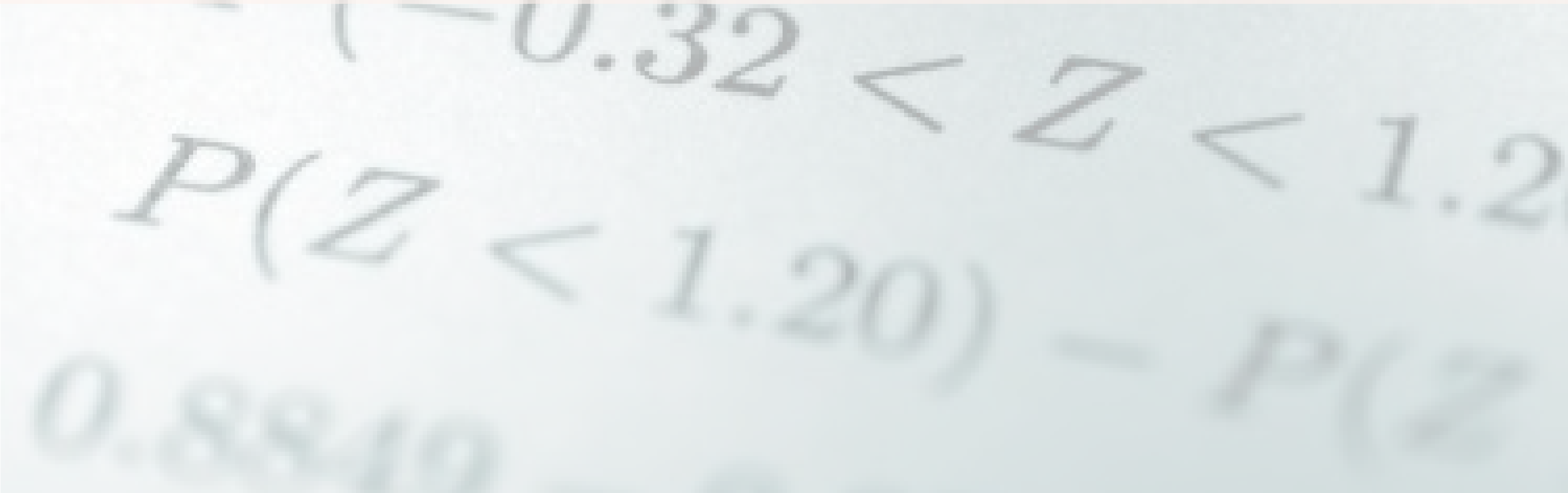
STA 4504
Categorical Data Analysis

Demetris Athienitis

University of Florida
Department of Statistics

Contents



1. Introduction

Methods for response variable whose measurement scale is a set of categories.

1.1 Categorical Response Data

The response variable will be categorical while the predictors/covariates can be either categorical/qualitative or quantitative as you have seen with regression models.

Definition 1.1 **(Categorical)** A categorical variable is one for which the measurement scale consists of a set of categories. Categorical variables can be

- Nominal: Unordered categories
- Ordinal: Ordered categories

Example 1.1 Categorical variables can be:

Nominal
- method of communication: text, call, phonetic, visual
- favorite music: rock, pop, country, indie, EDM, R&B, etc
- swipe: left, right

Ordinal
- political philosophy: liberal, moderate, conservative
- patient condition (excellent, good, fair, poor)

Remark 1.1. *Methods designed for ordinal variables utilize category ordering and thus they cannot be used for nominal variables.*

1.2 Probability Distributions for Categorical Data

For categorical data a very important distribution is the multinomial distribution of which the binomial is a special case for situations with a a binary outcome.

1.2.1 Bernoulli distribution

Imagine an experiment where the r.v. Y can take only two possible outcomes,

- success ($Y = 1$) with some probability π
- failure ($Y = 0$) with probability $1 - \pi$.

The p.m.f. of Y is

$$p(y) = \pi^y (1-\pi)^{1-y} \qquad y = 0, 1 \quad 0 \leq \pi \leq 1$$

and we denote this with $Y \sim \text{Bernoulli}(\pi)$ where $E(Y) = \pi$ and $V(Y) = \pi(1-\pi)$.

Example 1.2 A die is rolled and we are interested in whether the outcome is a 6 or not. Let,

$$Y = \begin{cases} 1 & \text{if outcome is 6} \\ 0 & \text{otherwise} \end{cases}$$

Then, $Y \sim \text{Bernoulli}(1/6)$ with mean $1/6$ and variance $5/36$.

1.2.2 Binomial distribution

If $Y_1, \ldots, Y_n$ correspond to n Bernoulli trials conducted where

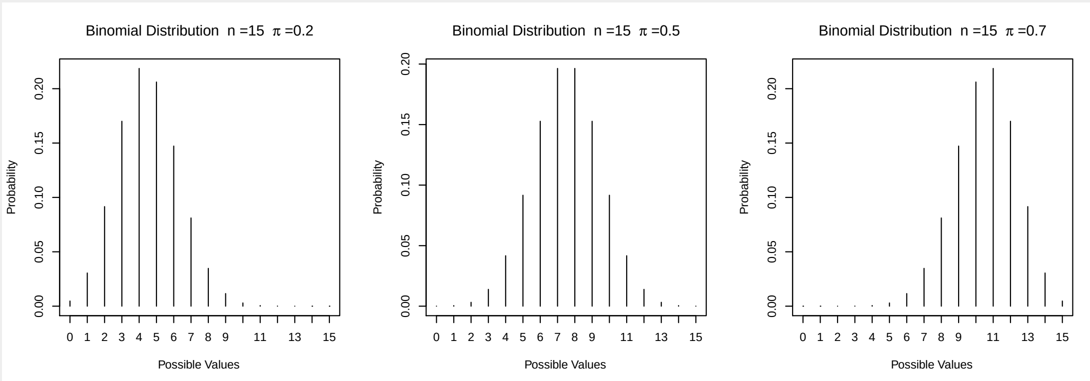
- the trials are independent
- each trial has identical probability of success π
- the r.v. Y is the total number of successes

then $Y = \sum_{i=1}^n Y_i \sim \text{Bin}(n, \pi)$ with p.m.f.

$$p(y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \quad y = 0, 1, \ldots, n$$

where $\binom{n}{y} = \frac{n!}{y!(n-y)!}$ with $E(Y) = n\pi$ and $V(Y) = n\pi(1-\pi)$. Note that ! is the "factorial" operator.

Example 1.3 The shape of 3 different binomials. Notice with $\pi = 0.5$ it is symmetric.



Example 1.4 A die is rolled 4 times and the number of 6s is observed (y).

y	$P(y)$
0	0.4823
1	0.3858
2	0.1157
3	0.0154
4	0.0008

In R, these were found using

```
dbinom(0:4,4,1/6)
```

Find the probability that there is at least one 6.

$$\begin{aligned} P(Y \geq 1) &= 1 - P(X < 1) \\ &= 1 - P(X = 0) \\ &= 0.5177 \end{aligned}$$

In R, one would simple use

```
1-pbinom(0,4,1/6)
```

Also, $E(Y) = 4(1/6) = 2/3$ and $V(Y) = 4(1/6)(5/6) = 5/9$.

Remark 1.2. Another variable of interest concerning experiments with binary outcomes is the proportion of successes $\hat{\pi} = Y/n$. Note that $\hat{\pi}$ is simply the r.v. Y multiplied by a constant, $1/n$. Hence,

$$E(\hat{\pi}) = E(Y/n) = \frac{n\pi}{n} = \pi$$

and

$$V(\hat{\pi}) = V(Y/n) = \frac{1}{n^2}V(Y) = \frac{\cancel{n}\pi(1-\pi)}{n^{\cancel{2}}} = \frac{\pi(1-\pi)}{n}$$

Remark 1.3. *Binomial distribution can be approximated by a normal distribution when* n *is large such that,* $n(\min\{\pi, 1-\pi\}) \geq 5$.

1.3 Statistical Inference for a Proportion

Parameters are often estimated using *maximum likelihood* (M.L.) That is, finding value of the parameters (of interest) that maximize the *likelihood function* or equivalently the log of the likelihood function.

Definition 1.2 (Likelihood Function) The probability of the observed data, expressed as a function of the parameter is called a likelihood function.

Definition 1.3 (MLE) The maximum likelihood estimator (MLE) is defined to be the parameter value, for which the likelihood function is maximized.

Example 1.5 Consider a widget that either works (success) or does not work (failure). Hence, if each attempt with the widget is identical and independent, the number of successes follows a $\text{Bin}(n, \pi)$.
Out of 10 attempts, 7 yielded a success. We use $\hat{\pi} = 7/10$.

$\text{Bin}(10, ?)$	$P(Y = 7)$
$\pi = 0.5$	0.1172
$\pi = 0.6$	0.2150
$\pi = 0.7$	0.2668
$\pi = 0.8$	0.2013

So by simple, but not thorough search, we saw that the outcome 7, was most "likely" if we had a $\text{Bin}(10, 0.7)$. Now lets be thorough:

1. Take the binomial p.m.f. but now treat is a function where π is the argument.

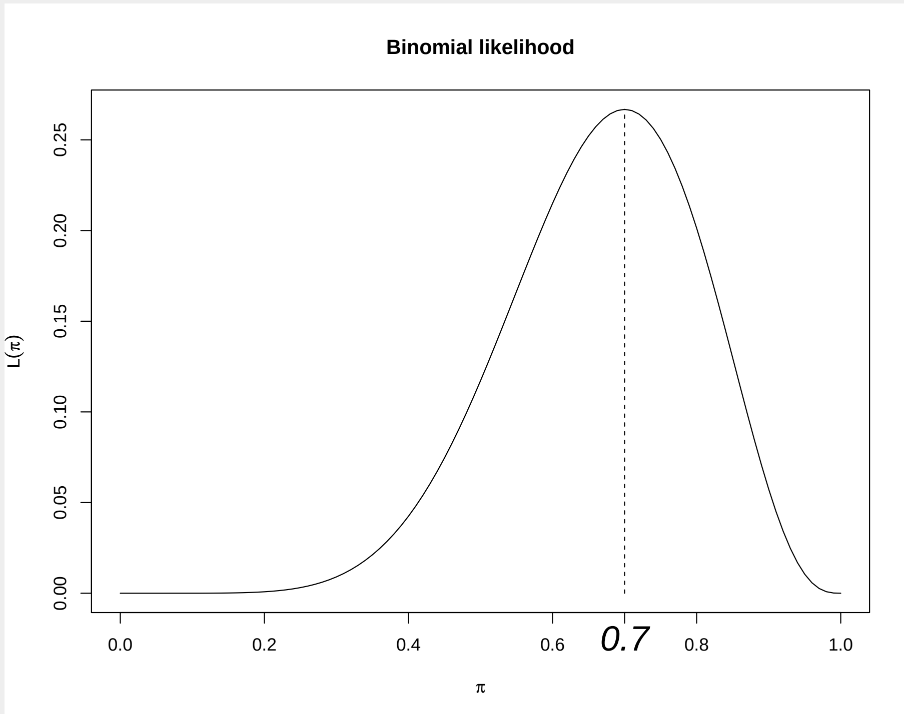
$$L(\pi) := \frac{n!}{y!(n-y)!}\pi^y(1-\pi)^{n-y}, \qquad y = 7, n = 10, \pi \in [0, 1]$$

2. To simplify lets take a look at the *log likehood*, where maximizing likehood is equivalent to maximizing log likehood.

$$l(\pi) := \log L(\pi) = \log\{n!\} - \log\{(n-y)!\} + y\log\{\pi\} + (n-y)\log\{1-\pi\}$$

3. Find maximum, take derivative and equate to 0.

$$\frac{dl(\pi)}{d\pi} = \frac{y}{\pi} - \frac{(n-y)}{1-\pi} = 0 \quad \Rightarrow \quad \hat{\pi} = \frac{y}{n} = \frac{7}{10}$$



1.3.1 Key facts

- If $y_1, y_2, \ldots, y_n$ are i.i.d. from a normal distribution, then

$$L(\mu, \sigma^2 | \boldsymbol{y}) = \prod_{i=1}^{n} f(\mu, \sigma^2 | y_i)$$

  where $f(.)$ is the p.d.f. The MLEs are then $\hat{\mu} = \bar{y}$ and $\hat{\sigma}^2 = \frac{1}{n} \sum (y_i - \bar{y})^2$

- In ordinary linear regression with Y being 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.

- Recall that $\hat{\pi}$ is *unbiased* with $E(\hat{\pi}) = \pi$ and *consistent* with $V(\hat{\pi}) \underset{n \to \infty}{\longrightarrow} 0$. MLEs are generally consistent.

- $\hat{\pi}$ is a sample mean for 0-1 data, so by the Central Limit Theorem, the sampling distribution is approximately normal for large n. Again, this is generally true for MLEs.

1.3.2 Inference Methodologies

Various significance tests exist and inverting them yields corresponding confidence intervals, values for the null hypothesis for which would fail to reject the null. Without loss of generality consider

$$\mathrm{H}_0 : \pi = \pi_0 \qquad \text{vs} \qquad \mathrm{H}_a : \pi \neq \pi_0$$

Let $p = \hat{\pi}$

Wald

Under the null,

$$TS = \frac{p - \pi_0}{\sqrt{p(1-p)/n}} \overset{\text{approx.}}{\sim} N(0,1)$$

which inverting yields the $100(1-\alpha)$ confidence interval (CI)

$$p \mp z_{1-\alpha/2}\sqrt{p(1-p)/n}$$

We fail to reject the null when

$$\left|\frac{p - \pi_0}{\sqrt{p(1-p)/n}}\right| < z_{1-\alpha/2}$$

Solving for π_0 we obtain the CI formula.

Remark 1.4. *Consider cases such as $p = 0$ or 1. Then the CI collapses to a singularity such as (0,0) or (1,1) and the CI can generally perform quite badly when n is relatively small, so other methods are advisable.*

R code 1.1 Within the "binom" package use:

```
binom.confint(y, n, conf.level = 0.95, methods = ``asymptotic'')
```

Score/Wilson

Being true to fully adopting the null hypothesis, π_0 is used in the standard error, so that under the null

$$TS = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}} \overset{\text{approx.}}{\sim} N(0,1)$$

We fail to reject the null when

$$\left|\frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}\right| < z_{1-\alpha/2}$$

Solving for π_0 requires the use of the quadratic formula and is a bit more complex and generally we let software solve for us.

R code 1.2 Within the "binom" package use:

```
binom.confint(y, n, methods = ``wilson'')
```

or with small adaptation (continuity correction) use

```
binom.confint(y, n, methods = ``prop.test'')
```

Other methods:

- Agresti-Coull, which has become the new norm.

  R code 1.3 Use
  ```
  binom.confint(y, n, methods = ``agresti-coull'')
  ```

- Clopper-Pearson a.k.a. “exact” which is recommended when n is small seeing how it is “exact”.

  R code 1.4 Use
  ```
  binom.confint(y, n, methods = ``exact'')
  ```

2. Contingency Tables

Analyzing tables involving frequency counts.

2.1 Introduction

2.1.1 Key Points

- X and Y are two categorical variables.
- X has I categories.
- Y has J categories.
- Display the IJ possible combinations of outcomes in a rectangular table having I rows for the categories of X and J columns for the categories of Y.

Definition 2.1 (Contingency table) A table of this form in which the cells contain frequency counts of outcomes is called a contingency table.

Example 2.1 (Physicians' Health Study) A study on Myocardial Infraction (MI) and treatment. We consider

- Y= heart attack: yes/no, response variable
- X= group: placebo/aspirin, explanatory variable

Group	MI	
	Yes	No
Placebo	189	10845
Aspirin	104	10933

2.1.2 Notation

- Let $\pi_{ij} = P(X = i, Y = j)$ probability that (X, Y) falls in the cell in row i and column j so that $\{\pi_{ij}\}$ form the *joint distribution* of X and Y such that

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij} = 1$$

- The *marginal distribution* of X is $\{\pi_{i+}\}$, which is obtained by $\pi_{i+} = \sum_{j=1}^{J}\pi_{ij}$. (Law of Total Probability)

- The *marginal distribution* of Y is $\{\pi_{+j}\}$, which is obtained by $\pi_{+j} = \sum_{i=1}^{I}\pi_{ij}$.

Example 2.2 In a 2×2 table.

		Y = 1	Y = 2	
X	1	π_{11}	π_{12}	π_{1+}
	2	π_{21}	π_{22}	π_{2+}
		π_{+1}	π_{+2}	1

- Similarly, let $\{n_{ij}\}, \{n_{i+}\}, \{n_{+j}\}$ denote the cell counts, row and column totals respectively.

Example 2.3 In a 2×2 table.

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

- Let

$$p_{ij} = \frac{n_{ij}}{n}$$

and

$$p_{i+} = \frac{n_{i+}}{n}, \qquad p_{+j} = \frac{n_{+j}}{n}$$

- It is informative to construct separate probability distributions for Y at each level of X. Such a distribution consists of conditional probabilities for Y given the level of X and is called a *conditional distribution*. That is,

$$\pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}} \qquad \text{estimated by} \qquad p_{j|i} = \frac{n_{ij}}{n_{i+}}$$

Example 2.4 (Physicians' Health Study ctd) Look at the probability of heart attack given the treatment group.

Group	MI		
	Yes	No	Total
Placebo	0.017	0.983	1
Aspirin	0.009	0.991	1

Remark 2.1. For many diseases there are tests to detect the disease but such tests are not fail proof. A 2×2 contingency table helps explore the effectiveness of the test. Let

- Y = *outcome of the test with* $\begin{cases} 1 & \textit{positive} \\ 2 & \textit{negative} \end{cases}.$
- X = *actual condition with* $\begin{cases} 1 & \textit{diseased} \\ 2 & \textit{not diseased} \end{cases}.$

The following two terms are important

- *Sensitivity:* $P(Y = 1|X = 1)$ *(True positive)*
- *Specificity:* $P(Y = 2|X = 2)$ *(True negative)*

2.1.3 Independence

Definition 2.2 (Independence) Variables X and Y are statistically independent if the true conditional distribution of Y is the same at each level of X.

That is,

$$\pi_{j|i} = \pi_{j|i'} \qquad \forall i, i'$$

and as a consequence

Lemma 2.1 X and Y are independent if and only if

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

Example 2.5 In a 2×2 table.

$$\begin{array}{cc|c|c|c} & & \multicolumn{2}{c}{Y} & \\ & & 1 & 2 & \\ \hline X & 1 & .42 & .28 & .7 \\ & 2 & .18 & .12 & .3 \\ \hline & & .6 & .4 & 1 \end{array}$$

2.2 Comparing Proportions in 2×2 Tables

Consider the conditional distributions, as in example 2.4, simplifying notation by using $\pi_i = \pi_{1|i}$.

		Y 1	2
X	1	π_1	$1-\pi_1$
	2	π_2	$1-\pi_2$

and interested in performing inference, on whether $\pi_1 = \pi_2$.
Before we begin we need to use

Lemma 2.2 (Delta Method) Assume that T_n is a statistic based on the data and θ is the parameter which T_n is trying to target such that

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

For a continuous function $g(\cdot)$, the asymptotic distribution of $g(T_n)$ is

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

by Taylor series expansion where $\sqrt{n}(g(T_n) - g(\theta)) \approx \sqrt{n}(T_n - \theta)g'(\theta)$

1. We can use what we learned in an introductory statistics course assuming the two levels of X are independent

$$p_1 - p_2 \mp z_{1-\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_{1+}} + \frac{p_2(1-p_2)}{n_{2+}}}$$

Example 2.6 From example 2.4, a 95% C.I. for $\pi_1 - \pi_2$

$$0.017 - 0.009 \mp 1.96\sqrt{\frac{0.017(0.983)}{11034} + \frac{0.009(0.991)}{11037}} \to (0.005, 0.011)$$

Note that if 0 was in the C.I. that would imply independence.

2. Another concept is

Definition 2.3 (Relative Risk) Relative Risk (R.R.) is defined as

$$R.R. = \frac{\pi_1}{\pi_2}$$

Example 2.7 From example 2.4, R.R.=1.82. Hence, the sample proportion of heart attacks was 82% higher for placebo group.

Note that $\log(R.R.) = \log(\pi_1) - \log(\pi_2)$ and the Delta Method allows us to find an asymptotic normal distribution for each $\log(\pi_i)$, and the linear combination of two asymptotic normal is still a normal. Therefore, a $100(1-\alpha)$ C.I. on $\log(\pi_1/\pi_2)$ is

$$\log\left(\frac{p_1}{p_2}\right) \mp z_{1-\alpha/2}\sqrt{\frac{1-p_1}{n_{1+}p_1} + \frac{1-p_2}{n_{2+}p_2}} \rightarrow (L, U)$$

and $100(1-\alpha)$ C.I. on π_1/π_2 is (e^L, e^U).

Example 2.8 From example 2.4, a 95% C.I. for $\log(\pi_1/\pi_2)$ ends up being $(0.3571, 0.8406)$ and hence for π_1/π_2

$$\left(e^{0.3571}, e^{0.8406}\right) \rightarrow (1.43, 2.31)$$

Note that if 1 was in the C.I. that would imply independence.

3. If we let redefine $Y = 1$ as a success and $Y = 2$ as a failure, the odds of success are

$$\text{odds(S)} = \begin{cases} \frac{\pi_1}{1-\pi_1} & X = 1 \\ \frac{\pi_2}{1-\pi_2} & X = 2 \end{cases}$$

Definition 2.4 (Odds Ratio) The Odds Ratio (O.R.) is the ratio of the odds of $Y = 1|X = 1$ to that of $Y = 1|X = 2$.

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

Example 2.9 From example 2.4,

$$\hat{\theta} = \frac{0.0171/0.9829}{0.0094/0.9906} = 1.83$$

The estimated odds of heart attack in placebo group are 1.83 times the odds of heart attack in the aspirin group.

Using the Delta Method the $100(1-\alpha)$ C.I. on $\log(\theta)$ is

$$\log\left(\hat{\theta}\right) \mp z_{1-\alpha/2}\sqrt{\frac{1}{n_{11}} + \frac{1}{n_{12}} + \frac{1}{n_{21}} + \frac{1}{n_{22}}} \rightarrow (L, U)$$

and $100(1-\alpha)$ C.I. on θ is (e^L, e^U).

Example 2.10 From example 2.4, a 95% C.I. for $\log(\theta)$

$$\log(1.83) \mp 1.96\sqrt{1/189 + 1/10845 + 1/104 + 1/10933} \rightarrow (0.365, 0.846)$$

and hence for θ, (1.44,2.33).

Properties:

- If $1 < \theta < \infty$, the odds of success are *higher* in row 1 than in row 2.
- If $0 < \theta < 1$, a success is *less* likely in row 1 than in row 2.
- $\theta = 1 \Leftrightarrow \log(\theta) = 0$. This also implies $\pi_1 = \pi_2$, hence independence.
- If rows are interchanged (or columns, but not both), $\theta \to 1/\theta$
- $\hat{\theta} = \frac{n_{11}n_{22}}{n_{12}n_{21}}$
- O.R. is valid for retrospective studies while R.R. and differencing are not. In retrospective studies, sampling is done within levels of Y, not to Y, and we cannot estimate $P(Y|X)$. We can estimate $P(X|Y)$ and hence θ, as θ treats rows and columns symmetrically.

$$\begin{aligned}\theta &= \frac{P(X=1|Y=1)/P(X=2|Y=1)}{P(X=1|Y=2)/P(X=2|Y=2)} \\ &= \cdots \\ &= \frac{P(Y=1|X=1)/P(Y=2|X=1)}{P(Y=1|X=2)/P(Y=2|X=2)}\end{aligned}$$

Example 2.11 (Case-control study in London Hospitals (Doll and Hill 1950))
Let,

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

$Y =$ 1 for lung cancer, 0 otherwise

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

This is a case-control study because the presence of lung cancer is considered "rare" so they found 709 individuals without lung cancer and then (using records) found 709 with lung cancer, and *then* looked at whether they smoked or not.

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

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

Remark 2.2.

- *When any values $n_{ij} \approx 0$, it is best to use $\{n_{ij} + 0.5\}$*
- *When π_1 and π_2 are close to zero then O.R. $\approx$ R.R.*

2.3 Testing Independence

To test whether X and Y we refer back to Lemma 2.1 that $\pi_{ij} = \pi_{i+}\pi_{+j}$. With any multinomial we have that the expected frequency of a cell is

$$\begin{aligned} \mu_{ij} &= n\pi_{ij} \\ &= n\pi_{i+}\pi_{+j} && \text{by ind.} \end{aligned}$$

The MLEs under the null of the expected frequencies are

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

2.3.1 Pearson Test

Testing

$$\mathrm{H}_0 : \mu_{ij} = \mu_{ij}^0 \overset{\text{ind.}}{=} \frac{n_{i+}n_{+j}}{n}, \quad \forall i, j$$

The *Pearson chi-square test statistic* under the assumption that $\hat{\mu}_{ij} > 5 \forall i, j$ is asymptotically

$$X^2 = \sum_{ij} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \underset{\text{approx.}}{\overset{\mathrm{H}_0}{\sim}} \chi^2_{(I-1)(J-1)} \tag{2.1}$$

with p-value $P\left(\chi^2_{(I-1)(J-1)} \geq X^2\right) < \alpha$ (the area to the right of the test statistic is less that α). More on the degrees of freedom later in equation (2.3).

Example 2.12 (Job Satisfaction) 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

http://users.stat.ufl.edu/~athienit/STA4504/Examples/job_sat.R

```
> job_test=chisq.test(job)
>job_test
```

```
data:  job
X-squared = 11.524, df = 9, p-value = 0.2415

Warning message:
In chisq.test(job) : Chi-squared approximation may be incorrect
```

Note that when we run the test we obtain a warning because many expected frequencies are < 5.

```
> round(job_test$expected,2)
        Dissat Little Moderate Very
<5        0.85   2.96    13.33 4.87
5k-15k    1.31   4.58    20.60 7.52
15k-25k   0.92   3.23    14.54 5.31
>25k      0.92   3.23    14.54 5.31
```

2.3.2 Likelihood-Ratio Test

The likelihood-ratio

$$\Lambda = \frac{\text{maximum likelihood when } H_0 \text{ is true}}{\text{maximum likelihood when parameters are unrestricted}}$$

Consider

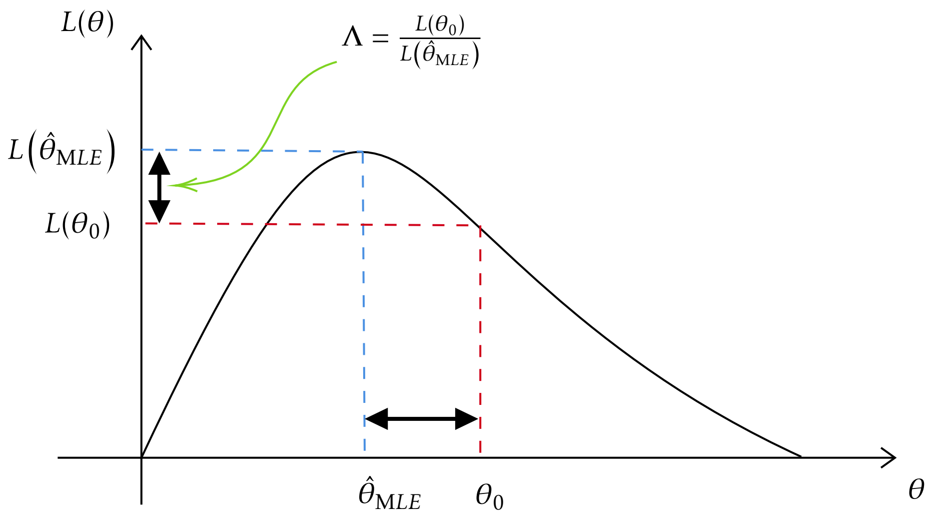
$$H_0 : \theta \in \Theta_0 \quad \text{vs} \quad H_1 : \theta \in \Theta_1$$

the likelihood ratio is given by

$$\Lambda = \frac{\sup_{\theta \in \Theta_0} L(\theta)}{\sup_{\theta \in (\Theta_0 \cup \Theta_1)} L(\theta)}$$

So if the ratio is close to 1 it implies that the estimated parameter(s) under the null are close in proximity to the unrestricted MLEs and hence null is plausible.

For example, assume we wish to test $\mathrm{H}_0 : \theta = \theta_0$. To determine if the null value θ_0 is plausible we will compare it to the maximum likelihood estimate $\hat{\theta}_{\mathrm{MLE}}$, by seeing how close the likelihood functions are at θ_0 and $\hat{\theta}_{\mathrm{MLE}}$.



The *likelihood ratio test (LRT) statistic* is asymptotically

$$G^2 = -2\log\Lambda \underset{\text{approx.}}{\overset{\mathrm{H}_0}{\sim}} \chi^2_{df} \tag{2.2}$$

and

$$\text{degrees of freedom} = \text{no. of parameters in general} - \text{no. of parameters under } \mathrm{H}_0 \tag{2.3}$$

Recall that multinomial pdf/likelihood function for an $I \times J$ table is

$$L(\pi_{ij}; n_{ij}) = \frac{n!}{n_{11}! \cdots n_{IJ}!} \pi_{11}^{n_{11}} \cdots \pi_{IJ}^{n_{IJ}}$$

Hence for a two-way contingency table and working with multinomials we have

$$\Lambda = \frac{\left(\frac{n_{i+}n_{+j}}{n^2}\right)^{n_{ij}}}{\left(\frac{n_{ij}}{n}\right)^{n_{ij}}}$$

We can ignore the constants up from since the play no role when maximizing. Recall $\hat{\mu}_{ij} = (n_{i+}n_{+j})/n$, so equation (2.2) becomes

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

and the df in equation (2.3)

- In general, there are IJ groupings in the multinomial with IJ, π_{ij}'s, hence $IJ - 1$ free parameters in general.
- Under H_0, $I - 1$ free π_{i+}'s and $J - 1$ free π_{+j}'s

and hence

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

Example 2.13 (Job Satisfaction continued) Performing the likelihood ratio test
http://users.stat.ufl.edu/~athienit/STA4504/Examples/job_sat.R

```
> library(DescTools)
> GTest(job)

data:  job
G = 13.467, X-squared df = 9, p-value = 0.1426
```

Remark 2.3.

- *As* $n \to \infty, X^2 \xrightarrow{d} \chi^2$ *faster than* $G^2 \xrightarrow{d} \chi^2$*, but they are usually similar and asymptotically equivalent, i.e.* $X^2 - G^2 \xrightarrow{d} 0$
- *These tests treat X and Y as nominal and reordering rows or columns has no effect. Methods for ordinal tests (section 2.5 of textbook) do exist.*

Once dependence is established, of interest is to determine which cells in the contingency table have higher or lower frequencies than expected (under independence). This is usually determined by observing the *standardized residuals* (deviations) of the observed counts, n_{ij}, to the expected counts $\hat{\mu}_{ij}$

Definition 2.5 (Standardized/Adjusted Residuals)

$$r_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}(1 - p_{i+})(1 - p_{+j})}}$$

which under H_0 behaves similar to $N(0,1)$. Hence, values exceeding 2 are indication of a lack of fit of H_0. Also, note the sign of the residual which describes the nature of the association.

Example 2.14 (Job Satisfaction continued) Residuals are:
http://users.stat.ufl.edu/~athienit/STA4504/Examples/job_sat.R

```
> round(job_test$stdres,4)
         Dissat  Little Moderate    Very
<5       1.4406  0.7305  -0.1606 -1.0792
5k-15k   0.7525  0.8716   0.6005 -1.7726
15k-25k -1.1171 -1.5211   0.2198  1.5098
>25k    -1.1171 -0.1574  -0.7327  1.5098
```

Remark 2.4. *The unstandardized (Pearson) residual is*

$$e_{ij} = \frac{n_{ij} - \hat{\mu}_{ij}}{\sqrt{\hat{\mu}_{ij}}}$$

tends to have a variance that is smaller than 1. Note that,

$$X^2 = \sum_{ij} e_{ij}^2$$

The deviance *residual that corresponds to G^2 is not discussed in this class.*

2.3.3 Partitioning Chi-squared

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

Lemma 2.3 Let $\chi^2_{\nu_1}$ and $\chi^2_{\nu_2}$ be independent. Then,

$$\chi^2_{\nu_1} + \chi^2_{\nu_2} \sim \chi^2_{\nu_1+\nu_2}$$

The G^2 statistic for testing independence can be partitioned into components representing certain aspects of the association. We refer the reader to the textbook for specifics.

Example 2.15 Consider the following data from a survey.

	Democrat	Independent	Republican
F	279	73	225
M	165	47	191

We have $G^2 = 7$ with $df = 2$. However the table can be partitioned into two tables

	Democrat	Independent
F	279	73
M	165	47

With $G^2 = 0.16$ and $df = 1$.

	Dem. and Ind.	Republican
F	352	225
M	212	191

With $G^2 = 6.84$ and $df = 1$.

Example 2.16 Consider example 2.12, with $G^2 = 13.47$ with $df = 9$ but partinioned as

Income	Job Satisfaction					
	Dissat	Little	Moderate	Very	G^2	df
Low					0.30	3
< 5k	2	4	13	3		
5k - 15k	2	6	22	4		
High					1.19	3
15k - 25k	0	1	15	8		
> 25k	0	3	13	8		
Low vs High					11.98	3
$< 15k$	4	10	35	7		
$> 15k$	0	4	28	16		
					13.47	9

2.3.4 Exact Inference

In this section we take a look at *Fisher's Exact Test* that does not implement an asymptotic distribution. It is exact for any sample size. It was first created and used for 2×2 tables but has since been extended.

With $\text{H}_0 : X, Y$ independent $\Leftrightarrow \theta = 1$ (odds ratio =1)

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

and treating the row and column totals as fixed, the exact null distribution of $\{n_{ij} | n_{1+}, n_{2+}, n_{+1}, n_{+2}\}$ is the *hypergeometric distribution*. In the 2×2 case the value of n_{11} completely determines the other 3 cells (since marginals are fixed).

$$
p(n_{11}) = \frac{\binom{n_{1+}}{n_{11}} \binom{n_{2+}}{n_{+1} - n_{11}}}{\binom{n}{n_{+1}}}, \qquad n_{11} \in \{\max(0, n_{+1} + n_{1+} - n), \ldots, \min(n_{+1}, n_{1+})\}
$$

The p-value is the sum of the hypergeometric probabilities for outcomes at least as favorable to the alternative hypothesis as the observed outcome.

Example 2.17 (Tea Testing) 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	Milk	3	1	4
	Tea	1	3	4
		4	4	8

Based on the marginals it is possible for $n_{11} = 0, 1, 2, 3, 4$ (not always the case).

R code 2.1 With software,

```
> cbind(0:4,dhyper(0:4,4,4,4))
     [,1]       [,2]
[1,]    0 0.01428571
[2,]    1 0.22857143
[3,]    2 0.51428571
[4,]    3 0.22857143
[5,]    4 0.01428571
```

To test

$$H_0 : \theta \leq 1 \quad \text{vs.} \quad H_a : \theta > 1$$

where the alternative is indicating that the lady can correctly guess better than simply guessing by chance, the p-value is thus

$$P(n_{11} \geq 3) = 0.243$$

With software,

```
> TeaTasting=matrix(c(3,1,1,3),2,2,byrow=T,
+ dimnames=list(Truth=c("Milk","Tea"),Guess=c("Milk","Tea")))
> TeaTasting
Guess
Truth Milk Tea
Milk 3 1
Tea  1 3

> fisher.test(TeaTasting,alternative="greater")
data: TeaTasting
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
0.3135693 Inf
sample estimates:
odds ratio
6.408309
```

The odds ratio in `fisher.test` is the ML odds ratio, not the unconditional one traditionally taught ($\hat{\theta} = 9$)

Example 2.18 (Job Satisfaction continued) For larger than 2×2 tables, In R

```
> fisher.test(job)
Fisher's Exact Test for Count Data

data:  job
p-value = 0.2315
alternative hypothesis: two.sided
```

Remark 2.5. For tables with ordinal variables please refer to "Analysis of Ordinal Data" by Alan Agresti. In addition, some methods you can review are:

- *section 2.5.1 of our textbook*
- *Goodman's gamma*
- *Kendall's tau b*

2.4 Three-Way Contingency Tables

2.4.1 Odds Ratios

Extending to three variables the goal is to examine the relationship between X and Y controlling (if significant) for a third variable Z.

Example 2.19 (Death Penalty) A $2 \times 2 \times 2$ table from data from Florida 1976-1987.

		Death Penalty		
Victim's Race	Defendant's Race	Yes	No	Percentage Yes
White	White	53	414	11.3
	Black	11	37	22.9
Black	White	0	16	0.0
	Black	4	139	2.8
Total	White	53	430	11.0
	Black	15	176	7.9

Let

- Y be the response whether they receive death penalty
- X be the defendant's race
- Z be the victim's race

The eastimated conditional odds ratios are

- Z = white, $\hat{\theta}_{XY(1)} = \frac{53\times37}{414\times11} = 0.43$ (0.42 after adding 0.5 to each cell)
- Z = black, $\hat{\theta}_{XY(2)} = \frac{0\times139}{16\times4} = 0$ (0.94 after adding 0.5 to each cell)

Controlling for victim's race, odds of receiving death penalty were lower for white defendants than for black defendants.
Ignoring victim's race, odds of death penalty higher for white defendants as

$$\hat{\theta}_{XY} = \frac{53\times176}{15\times430} = 1.45$$

This is an example of *Simpson's Paradox*.

Definition 2.6 (Simpson's paradox) When a marginal association can have different direction from the conditional associations is this is called *Simpson's paradox*.

Definition 2.7 (Conditional Independence) Variables X and Y are conditionally independent given Z if they are independent in each conditional table.

In a $2\times2\times K$ table this means

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

The converse however does not apply, as shown in the following example

Example 2.20 Data from clinical treatment yield

		Response		
Clinic	Treatment	Success	Failure	$\hat{\theta}$
1	A	18	12	1.0
	B	12	8	
2	A	2	8	1.0
	B	8	32	
Total	A	20	20	2.0
	B	20	40	

This also acts as an example of a symmetric property know as

Definition 2.8 (Homogeneous Association) A *homogeneous association* exists if the conditional odds ratios between X and Y are identical at all levels of Z.

2.4.2 Cochran-Mantel-Haenszel Test

The Cochran-Mantel-Haenszel (CMH) Test is used on $2\times2\times K$ tables to test

H_0: X and Y are conditionally independent given Z, i.e. $\theta_{XY(1)} = \cdots = \theta_{XY(K)} = 1$

Similar to Fisher's Exact Test, in the k-th partial table, the row totals are n_{1+k}, n_{2+k} and column totals are n_{+1k}, n_{+2k}. Given both these totals, n_{11k} has a hypergeometric distribution and that determines all other cell counts in the k-th partial table.

$$CMH = \frac{\left[\sum_{k=1}^{K} (n_{11k} - E(n_{11k}))\right]^2}{\sum_{k=1}^{K} V(n_{11k})} \underset{\text{approx.}}{\overset{\text{H}_0}{\sim}} \chi_1^2$$

where under independence,

$$E(n_{11k}) = \frac{n_{1+k} n_{+1k}}{n}$$
$$V(n_{11k}) = \frac{n_{1+k} n_{2+k} n_{+1k} n_{+2k}}{{n_{++k}}^2 (n_{++k} - 1)}$$

The Mantel-Haenszel estimator of that common value equals

$$\hat{\theta}_{MH} = \frac{\sum_{k=1}^{K} (n_{11k} n_{22k} / n_{++k})}{\sum_{k=1}^{K} (n_{12k} n_{21k} / n_{++k})}$$

The Delta Method can implemented to obtain the standard error of the $\log\left(\hat{\theta}_{MH}\right)$ but those calculations are ommitted here.

Remark 2.6.

- *This test is inappropriate when the association varies widely among the partial tables.*
- *If the true odds ratios are not identical but do not vary drastically, $\hat{\theta}_{MH}$ still provides a useful summary of the K conditional associations, i.e. the K conditional odds ratios.*

Example 2.21 Consider a $2 \times 2 \times 5$ table

```
> MIOC
, , Agegrp = 1
         OCuse
Status    Yes  No
  Case      4   2
  Control  62 224

, , Agegrp = 2
         OCuse
Status    Yes  No
  Case      9  12
  Control  33 390

, , Agegrp = 3
```

```
         OCuse
Status    Yes  No
  Case      4  33
  Control  26 330

, , Agegrp = 4
         OCuse
Status    Yes  No
  Case      6  65
  Control   9 362

, , Agegrp = 5
         OCuse
Status    Yes  No
  Case      6  93
  Control   5 301

> OR=function(matrix,adjust=TRUE){
+   if(adjust==TRUE){mat=matrix+0.5}
+   OR=(mat[1,1]*mat[2,2])/(mat[1,2]*mat[2,1])
+   return(OR)
+ }
> apply(MIOC,3,OR)
       1        2        3        4        5
6.465600 8.859104 1.675303 3.786661 3.810890
```

Since the five sample odds ratios do not very “drastically” we can proceed with the CMH test

```
> mantelhaen.test(MIOC)

Mantel-Haenszel chi-squared test with continuity correction

data:  MIOC
Mantel-Haenszel X-squared = 32.793, df = 1, p-value = 1.025e-08
alternative hypothesis: true common odds ratio is not equal to 1
95 percent confidence interval:
 2.426983 6.493688
sample estimates:
common odds ratio
         3.969895
```

http://users.stat.ufl.edu/~athienit/STA4504/Examples/CMH.R

Remark 2.7. The Breslow-Day Test also exists for testing homogeneity of odds ratios, not just for conditional independence.

3. Generalized Linear Models

Using models as the basis for analyzing associations, which can describe effects in more informative ways.

3.1 Components of a Generalized Linear Model (GLM)

1. **Random component**: Identifies the response variable Y and assumes a probability distribution for it. We will assume independent observations from the *exponential family* of distributions. We will primarily be looking at binomial and Poisson, but note that the Gaussian also falls in this family as do most of the "common" distributions.

2. **Systematic component**: Specifies the explanatory variables $(x_1, \dots, x_k)$ used as predictors in the model using a linear function of coefficients known as the *linear predictors*

$$\alpha + \beta_1 x_1 + \cdots + \beta_k x_k$$

3. **Link**: Describes the functional relation between the systematic component and expected value of the random component. It specifies how $\mu = E(Y)$ relates to explanatory variables in the linear predictor.

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

   The function $g(\cdot)$ is called the *link* function.

More about link functions

- Each potential probability distribution has one special function of the mean that is called its *natural parameter*. The link function that uses the natural parameter as $g(\mu)$ in the GLM is called the *canonical link*. (The benefit of using the canonical link is that the expected fisher-information matrix is the same as the observed matrix.)

- For the normal distribution, it is mean itself, i.e. identity link.

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

- For the Poisson, the natural parameter is the log of the mean. (Recall $\mu = \lambda$.)

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

- For the Bernoulli, the natural parameter is the logit of the mean. (Recall $\mu = \pi$.)

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

3.2 GLM for Binary Data

The distribution of a binary response is specified by probabilities

$$P(Y=1) = \pi \quad \text{and} \quad P(Y=0) = 1 - \pi$$

and for n independent and identical trails we end up with a binomial distribution.

3.2.1 Linear Probability Model

For simplicity, consider a single predictor x. Using an identity link,

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

For such a model probabilities may fall between 0 and 1 but for large or small enough values of x, the model may predict $\pi(x) < 0$ or $\pi(x) > 1$. Hence, this model is valid only for a finite range of predictor values. As such other links shall be used, such as *logit* and *probit*.

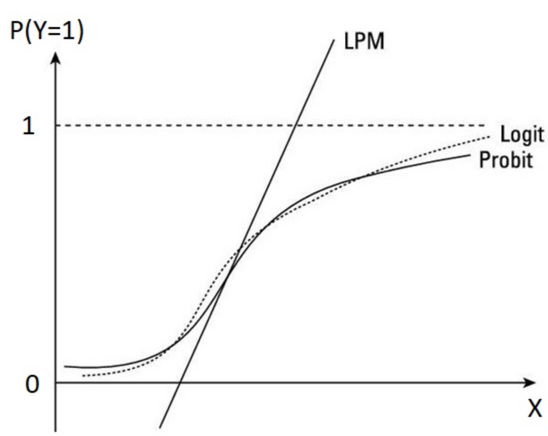


Figure 3.1: An example of a model with identity, logit and probit links

3.2.2 Logistic Regression Model

Using the logit link,

$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x \quad \Rightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} \tag{3.1}$$

That is,

$$\pi(x) = F_0(\alpha + \beta x) \quad \Rightarrow \quad F_0^{-1}(\pi(x)) = \alpha + \beta x$$

where

$$F_0(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

is the (standard) cdf of the *logistic* distribution. That is, the link function is the logistic's distribution quantile function (which is also the canonical link)

$$g(\cdot) \equiv F_0^{-1}(\cdot)$$

guaranteeing that $0 \leq \pi(x) \leq 1$. Although logistic regression will be covered more in depth in the next chapter some key points are:

- The parameter β determines the rate of increase or decrease of the curve and the magnitude of β determines how fast the curve increases or decreases.
- When $\beta > 0$, $\pi(x)$ increases as x increases.
- When $\beta < 0$, $\pi(x)$ decreases as x increases.

In the next chapter we will see that the $100(1-\alpha)\%$ C.I. on β is

$$\hat{\beta} \mp z_{1-\alpha/2}\left(s_{\hat{\beta}}\right)$$

where the estimate and standard error are provided by the software.

R code 3.1 A GLM is fitted using the

```
model=glm(formula,data,family)
```

where the `family` argument will specify the random component as well as the link function. Basic output is provided with `summary(model)` and CI created on the coefficients via `confint(model)`.
For a logistic regression, take for example

- When the response column is `y` is 0 or 1, use

```
glm(y~x,family=binomial,data=mydata)
```

- When there is a column grouping successes and one grouping failures, use

```
glm(cbind(Successes,Failures)~x,family=binomial,data=mydata)
```

Please see the `help(glm)` help file.

Alternative link: Probit Just as the logistic regression model utilized the logistic's distribution quantile function, an alternative is quantile function of the (standard) normal distribution

$$g(\cdot) \equiv \Phi^{-1}(\cdot)$$

which implies

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

The probit transform maps $\pi(x)$ so that the regression curve for $\pi(x)$ (or $1-\pi(x)$, when $\beta < 0$) has the appearance of the normal cdf with mean $\mu = -\alpha/\beta$ and standard deviation $\sigma = 1/|\beta|$.

Example 3.1 (Infant Malformation) A study was conducted about infant sex organ malformation and pregnant mother's alcohol consumption.

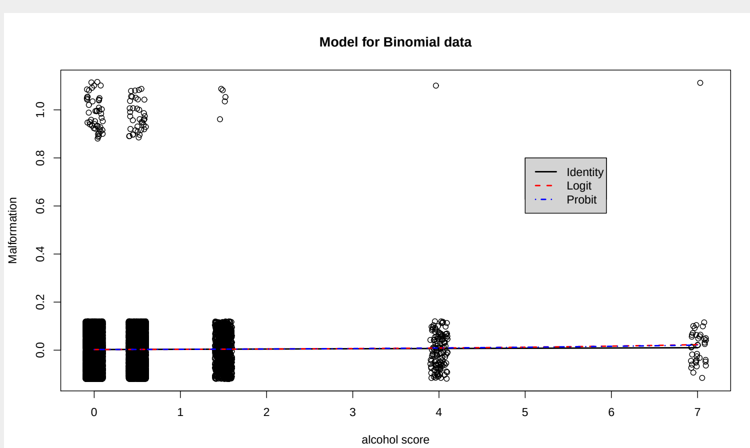
- Y = infant sex organ malformation (1 = present, 0 = absent)
- x = mother's alcohol consumption (avg drinks per day)

Consumption		Malformation	
Measured	Score	Absent	Present
0	0.0	17066	48
<1	0.5	14464	38
1-2	1.5	788	5
3-5	4.0	126	1
≥ 6	7.0	37	1

```
> malform.logit=glm(cbind(Present,Absent)~Alcohol,
+  family=binomial(link=logit))
> summary(malform.logit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9605     0.1154 -51.637   <2e-16 ***
Alcohol       0.3166     0.1254   2.523   0.0116 *
---
    Null deviance: 6.2020  on 4  degrees of freedom
Residual deviance: 1.9487  on 3  degrees of freedom
AIC: 24.576
```

The logistic regression model is

$$\text{logit}\,[\hat{\pi}(x)] = -5.9605 + 0.3166(\text{Score})$$

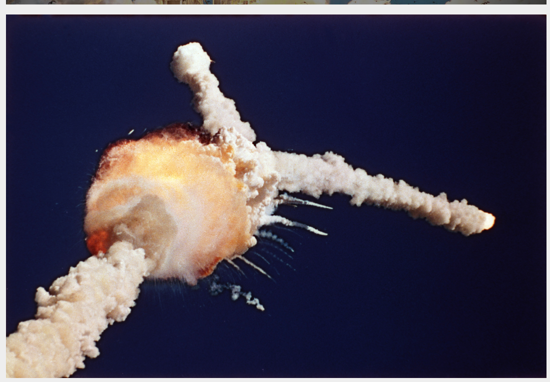


Note that in this example both the logistic and the linear model appear to be good fits. This is because whenever you "zoom" into to a part of a curve a linear relationship is adequate.
http://users.stat.ufl.edu/~athienit/STA4504/Examples/malformation.R

Remark 3.1. *If a logistic regression model is deemed an adequate fit then so will a probit model be deemed, i.e. when one is a good fit then so will the other, as seen in figure* 3.1

Example 3.2 (Challenger disaster) For the 23 space shuttle flights that occurred before the Challenger mission disaster in 1986, the data shows the temperature at the time of flight and whether at least one primary O-ring suffered thermal distress.

Flight	Temp	Failure
1	66	0
2	70	1
3	69	0
4	68	0
5	67	0
6	72	0
7	73	0
8	70	0
9	57	1
10	63	1
11	70	1
12	78	0
13	67	0
14	53	1
15	67	0
16	75	0
17	70	0
18	81	0
19	76	0
20	79	0
21	75	1
22	76	0
23	58	1



```
> preC.logit=glm(Failure~Temp,family=binomial(link=logit),data=preC)
> summary(preC.logit)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  15.0429     7.3786   2.039   0.0415 *
Temp         -0.2322     0.1082  -2.145   0.0320 *
---
    Null deviance: 28.267  on 22  degrees of freedom
```

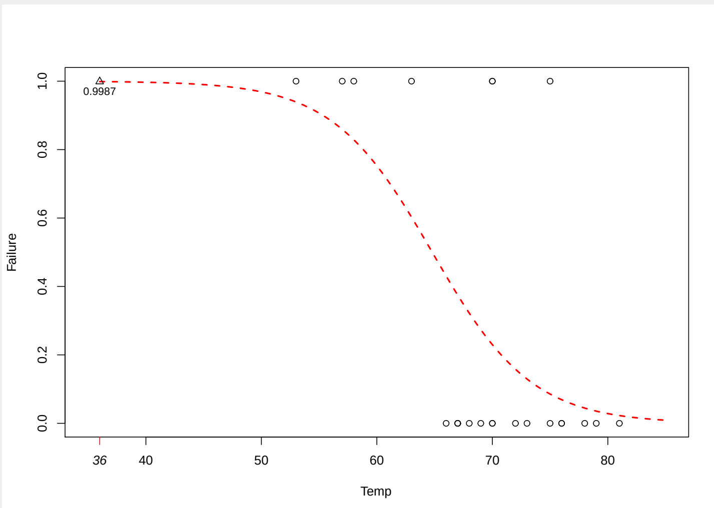
```
Residual deviance: 20.315  on 21  degrees of freedom
AIC: 24.315

> confint(preC.logit)
                  2.5 %      97.5 %
(Intercept)  3.3305848 34.34215133
Temp        -0.5154718 -0.06082076
```

The logistic regression model is

$$\text{logit}\,[\hat{\pi}(x)] = 15.0329 - 0.2322(\text{Temp.})$$

According to the report, the air temperature at the time of launch, 11:38 a.m. EST, was 36 degrees. This temperature was 15 degrees colder than any previous launch and the O-ring suffered catastrophic failure.

```
> predict.glm(preC.logit,newdata=data.frame(Temp=36),type="response")
        1
0.9987521
```



http://users.stat.ufl.edu/~athienit/STA4504/Examples/oring.R

3.3 GLM for Count Data

3.3.1 Modeling Counts

Many discrete response variables have counts as possible outcomes. The Poisson distribution is often used as a sampling model for counts.

Example 3.3 Data examples:

- For a sample of cities worldwide, each observation might be the number of automobile thefts in 2003.

- For a sample of silicon wafers used in computer chips, each observation might be the number of imperfections on wafer.

The Poisson probability mass function is

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

with $E(Y) = V(Y) = \mu$.

The (simple) Poisson *log-linear* is

$$\log(\mu) = \alpha + \beta x \quad \Rightarrow \quad \mu = e^{(\alpha+\beta x)} = e^{\alpha}(e^{\beta})^x$$

R code 3.2 For a poisson regression, take for example

```
glm(y∼x,family=poisson,data=mydata)
```

Please see the `help(glm)` help file.

Example 3.4 (Silicon Wafers) Let,

- Y = number of defects os silicon wafer.
- $x = 0$ if type A, 1 if type B.

A	8	7	6	6	3	4	7	2	3	4
B	9	9	8	14	8	13	11	5	7	6

```
> wafers.log=glm(defects~trt,family=poisson(link="log"),data=wafers)
> summary(wafers.log)
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 ***
---
    Null deviance: 27.857  on 19  degrees of freedom
Residual deviance: 16.268  on 18  degrees of freedom
AIC: 94.349

> confint(wafers.log)
                2.5 %    97.5 %
(Intercept) 1.3188383 1.8743819
trtB        0.2469096 0.9400962
```

The log-linear model is

$$\log[\mu(x)] = 1.6094 + 0.5878x$$

giving us

A: $\mu(0) = \exp(1.6094) = 5$

B: $\mu(1) = \exp(1.6094)\exp(0.5878) = 5 + 4 = 9$

http://users.stat.ufl.edu/~athienit/STA4504/Examples/wafers.R

3.3.2 Modeling Rates

There are situations when the counts have different bases and so an adjustment is necessary, that is we model the rate at which an event occurs.

Example 3.5 Consider two individuals given fishing nets and told they have 5 minutes to catch as many fish as possible. After, 5 minutes

- Person A catches 11
- Person B catches 20

Who "perfomed" better?

Person A

Person B

That is a trick question, because what if the sizes of the nets where different, then we need to account on how many fish per square inch of net, i.e. a rate. Also, if person A got 11 fish with that net that's amazing!

Let y be the count and t be the base

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

Hence,

$$\begin{aligned} \log\left(\frac{\mu}{t}\right) &= \log(\mu) - \log(t) = \alpha + \beta x \\ \Rightarrow \log(\mu) &= \alpha + \beta x + \log(t) \\ \Rightarrow \log(\mu) &= \alpha + \beta x + \underbrace{\beta_2}_{=1} \underbrace{x_2}_{\log(t)} \end{aligned}$$

All that is required is to add another "predictor" whose coefficient is set to 1, and the solve using *restricted maximum likelihood*. The term $\log(t)$ is called the *offset*.

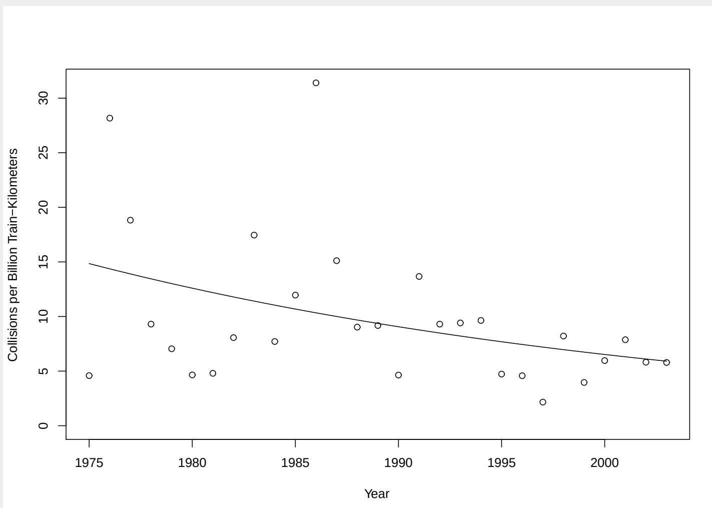
R code 3.3 Here when we fit the model we use `offset` argument

```
glm(y~x+offset(log(base)),...)
```

Example 3.6 (British Train Accidents) The first stationary gasoline engine developed by Carl Benz was a one-cylinder two-stroke unit which ran for the first time on New Year's Eve 1879. So consider the number of automobile accidents in 1879 compared to 2019. We need to adjust for the fact that there are more automobiles on the road and that they travel larger distances.
The same is true for the following Train-Road collision data, where an *offset* is needed. 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.

```
traincollisions
   Year  KM Train TrRd
1  2003 518     0    3
2  2002 516     1    3
3  2001 508     0    4
4  2000 503     1    3
5  1999 505     1    2
6  1998 487     0    4
7  1997 463     1    1
8  1996 437     2    2
9  1995 423     1    2
10 1994 415     2    4
11 1993 425     0    4
12 1992 430     1    4
13 1991 439     2    6
14 1990 431     1    2
15 1989 436     4    4
16 1988 443     2    4
17 1987 397     1    6
18 1986 414     2   13
19 1985 418     0    5
20 1984 389     5    3
21 1983 401     2    7
22 1982 372     2    3
23 1981 417     2    2
24 1980 430     2    2
25 1979 426     3    3
26 1978 430     2    4
27 1977 425     1    8
28 1976 426     2   12
29 1975 436     5    2

> trains.log=glm(TrRd~I(Year-1975)+offset(log(KM)),family=poisson(link=log),
```

```
+  data=traincollisions)
> summary(trains.log)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.21142    0.15892  -26.50  < 2e-16 ***
I(Year - 1975) -0.03292    0.01076   -3.06  0.00222 **
---
    Null deviance: 47.376  on 28  degrees of freedom
Residual deviance: 37.853  on 27  degrees of freedom
AIC: 133.52

> sum(resid(trains.log,type="pearson")^2)
[1] 42.19178
```



The model is

$$\log\left(\frac{\hat{\mu}}{t}\right) = -4.21142 - 0.03292x$$
$$\frac{\hat{\mu}}{t} = e^{-4.21142}e^{-0.03292x}$$
$$= (0.0148)(0.968)^x$$

Rate estimated to decrease by $1 - 0.968 = 0.032 = 3.2\%$ per year from from 1975 to 2003, i.e. the rate is 0.968 times the rate of the previous year.

- Est. rate for 1975 ($x = 0$) is 0.0148 per million km
- Est. rate for 2003 ($x = 28$) is 0.0059 per million km

http://users.stat.ufl.edu/~athienit/STA4504/Examples/trains.R

Example 3.7 (Airline Fatalities) Data obtained from MIT Airline Data Project and Wikipedia, provides information on fatalities, Available Seat Miles (ASM) and year

```
> air_deaths
   Fatalities     ASM Year
1        1828  829581 1995
2        2796  862621 1996
3        1768  884192 1997
4        1721  898359 1998
5        1150  945245 1999
6        1586  980769 2000
7        1539  953875 2001
8        1418  914901 2002
9        1233  922277 2003
10        767  998868 2004
11       1463 1028621 2005
12       1298 1027553 2006
13        981 1060116 2007
14        952 1040840 2008
15       1108  975307 2009
16       1130  991934 2010
17        828 1012597 2011
18        800 1012261 2012
19        459 1025616 2013
20       1328 1048107 2014
21        898 1090198 2015
22        629 1131694 2016
23        399 1168055 2017
```

Fitting a Poisson log-linear model with offset

```
> air.poisson=glm(Fatalities~I(Year-1995),family=poisson,data=air_deaths,
+ offset=log(ASM))
> summary(air.poisson)
Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.0541485  0.0101474 -596.62   <2e-16 ***
I(Year - 1995) -0.0638961  0.0009377  -68.14   <2e-16 ***
---
    Null deviance: 6595.9  on 22  degrees of freedom
Residual deviance: 1751.8  on 21  degrees of freedom
AIC: 1959.4
```

http://users.stat.ufl.edu/~athienit/STA4504/Examples/airline.R

3.4 Inference and Model Checking

3.4.1 Standard testing - Wald

Display 3.1 (Inference on parameters) Since parameter estimation is done via ML, and MLE's are asymptotically normal, inference is done in the traditional way. Let $\theta = (\alpha, \beta)$ denote the parameter vector

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(0, \frac{1}{I(\boldsymbol{\theta}_0)}\right)$$

where $(\boldsymbol{\theta}_0)$ is the *Fisher information* evaluated at $\boldsymbol{\theta}_0$ (not covered in this class). Therefore, hypotheses tests and confidence intervals for the parameter's are done accordingly.

To test $\mathrm{H}_0 : \beta = \beta_0$ you can create the test statistic

$$TS = \frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}} \overset{\mathrm{H}_0}{\sim} N(0,1)$$

and obtain p-value in traditional way. A $100(1-\alpha)\%$ CI on β can also be created

$$\hat{\beta} \mp z_{1-\alpha/2}\left(s_{\hat{\beta}}\right) \tag{3.2}$$

These methods can be extended to one-sided tests.

Example 3.8 (Infant malformatrion continued) From the output

```
> summary(malform.logit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9605     0.1154 -51.637   <2e-16 ***
Alcohol       0.3166     0.1254   2.523   0.0116 *
```

we can create a 95% C.I.

$$0.3166 \mp 1.96(0.1254) \longrightarrow (0.070816, 0.562384)$$

We will see functions that create C.I.'s but their default is not the Wald method.

3.4.2 Likelihood Ratio Test - Deviance

Goodness of Fit

Deviance is actually the likelihood ratio test for *goodness of model fit*, that is, equation (2.2) for

$$\mathrm{H}_0\text{: model adequately fits}$$

$$D(y;\hat{\mu}) := G^2 = -2[L(\hat{\mu};y) - L(y;y)] \xrightarrow[\mathrm{H}_0]{d} \chi^2_{df} \tag{3.3}$$

with p-value being $P\left(\chi^2_{df} \geq G^2\right)$ and where

- $L(\hat{\mu};y)$ is the log-likelihood of the fitted model
- $L(y;y)$ is the log-likelihood of the *saturated* model, that is the model that has a separate parameter for each observation giving a perfect fit but with 0 degrees of freedom (so no inference can be done within that model).
- df as in equation (2.3)

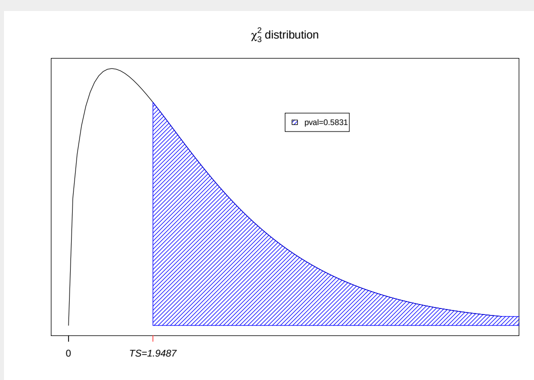
Display 3.2 (Goodness of fit) A *goodness of fit* can be used only in the number of predictor levels is fixed and relatively small to the overall sample size. Either, X^2 or G^2 can be used since to compare the observed counts to the values predicted by the fitted model.

Remark 3.2. *Goodness of fit can also be performed - preferred even - by using X^2, instead of G^2.*

Example 3.9 Revisiting some examples

(Infant Malformation) For example 3.1 a goodness of fit can be used (with either X^2 or G^2) as there are only 5 binomials and as more women are surveyed/sampled the number of binomials (rows of data) remains fixed.

```
...
Residual deviance: 1.9487  on 3  degrees of freedom
>  sum(resid(malform.logit,type=``pearson'')^2)
[1] 2.20523
```



$G^2 = 1.9487$ ($X^2 = 2.0523$) with $df = 3 \longrightarrow$ p-value $= 0.5831$

(Challenger disaster) A goodness of fit is not adequate as each row corresponds to a Bernoulli trial, that is a 0 or 1. As sample size increases so will the number of rows in the data.

Remark 3.3. *If the data is not grouped you may still perform a goodness of fit by*

- *grouping your predictor(s). For example, for temperature you could create groups 31-40, 41-50, … and then create scores such as 35, 45, … ensuring that the number of predictor levels remains relatively fixed.*

- *comparing current model to a "fuller" model rather than to a saturated model (fullest). A fuller model could be one with polynomial terms, interactions, etc.*

Parameter testing

Likelihood ratio test can be used to test $H_0 : \beta = \beta_0$ using deviances. To be specific the difference of two goodness of fit tests.

$$\begin{aligned} G^2 &= D(y;\hat{\mu}_0) - D(y;\hat{\mu}_1) \\ &= -2[L(\hat{\mu}_0;y) - L(y;y)] - (-2)[L(\hat{\mu}_1;y) - L(y;y)] \\ &= -2[L(\hat{\mu}_0;y) - L(\hat{\mu}_1;y)] \\ &\xrightarrow[H_0]{d} \chi^2_{df} \end{aligned} \tag{3.4}$$

where

- $L(\hat{\mu}_0;y)$ is the log-likelihood of the reduced (under the null) model
- $L(\hat{\mu}_1;y)$ is the log-likelihood of the fitted model
- df is the difference in degrees of freedom of the two models which corresponds to the dimension reduction of our coefficient parameter vector, in this case 1 as we are restricting one parameter $\beta = \beta_0$.

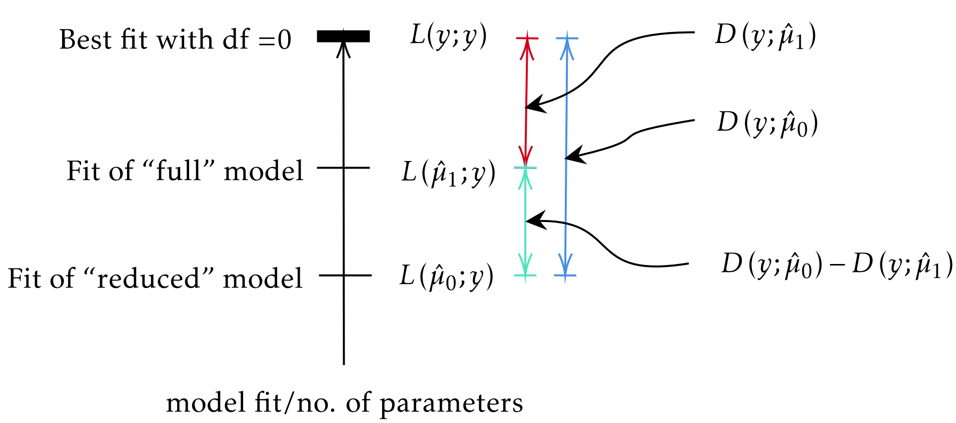


Figure 3.2: Illustration of how Deviances are used in LRTs

Remark 3.4. *The "Null Deviance" that is usually provided in R output is the deviance for the null*

$$H_0 : \beta = 0 \quad (\beta_i = 0\ \forall i \text{ for models with more than one predictor})$$

So that

$$\begin{aligned} \textit{Null Deviance} - \textit{Residual Deviance} &= D(y;\hat{\mu}_0) - D(y;\hat{\mu}_1) \\ &= G^2 \end{aligned}$$

which is the likelihood ratio test statistic.

For binomial and Poisson models

$$D(y;\hat{\mu}) = 2\sum_{i=1}^{n} y_i \log(y_i/\hat{\mu}_i)$$

The likelihood ratio test can be used to create a $100(1-\alpha)$ confidence interval on β. That is, finding all the null values β_0 for which would yield a test statistics with a large p-value. It is a bit more complicated than equation (3.2) so we use software.

R code 3.4 Use `confint(.)` to obtain the likelihood ratio confidence intervals.

Example 3.10 (Infant Malformation continued) We focus on testing $H_0 : \beta = 0$ via deviances.

```
> summary(malform.logit)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.9605     0.1154 -51.637   <2e-16 ***
Alcohol       0.3166     0.1254   2.523   0.0116 *
---
    Null deviance: 6.2020  on 4  degrees of freedom
Residual deviance: 1.9487  on 3  degrees of freedom
AIC: 24.576

> confint(malform.logit)
                  2.5 %     97.5 %
(Intercept) -6.19302366 -5.7396968
Alcohol      0.01868149  0.5234947
```

Note that this CI is different from the Wald CI done earlier of (0.070816, 0.562384). The test statistic from equation (3.4)

$$\text{Null deviance} - \text{Residual deviance} = 6.2020 - 1.9487 = 4.2533$$

with p-value

```
> 1-pchisq(4.2533,1)
[1] 0.03917414
```

and we reject the null.

Exercise 3.1 Do the same for the "Challenger" disaster and "Silicon wafers" examples.

3.5 Overdispersion

We know that
$$E(\chi^2_\nu) = \nu$$
so for a well fitting model we expect
$$X^2 \approx \text{Residual d.f.}$$
However, cases where
$$X^2 \gg \text{Residual d.f.}$$
are of concern. Could use G^2 (Residual Deviance) as an alternative but not as efficient in detecting overdispersion.

Reasons

1. Badly fitting model
   - omitted terms/variables
   - incorrect relationship (link)
   - outliers
2. Variation greater than predicted by model that leads to *overdispersion*
   - count data: $V(Y) > \mu$
   - binomial data: $V(Y) > n\pi(1-\pi)$

Causes of Overdispersion

- variability of experimental material - individual level variability
- correlation between individual responses, e.g. litters of rats
- cluster sampling, e.g. areas; schools; classes; children
- aggregate level data
- omitted unobserved variables
- excess zero counts (structural and sampling zeros)

Consequences With correct mean model we have consistent estimates of β but:

- incorrect standard errors
- selection of overly complex models

Remark 3.5. *Overdispersion is much more common for count data, especially due to the restriction by the Poisson model* $E(Y) = V(Y)$.

The two most popular methods for checking overdispersion are;

- Check whether $X^2 \gg df$, or $X^2/df \gg 1$,
- Fit a different model with additional parameters that allow variance to be greater and test the significance of those parameters
  - count data: Negative Binomial, parameter θ is introduced and estimated via MLE

$$V(Y) = \mu + \left(\frac{1}{\theta}\right)\mu^2 \tag{3.5}$$

  - binomial data: Beta-Binomial, parameter ρ is introduced and estimated via MLE

$$V(Y) = n\pi(1-\pi)[1+(n-1)\rho]$$

R code 3.5 Most common ways of fitting these models are

- Negative Binomial: `glm.nb(.)` in the `MASS` package
- Beta-Binomial: `betabinomial(.)` in the `VGAM` package

Example 3.11 (Homicide) 1308 individuals who where classified as “Black” or “White” were asked: ”How many people have you known personally that were victims of homicide?”

	Number of victims						
Race	0	1	2	3	4	5	6
Black	119	16	12	7	3	2	0
White	1070	60	14	4	0	0	1

```
> head(homicide)  #data entered in ``shorter'' format
  nvics  race Freq
1     0 Black  119
2     1 Black   16
3     2 Black   12
4     3 Black    7
5     4 Black    3
6     5 Black    2
> homicide=transform(homicide,race=relevel(race,"White"))
```

```
> hom.poi=glm(nvics~race,family=poisson(link="log"),
+ weights=Freq,data=homicide)
> summary(hom.poi)
.
.
    Null deviance: 962.80  on 10  degrees of freedom
Residual deviance: 844.71  on  9  degrees of freedom
```

Checking for overdispersion via $X^2/(df) \gg 1$ we first notice that the way the data was entered, the degrees of freedom is not 9 but actually 1308-2=1306

```
> sum(resid(hom.poi,type="pearson")^2)/
+  (sum(homicide$Freq)-length(hom.poi$coefficients))
[1] 1.745692
```

So some evidence of overdispersion is apparent. Now to find the negative binomial

```
> library(MASS)
> hom.nb=glm.nb(nvics~race,weights=Freq,data=homicide)
> summary(hom.nb)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.3832     0.1172 -20.335  < 2e-16 ***
raceBlack     1.7331     0.2385   7.268 3.66e-13 ***
---
    Null deviance: 471.57  on 10  degrees of freedom
Residual deviance: 412.60  on  9  degrees of freedom
AIC: 1001.8


              Theta:  0.2023
          Std. Err.:  0.0409

 2 x log-likelihood:  -995.7980
```

and the estimate of

$$\left(\frac{1}{\hat{\theta}}\right) \approx 5$$

seems substantial in equation (3.5). Much better now,

```
> sum(resid(hom.nb,type="pearson")^2)/
+  (sum(homicide$Freq)-length(hom.nb$coefficients))
[1] 1.090373
```

http://users.stat.ufl.edu/~athienit/STA4504/Examples/homicide.R

Example 3.12 (British Train Accidents continued) Checking for potential overdispersion, we are not quite sure if $X^2/df \gg 1$

```
> sum(resid(trains.log,type="pearson")^2)
[1] 42.19178
```

```
> sum(resid(trains.log,type="pearson")^2)/trains.log$df.residual
[1] 1.562658
```

So we fit a negative binomial,

```
> library(MASS)
> trains.nb=glm.nb(TrRd ~ I(Year-1975) + offset(log(KM)),
+  data=traincollisions)
> summary(trains.nb)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -4.19999    0.19584 -21.446  < 2e-16 ***
I(Year - 1975) -0.03367    0.01288  -2.615  0.00893 **
---
    Null deviance: 32.045  on 28  degrees of freedom
Residual deviance: 25.264  on 27  degrees of freedom
AIC: 132.69

              Theta:  10.12
          Std. Err.:  8.00

 2 x log-likelihood:  -126.69
```

Since, $\hat{\theta}+2se(\hat{\theta})\approx 26$ and hence $1/26\approx 0.0385$ is close to 0. Therefore the second term in equation (3.5) does not seem to be that significant and conclude no strong evidence of overdispersion.
http://users.stat.ufl.edu/~athienit/STA4504/Examples/trains.R

Example 3.13 (Airline Fatalities continued) Fit a negative binomial model due to potential overdispersion...why is there potential overdispersion?

```
> air.nb=glm.nb(Fatalities~I(Year-1995)+offset(log(ASM)),data=air_deaths)
> summary(air.nb)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -6.06375    0.10807 -56.110  < 2e-16 ***
I(Year - 1995) -0.06256    0.00843  -7.421 1.16e-13 ***
---
    Null deviance: 78.655  on 22  degrees of freedom
Residual deviance: 23.319  on 21  degrees of freedom
AIC: 334.03

              Theta:  14.09
          Std. Err.:  4.17

 2 x log-likelihood:  -328.03
```

we conclude that the rate is decreasing. As an exercise, interpret the magnitude of $\hat{\beta}$ per 1 year increase.

```
http://users.stat.ufl.edu/~athienit/STA4504/Examples/airline.R
```

Exercise 3.2 Check for overdispersion with the "Silicon wafers" example

Remark 3.6. *The Beta-Binomial is omitted here but an alternative method that does not use a likelihood approach but merely the structure between the mean and variance are the*

- *count data: Pseudo-Poisson*
- *binomial data: Pseudo-Binomial*

but as result likelihood ratio tests are not possible.

4. Logistic Regression

Closer look at logistic regression and reviewing the model fitting process.

4.1 Interpretation

We have seen the simple logistic regression model as in equation (3.1). That is

$$\operatorname{logit}[\pi(x)] = \alpha + \beta x \quad \Rightarrow \quad \pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

- $\beta > 0$, then $\pi(x) \uparrow$ as $x \uparrow$
- $\beta < 0$, then $\pi(x) \downarrow$ as $x \uparrow$
- $\beta = 0$, then $\pi(x) = e^{\alpha}/(1+e^{\alpha})$ which is a constant, with $\pi(x) > 0.5$ when $\alpha > 0$
- The rate of change in $\pi(x)$ (but taking derivatives) is $\beta\pi(x)[1-\pi(x)]$.

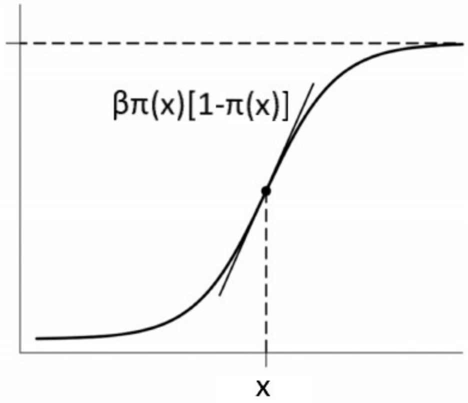


Figure 4.1: Rate of change

Note that the rate of change is maximized when $\pi(x) = 0.5$. This implies

$$\text{max rate of change is } \frac{\beta}{4} \quad \text{when} \quad x = \frac{-\alpha}{\beta}$$

This value of x is sometimes called the *median effective level* and it represents the level at which each outcome has a 50% chance.

- The term e^{β} is odds ratio for a 1 unit increase in x. The odds os success are
  - at x
$$\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x}$$
  - at $x+1$
$$\frac{\pi(x+1)}{1-\pi(x+1)} = e^{\alpha+\beta x}e^{\beta}$$

  Hence the odds ratio for $x+1$ versus x is

$$\text{OR} = \frac{\pi(x+1)/[1-\pi(x+1)]}{\pi(x)/[1-\pi(x)]} = e^{\beta}$$

- Parameters estimated via MLE are asymptotically normal.

Example 4.1 (Horseshoe crab) There are 173 female crabs for which we wish to model the presence or absence of male "satellites" dependent upon characteristics of the female horseshoe crabs.

$$Y_i = \begin{cases} 1 & \text{satellite present} \\ 0 & \text{otherwise} \end{cases}$$

Explanatory variables are: weight (in kg), width of shell, color (medium light, medium, medium dark, dark), and condition of spine (bad, good, excellent).

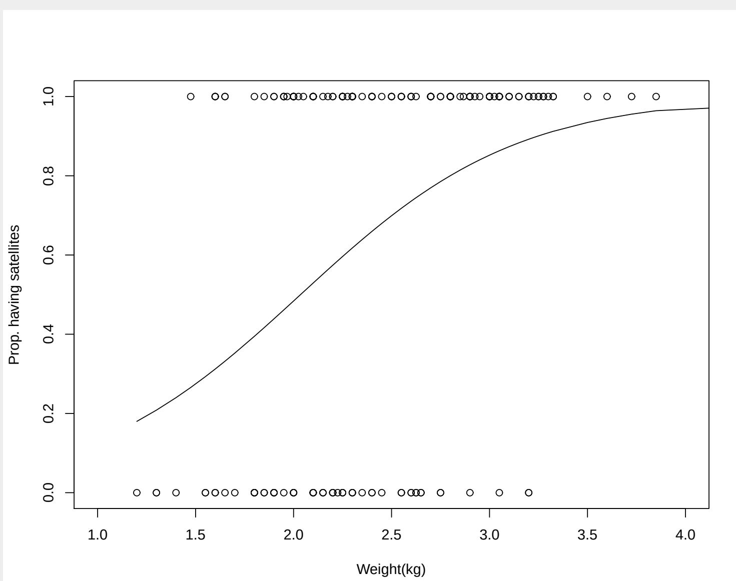
```
> fit=glm(y ~ weight, family=binomial(link=logit))
> summary(fit)
```

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 195.74  on 171  degrees of freedom
AIC: 199.74
```

The maximum likelihood fit is then $\text{logit}[\hat{\pi}(x)] = -3.6947 + 1.8151x$. Note that β is positive, implying that $\hat{\pi}(x) \uparrow$ as $x \uparrow$.

$$\hat{\pi}(x) = \frac{\exp(-3.6947 + 1.8151x)}{1 + \exp(-3.6947 + 1.8151x)}$$

- At the average weight of $x = \bar{x} = 2.44$, $\hat{\pi}(2.44) = 0.676$.
- The rate of change at $x = 2.44$ is $\hat{\beta}\hat{\pi}(1 - \hat{\pi}) = 1.8151(0.676)(0.324) = 0.398$.
- The estimated change in π per 1 kg increase is about 0.398 (in the neighborhood of the sample mean). However, the standard deviation of weight is $s_x = 0.58$ and hence talking about a 1 unit increase, i.e. 1 kg, may be too much of an increase and so the estimated change in π per 0.1 kg increase is about 0.0398.
- $\hat{\pi}(x) = 1/2$ when $x = \frac{-(-3.6947)}{1.8151} = 2.036$
- For a 1 kg increase in weight, the estimated odds of the presence of a satellite are multiplied by $\exp(1.8151) = 6.14169$. Consequently, for a 0.1 kg increase in weight, the estimated odds of the presence of a satellite are multiplied by $\exp(0.1(1.8151)) = 1.2$, i.e. the odds increase by 20%.



Part (I) of http://users.stat.ufl.edu/~athienit/STA4504/Examples/crab_u.R

4.2 Inference

We refer the reader to review Section 3.4 and we will expand from there. We have covered how to create confidence intervals on individual (coefficient) parameters, e.g. β, but now we expand to *linear combinations of parameters*.

Goal: Create a CI for $\pi(x)$.

1. First work with $\text{logit}[\hat{\pi}(x)] = \hat{\alpha} + \hat{\beta}x$, where we know (via MLE)

$$\hat{\alpha} \sim N\left(\alpha, \sigma_\alpha^2\right) \text{ and } \hat{\beta} \sim N\left(\beta, \sigma_\beta^2\right) \quad \Rightarrow \quad \underbrace{\hat{\alpha} + \hat{\beta}x}_{\text{logit}[\hat{\pi}(x)]} \sim N\left(\alpha + \beta x, \sigma_\alpha^2 + x^2\sigma_\beta^2 + 2x\sigma_{\alpha\beta}\right)$$

2. The $100(1-\alpha)\%$ CI for $\text{logit}[\pi(x)] = \alpha + \beta x$ is

$$\hat{\alpha} + \hat{\beta}x \mp z_{1-\alpha/2}\sqrt{s_\alpha^2 + x^2 s_\beta^2 + 2x s_{\alpha\beta}} \quad \rightarrow \quad (L, U) \tag{4.1}$$

where s_α^2 and s_β^2 are the estimated variances, and $s_{\alpha\beta}$ is the estimated covariance.

$$V\left(\hat{\alpha} + \hat{\beta}x\right) = V(\hat{\alpha}) + x^2 V(\hat{\beta}) + 2x\text{Cov}(\hat{\alpha}, \hat{\beta})$$

R code 4.1 Using software

- The variance-covariance matrix for all parameters can be found for `glm` objects by using `vcov(model)`
- The estimate and the standard error for $\text{logit}[\hat{\pi}(x)] = \hat{\alpha} + \hat{\beta}x$ can obtained using

```
predict.glm(model,newdata,type=``link'',...)
```

3. The $100(1-\alpha)\%$ CI for $\pi(x)$, using equation (4.1) is then

$$\left(\frac{e^L}{1+e^L}, \frac{e^U}{1+e^U}\right)$$

Remark 4.1. *We looked at CI for $\alpha + \beta x$, a linear combination of two parameters but this method can be extended to linear combinations of any length of parameters.*

Example 4.2 (Horseshoe crab continued) Test $\text{H}_0 : \beta = 0$ via

- Wald test given in the summary output (and CI could be derived)
- Likelihhod ratio test G^2 and prerrably corresponding CI

```
> confint(fit,"weight")
```

```
     2.5 %   97.5 %
  1.113790 2.597305
```

There are 6 female crabs with a weight of 2.4 kg (or 2400 g), of whom only 4 have at least one satellite. Using the model we construct a 95% CI for $\hat{\pi}(2.4)$, by first constructing the CI for logit$[\hat{\pi}(2.4)]$

```
> eta=predict(fit,newdata=data.frame(weight=2.4),type="link",se.fit=TRUE)
> eta
$fit
[1] 0.6616206

$se.fit
[1] 0.1780615
```

Note that the standard error is the same as if we directly use equation (4.1)

```
> sqrt(vcov(fit)[1,1]+2.4^2*vcov(fit)[2,2]+2*2.4*vcov(fit)[1,2])
[1] 0.1780615

> eta.ci=eta$fit+c(-1,1)*qnorm(0.975)*eta$se.fit
> eta.ci # This is (l,u) interval
[1] 0.3126265 1.0106148

> plogis(eta.ci) # This is (exp(l)/(1+exp(l)),exp(u)/(1+exp(u)))
[1] 0.5775262 0.7331404
```

Part (I) of http://users.stat.ufl.edu/~athienit/STA4504/Examples/crab_u.R

4.3 Multiple Logistic Regression

Just as in OLS regression, multiple regression can be used when multiple predictors $x_1, x_2, \ldots, x_k$ are available, yielding

$$\text{logit}[\pi(x)] = \alpha + \sum_{i=1}^{k} \beta_i x_i \quad \Leftrightarrow \quad \pi(x) = \frac{e^{\alpha + \sum_{i=1}^{k} \beta_i x_i}}{1 + e^{\alpha + \sum_{i=1}^{k} \beta_i x_i}}$$

Example 4.3 (Horseshoe crab continued) Next we introduce the color variable into the model by creating 3 indicator variables for the 4 levels of color. Let,

$$c_1 = \begin{cases} 1 & \text{medium light} \\ 0 & \text{o/w} \end{cases} \qquad c_2 = \begin{cases} 1 & \text{medium} \\ 0 & \text{o/w} \end{cases} \qquad c_3 = \begin{cases} 1 & \text{medium dark} \\ 0 & \text{o/w} \end{cases}$$

and hence $c_1 = c_2 = c_3 = 0$ indicates whether a female crab is dark (i.e. base group). The model is then

$$\text{logit}[\pi(x)] = \alpha + \beta_1 x + \beta_2 c_1 + \beta_3 c_2 + \beta_4 c_3$$

with

Color	logit $[\pi(x)]$
medium light	$(\alpha+\beta_2)+\beta_1 x$
medium	$(\alpha+\beta_3)+\beta_1 x$
medium dark	$(\alpha+\beta_4)+\beta_1 x$
dark	$\alpha+\beta_1 x$

```
> color=color - 1 # color now takes values 1,2,3,4
> color=factor(color) # treat color as a factor
> fit2.1=glm(y ~ weight + color, family=binomial(link=logit),
+ contrasts=list(color=contr.treatment(4,base=4,contrasts=TRUE)))
> summary(fit2.1)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -4.5266     1.0038  -4.510 6.50e-06 ***
weight        1.6928     0.3888   4.354 1.34e-05 ***
color1        1.2694     0.8488   1.495  0.13479
color2        1.4143     0.5449   2.595  0.00945 **
color3        1.0833     0.5884   1.841  0.06561 .
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 188.54  on 168  degrees of freedom
AIC: 198.54
```

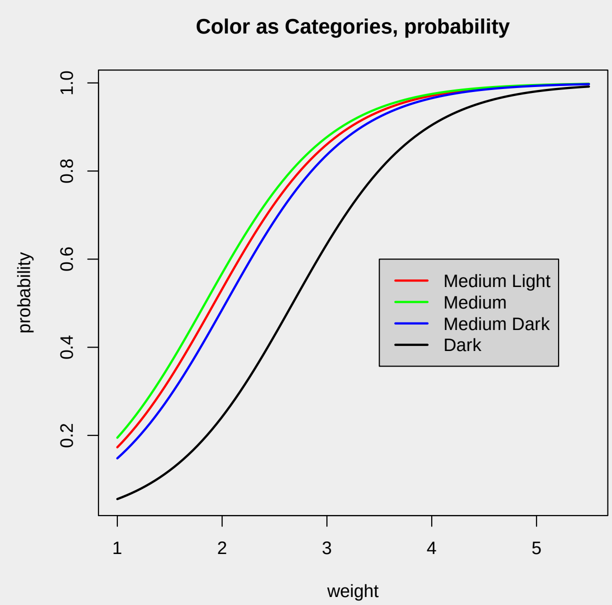


Figure 4.2: Probability curves

Part (II) subpart II1 of `http://users.stat.ufl.edu/~athienit/STA4504/Examples/crab_u.R`

In Section 3.4 we saw how to perform inference on a single parameter $\mathrm{H}_0 : \beta = \beta_0$ via

- Using standard normal test
$$\frac{\hat{\beta} - \beta_0}{s_{\hat{\beta}}} \overset{\mathrm{H}_0}{\sim} N(0,1)$$
- Using the likelihood ratio test G^2 in equation (3.4).

Now we extend the methodology in equation (3.4) to testing multiple parameters simultaneously. In the Horseshoe crab data there were 3 parameters for fitting color as a qualitative predictor, β_2, β_3 and β_4. If we wished to test if color at all was significant one would test

$$\mathrm{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0 \quad \text{vs} \quad \mathrm{H}_1 : \text{at least one } \beta \neq 0$$

which yields the *reduced model* (under null), μ_0 and the *full model*, μ_1

$$\begin{aligned} g(\mu_0) &= \alpha + \beta_1 x \\ g(\mu_1) &= \alpha + \beta_1 x + \beta_2 c_1 + \beta_3 c_2 + \beta_4 c_3 \end{aligned}$$

and by obtaining the deviances we can create the likelihood ratio test

$$G^2 = D(y; \hat{\mu}_0) - D(y; \hat{\mu}_1) \xrightarrow[\mathrm{H}_0]{d} \chi^2_{df}$$

where df is the difference in degrees of freedom of the two models which corresponds to the dimension reduction of our coefficient parameter vector, in this case $df = 3$ as we are restricting 3 parameter under the null.

Example 4.4 (Horseshoe crab continued) To test the significance of color, controlling for weight we must test $\mathrm{H}_0 : \beta_2 = \beta_3 = \beta_4 = 0$. The likelihood-ratio test (LRT) statistic is

$$\begin{aligned} G^2 &= D(y; \hat{\mu}_0) - D(y; \hat{\mu}_1) \\ &= 195.74 - 188.54 = 7.2 \end{aligned}$$

which when compared to a χ^2_3 produces a p-value of 0.06578905 ≈ 0.07 which at the 0.05 significance level might let us conclude that color is not significant.
However, looking at the individual test statistic values as well as the figure of probability curves we see that there is a more this problem that we will be addressing in the next chapter.
Part (II), subpart A of `http://users.stat.ufl.edu/~athienit/STA4504/Examples/crab_u.R`

Example 4.5 (Horseshoe crab continued) From figure 4.2 we notice than there are may be in fact be only two groups: dark and not dark.

```
> dark=ifelse(unclass(color)==4,1,0)
> fit2.2=glm(y ~ weight + dark, family=binomial(link=logit))
> summary(fit2.2)
Coefficients:
```

```
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.3134     0.8984  -3.688 0.000226 ***
weight        1.7292     0.3825   4.520 6.18e-06 ***
dark         -1.2954     0.5222  -2.481 0.013110 *
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 189.17  on 170  degrees of freedom
AIC: 195.17
```

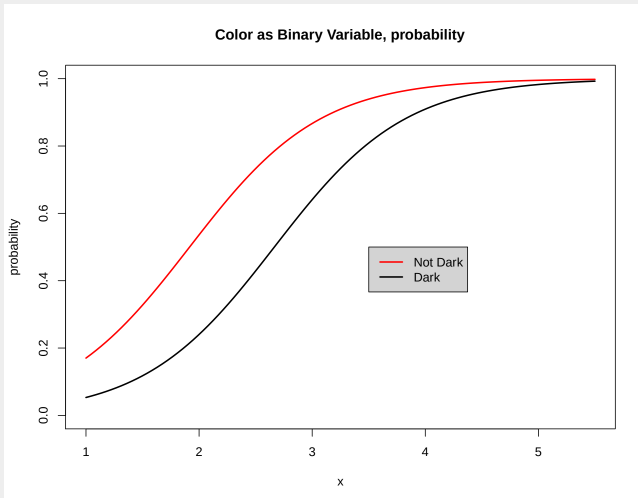


Figure 4.3: Probability curves

In example 4.4 the LRT for color yielded a p-value of 0.07. However, now testing dark vs non dark via $H_0 : \beta_2 = 0$ for this model

- Via normal test, p-value = 0.013110
- Via LRT, $G^2 = 195.74 - 189.17$ with 1 degree of freedom yields, p-value = 0.01039651

Part (II), subpart B of `http://users.stat.ufl.edu/~athienit/STA4504/Examples/crab_u.R`
Why did the p-value drop from 0.07 to about 0.01? Because we tested using a method that uses less degrees of freedom (1 instead of 3) and hence has more power in detecting significance.

4.3.1 Qualitative predictors

If a qualitative predictor is deemed significant, the next step is an investigation into the different levels. This yields situations where one might we to test linear combinations of parameters.

$$H_0 : \sum_{i=1}^{k} c_i \beta_i = \Delta_0 \tag{4.2}$$

for constants c_i.

Example 4.6 (Horseshoe crad continued) Testing $\beta_2 = 0, \beta_3 = 0$ and $\beta_4 = 0$ individually amounts to testing differences between each group to the base group

Color	$\text{logit}[\pi(x)]$
medium light	$(\alpha + \beta_2) + \beta_1 x$
medium	$(\alpha + \beta_3) + \beta_1 x$
medium dark	$(\alpha + \beta_4) + \beta_1 x$
dark	$\alpha + \beta_1 x$

In example 4.3, we note that there appear to be differences between medium and dark, and between medium dark and dark based on those tests. In addition, the estimated odds ratio comparing the following groups to dark at any fixed level of weight are

Comparison	OR
medium light vs dark	$\exp(\hat{\beta}_2) = \exp(1.2694) = 3.56$
medium vs dark	$\exp(\hat{\beta}_3) = \exp(1.4143) = 4.11$
medium dark vs dark	$\exp(\hat{\beta}_4) = \exp(1.0833) = 2.95$

To motivate the next section consider comparing two groups such as medium light vs. medium. We could always refit the model making one of these groups the new base group. Keeping with this model we this comparisons amounts to testing:

$$\mathrm{H}_0 : \beta_2 - \beta_3 = 0$$

a linear combination of the parameters

To test the null in (4.2), an option is to create a CI for $\sum c_i \beta_i$ using the asymptotic normality property and see whether Δ_0 is a plausible value or not.

$$\sum_{i=1}^{k} c_i \hat{\beta}_i \mp z_{1-\alpha/2} \sqrt{\hat{V}\left(\sum_{i=1}^{k} c_i \hat{\beta}_i\right)} \tag{4.3}$$

with the estimated variance obtained by the sum of the estimated pairwise covariances using the property that

$$\begin{aligned} V\left(\sum_{i=1}^{k} c_i \hat{\beta}_i\right) &= \sum_{i=1}^{k} \sum_{j=1}^{k} c_i c_j \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) \\ &= \sum_{i=1}^{k} c_i^2 V(\hat{\beta}_i) + 2 \sum_{i<j}\sum c_i c_j \mathrm{Cov}(\hat{\beta}_i, \hat{\beta}_j) \end{aligned}$$

This concept was used in equation (4.1) where $c = (1, x)$ and the parameter vector was (α, β), such that

$$(1, x)\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \alpha + \beta x$$

Example 4.7 (Horseshoe crab continued) The log odds ratio for comparing medium light vs medium at fixed levels of weight is $\beta_2 - \beta_3$. Using equation (4.3) with $c = (0,0,1,-1,0)$

$$(0,0,1,-1,0)\begin{pmatrix}\alpha\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix} = \beta_2 - \beta_3$$

we have that

$$\hat{\beta}_2 - \hat{\beta}_3 \mp z_{0.975}\sqrt{s^2_{\beta_2} + s^2_{\beta_3} - 2s_{\beta_2\beta_3}}$$

To be thorough though all 6 comparisons need to be made and the critical value adjusted via Bonferroni using $z_{1-\alpha/(2\times 6)}$.

Comparison	CI on
medium light vs dark	β_2
medium vs dark	β_3
medium dark vs dark	β_4
medium light vs medium	$\beta_2 - \beta_3$
medium light vs medium dark	$\beta_2 - \beta_4$
medium vs medium dark	$\beta_3 - \beta_4$

- Comparing medium light vs dark, the 95% CI on β_2, the log odds ratio, is

  $$1.2694 \mp 1.96(0.8488) \longrightarrow (-0.3943, 2.9331)$$

  which includes 0, hence CI on odds ratio will include 1.

  $\vdots$

Exercise 4.1 Perform all the CI's mentioned in the previous example.

Example 4.8 For the sake of practice let us compare dark vs non-dark using current model, for a fixed level of weight. Hence a CI on

$$\frac{(\alpha+\beta_2+\beta_1 x)+(\alpha+\beta_3+\beta_1 x)+(\alpha+\beta_4+\beta_1 x)}{3} - (\alpha+\beta_1 x) = \frac{1}{3}\beta_2 + \frac{1}{3}\beta_3 + \frac{1}{3}\beta_4$$

is needed. DO IN CLASS.

Example 4.9 (Florida Death Penalty) http://users.stat.ufl.edu/~athienit/STA4504/Examples/FL_death.R

4.3.2 Quantitative Treatment of Ordinal Factors

Qualitative variables can be

- nominal - no order
- ordinal - order

where ordinal variables can be treated as qualitative or quantitative.

Example 4.10 (Horseshoe crab continued) Consider example 4.3 where 3 dummy variables were created to distinguish the 4 levels of color: medium light, medium, medium dark and dark.
In the context of this problem "darkness" is of interest and hence color is ordinal, so a score can be created to reflect this

Color	Score
Medium Light	1
Medium	2
Medium dark	3
Dark	4

$$\text{logit}[\pi(x)] = \alpha + \beta_1 x + \beta_2 c$$

Referring to the (qualitative) model of example 4.3,

Color	logit $[\pi(x)]$ Qualitative	logit $[\pi(x)]$ Quantitative
medium light	$(\alpha+\beta_2)+\beta_1 x$	$(\alpha+\beta_2)+\beta_1 x$
medium	$(\alpha+\beta_3)+\beta_1 x$	$(\alpha+2\beta_2)+\beta_1 x$
medium dark	$(\alpha+\beta_4)+\beta_1 x$	$(\alpha+3\beta_2)+\beta_1 x$
dark	$\alpha+\beta_1 x$	$(\alpha+4\beta_2)+\beta_1 x$

Note that the qualitative model is a lot more flexible (as it has more parameters) in differentiating between groups, while the quantitative model assumes a systematic change between groups.

```
> linear=unclass(color)  #  convert back to integer levels
> fit2.3=glm(y ~ weight + linear, family=binomial(link=logit))
> summary(fit2.3)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -2.0316     1.1161  -1.820   0.0687 .
weight        1.6531     0.3825   4.322 1.55e-05 ***
linear       -0.5142     0.2234  -2.302   0.0213 *
---
    Null deviance: 225.76  on 172  degrees of freedom
Residual deviance: 190.27  on 170  degrees of freedom
AIC: 196.27
```

Testing the significance of color via $H_0 : \beta_2 = 0$ for this model

- Via normal test, p-value = 0.0213
- Via LRT, $G^2 = 195.74 - 190.27$ with 1 degree of freedom yields, p-value = 0.0193637

Part (II), subpart C of http://users.stat.ufl.edu/~athienit/STA4504/Examples/crab_u.R
To summarize in terms of the LRT for color

Color	df	LRT p-value
Qualitative	3	0.07
Binary (dark vs. non-dark)	1	0.01
Quantitative	1	0.02

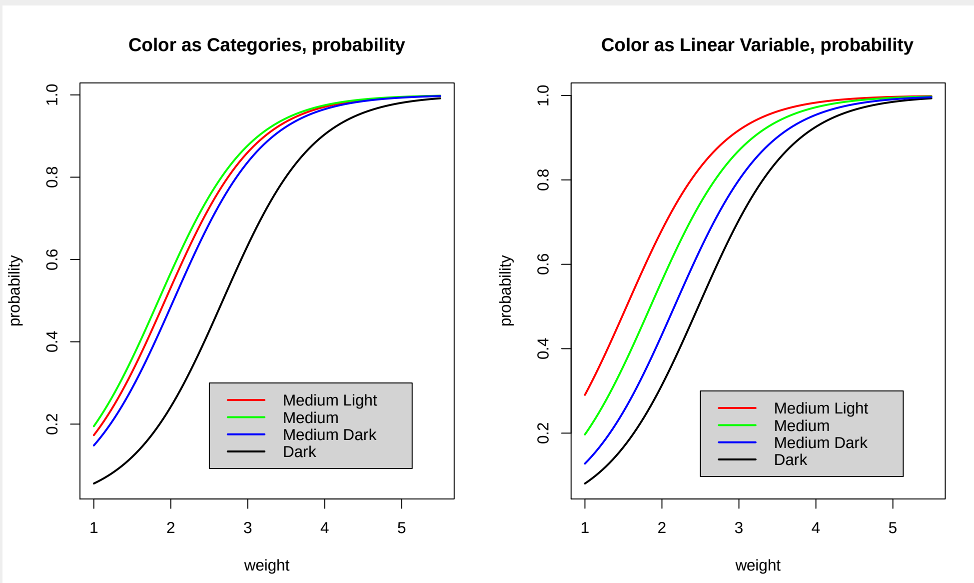


Figure 4.4: Probability curves

Remark 4.2. *To achieve more power in testing factors, it is best to use methodology, i.e. tests, that use fewer degrees of freedom.*

Exercise 4.2 Try to fit a quantitative model with a more representative score than $1, 2, 3, 4$, in order to obtain a p-value (for a LRT less) than 0.0193637

4.4 Summarizing Predictive Power

A naive way of summarizing predictive power is to calculate the correlation between observed responses and fitted responses.

Example 4.11 (Horseshoe crab continued) We look at the correlation between the observed values of $y = 0, 1$ and the fitted probabilities of the logistic regression models.

```
> cor(y,fitted(fit)) # weight
[1] 0.3955277
> cor(y,fitted(fit2)) # weight and color
[1] 0.4476282
> cor(y,fitted(fit2.2)) # weight and binary dark
[1] 0.3958138
> cor(y,fitted(fit2.3)) # weight and linear color
[1] 0.4385387
```

A more sophisticated method, similar to methods learned in other courses, is the (approximate) *leave-one-out cross-validation*, and producing classification tables

1. Fit the model to the data leaving out i^{th} observation

2. Use fitted model and the predictor settings of the i^{th} observation to compute response $\hat{\pi}(\boldsymbol{x}_i)$

3. Predict
$$\hat{y}_i = \begin{cases} 1 & \hat{\pi}(\boldsymbol{x}_i) > 0.50 =: \pi_0 \quad \text{(cutoff probability)} \\ 0 & \hat{\pi}(\boldsymbol{x}_i) \leq 0.50 \end{cases}$$
where the cutoff of 0.50 can be altered.

Example 4.12 (Horseshoe crab continued) Using the model with weight and (qualitative) color we obtain

	Predicted		
Actual	$\hat{y} = 0$	$\hat{y} = 1$	Total
$y = 0$	27	35	62
$y = 1$	17	94	111

$$\text{Sensitivity} = P(\hat{Y} = 1|Y = 1) = \frac{94}{111} \approx 0.847$$
$$\text{Specificity} = P(\hat{Y} = 0|Y = 0) = \frac{27}{62} \approx 0.435$$

and
$$P(\text{correct classification}) = \frac{94 + 27}{173} \approx 0.699$$

Part (III) of `http://users.stat.ufl.edu/~athienit/STA4504/Examples/crab_u.R`

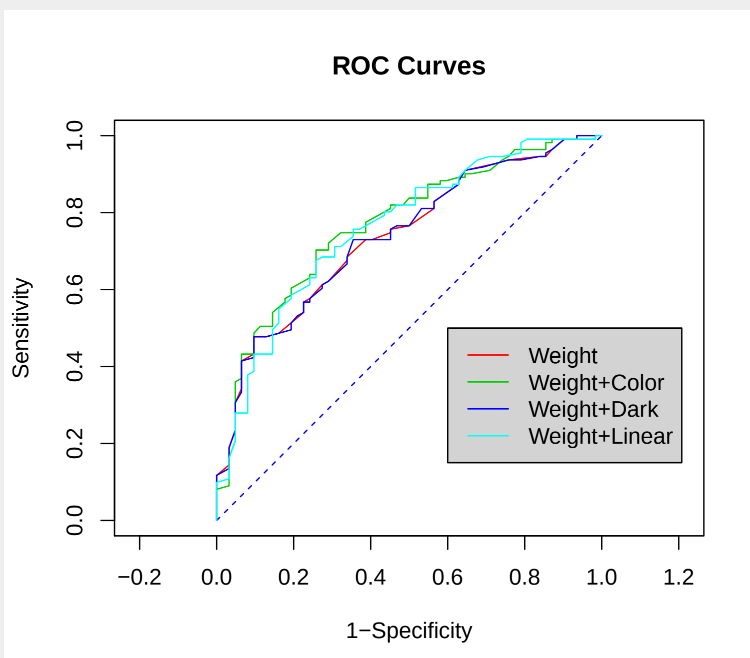
4.5 Receiver Operating Characteristic Curve

The *receiver operating characteristic* (ROC) curve plots the true positive rate, sensitivity, against false positive rate, 1-specificity, as the cutoff value π_0 varies from 0 to 1. It can also be thought of as a plot of the Power as a function of the Type I Error of the decision rule.

- 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 with larger c have better predictive power.
  - When $c = 1/2$ it is no better than random guessing.
- If feasible, use cross-validation.
- ROC curves should not be used with random predictors.

Example 4.13 (Horseshoe crab continued) The concordance indexes for some of the fitted models are

Model	Concordance
Weight	0.738
Weight and Color	0.769
Weight and Dark	0.738
Weight and Linear Color	0.761



Part (IV) of `http://users.stat.ufl.edu/~athienit/STA4504/Examples/crab_u.R`

5. Building Logistic Regression Models

Strategies in model selection and model checking.

5.1 Strategies

5.1.1 AIC and AICc

The Akaike information criterion (AIC) is an estimator of the relative quality of statistical models for a given set of data. Given a collection of models for the data, AIC estimates the quality of each model, relative to each of the other models. Thus, AIC provides a means for model selection.

$$\mathrm{AIC} = 2(k+1) - 2\log(\hat{L})$$

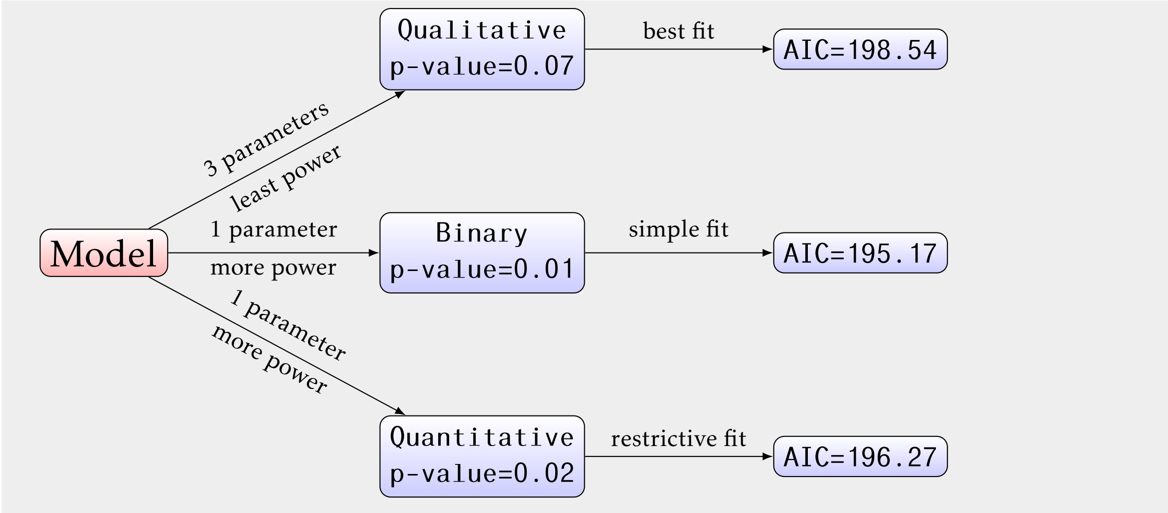
It is comprised of a *"penalizing"* function $2(k+1)$ that penalizes for complicated models with a large k value, and the maximum value of the likelihood function for the model. Hence, smaller values are desirable when comparing models.

When the sample size is small, there is a substantial probability that AIC will select models that have too many parameters. AICc was developed that includes a correction for small sample sizes. The formula for AICc depends upon the statistical model. Assuming that the model is univariate, is linear in its parameters, and has normally-distributed residuals (conditional upon regressors), then the formula for AICc is as follows.

$$\mathrm{AICc} = \mathrm{AIC} + \frac{2(k+1)^2 + 2(k+1)}{n-k-2}$$

Thus, AICc is essentially AIC with an extra penalty term for the number of parameters.

Example 5.1 (Horseshoe crab continued) Results best illustrated via



5.1.2 Multicollinearity

Multicollinearity is a phenomenon in which one predictor variable can be linearly predicted from the other predictors with a substantial degree of accuracy.

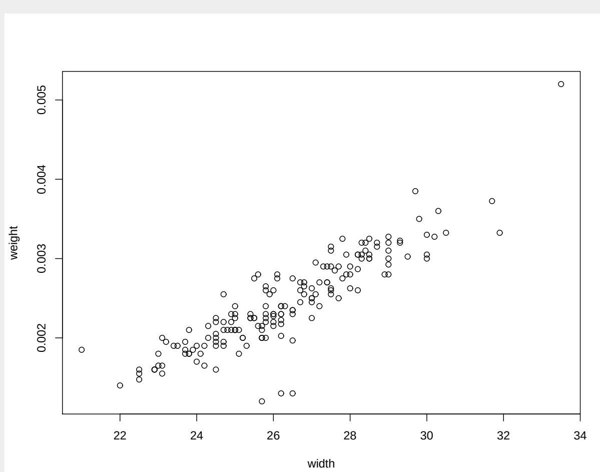
Effects:

- Coefficient estimates may change erratically in response to small changes in the model or the data.
- Coefficient standard errors are inflated.

Multicollinearity does not reduce the predictive power or reliability of the model as a whole, at least within the sample data set; it only affects calculations regarding individual predictors.

Example 5.2 (Horseshoe crab continued) Consider the weight and width of a crab that are likely to be correlated

```
> cor(weight,width)
[1] 0.8868715
```



and we could use either variable. However, we will see in this example it is best to use width.

```
> fit.we=glm(y ~ weight, family=binomial(link=logit))
> summary(fit.we)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -3.6947     0.8802  -4.198 2.70e-05 ***
weight        1.8151     0.3767   4.819 1.45e-06 ***
---
AIC: 199.74

> fit.wi=glm(y ~ width, family=binomial(link=logit))
> summary(fit.wi)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -12.3508     2.6287  -4.698 2.62e-06 ***
width         0.4972     0.1017   4.887 1.02e-06 ***
---
AIC: 198.45

> fit.wewi=glm(y ~ weight+width, family=binomial(link=logit))
> summary(fit.wewi)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.3547     3.5280  -2.652  0.00801 **
weight        0.8338     0.6716   1.241  0.21445
width         0.3068     0.1819   1.686  0.09177 .
---
AIC: 198.89
```

5.1.3 Stepwise Selection Algorithms

There are 3 common types of algorithms

- **Backward** - Start with a full model and *remove* 1 factor/predictor at a time, based on a criterion, until a stopping is reached.
- **Forward** - Start with a reduced simple model and *add* 1 factor/predictor at a time, based on a criterion, until a stopping is reached.
- **Both** - Start with any model (of varying complexity) and at each step add or remove a variable.

Common criteria include (but not limited to)

- AIC
- LRT p-values

Example 5.3 (Horseshoe crab continued) DONE IN CLASS. Part (V) of http://users.stat.ufl.edu/~athienit/STA4504/Examples/crab_u.R

Remark 5.1. *There is a study that suggests* ≥ 10 *outcomes of each type per model predictor (where dummy variables for qualitative predictors are considered individual predictors).*

Example 5.4 (Horseshoe crab) In this example there were 173 crabs, 111 had a male satellite while 62 did not. Hence, choosing the smaller count of the two

$$\frac{62}{10} \approx 6 \text{ predictors}$$

We noticed that a model with the 3-way interaction term was not estimable. In fact, based on this guideline, we probably should be attempting to fit some (if not all) of the 2-way interactions.

5.2 Model Checking

There 3 main ways of checking model fit

- Goodness of fit test. 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.
- Residuals.

Example 5.5 (Florida Death Penalty continued) In this example we will look at the first two points. In example 4.9 you were asked to perform a goodness of fit test as an exercise. Summarizing fit over 8 cells of table:

$$\begin{aligned}
X^2 &= \sum \frac{(\text{observed} - \text{fitted})^2}{\text{fitted}} = 0.20 \\
G^2 &= 2\sum (\text{observed}) \log\left(\frac{\text{observed}}{\text{fitted}}\right) = 0.38 \longleftarrow (\text{Residual Deviance}) \\
df &= \text{num. binomials} - \text{num. model params} = 4 - 3
\end{aligned}$$

For H_0: "model correctly specified", $G^2 = 0.38, df = 1$, p-value $= 0.54$. Hence, no evidence of lack of fit.
The 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". (Try it and look at df).

Remark 5.2.

- *X^2 usually recommended over G^2 for testing goodness of fit. Why?*
- *These tests only appropriate for grouped binary data with most ($\geq 80\%$) of fitted cell counts being "large" (e.g., $\hat{\mu}_i > 5$). In example 4.9 there were a two cells with fitted values of 0.18 and 3.82.*
- *For continuous predictors or many predictors with small fitted values, distributions of X^2 and G^2 are not well approximated by χ^2. For better approximations, try grouping data before applying X^2, G^2.*
  - *Hosmer-Lemeshow test forms groups using ranges of $\hat{\pi}$ values.*
  - *Or can try to group predictor values (if only 1 or 2 predictors).*

For obtaining residuals, notate at setting i of explanatory variables

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

Definition 5.1 **(Pearson residuals)** For a binomial GLM, the Pearson residuals are

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

The distribution of $e_i \overset{\text{approx}}{\sim} N(0, \nu)$ when model holds (and n_i large), but $\nu < 1$.

R code 5.1 Recall we use

```
residuals(model,type=``pearson'')
```

Definition 5.2 **(Standardized Pearson residual)** For a binomial GLM, the *standardized* Pearson residuals are

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

which correct for standard error so that $r_i \overset{\text{approx}}{\sim} N(0,1)$ and h_i is the i-th diagonal element of the "Hat" matrix (not covered here).

Therefore, values of $|r_i| > 2$ suggest a lack of fit.

R code 5.2 The function `rstandard()` provides standardized deviance residuals by default.

- For standardized Pearson residuals specify

```
rstandard(model,type=``pearson'')
```

- Standardized Deviance residuals are the default option

```
rstandard(model)
```

Example 5.6 (Berkeley Graduate Admissions) http://users.stat.ufl.edu/~athienit/STA4504/Examples/admissions.R

5.3 Effects of Sparse Data

As the term suggests, *sparse data* are when certain combinations of variables have no actual data or "limited" information. This can lead to parameter estimates being infinite (in value).

Example 5.7 Consider,

		S	F
X	1	8	2
	0	10	0

Fitting a simple logistic regression will yield the estimates odds ratio

$$e^{\hat{\beta}} = \frac{8 \times 0}{2 \times 10} = 0 \quad \Rightarrow \quad \hat{\beta} = \log(0) = -\infty$$

Infinite estimates exist when predictor values (x values) where $y = 1$ can be *separated* from predictor values where $y = 0$. This extends to multidimensional predictor space.

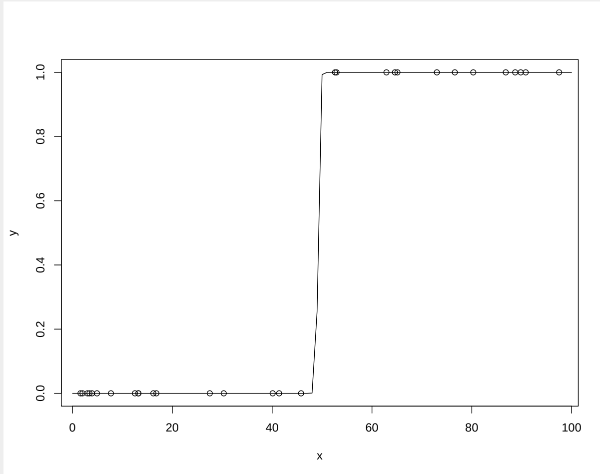
Example 5.8 Let

$$y = \begin{cases} 0 & x < 50 \\ 1 & x > 50 \end{cases}$$

with no values at $x = 50$.
Data were simulated at
http://users.stat.ufl.edu/~athienit/STA4504/Examples/sparse.R

```
> fit=glm(y~x,family=binomial)
Warning messages:
1: glm.fit: algorithm did not converge
2: glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(fit)
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -297.566 174094.706  -0.002    0.999
x                6.051   3542.717   0.002    0.999
---
    Null deviance: 4.1054e+01  on 29  degrees of freedom
Residual deviance: 5.0225e-09  on 28  degrees of freedom
AIC: 4
```

where although $\hat{\beta} = 6.051$ the standard error is 3542.717.



This is because the likelihood function has no point of inflection, that is, it keeps increasing as $\beta \uparrow$.

6. Multicategory Logit Models

Extensions of logistic regression for nominal and ordinal responses.

6.1 Logit Models for Nominal Responses

Let

$$\pi_j = P(Y = j), \quad j = 1, 2, \ldots, J$$

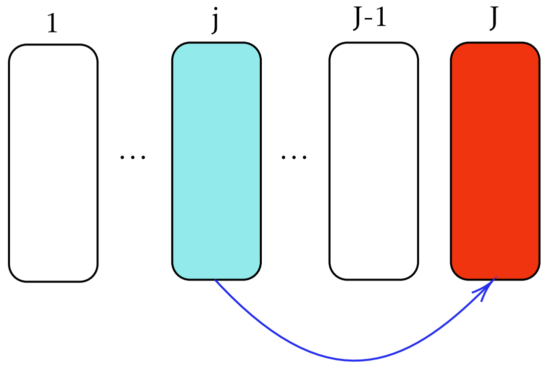
Conisder a binomial with two groups and two probabilities, $\pi_1, \pi_2 \ni \pi_1 + \pi_2 = 1$. A simple logistic model was

$$\log\left(\frac{\pi_1}{1-\pi_1}\right) = \log\left(\frac{\pi_1}{\pi_2}\right) = \alpha + \beta x$$

Baseline-category logits are similar but have the form

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

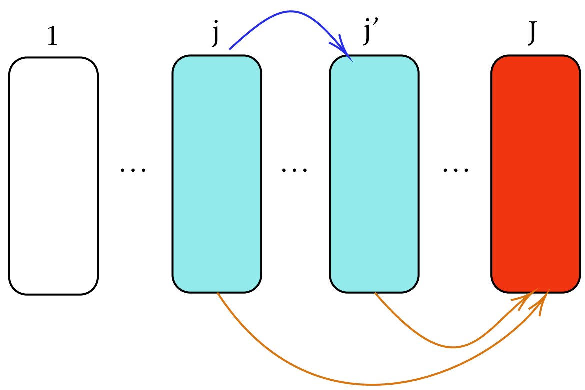
There is seperate set of parameters (α_j, β_j) for each logit. We compare the probability of being in group j, versus the baseline group J.



Hence,

$$\pi_j = \frac{e^{\alpha_j+\beta_j x}}{1+\sum_{i=1}^{J-1} e^{\alpha_i+\beta_i x}}, \qquad \pi_J = \frac{1}{1+\sum_{i=1}^{J-1} e^{\alpha_i+\beta_i x}}$$

but we can compare any two groups that where one group is not the baseline



$$\begin{aligned}\log\left(\frac{\pi_j}{\pi_{j'}}\right) &= \log\left(\frac{\pi_j/\pi_J}{\pi_{j'}/\pi_J}\right)\\ &= \log\left(\frac{\pi_j}{\pi_J}\right) - \log\left(\frac{\pi_{j'}}{\pi_J}\right)\\ &= (\alpha_j - \alpha_{j'}) + (\beta_j - \beta_{j'})x\end{aligned}$$

- Category used as baseline (i.e., category J) is arbitrary and does not affect model fit, since categories are nominal.
- The term e^{β_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.
- Could also use this model with ordinal response variables, but this would ignore information about ordering.

Example 6.1 (Job Satisfaction) Data 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

Consider x = income scores $(3, 10, 20, 30)$ and define VD=1, LD=2, MS=3, VS=4

```
> fit.blogit=vglm(cbind(VD,LD,MS,VS)~income,family=multinomial,data=dat)
> summary(fit.blogit)
Coefficients:
```

```
                 Estimate Std. Error  z value
(Intercept):1  0.563824   0.960138  0.58723
(Intercept):2  0.645091   0.668771  0.96459
(Intercept):3  1.818698   0.528955  3.43828
income:1      -0.198773   0.102096 -1.94693
income:2      -0.070502   0.036954 -1.90785
income:3      -0.046918   0.025519 -1.83858

Residual deviance: 4.17662 on 6 degrees of freedom
Log-likelihood: -16.71316 on 6 degrees of freedom
```

The prediction equations are

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_4}\right) = 0.564 - 0.199x$$
$$\log\left(\frac{\hat{\pi}_2}{\hat{\pi}_4}\right) = 0.645 - 0.071x$$
$$\log\left(\frac{\hat{\pi}_3}{\hat{\pi}_4}\right) = 1.819 - 0.047x$$

For each logit, the odds of being in a less satisfied category (instead of "very satisfied") decreases as income increases. ML estimates determine the effects for all pairs of categories. For example, comparing group 1 and 2, i.e. "dissatisfied" to "little dissatisfied"

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) = \log\left(\frac{\hat{\pi}_1}{\hat{\pi}_4}\right) - \log\left(\frac{\hat{\pi}_2}{\hat{\pi}_4}\right) = -0.081 - 0.128x$$

A global test of income effect is $\mathrm{H}_0 : \beta_1 = \beta_2 = \beta_3 = 0$.

```
> vglm(cbind(VD,LD,MS,VS)~1,family=multinomial,data=dat)
...
Degrees of Freedom: 12 Total; 9 Residual
Residual deviance: 13.4673
```

and hence

$$G^2 = 13.4673 - 4.17662 \quad df = 3 \quad \text{p-value of} 0.0257$$

http://users.stat.ufl.edu/~athienit/STA4504/Examples/jobsatis.R

Exercise 6.1 For the job satisfaction example, we obtained the logit for comparing "dissatisfied" to "little dissatisfied" to be

$$\log\left(\frac{\hat{\pi}_1}{\hat{\pi}_2}\right) = -0.081 - 0.128x$$

where $\hat{\beta}_1 - \hat{\beta}_2$=-0.128. Create a 95% confidence interval around $\beta_1 - \beta_2$ and interpret.

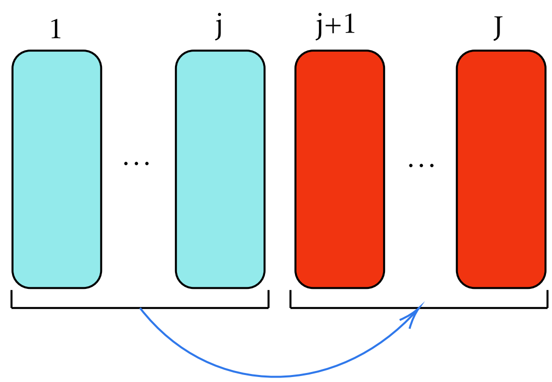
6.2 Cumulative Logit Models for Ordinal Responses

The cumulative logit probabilities are

$$P(Y \leq j) = \sum_{i=1}^{j} \pi_j, \quad j = 1, \ldots, J$$

and the *cumulative logit* model is

$$\begin{aligned} \text{logit}\left[P(Y \leq j)\right] &= \log\left(\frac{P(Y \leq j)}{1 - P(Y \leq j)}\right) \\ &= \log\left(\frac{P(Y \leq j)}{P(Y > j)}\right) \\ &= \alpha_j + \beta x, \quad j = 1, \ldots, J-1 \end{aligned}$$



$$P(Y \leq j) = \frac{e^{\alpha_j + \beta x}}{1 + e^{\alpha_j + \beta x}}, \quad j = 1, 2, \ldots J-1$$

- Separate intercept α_j for each cumulative logit
- Same (slope) coefficient β for each cumulative logit.
- The term e^{β} = multiplicative effect of 1-unit increase in x on odds that $(Y \leq j)$ instead of $(Y > j)$.

$$\begin{aligned} \frac{\text{odds}(Y \leq j | x_2)}{\text{odds}(Y \leq j | x_1)} &= \frac{e^{\alpha_j + \beta x_2}}{e^{\alpha_j + \beta x_1}} \\ &= e^{\beta(x_2 - x_1)} \\ &= e^{\beta}, \quad \text{when } x_2 = x_1 + 1 \end{aligned}$$

Also called *proportional odds* model.

Example 6.2 (Job Satisfaction continued) The model has form

$$\text{logit}\left[P(Y \leq j | x)\right] = \alpha_j + \beta x \quad j = 1, 2, 3$$

```
> fit.clogit1=vglm(cbind(VD,LD,MS,VS)~income,
+ family=cumulative(parallel=TRUE),data=dat)
> summary(fit.clogit1)
Coefficients:
               Estimate Std. Error z value
(Intercept):1 -2.473156   0.568376 -4.3513
(Intercept):2 -0.781728   0.373724 -2.0917
(Intercept):3  2.211091   0.445123  4.9674
income        -0.056347   0.020871 -2.6998

Residual deviance: 5.9527 on 8 degrees of freedom
Log-likelihood: -17.60121 on 8 degrees of freedom
```

The fitted model is

$$\text{logit}\left[\hat{P}(Y \le j|x)\right] = \hat{\alpha}_j - 0.056x \quad j = 1, 2, 3.$$

Hence the odds of response at low end of job satisfaction scale decrease as x increases, i.e. $\exp(-0.056) = 0.95$. Estimated odds of job satisfaction below any given level (instead of above it) multiply by 0.95 for a 1-unit increase in x (1-unit=\$1000). For a \$10,000 increase in income, the estimated odds multiply by $\exp(10(-0.056)) = 0.57$. (If we were to reverse the order of the responses, then $\hat{\beta} = +0.056$).

Odds ratio is the same between *same* two categories of x irrespective of cutoff region for response categories (to make response binary) as shown in the diagrams in the class notes.

In addition, the odds ratio is the same between categories $x = 10$ and $x = 20$, and $x = 20$ and $x = 30$ due to the same increment in x.

A goodness of fit test yields a p-value of

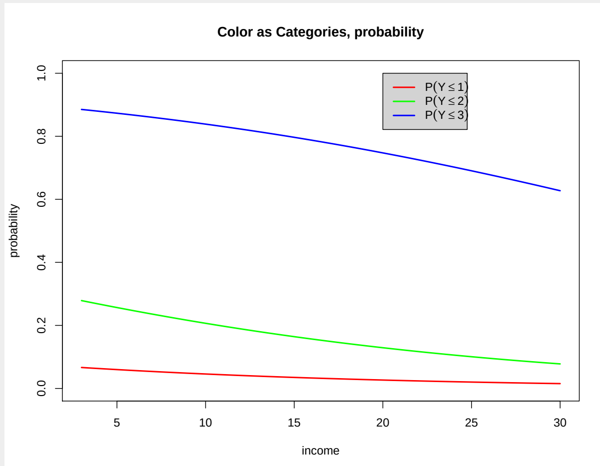
```
> 1-pchisq(deviance(fit.clogit1),df.residual(fit.clogit1))
[1] 0.6525305
```

so we conclude that the model is a good fit.
A test of H_0 : job satisfaction independent of income (i.e. $\beta = 0$ in cumulative logit model) yields

- A Wald z-stat of -2.6998 (or χ^2 of 7.17) and a p-value of 0.007.
- A LR statistic of $13.4673 - 5.9527 = 7.5146$ on 1 df and a p-value of 0.006. The null deviance was computed using

  ```
  > vglm(cbind(VD,LD,MS,VS)~1,
  + family=cumulative(parallel=TRUE),data=dat)
  Coefficients:
  (Intercept):1 (Intercept):2 (Intercept):3
      -3.218876     -1.563976      1.258955
  ```

```
Degrees of Freedom: 12 Total; 9 Residual
Residual deviance: 13.4673
Log-likelihood: -21.35851
```



A model with nonparallel lines, i.e. different β_j for $j = 1,2,3$ instead of one common slope, if fit but it does not differ from the parallel lines model (conclusion via LR test stat).

```
> fit.clogit2=vglm(cbind(VD,LD,MS,VS)~income,
+ family=cumulative(parallel=FALSE),data=dat)
> summary(fit.clogit2)
Coefficients:
              Estimate Std. Error z value
(Intercept):1 -1.74105   0.816828 -2.1315
(Intercept):2 -0.82432   0.449753 -1.8328
(Intercept):3  2.20524   0.515114  4.2811
income:1      -0.14443   0.091070 -1.5860
income:2      -0.05356   0.029750 -1.8003
income:3      -0.05603   0.024771 -2.2619

Residual deviance: 4.37717 on 6 degrees of freedom
Log-likelihood: -16.81344 on 6 degrees of freedom
```

- To test $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ via LRT, we use

  ```
  > 1-pchisq(13.4673-4.37717,3)
  [1] 0.02811625
  ```

  and conclude that at least one of the β's is significant.

- To test $H_0 : \beta_1 = \beta_2 = \beta_3$ via LRT, that is comparing the "parallel" model to the "non-parallel", we use

  ```
  > 1-pchisq(5.9527-4.37717,2)
  [1] 0.4548603
  ```

and conclude that at wwe should be using one common β, i.e. the “parallel” model.

http://users.stat.ufl.edu/~athienit/STA4504/Examples/jobsatis.R

Exercise 6.2 Instead of testing $H_0 : \beta_1 = \beta_2 = \beta_3$ via LRT, to determine whether to use the “non-parallel” model, obtain the AIC for each model, compare and conclude.

Example 6.3 (Political Ideology) An example with the following data yields

```
> ideow
  Gender      Party VLib SLib Mod SCon VCon
1 Female   Democrat   44   47 118   23   32
2 Female Republican   18   28  86   39   48
3   Male   Democrat   36   34  53   18   23
4   Male Republican   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)
Coefficients:
                Estimate Std. Error   z value
(Intercept):1   -1.45177    0.12284 -11.81819
(Intercept):2   -0.45834    0.10577  -4.33337
(Intercept):3    1.25499    0.11455  10.95598
(Intercept):4    2.08904    0.12916  16.17374
GenderMale      -0.11686    0.12681  -0.92147
PartyRepublican -0.96362    0.12936  -7.44917

Residual deviance: 15.05557 on 10 degrees of freedom
Log-likelihood: -47.41497 on 10 degrees of freedom
```

- First we perform a goodness of fit test with $G^2 = 15.056$ and 10 degrees of freedom to obtain a p-value of 0.13
- Testing for party effect (controlling for gender) we have
  - Wald: $z = -7.449$
  - LR: $71.902 - 15.056 = 56.846$ with df= 1. (Deviance of 71.902 was obtained by fitting model with only gender effect)

  Strong evidence that Republicans tend to be less liberal (more conservative) than Democrats (for each gender).

  Controlling for gender, estimated odds that a Republican’s response (i.e. going from $x_2 = 0$ to $x_2 = 1$, a 1-unit increase) is in liberal direction $(Y \le j)$ rather than conservative $(Y > j)$ are $\exp(-0.964) = 0.38$ times estimated odds for a Democrat.

(Equivalently, controlling for gender, estimated odds that a Democrat's response is in liberal direction rather than conservative exp(0.964) = 2.62 times estimated odds for a Republican.) The 95% CI for the odds ratio is (but best to use `confint`)

$$\exp(-0.964 \pm 1.96(0.129)) \to (0.30, 0.49)$$

- Testing for gender effect (controlling for party) we have a Wald statistic -0.921 indicating a lack of evidence.

However, before we simply drop the gender effect, we know from a previous example that there is a relationship between gender and party affiliation (see party affiliation example). It makes sense that an interaction may be present.

```
> ideo.cl2=vglm(cbind(VLib,SLib,Mod,SCon,VCon) ~ Gender*Party,
+               family=cumulative(parallel=TRUE), data=ideow)
> summary(ideo.cl2)
Coefficients:
                           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
PartyRepublican            -0.75621    0.16691  -4.53062
GenderMale:PartyRepublican -0.50913    0.25408  -2.00381

Residual deviance: 11.06338 on 9 degrees of freedom
Log-likelihood: -45.41887 on 9 degrees of freedom
```

Notice that the interaction term appears significant.

- Wald: $z = -2.004$ with p-value=0.04507
- LR: $15.056 - 11.063 = 3.993$ with df=1 and p-value=0.0457.

The goodness of fit test with $G^2 = 11.063$ residual deviance and df=9 wields a p-value of 0.2714153, a big improvement from 0.13 for the additive model. This is because the interaction takes into account the relationship between gender and party affiliation and how they affect political ideology.

Interpretation:

- Odds ratio
  - Estimated odds ratio for party effect (x_2), (allowing gender to differ) is

$$\begin{aligned} \exp(b_2) = \exp(-0.756) = 0.47 &\quad \text{when} x_1 = 0 \text{ (F)} \\ \exp(b_2 + b_3) = \exp(-0.756 - 0.509) = 0.28 &\quad \text{when} x_1 = 1 \text{ (M)} \end{aligned}$$

    * Estimated odds that a female Republican's response is in liberal direction rather than conservative are 0.47 times estimated odds for a female Democrat.

    * Estimated odds that a male Republican's response is in liberal direction rather than conservative are 0.28 times estimated odds for a male Democrat.

  - Estimated odds ratio for gender effect (x_1) is

$$\begin{aligned} \exp(b_1) = \exp(0.143) = 1.15 &\quad \text{when} x_2 = 0 \text{ (Dem)} \\ \exp(b_1 + b_3) = \exp(0.143 - 0.509) = 0.69 &\quad \text{when} x_2 = 1 \text{ (Rep)} \end{aligned}$$

    * Estimated odds that a male Democrat's response is in liberal direction rather than conservative are 1.15 times estimated odds for a female Democrat.
    * Estimated odds that a male Republican's response is in liberal direction rather than conservative are 0.69 times estimated odds for a female Republican.

- Probabilities

$$\hat{P}(Y \le j) = \frac{\exp(\hat{\alpha}_j + 0.143x_1 - 0.756x_2 - 0.509x_1x_2)}{1 + \exp(\hat{\alpha}_j + 0.143x_1 - 0.756x_2 - 0.509x_1x_2)}$$

  - $\hat{P}(Y = 1) = \hat{P}(Y \le 1)$. For $j = 1$ (very liberal) the probability for a male republican ($\hat{\alpha}_1 = -1.55, x_1 = 1, x_2 = 1$):

$$\hat{P}(Y = 1) = \frac{e^{-2.67}}{1 + e^{2.67}} = 0.065$$

  - Similarly, $\hat{P}(Y = 2) = \hat{P}(Y \le 2) - \hat{P}(Y \le 1)$, etc.
    Note $\hat{P}(Y = 5) = \hat{P}(Y \le 5) - \hat{P}(Y \le 4) = 1 - \hat{P}(Y \le 4)$.

http://users.stat.ufl.edu/~athienit/STA4504/Examples/pol_ideology.R

Exercise 6.3 Check the (cumulative probability conditions) whether a model with "non-parallel" systematic component is feasible.

8. Models for Matched Pairs

Methods for comparing categorical responses for two samples that have a natural pairing between each subject in one sample and a subject in the other sample.

8.1 McNemar's Test

Methods so far (e.g., X^2 and G^2 test of independence, CI for odds ratio, logistic regression) assume independent samples. Inappropriate for dependent samples (e.g., same subjects in each sample yielding matched pairs of responses).

Example 8.1 (Crossover Study: Drug vs Placebo) Consider 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

To reflect the dependence and looking at the full information

		Placebo S	Placebo F	
Drug	S	12	49	61
	F	10	15	25
		22	64	86

		Placebo S	Placebo F	
Drug	S	π_{11}	π_{12}	π_{1+}
	F	π_{21}	π_{22}	π_{2+}
		π_{+1}	π_{+2}	1

Definition 8.1 (Marginal Homogeneity) There is *marginal homogeneity* if

$$\pi_{1+} = \pi_{+1} \Leftrightarrow \pi_{12} = \pi_{21}$$

since

$$\pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21}$$

Under H_0: marginal homogeneity

$$\frac{\pi_{12}}{\pi_{12} + \pi_{21}} = \frac{1}{2}$$

and each of $n^\star = n_{12} + n_{21}$ observations has probability 1/2 of contributed to n_{12} and 1/2 of contributing to n_{21}

$$n_{12} \sim \text{Bin}(n^\star, 0.5) \quad \Rightarrow \quad z = \frac{n_{12} - n^\star/2}{\sqrt{n^\star\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \overset{\text{approx.}}{\sim} N(0,1)$$

and finding the two sided p-value as usual. However, using the normal approximation to the binomial we are assuming that $n^\star(1/2) > 5$. Some authors suggest > 10 or even > 25. Equivalent to a z-test you may see

$$z^2 = \frac{(n_{12} - n_{21})^2}{n_{12} + n_{21}} \sim \chi_1^2 \equiv [N(0,1)]^2$$

and the p-value being the area to the right (because we squared, only nonnegative values possible). To create a $100(1-\alpha)\%$ confidence interval for $\pi_{1+} - \pi_{+1}$ use

$$\underbrace{p_{1+} - p_{+1}}_{\frac{n_{12} - n_{21}}{n}} \mp z_{1-\alpha/2} \frac{1}{n} \sqrt{n_{12} + n_{21} - \frac{(n_{12} - n_{21})^2}{n}}$$

Remark 8.1. Depending on the situation, such as, if it is desirable n_{12} to be large then a 1-sided test of CI might yield some gain in power.

- *Hypothesis $H_a : \pi_{12} > \pi_{21}$, p-value=$P(Z \geq z)$ area to the right (using normal distribution).*
- *CI, use $+z_{1-\alpha}$*

R code 8.1 Use

```
mcnemar.test(x, y = NULL, correct = TRUE)
```

The continuity correction for using a continuous distribution to approximate the discrete binomial, is the default setting. Also recommended to use `mcnemar.exact{exact2x2}` which uses the exact Binomial test and does not require $n^\star(1/2) > 5$.

Example 8.2 (Crossover Study: Drug vs Placebo continued) Looking at the data again,

		Placebo S	Placebo F	
Drug	S	12	49	61
	F	10	15	25
		22	64	86

and

$$z = \frac{49-10}{\sqrt{49+10}} = 5.1 \qquad \text{and p-value} < 0.0001$$

Extremely strong evidence that probability of success is higher for drug than placebo. The 95% CI for $\pi_{1+} - \pi_{+1}$ is

$$\frac{49}{86} - \frac{10}{86} \mp 1.96 \frac{1}{86}\sqrt{49+10-\frac{(49-10)^2}{86}} \longrightarrow (0.31, 0.60)$$

and hence the probability of success under drug is larger than that under placebo.

```
> 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

> require(exact2x2)
> mcnemar.exact(crossover)

Exact McNemar test (with central confidence intervals)

data:  crossover
b = 49, c = 10, p-value = 2.706e-07
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
  2.451984 10.849724
sample estimates:
odds ratio
       4.9
```

Part (A) of http://users.stat.ufl.edu/~athienit/STA4504/Examples/crossover_gee.R

Remark 8.2. The derivation of the standard error for the CI is derived by the fact that

$$(n_{11}, n_{12}, n_{21}, n_{22}) \sim MN(n, \{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\})$$

and hence

$$V(n_{ij}) = n\pi_{ij}(1-\pi_{ij})$$
$$\mathrm{Cov}(n_{ij}, n_{i'j'}) = -n\pi_{ij}\pi_{i'j'} \quad (i \neq i' \text{ or } j \neq j')$$

Therefore,

$$\begin{aligned} V(p_{1+} - p_{+1}) &= V\left(\frac{n_{12}-n_{21}}{n}\right) = \frac{1}{n^2}V(n_{12}-n_{21}) \\ &= \frac{1}{n^2}[V(n_{12}) + V(n_{21}) - 2\mathrm{Cov}(n_{12}, n_{21})] \\ &= \cdots \\ &= \frac{1}{n}\left[\pi_{12} + \pi_{21} - (\pi_{12}-\pi_{21})^2\right] \end{aligned}$$

and hence

$$\hat{V}(p_{1+} - p_{+1}) = \cdots = \frac{1}{n^2}\left[n_{12} + n_{21} - \frac{(n_{12}-n_{21})^2}{n}\right]$$

Remark 8.3. [McNemarBowker tests.] For larger than 2×2 tables, $k \times k$ tables, McNemar's test is generalized as the McNemar-Bowker symmetry test for testing

$$H_0 : \pi_{ij} = \pi_{ji}, \quad \text{for all pairs.}$$

However, it may fail if there are 0's in certain locations in the matrix.

R code 8.2 Use nominalSymmetryTest{rcompapion}

```
nominalSymmetryTest(x, method = "fdr", digits = 3, ...)
```

For examples see https://rcompanion.org/handbook/H_05.html

8.2 Rater Agreement

In this section we wish to determine if two raters/reviewers are in agreement or not.

Example 8.3 (Movie reviews) Two movie reviewers give their opinion on 160 movies

	Reviewer 2			
Reviewer 1	Con	Mixed	Pro	Total
Con	24	8	13	45
Mixed	8	13	11	32
Pro	10	9	64	83
Total	42	30	88	160

8.2.1 Cohen's Kappa (unweighted)

Let $\pi_{ij} = P(R1 = i, R2 = j)$,

$$
\begin{aligned}
P(agree) &= \sum_i \pi_{ii} && \text{general case} \\
&= \sum_i \pi_{i+}\pi_{+i} && \text{if independence}
\end{aligned}
$$

Definition 8.2 (Cohen's Kappa)

$$\kappa = \frac{\sum_i \pi_{ii} - \sum_i \pi_{i+}\pi_{+i}}{1 - \sum_i \pi_{i+}\pi_{+i}}$$

where

- $\kappa = 0$ if agreement only equals that expected under independence.
- $\kappa = 1$ if perfect agreement.
- Denominator = maximum difference for numerator, attained if agreement is perfect, since perfect agreement implies $\sum_i \pi_{ii} = 1$.
- It is possible for the statistic to be negative, which implies that there is no effective agreement between the two raters or the agreement is worse than random.

Asymptotic normality can be established

$$\hat{\kappa} \overset{\mathrm{H}_0}{\sim} N\left(0, V(\hat{\kappa})\right)$$

and hence the standard error must first be found. Let,

- $\hat{\pi}_0 = \sum_i \hat{\pi}_{ii}$
- $\hat{\pi}_c = \sum_i \hat{\pi}_{i+}\hat{\pi}_{+i}$

$$
\begin{aligned}
\hat{V}(\hat{\kappa}) = \frac{1}{n(1-\hat{\pi}_c)^4} \Bigg\{ &\sum_i \hat{\pi}_{ii}\left[(1-\hat{\pi}_0) - (\hat{\pi}_{+i} + \hat{\pi}_{i+})(1-\hat{\pi}_0)\right]^2 \\
&+ (1-\hat{\pi}_0)^2 \sum\sum_{i \neq j} \hat{\pi}_{ij}(\hat{\pi}_{+i} + \hat{\pi}_{i+})^2 - (\hat{\pi}_0\hat{\pi}_c - 2\hat{\pi}_c + \hat{\pi}_0)^2 \Bigg\}
\end{aligned}
$$

R code 8.3 In R there are multiple packages such as `irr`, `psych`, `concord` that have their own functions and their own *weight* scheme.
We will use `cohen.kappa{psych}`.

Example 8.4 (Movie reviews continued) From the data,

- $\sum_i \hat{\pi}_{ii} = \frac{24+13+64}{160} = 0.63$
- $\sum_i \hat{\pi}_{i+}\hat{\pi}_{+i} = \frac{1}{160^2}(45 \times 42 + 32 \times 30 + 83 \times 88) = 0.40$

$$\hat{\kappa} = \frac{0.63 - 0.40}{1 - 0.40} = 0.39$$

Moderate agreement: difference between observed agreement and agreement expected under independence is about 40% of the maximum possible difference.

Inference To test $H_0 : \kappa = 0$

- Create test statistic
$$\frac{\hat{\kappa} - 0}{0.06} = 6.49$$
with a small p-value when finding the two tails on a $N(0,1)$.

- Create 95% CI
$$\hat{\kappa} \mp (1.96)(0.06) \quad \longrightarrow \quad (0.27, 0.51)$$

Calculation of standard error is left to software

```
> movie=matrix(c(24,8,10,8,13,9,13,11,64),3,3)
> dimnames(movie)=list(c("Con","Mixed","Pro"),c("Con","Mixed","Pro"))
> print(movie)
      Con Mixed Pro
Con    24     8  13
Mixed   8    13  11
Pro    10     9  64
>
> library(psych)
> cohen.kappa(movie)
Cohen Kappa and Weighted Kappa correlation coefficients
and confidence boundaries
                 lower estimate upper
unweighted kappa  0.27     0.39  0.51
weighted kappa    0.32     0.46  0.60

 Number of subjects = 160

> sqrt(cohen.kappa(movie)$var.kappa)
[1] 0.05979313
```

http://users.stat.ufl.edu/~athienit/STA4504/Examples/cohen_kappa.R

8.2.2 Cohen's Kappa (weighted)

Weighted kappa lets you count disagreements differently and is especially useful when codes are ordered. Three matrices are involved

- the matrix of observed scores, n_{ij}
- the matrix of expected scores based on independence, $m_{ij} = n_{i+}n_{+j}$,
- the weight matrix w_{ij}

Derivations of weighted kappa are sometimes expressed in terms of similarities, and sometimes in terms of dissimilarities. In the latter case, the weights on the diagonal are 1 and the weights off the diagonal are less than one. We omit the calculation and use software.

Example 8.5 (Movie reviews continued) Performing both unweighted and weighted versions

```
> cohen.kappa(movie)
Cohen Kappa and Weighted Kappa correlation coefficients
and confidence boundaries
                 lower estimate upper
unweighted kappa  0.27     0.39  0.51
weighted kappa    0.32     0.46  0.60

 Number of subjects = 160
```

with weight matrix

```
> cohen.kappa(movie)$weight
       Con Mixed  Pro
Con   1.00  0.75 0.00
Mixed 0.75  1.00 0.75
Pro   0.00  0.75 1.00
```

Notice that cells with 0.75 although they represent disagreement it is not as severe as disagreements with 0 weight.
http://users.stat.ufl.edu/~athienit/STA4504/Examples/cohen_kappa.R

Exercise 8.1 In cohen.kappa{psych} you can also create your own custom weights as an argument to the function. Repeat the previous example but use 0.5 instead on 0.75 in the weight matrix.

9. Models for Correlated, Clustered Responses

Expanding matched pairs to multiple matched sets, i.e. repeated measures.

9.1 Introduction

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, etc).
  - Fetuses from the same litter.

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.

We will use two approaches to such data: Observations $(Y_1, Y_2, \ldots, Y_T)$

- (In this chapter) Generalized Estimating Equations (GEE) to simultaneously fit marginal models on each (marginal) $E(Y_t), t = 0, \ldots T$.
- (in the next chapter) Generalized Linear Mixed Models (GLMM) to find random effect for the subject/block effect.

9.2 Generalized Estimating Equations

Focusing on GEE for Repeated Measures.

- Specify model in usual way by deciding what the random, component, link function and systematic components are.

- Select a *working correlation* matrix for best guess about correlation pattern between pairs of observations. That is the within-cluster correlation.

Example 9.1 For T repeated responses, *exchangeable* correlation matrix is

Time	1	2	$\cdots$	T
1	1	ρ	$\cdots$	ρ
2	ρ	1	$\cdots$	ρ
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
T	ρ	ρ	$\cdots$	1

When there is positive within-cluster correlation (as often is the case):

- The standard errors for *between-cluster* effects (such as different treatment groups) and standard errors of estimated means within clusters tends to be larger than when independent.
- The standard errors for *within-cluster* effects, such as a slope for a trend in the repeated measurements in a subject, tend to be smaller than when observations are independent.

Fitting method gives estimates that are consistent even if correlation structure is miss-specified. Adjusts standard errors to reflect actual observed dependence. Therefore, overly complicated structures are not encouraged. For other structures the reader is encouraged to review the literature.

Example 9.2 (Crossover Study: Drug vs Placebo continued) Going back to example 8.1

		Placebo S	Placebo F	
Drug	S	12	49	61
	F	10	15	25
		22	64	86

Fit the model

$$\operatorname{logit}[P(Y_t = 1)] = \alpha + \beta d, \qquad d = \begin{cases} 1 & \text{drug} \\ 0 & \text{placebo} \end{cases}$$

where $t = 1, 2$ represents the two time points, the two observations on each subject.

```
> head(crossm1)
  Subject   Treat Resp
1       1    Drug    1
2       1 Placebo    1
3       2    Drug    1
4       2 Placebo    1
5       3    Drug    1
6       3 Placebo    1
> tail(crossm1)
```

```
    Subject   Treat Resp
167      84    Drug    0
168      84 Placebo    0
169      85    Drug    0
170      85 Placebo    0
171      86    Drug    0
172      86 Placebo    0
> cross.gee1=gee(Resp ~ Treat, id=Subject, data=crossm1,family=binomial,
+ corstr="exchangeable")
> summary(cross.gee1)
 GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA

Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Exchangeable

Coefficients:
             Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) -1.067841  0.2485923 -4.295550   0.2471428 -4.320744
TreatDrug    1.959839  0.3798402  5.159639   0.3772338  5.195289

Estimated Scale Parameter:  1.011765

Working Correlation
           [,1]       [,2]
[1,]  1.0000000 -0.2140746
[2,] -0.2140746  1.0000000
```

Therefore, odds of Success with drug is estimated to be $e^{1.96} = 7.1$ times odds with placebo. The 95% CI for odds ratio (for marginal probabilities) is

$$e^{1.96 \mp (1.96)(0.377)} \longrightarrow (e^{1.22}, e^{2.70}) = (3.4, 14.9)$$

Part (B) of http://users.stat.ufl.edu/~athienit/STA4504/Examples/crossover_gee.R

Remark 9.1. *With $\hat{\rho} \approx 0$ it implies that there is no significant correlation between the "clustered" responses.*

Remark 9.2. *With cross-over designs it is important to allow enough time for the effects of the previous treatment not influence the results of the next treatment the unit will cross-over to.*

Remark 9.3. *With GEE approach, can also have "between-subject" explanatory variables. In* *the Drug vs Placebo, d was a variable monitored "within-subject" but we could have monitored "between-subject" gender and even order of treatment, e.g.*

$$sequence = \begin{cases} 1 & placebo\ then\ drug \\ 2 & drug\ then\ placebo \end{cases}$$

GEE is a known as *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 variables.

Example 9.3 (Depression) Consider the response on mental depression (normal, abnormal) measured three times (after 1, 2, and 4 weeks of treatment) with two drug treatments (standard, new) and two severity of initial diagnosis groups (mild, severe). Of interest is to find out if 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

Let

Y_t = response of randomly selected subject at time t (1 = normal, 0 = abnormal)

s = severity of initial diagnosis (1 = severe, 0 = mild)

d = drug (1 = new, 0 = std)

t = time (0, 1, 2), which is log2(weeks of trt)

Model:

$$\log\left[\frac{P(Y_t = 1)}{P(Y_t = 0)}\right] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (dt)$$

so that

$$\log\left[\frac{P(Y_t = 1)}{P(Y_t = 0)}\right] = \begin{cases} \alpha + \beta_1 s + \beta_3 t & \text{if } d = 0 \text{(standard drug)} \\ \alpha + \beta_2 + \beta_1 s + (\beta_3 + \beta_4)t & \text{if } d = 1 \text{(new drug)} \end{cases}$$

```
> dep.gee1=gee((response == "normal") ~ severity + drug*time,
+  id=subject, data=depression, family=binomial, corstr="exchangeable",
```

```
+  contrasts=list(drug=contr.treatment(2,base=2,contrasts=TRUE)))
> summary(dep.gee1)
Model:
 Link:                      Logit
 Variance to Mean Relation: Binomial
 Correlation Structure:     Exchangeable

Coefficients:
                  Estimate Naive S.E.    Naive z Robust S.E.   Robust z
(Intercept)    -0.02809866  0.1625499 -0.1728617   0.1741791 -0.1613205
severitysevere -1.31391033  0.1448627 -9.0700417   0.1459630 -9.0016667
drug1          -0.05926689  0.2205340 -0.2687427   0.2285569 -0.2593091
time            0.48246420  0.1141154  4.2278625   0.1199383  4.0226037
drug1:time      1.01719312  0.1877051  5.4191018   0.1877014  5.4192084

Estimated Scale Parameter:  0.985392
Number of Iterations:  5

Working Correlation
             [,1]         [,2]         [,3]
[1,]  1.000000000 -0.003432729 -0.003432729
[2,] -0.003432729  1.000000000 -0.003432729
[3,] -0.003432729 -0.003432729  1.000000000
```

Notice that β_4 is significant indicating very strong evidence of faster improvement for new drug.

Remarks:

- 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). However, 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:
  - standard drug ($d = 0$): $\hat{\beta}_3 = 0.48$
  - new drug ($d = 1$): $\hat{\beta}_3 + \hat{\beta}_4 = 1.50$
- Examined $s \times d$ and $s \times t$ interactions, but they were not statistically significant.
- Started with exchangeable working correlation, but estimated $\rho \approx 0$.

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.
http://users.stat.ufl.edu/~athienit/STA4504/Examples/depression.R

Remark 9.4. *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.

10. Random Effects: GLMM

Unlike marginal modeling, this chapter presents an alternative model type that has a term in the model for each cluster.

10.1 Generalized Linear Mixed Models

A GLMM with a random effects is able to account for having multiple responses per subject (or "cluster") by putting a subject term in model.

Binary response $Y_t = 0$ or 1. Let Y_{it} = response by subject i at time t. Model:

$$\text{logit}\left[P(Y_{it} = 1)\right] = \alpha_i + \beta x_{it}, \quad t = 1, \ldots, T$$

The intercept α_i varies by subject so that a heterogeneous population implies a highly variable $\{\alpha_i\}$.

Treating α_i as fixed is not possible because this model would yield at least n parameters, so the solution is to treat it as random, i.e. $\alpha_i \overset{\text{ind.}}{\sim} N(\alpha, \sigma^2)$ or equivalently

$$\alpha_i = \alpha + u_i, \quad u_i \sim N(0, \sigma^2)$$

Model:

$$\text{logit}\left[P(Y_{it} = 1)\right] = \alpha + u_i + \beta x_{it}, \quad t = 1, \ldots, T$$

Parameters α and β are *fixed effects* and $\{u_i\}$ are *random effects*.

$Y_{i1}, Y_{i2}, \ldots, Y_{iT}$ are conditionally independent given u_i, but marginally dependent. That is, responses within subject more alike than between subjects.

Remark 10.1. *Note that random effects $\{u_i\}$ are unobserved (not data), so software must "integrate out" $\{u_i\}$ to get likelihood function.*

Example 10.1 (Depression continued) Using the same data from example 9.3

$$\log\left[\frac{P(Y_t = 1)}{P(Y_t = 0)}\right] = u_i + \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (dt)$$

```
> dep.lme=glmer((response == "normal") ~ severity+drug*time+(1|subject),
+  data=depression, family=binomial,
+  contrasts=list(drug=contr.treatment(2,base=2,contrasts=TRUE)))
> summary(dep.lme)

     AIC      BIC   logLik deviance df.resid
  1173.9   1203.5   -581.0   1161.9     1014

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.2849 -0.8268  0.2326  0.7964  2.0181

Random effects:
 Groups  Name        Variance Std.Dev.
 subject (Intercept) 0.003231 0.05684
Number of obs: 1020, groups:  subject, 340

Fixed effects:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)    -0.02797    0.16406  -0.170    0.865
severitysevere -1.31488    0.15261  -8.616  < 2e-16 ***
drug1          -0.05967    0.22239  -0.268    0.788
time            0.48274    0.11566   4.174 3.00e-05 ***
drug1:time      1.01817    0.19150   5.317 1.06e-07 ***
---

Correlation of Fixed Effects:
            (Intr) svrtys drug1  time
severitysvr -0.389
drug1       -0.614 -0.005
time        -0.673 -0.123  0.524
drug1:time   0.462 -0.121 -0.742 -0.562
```

http://users.stat.ufl.edu/~athienit/STA4504/Examples/depression2.R
In this example, GLMM and GEE estimates and standard errors 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

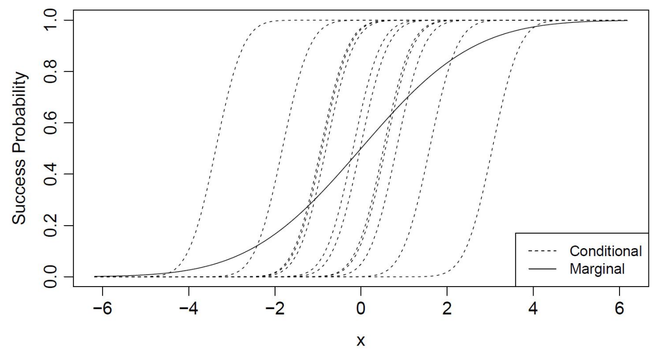
There appears to be little correlation between repeated measurements on subjects:

- $\hat{\rho} = -0.003 \approx 0$ 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 $\pm 1.96\sigma \approx \pm 0.11$. But $e^{\pm 0.11} \to (0.89, 1.12)$, so effect of u_i on odds is estimated to be small for most subjects.

10.2 Comparison with GEE

- When $\hat{\sigma} = 0$, estimates and standard errors same as treating repeated observations as independent.
- When $\hat{\sigma}$ is large, estimated β's from random effects logit model usually larger than from marginal model. They are estimating different things.



Example 10.2 (Teratology Overdispersions) Female rats on iron-deficient diets assigned to four groups:

1. placebo
2. iron injections on days 7 and 10
3. iron injections on days 0 and 7
4. iron injections weekly

Then they are made pregnant and sacrificed after 3 weeks. The response is whether fetus is dead or alive and the *cluster* is the litter.

Notation:

- GRP = group,
- LS = litter size,
- ND = number dead in litter

$$\text{logit}\left[P(\text{fetus } t \text{ in litter } i \text{ dead})\right] = \alpha + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4}$$

where

$$z_{ij} = \begin{cases} 1 & \text{if litter } i \text{ in group } j \\ 0 & \text{otherwise} \end{cases}$$

```
> terat$GRP=factor(terat$GRP)
> terat.binom=glm(cbind(ND,N-ND)~GRP, family=binomial, data=terat)
> summary(terat.binom)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.1440     0.1292   8.855  < 2e-16 ***
GRP2         -3.3225     0.3308 -10.043  < 2e-16 ***
GRP3         -4.4762     0.7311  -6.122 9.22e-10 ***
GRP4         -4.1297     0.4762  -8.672  < 2e-16 ***
---
    Null deviance: 509.43  on 57  degrees of freedom
Residual deviance: 173.45  on 54  degrees of freedom
AIC: 252.92

> 1-pchisq(173.45,df.residual(terat.binom)) # Gooodness of fit via LRT
[1] 1.876277e-14

> X2=sum(resid(terat.binom,type="pearson")^2);X2
[1] 154.707
> 1-pchisq(X2,df.residual(terat.binom)) # Gooodness of fit via Pearson
[1] 1.187217e-11

> X2/df.residual(terat.binom) # Evidence of overdispersion
[1] 2.864945
```

Results:

- Binomial model fits poorly ($X^2 = 154.7, G^2 = 173.5, df = 54$, p-value ≈ 0).
- 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) due to overdispersion.
- Possible solutions:
  - GEE: models marginal (population averaged) effect of treatment.
  - GLMM: models litter-specific effect.
  - At least two other approaches not discussed (thoroughly) in this class:
    * Quasi-binomial: simplified version of GEE.
    * Beta-binomial: parametric mixture model, analogous to negative-binomial for count data. Motivation similar to GLMM

```
> terat.gee <- gee((Resp == "Dead") ~ GRP, id = Litter,
+       data = teratbnry, family = binomial, corstr = "exchangeable")
> summary(terat.gee)
Summary of Residuals:
       Min         1Q     Median         3Q        Max
```

```
-0.7881637 -0.3568763  0.2118363  0.3420839  0.6431237

Coefficients:
              Estimate Naive S.E.   Naive z Robust S.E.  Robust z
(Intercept) -0.5889477  0.2317694 -2.541093   0.2966943 -1.985032
GRP2         1.2429690  0.4469084  2.781261   0.5612748  2.214546
GRP3         1.6997950  0.7248173  2.345136   0.8877114  1.914806
GRP4         1.9028396  0.5776533  3.294086   0.7226377  2.633186

Estimated Scale Parameter:  0.709622

> # Big working correlation matrix (17 x 17), but
> # all correlations equal with exchangeable struc:
> terat.gee$working.correlation[1,2]
[1] 0.8051211

> library(lme4)
> # Using grouped data
> terat.glmm <- glmer(cbind(ND, N-ND) ~ GRP + (1|Litter),
+         data = terat, family = binomial)
> # Using ungrouped binary data
> terat.glmm <-  glmer((Resp == "Dead") ~ GRP + (1|Litter),
+         data = teratbnry, family = binomial)
> summary(terat.glmm)

     AIC      BIC   logLik deviance df.resid
   445.9    468.0   -218.0    435.9      602

Scaled residuals:
    Min      1Q  Median      3Q     Max
-4.7821 -0.2431  0.1158  0.2673  2.8214

Random effects:
 Groups Name        Variance Std.Dev.
 Litter (Intercept) 2.284    1.511
Number of obs: 607, groups:  Litter, 58

Fixed effects:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.8094     0.3616  -5.004 5.62e-07 ***
GRP2          4.5396     0.7345   6.181 6.39e-10 ***
GRP3          5.8833     1.1754   5.005 5.58e-07 ***
GRP4          5.6062     0.9076   6.177 6.54e-10 ***
---

Correlation of Fixed Effects:
     (Intr) GRP2   GRP3
GRP2 -0.562
```

```
GRP3 -0.373  0.235
GRP4 -0.496  0.316  0.221
```

http://users.stat.ufl.edu/~athienit/STA4504/Examples/teratology.R

	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.

As a final note it seems that there are differences between groups 2,3,4 with the base group 1. As an exercise compare groups 2 and 3 for the GLMM model.

7. Loglinear Models

Loglinear models for contingency tables treat all variables as response variables, like multivariate analysis.

7.1 Loglinear for 2-way

7.1.1 $I \times J$

All variables are treated as responses, in that a set of variables is not used to model another variable but are interested in patterns of dependence and independence among the variables:

- Are the variables independent?
- The strength of associations
- Are there any interactions?

		Y			
		1	2	$\cdots$	J
X	1	n_{11}	n_{12}	$\cdots$	n_{1J}
	2	n_{21}	n_{22}	$\cdots$	n_{2J}
	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
	I	n_{I1}	n_{I2}	$\cdots$	n_{IJ}

Loglinear models treat cell counts as Poisson and use log link function. From Lemma 2.1 we have that

$$\mu_{ij} = n\pi_{ij} \overset{\text{ind.}}{=} n\pi_{i+}\pi_{+j} \Rightarrow \log\left(\mu_{ij}\right) = \underbrace{\log(n)}_{\lambda} + \underbrace{\log \pi_{i+}}_{\lambda_i^X} + \underbrace{\log \pi_{+j}}_{\lambda_j^Y}$$

- λ_i^X: effect of classification in row i ($I-1$ non-redundant parameters with the restriction of $\lambda_1^X = 0$ for base group)

- λ_j^Y: effect of classification in column j ($J-1$ non-redundant parameters with the restriction of $\lambda_1^Y = 0$ for base group))

The degrees of freedom are

$$df = \underbrace{\text{number of Poisson counts}}_{\text{number of cells in table}} - \text{number of parameters}$$

- For the independence model

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

  and hence

  $$df = \underbrace{IJ}_{\text{no. of cells}} - \underbrace{\left[\overbrace{1}^{\lambda} + \overbrace{(I-1)}^{\lambda_i^X} + \overbrace{(J-1)}^{\lambda_j^Y}\right]}_{\text{no. of parameters}} = (I-1)(J-1)$$

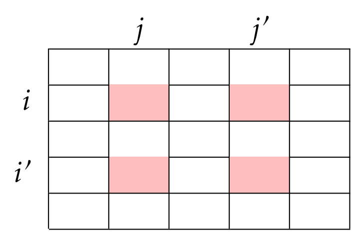
- For the saturated model

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

  and hence

  $$df = \underbrace{IJ}_{\text{no. of cells}} - \underbrace{\left[\overbrace{1}^{\lambda} + \overbrace{(I-1)}^{\lambda_i^X} + \overbrace{(J-1)}^{\lambda_j^Y} + \overbrace{(I-1)(J-1)}^{\lambda_{ij}^{XY}}\right]}_{\text{no. of parameters}} = 0$$

Log-odds-ratio comparing levels i and i' of X and j and j' of Y is



$$\begin{aligned}
\log\left(\frac{\mu_{ij}\mu_{i'j'}}{\mu_{ij'}\mu_{i'j}}\right) =& \log(\mu_{ij}) + \log(\mu_{i'j'}) - \log(\mu_{ij'}) - \log(\mu_{i'j}) \\
=& (\lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY}) + (\lambda + \lambda_{i'}^X + \lambda_{j'}^Y + \lambda_{i'j'}^{XY}) \\
& - (\lambda + \lambda_i^X + \lambda_{j'}^Y + \lambda_{ij'}^{XY}) - (\lambda + \lambda_{i'}^X + \lambda_j^Y + \lambda_{i'j}^{XY}) \\
=& \lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{ij'}^{XY} - \lambda_{i'j}^{XY}
\end{aligned}$$

- For the independence model, since all $\lambda_{ij}^{XY} = 0$ (they do not even exist), 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

$$\exp\left(\lambda_{ij}^{XY} + \lambda_{i'j'}^{XY} - \lambda_{ij'}^{XY} - \lambda_{i'j}^{XY}\right)$$

Substituting the MLEs of the saturated model (perfect fit) just reproduces the empirical odds ratio

$$\frac{n_{ij}n_{i'j'}}{n_{ij'}n_{i'j}}$$

Example 7.1 (Job Satisfaction) We are revisiting

- Example 2.12 where we tested independence via Pearson's X^2
- Example 6.1 where we fitted a baseline logit model
- Example 6.2 where we fitted a cumulative logit model

to fit

$$\log(\mu_{ij}) = \lambda + \lambda_i^I + \lambda_j^S \quad i = 1,2,3,\cancel{4}\ j = 1,2,3,\cancel{4}$$

which can be expressed as

$$\log(\mu_{ij}) = \lambda + \lambda_1^I z_{(10)} + \lambda_2^I z_{(20)} + \lambda_3^I z_{(30)} + \lambda_1^S w_{(LD)} + \lambda_2^S w_{(MS)} + \lambda_3^S w_{(VS)}$$

where

$$z_{(10)} = \begin{cases} 1 & \text{income score } = 10 \\ 0 & \text{otherwise} \end{cases}$$

and

$$w_{(LD)} = \begin{cases} 1 & \text{little dissatisfaction} \\ 0 & \text{otherwise} \end{cases}$$

and similarly for the rest. The independence model is

```
> jobsat.ind=glm(count~factor(income)+jobsat,
+  family=poisson(link=log),data=table.sat)
> summary(jobsat.ind)
Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -0.16705    0.53464  -0.312  0.75469
factor(income)10   0.43532    0.27362   1.591  0.11162
factor(income)20   0.08701    0.29516   0.295  0.76815
factor(income)30   0.08701    0.29516   0.295  0.76815
jobsatLD           1.25276    0.56694   2.210  0.02713 *
jobsatMS           2.75684    0.51563   5.347 8.96e-08 ***
jobsatVS           1.74920    0.54173   3.229  0.00124 **
---
    Null deviance: 90.242  on 15  degrees of freedom
Residual deviance: 13.467  on  9  degrees of freedom
AIC: 77.068
```

and performing a goodness of fit test is comparing this (independence) model to the saturated one, so hence the goodness of fit is the test of independence. That is, the goodness of fit tests

$$H_0 : \lambda_{ij}^{IS} = 0 \quad \forall i, j$$

```
> jobsat.sat=update(jobsat.ind,.~.+factor(income)*jobsat)
> anova(jobsat.ind,jobsat.sat,test="Chisq")
Analysis of Deviance Table

Model 1: count ~ factor(income) + jobsat
Model 2: count ~ factor(income) + jobsat + factor(income):jobsat
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         9     13.467
2         0      0.000  9   13.467   0.1426
```

and hence we conclude independence. Using the Independence model we can also obtain expected values under independence.

- Under example 2.12 with
  - $\hat{\mu}_{(3,D)} = \frac{22\times4}{104} = 0.846$
  - $\hat{\mu}_{(10,LD)} = \frac{34\times14}{104} = 4.5769$
- Under the independence model with
  - $\hat{\mu}_{(3,D)} = e^{-0.16705} = 0.846$
  - $\hat{\mu}_{(10,LD)} = e^{-0.16705+0.43532+1.25276} = 4.5769$

http://users.stat.ufl.edu/~athienit/STA4504/Examples/jobsatis_loglinear.R

7.1.2 $I \times 2$

Let $J = 2$, that is, $Y = 1, 2$ to only have two levels. Then, with $\pi_i := P(Y = i)$

$$\begin{aligned}
\log\left(\frac{\pi_1}{1-\pi_1}\right) = \log\left(\frac{n\pi_1}{n\pi_2}\right) = \log\left(\frac{\mu_{i1}}{\mu_{i2}}\right) &= \log(\mu_{i1}) - \log(\mu_{i2}) \\
&= (\lambda + \lambda_i^X + \lambda_1^Y + \lambda_{i1}^{XY}) - (\lambda + \lambda_i^X + \lambda_2^Y + \lambda_{i2}^{XY}) \\
&= (\lambda_1^Y - \cancelto{0}{\lambda_2^Y}) + (\lambda_{i1}^{XY} - \cancelto{0}{\lambda_{i2}^{XY}})
\end{aligned} \tag{7.1}$$

if we chose group 2 to be the base group then $\lambda_2^Y = \lambda_{i2}^{XY} = 0$.

Remark 7.1.

- *If group 1 was chosen as the base group then its corresponding parameters would be 0.*
- *If the independence model is used then all $\lambda^{XY} = 0$ and the formula simplifies.*

Example 7.2 (Belief in afterlife) Reconsider

	Belief	
Race	Yes	No
White	1339	300
Black	260	55
Other	88	22

Independence model

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y \quad i = 1,2,3 \;\; j = 1,2$$

```
> Race=rep(c("White","Black","Other"),each=2)
> Belief=rep(c("Yes","No"),3)
> count=c(1339,300,260,55,88,22)
> after=data.frame(Race,Belief,count)
> after=transform(after,Race=relevel(Race,"Other"))
> B_R=glm(count~Belief+Race,family=poisson(link=log),data=after)
> summary(B_R)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  3.00032    0.10611   28.28   <2e-16 ***
BeliefYes    1.49846    0.05697   26.30   <2e-16 ***
RaceBlack    1.05209    0.11075    9.50   <2e-16 ***
RaceWhite    2.70136    0.09849   27.43   <2e-16 ***
---

    Null deviance: 2849.21758  on 5  degrees of freedom
Residual deviance:    0.35649  on 2  degrees of freedom
AIC: 49.437
```

Note that the estimated odds (not odds ratio) of belief in the afterlife was $\exp(\hat{\lambda}_1^Y - 0) = \exp(1.49846) = 4.474793$ for each race.

Saturated model/Dependence model

$$\log(\mu_{ij}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \quad i = 1,2,3 \;\; j = 1,2$$

with

```
> BR=glm(count~Belief*Race,family=poisson(link=log),data=after)
> summary(BR)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)           3.0910     0.2132  14.498  < 2e-16 ***
BeliefYes             1.3863     0.2384   5.816 6.03e-09 ***
RaceBlack             0.9163     0.2523   3.632 0.000281 ***
RaceWhite             2.6127     0.2209  11.829  < 2e-16 ***
```

```
BeliefYes:RaceBlack   0.1671     0.2808   0.595 0.551889
BeliefYes:RaceWhite   0.1096     0.2468   0.444 0.656946
---
    Null deviance:  2.8492e+03  on 5  degrees of freedom
Residual deviance: -8.7930e-14  on 0  degrees of freedom
AIC: 53.081
```

We can test for independence by $H_0 : \lambda_{ij}^{XY} = 0\ \forall i,j$ by a likelihood ratio test using the difference of deviances. Notice that the model with the interaction is a saturated model, so the LR test is in fact a goodness of fit test for the independence model with

$$D_0 - D_1 = 0.35649 - 0$$

on df=2 and p-value= 0.8367, so we fail to reject H_0 and conclude independence between belief and race.

```
> anova(B_R,BR,test="Chisq")
Analysis of Deviance Table

Model 1: count ~ Belief + Race
Model 2: count ~ Belief * Race
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         2    0.35649
2         0    0.00000  2  0.35649   0.8367
```

http://users.stat.ufl.edu/~athienit/STA4504/Examples/afterlife.R

7.2 Loglinear for 3-way

Definition 7.1 **(Associations)** We review 5 types of associations

- X, Y, Z are *mutual independent*, (X, Y, Z) if $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$

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

- Y is *jointly independent* of X and Z, (XZ, Y) if $\pi_{ijk} = \pi_{+j+}\pi_{i+k}$

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

- X and Y are *conditionally independent* given Z, (XZ, YZ) if $\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$

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

- *Homogeneous association*, (XZ, XY, YZ) if two variables have the same association for all levels of the third, e.g. $\pi_{ij|k} = \pi_{ij|k'}$ same $\forall k, k'$

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

- *Non restricted association*, (saturated model) (XYZ)

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

Example 7.3 Consider a $2 \times 2 \times 2$ with X, Y conditional independence (XZ, YZ)

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

Hence,

- X and Y are conditionally independent given Z:

$$\log\left(\theta_{XY(k)}\right) = \log\left(\frac{\mu_{ijk}\mu_{i'j'k}}{\mu_{i'jk}\mu_{ij'k}}\right) = \cdots = 0 \Longrightarrow \theta_{XY(k)} = 1$$

- The $X - Z$ odds ratio is the same at all levels of Y:

$$\log\left(\theta_{X(j)Z}\right) = \log\left(\frac{\mu_{ijk}\mu_{i'jk'}}{\mu_{i'jk}\mu_{ijk'}}\right) = \cdots = \lambda_{11}^{XZ} + \lambda_{22}^{XZ} - \lambda_{12}^{XZ} - \lambda_{21}^{XZ}$$

  which does not depend on j.

- Similarly, $Y - Z$ odds ratio same at all levels of X. Model has no three-factor interaction.

Example 7.4 Consider the loglinear homogeneous association model denoted (XY,XZ,YZ).

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

Each pair of variables is conditionally dependent, but association (as measured by odds ratios) is the same at all levels of third variable.

Example 7.5 (Teen substance usage) A survey of 2276 high school seniors

```
> 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=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)

> summary(teens.AC.AM.CM)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)     5.63342    0.05970  94.361  < 2e-16 ***
alcyes          0.48772    0.07577   6.437 1.22e-10 ***
cigsyes        -1.88667    0.16270 -11.596  < 2e-16 ***
mjyes          -5.30904    0.47520 -11.172  < 2e-16 ***
alcyes:cigsyes  2.05453    0.17406  11.803  < 2e-16 ***
alcyes:mjyes    2.98601    0.46468   6.426 1.31e-10 ***
cigsyes:mjyes   2.84789    0.16384  17.382  < 2e-16 ***
---
    Null deviance: 2851.46098  on 7  degrees of freedom
Residual deviance:    0.37399  on 1  degrees of freedom
AIC: 63.417

> deviance(teens.AC.AM.CM)
[1] 0.3739859
> X2=sum(residuals(teens.AC.AM.CM,type="pearson")^2);X2
[1] 0.4011005
> 1-pchisq(X2,1)
[1] 0.5265215
```

The (AC,AM,CM) model fits well with $G^2 = 0.37$ (and $X^2 = 0.4$) on 1 df. Equivalently done via,

```
> 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 + alc:cigs:mj
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         1    0.37399
2         0    0.00000  1  0.37399   0.5408
```

Next we check if any 2-way interactions can be removed

```
> 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>          0.37  63.42
alc:cigs  1   187.75 248.80 187.38 < 2.2e-16 ***
alc:mj    1    92.02 153.06  91.64 < 2.2e-16 ***
cigs:mj   1   497.37 558.41 497.00 < 2.2e-16 ***
```

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    187.754
2         1      0.374  1   187.38 < 2.2e-16 ***
```

We can also get predicted counts under a variety of models and compare them to the actual data/saturated model

```
> 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

$$\frac{\hat{\mu}_{11k}\hat{\mu}_{22k}}{\hat{\mu}_{12k}\hat{\mu}_{21k}} = \exp(\hat{\lambda}_{11}^{AC} + \cancelto{0}{\hat{\lambda}_{22}^{AC}} - \cancelto{0}{\hat{\lambda}_{12}^{AC}} - \cancelto{0}{\hat{\lambda}_{21}^{AC}}) = e^{2.0545} = 7.8$$

A 95% CI is

$$e^{2.05 \mp (1.96)(0.174)} \longrightarrow (5.5, 11.0)$$

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 independent 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:

$$\frac{(909)(555)}{(439)(45.8)} = 25.14 \qquad \frac{(4.76)(180)}{(142)(0.24)} = 25.14$$

http://users.stat.ufl.edu/~athienit/STA4504/Examples/teens.R

Remark 7.2.

- *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.*
- *For modeling ordinal associations consider a 2-way table with assigned*
  - *row scores $u_1 \leq u_2 \leq \cdots \leq u_I$*
  - *column scores $v_1 \leq v_2 \leq \cdots \leq v_J$*

  and model

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

  where $\beta u_i v_j$ takes the role of λ_{ij}^{XY} but only 1 parameter is used, i.e. only 1 degree of freedom taken up, instead of $(I-1)(J-1)$

Checking residuals is always important and done in the usual way as with any GLM, however a new graphical visualization may also be useful

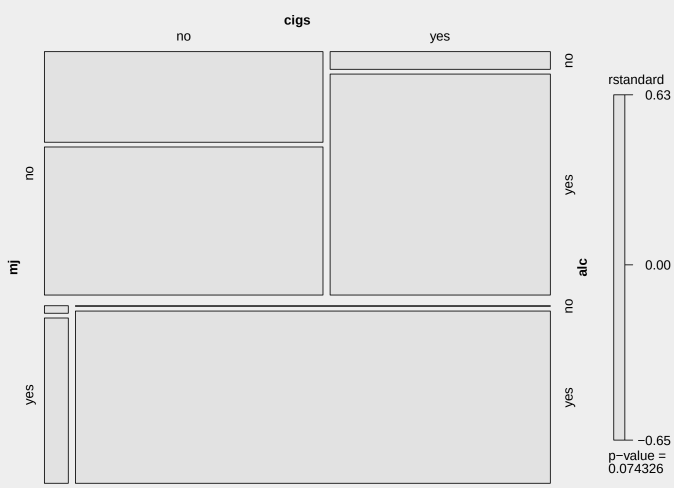
R code 7.1 In the `vcdExtra` package the function

```
mosaic(glm object,...)
```

is capable of a mosaic plot of the residuals, where the area of each tile is proportional to the corresponding cell entry, given the dimensions of previous splits.

Example 7.6 (Teen substance usage continued) Getting and visualizing the standardized deviance residuals

```
rstandard(teens.AC.AM.CM)
      1       2       3       4       5       6       7       8
 0.6332 -0.6334 -0.6347  0.6331 -0.6527  0.6317  0.5933 -0.6335
> mosaic(teens.AC.AM.CM,~mj+cigs+alc,residuals_type = "rstandard")
```



7.3 Loglinear-Logit Connection

We have already seen the connection in equation (7.1) which can be written as a logit model

$$\log\left(\frac{P(Y=1)}{1-P(Y=1)}\right) = \underbrace{(\lambda_1^Y - \cancelto{0}{\lambda_2^Y}\ )}_{\alpha} + \underbrace{(\lambda_{i1}^{XY} - \cancelto{0}{\lambda_{i2}^{XY}}\ )}_{\beta_i^X}$$

$$= \alpha + \beta_i^X$$

Consider the loglinear homogeneous association model denoted (XY, XZ, YZ).

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

Suppose Y is binary, treated as the response, and let

$$\pi_{ik} = P(Y = 1 | X = i, Z = k)$$

then

$$\begin{aligned}
\operatorname{logit}(\pi_{ik}) &= \log(\mu_{i1k}) - \log(\mu_{i2k}) \\
&= \cdots \\
&= \underbrace{(\lambda_1^Y - \cancelto{0}{\lambda_2^Y})}_{\alpha} + \underbrace{(\lambda_{i1}^{XY} - \cancelto{0}{\lambda_{i2}^{XY}})}_{\beta_i^X} + \underbrace{(\lambda_{1k}^{YZ} - \cancelto{0}{\lambda_{2k}^{YZ}})}_{\beta_k^Z} \\
&= \alpha + \beta_i^X + \beta_k^Z
\end{aligned}$$

an additive model with no XZ interaction.

When a "response" (say Y) exists and it has two levels then it is possible to fit a loglinear model and an *equivalent* logit model. We are not required to fit the equivalent model but we are exploring the special case.

Remark 7.3. The (XY, YZ) model also yields an additive logit model but for ML estimates, Deviances and degrees of freedom to match, the loglinear model must contain the most general interaction among variables that are explanatory in the logit model, those are X and Z. Therefore, the equivalent loglinear model must include XY (X linked to Y), the YZ (Z linked to Y), and the XZ (XZ linked to Y).

Some example with $\pi = P(Y = 1)$, predictors A, B, C (4-way table).:

- $\operatorname{logit}(\pi) = \alpha + \beta_i^A + \beta_j^B + \beta_k^C \longleftrightarrow (AY, BY, CY)$
- $\operatorname{logit}(\pi) = \alpha + \beta_i^A + \beta_j^B + \beta_k^C + \beta_{jk}^{BC} \longleftrightarrow (AY, BY, CY, BCY)$ *a.k.a.* (AY, BCY)

Remark 7.4.

- *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.*

Example 7.7 (Berkeley Graduate Admissions) Earlier we had fit a logit model for the probability of admission

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

with 12 binomial variates and 7 parameters, hence df = 5. Now we will take a look at the equivalent loglinear model (AG, AD, DG)

$$\log(\mu_{ijk}) = \lambda + \lambda_i^A + \lambda_j^G + \lambda_k^D + \lambda_{ij}^{AG} + \lambda_{ik}^{AD} + \lambda_{jk}^{DG}$$

with 24 independent Poisson variates and 19 parameters, hence df = 5.
Once we create the appropriate data frame

```
> head(berk2)
  Dept Gender Admit Freq
1    A   Male   Yes  512
2    A Female   Yes   89
3    B   Male   Yes  353
4    B Female   Yes   17
5    C   Male   Yes  120
6    C Female   Yes  202

> UCB.loglin=glm(Freq~Admit*Gender+Admit*Dept+Gender*Dept,family=poisson,
+  data=berk2)
> summary(UCB.loglin)
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)          3.59099    0.11659  30.801  < 2e-16 ***
AdmitYes             0.68192    0.09911   6.880 5.97e-12 ***
GenderMale           2.09846    0.11548  18.172  < 2e-16 ***
DeptB               -1.43464    0.23341  -6.146 7.93e-10 ***
DeptC                2.34983    0.12262  19.163  < 2e-16 ***
DeptD                1.90293    0.12557  15.154  < 2e-16 ***
DeptE                2.08467    0.12711  16.400  < 2e-16 ***
DeptF                2.17093    0.12798  16.963  < 2e-16 ***
AdmitYes:GenderMale -0.09987    0.08085  -1.235    0.217
AdmitYes:DeptB      -0.04340    0.10984  -0.395    0.693
AdmitYes:DeptC      -1.26260    0.10663 -11.841  < 2e-16 ***
AdmitYes:DeptD      -1.29461    0.10582 -12.234  < 2e-16 ***
AdmitYes:DeptE      -1.73931    0.12611 -13.792  < 2e-16 ***
AdmitYes:DeptF      -3.30648    0.16998 -19.452  < 2e-16 ***
GenderMale:DeptB     1.07482    0.22861   4.701 2.58e-06 ***
GenderMale:DeptC    -2.66513    0.12609 -21.137  < 2e-16 ***
GenderMale:DeptD    -1.95832    0.12734 -15.379  < 2e-16 ***
GenderMale:DeptE    -2.79519    0.13925 -20.073  < 2e-16 ***
GenderMale:DeptF    -2.00232    0.13571 -14.754  < 2e-16 ***
---
    Null deviance: 2650.095  on 23  degrees of freedom
Residual deviance:   20.204  on  5  degrees of freedom
AIC: 217.26
```

We note that $G^2 = 20.204$ is the same for both models and that the estimated odds (controlling for department) of admission for males compared to that of females is

- Logit model: $\exp(\hat{\beta}_1 - \hat{\beta}_2) = \exp(-0.09987) = 0.905$

- Loglinear model: $\exp(\hat{\lambda}_{11}^{AG} + \hat{\lambda}_{22}^{AG} - \hat{\lambda}_{12}^{AG} - \hat{\lambda}_{21}^{AG}) = \exp(-0.09987) = 0.905$

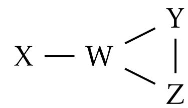
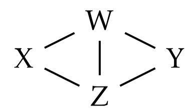
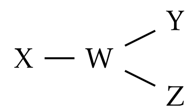
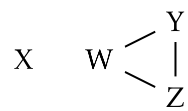
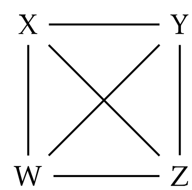
http://users.stat.ufl.edu/~athienit/STA4504/Examples/admissions_loglinear.R

7.4 Independence Graphs and Collapsibility

Independence graph is 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.

7.4.1 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)*	W — X — Y — Z
(X,WY,WZ,YZ) (X,WYZ)*	
(WX,YZ)*	W — X Y — Z
(WX,WY,WZ,XY,XZ,YZ) (WX,WY,WZ,XYZ) (WX,WYZ,XYZ) ...many others... (WXYZ)*	

* Graphical model

7.4.2 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 conditionally 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.

Example 7.8 (Teen substance usage) See example 7.5 where

- A = alcohol use
- C = cigarette use
- M = marijuana use

The model of AC conditional independence, (AM, CM), has independence graph

$$\text{A} — \text{M} — \text{C}$$

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.

$$\frac{(909.24)(142.16)}{(438.84)(4.76)} = \frac{(45.76)(179.84)}{(555.16)(0.24)} = \frac{(955)(322)}{(994)(5)} = 61.9$$

```
> exp(coef(teens.AM.CM)[5])
alcyes:mjyes
    61.87324

> AM.CM.fitted <- teens
> AM.CM.fitted[,,] <- predict(teens.AM.CM, type="response")
> AM.CM.fitted[,"yes",]
     alc
mj           yes         no
  yes 909.239583   4.760417
  no  438.840426 142.159574
> AM.CM.fitted[,"no",]
     alc
mj            yes          no
  yes  45.7604167   0.2395833
  no  555.1595745 179.8404255
> AM.CM.fitted[,"yes",] + AM.CM.fitted[,"no",]
     alc
mj    yes  no
  yes 955   5
  no  994 322
```

- 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.

$$\frac{(909.24)(0.24)}{(45.76)(4.76)} = \frac{(438.84)(179.84)}{(555.16)(142.16)} = 1$$

$$\frac{(1348.08)(180.08)}{(600.92)(146.92)} = 2.75 \neq 1$$

```
> AM.CM.fitted["yes",,]
     alc
cigs          yes          no
  yes 909.2395833   4.7604167
  no   45.7604167   0.2395833
> AM.CM.fitted["no",,]
     alc
cigs       yes       no
  yes 438.8404 142.1596
  no  555.1596 179.8404
> AM.CM.fitted["yes",,] + AM.CM.fitted["no",,]
     alc
cigs      yes      no
  yes 1348.08  146.92
  no   600.92  180.08
```

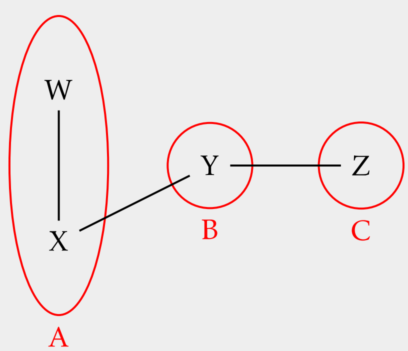
See Part II of http://users.stat.ufl.edu/~athienit/STA4504/Examples/teens.R

7.4.3 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.

A — B — C

Example 7.9 Consider the (WX, XY, YZ) model (drawn slightly differently)



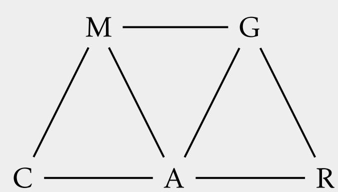
Then collapsing over Z:

- WX and XY associations are unchanged
- W and Y are still conditionally independent given X

Example 7.10 (Teen substance usage continued) In addition to the variables seen so far data exists on the race and gender of each teen.

```
> 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
```

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, 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        GMale
   5.9784142   -5.7507310   -3.0157544   -0.3895472    0.1358363
      ROther      ANo:CNo      ANo:MNo      CNo:MNo    ANo:GMale
  -2.6630477    2.0545341    3.0059195    2.8478892    0.2922863
  ANo:ROther    MNo:GMale GMale:ROther
   0.5934604   -0.2692945    0.1261850
> coef(AC.AM.CM)
(Intercept)         ANo         CNo         MNo     ANo:CNo     ANo:MNo
  6.8138656  -5.5282675  -3.0157544  -0.5248611   2.0545341   2.9860144
    CNo:MNo
  2.8478892
```

http://users.stat.ufl.edu/~athienit/STA4504/Examples/teens2.R

Bibliography

[1] A. Agresti. *An Introduction to Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley, 2018. ISBN: 9781119405269. URL: https://books.google.com/books?id=ukNxDwAAQBAJ.

[2] Brett Presnell. *Lecture notes for Introduction to Categorical Data Analysis*. Jan. 2012.

[3] Jin-Ting Zhang. *Lecture notes for to Categorical Data Analysis*. Jan. 2012.