# Chapter 1: Introduction
## 1.1 Categorical Response Data

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

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

### Example

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

$$x\text{'s} : \underbrace{\text{income,}}_{\text{continuous}} \underbrace{\text{gender, race,}}_{\text{categorical}} \underbrace{\text{education}}_{?}$$

## Two Types of Categorical Variables

Nominal: unordered categories

Ordinal: ordered categories

**Example**

Nominal
- transport to work (car, bus, bicycle, walk, other)
- favorite music (rock, hiphop, pop, classical, jazz, country, folk)

Ordinal
- patient condition (excellent, good, fair, poor)
- government spending (too high, about right, too low)

We pay special attention to

Binary variables: success or failure

for which nominal-ordinal distinction is unimportant.

## 1.2 Probability Distributions for Categorical Data

For categorical response data, the *binomial* distribution (and its generalization, the *multinomial* distribution) plays a role similar to that of the normal distribution for continuous responses.

### Binomial Distribution

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

$$
\begin{aligned}
P(y) &= \mathrm{Pr}(Y = y) \\
&= \frac{n!}{y!(n-y)!}\,\pi^{y}(1-\pi)^{n-y}, \quad y = 0, 1, 2, \ldots, n
\end{aligned}
$$

where “$y$ factorial” is given by

$$y! = y(y-1)(y-2)\cdots 1 \quad \text{with} \quad 0! = 1$$

# Example

Cola Preference (Coke, Pepsi)

Suppose $\pi = \text{Pr}(\text{Coke}) = 0.6$.

Sample $n = 3$ tasters; let $y =$ number preferring Coke among them.

$$P(y) = \frac{3!}{y!(3-y)!}(.6)^y(.4)^{3-y}$$

$$P(0) = \frac{3!}{0!3!}(.6)^0(.4)^3 = (.4)^3 = 0.064$$

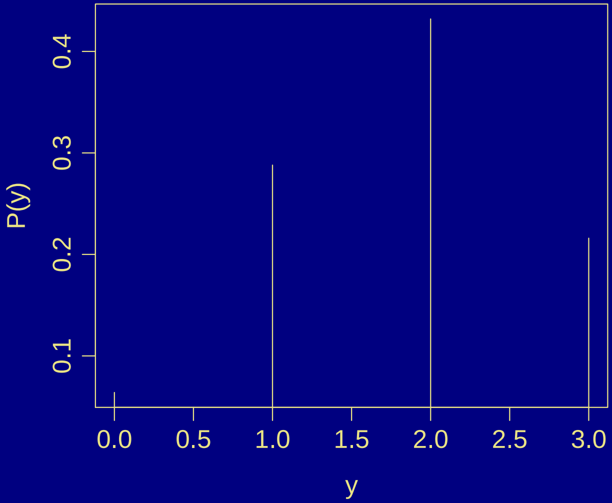
$$P(1) = \frac{3!}{1!2!}(.6)^1(.4)^2 = 3(.6)(.4)^2 = 0.288$$

| $y$ | $P(y)$ |
|---|---|
| 0 | 0.064 |
| 1 | 0.288 |
| 2 | 0.432 |
| 3 | 0.216 |
|  | 1 |

# R Code

```
> dbinom(0, 3, .6)

[1] 0.064

> dbinom(1, 3, .6)

[1] 0.288

> dbinom(0:3, 3, .6)

[1] 0.064 0.288 0.432 0.216

> cbind(0:3, dbinom(0:3, 3, .6))

     [,1]  [,2]
[1,]    0 0.064
[2,]    1 0.288
[3,]    2 0.432
[4,]    3 0.216
```

```
> plot(0:3, dbinom(0:3, 3, .6), type = "h",
       xlab = "y", ylab = "P(y)")
```



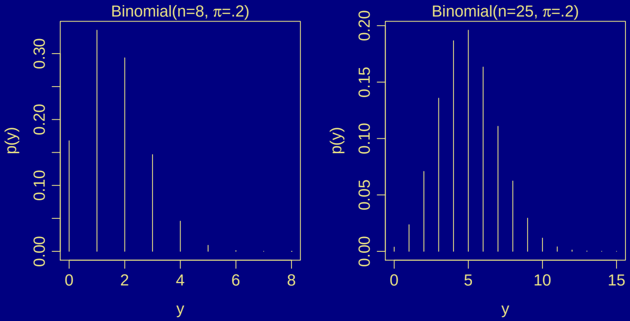
# Facts About the Binomial Distribution

- $E(Y) = n\pi$
- $\sigma^2 = \text{Var}(Y) = n\pi(1-\pi), \quad \sigma = \sqrt{n\pi(1-\pi)}$
- $p = \dfrac{Y}{n} =$ proportion of success (also denoted $\hat{\pi}$)

  $$E(p) = E\left(\frac{Y}{n}\right) = \pi \qquad \text{(mean of } p \text{ is } \pi\text{)}$$

  $$\sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}} \qquad \text{(std error of } p\text{)}$$

- Binomial distribution can be approximated by a normal distribution when $n$ is large ($n \cdot \min\{\pi, 1-\pi\} \geqslant 5$).


Binomial(n=8, π=.2)
p(y)
0.00
0.10
0.20
0.30
0
2
4
6
8
y
Binomial(n=25, π=.2)
p(y)
0.00
0.05
0.10
0.15
0.20
0
5
10
15
y

# 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

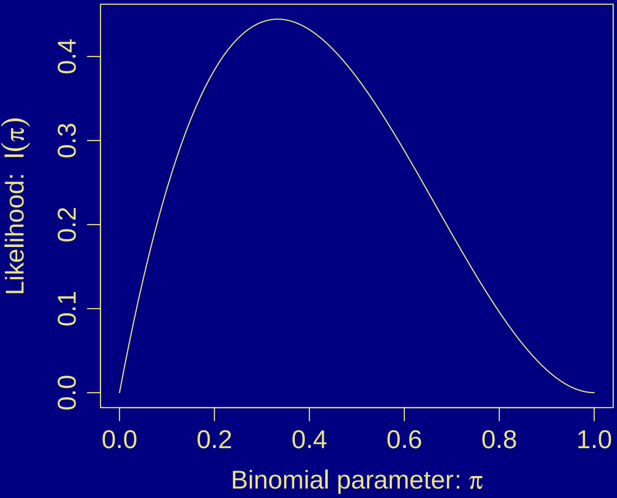
$$P(1) = \frac{3!}{1!2!}\pi^1(1-\pi)^2 = 3\pi(1-\pi)^2 =: \ell(\pi)$$

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

$$\pi = 0.1: \ \ell(0.1) = 3(.1)(.9)^2 = 0.243$$
$$\pi = 0.4: \ \ell(0.4) = 3(.4)(.6)^2 = 0.432$$
$$\pi = 0.6: \ \ell(0.6) = 3(.6)(.4)^2 = 0.288$$

## Plot of Binomial Likelihood Function when $n = 3$, $y = 1$

```
> curve(dbinom(1,3,x), xlim = c(0,1))
```



### Definition

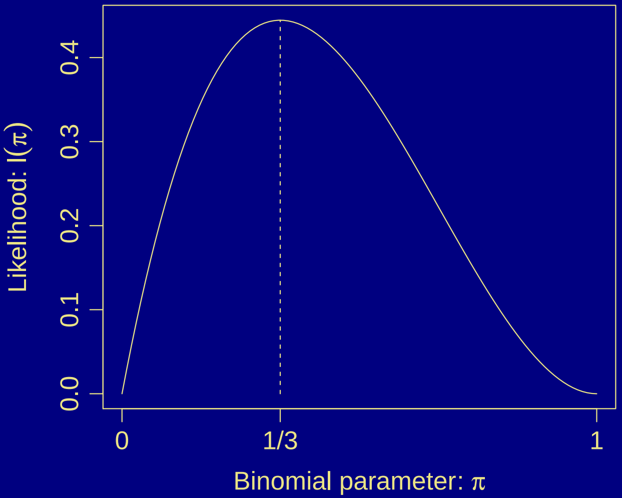
The *maximum likelihood estimate* (MLE) is the parameter value at which the likelihood function is maximized.

### Example

$\ell(\pi) = 3\pi(1-\pi)^2$ is maximized at $\hat{\pi} = 1/3 = 0.333$

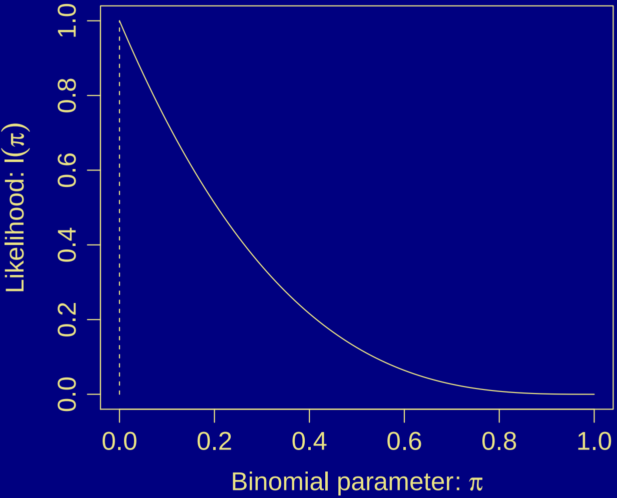
I.e., $y = 1$ success in $n = 3$ trials is more likely for $\pi = 1/3$ than for any other value of $\pi$.

# Plot of Binomial Likelihood Function when $n = 3$, $y = 1$



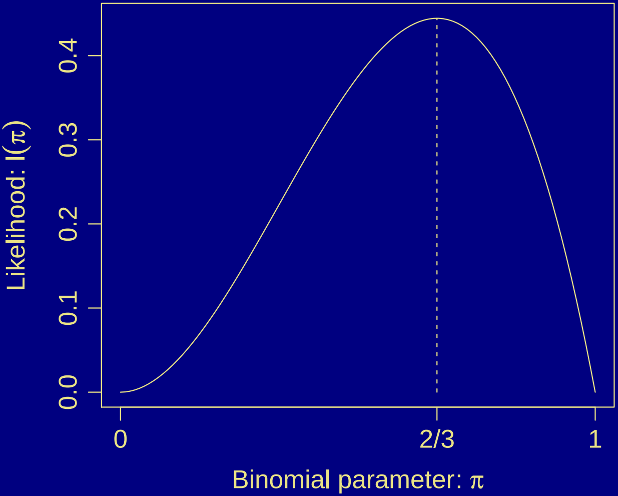
# Plot of Binomial Likelihood Function when $n = 3$, $y = 0$

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 $\hat{\pi} = \frac{0}{3} = 0$.



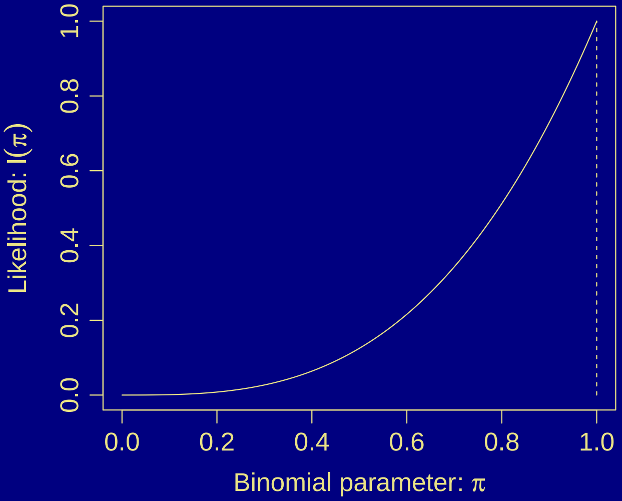
# Plot of Binomial Likelihood Function when $n = 3$, $y = 2$

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 $n = 3$, $y = 3$

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



# Facts About MLEs

- For binomial, MLE is $\hat{\pi} = \dfrac{y}{n} = p =$ 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 least squares estimators of the regression coefficients are also the 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 $\quad \hat{\pi} = p = \dfrac{y}{n}.$

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

- Note that $p$ is *unbiased* ($E(p) = \pi$) and that $\sigma(p) \downarrow 0$ as $n \uparrow \infty$. This implies that $p$ is a *consistent* estimator of $\pi$, i.e., $p \to \pi$ in probability.

  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

$$\mathrm{H}_0 : \pi = \pi_0 \qquad \text{vs} \qquad \mathrm{H}_a : \pi \neq \pi_0 \quad \text{(or 1-sided alternative)}$$

If $\mathrm{H}_0$ is true, then the sampling distribution of the test statistic

$$z = \frac{p - \pi_0}{\sigma(p)} = \frac{p - \pi_0}{\sqrt{\dfrac{\pi_0(1 - \pi_0)}{n}}}$$

is approximately $N(0, 1)$ for large samples: this is the *reference distribution*. Note that the null SE of $p$ was used to compute $z$.

### Definition

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

For the two-sided alternative hypothesis ($\pi \neq \pi_0$), use the two-tailed probability $\Pr(|Z| > |z|)$.

# Confidence Interval for Binomial Parameter

## Definition

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

For a 95% CI, $\alpha = 5\% = .05$ and $z_{\alpha/2} = z_{.025} = 1.96$, so take $\pm 1.96$ standard errors.

## Example

$$\theta = \pi : \qquad \text{MLE is } \hat{\theta} = \hat{\pi} = p$$

$$\sigma(p) = \sqrt{\frac{\pi(1-\pi)}{n}} \text{ estimated by } \text{SE} = \sqrt{\frac{p(1-p)}{n}}$$

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

## Example (in which the Wald interval collapses)

Estimate $\pi$, the population proportion of vegetarians.

For $n = 20$, suppose we observe $y = 0$.

$$p = \frac{0}{20} = 0$$

$$\text{95\% CI:} \quad 0 \pm 1.96 \sqrt{\frac{0 \times 1}{20}} = 0 \pm 0 = (0, 0)$$

## Remarks

- Wald intervals can perform poorly in categorical data analysis unless $n$ is quite large.
- Wald CI for $\pi$ collapses if $p = 0$ or 1.
- The actual coverage probability of the Wald interval can be much less than 0.95 when $\pi$ is 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

$$z = \frac{p - \pi_0}{\sqrt{\frac{p(1-p)}{n}}} \qquad \text{(denominator is estimated SE)}$$

### Definition

The *score test* and the *score CI* use null hypothesis value of the SE.

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

$$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$$

(denom is SE under $H_0$; known, not estimated)

## Example

$\pi =$ probability of being vegetarian.

$$n = 20, \qquad y = 0, \qquad p = \frac{0}{20} = 0$$

What values of $\pi_0$ satisfy

$$\frac{|0 - \pi_0|}{\sqrt{\frac{\pi_0(1-\pi_0)}{20}}} < 1.96 \qquad \text{i.e.,} \qquad |0 - \pi_0| < 1.96\sqrt{\frac{\pi_0(1-\pi_0)}{20}}$$

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 CI.

- When solving the quadratic, can show that midpoint of 95% score CI is

$$\frac{y + 1.96^2/2}{n + 1.96^2} \approx \frac{y+2}{n+4}.$$

- Can improve Wald CI $p \pm 1.96\sqrt{\frac{p(1-p)}{n}}$ by adding 2 successes and 2 failures before computing $p$ (“Agresti-Coull” method).

- For inference about proportions, *score* tests and CIs tend to perform better than *Wald*, in that the actual error rates are closer to their nominal levels.

- Another good approach uses the likelihood function directly. The CI it is the set of values of $\pi_0$ not rejected by the *likelihood ratio test*, i.e., the set of values of $\pi$ for which $\ell(\pi)$ is close to $\ell(\hat{\pi})$.

- For very small $n$, do inference using the exact binomial sampling distribution of the data, instead of the normal approximation.

# R Functions for Simple Binomial Tests and CIs

`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 to improve normal approx.

```
> 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
```

`prop.test` without continuity correction.

```
> 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 $H_a : \pi \neq 0.5$

```
> 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 ec
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.

### Physicians' Health Study (5 years)

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

| Group | MI | |
|---|---|---|
| | Y | N |
| Placebo | 189 | 10845 |
| Aspirin | 104 | 10933 |

With row totals:

| Group | MI | | |
|---|---|---|---|
| | Y | N | Total |
| Placebo | 189 | 10845 | 11034 |
| Aspirin | 104 | 10933 | 11037 |

# 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 $X$.

## Physicians' Health Study (ctd)

| Group | MI |  |  |
|---|---|---|---|
|  | Y | N | Total |
| Placebo | 0.017 | 0.983 | 1 |
| Aspirin | 0.009 | 0.991 | 1 |

Sample (or estimated) conditional probs for placebo group are

$$\frac{189}{11,034} = .017, \quad \frac{10,845}{11,034} = .983$$

Natural way to look at data when

$Y =$ response variable (e.g., heart attack: yes/no)

$X =$ explanatory variable (e.g., group: aspirin/placebo)

## Diagnostic Disease Tests

$Y =$ outcome of test: $1 =$ positive $2 =$ negative

$X =$ actual condition: $1 =$ diseased $2 =$ not diseased

|   |   | Y |   |
|---|---|---|---|
|   |   | 1 | 2 |
| X | 1 |   |   |
|   | 2 |   |   |

$$\text{sensitivity} = \Pr(Y = 1|X = 1)$$

$$\text{specificity} = \Pr(Y = 2|X = 2)$$

If you get a positive result, more relevant to you is $\Pr(X = 1|Y = 1)$. 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? Let

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

$$\pi_{i+} = \Pr(X = i) = \sum_j \pi_{ij} = \pi_{i1} + \cdots + \pi_{iJ}$$

$$\pi_{+j} = \Pr(Y = j) = \sum_i \pi_{ij} = \pi_{1j} + \cdots + \pi_{Ij}$$

$\{\pi_{ij}\}$ forms the *joint distribution* of X and Y.
$\{\pi_{i+}\}$ forms the *marginal* distribution of X.
$\{\pi_{+j}\}$ forms the *marginal* distribution of Y.

2 × 2 example:

|  |  | Y |  |  |
|---|---|---|---|---|
|  |  | 1 | 2 |  |
| X | 1 | $\pi_{11}$ | $\pi_{12}$ | $\pi_{1+}$ |
|  | 2 | $\pi_{21}$ | $\pi_{22}$ | $\pi_{2+}$ |
|  |  | $\pi_{+1}$ | $\pi_{+2}$ | 1 |

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:

|  |  | Y |  |  |
|---|---|---|---|---|
|  |  | 1 | 2 |  |
| X | 1 | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
|  | 2 | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
|  |  | $n_{+1}$ | $n_{+2}$ | $n$ |

# 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.

| | | Y | | |
|---|---|---|---|---|
| | | 1 | 2 | |
| X | 1 | .01 | .99 | 1 |
| | 2 | .01 | .99 | 1 |

Fact: $X$ and $Y$ are independent if and only if

$$\Pr(X = i, Y = j) = \Pr(X = i) \cdot \Pr(Y = j) \quad \text{for all } i \text{ and } j,$$

i.e., $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all $i$ and $j$.

$2 \times 2$ example:

|   |   | Y |   |   |
|---|---|---|---|---|
|   |   | 1 | 2 |   |
| X | 1 | .42 | .28 | .7 |
|   | 2 | .18 | .12 | .3 |
|   |   | .6 | .4 | 1 |

# 2.2 Comparing Proportions in 2 × 2 Tables

Conditional distributions:

| | | Y | |
|---|---|---|---|
| | | S | F |
| X | 1 | $\pi_1$ | $1-\pi_1$ |
| | 2 | $\pi_2$ | $1-\pi_2$ |

$$\hat{\pi}_1 - \hat{\pi}_2 = p_1 - p_2 \qquad \text{SE}(p_1 - p_2) = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

### Physicians’ Health Study (ctd)

$$p_1 = 0.017 \quad p_2 = 0.009 \quad p_1 - p_2 = 0.008$$

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

95% CI for $\pi_1 - \pi_2$: $.008 \pm 1.96(.0015) = .008 \pm .003 = (.005, .011)$
Apparently $\pi_1 - \pi_2 > 0$ (i.e., $\pi_1 > \pi_2$).

# Relative Risk

$$\text{relative risk} = \frac{\pi_1}{\pi_2}$$

### Physicians' Health Study (ctd)

*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 CI formula for relative risk.

  Example: 95% CI for RR in PHS is $(1.43, 2.31)$.

- Independence $\iff \frac{\pi_1}{\pi_2} = 1$

## 2.3 The Odds Ratio

|  |  | $S$ | $F$ |
|---|---|---|---|
| Group | 1 | $\pi_1$ | $1-\pi_1$ |
|  | 2 | $\pi_2$ | $1-\pi_2$ |

The *odds* of response S (instead of F) is $\dfrac{\Pr(S)}{\Pr(F)}$.

In the 2 × 2 table above: $\text{odds}(S) = \begin{cases} \dfrac{\pi_1}{1-\pi_1} & \text{in row 1} \\ \dfrac{\pi_2}{1-\pi_2} & \text{in row 2} \end{cases}$

Note

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

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

$$\text{odds}(S) = 3 \implies \Pr(S) = \frac{3}{1+3} = \frac{3}{4} \qquad \Pr(F) = \frac{1}{4}$$

$$\text{odds}(S) = \frac{1}{3} \implies \Pr(S) = \frac{1/3}{1+1/3} = \frac{1/3}{4/3} = \frac{1}{4} \qquad \Pr(F) = \frac{3}{4}$$

|  |  | $S$ | $F$ |
|---|---|---|---|
| Group | 1 | $\pi_1$ | $1-\pi_1$ |
|  | 2 | $\pi_2$ | $1-\pi_2$ |

### Definition (Odds Ratio)

*Odds Ratio:* $$\theta = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} = \frac{\pi_1(1-\pi_2)}{\pi_2(1-\pi_1)}$$

# Physicians Health Study (ctd)

| Group | MI | | |
|---|---|---|---|
| | Y | N | Total |
| Placebo | 189 | 10845 | 11034 |
| Aspirin | 104 | 10933 | 11037 |

Sample proportions:

$$\begin{array}{|c|c|}\hline p_1 & 1-p_1 \\ \hline p_2 & 1-p_2 \\ \hline\end{array} = \begin{array}{|c|c|}\hline 0.0171 & 0.9829 \\ \hline 0.0094 & 0.9906 \\ \hline\end{array} \begin{array}{c} 1.0 \\ 1.0 \end{array}$$

$$\text{sample odds} = \begin{cases} \dfrac{0.0171}{0.9829} = \dfrac{189}{10845} = 0.0174 & \text{placebo} \\ \dfrac{0.0094}{0.9906} = \dfrac{104}{10933} = 0.0095 & \text{aspirin} \end{cases}$$

$$\text{sample odds ratio} = \hat{\theta} = \frac{0.0174}{0.0095} = 1.83$$

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

# Properties of the Odds Ratio

- For counts

| S | F |
|---|---|
| $n_{11}$ | $n_{12}$ |
| $n_{21}$ | $n_{22}$ |

$$\hat{\theta} = \frac{n_{11}/n_{12}}{n_{21}/n_{22}} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \text{cross-product ratio}$$

- Treats X, Y symmetrically:

| MI | Group | |
|---|---|---|
| | Placebo | Aspirin |
| Y | 189 | 104 |
| N | 10845 | 10933 |

$$\Longrightarrow \hat{\theta} = 1.83$$

- Each odds $\geqslant 0$ and $\theta \geqslant 0$.

- $\theta = 1$ 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.,

  $$\theta = 2 \iff \log\theta = 0.7$$
  $$\theta = \frac{1}{2} \iff \log\theta = -0.7$$

- 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

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

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

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

## Physicians Health Study (ctd)

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

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

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

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

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

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,

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

# Review: Exponential and Natural Logarithm Functions

$\exp x = e^x$ (exponential function)

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

$e^x > 0$ 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 $e^{0.693} = 2$

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

X $=$ smoked $\geqslant$ 1 cigarette per day for at least 1 year

Y $=$ lung cancer

| Smoked | Cancer | |
|---|---|---|
| | Yes | No |
| Yes | 688 | 650 |
| No | 21 | 59 |
| Total | 709 | 709 |

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

Cannot estimate $\Pr(Y = \text{yes}|X)$. Cannot estimate

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

Cannot estimate $\pi_1/\pi_2$.

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

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

$$\hat{\theta} = \frac{\widehat{\Pr}(X = \text{yes}|Y = \text{yes})/\widehat{\Pr}(X = \text{no}|Y = \text{yes})}{\widehat{\Pr}(X = \text{yes}|Y = \text{no})/\widehat{\Pr}(X = \text{no}|Y = \text{no})}$$

$$= \frac{(688/709)/(21/709)}{(650/709)/(59/709)} = \frac{688 \times 59}{21 \times 650} = 2.97$$

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 $\theta \approx \pi_1/\pi_2 =$ relative risk, and can conclude that <u>probability</u> of lung cancer is ≈ 3 times as high for smokers as for non-smokers.

# 2.4 Testing Independence

## Job Satisfaction and Income

Data from General Social Survey (1991)

| Income | Job Satisfaction | | | | |
|---|---|---|---|---|---|
| | Dissat | Little | Moderate | Very | Total |
| $<$5K | 2 | 4 | 13 | 3 | 22 |
| 5K–15K | 2 | 6 | 22 | 4 | 34 |
| 15K–25K | 0 | 1 | 15 | 8 | 24 |
| $>$25K | 0 | 3 | 13 | 8 | 24 |
| Total | 4 | 14 | 63 | 23 | 104 |

$\mathsf{H}_0$ : X and Y independent vs $\mathsf{H}_a$ : X and Y dependent

$\mathsf{H}_0$ 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

$$\begin{aligned} \mu_{ij} &= \text{mean of dist. of cell count } n_{ij} \\ &= n\pi_{ij} \\ &= n\pi_{i+}\pi_{+j} \quad \text{under } \mathsf{H}_0 \end{aligned}$$

MLEs under $\mathsf{H}_0$ are

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

$\hat{\mu}_{ij}$ are called *estimated expected frequencies*.

# Chi-Squared Test of Independence

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

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

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

$$\text{df} = (I - 1)(J - 1)$$

where $I =$ number of rows, $J =$ number of columns.

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

(Table given on p. 343 of text.)

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

## Job Satisfaction and Income (ctd)

$$X^2 = 11.5$$

$$\text{df} = (I-1)(J-1) = 3 \times 3 = 9$$

$$\text{p-value} = \Pr(X^2 \geqslant 11.5) = 0.2415$$

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

# Likelihood-Ratio Test of Independence

Test statistic

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

$$= 2 \sum_{ij} n_{ij} \log \left( \frac{n_{ij}}{\hat{\mu}_{ij}} \right)$$

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

## Job Satisfaction and Income (ctd)

$G^2 = 13.47$

df $= 9$

p-value $= 0.1426$

# Degrees of Freedom for Chi-Squared Test

df for $X^2$ test = # parameters in general − # parameters under $H_0$

## Example (Degrees of Freedom for Chi-Squared Test of Indep)

Independence: $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$

$$\sum_{ij} \pi_{ij} = 1 \qquad \sum_{i} \pi_{i+} = 1 \qquad \sum_{j} \pi_{+j} = 1$$

- 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

$$\begin{aligned} df &= (IJ - 1) - \big[(I - 1) + (J - 1)\big] \\ &= (I - 1)(J - 1) \end{aligned}$$

## Remarks

- If all $n_{ij} = \hat{\mu}_{ij}$, then $X^2 = G^2 = 0$.
- As $n \uparrow$, $X^2 \xrightarrow{d} \chi^2$ faster than $G^2 \xrightarrow{d} \chi^2$, but $X^2$ and $G^2$ are usually similar if most $\hat{\mu}_{ij} \geqslant 5$.
- These tests treat $X$ and $Y$ as nominal: reordering rows or columns leaves $X^2$, $G^2$ 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, $r_{ij} \approx$ std normal $N(0, 1)$.

## Job Satisfaction and Income (ctd)

$$n_{44} = 8 \qquad \hat{\mu}_{44} = \frac{24 \times 23}{104} = 5.31$$

$$r_{44} = \frac{8 - 5.31}{\sqrt{5.31\left(1 - \frac{24}{104}\right)\left(1 - \frac{23}{104}\right)}} = 1.51$$

None of the cells show very strong evidence of association.

Standardized Residuals for Job Satisfaction Data

| Income | Dissat | Little | Moderate | Very |
|---|---|---|---|---|
| <5K | 1.44 | 0.73 | −0.16 | −1.08 |
| 5K–15K | 0.75 | 0.87 | 0.60 | −1.77 |
| 15K–25K | −1.12 | −1.52 | 0.22 | 1.51 |
| >25K | −1.12 | −0.16 | −0.73 | 1.51 |

## 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 contigency table of counts.

```
> 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)
> pag.tab
```

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

Once the data are saved as a table as above, we can easily convert them to an data frame, an R data structure with a column for each variable and a row for each observation.

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

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.

```
> 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

> pag.df$Freq <- c(762, 484, 327, 239, 468, 477)
```

Having created the data frame, we can generate the table using the `xtabs` function.

```
> 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

> xtabs(Freq ~ Gender + Party, data=pag.df)

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

The data could also be read from the columns of text file or a comma-separated (csv) file. The csv format provides an easy way to move data from a spreadsheet program into R or vice versa. The text or csv file should have a separate row for each combination of factor levels.

Thus a text file `Data/pag.txt` containing  can be read into an R dataframe via

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

Similarly, a csv file `Data/pag.csv` containing  can be read into an R dataframe via

```
> pag.df <- read.csv("Data/pag.csv")
```

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

# Computations on Tables

## Marginal Totals

```
> pag.tab

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

> 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

## Sample Proportions

Overall proportions $\{p_{ij}\}$ :

```
> prop.table(pag.tab)

        Party
Gender    Democrat Independent Republican
  Female 0.276387    0.118607   0.169750
  Male   0.175553    0.086688   0.173014

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

        Party
Gender    Democrat Independent Republican
  Female    0.276       0.119      0.170
  Male      0.176       0.087      0.173
```

# Computations on Tables

## Sample Proportions

Row proportions:

```
> prop.table(pag.tab, 1)
```

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

Column proportions:

```
> prop.table(pag.tab, 2)
```

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

# 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
```

```
> pag.chisq$observed

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

> pag.chisq$expected

        Party
Gender   Democrat Independent Republican
  Female   703.67      319.65     533.68
  Male     542.33      246.35     411.32

> with(pag.chisq, sum((observed - expected)^2/expected))

[1] 30.07
```

Unadjusted (or raw) Pearson residuals:

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

Standardized (or adjusted) Pearson residuals:

```
> pag.chisq$stdres
        Party
Gender    Democrat Independent Republican
  Female  4.50205     0.69945   -5.31595
  Male   -4.50205    -0.69945    5.31595
```

# 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:

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

- $G^2$ statistic for testing independence can be partitioned into components representing certain aspects of the association.
- Partition of $X^2$ is only approximate.
- Text discusses how to partition so that separate components are independent. This is required for $G^2$ to partition exactly.

# Job Satisfaction and Income (ctd)

| Income | Job Satisfaction | | | |
|---|---|---|---|---|
| | 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 |

Recall $X^2 = 11.52$, $G^2 = 13.47$, $\text{df} = 9$.

| Income | Job Satisfaction | | | |
|---|---|---|---|---|
| | Dissat | Little | Moderate | Very |
| $<$5K | 0.091 | 0.182 | 0.591 | 0.136 |
| 5K–15K | 0.059 | 0.176 | 0.647 | 0.118 |
| 15K–25K | 0.000 | 0.042 | 0.625 | 0.333 |
| $>$25K | 0.000 | 0.125 | 0.542 | 0.333 |

# Job Satisfaction and Income (ctd)

| Income | JobSat | | | | $X^2$ | $G^2$ | df |
|---|---|---|---|---|---|---|---|
| | VD | LS | MS | VS | | | |
| Low | | | | | | | |
| $<5$ | 2 | 4 | 13 | 3 | 0.30 | 0.30 | 3 |
| 5–15 | 2 | 6 | 22 | 4 | | | |
| High | | | | | | | |
| 15–25 | 0 | 1 | 15 | 8 | 1.14 | 1.19 | 3 |
| $>25$ | 0 | 3 | 13 | 8 | | | |
| Low vs High | | | | | | | |
| $<15$ | 4 | 10 | 35 | 7 | 10.32 | 11.98 | 3 |
| $>15$ | 0 | 4 | 28 | 16 | | | |
| Sum | | | | | | | |
| | | | | | 11.76 | 13.47 | 9 |

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

# Exact Inference for Small Samples

## 2 × 2 Case: Fisher's Exact Test

$H_0$: X, Y independent.

|  |  | Y |  |  |
|---|---|---|---|---|
|  |  | 1 | 2 |  |
| X | 1 | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
|  | 2 | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
|  |  | $n_{+1}$ | $n_{+2}$ | $n$ |

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

$$P(n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1}-n_{11}}}{\binom{n}{n_{+1}}}, \quad \text{where} \quad \binom{a}{b} = \frac{a!}{b!(a-b)!}$$

## 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. Asked to identify the 4 cups with milk poured first.

| | | Guess | | |
|---|---|---|---|---|
| | | Milk | Tea | |
| Poured First | Milk | ? | | 4 |
| | Tea | | | 4 |
| | | 4 | 4 | 8 |

$n_{11} = 0, 1, 2, 3, \text{or } 4.$

Under $H_0$, $\begin{array}{|c|c|}\hline 4 & 0 \\ \hline 0 & 4 \\ \hline \end{array}$ has probability

$$P(4) = \frac{\binom{4}{4}\binom{4}{4-4}}{\binom{8}{4}} = \frac{\frac{4!}{4!0!} \times \frac{4!}{0!4!}}{\frac{8!}{4!4!}} = \frac{4!4!}{8!} = \frac{1}{70} = 0.014$$

## Lady Tasting Tea (ctd)

Under $H_0$, $\begin{array}{|c|c|}\hline 3 & 1 \\ \hline 1 & 3 \\ \hline \end{array}$ 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))
```

| $n_{11}$ | $P(n_{11})$ |
|---|---|
| 0 | 0.014 |
| 1 | 0.229 |
| 2 | 0.514 |
| 3 | 0.229 |
| 4 | 0.014 |

For $2 \times 2$ tables,

$$H_0 : \text{indep} \iff H_0 : \theta = 1 \quad (\theta = \text{odds ratio})$$

To test $\quad H_0 : \theta = 1 \quad$ vs $\quad H_a : \theta > 1$

$$\text{p-value} = \Pr\left(\hat{\theta} \geqslant \hat{\theta}_{\text{obs}}\right) = \Pr\left(n_{11} \geqslant n_{11}^{\text{obs}}\right)$$

### Lady Tasting Tea (ctd)

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

| | | Guess | | |
|---|---|---|---|---|
| | | Milk | Tea | |
| Poured First | Milk | 3 | 1 | 4 |
| | Tea | 1 | 3 | 4 |
| | | 4 | 4 | 8 |

$$\boxed{n_{11} = 3}$$

$$\text{p-value} = \Pr(n_{11} \geqslant 3) = P(3) + P(4) = 0.229 + 0.014 = 0.243$$

Very little evidence against $H_0$.

```
> TeaTasting <-
    matrix(c(3, 1, 1, 3),
           nrow = 2,
           dimnames = list(Truth = c("Milk", "Tea"),
             Guess = c("Milk", "Tea")))
> TeaTasting <- as.table(TeaTasting)
> 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 = two-tail prob. of outcomes no more likely than that observed

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

p-value $= P(0) + P(1) + P(3) + P(4) = 0.486$

```
> 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 $\leqslant \alpha = .05$), then actual $\Pr(\text{type I error}) < \alpha$ because of discreteness (see text).
- Margins may be fixed by design, but test valid even if not.
- Fisher's exact test extends to $I \times J$ tables.

```
> sattab

          Job Satisfaction
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
```

## FL Death Penalty Cases

```
> data(deathpenalty)
> 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

> deathpenalty <-
    transform(deathpenalty,
              DeathPenalty = relevel(DeathPenalty, "Yes"),
              Defendant = relevel(Defendant, "White"),
              Victim = relevel(Victim, "White"))
```

# FL Death Penalty Cases (ctd)

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

, , DeathPenalty = Yes

       Defendant
Victim  White Black
  White    53    11
  Black     0     4

, , DeathPenalty = No

       Defendant
Victim  White Black
  White   414    37
  Black    16   139
```

## FL Death Penalty Cases (ctd)

```
> 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
```

# FL Death Penalty Cases (ctd)

$Y =$ death penalty (response var.)

$X =$ defendant's race (explanatory)

$Z =$ victim's race (control var.)

```
> round(100*prop.table(dpflat,1), 1)

                 DeathPenalty   Yes    No
Victim Defendant
White  White                   11.3  88.7
       Black                   22.9  77.1
Black  White                    0.0 100.0
       Black                    2.8  97.2
```

The tables

| Def | DeathPen | |
|---|---|---|
| | Yes | No |
| White | 53 | 414 |
| Black | 11 | 37 |

and

| Def | DeathPen | |
|---|---|---|
| | Yes | No |
| White | 0 | 16 |
| Black | 4 | 139 |

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

The (estimated) *conditional odds ratios* are:

$$Z = \text{white}: \quad \hat{\theta}_{XY(1)} = \frac{53 \times 37}{414 \times 11} = 0.43 \quad \text{(0.42 after add .5 to all cells)}$$

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

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.

| Def | DeathPen | |
|---|---|---|
| | Yes | No |
| White | 53 | 430 |
| Black | 15 | 176 |

$$\hat{\theta}_{XY} = 1.45$$

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 × 2 × 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:**

| Clinic | Treatment | Response | | |
|---|---|---|---|---|
| | | Y | | |
| Z | X | S | F | $\hat{\theta}$ |
| 1 | A | 18 | 12 | 1.0 |
| | B | 12 | 8 | |
| 2 | A | 2 | 8 | 1.0 |
| | B | 8 | 32 | |
| Marginal | A | 20 | 20 | 2.0 |
| | B | 20 | 40 | |

# Chapter 3: Generalized Linear Models

## 3.1 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 $\mu > 0$.

- $\log\left(\dfrac{\mu}{1-\mu}\right) = \alpha + \beta_1 x_1 + \cdots + \beta_k x_k$

  uses $g(\mu) = \log\left(\dfrac{\mu}{1-\mu}\right)$, the *logit* link. logit $=$ log(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 `glm()` function in R and of `proc genmod` in SAS.

# 3.2 GLMs for Binary Data

Suppose $Y = 1$ or 0 (“Bernoulli” or “binary” random variable).

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

This is binomial for $n = 1$ trial.

$$\begin{aligned} E(Y) &= \pi \\ \text{Var}(Y) &= \pi(1 - \pi) \end{aligned}$$

For an explanatory variable $x$, $\pi = \pi(x)$ varies as $x$ varies.

# Linear Probability Model

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

A GLM with binomial random component and identity link function.

$\text{Var}(Y) = \pi(x)\big[1 - \pi(x)\big]$ varies as $x$ varies, so least squares not optimal.

Use ML to fit this and other GLMs.

## Infant Malformation

$Y =$ infant sex organ malformation $\quad (1 =$ present, $\quad 0 =$ absent$)$

$x =$ mother's alcohol consumption (avg drinks per day)

| Alcohol Consumption | |
|---|---|
| Measured | Score |
| 0 | 0.0 |
| $< 1$ | 0.5 |
| 1–2 | 1.5 |
| 3–5 | 4.0 |
| $\geqslant 6$ | 7.0 |

## Infant Malformation (ctd)

```
> data(malformation)
> malformation

   Alcohol Malformation  Freq
1      0.0      Present    48
2      0.0       Absent 17066
3      0.5      Present    38
4      0.5       Absent 14464
5      1.5      Present     5
6      1.5       Absent   788
7      4.0      Present     1
8      4.0       Absent   126
9      7.0      Present     1
10     7.0       Absent    37
```

# Example (Infant Malformation (ctd))

```
> malform.tab <- xtabs(Freq ~ Alcohol + Malformation,
                       data=malformation)
> malform.tab

       Malformation
Alcohol Absent Present
    0    17066      48
    0.5  14464      38
    1.5    788       5
    4      126       1
    7       37       1

> round(100*prop.table(malform.tab, 1), 2)

       Malformation
Alcohol Absent Present
    0    99.72    0.28
    0.5  99.74    0.26
    1.5  99.37    0.63
    4    99.21    0.79
    7    97.37    2.63
```

## Infant Malformation (ctd)

To fit a glm to these (grouped binary) data, we first need to recast the data frame into a wide format.

```
> library(reshape2)
> malformwide <- dcast(malformation,
                        Alcohol ~ Malformation,
                        value.var="Freq")
> malformwide

  Alcohol Absent Present
1     0.0  17066      48
2     0.5  14464      38
3     1.5    788       5
4     4.0    126       1
5     7.0     37       1
```

# Infant Malformation: Linear Probability Model

Two ways to fit the same binomial model in R.

```
> malform.lin <-
    glm(cbind(Present,Absent) ~ Alcohol,
        family=binomial(link=make.link("identity")),
        data=malformwide)
> malformwide <-
    transform(malformwide, Total = Present + Absent)
> malform.lin.alt <-
    glm(Present/Total ~ Alcohol, weights=Total,
        family=binomial(link=make.link("identity")),
        data=malformwide)
> coef(malform.lin)

(Intercept)     Alcohol
  0.0025476   0.0010872


> summary(malform.lin)
```

```
Call:
glm(formula = cbind(Present, Absent) ~ Alcohol, family = binomia
    data = malformwide)

Deviance Residuals:
     1       2       3       4       5
 0.656  -1.049   0.863   0.130   0.828

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) 0.002548   0.000352    7.23  4.8e-13
Alcohol     0.001087   0.000832    1.31     0.19

(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
```

## Infant Malformation: Linear Probability Model (ctd)

Linear probability model for $\pi = \text{Pr}(\text{malformation})$ has ML fit

$$\hat{\pi} = \hat{\alpha} + \hat{\beta}x = \boxed{0.0025} + \boxed{0.0011}\,x$$

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

## Remarks

- ML estimates $\hat{\alpha}$ and $\hat{\beta}$ obtained by iterative numerical optimization.
- To test $H_0 : \beta = 0$ (independence), can use $z = \dfrac{\hat{\beta} - 0}{\mathrm{SE}(\hat{\beta})}$.

  For large $n$ has approx. std. normal dist. under $H_0$.

  - $z = \dfrac{\boxed{0.0011}}{\boxed{0.00083}} = \boxed{1.31}$ For $H_a : \beta \neq 0$, p-value $= \boxed{0.19}$
- Could use Pearson $X^2$ (or $G^2$) to test independence, but ignores ordering of rows.
- Alternative way to apply $X^2$ (or *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 *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.

In R, the default link for the `binomial` family is the logit.

## Infant Malformation: Logistic Regression Model

```
> malform.logit <- glm(cbind(Present,Absent) ~ Alcohol,
                                  family=binomial, data=malformwide)


> summary(malform.logit)
```

```
Call:
glm(formula = cbind(Present, Absent) ~ Alcohol, family = binomial,
    data = malformwide)

Deviance Residuals:
     1       2       3       4       5
 0.592  -0.880   0.886  -0.145   0.129

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -5.960      0.115  -51.64   <2e-16
Alcohol        0.317      0.125    2.52    0.012

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.2020  on 4  degrees of freedom
Residual deviance: 1.9487  on 3  degrees of freedom
AIC: 24.58

Number of Fisher Scoring iterations: 4
```

## Infant Malformation: Logistic Regression Model (ctd)

$$\text{logit}(\hat{\pi}) = \log\left(\frac{\hat{\pi}}{1-\hat{\pi}}\right) = \boxed{-5.96} + \boxed{0.32}\,x$$

- $\hat{\pi} \uparrow$ as $x \uparrow$.
- p-value $= \boxed{0.012}$ for $H_0 : \beta = 0$ vs $H_a : \beta \neq 0$.
- <u>But</u> p-value $= 0.3$ if delete single “present” obs. in $\geqslant 6$ drinks row!!

## Remarks

- Chap. 4 studies logistic regression model.
- For contingency table, can test $H_0$ : ”model correctly specified” with $X^2$ and $G^2$ test statistics using expected counts predicted by model.
  - Ex: $X^2 = 2.05$, $G^2 = 1.95$ 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 $x$.

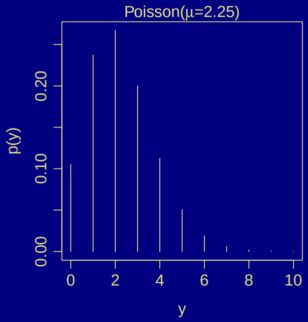
# 3.3 GLMs for Count Data

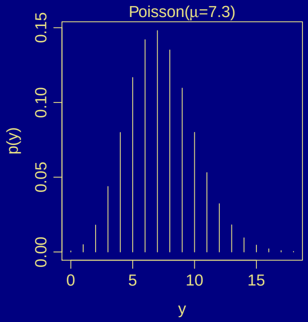
When $Y$ is a count $(0, 1, 2, 3, \ldots)$ usually assume a Poisson dist:

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

- $\mu = E(Y)$
- $\text{Var}(Y) = \mu, \quad \sigma = \sqrt{\mu}$
- 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.

```
> plot(0:10, dpois(0:10,2.25), type="h",
       xlab="y", ylab="p(y)", main="Poisson(mu=2.25)")
> plot(0:18, dpois(0:18,7.3), type="h",
       xlab="y", ylab="p(y)", main="Poisson(mu=7.3)")
```





# Poisson Regression for Count Data

Assume $Y$ has a Poisson dist., $x$ 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)

# Defects in Silicon Wafers

$Y$ = number defects on silicon wafer

$x$ = dummy var. for treatment ($0 = A$, $1 = B$)

```
> 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)
```

```
Call:
glm(formula = defects ~ trt, family = poisson(link = "identity")
    data = wafers)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-1.528  -0.762  -0.170   0.694   1.540

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    5.000      0.707    7.07  1.5e-12
trtB           4.000      1.183    3.38  0.00072

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 27.857  on 19  degrees of freedom
Residual deviance: 16.268  on 18  degrees of freedom
AIC: 94.35

Number of Fisher Scoring iterations: 3
```

```
Call:
glm(formula = defects ~ trt, family = poisson(link = "log"),
    data = wafers)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.528  -0.762  -0.170   0.694   1.540

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.609      0.141   11.38  < 2e-16
trtB           0.588      0.176    3.33  0.00086

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 27.857  on 19  degrees of freedom
Residual deviance: 16.268  on 18  degrees of freedom
AIC: 94.35

Number of Fisher Scoring iterations: 4
```

## Defects in Silicon Wafers (ctd)

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

$$\hat{\mu} = \boxed{5} + \boxed{4}\,x$$

$$x = 0: \quad \hat{\mu}_A = \boxed{5} \quad (= \bar{y}_A)$$

$$x = 1: \quad \hat{\mu}_B = \boxed{9} \quad (= \bar{y}_B)$$

$$\hat{\beta} = \boxed{4} = \hat{\mu}_B - \hat{\mu}_A \text{ has SE} = \boxed{1.18} \text{ (use for test and CI for } \beta\text{)}$$

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

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

$$x = 0: \quad \log \hat{\mu}_A = \boxed{1.609} \quad \hat{\mu}_A = e^{\boxed{1.609}} = 5$$

$$x = 1: \quad \log \hat{\mu}_B = \boxed{1.609} + \boxed{0.588} = \boxed{2.197} \quad \hat{\mu}_B = e^{\boxed{2.197}} = 9$$

# 3.4 Inference and Model Checking for GLMs

## Wald Test

Test $\mathrm{H}_0 : \beta = 0$

- $z = \dfrac{\hat{\beta}}{\mathrm{SE}}$ has approx. $N(0, 1)$ dist. under $\mathrm{H}_0$.
- For $\mathrm{H}_a : \beta \neq 0$ can also use Wald stat. $z^2 = \left(\dfrac{\hat{\beta}}{\mathrm{SE}}\right)^2$ approx. $\chi^2_1$.
- For $\mathrm{H}_0 : \beta = \beta_0$, use $z = \dfrac{\hat{\beta} - \beta_0}{\mathrm{SE}}$.
- Wald CI = set of $\beta_0$ values for which $|\hat{\beta} - \beta_0| / \mathrm{SE} < z_{\alpha/2}$, i.e.,

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

# Likelihood Ratio Test

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

$$\begin{aligned}
l_0 &= \text{maximized likelihood when } \beta = 0 \\
l_1 &= \text{maximized likelihood for arbitrary } \beta \\
\text{test stat} &= -2 \log\left(\frac{l_0}{l_1}\right) \\
&= -2\left[\log l_0 - \log l_1\right] \\
&= -2(L_0 - L_1)
\end{aligned}$$

where $L$ = maximized log-likelihood.

# Defects in Silicon Wafers (ctd)

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**

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

or

$$z^2 = 11.1 \quad df = 1 \quad \text{p-value} = 0.00086$$

**Likelihood-Ratio Test**

$$L_1 = -45.17 \qquad L_0 = -50.97$$

Test stat: $-2(L_0 - L_1) = 11.6$  df $= 1$  p-value $= 0.00066$

For binomial and Poisson GLMs, LR test statistics can be computed as a difference in *deviances* for two nested models. In R:

- the deviance for the fitted model is labelled as the "residual deviance." Analogous to the residual (or error) sum of squares in ordinary linear regression.

- the "null deviance" is the deviance for the model with intercept only (no predictors). Analogous to the total sum of squares.

- the difference between null and residual deviances is the LR statistic for testing the null hypothesis that all the regression coefficients (except for the intercept) equal 0. Analogous to the regression sum of squares.

- With only a single predictor $x$, the difference in null and residual deviances is the LR statistic for testing $\beta = 0$, e.g., $\boxed{27.9} - \boxed{16.3} = 11.6$.

The `drop1` function tests one coefficient at a time, while controlling for all other variables in the model. (Here there are no other variables in the model.)

```
> drop1(wafers.loglin, test="Chisq")

Single term deletions

Model:
defects ~ trt
       Df Deviance   AIC  LRT Pr(>Chi)
<none>        16.3  94.3
trt     1     27.9 103.9 11.6  0.00066
```

The `anova` function is useful for testing a sequence of nested models. Here only one fitted model is given, so it is tested against the null model with all coefficients equal to zero (intercept-only model).

```
> anova(wafers.loglin, test="Chisq")

Analysis of Deviance Table


Model: poisson, link: log


Response: defects


Terms added sequentially (first to last)



     Df Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL                    19       27.9
trt   1     11.6        18       16.3  0.00066
```

### 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 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.

## Wafer Defects

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

$$\beta = \log \mu_B - \log \mu_A = \log\left(\frac{\mu_B}{\mu_A}\right)$$

$$e^{\beta} = \frac{\mu_B}{\mu_A}$$

$$e^{\hat{\beta}} = e^{\boxed{0.5878}} = 1.8 = \frac{\hat{\mu}_B}{\hat{\mu}_A}$$

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

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

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

CI for $\beta$ based on LR test is $(0.247, 0.94)$.

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

```
> 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 $L_M$

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

*$H_0$: model M holds   vs   $H_a$: 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

$$X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

can used to test goodness-of-fit of model (i.e., $H_0$: model holds).

Under $H_0$, the distribution of $G^2$ (and $X^2$) is approx. $\chi^2$ with

$$\text{df} = \text{no. observations} - \text{no. model parameters}$$

## Wafer Defects (ctd)

$\hat{\mu}_i = 5$ for 10 obs in trt A

$\hat{\mu}_i = 9$ for 10 obs in trt B

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

deviance $G^2 = 16.3$

Pearson $X^2 = 16$

df $= 18$

These values of $G^2$ and $X^2$ do not contradict $H_0$: “model holds”, but in general we must be cautious about referring $G^2$ and/or $X^2$ to a chi-square dist. Usually we require

- $\hat{\mu}_i$s all large, and
- fixed df as $n \uparrow$ (happens with contingency tables but not here).

# Remarks

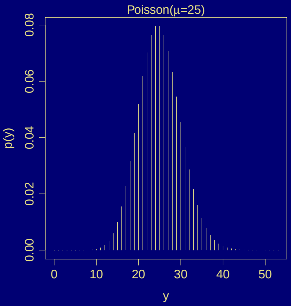
- For GLMs, can study lack-of-fit using residuals (later chapters).
- Count data often show *overdispersion* relative to a Poisson GLM.

  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

$$\mathrm{E}(Y) = \mu \qquad \mathrm{Var}(Y) = \mu + D\mu^2$$

where $D \geqslant 0$ is called a *dispersion parameter*.

As $D \downarrow 0$, neg. binom. → 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.

# Known Victims of Homicide

*Within the past 12 months, how many people have you known personally that were victims of homicide?*

```
> 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.

## Known Victims of Homicide (ctd)

You can safely ignore this slide if you wish.

```
> 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
```

## Known Victims of Homicide (ctd)

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

```
> ## homicide <-
> ##   transform(homicide, race = relevel(race, "White"))
> options(contrasts=c("contr.SAS","contr.poly"))
> 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 = Fre
    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 ta

    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
```

## Known Victims of Homicide (ctd)

In this example, both Poisson and neg. binom. model fits have

$$\log \hat{\mu} = \boxed{-2.38} + \boxed{1.73}\, x$$

$$e^{\hat{\beta}} = e^{\boxed{1.73}} = 5.7 \quad \left( = \frac{\bar{y}_B}{\bar{y}_W} = \frac{0.522}{0.092} \right)$$

But, SE for $\hat{\beta}$ is $\boxed{0.147}$ for Poisson, $\boxed{0.238}$ for neg. binom.

Wald 95% CI for $\beta$ is

$\boxed{1.73} \pm (1.96)(\boxed{0.147}) = (1.45, 2.02)$ for Poisson fit

$\boxed{1.73} \pm (1.96)(\boxed{0.238}) = (1.27, 2.2)$ for neg. binom. fit

Leads to 95% CI for $e^{\beta} = \mu_B/\mu_W$ of

$(e^{\boxed{1.45}}, e^{\boxed{2.02}}) = (4.25, 7.54)$ for Poisson

$(e^{\boxed{1.27}}, e^{\boxed{2.2}}) = (\boxed{3.55}, \boxed{9.03})$ for neg. binom.

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

```
> confint.default(hom.poi)

              2.5 %  97.5 %
(Intercept) -2.5736 -2.1928
raceBlack    1.4459  2.0204

> exp(confint.default(hom.poi))

               2.5 % 97.5 %
(Intercept) 0.076262 0.1116
raceBlack   4.245574 7.5414

> ## confint.default(hom.nb)
> exp(confint.default(hom.nb))

              2.5 %  97.5 %
(Intercept) 0.07332 0.11608
raceBlack   3.54571 9.02998
```

```
> confint(hom.poi)

              2.5 %  97.5 %
(Intercept) -2.5798 -2.1987
raceBlack    1.4437  2.0192

> exp(confint(hom.poi))

               2.5 %  97.5 %
(Intercept) 0.075788 0.11095
raceBlack   4.236333 7.53253

> ## confint(hom.nb)
> exp(confint(hom.nb))

              2.5 %  97.5 %
(Intercept) 0.07306 0.11573
raceBlack   3.57785 9.13164
```

## Known Victims of Homicide (ctd)

95% LR CIs for $e^{\beta} = \mu_B/\mu_W$ are:

- $(e^{\boxed{1.44}}, e^{\boxed{2.02}}) = (\boxed{4.24}, \boxed{7.53})$ for Poisson
- $(e^{\boxed{1.27}}, e^{\boxed{2.21}}) = (\boxed{3.58}, \boxed{9.13})$ for neg. binom.

## Remarks

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

  $$\widehat{\mathrm{Var}(Y)} = \hat{\mu} + \widehat{\mathrm{D}}\hat{\mu}^2 = \hat{\mu} + 4.94\hat{\mu}^2$$

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

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

  $$\widehat{\mathrm{Var}(Y)} = \hat{\mu} + \frac{\hat{\mu}^2}{\hat{\theta}} = \hat{\mu} + \frac{\hat{\mu}^2}{\boxed{0.2023}} = \hat{\mu} + \boxed{4.94}\,\hat{\mu}^2$$

- 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.

This slide and the next (output of `summary(hom.poi2)`) can be ignored.

```
> homicide2 <- homicide[rep(1:14, homicide$Freq),]
> homicide2$Freq <- NULL
> homicide2$ybar <- NULL
> head(homicide2)

    nvics  race
1       0 Black
1.1     0 Black
1.2     0 Black
1.3     0 Black
1.4     0 Black
1.5     0 Black

> hom.poi2 <-
    glm(nvics ~ race, data=homicide2, family=poisson)
> summary(hom.poi2)
```

```
Call:
glm(formula = nvics ~ race, family = poisson, data = homicide

Deviance Residuals:
   Min       1Q   Median       3Q      Max
 -1.02   -0.43   -0.43   -0.43     6.19

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 1307  degrees of freedom
Residual deviance: 844.71  on 1306  degrees of freedom
AIC: 1122

Number of Fisher Scoring iterations: 6
```

### Remarks

- When $Y$ is a count, overdispersion relative to Poisson is common. Safest strategy is to use neg. bin. model or some 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` and other packages contain code for fitting *ZIP* (zero-inflated Poisson) and *hurdle* 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 $y =$ count with base $t$. Sample rate is $\frac{y}{t}$.

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

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

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

$\log(t)$ is an *offset*.

See pp. 82–84 of text for discussion.

## 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.

```
> 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 =
    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
```

# British Train Accidents (ctd)

$$\log\left(\frac{\hat{\mu}}{t}\right) = \boxed{-4.21} - \boxed{0.0329}\,x$$

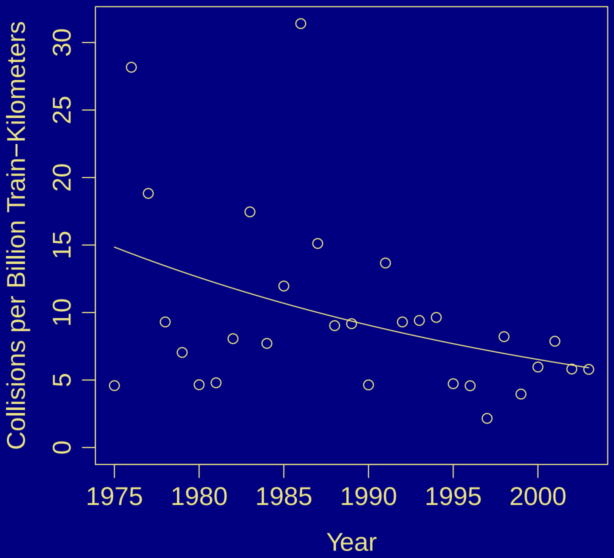
$$\frac{\hat{\mu}}{t} = \exp\left(\boxed{-4.21 - 0.0329x}\right) = e^{\boxed{-4.21}}\left(e^{\boxed{-0.0329}}\right)^{x}$$

$$= (0.0148)(0.968)^{x}$$

- Rate estimated to decrease by 1 − 0.968 = 0.032 = 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.

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

```
> attach(traincollisions)
> plot(Year, 1000*TrRd/KM, ylim=c(0,1000*max(TrRd/KM)),
       ylab="Collisions per Billion Train-Kilometers")
> curve(1000*exp(-4.21 - 0.0329*(x-1975)), add=TRUE)
> detach(traincollisions)
```


Collisions per Billion Train–Kilometers
0
5
10
15
20
25
30
1975
1980
1985
1990
1995
2000
Year

# 4. Logistic Regression

## Simple Logistic Regression

$$Y = 0 \text{ or } 1$$

$$\pi = \Pr(Y = 1)$$

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

Uses “logit” link for binomial $Y$. Equivalently,

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

where $\exp(\alpha + \beta x) = e^{\alpha + \beta x}$.

# 4.1 Interpreting the Logistic Regression Model 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 $\pi(x) = \dfrac{e^{\alpha}}{1 + e^{\alpha}}$ constant in $x$ (with $\pi > \frac{1}{2}$ if $\alpha > 0$).

- Curve can be approximated near a fixed point $x$ by a straight line describing rate of change in $\pi(x)$. Slope is $\beta\pi(x)\big[1 - \pi(x)\big]$. E.g.,

  - at $x$ with $\pi(x) = \dfrac{1}{2}$, slope $= \beta \cdot \dfrac{1}{2} \cdot \dfrac{1}{2} = \dfrac{\beta}{4}$

  - at $x$ with $\pi(x) = 0.1$ or $0.9$, slope $= \beta(0.1)(0.9) = 0.09\,\beta$

  - Steepest slope where $\pi(x) = \dfrac{1}{2}$

## 4.1 Interpreting the Logistic Regression Model II

- If $\pi(x) = \frac{1}{2}$ then

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \log\left(\frac{0.5}{0.5}\right) = \log(1) = 0 = \alpha + \beta x \implies x = \frac{-\alpha}{\beta}$$

- $\dfrac{1}{|\beta|} \approx$ dist. between $x$ values with $\pi = 0.5$ and $\pi = 0.75$ (or 0.25)
- ML fit obtained with iterative numerical methods.

## Horseshoe Crabs

Model the relationship between weight and the probability of having one or more “satellites” for female horseshoe crabs.

$$Y = \begin{cases} 1 & \text{if female crab has satellites} \\ 0 & \text{if no satellites} \end{cases}$$

$x =$ weight (kg) $\qquad \pi(x) =$ probability of at least one satellite

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

```
> data(horseshoecrabs)
> head(horseshoecrabs, 5)

  Color Spine Width Weight Satellites
1     2     3  28.3   3.05          8
2     3     3  22.5   1.55          0
3     1     1  26.0   2.30          9
4     3     3  24.8   2.10          0
5     3     3  26.0   2.60          4

> nrow(horseshoecrabs)

[1] 173
```

```
> summary(horseshoecrabs)

     Color          Spine          Width     
 Min.   :1.00   Min.   :1.00   Min.   :21.0  
 1st Qu.:2.00   1st Qu.:2.00   1st Qu.:24.9  
 Median :2.00   Median :3.00   Median :26.1  
 Mean   :2.44   Mean   :2.49   Mean   :26.3  
 3rd Qu.:3.00   3rd Qu.:3.00   3rd Qu.:27.7  
 Max.   :4.00   Max.   :3.00   Max.   :33.5  
     Weight       Satellites    
 Min.   :1.20   Min.   : 0.00  
 1st Qu.:2.00   1st Qu.: 0.00  
 Median :2.35   Median : 2.00  
 Mean   :2.44   Mean   : 2.92  
 3rd Qu.:2.85   3rd Qu.: 5.00  
 Max.   :5.20   Max.   :15.00  

> crabs.fit1 <- glm((Satellites > 0) ~ Weight,
                     family=binomial, data=horseshoecrabs)

> summary(crabs.fit1)
```

```
Call:
glm(formula = (Satellites > 0) ~ Weight, family = binomial, c

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      0.880   -4.20  2.7e-05
Weight         1.815      0.377    4.82  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
```

## Horseshoe Crabs: Fitted Logistic Regression on Weight

ML fit: $\text{logit}\left[\hat{\pi}(x)\right] = \boxed{-3.69} + \boxed{1.82}\, x$

i.e., $\hat{\pi}(x) = \boxed{\dfrac{\exp\left(-3.69 + 1.82x\right)}{1 + \exp\left(-3.69 + 1.82x\right)}}$

E.g., at $x = \bar{x} = 2.44$,

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

```
> xbar <- mean(horseshoecrabs$Weight)
> predict(crabs.fit1, data.frame(Weight=xbar), type="link")

      1
0.72913

> predict(crabs.fit1, data.frame(Weight=xbar),
          type="response")

      1
0.67461

> ab <- coef(crabs.fit1); ld50 <- -ab[1]/ab[2]
> names(ld50) <- NULL; ld50

[1] 2.0355

> predict(crabs.fit1,
          data.frame(Weight = ld50 + c(0, 0.1, 1)),
          type="response")

      1       2       3
0.50000 0.54525 0.85998
```

## Horseshoe Crabs: Fitted Logistic Regression on Weight

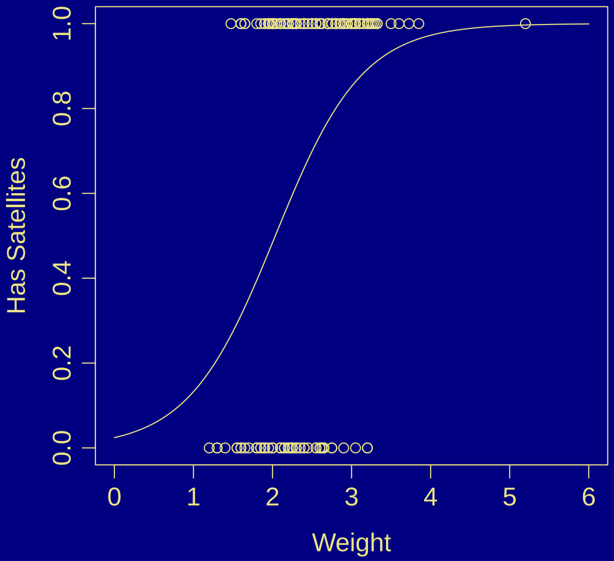
- $\hat{\beta} > 0$, so $\hat{\pi} \uparrow$ as $x \uparrow$
- $\hat{\pi} = \frac{1}{2}$ when $x = -\dfrac{\hat{\alpha}}{\hat{\beta}} = \boxed{\dfrac{3.69}{1.82}} = 2.04$
- $\hat{\pi} \approx \frac{3}{4}$ when $x = 2.04 + \dfrac{1}{\hat{\beta}} = 2.04 + \dfrac{1}{\boxed{1.82}} = 2.04 + 0.55 = 2.59$
- $\hat{\pi} \approx \frac{1}{4}$ when $x = 2.04 - 0.55 = 1.48$
- At $x = 2.04$, the estimated slope is

$$\hat{\beta}\hat{\pi}(1 - \hat{\pi}) = \frac{\hat{\beta}}{4} = \frac{\boxed{1.82}}{4} = 0.454,$$

i.e., for a small change in weight, $\Delta x$,

$$\hat{\pi}(2.04 + \Delta x) \approx \hat{\pi}(2.04) + 0.454\,(\Delta x) = 0.5 + 0.454\,(\Delta x)$$

```
> logit <- make.link("logit")
> ab <- coef(crabs.fit1)
> attach(horseshoecrabs)
> plot(Weight, (Satellites > 0), xlim=c(0,6), ylim=c(0,1),
       xlab="Weight", ylab="Has Satellites")
> curve(logit$linkinv(ab[1] + ab[2]*x), add=TRUE)
> detach(horseshoecrabs)
```


Has Satellites
0.0
0.2
0.4
0.6
0.8
1.0
0
1
2
3
4
5
6
Weight

# Horseshoe Crabs: Fitted Logistic Regression on Weight (ctd)

- Instantaneous rate of change of $\hat{\pi}(x)$ at $x = 2.04$ is the slope, 0.454 per kg change in weight. This means that for a small change of $\Delta x$ kg in weight, $\hat{\pi}$ changes by about 0.454 $(\Delta x)$.

  What is “small” here?

  Sample std dev of weights is $s = 0.58$; half the interquartile range is 0.43. Small should be small relative to these amounts.

| $\Delta x$ | $\hat{\pi}(2.04 + \Delta x)$ | $0.5 + (0.454)\,(\Delta x)$ | Approximation is |
|---|---|---|---|
| 0.1 | 0.545 | 0.545 | Good |
| 1.0 | 0.86 | 0.954 | Poor |

```
> sd(horseshoecrabs$Weight)

[1] 0.57703

> IQR(horseshoecrabs$Weight)/2

[1] 0.425
```

## Horseshoe Crabs: Fitted Logistic Regression on Weight (ctd)

- At $x = 5.2$ (max. obs. wt.), $\hat{\pi} = 0.997$, and est. slope is $\hat{\beta}\hat{\pi}(1 - \hat{\pi}) = \boxed{(1.82)(0.997)(0.003)} = 0.0058$.

  If $x$ increases by 0.1 kg, then $\hat{\pi}$ increases by $\approx (0.1)(0.0058) = 0.00058$.
- Rate of change of $\hat{\pi}(x)$ varies with $x$.
  E.g., it is $\boxed{0.454}$ at $x = 2.04$ and $\boxed{0.0058}$ at $x = 5.2$.

## 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)
- An alternative way to describe effect (not dependent on units of $x$) is

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

  For $x =$ weight, $LQ = 2.00$, $UQ = 2.85$.
  At $x = 2.00$, $\hat{\pi} = 0.483$; at $x = 2.85$, $\hat{\pi} = 0.814$.

  $\Longrightarrow$ $\hat{\pi}$ increases by 0.331 over middle half of $x$ values.

# Odds Ratio Interpretation

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

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

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

More generally,

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

If $\beta = 0$, then $e^{\alpha+\beta x} = e^{\alpha}$ and odds do not depend on $x$.

## Horseshoe Crabs (ctd)

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

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

$$e^{\boxed{(1.82)(0.1)}} = e^{0.182} = 1.20,$$

i.e., by $\boxed{20}$%.

# 4.2 Inference for Logistic Regression

## Confidence Intervals

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

### Horseshoe Crabs (ctd)

95% CI for $\beta$:

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

95% CI for $e^{\beta}$, multiplicative effect of a 1-kg increase in weight on odds:

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

95% CI for $e^{0.1\,\beta}$, multiplicative effect on odds of 100-gram increase, is

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

Odds estimated to increase by at least $\boxed{11}$ % and at most $\boxed{29}$ %.

## Remarks

- Safer to use LR CI than Wald CI.

  For crabs example, 95% LR CI for $e^{\beta}$ is (see next slide)

$$\left(e^{\boxed{1.11}}, e^{\boxed{2.60}}\right) = \boxed{(3.0, 13.4)}$$

- Can also construct CI for $\pi(x)$. The convenience function `predCI()` in the `icda` package does the calculation described in Section 4.2.6 of the text (see next slide).
  - For crabs data, at $x = 3.05$ (first crab), $\hat{\pi} = 0.863$.
    A 95% CI for $\pi$ at $x = 3.05$ is $(0.766, 0.924)$.
  - For crabs data, at $x = \bar{x} = 2.44$, $\hat{\pi} = \boxed{0.675}$.

    A 95% CI for $\pi$ at $x = 2.44$ is $\boxed{(0.592, 0.748)}$.

```
> confint(crabs.fit1)

              2.5 %  97.5 %
(Intercept) -5.5059 -2.0397
Weight       1.1138  2.5973

> exp(confint(crabs.fit1)[2,])

  2.5 %  97.5 %
 3.0459 13.4275

> crabs.predCI <- predCI(crabs.fit1)
> crabs.predCI[1,]

    fit     lwr     upr
0.86312 0.76606 0.92391

> xbar <- mean(horseshoecrabs$Weight)
> predCI(crabs.fit1, newdata=data.frame(Weight=xbar))

      fit     lwr     upr
1 0.67461 0.59213 0.74753
```

# Hypothesis Tests for $\beta$

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

**Wald Test**

$$z = \frac{\hat{\beta}}{\text{SE}} = \frac{\boxed{1.815}}{\boxed{0.377}} = \boxed{4.82} \quad \text{or} \quad z^2 = 23.2, \ \text{df} = 1 \text{ (chi-squared)}$$

p-value $< 0.0001$ : very strong evidence that $\pi \uparrow$ as weight $\uparrow$

**Likelihood ratio test**

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

When $\beta = \hat{\beta}, \ L_1 = -97.87$

Test stat : $-2(L_0 - L_1) = 30.02, \ \text{df} = 1$ (chi-sq)

p-value $< 0.0001$

```
> drop1(crabs.fit1, 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
> # anova(crabs.fit1, test="Chisq")
```

## 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 $M_1$,

$$\begin{aligned}\text{LR statistic} &= -2(L_0 - L_1)\\ &= -2\big[(L_0 - L_S) - (L_1 - L_S)\big]\\ &= \big[-2(L_0 - L_S)\big] - \big[-2(L_1 - L_S)\big]\\ &= \text{difference of (residual) deviances for two models}\end{aligned}$$

## Horseshoe Crabs (ctd)

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

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

diff. of deviances $= \boxed{225.76} - \boxed{195.74} = 30.02 =$ LR stat.

# 4.3–4.4 Multiple Logistic Regression

Y binary, $\pi = \text{Pr}(Y = 1)$

$x_1, x_2, \ldots, x_k$ 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

## Horseshoe Crabs: Logistic Regression on Color and Weight

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

$x =$ Weight

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

$$c_2 = \begin{cases} 1 & \text{medium} \\ 0 & \text{o/w} \end{cases} \qquad c_3 = \begin{cases} 1 & \text{dark med} \\ 0 & \text{o/w} \end{cases} \qquad c_4 = \begin{cases} 1 & \text{dark} \\ 0 & \text{o/w} \end{cases}$$

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.

### Remark

To match textbook's dummy variables ($c_1$, $c_2$, $c_3$), use

```
> options(contrasts=c("contr.SAS","contr.poly"))
```

We are using R's default, which is

```
> options(contrasts=c("contr.treatment","contr.poly"))
```

Textbook also uses crab width instead of weight.

```
> horseshoecrabs <-
    transform(horseshoecrabs, C = as.factor(Color))
> levels(horseshoecrabs$C)

[1] "1" "2" "3" "4"

> crabs.fit2 <-
    glm((Satellites > 0) ~ C + Weight, family=binomial,
        data=horseshoecrabs)


> summary(crabs.fit2)


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


Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.191  -1.014   0.510   0.868   2.075
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.257      1.198   -2.72   0.0066
C2             0.145      0.736    0.20   0.8441
C3            -0.186      0.775   -0.24   0.8102
C4            -1.269      0.849   -1.50   0.1348
Weight         1.693      0.389    4.35  1.3e-05

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 188.54  on 168  degrees of freedom
AIC: 198.5

Number of Fisher Scoring iterations: 4
```

# Horseshoecrabs: Logistic Regression on Color and Weight (ctd)

Model:

$$\text{logit}\left[\Pr(Y=1)\right] = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x$$

has ML fit

$$\text{logit}(\hat{\pi}) = \boxed{-3.26 + 0.14 c_2 - 0.19 c_3 - 1.27 c_4 + 1.69 x}$$

- For light med. female ($c_2 = c_3 = c_4 = 0$),

$$\text{logit}(\hat{\pi}) = \boxed{-3.26 + 1.69 x}$$

At $x = \bar{x} = 2.44$,

$$\hat{\pi} = \boxed{\frac{\exp\{-3.26 + 1.69(2.44)\}}{1 + \exp\{-3.26 + 1.69(2.44)\}}} = 0.704$$

## Horseshoecrabs: Logistic Regression on Color and Weight (ctd)

- For medium female ($c_2 = 1$, $c_3 = c_4 = 0$),

$$\text{logit}(\hat{\pi}) = \boxed{-3.26 + 0.14 + 1.69x} = -3.11 + 1.69x$$

  At $x = \bar{x} = 2.44$, $\hat{\pi} = 0.734$.

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

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

  Estimated <u>odds</u> a medium color female has satellites are $\boxed{1.16}$ times estimated odds for a light med. female of the same weight.

  E.g., at $x = 2.44$,

$$\frac{\text{odds for medium}}{\text{odds for light-med}} = \frac{0.734/0.266}{0.704/0.296} = 1.16$$

## Horseshoecrabs: Logistic Regression on Color and Weight (ctd)

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

$$\hat{\beta}_4 - \hat{\beta}_2 = -1.269 - 0.145 = -1.41 \qquad e^{-1.41} = 0.243$$

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

Equivalently,

$$0.145 - (-1.269) = 1.41 \qquad e^{1.41} = 4.11 \quad (= 1/0.243)$$

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

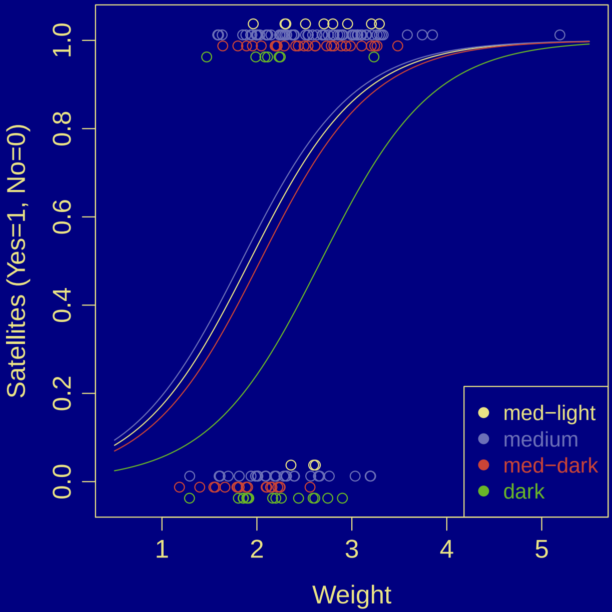
## Horseshoecrabs: Logistic Regression on Color and Weight (ctd)

- Model assumes no interaction between color and weight effects.

  Coef. of $x =$ 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.


Satellites (Yes=1, No=0)
0.0
0.2
0.4
0.6
0.8
1.0
1
2
3
4
5
Weight
● med−light
● medium
● med−dark
● dark

# Horseshoecrabs: Logistic Regression on Color and Weight (ctd)

- Do we need color in the model?

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{(given weight, Y indep. of color)}$$

Likelihood-ratio statistic

$$-2(L_0 - L_1) = -2\big[(-97.9) - (-94.3)\big] = 7.19$$

or

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

$$\text{df} = 171 - 168 = 3 \qquad \text{p-value} = 0.066$$

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

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

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

Analysis of Deviance Table


Model 1: (Satellites > 0) ~ Weight
Model 2: (Satellites > 0) ~ C + Weight
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       171        196
2       168        188  3     7.19    0.066

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

Single term deletions


Model:
(Satellites > 0) ~ C + Weight
       Df Deviance AIC   LRT Pr(>Chi)
<none>         188 198
C       3      196 200  7.19    0.066
Weight  1      212 220 23.52  1.2e-06
```

## Horseshoe Crabs: Logistic Regression on Color and Weight (ctd)

Other simple models are also adequate.

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

suggests

$$\text{logit}(\pi) = \alpha + \beta_1 z + \beta_2 x, \qquad z = \begin{cases} 1, & \text{dark} \\ 0, & \text{o/w} \end{cases}$$

ML gives $\hat{\beta}_1 = -1.295$ (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.fit3 <-
    glm((Satellites > 0) ~ I(Color == 4) + Weight,
        family=binomial, data=horseshoecrabs)
> summary(crabs.fit3)

Call:
glm(formula = (Satellites > 0) ~ I(Color == 4) + Weight, fami
    data = horseshoecrabs)


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.fit3, crabs.fit2, 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 Pr(>Chi)
1       170        189
2       168        188  2    0.629     0.73
```

## Horseshoe Crabs: Logistic Regression on Color and Weight (ctd)

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

$$\mathrm{H}_0 : \text{simple model} \quad \text{vs} \quad \mathrm{H}_a : \text{more complex model}$$

Note $\mathrm{H}_0$ is $\beta_2 = \beta_3 = 0$ in more complex model.

$$\text{LR stat} = \text{diff. in deviances} = 189.17 - 188.54 = 0.63$$

$$\text{df} = 170 - 168 = 2 \qquad \text{p-value} = 0.73$$

Simpler model appears to be adequate.

## Horseshoe Crabs: Logistic Regression on Color and Weight (ctd)

How about interaction?

$$\text{logit}(\pi) = \alpha + \beta_2 c_2 + \beta_3 c_3 + \beta_4 c_4 + \beta x + \gamma_2 c_2 x + \gamma_3 c_3 x + \gamma_4 c_4 x$$

| Color | Dummies | Weight Coef |
|---|---|---|
| light-med | $c_2 = c_3 = c_4 = 0$ | $\beta$ |
| medium | $c_2 = 1, c_3 = c_4 = 0$ | $\beta + \gamma_2$ |
| dark-med | $c_3 = 1, c_2 = c_4 = 0$ | $\beta + \gamma_3$ |
| dark | $c_4 = 1, c_2 = c_3 = 0$ | $\beta + \gamma_4$ |

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

LR stat $= 188.54 - 181.66 = 6.89$ df $= 3$ p-value $= 0.076$

Weak evidence of interaction.

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

```
> crabs.fit4 <-
    update(crabs.fit2, . ~ C*Weight)
> deviance(crabs.fit4)

[1] 181.66

> anova(crabs.fit2, crabs.fit4, 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 Pr(>Chi)
1       168        188
2       165        182  3     6.89    0.076
```

```
> drop1(crabs.fit4, test="Chisq")

Single term deletions

Model:
(Satellites > 0) ~ C + Weight + C:Weight
         Df Deviance AIC  LRT Pr(>Chi)
<none>           182 198
C:Weight  3      188 198 6.89    0.076
```

# Quantitative Treatment of Ordinal Factors

Models with dummy variables for a factor treat that factor as qualitative (nominal), i.e., order is ignored.

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

## Horseshoe Crabs: Logistic Regression on Color and Weight (ctd)

Recall that color was originally coded with numerical scores $(1, 2, 3, 4)$. Model:

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

```
> crabs.fit5 <-
    glm((Satellites > 0) ~ Weight + Color,
        family=binomial, data=horseshoecrabs)

> summary(crabs.fit5)
```

```
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
```

## Horseshoe Crabs: Logistic Regression on Color and Weight (ctd)

ML estimates and SEs are

$$\hat{\alpha} = -2.03\ (1.12) \qquad \hat{\beta}_1 = 1.65\ (0.38) \qquad \hat{\beta}_2 = -0.51\ (0.22)$$

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

$\hat{\pi} \downarrow$ as Color $\uparrow$, 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

## Horseshoe Crabs: Logistic Regression on Color and Weight (ctd)

Does model treating color as nominal fit as well as model treating it as qualitative?

$H_0$ : simpler (ordinal) model holds

$H_a$ : more complex (nominal) model holds

$$\begin{aligned} \text{LR stat} &= -2(L_0 - L_1) \\ &= \text{diff in deviances} \\ &= 190.27 - 188.54 = 1.73, \qquad \text{df} = 2 \end{aligned}$$

Do not reject $H_0$. Simpler model appears to be adequate.

```
> anova(crabs.fit5, crabs.fit2, test="Chisq")

Analysis of Deviance Table

Model 1: (Satellites > 0) ~ Weight + Color
Model 2: (Satellites > 0) ~ C + Weight
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       170        190
2       168        188  2     1.73     0.42
```

## FL Death Penalty Revisited

```
> 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.

```
> 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
> library(reshape2)
> dp <- melt(deathpenalty)
> dpwide <- dcast(dp, ... ~ DeathPenalty)
> 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

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

> summary(dp.fit1)
```

```
Call:
glm(formula = cbind(Yes, No) ~ Defendant + Victim, family = binomia
    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    0.018
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
AIC: 19.3

Number of Fisher Scoring iterations: 4
```

# FL Death Penalty Revisited

$$\pi = \Pr(Y = \text{yes}) \quad \text{death penalty}$$

$$v = \begin{cases} 1, & \text{victim black} \\ 0, & \text{victim white} \end{cases} \qquad d = \begin{cases} 1, & \text{defendant black} \\ 0, & \text{defendant white} \end{cases}$$

Model:

$$\text{logit}(\pi) = \alpha + \beta_1 d + \beta_2 v$$

ML fit:

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

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

95% CI for odds-ratio is

$$e^{0.868 \pm 1.96(0.367)} = (e^{0.148}, e^{1.59}) = (1.16, 4.89)$$

## Remarks

- No interaction term means estimated odds ratio between $Y$ and
  - $\mathrm{d}$ same at each level of $v$ ($e^{0.868} = 2.38$)
  - $v$ same at each level of $\mathrm{d}$ ($e^{-2.40} = 0.09$)

    For white vic vs black vic: $e^{2.40} = \frac{1}{0.09} = 11.1$

  Homogeneous association: odds ratio does not depend on level of other explanatory variable.

- Test $\mathrm{H}_0 : \beta_1 = 0$ ($Y$ cond. indep. of $\mathrm{d}$ given $v$) vs $\mathrm{H}_a : \beta_1 \neq 0$

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

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

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

```
> drop1(dp.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

> dp.fit2 <- update(dp.fit1, . ~ Victim)
> deviance(dp.fit2)

[1] 5.394

> df.residual(dp.fit2)

[1] 2
```

## Multi-Center Trials

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

| Center | Treatment | Response | |
|---|---|---|---|
| | | S | F |
| 1 | 1 | | |
| | 2 | | |
| 2 | 1 | | |
| | 2 | | |
| ⋮ | ⋮ | ⋮ | |
| K | 1 | | |
| | 2 | | |

$$\text{logit}\big[\Pr(Y = 1)\big] = \alpha + \beta_2 c_2 + \cdots + \beta_K c_K + \beta x$$

Assumes odds ratio $e^{\beta}$ is the same for each center.

A model like this is commonly expressed in the form

$$\text{logit}\big[\Pr(Y = 1)\big] = \alpha + \beta_i^c + \beta x$$

$\beta_i^c$ 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 Review: Binomial Distribution

- Recognize.
- Compute probabilities, mean, sd.
- Wald test and CI for a single proportion.
- Score test and CI 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: $\dfrac{\pi_1}{\pi_2}$
  - Odds ratio: $\theta = \dfrac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$
    - 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 $\sigma = \sqrt{2\,\text{df}}$.

# Exam 1 Review: Testing Independence in $\mathrm{I} \times \mathrm{J}$ Tables II

- Examining sources of dependence
  - Adjusted residuals (analyze dependence)
    - Should be approx. $\mathrm{N}(0, 1)$ if all expected freqs $\geqslant 5$.
    - If so, then $|\text{adj resid}| \geqslant 2$ (or 3 in big tables) meaningful.
  - Partitioning chi-square

    With a correct partition

    - $\mathrm{G}^2$ stats add up ($\mathrm{X}^2$ 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 row and column totals (i.e., treats them as fixed).
- 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

$$\underbrace{g(\mu)}_{\text{link}} = \underbrace{\alpha + \beta_1 x_1 + \cdots + \beta_k x_k}_{\text{linear predictor}}$$

- Common link functions

$$\text{Identity: } g(\mu) = \mu$$

$$\text{Log: } g(\mu) = \log(\mu)$$

$$\text{Logit: } g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$$

# 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
  - 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

  $$\log\left(\frac{\mu}{t}\right) = \alpha + \beta x \implies \log(\mu) = \underbrace{\log(t)}_{\text{offset}} + \alpha + \beta x$$

- 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

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

$$\pi = \text{prob. of ”success”} \qquad \frac{\pi}{1-\pi} = \text{odds of success}$$

- 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 $\beta\pi(x)\big[1-\pi(x)\big]$. Steepest at $x = -\alpha/\beta$ where $\pi = 1/2$ and slope $= \beta/4$. Flattest when $\pi(x)$ close to 0 or 1.
- $\dfrac{1}{|\beta|} \approx$ dist. between $x$ values with $\pi = 0.5$ and $\pi = 0.75$ (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:

  Test stat: $$z = \frac{\hat{\beta} - \beta_0}{\text{SE}}$$

  CI: $$\hat{\beta} \pm z_{\alpha/2}\,\text{SE}$$

- 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 $H_0 : \beta = 0$:

$$\begin{aligned}\text{LR test statistic} &= -2[L_0 - L_1] \\ &= \text{deviance}_0 - \text{deviance}_1\end{aligned}$$

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 $\hat{\alpha}$, $\hat{\beta}$

deviance$_0$ = ”null deviance” in R

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 $M_0 \subset M_1$

  $$\begin{aligned} \text{LR test statistic} &= -2[L_0 - L_1] \\ &= \text{deviance}_0 - \text{deviance}_1 \end{aligned}$$

  $$\begin{aligned} \text{df} &= \text{num. free params in } M_1 - \text{num. free params in } M_0 \\ &= \text{residual df for } M_0 - \text{residual df for } M_1 \end{aligned}$$

  $L_0$ = maximized log-likelihood for $M_0$
  $L_1$ = maximized log-likelihood for $M_1$

  deviance$_0$ = (residual) deviance for $M_0$
  deviance$_1$ = (residual) deviance for $M_1$

# Exam 1: Time and Place

Tuesday, Feb 21, 2012
8:30 a.m. – 9:25 a.m.
Griffin-Floyd Hall (FLO)
Room 100

# Ch 5: Building Logistic Regression Models

- Model selection
- Model checking
- Problems w/ sparse categorical data (estimators may be infinite)

# 5.1 Strategies in Model Selection

## 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 $s_1$, $s_2$

```
> horseshoecrabs <-
    transform(horseshoecrabs,
              C = as.factor(Color),
              S = as.factor(Spine))
> options(contrasts=c("contr.SAS","contr.poly"))
> 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, fami
    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   -2.42   0.0157
C1             1.609      0.936    1.72   0.0855
C2             1.506      0.567    2.66   0.0079
C3             1.120      0.593    1.89   0.0591
S1            -0.400      0.503   -0.80   0.4259
S2            -0.496      0.629   -0.79   0.4302
Weight         0.826      0.704    1.17   0.2407
Width          0.263      0.195    1.35   0.1779
```

```
(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
```

## Horseshoe Crab Study

Consider model for crabs:

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

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

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

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

Strong evidence at least one predictor assoc. w/ presence of satellites.

But look back at Wald tests for partial effects of weight and width.
Or better, look at LR tests of all partial effects (next slide).

```
> drop1(crabs.fitall, test="Chisq")

Single term deletions

Model:
(Satellites > 0) ~ C + S + Weight + Width
       Df Deviance AIC  LRT Pr(>Chi)
<none>         185 201
C       3      193 203 7.60    0.055
S       2      186 198 1.01    0.604
Weight  1      187 201 1.41    0.235
Width   1      187 201 1.80    0.180
```

# Multicollinearity

Multicollinearity (strong correlations among predictors) plays havoc with GLMs just as it does with LMs.

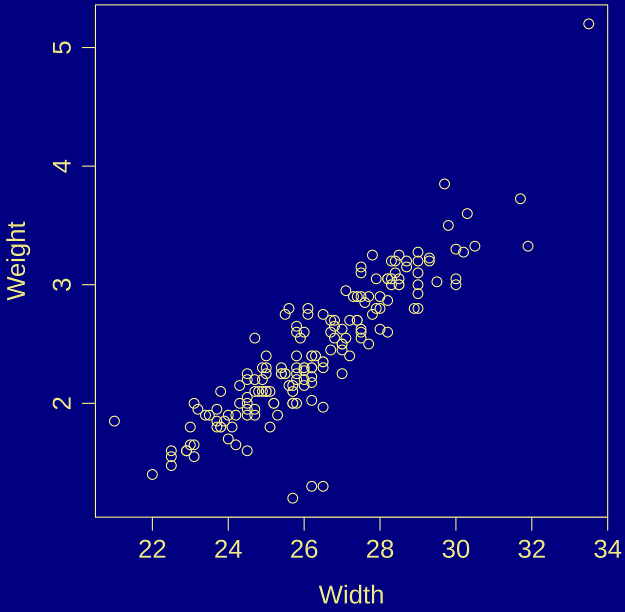
E.g., $\mathrm{Corr}(\text{width}, \text{weight}) = 0.89$.

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

```
> attach(horseshoecrabs)
> cor(Weight, Width)

[1] 0.88687
```

```
> plot(Width, Weight)
```

```
> detach(horseshoecrabs)
```


Weight
5
4
3
2
22
24
26
28
30
32
34
Width

# Backward Elimination

- Use $W =$ width, $C =$ color, $S =$ 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 begin by testing all at one time as on next slide.

```
> 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

Model 1: (Satellites > 0) ~ C + S + Width
Model 2: (Satellites > 0) ~ C * S * Width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       166        187
2       152        170 14      16.2      0.3
```

## Horseshoe Crabs: Backward Elimination

H$_0$: Model $C + S + W$ holds (has 3 parameters for $C$, 2 for $S$, 1 for $W$)

H$_a$: 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 $= 186.6 - 170.4 = \boxed{16.2}$

df $= 166 - 152 = \boxed{14}$ $\boxed{\text{p-value} = 0.30}$

Simpler model $C + S + W$ appears to be adequate.

## Remark

df $= 14$ on previous slide is unexpected. Model $C * S * W$ has
$3 \times 2 = \underline{6}$ parameters for $C \times S$ interaction,
$3 \times 1 = \underline{3}$ for $C \times W$,
$2 \times 1 = \underline{2}$ for $S \times W$, and
$3 \times 2 \times 1 = \underline{6}$ for $C \times S \times W$,
so $6 + 3 + 2 + 6 = \underline{17}$ more parameters than model $C + S + W$.
However, 3 combinations of $C$ and $S$ have only one obs. each, so 3 of the $C \times S \times W$ interaction coef.'s cannot be estimated.

```
> with(horseshoecrabs, table(C,S))

   S
C    1  2  3
  1  9  2  1
  2 24  8 63
  3  3  4 37
  4  1  1 20
```

### Remark

In this example, we end up with the same model if we eliminate higher order interactions 1 at a time. Try the following sequence of commands to see this.

```
> drop1(crabs.fit1, test="Chisq")
> crabs.fit1a <-
    update(crabs.fit1, . ~ . - C:S:Width)
> drop1(crabs.fit1a, test="Chisq")
> crabs.fit1b <- update(crabs.fit1a, . ~ . - S:Width)
> drop1(crabs.fit1b, test="Chisq")
> crabs.fit1c <- update(crabs.fit1b, . ~ . - C:Width)
> drop1(crabs.fit1c, test="Chisq")
```

## Horseshoe Crabs: Backward Elimination (ctd)

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

$$\text{diff. in deviances} = 187.46 - 186.61 = 0.85, \qquad \text{df} = 2$$

```
> drop1(crabs.fit2, test="Chisq")
Single term deletions

Model:
(Satellites > 0) ~ C + S + Width
       Df Deviance AIC   LRT Pr(>Chi)
<none>         187 201
C       3      194 202  7.81     0.05
S       2      188 198  0.85     0.66
Width   1      209 221 22.22  2.4e-06
```

```
> ## crabs.fit3 <- update(crabs.fit2, . ~ . - S)
> crabs.fit3 <- update(crabs.fit2, . ~ C + Width)
> deviance(crabs.fit3)

[1] 187.46

> deviance(crabs.fit2)

[1] 186.61

> anova(crabs.fit3, crabs.fit2, test="Chisq")

Analysis of Deviance Table

Model 1: (Satellites > 0) ~ C + Width
Model 2: (Satellites > 0) ~ C + S + Width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       168        188
2       166        187  2    0.845     0.66
```

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

Single term deletions

Model:
(Satellites > 0) ~ C + Width
       Df Deviance AIC  LRT Pr(>Chi)
<none>         188 198
C       3      194 198  7.0    0.072
Width   1      212 220 24.6    7e-07

> 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)
```

```
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 187.46  on 168  degrees of freedom
AIC: 197.5

Number of Fisher Scoring iterations: 4
```

## Horseshoe Crabs: Backward Elimination (ctd)

Results in model fit

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

Forcing $\beta_1 = \beta_2 = \beta_3$ gives

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

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

Model 1: (Satellites > 0) ~ I(C == "4") + Width
Model 2: (Satellites > 0) ~ C + Width
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       170        188
2       168        188  2    0.501     0.78

> summary(crabs.fit4)
```

```
Call:
glm(formula = (Satellites > 0) ~ I(C == "4") + Width, family
    data = horseshoecrabs)

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-2.082  -0.993   0.527   0.861   2.155

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)       -12.980      2.727   -4.76  1.9e-06
I(C == "4")FALSE    1.301      0.526    2.47    0.013
Width               0.478      0.104    4.59  4.4e-06

(Dispersion parameter for binomial family taken to be 1)
```

## Horseshoe Crabs Study

Conclude:

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

  $$95\%\text{CI}: \boxed{e^{1.301 \pm 1.96(0.526)} = \left(e^{0.270}, e^{2.33}\right) = (1.3, 10.3)}$$

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

  $$95\%\text{CI}: \boxed{e^{0.478 \pm 1.96(0.104)} = \left(e^{0.274}, e^{0.682}\right) = (1.3, 2.0)}$$

# Criteria for Selecting a Model I

- Use theory, other research as guide.
- Parsimony (simplicity) is good.
- Can use a model selection criterion to choose among models. Most popular is Akaiki information criterion (AIC).

  Choose model with minimum AIC where

  $$\text{AIC} = -2\mathrm{L} + 2(\text{number of model parameters})$$

  with $\mathrm{L} =$ log-likelihood.
- For exploratory purposes, can use automated procedure such as backward elimination, but not generally recommended.

  R function `step()` will do stepwise selection procedures (forward, backward, or both).

# Criteria for Selecting a Model II

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

  Example: $n = 1000$, $(Y = 1)$ 30 times, $(Y = 0)$ 970 times

  $$\text{Model should contain} \leqslant \tfrac{30}{10} = 3 \text{ predictors.}$$

  Example: $n = 173$ crabs, $(Y = 1)$ 111 crabs, $(Y = 0)$ 62 crabs

  $$\text{Use} \leqslant \tfrac{62}{10} \approx 6 \text{ predictors.}$$

- 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 $\hat{\pi}$.

| Model | Correlation |
|---|---|
| color | 0.285 |
| width | 0.402 |
| color + width | 0.452 |
| dark + width | 0.447 |

```
> crabs.color <- glm((Satellites > 0) ~ C, family=binomial,
                      data=horseshoecrabs)
> crabs.width <- update(crabs.color, . ~ Width)
> crabs.color.width <- update(crabs.color, . ~ C + Width)
> crabs.dark.width <-
     update(crabs.color, . ~ I(C == "4") + Width)
> y <- as.numeric(horseshoecrabs$Satellites > 0)
> pihat <- predict(crabs.color, type="response")
> cor(y,pihat)

[1] 0.28526

> pihat <- predict(crabs.width, type="response")
> cor(y,pihat)

[1] 0.40198

> pihat <- predict(crabs.color.width, type="response")
> cor(y,pihat)

[1] 0.45221

> pihat <- predict(crabs.dark.width, type="response")
> cor(y,pihat)

[1] 0.44697
```

# Summarizing Predictive Power

## Classification Tables

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

### Horseshoe Crabs with Width and Color as Predictors

|  | Predicted |  |  |
|---|---|---|---|
| Actual | $\hat{y} = 1$ | $\hat{y} = 0$ | Total |
| $y = 1$ | 96 | 15 | 111 |
| $y = 0$ | 31 | 31 | 62 |

$$\text{Sensitivity} = \Pr(\widehat{Y} = 1 | Y = 1) \approx \frac{96}{111} = 0.86$$

$$\text{Specificity} = \Pr(\widehat{Y} = 0 | Y = 0) \approx \frac{31}{62} = 0.50$$

$$\Pr(\text{correct classification}) \approx \frac{96 + 31}{173} = 0.73$$

```
> pihat <- predict(crabs.color.width, type="response")
> yhat <- as.numeric(pihat > 0.50)
> y <- as.numeric(horseshoecrabs$Satellites > 0)
> 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
```

## Remark

Table 5.3 in text actually produced by (approximate) leave-one-out *cross-validation*, which gives more realistic estimates. For $i = 1, \ldots, n$:

1. Fit the model to the data leaving out $i$th obs.
2. Use fitted model and $x_i$ to compute $\hat{\pi}_{(i)}$.
3. Predict $\hat{y}_i = 1$ if $\hat{\pi}_{(i)} > 0.50$ and $\hat{y}_i = 0$ if $\hat{\pi}_{(i)} < 0.50$.

| | Predicted | | |
|---|---|---|---|
| Actual | $\hat{y} = 1$ | $\hat{y} = 0$ | Total |
| $y = 1$ | 94 | 17 | 111 |
| $y = 0$ | 34 | 28 | 62 |

$$\text{Sensitivity} = \Pr(\widehat{Y} = 1 | Y = 1) \approx \frac{94}{111} = 0.85$$

$$\text{Specificity} = \Pr(\widehat{Y} = 0 | Y = 0) \approx \frac{28}{62} = 0.45$$

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

```
> pihat <- vector(length=173)
> for (i in 1:173) {
    pihat[i] <-
      predict(update(crabs.color.width, subset=-i),
              newdata=horseshoecrabs[i,], type="response")
  }

> yhat <- as.numeric(pihat > 0.50)
> y <- as.numeric(horseshoecrabs$Satellites > 0)
> confusion <- table(y, yhat)
> confusion

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

```
> prop.table(confusion, 1)

   yhat
y         0       1
  0 0.45161 0.54839
  1 0.15315 0.84685

> sum(diag(confusion))/sum(confusion)

[1] 0.7052

> yhat <- as.numeric(pihat > 0.64)
> table(y,yhat)

   yhat
y    0  1
  0 42 20
  1 37 74
```

Could use cut-offs other than $\pi_0 = 0.5$. E.g., for the crabs data, $\pi_0 = \frac{111}{173} = 0.64$ ($\hat{\pi}$ for intercept-only model).

|  | Predicted |  |  |
|---|---|---|---|
| Actual | $\hat{y} = 1$ | $\hat{y} = 0$ | Total |
| $y = 1$ | 74 | 37 | 111 |
| $y = 0$ | 20 | 42 | 62 |

$$\text{Sensitivity} = \Pr(\widehat{Y} = 1 | Y = 1) \approx \frac{74}{111} = 0.67$$

$$\text{Specificity} = \Pr(\widehat{Y} = 0 | Y = 0) \approx \frac{42}{62} = 0.68$$

$$\Pr(\text{correct classification}) \approx \frac{74 + 42}{173} = 0.67$$

Note: As cutoff $\pi_0$ increases, sensitivity decreases and specificity increases.

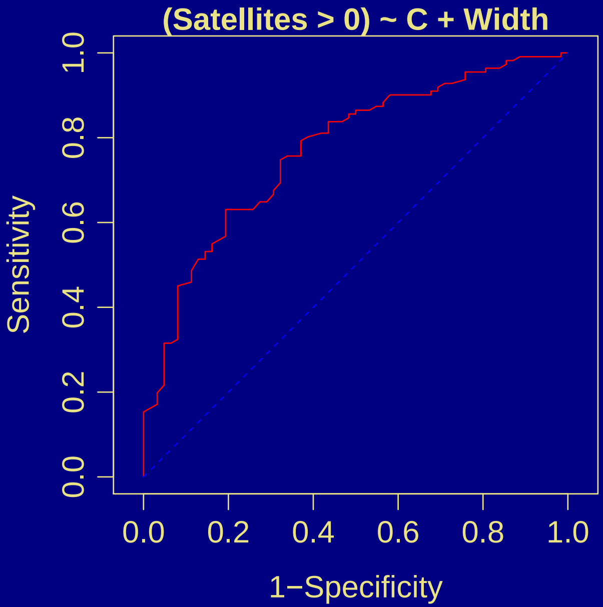
## Receiver Operating Characteristic (ROC) Curve

The *receiver operating characteristic* (ROC) curve plots sensitivity against 1 $-$ specificity as the cutoff $\pi_0$ varies from 0 to 1.

- The higher the sensitivity for a given specificity, the better, so a model with a higher ROC curve is preferred to one with a lower ROC curve.
- The area under the ROC curve is a measure of predictive power, called the *concordance index*, $c$.
  - Models w/ bigger $c$ have better predictive power.
  - $c = 1/2$ is no better than random guessing.
- If feasible, use cross-validation.

The slide after the next shows ROC curve for horseshoecrab data using color and width as predictors ($c = 0.77$).

```
> library(epicalc)
> lroc(crabs.width, graph=FALSE)$auc

[1] 0.74244

> lroc(crabs.color, graph=FALSE)$auc

[1] 0.63862

> lroc(crabs.color.width, graph=FALSE)$auc

[1] 0.77136

> lroc(crabs.dark.width, graph=FALSE)$auc

[1] 0.77201


> lroc(crabs.color.width, grid=FALSE, title=TRUE)
```


(Satellites > 0) ~ C + Width
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
1−Specificity
0.0
0.2
0.4
0.6
0.8
1.0

# 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 generally 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

| | | Death Penalty | | |
|---|---|---|---|---|
| Victim | Defendant | Yes | No | n |
| Black | Black | 4 | 139 | 143 |
| | White | 0 | 16 | 16 |
| White | Black | 11 | 37 | 48 |
| | White | 53 | 414 | 467 |

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

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

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

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

## Florida Death Penalty Data (ctd)

Fitted counts for 467 cases with $d = v = 0$:

$$\text{"Yes"}: \boxed{467 \times 0.113} = 52.8 \qquad \text{"No"}: \boxed{467 \times 0.887} = 414.2$$

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

Summarizing fit over 8 cells of table:

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

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

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

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

```
> formula(dp.fit1)

cbind(Yes, No) ~ Defendant + Victim

> deviance(dp.fit1)

[1] 0.37984

> df.residual(dp.fit1)

[1] 1

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

[1] 0.53769

> chisqstat(dp.fit1)

[1] 0.19779

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

[1] 0.65651
```

## Remarks

- Model assumes lack of interaction between $d$ and $v$ in effects on $Y$ (homogeneous association). Adding interaction term gives saturated model, so goodness-of-fit test in this example is a test of $H_0$: “no interaction”. (Compare next slide to previous.)
- $X^2$ usually recommended over $G^2$ for testing goodness of fit.
- These tests only appropriate for grouped binary data with most ($\geqslant$ 80%) fitted cell counts “large” (e.g., $\hat{\mu}_i \geqslant 5$).
  - Questionable (?) in death penalty example, where $\hat{\mu} = 0.18$ for ($v =$ bl, $d =$ wh, $Y =$ yes) and $\hat{\mu} = 3.82$ for ($v =$ wh, $d =$ bl, $Y =$ yes).
- 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).

```
> dp.saturated <- update(dp.fit1, . ~ Defendant*Victim)
> anova(dp.fit1, dp.saturated, test="LRT")

Analysis of Deviance Table


Model 1: cbind(Yes, No) ~ Defendant + Victim
Model 2: cbind(Yes, No) ~ Defendant + Victim + Defendant:Vict
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1       0.38
2         0       0.00  1     0.38     0.54

> anova(dp.fit1, dp.saturated, test="Rao")

Analysis of Deviance Table


Model 1: cbind(Yes, No) ~ Defendant + Victim
Model 2: cbind(Yes, No) ~ Defendant + Victim + Defendant:Vict
  Resid. Df Resid. Dev Df Deviance   Rao Pr(>Chi)
1         1       0.38
2         0       0.00  1     0.38 0.198     0.66
```

## Residuals for Logit Models

At setting $i$ of explanatory variables, let

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

### Definition (Pearson residuals)

For a binomial GLM, the *Pearson residuals* are

$$e_i = \frac{y_i - n_i\hat{\pi}_i}{\sqrt{n_i\hat{\pi}_i(1-\hat{\pi}_i)}} \qquad \left(\text{note that } X^2 = \sum_i e_i^2\right)$$

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

## Definition (Standardized Pearson residual)

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

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

where $h_i$ is the “leverage” of the $i$th obs.

- A.K.A. “adjusted” Pearson residual.
- $r_i$ approx. $N(0, 1)$ when model holds (and $n_i$ large).
  $|r_i| > 2$ or 3 (approx.) suggests lack of fit.
- R function `rstandard()` provides standardized deviance residuals by default. For standardized Pearson residuals specify `type="pearson"`.

## Example (Berkeley Graduate Admissions)

Data on p. 237 of text.

$$\begin{aligned} Y &= \text{admitted into grad school at UC Berkeley (1=yes, 0=no)} \\ G &= \text{gender (g=1 female, g=0 male)} \\ D &= \text{dept (A, B, C, D, E, F)} \end{aligned}$$

$$d_1 = \begin{cases} 1, & \text{dept B,} \\ 0, & \text{o/w,} \end{cases} \qquad \ldots, \qquad d_5 = \begin{cases} 1, & \text{dept F,} \\ 0, & \text{o/w.} \end{cases}$$

For dept. A, $d_1 = \cdots = d_5 = 0$.

- Model

  $$\text{logit}\big[\Pr(Y = 1)\big] = \alpha + \beta_1 d_1 + \cdots + \beta_5 d_5 + \beta_6 g$$

  seems to fit poorly ($G^2 = 20.2$, $X^2 = 18.8$, $\text{df} = 5$). Apparently there is gender × dept interaction.

```
> 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
```

Ignoring department is misleading (Simpson’s paradox):

```
> margin.table(UCBAdmissions, 2:1)

        Admit
Gender   Admitted Rejected
  Male       1198     1493
  Female      557     1278

> round(prop.table(margin.table(UCBAdmissions, 2:1), 1), 3)

        Admit
Gender   Admitted Rejected
  Male      0.445    0.555
  Female    0.304    0.696

> oddsratio(margin.table(UCBAdmissions, 2:1))

[1] 1.8411
```

```
> 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

   Gender Dept Admitted Rejected
1    Male    A      512      313
2    Male    B      353      207
3    Male    C      120      205
4    Male    D      138      279
5    Male    E       53      138
6    Male    F       22      351
7  Female    A       89       19
8  Female    B       17        8
9  Female    C      202      391
10 Female    D      131      244
11 Female    E       94      299
12 Female    F       24      317
```

```
> options(contrasts=c("contr.treatment","contr.poly"))
> UCB.fit1 <- glm(cbind(Admitted,Rejected) ~ Dept + Gender,
                  family=binomial, data=UCBw)

> summary(UCB.fit1)
```

```
Call:
glm(formula = cbind(Admitted, Rejected) ~ Dept + Gender, fami
    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)    0.5821     0.0690    8.44   <2e-16
DeptB         -0.0434     0.1098   -0.40     0.69
DeptC         -1.2626     0.1066  -11.84   <2e-16
DeptD         -1.2946     0.1058  -12.23   <2e-16
DeptE         -1.7393     0.1261  -13.79   <2e-16
DeptF         -3.3065     0.1700  -19.45   <2e-16
GenderFemale   0.0999     0.0808    1.24     0.22
```

```
(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
```

```
> chisqstat(UCB.fit1)

[1] 18.824

> df.residual(UCB.fit1)

[1] 5

> pchisq(chisqstat(UCB.fit1), df.residual(UCB.fit1),
         lower.tail=FALSE)

[1] 0.0020725

> UCB.fit1.stdres <- rstandard(UCB.fit1, type="pearson")
> round(UCB.fit1.stdres, 2)

    1     2     3     4     5     6     7     8     9
-4.03 -0.28  1.88  0.14  1.63 -0.30  4.03  0.28 -1.88
   10    11    12
-0.14 -1.63  0.30
```

```
> cbind(UCBw, "stdres" = round(UCB.fit1.stdres, 2))

   Gender Dept Admitted Rejected stdres
1    Male    A      512      313  -4.03
2    Male    B      353      207  -0.28
3    Male    C      120      205   1.88
4    Male    D      138      279   0.14
5    Male    E       53      138   1.63
6    Male    F       22      351  -0.30
7  Female    A       89       19   4.03
8  Female    B       17        8   0.28
9  Female    C      202      391  -1.88
10 Female    D      131      244  -0.14
11 Female    E       94      299  -1.63
12 Female    F       24      317   0.30
```

## Example (Berkeley Admissions Ctd)

- Standardized resids suggest Dept. A as main source of lack of fit.
- Leaving out Dept. A, model with no interaction and no gender effect fits well ($G^2 = 2.68$, $X^2 = 2.69$, $\text{df} = 5$).
- 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.,

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

where $\beta_i^X$ 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 = bino
    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 Pr(>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
```

# 5.3 Effects of Sparse Data

Caution: Parameter estimates in logistic regression can be infinite.

Example:

| | | S | F |
|---|---|---|---|
| X | 1 | 8 | 2 |
| | 0 | 10 | 0 |

Model:

$$\log\left(\frac{\Pr(S)}{\Pr(F)}\right) = \alpha + \beta x \implies e^{\hat{\beta}} = \text{sample odds-ratio} = \frac{8 \times 0}{2 \times 10} = 0$$

$$\hat{\beta} = \log(0) = -\infty$$

Example: Text p. 155 for multi-center trial (5 ctrs, each w/ 2 × 2 table). Two centers had no successes under either treatment arm, so estimate of center effect for these two centers is $-\infty$.

Infinite estimates exist when $x$-values where $y = 1$ can be “separated” from $x$-values where $y = 0$.

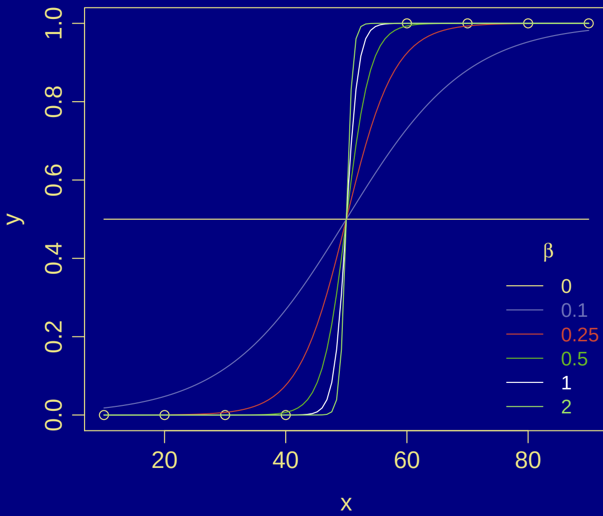
Example: $y = 0$ for $x < 50$ and $y = 1$ for $x > 50$.

$$\text{logit}\Big[\Pr(Y = 1)\Big] = \alpha + \beta x$$

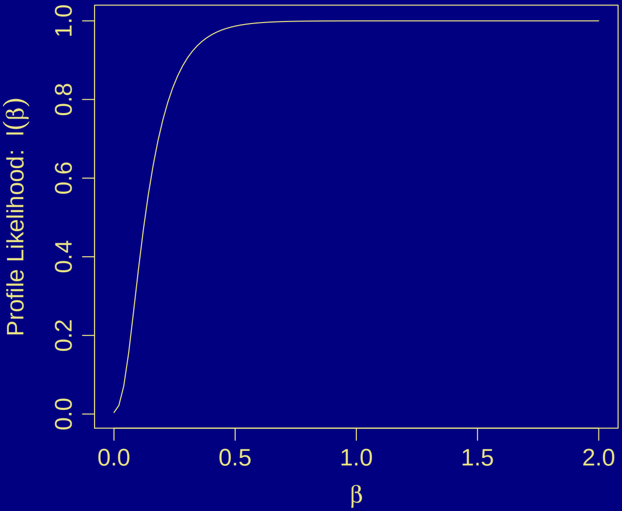
has $\hat{\beta} = \infty$ (roughly speaking).

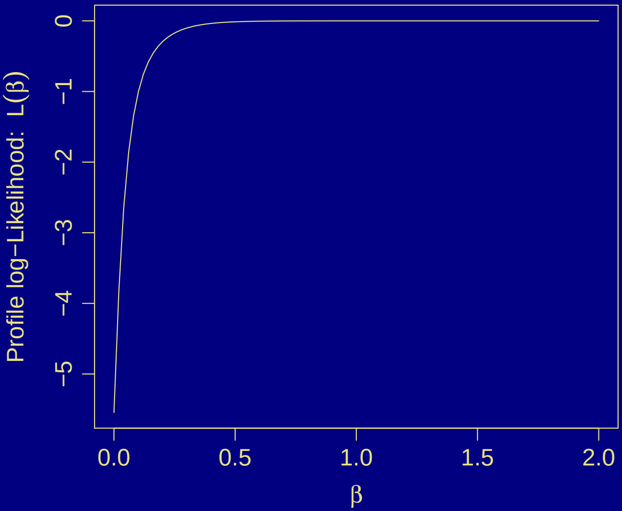
Software may not realize this!

- 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.



1.0
0.8
0.6
0.4
0.2
0.0
y
20
40
60
80
x
β
0
0.1
0.25
0.5
1
2


Profile Likelihood: l(β)
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.5
1.0
1.5
2.0
β


Profile log−Likelihood: L(β)
0
−1
−2
−3
−4
−5
0.0
0.5
1.0
1.5
2.0
β

# 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 $Y$.

In R, we will fit these models using the `VGAM` package.

# 6.1 Logit Models for Nominal Responses

Let $\pi_j = \Pr(Y = j)$, $j = 1, 2, \ldots, J$.

Baseline-category logits are

$$\log\left(\frac{\pi_j}{\pi_J}\right), \qquad j = 1, 2, \ldots, J - 1.$$

Baseline-category logit model has form

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x, \qquad j = 1, 2, \ldots, J - 1.$$

Separate set of parameters $(\alpha_j, \beta_j)$ for each logit.

In R, use `vglm` function w/ `multinomial` family from `VGAM` package.

Note:

- Category used as baseline (i.e., category $J$) is arbitrary and does not affect model fit.
  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 conditional odds of response $j$ given that response is one of $j$ or $J$.
  I.e., on the odds of $j$ vs the baseline $J$.

- Could also use this model with ordinal response variables, but this would ignore information about ordering.

## Example (Income and Job Satisfaction from 1991 GSS)

| Income | Job Satisfaction | | | |
|---|---|---|---|---|
| | 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 |

Using $x =$ income scores (3, 10, 20, 35), we fit the model

$$\log\left(\frac{\pi_j}{\pi_4}\right) = \alpha_j + \beta_j x, \qquad j = 1, 2, 3,$$

for $J = 4$ job satisfaction categories.

```
> data(jobsatisfaction)
> head(jobsatisfaction)
  Gender Income JobSat Freq
1      F      3      1    1
2      F     10      1    2
3      F     20      1    0
4      F     35      1    0
5      M      3      1    1
6      M     10      1    0
> jobsatisfaction <-
    transform(jobsatisfaction, JobSat = factor(JobSat,
                   labels = c("Diss","Little","Mod","Very"),
                   ordered = TRUE))
```

```
> library(reshape2)
> jobsatw <- dcast(jobsatisfaction, Income ~ JobSat, sum,
                   value.var = "Freq")
> jobsatw

  Income Diss Little Mod Very
1      3    2      4  13    3
2     10    2      6  22    4
3     20    0      1  15    8
4     35    0      3  13    8
```

```
> library(VGAM)
> jobsat.fit1 <-
    vglm(cbind(Diss,Little,Mod,Very) ~ Income,
         family=multinomial, data=jobsatw)
> coef(jobsat.fit1)

(Intercept):1 (Intercept):2 (Intercept):3
     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, famil
    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:
              Estimate Std. Error z 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 ($x$ = income score):

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_4}\right) = \boxed{0.430 - 0.185x}$$

$$\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_4}\right) = \boxed{0.456 - 0.054x}$$

$$\log\left(\frac{\hat{\pi}_3}{\hat{\pi}_4}\right) = \boxed{1.704 - 0.037x}$$

Note:

- For each logit, estimated odds of being in less satisfied category (vs very satisfied) decrease as $x$ = income increases.
- Estimated odds of being “very dissatisfied” vs “very satisfied” multiplied by $\boxed{e^{-0.185}} = 0.83$ 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

  $$\boxed{e^{(10)(-0.185)} = e^{-1.85}} = 0.16$$

  e.g., at $x = 20$, the estimated odds of being “very dissatisfied” instead of “very satisfied” are just 0.16 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 baseline-category logit model is

$$\pi_j = \frac{e^{\alpha_j + \beta_j x}}{1 + e^{\alpha_1 + \beta_1 x} + \cdots + e^{\alpha_{J-1} + \beta_{J-1} x}}, \quad j = 1, 2, \ldots, J-1,$$

$$\pi_J = \frac{1}{1 + e^{\alpha_1 + \beta_1 x} + \cdots + e^{\alpha_{J-1} + \beta_{J-1} x}}.$$

Check that

$$\frac{\pi_j}{\pi_J} = e^{\alpha_j + \beta_j x} \quad \Longrightarrow \quad \log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x$$

and

$$\sum_{j=1}^{J} \pi_j = 1.$$

## Example (Job Satisfaction)

$$\hat{\pi}_1 = \frac{e^{0.430-0.185x}}{1+e^{0.430-0.185x}+e^{0.456-0.054x}+e^{1.704-0.037x}}$$

$$\hat{\pi}_2 = \frac{e^{0.456-0.054x}}{1+e^{0.430-0.185x}+e^{0.456-0.054x}+e^{1.704-0.037x}}$$

$$\hat{\pi}_3 = \frac{e^{1.704-0.037x}}{1+e^{0.430-0.185x}+e^{0.456-0.054x}+e^{1.704-0.037x}}$$

$$\hat{\pi}_4 = \frac{1}{1+e^{0.430-0.185x}+e^{0.456-0.054x}+e^{1.704-0.037x}}$$

E.g., at $x = 35$, estimated probability of being “very satisfied” is

$$\hat{\pi}_4 = \frac{1}{1+e^{0.430-0.185(35)}+e^{0.456-0.054(35)}+e^{1.704-0.037(35)}} = 0.367$$

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.,

$$\begin{aligned}\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) &= \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_4}\right) - \log\left(\frac{\hat{\pi}_2}{\hat{\pi}_4}\right)\\ &= (0.430 - 0.185x) - (0.456 - 0.054x)\\ &= -0.026 - 0.131x\end{aligned}$$

- 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". No evidence against the model.

  There are $\boxed{3 \times 4 = 12}$ logits to estimate (3 baseline category logits at each of 4 income levels), so the saturated model has $\boxed{12}$ parameters. The fitted model has $\boxed{6}$ parameters, so df $= \boxed{12 - 6 = 6}$.

- Inference uses usual methods
  - Wald CI for $\beta_j$ is $\hat{\beta}_j \pm z_{\alpha/2}\,\mathrm{SE}$.
  - Wald test of $H_0 : \beta_j = 0$ uses $z = \dfrac{\hat{\beta}_j}{\mathrm{SE}}$ or $z^2 \sim \chi^2_1$.
  - For small $n$, better to use LR test and LR CI, if available.
- However, unlikely to be interested in a single coefficient, because even a single numerical $x$ has $J - 1$ coefficients.

  More common to compare nested models where some variable(s) are included/excluded. LR tests best for this.

## 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.

$$\text{deviance}_0 - \text{deviance}_1 = \boxed{13.47 - 4.66} = 8.81$$

$$\text{df} = \boxed{6 - 3 = 9 - 6 = 3}$$

$$\text{p-value} = 0.032$$

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 $G^2 = 13.47$, df $= 9$, p-value $= 0.1426$). What is different here that made this possible?

```
> 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

> pchisq(deviance(jobsat.fit2) - deviance(jobsat.fit1), 3,
         lower.tail=FALSE)

[1] 0.031937


> summary(jobsat.fit2)
```

```
Call:
vglm(formula = cbind(Diss, Little, Mod, Very) ~ 1, family = m
    data = jobsatw)

Coefficients:
              Estimate Std. Error z value
(Intercept):1   -1.749      0.542   -3.23
(Intercept):2   -0.496      0.339   -1.46
(Intercept):3    1.008      0.244    4.14

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: 13.467 on 9 degrees of freedom

Log-likelihood: -21.359 on 9 degrees of freedom
```

## 6.2 Cumulative Logit Models for Ordinal Responses

The cumulative probabilities are

$$\Pr(Y \leqslant j) = \pi_1 + \cdots + \pi_j, \qquad j = 1, 2, \ldots, J.$$

The cumulative logits are

$$\begin{aligned} \operatorname{logit}\big[\Pr(Y \leqslant j)\big] &= \log\left(\frac{\Pr(Y \leqslant j)}{1 - \Pr(Y \leqslant j)}\right) = \log\left(\frac{\Pr(Y \leqslant j)}{\Pr(Y > j)}\right) \\ &= \log\left(\frac{\pi_1 + \cdots + \pi_j}{\pi_{j+1} + \cdots + \pi_J}\right), \qquad j = 1, \ldots, J-1. \end{aligned}$$

Cumulative logit model has form

$$\operatorname{logit}\big[\Pr(Y \leqslant j)\big] = \alpha_j + \beta x, \qquad j = 1, \ldots, J-1.$$

Note:

- 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 $(Y \leqslant j)$ (instead of $(Y > j)$).

  $$\begin{aligned} \frac{\text{odds}(Y \leqslant j|x_2)}{\text{odds}(Y \leqslant j|x_1)} &= e^{\beta(x_2 - x_1)} \\ &= e^{\beta} \quad \text{when } x_2 = x_1 + 1. \end{aligned}$$

  Also called proportional odds model.
- In R, use `vglm` function w/ `cumulative` family from `VGAM` package.

## Example (Income and Job Satisfaction from 1991 GSS)

| Income | Job Satisfaction | | | |
|---|---|---|---|---|
| | 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 |

Using $x =$ income scores (3, 10, 20, 35), cumulative logit model fit is

$$\text{logit}\left[\widehat{\Pr}(Y \leqslant j)\right] = \hat{\alpha}_j + \hat{\beta}x = \boxed{\hat{\alpha}_j - 0.0449x}, \qquad j = 1, 2, 3.$$

Odds of response at low end of job satisfaction scale decreases as income increases.

Contingency table data. Model fits well: deviance $= \boxed{6.75}$, df $= \boxed{8}$.

```
> 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, famil
    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
```

```
Coefficients:
              Estimate Std. Error z value
(Intercept):1  -2.5829     0.5584   -4.63
(Intercept):2  -0.8970     0.3550   -2.53
(Intercept):3   2.0751     0.4158    4.99
Income         -0.0449     0.0175   -2.56

Number of linear predictors:  3

Names of linear predictors:
logit(P[Y<=1]), logit(P[Y<=2]), logit(P[Y<=3])

Dispersion Parameter for cumulative family:   1

Residual deviance: 6.749 on 8 degrees of freedom
```

Estimated odds of satisfaction below any given level multiplied by

$$e^{\hat{\beta}} = \boxed{e^{-0.0449}} = 0.96$$

for each 1K increase in income (but $x = 3, 10, 20, 35$).

For 10K increase in income, estimated odds multiplied by

$$e^{10\hat{\beta}} = \boxed{e^{(10)(-0.0449)} = e^{-0.449}} = 0.64,$$

e.g., at \$20K income, estimated odds of satsifaction 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:

$$\hat{\beta} = 0.0449, \qquad e^{\hat{\beta}} = 1.046 = 1/0.96.$$

```
> jobsat.cl1r <-
    vglm(cbind(Very,Mod,Little,Diss) ~ Income,
         family=cumulative(parallel=TRUE), data=jobsatw)
> coef(jobsat.cl1r)

(Intercept):1 (Intercept):2 (Intercept):3
    -2.075060      0.896979      2.582873
       Income
     0.044859
```

To test $\text{H}_0 : \beta = 0$ (job satisfaction indep. of income):

$$\text{Wald:} \quad z = \frac{\hat{\beta} - 0}{\text{SE}} = \boxed{\frac{-0.0449}{0.0175} = -2.56} \quad (z^2 = 6.57, \quad \text{df} = 1)$$

p-value $= 0.0105$

$$\text{LR:} \quad \text{deviance}_0 - \text{deviance}_1 = \boxed{13.47 - 6.75 = 6.72} \quad (\text{df} = 1)$$

p-value $= 0.0095$

```
> 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
```

### Remark

Test based on cumlative logit (CL) model treats $Y$ as ordinal, and yielded stronger evidence of association (p-value $\approx$ 0.01) than obtained when we treated:

- $Y$ as nominal (BCL model): $\log\left(\dfrac{\pi_j}{\pi_4}\right) = \alpha_j + \beta_j x$.

  Recall p-value $=$ 0.032 for LR test (df $=$ 3).
- $X, Y$ both as nominal: Pearson's chi-square test of indep. had $X^2 = 11.5$, df $= 9$, p-value $= 0.24$.

  Alternatively, $G^2 = 13.47$, p-value $= 0.14$ ($G^2$ here equivalent to LR test of all $\beta_j = 0$ in BCL model w/ dummies for income).

The BCL and CL models also allow us to control for other variables, mix quantitative and qualitative predictors, interaction terms, etc.

## Political Ideology and Party Affiliation (GSS)

| | | Ideology | | | | |
|---|---|---|---|---|---|---|
| Gender | Party | VLib | SLib | Mod | SCon | VCon |
| Female | Dem | 44 | 47 | 118 | 23 | 32 |
| | Rep | 18 | 28 | 86 | 39 | 48 |
| Male | Dem | 36 | 34 | 53 | 18 | 23 |
| | Rep | 12 | 18 | 62 | 45 | 51 |

$$\begin{aligned} Y &= \text{political ideology } (\text{very liberal, slightly liberal, moderate,} \\ &\qquad\qquad \text{slightly conservative, very conservative}) \\ x_1 &= \text{gender } (1 = \text{M}, 0 = \text{F}) \\ x_2 &= \text{political party } (1 = \text{Rep}, 0 = \text{Dem}) \end{aligned}$$

Cumulative Logit Model:

$$\text{logit}\big[\Pr(Y \leqslant j)\big] = \alpha_j + \beta_1 x_1 + \beta_2 x_2, \qquad j = 1, 2, 3, 4.$$

```
> 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

Coefficients:
              Estimate Std. Error z 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

> deviance(ideo.cl1)
[1] 15.056
> df.residual(ideo.cl1)
[1] 10
> pchisq(deviance(ideo.cl1), df.residual(ideo.cl1),
         lower.tail = FALSE)
[1] 0.13005
```

# Political Ideology and Party Affiliation (ctd)

Cumulative logit model fit:

$$\text{logit}\left[\widehat{\Pr}(Y \leqslant j)\right] = \boxed{\hat{\alpha}_j - 0.117x_1 - 0.964x_2}, \qquad j = 1, 2, 3, 4.$$

- Controlling for gender, estimated odds that a Rep's response is in liberal direction $(Y \leqslant j)$ rather than conservative $(Y > j)$ are $\boxed{e^{-0.964}} = 0.38$ times estimated odds for a Dem.
  - Equivalently: controlling for gender, estimated odds that a Dem's response is in liberal direction $(Y \leqslant j)$ rather than conservative $(Y > j)$ are $\boxed{e^{0.964}} = 2.62$ times estimated odds for a Rep.
  - Statement holds for all $j = 1, 2, 3, 4$.
- 95% CI for true odds ratio is

$$\boxed{e^{-0.964 \pm (1.96)(0.129)}} = (0.30, 0.49)$$

# Political Ideology and Party Affiliation (ctd)

- Contingency table data. No evidence of lack of fit:

  $$\text{deviance} = 15.1, \quad \text{df} = 10, \quad \text{p-value} = 0.13$$

- Test for party effect (controlling for gender), i.e., $H_0 : \boxed{\beta_2 = 0}$

  $$\text{Wald: } z = \boxed{\frac{-0.964}{0.129} = -7.45} \qquad (z^2 = 55.49)$$

  $$\text{LR: deviance}_0 - \text{deviance}_1 = \boxed{71.9 - 15.1 = 56.8}, \quad \text{df} = \boxed{1}$$

  $$\text{p-value} < 0.0001 \quad \text{(either test)}$$

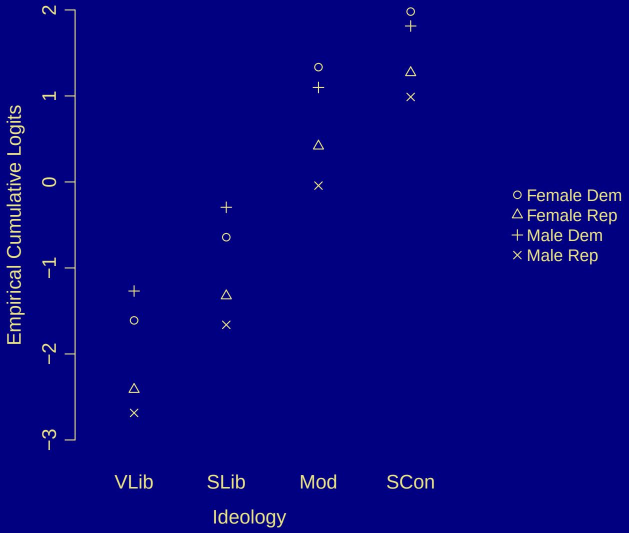
  Strong evidence that Republicans tend to be less liberal (more conservative) than Democrats (for each gender).

- No evidence of gender effect (controlling for party). (p-value $\approx 0.36$ using either Wald or LR test).

```
> 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

> deviance(ideo.cl2) - deviance(ideo.cl1)

[1] 56.847

> pchisq(deviance(ideo.cl2) - deviance(ideo.cl1),
         df.residual(ideo.cl2) - df.residual(ideo.cl1),
         lower.tail=FALSE)

[1] 4.711e-14
```

# Party-Gender interaction?

```
> 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

> ideo.csum <- t(apply(ideow[,-(1:2)], 1, cumsum))
> ideo.csum

  VLib SLib Mod SCon VCon
1   44   91 209  232  264
2   18   46 132  171  219
3   36   70 123  141  164
4   12   30  92  137  188

> ideo.cprop <- ideo.csum[,1:4]/ideo.csum[,5]
> ideo.ecl <- qlogis(ideo.cprop) # empirical cumul. logits
```


Empirical Cumulative Logits
2
1
0
−1
−2
−3
VLib
SLib
Mod
SCon
Ideology
○ Female Dem
△ Female Rep
+ Male Dem
× Male Rep

```
> ideo.cl3 <-
    vglm(cbind(VLib,SLib,Mod,SCon,VCon) ~ Gender*Party,
         family=cumulative(parallel=TRUE), data=ideow)
> coef(summary(ideo.cl3))

                     Estimate Std. Error   z value
(Intercept):1        -1.55209    0.13353 -11.62339
(Intercept):2        -0.55499    0.11703  -4.74225
(Intercept):3         1.16465    0.12337   9.44006
(Intercept):4         2.00121    0.13682  14.62633
GenderMale            0.14308    0.17936   0.79772
PartyRep             -0.75621    0.16691  -4.53062
GenderMale:PartyRep -0.50913    0.25408  -2.00381
```

```
> deviance(ideo.cl3)

[1] 11.063

> df.residual(ideo.cl3)

[1] 9

> deviance(ideo.cl1) - deviance(ideo.cl3)

[1] 3.9922

> pchisq(deviance(ideo.cl1) - deviance(ideo.cl3),
         df.residual(ideo.cl1) - df.residual(ideo.cl3),
         lower.tail=FALSE)

[1] 0.045712
```

# Political Ideology and Party Affiliation (w/ Interaction)

Plot of empirical logits suggest interaction between party and gender. Model with interaction is

$$\text{logit}\big[\Pr(Y \leqslant j)\big] = \alpha_j + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2, \quad j = 1, 2, 3, 4$$

- ML fit:

$$\text{logit}\big[\widehat{\Pr}(Y \leqslant j)\big] = \boxed{\hat{\alpha}_j + 0.143 x_1 - 0.756 x_2 - 0.509 x_1 x_2}$$

- Test for party × gender interaction ($H_0 : \beta_3 = 0$):

$$\text{LR: deviance}_0 - \text{deviance}_1 = \boxed{15.1 - 11.1 = 3.99}$$

$$\text{df} = \boxed{1} \qquad \text{p-value} = \boxed{0.046}$$

Some evidence (significant at 0.05 level) that effect of Party varies with Gender (and vice versa).

## Political Ideology and Party Affiliation (w/ Interaction) (ctd)

- Estimated odds ratio for party effect ($x_2$) is

$$\begin{aligned} e^{-0.756} &= 0.47 \quad \text{when } x_1 = 0 \text{ (F)} \\ e^{-0.756-0.509} = e^{-1.265} &= 0.28 \quad \text{when } x_1 = 1 \text{ (M)} \end{aligned}$$

- Estimated odds ratio for gender effect ($x_1$) is

$$\begin{aligned} e^{0.143} &= 1.15 \quad \text{when } x_2 = 0 \text{ (Dem)} \\ e^{0.143-0.509} = e^{-0.336} &= 0.69 \quad \text{when } x_2 = 1 \text{ (Rep)} \end{aligned}$$

Among Dems, males tend to be more liberal than females.
Among Reps, males tend to be more conservative than females.

## Political Ideology and Party Affiliation (w/ Interaction) (ctd)

- $\widehat{\Pr}(Y = 1)$ (very liberal) for male and female Republicans:

  $$\widehat{\Pr}(Y \leqslant j) = \frac{\exp\left(\hat{\alpha}_j + 0.143x_1 - 0.756x_2 - 0.509x_1x_2\right)}{1 + \exp\left(\hat{\alpha}_j + 0.143x_1 - 0.756x_2 - 0.509x_1x_2\right)}$$

  For $j = 1$, $\hat{\alpha}_1 = -1.55$.

  - Male Republicans ($x_1 = 1$, $x_2 = 1$):

    $$\widehat{\Pr}(Y = 1) = \frac{e^{-1.55+0.143-0.756-0.509}}{1 + e^{-1.55+0.143-0.756-0.509}} = \frac{e^{-2.67}}{1 + e^{-2.67}} = 0.065$$

  - Female Republicans ($x_1 = 0$, $x_2 = 1$):

    $$\widehat{\Pr}(Y = 1) = \frac{e^{-1.55-0.756}}{1 + e^{-1.55-0.756}} = \frac{e^{-2.31}}{1 + e^{-2.31}} = 0.090$$

- Similarly, $\widehat{\Pr}(Y = 2) = \widehat{\Pr}(Y \leqslant 2) - \widehat{\Pr}(Y \leqslant 1)$, etc.

  Note $\widehat{\Pr}(Y = 5) = \widehat{\Pr}(Y \leqslant 5) - \widehat{\Pr}(Y \leqslant 4) = 1 - \widehat{\Pr}(Y \leqslant 4)$.

### Remarks

- Reversing order of response categories changes signs of “slope” estimates (cumulative odds ratio $\mapsto 1/$cumulative odds ratio).
- For ordinal response, only two sensible orderings.

# Ch 8: Models for Matched Pairs

## 8.1 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.

| Treatment | S | F | Total |
|---:|---|---|---|
| Drug | 61 | 25 | 86 |
| Placebo | 22 | 64 | 86 |

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 yielding matched pairs of responses).

# Example (Crossover Study: Drug vs Placebo II)

To reflect dependence, display data as 86 obs rather than 2 $\times$ 86 obs.

| | | Placebo | | |
|---|---|---|---|---|
| | | S | F | |
| Drug | S | 12 | 49 | 61 |
| | F | 10 | 15 | 25 |
| | | 22 | 64 | 86 |

Population probabilities:

| | | Placebo | | |
|---|---|---|---|---|
| | | S | F | |
| Drug | S | $\pi_{11}$ | $\pi_{12}$ | $\pi_{1+}$ |
| | F | $\pi_{21}$ | $\pi_{22}$ | $\pi_{2+}$ |
| | | $\pi_{+1}$ | $\pi_{+2}$ | 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 $n_{21}$:

$$n_{12} \sim \text{Bin}\left(n^*, \frac{1}{2}\right), \quad \text{mean} = \frac{n^*}{2}, \quad \text{std dev} = \sqrt{n^*\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}$$

By normal approx. to binomial, for large $n^*$,

$$z = \frac{n_{12} - n^*/2}{\sqrt{n^*\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \sim N(0, 1)$$

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)

| | | Placebo | | | |
|---|---|---|---|---|---|
| | | S | F | | |
| Drug | S | 12 | 49 | 61 | (71%) |
| | F | 10 | 15 | 25 | |
| | | 22 | 64 | 86 | |
| | | (26%) | | | |

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{49 - 10}{\sqrt{49 + 10}} = 5.1 \qquad (z^2 = 25.8, \text{df} = 1)$$

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, $p_{1+} - p_{+1}$.

$$p_{1+} - p_{+1} = \frac{n_{1+}}{n} - \frac{n_{+1}}{n} = \frac{n_{12} - n_{21}}{n}$$

$$\mathrm{SE} = \frac{1}{n}\sqrt{n_{12} + n_{21} - \frac{(n_{12} - n_{21})^2}{n}}$$

### Example (Crossover Study: Drug vs Placebo IV)

$$\begin{array}{|c|c|c}
\hline
n_{11} & n_{12} & \\
\hline
n_{21} & n_{22} & \\
\hline
\multicolumn{2}{c|}{} & \text{n}
\end{array}
\quad = \quad
\begin{array}{|c|c|c}
\hline
12 & 49 & \\
\hline
10 & 15 & \\
\hline
\multicolumn{2}{c|}{} & 86
\end{array}$$

$$p_{1+} - p_{+1} = \frac{49 - 10}{86} = \frac{39}{86} = 0.453$$

$$\mathrm{SE} = \frac{1}{86}\sqrt{49 + 10 - \frac{(49 - 10)^2}{86}} = 0.075$$

$$95\%\mathrm{CI}: \quad 0.453 \pm (1.96)(0.075) = 0.453 \pm 0.146 = (0.31, 0.60)$$

## Aside: How is the SE derived?

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim \text{MN}\big(n, (\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22})\big)$$

$$\Longrightarrow \begin{cases} \text{Var}(n_{ij}) = n\pi_{ij}(1-\pi_{ij}) \\ \text{Cov}(n_{ij}, n_{i',j'}) = -n\pi_{ij}\pi_{i'j'} \quad \text{if } i \neq i' \text{ or } j \neq j' \end{cases}$$

$$\begin{aligned}
\text{Var}(p_{1+} - p_{+1}) &= \text{Var}\left(\frac{n_{12} - n_{21}}{n}\right) = \frac{\text{Var}(n_{12} - n_{21})}{n^2} \\
&= \frac{\text{Var}(n_{12}) + \text{Var}(n_{21}) - 2\,\text{Cov}(n_{12}, n_{21})}{n^2} \\
&= \frac{n\pi_{12}(1-\pi_{12}) + n\pi_{21}(1-\pi_{21}) + 2n\pi_{12}\pi_{21}}{n^2} \\
&= \frac{\pi_{12} + \pi_{21} - (\pi_{12}^2 - 2\pi_{12}\pi_{21} + \pi_{21}^2)}{n} \\
&= \frac{\pi_{12} + \pi_{21} - (\pi_{12} - \pi_{21})^2}{n} \quad \text{(ctd next frame)}
\end{aligned}$$

$$\mathrm{Var}(p_{1+} - p_{+1}) = \frac{\pi_{12} + \pi_{21} - (\pi_{12} - \pi_{21})^2}{n}$$

$$\widehat{\mathrm{Var}}(p_{1+} - p_{+1}) = \frac{p_{12} + p_{21} - (p_{12} - p_{21})^2}{n}$$

$$= \frac{\frac{n_{12}}{n} + \frac{n_{21}}{n} - \left(\frac{n_{12}}{n} - \frac{n_{21}}{n}\right)^2}{n}$$

$$= \frac{\frac{n_{12}}{n} + \frac{n_{21}}{n} - \frac{(n_{12} - n_{21})^2}{n^2}}{n} \times \frac{n}{n}$$

$$= \frac{n_{12} + n_{21} - \frac{(n_{12} - n_{21})^2}{n}}{n^2}$$

Another way:

$$\text{Var}(p_{1+} - p_{+1}) = \text{Var}(p_{1+}) + \text{Var}(p_{+1}) - 2\,\text{Cov}(p_{1+}, p_{+1})$$

$$\text{Var}(p_{1+}) = \frac{\pi_{1+}(1 - \pi_{1+})}{n}, \qquad \text{Var}(p_{+1}) = \frac{\pi_{+1}(1 - \pi_{+1})}{n},$$

$$\begin{aligned}
\text{Cov}(p_{1+}, p_{+1}) &= \text{Cov}\left(\frac{n_{1+}}{n}, \frac{n_{+1}}{n}\right) = \text{Cov}\left(\frac{n_{11} + n_{12}}{n}, \frac{n_{11} + n_{21}}{n}\right) \\
&= \frac{1}{n^2}\,\text{Cov}(n_{11} + n_{12},\ n_{11} + n_{21}) \\
&= \frac{1}{n^2}\left[\text{Var}(n_{11}) + \text{Cov}(n_{11}, n_{21}) + \text{Cov}(n_{12}, n_{11}) + \text{Cov}(n_{12}, n_{21})\right] \\
&= \frac{1}{n^2}\left[n\pi_{11}(1 - \pi_{11}) - n\pi_{11}\pi_{21} - n\pi_{12}\pi_{11} - n\pi_{12}\pi_{21}\right] \\
&= \frac{1}{n}\left[\pi_{11}(\underbrace{1 - \pi_{11} - \pi_{12} - \pi_{21}}_{\pi_{22}}) - \pi_{12}\pi_{21}\right] \\
&= \frac{\pi_{11}\pi_{22} - \pi_{12}\pi_{21}}{n}
\end{aligned}$$

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,

$$\widehat{\text{Var}}(p_{1+} - p_{+1}) \\ = \frac{1}{n}\left[p_{1+}(1 - p_{1+}) + p_{+1}(1 - p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21})\right]$$

After algebra, this simplifies to expression given before.

```
> 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
```

# 8.5 Rater Agreement

## Example (Movie Reviews by Siskel and Ebert)

| | | Ebert | | |
|---|---|---|---|---|
| Siskel | Con | Mixed | Pro | Total |
| Con | 24 | 8 | 13 | 45 |
| Mixed | 8 | 13 | 11 | 32 |
| Pro | 10 | 9 | 64 | 83 |
| Total | 42 | 30 | 88 | 160 |

How strong is their agreement?

## 8.5.5 Cohen's Kappa

Let $\pi_{ij} = \Pr(S = i, E = j)$.

$$\begin{aligned}\Pr(\text{agree}) &= \pi_{11} + \pi_{22} + \pi_{33} = \sum_i \pi_{ii} \\ &= 1 \text{ if perfect agreement}\end{aligned}$$

If ratings are independent, then $\pi_{ii} = \pi_{i+}\pi_{+i}$ and

$$\Pr(\text{agree}|\text{indep}) = \sum_i \pi_{i+}\pi_{+i}$$

Cohen's kappa is

$$\kappa = \frac{\Pr(\text{agree}) - \Pr(\text{agree}|\text{indep})}{1 - \Pr(\text{agree}|\text{indep})} = \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+}\pi_{+i}}{1 - \sum_i \pi_{i+}\pi_{+i}}$$

Note:

- $\kappa = 0$ if agreement only equals that expected under independence.
- $\kappa = 1$ if perfect agreement.
- Demoninator = maximum difference for numerator, attained if agreement is perfect.

## Example (Siskel and Ebert (ctd))

$$\sum_i \hat{\pi}_{ii} = \frac{24 + 13 + 64}{160} = 0.63$$

$$\sum_i \hat{\pi}_{i+}\hat{\pi}_{+i} = \left(\frac{45}{160}\right)\left(\frac{42}{160}\right) + \left(\frac{32}{160}\right)\left(\frac{30}{160}\right) + \left(\frac{83}{160}\right)\left(\frac{88}{160}\right)$$

$$= 0.40$$

$$\hat{\kappa} = \frac{0.63 - 0.40}{1 - 0.40} = 0.39$$

Moderate agreement: difference between observed agreement and agreement expected under independence is about 40% of the maximum possible difference.

- 95% CI for $\kappa$:

$$\hat{\kappa} \pm 1.96\,\mathrm{SE} = 0.39 \pm (1.96)(0.06) = 0.39 \pm 0.12 = (0.27, 0.51)$$

- For $\mathrm{H}_0 : \kappa = 0$,

$$z = \frac{\hat{\kappa}}{\mathrm{SE}} = \frac{0.39}{0.06} = 6.49$$

  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).

```
> 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] 0.059917
```

# Ch 9: Models for Correlated, Clustered Responses

Correlated responses occur in several ways, including:

- Repeated measures/longitudinal studies: repeated observations on each subject.
- Multiple, matched sets of subjects.
  - Children in the same family.
  - Children in the same elementary school class (children within class, class within school, school within district, . . . ).
  - Fetuses from the same litter of mice.

Usual model forms 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) in order to get appropriate standard errors and valid inferences.

# 9.2 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

$$\begin{array}{cc}
\text{Time} & \begin{array}{cccc} 1 & 2 & \cdots & T \end{array} \\
\begin{array}{c} 1 \\ 2 \\ \vdots \\ T \end{array} & \begin{pmatrix} 1 & \rho & \ldots & \rho \\ \rho & 1 & \ldots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \ldots & 1 \end{pmatrix}
\end{array}$$

- Fitting method gives estimates that are consistent even if correlation structure is misspecified. Adjusts standard errors to reflect actual observed depedendence.

- Available in R package `gee` and others.

# Crossover Study: Drug vs Placebo V

| | | Placebo | | |
|---|---|---|---|---|
| | | S | F | |
| Drug | S | 12 | 49 | 61 |
| | F | 10 | 15 | 25 |
| | | 22 | 64 | 86 |

Model:

$$\text{logit}\big[\Pr(Y_t = 1)\big] = \alpha + \beta t, \qquad t = \begin{cases} 1, & \text{drug} \\ 0, & \text{placebo} \end{cases}$$

GEE fit:

$$\text{logit}\big[\Pr(Y_t = 1)\big] = -1.07 + 1.96t, \qquad \text{SE}(\hat{\beta}) = 0.377 \text{ (“robust”)}$$

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 $\hat{\theta} = (61/25)/(22/64) = 7.1$ ($\log \hat{\theta} = 1.96$).
- With GEE approach, can also have “between-subject” explanatory variables, e.g., gender, order of treatments.

GEE is a “quasi-likelihood” method.

- 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.,

  $$\text{logit}\left[\Pr(Y_{it} = 1)\right] = \alpha_i + \beta t \qquad (\alpha_i = \text{effect for subject } i)$$

  $\{\alpha_i\}$ commonly treated as “random effects” having a normal distribution (Ch 10).

```
> 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
```

```
> 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)

   Subject variable value
1        1     Drug     1
87       1  Placebo     1
2        2     Drug     1
88       2  Placebo     1
3        3     Drug     1
89       3  Placebo     1

> names(crossm)[2:3] <- c("Treat","Resp")
> names(crossm)

[1] "Subject" "Treat"   "Resp"

> crossm <-
    transform(crossm, Treat=relevel(Treat, "Placebo"))
```

Note: the `gee` function has the annoying habit of printing out the starting values used in its iterative algorithm. These values, obtained from `glm`, are not the actual GEE estimates (unless the working correlation structure is `independence`) and should be ignored.

```
> cross.gee1 <-
    gee(Resp ~ Treat, id=Subject, data=crossm,
        family=binomial, corstr="exchangeable")

(Intercept)   TreatDrug
    -1.0678      1.9598

> cross.gee2 <-
    gee(Resp ~ Treat, id=Subject, data=crossm,
        family=binomial, corstr="independence")

(Intercept)   TreatDrug
    -1.0678      1.9598
```

```
> summary(cross.gee1)

 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
 gee S-function, version 4.13 modified 98/01/27 (1998)

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Exchangeable

Call:
gee(formula = Resp ~ Treat, id = Subject, data = crossm, fami
    corstr = "exchangeable")

Summary of Residuals:
     Min        1Q   Median        3Q       Max
-0.70930 -0.25581 -0.25581  0.29070  0.74419
```

```
Coefficients:
            Estimate Naive S.E. Naive z Robust S.E.
(Intercept)  -1.0678    0.24859 -4.2956     0.24714
TreatDrug     1.9598    0.37984  5.1596     0.37723
            Robust z
(Intercept)  -4.3207
TreatDrug     5.1953

Estimated Scale Parameter:  1.0118
Number of Iterations:  1

Working Correlation
         [,1]     [,2]
[1,]  1.00000 -0.21407
[2,] -0.21407  1.00000
```

```
> coef(summary(cross.gee1))

            Estimate Naive S.E. Naive z Robust S.E.
(Intercept)  -1.0678    0.24859 -4.2956     0.24714
TreatDrug     1.9598    0.37984  5.1596     0.37723
            Robust z
(Intercept)  -4.3207
TreatDrug     5.1953

> coef(summary(cross.gee2))

            Estimate Naive S.E. Naive z Robust S.E.
(Intercept)  -1.0678    0.24859 -4.2956     0.24714
TreatDrug     1.9598    0.34475  5.6848     0.37723
            Robust z
(Intercept)  -4.3207
TreatDrug     5.1953
```

## 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?

| | Time | Response Pattern | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | A | A | A | A | N | N | N | N |
| | 1 | A | A | N | N | A | A | N | N |
| | 2 | A | N | A | N | A | N | A | N |
| Severity | Drug | | | | | | | | |
| Mild | Std | 6 | 15 | 4 | 14 | 3 | 9 | 13 | 16 |
| | New | 0 | 9 | 2 | 22 | 0 | 6 | 0 | 31 |
| Severe | Std | 28 | 27 | 15 | 9 | 9 | 8 | 2 | 2 |
| | New | 6 | 32 | 5 | 31 | 2 | 5 | 2 | 7 |

## Example (Depression II)

| | | Sample Proportion Normal | | |
|---|---|---|---|---|
| Severity | Drug | Week 1 | Week 2 | Week 4 |
| Mild | Std | 0.51 | 0.59 | 0.68 |
| | New | 0.53 | 0.79 | 0.97 |
| Severe | Std | 0.21 | 0.28 | 0.46 |
| | New | 0.18 | 0.50 | 0.83 |

E.g., 0.51 $= (3 + 9 + 13 + 16)/(6 + 15 + 4 + 14 + 3 + 9 + 13 + 16)$

Let

$Y_t$ = resp of randomly selected subj at time $t$ (1 = norm, 0 = abnor)

$s$ = severity of initial diagnosis (1 = severe, 0 = mild)

$d$ = drug (1 = new, 0 = std)

$t$ = time (0, 1, 2), which is $\log_2$(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:

$$\log\left\{\frac{\Pr(Y_t = 1)}{\Pr(Y_t = 0)}\right\} = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt$$

$$\text{time effect} = \begin{cases} \beta_3 & \text{if } d = 0 \text{ (std drug)} \\ \beta_3 + \beta_4 & \text{if } d = 1 \text{ (new drug)} \end{cases}$$

GEE estimates: $\hat{\alpha} = -0.028$

$\hat{\beta}_1 = -1.31$ $\hat{\beta}_2 = -0.06$ $\hat{\beta}_3 = 0.48$ $\hat{\beta}_4 = 1.02$

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 $e^{\hat{\beta}_2 + 2\hat{\beta}_4} = 7.2$.
- Estimated time effects are

$$\begin{aligned} \text{std drug}(d = 0): &\quad \hat{\beta}_3 = 0.48 \\ \text{new drug}(d = 1): &\quad \hat{\beta}_3 + \hat{\beta}_4 = 1.50 \end{aligned}$$

- 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.

```
> 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
 gee 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:
                Estimate Naive S.E.  Naive z
(Intercept)    -0.027988    0.16271 -0.17202
severitysevere -1.313911    0.14534 -9.04006
drugnew        -0.059604    0.22058 -0.27021
time            0.482412    0.11392  4.23457
drugnew:time    1.017445    0.18741  5.42889
               Robust S.E. Robust z
(Intercept)        0.17419 -0.16068
severitysevere     0.14598 -9.00034
drugnew            0.22854 -0.26080
time               0.11994  4.02228
drugnew:time       0.18769  5.42077
```

```
Estimated Scale Parameter:  0.98541
Number of Iterations:  1

Working Correlation
     [,1] [,2] [,3]
[1,]    1    0    0
[2,]    0    1    0
[3,]    0    0    1
```

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.

```
> 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

           [,1]       [,2]       [,3]
[1,]  1.0000000 -0.0034327 -0.0034327
[2,] -0.0034327  1.0000000 -0.0034327
[3,] -0.0034327 -0.0034327  1.0000000
```

```
> coef(summary(dep.gee1))[,c(1,2,4)]

                Estimate Naive S.E. Robust S.E.
(Intercept)    -0.027988    0.16271     0.17419
severitysevere -1.313911    0.14534     0.14598
drugnew        -0.059604    0.22058     0.22854
time            0.482412    0.11392     0.11994
drugnew:time    1.017445    0.18741     0.18769

> coef(summary(dep.gee2))[,c(1,2,4)]

                Estimate Naive S.E. Robust S.E.
(Intercept)    -0.028099    0.16255     0.17418
severitysevere -1.313910    0.14486     0.14596
drugnew        -0.059267    0.22053     0.22856
time            0.482464    0.11412     0.11994
drugnew:time    1.017193    0.18771     0.18770
```

```
> dep.gee3 <-
    gee((response == "normal") ~ (severity + drug)*time,
        id=subject, data=depression, family=binomial)

        (Intercept)      severitysevere
           0.073547           -1.528703
            drugnew                time
          -0.055304            0.358728
severitysevere:time        drugnew:time
           0.235006            1.001094

> round(coef(summary(dep.gee3))[,"Robust z"],2)

        (Intercept)      severitysevere
               0.37               -6.55
            drugnew                time
              -0.24                2.31
severitysevere:time        drugnew:time
               1.29                5.33
```

Note:

- GEE have been generalized to multivariate categorical responses, but software is still limited.

  SAS's PROC GENMOD will do GEE for cumulative logit models, but, at last check, only with independence working correlations.

  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).

$$E(Y_t) = \text{Pr}(Y_t = 1)$$

Model: $\text{logit}\big[\text{Pr}(Y_t = 1)\big] = \alpha + \beta x_t,$

$x_t =$ 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 $t$.

$$\text{Model: } \operatorname{logit}\big[\Pr(Y_{it} = 1)\big] = \alpha_i + \beta x_{it}, \quad t = 1, \ldots, T$$

Intercept $\alpha_i$ varies by subject.

$$\begin{array}{rcl} \text{large positive } \alpha_i & \Longrightarrow & \text{large } \Pr(Y_{it} = 1) \text{ each } t \\ \text{large negative } \alpha_i & \Longrightarrow & \text{small } \Pr(Y_{it} = 1) \text{ each } t \end{array}$$

Heterogeneous population $\Longrightarrow$ 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.

$$\text{Model: logit}\big[\Pr(Y_{it} = 1)\big] = u_i + \alpha + \beta x_{it}$$

$\{u_i\}$ 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)

| | Time | Response Pattern | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | 0 | A | A | A | A | N | N | N | N |
| | 1 | A | A | N | N | A | A | N | N |
| | 2 | A | N | A | N | A | N | A | N |
| Severity | Drug | | | | | | | | |
| Mild | Std | 6 | 15 | 4 | 14 | 3 | 9 | 13 | 16 |
| | New | 0 | 9 | 2 | 22 | 0 | 6 | 0 | 31 |
| Severe | Std | 28 | 27 | 15 | 9 | 9 | 8 | 2 | 2 |
| | New | 6 | 32 | 5 | 31 | 2 | 5 | 2 | 7 |

Previously used GEE to fit “marginal model”

$$\text{logit}\left[\Pr(Y_t = 1)\right] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt,$$

$Y_t = 1$ (normal); $s = 0, 1$ (initial diagnosis.: mild vs severe);
$t = \log_2(\text{wks on trt})$; $d = 0, 1$ (drug: std vs new).

Now use ML to fit “random effects model” (a.k.a., “mixed model”)

$$\text{logit}\big[\Pr(Y_{it} = 1)\big] = u_i + \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 dt.$$

Assume $\{u_i\}$ indep. $N(0, \sigma^2)$. Need to estimate $\sigma^2$.

MLEs: $\hat{\sigma} = 0.057$ ($\hat{\sigma}^2 = 0.00323$), $\hat{\alpha} = -0.028$

$$\hat{\beta}_1 = -1.31 \qquad \hat{\beta}_2 = -0.06 \qquad \hat{\beta}_3 = 0.48 \qquad \hat{\beta}_4 = 1.02$$

```
> 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 approximati
Formula: (response == "normal") ~ severity + drug * time + (1
   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:

| | GLMM | | GEE | |
|---|---|---|---|---|
| | Est | SE | Est | SE |
| alpha | −0.03 | 0.16 | −0.03 | 0.17 |
| beta.1 | −1.31 | 0.15 | −1.31 | 0.15 |
| beta.2 | −0.06 | 0.22 | −0.06 | 0.23 |
| beta.3 | 0.48 | 0.11 | 0.48 | 0.12 |
| beta.4 | 1.02 | 0.19 | 1.02 | 0.19 |

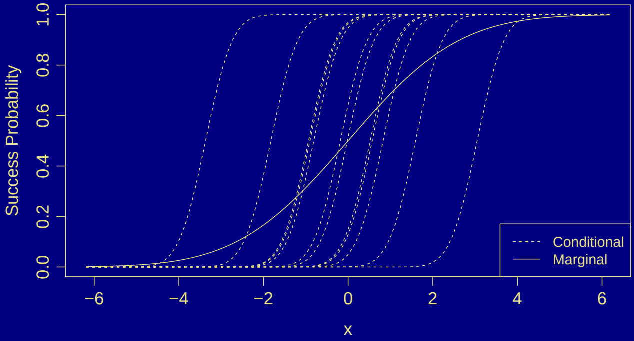
Why? Because there appears to be little correlation between repeated measurements on subjects:

- $\hat{\rho} = -0.003 \approx 0$ 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.)



## Teratology Overdispersion

- Female rats on iron-deficient diets assigned to four groups:

| | |
|---|---|
| Gp 1: | placebo |
| Gp 2: | iron injections on days 7 and 10 |
| Gp 3: | iron injections on days 0 and 7 |
| Gp 4: | iron injections weekly |

- Made pregnant and sacrificed after 3 weeks.
- Response: fetus dead or alive. Data on next frame. (Gp = group, LS = litter size, ND = number dead in litter).
- Cluster $=$ litter.
- $\pi_{it} = \text{Pr}(\text{fetus } t \text{ in litter } i \text{ dead})$.
- Model: $\text{logit}(\pi_{it}) = \alpha + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4}$, where

$$z_{ij} = \begin{cases} 1, & \text{if litter } i \text{ in trt gp } j, \\ 0, & \text{o/w}, \end{cases} \qquad j = 2, 3, 4.$$

# Teratology Overdispersion (ctd)

| Gp | LS | ND |
|---|---|---|
| 1 | 10 | 1 |
| 1 | 11 | 4 |
| 1 | 12 | 9 |
| 1 | 4 | 4 |
| 1 | 10 | 10 |
| 1 | 11 | 9 |
| 1 | 9 | 9 |
| 1 | 11 | 11 |
| 1 | 10 | 10 |
| 1 | 10 | 7 |
| 1 | 12 | 12 |
| 1 | 10 | 9 |
| 1 | 8 | 8 |
| 1 | 11 | 9 |
| 1 | 6 | 4 |
| 1 | 9 | 7 |
| 1 | 14 | 14 |
| 1 | 12 | 7 |
| 1 | 11 | 9 |
| 1 | 13 | 8 |
| 1 | 14 | 5 |
| 1 | 10 | 10 |
| 1 | 12 | 10 |
| 1 | 13 | 8 |
| 1 | 10 | 10 |
| 1 | 14 | 3 |
| 1 | 13 | 13 |
| 1 | 4 | 3 |
| 1 | 8 | 8 |
| 1 | 13 | 5 |
| 1 | 12 | 12 |
| 2 | 10 | 1 |
| 2 | 3 | 1 |
| 2 | 13 | 1 |
| 2 | 12 | 0 |
| 2 | 14 | 4 |
| 2 | 9 | 2 |
| 2 | 13 | 2 |
| 2 | 16 | 1 |
| 2 | 11 | 0 |
| 2 | 4 | 0 |
| 2 | 1 | 0 |
| 2 | 12 | 0 |
| 3 | 8 | 0 |
| 3 | 11 | 1 |
| 3 | 14 | 0 |
| 3 | 14 | 1 |
| 3 | 11 | 0 |
| 4 | 3 | 0 |
| 4 | 13 | 0 |
| 4 | 9 | 2 |
| 4 | 17 | 2 |
| 4 | 15 | 0 |
| 4 | 2 | 0 |
| 4 | 14 | 1 |
| 4 | 8 | 0 |
| 4 | 6 | 0 |
| 4 | 17 | 0 |

```
> data(teratology)
> ## Data also include HB = mother's hemoglobin level
> head(teratology)

   N  R  HB GRP
1 10  1 4.1   1
2 11  4 3.2   1
3 12  9 4.7   1
4  4  4 3.5   1
5 10 10 3.2   1
6 11  9 5.9   1

> terat.binom <-
    glm(cbind(R, N-R) ~ GRP, family = binomial,
        data = teratology)
> chisqstat(terat.binom)

[1] 154.71


> summary(terat.binom)
```

```
Call:
glm(formula = cbind(R, N - R) ~ GRP, family = binomial, data

Deviance Residuals:
   Min       1Q   Median       3Q      Max
-4.430  -0.975  -0.028   1.402   2.783

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)    1.144      0.129    8.85  < 2e-16
GRP2          -3.323      0.331  -10.04  < 2e-16
GRP3          -4.476      0.731   -6.12  9.2e-10
GRP4          -4.130      0.476   -8.67  < 2e-16

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 509.43  on 57  degrees of freedom
Residual deviance: 173.45  on 54  degrees of freedom
AIC: 252.9
```

## Teratology Overdispersion (ctd)

- Binomial model fits poorly ($X^2 = 154.7$, $G^2 = 173.5$, df $= 54$).
- There is inter-litter variability that cannot be accounted for in a binomial model by treatment group alone.
  - Fetuses are more alike within litters than across litters, even within the same treatment group.
- Standard errors invalid (too small).
- Possible solutions:
  - GEE: models marginal (population averaged) effect of trt.
  - GLMM: models litter-specific effect.
  - At least two other approaches not discussed in this class:
    - Quasi-binomial: simplified version of GEE.
    - Beta-binomial: parametric mixture model, analogous to negative-binomial for count data (Ch 3). Motivation similar to GLMM.

```
> ## Adding explicit litter variable for remaining analyses:
> teratology$Litter <- as.factor(1:nrow(teratology))
> ## Need data in ungrouped (binary) format for GEE (???):
> teratbnry <- teratology
> teratbnry$N <- teratbnry$R <- NULL
> teratbnry <-
    teratbnry[rep(1:nrow(teratology), teratology$N),]
> rownames(teratbnry) <- NULL  # cleaning up row names
> teratbnry$Response <-
    with(teratology,
         unlist(apply(cbind(R, N-R), 1,
                function(x) rep(c("Dead","Alive"), x))))
> head(teratbnry, 4)

   HB GRP Litter Response
1 4.1   1      1     Dead
2 4.1   1      1    Alive
3 4.1   1      1    Alive
4 4.1   1      1    Alive
```

```
> library(gee)
> terat.gee <-
    gee((Response == "Dead") ~ GRP, id = Litter,
        data = teratbnry, family = binomial,
        corstr = "exchangeable")

(Intercept)        GRP2        GRP3        GRP4
     1.1440     -3.3225     -4.4762     -4.1297

> coef(summary(terat.gee))[,c("Estimate","Robust S.E.")]

            Estimate Robust S.E.
(Intercept)   1.2115     0.26956
GRP2         -3.3692     0.43042
GRP3         -4.5837     0.62354
GRP4         -4.2474     0.60479

> ## Big working correlation matrix (17 x 17), but
> ## all correlations equal with exchangeable struc:
> terat.gee$working.correlation[1,2]

[1] 0.18534
```

```
> ## glmer can use grouped or ungrouped data.
> library(lme4)
> ## Using grouped data
> terat.glmm <-
    glmer(cbind(R, N-R) ~ GRP + (1|Litter),
          data = teratology, family = binomial)
> ## Using ungrouped binary data
> terat.glmm <-
    glmer((Response == "Dead") ~ GRP + (1|Litter),
          data = teratbnry, family = binomial)
> coef(summary(terat.glmm))

            Estimate Std. Error z value   Pr(>|z|)
(Intercept)   1.8095    0.32858  5.5070 3.6505e-08
GRP2         -4.5398    0.67787 -6.6972 2.1242e-11
GRP3         -5.8838    1.17637 -5.0017 5.6840e-07
GRP4         -5.6068    0.86188 -6.5054 7.7510e-11
```

# Teratology Overdispersion (ctd)

| | Binomial ML | GEE | GLMM |
|---|---|---|---|
| (Intercept) | 1.14 (0.13) | 1.21 (0.27) | 1.81 (0.33) |
| GRP2 | -3.32 (0.33) | -3.37 (0.43) | -4.54 (0.68) |
| GRP3 | -4.48 (0.73) | -4.58 (0.62) | -5.88 (1.18) |
| GRP4 | -4.13 (0.48) | -4.25 (0.6) | -5.61 (0.86) |

- SEs for binomial ML fit invalid (because of lack of fit)
- GEE estimates are similar to binomial but with larger SEs.
  - Estimate marginal (population averaged) effects.
- GLMM estimates are larger in magnitude.
  - Estimate conditional (within litter) effects.

# 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 alchohol? (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

$$\Pr(X = i, Y = j) = \Pr(X = i)\Pr(Y = j) \quad \text{for all } i, j$$

i.e., $\pi_{ij} = \pi_{i+}\pi_{+j}$

For expected cell frequencies,

$$\begin{aligned} \mu_{ij} &= n\pi_{ij} && \text{(general form)} \\ &= n\pi_{i+}\pi_{+j} && \text{(under independence)} \end{aligned}$$

$$\Longrightarrow \log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

$\lambda_i^X$ : effect of classification in row $i$ $(I - 1$ nonredundant parameters)

$\lambda_j^Y$ : effect of classification in col $j$ $(J - 1$ nonredundant parameters)

*Loglinear model of independence*: treats $X$ and $Y$ symmetrically.
Unlike, e.g., logistic regr where $Y$ = response, $X$ = explanatory.

Note: For a Poisson loglinear model,

df $=$ number of Poisson counts $-$ 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: $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y \quad (df = (I-1)(J-1))$

$$\begin{aligned} \text{no. cells} &= IJ \\ \text{no. parameters} &= 1 + (I-1) + (J-1) = I + J - 1 \\ df &= IJ - (I + J - 1) = (I-1)(J-1) \end{aligned}$$

- Saturated model: $\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad$ (df = 0)

| Parameter | Nonredundant |
|---|---|
| $\lambda$ | 1 |
| $\lambda_i^X$ | $I-1$ |
| $\lambda_j^Y$ | $J-1$ |
| $\lambda_{ij}^{XY}$ | $(I-1)(J-1)$ |
| | Total: $IJ$ |

Note: Log-odds-ratio comparing levels $i$ and $i'$ of $X$ and $j$ and $j'$ of $Y$ is

$$\begin{aligned}\log\left(\frac{\mu_{ij}\,\mu_{i'j'}}{\mu_{ij'}\mu_{i'j}}\right) &= \log\mu_{ij} + \log\mu_{i'j'} - \log\mu_{ij'} - \log\mu_{i'j} \\ &= \left(\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}\right) + \left(\lambda + \lambda_{i'}^X + \lambda_{j'}^Y + \lambda_{i'j'}^{XY}\right) \\ &\quad - \left(\lambda + \lambda_i^X + \lambda_{j'}^Y + \lambda_{ij'}^{XY}\right) - \left(\lambda + \lambda_{i'}^X + \lambda_j^Y + \lambda_{i'j}^{XY}\right) \\ &= \lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{ij'}^{XY} - \lambda_{i'j}^{XY}.\end{aligned}$$

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 of the parameters of the loglinear model, is

$$\frac{\mu_{ij}\,\mu_{i'j'}}{\mu_{ij'}\mu_{i'j}} = \exp\left\{\lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{ij'}^{XY} - \lambda_{i'j}^{XY}\right\}.$$

Substituting the MLEs of the saturated model (perfect fit) just reproduces the empirical odds ratio $\frac{n_{ii'}n_{jj'}}{n_{ij'}n_{i'j}}$.

# Income and Job Satisfaction

| Income | Job Satisfaction | | | |
|---|---|---|---|---|
| | 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 |

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

$$\log\left(\frac{\pi_j}{\pi_4}\right) = \alpha_j + \beta_j x, \qquad j = 1, 2, 3.$$

and later, cumulative logit model

$$\text{logit}\big[\Pr(Y \leqslant j)\big] = \alpha_j + \beta x, \qquad j = 1, 2, 3.$$

Using dummy variables, the model

$$\log(\mu_{ij}) = \lambda + \lambda_i^I + \lambda_j^S$$

can be expressed as

$$\log(\mu_{ij}) = \lambda + \lambda_1^I z_1 + \lambda_2^I z_2 + \lambda_3^I z_3 + \lambda_1^S w_1 + \lambda_2^S w_2 + \lambda_3^S w_3$$

where we take $\lambda_4^I = \lambda_4^S = 0$ and

$$z_1 = \begin{cases} 1, & \text{inc} < 5\text{K}, \\ 0, & \text{otherwise}, \end{cases} \qquad w_1 = \begin{cases} 1, & \text{very dissat} \\ 0, & \text{otherwise}, \end{cases}$$

$$z_2 = \begin{cases} 1, & 5\text{K} \leqslant \text{inc} < 15\text{K}, \\ 0, & \text{otherwise}, \end{cases} \qquad w_2 = \begin{cases} 1, & \text{a little sat.} \\ 0, & \text{otherwise}, \end{cases}$$

$$z_3 = \begin{cases} 1, & 15\text{K} \leqslant \text{inc} < 25\text{K}, \\ 0, & \text{otherwise}, \end{cases} \qquad w_3 = \begin{cases} 1, & \text{moderately sat.} \\ 0, & \text{otherwise}, \end{cases}$$

```
> sattab

          Job Satisfaction
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

> jobsat <- as.data.frame(sattab)
> names(jobsat)

[1] "Income"           "Job.Satisfaction"
[3] "Freq"

> names(jobsat)[2] <- "Satis"
> jobsat
     Income    Satis Freq
1       <5K   Dissat    2
2   5K--15K   Dissat    2
3  15K--25K   Dissat    0
4      >25K   Dissat    0
5       <5K   Little    4
6   5K--15K   Little    6
7  15K--25K   Little    1
8      >25K   Little    3
9       <5K Moderate   13
10  5K--15K Moderate   22
11 15K--25K Moderate   15
12     >25K Moderate   13
13      <5K     Very    3
14  5K--15K     Very    4
15 15K--25K     Very    8
16     >25K     Very    8
```

```
> levels(jobsat$Income)

[1] "<5K"      "5K--15K"  "15K--25K" ">25K"

> levels(jobsat$Satis)

[1] "Dissat"   "Little"   "Moderate" "Very"

> options(contrasts=c("contr.SAS","contr.poly"))
> jobsat.indep <-
    glm(Freq ~ Income + Satis, family=poisson,
        data=jobsat)
```

```
> summary(jobsat.indep)

Call:
glm(formula = Freq ~ Income + Satis, family = poisson, data =

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
SatisDissat    -1.75e+00   5.42e-01   -3.23   0.0012
SatisLittle    -4.96e-01   3.39e-01   -1.46   0.1431
SatisModerate   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 Pr(>Chi)
1         9       13.5
2         0        0.0  9     13.5     0.14

> ## Set contrasts back to R defaults
> options(contrasts=c("contr.treatment","contr.poly"))
```

# 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 model

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

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:

$$\log(\theta_{X(j)Z}) = \underbrace{\lambda_{11}^{XZ} + \lambda_{22}^{XZ} - \lambda_{12}^{XZ} - \lambda_{21}^{XZ}}_{\text{does not depend on } j}$$

Similarly, Y-Z odds ratio same at all levels of X. Model has no three-factor interaction.

Ex. Consider the loglinear model

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}.$$

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.

```
> 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
```

```
> 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

> teens.df <-
    transform(teens.df,
              cigs = relevel(cigs, "no"),
              alc = relevel(alc, "no"),
              mj = relevel(mj, "no"))
```

```
> 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, famil
    data = teens.df)

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
cigsyes         -1.8867     0.1627  -11.60  < 2e-16
mjyes           -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
cigsyes: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
```

The (AC,AM,CM) model fits well: $G^2 = 0.37$ (and $X^2 = 0.4$) on 1 df.

```
> df.residual(teens.AC.AM.CM)
[1] 1
> deviance(teens.AC.AM.CM)
[1] 0.37399
> chisqstat(teens.AC.AM.CM)
[1] 0.4011
```

Note: As a LRT, goodness-of-fit on previous slide is comparing to saturated model.

```
> teens.ACM <- update(teens.AC.AM.CM, . ~ alc*cigs*mj)
> 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 Df Deviance Pr(>Chi)
1         1      0.374
2         0      0.000  1    0.374     0.54
```

And none of the interaction terms can be dropped:

```
> drop1(teens.AC.AM.CM, test="Chisq")

Single term deletions

Model:
Freq ~ alc * cigs + alc * mj + cigs * mj
         Df Deviance AIC LRT Pr(>Chi)
<none>                 0  63
alc:cigs  1      188 249 187   <2e-16
alc:mj    1       92 153  92   <2e-16
cigs:mj   1      497 558 497   <2e-16
```

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:

```
> 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


Model 1: Freq ~ alc + mj + cigs + alc:mj + mj:cigs
Model 2: Freq ~ alc * cigs + alc * mj + cigs * mj
  Resid. Df Resid. Dev Df Deviance Pr(>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.

```
> 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.CM, 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"))
```

```
> table.7.4 <- signif(table.7.4, 3)
> table.7.4 <-
   cbind(teens.df[,c("alc","cigs","mj")],
         table.7.4)
> names(table.7.4) <-
    c("alc","cigs","mj",
      "(A,C,M)","(AC,M)","(AM,CM)","(AC,AM,CM)","(ACM)")
```

```
> table.7.4
  alc cigs  mj (A,C,M) (AC,M) (AM,CM) (AC,AM,CM) (ACM)
1 yes  yes yes   540.0  611.0  909.00     910.00   911
2 yes  yes  no   740.0  838.0  439.00     539.00   538
3 yes   no yes   282.0  211.0   45.80      44.60    44
4 yes   no  no   387.0  289.0  555.00     455.00   456
5  no  yes yes    90.6   19.4    4.76       3.62     3
6  no  yes  no   124.0   26.6  142.00      42.40    43
7  no   no yes    47.3  119.0    0.24       1.38     2
8  no   no  no    64.9  162.0  180.00     280.00   279
```

In (AC,AM,CM) model, AC odds-ratio is the same at each level of M. With 1 $=$ yes and 2 $=$ no for each variable, the estimated conditional AC odds ratio is

$$\frac{\hat{\mu}_{11k}\hat{\mu}_{22k}}{\hat{\mu}_{12k}\hat{\mu}_{21k}} = \text{exp}\left(\hat{\lambda}^{AC}_{11} + \hat{\lambda}^{AC}_{22} - \hat{\lambda}^{AC}_{12} - \hat{\lambda}^{AC}_{21}\right) = e^{2.0545} = 7.8$$

A 95% CI is

$$e^{2.05 \pm (1.96)(0.174)} = \left(e^{1.71}, e^{2.40}\right) = (5.5, 11.0)$$

The commons odds-ratio is reflected in the fitted values for the model:

$$\frac{(910)(1.38)}{(44.6)(3.62)} = 7.8 \qquad \frac{(539)(280)}{(455)(42.4)} = 7.8$$

Similar results hold for AM and CM conditional odds-ratios in this model.

In (AM,CM) model, $\lambda_{ij}^{AC} = 0$, 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.

$$\frac{(909)(0.24)}{(45.8)(4.76)} = 1 \qquad \frac{(439)(180)}{(555)(142)} = 1$$

The AM odds-ratio is not 1, but it is the same at each level of C:

$$\frac{(909)(142)}{(439)(4.76)} = 61.87 \qquad \frac{(45.8)(180)}{(555)(0.24)} = 61.87$$

Similarly, the CM odds-ratio is the same at each level of A:

$$\frac{(909)(555)}{(439)(45.8)} = 25.14 \qquad \frac{(4.76)(180)}{(142)(0.24)} = 25.14$$

Standardized residuals may help understand lack of fit.
Text uses standardized Pearson residuals.
`rstandard()` computes standardized deviance resids. by default
but has `type = "pearson"` option.

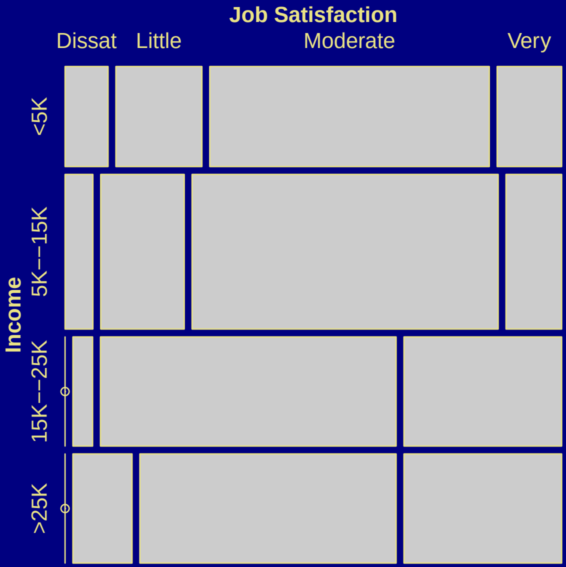
See Section 7.2.2 for example and discussion.

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.

# Mosaic Plots: Two-Way Tables



The previous plot was produced by the commands

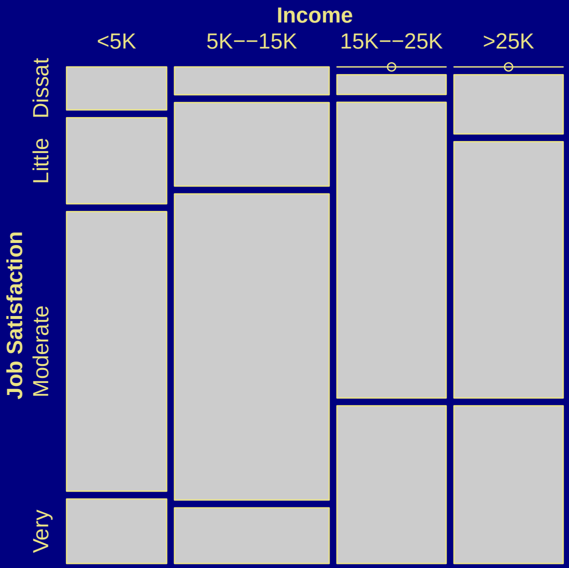
```
> library(vcd)
> mosaic(sattab)
```

The same plot could have been produced with

```
> mosaic(~ Income + Satis, data = jobsat)
```

You might prefer to view the plot with a different orientation:

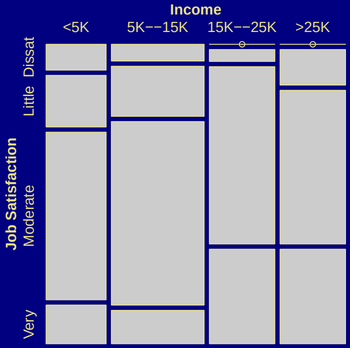
```
> mosaic(sattab, split_vertical = TRUE)
```


Income
<5K
5K−−15K
15K−−25K
>25K
Job Satisfaction
Dissat
Little
Moderate
Very

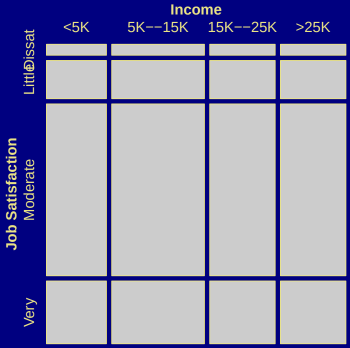
Recall:

```
> (sat.chisq <- chisq.test(sattab))

        Pearson's Chi-squared test


data:  sattab
X-squared = 11.524, df = 9, p-value = 0.2415

> round(sat.chisq$expected, 1)

          Job Satisfaction
Income     Dissat Little Moderate Very
  <5K         0.8    3.0     13.3  4.9
  5K--15K     1.3    4.6     20.6  7.5
  15K--25K    0.9    3.2     14.5  5.3
  >25K        0.9    3.2     14.5  5.3

> mosaic(sattab, split_vertical = TRUE, main = "Observed")
> mosaic(sattab, split_vertical = TRUE, type = "expected",
         main = "Expected")
```

## Observed



## Expected



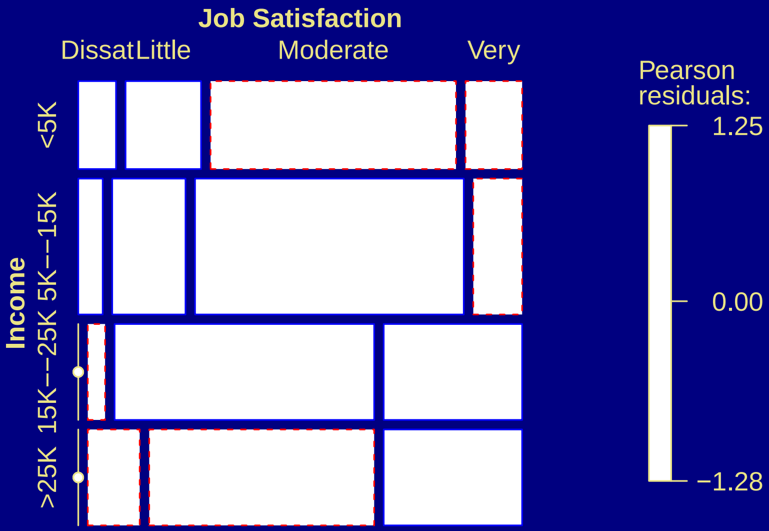
```
> round(sat.chisq$stdres, 1)

           Job Satisfaction
Income      Dissat Little Moderate Very
  <5K          1.4    0.7     -0.2 -1.1
  5K--15K      0.8    0.9      0.6 -1.8
  15K--25K    -1.1   -1.5      0.2  1.5
  >25K        -1.1   -0.2     -0.7  1.5
```

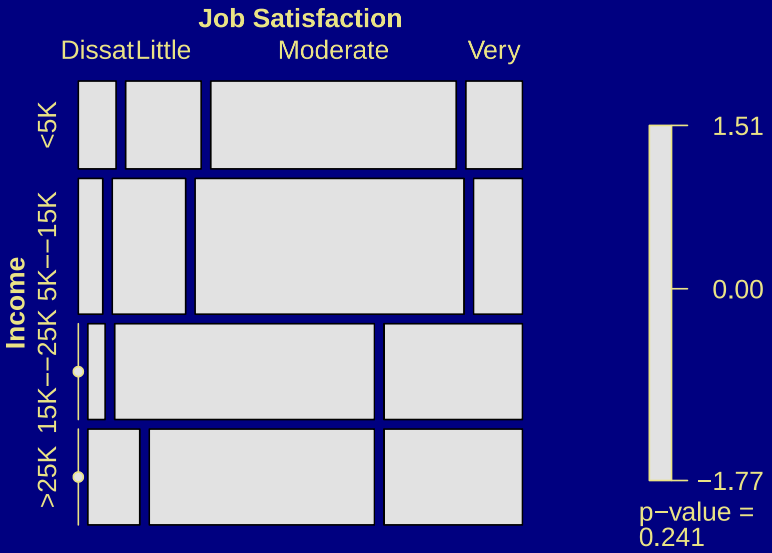
Same as the standardized (i.e., ”adjusted”) Pearson residuals from fitting loglinear model of independence:

```
> round(rstandard(jobsat.indep, type = "pearson"), 1)

   1    2    3    4    5    6    7    8    9   10   11
 1.4  0.8 -1.1 -1.1  0.7  0.9 -1.5 -0.2 -0.2  0.6  0.2
  12   13   14   15   16
-0.7 -1.1 -1.8  1.5  1.5
```

This example isn't the best here because Pearson's chi-square test does not provide any evidence against independence.

```
> mosaic(sattab, gp = shading_Friendly)

> mosaic(sattab, residuals = sat.chisq$stdres,
         gp = shading_hcl,
         gp_args = list(p.value = sat.chisq$p.value,
                        interpolate = c(2,4)))
```


Job Satisfaction
Dissat
Little
Moderate
Very
Income
<5K
5K−−15K
15K−−25K
>25K
Pearson
residuals:
1.25
0.00
−1.28


Job Satisfaction
Dissat
Little
Moderate
Very
Income
<5K
5K−−15K
15K−−25K
>25K
1.51
0.00
−1.77
p−value =
0.241

## Hair and Eye Color

Data from `vcd` package.

```
> ftable(Eye ~ Sex + Hair, data = HairEyeColor)

            Eye Brown Blue Hazel Green
Sex    Hair
Male   Black       32   11    10     3
       Brown       53   50    25    15
       Red         10   10     7     7
       Blond        3   30     5     8
Female Black       36    9     5     2
       Brown       66   34    29    14
       Red         16    7     7     7
       Blond        4   64     5     8
```

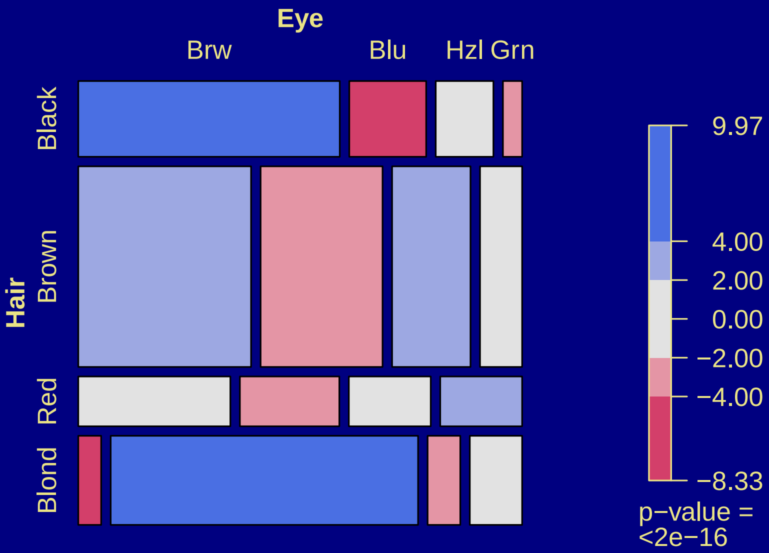
# Hair and Eye Color (ctd)

Collapsing across Sex.

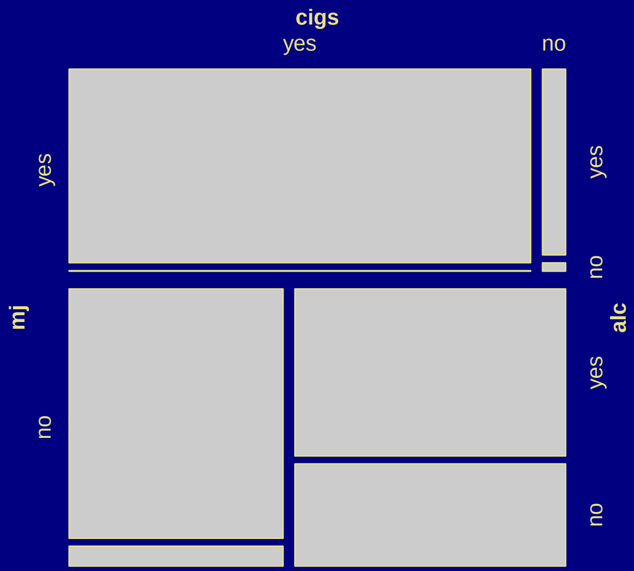
```
> haireye <- margin.table(HairEyeColor, 1:2)
> haireye

       Eye
Hair    Brown Blue Hazel Green
  Black    68   20    15     5
  Brown   119   84    54    29
  Red      26   17    14    14
  Blond     7   94    10    16

> (he.chisq <- chisq.test(haireye))

        Pearson's Chi-squared test

data:  haireye
X-squared = 138.29, df = 9, p-value < 2.2e-16
```

## Hair and Eye Color (ctd)

```
> mosaic(haireye, residuals = he.chisq$stdres,
      gp = shading_hcl,
      gp_args = list(p.value = he.chisq$p.value,
                     interpolate = c(2,4)),
      labeling_args = list(abbreviate_labs = c(Eye = 3)))
```


Eye
Brw
Blu
Hzl
Grn
Hair
Black
Brown
Red
Blond
9.97
4.00
2.00
0.00
−2.00
−4.00
−8.33
p−value =
<2e−16

# Teen Survey Data



The previous plot was produced by the commands

```
> mosaic(teens)
```

Compare to

```
> ftable(round(prop.table(teens), 3))

         alc   yes    no
mj  cigs
yes yes        0.400 0.001
    no         0.019 0.001
no  yes        0.236 0.019
    no         0.200 0.123
```

The same plot could have been produced by either of the commands:

```
> mosaic(~ mj + cigs + alc, data = teens)
> mosaic(~ mj + cigs + alc, data = teens.df)
```

Changing the order of the terms in the formula has the expected effect.

## Standardized residuals from two loglinear models.

```
> table.7.8 <- teens.df[,c("alc","cigs","mj","Freq")]
> table.7.8 <- cbind(table.7.8,
    round(predict(teens.AM.CM, type = "response"),1))
> table.7.8 <- cbind(table.7.8,
    round(rstandard(teens.AM.CM, type = "pearson"),2))
> table.7.8 <- cbind(table.7.8,
    round(predict(teens.AC.AM.CM, type = "response"),1))
> table.7.8 <- cbind(table.7.8,
    round(rstandard(teens.AC.AM.CM, type = "pearson"),2))
> names(table.7.8) <-
    c("A","C","M","Obs","(AM,CM)","StdRes",
      "(AC,AM,CM)","StdRes")
```

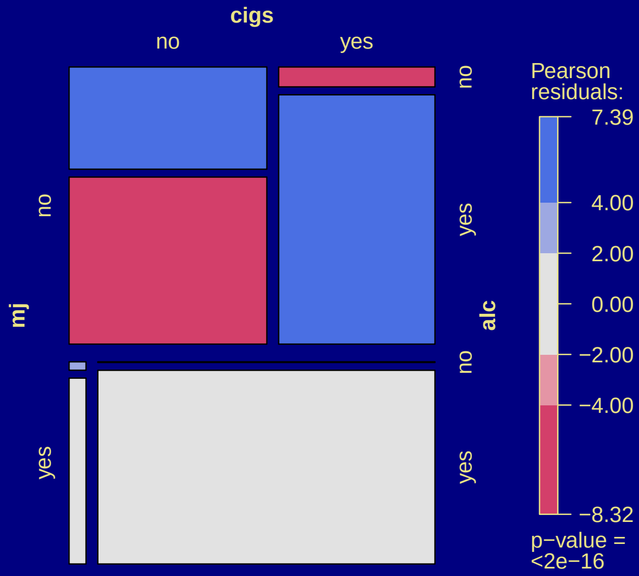
```
> table.7.8

    A   C   M Obs (AM,CM) StdRes (AC,AM,CM) StdRes
1 yes yes yes 911   909.2    3.7      910.4   0.63
2 yes yes  no 538   438.8   12.8      538.6  -0.63
3 yes  no yes  44    45.8   -3.7       44.6  -0.63
4 yes  no  no 456   555.2  -12.8      455.4   0.63
5  no yes yes   3     4.8   -3.7        3.6  -0.63
6  no yes  no  43   142.2  -12.8       42.4   0.63
7  no  no yes   2     0.2    3.7        1.4   0.63
8  no  no  no 279   179.8   12.8      279.6  -0.63
```

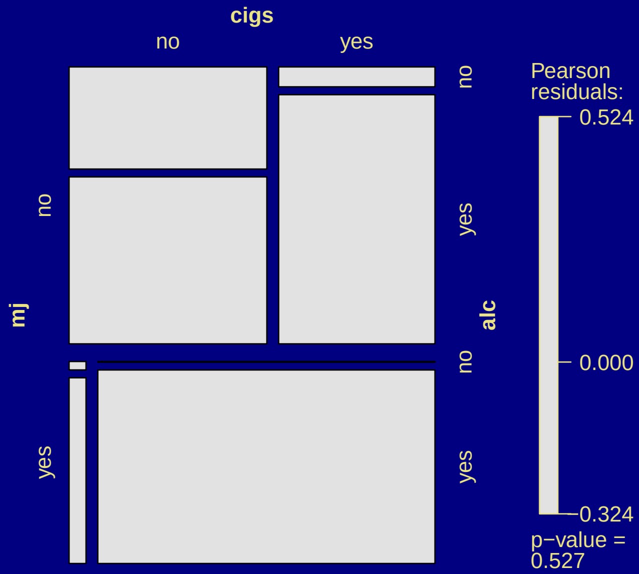
- Number nonredundant *standardized* residuals = residual df.
  - Model (AM,CM): Residual df = 2
  - Model (AC,AM,CM): Residual df = 1

`vcdExtra` package works from fitted loglinear model.
Uses *unadjusted* Pearson residuals, or optionally, standardized *deviance* residuals.

Here is the default, using unadjusted Pearson residuals:

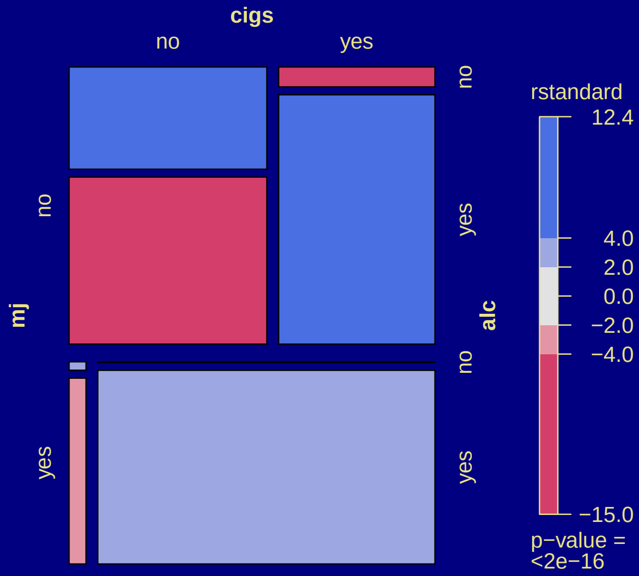
```
> library(vcdExtra)
> mosaic(teens.AM.CM, ~ mj + cigs + alc)

> mosaic(teens.AC.AM.CM, ~ mj + cigs + alc)
```

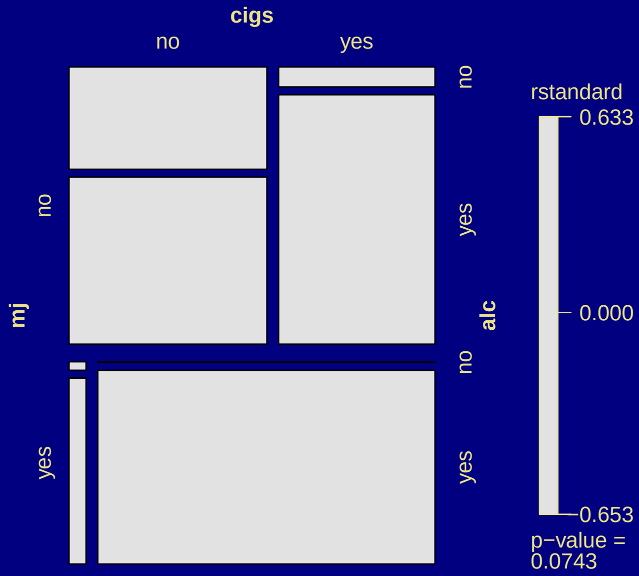

cigs
no
yes
mj
no
yes
alc
no
yes
no
yes
Pearson residuals:
7.39
4.00
2.00
0.00
−2.00
−4.00
−8.32
p−value = <2e−16


cigs
no
yes
mj
no
yes
alc
no
yes
no
yes
Pearson residuals:
0.524
0.000
−0.324
p−value = 0.527

With the standardized deviance residuals:

```
> mosaic(teens.AM.CM,  ~ mj + cigs + alc,
         residuals_type = "rstandard")

> mosaic(teens.AC.AM.CM,  ~ mj + cigs + alc,
         residuals_type = "rstandard")
```

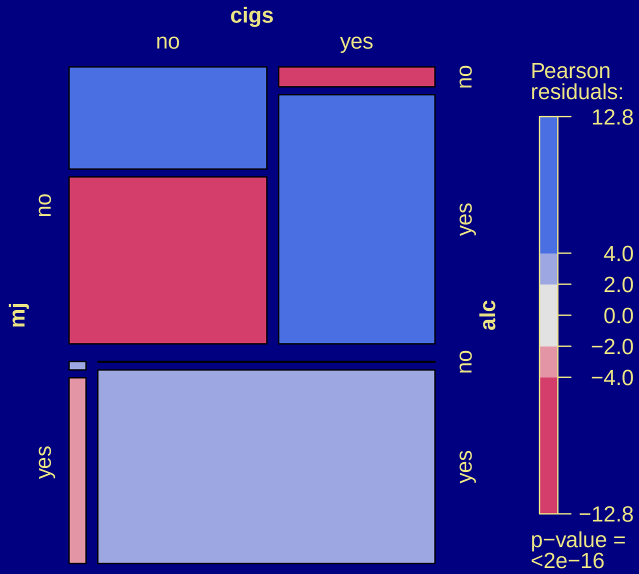

cigs
no
yes
mj
no
yes
alc
no
yes
no
yes
rstandard
12.4
4.0
2.0
0.0
−2.0
−4.0
−15.0
p−value =
<2e−16

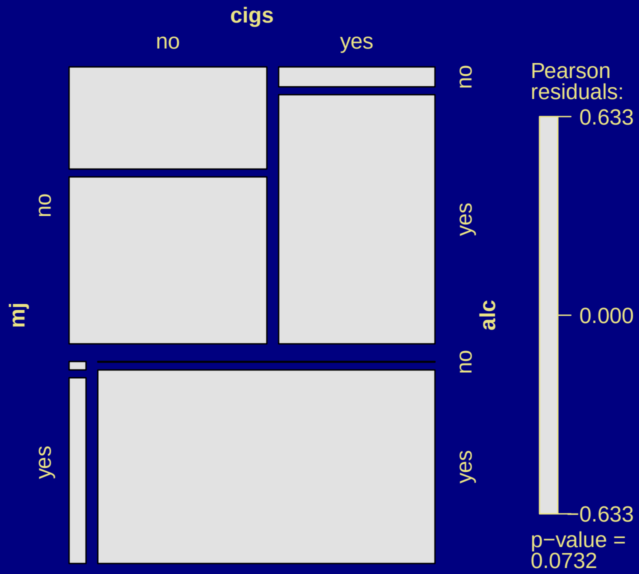

cigs
no
yes
mj
no
yes
alc
no
yes
no
yes
rstandard
0.633
0.000
−0.653
p−value =
0.0743

And finally, with the standardized Pearson residuals (note that the title on the legend is not correct):

```
> mosaic(teens.AM.CM, ~ mj + cigs + alc,
    residuals = rstandard(teens.AM.CM, type = "pearson"))

> mosaic(teens.AC.AM.CM, ~ mj + cigs + alc,
    residuals = rstandard(teens.AC.AM.CM, type = "pearson"))
```

I have suggested a patch to make the selection of Pearson vs deviance and non-standardized vs standardized residuals more straightforward.


cigs
no
yes
mj
no
yes
alc
no
yes
no
yes
Pearson
residuals:
12.8
4.0
2.0
0.0
−2.0
−4.0
−12.8
p−value =
<2e−16


cigs
no
yes
mj
no
yes
alc
no
yes
no
yes
Pearson
residuals:
0.633
0.000
−0.633
p−value =
0.0732

# 7.3 The Loglinear-Logit Connection

The loglinear model $(XY, XZ, YZ)$, i.e.,

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ},$$

- treats variables symmetrically
- permits association for each pair of vars.
- allows no three-factor association (i.e., implies homogeneous association)

Suppose $Y$ is binary and let

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

Treat $Y$ as response. If model $(XY, XZ, YZ)$ holds, then

$$\begin{aligned}
\text{logit}(\pi_{ik}) &= \log\left(\frac{\pi_{ik}}{1-\pi_{ik}}\right) = \log\left(\frac{P(Y=1|X=i, Z=k)}{P(Y=2|X=i, Z=k)}\right) \\
&= \log(\mu_{i1k}) - \log(\mu_{i2k}) \\
&= (\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{i1}^{XY} + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ}) \\
&\qquad - (\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{i2}^{XY} + \lambda_{ik}^{XZ} + \lambda_{2k}^{YZ}) \\
&= \underbrace{(\lambda_1^Y - \lambda_2^Y)}_{\alpha} + \underbrace{(\lambda_{i1}^{XY} - \lambda_{i2}^{XY})}_{\beta_i^X} + \underbrace{(\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ})}_{\beta_k^Z} \\
&= \alpha + \beta_i^X + \beta_k^Z
\end{aligned}$$

i.e., logit model for $Y$ has additive main effects and no interaction.

## UCB Admissions

Recall the UCB admissions data.

```
      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
```

Let $A =$ admission (yes/no) be response var. Logit model:

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

The corresponding loglinear model is $(AG, AD, DG)$:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^A + \lambda_j^G + \lambda_k^D + \lambda_{ij}^{AG} + \lambda_{ik}^{AD} + \lambda_{jk}^{DG}$$

## UCB Admissions (ctd)

Both models have deviance $G^2 = 20.20$ (df $= 5$):

```
> UCB.logit <-
    glm(cbind(Admitted, Rejected) ~ Gender + Dept,
              family = binomial, data = UCBw)
> c(deviance(UCB.logit), df.residual(UCB.logit))

[1] 20.204  5.000

> UCB.loglin <-
    glm(Freq ~ Admit*Gender + Admit*Dept + Gender*Dept,
        family = poisson, data = UCBdf)
> c(deviance(UCB.loglin), df.residual(UCB.loglin))

[1] 20.204  5.000
```

## UCB Admissions (ctd)

The df for testing fit are the same for each model:

Logit model Treats table as $\boxed{12}$ indep. binomial variates on response $A$ at 12 combinations of levels of $\mathrm{D}$ and $\mathrm{G}$:

$$\begin{aligned} \text{no. obs.} &= \boxed{12} \\ \text{no. param.} &= \boxed{1+1+5=7} \\ \text{(residual) df} &= \boxed{12-7=5} \end{aligned}$$

Loglinear model Treats table as 24 indep. Poisson variates:

$$\begin{aligned} \text{no. obs.} &= \boxed{24} \\ \text{no. param.} &= \boxed{1+1+1+5+1+5+5=19} \\ \text{(residual) df} &= \boxed{24-19=5} \end{aligned}$$

## UCB Admissions (ctd)

Controlling for $\mathrm{D}$ (department), estimated odds ratio for effected of $\mathrm{G}$ on $A$ (odds of admission for males divided by odds for females), is

$$\exp(\hat{\beta}_1^{\mathrm{G}} - \hat{\beta}_2^{\mathrm{G}}) = \boxed{e^{0-0.0999}} = .905$$

Identical to

$$\exp\left(\hat{\lambda}_{11}^{AG} + \hat{\lambda}_{22}^{AG} - \hat{\lambda}_{12}^{AG} - \hat{\lambda}_{21}^{AG}\right) = \boxed{\exp\left(0 - 0.0999 - 0 - 0\right)}$$

```
> coef(UCB.logit)

 (Intercept) GenderFemale        DeptB        DeptC
    0.582051     0.099870    -0.043398    -1.262598
       DeptD        DeptE        DeptF
   -1.294606    -1.739306    -3.306480
```

```
> coef(UCB.loglin)
               (Intercept)              AdmitRejected
                  6.271499                  -0.582051
              GenderFemale                      DeptB
                 -1.998588                  -0.403220
                     DeptC                      DeptD
                 -1.577903                  -1.350005
                     DeptE                      DeptF
                 -2.449820                  -3.137871
AdmitRejected:GenderFemale        AdmitRejected:DeptB
                 -0.099870                   0.043398
       AdmitRejected:DeptC        AdmitRejected:DeptD
                  1.262598                   1.294606
       AdmitRejected:DeptE        AdmitRejected:DeptF
                  1.739306                   3.306480
        GenderFemale:DeptB         GenderFemale:DeptC
                 -1.074820                   2.665133
        GenderFemale:DeptD         GenderFemale:DeptE
                  1.958324                   2.795186
        GenderFemale:DeptF
                  2.002319
```

For a given logit model, equivalent loglinear model (same goodness of fit, df, fitted values, etc) has:

- interactions of $Y$ with explanatory variables implied by logit model;
- and the fullest interaction term among explanatory variables

Example

$\pi = P(Y = 1)$, predictors $A, B, C$ (4-way table).

Logit model

$$\text{logit}(\pi) = \alpha + \beta_i^A + \beta_j^B + \beta_k^C$$

corresponds to loglinear model $\boxed{(AY, BY, CY, ABC)}$.

Logit model

$$\text{logit}(\pi) = \alpha + \beta_i^A + \beta_j^B + \beta_k^C + \beta_{jk}^{BC}$$

corresponds to loglinear model $\boxed{(AY, BCY, ABC)}$.

### Remarks

- When there is a single binary response, it is simpler to approach data directly using logit models.
- Similar remarks hold for a multi-category response $Y$:
  - Baseline-category logit model has a matching loglinear model.
  - With a single response, it is simpler to use the baseline-category logit model.
- Loglinear models have advantage of generality — can handle multiple responses, some of which may have more than two outcome categories.

# 7.4 Independence Graphs and Collapsibility

Independence graph: a graphical representation for conditional independence.

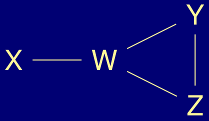
- Vertices (or nodes) represent variables.
- Connected by edges: a missing edge between two variables represents a conditional independence between the variables.
- Different models may produce the same graph.
- *Graphical models*: subclass of loglinear models
  - Within this class there is a unique model for each independence graph.
  - For any group of variables having no missing edges, graphical model contains the highest order interaction term for those variables.

# Independence Graphs for a 4-Way Table (Variables $W$, $X$, $Y$, $Z$)

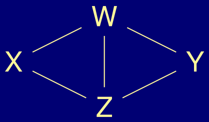
Model(s)

Graph

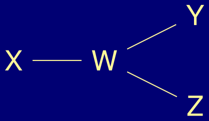
$(WX, WY, WZ, YZ)$
$(WX, WYZ)^*$



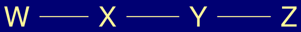
$(WX, WY, WZ, XZ, YZ)$
$(WX, XZ, WYZ)$
$(WXZ, WY, YZ)$
$(WXZ, WYZ)^*$



$(WX, WY, WZ)^*$



$(WX, XY, YZ)^*$



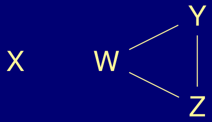
* Graphical models.

# Independence Graphs for a 4-Way Table (ctd)

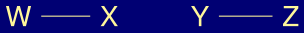
Model(s)

Graph

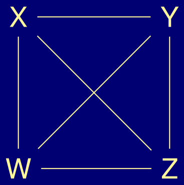
$(X, WY, WZ, YZ)$
$(X, WYZ)^*$



$(WX, YZ)^*$



$(WX, WY, WZ, XY, XZ, YZ)$
$(WX, WY, WZ, XYZ)$
$(WX, WYZ, XYZ)$
. . . many others . . .
$(WXYZ)^*$

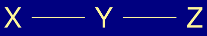


* Graphical models.

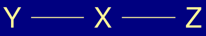
# Collapsibility Conditions for Three-Way Tables

*For a three-way table, the $XY$ marginal and conditional odds ratios are identical if either $Z$ and $X$ are conditonally independent or if $Z$ and $Y$ are conditionally independent.*

- Conditions say control variable $Z$ is either:
  - conditionally independent of $X$ given $Y$, as in model $(XY, YZ)$;
  - or conditionally independent of $Y$ given $X$, as in $(XY, XZ)$.
- I.e., $XY$ association is identical in the partial tables and the marginal table for models with independence graphs



X —— Y —— Z



Y —— X —— Z

or even simpler models.

## Teen Survey

$A =$ alcohol use, $C =$ cigarette use, $M =$ marijuana use.

The model of $AC$ conditional independence, $(AM, CM)$, has independence graph

$$A \text{ --- } M \text{ --- } C$$

Consider $AM$ association, treating $C$ as control variable.
Since $C$ is conditionally independent of $A$, the $AM$ conditional odds ratios are the same as the $AM$ marginal odds ratio collapsed over $C$.

$$\frac{(909.24)(142.16)}{(438.84)(4.76)} = \frac{(45.76)(179.84)}{(555.16)(0.24)} = \frac{(955)(322)}{(994)(5)} = 61.9$$

See Tables 7.4 and 7.5, or next slide.

```
> exp(coef(teens.AM.CM)[5])

alcyes:mjyes
      61.873
```

```
> AM.CM.fitted <- teens
> AM.CM.fitted[,,] <- predict(teens.AM.CM, type="response")
> AM.CM.fitted[,"yes",]

     alc
mj        yes       no
  yes 909.24   4.7604
  no  438.84 142.1596

> AM.CM.fitted[,"no",]

     alc
mj        yes        no
  yes  45.76   0.23958
  no  555.16 179.84043

> AM.CM.fitted[,"yes",] + AM.CM.fitted[,"no",]

     alc
mj    yes  no
  yes 955   5
  no  994 322
```

## Teen Survey

- Similarly, $CM$ association is collapsible over $A$.
- The $AC$ association is not collapsible, because $M$ is conditionally dependent with both $A$ and $C$ in model $(AM, CM)$.

  Thus, $A$ and $C$ may be marginally dependent, even though conditionally independent.

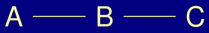
$$\frac{(909.24)(0.24)}{(45.76)(4.76)} = \frac{(438.84)(179.84)}{(555.16)(142.16)} = \boxed{1}$$

$$\frac{(1348.08)(180.08)}{(600.92)(146.92)} = 2.75 \neq 1$$

```
> AM.CM.fitted["yes",,]

     alc
cigs      yes      no
  yes 909.24 4.76042
  no    45.76 0.23958

> AM.CM.fitted["no",,]

     alc
cigs      yes     no
  yes 438.84 142.16
  no  555.16 179.84

> AM.CM.fitted["yes",,] + AM.CM.fitted["no",,]

     alc
cigs       yes     no
  yes 1348.08 146.92
  no    600.92 180.08
```

# Collapsibility Conditions for Multiway Tables

*If the variables in a model for a multiway table partition into three mutually exclusive subsets, A, B, C, such that B separates A and C (that is, if the model does not contain parameters linking variables from A directly to variables from C), then when the table is collapsed over the variables in C, model parameters relating variables in A and model parameters relating variables in A with variables in B are unchanged.*

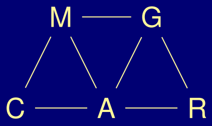


## Teen Survey Data

```
> data(teens)
> ftable(R + G + M ~ A + C, data = teens)

        R  White               Other
        G Female     Male     Female     Male
        M    Yes  No  Yes  No    Yes  No  Yes  No
A   C
Yes Yes      405 268  453 228     23  23   30  19
    No        13 218   28 201      2  19    1  18
No  Yes        1  17    1  17      0   1    1   8
    No         1 117    1 133      0  12    0  17
```

## Teen Survey Data (ctd)

Text suggests loglinear model (AC, AM, CM, AG, AR, GM, GR).



The set $\{A, M\}$ separates sets $\{C\}$ and $\{G, R\}$.
I.e., C is conditionally independent of G and R given M and A.
Thus (as verified on the next slide):

> *Collapsing over G and R, the conditional associations between C and M and between C and A are the same as with the model (AC, AM, CM) fitted earlier.*

```
> teens.df <- as.data.frame(teens)
> ACM <- margin.table(teens, 1:3)
> ACM.df <- as.data.frame(ACM)
> teens.m6 <-
    glm(Freq ~ A*C + A*M + C*M + A*G + A*R + G*M + G*R,
        family = poisson, data = teens.df)
> AC.AM.CM <- glm(Freq ~ A*C + A*M + C*M,
                  family = poisson, data = ACM.df)
> coef(teens.m6)

 (Intercept)          ANo          CNo          MNo
     5.97841     -5.75073     -3.01575     -0.38955
       GMale       ROther      ANo:CNo      ANo:MNo
     0.13584     -2.66305      2.05453      3.00592
     CNo:MNo    ANo:GMale   ANo:ROther    MNo:GMale
     2.84789      0.29229      0.59346     -0.26929
GMale:ROther
     0.12619

> coef(AC.AM.CM)

(Intercept)         ANo         CNo         MNo
    6.81387    -5.52827    -3.01575    -0.52486
    ANo:CNo     ANo:MNo     CNo:MNo
    2.05453     2.98601     2.84789
```

## Exam 2: Time and Place

Tuesday, Apr 24, 2012
8:30 a.m. – 10:25 a.m.
Room 100 Griffin-Floyd Hall (FLO 100)

# Exam 2 Review: Building Logistic Regression Models

## Model Selection

- LR tests to compare nested models.
  - $-2(\mathrm{L}_0 - \mathrm{L}_1) = \text{deviance}_0 - \text{deviance}_1$
  - df $=$ diff. in no. nonredundant params $=$ diff. in residual df's
  - Wald tests can also be used, but LR generally preferred.
- AIC.
- Measures of predictive power.
  - Classification table (a.k.a., confusion matrix).
  - Cross-validation.
  - ROC curve, concordance index (area under ROC curve).
  - 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
  - $X^2$ (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 $X^2$ 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:

$$\log\left(\frac{\pi_j}{\pi_J}\right) = \alpha_j + \beta_j x, \qquad j = 1, 2, \ldots, J - 1.$$

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.

- Usual inferential procedures apply
  - $X^2$ and/or $G^2$ for gof in contingency tables.
  - LR tests.
  - Wald tests and CIs.
- Estimated probs calculated from

$$\pi_j = \frac{e^{\alpha_j + \beta_j x}}{1 + e^{\alpha_1 + \beta_1 x} + \cdots + e^{\alpha_{J-1} + \beta_{J-1} x}}, \quad j = 1, 2, \ldots, J-1,$$

$$\pi_J = \frac{1}{1 + e^{\alpha_1 + \beta_1 x} + \cdots + e^{\alpha_{J-1} + \beta_{J-1} x}},$$

# Exam 2 Review: Cumulative Logit Model

For ordinal response $Y$ with $J > 2$ categories.

Model:

$$\text{logit}\big[\Pr(Y \leqslant j)\big] = \alpha_j + \beta x, \qquad j = 1, \ldots, J-1.$$

- 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 $(Y \leqslant j)$ (instead of $(Y > j)$).
- Reversing ordering of $Y$ changes sign of $\beta$.
- Usual inferential methods apply. Takes avantage 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.

| | | Resp 2 | | |
|---|---|---|---|---|
| | | S | F | |
| Resp 1 | S | $n_{11}$ | $n_{12}$ | $n_{1+}$ |
| | F | $n_{21}$ | $n_{22}$ | $n_{2+}$ |
| | | $n_{+1}$ | $n_{+2}$ | n |

|  |  | Resp 2 |  |  |
|---|---|---|---|---|
|  |  | S | F |  |
| Resp 1 | S | $\pi_{11}$ | $\pi_{12}$ | $\pi_{1+}$ |
|  | F | $\pi_{21}$ | $\pi_{22}$ | $\pi_{2+}$ |
|  |  | $\pi_{+1}$ | $\pi_{+2}$ | 1 |

- Want to test $H_0 : \pi_{1+} = \pi_{+1}$ (marginal homogeneity).

- McNemar's test:

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \underset{H_0}{\sim} N(0,1) \qquad \left(\text{or} \quad z^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \underset{H_0}{\sim} \chi^2_1\right)$$

- CI for $\pi_{1+} - \pi_{+1}$

$$p_{1+} - p_{+1} = \frac{n_{1+}}{n} - \frac{n_{+1}}{n}$$

$$\text{SE} = \frac{1}{n}\sqrt{n_{12} + n_{21} - \frac{(n_{12} - n_{21})^2}{n}}$$

# Exam 2 Review: Measuring Agreement

Suppose rating on a 4-point scale.

| | | Rater 2 | | | | |
|---|---|---|---|---|---|---|
| | | 1 | 2 | 3 | 4 | |
| Rater 1 | 1 | $n_{11}$ | $n_{12}$ | $n_{13}$ | $n_{14}$ | $n_{1+}$ |
| | 2 | $n_{21}$ | $n_{22}$ | $n_{23}$ | $n_{24}$ | $n_{2+}$ |
| | 3 | $n_{31}$ | $n_{32}$ | $n_{33}$ | $n_{34}$ | $n_{3+}$ |
| | 4 | $n_{41}$ | $n_{42}$ | $n_{43}$ | $n_{44}$ | $n_{4+}$ |
| | | $n_{+1}$ | $n_{+2}$ | $n_{+3}$ | $n_{+4}$ | $n$ |

Cohen's kappa measures agreement as departure from independence in direction of perfect agreement:

$$\kappa = \frac{\text{Pr}(\text{agree}) - \text{Pr}(\text{agree}|\text{indep})}{1 - \text{Pr}(\text{agree}|\text{indep})} = \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+}\pi_{+i}}{1 - \sum_i \pi_{i+}\pi_{+i}}$$

- $\kappa = 0$ 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 working correlation structure misspecified.
- (Robust) standard errors adjusted to reflect actual observed depedendence, even if form of working correlation is wrong.
- “Quasi-likelihood” method. No particular form assumed for joint distribution of responses.

# Exam 2 Review: Random Effects Models (GLMM)

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 $\mathrm{N}(0, \sigma^2)$ distribution, $\sigma^2$ unknown.
- $\sigma^2 = 0$ means responses independent (not usually expected with repeated measures).

# Exam 2 Review: Contrasting GEE and GLMM

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.

# Exam 2 Review: Loglinear Models

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 $n_{ijk}$, $1 \leqslant i \leqslant I$, $1 \leqslant j \leqslant J$, $1 \leqslant k \leqslant K$ as indep. 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 condionally 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 $= (IJ-1)(K-1)$.

Loglinear model related topics:

- Mosaic plots.
- Logit-loglinear connection.
- Independence graphs and collapsibility conditions.

# Small-Area Estimation of Binomial Proportions

Wish to estimate parameters for a large number of geographical areas when the individual areas may have relatively few observations.

E.g., wish to estimate county-level rates of health ins. coverage from a national or statewide survey. Most counties have few observations.

One approach:

- Fit a random effects model treating each small area as a cluster.
- Estimate (or “predict”) random effect(s) for each small area using the mode of its (estimated) conditional distribution given the data.
- Combine with estimates of fixed effects to estimate (or “predict”) small-area means or proportions.

# 2008 Presidential Election Survey

Simulated survey of 2000 voters. For each state $i$ (including DC):

$T_i$ = sample size: proportional to number of actual voters

$y_i$ = number in sample favoring Obama $\sim \text{Bin}(T_i, \pi_i)$

$\pi_i$ = success probability: true proportion that voted for Obama

- Fixed effects model:

$$\text{logit}(\pi_i) = \beta_i, \quad i = 1, \ldots, 51$$

  - Saturated (51 parameters for 51 binomial observations).
  - MLEs are sample proportions for each state.

- Logit model with random intercept:

$$\text{logit}(\pi_i) = u_i + \alpha, \qquad u_1, \ldots, u_{51} \sim \text{i.i.d. } N(0, \sigma^2)$$

  - Two parameters: $\alpha$ and $\sigma^2$.

## 2008 Presidential Election Survey (ctd)

- MLEs for random effects model: $\hat{\alpha} = 0.042$, $\hat{\sigma} = 0.344$
- $\frac{e^{\hat{\alpha}}}{1+e^{\hat{\alpha}}} = 0.511$, close to combined sample proportion of 0.52.
- $\hat{\pi}_i$ from fitted random effects model shown on next slide.
  They are closer than sample proportions to the true $\pi_i$:

  $$\text{RMSE} = \begin{cases} 0.132, & \text{for sample proportions} \\ 0.085, & \text{for model estimates} \end{cases}$$

  RMSE is 35.4% smaller.
- How does it work?
  - $\hat{\pi}_i$ closer to overall proportion than $p_i$ ("shrinkage").
    $\hat{\pi}_i$ varies from 0.414 (MS) to 0.635 (NY), while
    $p_i$ varies from 0.167 (NE) to 1 (HI, VT).
  - The difference between $\hat{\pi}_i$ and $p_i$ tends to be small when $T_i$ is large and larger when $T_i$ is small. This is sensible.

```
     T p.true p.smpl p.hat    MS  20  0.430  0.250 0.414
AK   5  0.379  0.200 0.471    MT   7  0.471  0.286 0.472
AL  29  0.387  0.379 0.450    NC  66  0.497  0.500 0.504
AR  17  0.389  0.235 0.419    ND   5  0.445  0.800 0.548
AZ  35  0.449  0.514 0.512    NE  12  0.416  0.167 0.421
CA 207  0.609  0.589 0.578    NH  11  0.541  0.364 0.475
CO  37  0.537  0.541 0.526    NJ  59  0.571  0.661 0.606
CT  25  0.606  0.640 0.565    NM  13  0.569  0.769 0.582
DC   4  0.925  0.750 0.536    NV  13  0.552  0.462 0.497
DE   6  0.619  0.667 0.534    NY 116  0.629  0.672 0.635
FL 128  0.509  0.492 0.496    OH  87  0.514  0.414 0.441
GA  60  0.469  0.367 0.419    OK  22  0.344  0.318 0.435
HI   7  0.718  1.000 0.593    OR  28  0.568  0.536 0.522
IA  23  0.539  0.565 0.533    PA  92  0.545  0.543 0.535
ID  10  0.359  0.400 0.485    RI   7  0.629  0.714 0.545
IL  84  0.618  0.655 0.613    SC  29  0.449  0.448 0.482
IN  42  0.498  0.643 0.583    SD   6  0.448  0.833 0.559
KS  19  0.415  0.421 0.478    TN  40  0.418  0.425 0.464
KY  28  0.412  0.429 0.473    TX 123  0.436  0.390 0.416
LA  30  0.399  0.500 0.506    UT  15  0.342  0.267 0.436
MA  47  0.618  0.596 0.560    VA  57  0.526  0.491 0.498
MD  40  0.619  0.575 0.545    VT   5  0.675  1.000 0.573
ME  11  0.577  0.636 0.541    WA  46  0.573  0.587 0.554
MI  76  0.573  0.579 0.558    WI  45  0.562  0.489 0.498
MN  44  0.541  0.432 0.466    WV  11  0.425  0.545 0.519
MO  45  0.492  0.444 0.473    WY   4  0.325  0.250 0.483
```

## Remarks

- Specific results of simulation would change if we redid it, but they would be fairly similar.
- Method “borrows strength” from all small areas to improve estimation.
- Typically “shrinks” towards overall average or proportion. Shrinkage is more pronounced for areas with small sample sizes.
- Bias-variance tradeoff.

The next few slides show the R commands that we used to create the simulated survey data, fit the random intercept model, and compute some of the various summaries. Note the use of the extractor functions `fitted()`, `fixef()`, and `VarCorr()`.

```
> datasite <- "http://www.stat.ufl.edu/"
> datadir <- "~presnell/Courses/sta4504-2012sp/Var/"
> datafile <- paste(datasite, datadir, "obama.txt", sep = "")
> obama <- read.table(datafile, header=TRUE)
> names(obama)

[1] "State"  "T"      "p.true"

> obama$y <- with(obama, rbinom(length(State), T, p.true))

> obama <- transform(obama, p.smpl = y/T)
> library(lme4)
> obama.fit <- glmer(cbind(y, T - y) ~ (1|State),
                     data = obama, family = binomial)
> obama$p.hat <- fitted(obama.fit)
> head(obama, 2)

  State  T p.true  y  p.smpl   p.hat
1    AK  5  0.379  1 0.20000 0.47069
2    AL 29  0.387 11 0.37931 0.45014
```

```
> summary(obama.fit)

Generalized linear mixed model fit by the Laplace approximati
Formula: cbind(y, T - y) ~ (1 | State)
   Data: obama
 AIC BIC logLik deviance
 103 107  -49.7     99.5
Random effects:
 Groups Name        Variance Std.Dev.
 State  (Intercept) 0.118    0.344
Number of obs: 51, groups: State, 51

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   0.0424     0.0731    0.58     0.56
```

```
> (alpha <- fixef(obama.fit, drop = TRUE))
(Intercept)
   0.042428
> (sigma2 <- VarCorr(obama.fit)$State[1,1])
[1] 0.11808
> (sigma <- sqrt(sigma2))
[1] 0.34362
```

```
> (p.smpl.combined <- with(obama, sum(y)/sum(T)))

[1] 0.51952

> (rmse.p.smpl <-
     with(obama, sqrt(mean((p.smpl - p.true)^2))))

[1] 0.13229

> (rmse.p.hat <-
     with(obama, sqrt(mean((p.hat - p.true)^2))))

[1] 0.085448
```