

The **conditional** distribution of Y given that X is in row i is defined by

$$\pi_{j|i} = P(Y \text{ in col } j \mid X \text{ in row } i) = \frac{\pi_{ij}}{\pi_{i+}}$$

2.1.2 Descriptive Statistics

Let $n_{ij} =$ *count in row* i *and col* j and $n = \sum_i \sum_j n_{ij}$.

The **margins** of the table:

$$n_{i+} = \sum_j n_{ij}, \quad n_{+j} = \sum_i n_{ij}$$

Natural Estimation

1. Natural estimate of π_{ij}: $\hat{\pi}_{ij} = \frac{n_{ij}}{n}$
2. Similarly marginals: $\hat{\pi}_{i+} = \sum_j \hat{\pi}_{ij} = \frac{n_{i+}}{n}$; $\hat{\pi}_{+j} = \sum_i p_{ij} = \frac{n_{+j}}{n}$
3. And conditionals: $\hat{\pi}_{j|i} = \frac{\hat{\pi}_{ij}}{\hat{\pi}_{i+}} = \frac{n_{ij}}{n_{i+}}$

2.1.3 Sampling Models (Examples)

Possible joint distributions for counts in $I \times J$ table:

1. Poisson (random total): Y_{ij} = count in cell (i, j),

$$Y_{ij} \sim \text{Poisson}(\mu_{ij})$$

and the Y_{ij}s are independent.

2. Multinomial (fixed total n): N_{ij} = count in cell (i, j),

$$\{N_{ij}\} \sim \text{multinomial}(n, \{\pi_{ij}\})$$

3. Independent Multinomial: Assume n_{i+} (row totals n_i) are fixed,

$$\left.\begin{array}{c} \{N_{1j}\}_{j=1}^J \sim \text{multinomial}(n_1, \{\pi_{j|1}\}_{j=1}^J) \\ \vdots \\ \{N_{Ij}\}_{j=1}^J \sim \text{multinomial}(n_I, \{\pi_{j|I}\}_{j=1}^J) \end{array}\right\}$$

(When $J = 2$, this is independent binomial sampling, for which we may just write π_i for $\{\pi_{1|i}, \pi_{2|i}\}$.)

2.1.4 Independent / Homogeneity

Definition 2.1 (independent)

If both X and Y are random, they are **independent** if

$$\pi_{ij} = \pi_{i+}\pi_{+j}, \ \forall i, j$$

which implies $\pi_{j|i} = \frac{\pi_{i+}\pi_{+j}}{\pi_{i+}} = \pi_{+j}, \forall i, j$. That is, $\pi_{j|i}$ doesn't depend on i and is the same as the marginal distribution of Y. (Intuitively, knowing X tells nothing about Y.)

♣

Definition 2.2 (homogeneity)

Even if X is not really random, the condition that $\pi_{j|i} = \pi_{+j}, \forall i, j$ is called **homogeneity**. This might still be relevant in a situation where X is deliberately chosen and Y is observed as a response.

♣

2.1.5 Measuring Inhomogeneity

Homogeneity is the condition $\pi_1 = \pi_2$. We can measure inhomogeneity by three different measures:

n_{11}	n_{12}
n_{21}	n_{22}

Y_1	$n_1 - Y_1$
Y_2	$n_2 - Y_2$

where $Y_i \sim$ indep. binomial(n_i, π_i). This regards row totals as fixed.

1. **difference of proportions**:

$$\pi_1 - \pi_2$$

The estimation is

$$\hat{\pi}_1 - \hat{\pi}_2 = \frac{y_1}{n_1} - \frac{y_2}{n_2}$$

The approx $(1-\alpha)100\%$ confidence interval is:

$$\hat{\pi}_1 - \hat{\pi}_2 \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{\pi}_1(1-\hat{\pi}_1)}{n_1} + \frac{\hat{\pi}_2(1-\hat{\pi}_2)}{n_2}}$$

(Problematic if π_1 and π_2 are near 0 or 1.)

2. **relative risk**:

$$RR = \frac{\pi_1}{\pi_2}$$

The estimation is

$$r = \frac{\hat{\pi}_1}{\hat{\pi}_2} = \frac{y_1/n_1}{y_2/n_2}$$

The approx $(1-\alpha)100\%$ confidence interval of $\ln RR$ is:

$$\ln r \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1-\hat{\pi}_1}{y_1} + \frac{1-\hat{\pi}_2}{y_2}}$$

3. **odds ratio**:

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$$

When $\theta = 1$, we can say there is no association.

The **odds** for a probability π is $\Omega = \frac{\pi}{1-\pi}$. Note $\pi = \frac{\Omega}{1+\Omega}$.

(In the multinomial model: $\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}$ ("cross-product ratio"); in Poisson model: $\theta = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}$)

The usual (unrestricted) estimates

$$\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$$

The approx $(1-\alpha)100\%$ confidence interval for $\ln\theta$ is

$$\ln\hat{\theta} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

Useful properties of odds ratio:

(1) Interchanging rows (or cols) changes θ to $\frac{1}{\theta}$.

(2) Interchanging X and Y doesn't change θ.

(3) Multiplying a row (or col) by a factor doesn't change $\hat{\theta}$.

(4) Relationship to relative risk: $\theta = RR \cdot \frac{1-\pi_2}{1-\pi_1}$. ($\theta$ and RR are similar if both π_1 and π_2 are small.)

2.1.6 Delta Method

It is easy to obtain approximate CI for a mean based on a sample mean by using the Central Limit Theorem and a consistent estimate of standard error.

But the log Odds Ratio and log Relative Risk are transformed means. How were their CI's derived? And why take logs?

Suppose a statistic T_n and parameter θ such that

$$\sqrt{n}\left(T_n - \theta\right) \xrightarrow[n\to\infty]{d} N\left(0, \sigma^2\right)$$

(e.g. T_n might be a sample mean from a sample of size n with population mean θ and variance σ^2)

We want a CI for $g(\theta)$, for some smooth g.

The Taylor expand at T_n is

$$g(\theta) \approx g(T_n) + g'(T_n)(\theta - T_n)$$

So,

$$\sqrt{n}(g(T_n) - g(\theta)) \approx g'(T_n)\sqrt{n}(T_n - \theta) \xrightarrow[n\to\infty]{d} N\left(0, (g'(T_n))^2\sigma^2\right)$$

(This is useful only if $g'(T_n) \neq 0$) Hence, when n is large

$$\sqrt{n}\frac{g(T_n) - g(\theta)}{|g'(T_n)|\sigma} \quad \dot{\sim} \quad N(0,1)$$

which suggests this approximate CI for $g(\theta)$:

$$g(T_n) \pm z_{\frac{\alpha}{2}} \frac{|g'(T_n)|\sigma}{\sqrt{n}}$$

2.1.7 Testing Independence by Residuals: X^2 Test (Pearson)

Let $\mu_{ij} = \mathbb{E}(N_{ij}) = n\pi_{ij}$. Under $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, $\forall i, j$

$$\mu_{ij} = n\pi_{ij} = n\pi_{i+}\pi_{+j}$$

Under H_0, can show the MLEs are

$$\hat{\mu}_{ij} = n\hat{\pi}_{i+}\hat{\pi}_{+j} = n\left(\frac{n_{i+}}{n}\right)\left(\frac{n_{+j}}{n}\right) = \frac{n_{i+}n_{+j}}{n}$$

(assuming no empty rows or cols)

Residuals:

1. Raw: $n_{ij} - \hat{\mu}_{ij}$
2. Pearson: $e_{ij} = \frac{n_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$. $X^2 = \sum_i \sum_j e_{ij}^2$.
3. Standardized: $r_{ij} = \frac{n_{ij}-\hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1-\hat{\pi}_{i+})(1-\hat{\pi}_{+j})}}$

Usage: Look for Pearson or standardized residuals with absolute value *exceeding 2 or 3*. These suggest the reason for significant dependence.

Remark: Under independence, both Pearson and standardized residuals are asymp. normal, but only standardized has asymp. variance equal to 1.

Definition 2.3 (X^2 Test: Pearson χ^2 Test (Score Test))

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \overset{\cdot}{\underset{H_0}{\sim}} \chi^2_{(I-1)(J-1)}$$

Note:

$$\begin{aligned}(I-1)(J-1) &= (IJ-1) - ((I-1)+(J-1)) \\ &= \text{ total \# params. } - \# \text{ params. under } H_0\end{aligned}$$

Reject H_0 if

$$X^2 > \chi^2_{(I-1)(J-1)}(\alpha)$$

(or use P-value)

♣

Example 2.1 Testing independence is equivalent to testing homogeneity in the indep. binomial model:

$$H_0 : \pi_1 = \pi_2$$

Can show

$$X^2 = z^2$$

where

$$z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}(1-\hat{\pi})\,(1/n_1 + 1/n_2)}} \quad \hat{\pi} = \frac{y_1 + y_2}{n_1 + n_2}$$

2.1.8 Testing Independence: G^2 Test (Likelihood Ratio)

Definition 2.4 (G^2 Test: Likelihood Ratio χ^2 Test)

$$G^2 = 2\sum_{ij} n_{ij} \ln \frac{n_{ij}}{\hat{\mu}_{ij}} \overset{\cdot}{\underset{H_0}{\sim}} \chi^2_{(I-1)(J-1)}$$

Reject H_0 if

$$G^2 > \chi^2_{(I-1)(J-1)}(\alpha)$$

(or use P-value)

(Convention: $0 \ln 0 = 0$)

♣

Comparison:

1. X^2 and G^2 are asymptotically equivalent under H_0
2. The X^2 tends to be better.

Remark: The X^2 and G^2 tests are not necessarily compatible with the Wald CIs. For example,

$$\text{reject } H_0 \not\Leftrightarrow \text{odds ratio } \theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = 1 \text{ not in Wald CI}$$

2.1.9 Testing Independence: Fisher's Exact Test

When cell counts are small, the X^2 and G^2 independence tests are not recommended: The χ^2 approximations are poor. In this section we introduce a *Fisher's Exact Test*.

Consider a 2×2 table with row and col totals fixed:

		Y		
X		N_{11}	N_{12}	n_{1+}
		N_{21}	N_{22}	n_{2+}
		n_{+1}	n_{+2}	n

Note: Any cell count, say N_{11}, determines the whole table.

Can show that, under H_0 : independence, N_{11} is (conditionally) hypergeometric:

$$\mathrm{P}_{H_0}(N_{11} = t) = \frac{\binom{n_{1+}}{t}\binom{n_{2+}}{n_{+1}-t}}{\binom{n}{n_{+1}}}$$

In terms of odds ratio $\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$, independence is

$$H_0 : \theta = 1$$

Possible alternatives:

$$H_\alpha : \theta > 1 \quad \Rightarrow \quad N_{11} \text{ tends larger}$$

$$H_\alpha : \theta < 1 \quad \Rightarrow \quad N_{11} \text{ tends smaller}$$

$$H_\alpha : \theta \neq 1 \quad \Rightarrow \quad N_{11} \text{ tends larger or smaller}$$

For $H_\alpha : \theta > 1$, the (one-sided) p-value is $\mathrm{P}_{H_0}(N_{11} \geq t_0)$, where $t_0 = n_{11}$ is the observed value of N_{11}.

Remarks: Could use mid p-values instead; Implemented in R function fisher.test(); Can be extended to $I \times J$ tables (with some computational difficulty).

2.2 Conditional Association in Three-Way Tables

Add a third categorical variable Z.

Example 2.2 Is a drug more effective at curing a disease among younger patients than among older? X = drug or placebo; Y = disease cured or not; Z = age group (young, old).

2.2.1 Conditional Association

Z may be called a **stratification variable**. We are interested in the distribution of (X, Y) *conditional* on Z.

Definition 2.5 (partial table)

Each Z category defines a **partial table** for X and Y. ♣

Example 2.3 When $Z = 1, 2$ and X, Y are binary ($2 \times 2 \times 2$ table):

$$Z = 1: \quad \begin{array}{c|cc} X \,\backslash\, Y & & \\ & n_{111} & n_{121} \\ & n_{211} & n_{221} \end{array} \qquad Z = 2: \quad \begin{array}{c|cc} X \,\backslash\, Y & & \\ & n_{112} & n_{122} \\ & n_{212} & n_{222} \end{array}$$

These represent **conditional associations**.

Definition 2.6 (marginal table)

The **marginal table** sums the partial tables: ♣

$$\begin{array}{c|cc} X \,\backslash\, Y & & \\ & n_{11+} & n_{12+} \\ & n_{21+} & n_{22+} \end{array}$$

This represents the **marginal association** (ignoring Z).

In general, let $\mu_{ijk} =$ *expected count in row i, col j, table k.*

The **conditional odds ratios**,

$$\theta_{XY(k)} = \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}$$

which are estimated by

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

The **marginal odds ratio**

$$\theta_{XY} = \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}}$$

is estimated from the marginal table.

2.2.2 Simpson's Paradox

Some counter-intuitive but possible situations:

1. There are conditional associations ($\theta_{XY(k)} \neq 1$) but no marginal association ($\theta_{XY} = 1$)
2. There is a marginal association ($\theta_{XY} \neq 1$) but no conditional associations ($\theta_{XY(k)} = 1$)
3. **Simpson's paradox**: The conditional associations are in the opposite direction from the marginal, e.g. $\theta_{XY(k)} > 1, \theta_{XY} < 1$

	Full Population, N = 52			Men (M), N = 20			Women (¬M), N = 32		
	Success (S)	Failure (¬S)	Success Rate	Success	Failure	Success Rate	Success	Failure	Success Rate
Treatment (T)	20	20	50%	8	5	≈ 61%	12	15	≈ 44%
Control (¬T)	6	6	50%	4	3	≈ 57%	2	3	≈ 40%

TABLE 1: Simpson's Paradox: the type of association at the population level (positive, negative, independent) changes at the level of subpopulations. Numbers taken from Simpson's original example (1951).

Figure 2.1: Simpson's paradox

2.2.3 Conditional Independence, Marginal Independence

Definition 2.7 (conditionally independent given Z, marginal independent)

We also call X and Y are **conditionally independent given** $Z = k$ if $\theta_{XY(k)} = 1$. If this is true for all k, X and Y are **conditionally independent given** Z. Not the same to "X and Y are **marginal independent** if $\theta_{XY} = 1$".

♣

Proposition 2.1

For multinomial sampling, can show that conditional independence is

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}, \quad \forall i, j, k$$

♠

2.2.4 Homogeneous Association

Definition 2.8

Let Z have K categories. X and Y have **homogeneous association** over Z if

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

(Conditional independence is a special case.)

Chapter 3 Generalized Linear Models

3.1 Introduction

A linear model $Y = \alpha + \sum_{i=1}^{p} \beta_i x_i + \varepsilon$ is usually not appropriate if Y is binary or a count.

3.1.1 Definition

We seek to model independent observations $Y_1, ..., Y_n$ of a **response variable**, in terms of corresponding vectors $\vec{x}_i = (x_{i1}, ..., x_{ip}), i = 1, ..., n$ of values of p **explanatory variables**.

(1) **Random component:** density of Y_i from a **natural exponential family**

$$f(y_i; \theta_i) = a(\theta_i)b(y_i)\exp(y_i Q(\theta_i))$$

where $Q(\theta_i)$ is the **natural parameter**.

(**Fact:** Since Y_i is from a natural exponential family, its distribution is completely determined by its mean μ_i. In particular, $\text{Var}(Y_i)$ is a function of μ_i.)

(2) **Systematic component:** the **linear predictor**

$$\eta_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

with parameters $\alpha, \beta_1, ...\beta_p$ (**coefficients**)

Y_i will depend on $\vec{x}_i$ only through η_i.

(3) **Link function:** monotonic, differentiable g such that $g(\mu_i) = \eta_i$, that is

$$g(\mu_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \quad \text{where } \mu_i = \mathbb{E}(Y_i)$$

(Note: Ordinary linear models use the identity link: $g(\mu) = \mu$, which means $\mu_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$.)

Definition 3.1 (Canonical Link)

The **canonical link** satisfies

$$Q(\theta_i) = g(\mu_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

♣

Let F be a continuous and invertible c.d.f. on the real line. A reasonable link might be

$$g(\pi) = F^{-1}(\pi)$$

since it transforms interval $(0, 1)$ to the whole real line.

Definition 3.2 (Probit Regression)

Using the c.d.f. Φ for a standard normal is called **probit regression**. ♣

3.1.2 Fitting GLMs

Usually by maximum likelihood: find

$$\hat{\alpha}, \hat{\beta}_1, \ldots, \hat{\beta}_p$$

maximizing

$$\prod_{i=1}^{n} f\left(y_i; \theta_i\right)$$

Explicit solutions exist only in special cases, so need numerical methods: e.g. Newton-Raphson, Fisher Scoring.

3.2 Binary and Binomial Responses

3.2.1 Binary Regression

Example 3.1 (Binary Regression)

$$Y_i \sim \text{Bernoulli}\left(\pi_i\right) \quad \left(\theta_i = \pi_i\right)$$

$$\begin{aligned} f\left(y_i; \pi_i\right) &= \begin{cases} 1 - \pi_i & y_i = 0 \\ \pi_i & y_i = 1 \end{cases} \\ &= \pi_i^{y_i} \left(1 - \pi_i\right)^{1 - y_i} \\ &= \left(1 - \pi_i\right) \left(\frac{\pi_i}{1 - \pi_i}\right)^{y_i} \\ &= \left(1 - \pi_i\right) \exp\left(y_i \ln\left(\frac{\pi_i}{1 - \pi_i}\right)\right) \end{aligned}$$

So $a(\pi) = 1 - \pi, b(y) = 1$, and

$$Q(\pi) = \ln\left(\frac{\pi}{1 - \pi}\right) = \text{logit}(\pi)$$

The natural parameter is the **log odds**.

Note: $\mu_i = E\left(Y_i\right) = \pi_i$. Hence, we can write $\pi_i(\vec{x}_i)$ as a response to

- **Identity Link:**

$$\pi(\vec{x}_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

- **Log Link:**

$$\ln(\pi(\vec{x}_i)) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

- **Canonical link: (logistic regression)**

$$\text{logit}(\pi(\vec{x}_i)) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

Specifically, when $p = 1$,

$$\begin{aligned} \text{logit}(\pi(x)) = \alpha + \beta x \quad &\Leftrightarrow \quad \text{odds}(\pi(x)) = e^{\alpha+\beta x} \\ &\Leftrightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}} \end{aligned}$$

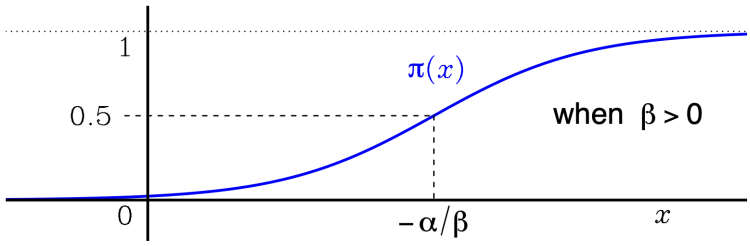

Figure 3.1: $\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$

3.2.2 Grouped Data: Binomial Response

If several observations have the same $\vec{x}$ ("replications"), then they have the same $\pi(\vec{x})$.

Summing binary (0/1) observations with the same $\vec{x}$ gives **grouped** data:

$$Y_i \sim \text{binomial}\left(n_i, \pi\left(\vec{x}_i\right)\right)$$

where "i" now refers to the i^{th} group (of n_i binary obs.).

Note: Both Y_i and n_i (or $n_i - Y_i$) must be included in the data.

Remarks:

1. Whether data are grouped or ungrouped, fitting with maximum likelihood gives the same results.
2. Technically, the binomial GLM should use $\bar{Y}_i = \frac{Y_i}{n_i}$ as the responses, and use an *exponential dispersion family form* for the density.

For 2×2 Tables

$x = 1$	Y_1	$n_1 - Y_1$
$x = 0$	Y_2	$n_2 - Y_2$

Note: Can regard as grouped data with two groups.

A binomial regression model (with $x = 0$ or 1) is equivalent to the independent binomial model:

$$\left.\begin{aligned} Y_1 &\sim \text{binomial}\left(n_1, \pi_1 = \pi(1)\right) \\ Y_2 &\sim \text{binomial}\left(n_2, \pi_2 = \pi(0)\right) \end{aligned}\right\} \text{ independent}$$

For logistic regression:

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

so the odds ratio is

$$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \exp\left(\text{logit}\left(\pi_1\right) - \text{logit}\left(\pi_2\right)\right)$$
$$= \exp(\alpha + \beta \cdot 1 - (\alpha + \beta \cdot 0)) = e^{\beta}$$

So β is the **log odds ratio**.

3.3 Count Responses

For binomial data, the maximum possible count is known (for each observation). What if there are no known maximum counts? Counts of independently-occurring incidents (without any maximum) are often modeled using the Poisson distribution.

3.3.1 Poisson Regression

Example 3.2 (Poisson Regression)

$$Y_i \sim \text{Poisson}\left(\mu_i\right) \quad (\theta_i = \mu_i)$$

Note: $\mu_i = E\left(Y_i\right) = \text{Var}(Y_i)$

$$f\left(y_i; \mu_i\right) = \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}$$
$$= e^{-\mu_i} \frac{1}{y_i!} \exp\left(y_i \ln \mu_i\right)$$

So $a(\mu) = e^{-\mu}, b(y) = \frac{1}{y!}$

$$Q(\mu) = \ln \mu$$

The natural parameter is the log-mean.

Canonical link:

$$\ln \mu(\vec{x}_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

which gives the **(Poisson) loglinear model**.

Specifically, when $p = 1$,

$$\ln \mu(x) = \alpha + \beta x$$
$$\Leftrightarrow \mu(x) = e^{\alpha + \beta x} = e^{\alpha}(e^{\beta})^x$$

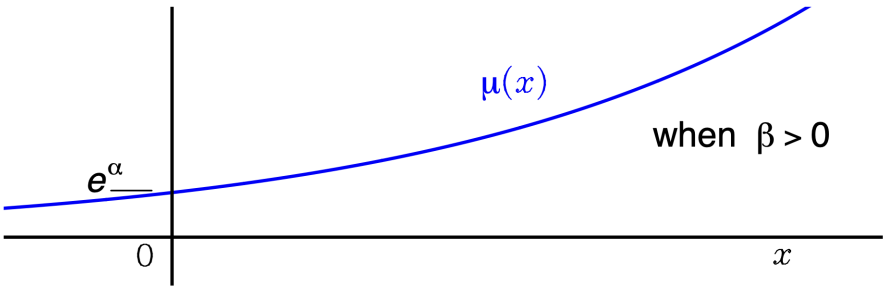

Figure 3.2: $\mu(x) = e^{\alpha}(e^{\beta})^{x}$

3.3.2 Rate Models

$E(Y_i) = \mu_i$ is sometimes expected to be proportional to another observed variable $t_i > 0$:

$$\mu_i = \lambda_i t_i$$

e.g.

$$Y_i = \text{ cases of rare disease in nation } i$$

$$t_i = \text{ national population (known)}$$

$$\lambda_i = \text{ disease } \textbf{rate} \text{ (unknown)}$$

(t could alternatively be a temporal or spatial extent)

Canonical link:

$$\begin{aligned} \ln \mu_i &= \ln \lambda_i + \ln t_i \\ &= \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \ln t_i \end{aligned}$$

where λ_i works as linear predictor, $\ln \lambda_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$

Note: $\ln t_i$ has no coefficient. We call $\ln t_i$ an **offset**.

For 2×2 Tables

	$x_2 = 1$	$x_2 = 0$
$x = 1$	Y_{11}	Y_{12}
$x = 0$	Y_{21}	Y_{22}

$$\{Y_{ij}\} \sim \text{ indep. Poisson } (\{\mu_{ij}\})$$

The full loglinear regression model can be parameterized as

$$\ln \mu_{ij} = \alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

(Can solve for $\alpha, \beta_1, \beta_2, \beta_3$ in terms of the μ_{ij}.)

Recall relation to multinomial:

$$\{Y_{ij}\} \mid \sum_{ij} Y_{ij} = n \sim \text{multinomial}\,(n, \{\pi_{ij}\})$$

$$\pi_{ij} = \frac{\mu_{ij}}{\mu_{11} + \mu_{12} + \mu_{21} + \mu_{22}}$$

Recall odds ratio:

$$\theta = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = \frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}$$

Can show $\theta = 1$ (i.e., no association) is equivalent to $\beta_3 = 0$ (i.e., no interaction term):

$$\ln \mu_{ij} = \alpha + \beta_1 x_1 + \beta_2 x_2$$

3.4 Coefficient and Model Inferences

Matrix Forms

We can write the **linear predictor** of GLM

$$\eta_i = \alpha + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}, \quad i = 1, ..., N$$

in vector form:

$$\boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta}$$

where $\boldsymbol{\eta} = [\eta_1, \cdots, \eta_N]^T$, $\boldsymbol{\beta} = [\alpha, \beta_1, \cdots, \beta_p]^T$, and the model matrix $\mathbf{X}$ has i^{th} row $[1, x_{i1}, \cdots, x_{ip}]$.

Let the MLE of $\boldsymbol{\beta}$ be

$$\hat{\boldsymbol{\beta}} = \left[\hat{\alpha}, \hat{\beta}_1, \cdots, \hat{\beta}_p\right]^T$$

3.4.1 Wald Inference

3.4.1.1 (Fisher) Information Matrix

Definition 3.3 ((Fisher) Information Matrix)

The (Fisher) information matrix for $\boldsymbol{\beta} \in \mathbb{R}^{p+1}$ is a $(p+1) \times (p+1)$ matrix

$$\boldsymbol{\mathcal{J}}$$

with element (h, j) being

$$\mathbb{E}\left(-\frac{\partial^2 L(\boldsymbol{\beta})}{\partial\beta_h \partial\beta_j}\right)$$

♣

For a GLM, the information matrix becomes

$$\boldsymbol{\mathcal{J}} = \mathbf{X}^T\mathbf{W}\mathbf{X}$$

where $\mathbf{W} = \text{diag}\,(w_1, \cdots, w_N)$ with

$$w_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \cdot \frac{1}{\text{var}(Y_i)}$$

Recall $\mu_i = \mathbb{E}(Y_i)$.

Example 3.3 Logistic Regression

$$\mu_i = E\,(Y_i) = n_i \pi_i \quad \text{var}\,(Y_i) = n_i \pi_i\,(1 - \pi_i)$$

$$\eta_i = \text{logit}\,(\pi_i) = \ln \pi_i - \ln\,(1 - \pi_i)$$

Then

$$\begin{aligned}\frac{\partial \mu_i}{\partial \eta_i} &= \frac{\partial \mu_i}{\partial \pi_i} \cdot \frac{\partial \pi_i}{\partial \eta_i} = n_i \cdot \left(\frac{\partial \eta_i}{\partial \pi_i}\right)^{-1} \\ &= n_i \left(\frac{1}{\pi_i} + \frac{1}{1 - \pi_i}\right)^{-1} = n_i \pi_i\,(1 - \pi_i)\end{aligned}$$

Thus

$$w_i = (n_i \pi_i\,(1 - \pi_i))^2 \cdot \frac{1}{n_i \pi_i\,(1 - \pi_i)} = n_i \pi_i\,(1 - \pi_i)$$

3.4.1.2 Wald Inference

Under regularity conditions, as $N \to \infty$, the distribution of $\hat{\boldsymbol{\beta}}$ is approximately multivariate normal with mean vector $\boldsymbol{\beta}$ and covariance matrix $\boldsymbol{\mathcal{J}}^{-1}$ (a proposition in MLE part):

$$\hat{\boldsymbol{\beta}} \quad \dot{\sim} \quad N\left(\boldsymbol{\beta}, \boldsymbol{\mathcal{J}}^{-1}\right)$$

So the asymptotic covariance of $\hat{\boldsymbol{\beta}}$ is

$$\boldsymbol{\mathcal{J}}^{-1} = \left(\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\right)^{-1}$$

which is estimated as

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1}$$

where $\hat{\boldsymbol{W}}$ is $\boldsymbol{W}$ estimated using $\hat{\boldsymbol{\beta}}$ for $\boldsymbol{\beta}$.

In particular, the element $\hat{\beta}_j$ of $\hat{\boldsymbol{\beta}}$ is asymptotically normal with asymptotic variance

$$\widehat{\text{var}}\left(\hat{\beta}_j\right) = (j+1) \text{ st diagonal element of } \widehat{\text{cov}}(\hat{\boldsymbol{\beta}})$$

The Wald z statistic for testing $H_0 : \beta_j = \beta_{j0}$ is

$$z_W = \frac{\hat{\beta}_j - \beta_{j0}}{SE\left(\hat{\beta}_j\right)} \underset{\text{H}_0}{\dot{\sim}} N(0,1)$$

where $SE\left(\hat{\beta}_j\right) = \sqrt{\widehat{\text{var}}\left(\hat{\beta}_j\right)}$.

Also Wald Cls:

$$\hat{\beta}_j \pm z_{\alpha/2} \cdot SE\left(\hat{\beta}_j\right)$$

3.4.2 Deviance and Likelihood-Ratio Test

3.4.2.1 Deviance and Goodness of Fit

Then it can be shown that $\boldsymbol{\mu} = \boldsymbol{y}$ maximizes L. It follows that

$$L(\boldsymbol{y}; \boldsymbol{y}) \geq L(\hat{\boldsymbol{\mu}}; \boldsymbol{y})$$

where $\hat{\boldsymbol{\mu}}$ is the MLE of $\boldsymbol{\mu}$ (when it exists) for the GLM. The unrestricted case, in which each observation has its own mean, is called the **saturated model**.

Definition 3.4 (Deviance)

The **deviance** of the GLM is

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2(L(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y}))$$

Note: $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ is the likelihood-ratio test (LRT) chi-squared statistic G^2 for

- H_0 : the GLM is correct
- H_a : the GLM is incorrect (but the saturated model is correct)

♣

The deviance is associated with degrees of freedom

$$\begin{aligned} \mathrm{df} &= \#\text{ means in sat. model } - \#\text{ params. in GLM} \\ &= N - (p+1) \quad \text{(usually)} \end{aligned}$$

For a 2×2 table, for the independent binomial model under homogeneity ($\pi_1 = \pi_2$), $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$ is G^2 for testing homogeneity. The notion for model M's deviance is $G^2(M)$, that is a model M's deviance is

$$G^2(M) = D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = -2(L(\hat{\boldsymbol{\mu}}; \boldsymbol{y}) - L(\boldsymbol{y}; \boldsymbol{y}))$$

3.4.2.2 Goodness of Fit Test / Likelihood-Ratio Test

Under certain asymptotic conditions,

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) \underset{\mathrm{H_0}}{\dot{\sim}} \chi^2_{\mathrm{df}}$$

and tends larger under H_α.

So reject correctness of the GLM if

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) > \chi^2_{\mathrm{df}}(\alpha)$$

(or use a P-value).

Warning: Chi-squared approximation can be poor. The chi-squared approximation (under H_0) is adequate if

all

$$\mu_i = n_i\pi_i \quad \text{and} \quad n_i - \mu_i = n_i\,(1-\pi_i)$$

are sufficiently large.

The chi-squared approximation is never valid for binary responses ($n_i = 1$, i.e. ungrouped data). Indeed, in that case, the deviance is completely useless for model checking.

Example 3.4 Poisson Case For a (Poisson) loglinear model, $L(\boldsymbol{\mu};\boldsymbol{y}) = \sum_i \ln\left(\frac{\mu_i^{y_i}}{y_i!}e^{-\mu_i}\right)$, we can show

$$\begin{aligned} D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) &= 2\sum_i \left(y_i \ln\frac{y_i}{\hat{\mu}_i} - y_i + \hat{\mu}_i\right) \\ &= 2\sum_i y_i \ln\frac{y_i}{\hat{\mu}_i} \end{aligned}$$

Remark:

- The chi-squared approximation (under H_0) is adequate if all μ_i are sufficiently large.
- These formulas also apply to loglinear rate models (with rate variable t_i), for which $\mu_i = \lambda_i t_i$, $\hat{\mu}_i = \hat{\lambda}_i t_i$, where $\hat{\lambda}_i$ is the MLE of rate λ_i.

Example 3.5 Binomial Case $Y_i \sim \text{binomial}(n_i, \pi_i)$, $L(\boldsymbol{\mu};\boldsymbol{y}) = \sum_i \ln\left(\frac{\mu_i}{n_i}\right)^{y_i}\left(1-\frac{\mu_i}{n_i}\right)^{n_i-y_i}$

$$D(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\sum_i y_i \ln\frac{y_i}{\hat{\mu}_i} + 2\sum_i (n_i - y_i)\ln\frac{n_i - y_i}{n_i - \hat{\mu}_i}$$

where $\hat{\mu}_i = n_i\hat{\pi}_i$. (Convention: $0\ln 0 = 0$)

Remark: If the data is $N \times 2$, this deviance is the same as the deviance for the Poisson model with $2N$ observations.

3.4.3 Nested Model Comparison

"Nested" means that one model is a subset of another.

Definition 3.5 (Nested Model)

Model M_0 is **nested** in Model M_1 if the parameters in Model M_0 are a subset of the parameters in Model M_1. E.g.

$$\begin{aligned} M_0: &\quad g(\mu_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_{p_0} x_{ip_0} \\ M_1: &\quad g(\mu_i) = \alpha + \beta_1 x_{i1} + \cdots + \beta_{p_1} x_{ip_1} \end{aligned}$$

where $p_0 < p_1$. That is, $\boldsymbol{\mu}$ is more restricted under M_0 than under M_1. ♣

Let $\hat{\boldsymbol{\mu}}_0$ be the MLE under M_0 and $\hat{\boldsymbol{\mu}}_1$ be the MLE under M_1.

For testing

$$H_0 : M_0 \text{ true} \quad H_a : M_1 \text{ true, but not } M_0$$

the LRT chi-squared statistic is

$$\begin{aligned} & -2\left(L\left(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y}\right) - L\left(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y}\right)\right) \\ =& -2\left(L\left(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y}\right) - L(\boldsymbol{y}; \boldsymbol{y})\right) - \left[-2\left(L\left(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y}\right) - L(\boldsymbol{y}; \boldsymbol{y})\right)\right] \\ =& D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0\right) - D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1\right) \end{aligned}$$

which is always non-negative.

If the chi-squared approximation is adequate, reject H_0 if

$$D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0\right) - D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1\right) > \chi^2_{\mathrm{df}}(\alpha)$$

where

$$\mathrm{df} = \text{effective \# params. in } M_1 - \text{ effective \# params. in } M_0$$

Remark: The chi-squared approximation is often adequate here even when it isn't adequate for the saturated model (provided M_1 is not too close to saturated).

Notation: For comparing null model M_0 to larger model M_1, denote the LRT chi-squared statistic as

$$G^2(M_0 \mid M_1) = D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_0\right) - D\left(\boldsymbol{y}; \hat{\boldsymbol{\mu}}_1\right) = -2\left(L\left(\hat{\boldsymbol{\mu}}_0; \boldsymbol{y}\right) - L\left(\hat{\boldsymbol{\mu}}_1; \boldsymbol{y}\right)\right)$$

Definition 3.6 (Profile Likelihood CIs)

Say we want a Cl for a parameter β in a model M_1.

Let

$$M_0\left(\beta_0\right) = \text{ same model, except } \beta \text{ is fixed at } \beta_0$$

Then the LRT tests

$$H_0 : \beta = \beta_0 \quad H_a : \beta \neq \beta_0$$

and produces a P-value (from chi-squared approximation).

Then

$$\{\beta_0 : P\text{-value } > \alpha\}$$

is an approx. $(1-\alpha)100\%$ confidence set (usually a Cl) for β.

This interval (based on the test inversion idea) is a **profile likelihood CI**.

♣

3.4.4 Residuals

As in linear regression, residuals provide a way to examine lack of fit: patterns of departure from the model, and "outliers." Recall $\hat{\mu}_i = \mathrm{MLE}$ of $E\left(Y_i\right)$. The raw residuals $y_i - \hat{\mu}_i$ have unequal variances, making it difficult to use them to examine lack of fit.

Definition 3.7 (Pearson Residuals)

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\nu\left(\hat{\mu}_i\right)}}$$

where $\nu(\mu) = \text{var}(Y)$.

When $\text{var}(Y) = \mu$, eg: Poisson

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

♣

Definition 3.8 (Deviance Residuals)

The deviance can be written in terms of the sum of contributions from each observation $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) = \sum_{i=1}^{N} d_i$, where $d_i = -2\left(L\left(\hat{\mu}_i; y_i\right) - L\left(y_i; y_i\right)\right)$ is non-negative.

The i^{th} **deviance residual** is

$$\text{sign}\left(y_i - \hat{\mu}_i\right) \cdot \sqrt{d_i}$$

♣

Problem: Neither Pearson nor deviance residuals are truly standardized. Their variances tend to be less than 1, and often unequal. Need a type of residual that (approximately) fixes these problems.

Definition 3.9 (Standardized Residuals)

$$r_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\nu\left(\hat{\mu}_i\right)\left(1 - \hat{h}_i\right)}} = \frac{e_i}{\sqrt{1 - \hat{h}_i}}$$

where $\hat{h}_i$ is the i^{th} **leverage**.

The **leverages** are the diagonal elements of the (estimated) **hat matrix**

$$\hat{\boldsymbol{H}}_{at} = \hat{\boldsymbol{W}}^{1/2} \boldsymbol{X} \left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \hat{\boldsymbol{W}}^{1/2}$$

where $\hat{\boldsymbol{W}}^{1/2}$ is the diagonal matrix of square roots of the diagonal elements of $\hat{\boldsymbol{W}}$.

♣

Remark: A leverage measures "potential influence" of its observation — how sensitively the fit depends on it.

3.4.5 Overdispersion

Recall: In our GLMs, $\text{var}(Y)$ is a function of $\mu = E(Y)$.

Definition 3.10 (Overdispersion)

Overdispersion is when $\text{var}(Y)$ for the data appears larger than the fitted μ predicts.

This is common in Poisson regression, and sometimes in binomial.

(**Underdispersion**, in which $\text{var}(Y)$ is smaller than predicted, is rare.)

Overdispersion is a type of lack of fit, and may cause a goodness of fit test to reject.

♣

Under the GLM, let $\nu^*(\mu) = \text{var}(Y)$, the **overdispersion** is when

$$\text{actual var}(Y) > \nu^*(\mu)$$

Possible causes of overdispersion:

1. heterogeneity among observations (variations in μ not captured by the model)
2. lurking variables (unused predictors, possibly unknown)
3. correlations among observations (e.g. clustering)

One remedy: quasi-likelihood

Use a **quasi-likelihood** having an additional **dispersion parameter** $\phi > 0$ that scales the variance:

$$\nu_\phi(\mu) = \phi\nu^*(\mu)$$

E.g., for Poisson, $\nu_\phi(\mu) = \phi\mu$. Then $\phi > 1$ represents overdispersion.

Under quasi-likelihood, the estimate of $\boldsymbol{\beta}$ (and thus $\boldsymbol{\mu}$) remains unchanged: $\hat{\boldsymbol{\beta}}$ is still the MLE from the original model.

Usually ϕ is estimated as

$$\hat{\phi} = \frac{X^2}{N - p'}, \quad \text{where } X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\nu^*(\hat{\mu}_i)}$$

and p' is the effective number of parameters in $\boldsymbol{\beta}$ (usually $p + 1$).

We can use the $\hat{\phi}$ to scale the GLM asymptotic covariance of $\hat{\boldsymbol{\beta}}$. In particular, the new (adjusted) $SE\left(\hat{\beta}_j\right)$ is just the GLM version multiplied by $\sqrt{\hat{\phi}}$. Then use the new standard errors for Wald-type inference.

Remark: Similar adjustments apply to likelihood-type and score-type inferences.

Chapter 4 Binary-Response Regression except Logistic Regression

4.1 Probit Model

4.1.1 Probit Link

The probit link is

$$g(\pi) = \Phi^{-1}(\pi)$$

where $\Phi =$ c.d.f. of $N(0,1)$. Using the probit link is **probit regression**.

4.1.2 Motivation: Latent Response

Suppose there is just a single explanatory variable x, and it is quantitative.

Consider ordinary simple linear regression with a response Y^* :

$$Y^* \sim N(\alpha + \beta x, 1)$$

assuming variance σ^2 is equal to 1.

Suppose Y^* is latent (unobserved), but we observe

$$Y = \begin{cases} 1 & \text{if} \quad Y^* > 0 \\ 0 & \text{if} \quad Y^* \leq 0 \end{cases}$$

Then

$$\begin{aligned} \pi(x) &= \mathrm{P}(Y = 1 \mid x) \\ &= \mathrm{P}\left(Y^* > 0 \mid x\right) \\ &= \mathrm{P}\left(Y^* - \alpha - \beta x > -\alpha - \beta x \mid x\right) \\ &= \mathrm{P}(Z > -\alpha - \beta x) \quad (Z \text{ standard normal}) \\ &= \mathrm{P}(Z < \alpha + \beta x) \quad \text{(symmetry)} \\ &= \Phi(\alpha + \beta x) \end{aligned}$$

So Y satisfies a probit regression.

4.1.3 Properties and Interpretation

Again consider one quantitative x variable:

$$\pi(x) = \Phi(\alpha + \beta x)$$

Then

$$\frac{d\pi(x)}{dx} = \beta\phi(\alpha + \beta x)$$

where $\phi =$ density of $N(0, 1)$.

If $\beta \neq 0$, slope is steepest at

$$x = -\frac{\alpha}{\beta} \quad (\text{where } \alpha + \beta x = 0)$$

Relation to latent response Y^* :

a unit increase in x implies an increase of β in $E(Y^*)$ (which is a decrease if $\beta < 0$)

This assumes $\text{var}(Y^*) = 1$. (Otherwise, the increase is in standard deviation units - later.)

Sign of β indicates direction of $X - Y^*$ relationship.

4.1.4 Model Fitting and Checking

Usually fit by maximum likelihood to get $\hat{\boldsymbol{\beta}}$.

Can use deviance (or Pearson X^2) to check fit.

Residuals and influence measures are also available. Fitted values are defined as usual.

Remark: Fisher scoring and Newton-Raphson algorithms are different for probit regression, and can cause software to report different standard errors.

4.1.5 Inference

Assume the general (binomial) form for the data:

$$Y_i \sim \text{ indep. binomial } (n_i, \pi(\boldsymbol{x}_i))$$

$$\Phi^{-1}(\pi(\boldsymbol{x}_i)) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

$$\boldsymbol{x}_i^T = i^{\text{th}} \text{ row of } \boldsymbol{X} \quad \boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$$

The estimated asymptotic covariance (obtained by inverting estimated information):

$$\widehat{\text{cov}}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^T\hat{\boldsymbol{W}}\boldsymbol{X}\right)^{-1}$$

with

$$\hat{\boldsymbol{W}} = \text{diag}\left(n_i \frac{(\phi(\hat{\eta}_i))^2}{\Phi(\hat{\eta}_i)(1 - \Phi(\hat{\eta}_i))}\right)$$

where

$$\hat{\eta}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$$

This can be used for Wald inference. Likelihood-ratio and score inference are also available.

4.1.6 Variance $\sigma^2 \neq 1$

Suppose we allow the latent variable to have an unknown variance σ^2, so that

$$Y^* \sim N\left(\tilde{\alpha} + \tilde{\beta}x, \sigma^2\right)$$

where $\sigma^2 > 0$ is arbitrary.

As before, observe

$$Y = \begin{cases} 1 & \text{if} \quad Y^* > 0 \\ 0 & \text{if} \quad Y^* \leq 0 \end{cases}$$

It turns out that $\tilde{\alpha}$, $\tilde{\beta}$, and σ^2 are not identifiable, so can't be estimated ...

$$\begin{aligned} \pi(x) &= \mathrm{P}(Y = 1 \mid x) = \mathrm{P}\left(Y^* > 0 \mid x\right) \\ &= \mathrm{P}\left(\frac{Y^* - \tilde{\alpha} - \tilde{\beta}x}{\sigma} > \frac{0 - \tilde{\alpha} - \tilde{\beta}x}{\sigma} \mid x\right) \\ &= \mathrm{P}(Z > -\tilde{\alpha}/\sigma - \tilde{\beta}x/\sigma) \quad (Z \text{ standard normal}) \\ &= \mathrm{P}(Z < \tilde{\alpha}/\sigma + \tilde{\beta}x/\sigma) \quad \text{(symmetry)} \\ &= \Phi(\underbrace{\tilde{\alpha}/\sigma}_{\alpha} + \underbrace{(\tilde{\beta}/\sigma)}_{\beta} x) \end{aligned}$$

Probit regression estimates α and β, not $\tilde{\alpha}$ and $\tilde{\beta}$.

Thus, when $\sigma^2 \neq 1$, α and β must be interpreted in standard deviation units.

Eg: Increasing x by 1

$$\text{increases } E\left(Y^*\right) \text{ by } \tilde{\beta} = \sigma\beta$$

i.e., by β standard deviations.

Again, $\tilde{\alpha}$, $\tilde{\beta}$, and σ^2 are not identified, so can't be estimated.

4.1.7 Symmetry Property

Logit and probit links have the symmetry property

$$g(\pi) = -g(1 - \pi)$$

This means both tails exhibit the same asymptotic behavior.

An implication: If

$$Y_i \sim \text{binomial}\,(n_i, \cdot)$$

follows a logistic (or probit) model with parameter $\boldsymbol{\beta}$, then

$$n_i - Y_i$$

follows a logistic (or probit) model with parameter $-\boldsymbol{\beta}$. This property is convenient, since the model is equivalent no matter which type of outcome is designated as the "success" and which the "failure." However, sometimes an asymmetric model fits the data better.

4.2 Complementary Log-Log Model

The complementary log-log link is

$$g(\pi) = \ln(-\ln(1-\pi))$$

The corresponding "response curve":

$$\pi(\eta) = g^{-1}(\eta) = 1 - e^{-e^{\eta}}$$

Property: When $\eta = \alpha + \beta x$,

$$(1-\pi(x_2)) = (1-\pi(x_1))^{e^{\beta(x_2-x_1)}}$$

Remarks:

- This link does not have the symmetry property - the model does depend on which type of outcome is the "success."
- There is also a log-log link, which is obtained by interchanging the roles of "success" and "failure."

Chapter 5 Logistic Regression

5.1 Parameter Interpretation of (Simple) Logistic Regression

Suppose we observe $0/1$ (Bernoulli) response Y and quantitative explanatory variable X. Let

$$\pi(x) = \mathrm{P}(Y = 1 \mid X = x) \quad \in (0, 1) \text{ for all } x$$

The (simple) logistic regression of Y on X uses

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}$$

i.e.

$$\text{logit}(\pi(x)) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

5.1.1 Interpretation β

Note: "odds of $Y = 1$" $= \frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x}$

Odds ratio for $Y = 1$ at $x + 1$ versus at x: $\frac{e^{\alpha+\beta(x+1)}}{e^{\alpha+\beta x}} = e^{\beta}$, which doesn't depend on x.

So β is the **log-odds ratio for the effect of increasing X by one unit**.

We can also show

$$\frac{d}{dx}\pi(x) = \beta\pi(x)(1 - \pi(x))$$

The **median effective level** is the $x = -\frac{\alpha}{\beta}$ (if $\beta \neq 0$) at which $\pi(x) = \frac{1}{2}$ and the $\pi(x)$ is the steepest.

β determines nature of $X - Y$ relationship:

- $\beta > 0$: increasing X increases prob. $Y = 1$
- $\beta = 0$: no relationship
- $\beta < 0$: increasing X decreases prob. $Y = 1$ Can show

5.1.2 Interpretation α

Interpreting α can be more difficult: it's the log-odds when $x = 0$.

If a mean-centered version of X is used ($\mathbb{E}X = 0$), α is the log-odds at the sample mean of the original X.

5.2 Inference

Assume independently-sampled data pairs $(x_i, y_i), i = 1, ..., N$. Let $\hat{\alpha}, \hat{\beta}$ be the MLEs.

5.2.1 Testing

We still use three testing methods: 1. Wald; 2. Likelihood ratio; 3. Score ("Rao").

Eg: Wald z-statistic for

$$H_0 : \beta = 0 \quad H_a : \beta \neq 0$$

is

$$\frac{\hat{\beta}}{SE(\hat{\beta})} \underset{H_0}{\dot{\sim}} N(0,1)$$

where the square of SE comes from the estimated asymptotic covariance matrix (estimated inverse information matrix).

Similarly, we can form a Wald Cl for β: (L, U).

Then, odds ratio e^{β} for increasing x by one unit is estimated by $e^{\hat{\beta}}$ and has CI (e^L, e^U).

5.2.2 Estimation

Estimated logistic curve:

$$\hat{\pi}(x) = \text{logit}^{-1}(\hat{\alpha} + \hat{\beta}x)$$

with estimated slope

$$\hat{\beta}\hat{\pi}(x)(1 - \hat{\pi}(x))$$

So the estimated median effective level (x such that $\hat{\pi}(x) = 1/2$) is $-\frac{\hat{\alpha}}{\hat{\beta}}$ if $\hat{\beta} \neq 0$.

For possible X value x_0, we estimate $\pi(x_0) = P(Y = 1 \mid X = x_0)$ with $\hat{\pi}(x_0)$.

Note: $SE\left(\text{logit}\,\hat{\pi}\,(x_0)\right)$ is the square root of

$$\begin{aligned}\widehat{\text{var}}\left(\text{logit}\,\hat{\pi}\,(x_0)\right) &= \widehat{\text{var}}\left(\hat{\alpha} + \hat{\beta}x_0\right) \\ &= \widehat{\text{var}}(\hat{\alpha}) + x_0^2\widehat{\text{var}}(\hat{\beta}) + 2x_0\widehat{\text{cov}}(\hat{\alpha}, \hat{\beta})\end{aligned}$$

where the estimated variances and covariance are from the estimated asymptotic covariance matrix.

The Wald Cl

$$\text{logit}\,\hat{\pi}\,(x_0) \pm z_{\alpha/2}SE\left(\text{logit}\,\hat{\pi}\,(x_0)\right)$$

can be transformed (inverse logit) to a Wald Cl for $\pi\,(x_0)$.

5.2.3 Testing Goodness of Fit; Remarks on Grouping

Various strategies:

- Test an added higher-order term, e.g. test $\beta_2 = 0$ in the quadratic

$$\alpha + \beta_1 x + \beta_2 x^2$$

- If there are **replicates** (repeated x values), use **grouped** (binomial) data to test using the deviance $D(\boldsymbol{y}; \hat{\boldsymbol{\mu}})$

  Remarks on grouping:

  1. MLEs and likelihood-based inference for parameters remain the same, since grouping doesn't change the kernel of the likelihood.
  2. The deviance changes, since the saturated model for the grouped data is different.
  3. But deviance-based comparisons between nested sub-models do not change, since the saturated model log-likelihood cancels out.
  4. A deviance-based goodness of fit test may be valid for the grouped data, provided the values of $\mathbb{E}(Y_i)$ and $n_i - \mathbb{E}(Y_i)$ are not too small.
  5. Note: "Fitted values" from R are still the probabilities $\hat{\pi}(x_i)$ (rather than estimated means $n_i\hat{\pi}(x_i)$).

5.3 Categorical Predictors

Consider categorical X with I categories.

After grouping, we obtain an $I \times 2$ table:

		success	failure
X category	1	Y_1	$n_1 - Y_1$
	$\vdots$	$\vdots$	$\vdots$
	$\vdots$	$\vdots$	$\vdots$
	I	Y_I	$n_I - Y_I$

Then Y_i are i.i.d. variables following the binomial distribution.

$$Y_i \sim \text{binomial}\,(n_i, \pi_i) \quad i = 1, \ldots, I$$

5.3.1 Interpretation of β

Code X using $I - 1$ indicator variables $\tilde{X}_i = [\tilde{x}_{i2}, \tilde{x}_{i3}, \cdots, \tilde{x}_{iI}]$ (No indicator variable for row 1 to avoid redundancy.):

$$\tilde{x}_{ij} = \begin{cases} 1, & \text{if } j = i \\ 0, & \text{otherwise} \end{cases}$$

For logistic regression,

$$\text{logit}(\pi_i) = \alpha + \beta_2 \tilde{x}_{i2} + \cdots + \beta_I \tilde{x}_{iI}$$

$$= \alpha + \beta_i$$

if we define $\beta_1 \equiv 0$. Then

$$\alpha = \text{logit}(\pi_1)$$

$$\beta_i = \text{logit}(\pi_i) - \text{logit}(\pi_1) = \ln\left(\frac{\pi_i/(1-\pi_i)}{\pi_1/(1-\pi_1)}\right)$$

$$= \text{ log-odds ratio from category } i \text{ to category } 1$$

Similarly,

$$\beta_i - \beta_{i'} = \text{log-odds ratio from category } i \text{ to category } i'$$

5.3.2 MLE Estimations

MLEs of the π_i s are

$$\hat{\pi}_i = p_i = \frac{y_i}{n_i}$$

and MLEs of α and β_i are obtained from the empirical (or sample) logits

$$\text{logit}(\hat{\pi}_i)$$

Letting $\beta_1 \equiv 0 \equiv \hat{\beta}_1$, we find

$$\hat{\alpha} = \text{logit}(\hat{\pi}_1) \quad \hat{\beta}_i = \text{logit}(\hat{\pi}_i) - \hat{\alpha}$$

For $i < i'$,

$$\hat{\beta}_i - \hat{\beta}_{i'} \text{ is the empirical log-odds ratio for the sub-table with only rows } i \text{ and } i'.$$

Note:

1. If $y_i = 0$ or $y_i = n_i$, then $\text{logit}(\hat{\pi}_i)$ does not exist, and thus neither does the MLE for the logistic regression.
2. This model is saturated, so a deviance-based goodness of fit test is not available.

5.3.3 Testing

Testing for any X effect is testing

$$H_0 : \beta_2 = \cdots = \beta_I = 0 \quad (\equiv \beta_1)$$

$$H_a : \text{not all } \beta_i \text{ s equal (including } \beta_1)$$

A LRT based on

$$M_0 : \text{logit}\,(\pi_i) = \alpha \quad M_1 : \text{logit}\,(\pi_i) = \alpha + \beta_i$$

uses $G^2\,(M_0 \mid M_1)$, which equals the G^2 statistic for testing independence/homogeneity in the $I \times 2$ table.

Alternative: The Pearson X^2 statistic.

Remark: When X is ordinal, can sometimes improve power by assigning numerical scores x_i^* to the X levels and then using the linear logit model

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

5.4 Multiple Logistic Regression

Allow p explanatory variables:

$$\text{logit}(\pi(\boldsymbol{x})) = \alpha + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$\pi(\boldsymbol{x}) = \frac{e^{\alpha+\beta_1 x_1+\cdots+\beta_p x_p}}{1 + e^{\alpha+\beta_1 x_1+\cdots+\beta_p x_p}}$$

"odds of $Y = 1$" $= \frac{\pi(\boldsymbol{x})}{1-\pi(\boldsymbol{x})} = e^{\alpha+\beta_1 x_1+\cdots+\beta_p x_p}$.

Increasing just x_j by 1 increases the logit by β_j, so

$$e^{\beta_j} = \text{ odds ratio at } x_j + 1 \text{ relative to } x_j$$

(all other variables held fixed)

More generally, consider adding Δx_j to $x_j, j = 1, \ldots, p$. Using matrix notation let

$$\mathbf{x} = \begin{pmatrix} 1 \\ x_1 \\ \vdots \\ x_p \end{pmatrix} \quad \boldsymbol{\delta} = \begin{pmatrix} 0 \\ \Delta x_1 \\ \vdots \\ \Delta x_p \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \alpha \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}$$

- In this notation, $\text{logit}(\pi(\mathbf{x})) = \mathbf{x}^T\boldsymbol{\beta}$
- Adding $\boldsymbol{\delta}$ to x adds $\boldsymbol{\delta}^T\boldsymbol{\beta}$ to the logit, so

$$e^{\boldsymbol{\delta}^T\boldsymbol{\beta}} = \text{ odds ratio at } \mathbf{x} + \boldsymbol{\delta} \text{ relative to } \mathbf{x}$$

5.5 Matrix-Vector Formulation of Estimation

$$Y_i \sim \text{ indep. binomial } (n_i, \pi\,(\boldsymbol{x}_i))$$

where $\boldsymbol{x}_i$ is the column vector of explanatory variable values for observation i, and (typically) a 1 for the intercept, and

$$\text{logit}\left(\pi\left(\boldsymbol{x}_i\right)\right) = \eta_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$$

Let

$$\begin{aligned}\boldsymbol{Y} &= \text{ column vector with } i \text{ th element } Y_i\\ \boldsymbol{X} &= \text{ matrix with } \boldsymbol{x}_i^T \text{ as } i \text{ th row}\\ \boldsymbol{\eta} &= \text{ column vector with } i \text{ th element } \eta_i\end{aligned}$$

Log-Likelihood

Note:$\boldsymbol{\eta} = \boldsymbol{X}\boldsymbol{\beta}$, $\pi\left(\boldsymbol{x}_i\right) = \frac{e^{\eta_i}}{1+e^{\eta_i}}$, and

$$\begin{aligned}l(\boldsymbol{\beta}) &= \prod_i \binom{n_i}{y_i} \pi\left(\boldsymbol{x}_i\right)^{y_i}\left(1-\pi\left(\boldsymbol{x}_i\right)\right)^{n_i-y_i}\\ L(\boldsymbol{\beta}) &= \sum_i \left[y_i \ln \pi\left(\boldsymbol{x}_i\right) + \left(n_i - y_i\right)\ln(1-\pi\left(\boldsymbol{x}_i\right))\right]\\ &= \sum_i \left[y_i(\eta_i - \ln\left(1+e^{\eta_i}\right)) - \left(n_i-y_i\right)\ln\left(1+e^{\eta_i}\right)\right]\\ &= \sum_i \left[y_i\eta_i - n_i\ln\left(1+e^{\eta_i}\right)\right]\end{aligned}$$

The log-likelihood (binomial response) is

$$L(\boldsymbol{\beta}) = \boldsymbol{y}^T\boldsymbol{X}\boldsymbol{\beta} - \sum_i n_i \ln\left(1+e^{\eta_i}\right)$$

where $\boldsymbol{y}$ is the observed version of $\boldsymbol{Y}$.

Score Vector

The score vector is

$$\begin{aligned}\boldsymbol{u}(\boldsymbol{\beta}) = \nabla L(\boldsymbol{\beta}) &= \boldsymbol{X}^T\boldsymbol{y} - \sum_i n_i\boldsymbol{x}_i\pi\left(\boldsymbol{x}_i\right)\\ &= \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{X}^T\boldsymbol{\mu}\end{aligned}$$

where $\boldsymbol{\mu} = E(\boldsymbol{Y})$

This equals $\mathbf{0}$ at the MLE, giving the likelihood equations

$$\boldsymbol{X}^T\boldsymbol{y} = \boldsymbol{X}^T\hat{\boldsymbol{\mu}}$$

(Note: $\boldsymbol{y} - \hat{\boldsymbol{\mu}}$ is orthogonal to the columns of $\boldsymbol{X}$, just like in linear regression.)

Information Matrix and Covariance

The second derivative matrix is

$$\begin{aligned}\nabla^2 L(\boldsymbol{\beta}) &= -\sum_i n_i \boldsymbol{x}_i \boldsymbol{x}_i^T \pi\left(\boldsymbol{x}_i\right)\left(1-\pi\left(\boldsymbol{x}_i\right)\right) \\ &= -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}\end{aligned}$$

where

$$\boldsymbol{W}=\operatorname{diag}\left(n_i \pi\left(\boldsymbol{x}_i\right)\left(1-\pi\left(\boldsymbol{x}_i\right)\right)\right)=\operatorname{diag}\left(\operatorname{var}\left(Y_i\right)\right)$$

Since this is not random (doesn't depend on $\boldsymbol{Y}$), the information matrix is

$$\mathcal{J}=-\nabla^2 L(\boldsymbol{\beta})=\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}$$

and

$$\widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}})=\left(\boldsymbol{X}^T \hat{\boldsymbol{W}} \boldsymbol{X}\right)^{-1}$$

where $\hat{\boldsymbol{W}}=\operatorname{diag}\left(n_i \hat{\pi}\left(\boldsymbol{x}_i\right)\left(1-\hat{\pi}\left(\boldsymbol{x}_i\right)\right)\right)$.

Wald Test

The standard error of $\operatorname{logit}(\hat{\pi}(\boldsymbol{x}))=\boldsymbol{x}^T \hat{\boldsymbol{\beta}}$ is then

$$\sqrt{\boldsymbol{x}^T \widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}}$$

and a Wald Cl for $\operatorname{logit}(\pi(\boldsymbol{x}))$ is

$$\boldsymbol{x}^T \hat{\boldsymbol{\beta}} \pm z_{\alpha / 2} \sqrt{\boldsymbol{x}^T \widehat{\operatorname{cov}}(\hat{\boldsymbol{\beta}}) \boldsymbol{x}}$$

which can be transformed to a Cl for $\pi(\boldsymbol{x})$.

5.6 Variable Selection

Goal: Choose the simplest model (fewest explanatory variables) that still explains the data well.

5.6.1 Collinearity

When there is **collinearity** (i.e., strong linear relationships among the X variables), the effect of one X variable may mask the effect of another.

Suggests that we should consider adding or removing variables *one-at-a-time*.

Definition 5.1 (Stepwise Procedures)

- **forward selection:** starting with none, add the "best" variable at each step, until the model stops improving
- **backward elimination:** starting with all, remove the "worst" variable at each step, stopping before the model becomes inadequate
- **stepwise selection:** perform forward selection, but also allow removal of variables (if appropriate) at each step

♣

Notes:

- If possible, consider interaction terms, not just main effects.
  (An interaction can be in the model only if the corresponding lower-order interactions/main effects are.)
- When evaluating a categorical variable (or its interactions), all of its indicator variables must be added/removed together.

5.6.2 Akaike Information Criterion (AIC)

Definition 5.2 (Akaike Information Criterion (AIC))

$$\text{AIC} = -2(\max \text{ log-likelihood} - \text{effective \# parameters})$$

Equivalent to an adjusted deviance:

$$D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) - 2 \cdot \mathbf{rdf} = D(\boldsymbol{y}; \hat{\boldsymbol{\mu}}) + 2(p+1) - 2n$$

where **rdf** is the (residual) degrees of freedom ($\text{rdf} = n$ - effective number of parameters), and we can ignore the constant $-2n$ when comparing models.

♣

Note: First term penalizes lack of fit, second term penalizes complexity (too many variables). We choose model with the smallest AIC.

5.7 Diagnostic

We have already seen general goodness-of-fit tests, but there are fit-related questions that concern individual observations:

1. Does the model fit all of the observations well?
2. Is the fit especially sensitive to certain observations?

5.7.1 Residuals

Definition 5.3 (Residuals)

1. **Pearson**:

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

The sum of their squares is a kind of generalized Pearson X^2 statistic.

2. **Deviance**:

$$\operatorname{sign}\left(y_i - n_i\hat{\pi}_i\right)\sqrt{d_i}$$

where

$$d_i = 2\left(y_i \ln\frac{y_i}{n_i\hat{\pi}_i} + \left(n_i - y_i\right)\ln\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)$$

The sum of their squares is the deviance $\left(G^2\right)$.

3. **Standardized**:

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

where $\hat{h}_i$ is the i th leverage (from the hat matrix) Idea: Divide the raw residual by (approximately) its standard deviation. Advantage: Closer to $N(0,1)$, when model fits.

♣

Usage: Check for unusual observations (outliers) or trends (when plotted versus unused variables).

Warning: For ungrouped data ($Y = 0$ or 1), residuals tend not to be very useful.

5.7.2 Influence

- Leverages $\hat{h}_i$ measure **potential influence**.

  Outliers having high leverage are particularly influential.

  (Remark: Unlike in linear regression, observations outlying in the X variables do not necessarily have high leverage.)

- **Cook's Distance**: for each observation, an overall measure of change in $\hat{\boldsymbol{\beta}}$ when that observation is removed.

  Generally indicates a problem if > 1.

  (Agresti instead uses a quantity c that is larger by a constant factor.)

- **Dfbeta**: for each parameter and each observation, the (standardized) change in the parameter estimate when the observation is removed.

  Indicates a problem if its magnitude is large, e.g. > 2 or > 3, since standardized. (But even just > 1

could indicate undue influence.)

5.8 Predictive Model Metrics

Want to evaluate

1. Strength of association (like R^2 in linear regression)
2. Performance of model as a classifier

5.8.1 Strength of Association

Definition 5.4 (Correlation Measure)

$$R(\boldsymbol{y}, \hat{\boldsymbol{\mu}}) = \text{sample correlation between}$$

$$y_i \text{ s and fitted values } \hat{\mu}_i$$

Agresti assumes data are in ungrouped format:

$$y_i = 0 \text{ or } 1 \quad \hat{\mu}_i = \hat{\pi}_i$$

Note: Similar to $\sqrt{R^2}$ in linear regression, where R^2 is the coefficient of determination.

Definition 5.5 (Likelihood Measures)

Consider a model M (that has an intercept).

Model	Maximum log-likelihood
intercept-only	L_0
M	L_M
saturated	L_S

$$(\text{so } L_0 \leq L_M \leq L_S)$$

Then use

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

with larger values indicating M better relative to intercept-only model.

Note: This likelihood measure may depend on the definition of the saturated model.

Recall: The saturated models are generally different for ungrouped (binary) and grouped (binomial) formats of the data.

Call the likelihood measure

$$D \text{ for the ungrouped (binary) format}$$

D^* for the grouped (binomial) format

Agresti recommends D over D^*.

♣

5.8.2 Performance of Model as a Classifier

Definition 5.6 (Classifier)

A logistic regression fit

$$\hat{\pi}(\boldsymbol{x}) = \text{ estimate of } \mathrm{P}(\text{ success } \mid \boldsymbol{x})$$

can be used as a **classifier**:

$$\hat{y} = \begin{cases} 1 \text{ (success)} & \text{if } \hat{\pi}(\boldsymbol{x}) > \pi_0 \\ 0 \text{ (failure)} & \text{if } \hat{\pi}(\boldsymbol{x}) \leq \pi_0 \end{cases}$$

for some cutoff π_0.

Could take $\pi_0 = 0.5$, or perhaps $\pi_0 =$ the observed fraction of successes.

♣

Definition 5.7 (Classification Table)

A classification table is a 2×2 contingency table of actual (binary) response y versus classified $\hat{y}$:

	$\hat{Y}=1$	$\hat{Y}=0$
$Y=1$		
$Y=0$		

Can be used to estimate

$$\textbf{sensitivity} = \mathrm{P}(\hat{Y} = 1 \mid Y = 1)$$

$$\textbf{specificity} = \mathrm{P}(\hat{Y} = 0 \mid Y = 0)$$

(These don't depend on the marginal distribution of Y)

♣

The proportion correct is

$$\mathrm{P}(Y = 1, \hat{Y} = 1) + \mathrm{P}(Y = 0, \hat{Y} = 0)$$

(Prob. of correct classification) and the error rate is its complement:

$$\mathrm{P}(Y = 1, \hat{Y} = 0) + \mathrm{P}(Y = 0, \hat{Y} = 1)$$

Note: These depend on the marginal distribution of Y (unlike sensitivity and specificity).

Therefore, they cannot be estimated from retrospectively sampled data alone.

- **Problem**: If same data is used to fit the model and to construct the table, performance estimates can be too optimistic.
- **Remedy**: Use leave-one-out cross validation, in which each observation is classified according to a model fit without it. Then build the table using the cross-validated predictions.

5.8.3 Receiver Operating Characteristic (ROC) Curves

Idea: Want to evaluate classification performance of a model without having to choose a cutoff π_0.

Plot sensitivity vs. $1-$specificity for all possible cutoffs π_0.

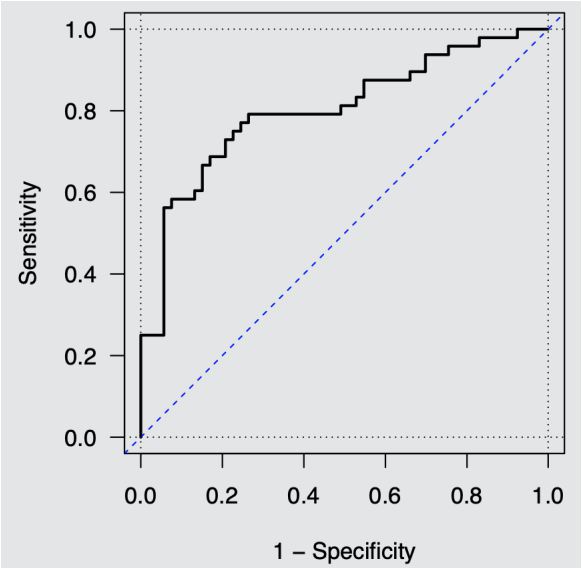

Figure 5.1: Example of ROC

Note: ROC curve

- is non-decreasing
- goes from $(0, 0)$ to $(1, 1)$
- is often above the $45°$ line (which represents random guessing.)
- indicates a better classifier if it is higher

The area under the ROC curve, a.k.a. the **"concordance index,"** measures predictive power.