Exploring Categorical Data and Contingency Tables

David B King, Ph.D.

November 29, 2015

Contingency Tables

Contingency Table

A **contingency table** is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables. The table entries must be integer counts only and are heavily used in survey research, business intelligence, engineering and scientific research. They provide a basic picture of the interrelation between two or many variables and can help find interactions between them.

Some examples of contingency tables in R are:

```
occupationalStatus
Titanic
UCBAdmissions
crimtab
#Agresti's Migration data:
count <- c(11607,100,366,124,87, 13677,515,302,
172, 225,17819, 270,63,176,286,10192)
region = c("NE", "MW", "S", "W")
row = gl(4, 4, labels = region)
col = gl(4, 1, length = 16, labels = region)
dat = data.frame(row=row,col=col,count=count)
T=xtabs(count ~ row + col, data=dat)
```

Probability Distribution Functions

Most theory for contingency table is based upon the **binomial distribution** or the **hypergeometric distribution**. The following table describes four distributions related to the number of successes in a sequence of draws:

	With replacements	No replacements
Given number of draws	binomial distribution	hypergeometric distribution
Given number of failures	negative binomial distribution	negative hypergeometric distribution

Binomial Distribution

If X is a random variable representing the number of success in n draws where the probability p of success is fixed then X has a binomial distribution with parameters n and p, denoted $X \sim \text{Bin}(n, p)$ and

$$P(X = x|n, p) = \binom{x}{n} p^x (1-p)^{n-x}$$

$$\begin{aligned}
\text{E}[X] &= np \\
\text{Var}[X] &= npq = np(1-p) \\
\hat{p} &= X/n \\
\text{E}[\hat{p}] &= \frac{\text{E}[X]}{n} = \frac{np}{n} = p \\
\text{Var}[\hat{p}] &= \frac{\text{Var}[X]}{n^2} = \frac{npq}{n^2} = \frac{p(1-p)}{n}
\end{aligned}$$

Probability Distribution Functions

Hypergeometric Distribution

If X is a random variable representing the number of successes drawn out of a pot where we make n draws without replacement and the pot contains K total successes out of total population of N then X has a hypergeometric distribution with parameters n, K and N, denoted $X \sim \text{Hyper}(n, K, N)$ and

$$P(X = k|n, K, N) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

$$\mathrm{E}[X] = n\frac{K}{N}$$

$$\mathrm{Var}[X] = n\frac{K}{N}\frac{N-K}{N}\frac{N-n}{N-1}$$

Probability Distribution Functions

- For the binomial distribution, if n is large and p is not close to 0 or 1 then the large sample Normal approximation to the Binomial distribution holds and

$$P(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right).$$

- Similarly for the Hypergeometric distribution, if n is large and $\{K, N\}$ are even larger compared with n then

$$P(X \leq x) \approx \Phi\left(\frac{x - np}{\sqrt{np(1-p)}}\right).$$

with $p = \frac{K}{N}$.

Confidence Intervals for $\hat{p}$

Large sample asymptotics for $\hat{p}$ rely on the fact that as $n \to \infty$

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, pq)$$

This puts the formula for the two-sided $(1 - \alpha) \times 100\%$ confidence interval for p as

$$\left[\hat{p} - z_{1-\alpha/2}\frac{\hat{p}\hat{q}}{n}, \hat{p} + z_{1-\alpha/2}\frac{\hat{p}\hat{q}}{n}\right]$$

Or, alternatively the one sided confidence intervals as

$$\left[\hat{p} - z_{1-\alpha}\frac{\hat{p}\hat{q}}{n}, 1\right] \text{ or}$$

$$\left[0, \hat{p} + z_{1-\alpha}\frac{\hat{p}\hat{q}}{n}\right].$$

Functions of Statistics and the Delta Theorem

Functions of a statistic are also statistics, moreover if $f(x)$ is at least twice differentiable then by Taylor's Theorem

$$f(x) = f(\mu) + f'(\mu)(x-\mu) + f''(\xi)\frac{(x-\mu)^2}{2!}$$

for some $\xi \in [\mu, x]$. Hence for any random variable X

$$\begin{aligned}\mathrm{Var}[f(X)] &= \mathrm{Var}[f(\mu) + f'(\mu)(X-\mu) + f''(\xi)\frac{(X-\mu)^2}{2!}] \\ &= [f'(\mu)]^2\mathrm{Var}(X) + \frac{[f''(\xi)]^2}{4}\kappa_4 \approx [f'(\mu)]^2\sigma^2\end{aligned}$$

Consequently, if it is known that

$$\sqrt{n}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma^2)$$

then the Delta Method says that

$$\boxed{\sqrt{n}(f(\hat{\theta}) - f(\theta)) \xrightarrow{d} N(0, [f'(\theta)]^2\sigma^2).}$$

The Delta Theorem

Proof of Delta Method: Since $g'(x)$ is continuous we have by the mean value theorem

$$g(\hat{\theta}) = g(\theta) + g'(\tilde{\theta})(\hat{\theta} - \theta)$$

where $\hat{\theta} < \tilde{\theta} < \theta$. Now as $n \to \infty$ then if $\hat{\theta} \xrightarrow{p} \theta$ then $\tilde{\theta} \xrightarrow{p} \theta$. Moreover, since

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] = g'(\tilde{\theta})\sqrt{n}[\hat{\theta} - \theta]$$

and

$$\sqrt{n}[\hat{\theta} - \theta] \xrightarrow{d} N(0, \sigma^2)$$

it follows that

$$\sqrt{n}[g(\hat{\theta}) - g(\theta)] \xrightarrow{d} N(0, [g'(\theta)]^2\sigma^2).$$

Example of the Delta Theorem

- Let $f(x) = \sin^{-1}(\sqrt{x})$, then $f'(x) = \frac{1}{2\sqrt{x}\sqrt{1-(\sqrt{x})^2}} = \frac{1}{2\sqrt{x}\sqrt{1-x}}$ and $[f'(x)]^2 = \frac{1}{4x(1-x)}$
- Consider the asymptotic distribution of $\sqrt{n}(f(\hat{p}) - f(p))$ for $\hat{p}$ the estimator of the binomial proportion.
- Since the large sample approximation to the Binomial distribution ensures that

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, p(1-p))$$

- It follows by the Delta Method that

$$\begin{aligned}\sqrt{n}(f(\hat{p}) - f(p)) &\xrightarrow{d} N(0, [f'(p)]^2 p(1-p)) \\ &= N\left(0, \frac{p(1-p)}{4p(1-p)}\right) = N\left(0, \frac{1}{4}\right)\end{aligned}$$

Hence the transform $f(x) = \sin^{-1}(\sqrt{x})$ is called the **variance stabalizing** transform for the binomial proportion, because the variance of this statistic does not depend on p!!

Measures of Effect for Categorical Data

1. Probability or risk difference: $\hat{p}_1 - \hat{p}_2$
2. Probability or risk ratio: $\hat{p}_1 / \hat{p}_2$
3. Log risk ratio: $\log(\hat{p}_1) - \log(\hat{p}_2)$
4. Odds: $\frac{\hat{p}}{(1-\hat{p})}$
5. Log Odds: $\log\left(\frac{\hat{p}}{(1-\hat{p})}\right)$
6. Odds Ratio: $\frac{\frac{\hat{p}_1}{(1-\hat{p}_1)}}{\frac{\hat{p}_2}{(1-\hat{p}_2)}}$.
7. Log Odds Ratio: $\theta = \log\left(\frac{\hat{p}_1}{(1-\hat{p}_1)}\right) - \log\left(\frac{\hat{p}_2}{(1-\hat{p}_2)}\right)$.

The Odds Function

1. If $p \in [0,1]$ is a probability, the odds of p is the function

$$\text{Odds} = f(p) = \frac{p}{(1-p)}$$

2. The odds function is a continuously differential mapping $f : [0,1] \mapsto [0,\infty)$ and the inverse function $f^{-1} : [0,\infty) \mapsto [0,1]$ is given by

$$p = f^{-1}(\text{Odds}) = \frac{\text{Odds}}{(\text{Odds}+1)}$$

3. When $p = 1/2 \Longrightarrow \text{Odds} = 1$.
4. Since large sample approximation to the binomial distribution ensures that

$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, pq).$$

5. Application of the Delta Method ensures that

$$\begin{aligned}\sqrt{n}(\widehat{\text{Odds}} - \text{Odds}) &= \sqrt{n}(f(\hat{p}) - f(p)) \xrightarrow{d} N(0, [f'(p)]^2 pq) \\ &= N\left(0, \left[\frac{(1-p)(1) - (p)(-1)}{(1-p)^2}\right]^2 pq\right) = N\left(0, \frac{p}{q^3}\right).\end{aligned}$$

The Log(Odds) or Logit Function

1. If $p \in [0,1]$ is a probability, the logit of p is the function
$$\log(\text{Odds}) = f(p) = \log\left(\frac{p}{1-p}\right)$$
2. The derivative is $f'(p) = \left[\frac{1}{p/(1-p)}\right]\left[\frac{1}{(1-p)^2}\right] = \frac{1}{p(1-p)}$.
3. The logit function is a bijective continuously differential mapping $f : [0,1] \mapsto (-\infty,\infty)$ and the inverse function $f^{-1} : (-\infty,\infty) \mapsto [0,1]$ is given by
$$p = f^{-1}(x) = \frac{\exp(x)}{(\exp(x)+1)}$$
4. When $p = 1/2 \Longrightarrow \text{Logit}(p) = 0$.
5. Since large sample approximation to the binomial distribution ensures that
$$\sqrt{n}(\hat{p} - p) \xrightarrow{d} N(0, pq).$$
6. Application of the Delta Method ensures that
$$\begin{aligned}\sqrt{n}(\widehat{\theta} - \theta) &= \sqrt{n}(\text{logit}(\hat{p}) - \text{logit}(p)) \xrightarrow{d} N(0, [f'(p)]^2 pq) \\ &= N\left(0, \left[\frac{1}{p^2(1-p)^2}\right] pq\right) = N\left(0, \frac{1}{pq}\right).\end{aligned}$$

Measures of Association

Large Sample Estimator for Risk Difference

$$(\hat{p}_1 - \hat{p}_2) \sim N\left(p_1 - p_2, \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}\right)$$

$$2\,\text{sided confidence interval}: \left[(\hat{p}_1 - \hat{p}_2) - z_{1-\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}, (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha/2}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right]$$

$$1\,\text{sided confidence intervals}: \left[(\hat{p}_1 - \hat{p}_2) - z_{1-\alpha}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}, 1\right] \text{ or}$$

$$\left[0, (\hat{p}_1 - \hat{p}_2) + z_{1-\alpha}\sqrt{\frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}}\right]$$

prop.test()

Measures of Association

Large Sample Estimator for Risk Ratio

	Disease		
Treatment	Present	Absent	Total
1	a	b	a+b
2	c	d	c+d
Total	a+c	b+d	n

$$\hat{RR} = \frac{\hat{p}_{\text{Disease 1}}}{\hat{p}_{\text{Disease 2}}} = \frac{a/(a+b)}{c/(c+d)}$$

$$\ln\left(\hat{RR}\right) = \ln\left(\hat{p}_1\right) - \ln\left(\hat{p}_2\right)$$

By the delta method

$$Var[\ln(\hat{p}_1)] = \left[\frac{1}{\hat{p}_1}\right]^2 Var[\hat{p}_1] = \left[\frac{1}{\hat{p}_1}\right]^2 \left[\frac{\hat{p}_1\hat{q}_1}{n_1}\right] = \left[\frac{\hat{q}_1}{\hat{p}_1 n_1}\right] = \left[\frac{b}{an_1}\right]$$

Similarly, $Var[\ln(\hat{p}_2)] = \left[\frac{d}{cn_2}\right]$

2-sided confidence interval on Ln(RR): $\left[\ln(\hat{RR}) - z_{1-\alpha/2}\sqrt{\frac{b}{an_1}+\frac{d}{cn_2}}, \ln(\hat{RR}) + z_{1-\alpha/2}\sqrt{\frac{b}{an_1}+\frac{d}{cn_2}}\right]$

2-sided confidence interval on RR: $\exp\left[\ln(\hat{RR}) - z_{1-\alpha/2}\sqrt{\frac{b}{an_1}+\frac{d}{cn_2}}, \ln(\hat{RR}) + z_{1-\alpha/2}\sqrt{\frac{b}{an_1}+\frac{d}{cn_2}}\right]$

Measures of Association

Large Sample Estimator for Odds

	Disease		
Treatment	Present	Absent	Total
1	a	b	a+b
2	c	d	c+d
Total	a+c	b+d	n

$$Odds_1 = \frac{p_1}{(1-p_1)} = \frac{a/(a+b)}{b/(a+b)} = \frac{a}{b}$$

$$Odds_2 = \frac{p_2}{(1-p_2)} = \frac{c/(c+d)}{d/(c+d)} = \frac{c}{d}$$

$$\text{Log}(\text{Odds}_1) = \log(a) - \log(b); \quad \text{Log}(\text{Odds}_2) = \log(c) - \log(d)$$

$$\text{Var}[\text{Log}(\text{Odds})] = \text{Var}\left[\text{Log}\left(\frac{\hat{p}}{1-\hat{p}}\right)\right] = \left[\frac{1}{\frac{\hat{p}}{1-\hat{p}}}\left[\frac{(1-\hat{p})1-\hat{p}(-1)}{(1-\hat{p})^2}\right]\right]^2\left[\frac{\hat{p}\hat{q}}{n}\right]$$

$$= \left[\frac{1}{\hat{p}\hat{q}}\right]^2\left[\frac{\hat{p}\hat{q}}{n}\right] = \frac{1}{n\hat{p}\hat{q}}$$

Measures of Association

Large Sample Estimator for Odds

	Disease		
Treatment	Present	Absent	Total
1	a	b	a+b
2	c	d	c+d
Total	a+c	b+d	n

Harmonic Sum

$$Log(Odds_1) = \log\left(\frac{a}{b}\right); \text{ and similarly } Log(Odds_2) = \log\left(\frac{c}{d}\right)$$

$$Var[\log(Odds_1)] = \frac{1}{n_1 \hat{p}_1 \hat{q}_1} = \frac{1}{(a+b)\left(\frac{a}{a+b}\right)\left(\frac{b}{a+b}\right)} = \frac{(a+b)}{ab} = \frac{1}{a} + \frac{1}{b} \text{ and similarly}$$

$$Var[\log(Odds_2)] = \frac{1}{c} + \frac{1}{d}$$

Two Sided Confidence Interval for Log($Odds_1$):

$$\left[\log\left(\frac{a}{b}\right) - z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}}, \log\left(\frac{a}{b}\right) + z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}}\right]$$

Measures of Association

The Odds Ratio

	Disease		
Treatment	Present	Absent	Total
Drug	a	b	a+b
Placebo	c	d	c+d
Total	a+c	b+d	n

The **Disease Odds Ratio** is the odds in favor of the disease for the drug group divided by the odds in favor of the disease for the unexposed group

$$\hat{OR}_{\text{Disease}} = \frac{\text{Odds}_{\text{disease}|\text{drug}}}{\text{Odds}_{\text{disease}|\text{placebo}}} = \frac{P(disease|\,drug)/(1-P(disease|\,drug))}{P(disease|\,placebo)/(1-P(disease|\,placebo))} = \frac{\text{a/b}}{\text{c/d}} = \frac{\text{ad}}{\text{bc}}$$

The **Exposure Odds Ratio** is the odds in favor of being exposed to drug for disease group divided by the odds in favor of being exposed to drug for the non diseased group

$$\hat{OR}_{\text{Exposure}} = \frac{\text{Odds}_{\text{drug}|\text{disease}}}{\text{Odds}_{\text{drug}|\text{no disease}}} = \frac{\text{P(drug}\,|\,\text{disease)}/(1-\text{P(drug}\,|\,\text{disease))}}{\text{P(drug}\,|\,\text{no disease)}/(1-\text{P(drug}\,|\,\text{no disease))}} = \frac{a/c}{b/d} = \frac{ad}{bc}$$

THE NICE PROPERTY OF ODDS RATIOS: $\hat{OR}_{Disease} = \hat{OR}_{Exposure} = \hat{OR} = \frac{ad}{bc} = \hat{\theta}$

Measures of Association

Large Sample Confidence Intervals for Odds Ratios

	Disease		
Treatment	Present	Absent	Total
Drug	a	b	a+b
Placebo	c	d	c+d
Total	a+c	b+d	n

$$\hat{OR} = \hat{\theta} = \frac{ad}{bc}$$

$$Log(\hat{\theta}) = \log\left(\frac{ad}{bc}\right)$$

Harmonic Sum of entire table! NICE!

$$Var\left(\log(\hat{\theta})\right) = \frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}$$

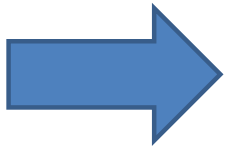
$$\left[\log\left(\frac{ad}{bc}\right) - z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}, \log\left(\frac{ad}{bc}\right) + z_{1-\alpha/2}\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}}\right]$$

Measures of Association

Three Way Tables

Example 1: Death Penalty Data

```
> library(faraway)
> ?death
> data(death)
> death
   y penalty victim defend
1  19     yes      w      w
2 132      no      w      w
3   0     yes      b      w
4   9      no      b      w
5  11     yes      w      b
6  52      no      w      b
7   6     yes      b      b
8  97      no      b      b
```

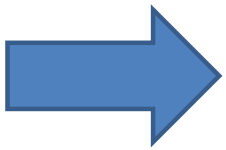


```
> Table=xtabs(y~victim+defend+penalty,data=death)
> Table
, , penalty = no

      defend
victim   b   w
    b  97   9
    w  52 132

, , penalty = yes

      defend
victim   b   w
    b   6   0
    w  11  19
```

Measures of Association

Example 2: Clinical Trial Data

```
> clinic=gl(2,4)
> clinic
[1] 1 1 1 1 2 2 2 2

> treatment=gl(2,2,length=8,labels=c("A","B"))
> treatment
[1] A A B B A A B B

> Result=gl(2,1,length=8,labels=c("Success","Fail"))
> Result
[1] Success Fail    Success Fail    Success Fail    Success Fail
Levels: Success Fail

> count=c(18,12,12,8,2,8,8,32)

> T=data.frame(Clinic=clinic,Trt=treatment,Result=Result,Count=count)
> T
  Clinic Trt  Result Count
1      1   A Success    18
2      1   A    Fail    12
3      1   B Success    12
4      1   B    Fail     8
5      2   A Success     2
6      2   A    Fail     8
7      2   B Success     8
8      2   B    Fail    32
```



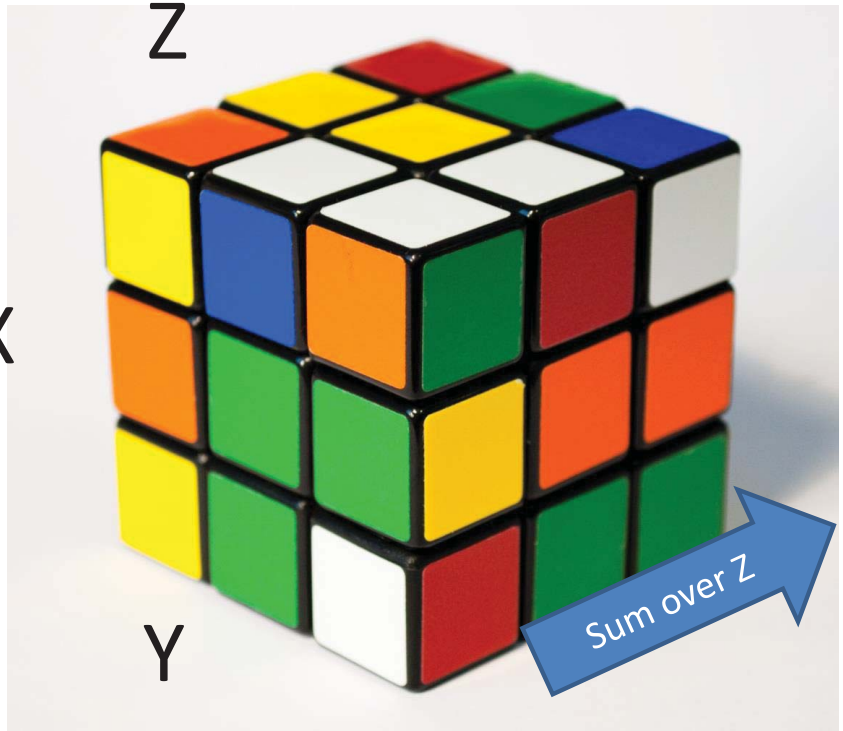
```
> Tab=xtabs(Count~Clinic+Trt+Result,data=T)
> Tab
, , Result = Success

      Trt
Clinic  A  B
     1 18 12
     2  2  8

, , Result = Fail

      Trt
Clinic  A  B
     1 12  8
     2  8 32
```

Measures of Association

Three Way Tables

Three-Way Tables (like array's) are like Rubik's Cubes

Marginal Tables result from summing over one dimension



Marginal (XY) Table
(Summing over Z)

X

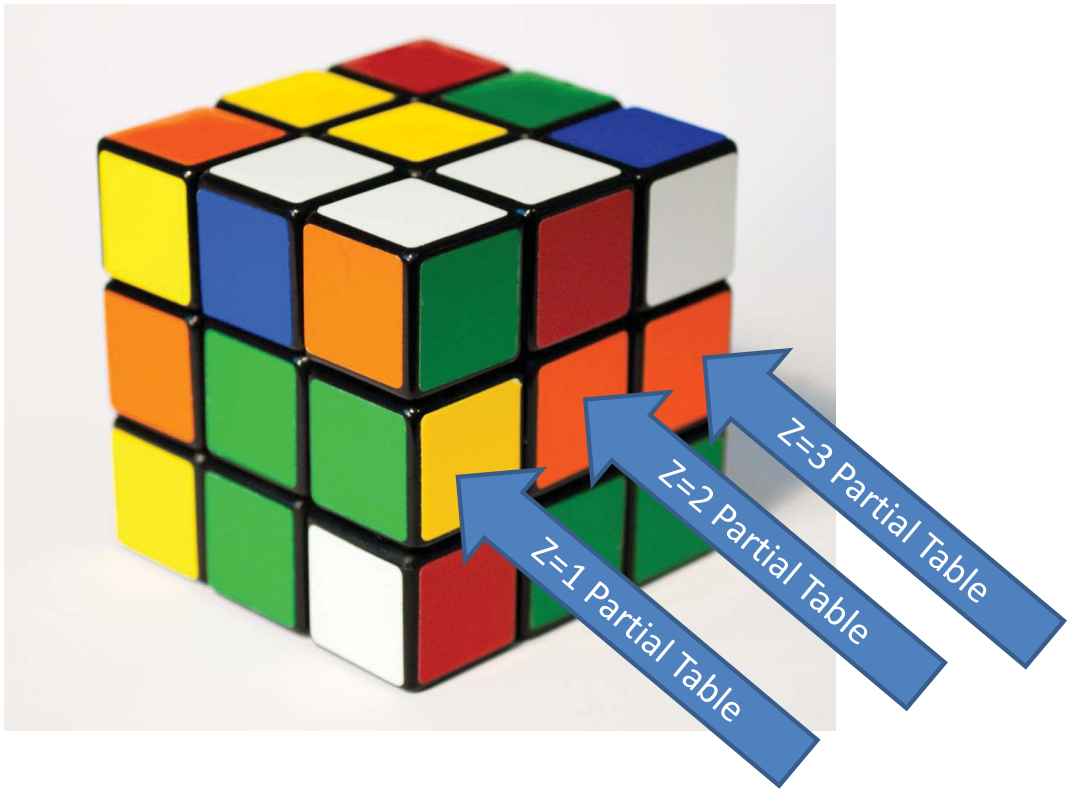
a	b	c
d	e	f
g	h	i

Y

Marginal XZ table = sum over Y
Marginal YZ table = sum over X
.....

Measures of Association

Partial Tables are two way tables which cross classify two variables at different levels of a third variables. Such as the two way table cross-classifying X and Y at separate levels of Z. The Cross-sectional "Slices" are called **partial tables**.



Measures of Association

Conditional Association versus Marginal Associations

Definition

A **conditional association** is an association which is found between two variables, say X and Y, at a certain value of a third variable Z. **Conditional associations are found for partial tables**.

Definition

A **marginal association** is an association which is found between two variables, say X and Y when one sums over the other variables. **Marginal associations are found for the marginal tables**.

Measures of Association

Conditional Association versus Marginal Associations

Death Penalty Example

```
> dpflat <- ftable(penalty ~ victim + defend,data=Table)
> dpflat
          penalty  no yes
victim defend
b      b          97   6
       w           9   0
w      b          52  11
       w         132  19
```

Z = Victim's race
X = Defendant's race
Y = Death Penalty (yes, no)

Conditional Tables and Conditional Odd's Ratio's

Z= Victim Race = Black

	Death Penalty		
Defendant's Race	No	Yes	Total
Black	97	6	103
White	9	0	9
Total	106	6	112

Z= Victim Race = White

	Death Penalty		
Defendant's Race	No	Yes	Total
Black	52	11	63
White	132	19	151
Total	184	30	214

$$\hat{\theta}_{XY(Z=1)} = \frac{ad}{bc} = \frac{(97)(0)}{(9)(6)} = 0$$

$$\hat{\theta}_{XY(Z=2)} = \frac{ad}{bc} = \frac{(52)(19)}{(132)(11)} = 0.43$$

Measures of Association

Conditional Association versus Marginal Associations

Death Penalty Example

Conditional Tables and Conditional Odd's Ratio's

Z= Victim Race = Black

Defendant's Race	Death Penalty No	Yes	Total
Black	97	6	103
White	9	0	9
Total	106	6	112

Z= Victim Race = White

Defendant's Race	Death Penalty No	Yes	Total
Black	52	11	63
White	132	19	151
Total	184	30	214

$$\hat{\theta}_{XY(Z=1)} = \frac{ad}{bc} = \frac{(97)(0)}{(9)(6)} = 0 \qquad \hat{\theta}_{XY(Z=2)} = \frac{ad}{bc} = \frac{(52)(19)}{(132)(11)} = 0.43$$

Marginal XY Table and Marginal Odd's Ratio's

MARGINAL TABLE

Defendant's Race	Death Penalty No	Yes	Total
Black	149	17	166
White	141	19	160
Total	290	36	326

$$\hat{\theta}_{XY} = \frac{ad}{bc} = \frac{(149)(19)}{(141)(17)} = 1.45$$

Measures of Association

Conditional Association versus Marginal Associations

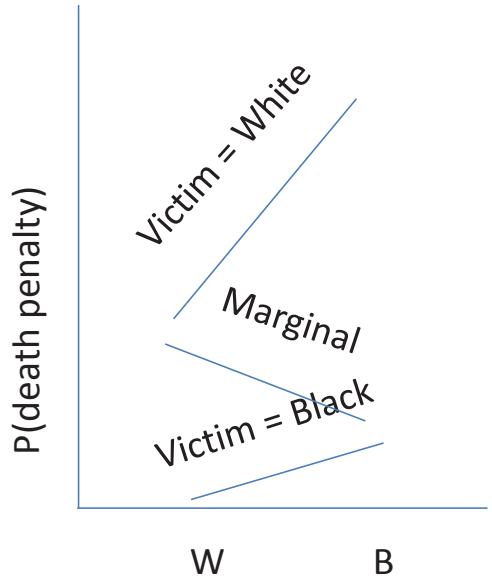
Death Penalty Example

What's going on?

When the victim was black, the death penalty was 2.8% more for black Defendant's than white defendant's.

When the victim was white, the death penalty was impose 11.6% more Often for black defendant's than white defendant's.

However, if we ignore the race of the victim
the percentage of death penalty verdict's for
black defendant's is lower than for white defendant's



Definition

Simpson's Paradox
The marginal associations can have a different direction than the conditional associations.

Moral: can be dangerous to "collapse" contingency tables.

Example 2 of Simpson's Paradox

Alaska Airlines

Destination	# of Arrivals	Arrivals on Time	% On Time
Los Angeles	559	497	88.9%
Phoenix	233	221	94.8%
San Diego	232	212	91.4%
San Francisco	605	503	83.1%
Seattle	2146	1841	85.8%
Five-Airport Totals	3775	3274	**86.7%**

America West Airlines

Destination	# of Arrivals	Arrivals On Time	% On Time
Los Angeles	811	694	85.6%
Phoenix	5255	4840	92.1%
San Diego	448	383	85.5%
San Francisco	449	320	71.3%
Seattle	262	201	76.7%
Five-Airport Totals	7225	6438	**89.1%**

In 5 of the major destination cities Alaska Airlines had better on time performance than America West

Alaska Airlines > America West

Example 2 of Simpson's Paradox

Let's give our summary to the CEO by summing over airport destination and just report the marginal totals

Marginal Total	Flight Arrivals	Arrivals On Time	% On Time
America-West	7225	6438	**89.10%**
Alaska Airlines	3775	3274	**86.70%**

Oops different conclusion: America West > Alaska Airlines

What happened, is this some kind of trick? How can Alaska Airlines be better at every destination category but be worse when you sum over every destination?

Simpson's Paradox: An abrupt change between the conditional Associations and the marginal associations that occur when one sums over and ignores **a hidden and significant lurking variable.**

Example 2 of Simpson's Paradox

A **lurking variable** is another variable which affects the associations between other variables.

Lurking variable: City Destination

Airline Flights	Los Angeles	Phoenix	San Diego	San Francisco	Seattle	Total
America-West Flights	811	5255	448	449	262	7225
America-West Percentage	11%	73%	6%	6%	4%	100%
Alaska Airlines Flights	559	233	232	605	2146	3775
Alaska Airlines Percentage	15%	6%	6%	16%	57%	100%

Lurking variable is that 73% of all flights by America West were to Phoenix (fair weather).

Whereas 57% of all Alaska Airlines flights were to Seattle (inclement weather).

Since there were vast disparities between the city destinations we get disparities in marginal results.

Example 2 of Simpson's Paradox

Definition (Homogeneous Association)
Let K denote the number of categories for Z. When X and Y are both binary, there is homogeneous XY association when

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

Definition (Conditional Independence)
X and Y are conditionally independent given Z if they are independent in each partial table.

$$\theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)} = 1.0$$

Generalized Linear Models (GLM)

Generalized Linear Models (GLM)

GLM's have 3 components: Random Component, Systematic Component, Link Function

Random Component: Identifies the probability density function (PDF) of Y

Example 1: Y is (1,0)

$Y_i \sim Binomial(p, n)$ or

$Y_i \sim Binomial(p, 1) = Bernoulli(p)$

Example 2: Y is integer

$Y_i = Poisson(\mu)$

Systematic Component: The systematic component is the right hand side of the model and it identifies the explanatory variables chosen

$$\eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_K X_K$$

η = Linear Predictorinvolving the linear combination of the X's

Link Function: The link function specifies how the mean $\mu = E[Y]$ relates to the linear predictor η

$$g(\mu) = \eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_K X_K$$

Generalized Linear Models (GLM)

Generalized Linear Models (GLM)

Binary Data

Logistic Regression:

$$Y_i \sim Bernoulli(p_i)$$

Random Component

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

Link Function

Systematic Component

Count Data

Poisson Regression:

$$Y_i \sim Poisson(\mu_i)$$

Random Component

$$\log(\mu_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

Link Function

Systematic Component

Generalized Linear Models (GLM)

Count + Person Hour Data

Poisson Regression with an Offset Term:

$$Y_i \sim Poisson(\mu_i)$$

$$\log(\mu_i) = Log(T_i) + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

Offset is utilized in situations where we have count data but Person Hours at risk is also known (See Fitzmaurice chapter 11).

Polychotomous (Multi-Catgory) Data

Y can be in one of K categories. $Y_i = \{1, 2, \ldots, K\}$

For k = 1, 2, ..., K-1 let

$$F_{ik} = P(Y_i \le k), \quad \text{and} \quad F_{iK} = 1 - \sum_{k=1}^{K-1} F_{ik}$$

Proportional Odds Model:

$$Y_i \sim \text{Multinom}(1, F_{i1}, F_{i2}, \ldots, F_{iK})$$

$$\text{logit}(F_{ik}) = \log\left(\frac{F_{ik}}{1 - F_{ik}}\right) = \alpha_k + \beta_1 X_{1i} + \beta_1 X_{2i} + \cdots + \beta_p X_{pi}$$

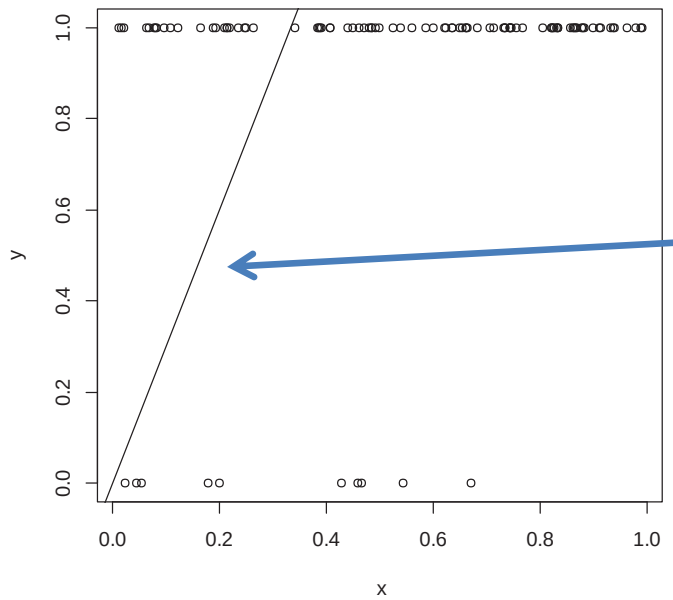
Generalized Linear Models (GLM)

Generalized Linear Models (GLM)

For Binary Data

Y is Binary: Democrat / Republican
Disease / No Disease
0 / 1

$$Y_i \sim Benoulli(P(x))$$



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

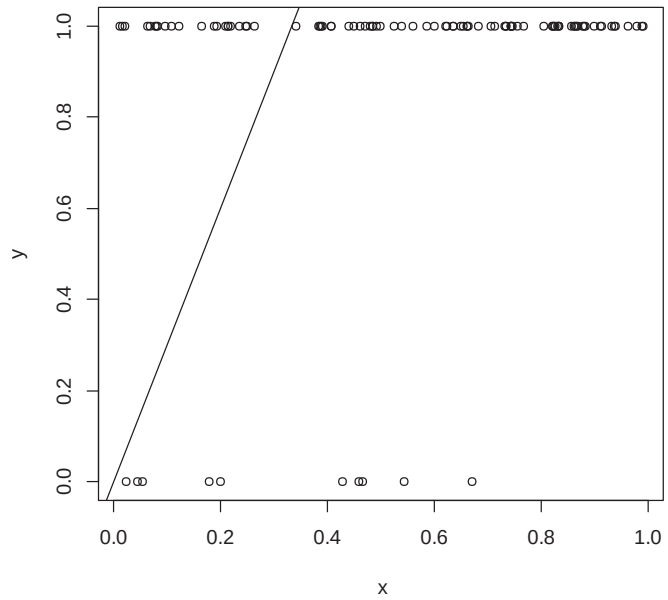
Link Function
= Identity

What's wrong with this equation? $P = -1 + 5x$

Generalized Linear Models (GLM)

Generalized Linear Models (GLM)

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



Problem with Identity link is that

P(x) >1

P(x)<0

Big Problem

Generalized Linear Models (GLM)

GLM for Binary Data

$$Y = 1 \text{ or } 0 \text{ (Bernoulli random variable)}$$

$$Y \sim \text{Bin}(1, P)$$

$$P(Y = 1) = P, \quad P(Y = 0) = 1 - P$$

$$E(Y) = P$$

$$\text{Var}(Y) = P(1 - P)$$

If P= f(x), then Var(Y) = f(x)(1-f(x)) so variance changes with X

➔ Var(Y) is not constant, So cannot use Least Squares Approach!!

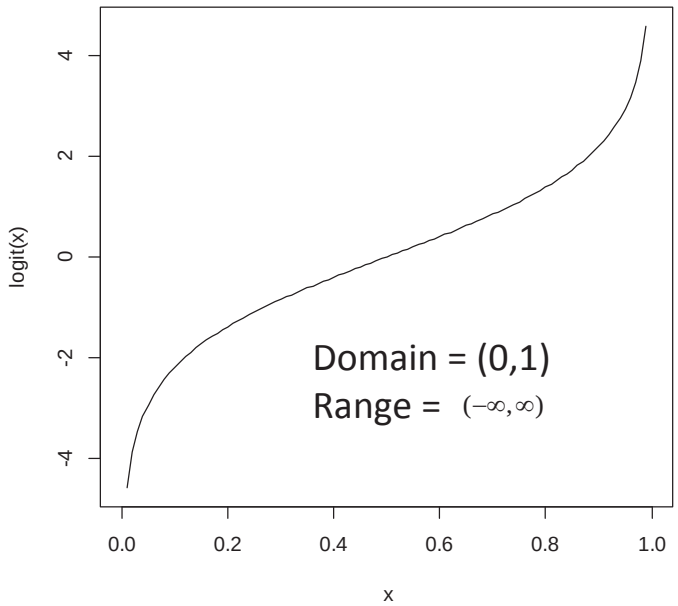
GLM approaches use Max Likelihood approaches
(e.g. Newton-Raphson to estimate parameters)

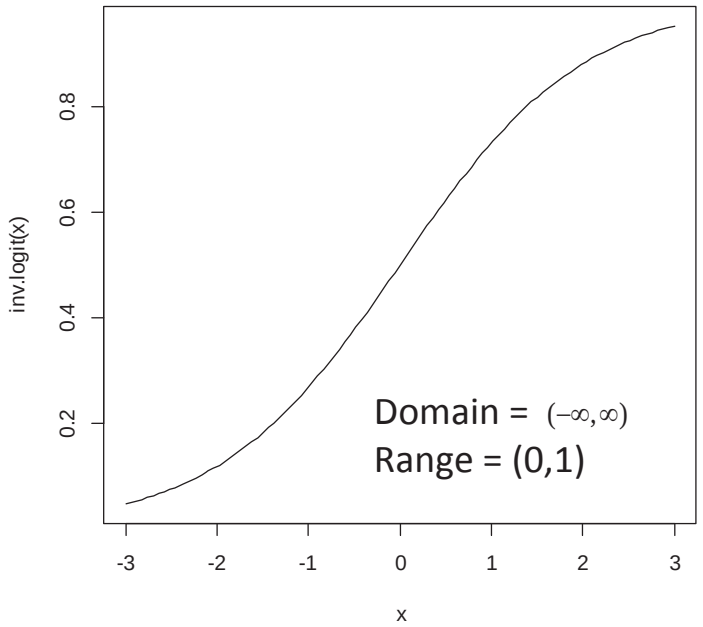
Generalized Linear Models (GLM)

The Logit Function

$$\text{logit(p)} = \log\left(\frac{p}{1-p}\right) = \log(Odds)$$

$$\text{inv.logit(x)} = \frac{\exp(x)}{1+\exp(x)}$$






```
> curve(inv.logit(x),from=-3,to=3)
> curve(logit(x),from=0,to=1)
```

```
library(boot)
?logit
```

Generalized Linear Models (GLM)

Logistic Regression

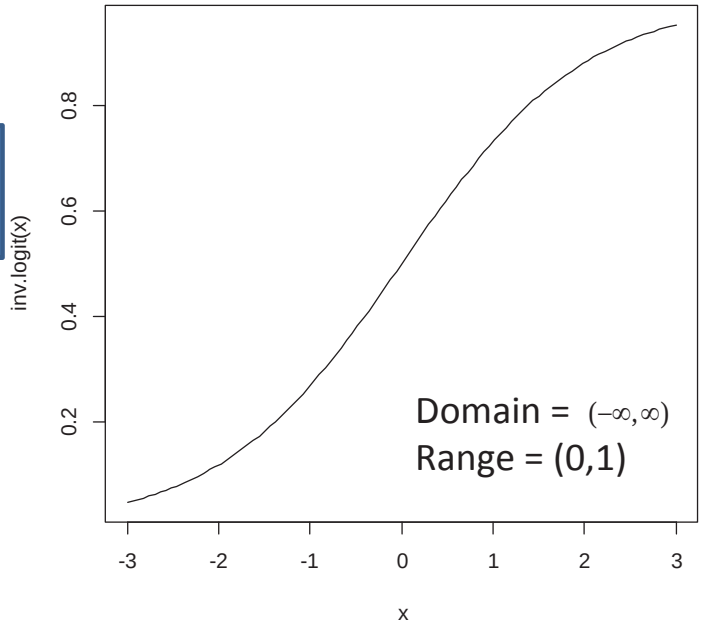
$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \eta = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

η = linear predictor

$$P = \text{inv.logit}(\eta) = \frac{\exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K)}{1 + \exp(\alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K)}$$

Output = any real number in (0,1)

Input = any real number



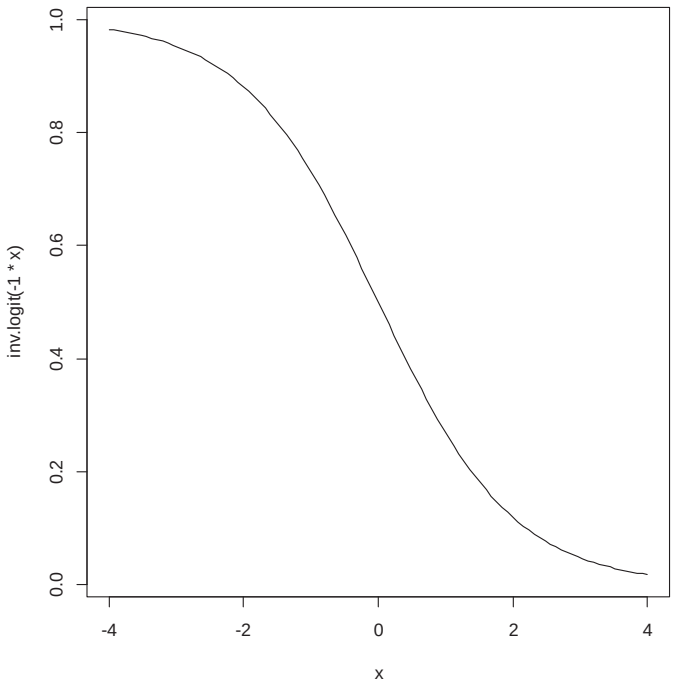
Generalized Linear Models (GLM)

Simple Logistic Regression with 1 X variable

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

$$P = \text{inv.logit}(\alpha + \beta x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

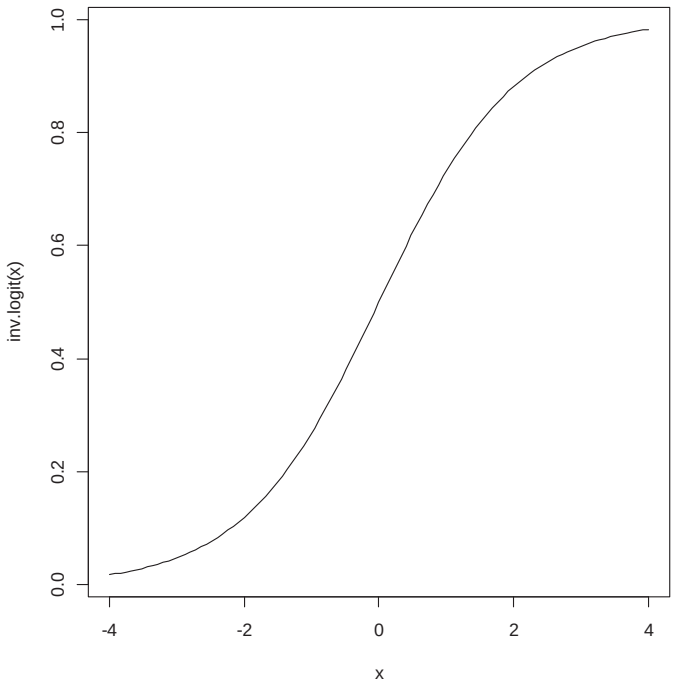
Case shown is for $\beta < 0$



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

$$P = \text{inv.logit}(\alpha + \beta x) = \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)}$$

Case shown is for $\beta > 0$



Generalized Linear Models (GLM)

Generalized Linear Models in R

Generalized linear models are fit using the **glm( )** function. The form of the **glm** function is

glm(*formula***, family=***familytype***(link=***linkfunction***), data=)**

Family	Default Link Function
binomial	(link = "logit")
gaussian	(link = "identity")
Gamma	(link = "inverse")
inverse.gaussian	(link = "1/mu^2")
poisson	(link = "log")
quasi	(link = "identity", variance = "constant")
quasibinomial	(link = "logit")
quasipoisson	(link = "log")

See **help(glm)** for other modeling options. See **help(family)** for other allowable link functions for each family.

Logistic Regression Example 1

```
library(reshape2)
Alcohol=rep(c(0,0.5,1.5,4,7),each=2)
Malformation=gl(2,1,length=10,labels=c("Present","Absent"))
Freq = c(48,17066,38,14464,5,788,1,126,1,37)
acat=ordered(Alcohol,labels=c("0","<1","1-2","3-5",">=6"))
malform = data.frame(Alcohol=Alcohol,Malformation=Malformation,
Freq=Freq,acat=acat)
malform.wd=reshape(malform,idvar=c("acat","Alcohol"),
timevar="Malformation",direction="wide")
names(malform.wd)[3:4] <- c("Present","Absent")
rownames(malform.wd)=1:5
malform.wd <-transform(malform.wd,Total=Present+Absent,
PctPresent=100*Present/(Present+Absent))
print(malform.wd)
```

Generalized Linear Models (GLM)

- fit=glm(cbind(Present,Absent)~Alcohol,data=malform.wd,family=binomial(link="logit"))
- summary(fit)

```
Call:
glm(formula = cbind(Present, Total) ~ Alcohol, family = binomial(link = "logit"),
    data = malform.wd)

Deviance Residuals:
     1       2       3       4       5
0.5872  -0.8752  0.8859  -0.1400  0.1233

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9625     0.1154 -51.669  <2e-16 ***
Alcohol       0.3139     0.1253   2.506  0.0122 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6.1332  on 4  degrees of freedom
Residual deviance: 1.9303  on 3  degrees of freedom
AIC: 24.558

Number of Fisher Scoring iterations: 4
```

Logistic Regression Model

$$Y_i \sim Bin(n_i, P_i)$$

$$\text{logit}(P_i) = \beta_0 + \beta_1 X_{1i}$$

$$\hat{\beta}_0 \approx -6, \hat{\beta}_1 \approx 0.3$$

Residual deviance < degrees of freedom means no overdispersion

We'll get to what overdispersion is later....

Generalized Linear Models (GLM)

```
> coef(fit)
(Intercept)     Alcohol
 -5.9625316   0.3139259
```

What do the coefficients of the model mean?

Linear predictor

(X = Alcohol)

$$\log(ODDS) = \log\left(\frac{P}{1-P}\right) = -5.96 + 0.313X$$

$$ODDS = \frac{P}{1-P} = \exp(-5.96)\exp(0.313X)$$

Cannot substitute the word "probability"

Take exp() on both sides

"For every unit increase in X, the odds of infant malformation Increase by a multiplicative factor of exp(0.313)"

Cannot substitute the words "increase by exp(0.313)" must say multiplicative factor

Generalized Linear Models (GLM)

```
> data(menarche)
> menarche
    Age Total Menarche
1  9.21   376        0
2 10.21   200        0
3 10.58    93        0
4 10.83   120        2
5 11.08    90        2
6 11.33    88        5
7 11.58   105       10
8 11.83   111       17
9 12.08   100       16
10 12.33   93       29
11 12.58  100       39
12 12.83  108       51
13 13.08   99       47
14 13.33  106       67
15 13.58  105       81
16 13.83  117       88
17 14.08   98       79
18 14.33   97       90
19 14.58  120      113
20 14.83  102       95
21 15.08  122      117
22 15.33  111      107
23 15.58   94       92
24 15.83  114      112
25 17.58 1049     1049
```

Example # 2 of Logistic Regression in R

Seems like as Age get's bigger

Proportion of Menarche goes up

xyplot(Menarche/Total~Age,data=menarche)

Generalized Linear Models (GLM)

- fit=glm(cbind(Menarche,Total-Menarche)~Age,data=menarche,family=binomial(link="logit"))
- summary(fit)

```
Call:
glm(formula = cbind(Menarche, Total) ~ Age, family = binomial(link = "logit"),
    data = menarche)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-8.593  -3.285   1.692   3.330   4.447

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.0929     0.1817  -28.03   <2e-16 ***
Age           0.3064     0.0118   25.96   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1211.58  on 24  degrees of freedom
Residual deviance:  410.56  on 23  degrees of freedom
AIC: 528.4

Number of Fisher Scoring iterations: 5
```

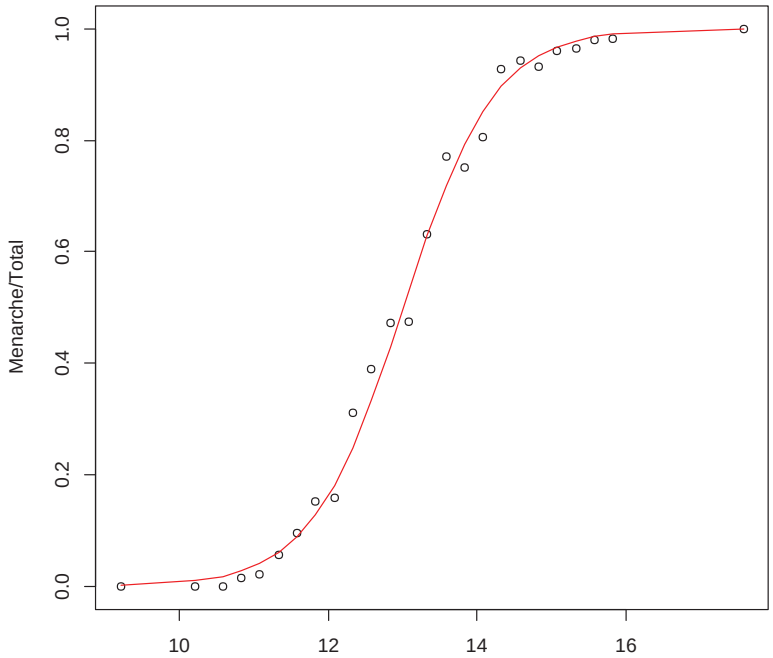
Residual deviance >> degrees of freedom means overdispersion

We'll get to what overdispersion is later....

Generalized Linear Models (GLM)

- plot(Menarche/Total ~ Age, data=menarche)
- lines(menarcheAge, fitfitted, type="l", col="red")
- title(main="Menarche Data with Fitted Logistic Regression Line")

```
> coef(fit)
(Intercept)         Age
 -21.226395    1.631968
```




```
> exp(coef(fit))
 (Intercept)          Age
6.046358e-10 5.113931e+00
```

$$\log(ODDS) = \log\left(\frac{P}{1-P}\right) = -21.22 + 1.63X$$

$$ODDS = \frac{P}{1-P} = \exp(-21.22)\exp(1.63X)$$

$$\left(\frac{ODDS(X+1)}{ODDS(X)}\right) = \frac{\exp(-21.22)\exp(1.63(X+1))}{\exp(-21.22)\exp(1.63(X))} = \exp(1.63) = 5.11$$

For every unit increase in age the odds of menarche increase by a multiplicative factor of 5.11

Inference for Parameters

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

1. ML estimates $\{\hat{\alpha}, \hat{\beta}\}$ by Newton-Raphson maximization iteration.
2. To test $H_0 : \beta = 0$ (independence) we can use the asymptotic property of MLE estimates which ensures that

$$Z = \frac{\hat{\beta} - 0}{\text{se}(\hat{\beta})} \sim N(0, 1)$$

3. For the menarche data $Z = \frac{1.63-0}{0.058} = 27.68 \Longrightarrow$ P-val $< 1.6 \times 10^{-16}$.
4. Could have used Pearson's χ^2 (or deviance G^2) to test independence.
5. Confidence interval of β is $\hat{\beta} \pm Z_{1-\alpha/2}\text{se}(\hat{\beta}) = (1.51, 1.75) = (L, U)$
6. This implies confidence interval on odds is $(\exp(L), \exp(U)) = (4.55, 5.73)$.

Generalized Linear Models (GLM)

Comparing Wald Tests and LR Tests

```
> fit=glm(cbind(Menarche,Total-
Menarach)~Age,data=menarche,family=binomial(link="logit"))
> summary(fit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -21.22639    0.77068  -27.54   <2e-16 ***
Age           1.63197    0.05895   27.68   <2e-16 ***

 Null deviance: 3693.884  on 24  degrees of freedom
Residual deviance:   26.703  on 23  degrees of freedom
AIC: 114.76

Number of Fisher Scoring iterations: 4
```

Null Deviance $= -2\ell_0$

Residual Deviance $= -2\ell_1$

# of iterations in maximization procedure

Likelihood Ratio test for H0: $\beta = 0$

$$LR = -2(\ell_0 - \ell_1) = \text{Null Deviance} - \text{Resid Deviance} \sim \chi^2_{24\text{-}23}$$
$$= 3693.884 - 26.703 = \text{fit\$null.deviance} - \text{fit\$deviance}$$
$$= 3667.18$$

Wald test for H0: $\beta = 0$

$$\text{Wald} = Z^{*2} = \left(\frac{\hat{\beta}}{se(\hat{\beta})}\right)^2 = \left(\frac{1.63197}{0.02895}\right)^2 = 766.4022$$

(Wald relies on numeric computation of second derivatives)

Both Wald and LR have roughly the same P-value in this case

Old School Approach to Modeling Probability

In the old days before we had invented GLM the main tool in the tool bag was ordinary regression.

1. Suppose $y \sim \text{Bin}(n, p(x))$ and we model probability as a linear function of x

$$\hat{p(x)} = \hat{\alpha} + \hat{\beta}x$$

2. Then the variance of y is

$$\text{Var}(y) = np(x)(1 - p(x))$$

3. This violates the homoscedastic assumption of constant variance!!
4. The old school method was to use the **variance stabailizing** transform for the binomial probability so that

$$\text{model}: \sin^{-1}(\sqrt{p(x)}) = \alpha + \beta x + \epsilon$$

This way $\text{Var}(Y) = 1/4n \Longrightarrow$ Constant!!

Example of Old School Approach

```
fit1=lm(asin(sqrt(Present/Total))~Alcohol,data=malform.wd)
summary(fit1)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.047857   0.008870   5.396  0.01248 *
Alcohol     0.015092   0.002414   6.252  0.00826 **
---
Residual standard error: 0.01401 on 3 degrees of freedom
Multiple R-squared:  0.9287,Adjusted R-squared:  0.905
F-statistic: 39.08 on 1 and 3 DF,  p-value: 0.008257
phat = (sin(0.047857 + 0.015092 * Alcohol))^2
attach(malform.wd)
f = function(x){ (sin(0.047857 + 0.015092 *x))^2}
curve(f(x),from=min(Alcohol),to=max(Alcohol),ylim=c(0,0.03),col="red")
points(unique(Alcohol),Present/Total)
```

Generalized Linear Models (GLM)

Building Large Logistic Regression Models

$$Y_i \sim Bin(n_i, p)$$

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_K X_K$$

- Two competing goals: Model fit vs Model Simplicity
- Be careful of multi-collinearity
- Stepwise variable selection (Forward selection, backward elimination, stepwise)
- See demo(glm.vr)
- Significance: practical vs statistical

- Akaike information criterion: AIC

AIC = -2[maximized log-likelihood – number of model parameters

AIC is a compromise between model simplicity and model fit

AIC penalizes over fit models and we want small AIC

Generalized Linear Models (GLM)

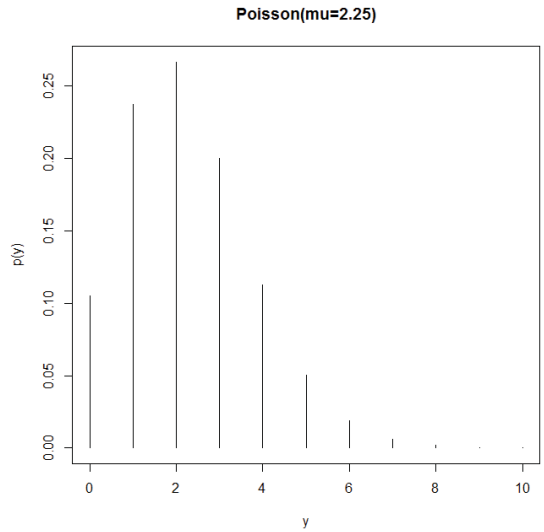
GLMs for Count Data

- When Y is a count (0, 1, 2, 3, . . .) we usually assume a Poisson dist:

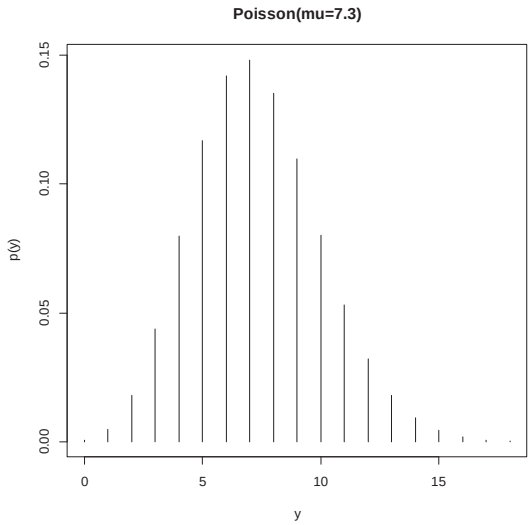
$$P(Y = y) = \frac{\mu^y e^{-y}}{y!}, \quad y = 0,1,2,....$$

$$E[Y] = \mu, \quad \text{Var}[Y] = \mu \quad \text{Mean and Variance are the same!!}$$

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)")



Generalized Linear Models (GLM)

GLMs for Count Data

Poisson Regression Model:

Random Component

$$Y_i \sim Poisson(\mu_i)$$

Log Link Function

$$\log(\mu_i) = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

Identity Link Function

$$\mu_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_K X_{Ki}$$

Log Link is the "Canonical Link" or Natural Link function for the Poisson Dist.

$$P(Y_i \mid \mu_i) = L(\mu_i) = \prod_{i=1}^{n} \frac{\exp(-\mu_i)\mu_i^{Y_i}}{Y_i!}$$

$$\ell(\mu_i) = \log(L(\mu_i)) = \sum_{i=1}^{n} \left(Y_i \log(\mu_i) - \mu_i\right) + C$$

Generalized Linear Models (GLM)

Example 1 of Poisson Regression: (Defects in Silicon Wafers)

```
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="log"),data=wafers)
wafers.loglin <- glm(defects ~ trt,family=poisson(link="log"),data=wafers)
summary(wafers.loglin)
```

```
Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.5280  -0.7622  -0.1699   0.6938   1.5399

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.6094     0.1414  11.380  < 2e-16 ***
trtB          0.5878     0.1764   3.332 0.000861 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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: 94.349

Number of Fisher Scoring iterations: 4
```

Poisson Regression Model with log Link function:

$$Y_i \sim Poisson(\mu_i)$$

$$\log(\mu_i) = \alpha + \beta_1 X_{1i}$$

Generalized Linear Models (GLM)

Example of Poisson Regression: (Defects in Silicon Wafers)

```
> summary(wafers.lin)

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

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.5280  -0.7622  -0.1699   0.6938   1.5399

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   5.0000     0.7071   7.071 1.54e-12 ***
trtB          4.0000     1.1832   3.381 0.000723 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(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: 94.349

Number of Fisher Scoring iterations: 3
```

Poisson Regression Model with identity link function:

$$Y_i \sim Poisson(\mu_i)$$

$$\mu_i = \alpha + \beta_1 X_{1i}$$

Generalized Linear Models (GLM)

Example of Poisson Regression: (Defects in Silicon Wafers)

For Identity Link

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

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

$$x = 0: \ \hat{\mu}_A = 5.0 \ (= \bar{y}_A)$$

$$x = 1: \ \hat{\mu}_B = 9.0 \ (= \bar{y}_B)$$

$$\hat{\beta} = 4.0 = \hat{\mu}_B - \hat{\mu}_A \text{ has SE} = 1.18$$

For Log Link

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

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

$$x = 0: \ \log(\hat{\mu}_A) = 1.609 \ \hat{\mu}_A = e^{1.609} = 5.0 \ (= \bar{y}_A) \text{ good}$$

$$x = 1: \ \log(\hat{\mu}_B) = 2.197 \ \hat{\mu}_B = e^{2.197} = 9.0 \ (= \bar{y}_B) \text{ good}$$

$$\hat{\beta} = 4.0 = \hat{\mu}_B - \hat{\mu}_A \text{ has SE} = 1.18$$

Generalized Linear Models (GLM)

Inference for Poisson Regression works the same way as Logistic Regression

There are 2 types of tests for H0: $\beta = 0$

WALD TYPE: Relies on asymptotic property that as n → infinity

$$Z^* = \frac{\hat{\beta}}{SE(\hat{\beta})} \sim N(0,1) \text{ so } \quad \text{Wald} = Z^{*2} \sim \chi^2_1$$

$\hat{\beta}$ found from maximizing log-likelihood function = $\ell(\text{data} \mid \beta) = \log(L(\text{data} \mid \beta))$

So $\hat{\beta}$ found by solving $\frac{d\ell(\text{data} \mid \beta)}{d\beta} = 0$

$SE(\hat{\beta}) = \sqrt{Var(\hat{\beta})} = \sqrt{\frac{1}{I_n(\hat{\beta}_{MLE})}}$ where

$\hat{I}_n(\hat{\beta}_{MLE})$ is the est of fisher information $I_n(\hat{\beta}_{MLE}) \equiv -E\left[\frac{d\ell^2(\text{data} \mid \beta)}{d\beta^2}\right]$

LIKELIHOOD RATIO TYPE:

The Likelihood Ratio test compares models:

$\text{Model 1: } \log(\mu) = \alpha + \beta x \text{ (Full Model) } \log\text{-likelihood} = \ell_1$

$\text{Model 0: } \log(\mu) = \alpha \text{ (Null Model) } \log\text{-likelihood} = \ell_0$

$$-2(\ell_0 - \ell_1) \sim \chi^2_{dm_1 - dm_0}$$

- 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 confidence intervals for tests of H0 $\beta = \beta_0$ see the the confint() function in R

Generalized Linear Models (GLM)

```
> 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     27.857
trt   1   11.589        18     16.268 0.0006633 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Sequentially adds variables to the model and uses LR to test for significance

```
> drop1(wafers.loglin)
Single term deletions

Model:
defects ~ trt
       Df Deviance     AIC
<none>      16.268
94.349
trt    1   27.857 103.938
```

Cannot Drop treatment from the model!
The variable is significant

Generalized Linear Models (GLM)

Example of Poisson Regression: (Defects in Silicon Wafers)

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

$$e^{\beta} = \frac{\mu_B}{\mu_A} \qquad e^{\hat{\beta}} = e^{0.5878} = \frac{\hat{\mu}_B}{\hat{\mu}_A} = 1.8$$

WALD TYPE CI:

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

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

LR TYPE CI:

```
> confint(wafers.loglin)
Waiting for profiling to be done...
                2.5 %    97.5 %
(Intercept) 1.3188383 1.8743819
trtB        0.2469096 0.9400962
```

$$95\%\,\text{CI for } \beta \text{ is: } \quad (0.246, 0.940)$$

$$95\%\,\text{CI for } e^{\beta} \text{ is: } \quad (e^{0.246}, e^{0.940}) = (1.28, 2.56)$$

Generalized Linear Models (GLM)

Example of Poisson Regression with Offset

Smoking	BP	Behavior	CHD	Personyears
0	0	0	20	5268.2
10	0	0	16	2542
20	0	0	13	1140.7
30	0	0	3	614.6
0	0	1	41	4451.1
10	0	1	24	2243.5
20	0	1	27	1153.6
30	0	1	17	925
0	1	0	8	1366.8
10	1	0	9	497
20	1	0	3	238.1
30	1	0	7	146.3
0	1	1	29	1251.9
10	1	1	21	640
20	1	1	7	374.5
30	1	1	12	338.2

Y = # of cases of coronary heart disease (CHD)

T = Total Person-Years of follow up in group (Offset)

X1 = Blood Pressure (BP)
X2 = Behavior (type A or B)
X3 = Smoking Category.

Model:

$$Y_i \sim Poisson(\mu_i)$$

$$\log(\mu_i/T_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

$$\Rightarrow \log(\mu_i) = \log(T_i) + \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

Generalized Linear Models (GLM)

Read in CHD Data from CSV File:

```
> setwd("C:/Users/David/Longitudinal Data Analysis/")
> getwd()
[1] "C:/Users/David/Longitudinal Data Analysis"
> chd=read.csv("chd.csv")
```

Model wth no offset term:

$$Y_i \sim Poisson(\mu_i)$$
$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

```
fit=glm(CHD~Smoking+BP+Behavior,data=chd,family=poisson())
```

Model with offset term:

$$Y_i \sim Poisson(\mu_i)$$
$$\log(\mu_i/T_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$
$$\Rightarrow \log(\mu_i) = \log(T_i) + \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

```
> fit1=glm(CHD~Smoking+BP+Behavior+offset(log(Personyears)),data=chd,family=poisson(link="log"))
> summary(fit1)
```

Generalized Linear Models (GLM)

```
> fit1=glm(CHD~Smoking+BP+Behavior+offset(log(Personyears)),data=chd,family=poisson(link="log"))
> summary(fit1)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.8038  -0.7615  -0.2933   1.0110   1.8821

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.420153   0.130813 -41.434  < 2e-16 ***
Smoking      0.027344   0.005614   4.871 1.11e-06 ***
BP           0.753377   0.129240   5.829 5.57e-09 ***
Behavior     0.752555   0.136202   5.525 3.29e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 119.06  on 15  degrees of freedom
Residual deviance:  21.24  on 12  degrees of freedom
AIC: 99.546

Number of Fisher Scoring iterations: 4
```

When residual deviance > degrees of freedom this Is evidence of OVERDISPERSION

Generalized Linear Models (GLM)

Predictions with GLM's

Logistic Regression: $\text{logit}(\hat{P}_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_p X_{pi}$

Poisson Regression: $\log(\hat{\mu}_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_p X_{pi}$

Many types of predictions with **Generalized Linear Models (GLMs)**

To obtain predictions use:

predict(object, newdata , type = c("link", "response", "terms"))

1. predict(object, newdata , type = "link") OR predict(object, newdata) gives you

   The linear predictor $\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_p X_{pi}$

2. predict(object, newdata , type = "response") gives you

   Prediction on the probability or count scale (on the scale of the prediction)

$$\hat{P} = \frac{\exp(\hat{\eta})}{1+\exp(\hat{\eta})} \quad \text{Or} \quad \hat{\mu} = \exp(\hat{\eta})$$

Generalized Linear Models (GLM)

Predictions with GLM's

Logistic Regression: $\text{logit}(\hat{P}_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_p X_{pi}$

Poisson Regression: $\log(\hat{\mu}_i) = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_p X_{pi}$

Many types of predictions with **Generalized Linear Models (GLMs)**

To obtain predictions use:

predict(object, newdata , type = c("link", "response", "terms"))

1. predict(object, newdata , type = "link") OR predict(object, newdata) gives you

   The linear predictor $\hat{\eta} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \cdots + \hat{\beta}_p X_{pi}$

2. predict(object, newdata , type = "response") gives you

   Prediction on the probability or count scale (on the scale of the prediction)

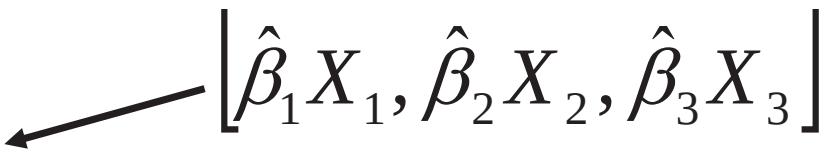
$$\hat{P} = \frac{\exp(\hat{\eta})}{1+\exp(\hat{\eta})} \quad \text{Or} \quad \hat{\mu} = \exp(\hat{\eta})$$

Generalized Linear Models (GLM)

3. predict(object, newdata , type = “terms") gives you

   Matrix continaing each variables contribution to the prediction on the linear predictor scale

```
> predict(fit1,type="terms")
     Smoking        BP   Behavior
1  -0.4101612 -0.3766883 -0.3762774
2  -0.1367204 -0.3766883 -0.3762774
3   0.1367204 -0.3766883 -0.3762774
4   0.4101612 -0.3766883 -0.3762774
5  -0.4101612 -0.3766883  0.3762774
6  -0.1367204 -0.3766883  0.3762774
7   0.1367204 -0.3766883  0.3762774
8   0.4101612 -0.3766883  0.3762774
9  -0.4101612  0.3766883 -0.3762774
10 -0.1367204  0.3766883 -0.3762774
11  0.1367204  0.3766883 -0.3762774
12  0.4101612  0.3766883 -0.3762774
13 -0.4101612  0.3766883  0.3762774
14 -0.1367204  0.3766883  0.3762774
15  0.1367204  0.3766883  0.3762774
16  0.4101612  0.3766883  0.3762774
attr(,"constant")
[1] -4.257026
```



Generalized Linear Models (GLM)

What is Overdispersion?

Overdispersion is a condition where the where the actual data response Y has more variation than can be accounted for in the statistical model. An indication that a model has an overdispersion issue is when **the residual deviance is much larger than the residual degrees of freedom**.

Example: In Logistic Regression We Assume:

$$Y_i \sim Bin(n_i, p_i)$$

So under the assumptions of the model:

Variance function

$$Var(Y_i) = n_i p_i (1 - p_i) = v(p_i)$$

Overdispersion exists if

$$Var(Y_i) = \phi n_i p_i (1 - p_i) = \phi v(p_i)$$

$$\phi > 1$$

Dispersion parameter

Generalized Linear Models (GLM)

Another Example: In Poisson Regression We Assume:

$$Y_i \sim Pois(\mu_i)$$

So under the model we would expect:

$$Var(Y_i) = \mu_i = v(\mu_i)$$ Variance = Mean

Overdispersion exists if

$$Var(Y_i) = \phi\mu_i = \phi v(\mu_i)$$

$$\phi > 1$$

Dispersion parameter

Generalized Linear Models (GLM)

Modeling Overdispersion

One method of accounting for overdispersion is to add an extra source of variability to the linear predictor

$$Y_i \sim Poisson(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

$$Y_i \sim Bin(n_i, p_i)$$

$$\text{logit}(p_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Where,

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{OR} \quad \exp(\varepsilon_i) \sim Gamma(1, \theta)$$

$$Var(Y_i) = \mu_i + \phi\mu_i^2$$

Generalized Linear Models (GLM)

Modeling Overdispersion

When you account for overdispersion by adding an extra source of variability to the linear predictor in poisson regression we have

$$Y_i \sim Poisson(\mu_i)$$

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

Where, $\varepsilon_i \sim N(0, \sigma^2)$ we get

$$Var(Y_i) = \mu_i + (e^{\sigma^2} - 1)\mu_i^2$$

If we assume, $\exp(\varepsilon_i) \sim Gamma(1, \phi)$ we get

$$Var(Y_i) = \mu_i + \phi\mu_i^2$$

Most general case

ϕ Is the dispersion parameter

Generalized Linear Models (GLM)

Modeling Overdispersion in R

To model overdispersion in this way using R we use “family =quasipoisson()” or “family = quasibinom()” in the glm() function

```
> fit2=glm(CHD~Smoking+BP+Behavior+offset(log(Personyears)),data=chd,family=quasipoisson(link="log"))
> summary(fit2)


Deviance Residuals:
   Min      1Q  Median      3Q     Max
-1.8038  -0.7615  -0.2933   1.0110   1.8821

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.420153   0.177707 -30.500 9.68e-13 ***
Smoking      0.027344   0.007627   3.585  0.00375 **
BP           0.753377   0.175570   4.291  0.00105 **
Behavior     0.752555   0.185027   4.067  0.00156 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 1.845468)

    Null deviance: 119.06  on 15  degrees of freedom
Residual deviance:  21.24  on 12  degrees of freedom
Number of Fisher Scoring iterations: 4
```

Generalized Linear Models (GLM)

Modeling Overdispersion with Binomial Model

```
> fit=glm(cbind(Menarche,Total-
Menarche)~Age,data=menarche,family=quasibinomial(link="logit"))
> summary(fit)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.0363  -0.9953  -0.4900   0.7780   1.3675

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -21.22639    0.75151  -28.25   <2e-16 ***
Age           1.63197    0.05749   28.39   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasibinomial family taken to be 0.9508657)

    Null deviance: 3693.884  on 24  degrees of freedom
Residual deviance:   26.703  on 23  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 4
```

Exploring Associations with Log Linear Models

Log Linear Model Theory for Contingency Tables

Exploring Associations with Log Linear Models

Log-Linear Models for Contingency Tables

Log-Linear modeling is a POWERFUL THEORY for categorical data analysis
Recall: Log-Linear Models for Poisson models (GLM with log link)

Log-linear models can be used to describe associations and interaction Patterns among categorical variables in contingency tables

➔ Logit Models are equivalent to certain Log-linear models.

Start with an (R x C) contingency table:

Cell Counts:

	Y					
X	A	B	C	D	...	Total
1	O_{11}	O_{12}	O_{13}	O_{14}	...	n_{1+}
2	O_{21}	O_{22}	O_{23}	O_{24}	...	n_{2+}
3	O_{31}	O_{32}	O_{33}	O_{34}	...	n_{3+}
4	O_{41}	O_{42}	O_{43}	O_{44}	...	n_{4+}
...	...	...	...	...	...	...
Total	n_{+1}	n_{+1}	n_{+1}	n_{+1}	...	n

Cell Probabilities:

	Y					
X	A	B	C	D	...	Total
1	Π_{11}	Π_{12}	Π_{13}	Π_{14}	...	Π_{1+}
2	Π_{21}	Π_{22}	Π_{23}	Π_{24}	...	Π_{2+}
3	Π_{31}	Π_{32}	Π_{33}	Π_{34}	...	Π_{3+}
4	Π_{41}	Π_{42}	Π_{43}	Π_{44}	...	Π_{4+}
...	...	...	...	...	...	...
Total	Π_{+1}	Π_{+1}	Π_{+1}	Π_{+1}	...	1

Exploring Associations with Log Linear Models

Log-Linear Models for Contingency Tables

	Y					
X	A	B	C	D	...	Total
1	O_{11}	O_{12}	O_{13}	O_{14}	...	n_{1+}
2	O_{21}	O_{22}	O_{23}	O_{24}	...	n_{2+}
3	O_{31}	O_{32}	O_{33}	O_{34}	...	n_{3+}
4	O_{41}	O_{42}	O_{43}	O_{44}	...	n_{4+}
...	...	...	...	...	...	...
Total	n_{+1}	n_{+1}	n_{+1}	n_{+1}	...	n

	Y					
X	A	B	C	D	...	Total
1	Π_{11}	Π_{12}	Π_{13}	Π_{14}	...	Π_{1+}
2	Π_{21}	Π_{22}	Π_{23}	Π_{24}	...	Π_{2+}
3	Π_{31}	Π_{32}	Π_{33}	Π_{34}	...	Π_{3+}
4	Π_{41}	Π_{42}	Π_{43}	Π_{44}	...	Π_{4+}
...	...	...	...	...	...	...
Total	Π_{+1}	Π_{+1}	Π_{+1}	Π_{+1}	...	1

Independence of rows and columns implies that

$\pi_{ij} = \pi_{i+}\pi_{+j}$ (If X and Y are independent)

Cell Counts under independence

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

Log-Linear models establish models for counts rather than proportions

$\text{Log-linear models } E_{ij} = \mu_{ij} = E[O_{ij}]$ rather than π_{ij}

Exploring Associations with Log Linear Models

Log-Linear Models for Contingency Tables

Back to the INDEPENDENCE MODEL

$$\mu_{ij} = n\pi_{i+}\pi_{+j} \quad \text{(Multiplicative)}$$

However,

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y$$

Independence Log-Linear Model

Sample size term

Row Term

Column Term

$$\lambda_i^X = \text{row effect of classification}$$

$$\lambda_j^Y = \text{column effect of classification}$$

H0: Independence translate into the above model holding

Fitted values (under H0) = $\frac{n_{i+}n_{+j}}{n}$

X^2 and G^2 both test the Goodness of Fit of the Above Model!!

Exploring Associations with Log Linear Models

Log-Linear Models for Contingency Tables

Now consider the fully statistically dependent or SATURATED MODEL

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

1 parameter	R-1 parameters	C-1 parameters	(R-1)(C-1) parameters	Total parameters= (R)(C)

The above model is called a **saturated model** since number of parameters = (# Rows)(# Columns) = Fits data exactly (so no degrees of freedom to estimate errors).

The λ_{ij}^{XY} reflect interactions: **the effect of one variable on the cell count depends on the level of the other variable.**

λ_{ij}^{XY} are deviations from the independence of X an Y

Independence ➔ $\lambda_{ij}^{XY} = 0$

Exploring Associations with Log Linear Models

Consider the 2 x 2 table

$$\log(\theta) = \log\left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\right) = \log(\mu_{11}) + \log(\mu_{22}) - \log(\mu_{12}) - \log(\mu_{21})$$

Log(ODDS RATIO)

Plug In:

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}$$

$$\log(\theta) = \log\left(\frac{\mu_{11}\mu_{22}}{\mu_{12}\mu_{21}}\right) = (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_{22}^{XY}) - (\lambda + \lambda_1^X + \lambda_2^Y + \lambda_{12}^{XY}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_{21}^{XY})$$

$$= 2(\lambda - \lambda) + (\lambda_1^X - \lambda_1^X) + (\lambda_2^X - \lambda_2^X) + (\lambda_1^Y - \lambda_1^Y) + (\lambda_2^Y - \lambda_2^Y) + (\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY})$$

The constant and linear terms always cancel out in odds ratio calculations

$$\log(\theta) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$

Only the interaction terms survive

➔ The constant and linear terms always go away when computing the odds ratio

➔ The λ_{ij}^{XY} determine $\log(\theta)$

➔ When $\lambda_{ij}^{XY} = 0$ for all i and j then Odds Ratio = 1 (XY independence)

Chi-Square test for independence is equivalent to test that $\lambda_{ij}^{XY} = 0$!!!

Exploring Associations with Log Linear Models

Example 1: Alcohol, Cigarette and Marijuanna Use

A survey was conducted by Wright State University School of Medicine and United Health Services in Dayton, Ohio. The survey asked high school seniors Whether they had ever used alcohol, cigarettes, or marijuana. (Agresti, 2007)

```
count=c(911,538,44,456,3,43,2,279)
A=gl(2,4,labels=c("yes","no"))
C=gl(2,2,8,labels=c("yes","no"))
M=gl(2,1,8,labels=c("yes","no"))

D=data.frame(A=A,C=C,M=M,count=count)
T=xtabs(count~A+C+M,data=D)
```

```
> D
   A   C   M count
1 yes yes yes   911
2 yes yes  no   538
3 yes  no yes    44
4 yes  no  no   456
5  no yes yes     3
6  no yes  no    43
7  no  no yes     2
8  no  no  no   279
```

```
> F=ftable(T)
> F
```

		Marijuana Use	
		yes	no
Alcohol	Cigs		
yes	yes	911	538
	no	44	456
no	yes	3	43
	no	2	279

3-way Contingency Table with dimension = 2 X 2 X 2

Exploring Associations with Log Linear Models

Example 1: Alcohol, Cigarette and Marijuanna Use

X = Alcohol use (yes, no) Y = Cigarette use (yes, no) Z = Marijuana use (yes, no)

Full Saturated Model: $$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ}$$

Two ways to fit this model in R:

loglin() in the stats package:

- satfit=loglin(T,margin=list(c(1,2,3)),fit=TRUE,param=TRUE)
- ftable(satfit$fit, row.vars = 1:3)

A	C	M	$\hat{Y}$
yes	yes	yes	911
		no	538
	no	yes	44
		no	456
no	yes	yes	3
		no	43
	no	yes	2
		no	279

loglm() in the MASS package:

- library(MASS)
- satmodel=loglm(~A*C*M,data=T)
- satmodel$param

Because saturated model has a parameter for every cell in contingency table it fits the data exactly!!

Saturated models generally OVERFIT the data!!

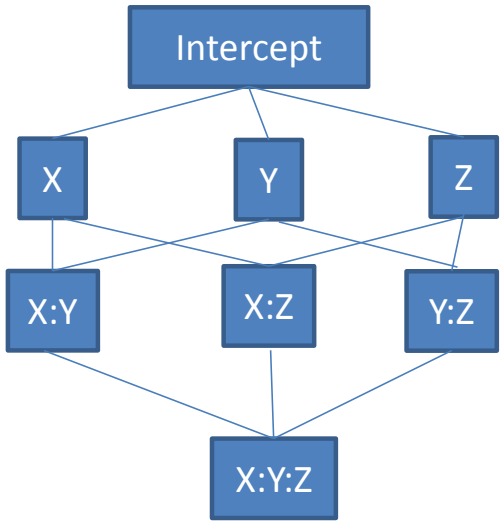
Exploring Associations with Log Linear Models

Rules of Hierarchy

Just as in regression modeling, the rules of building models obey the rules of hierarchy

- If a linear term for any variable is in the model then the constant intercept must be in model
- If interaction term XY is in the model then the linear terms X and Y as well as the interaction
- If the three way interaction terms XYZ are in model then the two way interactions (XY, YZ,XZ), the linear terms (X,Y,Z) and the intercept are in model.

HASSE DIAGRAM DEPICTION OF HIERACHY:

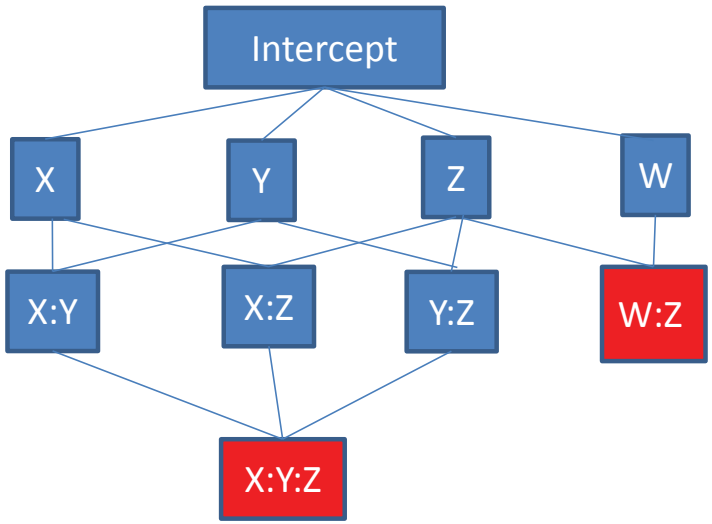


Exploring Associations with Log Linear Models

Short hand notation for models following rules of hierarchy

Because all models obey the rules of hierarchy, we can use short hand notation where we **only list the highest order terms when labeling a model**:

FOR EXAMPLE IN THE HASSE DIAGRAM:



This short-hand label for this model is **the (XYZ, WZ) model**

The model has this form mathematically:

$$\log(\mu_{ijkh}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_h^W + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ} + \lambda_{kh}^{WZ}$$

Exploring Associations with Log Linear Models

The Parameters of the Saturated Model

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ}$$

```
> satfit$param
$`(Intercept)`

[1] 4.274157
$A
    yes      no
 1.478114 -1.478114
$C
     yes       no
 0.2163966 -0.2163966
$M
    yes      no
-1.176534  1.176534
$A.C
    C
A          yes         no
  yes  0.5825332 -0.5825332
  no  -0.5825332  0.5825332
```

λ

λ_i^X 1 redundant parameter

λ_j^Y 1 redundant parameter

λ_k^Z 1 redundant parameter

λ_{ij}^{XY} 1 redundant parameter

```
$A.M
    M
A          yes         no
  yes  0.7236292 -0.7236292
  no  -0.7236292  0.7236292
$C.M
    M
C          yes         no
  yes  0.642558 -0.642558
  no  -0.642558  0.642558
$A.C.M
, , M = yes
    C
A          yes         no
  yes  0.07368884 -0.07368884
  no  -0.07368884  0.07368884
, , M = no
    C
A          yes         no
  yes -0.07368884  0.07368884
  no   0.07368884 -0.07368884
```

λ_{ik}^{XZ} 1 redundant parameter

λ_{jk}^{YZ} 1 redundant parameter

λ_{ijk}^{XYZ} 1 redundant parameter

Exploring Associations with Log Linear Models

The Interesting Log-Linear Models

Model	Equation	Resid DF	Comment
(XYZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ}$	0	**Saturated model**
(XY,XZ,YZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$	1	All pairwise interactions (Homogeneous assoc.)
(XY,XZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XY}$	2	No YZ interaction (Homogeneous assoc.)
(XY,YZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$	2	No XZ interaction (Homogeneous assoc.)
(YZ,XZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	2	No XY interaction (Homogeneous assoc.)
(XY,Z)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	3	Only XY interaction (Homogeneous assoc.)
(XZ,Y)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$	3	Only XZ interaction (Homogeneous assoc.)
(YZ,X)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$	3	Only YZ interaction (Homogeneous assoc.)
(X,Y,Z)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	4	**INDEPENDENCE MODEL!!**

Exploring Associations with Log Linear Models

The Interesting Log-Linear Models

- The last model, **with only linear terms**, is called the **INDEPENDENCE MODEL** because the conditional odds ratio for any pair of variables is zero when you have only linear terms in the model. For example,

$$\log(\theta_{XY(k)}) = \log\left(\frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}\right) = \log(\mu_{11k}) + \log(\mu_{22k}) - \log(\mu_{12k}) - \log(\mu_{21k})$$

$$= (\lambda + \lambda_1^X + \lambda_1^Y) + (\lambda + \lambda_2^X + \lambda_2^Y) - (\lambda + \lambda_1^X + \lambda_2^Y) - (\lambda + \lambda_2^X + \lambda_1^Y) = 0.$$

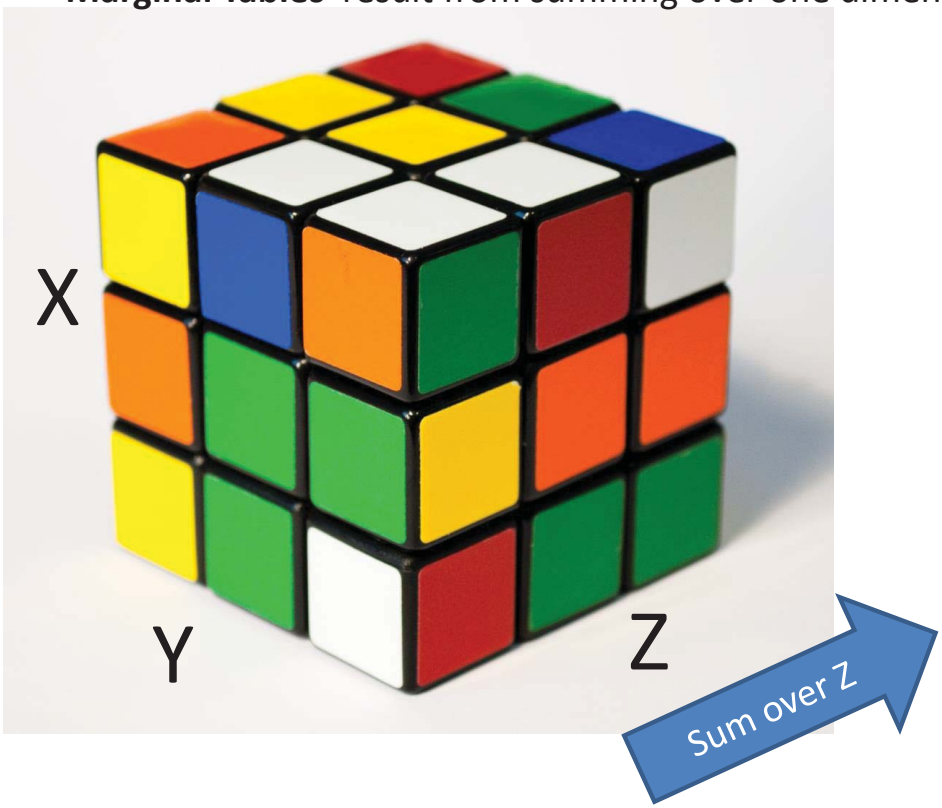
- Generally the **INDEPENDENCE MODEL, rarely holds true in real data.**
- Because generally there are row and column effects for every variable in every practical study we generally want at least a linear effect term for every variable at a minimum. Because of this the independence model is the smallest model we would ever consider.
- Any model without a three-way interaction term (or higher) λ_{ijk}^{XYZ} is called a **HOMOGENEOUS ASSOCIATION** model.
- Why are interactions like λ_{ij}^{XY} measures of **CONDITIONAL DEPENDENCE**? We'll get to that but first let's recall some blast from the past info...

Exploring Associations with Log Linear Models

RECALL: Three Way Tables FLASH FROM THE PAST

Three-Way Tables (like array's) are like Rubik's Cubes

Marginal Tables result from summing over one dimension



Marginal (XY) Table
(Summing over Z)

X

a	b	c
d	e	f
g	h	i

Y

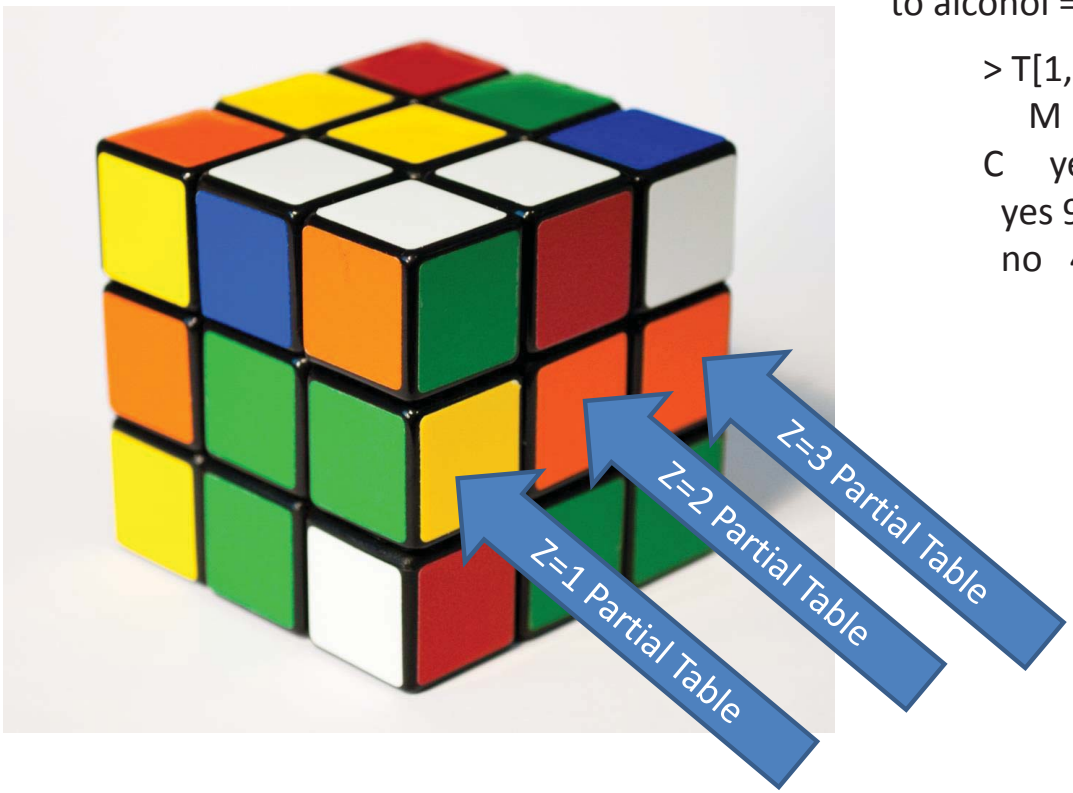
Marginal XZ table = sum over Y
Marginal YZ table = sum over X
Example:

- margin.table(T,c(1,2))

```
   C
A    yes  no
 yes 1449 500
 no   46 281
```

Exploring Associations with Log Linear Models

Partial Tables are two way tables which cross classify two variables at different levels of a third variables. Such as the two way table cross-classifying X and Y at separate levels of Z. The cross-sectional "Slices" are called **partial tables**.



Example:
The partial table corresponding to alcohol = Yes, or X ==1

```
> T[1,,]
    M
C    yes  no
 yes 911 538
 no   44 456
```

Exploring Associations with Log Linear Models

Recall Conditional Association versus Marginal Associations

Definition

A **conditional association** is an association which is found between two variables, say X and Y, at a certain value of a third variable Z. **Conditional associations are found for partial tables**.

$\theta_{XY(k)}$ is the measure of conditional association between X&Y given that Z = k (i.e. Association on the Z=k partial table "slice")

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

Definition

A **marginal association** is an association which is found between two variables, say X and Y when one sums over the other variables. **Marginal associations are found for the marginal tables**.

θ_{XY+} is the measure of marginal association between X&Y summing over Z

RECALL: Simpson's Paradox: $\theta_{XY+} \neq \theta_{XY(k)}$

Exploring Associations with Log Linear Models

Ok, What's so special about the Interaction Terms Like λ_{ij}^{XY} ?

Consider the all pairwise interaction model **(XY, XZ, YZ)**

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$$

$$\log(\theta_{XY(k)}) = \log\left(\frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}}\right) = \text{linear terms cancel} + (\lambda_{ik}^{XZ} \text{ terms cancel}) + (\lambda_{jk}^{YZ} \text{ terms cancel}) +$$

$$+\left(\lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}\right) \ldots \text{ so we see that}$$

$$\log(\theta_{XY(k)}) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$

Same value regardless of (k) index!!

Moral of the story .... λ_{ij}^{XY} Measures the **homogeneous conditional association for each of the k=K conditional Z=k "slices" of the overall contingency table.**

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

- True for the other conditional associations too:

$$\theta_{XZ(1)} = \theta_{XZ(2)} = \cdots = \theta_{XZ(J)}$$

$$\theta_{YZ(1)} = \theta_{YZ(2)} = \cdots = \theta_{YZ(I)}$$

- But the conditional associations are not in general the same: $\theta_{XY(k)} \neq \theta_{YZ(i)} \neq \theta_{XZ(j)}$

Exploring Associations with Log Linear Models

Moral of the Story

- Any model without the three-way interaction term (XYZ) is a **HOMOGENEOUS ASSOCIATION** model
- Let's now consider the model **(XY, XZ):**

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XY}$$

- Because this model has the terms λ_{ij}^{XY} and λ_{ik}^{XZ} we **still have homogeneous XY and XZ association, i.e.**

$$\log(\theta_{XY(k)}) = \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY}$$
$$\log(\theta_{XZ(j)}) = \lambda_{11}^{XZ} + \lambda_{22}^{XZ} - \lambda_{12}^{XZ} - \lambda_{21}^{XZ}$$

- But this model is **MISSING the** λ_{jk}^{YZ} **term, so**

$$\log(\theta_{YZ(i)}) = \lambda_{11}^{YZ} + \lambda_{22}^{YZ} - \lambda_{12}^{YZ} - \lambda_{21}^{YZ} = 0 \Rightarrow \theta_{YZ(i)} = 1$$
$$\theta_{YZ(1)} = \theta_{YZ(2)} = \cdots = \theta_{YZ(I)} = 1$$

- Thus the **(XY, XZ) model has homogeneous conditional XY and XZ association but the model says Y&Z are conditionally independent given X**

Exploring Associations with Log Linear Models

The Interesting Log-Linear Models

Model	Equation	Interpretation
(XYZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY} + \lambda_{ijk}^{XYZ}$	**Saturated model** **No homogeneous assoc.**
(XY,XZ,YZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$	Homogeneous Associations for every pair of variables
(XY,XZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{ij}^{XY}$	Homogeneous associations YZ conditional independence given X
(XY,YZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$	Homogeneous associations XZ conditional independence given Y
(YZ,XZ)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}$	Homogeneous associations XY conditional independence given Z
(XY,Z)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY}$	Homogeneous associations YZ & XZ conditional independence
(XZ,Y)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ}$	Homogeneous associations XY & YZ conditional independence
(YZ,X)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ}$	Homogeneous associations XY & XZ conditional independence
(X,Y,Z)	$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z$	**INDEPENDENCE MODEL!!**

Exploring Associations with Log Linear Models

Log-Linear Models for Contingency Tables

- We like non-saturated models because they are easier to interpret
- Odds ratios tell us about conditional associations
- The meaning of the model **depends on terms that are not in the model just as much as the terms that are.**
- Three way interactions are difficult to interpret. A model with 3-way interactions destroys the possibility of homogeneous associations.

LOG-LINEAR models find associations

➔ Convert parameter estimates into interesting conditional odds ratios

➔ We use hierarchical models

Caution: watch interpretation of lower order terms when in the presence of lower order terms. Generally when focusing on the meaning of a model focus on the higher order terms.

Exploring Associations with Log Linear Models

How to fit these models using loglm() and update() in R

Model		
(XYZ)	Start with the saturated model:	satmodel=loglm(~A*C*M,data=T)
(XY,XZ,YZ)	Take away 3-way interaction from sat model:	model1=update(satmodel,.~.-A:C:M)
(XY,XZ)	Take away C:M interaction from model 1:	model2=update(model1,.~.-C:M)
(XY,YZ)	Take away A:M interaction from model 1:	model3=update(model1,.~.-A:M)
(YZ,XZ)	Take away A:C interaction from model 1:	model4=update(model1,.~.-A:C)
(XY,Z)	Take away (C:M+A:M) interaction from model 1:	model5=update(model1,.~.-(C:M+A:M))
(XZ,Y)	Alternatively, build model from scratch	model6=loglm(~A*M+C,data=T)
(YZ,X)	Build model from scratch	model7=loglm(~C*M+A,data=T)
(X,Y,Z)	Independence model:	indepmodel=loglm(~A+C+M,data=T)

Exploring Associations with Log Linear Models

How to fit these models using loglin() in R

Model	
(XYZ)	satfit=loglin(T,margin=list(1:3),fit=TRUE,param=TRUE)
(XY,XZ,YZ)	fit1=loglin(T,margin=list(c(1,2),c(1,3),c(2,3)),fit=TRUE,param=TRUE)
(XY,XZ)	fit2=loglin(T,margin=list(c(1,2),c(1,3)),fit=TRUE,param=TRUE)
(XY,YZ)	fit3=loglin(T,margin=list(c(1,2),c(2,3)),fit=TRUE,param=TRUE)
(YZ,XZ)	fit4=loglin(T,margin=list(c(1,3),c(2,3)),fit=TRUE,param=TRUE)
(XY,Z)	fit5=loglin(T,margin=list(c(1,2),c(3)),fit=TRUE,param=TRUE)
(XZ,Y)	fit6=loglin(T,margin=list(c(1,3),c(2)),fit=TRUE,param=TRUE)
(YZ,X)	fit7=loglin(T,margin=list(c(2,3),1),fit=TRUE,param=TRUE)
(X,Y,Z)	indepfit=loglin(T,margin=list(1,2,3),fit=TRUE,param=TRUE)

Exploring Associations with Log Linear Models

Compare model fits

```
FIT=as.data.frame(ftable(satfit$fit,row.vars=1:3))
y1=c(fit1$fit)
y2=c(fit2$fit)
y3=c(fit3$fit)
y4=c(fit4$fit)
y5=c(fit5$fit)
y6=c(fit6$fit)
y7=c(fit7$fit)
y8=c(indepfit$fit)
FIT=cbind(FIT,y1,y2,y3,y4,y5,y6,y7,y8)
names(FIT)[4:12]<-c("(XYZ)","(XY,XZ,YZ)","(XY,XZ)","(XY,YZ)","(YZ,XZ)","(XY,Z)","(XZ,Y)","(YZ,X)","(X,Y,Z)")
```

			MODEL FIT								
A	C	M	(XYZ)	(XY,XZ,YZ)	(XY,XZ)	(XY,YZ)	(YZ,XZ)	(XY,Z)	(XZ,Y)	(YZ,X)	(X,Y,Z)
yes	yes	yes	911	910.4	710.0	885.9	909.2	611.2	627.3	782.7	540.0
no	yes	yes	3	3.6	0.7	28.1	4.8	19.4	3.3	131.3	90.6
yes	no	yes	44	44.6	245.0	29.4	45.8	210.9	327.7	39.4	282.1
no	no	yes	2	1.4	4.3	16.6	0.2	118.5	1.7	6.6	47.3
yes	yes	no	538	538.6	739.0	563.1	438.8	837.8	652.9	497.5	740.2
no	yes	no	43	42.4	45.3	17.9	142.2	26.6	211.5	83.5	124.2
yes	no	no	456	455.4	255.0	470.6	555.2	289.1	341.1	629.4	386.7
no	no	no	279	279.6	276.7	264.4	179.8	162.5	110.5	105.6	64.9

Saturated model fit = raw data

(XY,XZ,YZ) model fit is pretty close

(XZ,YZ) model fit is also pretty close

It seems like models (XY,XZ,YZ) and models (YZ,XZ) both fit the data pretty well

Exploring Associations with Log Linear Models

Illustration of Conditional Associations with (YZ,XZ) model

The fit $\hat{Y}$ of the (YZ,XZ) model can be obtained by:

```
> U=fit4$fit
> U
, , M = yes
    C
A          yes        no
 yes 909.2395833  45.7604167
 no    4.7604167   0.2395833

, , M = no
    C
A          yes        no
 yes 438.8404255 555.1595745
 no  142.1595745 179.8404255
```

The fit of a loglin object is a contingency table

The XY(1) partial table is:

```
> U[,,1]
    C
A          yes        no
 yes 909.239583 45.7604167
 no    4.760417  0.2395833
```

$$\theta_{XY(1)} = \frac{(909.24)(0.24)}{(4.76)(45.76)} = 1.0$$

The XY(2) partial table is:

```
> U[,,2]
    C
A        yes      no
 yes 438.8404 555.1596
 no  142.1596 179.8404
```

$$\theta_{XY(2)} = \frac{(438.84)(179.84)}{(142.15)(555.15)} = 1.0$$

Under Model, $\theta_{XY(1)} = \theta_{XY(2)} = 1.0 \Rightarrow \text{X \& Y are conditionally independent given Z}$

Exploring Associations with Log Linear Models

Illustration of Conditional Associations with (YZ,XZ) model

The XZ(1) partial table is:

```
> U[,1,]
    M
A       yes      no
  yes 909.239583 438.8404
  no    4.760417 142.1596
```

$$\theta_{XZ(1)} = \frac{(909.24)(142.16)}{(438.84)(4.76)} = 61.9$$

The XZ(2) partial table is:

```
> U[,2,]
    M
A       yes      no
  yes 45.7604167 555.1596
  no   0.2395833 179.8404
```

$$\theta_{XZ(2)} = \frac{(45.76)(179.84)}{(0.24)(555.15)} = 61.9$$

$$\theta_{XZ(1)} = \theta_{XZ(2)} = 61.9 \text{ (Homogeneous Conditional Association)}$$

The YZ(1) partial table is:

```
➢ U[1,,]
    M
C       yes      no
  yes 909.23958 438.8404
  no   45.76042 555.1596
```

$$\theta_{YZ(1)} = \frac{(909.24)(555.16)}{(438.84)(45.76)} = 25.1$$

The YZ(2) partial table is:

```
> U[2,,]
    M
C       yes      no
  yes 4.7604167 142.1596
  no  0.2395833 179.8404
```

$$\theta_{YZ(2)} = \frac{(4.76)(179.84)}{(142.15)(0.24)} = 25.1$$

$$\theta_{YZ(1)} = \theta_{YZ(2)} = 25.1 \text{ (Homogeneous Conditional Association)}$$

Exploring Associations with Log Linear Models

But Does Conditional Association = Marginal Association under (YZ,XZ) model?

The XY(1) partial table is:

```
> U[,,1]
    C
A       yes       no
  yes 909.239583 45.7604167
  no    4.760417  0.2395833
```

$$\theta_{XY(1)} = \frac{(909.24)(0.24)}{(4.76)(45.76)} = 1.0$$

The XY(2) partial table is:

```
> U[,,2]
    C
A       yes      no
  yes 438.8404 555.1596
  no  142.1596 179.8404
```

$$\theta_{XY(2)} = \frac{(438.84)(179.84)}{(142.15)(555.15)} = 1.0$$

The XY marginal table is:

```
> margin.table(U,c(1,2))
    C
A       yes     no
  yes 1348.08  600.92
  no   146.92  180.08
```

$$\theta_{XY(+)} = \frac{(1348.08)(180.08)}{(146.92)(600.92)} = 2.7$$

NOPE!! Simpson's Paradox still holds true

$$\theta_{XY+} \neq \theta_{XY(k)}$$

Sometimes, however **Conditional Association = Marginal Association and we will learn when this is true soon.**

Exploring Associations with Log Linear Models

Other way to compute conditional association

Fit$param returns the parameters

```
> fit4$param$A.M
     M
A          yes        no
  yes  1.031272 -1.031272
  no  -1.031272  1.031272
```

λ_{ik}^{XZ}

1 redundant parameter

Recall:

$$\log(\theta_{XZ(j)}) = \lambda_{11}^{XZ} + \lambda_{22}^{XZ} - \lambda_{12}^{XZ} - \lambda_{21}^{XZ}$$

Thus,

$$\theta_{XZ(j)} = \exp(\lambda_{11}^{XZ} + \lambda_{22}^{XZ} - \lambda_{12}^{XZ} - \lambda_{21}^{XZ})$$

$\theta_{XZ(j)} =$

```
exp(sum(abs(fit4$param$A.M)))
[1] 61.87324
```

Exploring Associations with Log Linear Models

How do you determine which model is the best model?

			MODEL FIT								
A	C	M	(XYZ)	(XY,XZ,YZ)	(XY,XZ)	(XY,YZ)	(YZ,XZ)	(XY,Z)	(XZ,Y)	(YZ,X)	(X,Y,Z)
yes	yes	yes	911	910.4	710.0	885.9	909.2	611.2	627.3	782.7	540.0
no	yes	yes	3	3.6	0.7	28.1	4.8	19.4	3.3	131.3	90.6
yes	no	yes	44	44.6	245.0	29.4	45.8	210.9	327.7	39.4	282.1
no	no	yes	2	1.4	4.3	16.6	0.2	118.5	1.7	6.6	47.3
yes	yes	no	538	538.6	739.0	563.1	438.8	837.8	652.9	497.5	740.2
no	yes	no	43	42.4	45.3	17.9	142.2	26.6	211.5	83.5	124.2
yes	no	no	456	455.4	255.0	470.6	555.2	289.1	341.1	629.4	386.7
no	no	no	279	279.6	276.7	264.4	179.8	162.5	110.5	105.6	64.9

Saturated model fit = raw data

(XY,XZ,YZ) model fit is pretty close

(XZ,YZ) model fit is also pretty close

It seemed like models (XY,XZ,YZ) and models (YZ,XZ) both fit the data pretty well

Goodness of Fit Tests are based upon either

Deviance G^2

$$G^2 = 2\sum n_{ijk} \log\left(\frac{n_{ijk}}{\hat{\mu}_{ijk}}\right)$$

Pearson's X^2

$$X^2 = \sum \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}$$

Both compare the fitted count $\hat{\mu}_{ijk}$ to the actual count n_{ijk}

Exploring Associations with Log Linear Models

How do you determine which model is the best model?

- lrt=c(satfit$lrt,fit1$lrt,fit2$lrt,fit3$lrt,fit4$lrt,fit5$lrt,fit6$lrt,fit7$lrt,indepfit$lrt)
- pearson=c(satfit$pearson,fit1$pearson,fit2$pearson,fit3$pearson,fit4$pearson, fit5$pearson,fit6$pearson,fit7$pearson,indepfit$pearson)
- df=c(satfit$df,fit1$df,fit2$df,fit3$df,fit4$df,fit5$df,fit6$df,fit7$df,indepfit$df)
- pval=1-pchisq(lrt,df)
- comparison=data.frame(G2=lrt,X2=pearson,df=df,pval=pval)
- rownames(comparison)<- c("(XYZ)","(XY,XZ,YZ)","(XY,XZ)","(XY,YZ)","(YZ,XZ)","(XY,Z)","(XZ,Y)","(YZ,X)","(X,Y,Z)")
- comparison

Model	G2	X2	df	pval
(XYZ)	0	0	0	NA
(XY,XZ,YZ)	0.37	0.40	1	0.54084
(XY,XZ)	497.37	443.76	2	0
(XY,YZ)	92.02	80.81	2	0
(YZ,XZ)	187.75	177.61	2	0
(XY,Z)	843.83	704.91	3	0
(XZ,Y)	939.56	824.16	3	0
(YZ,X)	534.21	505.60	3	0
(X,Y,Z)	1286.02	1411.39	4	0

- Models which have large G2 and X2 values for a given df indicate poor fits
- Want p-value large ➔ indicates that fit is not statistically different from sat. model

- The (XY,XZ,YZ) model is the best fit!!

Exploring Associations with Log Linear Models

How do you determine which model is the best model?

Could also compare the two models using the anova() funtion

```
> anova(model4,model1)
LR tests for hierarchical log-linear models

Model 1:
 . ~ C + A + M
Model 2:
 . ~ C + A + M

          Deviance df  Delta(Dev) Delta(df) P(> Delta(Dev)
Model 1   187.7543029  2
Model 2     0.3739859  1 187.3803170        1       0.00000
Saturated   0.0000000  0   0.3739859        1       0.54084
```

Exploring Associations with Log Linear Models

What make log-linear models so special?

Answer: **The correspondence between Log-linear and Logistic Models.**

Example: Consider the (XY,XZ,YZ) model:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$$

Suppose we consider Y=cigarette use as a Binary response variable. Let us treat the variables X and Z as explanatory variables in a **logistic regression**. In a logistic regression we are interested in:

$$Logit(P(Y=1)) = \log\left[\frac{P(Y=1)}{1-P(Y=1)}\right] = \log\left[\frac{P(Y=1 \mid X=i, Z=k)}{P(Y=2 \mid X=i, Z=k)}\right]$$

$$= \log\left[\frac{\mu_{i1k}}{\mu_{i2k}}\right] = \log(\mu_{i1k}) - \log(\mu_{i2k})$$

Plug in Log-linear formula above

$$= (\lambda + \lambda_i^X + \lambda_1^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{1k}^{YZ} + \lambda_{i1}^{XY}) - (\lambda + \lambda_i^X + \lambda_2^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{2k}^{YZ} + \lambda_{i2}^{XY})$$

$$Logit(P(Y=1)) = (\lambda_1^Y - \lambda_2^Y) + (\lambda_{i1}^{XY} - \lambda_{i2}^{XY}) + (\lambda_{1k}^{YZ} - \lambda_{2k}^{YZ})$$

Constant | Depends on X | Depends on Z

Re-parameterize: $$Logit(P(Y=1)) = \alpha + \beta_i^X + \beta_j^Z$$

Exploring Associations with Log Linear Models

The correspondence between Log-linear and Logistic Models.

So the logistic model

$$Logit(P(Y=1)) = \alpha + \beta_i^X + \beta_j^Z$$

is equivalent to the (XY,XZ,YZ) log-linear model:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$$

but it is also equivalent to the (XY,YZ) log-linear model:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{jk}^{YZ} + \lambda_{ij}^{XY}$$

This model is missing associations & dependencies among the explanatory variabibles (XZ)

- For every logistic regression model, there is an equivalent log-linear model

$$\text{Logistic Regression} \subset \text{Log-linear Regression}$$

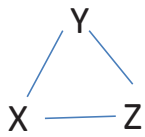
- But logistic regression models cannot explain relationships among the explanatory variables like the log-linear models can.

Exploring Associations with Log Linear Models

Independence Graphs

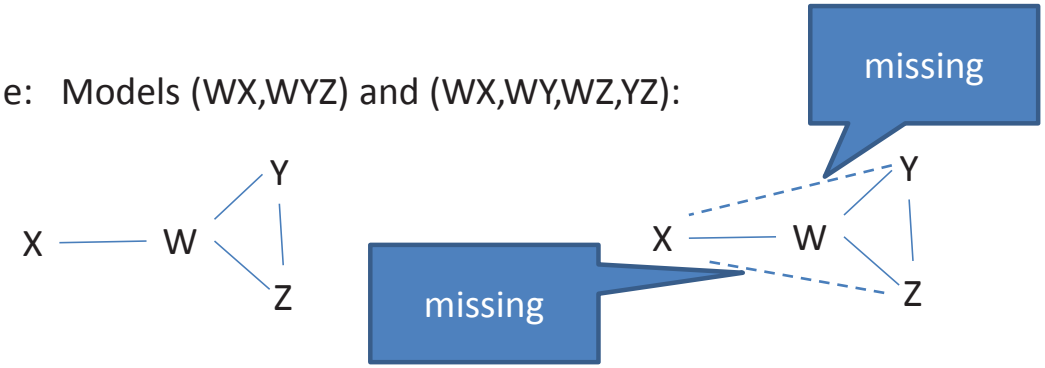
- For every log-linear model there is an associated independence graph which has a set of vertices, with each vertex representing a variable.
- Any two variables either are or are not connected by an edge (a line). A variable has an edge if there is an association between variables.
- A missing edge represents conditional independence between the two corresponding variables.

Example: Models (XYZ) and (XY,XZ,YZ):



Model (XY,YZ) : X — Y — Z

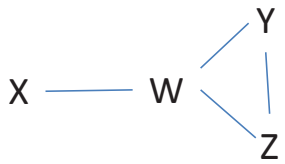
4-variable example: Models (WX,WYZ) and (WX,WY,WZ,YZ):



Exploring Associations with Log Linear Models

Independence Graphs

4-variable example: Models (WX,WYZ) and (WX,WY,WZ,YZ):



Definition

- A **path** in an independence graph is a sequence of edges leading from one variable to another.

- Two variables X and Y are said to be **separated by a subset of variables** if all paths connecting X & Y intersect that subset.

Example: W separates X &Y in the above graph
The subset {W,Z} also separates X&Y.

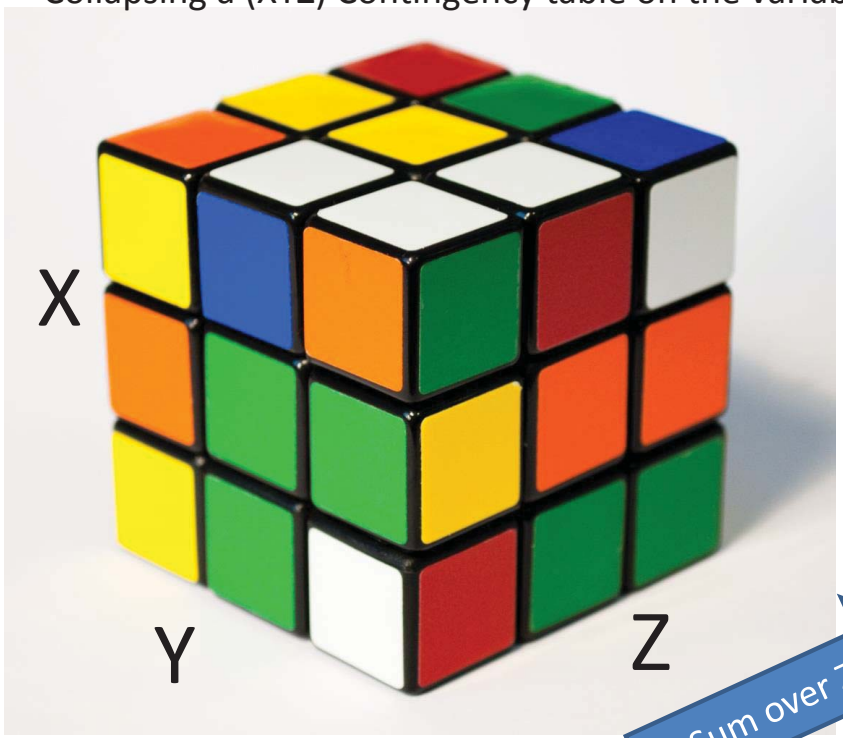
- **Two variables are conditionally independent** given any subset of variables that separates them. For example,
  X and Y are conditionally independent given W and Z.
  X and Y are also conditionally independent given W alone.

Exploring Associations with Log Linear Models

Independence Graphs and Contingency Table Collapsibility

- People want to collapse contingency tables because it makes things simpler.
- Collapsing a Multi-Way table entails summing over one of the variables to eliminate a dimension

Collapsing a (XYZ) Contingency table on the variable Z:



Marginal (XY) Table
(Summing over Z)

X	a	b	c
	d	e	f
	g	h	i
		Y	

Exploring Associations with Log Linear Models

Independence Graphs and Contingency Table Collapsibility

- A contingency table is collapsible on some variable Z, if the marginal and the conditional odds ratios for all variables (other than Z) are the same.

3-Way Table Collapsibility Conditions:

- For 3-way tables, XY marginal and conditional odds ratios are identical if EITHER X & Z are conditionally independent OR if Y&Z are conditionally independent.

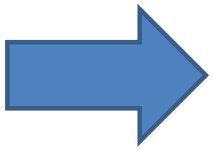
For (XY,YZ) model:

X —— Y——Z

$\theta_{XY(k)} = \theta_{XY(+)}$

$\theta_{YZ(k)} = \theta_{YZ(+)}$

But, $\theta_{XZ(k)} \neq \theta_{XZ(+)}$



Can collapse on X
Can collapse on Z

But cannot collapse on Y

Multi-Way Table Collapsibility Conditions:

Suppose that variables in a model for a multi-way table partition into Three mutually exclusive subsets, A,B, C such that B separates A from C.

A —— B——C

When we collapse the table over variables in C, the model parameters for A and the model parameters relating A & B remain unchanged.