1.1 Categorical Response Data

- Methods for response variable (a.k.a. outcome, dependent variable) $Y$ having a scale that is a set of categories.

- Explanatory variables (a.k.a. predictors, covariates, independent variables) may be categorical or continuous or both. Generically denoted $x_1, x_2, \text{etc}$.

**Example**

$Y = \text{vote in election (Dem, Rep, Indep)}$

$x's: \text{income, gender, race, education}$

Two Types of Categorical Variables

- **Nominal**: unordered categories
  - Example
    - patient condition (excellent, good, fair, poor)
    - transport to work (car, bus, bicycle, walk, other)

- **Ordinal**: ordered categories
  - Example
    - government spending (too high, about right, too low)
    - favorite music (rock, hiphop, pop, classical, jazz, country, folk)

We pay special attention to

Binary variables: success or failure

for which nominal-ordinal distinction is unimportant.

1.2 Probability Distributions for Categorical Data

The binomial distribution (and its generalization, the multinomial distribution) plays a similar role to that of the normal distribution for continuous response.

**Binomial Distribution**

- $n$ Bernoulli trials: two possible outcomes for each trial (success, failure)
  - $\pi = \Pr(\text{success}), 1 - \pi = \Pr(\text{failure})$, for each trial
  - trials are independent
  - $Y = \text{number of successes out of } n \text{ trials}$

$Y$ has a binomial distribution

When each trial has more than 2 possible outcomes, the joint distribution of the counts of outcomes in the various categories is a multinomial distribution (see text).
Probability Function of Binomial Distribution

\[ P(y) = \Pr(Y = y) = \frac{n!}{y!(n-y)!} \pi^y (1 - \pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n \]

where "y factorial" is given by
\[ y! = y(y-1)(y-2) \cdots 1 \quad \text{with} \quad 0! = 1 \]

Example
Vote (Dem, Rep)
Suppose \( \pi = Pr(\text{Dem}) = 0.6 \).
Sample \( n = 3 \) voters, let \( y \) = number of Dem votes among them.

\[
P(y) = \frac{3!}{y!(3-y)!} (0.6)^y (0.4)^{3-y}
\]

\[
P(0) = \frac{3!}{0!3!} (0.6)^0 (0.4)^3 = (0.4)^3 = 0.064
\]

\[
P(1) = \frac{3!}{1!2!} (0.6)^1 (0.4)^2 = 3(0.6)(0.4)^2 = 0.288
\]

\[
\begin{array}{c|c}
y & P(y) \\
0 & 0.064 \\
1 & 0.288 \\
2 & 0.432 \\
3 & 0.216 \\
\hline
1 & 1 \end{array}
\]

R Code
```r
> dbinom(0, 3, .4)
[1] 0.216
> dbinom(1, 3, .4)
[1] 0.432
> dbinom(0:3, 3, .4)
[1] 0.216 0.432 0.288 0.064
> cbind(0:3, dbinom(0:3, 3, .4))
     [,1] [,2]
[1,] 0 0.216
[2,] 1 0.432
[3,] 2 0.288
[4,] 3 0.064
```
Facts About the Binomial Distribution

- $E(Y) = n\pi$
- $\sigma^2 = \text{Var}(Y) = n\pi(1 − \pi), \; \sigma = \sqrt{n\pi(1 − \pi)}$
- $p = \frac{Y}{n} = \text{proportion of success (also denoted } \hat{\pi})$
- $E(p) = E\left(\frac{Y}{n}\right) = \pi \quad \text{(mean of } p \text{ is } \pi)$
- $\sigma(p) = \sqrt{\frac{\pi(1 − \pi)}{n}} \quad \text{(std error of } p)$

- Binomial distribution can be approximated by a normal distribution when $n$ is large ($n \cdot \min\{p, 1 − p\} \geq 5$).
1.3 Statistical Inference for a Proportion

Parameters are often estimated using maximum likelihood (ML).

Definition

The likelihood function is the probability of the observed data, expressed as a function of the parameter value.

Example

Binomial, \( n = 3 \), observe \( y = 1 \). Then

\[
p(1) = \frac{3!}{1!2!} \pi(1 - \pi)^2 = 3\pi(1 - \pi)^2 =: l(\pi)
\]

is the likelihood function, defined for \( \pi \) between 0 and 1.

\[
\begin{align*}
\pi &= 0.1 : l(0.1) = 3(0.1)(0.9)^2 = 0.243 \\
\pi &= 0.4 : l(0.4) = 3(0.4)(0.6)^2 = 0.432 \\
\pi &= 0.6 : l(0.6) = 3(0.6)(0.4)^2 = 0.288
\end{align*}
\]

Definition

The maximum likelihood estimate (MLE) is the parameter value at which the likelihood function is maximized.

Example

\[
l(\pi) = 3\pi(1 - \pi)^2
\]

is maximized at \( \hat{\pi} = 1/3 = 0.333 \)

I.e., \( y = 1 \) success in \( n = 3 \) trials is more likely for \( \pi = 1/3 \) than for any other value of \( \pi \).
Plot of Binomial Likelihood Function when $n = 3, y = 1$

Naturally, the likelihood function and the MLE depend on the data. If we observe $y = 0$ successes in $n = 3$ trials, then the MLE is $\hat{\pi} = \frac{0}{3} = 0$.

Plot of Binomial Likelihood Function when $n = 3, y = 0$

Plot of Binomial Likelihood Function when $n = 3, y = 2$

If we observe $y = 2$ successes in $n = 3$ trials, then MLE is $\hat{\pi} = \frac{2}{3} = 0.667$. 

Plot of Binomial Likelihood Function when $n = 3, y = 2$
Plot of Binomial Likelihood Function when $n = 3$, $y = 3$

If we observe $y = 3$ successes in $n = 3$ trials, then MLE is $\hat{\pi} = \frac{3}{3} = 1.$

Facts About MLEs

- For binomial, $\hat{\pi} = \frac{y}{n} =$ sample proportion of successes.
- If $y_1, y_2, \ldots, y_n$ are independent observations from a fixed normal distribution, then the MLE of the underlying mean $\mu$ is $\hat{\mu} = \bar{y}$ (sample mean). Same is true for Poisson distribution.
- In ordinary linear regression with $Y \sim$ normal, the usual least squares estimates are MLEs.
- For large sample size $n$, MLEs are optimal (no other estimator has smaller mean squared error: variance plus squared bias). This is true in fairly broad generality.
- For large $n$, the sampling distribution of the MLE is approximately normal. Again, this is true in fairly broad generality.

ML Inference for a Binomial Success Probability

MLE of $\pi$ is $\hat{\pi} = p = \frac{y}{n}$.

Recall $E(p) = \pi$, $\sigma(p) = \sqrt{\frac{\pi(1 - \pi)}{n}}$.

- Note that $\sigma(p) \downarrow 0$ as $n \uparrow \infty$, so $p \to \pi$ in probability by law of large numbers. Say that $p$ is a consistent estimator of $\pi$.

MLEs are generally consistent.

- $p$ is a sample mean for 0-1 data, so by the Central Limit Theorem, the sampling distribution of $p$ is approximately normal for large $n$.

Again, this is generally true for MLEs.
Significance Test for Binomial Parameter

\[ H_0 : \pi = \pi_0 \]
\[ H_a : \pi \neq \pi_0 \quad \text{(or 1-sided alternative)} \]

If \( H_0 \) is true, then the test statistic

\[ z = \frac{p - \pi_0}{\sigma(p)} = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \]

has a large-sample standard normal \((N(0, 1))\) null distribution (this is the reference distribution for the test). Note that we used the null s.e. for the test.

Definition

p-value = prob of results at least as extreme as observed (if null were true)

For the two-sided alternative hypothesis above, we use the two-tailed probability \( Pr(|Z| > |z| \mid H_0) \).

Confidence Interval for Binomial Parameter

Definition

The Wald CI for a parameter \( \theta \) is \( \hat{\theta} \pm z_{\alpha/2} \text{SE} \), where SE is the estimated standard error of \( \hat{\theta} \).

For a 95\% CI, \( \alpha = 5\% = 0.05 \) and \( z_{\alpha/2} = z_{0.025} = 1.96 \), so take \( \pm 1.96 \) standard errors.

Example

\( \theta = \pi : \) MLE is \( \hat{\pi} = \hat{\theta} = p \)
\( \sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}} \) estimated by \( \text{SE} = \sqrt{\frac{p(1-p)}{n}} \)

95\% CI for \( \pi : p \pm 1.96 \sqrt{\frac{p(1-p)}{n}} \)

Example (Wald Interval Collapsing)

Estimate \( \pi \), the population proportion of vegetarians.

For \( n = 20 \), suppose we observe \( y = 0 \).

\( p = \frac{0}{20} = 0 \)

95\% CI: \( 0 \pm 1.96 \sqrt{\frac{0(1-0)}{20}} = 0 \pm 0 = (0, 0) \)
Remarks

- Wald intervals often have poor performance in categorical data analysis unless \( n \) is quite large.
- Wald CI for \( \pi \) collapses if \( p = 0 \) or 1.
- Actual coverage probability much less than 0.95 if \( \pi \) close to 0 or 1.
- Wald 95% CI is the set of \( \pi_0 \) values with p-value > .05 when testing
  \[ H_0 : \pi = \pi_0 \quad \text{vs} \quad H_a : \pi \neq \pi_0 \]
  using the test statistic
  \[ z = \frac{p - \pi_0}{\sqrt{\frac{p(1-p)}{n}}} \]
  (denominator is estimated std error)

Definition

The score test and score CI use null hypothesis value of std error.

E.g., score 95% CI is the set of \( \pi_0 \) values for which p-value > .05 when testing
  \[ H_0 : \pi = \pi_0 \quad \text{vs} \quad H_a : \pi \neq \pi_0 \]
  using the test statistic
  \[ z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1-\pi_0)}{n}}} \]
  (null std error in denominator; known, not est'd)

Example

\( \pi \) = probability of being vegetarian.

\( n = 20, \quad y = 0, \quad p = \frac{0}{20} = 0 \)

What values of \( \pi_0 \) satisfy
  \[ \frac{|0 - \pi_0|}{\pi_0(1-\pi_0)} < 1.96 \quad \text{i.e.,} \quad |0 - \pi_0| < 1.96 \sqrt{\frac{\pi_0(1-\pi_0)}{20}} \]
Get equality at \( \pi_0 = 0 \) and \( \pi_0 = .16 \) (solve quadratic equation).
Inequality is satisfied for all values of \( \pi_0 \) between 0 and .16.
So 95% score CI for \( \pi \) is \((0, .16)\), more sensible than Wald interval.
When solving the quadratic, can show that midpoint of 95% score CI is
\[ y + \frac{(1.92^2)/2}{n + 1.96^2} \approx y + \frac{2}{n + 4}. \]

Wald CI \( p \pm 1.96 \sqrt{\frac{p(1-p)}{n}} \) also works reasonably well if we add 2 successes and 2 failures before computing \( p \) and using the formulas. This is the "Agresti-Coull" method.

For inference about proportions, score methods tend to perform better than Wald methods, in terms of having actual error rates closer to the nominal levels.

Another good test and CI use the likelihood function (e.g., the CI is the set of values of \( \pi \) for which \( l(\pi) \) is close to \( l(\hat{\pi}) \), i.e., the set of values of \( \pi_0 \) not rejected in a likelihood ratio test.

For small \( n \), can also do inference using the actual binomial sampling distribution of the data instead its normal approximation.

---

R Functions for Simple Binomial Tests and CIs

\texttt{prop.test} computes score test and CI.

- Default test is for \( H_0 : \pi = 0.5 \) vs \( H_a : \pi \neq 0.5 \)
- Uses continuity correction by default.

\texttt{prop.test(0, 20)}

1-sample proportions test with continuity correction
data: 0 out of 20, null probability 0.5
X-squared = 18.05, df = 1, p-value = 2.152e-05
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.00000 0.20045
sample estimates:
p
0

\texttt{prop.test(0, 20, correct=FALSE)}

1-sample proportions test without continuity correction
data: 0 out of 20, null probability 0.5
X-squared = 20, df = 1, p-value = 7.744e-06
alternative hypothesis: true p is not equal to 0.5
95 percent confidence interval:
0.00000 0.16113
sample estimates:
p
0
binom.test does exact test and corresponding exact CI.

- Default test is for $H_0: \pi = 0.5$ vs $H_a: \pi \neq 0.5$
- Uses continuity correction by default.

```r
> binom.test(0,20)

Exact binomial test

data:  0 and 20
number of successes = 0, number of trials = 20,
p-value = 1.907e-06
alternative hypothesis: true probability of success is not equal to 0.5
95 percent confidence interval:
 0.00000 0.16843
sample estimates:
  probability of success
        0
```

2. Contingency Tables

Two-Way Contingency Tables

*Contingency table:* cells contain counts of outcomes.

A two-way table with $I$ rows and $J$ columns is called an $I \times J$ table.

Example (Physicians Health Study (5 years))

Myocardial Infarction (MI) = heart attack. $2 \times 2$ table.

<table>
<thead>
<tr>
<th>Group</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
</tr>
<tr>
<td>Placebo</td>
<td>189</td>
</tr>
<tr>
<td>Aspirin</td>
<td>104</td>
</tr>
</tbody>
</table>

Still $2 \times 2$:

<table>
<thead>
<tr>
<th>Group</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Placebo</td>
<td>189</td>
</tr>
<tr>
<td>Aspirin</td>
<td>104</td>
</tr>
</tbody>
</table>

Conditional Distributions

A *conditional distribution* of $Y$ given $X$ refers to the probability distribution of $Y$ when we restrict attention to a fixed level of $X$.

Example (Physicians Health Study)

<table>
<thead>
<tr>
<th>Group</th>
<th>MI</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Y</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td>Placebo</td>
<td>0.017</td>
</tr>
<tr>
<td>Aspirin</td>
<td>0.009</td>
</tr>
</tbody>
</table>

Sample (or estimated) conditional probs for placebo group are

\[
0.017 = \frac{189}{11,034}, \quad 0.983 = \frac{10,845}{11,034}
\]

Natural way to look at data when

- $Y$ = response variable (e.g., heart attack: yes/no)
- $X$ = explanatory variable (e.g., group: aspirin/placebo)
Example (Diagnostic Disease Tests)

- **Y** = outcome of test: 1 = positive, 2 = negative
- **X** = actual condition: 1 = diseased, 2 = not diseased

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- sensitivity = \( \Pr(Y = 1 \mid X = 1) \)
- specificity = \( \Pr(Y = 2 \mid X = 2) \)

If you get a positive result, more relevant to you is \( \Pr(X = 1 \mid Y = 1) \). If disease is relatively rare, this may be low even if sensitivity and specificity are high (see pp. 23–24 of text for an example).

Joint and Marginal Distributions

What if \( X \) and \( Y \) are both *response* variables?

- \( \pi_{ij} = \Pr(X = i, Y = j), \quad i = 1, \ldots, I, \quad j = 1, \ldots, J \)

\( \{\pi_{ij}\} \) forms the *joint distribution* of \( X \) and \( Y \).

\[
\begin{align*}
\Pr(X = i) &= \pi_{i+} = \sum_j \pi_{ij} = \pi_{i1} + \cdots + \pi_{iJ} \\
\Pr(Y = j) &= \pi_{+j} = \sum_i \pi_{ij} = \pi_{1j} + \cdots + \pi_{IJ}
\end{align*}
\]

\( \{\pi_{i+}\} \) form the *marginal distribution* of \( X \).

\( \{\pi_{+j}\} \) form the *marginal distribution* of \( Y \).

2 × 2 example:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Sample cell counts: \( \{n_{ij}\} \)

Cell proportions: \( \{p_{ij}\} \)

\[
p_{ij} = \frac{n_{ij}}{n} \quad \text{where} \quad n = \sum_i \sum_j n_{ij}
\]

2 × 2 example:

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Independence

Definition (Statistical Independence)

X and Y are statistically independent if the true conditional distribution of Y is the same at each level of X.

2 × 2 example. Rows represent conditional distributions of Y given X.

\[
\begin{array}{c|cc|c}
& 1 & 2 & 1 \\
\hline
X & \cdot & \cdot & \cdot \\
1 & .01 & .99 & 1 \\
2 & .01 & .99 & 1 \\
\end{array}
\]

Fact: X and Y are independent if and only if

\[
Pr(X = i, Y = j) = Pr(X = i) \cdot Pr(Y = j)
\]

for all i and j,

i.e., \( \pi_{ij} = \pi_i + \pi_j \) for all i and j.

2 × 2 example:

\[
\begin{array}{c|ccc|c}
& 1 & 2 & .7 \\
\hline
X & \cdot & \cdot & \cdot \\
1 & .42 & .28 & .7 \\
2 & .18 & .12 & .3 \\
\hline
& .6 & .4 & 1 \\
\end{array}
\]

Comparing Proportions in 2 × 2 Tables

Conditional distributions:

\[
\begin{array}{c|cc|c}
& S & F \\
\hline
X & \pi_1 & \pi_2 \\
1 & 1 - \pi_1 & 1 - \pi_1 \\
2 & \pi_2 & \pi_2 \\
\end{array}
\]

\( \hat{\pi}_1 - \hat{\pi}_2 = p_1 - p_2 \)

\[
SE(p_1 - p_2) = \sqrt{\frac{p_1(1 - p_1)}{n_1} + \frac{p_2(1 - p_2)}{n_2}}
\]

Example (Physicians Health Study)

\[
p_1 = 0.017 \quad p_2 = 0.009 \quad p_1 - p_2 = 0.008
\]

\[
SE = \sqrt{\frac{0.017 \times 0.983}{11034} + \frac{0.009 \times 0.991}{11037}} = 0.0015
\]

95% CI for \( \pi_1 - \pi_2 \): .008 ± 1.96(.0015) = .008 ± .003 = (.005, .011)

Apparently \( \pi_1 - \pi_2 > 0 \) (i.e., \( \pi_1 > \pi_2 \)).
Relative Risk

\[ \text{relative risk} = \frac{\pi_1}{\pi_2} \]

Example (Physicians Health Study)

Sample relative risk in the Physicians Health Study is

\[ \frac{p_1}{p_2} = \frac{0.017}{0.009} = 1.82 \]

Sample proportion of heart attacks was 82% higher for placebo group.

▶ See p. 58 of text for SE of sample relative risk.
▶ Use SE to form CI for \( \pi_1/\pi_2 \).

Example: 95% CI for RR in PHS is \((1.43, 2.31)\).
▶ Independence \( \iff \frac{\pi_1}{\pi_2} = 1 \)

Odds Ratio

<table>
<thead>
<tr>
<th>Group</th>
<th>S</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>( \pi_1 )</td>
<td>( 1 - \pi_1 )</td>
</tr>
<tr>
<td>2</td>
<td>( \pi_2 )</td>
<td>( 1 - \pi_2 )</td>
</tr>
</tbody>
</table>

The odds of response S (instead of F) is \( \frac{\Pr(S)}{\Pr(F)} \).

In the \( 2 \times 2 \) table above:

\[ \text{odds}(S) = \begin{cases} \frac{\pi_1}{1 - \pi_1} & \text{in row 1} \\ \frac{\pi_2}{1 - \pi_2} & \text{in row 2} \end{cases} \]

Note

▶ if odds = 3, then S is three times as likely as F;
▶ if odds = \( \frac{1}{3} \), then F is three times as likely as S.

\[
\Pr(S) = \frac{\text{odds}(S)}{1 + \text{odds}(S)}
\]

\[ \text{odds}(S) = 3 \implies \Pr(S) = \frac{3}{1 + 3} = \frac{3}{4} \quad \Pr(F) = \frac{1}{4} \]

\[ \text{odds}(S) = \frac{1}{3} \implies \Pr(S) = \frac{1/3}{1 + 1/3} = \frac{1/3}{4/3} = \frac{1}{4} \quad \Pr(F) = \frac{3}{4} \]
Definition (Odds Ratio)

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

Example (Physicians Health Study)

<table>
<thead>
<tr>
<th>Group</th>
<th>MI</th>
<th>Y</th>
<th>N</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td></td>
<td>189</td>
<td>10845</td>
<td>11034</td>
</tr>
<tr>
<td>Aspirin</td>
<td></td>
<td>104</td>
<td>10933</td>
<td>11037</td>
</tr>
</tbody>
</table>

Sample proportions:

\[ \begin{align*}
  p_1 &= 0.0171 \\
  1 - p_1 &= 0.9829 \\
  p_2 &= 0.0094 \\
  1 - p_2 &= 0.9906
\end{align*} \]

Sample odds:

\[ \begin{align*}
  \text{placebo} &= \frac{0.0171}{0.9829} = 0.0174 \\
  \text{aspirin} &= \frac{0.0094}{0.9906} = 0.0095
\end{align*} \]

Sample odds ratio:

\[ \hat{\theta} = \frac{0.0174}{0.0095} = 1.83 \]

Estimate odds of heart attack in placebo group to be 1.83 times odds in aspirin group.

Properties of the Odds Ratio

- For counts

\[ \hat{\theta} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} \]

- Treats X, Y symmetrically:

\[ \begin{align*}
  \text{Placebo} & \quad 189 & 10845 \\
  \text{Aspirin} & \quad 104 & 10933
\end{align*} \]

\[ \hat{\theta} = 1.83 \]

- Each odds \( \geq 0 \) and \( \theta \geq 0 \).

- \( \theta = 1 \) when \( \pi_1 = \pi_2 \); i.e., when response independent of group.

- The further \( \theta \) is from 1, the stronger the association.

(For Y = lung cancer, some studies have \( \theta \approx 10 \) for X = smoking, \( \theta \approx 2 \) for X = passive smoking.)
► If rows are interchanged (or if columns are interchanged), $\theta \mapsto 1/\theta$.

For example, a value of $\theta = 1/5$ indicates the same strength of association as $\theta = 5$, but in the opposite direction.

$\theta = 1 \iff \log \theta = 0$

The log odds ratio ($\log \theta$) is symmetric about 0, e.g.,

$\theta = 2 \iff \log \theta = 0.7$

$\theta = 1/2 \iff \log \theta = -0.7$

► Sampling distribution of $\hat{\theta}$ is skewed to the right.

Normal approximation is good only if $n$ is very large.

► Sampling distribution of $\log \hat{\theta}$ is closer to normal, so construct CI for $\log \theta$ and then exponentiate endpoints to get CI for $\theta$.

Note: We use “natural logs” (with base $e = 2.718\ldots$) LN on most calculators.

---

**A Confidence Interval for the Odds Ratio**

Large-sample (asymptotic) SE of $\log \hat{\theta}$ is

$$SE(\log \hat{\theta}) = \sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}}$$

CI for $\log \theta$: $(L, U) = \log \hat{\theta} \pm z_{\alpha/2} \times SE(\log \hat{\theta})$

CI for $\theta$: $(e^L, e^U)$.

---

**Example (Physicians Health Study)**

$$\hat{\theta} = \frac{189 \times 10933}{104 \times 10845} = 1.83$$

$$\log \hat{\theta} = \log(1.83) = 0.605$$

$$SE(\hat{\theta}) = \sqrt{\frac{1}{189} + \frac{1}{10845} + \frac{1}{104} + \frac{1}{10933}} = 0.123$$

95% CI for $\log \theta$: $0.605 \pm 1.96(0.123) = (0.365, 0.846)$

95% CI for $\theta$: $(e^{0.365}, e^{0.846}) = (1.44, 2.33)$

Apparently $\theta > 1$. 
Remarks

- $\hat{\theta}$ not midpoint of CI because of skewness
- Better estimate if we use $\{n_{ij} + 0.5\}$. Especially if any $n_{ij} = 0$.
- When $\pi_1$ and $\pi_2$ close to zero,

$$\theta = \frac{\pi_1}{1 - \pi_1} \approx \frac{\pi_1}{\pi_2} = \text{relative risk}$$

Review: Exponential and Natural Logarithm Functions

$$\exp x = e^x \quad \text{(exponential function)}$$

$$e^0 = 1 \quad e^1 = 2.718 \ldots \quad e^{-1} = \frac{1}{e} = 0.368$$

$e^x > 0$ for all $x$

Exponential function is the antilog for the natural logarithm $\ln = \log_e$

$$e^x = y \iff \log_e(y) = x$$

$e^0 = 1$ means $\log_e(1) = 0$

$e^1 = 2.718$ means $\log_e(2.718) = 1$

$e^{-1} = 0.368$ means $\log_e(0.368) = -1$

$\log_e(2) = 0.693$ means $e^{0.693} = 2$

Example

Case-control study in London Hospitals (Doll and Hill, 1950)

$X = \text{smoked } \geq 1 \text{ cigarette per day for at least 1 year}$

$Y = \text{lung cancer}$

<table>
<thead>
<tr>
<th></th>
<th>Smoked</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Yes</td>
<td>688</td>
<td>650</td>
</tr>
<tr>
<td>No</td>
<td>21</td>
<td>59</td>
</tr>
<tr>
<td>Total</td>
<td>709</td>
<td>709</td>
</tr>
</tbody>
</table>

Case-control studies are “retrospective.” Binomial sampling model applies to $X$ (sampled within levels of $Y$), not to $Y$.

Cannot estimate $\Pr(Y = \text{yes}|X)$, nor

$$\pi_1 - \pi_2 = \Pr(Y = \text{yes}|X = \text{yes}) - \Pr(Y = \text{yes}|X = \text{no})$$

nor $\pi_1 / \pi_2$. 
Example (ctd)

We can estimate \( \Pr(X|Y) \) so we can estimate \( \theta \) (recall that \( \theta \) treats rows and columns symmetrically).

\[
\hat{\theta} = \frac{\Pr(X = \text{yes}|Y = \text{yes})/\Pr(X = \text{no}|Y = \text{yes})}{\Pr(X = \text{yes}|Y = \text{no})/\Pr(X = \text{no}|Y = \text{no})} = \frac{688/709}{21/709} = 2.97
\]

Odds of lung cancer for smokers estimated to be about 3 times the odds for non-smokers.

If \( \Pr(Y = \text{yes}|X) \) is near 0 (lung cancer rare in both groups), then \( \theta \approx \pi_1/\pi_2 = \text{relative risk} \), and can conclude that probability of lung cancer is \( \approx 3 \) times as high for smokers as for non-smokers.

Testing Independence

Example (Job Satisfaction and Income)

Data from General Social Survey (1991)

<table>
<thead>
<tr>
<th>Income</th>
<th>Job Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dissat</td>
</tr>
<tr>
<td>&lt;5K</td>
<td>2</td>
</tr>
<tr>
<td>5K–15K</td>
<td>2</td>
</tr>
<tr>
<td>15K–25K</td>
<td>0</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>4</td>
</tr>
</tbody>
</table>

\( H_0 : X \) and \( Y \) independent vs \( H_a : X \) and \( Y \) dependent

\( H_0 \) means that for all \((i, j)\)

\[
\Pr(X = i, Y = j) = \Pr(X = i) \Pr(Y = j)
\]

\( \pi_{ij} = \pi_i \pi_j \)

Expected frequency is

\[
\mu_{ij} = \text{mean of dist. of cell count } n_{ij}
\]

\[
= n \pi_{ij}
\]

\[
= n \pi_i \pi_j \quad \text{under } H_0
\]

MLEs under \( H_0 \) are

\[
\hat{\mu}_{ij} = n \pi_i \pi_j
\]

\[
= n \left( \frac{n_i}{n} \right) \left( \frac{n_j}{n} \right) = \frac{n_i n_j}{n}
\]

\( \hat{\mu}_{ij} \) are called estimated expected frequencies.
Chi-Squared Test of Independence

Usual test statistic is Pearson’s chi-squared statistic:

\[ X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \sum_{\text{all cells}} \frac{\text{observed} - \text{expected}}{\text{expected}} \]

\( X^2 \) has a large-sample chi-squared dist. under \( H_0 \), with

\[ df = (I - 1)(J - 1) \]

where \( I \) = number of rows, \( J \) = number of columns.

p-value = \( Pr(X^2 \geq X^2_{\text{obs}}) \) = right-tail prob

(Table given on p. 343 of text.)

Note: chi-squared dist. has \( \mu = df \), \( \sigma = \sqrt{2 \times df} \) and becomes more bell-shaped as \( df \) increases.

Example (Job Satisfaction and Income)

\[ X^2 = 11.5 \]
\[ df = (I - 1)(J - 1) = 3 \times 3 = 9 \]
\[ \text{p-value} = Pr(X^2 \geq 11.5) = 0.2415 \]

The evidence against \( H_0 \) is weak: it is plausible that job satisfaction and income are independent.

Likelihood-Ratio Test of Independence

Test statistic

\[ G^2 = -2 \log \left( \frac{\text{maximized likelihood when } H_0 \text{ true}}{\text{maximized likelihood generally}} \right) \]

\[ = 2 \sum \frac{n_{ij} \log \left( \frac{n_{ij}}{\hat{\mu}_{ij}} \right)}{\hat{\mu}_{ij}} \]

Dist. of \( G^2 \) under \( H_0 \) is also approx. chi-squared \( df = (I - 1)(J - 1) \).

Example (Job Satisfaction and Income)

\[ G^2 = 13.47 \]
\[ df = 9 \]
\[ \text{p-value} = 0.1426 \]
Degrees of Freedom for Chi-Squared Test

\[ \text{df for } \chi^2 \text{ test} = \# \text{ parameters in general} - \# \text{ parameters under } H_0 \]

Example (Chi-squared test of independence)

Independence: \( H_0 : \pi_{ij} = \pi_i + \pi_j \)

\[ \sum_{ij} \pi_{ij} = 1 \quad \sum_i \pi_i = 1 \quad \sum_j \pi_j = 1 \]

- In general there are \( IJ - 1 \) free parameters: If we know \( IJ - 1 \) of the \( \pi_{ij} \), then we know the last one because they must add to 1.
- Under \( H_0 \), there are \((I - 1) + (J - 1)\) free parameters: \((I - 1)\) free \( \pi_i + \) and \((J - 1)\) free \( \pi_j + \). These determine the \( \pi_{ij} \) under \( H_0 \).

Thus

\[ \text{df} = (IJ - 1) - [(I - 1) + (J - 1)] = (I - 1)(J - 1) \]

Remarks

- If all \( n_{ij} = \hat{\mu}_{ij} \), then \( \chi^2 = G^2 = 0 \).
- As \( n \uparrow \), \( \chi^2 \xrightarrow{d} \chi^2 \) faster than \( G^2 \xrightarrow{d} \chi^2 \), but \( \chi^2 \) and \( G^2 \) are usually similar if most \( \hat{\mu}_{ij} \geq 5 \).
- These tests treat X and Y as nominal: reordering rows or columns leaves \( \chi^2 \), \( G^2 \) unchanged.

Sec. 2.5 (we skip) presents ordinal tests. We re-analyze the job sat data with an ordinal model in Ch. 6 (more powerful test, much smaller p-value).

Residuals

Definition (Standardized (or Adjusted) Residuals)

\[ r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_i +)(1 - p_j +)}} \]

Under \( H_0 \) : independence, \( r_{ij} \approx \text{std normal } N(0, 1) \).

Example (Job Satisfaction and Income)

\[ n_{44} = 8 \quad \hat{\mu}_{44} = \frac{24 \times 23}{104} = 5.31 \]

\[ r_{44} = \frac{8 - 5.31}{\sqrt{5.31(1 - \frac{24}{104})(1 - \frac{23}{104})}} = 1.51 \]

None of the cells show very strong evidence of association.
Standardized Residuals for Job Satisfaction Data

<table>
<thead>
<tr>
<th>Income</th>
<th>Dissat</th>
<th>Little</th>
<th>Moderate</th>
<th>Very</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5K</td>
<td>1.44</td>
<td>0.73</td>
<td>-0.16</td>
<td>-1.08</td>
</tr>
<tr>
<td>5K–15K</td>
<td>0.75</td>
<td>0.87</td>
<td>0.60</td>
<td>-1.77</td>
</tr>
<tr>
<td>15K–25K</td>
<td>-1.12</td>
<td>-1.52</td>
<td>0.22</td>
<td>1.51</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>-1.12</td>
<td>-0.16</td>
<td>-0.73</td>
<td>1.51</td>
</tr>
</tbody>
</table>

Getting Tabled Data into R

There are many ways to enter contingency table data into R. With a simple two-way table, perhaps the easiest is to enter the data as matrix of counts. We will illustrate with Example 2.44 from the text concerning Party Affiliation by Gender (pag). Note that by default a matrix is read by columns. The `as.table()` function lets R know that the matrix represents a contingency table of counts.

```R
> pag.tab <- matrix(c(762, 484, 327, 239, 468, 477), nrow=2)
> dimnames(pag.tab) <-
    list(Gender=c("Female","Male"),
         Party=c("Democrat","Independent","Republican"))
> pag.tab <- as.table(pag.tab)

Gender   Party            Freq
Female   Democrat         762
         Independent      327
         Republican       468
Male     Democrat         484
         Independent      239
         Republican       477
```

Once the data are saved as a table as above, we can easily convert them to a data frame:

```R
> pag.df <- as.data.frame(pag.tab)

Gender  Party  Freq
1 Female Democrat  762
2   Male Democrat  484
3 Female Independent 327
4   Male Independent 239
5 Female Republican 468
6   Male Republican 477
```
Alternatively, we could create the data frame first, with a row for each combination of factor levels. Here the expand.grid function can save us some work.

\begin{verbatim}
> pag.df <- expand.grid(Gender=c("Female","Male"),
    Party=c("Democrat","Independent","Republican"))
> pag.df
   Gender Party
   1 Female Democrat
   2 Male Democrat
   3 Female Independent
   4 Male Independent
   5 Female Republican
   6 Male Republican
\end{verbatim}

\begin{verbatim}
> pag.df$Freq <- c(762, 484, 327, 239, 468, 477)
> pag.df
   Gender Party  Freq
   1 Female Democrat 762
   2 Male Democrat  484
   3 Female Independent 327
   4 Male Independent 239
   5 Female Republican 468
   6 Male Republican  477
\end{verbatim}

\begin{verbatim}
> xtabs(Freq ~ Gender + Party, data=pag.df)

             Party
Gender Democrat Independent Republican
Female   762      327        468
Male     484      239        477
\end{verbatim}

The data could also be read from the columns of a text file or a comma-separated (csv) file, which could be created with a text editor or a spreadsheet program. The text or csv file should have a separate row for each combination of factor levels.

Thus a text file Data/pag.txt containing

\begin{verbatim}
Gender       Party  Freq
Female Democrat 762
Male  Democrat  484
Female Independent 327
Male Independent 239
Female Republican 468
Male  Republican 477
\end{verbatim}

can be read into an R dataframe via

\begin{verbatim}
> pag.df <- read.table("Data/pag.txt", header=TRUE)
\end{verbatim}
Similarly, a csv file Data/pag.csv containing

"Gender","Party","Freq"
"Female","Democrat",762
"Male","Democrat",484
"Female","Independent",327
"Male","Independent",239
"Female","Republican",468
"Male","Republican",477

can be read into an R dataframe via

> pag.df <- read.csv("Data/pag.csv", header=TRUE)

See the R help for read.table and read.csv for more information.

Count data entered or read as rows in a dataframe can be converted to
a table using the xtabs function.

> names(pag.df)
[1] "Gender" "Party" "Freq"
> pag.tab <- xtabs(Freq ~ Gender + Party, data=pag.df)
> pag.tab


Computations on Tables
Marginal Totals

> margin.table(pag.tab, 1)

Gender
Female   Male
1557    1200
> margin.table(pag.tab, 2)

Party
Democrat Independent Republican
1246    566    945
> addmargins(pag.tab)

Party
Gender Democrat Independent Republican Sum
Female   762    327    468    1557
Male     484    239    477    1200
Sum      1246    566    945    2757
Computations on Tables
Conditional Distributions

> prop.table(pag.tab, 1)

    Party
Gender Democrat Independent Republican
Female 0.48940 0.21002 0.30058
Male 0.40333 0.19917 0.39750

> prop.table(pag.tab, 2)

    Party
Gender Democrat Independent Republican
Female 0.61156 0.57774 0.49524
Male 0.38844 0.42226 0.50476

> # round(prop.table(pag.tab, 1), 3)

Computations on Tables
Chi-Square Test for Independence

> chisq.test(pag.tab)

Pearson's Chi-squared test
data: pag.tab
X-squared = 30.07, df = 2, p-value = 2.954e-07

> pag.chisq <- chisq.test(pag.tab)
> names(pag.chisq)
[1] "statistic" "parameter" "p.value" "method"
[5] "data.name" "observed" "expected" "residuals"
[9] "stdres"
> pag.chisq$statistic
X-squared
30.07
> pag.chisq$parameter
df
2
> pag.chisq$p.value
[1] 2.9536e-07
The residuals computed by `chisq.test` are the unadjusted (raw) Pearson residuals:

```r
> pag.chisq$residuals
    Party
Gender Democrat Independent Republican
Female  2.19886  0.41137 -2.84324
Male   -2.50467 -0.46858  3.23867
```

To convert these to standardized Pearson residuals we must divide by \( \sqrt{(1 - p_{ij})(1 - p_{i+})(1 - p_{+j})} \), so we need the marginal proportions.

```r
> n <- sum(pag.tab)
> n.gender <- margin.table(pag.tab, 1)
> n.party <- margin.table(pag.tab, 2)
```
Check:

\[
> \frac{\text{n.gender} \odot \text{n.party}}{\text{n}}
\]

<table>
<thead>
<tr>
<th>Party</th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>703.67</td>
<td>319.65</td>
<td>533.68</td>
</tr>
<tr>
<td>Male</td>
<td>542.33</td>
<td>246.35</td>
<td>411.32</td>
</tr>
</tbody>
</table>

> pag.chisq$expected

<table>
<thead>
<tr>
<th>Party</th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>703.67</td>
<td>319.65</td>
<td>533.68</td>
</tr>
<tr>
<td>Male</td>
<td>542.33</td>
<td>246.35</td>
<td>411.32</td>
</tr>
</tbody>
</table>

The standardized (or adjusted) residuals are:

\[
> \frac{\text{pag.chisq$residual}}{\sqrt{\left(\frac{1-\text{n.gender}}{\text{n}}\right) \odot \left(\frac{1-\text{n.party}}{\text{n}}\right)}}
\]

<table>
<thead>
<tr>
<th>Party</th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>4.50205</td>
<td>0.69945</td>
<td>-5.31595</td>
</tr>
<tr>
<td>Male</td>
<td>-4.50205</td>
<td>-0.69945</td>
<td>5.31595</td>
</tr>
</tbody>
</table>

It is not difficult to write a simple function to do these computations. Put the following code into a text file named R/myadjresids.R:

```r
myadjresids <- function(x) {
  n <- sum(x)
  n1 <- margin.table(x,1)
  n2 <- margin.table(x,2)
  expected <- (n1 %o% n2) / n
  adj <- (1 - n1/n) %o% (1 - n2/n)
  (x - expected) / sqrt(expected * adj)
}
```

Then source the file into R and use the function whenever you need it:

```r
> source("R/myadjresids.R")
> myadjresids(pag.tab)
```

<table>
<thead>
<tr>
<th>Party</th>
<th>Democrat</th>
<th>Independent</th>
<th>Republican</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>4.50205</td>
<td>0.69945</td>
<td>-5.31595</td>
</tr>
<tr>
<td>Male</td>
<td>-4.50205</td>
<td>-0.69945</td>
<td>5.31595</td>
</tr>
</tbody>
</table>
Partitioning Chi-squared

The sum of two independent chi-squared random variables has a chi-squared distribution with df equal to the sum of the df of the two components. Symbolically:

$$\chi^2_a, \chi^2_b \text{ independent } \implies \chi^2_a + \chi^2_b \sim \chi^2_{a+b}$$

- G² statistic for testing independence can be partitioned into components representing certain aspects of the association.
- Partition of $\chi^2$ is only approximate.
- Text gives guidelines on how to partition so that separate components are independent. This is needed for $G^2$ to partition exactly.

Example (Job Satisfaction and Income)

<table>
<thead>
<tr>
<th>Income</th>
<th>Job Satisfaction</th>
<th>Dissat</th>
<th>Little</th>
<th>Moderate</th>
<th>Very</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5K</td>
<td></td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>5K–15K</td>
<td></td>
<td>2</td>
<td>6</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>15K–25K</td>
<td></td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>&gt;25K</td>
<td></td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>8</td>
</tr>
</tbody>
</table>

Recall $\chi^2 = 11.52$, $G^2 = 13.47$, df = 9.

<table>
<thead>
<tr>
<th>Income</th>
<th>Job Satisfaction</th>
<th>Dissat</th>
<th>Little</th>
<th>Moderate</th>
<th>Very</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5K</td>
<td></td>
<td>0.091</td>
<td>0.182</td>
<td>0.591</td>
<td>0.136</td>
</tr>
<tr>
<td>5K–15K</td>
<td></td>
<td>0.059</td>
<td>0.176</td>
<td>0.647</td>
<td>0.118</td>
</tr>
<tr>
<td>15K–25K</td>
<td></td>
<td>0.000</td>
<td>0.042</td>
<td>0.625</td>
<td>0.333</td>
</tr>
<tr>
<td>&gt;25K</td>
<td></td>
<td>0.000</td>
<td>0.125</td>
<td>0.542</td>
<td>0.333</td>
</tr>
</tbody>
</table>

Example (Job Satisfaction and Income (ctd))

<table>
<thead>
<tr>
<th>Income</th>
<th>JobSat</th>
<th>VD</th>
<th>LS</th>
<th>MS</th>
<th>VS</th>
<th>$\chi^2$</th>
<th>$G^2$</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;5</td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>3</td>
<td></td>
<td>0.30</td>
<td>0.30</td>
<td>3</td>
</tr>
<tr>
<td>5–15</td>
<td>2</td>
<td>6</td>
<td>22</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15–25</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>8</td>
<td></td>
<td>1.14</td>
<td>1.19</td>
<td>3</td>
</tr>
<tr>
<td>&gt;25</td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low vs High</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt;15</td>
<td>4</td>
<td>10</td>
<td>35</td>
<td>7</td>
<td></td>
<td>10.32</td>
<td>11.98</td>
<td>3</td>
</tr>
<tr>
<td>&gt;15</td>
<td>0</td>
<td>4</td>
<td>28</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sum</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>11.76</td>
<td>13.47</td>
<td>9</td>
</tr>
</tbody>
</table>

- Job satis. appears to depend on whether income > or < $15K.
- For $\chi^2$, note $0.30 + 1.14 + 10.32 = 11.76 \neq 11.52$. 
Exact Inference for Small Samples

2 × 2 Case: Fisher’s Exact Test

H₀: X, Y independent.

\[
\begin{array}{c|cc|c}
 & 1 & 2 & \text{Y} \\
\hline
X & n_{11} & n_{12} & n_{1+} \\
2 & n_{21} & n_{22} & n_{2+} \\
\hline
n_{+1} & n_{+2} & n \\
\end{array}
\]

Treating the row and column totals as fixed, the exact null distribution of \( \{n_{ij}\} \) is the hypergeometric distribution:

\[
P(n_{11}) = \frac{\binom{n_{1+} - n_{11}}{n_{11}} \binom{n_{2+} - n_{11}}{n_{2+} - n_{11}}}{\binom{n}{n_{11}}}
\]

where \( \binom{a}{b} = \frac{a!}{b!(a-b)!} \)

Example (Lady Tasting Tea (Fisher))

The lady is told that milk was poured first in 4 cups and tea first in the other 4. Order of tasting is randomized.

<table>
<thead>
<tr>
<th>Guess</th>
<th>Milk</th>
<th>Tea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poured First</td>
<td>?</td>
<td>4</td>
</tr>
<tr>
<td>Milk</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>Tea</td>
<td>4</td>
<td>8</td>
</tr>
</tbody>
</table>

Under H₀, \( n_{11} = 0, 1, 2, 3, \) or 4.

Under H₀, \( \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} \) has probability

\[
P(4) = \frac{\binom{4}{4} \binom{4}{4-4}}{\binom{8}{4}} = \frac{4! \cdot 4!}{8!} = \frac{4! \cdot 4!}{8!} = \frac{1}{70} = 0.014
\]

Example (Lady Tasting Tea (ctd))

Under H₀, \( \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \) has probability

\[
P(3) = \frac{\binom{4}{3} \binom{4}{1}}{\binom{8}{4}} = \frac{16}{70} = 0.229
\]

\[
cbind(0:4, dhyper(0:4, 4, 4, 4))
\]

<table>
<thead>
<tr>
<th>( n_{11} )</th>
<th>( P(n_{11}) )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.014</td>
</tr>
<tr>
<td>1</td>
<td>0.229</td>
</tr>
<tr>
<td>2</td>
<td>0.514</td>
</tr>
<tr>
<td>3</td>
<td>0.229</td>
</tr>
<tr>
<td>4</td>
<td>0.014</td>
</tr>
</tbody>
</table>
For 2 $\times$ 2 tables,

\[ H_0 : \text{indep} \iff H_0 : \theta = 1 \ (\theta = \text{odds ratio}) \]

To test \( H_0 : \theta = 1 \) vs \( H_a : \theta > 1 \)

\[ p-value = \Pr(\hat{\theta} \geq \hat{\theta}_{obs}) = \Pr(n_{11} \geq n_{11}^{obs}) \]

Example (Lady Tasting Tea (ctd))

Lady guesses correctly on 3 of the milk-first cups and 3 of the tea-first:

<table>
<thead>
<tr>
<th>Poured First</th>
<th>Milk</th>
<th>Tea</th>
</tr>
</thead>
<tbody>
<tr>
<td>Milk</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Tea</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>4</td>
</tr>
</tbody>
</table>

\[ n_{11} = 3 \]

\[ p-value = \Pr(n_{11} \geq 3) = P(3) + P(4) = 0.229 + 0.014 = 0.243 \]

Very little evidence against \( H_0 \).

```r
> TeaTasting <- matrix(c(3, 1, 1, 3), nrow = 2, 
                      dimnames = list(Truth = c("Milk", "Tea"),
                                      Guess = c("Milk", "Tea")))
> fisher.test(TeaTasting, alternative = "greater")
Fisher's Exact Test for Count Data
data: TeaTasting
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
0.31357 Inf
sample estimates:
odds ratio
6.4083
```

To test \( H_0 : \theta = 1 \) vs \( H_a : \theta \neq 1 \)

\[ p-value = \text{two-tail prob. of outcomes no more likely than obs.} \]

In the lady tasting tea example, the \( p \)-value for the two-tailed test is

\[ p-value = P(0) + P(1) + P(3) + P(4) = 0.486 \]

```r
> fisher.test(TeaTasting, alternative = "two.sided")
Fisher's Exact Test for Count Data
data: TeaTasting
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
0.21173 621.93375
sample estimates:
odds ratio
6.4083
```
Remarks

▶ If formal test, e.g., reject $H_0$ if p-value $\leq \alpha = .05$, actual $\Pr(\text{type I error}) < .05$ because of discreteness (see text).
▶ Margins may be fixed by design, but test valid even if not.
▶ Fisher's exact test extends to $1 \times J$ tables.

```R
> sattab
    Income  Dissat Little Moderate Very
<5K 2 4 13 3
5K--15K 2 6 22 4
15K--25K 0 1 15 8
>25K 0 3 13 8

> fisher.test(sattab)
Fisher's Exact Test for Count Data
data: sattab
p-value = 0.2315
alternative hypothesis: two.sided
```

Three-Way Contingency Tables

Example (FL Death Penalty Cases)

```R
> data(deathpenalty)
> dp <- xtabs(Freq ~ Victim + Defendant + DeathPenalty,
data=deathpenalty)
> dpflat <- ftable(DeathPenalty ~ Victim + Defendant,
data=dp)
> dpflat
```

```
DeathPenalty Yes No
Victim Defendant
White White 53 414
Black 11 37
Black White 0 16
Black 4 139
```
Example (FL Death Penalty Cases (ctd))

\[ Y = \text{death penalty (response var.)} \]

\[ X = \text{defendant's race (explanatory)} \]

\[ Z = \text{victim's race (control var.)} \]

\[
> \text{round(100*prop.table(dpflat,1), 1)}
\]

<table>
<thead>
<tr>
<th>Victim</th>
<th>Defendant</th>
<th>DeathPenalty</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
<tbody>
<tr>
<td>White</td>
<td>White</td>
<td>11.3</td>
<td>88.7</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>22.9</td>
<td>77.1</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td>White</td>
<td>0.0</td>
<td>100.0</td>
<td></td>
</tr>
<tr>
<td>Black</td>
<td></td>
<td>2.8</td>
<td>97.2</td>
<td></td>
</tr>
</tbody>
</table>

The tables:

\[
\begin{array}{llll}
\text{Def} & \text{DeathPen} & \text{Def} & \text{DeathPen} \\
\text{Yes} & \text{No} & \text{Yes} & \text{No} \\
\text{White} & 53 & 414 & \text{White} & 0 & 16 \\
\text{Black} & 11 & 37 & \text{Black} & 4 & 139 \\
\end{array}
\]

are called **partial tables**. They **control for** \( Z \) (hold it constant).

The (estimated) **conditional odds ratios** are:

\[
Z = \text{white} : \hat{\theta}_{XY(1)} = \frac{53 \times 37}{414 \times 11} = 0.43
\]

\[
Z = \text{black} : \hat{\theta}_{XY(2)} = \frac{0 \times 139}{16 \times 4} = 0 \quad (0.94 \text{ after add .5 to all cells})
\]

Controlling for victim’s race, odds of receiving death penalty were **lower** for white defendants than for black defendants.

Adding the partial tables gives **XY marginal table**.

\[
\begin{array}{llll}
\text{Def} & \text{DeathPen} & \text{Def} & \text{DeathPen} \\
\text{Yes} & \text{No} & \text{Yes} & \text{No} \\
\text{White} & 53 & 430 & \hat{\theta}_{XY} = 1.45 \\
\text{Black} & 15 & 176 & \\
\end{array}
\]

Ignoring victim’s race, odds of death penalty **higher** for white defendants.

Definition (Simpson’s Paradox)

All partial tables show reverse association from that in marginal table.

▶ **Cause**?

▶ **Moral**: can be dangerous to “collapse” contingency tables.
Definition (Conditional Independence)

X and Y are conditionally independent given Z if they are independent in each partial table.

In a $2 \times 2 \times K$ table this means

$$\theta_{XY(1)} = \cdots = \theta_{XY(K)} = 1.0$$

Remark

Conditional independence does not imply that X and Y are independent in the marginal two-way table.

Example:

<table>
<thead>
<tr>
<th>Clinic (Z)</th>
<th>Treatment (X)</th>
<th>Response (Y)</th>
<th>F</th>
<th>(\hat{\theta})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>18</td>
<td>12</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>12</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>2</td>
<td>8</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>8</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>Marginal</td>
<td>A</td>
<td>20</td>
<td>20</td>
<td>2.0</td>
</tr>
<tr>
<td></td>
<td>B</td>
<td>20</td>
<td>40</td>
<td></td>
</tr>
</tbody>
</table>

Generalized Linear Models

Components of a GLM

1. **Random Component**

   Identify response variable $Y$.

   Assume independent observations $y_1, \ldots, y_n$ from particular family of distributions, e.g., Poisson or binomial.

2. **Systematic Component**

   Model how $\mu = E(Y)$ depends on explanatory variables $x_1, \ldots, x_k$.

   - **Linear predictor**: $\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$.
   - **Link function**: Assume that $\mu = E(Y)$ satisfies

     $$g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

     $g$ is the Link function.
Examples

- \( \log(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \) uses \( g(\mu) = \log(\mu) \).
  - The log link is often used for a "count" response for which \( \mu > 0 \).
- \( \log \left( \frac{\mu}{1 - \mu} \right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \)
  - uses \( g(\mu) = \log \left( \frac{\mu}{1 - \mu} \right) \), the logit link. \( \text{logit} = \log(\text{odds}) \).
  - Often used for binomial, with \( \mu = \pi \) between 0 and 1.
- \( \mu = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \) uses \( g(\mu) = \mu \), the identity link, e.g., ordinary regression for a normal response.

Remarks

- A GLM generalizes ordinary regression by
  - permitting \( Y \) to have a nonnormal dist.
  - permitting modeling of \( g(\mu) \) rather than \( \mu \).
- The same ML (maximum likelihood) fitting procedure applies to all GLMs. It is the basis of the \texttt{glm()} function in R and of \texttt{proc genmod} in SAS.

GLMs for Binary Data

Suppose \( Y = 1 \) or 0 ("Bernoulli" or "binary" random variable).

Let \( P(Y = 1) = \pi \), \( P(Y = 0) = 1 - \pi \).

This is binomial for \( n = 1 \) trial.

\[
\begin{align*}
E(Y) &= \pi \\
\text{Var}(Y) &= \pi(1 - \pi)
\end{align*}
\]

For an explanatory variable \( x \), \( \pi = \pi(x) \) varies as \( x \) varies.
GLMs for Binary Data
Linear Probability Model

\[ \pi(x) = \alpha + \beta x \]

A GLM with binomial random component and identity link function.

\[ \text{Var}(Y) = \pi(x) \left[ 1 - \pi(x) \right] \]

varies as \( x \) varies, so least squares not optimal.

Use ML to fit this and other GLMs.

Example (Infant Malformation)

\[ Y = \text{infant sex organ malformation} \quad (1 = \text{present}, \quad 0 = \text{absent}) \]

\[ x = \text{mother's alcohol consumption (avg drinks per day)} \]

<table>
<thead>
<tr>
<th>Alcohol Consumption</th>
<th>Measured Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.0</td>
</tr>
<tr>
<td>&lt; 1</td>
<td>0.5</td>
</tr>
<tr>
<td>1–2</td>
<td>1.5</td>
</tr>
<tr>
<td>3–5</td>
<td>4.0</td>
</tr>
<tr>
<td>( \geq 6 )</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Example (Infant Malformation (ctd))

\[ > \text{data(malformation)} \]
\[ > \text{malformation} \]

<table>
<thead>
<tr>
<th>Alcohol Malformation</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0</td>
</tr>
<tr>
<td>2</td>
<td>0.0</td>
</tr>
<tr>
<td>3</td>
<td>0.5</td>
</tr>
<tr>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td>6</td>
<td>1.5</td>
</tr>
<tr>
<td>7</td>
<td>4.0</td>
</tr>
<tr>
<td>8</td>
<td>4.0</td>
</tr>
<tr>
<td>9</td>
<td>7.0</td>
</tr>
<tr>
<td>10</td>
<td>7.0</td>
</tr>
</tbody>
</table>

Linear probability model for \( \pi = \text{Pr(malformation present)} \) has fit
Example (Infant Malformation (ctd))
> malform.wd <-
    reshape(malformation,
        idvar="Alcohol", timevar="Malformation",
        direction="wide")
> malform.wd
Alcohol  Freq.Present  Freq.Absent
1     0.0          48    17066
3     0.5          38    14464
5     1.5           5    788
7     4.0           1    126
9     7.0           1    37
> names(malform.wd)[2:3] <- c("Present","Absent")
> rownames(malform.wd) <- 1:5

Example (Infant Malformation (ctd))
> malform.wd <-
    transform(malform.wd,
        Total = Present + Absent,
        PctPresent = 100*Present/(Present + Absent))
> malform.wd
Alcohol  Present  Absent  Total  PctPresent
1     0.0          48    17066  0.28047
2     0.5          38    14464  0.26203
3     1.5           5    788   0.63052
4     4.0           1    126   0.78740
5     7.0           1    37    2.63158

Example (Infant Malformation (ctd))
Two ways to fit the same binomial model in R.
> malform.linear <-
    glm(cbind(Present,Absent) ~ Alcohol,
        family=binomial(link=make.link("identity")),
        data=malform.wd)
> malform.linear <-
    glm(Present/Total ~ Alcohol, weights=Total,
        family=binomial(link=make.link("identity")),
        data=malform.wd)
> coef(malform.linear)
(Intercept)     Alcohol
  0.0025476   0.0010872
> summary(malform.linear)
Example (Infant Malformation (ctd))

Linear probability model for \( \pi = \Pr(\text{malform. pres.}) \) has ML fit

\[
\hat{\pi} = \hat{\alpha} + \hat{\beta} x = 0.0025 + 0.0011 x
\]

- At \( x = 0 \), \( \hat{\pi} = \hat{\alpha} = 0.0025 \).
- \( \hat{\pi} \) increases by \( \hat{\beta} = 0.0011 \) for each 1-unit increase in alcohol consumption.

Call:

\[
\text{glm(formula = Present/Total ~ Alcohol, family = binomial(link = \text{make.link("identity")}), data = malform.wd, weights = Total)}
\]

Deviance Residuals:

\[
\begin{array}{ccccc}
1 & 2 & 3 & 4 & 5 \\
0.656 & -1.049 & 0.863 & 0.130 & 0.828 \\
\end{array}
\]

Coefficients:

\[
\begin{array}{cccc}
\text{Estimate} & \text{Std. Error} & z \text{ value} & \Pr(>|z|) \\
\text{(Intercept)} & 0.002548 & 0.000352 & 7.23 & 4.8e-13 \\
\text{Alcohol} & 0.001087 & 0.000832 & 1.31 & 0.19 \\
\end{array}
\]

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 6.2020 on 4 degrees of freedom
Residual deviance: 2.9795 on 3 degrees of freedom
AIC: 25.61

Number of Fisher Scoring iterations: 10

Remarks

- ML estimates \( \hat{\alpha} \) and \( \hat{\beta} \) obtained by iterative numerical optimization.
- To test \( H_0 : \beta = 0 \) (independence), can use \( z = \frac{\hat{\beta} - 0}{\text{SE}(\hat{\beta})} \).
  - For large \( n \), has approx. std. normal dist. under \( H_0 \).
    - Ex: \( z = \frac{0.0011}{0.00083} = 1.31 \) For \( H_{\alpha : \beta \neq 0} \), \( p \text{-value} = 0.192 \)
- Could use Pearson \( X^2 \) (or \( G^2 \)) to test independence, but ignores ordering of rows.
- Alternative way to apply \( X^2 \) (or \( \text{deviance } G^2 \)) is to test fit of model: compares observed counts to values predicted by fitted model.
- Same fit results if we enter 5 binomial “success counts” or the 32574 individual binary responses of 0 (failure) or 1 (success).
- Problem: Model \( \pi(x) = \alpha + \beta x \) can give \( \hat{\pi} > 1 \) or \( \hat{\pi} < 0 \).
  - More realistic models take \( \pi(x) \) to be \textit{nonlinear} in \( x \).
Logistic Regression Model

\[ \log \left( \frac{\pi}{1 - \pi} \right) = \alpha + \beta x \]

is a GLM for binomial Y with logit link.

Example (Infant Malformation)

\[ \logit(\hat{\pi}) = \log \left( \frac{\hat{\pi}}{1 - \hat{\pi}} \right) = -5.96 + 0.32x \]

▶ \( \hat{\pi} \) ↑ as \( x \) ↑.

▶ p-value = 0.012 for \( H_0 : \beta = 0 \) vs \( H_a : \beta \neq 0 \).

▶ But p-value = 0.3 if delete single “present” obs. in \( \geq 6 \) drinks row!!

Remarks

▶ Chap. 4 studies logistic regression model.

▶ For contingency table, can test \( H_0 : \) “model correctly specified” with \( \chi^2 \) and \( G^2 \) test statistics using expected counts predicted by model.

▶ Ex: \( \chi^2 = 2.05, \ G^2 = 1.95 \) for \( H_0 : \) logistic model correct.

df = 3 = (5 binomial obs) − (2 parameters)

p-value large, no evidence against \( H_0 \).

▶ Both linear probability model and logistic regression model fit infant malformation data adequately. How is this possible?

logistic \( \approx \) linear when \( \hat{\pi} \) near 0 for all observed \( x \).

Ditto when \( \hat{\pi} \) near 1 for all observed \( x \).

GLMs for Count Data

When \( Y \) is a count \((0, 1, 2, 3, \ldots)\) usually assume a Poisson dist:

\[ P(y) = \frac{\mu^y e^{-\mu}}{y!}, \quad y = 0, 1, 2, \ldots \]

▶ \( \mu = E(Y) \)

▶ \( \text{Var}(Y) = \mu, \quad \sigma = \sqrt{\mu} \)

▶ In practice often \( \sigma^2 > \mu \), i.e., variation greater than predicted by Poisson (overdispersion).

▶ Negative binomial dist. has separate parameter for \( \sigma^2 \) and allows for overdispersion.
Poisson Regression for Count Data

Assume $Y$ has a Poisson distribution, $x$ an explanatory variable.

Model:

$$\mu = \alpha + \beta x$$  

identity link

or

$$\log(\mu) = \alpha + \beta x$$  

log link

Loglinear models use Poisson with log link (details in Ch. 7)

Example (Defects in Silicon Wafers)

$Y$ = number defects on silicon wafer
$x$ = dummy var. for treatment ($0 = A$, $1 = B$)

```r
A <- c(8,7,6,6,3,4,7,2,3,4)
B <- c(9,9,8,14,8,13,11,5,7,6)
trt <- factor(rep(c("A","B"), each=10))
wafers <- data.frame(trt=trt, defects=c(A,B))
wafers.lin <- glm(defects ~ trt, family=poisson(link="identity"), data=wafers)
wafers.loglin <- glm(defects ~ trt, family=poisson(link="log"), data=wafers)
```

> summary(wafers.lin)
> summary(wafers.loglin)
Example (Defects in Silicon Wafers (ctd))

For model $\mu = \alpha + \beta x$ (identity link)

$\hat{\mu} = 5.0 + 4.0x$

$x = 0: \quad \hat{\mu}_A = 5.0$ (= $y_A$)
$x = 1: \quad \hat{\mu}_B = 9.0$ (= $y_B$)

$\hat{\beta} = 4.0 = \hat{\mu}_B - \hat{\mu}_A$ has SE = 1.18 (use for test and CI for $\beta$)

For loglinear model $\log(\mu) = \alpha + \beta x$

$\log(\hat{\mu}) = 1.609 + 0.588x$

$x = 0: \quad \log \hat{\mu}_A = 1.609 \quad \hat{\mu}_A = e^{1.609} = 5.0$
$x = 1: \quad \log \hat{\mu}_B = 1.609 + 0.588 = 2.197 \quad \hat{\mu}_B = e^{2.197} = 9.0$
Inference for GLM Parameters

Wald Test

Test $H_0 : \beta = 0$

$z = \frac{\hat{\beta}}{SE}$ has approx. $N(0, 1)$ dist. under $H_0$.

For $H_a : \beta \neq 0$ can also use Wald stat. $z^2 = \left( \frac{\hat{\beta}}{SE} \right)^2$ approx. $\chi^2_1$.

For $H_0 : \beta = \beta_0$, use $z = \frac{\hat{\beta} - \beta_0}{SE}$.

CI = set of $\beta_0$ values for $H_0 : \beta = \beta_0$ s.t. $|\hat{\beta} - \beta_0|/SE < z_{\alpha/2}$, i.e.,

$\hat{\beta} \pm z_{\alpha/2} SE$

Likelihood Ratio Test

Test $H_0 : \beta = 0$ vs $H_a : \beta \neq 0$

$l_0 = \text{maximized likelihood when } \beta = 0$

$l_1 = \text{maximized likelihood for arbitrary } \beta$

test stat $= -2 \log \left( \frac{l_0}{l_1} \right)$

$= -2[\log l_0 - \log l_1]$

$= -2(L_0 - L_1)$

where $L = \text{maximized log-likelihood}$.

Example (Defects in Silicon Wafers)

Log-linear model: $\log(\mu) = \alpha + \beta x$.

$\beta = \log \mu_B - \log \mu_A$

$H_0 : \mu_A = \mu_B \iff \beta = 0$

Wald Test

$z = \frac{\hat{\beta}}{SE} = \frac{0.588}{0.176} = 3.33$  $p$-value $= 2 \times 0.00043 = 0.00086$

or

$z^2 = 11.1$  $df = 1$  $p$-value $= 0.00086$

Likelihood-Ratio Test

$L_1 = -45.17$  $L_0 = -50.97$

Test stat: $-2(L_0 - L_1) = 11.6$  $df = 1$  $p$-value $= 0.00066$
```r
> # anova(wafers.loglin, test="Chisq")
> drop1(wafers.loglin, test="Chisq")
```

**Single term deletions**

**Model:**
```
defects ~ trt
```

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
<th>LRT</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>16.3</td>
<td>94.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>trt</td>
<td>1</td>
<td>27.9</td>
<td>103.9</td>
<td>11.6</td>
</tr>
</tbody>
</table>

**Remarks**

- For very large $n$, Wald and LR tests are approx. equivalent, but for small to moderate $n$ the LR test is more reliable and powerful.
- LR statistic also equals difference in “deviances” (which are goodness-of-fit statistics).
  
  Ex: $27.857 - 16.268 = 11.589$
- LR method also extends to CIs: $(1 - \alpha) \times 100\%$ CI is set of $\beta_0$ for which $p$-value $> \alpha$ in LR test of $H_0: \beta = \beta_0$. Computed by `confint()` function in R.

**Example (Wafer Defects)**

\[
\beta = \log_2 \mu_B - \log_2 \mu_A = \log \left( \frac{\mu_B}{\mu_A} \right)
\]

\[
e^\beta = \frac{\mu_B}{\mu_A}
\]

\[
e^{\hat{\beta}} = e^{0.5878} = 1.8 = \frac{\hat{\mu}_B}{\hat{\mu}_A}
\]

95% CI for $\beta$: $0.588 \pm (1.96)(0.176) = (0.242, 0.933)$

95% CI for $e^{\beta}$: $(e^{0.242}, e^{0.933}) = (1.27, 2.54)$

We are 95% confident that $\mu_B$ is from 1.27 to 2.54 times as large as $\mu_A$. 

CI for $e^\beta = \mu_B/\mu_A$ based on LR test is $(e^{0.247}, e^{0.94}) = (1.28, 2.56)$. 

We are 95% confident that $\mu_B$ is from 1.27 to 2.54 times as large as $\mu_A$.
> wafCI.LR <- confint(wafers.loglin)
> wafCI.Wald <- confint.default(wafers.loglin)
> wafCI.LR
2.5 % 97.5 %
(Intercept) 1.31884 1.8744
trtB 0.24691 0.9401
> exp(wafCI.LR)
2.5 % 97.5 %
(Intercept) 3.7391 6.5168
trtB 1.2801 2.5602
> wafCI.Wald
2.5 % 97.5 %
(Intercept) 1.33226 1.88662
trtB 0.24208 0.93349

Deviance

The saturated model has a separate parameter for each observation
and fits the data perfectly: \( \hat{\mu}_i = y_i \).

For a model M with maximized log-likelihood \( L_M \)

\[
\text{deviance} = -2(L_M - L_S)
\]

where S is the saturated model

The deviance is the LR stat. for comparing model M to the saturated
model S, i.e., for

\[ H_0: \text{model M holds} \quad \text{vs} \quad H_a: \text{saturated model} \]

Tests that all parameters in S but not in M are equal to 0.

For binomial and Poisson models for counts

\[
\text{deviance} = G^2 = 2 \sum y_i \log \left( \frac{y_i}{\hat{\mu}_i} \right)
\]

where the \( \hat{\mu}_i \)s computed for M. (Sum is over success and failure counts
for binomial.)

When the \( \hat{\mu}_i \) are large and the number of predictor settings is fixed, \( G^2 \)
and Pearson’s chi-square statistic

\[
\chi^2 = \sum \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}
\]

are used to test goodness-of-fit of model (i.e., \( H_0: \text{model holds} \)).

Their distribution is approx. \( \chi^2 \) with

\[ df = \text{no. observations} - \text{no. model parameters} \]
Example (Wafer Defects)

\[ \hat{\mu}_i = 5 \text{ for 10 obs in trt A} \]

\[ \hat{\mu}_i = 9 \text{ for 10 obs in trt B} \]

For loglinear model, \( \log \mu = \alpha + \beta x \):

deviance \( G^2 = 16.3 \)

Pearson \( X^2 = 16.0 \)

df = 18

These values of \( G^2 \) and \( X^2 \) do not contradict \( H_0 \): "model holds" but be cautious about referring them to a chi-square dist. in situations like this. The usual theory requires

- \( \hat{\mu}_i \)s are all large, and
- fixed df as \( n \uparrow \) (happens with contingency tables).

Remarks

- For GLMs, can study lack-of-fit using residuals (later chapters).

- Count data often show overdispersion relative to a Poisson GLM.

  i.e., at fixed \( x \), sample variance > mean, whereas variance = mean in Poisson.

  Overdispersion may be caused by subject heterogeneity.

  Ex: \( Y = \) no. times attended religious services in past year.

  Suppose \( \mu = 25 \). Is \( \sigma^2 = 25 \) (\( \sigma = 5 \))?

Negative Binomial Regression

More flexible model for count data that has

\[ E(Y) = \mu \quad \text{Var}(Y) = \mu + D\mu^2 \]

where \( D \geq 0 \) is called a dispersion parameter.

As \( D \downarrow 0 \), neg. binom. \( \rightarrow \) Poisson.

(Can derive neg. binom. as a "gamma mixture of Poissons", where the Poisson mean varies according to a gamma dist.)

Negative binomial regression models can be fit using the VGAM package for R. Also the MASS package, and probably some others.
Within the past 12 months, how many people have you known personally that were victims of homicide?

```r
> homicide <-
data.frame(nvics=rep(0:6, 2),
  race=rep(c("Black","White"), each=7),
  freq=c(119,16,12,7,3,2,0,1070,60,14,4,0,0,1))
> xtabs(freq ~ race + nvics, data=homicide)

    nvics
    race 0 1 2 3 4 5 6
   Black 119 16 12 7 3 2 0
   White 1070 60 14 4 0 0 1

Black: \( n = 159, \ \bar{y} = 0.52, \ s^2 = 1.14 \)
White: \( n = 1149, \ \bar{y} = 0.09, \ s^2 = 0.16 \)

At these sample sizes, very unusual to see such large discrepancies between \( \bar{y} \) and \( s^2 \) if the samples drawn from Poisson distributions.
```

You can safely ignore this slide if you wish.

```r
> n <- with(homicide, tapply(freq, race, sum))
> ybar <- by(homicide, homicide$race,
  function(x) weighted.mean(x$nvics, x$freq))
> homicide$ybar <- rep(ybar, each=7)
> s2 <-
  by(homicide, homicide$race,
    function(x) weighted.mean((x$nvics - x$ybar)^2, x$freq))
> cbind(n, ybar, s2)

    n ybar s2
   Black 159 0.522013 1.14260
   White 1149 0.092254 0.15511
> NA
```

Model: \( \log(\mu) = \alpha + \beta x \)

```r
> homicide <-
transform(homicide, race = relevel(race, "White"))
> hom.poi <-
glm(nvics ~ race, data=homicide, weights=freq,
  family=poisson)
> library(MASS)
> hom.nb <-
glm.nb(nvics ~ race, data=homicide, weights=freq)
> summary(hom.poi)
> summary(hom.nb)
```
Call: glm(formula = nvics ~ race, family = poisson, data = homicide, weights = freq)

Deviance Residuals:
Min 1Q Median 3Q Max
-14.05 0.00 5.26 6.22 13.31

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.3832 0.0971 -24.5 <2e-16
raceBlack 1.7331 0.1466 11.8 <2e-16

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 962.80 on 10 degrees of freedom
Residual deviance: 844.71 on 9 degrees of freedom
AIC: 1122

Number of Fisher Scoring iterations: 6

Call:
glm.nb(formula = nvics ~ race, data = homicide, weights = freq, init.theta = 0.2023119205, link = log)

Deviance Residuals:
Min 1Q Median 3Q Max
-12.75 0.00 2.09 3.28 9.11

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.383 0.117 -20.33 < 2e-16
raceBlack 1.733 0.238 7.27 3.7e-13

(Dispersion parameter for Negative Binomial(0.2023) family taken to be 1)

Null deviance: 471.57 on 10 degrees of freedom
Residual deviance: 412.60 on 9 degrees of freedom
AIC: 1002

Number of Fisher Scoring iterations: 1

Theta: 0.2023
Std. Err.: 0.0409
2 x log-likelihood: -995.7980
For both Poisson and neg. binom. model

\[ \log \hat{\mu} = -2.38 + 1.73x \]
\[ e^{1.73} = 5.7 \]
\[ \frac{\hat{y}_B}{\hat{y}_W} = 0.522 \]

But, SE for \( \hat{\beta} = 1.73 \) is 0.147 for Poisson, 0.238 for neg. binom.

Wald 95% CI for \( \hat{\beta} \) is
\[ 1.73 \pm (1.96)(0.147) = (1.45, 2.02) \] for Poisson
\[ 1.73 \pm (1.96)(0.238) = (1.27, 2.20) \] for neg. binom.

Leads to 95% CI for \( e^\beta = \mu_B/\mu_W \) of
\[ (e^{1.45}, e^{2.02}) = (4.24, 7.54) \] for Poisson
\[ (e^{1.27}, e^{2.20}) = (3.54, 9.03) \] for neg. binom.

In accounting for overdispersion, neg. binom. model gives wider CIs.

LR CIs are \((e^{1.44}, e^{2.02}) = (4.23, 7.53)\) for Poisson
and \((e^{1.27}, e^{2.21}) = (3.58, 9.13)\) for neg. binom.

\[
\begin{align*}
> \text{confint.default(hom.poi)} \\
2.5 \% & 97.5 \% \\
(\text{Intercept}) & \text{-2.5736 -2.1928} \\
\text{raceBlack} & 1.4459 2.0204 \\
> \text{exp(confint.default(hom.poi))} \\
2.5 \% & 97.5 \% \\
(\text{Intercept}) & 0.076262 0.1116 \\
\text{raceBlack} & 4.245574 7.5414 \\
> \text{## confint.default(hom.nb)} \\
> \text{exp(confint.default(hom.nb))} \\
2.5 \% & 97.5 \% \\
(\text{Intercept}) & 0.07332 0.11608 \\
\text{raceBlack} & 3.54571 9.02988 \\
\end{align*}
\]
Remarks

- For negative binomial model, estimated value of $D$ is $\hat{D} = 4.94$ (SE = 1.00).

\[
\text{Var}(Y) = \hat{\mu} + \hat{D}\hat{\mu}^2 = \hat{\mu} + 4.94\hat{\mu}^2
\]

Strong evidence of overdispersion ($D \neq 0$).

- Note that glm.nb returns $\hat{\theta} = 1/\hat{D} = 0.2023$ (SE = 0.0409).

\[
\text{Var}(Y) = \hat{\mu} + \frac{\hat{\mu}^2}{\hat{\theta}} = \hat{\mu} + \frac{4.94\hat{\mu}^2}{0.2023}
\]

- Output degrees of freedom for deviance are wrong because we used weights=freq instead of a data frame with 159 + 1149 = 1308 rows. Fitted model unchanged.

```r
homicide2 <- homicide[rep(1:14, homicide$freq),-3]
> hom.poi2 <- glm(nvics ~ race, data=hom2, family=poisson)
```

Remarks

- When $Y$ is a count, overdispersion relative to Poisson is common. Safest strategy is to use neg. bin. model, or other method that allows for overdispersion (e.g., quasi-Poisson GLM).

- May also have zero-inflated counts (excess of zeros relative to Poisson distribution). VGAM package (and others) contains code for fitting ZIP (zero-inflated Poisson) and related models.

Models for Rates

When $y_i$ have different bases (e.g., number murders for cities with different pop. sizes) more relevant to model rate at which events occur.

Let $y = \text{count with base } t$. Sample rate is $\frac{y}{t}$.

\[
E\left(\frac{Y}{t}\right) = \frac{\mu}{t}
\]

Loglinear model $\log\left(\frac{\mu}{t}\right) = \alpha + \beta x$

i.e., $\log(\mu) - \log(t) = \alpha + \beta x$.

$log(t)$ is an offset.

See pp. 82–84 of text for discussion.
Example (British Train Accidents over Time)

Have collisions between trains and road vehicles become more prevalent over time?

Total number of train-km (in millions) varies from year to year.

Model annual rate of train-road collisions per million train-km with $t =$ annual no. of train-km and $x =$ no. of years since 1975.

```r
> data(traincollisions)
> trains.loglin <-
    glm(TrRd ~ I(Year-1975), offset = log(KM),
        family=poisson, data=traincollisions)
> summary(trains.loglin)
```

```
Call:
glm(formula = TrRd ~ I(Year - 1975), family = poisson, data = traincollisions,
    offset = log(KM))

Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.058   -0.783   -0.083    0.377    3.387

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.2114     0.1589  -26.50  <2e-16
I(Year - 1975) -0.0329     0.0108   -3.06  0.0022

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 47.376 on 28 degrees of freedom
Residual deviance: 37.853 on 27 degrees of freedom
AIC: 133.5

Number of Fisher Scoring iterations: 5
```

Example (British Train Accidents (ctd))

\[
\log\left(\frac{\hat{\mu}}{t}\right) = -4.21 - 0.0329x
\]

\[
\frac{\hat{\mu}}{t} = e^{-4.21-0.0329x} = e^{-4.21}(e^{-0.0329})^x = (0.0148)(0.968)^x
\]

- Rate estimated to decrease by 3.2% per yr from 1975 to 2003.
- Est. rate for 1975 ($x = 0$) is 0.0148 per million km (15 per billion).
- Est. rate for 2003 ($x = 28$) is 0.0059 per million km (6 per billion).
- Overdispersion? Try negative binomial. Similar fit w/ SEs and p-values slightly larger.

```r
> trains.nb <-
    glm.nb(TrRd ~ I(Year-1975) + offset(log(KM)),
        data=traincollisions)
```

```
Call:
glm.nb(formula = TrRd ~ I(Year - 1975) + offset(log(KM)),
    data = traincollisions)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
  -2.058   -0.783   -0.083    0.377    3.387

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.2114     0.1589  -26.50  <2e-16
I(Year - 1975) -0.0329     0.0108   -3.06  0.0022

(Dispersion parameter for negative binomial family taken to be 6.9334)

    Null deviance: 47.376 on 28 degrees of freedom
Residual deviance: 37.853 on 27 degrees of freedom
AIC: 119.5

Number of Fisher Scoring iterations: 10
```
4. Logistic Regression
Simple Logistic Regression

\[ Y = 0 \text{ or } 1 \]
\[ \pi = \Pr(Y = 1) \]
\[ \logit[\pi(x)] = \log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \alpha + \beta x \]

Uses “logit” link for binomial \( Y \). Equivalently,
\[ \pi(x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \]

where \( \exp(\alpha + \beta x) = e^{\alpha + \beta x} \).
Properties of Simple Logistic Regression I

- If $\beta > 0$, then $\pi(x)$ increases as $x$ increases.
  If $\beta < 0$, then $\pi(x)$ decreases as $x$ increases.

- If $\beta = 0$, then $\pi(x) = \frac{e^{\alpha}}{1 + e^{\alpha}}$ constant in $x$ (with $\pi > \frac{1}{2}$ if $\alpha > 0$).

- Curve can be approximated at a fixed point $x$ by a straight line describing rate of change in $\pi(x)$. Slope is $\beta \pi(x) \left[1 - \pi(x)\right]$. E.g.,
  - at $x$ with $\pi(x) = \frac{1}{2}$, slope $= \beta \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{\beta}{4}$
  - at $x$ with $\pi(x) = 0.1$ or $0.9$, slope $= \beta (0.1)(0.9) = 0.09 \beta$
  - Steepest slope where $\pi(x) = \frac{1}{2}$

Properties of Simple Logistic Regression II

- If $\pi(x) = \frac{1}{2}$ then
  \[
  \log \left( \frac{\pi(x)}{1 - \pi(x)} \right) = \log \left( \frac{0.5}{0.5} \right) = \log(1) = 0 = \alpha + \beta x \implies x = -\frac{\alpha}{\beta}
  \]
  \[
  \frac{1}{|\beta|} \approx \text{dist. between } x \text{ values with } \pi = 0.5 \text{ and } \pi = 0.75 \text{ (or 0.25)}
  \]
  - ML fit obtained with iterative numerical methods.

Example (Horseshoe Crabs)

- $Y = \begin{cases} 1 & \text{if female crab has satellites} \\ 0 & \text{if no satellites} \end{cases}$
- $x = \text{weight (kg)} \quad (\bar{x} = 2.44, \ s = 0.58)$
- $n = 173$
- ML fit: $\text{logit}[\hat{\pi}(x)] = -3.69 + 1.82x$
- i.e., $\hat{\pi}(x) = \frac{\exp(-3.69 + 1.82x)}{1 + \exp(-3.69 + 1.82x)}$

```r
> data(horseshoecrabs)
> crabs.logit <- glm((Satellites > 0) ~ Weight, family=binomial, 
data=horseshoecrabs)
> summary(crabs.logit)
```
Call:
  glm(formula = (Satellites > 0) ~ Weight, family = binomial, data = horseshoecrabs)

Deviance Residuals:
            Min       1Q     Median       3Q      Max
-2.111   -1.075     0.543     0.912    1.629

Coefficients:
             Estimate Std. Error    z value  Pr(>|z|)
(Intercept)  -3.695   0.8803  -4.2025   2.7e-05
Weight       1.815   0.3771   4.8174   1.4e-06

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 225.76 on 172 degrees of freedom
Residual deviance: 195.74 on 171 degrees of freedom
AIC: 199.7

Number of Fisher Scoring iterations: 4

xbar <- mean(horseshoecrabs$Weight)
xbar
[1] 2.4372
sd(horseshoecrabs$Weight)
[1] 0.57703
predict(crabs.logit, data.frame(Weight=xbar), type="link")
1
0.72913
predict(crabs.logit, data.frame(Weight=xbar), type="response")
1
0.67461

► $\hat{\beta} > 0$, so $\hat{\pi}$ ↑ as $x$ ↑

► At $x = \bar{x} = 2.44$,

$$\hat{\pi} = \frac{\exp(-3.69 + 1.82(2.44))}{1 + \exp(-3.69 + 1.82(2.44))} = \frac{e^{0.729}}{1 + e^{0.729}} = \frac{2.07}{3.07} \approx 0.675$$

► $\hat{\pi} = \frac{1}{1 + e^{\hat{\beta}}}$ when $x = -\frac{\hat{\alpha}}{\hat{\beta}} = 3.69 \approx 1.82 = 2.04$

► At $x = 2.04$, estimated slope is $\hat{\beta}\hat{\pi}(1 - \hat{\pi}) = \frac{\hat{\beta}}{\hat{\pi}} = 0.45$. If $x$ increases by 1 kg, then $\hat{\pi}$ increases by $\approx 0.454$.

HOWEVER, $s = 0.58$ for weight, and 1-kg change too large for approximation to be good (actual $\hat{\pi} = 0.86$ at $x = 3.04$).

if $x$ increases by 0.1 kg, then $\hat{\pi}$ increases by $\approx 0.045$ (actual $\hat{\pi} = 0.545$ at $x = 2.14$).
At $x = 5.2$ (max. obs. wt.), $\hat{\pi} = 0.997$, and est. slope is $(1.82)(0.997)(0.003) = 0.0058$.

If $x$ increases by 0.1 kg, then $\hat{\pi}$ increases by $\approx (0.1)(0.0058) = 0.00058$.

- Rate of change varies with $x$.

```r
> attach(horseshoecrabs)
> plot(Weight, (Satellites > 0), xlim=c(0,6), ylim=c(0,1), xlab="Weight", ylab="Has Satellites")
> curve(plogis(-3.69+1.82*x), add=TRUE)
> detach(horseshoecrabs)
```

Remarks

- Fitting linear probability model (binomial w/ identity link) fails in the crabs example.
- If we assume $Y \sim$ Normal and fit linear model $\mu = \alpha + \beta x$,
  
  $\hat{\mu} = -0.415 + 0.323x$

  At $x = 5.2$, $\hat{\mu} = 1.53$ !!! (estimated prob. of satellites)

- Alternative way to describe effect (not dependent on units) is, e.g.,

  $\hat{\pi}(UQ) - \hat{\pi}(LQ)$

Example (Horseshoe Crabs)

For $x =$ weight, $LQ = 2.00$, $UQ = 2.85$.
At $x = 2.00$, $\hat{\pi} = 0.483$; at $x = 2.85$, $\hat{\pi} = 0.814$.
$\hat{\pi}$ increases by 0.33 over middle half of $x$ values.
Odds Ratio Interpretation

Since $\log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta x$, odds are

$$\frac{\pi}{1-\pi} = \begin{cases} e^{\alpha+\beta x} & \text{at } x \\ e^{\alpha+\beta(x+1)} = e^\beta e^{\alpha+\beta x} & \text{at } x + 1 \end{cases}$$

$$\Rightarrow \frac{\text{odds at } (x+1)}{\text{odds at } x} = \frac{e^\beta e^{\alpha+\beta x}}{e^{\alpha+\beta x}} = e^\beta$$

More generally,

$$\frac{\text{odds at } (x + \Delta x)}{\text{odds at } x} = \frac{e^{\alpha+\beta(x+\Delta x)}}{e^{\alpha+\beta x}} = e^{\beta\Delta x}$$

If $\beta = 0$, then $e^\beta = 1$ and odds do not depend on $x$.

Example (Horseshoe Crabs)

$\hat{\beta} = 1.82 \implies e^{\hat{\beta}} = e^{1.82} = 6.1$

Estimated odds of having at least one satellite increase by a factor of 6.1 for each 1 kg increase in weight.

If weight increases by 0.1 kg, then estimated odds increase by factor

$$e^{(1.82)(0.1)} = e^{0.18} = 1.2$$

Inference for Simple Logistic Regression

Confidence Interval

Wald $(1 - \alpha)100\%$ CI for $\beta$ is $\hat{\beta} \pm \frac{z_{\alpha/2}}{SE}$

Example (Horseshoe Crabs)

95% CI for $\beta$:

$$1.82 \pm (1.96)(0.377) = 1.82 \pm 0.74 = (1.08, 2.55)$$

95% CI for $e^\beta$, multiplicative effect on odds of 1-unit increase in $x$ is

$$\left(e^{1.08}, e^{2.55}\right) = (2.9, 12.9)$$

95% CI for $e^{0.1\beta}$ is

$$\left(e^{0.108}, e^{0.255}\right) = (1.11, 1.29)$$

Odds estimated to increase at least 11%, at most 29%.
Remarks
▶ Safer to use LR CI than Wald CI.
   For crabs example, LR CI for $e^\beta$ is
   $$(e^{1.11}, e^{2.60}) = (3.0, 13.4)$$
▶ Can also construct CI for $\pi(x)$. The convenience function
   `predCI()` in the `icda` does the calculation described in
   Section 4.2.6 of the text.
   For crabs data, at $x = 3.05$ (first crab), $\hat{\pi} = 0.863$.
   A 95% CI for $\pi$ at $x = 3.05$ is
   $$(0.766, 0.924)$$

```r
> crabs.predCI <- predCI(crabs.logit)
> crabs.predCI[,]

fit  lwr  upr
0.86312 0.76606 0.92391

> predCI(crabs.logit, newdata=data.frame(Weight=2.44))

fit  lwr  upr
1 0.67573 0.59321 0.74861
```

Inference for Simple Logistic Regression
Significance Tests for $\beta$ (1)

$H_0 : \beta = 0$ states that $Y$ indep. of $X$ (i.e., $\pi(x)$ constant in $x$)

$H_a : \beta \neq 0$

Wald Test

$$z = \frac{\hat{\beta}}{SE} = \frac{1.815}{0.377} = 4.82$$

or $z^2 = 23.2$, df = 1 (chi-squared).

p-value < 0.0001: very strong evidence that $\pi$ increases with weight.
Inference for Simple Logistic Regression
Significance Tests for \( \beta \) (2)

**Likelihood ratio test**

When \( \beta = 0 \), \( L_0 = -112.88 \) (log-likelihood under \( H_0 \))

When \( \beta = \hat{\beta} \), \( L_1 = -97.87 \)

Test statistic: \(-2(L_0 - L_1) = 30.02\), df = 1 (chi-sq), p-value < 0.0001

```r
> drop1(crabs.logit, test="Chisq")
Single term deletions
Model: (Satellites > 0) ~ Weight
Df Deviance AIC LRT Pr(Chi)
<none> 196 200
Weight 1 226 228 30 4.3e-08
```

**Remark**

Recall for a model \( M \),

\[
\text{deviance} = -2(L_M - L_S)
\]

\( L_S \) is log-likelihood under saturated model (perfect fit).

To compare model \( M_0 \) with more complex model \( M_1 \),

\[
\text{LR statistic} = -2(L_0 - L_1) = -2(L_0 - L_S) - (L_1 - L_S)
\]

\[
= \text{diff. of deviances}
\]

**Example (Horseshoe Crabs)**

Model: \( \logit[\pi(x)] = \alpha + \beta x \) (this is \( M_1 \))

\( H_0 : \beta = 0 \implies \logit[\pi(x)] = \alpha \) (this is \( M_0 \))

\[
\text{diff. of deviances} = 225.76 - 195.74 = 30.02 = \text{LR stat.}
\]
Multiple Logistic Regression

\(Y\) binary, \(\pi = \Pr(Y = 1)\)

\(x_1, x_2, \ldots, x_k\) can be quantitative, qualitative (dummy variables), or both.

Model form is

\[
\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k
\]

or equivalently

\[
\pi = \frac{\exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k)}
\]

\(\beta_i\) = partial effect of \(x_i\) controlling for other variables in model

\(e^{\beta_i}\) = cond. odds ratio at \(x_i + 1\) vs at \(x_i\) keeping other \(x\)'s fixed

= multi. effect on odds of 1-unit incr. in \(x_i\), w/ other \(x\)'s fixed

Example (Horseshoe Crabs)

\(Y = \)
\[
\begin{cases} 
1 & \text{sampled female has at least 1 satellite} \\
0 & \text{sampled female has no satellites} 
\end{cases}
\]

\(x = \) Weight

\(c = \) Color (qualitative w/ 4 categories)

\[
\begin{align*}
  c_2 &= \begin{cases} 
    1 & \text{medium} \\
    0 & \text{o/w}
  \end{cases} \\
  c_3 &= \begin{cases} 
    1 & \text{dark med} \\
    0 & \text{o/w}
  \end{cases} \\
  c_4 &= \begin{cases} 
    1 & \text{dark} \\
    0 & \text{o/w}
  \end{cases}
\end{align*}
\]

For "light medium" crabs, \(c_2 = c_3 = c_4 = 0\).

Original data set had color coded 1–4 for "light med", "medium", "dark med", and "dark". R interprets this as a numeric variable, so we must convert it to factor.

\[
> \text{horseshoecrabs} \leftarrow \text{transform(horseshoecrabs, C = as.factor(Color))}
\]
\[
> \text{levels(horseshoecrabs$C)}
\]
\[
[1] \ "1" \ "2" \ "3" \ "4"
\]
\[
> \text{crabs.fit1 <- glm(Satellites > 0 ~ C + Weight, family=binomial, data=horseshoecrabs)}
\]
\[
> \text{summary(crabs.fit1)}
\]
Call:

\[
\text{glm(formula = (Satellites > 0) ~ C + Weight, family = binomial, data = horseshoecrabs)}
\]

Deviance Residuals:

\[
\begin{array}{cccc}
\text{Min} & \text{IQR} & \text{Median} & \text{3Q} & \text{Max} \\
-2.191 & -1.014 & 0.510 & 0.868 & 2.075
\end{array}
\]

Coefficients:
Example (Horseshoecrabs (ctd))

Model:
\[
\logit \left[ \Pr(Y = 1) \right] = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x
\]

has ML fit
\[
\logit(\hat{\pi}) = -3.26 + 0.14c_2 - 0.19c_3 - 1.27c_4 + 1.69x
\]

For light med. female \((c_2 = c_3 = c_4 = 0)\),
\[
\logit(\hat{\pi}) = -3.26 + 1.69x
\]

At \(x = x = 2.44\),
\[
\hat{\pi} = \frac{e^{-3.26 + (1.69)(2.44)}}{1 + e^{-3.26 + (1.69)(2.44)}} = 0.71
\]

Example (Horseshoecrabs (ctd))

For medium female \((c_2 = 1, c_3 = c_4 = 0)\),
\[
\logit(\hat{\pi}) = -3.26 + (0.14)(1) + 1.69x = -3.11 + 1.69x
\]

At \(x = x = 2.44\), \(\hat{\pi} = 0.73\).

At each weight, estimate medium females more likely than light med. to have satellites:

\[
\hat{\beta}_2 = 0.145 \implies e^{0.145} = 1.16
\]

Estimated odds a medium female has satellites are 1.16 times estimated odds for a light med. female of the same weight.

E.g., at \(x = 2.44\),
\[
\frac{\text{odds for medium}}{\text{odds for light-med}} = \frac{0.73/0.27}{0.71/0.29} = 1.16
\]
Example (Horseshoecrabs (ctd))

- How do we compare, e.g., dark \((c_2 = c_3 = 0, c_4 = 1)\) to medium \((c_2 = 1, c_3 = c_4 = 0)\)?

\[-1.27 - 0.14 = -1.41 \quad e^{-1.41} = 0.24\]

Estimated odds a dark crab has satellites are 0.24 times estimated odds a medium crab of same weight has satellites.

Equivalently,

\[0.14 - (-1.27) = 1.41 \quad e^{1.41} = 4.11 \quad (= 1/0.24)\]

Estimated odds a medium crab has satellites are 4.11 times estimated odds a dark crab of same weight has satellites.

Example (Horseshoecrabs (ctd))

- Model assumes no interaction between color and weight effects.

Coef. of \(x = \text{Weight}\) is same for each color \((\hat{\beta} = 1.69)\).

For fixed color, estimated odds of satellites at weight \((x + 1)\) is \(e^{1.69} = 5.4\) times estimated odds at weight \(x\).

Curves have same shape across colors, but shifted left or right.
Example (Horseshoe Crabs (ctd))

Do we need color in the model?

\[ H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \] (given weight, \( Y \) independent of color)

Likelihood-ratio statistic

\[ -2(L_0 - L_1) = -2\left[(-97.9) - (-94.3)\right] = 7.19 \]

or

\[ \text{diff. of deviances} = 195.7 - 188.54 = 7.19 \]

\[ \text{df} = 171 - 168 = 3 \quad \text{p-value} = 0.066 \]

Some evidence (not strong) of a color effect given weight.

There is strong evidence of weight effect (\( \hat{\beta} = 1.69 \) has SE = 0.39).

\[
\text{logit}(\hat{\pi}) = \begin{cases} 
-3.26 + 1.69x, & \text{med-light} \\
-3.11 + 1.69x, & \text{med} \\
-3.44 + 1.69x, & \text{med-dark} \\
-4.53 + 1.69x, & \text{dark} 
\end{cases}
\]

suggests

\[
\text{logit}(\pi) = \alpha + \beta_1z + \beta_2x, \quad z = \begin{cases} 
1, & \text{dark} \\
0, & \text{o/w} 
\end{cases}
\]

Fitting gives \( \beta_1 = -1.295 \) (SE = 0.522).

Estimated odds of satellites for a dark crab is \( e^{-1.295} = 0.27 \) times estimated odds a non-dark crab of the same weight.
> crabs.fit2 <-
glm((Satellites > 0) ~ I(Color == 4) + Weight,
    family=binomial, data=horseshoe.crabs)
> summary(crabs.fit2)
Call:
glm(formula = (Satellites > 0) ~ I(Color == 4) + Weight, family = binomial,
    data = horseshoe.crabs)
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.155   -1.023    0.513    0.848    2.087
Coefficients:        Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.313      0.898  -3.69  0.00023
I(Color == 4)TRUE  -1.295      0.522  -2.48  0.01311
Weight               1.729      0.383   4.52  6.2e-06

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 225.76 on 172 degrees of freedom
Residual deviance: 189.17 on 170 degrees of freedom
AIC: 195.2

Number of Fisher Scoring iterations: 4

> anova(crabs.fit2, crabs.fit1, test="Chisq")
Analysis of Deviance Table
Model 1: (Satellites > 0) ~ I(Color == 4) + Weight
Model 2: (Satellites > 0) ~ C + Weight
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1        170    189
2        168    188   2   0.629    0.73

Example (Horseshoe Crabs (ctd))

Compare model with 1 dummy for color to full model with 3 dummies.

H<sub>0</sub>: simple model vs H<sub>a</sub>: more complex model

Note H<sub>0</sub> is β<sub>2</sub> = β<sub>3</sub> = 0 in more complex model.

LR stat = diff. in deviances = 189.17 − 188.54 = 0.63
df = 170 − 168 = 2 p-value = 0.73

Simpler model is adequate.
Example (Horseshoe Crabs (ctd))

How about interaction?

\[
\logit(\pi) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x + \gamma_2 c_2 x + \gamma_3 c_3 x + \gamma_4 c_4 x
\]

<table>
<thead>
<tr>
<th>Color</th>
<th>Dummies</th>
<th>Weight Coef</th>
</tr>
</thead>
<tbody>
<tr>
<td>light-med</td>
<td>(c_2 = c_3 = c_4 = 0)</td>
<td>(\beta)</td>
</tr>
<tr>
<td>medium</td>
<td>(c_2 = 1, c_3 = c_4 = 0)</td>
<td>(\beta + \gamma_2)</td>
</tr>
<tr>
<td>dark-med</td>
<td>(c_3 = 1, c_2 = c_4 = 0)</td>
<td>(\beta + \gamma_3)</td>
</tr>
<tr>
<td>dark</td>
<td>(c_4 = 1, c_2 = c_3 = 0)</td>
<td>(\beta + \gamma_4)</td>
</tr>
</tbody>
</table>

Testing \(H_0\): no interaction (\(\gamma_2 = \gamma_3 = \gamma_4 = 0\))

\[
\text{LR stat} = 188.54 - 181.66 = 6.89 \quad \text{df} = 3 \quad \text{p-value} = 0.076
\]

Weak evidence of interaction.

For easier interpretation, use simpler model (no interaction).

```r
> crabs.fit3 <- update(crabs.fit1, . ~ C*Weight)
> deviance(crabs.fit3)
[1] 181.66
> anova(crabs.fit1, crabs.fit3, test="Chisq")
Analysis of Deviance Table

Model 1: (Satellites > 0) ~ C + Weight
Model 2: (Satellites > 0) ~ C + Weight + C:Weight

| Resid. Df | Resid. Dev | Df | Deviance | P(>|Chi|) |
|-----------|------------|----|----------|----------|
| 1          | 188        |    |          |          |
| 2          | 182        | 3  | 6.89     | 0.076    |

> drop1(crabs.fit3, test="Chisq")

Single term deletions

Model:
(Satellites > 0) ~ C + Weight + C:Weight

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
<th>LRT</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>182 198</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>C:Weight</td>
<td>3 188 198 6.89</td>
<td>0.076</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Logistic Regression

Ordinal Factors

Models with dummy variables treat color as qualitative (nominal).

To treat as quantitative, assign scores such as $(1, 2, 3, 4)$ and model trend.

\[
\text{logit}(\pi) = \alpha + \beta_1 x_1 + \beta_2 x_2, \quad x_1: \text{weight}, \quad x_2: \text{color}
\]

ML estimates and SEs are

\[
\hat{\alpha} = -2.03 (1.12) \quad \hat{\beta}_1 = 1.65 (0.38) \quad \hat{\beta}_2 = -0.51 (0.22)
\]

\[
\text{logit}(\pi) = -2.03 + 1.65x_1 - 0.51x_2
\]

\(\hat{\pi}\) ↓ as Color ↑, controlling for weight.

Controlling for weight, odds of having at least one satellite estimated to decrease by a factor of

\[e^{-0.51} = 0.60\]

for each 1-category increase in shell darkness

```r
> crabs.fit4 <-
glm((Satellites > 0) ~ Weight + Color,
      family=binomial, data=horseshoecrabs)
> summary(crabs.fit4)
Call:
  glm(formula = (Satellites > 0) ~ Weight + Color,
      family = binomial, data = horseshoecrabs)
Deviance Residuals:
     Min       1Q   Median       3Q      Max
-2.160   -1.000    0.524    0.882    1.911

Coefficients:        Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.032     1.116   -1.82   0.069
Weight          1.653     0.382    4.32  1.5e-05
Color          -0.514     0.223   -2.30   0.021

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 190.27 on 170  degrees of freedom
AIC: 196.3

Number of Fisher Scoring iterations: 4

> anova(crabs.fit4, crabs.fit1, test="Chisq")
Analysis of Deviance Table

Model 1: (Satellites > 0) ~ Weight + Color
Model 2: (Satellites > 0) ~ C + Weight

Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1    170      190
2    168      188   2    1.73    0.42
```
Does model treating color as nominal fit as well as model treating it as qualitative?

\[ H_0 : \text{simpler (ordinal) model holds} \]

\[ H_a : \text{more complex (nominal) model holds} \]

\[
\text{LR stat} = -2(L_0 - L_1) \\
= \text{diff in deviances} \\
= 190.27 - 188.54 = 1.73, \quad \text{df} = 2
\]

Do not reject \( H_0 \). Simpler model appears to be adequate.

---

**Qualitative Predictors: FL Death Penalty Revisited I**

```r
> dpflat
DeathPenalty Yes No
Victim Defendant
White White 53 414
Black 11 37
Black White 0 16
Black 4 139
```

Modeling approach: take death penalty (Yes/No) as response, race of defendant and race of victim as explanatory variables.

---

**Qualitative Predictors: FL Death Penalty Revisited II**

```r
> deathpenalty
DeathPenalty Defendant Victim Freq
1 Yes White White 53
2 No White White 414
3 Yes Black White 11
4 No Black White 37
5 Yes White Black 0
6 No White Black 16
7 Yes Black Black 4
8 No Black Black 139
```
Qualitative Predictors: FL Death Penalty Revisited III

```r
> library(reshape2)
> dp <- melt(deathpenalty)
> dpwide <- dcast(dp, Defendant ~ Victim)
> dpwide

Defendant Victim variable Yes No
1 White White Freq 53 414
2 White Black Freq 0 16
3 Black White Freq 11 37
4 Black Black Freq 4 139
```

```r
> dep.fit1 <-
  glm(cbind(Yes, No) ~ Defendant + Victim, family=binomial, data=dpwide)
> summary(dep.fit1)
```

Qualitative Predictors: FL Death Penalty Revisited IV

```
Call:
  glm(formula = cbind(Yes, No) ~ Defendant + Victim, family = binomial, data = dpwide)
Deviance Residuals:
  1     2     3     4
  0.0266 -0.6054 -0.0623  0.0938
Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)              -2.059     0.146  -14.12  < 2e-16
DefendantBlack           0.868     0.367   2.36  0.0184  
VictimBlack              -2.404     0.601  -4.00  6.2e-05

(Dispersion parameter for binomial family taken to be 1)
Null deviance: 22.26591 on 3 degrees of freedom
Residual deviance: 0.37984 on 1 degrees of freedom
```

Qualitative Predictors: FL Death Penalty Revisited V

```
AIC: 19.3
Number of Fisher Scoring iterations: 4
> drop1(dep.fit1, test="Chisq")
Single term deletions

Model:
cbind(Yes, No) ~ Defendant + Victim
 Df Deviance AIC LRT Pr(Chi)
<none> 0.38 19.3
Defendant 1 5.39 22.3 5.01 0.025
Victim 1 20.73 37.6 20.35 6.4e-06
```
Qualitative Predictors: FL Death Penalty Revisited VI

\[ \pi = \Pr(Y = \text{yes}) \] death penalty

\[ \nu = \begin{cases} 
1, & \text{victim black} \\
0, & \text{victim white}
\end{cases} \]

\[ d = \begin{cases} 
1, & \text{defendant black} \\
0, & \text{defendant white}
\end{cases} \]

Model:

\[ \logit(\pi) = \alpha + \beta_1 d + \beta_2 \nu \]

ML fit:

\[ \logit(\hat{\pi}) = -2.06 + 0.87 d - 2.40 \nu \]

Controlling for race of victim, estimated odds of death penalty for black def is \( e^{0.87} = 2.38 \) times estimated odd for white def.

95% CI for odds-ratio is \( e^{0.868 \pm 1.96(0.367)} = (e^{0.148}, e^{1.59}) = (1.16, 4.89) \)

Qualitative Predictors: FL Death Penalty Revisited VII

Remarks

▶ No interaction term means estimated odds ratio between \( Y \) and 
▶ \( d \) same at each level of \( \nu \) (\( e^{2.40} = 11.1 \))
▶ \( \nu \) same at each level of \( d \) (\( e^{-2.40} = 0.09 \))

Homogeneous association: same odds ratio at each level of other variable.

▶ Test \( H_0 : \beta_1 = 0 \) (Y cond. indep. of \( d \) given \( \nu \)) vs \( H_a : \beta_1 \neq 0 \)

\[ z = \frac{\hat{\beta}}{SE} = \frac{0.868}{0.367} = 2.36 \quad p\text{-value} = 0.018 \]

Evidence that controlling for race of victim, death penalty more likely for black defendants than white.

\[ \text{LR stat} = 5.39 - 0.38 = 5.01 \quad df = 1 \quad p\text{-value} = 0.025 \]

A common application for logistic regression on multiple 2 × 2 tables is multi-center clinical trials:

<table>
<thead>
<tr>
<th>Center</th>
<th>Treatment</th>
<th>Response</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>S</td>
<td>F</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>K</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

\[ \logit[\Pr(Y = 1)] = \alpha + \beta_2 c_2 + \cdots + \beta_K c_K + \beta x \]

Assumes odds ratio \( e^\beta \) is the same for each center.
A model like this is commonly expressed in the form

\[
\text{logit}[\Pr(Y = 1)] = \alpha + \beta_c^i + \beta x
\]

\(\beta_c^i\) is effect for center \(i\) (relative to first center).

To test \(H_0 : \beta = 0\) (no treatment effect) for several \(2 \times 2\) tables, could use

- likelihood-ratio test
- Wald test
- Cochran-Mantel-Haenszel test (p. 114)
- generalization of Fisher’s exact test (pp. 158–159) (useful for small samples)

---

Exam 1: Time and Place

Thursday, Feb 24, 2011
8:20 p.m. – 10:10 p.m.
Griffin-Floyd Hall (FLO)
Room 100

Exam 1 Review: Binomial Distribution

- Recognize.
- Compute probs, mean, sd.
- Wald and score tests and CIs for a single proportion.
Exam 1 Review: Likelihood

▶ What is the likelihood function?
▶ What is MLE?

Exam 1 Review: Contingency Tables

▶ Joint, marginal, and conditional distributions.
▶ INDEPENDENCE.
▶ Sensitivity/specificity.
▶ Probability and ODDS.

Exam 1 Review: 2 × 2 Tables

▶ Measures of dependence:
  ▶ Diff in proportions: \( \pi_1 - \pi_2 \)
  ▶ Relative risk: \( \frac{\pi_1}{\pi_2} \)
  ▶ Odds ratio: \( \theta = \frac{\pi_1/(1 - \pi_1)}{\pi_2/(1 - \pi_2)} \)
    ▶ When are odds ratio and relative risk similar?
    ▶ Retrospective study: can estimate \( \theta \), not others.
▶ CIs for \( \pi_1 - \pi_2 \) and \( \theta \).
Exam 1 Review: Testing Independence in $I \times J$ Tables I

- Estimated expected frequencies.
- Pearson’s chi-squared statistics: $X^2$
- Likelihood-ratio statistic: $G^2$
- $df = (I-1)(J-1)$
- Chi-square dist. has $\mu = df$ and $\sigma = \sqrt{2df}$.

Exam 1 Review: Testing Independence in $I \times J$ Tables II

- Examining sources of dependence
  - Adjusted residuals (analyze dependence)
    - Should be approx. $N(0, 1)$ if all expected freqs $\geq 5$.
    - If so, then $|\text{adj resid}| \geq 2$ (or 3 in big tables) meaningful.
  - Partitioning chi-square
    - With a correct partition
      - $G^2$ stats add up ($X^2$ approximately so)
      - df’s add up to total df

Exam 1 Review: Fisher’s Exact Test

- Test of independence ($\theta = 1$) in $2 \times 2$ tables.
- Conditions on (holds fixed) row and column totals.
- Dist. of $n_{11}$ under independence is hypergeometric.
  - Expected value is $\frac{n_{1+}n_{+1}}{n}$
  - Large values of $n_{11}$ suggest $\theta > 1$; small values suggest $\theta < 1$.
  - Can be extended to $I \times J$ tables.
Exam 1 Review: Three-Way Contingency Tables

- Three variables: X, Y, Z
- Partial tables
  - Hold Z fixed
- Conditional odds ratios
- Simpson’s paradox
- Conditional independence: X and Y indep. in each partial table.

Exam 1 Review: Generalized Linear Models I

- Random component: form of distribution for Y
- Systematic component
  \[ g(\mu) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k \]
- Common link functions
  - Identity: \( g(\mu) = \mu \)
  - Log: \( g(\mu) = \log(\mu) \)
  - Logit: \( g(\mu) = \log\left(\frac{\mu}{1 - \mu}\right) \)

Exam 1 Review: Generalized Linear Models II

- Compute MLEs with iterative numerical algorithm.
- Test hypotheses about parameters using Wald or LR tests.
- CIs also based on Wald or LR tests.
- Special case: ordinary linear regression
  - random component: Y is normally distributed
  - link: identity link
Exam 1 Review: GLMs for Binary Data I

- Ordinary linear model inappropriate for binary data
  - Binary response not normally distributed
  - \( \text{Var}(Y) \) depends on \( \pi(x) \) so least squares not optimal
  - Identity link may give estimated probabilities that are negative or greater than one.

- Linear probability model
  - Binomial random component with identity link
  - Advantage of identity link: easy interpretation of \( \beta \)
  - Disadvantage of identity link: may give estimated probabilities that are negative or greater than one

Exam 1 Review: GLMs for Binary Data II

- Logistic regression model
  - Binomial random component with logit link
  - Logit link respects bounds on probabilities: must be between 0 and 1
  - Interpret \( \beta \) in terms of odds and odds ratios.

Exam 1 Review: GLMs for Count Data I

- Poisson log-linear model
  - Random component: Poisson distribution
  - Link: log
  - In simple Poisson log-linear regression model
  \[
  \log \mu = \alpha + \beta x
  \]
  the mean is multiplied by a factor of \( e^\beta \) for each 1-unit increase in \( x \).
Exam 1 Review: GLMs for Count Data II

▶ Often have different bases for counts: need to model rate.
With log link, this leads to an offset.

If \( t \) is base for count \( Y \), systematic component is

\[
\log \left( \frac{Y}{t} \right) = \alpha + \beta x \quad \implies \quad \log(Y) = \log(t) + \alpha + \beta x
\]

▶ Overdispersion is common with count data
Poisson random component has \( \text{Var}(Y) = \mu \).
Often have \( \text{Var}(Y) > \mu \) due to subject heterogeneity or other source(s) of unexplained variation.

▶ One way to address overdispersion: use negative binomial distribution as random component instead of Poisson.

Exam 1 Review: Simple Logistic Regression Model

▶ Binomial GLM with logit link and a single numerical explan. variable

\[
\log \left( \frac{\pi}{1 - \pi} \right) = \alpha + \beta x \quad \text{i.e.} \quad \pi = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}}
\]

\( \pi \) = prob. of "success"  \( \frac{\pi}{1 - \pi} \) = odds of success

▶ Odds of success multiplied by \( e^\beta \) for each 1-unit increase in \( x \).
Multiplied by \( e^{\beta \Delta x} \) if \( x \) changed by amount \( \Delta x \).

▶ \( \pi = 1/2 \) when \( x = -\alpha/\beta \).

▶ Rate of change (slope) of \( \pi \) at a fixed point \( x \) is \( \beta \pi(x) \left[ 1 - \pi(x) \right] \).
Steepest at \( x = -\alpha/\beta \) where \( \pi = 1/2 \) and slope = \( \beta/4 \).
Flattest when \( \pi(x) \) close to 0 or 1.

\[
\frac{1}{|\beta|} \approx \text{dist. between } x \text{ values with } \pi = 0.5 \text{ and } \pi = 0.75 \text{ (or 0.25)}
\]

Exam 1 Review: Inference in Simple Logistic Regression I

▶ Inference about \( \beta \) using Wald and LR tests and CIs.
LR methods preferred.

▶ Wald test and CI have usual form:

\[
\text{Test stat:} \quad z = \frac{\hat{\beta} - \beta_0}{\text{SE}}
\]

\[
\text{CI:} \quad \hat{\beta} \pm z_{\alpha/2} \times \text{SE}
\]

▶ CI for \( e^\beta \): first compute CI \((L, U)\) for \( \beta \), then take \((e^L, e^U)\).
Exam 1 Review: Inference in Simple Logistic Regression II

- LR test of $H_0: \beta = 0$:
  
  \[
  \text{LR test statistic} = -2[L_0 - L_1] = \text{deviance}_0 - \text{deviance}_1
  \]
  
  \[\text{df} = 1\]
  
  $L_0 = \log$-likelihood maximized over $\alpha$ with $\beta = 0$
  
  $L_1 = \log$-likelihood maximized over $\alpha$ and $\beta$
  
  $= \log$-likelihood at MLEs $\hat{\alpha}, \hat{\beta}$
  
  $\text{deviance}_0 =$ "null deviance" in R
  
  $\text{deviance}_1 =$ "residual deviance" in R

---

Exam 1 Review: Multiple Logistic Regression

- Logistic regression with multiple explanatory variables
  
  \[
  \log \left( \frac{\pi}{1 - \pi} \right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k
  \]
  
  i.e.,
  
  $\pi = \frac{\exp(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)}{1 + \exp(\alpha + \beta_1 x_1 + \cdots + \beta_k x_k)}$
  
  - $\beta_i =$ partial effect of $x_i$ controlling for other variables in model
  
  $e^{\beta_i} =$ cond. odds ratio at $x_i + 1$ vs at $x_i$ keeping other $x$'s fixed
  
  = multi. effect on odds of 1-unit incr. in $x_i$, w/ other $x$'s fixed
  
  - Model may include dummies for qualitative explan. vars.
  
  - If $x_1$ is the dummy for a 2-level factor, then no interaction with other explan. vars implies homogeneous assoc: odds ratio between $Y$ and $x_1$ is the same ($e^{\beta_1}$) at any fixed level of other explan. vars.

---

Exam 1 Review: Inference for Multiple Logistic Regression

- Usual Wald tests and CIs for individual $\beta_j$s
  
  - LR test to compare reduced model $M_0$ to full model $M_1$
    
    $H_0: M_0$ holds, where $M_0 \subset M_1$
    
    \[
    \text{LR test statistic} = -2[L_0 - L_1] = \text{deviance}_0 - \text{deviance}_1
    \]
    
    \[\text{df} = \text{num. free params in } M_1 - \text{num. free params in } M_0\]
    
    $L_0 =$ maximized log-likelihood for $M_0$
    
    $L_1 =$ maximized log-likelihood for $M_1$
    
    $\text{deviance}_0 =$ (residual) deviance for $M_0$
    
    $\text{deviance}_1 =$ (residual) deviance for $M_1$
Chapter 5: Building Logistic Regression Models

- Model selection
- Model checking
- Be careful with “sparse” categorical data (estimators may take value $\infty$ or $-\infty$)

Model Selection with Many Predictors

Example (Horseshoe Crab Study)

$Y =$ whether female crab has satellites ($1 =$ yes, $0 =$ No).

Explanatory variables:
- Weight
- Width
- Color (ML, M, MD, D) w/ dummy vars $c_1, c_2, c_3$
- Spine condition (3 categories) w/ dummy vars $s_1, s_2$

Consider model for crabs:

$$\text{logit}[\Pr(Y = 1)] = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 s_1 + \beta_4 s_2 + \beta_6 \text{weight} + \beta_7 \text{width}$$

LR test of $H_0 : \beta_1 = \beta_2 = \cdots = \beta_7 = 0$ has test statistic

$$-2(L_0 - L_1) = \text{difference of deviances} = 225.8 - 185.2 = 40.6$$

$$\text{df} = 7 \quad \text{p-value} < 0.0001$$

Strong evidence at least one predictor has an effect.

But look at Wald tests of individual effects, e.g., weight.
> horseshoecrabs <-
  transform(horseshoecrabs,
    C = as.factor(Color),
    S = as.factor(Spine))
> horseshoecrabs <-
  transform(horseshoecrabs,
    C = relevel(C, "4"),
    S = relevel(S, "3"))
> crabs.fitall <-
glm((Satellites > 0) ~ C + S + Weight + Width,
    family=binomial, data=horseshoecrabs)

> summary(crabs.fitall)

Call:
glm(formula = (Satellites > 0) ~ C + S + Weight + Width, family = binomial, data = horseshoecrabs)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.198   -0.942    0.485    0.849    2.120

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.273    3.838  -2.420  0.01570 
   C1         1.609    0.936   1.720  0.08550 
   C2         1.506    0.567   2.660  0.00790 
   C3         1.120    0.593   1.890  0.05910 
   S1         0.040    0.503  -0.080  0.93390 
   S2        -0.496    0.629  -0.790  0.42510 
Weight       0.826    0.704   1.170  0.24070 
Width       0.263    0.195   1.350  0.17788 

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 225.76 on 172 degrees of freedom
Residual deviance: 185.20 on 165 degrees of freedom
AIC: 201.2

Number of Fisher Scoring iterations: 4
Multicollinearity (strong correlations among predictors) plays havoc with LMs and GLMs too.

E.g., Corr(weight, width) = 0.89.

Is partial effect of either one relevant?
Sufficient to pick one of these for a model.

```r
> attach(horseshoecrabs)
> cor(Weight, Width)
[1] 0.88687
> plot(Width, Weight)
> detach(horseshoecrabs)
```

Backward Elimination

- Use $W = \text{width}$, $C = \text{color}$, $S = \text{spine}$ as predictors.
- Start with complex model, including all interactions, say.
- Drop “least significant” (i.e., largest p-value) variable among highest-order terms.
- Refit model.
- Continue until all variables left are “significant”.

Note: If testing many interactions, simpler and possibly better to test at one time as a group of terms.
Example (Horseshoe Crabs I)

H₀: Model C + S + W holds (has 3 parameters for C, 2 for S, 1 for W)

Hₐ: Model C * S * W holds, where

\[
C + S + W = C + S + W + (C \times S) + (C \times W) + (S \times W) + (C \times S \times W)
\]

LR stat = diff. in deviances = 186.6 - 170.4 = 16.2

df = 166 - 152 = 14

p-value = 0.30

Simpler model C + S + W is adequate.

> crabs.fit1 <-
glm((Satellites > 0) ~ C*S*Width,
family=binomial, data=horseshoecrabs)

> crabs.fit2 <- update(crabs.fit1, . ~ C + S + Width)

> anova(crabs.fit2, crabs.fit1, test="Chisq")

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
<th>LRT</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>187</td>
<td>201</td>
<td></td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>194</td>
<td>202</td>
<td>7.81</td>
</tr>
<tr>
<td>S</td>
<td>2</td>
<td>188</td>
<td>198</td>
<td>0.85</td>
</tr>
<tr>
<td>Width</td>
<td>1</td>
<td>209</td>
<td>221</td>
<td>22.22</td>
</tr>
</tbody>
</table>

Example (Horseshoe Crabs II)

At next stage, S can be dropped from model C + S + W:

\[
\text{diff. in deviances} = 187.46 - 186.61 = 0.85, \quad df = 2
\]

> drop1(crabs.fit2, test="Chisq")

Single term deletions

<table>
<thead>
<tr>
<th>Model: (Satellites &gt; 0) ~ C + S + Width</th>
<th>Df Deviance AIC</th>
<th>LRT</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>187</td>
<td>201</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>3</td>
<td>194</td>
<td>202</td>
</tr>
<tr>
<td>S</td>
<td>2</td>
<td>188</td>
<td>198</td>
</tr>
<tr>
<td>Width</td>
<td>1</td>
<td>209</td>
<td>221</td>
</tr>
</tbody>
</table>
crabs.fit3 <- update(crabs.fit2, . ~ C + Width)
## crabs.fit3 <- update(crabs.fit2, . ~ . - S)
deviance(crabs.fit3)
[1] 187.46
deviance(crabs.fit2)
[1] 186.61
anova(crabs.fit3, crabs.fit2, test="Chisq")
Analysis of Deviance Table

| Resid. Df | Resid. Dev | Df | Deviance | P(>|Chi|) |
|-----------|------------|----|----------|----------|
| 1          | 188        |    |          |          |
| 2          | 187        | 2  | 0.845    | 0.66     |

drop1(crabs.fit3, test="Chisq")

Single term deletions

<table>
<thead>
<tr>
<th>Model: (Satellites &gt; 0) ~ C + Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Df Deviance AIC LRT Pr(Chi)</td>
</tr>
<tr>
<td>&lt;none&gt; 188 198</td>
</tr>
<tr>
<td>C 3 194 198 7.0 0.072</td>
</tr>
<tr>
<td>Width 1 212 220 24.6 7e-07</td>
</tr>
</tbody>
</table>

summary(crabs.fit3)

Call:
glm(formula = (Satellites > 0) ~ C + Width, family = binomial, data = horseshoecrabs)

Deviance Residuals:
Min 1Q Median 3Q Max
-2.112 -0.985 0.524 0.851 2.141

Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) -12.715 2.762 -4.60 4.1e-06 |
| C1 1.330 0.853 1.56 0.119 |
| C2 1.402 0.548 2.56 0.011 |
| C3 1.106 0.592 1.87 0.062 |
| Width 0.468 0.106 4.43 9.3e-06 |

(Dispersion parameter for binomial family taken to be 1)
Example (Horseshoe Crabs III)

Results in model fit

\[
\logit(\hat{\pi}) = -12.7 + 1.3c_1 + 1.4c_2 + 1.1c_3 + 0.47 \text{ width}
\]

Forcing \( \beta_1 = \beta_2 = \beta_3 \) gives

\[
\logit(\hat{\pi}) = -13.0 + 1.3c + 0.48 \text{ width}
\]

where \( c = \begin{cases} 
1, & \text{if color ML, M, MD}, \\
0, & \text{if color D.}
\end{cases} \)

> crabs.fit4 <- update(crabs.fit3, . ~ I(C != "4") + Width)
> anova(crabs.fit4, crabs.fit3, test="Chisq")

Analysis of Deviance Table

<table>
<thead>
<tr>
<th>Model: (Satellites &gt; 0) ~ I(C != &quot;4&quot;) + Width</th>
<th>Model 2: (Satellites &gt; 0) ~ C + Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resid. Df Resid. Dev Df Deviance P(&gt;</td>
<td>Chi</td>
</tr>
<tr>
<td>1 170 188</td>
<td>2 168 188 2 0.501 0.78</td>
</tr>
</tbody>
</table>
Example (Horseshoe Crabs IV)

Conclude:
- Controlling for width, estimated odds of satellite for nondark crabs equal $e^{1.3} = 3.7$ times est'd odds for dark crabs.

  $\text{95\%CI} : e^{1.3 \pm 1.96(0.52)} = (e^{0.270}, e^{2.33}) = (1.3, 10.3)$

- Given color (nondark or dark), est'd odds of satellite multiplied by $e^{0.478} = 1.6$ for each 1 cm increase in width.

  $\text{95\%CI} : e^{0.478 \pm 1.96(0.104)} = (e^{0.274}, e^{0.682}) = (1.3, 2.0)$
Criteria for Selecting a Model I

- Use theory, other research as guide.
- Parsimony (simplicity) is good.
- Can use some criterion to choose among set of models.
  Most popular criterion is Akaike information criterion (AIC).

Choose model with minimum AIC where
\[ AIC = -2L + 2\text{(number of model parameters)} \]

with \( L \) = log-likelihood.

- For exploratory purposes, can use automated procedure such as backward elimination.
  R function `step()` will do stepwise selection procedures (forward, backward, or both).

Criteria for Selecting a Model II

- One published simulation study suggests \( \geq 10 \) outcomes of each type (S or F) per “predictor” (count dummy variables for factors).

  Example: \( n = 1000, \ (Y = 1) \ 30 \text{ times,} \ (Y = 0) \ 970 \text{ times} \)
  Model should contain \( \leq \frac{30}{10} = 3 \) predictors.

  Example: \( n = 173 \text{ crabs,} \ (Y = 1) \ 111 \text{ crabs,} \ (Y = 0) \ 62 \text{ crabs} \)
  Use \( \leq \frac{62}{173} \approx 6 \) predictors.

- Can further check fit with residuals for grouped data, influence measures, cross validation.

Summarizing Predictive Power
A Correlation

For binary \( Y \), can summarize predictive power with sample correlation of \( Y \) and \( \hat{\pi} \).

<table>
<thead>
<tr>
<th>Model</th>
<th>Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>color</td>
<td>0.285</td>
</tr>
<tr>
<td>width</td>
<td>0.402</td>
</tr>
<tr>
<td>color + width</td>
<td>0.452</td>
</tr>
<tr>
<td>dark + width</td>
<td>0.447</td>
</tr>
</tbody>
</table>
Summarizing Predictive Power

Classification Tables

Predict $\hat{y} = 1$ if $\hat{\pi} > 0.50$ and $\hat{y} = 0$ if $\hat{\pi} < 0.50$.

Example: Horseshoe crabs with width and color as predictors

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$y = 1$</td>
<td>$y = 0$</td>
<td>Total</td>
</tr>
<tr>
<td>$y = 1$</td>
<td>94</td>
<td>17</td>
<td>111</td>
</tr>
<tr>
<td>$y = 0$</td>
<td>34</td>
<td>28</td>
<td>62</td>
</tr>
</tbody>
</table>

Sensitivity = $\Pr(\hat{Y} = 1|Y = 1) \approx \frac{94}{94 + 17} = 0.85$

Specificity = $\Pr(\hat{Y} = 0|Y = 0) \approx \frac{28}{28 + 34} = 0.45$

$\Pr($correct classification$) \approx \frac{94 + 28}{173} = 0.705$
Note:

- Table 5.3 in text actually produced by cross-validation (or more precisely by an approximation of cross-validation).

- Could use cut-off other than $\pi_0 = 0.5$, e.g., $\pi_0 = \frac{111}{173} = 0.64$.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual $\hat{y} = 1$</th>
<th>Actual $\hat{y} = 0$</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$y = 1$</td>
<td>74</td>
<td>37</td>
<td>111</td>
</tr>
<tr>
<td>$y = 0$</td>
<td>20</td>
<td>42</td>
<td>62</td>
</tr>
</tbody>
</table>

- See also receiver operating characteristic (ROC) curve.

```r
> formula(crabs.fit3)
(Satellites > 0) ~ C + Width
> pihat <- predict(crabs.fit3, type="response")
> y <- as.numeric(horseshoecrabs$Satellites > 0)
> yhat <- as.numeric(pihat > 0.50)
> table(y, yhat)

  yhat
 y   0  1
 0  31 31
 1  15 96
> addmargins(table(y, yhat), 2)

  yhat
 y   0  1 Sum
 0  31 31  62
 1  15 96 111
```

```r
> pihat <- vector(length=173)
> for (i in 1:173) {
  pihat[i] <- predict(update(crabs.fit3, subset=-i), newdata=horseshoecrabs[i,], type="response")
}
> y <- as.numeric(horseshoecrabs$Satellites > 0)
> yhat <- as.numeric(pihat > 0.50)
> confusion <- table(y, yhat)
> confusion

  yhat
 y   0  1
 0  28 34
 1  17 94
```

```r
> pihat <- vector(length=173)
> for (i in 1:173) {
  pihat[i] <- predict(update(crabs.fit3, subset=-i), newdata=horseshoecrabs[i,], type="response")
}
> y <- as.numeric(horseshoecrabs$Satellites > 0)
> yhat <- as.numeric(pihat > 0.50)
> confusion <- table(y, yhat)
> confusion

  yhat
 y   0  1
 0  28 34
 1  17 94
```
Model Checking

Is the chosen model adequate?

- Goodness of fit test.
  
  Note that tests using deviance $G^2$ and Pearson's chi-square $X^2$ are limited to “non-sparse” contingency tables.

- Check whether fit improves by adding other predictors or interactions between predictors.
  
  LR statistic (change in deviance) is useful for comparing models even when $G^2$ is not valid as an overall test of fit.

- Residuals.

Florida Death Penalty Data I

<table>
<thead>
<tr>
<th>Victim</th>
<th>Defendant</th>
<th>Death Penalty</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>Black</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>White</td>
<td>Black</td>
<td>Yes</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>Yes</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

Model fit with $d = \begin{cases} 1, & \text{black def} \\ 0, & \text{white def} \end{cases}$ and $v = \begin{cases} 1, & \text{black vic} \\ 0, & \text{white vic} \end{cases}$

$$\text{logit}(\hat{\pi}) = -2.06 + 0.87d - 2.40v$$

$$\hat{\pi} = \frac{e^{-2.06 + 0.87d - 2.40v}}{1 + e^{-2.06 + 0.87d - 2.40v}}$$

E.g., for 467 cases with $d = v = 0$: $\hat{\pi} = \frac{e^{-2.06}}{1 + e^{-2.06}} = 0.113$. 

Model Checking

Is the chosen model adequate?

- Goodness of fit test.

  Note that tests using deviance $G^2$ and Pearson's chi-square $X^2$ are limited to “non-sparse” contingency tables.

- Check whether fit improves by adding other predictors or interactions between predictors.
  
  LR statistic (change in deviance) is useful for comparing models even when $G^2$ is not valid as an overall test of fit.

- Residuals.

Florida Death Penalty Data I

<table>
<thead>
<tr>
<th>Victim</th>
<th>Defendant</th>
<th>Death Penalty</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>Black</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>White</td>
<td>Black</td>
<td>Yes</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>Yes</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

Model fit with $d = \begin{cases} 1, & \text{black def} \\ 0, & \text{white def} \end{cases}$ and $v = \begin{cases} 1, & \text{black vic} \\ 0, & \text{white vic} \end{cases}$

$$\text{logit}(\hat{\pi}) = -2.06 + 0.87d - 2.40v$$

$$\hat{\pi} = \frac{e^{-2.06 + 0.87d - 2.40v}}{1 + e^{-2.06 + 0.87d - 2.40v}}$$

E.g., for 467 cases with $d = v = 0$: $\hat{\pi} = \frac{e^{-2.06}}{1 + e^{-2.06}} = 0.113$. 

Model Checking

Is the chosen model adequate?

- Goodness of fit test.

  Note that tests using deviance $G^2$ and Pearson's chi-square $X^2$ are limited to “non-sparse” contingency tables.

- Check whether fit improves by adding other predictors or interactions between predictors.
  
  LR statistic (change in deviance) is useful for comparing models even when $G^2$ is not valid as an overall test of fit.

- Residuals.

Florida Death Penalty Data I

<table>
<thead>
<tr>
<th>Victim</th>
<th>Defendant</th>
<th>Death Penalty</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Black</td>
<td>Black</td>
<td>Yes</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>Yes</td>
<td>0</td>
</tr>
<tr>
<td>White</td>
<td>Black</td>
<td>Yes</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
<tr>
<td>White</td>
<td></td>
<td>Yes</td>
<td>53</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

Model fit with $d = \begin{cases} 1, & \text{black def} \\ 0, & \text{white def} \end{cases}$ and $v = \begin{cases} 1, & \text{black vic} \\ 0, & \text{white vic} \end{cases}$

$$\text{logit}(\hat{\pi}) = -2.06 + 0.87d - 2.40v$$

$$\hat{\pi} = \frac{e^{-2.06 + 0.87d - 2.40v}}{1 + e^{-2.06 + 0.87d - 2.40v}}$$

E.g., for 467 cases with $d = v = 0$: $\hat{\pi} = \frac{e^{-2.06}}{1 + e^{-2.06}} = 0.113$. 

Model Checking

Is the chosen model adequate?

- Goodness of fit test.

  Note that tests using deviance $G^2$ and Pearson's chi-square $X^2$ are limited to “non-sparse” contingency tables.

- Check whether fit improves by adding other predictors or interactions between predictors.
  
  LR statistic (change in deviance) is useful for comparing models even when $G^2$ is not valid as an overall test of fit.

- Residuals.
Florida Death Penalty Data II

Fitted counts for 467 cases with \( d = v = 0 \):

"Yes": \( 467 \times 0.113 = 52.8 \)  
"No": \( 467 \times 0.887 = 414.2 \)

Corresponding observed counts are 53 “yes” and 414 “no”.

Summarizing fit over 8 cells of table:

\[
X^2 = \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}} = 0.20
\]

\[
G^2 = 2 \sum \left( \frac{\text{observed}}{\text{fitted}} \log \frac{\text{observed}}{\text{fitted}} \right) = 0.38 = \text{deviance}
\]

\( df = \text{num. binomials} - \text{num. model params} = 4 - 3 = 1 \)

For \( H_0: \) “model correctly specified”, \( G^2 = 0.38, df = 1, p-value = 0.54 \). No evidence of lack of fit.

> formula(dep.fit1)

\[ \text{cbind(Yes, No) ~ Defendant + Victim} \]

> deviance(dep.fit1)

[1] 0.37984

> df.residual(dep.fit1)

[1] 1

> pchisq(deviance(dep.fit1), 1, lower.tail=FALSE)

[1] 0.53769

> chisqstat(dep.fit1)

[1] 0.19779

> pchisq(chisqstat(dep.fit1), 1, lower.tail=FALSE)

[1] 0.65651

Remarks

- Model assumes lack of interaction between \( d \) and \( v \) in effects on \( Y \). Adding interaction term gives saturated model, so goodness of fit test in this example is a test of \( H_0: \) "no interaction".
- \( X^2 \) usually recommended over \( G^2 \) for testing goodness of fit.
- These tests only appropriate for grouped binary data with most (\( \geq 80\% \)) fitted cell counts “large” (e.g., \( \hat{\mu}_i \geq 5 \)).
- For continuous predictors or many predictors with small \( \hat{\mu}_i \), distributions of \( X^2 \) and \( G^2 \) are not well approximated by \( \chi^2 \). For better approx., can try grouping data before applying \( X^2, G^2 \).
  - Hosmer-Lemeshow test forms groups using ranges of \( \hat{\pi} \) values. Implemented in R packages LDdiag and MKmisc and perhaps others.
  - Or can try to group predictor values (if only 1 or 2 predictors).
Residuals for Logit Models

At setting $i$ of explanatory variables, let

- $y_i =$ number of successes
- $n_i =$ number of trials (preferably "large")
- $\hat{\pi}_i =$ estimated probability of success based on ML fit of model

**Definition (Pearson residuals)**

For a binomial GLM, the *Pearson residuals* are

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

(note that $X^2 = \sum_i e_i^2$)

- Dist. of $e_i$ is approx. $N(0, \nu)$ when model holds, but $\nu < 1$.
- Use R function `residuals()` with option `type="pearson"`.

**Definition (Standardized Pearson residual)**

For a binomial GLM, the *standardized Pearson residuals* are

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

where $h_i$ is the “leverage” of the $i$th obs.

- $r_i$ approx. $N(0, 1)$ when model holds.
- $|r_i| > 2$ or 3 (approx.) suggests lack of fit.
- R function `rstandard()` provides standardized deviance residuals.
- Option for standardized Pearson residuals requested.
- Currently implemented by `rstandard2()` in icda package.

**Example (Berkeley Graduate Admissions)**

Data on p. 237 of text.

- $Y =$ admitted into grad school at UC Berkeley ($1=$yes, $0=$no)
- $G =$ gender ($g=1$ female, $g=0$ male)
- $D =$ dept ($A$, $B$, $C$, $D$, $E$, $F$)

- $d_1 = \begin{cases} 1, & \text{dept B}, \\ 0, & \text{o/w}, \end{cases}$ \quad \cdots \quad d_5 = \begin{cases} 1, & \text{dept F}, \\ 0, & \text{o/w}. \end{cases}$

For dept. $A$, $d_1 = \cdots = d_5 = 0$.

- Model

$$ \logit[\Pr(Y = 1)] = \alpha + \beta_1 d_1 + \cdots + \beta_5 d_5 + \beta_6 g $$

seems to fit poorly ($G^2 = 20.2$, $\chi^2 = 18.8$, df = 5).

Apparently there is gender $\times$ dept interaction.
```r
> data(UCBAdmissions)
> is.table(UCBAdmissions)
[1] TRUE
> dimnames(UCBAdmissions)
$Admit
[1] "Admitted" "Rejected"
$Gender
[1] "Male" "Female"
$Dept
[1] "A" "B" "C" "D" "E" "F"

> ftable(UCBAdmissions,
row.vars="Dept", col.vars=c("Gender","Admit"))
Gender   Male   Female
Admit     Admitted Rejected Admitted Rejected
Dept
A     512     313     89     19
B     353     207     17      8
C     120     205    202    391
D     138     279    131    244
E      53     138     94    299
F      22     351     24    317

> UCBdf <- as.data.frame(UCBAdmissions)
> head(UCBdf)
Admit Gender Dept Freq
1 Admitted    Male  A   512
2  Rejected   Male  A    313
3 Admitted Female  A     89
4  Rejected Female  A    19
5 Admitted    Male  B   353
6  Rejected   Male  B    207
```
> library(reshape2)
> UCBw <- dcast(UCBdf, Gender + Dept ~ Admit, value_var="Freq")
> UCBw

<table>
<thead>
<tr>
<th>Gender</th>
<th>Dept</th>
<th>Admitted</th>
<th>Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>A</td>
<td>512</td>
<td>313</td>
</tr>
<tr>
<td>Male</td>
<td>B</td>
<td>353</td>
<td>207</td>
</tr>
<tr>
<td>Male</td>
<td>C</td>
<td>120</td>
<td>205</td>
</tr>
<tr>
<td>Male</td>
<td>D</td>
<td>138</td>
<td>279</td>
</tr>
<tr>
<td>Male</td>
<td>E</td>
<td>53</td>
<td>138</td>
</tr>
<tr>
<td>Male</td>
<td>F</td>
<td>22</td>
<td>351</td>
</tr>
<tr>
<td>Female</td>
<td>A</td>
<td>89</td>
<td>19</td>
</tr>
<tr>
<td>Female</td>
<td>B</td>
<td>17</td>
<td>8</td>
</tr>
<tr>
<td>Female</td>
<td>C</td>
<td>202</td>
<td>391</td>
</tr>
<tr>
<td>Female</td>
<td>D</td>
<td>131</td>
<td>244</td>
</tr>
<tr>
<td>Female</td>
<td>E</td>
<td>94</td>
<td>299</td>
</tr>
<tr>
<td>Female</td>
<td>F</td>
<td>24</td>
<td>317</td>
</tr>
</tbody>
</table>

> UCB.fit1 <- glm(cbind(Admitted,Rejected) ~ Dept + Gender, family=binomial, data=UCBw)
> summary(UCB.fit1)

Call:  
glm(formula = cbind(Admitted, Rejected) ~ Dept + Gender, family = binomial, data = UCBw)

Deviance Residuals:

1          2          3          4          5          6
-1.249     -0.056     1.253      0.083      1.221     -0.208     
7          8          9          10         11         12         
3.719      0.271     -0.924     -0.086     -0.851     0.205      

Coefficients:

Estimate Std. Error  z value  Pr(>|z|)  
(Intercept) 0.58210   0.0690  8.440 <2e-16  
DeptB -0.04340   0.1098  -0.400  0.690  
DeptC -1.26260   0.1066 -11.840 <2e-16  
DeptD -1.29460   0.1058 -12.230 <2e-16  
DeptE -1.73930   0.1261 -13.790 <2e-16  
DeptF -3.30650   0.1700 -19.450 <2e-16  
GenderFemale 0.09990   0.0808  1.240  0.220  

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 877.056 on 11 degrees of freedom
Residual deviance: 20.204 on 5 degrees of freedom
AIC: 103.1

Number of Fisher Scoring iterations: 4
Example (Berkeley Admissions Ctd)

- Standardized residuals suggest Dept. A as main source of lack of fit.
- Leaving out Dept. A, model with no interaction and no gender effect fits well ($G^2 = 2.68, X^2 = 2.69, df = 5$).
- In Dept. A, sample odds-ratio of admission for females vs males is $\hat{\theta} = 2.86$ (odds of admission higher for females).

Note: Alternative way to express model with qualitative factors is, e.g.,

$$\text{logit}[\Pr(Y = 1)] = \alpha + \beta_X i + \beta_Z k,$$

where $\hat{\beta}_i^X$ is effect of classification in category $i$ of $X$.

```
> UCB.fit2 <- glm(cbind(Admitted,Rejected) ~ Dept,
family=binomial, data=UCBw, 
subset=(Dept != "A"))
> summary(UCB.fit2)
```

```
Call:
glm(formula = cbind(Admitted, Rejected) ~ Dept, family = binomial, 
data = UCBw, subset = (Dept != "A"))

Deviance Residuals:

2 3 4 5 6 8
-0.104 0.695 -0.376 0.812 -0.434 0.498

9 10 11 12
-0.518 0.395 -0.575 0.442

Coefficients:

Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.5429 0.0858 6.33 2.4e-10
DeptC -1.1586 0.1102 -10.52 < 2e-16
DeptD -1.2077 0.1139 -10.60 < 2e-16
DeptE -1.6324 0.1282 -12.73 < 2e-16
DeptF -3.2185 0.1749 -18.40 < 2e-16

(Dispersion parameter for binomial family taken to be 1)
```
Null deviance: 539.4581 on 9 degrees of freedom
Residual deviance: 2.6815 on 5 degrees of freedom
AIC: 69.92

Number of Fisher Scoring iterations: 3

> chisqstat(UCB.fit2)
[1] 2.6904

> UCB.fit3 <- update(UCB.fit2, . ~ Dept + Gender)
> anova(UCB.fit2, UCB.fit3, test="Chisq")

Analysis of Deviance Table

Model 1: cbind(Admitted, Rejected) ~ Dept
Model 2: cbind(Admitted, Rejected) ~ Dept + Gender
   Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1     5       2.68
2     4       2.56  1   0.125    0.72

> UCBAdmissions[,,"A"]

Gender
Admit     Male Female
Admitted  512   89
Rejected  313   19

> oddsratio(UCBAdmissions[,,"A"])
[1] 0.34921
> 1/oddsratio(UCBAdmissions[,,"A"])
[1] 2.8636
Sparse Data

Caution: Parameter estimates in logistic regression can be infinite.

Example:

<table>
<thead>
<tr>
<th>X</th>
<th>S</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
</tbody>
</table>

Model:

\[
\log \left( \frac{\Pr(S)}{\Pr(F)} \right) = \alpha + \beta x
\]

\[
e^\hat{\beta} = \text{odds-ratio} = \frac{8 \times 0}{2 \times 10}
\]

\[
\hat{\beta} = \text{log-odds-ratio} = -\infty
\]

Example: Text p. 155 for multi-center trial (5 ctrs, each with 2x2 table). Two centers had no successes under either treatment arm, so estimate of center effect is \(-\infty\).

Infinite estimates exist when \(x\)-values where \(y = 1\) can be "separated" from \(x\)-values where \(y = 0\).

Example: \(y = 0\) for \(x < 50\) and \(y = 1\) for \(x > 50\).

\[
\logit[\Pr(Y = 1)] = \alpha + \beta x
\]

has \(\hat{\beta} = \infty\) (roughly speaking).

Software may not realize this!

- SAS PROC GENMOD: \(\hat{\beta} = 3.84\), SE = 15601054
- SAS PROC LOGISTIC gives warning.
- SPSS: \(\hat{\beta} = 1.83\), SE = 674.8
- R: \(\hat{\beta} = 2.363\), SE = 5805, with warning.

Ch 6: Multicategory Logit Models

\(Y\) has \(J\) categories, \(J > 2\).

Extensions of logistic regression for nominal and ordinal \(Y\) assume a multinomial distribution for \(Y\).
Let $\pi_j = \Pr(Y = j)$, $j = 1, 2, \ldots, J$. 

Baseline-category logits are

$$\log\left(\frac{\pi_j}{\pi_J}\right), \quad j = 1, 2, \ldots, J - 1.$$ 

Baseline-category logit model has form

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x, \quad j = 1, 2, \ldots, J - 1.$$ 

Separate set of parameters $(\alpha_j, \beta_j)$ for each logit.

Note:
- Category used as baseline (i.e., category $J$) is arbitrary. Important because order of categories for nominal response is arbitrary.
- $e^{\beta_j}$ is the multiplicative effect of a 1-unit increase in $x$ on the odds of response $j$ instead of response $J$.
- Could also use this model with ordinal response variables, but this would ignore information about ordering.

Example (Income and Job Satisfaction from 1991 GSS)

<table>
<thead>
<tr>
<th>Income</th>
<th>Job Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dissat</td>
</tr>
<tr>
<td>$&lt;5K$</td>
<td>2</td>
</tr>
<tr>
<td>5K–15K</td>
<td>2</td>
</tr>
<tr>
<td>15K–25K</td>
<td>0</td>
</tr>
<tr>
<td>$&gt;25K$</td>
<td>0</td>
</tr>
</tbody>
</table>

Using $x = $ income scores (3, 10, 20, 35), we fit the model

$$\log\left(\frac{\pi_j}{\pi_4}\right) = \alpha_j + \beta_j x, \quad j = 1, 2, 3.$$ 

for $J = 4$ job satisfaction categories.

Use `vglm` function w/ multinomial family, both from `VGAM` package.
```r
> data(jobsat)
> head(jobsat)

<table>
<thead>
<tr>
<th>Gender</th>
<th>Income</th>
<th>JobSat</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>F</td>
<td>20</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>F</td>
<td>35</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>M</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>M</td>
<td>10</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

> jobsat <-
  transform(jobsat,
    JobSat = factor(JobSat,
      labels=c("Diss","Little","Mod","Very"),
      ordered=TRUE))

> library(reshape2)
> jobsatw <- dcast(jobsat, Income ~ JobSat, sum,
  value_var = "Freq")
> jobsatw

<table>
<thead>
<tr>
<th>Income</th>
<th>Diss</th>
<th>Little</th>
<th>Mod</th>
<th>Very</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>2</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>2</td>
<td>10</td>
<td>2</td>
<td>6</td>
<td>22</td>
</tr>
<tr>
<td>3</td>
<td>20</td>
<td>0</td>
<td>1</td>
<td>15</td>
</tr>
<tr>
<td>4</td>
<td>35</td>
<td>0</td>
<td>3</td>
<td>13</td>
</tr>
</tbody>
</table>

> library(VGAM)
> jobsat.fit1 <-
  vglm(cbind(Diss,Little,Mod,Very) ~ Income,
    family=multinomial, data=jobsatw)
> coef(jobsat.fit1)

  0.429801  0.456275  1.703929
Income:1 Income:2 Income:3
  -0.185368  -0.054412  -0.037385
```
> summary(jobsat.fit1)

Call:
vglm(formula = cbind(Diss, Little, Mod, Very) ~ Income, family = multinomial, 
data = jobsatw)

Pearson Residuals:
  log(mu[,1]/mu[,4])   log(mu[,2]/mu[,4])
  1  -0.311         0.129
  2   0.700         0.554
  3  -0.590        -1.428
  4  -0.132         0.702

log(mu[,3]/mu[,4])
  1  -0.1597
  2   0.3435
  3  -0.3038
  4   0.0489

Coefficients:

  Value Std. Error t value
(Intercept):1  0.4298  0.9448   0.455
(Intercept):2  0.4563  0.6209   0.735
(Intercept):3  1.7039  0.4811  3.542
Income:1     -0.1854  0.1025  -1.808
Income:2     -0.0544  0.0311  -1.748
Income:3     -0.0374  0.0209  -1.790

Number of linear predictors:  3

Names of linear predictors:
  log(mu[,1]/mu[,4]), log(mu[,2]/mu[,4]), log(mu[,3]/mu[,4])

Dispersion Parameter for multinomial family:  1

Residual Deviance:  4.658  on  6 degrees of freedom

Log-likelihood:  -16.954  on  6 degrees of freedom

Number of Iterations:  5
Example (Income and Job Satisfaction)

Prediction equations:

\[
\log \left( \frac{\hat{\pi}_1}{\hat{\pi}_4} \right) = 0.43 - 0.19x \\
\log \left( \frac{\hat{\pi}_2}{\hat{\pi}_4} \right) = 0.46 - 0.05x \\
\log \left( \frac{\hat{\pi}_3}{\hat{\pi}_4} \right) = 1.70 - 0.04x 
\]

Note:

▶ For each logit, estimated odds of being in less satisfied category (instead of very satisfied) decrease as \(x = \) income increases.

▶ Estimated odd of being “very dissatisfied” instead of “very satisfied” multiplied by \(e^{-0.19} = 0.83\) for each 1K increase in income.

▶ For a 10K increase in income (e.g., from row 2 to row 3), estimated odds are multiplied by \(e^{(10)(-0.19)} = e^{-1.9} = 0.16\)

   e.g., at \(x = 20\), the estimated odds of being “very dissatisfied” instead of “very satisfied” are just 0.14 times the corresponding odds at \(x = 10\).

▶ Model treats \(Y = \) job satisfaction as qualitative (nominal), but \(Y\) is ordinal. (Later we will consider a model that treats \(Y\) as ordinal.)

---

Estimating Response Probabilities

Equivalent form of model is

\[
\pi_j = \frac{e^{\alpha_j + \beta_jx}}{1 + e^{\alpha_1 + \beta_1x} + \cdots + e^{\alpha_{j-1} + \beta_{j-1}x}}, \quad j = 1, 2, \ldots, J-1, \\
\pi_J = \frac{1}{1 + e^{\alpha_1 + \beta_1x} + \cdots + e^{\alpha_{J-1} + \beta_{J-1}x}}
\]

Then

\[
\frac{\pi_j}{\pi_J} = e^{\alpha_j + \beta_jx} \\
\log \left( \frac{\pi_j}{\pi_J} \right) = \alpha_j + \beta_jx
\]

Note \(\sum_{j=1}^{J} \pi_j = 1\).
Example (Job Satisfaction)

\[ \hat{\pi}_1 = \frac{e^{0.43 - 0.19x}}{1 + e^{0.43 - 0.19x} + e^{0.46 - 0.05x} + e^{1.70 - 0.04x}} \]

\[ \hat{\pi}_2 = \frac{e^{0.46 - 0.05x}}{1 + e^{0.43 - 0.19x} + e^{0.46 - 0.05x} + e^{1.70 - 0.04x}} \]

\[ \hat{\pi}_3 = \frac{e^{1.70 - 0.04x}}{1 + e^{0.43 - 0.19x} + e^{0.46 - 0.05x} + e^{1.70 - 0.04x}} \]

\[ \hat{\pi}_4 = \frac{1}{1 + e^{0.43 - 0.19x} + e^{0.46 - 0.05x} + e^{1.70 - 0.04x}} \]

E.g., at \( x = 35 \), estimated probability of being “very satisfied” is

\[ \hat{\pi}_4 = \frac{1}{1 + e^{0.43 - 0.19(35)} + e^{0.46 - 0.05(35)} + e^{1.70 - 0.04(35)}} = 0.367 \]

Similarly, \( \hat{\pi}_1 = 0.001, \hat{\pi}_2 = 0.086, \hat{\pi}_3 = 0.545, \) and

\[ \hat{\pi}_1 + \hat{\pi}_2 + \hat{\pi}_3 + \hat{\pi}_4 = 1. \]

▶ MLEs determine estimated effects for all pairs of categories, e.g.,

\[ \log \left( \frac{\hat{\pi}_1}{\hat{\pi}_2} \right) = \log \left( \frac{\hat{\pi}_1}{\hat{\pi}_4} \right) - \log \left( \frac{\hat{\pi}_2}{\hat{\pi}_4} \right) \]

\[ = (0.43 - 0.19x) - (0.46 - 0.05x) \]

\[ = -0.026 - 0.131x \]

▶ Contingency table data, so can test goodness of fit.

(Residual) deviance is LR test statistic for comparing fitted model to saturated model.

Deviance = 4.66, df = 6, p-value = 0.59 for \( H_0: \) “model holds with linear trends for income”.

There are \( 3 \times 4 = 12 \) logits to estimate (3 baseline category logits at each of 4 income levels), so the saturated model has 12 parameters. The fitted model has 6 parameters, so df = 12 − 6 = 6.

▶ Inference uses usual methods

▶ Wald CI for \( \beta_j \) is \( \hat{\beta}_j \pm z_{\alpha/2} \) SE.

▶ Wald test of \( H_0: \beta_j = 0 \) uses \( z = \frac{\hat{\beta}_j}{\text{SE}} \) or \( z^2 \sim \chi^2_1 \).

▶ For small \( n \), better to use LR test and LR CI, if available.
Example (Job Satisfaction)

Overall “global” test of income effect

\[ H_0 : \beta_1 = \beta_2 = \beta_2 = 0 \]

LR test obtained by fitting simpler intercept only model (implies job satisfaction independent of income) to get null deviance. LR test stat is difference in deviances. Df is difference in number of parameters, or equivalently, difference in (residual) df.

\[
\text{deviance}_0 - \text{deviance}_1 = 13.47 - 4.66 = 8.81 \\
df = 6 - 3 = 9 - 6 = 3 \\
p\text{-value} = 0.032
\]

Evidence (p-value < .05) of dependence between job sat. and income.

Note that conclusion differs from that obtained with a simple chi-square test of independence (even using LR statistic \(G^2 = 13.47, \text{df} = 9, p\text{-value} = 0.1426\)). What is different here that made this possible?

```r
> jobsat.fit2 <-
  vglm(cbind(Diss,Little,Mod,Very) ~ 1,
       family=multinomial, data=jobsatw)
> deviance(jobsat.fit2)
[1] 13.467
> df.residual(jobsat.fit2)
[1] 9
```

Cumulative Logit Models for Ordinal Responses

The cumulative probabilities are

\[ \Pr(Y \leq j) = \pi_1 + \cdots + \pi_j, \quad j = 1, 2, \ldots, J. \]

The cumulative logits are

\[
\logit[\Pr(Y \leq j)] = \log \left( \frac{\Pr(Y \leq j)}{1 - \Pr(Y \leq j)} \right) = \log \left( \frac{\Pr(Y \leq j)}{\Pr(Y > j)} \right) = \log \left( \frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J} \right), \quad j = 1, \ldots, J - 1.
\]

Cumulative logit model has form

\[
\logit[\Pr(Y \leq j)] = \alpha_j + \beta x, \quad j = 1, \ldots, J - 1.
\]
Note:

- separate intercept $\alpha_j$ for each cumulative logit
- same slope $\beta$ for each cumulative logit
- $e^{\beta} = \text{multiplicative effect of 1-unit increase in } x \text{ on odds that } (Y \leq j) \text{ (instead of } (Y > j)).$

\[
\frac{\text{odds}(Y \leq j | x_2)}{\text{odds}(Y \leq j | x_1)} = e^{\beta(x_2-x_1)} = e^{\beta} \text{ when } x_2 = x_1 + 1.
\]

Also called proportional odds model.

- Use `vglm` function w/ cumulative family, both from `VGAM` package.

Example (Income and Job Satisfaction from 1991 GSS)

<table>
<thead>
<tr>
<th>Income</th>
<th>Job Satisfaction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dissat</td>
</tr>
<tr>
<td>&lt;5K</td>
<td>2</td>
</tr>
<tr>
<td>5K–15K</td>
<td>2</td>
</tr>
<tr>
<td>15K–25K</td>
<td>0</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>0</td>
</tr>
</tbody>
</table>

Using $x =$ income scores (3, 10, 20, 35), we fit the model

\[
\logit \left[ \Pr(Y \leq j) \right] = \hat{\alpha}_j + \hat{\beta}_j x = \hat{\alpha}_j - 0.045x, \quad j = 1, 2, 3.
\]

Odds of response at low end of job satisfaction scale decreases as income increases.

Contingency table data. Model fits well: deviance $= 6.75$, df $= 8$. 

```r
> jobsat.cl1 <- vglm(cbind(Diss,Little,Mod,Very) ~ Income,
                   family=cumulative(parallel=TRUE), data=jobsatw)
> summary(jobsat.cl1)

Call: vglm(formula = cbind(Diss, Little, Mod, Very) ~ Income, family = cumulative(parallel = TRUE), data = jobsatw)

Pearson Residuals: 
logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3])
1  0.583  -0.0385  -0.178
2  0.300   0.2608   0.696
3 -0.675  -1.1793  -0.960
4 -0.782   1.1186   0.334
```
Estimated odds of satisfaction below any given level multiplied by
\[ e^{\hat{\beta}} = e^{-0.045} = 0.96 \]
for each 1K increase in income (but \( x = 3, 10, 20, 35 \)).

For 10K increase in income, estimated odds multiplied by
\[ e^{10\hat{\beta}} = e^{(10)(-0.045)} = 0.64, \]
e.g., at $20K income, estimated odds of satisfaction below any given level is 0.64 times the odds at $10K income.

Remark
If reverse ordering of response, \( \hat{\beta} \) changes sign but has same SE.
With very satisfied < moderately satisfied < little dissatisfied <
very dissatisfied:
\[ \hat{\beta} = 0.045, \quad e^{\hat{\beta}} = 1.046 = 1/0.96. \]
To test $H_0: \beta = 0$ (job satisfaction indep. of income):

Wald: $z = \frac{\hat{\beta} - 0}{SE} = \frac{-0.0449}{0.0175} = -2.56$ ($z^2 = 6.57$, $df = 1$)

$p$-value = 0.0105

LR: $\text{deviance}_0 - \text{deviance}_1 = 13.47 - 6.75 = 6.72$ ($df = 1$)

$p$-value = 0.0095

Stronger evidence of association than obtained if we treat:

- $Y$ as nominal (BCL model): $\log \left( \frac{\pi_j}{\pi_4} \right) = \alpha_j + \beta_j x$.
  Recall $p$-value = 0.032 for LR test.

- $X, Y$ both as nominal: Pearson test of indep. had $X^2 = 11.5, df = 9$,
  $p$-value = 0.24 ($G^2 = 13.47, p$-value = 0.14).

$G^2$ same as testing all $\beta_j = 0$ in BCL model w/ dummies for income.

With BCL or cumulative logit models, can mix quantitative and qualitative predictors, interaction terms, etc.

```r
> jobsat.cl0 <-
vglm(cbind(Diss,Little,Mod,Very) ~ 1,
family=cumulative(parallel=TRUE), data=jobsatw)
> deviance(jobsat.cl0)
[1] 13.467
> deviance(jobsat.cl1)
[1] 6.7494
> pchisq(deviance(jobsat.cl0) - deviance(jobsat.cl1), 1,
lower.tail=FALSE)
[1] 0.009545
```

Example (Political Ideology and Party Affiliation (GSS))

$Y =$ political ideology (1 = very liberal, . . ., 5 = very conservative)

$x_1 =$ gender (1 = M, 0 = F)

$x_2 =$ political party (1 = Rep, 0 = Dem)

ML fit:

$$\text{logit} \left[ \hat{P}(Y \leq j) \right] = \hat{\alpha}_j - 0.117x_1 - 0.964x_2, \quad j = 1, 2, 3, 4.$$ 

Controlling for gender, estimated odds that a Republican’s response is in liberal direction ($Y \leq j$) rather than conservative ($Y > j$) are

$e^{-0.964} = 0.38$ times estimated odds for a Democrat.

Equivalently: controlling for gender, estimated odds that a Democrat’s response is in liberal direction ($Y \leq j$) rather than conservative ($Y > j$) are $e^{0.964} = 2.62$ times estimated odds for a Republican.

Same for all $j = 1, 2, 3, 4$. 

95% CI for true odds ratio is
\[ e^{-0.964 \pm (1.96)(0.129)} = (0.30, 0.49) \]

Contingency table data. Little evidence of lack of fit:
\[
\text{deviance} = 15.1, \quad \text{df} = 10, \quad \text{p-value} = 0.13
\]

LR test of \( H_0 : \beta_2 = 0 \) (no party effect, given gender)
\[
\text{deviance}_0 - \text{deviance}_1 = 71.9 - 15.1 = 56.8, \quad \text{df} = 1
\]
\[ \text{p-value} < 0.0001 \]
Strong evidence that Republicans tend to be less liberal (more conservative) than Democrats (for each gender).

Little evidence of gender effect (controlling for party).

Interaction? ML fit of model permitting interaction is
\[
\logit [\hat{Pr}(Y \leq j)] = \hat{\alpha}_j + 0.143x_1 - 0.756x_2 - 0.509x_1x_2
\]
For \( H_0 : \beta_3 = 0 \)
\[
\text{LR stat.} = 3.99, \quad \text{df} = 1, \quad \text{p-value} = 0.046
\]
Evidence that effect of Party depends on Gender (and vice versa).

Estimated odds ratio for party effect (\( x_2 \)) is
\[ e^{-0.756} = 0.47 \quad \text{when} \quad x_1 = 0 \ (F) \]
\[ e^{-0.756 - 0.509} = e^{-1.265} = 0.28 \quad \text{when} \quad x_1 = 1 \ (M) \]

Estimated odds ratio for gender effect (\( x_1 \)) is
\[ e^{0.143} = 1.15 \quad \text{when} \quad x_2 = 0 \ (\text{Dem}) \]
\[ e^{0.143 - 0.509} = e^{-0.336} = 0.69 \quad \text{when} \quad x_2 = 1 \ (\text{Rep}) \]
Among Dems, males tend to be more liberal than females.
Among Reps, males tend to be more conservative than females.

Find \( \hat{Pr}(Y = 1) \) (very liberal) for male Republicans, female Republicans.
\[
\hat{Pr}(Y \leq j) = \frac{\exp(\hat{\alpha}_j + 0.143x_1 - 0.756x_2 - 0.509x_1x_2)}{1 + \exp(\hat{\alpha}_j + 0.143x_1 - 0.756x_2 - 0.509x_1x_2)}
\]
For \( j = 1, \hat{\alpha}_1 = -1.55. \)

Male Republicans \((x_1 = 1, \ x_2 = 1)\):
\[
\hat{Pr}(Y = 1) = \frac{e^{-1.55+0.143-0.756-0.509}}{1 + e^{-1.55+0.143-0.756-0.509}} = \frac{e^{-2.67}}{1 + e^{-2.67}} = 0.065
\]

Female Republicans \((x_1 = 0, \ x_2 = 1)\):
\[
\hat{Pr}(Y = 1) = \frac{e^{-1.55-0.756}}{1 + e^{-1.55-0.756}} = \frac{e^{-2.31}}{1 + e^{-2.31}} = 0.090
\]

Similarly, \( \hat{Pr}(Y = 2) = \hat{Pr}(Y \leq 2) - \hat{Pr}(Y \leq 1) \), etc.
Note \( \hat{Pr}(Y = 5) = \hat{Pr}(Y \leq 5) - \hat{Pr}(Y \leq 4) = 1 - \hat{Pr}(Y \leq 4) \).
Remarks

- Reversing order of response categories changes signs of “slope” estimates (odds ratio $\rightarrow \frac{1}{\text{odds ratio}}$).
- For ordinal response, no other reordering sensible.

```r
> data(ideology)
> head(ideology)

  Party Gender Ideology Freq
  1  Dem Female  VLib  44
  2  Rep Female  VLib  18
  3  Dem  Male  VLib  36
  4  Rep   Male  VLib  12
  5  Dem Female  SLib  47
  6  Rep Female  SLib  28

> library(reshape2)
> ideow <- dcast(ideology, Gender + Party ~ Ideology,
                   value_var="Freq")

> ideow
   Gender Party   VLib   SLib   Mod   SCon   VCon
  1 Female Dem  44  47  118  23  32
  2 Female Rep  18  28  86  39  48
  3  Male Dem  36  34  53  18  23
  4  Male Rep  12  18  62  45  51

> library(VGAM)
> ideo.cl1 <-
  vglm(cbind(VLib,SLib,Mod,SCon,VCon) ~ Gender + Party,
       family=cumulative(parallel=TRUE), data=ideow)
```
summary(ideo.cl1)

Call:
vglm(formula = cbind(VLib, SLib, Mod, SCon, VCon) ~ Gender +
     Party, family = cumulative(parallel = TRUE), data = ideow)

Pearson Residuals:
  logit(P[Y<=1]) logit(P[Y<=2]) logit(P[Y<=3])
1 -0.638 -1.499  1.318
2 -0.216  0.540  0.641
3  1.288  1.668 -0.346
4 -0.363 -0.272 -1.752

  logit(P[Y<=4])
1 -0.844
2  0.684
3 -0.869
4  0.528

Coefficients:
     Value Std. Error t value
(Intercept):1 -1.452  0.123 -11.818
(Intercept):2 -0.458  0.106 -4.333
(Intercept):3  1.255  0.115 10.956
(Intercept):4  2.089  0.129 16.174
  GenderMale -0.117  0.127 -0.921
  PartyRep  -0.964  0.129 -7.449

Number of linear predictors: 4

Names of linear predictors:
  logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3]), logit(P[Y<=4])

Dispersion Parameter for cumulative family: 1

Residual Deviance: 15.056 on 10 degrees of freedom

Log-likelihood: -47.415 on 10 degrees of freedom

pchisq(deviance(ideo.cl1),10,lower.tail=FALSE)
[1] 0.13005

ideo.cl2 <-
vglm(cbind(VLib, SLib, Mod, SCon, VCon) ~ Gender,
     family=cumulative(parallel=TRUE), data=ideow)

deviance(ideo.cl2)
[1] 71.902

df.residual(ideo.cl2)
[1] 11

pchisq(deviance(ideo.cl2) - deviance(ideo.cl1), 1,
       lower.tail=FALSE)
[1] 4.711e-14
Ch 8: Models for Matched Pairs
McNemar’s Test

Example (Crossover Study: Drug vs Placebo I)

86 subjects. Randomly assign each to either “drug then placebo” or “placebo then drug”. Binary response (S,F) for each.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>S</th>
<th>F</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Drug</td>
<td>61</td>
<td>25</td>
<td>86</td>
</tr>
<tr>
<td>Placebo</td>
<td>22</td>
<td>64</td>
<td>86</td>
</tr>
</tbody>
</table>

Methods so far (e.g., $X^2$ and $G^2$ test of indep, CI for $\theta$, logistic regr) assume independent samples. Inappropriate for dependent samples (e.g., same subjects in each sample, which yield matched pairs).
Example (Crossover Study: Drug vs Placebo II)

To reflect dependence, display data as 86 obs rather than $2 \times 86$ obs.

<table>
<thead>
<tr>
<th>Drug</th>
<th>S</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>12</td>
<td>49</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>64</td>
</tr>
</tbody>
</table>

Population probabilities:

<table>
<thead>
<tr>
<th>Drug</th>
<th>S</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>$\pi_{11}$</td>
<td>$\pi_{12}$</td>
</tr>
<tr>
<td>F</td>
<td>$\pi_{21}$</td>
<td>$\pi_{22}$</td>
</tr>
</tbody>
</table>

$\pi_{+1} = \pi_{1+}$

Compare dependent samples by making inference about $\pi_{1+} - \pi_{+1}$.

There is *marginal homogeneity* if $\pi_{1+} = \pi_{+1}$.

Under $H_0$: marginal homogeneity,

$$\frac{\pi_{12}}{\pi_{12} + \pi_{21}} = \frac{1}{2}.$$  

Under $H_0$, each of $n^* = n_{12} + n_{21}$ observations has probability $1/2$ of contributed to $n_{12}$ and $1/2$ of contributing to $n_{21}$:

$$n_{12} \sim \text{Bin}(n^*, \frac{1}{2}), \quad \text{mean} = \frac{n^*}{2}, \quad \text{std dev} = \sqrt{n^* \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)}$$

By normal approx. to binomial, for large $n^*$,

$$z = \frac{n_{12} - n^*/2}{\sqrt{n^* \left(\frac{1}{2}\right) \left(\frac{1}{2}\right)}} \sim N(0,1)$$

or equivalently

$$z^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi^2_1$$

Called McNemar’s test.

Example (Crossover Study: Drug vs Placebo III)

<table>
<thead>
<tr>
<th>Drug</th>
<th>S</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>12</td>
<td>49</td>
</tr>
<tr>
<td>F</td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>64</td>
</tr>
</tbody>
</table>

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{49 - 10}{\sqrt{49 + 10}} = 5.1 \quad (z^2 = 25.8, \text{df} = 1)$$

p-value < 0.0001 for $H_0 : \pi_{1+} = \pi_{+1}$ vs $H_a : \pi_{1+} \neq \pi_{+1}$.

Extremely strong evidence that probability of success is higher for drug than placebo.
CI for $\pi_1 - \pi_{+1}$

Estimate $\pi_1 - \pi_{+1}$ by diff. of sample proportions, $p_{1+} - p_{+1}$.

$$p_{1+} - p_{+1} = \frac{n_{1+} - n_{+1}}{n} = \frac{n_{12} - n_{21}}{n}$$

$$SE = \frac{1}{n} \sqrt{n_{12} + n_{21} - \frac{(n_{12} - n_{21})^2}{n}}$$

Example (Crossover Study: Drug vs Placebo IV)

\[
\begin{array}{c|c|c}
 n_{11} & n_{12} & 12 \\
 n_{21} & n_{22} & 49 \\
\hline
 n & 10 \\
\end{array}
\]

$$p_{1+} - p_{+1} = \frac{49 - 10}{86} = \frac{39}{86} = 0.453$$

$$SE = \frac{1}{86} \sqrt{49 + 10 - \frac{(49 - 10)^2}{86}} = 0.075$$

95% CI: $0.453 \pm (1.96)(0.075) = 0.453 \pm 0.146 = (0.31, 0.60)$

Aside: How is the SE derived?

\((n_{11}, n_{12}, n_{21}, n_{22}) \sim MN(n, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}))\)

\[\Rightarrow \begin{cases} Var(n_{ij}) = n\pi_{ij}(1 - \pi_{ij}) \\ Cov(n_{ij}, n_{i,j'}) = -n\pi_{ij}\pi_{i,j'} & \text{if } i \neq i' \text{ or } j \neq j' \end{cases} \]

\[Var(p_{1+} - p_{+1}) = Var\left(\frac{n_{12} - n_{21}}{n}\right) = \frac{Var(n_{12} - n_{21})}{n^2} \]

\[= Var(n_{12}) + Var(n_{21}) - 2 Cov(n_{12}, n_{21}) \]

\[= \frac{n\pi_{12}(1 - \pi_{12}) + n\pi_{21}(1 - \pi_{21}) + 2n\pi_{12}\pi_{21}}{n^2} \]

\[= \frac{\pi_{12} + \pi_{21} - (\pi_{12}^2 - 2\pi_{12}\pi_{21} + \pi_{21}^2)}{n} \]

\[= \frac{\pi_{12} + \pi_{21} - (\pi_{12} - \pi_{21})^2}{n} \]

(.ctd next frame)
Another way:
\[
\text{Var}(p_{1+} - p_{1+}) = \text{Var}(p_{1+}) + \text{Var}(p_{1+}) - 2 \text{Cov}(p_{1+}, p_{1+})
\]
\[
\text{Var}(p_{1+}) = \frac{\pi_1(1-\pi_1)}{n}, \quad \text{Var}(p_{1+}) = \frac{\pi_{1+}(1-\pi_{1+})}{n},
\]
\[
\text{Cov}(p_{1+}, p_{1+}) = \text{Cov}\left(\frac{n_{1+}}{n}, \frac{n_{1+}}{n}\right) = \text{Cov}\left(\frac{n_{11}+n_{12}}{n}, \frac{n_{11}+n_{21}}{n}\right).
\]
\[
\text{Cov}(p_{1+}, p_{1+}) = \frac{1}{n^2} \text{Cov}(n_{11} + n_{12}, n_{11} + n_{21})
\]
\[
\text{Cov}(p_{1+}, p_{1+}) = \frac{1}{n^2} \left[ \text{Var}(n_{11}) + \text{Cov}(n_{11}, n_{21}) + \text{Cov}(n_{12}, n_{11}) + \text{Cov}(n_{12}, n_{21}) \right]
\]
\[
\text{Cov}(p_{1+}, p_{1+}) = \frac{1}{n^2} \left[ n\pi_{11}(1-\pi_{11}) - n\pi_{11}\pi_{21} - n\pi_{12}\pi_{11} - n\pi_{12}\pi_{21} \right]
\]
\[
\text{Cov}(p_{1+}, p_{1+}) = \frac{1}{n} \left[ \pi_{11}(1-\pi_{11}) - \pi_{12}\pi_{21} \right]
\]
\[
\text{Cov}(p_{1+}, p_{1+}) = \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{n}
\]

Thus,
\[
\text{Var}(p_{1+} - p_{1+})
\]
\[
= \frac{1}{n} \left[ \pi_{1+}(1-\pi_{1+}) + \pi_{1+}(1-\pi_{1+}) - 2(\pi_{11}\pi_{22} - \pi_{12}\pi_{21}) \right]
\]

Often matched-pairs exhibit positive association (odds-ratio greater than 1), i.e., \( \pi_{11}\pi_{22} > \pi_{12}\pi_{21} \), so covariance term is negative. Compare to two independent samples of size \( n \) each.

Continuing,
\[
\text{Var}(p_{1+} - p_{1+})
\]
\[
= \frac{1}{n} \left[ p_{1+}(1-p_{1+}) + p_{1+}(1-p_{1+}) - 2(p_{11}p_{22} - p_{12}p_{21}) \right]
\]

After algebra, this simplifies to expression given before.

```r
> crossover <-
  matrix(c(12, 10, 49, 15), nrow=2,
  dimnames=list(Drug=c("S","F"), Placebo=c("S","F")))
> crossover <- as.table(crossover)
> crossover
Placebo
Drug  S  F
  S 12  49
  F 10  15

> mcnemar.test(crossover, correct = FALSE)
McNemar's Chi-squared test
data: crossover
McNemar's chi-squared = 25.78, df = 1, p-value =
3.827e-07
```
Sec 8.5.5 Measuring Agreement

Example (Movie Reviews by Siskel and Ebert)

<table>
<thead>
<tr>
<th></th>
<th>Ebert</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Con</td>
<td>Mixed</td>
<td>Pro</td>
<td>Total</td>
</tr>
<tr>
<td>Siskel</td>
<td>Con</td>
<td>24</td>
<td>8</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>Mixed</td>
<td>8</td>
<td>13</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Pro</td>
<td>10</td>
<td>9</td>
<td>64</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>42</td>
<td>30</td>
<td>88</td>
</tr>
</tbody>
</table>

How strong is their agreement?

Cohen's Kappa

Let $\pi_{ij} = \Pr(S = i, E = j)$.

$$
\Pr(\text{agree}) = \pi_{11} + \pi_{22} + \pi_{33} = \sum_i \pi_{ii}
$$

$= 1$ if perfect agreement

If ratings are independent, then $\pi_{ii} = \pi_i = \pi_{i+}$ and

$$
\Pr(\text{agree|indep}) = \sum_i \pi_i = \sum_i \pi_{i+} = \sum_i \pi_{+i}
$$

Cohen's kappa is

$$
\kappa = \frac{\Pr(\text{agree}) - \Pr(\text{agree|indep})}{1 - \Pr(\text{agree|indep})} = \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+} = \sum_i \pi_{+i}}{1 - \sum_i \pi_{i+} + \sum_i \pi_{+i}}
$$

Note:

- $\kappa = 0$ if agreement only equals that expected under independence.
- $\kappa = 1$ if perfect agreement.
- Denominator = maximum difference for numerator, attained if agreement is perfect.
Example (Siskel and Ebert (ctd))

\[
\sum_{i} \hat{\pi}_{ii} = \frac{24 + 13 + 64}{160} = 0.63
\]

\[
\sum_{i} \hat{\pi}_{i+} \hat{\pi}_{+j} = \left( \frac{45}{160} \right) \left( \frac{42}{160} \right) + \left( \frac{32}{160} \right) \left( \frac{30}{160} \right) + \left( \frac{83}{160} \right) \left( \frac{88}{160} \right)
= 0.40
\]

\[
\hat{\kappa} = \frac{0.63 - 0.40}{1 - 0.40} = 0.39
\]

Moderate agreement: difference between observed agreement and agreement expected under independence is about 40% of the maximum possible difference.

- 95% CI for \( \kappa \):
  \[
  \hat{\kappa} \pm 1.96 \text{SE} = 0.39 \pm (1.96)(0.06) = 0.39 \pm 0.12 = (0.27, 0.51)
  \]

- For \( H_0 : \kappa = 0 \),
  \[
  z = \frac{\hat{\kappa}}{\text{SE}} = \frac{0.39}{0.06} = 6.49
  \]
  Very strong evidence that agreement is better than "chance".

- A very simple `cohens.kappa()` is in the icda package. More sophisticated versions can be found in several packages on CRAN (e.g., irr, concord, and psy).

```r
> data(moviereviews)
> moviereviews

  Ebert
  Siskel  Con Mixed Pro
  Con 24 8 13
  Mixed 8 13 11
  Pro 10 9 64
> cohens.kappa(moviereviews)

$kappa
[1] 0.38884

$SE
[1] 0.05992
```
The usual models apply (e.g., logistic regression for binary response, cumulative logit for ordinal response), but model fitting must account for dependence (e.g., from repeated measures on subjects).

**Generalized Estimating Equations (GEE) for Repeated Measures**

- Specify model in usual way.
- Select a “working correlation” matrix for best guess about correlation pattern between pairs of observations.

| Ex: For T repeated responses, exchangeable correlation matrix is |
|---|---|---|---|
| Time | 1 | 2 | ... | T |
| 1 | 1 | ρ | ... | ρ |
| 2 | ρ | 1 | ... | ρ |
| ... | ... | ... | ... | ... |
| T | ρ | ρ | ... | 1 |

- Fitting method gives estimates that are good even if correlation structure is misspecified. Adjusts standard errors to reflect actual observed dependencies.
- Available in R package `gee` and others.

**Example (Crossover Study: Drug vs Placebo V)**

<table>
<thead>
<tr>
<th>Drug</th>
<th>S</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Placebo</td>
<td>12</td>
<td>49</td>
</tr>
<tr>
<td></td>
<td>10</td>
<td>15</td>
</tr>
<tr>
<td></td>
<td>22</td>
<td>64</td>
</tr>
</tbody>
</table>

Model:

\[
\text{logit}[\Pr(Y_t = 1)] = \alpha + \beta t, \quad t = \begin{cases} 
1, & \text{drug} \\
0, & \text{placebo}
\end{cases}
\]

GEE fit:

\[
\text{logit}[\Pr(Y_t = 1)] = -1.07 + 1.96t, \quad \text{SE}(\beta) = 0.377 \text{ ("robust")}
\]

Odds of S w/ drug estimated to be \(e^{1.96} = 7.1\) times odds w/ placebo. 95% CI for odds ratio (for marginal probabilities) is

\[
e^{1.96 \pm (1.96)(0.377)} = (e^{1.22}, e^{2.70}) = (3.4, 14.9)
\]
Note:

- Sample marginal odds ratio is $\hat{\theta} = \frac{61/25}{22/64} = 7.1$ ($\log \hat{\theta} = 1.96$). Model is saturated.

- With GEE approach, can also have "between-subject" explanatory variables, e.g., gender, order of treatments.

GEE is a "quasi-likelihood" method.

- Assumes a distribution (e.g., binomial) for $Y_1$, for $Y_2$, ..., for $Y_T$ (marginal distributions).

- No particular form assumed for joint distribution of $(Y_1, Y_2, \ldots, Y_T)$

- Hence, no likelihood function, no LR inference (LR test, LR CI).

- For responses $(Y_1, Y_2, \ldots, Y_T)$ at $T$ times, we consider marginal model that describes each $Y_t$ in terms of explanatory var’s.

- Alternative conditional model put terms in model for subjects, effects apply conditional on subject, e.g.,

  $\logit \left[ \Pr(Y_{it} = 1) \right] = \alpha_i + \beta_t$ \hspace{1cm} ($\alpha_i =$ effect for subject $i$)

  $\{\alpha_i\}$ commonly treated as “random effects” having a normal distribution (Ch 10).

```r
> library(gee)
> crossover
   Placebo Drug S F
   S 12 49
   F 10 15
> cross.df <- data.frame(crossover)
> cross.df <-
  transform(cross.df,
    Drug = as.numeric(Drug=="S"),
    Placebo = as.numeric(Placebo=="S"))
> cross.df
   Drug Placebo Freq
   1    1       1  12
   2    0       1  10
   3    1       0  49
   4    0       0  15
```
```r
> Freq <- cross.df$Freq
> cross.df$Freq <- NULL
> cross.df <- cross.df[rep(1:4, Freq),]
> rm(Freq)
> head(cross.df)

Drug Placebo
1 1 1
1.1 1 1
1.2 1 1
1.3 1 1
1.4 1 1
1.5 1 1

> rownames(cross.df) <- NULL
> head(cross.df)

Drug Placebo
1 1 1
2 1 1
3 1 1
4 1 1
5 1 1
6 1 1

> xtabs(~ Drug + Placebo, cross.df)

Placebo
Drug 0 1
0 15 10
1 49 12

> dim(cross.df)
[1] 86 2
> cross.df$Subject <- factor(1:86)
> crossm <- melt(cross.df)
> head(crossm)

Subject variable value
1 1 Drug 1
2 2 Drug 1
3 3 Drug 1
4 4 Drug 1
5 5 Drug 1
6 6 Drug 1
```
> ## VERY IMPORTANT: Data should be ordered by "cluster"
> crossm <- crossm[order(crossm$Subject),]
> head(crossm)
>
<table>
<thead>
<tr>
<th>Subject</th>
<th>variable</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Drug</td>
<td>1</td>
</tr>
<tr>
<td>87</td>
<td>Placebo</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Drug</td>
<td>1</td>
</tr>
<tr>
<td>88</td>
<td>Placebo</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Drug</td>
<td>1</td>
</tr>
<tr>
<td>89</td>
<td>Placebo</td>
<td>1</td>
</tr>
</tbody>
</table>
>
> crossm <-
> transform(crossm, variable=relevel(variable, "Placebo"))

> cross.gee1 <-
> gee(value ~ variable, id=Subject, data=crossm,
> family=binomial, corstr="exchangeable")
>
<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>variableDrug</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.0678</td>
<td>1.9598</td>
</tr>
</tbody>
</table>

> cross.gee2 <-
> gee(value ~ variable, id=Subject, data=crossm,
> family=binomial, corstr="independence")
>
<table>
<thead>
<tr>
<th>(Intercept)</th>
<th>variableDrug</th>
</tr>
</thead>
<tbody>
<tr>
<td>-1.0678</td>
<td>1.9598</td>
</tr>
</tbody>
</table>

> summary(cross.gee1)

GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gg S-function, version 4.13 modified 98/01/27 (1998)

Model:        
Link:         Logit
Variance to Mean Relation: Binomial
Correlation Structure:   Exchangeable

Call:        
gee(formula = value ~ variable, id = Subject, data = crossm,
family = binomial, corstr = "exchangeable")

Summary of Residuals:

<table>
<thead>
<tr>
<th>Min</th>
<th>1Q</th>
<th>Median</th>
<th>3Q</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>-0.70930</td>
<td>-0.25581</td>
<td>-0.25581</td>
<td>0.29070</td>
<td>0.74419</td>
</tr>
</tbody>
</table>
Coefficients:

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Naive S.E.</th>
<th>Naive z</th>
<th>Robust S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-1.0678</td>
<td>0.24859</td>
<td>-4.2956</td>
<td>0.24714</td>
</tr>
<tr>
<td>variableDrug</td>
<td>1.9598</td>
<td>0.37984</td>
<td>5.1596</td>
<td>0.37723</td>
</tr>
</tbody>
</table>

Robust z

<table>
<thead>
<tr>
<th></th>
<th>Estimate</th>
<th>Naive S.E.</th>
<th>Naive z</th>
<th>Robust S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-4.3207</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>variableDrug</td>
<td>5.1953</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Estimated Scale Parameter: 1.0118
Number of Iterations: 1

Working Correlation

```
[,1] [,2]
[1,] 1.00000 -0.21407
[2,] -0.21407 1.00000
```

Example (Depression I)

\( y = \) response on mental depression (normal, abnormal)
measured three times (after 1, 2, and 4 wks of treatment)
two drug treatments (standard, new)
two severity of initial diagnosis groups (mild, severe)
Is the rate of improvement better with the new drug?

<table>
<thead>
<tr>
<th>Time</th>
<th>Response Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A A A A N N N N</td>
</tr>
<tr>
<td>1</td>
<td>A A A N A A N N</td>
</tr>
<tr>
<td>2</td>
<td>A N A N A N A N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Severity</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mild</td>
<td>Std</td>
</tr>
<tr>
<td></td>
<td>6 15 14 3 9 13 16</td>
</tr>
<tr>
<td></td>
<td>New</td>
</tr>
<tr>
<td></td>
<td>0 9 2 22 0 6 0 31</td>
</tr>
<tr>
<td>Severe</td>
<td>Std</td>
</tr>
<tr>
<td></td>
<td>28 27 15 9 9 8 2 2</td>
</tr>
<tr>
<td></td>
<td>New</td>
</tr>
<tr>
<td></td>
<td>6 32 5 31 2 5 2 7</td>
</tr>
</tbody>
</table>
### Example (Depression II)

<table>
<thead>
<tr>
<th>Severity</th>
<th>Drug</th>
<th>Week 1</th>
<th>Week 2</th>
<th>Week 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mild</td>
<td>Std</td>
<td>0.51</td>
<td>0.59</td>
<td>0.68</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>0.53</td>
<td>0.79</td>
<td>0.97</td>
</tr>
<tr>
<td>Severe</td>
<td>Std</td>
<td>0.21</td>
<td>0.28</td>
<td>0.46</td>
</tr>
<tr>
<td></td>
<td>New</td>
<td>0.18</td>
<td>0.50</td>
<td>0.83</td>
</tr>
</tbody>
</table>

E.g., $0.51 = \frac{3 + 9 + 13 + 16}{6 + 15 + 4 + 14 + 3 + 9 + 13 + 16}$

---

Let

$Y_t = \text{resp of randomly selected subj at time } t$ ($1 = \text{norm}, 0 = \text{abnor}$)

$s = \text{severity of initial diagnosis} (1 = \text{severe}, 0 = \text{mild})$

$d = \text{drug} (1 = \text{new}, 0 = \text{std})$

$t = \text{time} (0, 1, 2), \text{which is } \log_2(\text{weeks of trt})$

Model:

$$\log \left\{ \frac{\Pr(Y_t = 1)}{\Pr(Y_t = 0)} \right\} = \alpha + \beta_1 s + \beta_2 d + \beta_3 t$$

Assumes same rate of change $\beta_3$ over time for each $(s, d)$ combination. Unrealistic?

More realistic model permits time effect to differ by drug:

$$\log \left\{ \frac{\Pr(Y_t = 1)}{\Pr(Y_t = 0)} \right\} = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt$$

Time effect = \begin{cases} 
\beta_3 & \text{if } d = 0 \text{ (std drug)} \\
\beta_3 + \beta_4 & \text{if } d = 1 \text{ (new drug)}
\end{cases}

GEE estimates: \begin{align*}
\hat{\alpha} &= -0.28 \\
\hat{\beta}_1 &= -1.31 \\
\hat{\beta}_2 &= -0.06 \\
\hat{\beta}_3 &= 0.48 \\
\hat{\beta}_4 &= 1.02
\end{align*}

Test of $H_0$: no interaction ($\beta_4 = 0$) has

$$z = \frac{\hat{\beta}_4}{\text{SE}} = \frac{1.02}{0.188} = 5.42 \quad (z^2 = 29.4, \text{df} = 1)$$

Very strong evidence of faster improvement for new drug.
When initial diagnosis is severe, estimated odds of normal response are $e^{-1.31} = 0.27$ times estimated odds when initial diagnosis is mild, at each $d \times t$ combination.

$\hat{\beta}_2 = -0.06$ is drug effect only at $t = 0$. $e^{-0.06} = 0.94 \approx 1$, so essentially no drug effect at $t = 0$ (after 1 week).

Drug effect at end of study ($t = 2$) estimated to be $e^{\hat{\beta}_2 + \hat{\beta}_4} = 7.2$.

Estimated time effects are

- std drug ($d = 0$): $\hat{\beta}_3 = 0.48$
- new drug ($d = 1$): $\hat{\beta}_3 + \hat{\beta}_4 = 1.50$

Examined $s \times d$ and $s \times t$ interactions, but they were not statistically significant.

Started w/ exchangeable working correlation, but est'd $\rho$ close to 0.

```r
> library(gee)
> data(depression)
> head(depression)
subject severity drug time response
1 1 mild std 0 normal
2 1 mild std 1 normal
3 1 mild std 2 normal
4 2 mild std 0 normal
5 2 mild std 1 normal
6 2 mild std 2 normal

> dep.gee1 <- 
  gee((response == "normal") ~ severity + drug*time, 
    id=subject, data=depression, family=binomial)

(Intercept) severitysevere drugnew
-0.027988 -1.313911 -0.059604
time drugnew:time
0.482412 1.017445

> summary(dep.gee1)
```

```
GEE: GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
gge S-function, version 4.13 modified 98/01/27 (1998)

Model:
  Link: Logit
  Variance to Mean Relation: Binomial
  Correlation Structure: Independent

Call:
gee(formula = (response == "normal") ~ severity + drug * time, 
    id = subject, data = depression, family = binomial)

Summary of Residuals:
  Min 1Q Median 3Q Max
-0.948442 -0.406833 0.051558 0.388310 0.802422
```
Coefficients:

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Naive S.E.</th>
<th>Naive z</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>-0.027988</td>
<td>0.16271</td>
<td>-0.17202</td>
</tr>
<tr>
<td>severitysevere</td>
<td>-1.313911</td>
<td>0.14534</td>
<td>-9.04006</td>
</tr>
<tr>
<td>drugnew</td>
<td>-0.059604</td>
<td>0.22058</td>
<td>-0.27021</td>
</tr>
<tr>
<td>time</td>
<td>0.482412</td>
<td>0.11392</td>
<td>4.23457</td>
</tr>
<tr>
<td>drugnew:time</td>
<td>1.017445</td>
<td>0.18741</td>
<td>5.42889</td>
</tr>
</tbody>
</table>

Robust S.E. Robust z

<table>
<thead>
<tr>
<th>Term</th>
<th>Robust S.E.</th>
<th>Robust z</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Intercept)</td>
<td>0.17419</td>
<td>-0.16068</td>
</tr>
<tr>
<td>severitysevere</td>
<td>0.14598</td>
<td>-9.00034</td>
</tr>
<tr>
<td>drugnew</td>
<td>0.22854</td>
<td>-0.26080</td>
</tr>
<tr>
<td>time</td>
<td>0.11994</td>
<td>4.02228</td>
</tr>
<tr>
<td>drugnew:time</td>
<td>0.18769</td>
<td>5.42077</td>
</tr>
</tbody>
</table>

Estimated Scale Parameter: 0.98541
Number of Iterations: 1

Working Correlation

<table>
<thead>
<tr>
<th>[,1]</th>
<th>[,2]</th>
<th>[,3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>[2,]</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>[3,]</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

By way of illustration, the next few frames show bits and pieces of some other gee fits to these data. Note that the working correlation matrix can be "independence" (default), "exchangeable", "AR-M", "stat M dep", "non_stat M dep", "unstructured", and "fixed". See the help for gee for details.

```r
> dep.gee2 <-
  gee(response == "normal" ~ severity + drug*time, 
        id=subject, data=depression, family=binomial, 
        corstr="exchangeable")

(Intercept) severitysevere drugnew
-0.027988   -1.313911   -0.059604

time drugnew:time
0.482412    1.017445

> dep.gee2$working.correlation

<table>
<thead>
<tr>
<th>[,1]</th>
<th>[,2]</th>
<th>[,3]</th>
</tr>
</thead>
<tbody>
<tr>
<td>[1,]</td>
<td>1.0000000</td>
<td>-0.0034327</td>
</tr>
<tr>
<td>[2,]</td>
<td>-0.0034327</td>
<td>1.0000000</td>
</tr>
<tr>
<td>[3,]</td>
<td>-0.0034327</td>
<td>-0.0034327</td>
</tr>
</tbody>
</table>
```
> dep.gee3 <-
gee(response == "normal" ~ (severity + drug)*time,
    id=subject, data=depression, family=binomial)

(Intercept) severity severe
  0.073547    -1.528703
drug new    time
-0.055304    0.358728
severity severe:time drug new:time
  0.235006    1.001094

> round(coef(summary(dep.gee3))[, "Robust z"], 2)

(Intercept) severity severe
  0.37       -6.55
drug new    time
-0.24       2.31
severity severe:time drug new:time
  1.29       5.33

Note:
▶ GEE have been generalized to multivariate response, but not much available in software.

SAS’s PROC GENMOD will do GEE for cumulative logit models, but only with independence working correlations (check this).

See insomnia study in Section 9.3.2 for an example.

▶ Missing data is not uncommon and can be very problematic unless missing completely at random (MCAR): missingness unrelated to response or any explanatory variables.

Missing at random (MAR) means missingness unrelated to response after controlling for explanatory variables. Methods exist to handle this and some other forms of missingness.

Ignoring missing data leads to biased estimates.

Analyzing Repeated Measurements and Other Clustered Data

Observations \((Y_1, Y_2, \ldots, Y_T)\) (e.g., \(T\) times).

1. Marginal Models (Ch. 9)

Simultaneously model each (marginal) \(E(Y_t)\), \(t = 1, \ldots, T\).
Get standard errors that account for the actual dependence using method such as GEE (generalized estimating equations).

Ex. Binary response \(Y_t = 0\ or 1, t = 1, 2\) (matched pairs).

\[ E(Y_t) = Pr(Y_t = 1) \]

Model: \( \text{logit}[Pr(Y_t = 1)] = \alpha + \beta x_t \),

\( x_t \) = value of explan. var. for \( t^{th} \) obs.

Depression example (matched triplets): some explanatory variables constant across \( t \) (severity and drug), others vary (time).
2. Random Effects Models (Ch. 10)

Account for having multiple responses per subject (or “cluster”) by putting a subject term in model.

Ex. Binary response $Y_t = 0$ or $1$.

Let $Y_{it} =$ response by subject $i$ at time $t$.

Model: logit $[\Pr(Y_{it} = 1)] = \alpha_i + \beta x_{it}, \quad t = 1, \ldots, T$

Intercept $\alpha_i$ varies by subject.

large positive $\alpha_i \implies$ large $\Pr(Y_{it} = 1)$ each $t$

large negative $\alpha_i \implies$ small $\Pr(Y_{it} = 1)$ each $t$

Heterogeneous population $\implies$ highly variable $\{\alpha_i\}$.

Problem: number of parameters $>$ number of subjects.

Solution: treat $\{\alpha_i\}$ as random rather than parameters (fixed).

Assume a distribution for $\{\alpha_i\}$, e.g., $\alpha_i \sim N(\alpha, \sigma^2)$, i.e.,

$\alpha_i = \alpha + u_i, \quad u_i \sim N(0, \sigma^2)$

where $\alpha$ is a fixed, unknown parameter.

Model: logit $[\Pr(Y_{it} = 1)] = u_i + \alpha + \beta x_{it}$

$\{u_i\}$ are random effects.

Parameters $\alpha$ and $\beta$ are fixed effects.

$Y_{i1}, Y_{i2}, \ldots, Y_{iT}$ conditionally independent given $u_i$.

But marginally dependent: responses within subject more alike than between subjects.

A generalized linear mixed model (GLMM) is a GLM with both fixed and random effects.

Note that random effects $\{u_i\}$ are unobserved (not data).

Software must “integrate out” $\{u_i\}$ to get likelihood fcn, MLEs $\hat{\alpha}, \hat{\beta}$, SE's. Also estimate $\sigma^2$ and can “predict” $\{u_i\}$.

Example (Depression Study)

<table>
<thead>
<tr>
<th>Time</th>
<th>Response Pattern</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>A A A A N N N N</td>
</tr>
<tr>
<td>1</td>
<td>A A N N A A N N</td>
</tr>
<tr>
<td>2</td>
<td>A N A N A N A N</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Severity</th>
<th>Drug</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mild</td>
<td>Std</td>
</tr>
<tr>
<td>6 15 14 3 9 13 16</td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>0 9 2 22 0 6 0 31</td>
</tr>
<tr>
<td>Severe</td>
<td>Std</td>
</tr>
<tr>
<td>28 27 15 9 9 8 2 2</td>
<td></td>
</tr>
<tr>
<td>New</td>
<td>6 32 5 31 2 5 2 7</td>
</tr>
</tbody>
</table>

Previously used GEE to fit “marginal model”

$\text{logit} [\Pr(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt,$

$Y_t = 1$ (normal); $s = 0, 1$ (initial diagnosis.: mild vs severe); $t = \log_2(\text{wks on trt}); \quad d = 0, 1$ (drug: std vs new).
Now use ML to fit "random effects model" (a.k.a., "mixed model")

\[
\logit(\Pr(Y_{it} = 1)) = u_i + \alpha + \beta_1s + \beta_2d + \beta_3t + \beta_4dt.
\]

Assume \{u_i\} indep. \(N(0, \sigma^2)\). Need to estimate \(\sigma^2\).

MLEs: \(\hat{\sigma} = 0.057\) (\(\hat{\sigma}^2 = 0.00323\)), \(\hat{\alpha} = -0.028\)
\(\hat{\beta}_1 = -1.31\) \(\hat{\beta}_2 = -0.06\) \(\hat{\beta}_3 = 0.48\) \(\hat{\beta}_4 = 1.02\)

```r
> library(lme4)
> data(depression)
> head(depression)
subject severity drug time response
1 1 mild std 0 normal
2 1 mild std 1 normal
3 1 mild std 2 normal
4 2 mild std 0 normal
5 2 mild std 1 normal
6 2 mild std 2 normal

> dep.lme4.1 <-
glmer((response == "normal") ~ severity + drug*time + (1 | subject),
family = binomial, data = depression)

> summary(dep.lme4.1)
```

Generalized linear mixed model fit by the Laplace approximation
Formula: (response == "normal") ~ severity + drug * time + (1
| subject) Data: depression

AIC BIC logLik deviance
1174 1204 -581 1162

Random effects:
Groups Name Variance Std.Dev.
subject (Intercept) 0.00323 0.0568
Number of obs: 1020, groups: subject, 340

Fixed effects:

|                | Estimate | Std. Error | z value | Pr(>|z|) |
|----------------|----------|------------|---------|----------|
| (Intercept)    | -0.0280  | 0.1640     | -0.17   | 0.86     |
| severitysevere | -1.3149  | 0.1466     | -8.97   | < 2e-16  |
| drugnew        | -0.0597  | 0.2223     | -0.27   | 0.79     |
| time           | 0.4827   | 0.1148     | 4.21    | 2.6e-05  |
| drugnew:time   | 1.0181   | 0.1888     | 5.39    | 7.0e-08  |
Correlation of Fixed Effects:
(Intr) svrtys drugnw time
severitysvr -0.403
drugnew -0.614 -0.010
time -0.679 -0.094 0.530
drugnew:tim 0.468 -0.079 -0.750 -0.595

In this example, GLMM and GEE estimates and SE’s for fixed effects are nearly identical:

<table>
<thead>
<tr>
<th></th>
<th>GLMM</th>
<th>GEE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Est</td>
<td>SE</td>
</tr>
<tr>
<td>alpha</td>
<td>-0.03</td>
<td>0.16</td>
</tr>
<tr>
<td>beta.1</td>
<td>-1.31</td>
<td>0.15</td>
</tr>
<tr>
<td>beta.2</td>
<td>-0.06</td>
<td>0.22</td>
</tr>
<tr>
<td>beta.3</td>
<td>0.48</td>
<td>0.11</td>
</tr>
<tr>
<td>beta.4</td>
<td>1.02</td>
<td>0.19</td>
</tr>
</tbody>
</table>

Why? Because there appears to be little correlation between repeated measurements on subjects:

▶ $\hat{\rho} = -0.003 \approx 0$ in GEE with exchangeable working correlation.

▶ $\hat{\sigma} = 0.057 \approx 0$ in GLMM. According to model, 95% of all individuals will have $u_i$ between $\pm 1.96\sigma$. Estimate this as $\pm 1.96(0.057) = \pm 0.11$. But $e^{-0.11} = 0.89$ and $e^{0.11} = 1.12$, so effect of $u_i$ on odds is estimated to be small for most subjects.

Note:

▶ When $\hat{\sigma} = 0$, estimates and SEs same as treating repeated observations as independent.

▶ When $\hat{\sigma}$ is large, estimated $\beta$s from random effects logit model usually larger than from marginal model. They are estimating different things: see figure below. (Details in Sec. 10.1.4 of text.)
Ch 7: Loglinear Models

- Logistic regression and other models in Ch 3–6, 8–10 distinguish between a response variable $Y$ and explanatory vars $x_1, x_2$, etc.

- Loglinear models for contingency tables treat all variables as response variables, like multivariate analysis.

**Ex:** Survey of high school seniors (see text):

- $Y_1$: used alcohol? (yes, no)
- $Y_2$: cigarettes? (yes, no)
- $Y_3$: marijuana? (yes, no)

Interested in patterns of dependence and independence among the three variables:

- Any variables independent?
- Strength of associations?
- Interactions?

Loglinear models treat cell counts as Poisson and use log link fcn.

**Motivation:** In $I \times J$ table, $X$ and $Y$ are independent if

$$Pr(X = i, Y = j) = Pr(X = i) \cdot Pr(Y = j) \quad \text{for all } i, j$$

i.e., $\pi_{ij} = \pi_i + \pi_j$

For expected cell frequencies,

$$\mu_{ij} = n \pi_{ij} \quad \text{(general form)}$$

$$= n \pi_i + \pi_j \quad \text{(under independence)}$$

$$\implies \log(\mu_{ij}) = \lambda + \lambda_X^i + \lambda_Y^j$$

$\lambda_X^i$: effect of classification in row $i$  \hspace{1cm} (I−1 nonredundant parameters)

$\lambda_Y^j$: effect of classification in col $j$ \hspace{1cm} (J−1 nonredundant parameters)

**Loglinear model of independence:** treats $X$ and $Y$ symmetrically. Unlike, e.g., logistic regr where $Y = \text{response, } X = \text{explanatory.}$

**Note:** For a Poisson loglinear model,

$$df = \text{number of Poisson counts} - \text{number of parameters}$$

Here number of Poisson counts = number cells in table.

Think of dummy variables for each variable. Number of dummies is one less than number of levels of variable. Products of dummy variables correspond to “interaction” terms.
For an $I \times J$ contingency table:

- **Indep. model:** $\log(\mu_{ij}) = \lambda + \lambda_X^i + \lambda_Y^j$ (df = $(I-1)(J-1)$)
  
  no. cells = $IJ$

  no. parameters = $1 + (I-1) + (J-1) = I + J - 1$

  df = $IJ - (I+J-1) = (I-1)(J-1)$

- **Saturated model:** $\log(\mu_{ij}) = \lambda + \lambda_X^i + \lambda_Y^j + \lambda_{XY}^{ij}$ (df = 0)

  Parameter Nonredundant

  $\lambda$ 1

  $\lambda_X^i$ $I-1$

  $\lambda_Y^j$ $J-1$

  $\lambda_{XY}^{ij}$ $(I-1)(J-1)$

  Total: $IJ$

Note: Log-odds-ratio comparing levels $i$ and $i'$ of $X$ and $j$ and $j'$ of $Y$ is

$$\log\left( \frac{\mu_{ij}\mu_{i'j'}}{\mu_{ij'}\mu_{i'j}} \right) = \log \mu_{ij} + \log \mu_{i'j'} - \log \mu_{ij'} - \log \mu_{i'j}$$

$$= (\lambda + \lambda_X^i + \lambda_Y^j + \lambda_{XY}^{ij}) + (\lambda + \lambda_X^{i'} + \lambda_Y^{j'} + \lambda_{XY}^{i'j'}) - (\lambda + \lambda_X^i + \lambda_Y^{j'} + \lambda_{XY}^{ij'}) - (\lambda + \lambda_X^{i'} + \lambda_Y^j + \lambda_{XY}^{i'j})$$

$$= \lambda_{XY}^{ij} + \lambda_{XY}^{i'j'} - \lambda_{XY}^{ij'} - \lambda_{XY}^{i'j}.$$ 

For the independence model this is 0, and the odds-ratio is $e^0 = 1$.

For the saturated model, the odds-ratio, expressed in terms of the parameters of the loglinear model, is

$$\frac{\mu_{ij}\mu_{i'j'}}{\mu_{ij'}\mu_{i'j}} = \exp\{\lambda_{XY}^{ij} + \lambda_{XY}^{i'j'} - \lambda_{XY}^{ij'} - \lambda_{XY}^{i'j}\}.$$

Substituting the MLEs of the saturated model (perfect fit) just reproduces the empirical odds ratio $\frac{n_{ij}/n_{i'j'}}{n_{ij'}/n_{i'j'}}$.

Example (Income and Job Satisfaction)

<table>
<thead>
<tr>
<th>Income</th>
<th>Dissat</th>
<th>Little</th>
<th>Moderate</th>
<th>Very</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5K</td>
<td>2</td>
<td>4</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>5K–15K</td>
<td>2</td>
<td>6</td>
<td>22</td>
<td>4</td>
</tr>
<tr>
<td>15K–25K</td>
<td>0</td>
<td>1</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>0</td>
<td>3</td>
<td>13</td>
<td>8</td>
</tr>
</tbody>
</table>

Originally used Pearson's chisquare test: $X^2 = 11.5$, df = 9 ($G^2 = 13.5$).

With income scores $x = 3, 10, 20, 35$, used VGAM package to fit baseline category logit model

$$\log\left( \frac{\pi_j}{\pi_4} \right) = \alpha_j + \beta_j x, \quad j = 1, 2, 3.$$ 

and later, cumulative logit model

$$\logit[\Pr(Y \leq j)] = \alpha_j + \beta x, \quad j = 1, 2, 3.$$
Using dummy variables, the model

\[ \log(\mu_{ij}) = \lambda + \lambda_1 I_i + \lambda_2 S_j \]

can be expressed as

\[ \log(\mu_{ij}) = \lambda + \lambda_1 I_1 z_1 + \lambda_2 I_2 z_2 + \lambda_3 I_3 z_3 + \lambda_4 S_1 w_1 + \lambda_5 S_2 w_2 + \lambda_6 S_3 w_3 \]

where we take \( \lambda_4 = \lambda_5 = 0 \) and

\[
\begin{align*}
z_1 &= \begin{cases} 1, & \text{inc} < 5K, \\ 0, & \text{otherwise}, \end{cases} \\
z_2 &= \begin{cases} 1, & 5K \leq \text{inc} < 15K, \\ 0, & \text{otherwise}, \end{cases} \\
z_3 &= \begin{cases} 1, & 15K \leq \text{inc} < 25K, \\ 0, & \text{otherwise}, \end{cases} \\
w_1 &= \begin{cases} 1, & \text{dissatisfied} \\ 0, & \text{otherwise}, \end{cases} \\
w_2 &= \begin{cases} 1, & \text{little dissat.} \\ 0, & \text{otherwise}, \end{cases} \\
w_3 &= \begin{cases} 1, & \text{moderately sat.} \\ 0, & \text{otherwise}, \end{cases}
\end{align*}
\]

```r
> jobsat <- matrix(c(2,2,0,0, 4,6,1,3, 13,22,15,13, 3,4,8,8), nrow=4)
> dimnames(jobsat) <- list(income = c("<5K","5K-15K","15K-25K",">25K"),
                        satis = c("VeryD","LittleD","ModerateS","VeryS"))
> jobsat <- as.table(jobsat)
> jobsat

<table>
<thead>
<tr>
<th>income</th>
<th>satis</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5K</td>
<td>VeryD</td>
<td>2</td>
</tr>
<tr>
<td>5K-15K</td>
<td>VeryD</td>
<td>2</td>
</tr>
<tr>
<td>15K-25K</td>
<td>VeryD</td>
<td>0</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>VeryD</td>
<td>0</td>
</tr>
<tr>
<td>&lt;5K</td>
<td>LittleD</td>
<td>4</td>
</tr>
<tr>
<td>5K-15K</td>
<td>LittleD</td>
<td>6</td>
</tr>
<tr>
<td>15K-25K</td>
<td>LittleD</td>
<td>1</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>LittleD</td>
<td>3</td>
</tr>
<tr>
<td>&lt;5K</td>
<td>ModerateS</td>
<td>13</td>
</tr>
<tr>
<td>5K-15K</td>
<td>ModerateS</td>
<td>22</td>
</tr>
<tr>
<td>15K-25K</td>
<td>ModerateS</td>
<td>15</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>ModerateS</td>
<td>13</td>
</tr>
<tr>
<td>&lt;5K</td>
<td>VeryS</td>
<td>3</td>
</tr>
<tr>
<td>5K-15K</td>
<td>VeryS</td>
<td>4</td>
</tr>
<tr>
<td>15K-25K</td>
<td>VeryS</td>
<td>8</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>VeryS</td>
<td>8</td>
</tr>
</tbody>
</table>
```

```r
> jobsat.df <- as.data.frame(jobsat)

<table>
<thead>
<tr>
<th>income</th>
<th>satis</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;5K</td>
<td>VeryD</td>
<td>2</td>
</tr>
<tr>
<td>5K-15K</td>
<td>VeryD</td>
<td>2</td>
</tr>
<tr>
<td>15K-25K</td>
<td>VeryD</td>
<td>0</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>VeryD</td>
<td>0</td>
</tr>
<tr>
<td>&lt;5K</td>
<td>LittleD</td>
<td>4</td>
</tr>
<tr>
<td>5K-15K</td>
<td>LittleD</td>
<td>6</td>
</tr>
<tr>
<td>15K-25K</td>
<td>LittleD</td>
<td>1</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>LittleD</td>
<td>3</td>
</tr>
<tr>
<td>&lt;5K</td>
<td>ModerateS</td>
<td>13</td>
</tr>
<tr>
<td>5K-15K</td>
<td>ModerateS</td>
<td>22</td>
</tr>
<tr>
<td>15K-25K</td>
<td>ModerateS</td>
<td>15</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>ModerateS</td>
<td>13</td>
</tr>
<tr>
<td>&lt;5K</td>
<td>VeryS</td>
<td>3</td>
</tr>
<tr>
<td>5K-15K</td>
<td>VeryS</td>
<td>4</td>
</tr>
<tr>
<td>15K-25K</td>
<td>VeryS</td>
<td>8</td>
</tr>
<tr>
<td>&gt;25K</td>
<td>VeryS</td>
<td>8</td>
</tr>
</tbody>
</table>
```
```r
> levels(jobsat.df$income)
[1] "<5K" "5K-15K" "15K-25K" ">25K"
> levels(jobsat.df$satis)
[1] "VeryD" "LittleD" "ModerateS" "VeryS"
> jobsat.df <-
transform(jobsat.df,
  income = relevel(income, ">25K"),
  satis = relevel(satis, "VeryS"))
> jobsat.indep <-
glm(Freq ~ income + satis, family=poisson,
data=jobsat.df)

> summary(jobsat.indep)
Call:
glm(formula = Freq ~ income + satis, family = poisson, data = jobsat.df)
Deviance Residuals:
       Min        1Q   Median        3Q       Max
-1.4547 -1.0228  0.0152  0.5880  1.0862
Coefficients:
                                    Estimate Std. Error z value Pr(>|z|)
(Intercept)                          1.67e+00  2.75e-01   6.07  1.3e-09
income<5K                            -8.70e-02  2.95e-01  -0.29  0.7682
income5K-15K                         3.48e-01  2.67e-01   1.31  0.1914
income15K-25K                        3.91e-15  2.89e-01  -0.00  1.0000
satisVeryD                           -1.75e+00  5.42e-01  -3.23  0.0012
satisLittleD                         -4.96e-01  3.39e-01  -1.46  0.1431
satisModerateS                       1.01e+00  2.44e-01   4.14  3.5e-05

(Dispersion parameter for poisson family taken to be 1)

   Null deviance: 90.242 on 15 degrees of freedom
   Residual deviance: 13.467 on 9 degrees of freedom
   AIC: 77.07

Number of Fisher Scoring iterations: 5
NA
> chisqstat(jobsat.indep)
[1] 11.524
```
> jobsat.saturated <- update(jobsat.indep, . ~ income*satis)
> anova(jobsat.indep, jobsat.saturated, test="Chisq")

Analysis of Deviance Table

Model 1: Freq ~ income + satis
Model 2: Freq ~ income + satis + income:satis

| Resid. Df | Resid. Dev | Df | Deviance | P(>|Chi|) |
|-----------|------------|----|----------|----------|
| 1         | 9          | 1  | 13.5     |          |
| 2         | 0          | 9  | 0.0      | 13.5     | 0.14     |

Loglinear Models for Three-Way Tables

Here two-factor terms represent conditional log odds ratios at a fixed level of the third variable.

Ex. 2 \times 2 \times 2 table. Consider the loglinear model

$$\log(\mu_{ijk}) = \lambda + \lambda_X^i + \lambda_Y^j + \lambda_Z^k + \lambda_{XY}^{ij} + \lambda_{XZ}^{ik} + \lambda_{YZ}^{jk}. $$

Called the model of X-Y conditional independence; denoted (XZ, YZ).

- X and Y are conditionally independent, given Z:
  $$\log(\theta_{XY|k}) = 0 \implies \theta_{XY|k} = 1$$

- the X-Z odds ratio is the same at all levels of Y:
  $$\log(\theta_{X|jZ}) = \frac{\lambda_X^Z + \lambda_{XZ}^{1Z} - \lambda_{XZ}^{2Z}}{\lambda_{Y}^Z}$$

  (does not depend on )

Similarly, Y-Z odds ratio same at all levels of X. Model has no three-factor interaction.

Ex. Consider the loglinear model

$$\log(\mu_{ijk}) = \lambda + \lambda_X^i + \lambda_Y^j + \lambda_Z^k + \lambda_{XY}^{ij} + \lambda_{XZ}^{ik} + \lambda_{YZ}^{jk}. $$

Each pair of variables is conditionally dependent, but association (as measured by odds ratios) is the same at all levels of third variable.

Called the model of homogeneous association (or model of no three-factor interaction; denoted (XY, XZ, YZ).
Ex. Survey of 2276 high school seniors.

```r
> teens <-
array(c(911, 44, 3, 2, 538, 456, 43, 279),
dim = c(2, 2, 2),
dimnames = list(cigs=c("yes", "no"),
              alc=c("yes", "no"), mj=c("yes", "no")))
> ## Next line just for Table 7.4. Not required.
> teens <- aperm(teens, c(3, 1, 2))
> teens <- as.table(teens)
> ftable(teens, row.vars=c("alc", "cigs"))

mj  yes  no
alc  cigs
yes yes 911 538
   no  44 456
no yes  3  43
  no  2 279
```

```r
> teens.df <- as.data.frame(teens)
> teens.df
mj  cigs  alc  Freq
1  yes  yes  yes  911
2  no  yes  yes  538
3  yes  no  yes   44
4  no  no  yes  456
5  yes  yes  no   3
6  no  yes  no  43
7  yes  no  no   2
8  no  no  no  279
```

```r
> teens.AC.AM.CM <-
glm(Freq ~ alc*cigs + alc*mj + cigs*mj,
    family=poisson, data=teens.df)
> ### Another way:
> ## teens.AC.AM.CM <-
> ## glm(Freq ~ alc*cigs*mj - alc:cigs:mj,
> ##      family=poisson, data=teens.df)
> summary(teens.AC.AM.CM)
```

Call:
```
glm(formula = Freq ~ alc * cigs + alc * mj + cigs * mj,
     family = poisson, data = teens.df)
```

Deviance Residuals:
```
    1     2     3     4     5     6
  0.0204 -0.0266 -0.0926  0.0289 -0.3343  0.0945
```
```
         7
  0.4913
```
```
         8
 -0.0369
```
Coefficients:

| Estimate | Std. Error | z value | Pr(>|z|) |
|----------|------------|---------|----------|
| (Intercept) | 5.6334 | 0.0597 | 94.36 | < 2e-16 |
| alcyes | 0.4877 | 0.0758 | 6.44 | 1.2e-10 |
| cigysyes | -1.8867 | 0.1627 | -11.60 | < 2e-16 |
| mjesyes | -5.3090 | 0.4752 | -11.17 | < 2e-16 |
| alcyes:cigsyes | 2.0545 | 0.1741 | 11.80 | < 2e-16 |
| alcyes:mjyes | 2.9860 | 0.4647 | 6.43 | 1.3e-10 |
| cigysyes:mjyes | 2.8479 | 0.1638 | 17.38 | < 2e-16 |

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 2851.46098 on 7 degrees of freedom
Residual deviance: 0.37399 on 1 degrees of freedom
AIC: 63.42
Number of Fisher Scoring iterations: 4

The (AC,AM,CM) model fits well:

\[ G^2 = 0.37 \text{ (and } X^2 = 0.4) \text{ on 1 df.} \]

\[ \text{df.residual(teens.AC.AM.CM)} \]
[1] 1

\[ \text{deviance(teens.AC.AM.CM)} \]
[1] 0.37399

\[ \text{chisqstat(teens.AC.AM.CM)} \]
[1] 0.4011

Note: As a LRT, goodness-of-fit on previous slide is comparing to
saturated model.

\[ \text{teens.ACM <- update(teens.AC.AM.CM, . ~ alc*cigs*mj)} \]
\[ \text{anova(teens.AC.AM.CM, teens.ACM, test="Chisq")} \]

Analysis of Deviance Table

Model 1: Freq ~ alc * cigs + alc * mj + cigs * mj
Model 2: Freq ~ alc + cigs + mj + alc:cigs + alc:mj + cigs:mj

| Resid. Df | Resid. Dev | Dev Df | Deviance | P(>|Chi|) |
|-----------|------------|--------|----------|----------|
| 1          | 0.374      | 1      | 0.374    | 0.54     |
And none of the interaction terms can be dropped:

```r
> drop1(teens.AC.AM.CM, test="Chisq")
```

Single term deletions

Model:

Freq ~ alc * cigs + alc * mj + cigs * mj  

<table>
<thead>
<tr>
<th>Df</th>
<th>Deviance</th>
<th>AIC</th>
<th>LRT</th>
<th>Pr(Chi)</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;none&gt;</td>
<td>0</td>
<td>63</td>
<td></td>
<td></td>
</tr>
<tr>
<td>alc:cigs</td>
<td>1</td>
<td>188</td>
<td>249</td>
<td>187</td>
</tr>
<tr>
<td>alc:mj</td>
<td>1</td>
<td>92</td>
<td>153</td>
<td>92</td>
</tr>
<tr>
<td>cigs:mj</td>
<td>1</td>
<td>497</td>
<td>558</td>
<td>497</td>
</tr>
</tbody>
</table>

Note: `drop1()` does LRTs comparing to simpler models. Test statistic is the usual

\[-2(L_0 - L_1) = \text{deviance}_0 - \text{deviance}_1\]

and df is difference in number of nonredundant parameters.

E.g., to test for conditional independence of A and C given M:

```r
> teens.AM.CM <- update(teens.AC.AM.CM, . ~ alc*mj + cigs*mj)
> anova(teens.AM.CM, teens.AC.AM.CM, test="Chisq")
```

Analysis of Deviance Table

| Resid. Df | Resid. Dev | Df | Deviance | P(>|Chi|) |
|----------|------------|----|----------|---------|
| 1 | 2 | 187.8 | | |
| 2 | 1 | 0.4 | 1 | 187 | <2e-16 |

Table 7.4 gives fitted values for several different models fit to these data.

```r
> teens.AM.CM <- update(teens.AC.AM.CM, . ~ alc*mj + cigs*mj)
> teens.AC.M <- update(teens.AC.AM.CM, . ~ alc*cigs + mj)
> teens.A.C.M <- update(teens.AC.AM.CM, . ~ alc + cigs + mj)
> teens.ACM <- update(teens.AC.AM.CM, . ~ alc*cigs*mj)
> table.7.4 <- data.frame(predict(teens.A.C.M, type="response"))
> table.7.4 <- cbind(table.7.4, predict(teens.AC.M, type="response"))
> table.7.4 <- cbind(table.7.4, predict(teens.AM.CN, type="response"))
> table.7.4 <- cbind(table.7.4, predict(teens.AC.AM.CM, type="response"))
> table.7.4 <- cbind(table.7.4, predict(teens.ACM, type="response"))
```
In (AC,AM,CM) model, AC odds-ratio is the same at each level of M. With 1 = yes and 2 = no for each var., est.'d cond. AC odds ratio is

$$\hat{\mu}_{11k}/\hat{\mu}_{22k} = \exp(\hat{\lambda}_{11} + \hat{\lambda}_{22} - \hat{\lambda}_{12} - \hat{\lambda}_{21}) = e^{2.0545} = 7.8$$

A 95% CI is

$$e^{2.05 \pm (1.96)(0.174)} = (e^{1.71}, e^{2.40}) = (5.5, 11.0)$$

The commons odds-ratio is reflected in the fitted values for the model:

$$\frac{(910)(1.38)}{44.6(3.62)} = 7.8 \quad \frac{(539)(280)}{455(42.4)} = 7.8$$

Similar results hold for AM and CM conditional odds-ratios in this model.
In (AM,CM) model, $\lambda_{ij}^{TM} = 0$, and conditional AC odds-ratio (given M) is $e^0 = 1$ at each level of M, i.e., A and C are conditionally indep. given M.

Again, this is reflected in the fitted values for this model.

\[
\frac{(909)(0.24)}{(45.8)(4.76)} = 1 \quad \frac{(439)(180)}{(555)(142)} = 1
\]

The AM odds-ratio is not 1, but it is the same at each level of C:

\[
\frac{(909)(142)}{(439)(4.76)} = 61.87 \quad \frac{(45.8)(180)}{(555)(0.24)} = 61.87
\]

Similarly, the CM odds-ratio is the same at each level of A:

\[
\frac{(909)(555)}{(439)(45.8)} = 25.14 \quad \frac{(4.76)(180)}{(142)(0.24)} = 25.14
\]

Standardized residuals may help understand lack of fit.
Text uses standardized Pearson residuals.
\texttt{rstandard()} computes only standardized deviance resids.
\texttt{rstandard2()} computes in course package does both.
Next version of R will incorporate this into \texttt{rstandard()}.

See Section 7.2.2 for example and discussion.

Note:

- Loglinear models extend to any number of dimensions.
- Loglinear models treat all variables symmetrically.

Logistic regression models treat Y as response and other variables as explanatory. More natural approach when there is a single response.
Exam 2: Time and Place

Tuesday, Apr 19, 2011
8:30 a.m. – 10:25 a.m.
Room 018 Matherly Hall (MAT 018)

Exam 2 Review: Building Logistic Regression Models

Model Selection

- LR tests to compare nested models.
  - \(-2(L_0 - L_1) = \text{deviance}_0 - \text{deviance}_1\)
  - df = diff. in no. nonredundant params = diff. in residual df's
  - Wald tests ok for single parameter, but LR generally preferred.
- AIC.
- Measures of predictive power.
  - Classification table (a.k.a., confusion matrix).
  - ROC curves (not discussed).
  - Correlation between Y and \(\hat{\pi}\) (meh).

- Multicollinearity (correlated explanatory variables) problematic (big SEs, hard to pick model).
- Automated backward elimination or forward selection generally not recommended (multiple testing).
- Parsimony (simplicity) good, but use care and judgement in choosing model. Keep research questions and subject area expertise in mind.
Exam 2 Review: Building Logistic Regression Models

Model Checking

- Goodness-of-fit tests
  - $\chi^2$ (chi-square statistic) or $G^2$ (deviance)
  - Compares fitted model to saturated model (e.g., the data).
  - df = num. binomials − num. model params
  - Use for contingency tables with few expected counts < 5.
  - For “sparse” data, chi-square approx. poor for $\chi^2$ and $G^2$.
    - May try grouping observations to reduce sparsity:
      - by partitioning numeric predictor(s). E.g., for horseshoe crab width,
        | Range   | 20–24 | 24–26 | 26–28 | 28–34 |
        | Score   | 22    | 25    | 27    | 31    |
      - by partitioning $\hat{\pi}$ (Hosmer-Lemeshow)

- Use LR test to check whether fit improves significantly when other predictors or interactions are added.
  - LR test ok even when deviance alone invalid for gof (sparse data).

- Standardized residuals.
  - Residual standardized by dividing by SE.
  - Examine where lack of fit occurs.
  - Values $< -2$ or $> 2$ suggest lack of fit in small tables.
  - Values $< -3$ or $> 3$ very strong evidence for lack of fit.

- Sparse data and/or too many terms in model may lead to
  - infinite MLEs
  - very large SEs
  - bad Wald tests and CIs

Exam 2 Review: Baseline-Category Logit Model

For response $Y$ with $J > 2$ categories.

$\pi_j = \Pr(Y = j), \quad j = 1, \ldots, J.$

Model:

$$\log \left( \frac{\pi_j}{\pi_j} \right) = \alpha_j + \beta_j x, \quad j = 1, 2, \ldots, J - 1.$$ 

Separate set of parameters $(\alpha_j, \beta_j)$ for each logit.

- Used for nominal response.
- Ok for ordinal response, but ignores ordering.
- Choice of category for baseline not important.
Exam 2 Review: Cumulative Logit Model

For ordinal response \( Y \) with \( J > 2 \) categories.

Model:

\[
\logit[\Pr(Y \leq j)] = \alpha_j + \beta x,
\]

\( j = 1, \ldots, J - 1 \).

- Separate intercept \( \alpha_j \) for each cumulative logit
- Same slope \( \beta \) for each cumulative logit
- \( e^\beta \) = multiplicative effect of 1-unit increase in \( x \) on odds that \( Y \leq j \) (instead of \( Y > j \)).
- Reversing ordering of \( Y \) changes sign of \( \beta \).
- Usual inferential methods apply. Takes advantage of ordering in \( Y \).

Exam 2 Review: Models for Matched Pairs

McNemar's Test

Two binary responses from each subject or matched pair. E.g.,

- measure response at two different times
- husband and wife answer same question

Simplest kind of dependent response.

<table>
<thead>
<tr>
<th>Resp 1</th>
<th>Resp 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>F</td>
<td>( n_{12} )</td>
</tr>
<tr>
<td>( n_{+1} )</td>
<td>( n_{2+} )</td>
</tr>
</tbody>
</table>

\( n \)
Exam 2 Review: Measuring Agreement
Suppose rating on a 4-point scale.

Cohen’s kappa measures agreement as departure from independence in direction of perfect agreement:

\[ \kappa = \frac{\Pr(\text{agree}) - \Pr(\text{agree}\mid\text{indep})}{1 - \Pr(\text{agree}\mid\text{indep})} = \frac{\sum \pi_{ii} - \sum \pi_{i+} \pi_{+i}}{1 - \sum \pi_{i+} \pi_{+i}} \]

- \( \kappa = 0 \) if agreement only equals that expected under independence.
- \( \kappa = 1 \) if perfect agreement.

Exam 2 Review: Generalized Estimating Equations (GEE)

GEE used for correlated responses (repeated measurements/clustered data).

- Specify (marginal) model for individual responses in usual way.
- Select a “working correlation” matrix (independence, exchangeable, etc).
- GEE parameter estimates consistent even if correlation structure.
- (Robust) standard errors adjusted to reflect actual observed dependence, even if form of working correlation is wrong.
- “Quasi-likelihood” method. No particular form assumed for joint distribution of responses.
Random (or mixed) effects models also useful for correlated responses.

- Add subject specific terms to model.
- Subject specific terms modeled as unobserved random variables (random effects).
- Usually assume random effects follow $N(0, \sigma^2)$ distribution, $\sigma^2$ unknown.
- $\sigma^2 = 0$ means responses independent (not usually expected with repeated measures).

In a repeated measures context:

- GLMM is a conditional (subject specific) approach: fixed effect $\beta$ represents effect of change in $x$ on an individual subject's response.
- GEE models marginal effects: $\beta$ represents population average effect of changing $x$.
- When $\sigma^2$ large in GLMM (or responses highly correlated in GEE), fixed effects coefficients ($\beta$'s) in conditional model (GLMM) usually larger in magnitude than in marginal model (GEE).
- GLMM completely specifies joint distribution of responses: likelihood methods apply.
- GEE does not assume a specific form for the distribution of responses: not a likelihood-based method.

Used to study dependence structure in contingency tables.

- Multivariate analysis for contingency tables:
  - All variables treated on an equal footing.
  - No distinction between response and explanatory variables.
- Loglinear models are fit by treating cell frequencies as independent Poisson responses.

E.g., for $I \times J \times K$ three-way table:

- Variable $X$ has $I$ levels, $Y$ has $J$ levels, $Z$ has $K$ levels.
- Treat $n_{ijk}, 1 \leq i \leq I, 1 \leq j \leq J, 1 \leq k \leq K$ as independent Poisson counts.
- Fit Poisson GLM with log link on $\mu_{ijk}$, with $n_{ijk}$ as response, and $X, Y, Z$ as predictors, generally with interactions.
Use LRT to compare nested models.

Use $X^2$ or $G^2$ to test goodness-of-fit.

Looking for simplest model that explains data adequately.

Don’t depend only on formal tests: statistically significant terms may be practically unimportant (see Section 7.2.8).

Some models for $I \times J \times K$ three-way table:

- (XYZ)
  \[
  \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ijk}^{XYZ}
  \]
  - Saturated model, fits cell counts perfectly.
  - Residual df = 0.

- (XY, XZ, YZ)
  \[
  \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}
  \]
  - homogeneous assoc (no 3-factor interaction): conditional odds-ratio for any pair of variables is constant across levels of 3rd var.
  - Residual df = $(I-1)(J-1)(K-1)$.

- (XZ, YZ)
  \[
  \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}
  \]
  - X and Y conditionally independent given Z.
  - Homogeneous XZ association. Homogeneous YZ association.
  - Residual df = $(I-1)(J-1)K$.

- (XY, Z)
  \[
  \log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}
  \]
  - Z independent of X and Y.
  - Residual df = $(IJ-1)(K-1)$.