Generalized Linear Models Part 1

Prof. dr. Helena Geys

Hasselt University

2020-2021

Chapter 1: A Gentle Introduction

This course will focus on categorical "outcome" data.
Example: Clinical trials for patients with cancer

	RESPONSE		
	Not Cured	Cured	TOTAL
Treatment A	49	26	75
Treatment B	64	11	75
TOTAL	113	37	150

- Randomly assign a patient to one of two treatments (A or B) and see if patient is cured from disease
- categorical outcome=response (cured/not cured)
- **Question:**
  Is treatment associated with response?

- We will spend a lot of time talking about (2×2) tables
- Much of the statistical theory is more easily seen in (2×2) tables, and then generalizes to more complicated problems:
  - More than 2 levels of response (not cured, partially cured, cured) which may be ordered or not ordered.
  - More than 2 levels of treatment (A,B,C,D,...)
  - Other factors or *covariates* that influence response (age, sex, etc.)
- (2×2) tables may arise from different sampling plans (see below)

Sampling Plans

Prospective Study

Some more Examples of (2×2) Tables

- Cold incidence among French Skiers (Pauling, Proceedings of the national Academy of Sciences, 1971)

	OUTCOME		
	COLD	NO COLD	TOTAL
VITAMIN C	17	122	**139**
NO VITAMIN C	31	109	**140**
TOTAL	48	231	279

- Number on each *treatment* fixed by design.
- Individuals are randomized to treatment 1 or treatment 2
- **Question of interest:**
  Does treatment affect outcome?

Sampling Plans

Retrospective Study (Case-Control Study)

- Alcohol Consumption and occurence of esophageal cancer (Tuyms et al., Bulletin of Cancer, 1974)
- Not ethical to randomize patients in a prospective study

	STATUS		
	CASE	CONTROL	TOTAL
80+ (g/day)	96	109	205
0-79 (g/day)	104	666	770
TOTAL	**200**	**775**	975

- Number of *cases and controls* (outcomes) are fixed by design and exposures are random.
- **Question of interest:**
  Does alcohol exposure vary among cases and controls?

Sampling Plans

Cross-sectional or Prevalence Study

- General Social Survey (1984 SPSS Manual)

	JOB SATISFACTION		
	DISSAT	SATIS	TOTAL
$<$ \$15.000	104	391	495
$\geq$ \$15.000	66	340	406
TOTAL	170	731	**901**

- Sample subjects (*total number fixed*) and then cross-classify them on the basis of 2 variables.
- **Question of interest:**
  Is there an association between job satisfaction and income?

Categorical Response Data

Categorical variable - variable for which the measurement scale consists of a set of categories.

Example:

- in social sciences: political philosophy - "liberal", "moderate" or conservative.
- in health sciences: diagnostic test for Aldzheimer disease - "symptoms absent", "symptoms present".
- in behavioural sciences: "schizophrenia", "depression", ...
- in zoology: "fish", "invertebrate", "reptile", ...
- ...

Note: One and only one category should apply to each subject.

Categorical Response Data

Nominal/Ordinal Scale Distinction

Ordinal Variable - categorical variable having ordered scales
Examples:

- attitude towards legalization of abortion (disapprove in all cases, approve only in certain cases, approve in all cases)
- response to a medical treatment (excellent, good, fair, poor)

Nominal Variable - categorical variable having unordered scales
Examples:

- Religious affiliation: Catholic, Jewish, Protestant, other
- Type of Music: classical, country, folk, jazz, rock, other

Problem: Which scale of measurement is most appropriate for the following variables?

- Pain (none, moderate, severe)
- Race (Black, Caucasian, Other)

Categorical Response Data

Remark

Variables measured on the nominal or ordinal scale are often referred to as **qualitative variables**, because the measurement consists only of ordered or unordered discrete categories. In contrast, **quantitative variables** have underlying continuity; that is they can take on any value on the measurement scale (e.g. IQ, temperature, . . .).

Exact Test Statistics and Confidence Intervals

Unfortunately, many (phase II) trials have small samples, and the above asymptotic test statistics and confidence intervals have very poor properties in small samples (A 95% confidence interval may only have 80% coverage). In this situation, "Exact test statistics and Confidence Intervals" can be obtained.

Exact Test Statistics and Confidence Intervals

One-sided Exact Test Statistic

- Suppose you want to test

$$H_0 : p = p_0$$

versus

$$H_1 : p > p_0$$

- Test Statistic

$$Y = \text{ the number of successes out of } n \text{ trials}$$

Under H_0:

$$Y \sim Bin(n, p_0)$$

- When would you tend to reject $H_0 : p = p_0$ in favor of $H_1 : p > p_0$?

Exact Test Statistics and Confidence Intervals

Answer

Under $H_0 : p = p_0$ you would expect $Y \sim np_0$
Under $H_1 : p > p_0$ you would expect $Y > np_0$,
i.e. you would expect Y to be "large" under the alternative.

Exact Test Statistics and Confidence Intervals

Exact one-sided p-value

If you observe y successes, the exact p-value is

$$\begin{aligned} p\text{-value} &= P(Y \geq y | H_0 : p = p_0) \\ &= \sum_{j=y}^{n} \binom{n}{j} p_0^j (1-p_0)^{n-j} \end{aligned}$$

Exact Test Statistics and Confidence Intervals

Exact one-sided p-value

Other one-sided p-value

Suppose you want to test

$$H_0 : p = p_0$$

versus

$$H_1 : p < p_0$$

The exact p-value is:

$$\begin{aligned} p\text{-value} &= P(Y \leq y | H_0 : p = p_0) \\ &= \sum_{j=0}^{y} \binom{n}{j} p_0^j (1 - p_0)^{n-j} \end{aligned}$$

Exact Test Statistics and Confidence Intervals

Two-sided exact p-value

The general definition of a 2-sided exact p-value is

$$P\left[\begin{array}{l}\text{seeing a result that is more extreme} \\ \text{than the observed result in either direction}\end{array} \Big| H_0\right]$$

It is easy to calculate a 2-sided p-value for a symmetric distribution. However, the 2-sided exact p-value is trickier when the binomial distribution is not symmetric:

Exact Test Statistics and Confidence Intervals

Two-sided exact p-value

1. Calculate the probability of the observed result under the null:

$$\pi = \binom{n}{y} p_0^y (1-p_0)^{n-y}$$

2. Calculate the probabilities of all $n+1$ values that Y can take on:

$$\pi_j = \binom{n}{j} p_0^j (1-p_0)^{n-j},$$

$j = 0, \ldots, n$

3. Sum the probabilities π_j in (2) that are less than or equal to the observed probability π in (1):

$$p\text{-value} = \sum_{j=0}^{n} \pi_j I(\pi_j \leq \pi)$$

where

$$I(\pi_j \leq \pi) = \begin{cases} 1 & \text{if as likely or less likely} \\ 0 & \text{if more likely} \end{cases}$$

Exact Test Statistics and Confidence Intervals

Problem

Suppose $n = 5$, hypothesize $p = 0.4$ and we observe $y = 3$ successes. Calculate the exact p-values for

1. $H_1 : p > 0.4$
2. $H_1 : p \neq 0.4$

Exact Test Statistics and Confidence Intervals

Solution

$$
\begin{aligned}
P(Y=0) &= 0.0777 \\
P(Y=1) &= 0.2592 \\
P(Y=2) &= 0.3456 \\
P(Y=3) &= 0.2304 \\
P(Y=4) &= 0.0768 \\
P(Y=5) &= 0.0102
\end{aligned}
$$

1. $p = 0.2304 + 0.0768 + 0.0102 = 0.3174$
2. $p = 0.0777 + 0.2304 + 0.0768 + 0.0102 = 0.3951$

Remark that it is no longer true that the p-value for a two-sided test is twice the p-value for the one-sided test.

Chapter 2:
Two-Contingency Tables

Probability Structure

Table Representation

Contingency Table:
A tabel in which the cells contain frequency counts of outcomes
Two-way Table:
A contingency table that cross classifies two variables
$I \times J$ **Table:**
A two-way table having I rows and J columns

Example of a 2×2 Table

- Cross classification of sample of Americans according to their gender and their opinion about afterlife.

	Belief in Afterlife		
Gender	Yes	No or Undecided	Total
Females	435	147	582
Males	375	134	509
Total	810	281	1091

- Is one sex more likely than the other to believe in an afterlife, or is belief in an afterlife independent of gender?

Joint, Marginal and Conditional Probabilities

- joint distribution

$$\pi_{ij} = P(X = i, Y = j)$$

- marginal distribution

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

- conditional distribution of $Y|X$

$$\pi_{j|i} = \pi_{ij}/\pi_{i+}$$

- Similar notation for **samples** with $p_{ij} = n_{ij}/n$

Notation for Joint, Conditional and Marginal Probabilities in a 2×2 Table

	Column		
Row	1	2	Total
1	p_{11} $(p_{1\|1})$	p_{12} $(p_{2\|1})$	p_{1+} (1.0)
1	p_{21} $(p_{1\|2})$	p_{22} $(p_{2\|2})$	p_{2+} (1.0)
Total	p_{+1}	p_{+2}	1.0

Example (after-life):

$$p_{11} = \frac{435}{1091} = 0.399$$

$$p_{1+} = \frac{582}{1091} = 0.533$$

Independence

Two variables are *statistically independent* if the joint probabilities equal the product of their marginal probabilities:

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

or if the conditional distributions of Y are identical at each level of X:

$$\pi_{j|i} = \pi_{+j} \quad \forall i \quad \text{or} \quad \pi_{j|1} = \cdots = \pi_{j|I} \quad \forall j$$

Comparing Proportions in 2×2 Tables

Example: Aspirin and Heart Attacks in Doctors

Agresti (1996)

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

- Prospective study (Clinical Trial)
- Number on each treatment fixed by design
- Outcome is succes or failure
- Rows are independent binomials
- Overall probability of heart attack in Doctors is low:

$$\frac{293}{22071} = 1.33\%$$

The disease is "rare".

General Notation

In general, we can form the following (2×2) table:

	Outcome		
Treatment	1	2	Total
1	Y_1	$n_1 - Y_1$	n_1
2	Y_2	$n_2 - Y_2$	n_2

- Individuals are given treatment 1 or treatment 2
- Outcome is success or failure

Facts about the distribution

- n_1 and n_2 are fixed by design
- Y_1 and Y_2 are independent with distributions:

$$Y_1 \sim Bin(n_1, \pi_1)$$

$$Y_2 \sim Bin(n_2, \pi_2)$$

Questions of interest (all the same)

- Does treatment affect outcome?
- Are treatment and outcome associated?
- Is the probability of success the same on both treatments?

Hypotheses

- The null hypothesis is

$$H_0 : \pi_1 = \pi_2 = \pi$$

- and the alternative is

$$H_1 : \pi_1 \neq \pi_2$$

Quantifying treatment differences

- When $\pi_1 \neq \pi_2$, we want to quantify how the two probabilities are different.
- In other words, we want a single measure of how the treatments differ. The exact interpretation of these measures will be deferred to later.
- The measures:
  - Difference of proportions (Risk difference) $(\pi_1 - \pi_2)$
  - Relative Risk (Risk Ratio) (π_1/π_2)
  - Odds Ratio (Relative Odds) $(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)})$

Interludium: General formula for variance of treatment difference

- Intuitively, the estimators for π_1 and π_2 should be the proportion of successes in the two groups, i.e.}

$$\begin{aligned} p_1 &= \frac{Y_1}{n_1} \\ p_2 &= \frac{Y_2}{n_2} \end{aligned}$$

  These are the MLE's. But, you can go through a lot of statistical theory to show that these are the MLE's.

- The MLE of a treatment difference

$$g(\pi_1) - g(\pi_2)$$

  is then

$$g(p_1) - g(p_2).$$

Interludium: General formula for variance of treatment difference

- Recall, the variance of a difference of two *independent* random variables is

$$Var[g(\pi_1) - g(\pi_2)] = Var[g(\pi_1)] + Var[g(\pi_2)]$$

- Then, to obtain the large sample variance, we can apply the delta method to $g(\pi_1)$ to get $Var[g(\pi_1)]$ and to $g(\pi_2)$ to get $Var[g(\pi_2)]$ and then sum the two.
- The results (for the estimates) are summarized in the following table:

TREATMENT	ESTIMATE	Var(ESTIMATE)
RISK DIFF	$p_1 - p_2$	$\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$
log(RR)	$\log(p_1/p_2)$	$\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}$
log(OR)	$\log(\frac{p_1/(1-p_1)}{p_2/(1-p_2)})$	$\left[\frac{1}{n_1 p_1} + \frac{1}{n_1(1-p_1)}\right] + \left[\frac{1}{n_2 p_2} + \frac{1}{n_2(1-p_2)}\right.$

Note

The estimated variance of the log-odds ratio can also be written as:

$$\frac{1}{y_1} + \frac{1}{n_1 - y_1} + \frac{1}{y_2} + \frac{1}{n_2 - y_2}$$

Difference of Proportions or Risk Difference

- The *risk difference* is the difference between the "success" probabilities for the two groups:

$$\pi_1 - \pi_2$$

- The *estimated risk difference* is $p_1 - p_2$
- Since Y_1 and Y_2 are independent binomials, we know (see interludium) that the estimated variance of the risk difference is:

$$Var(p_1 - p_2) = \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}$$

- Therefore, to test H_0 or to construct confidence intervals, we can use the following statistic:

$$Z = \frac{(p_1 - p_2) - 0}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim N(0,1) \text{ (WALD)}$$

Example - Aspirin use and heart attack

For the aspirine intake example, construct a 95% confidence interval for the true difference $\pi_1 - \pi_2$.
Can we conclude that the intake of aspirin appears to diminish the risk of myocardial infarction?

Solution

$p_1 = 189/11034 = 0.0171$
$p_2 = 104/11037 = 0.0094$
$p_1 - p_2 = 0.0077$
$\sqrt{\frac{(.0171)(.9829)}{11034} + \frac{(.0094)(.9906)}{11037}} = 0.0015$
Thus, a 95% confidence interval for the true difference is:

$$[(p_1 - p_2) - 1.96 * \sqrt{Var(p_1 - p_2)}; (p_1 - p_2) + 1.96 * \sqrt{Var(p_1 - p_2)}]$$

$$= [0.0077 - 1.96 * 0.0015; 0.0077 + 1.96 * 0.0015]$$

$$= [0.005; 0.011].$$

Example - Aspirin use and heart attack

SAS Program

```
data m.aspirin;
input aspirin outcome count;
cards;
0 1 189
0 0 10845
1 1 104
1 0 10933
;
run;

proc freq data=m.aspirin order=data;
tables aspirin*outcome/riskdiff ;
weight count;
run;
```

Example - Aspirin use and heart attack

SAS Output

	Risk	ASE	(Asymptotic)		(Exact)	
Row 1	0.017	0.001	0.015	0.020	0.015	0.020
Row 2	0.009	0.001	0.008	0.011	0.008	0.011
Total	0.013	0.001	0.012	0.015	0.012	0.015
Difference (Row 1 - Row 2)	0.008	0.002	0.005	0.011		

Example - Aspirin use and heart attack

Interpretation of Risk Difference

The risk difference has the interpretation that the excess risk of a MI on placebo is 0.0077. This "fraction" is not very meaningful for rare diseases, but stated in terms of subjects, we can say that we would expect 77 more MIs in 10000 placebo subjects than in 10000 aspirin users.

Relative Risk

- The *relative risk (RR)* is the ratio of the "success" probabilities for the two groups:

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

  - Can take any nonnegative real number
  - Skewed sampling distribution
  - The log-relative risk is often used to alleviate the restrictions that the relative risk must be positive, i.e.

$$\log(RR) = \log(\frac{\pi_1}{\pi_2}) = \log(\pi_1) - \log(\pi_2),$$

    where

$$-\infty \leq \log(RR) \leq \infty.$$

  - Less skewed sampling distribution, closer to normality.
  - Therefore, best to construct c.i. for $\log(RR)$ and then transform back (by exponentiating) to obtain c.i. for RR.

Relative Risk

- The estimated $\log(RR)$ is $\log(p_1/p_2)$.
- The estimated variance of $\log(p_1/p_2)$ is (see interludium):

$$\frac{1-p_1}{n_1 p_1} + \frac{1-p_2}{n_2 p_2}$$

.

Example - Aspirin use and heart attack

For the aspirine intake example, construct a 95% confidence interval for the true relative risk $\frac{\pi_1}{\pi_2}$.

Example - Aspirin use and heart attack

Solution

$p_1 = 189/11034 = 0.0171$
$p_2 = 104/11037 = 0.0094$
$RR = p_1/p_2 = 1.818$
$\log(RR) = 0.598$
The estimated standard error for $\log(RR)$ is 0.1212.
Hence, a 95% confidence interval for $\log(RR)$ is [0.360;0.836].
Thus, a 95% confidence interval for the true RR is obtained by exponentiating the interval for $\log(RR)$:

$$[1.434; 2.306].$$

Example - Aspirin use and heart attack

SAS program

```
proc freq data=m.aspirin order=data;
tables aspirin*outcome/relrisk;
weight count;
run;
```

Example - Aspirin use and heart attack

SAS output

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Bounds	
Case-Control (odds ratio)	1.832	1.440	2.331
Cohort (Col1 Risk)	1.818	1.433	2.306
Cohort (Col2 Risk)	0.992	0.989	0.995

Sample Size = 22071

Example - Aspirin use and heart attack

Interpretation of Relative Risk

The relative risk has the interpretation that individuals on placebo have almost twice (1.8) the risk (or probability) of a heart attack than individuals on Aspirin.

Odds Ratio

- The *odds ratio (OR)* is the ratio of the "odds" of success versus failure for the two groups:

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

  - Can take any nonnegative real number
  - Skewed sampling distribution
  - The log-odds ratio is often used to alleviate the restrictions that the odds ratio must be positive, i.e.

$$\begin{aligned}\log(OR) &= \log\left(\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}\right)\\ &= \log\left(\frac{\pi_1}{1-\pi_1}\right) - \log\left(\frac{\pi_2}{1-\pi_2}\right)\\ &= logit(\pi_1) - logit(\pi_2),\end{aligned}$$

    where

$$-\infty \leq \log(OR) \leq \infty.$$

  - Less skewed sampling distribution, closer to normality.
  - Therefore, best to construct c.i. for $\log(OR)$ and then transform back

Odds Ratio

- The estimated OR is:
$$\frac{p_1/(1-p_1)}{p_2/(1-p_2)}.$$
- The estimated variance of $\log\left(\frac{p_1/(1-p_1)}{p_2/(1-p_2)}\right)$ is (see interludium):
$$\left[\frac{1}{y_1}+\frac{1}{n_1-y_1}\right]+\left[\frac{1}{y_2}+\frac{1}{n_2-y_2}\right]$$

Odds Ratio

Further Properties

- OR=1 corresponds with independence
- When $1 < OR < \infty$, the odds of success are higher in group 1 than in group 2. Thus, subjects in the first group are more likely to have successes than subjects in group 2, that is $\pi_1 > \pi_2$.
- When $0 < OR < 1$, the odds of success are smaller in group 1 than in group 2. Thus, subjects in the first group are less likely to have successes than subjects in group 2, that is $\pi_1 < \pi_2$.
- Values of OR farther from 1 in a given direction represent stronger level of association.

Odds Ratio

Further Properties

- When order of rows or columns is reversed, the new value of OR is the inverse of the original value.
- When the orientation of the table is reversed (rows become columns and vice versa), the OR does not change. This in contrast to the relative risk, which does not treat the variables symmetrically!!
- When cell counts are very small or any zero cell counts occur, it is preferred to use an amended estimator for the true OR, by adding $1/2$ to each cell count.

Example - Aspirin use and heart attack

For the aspirine intake example, construct a 95% confidence interval for the true odds rato $\frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)}$.

Example - Aspirin use and heart attack

Solution

$p_1 = 189/11034 = 0.0171$
$p_2 = 104/11037 = 0.0094$
$OR = \frac{p_1/(1-p_1)}{p_2/(1-p_2)} = 1.832$
$\log(OR) = 0.605$
The estimated standard error for $\log(OR)$ is 0.1228.
Hence, a 95% confidence interval for $\log(OR)$ is [0.364;0.846].
Thus, a 95% confidence interval for the true OR is obtained by exponentiating the interval for $\log(OR)$:

$$[1.440; 2.331].$$

(cfr. SAS output!)

Example - Aspirin use and heart attack

Interpretation of Odds Ratio

The odds ratio has the interpretation that individuals on placebo have almost twice (1.8) the odds of a heart attack versus no heart attack than individuals on Aspirin.

Example - Aspirin use and heart attack

Summary of results

Estimates and Test Statistics

Parameter	Estimate	Estimated Standard Error	Z-Statistic (Est/SE)
Risk Diff	0.0077	0.00154	5.00
$\log(RR)$	0.598	0.1212	4.934
$\log(OR)$	0.605	0.1228	4.927

- In each case, we reject the null, and the Z-statistic is about 5.
- The WALD test statistic using the Risk difference, log OR and log RR are slightly different.

Confidence intervals

Parameter	Estimate	95% CI
Risk Diff	0.0077	[0.005;0.011]
RR	1.818	[1.433;2.306]
OR	1.832	[1.440;2.331]

- For the OR and the RR, we exponentiated the 95% confidence intervals for the log(OR) and log(RR), respectively.
- None of the confidence intervals contain the null value for no association (0 for the risk difference, 1 for the OR and RR).

Relationship Odds Ratio and Relative Risk

- Recall,

$$
\begin{aligned}
OR &= \frac{p_1/(1-p_1)}{p_2/(1-p_2)} \\
&= \left(\frac{p_1}{p_2}\right)\left[\frac{1-p_2}{1-p_1}\right] \\
&= RR\left[\frac{1-p_2}{1-p_1}\right]
\end{aligned}
$$

Relationship Odds Ratio and Relative Risk

- When the disease is "rare" (such as in the aspirin example),

$$\left[\frac{1-p_2}{1-p_1}\right] \approx 1, \text{ and } OR \approx RR$$

- In the example, the estimated values for OR and RR are 1.832 and 1.818 respectively, i.e. they are almost identical.

Chi-squared Tests of Independence in a 2-way Contingency Table

The null hypothesis of statistical independence in a general $I \times J$ 2-way table has following form:

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

for all i and j.

How do we proceed?

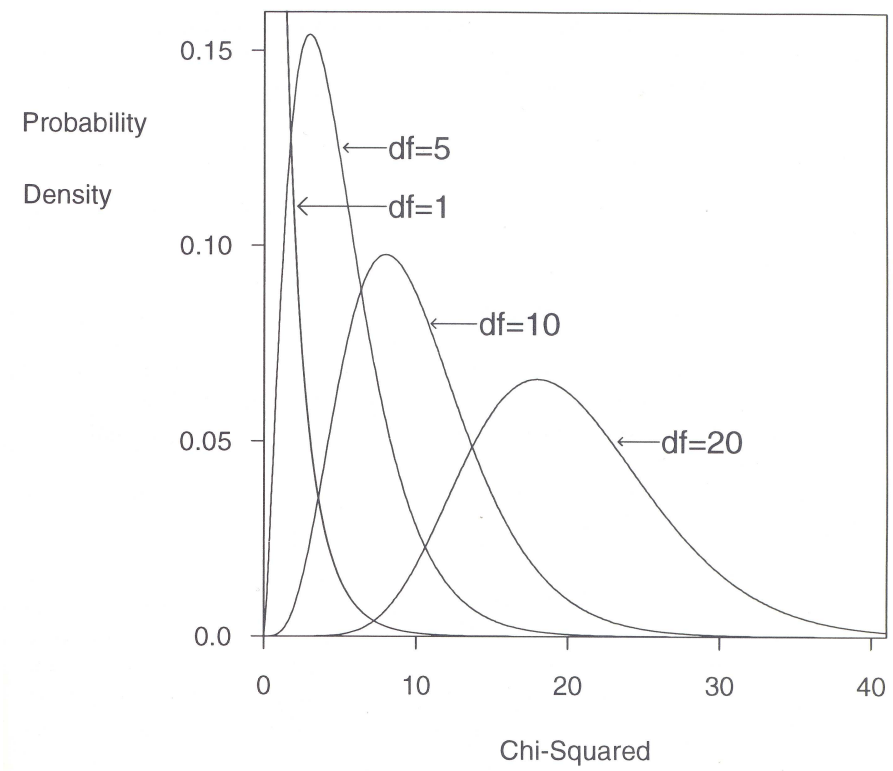
Intuitive Outline:

- In order to judge whether the data contradict H_0, we will compare the sample cell counts to the expected frequencies under the H_0.
- The larger the "*Observed - Expected*" differences, the stronger the evidence against H_0.
- The test statistics used to make such comparisons have large-sample *chi-squared distributions*.

The chi-squared distribution

- Specified by its *degrees of freedom* (df).
  - Mean of chi-squared distribution$=df$.
  - Variance of chi-squared distribution$=2df$.
- Defined only for nonnegative values.
- Skewed to the right.
- Becomes more "bell-shaped" and is concentrated around larger values, as df increases.

Examples of chi-squared distributions



Pearson Statistic

The *Pearson chi-squared statistic* for testing H_0 is

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where

- O_{ij} is the observed cell count,
- E_{ij} is the estimated expected count under the null hypothesis

in the ij-th cell of a $(I \times J)$ table.

Properties

- Pearson's Chi-Square measures the discrepancy between the observed counts, and the estimated expected count under the null.
- If they are similar, you expect the statistic to be small, and for us not to reject the null.
- The minimum value of 0 is obtained when all O_{ij}'s are equal to the E_{ij}'s.
- The X^2-statistic has approximately a chi-squared distribution for "large" sample sizes (where "large" means $E_{ij} \geq 5$) with $(I-1)(J-1)$ degrees of freedom.

Likelihood Ratio Statistic

- An alternative test statistic for testing H_0.
- The test determines the parameter values that maximize the likelihood function under the assumption that H_0 is true.
- It also determines the values that maximize it under the more general condition that H_0 may or may not be true.
- The test is based on the ratio of the maximized likelihoods:

$$\Lambda = \frac{\text{maximum likelihood when parameters satisfy } H_0}{\text{maximum likelihood when parameters are unrestricted}}.$$

Properties of Λ:

- Λ cannot exceed 1.
- Λ far below 1 indicates that the maximum likelihood is much larger when the parameters are not forced to satisfy H_0. Therefore it indicates strong evidence against H_0.

Likelihood Ratio Test Statistic

The likelihood ratio statistic equals:

$$-2\log(\Lambda).$$

- Non-negative
- "Small" values of Λ yield "large" values of $-2\log(\Lambda)$.

Likelihood ratio statistic for 2-way contingency tables: One can show that the likelihood ratio statistic for 2-way contingency tables can be written as (trust me!):

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} O_{ij}\log\left(\frac{O_{ij}}{E_{ij}}\right).$$

Likelihood Ratio Test Statistic

- Minimum value of 0 is attained when all $O_{ij} = E_{ij}$.
- Larger values provide stronger evidence against H_0.
- Approximate chi-squared sampling distribution with $df = (I-1)(J-1)$

Notes

- Pearson χ^2 and likelihood-ratio G^2 share many properties and commonly yield the same conclusions.
- When H_0 is true and sample cell counts large, the 2 statistics have the same chi-squared distribution and their numerical values are similar.
- Each statistic has advantages and disadvantages (see later).

Expected Cell Counts under H_0

Under the null hypothesis of independence, the expected cell counts can be estimated by:

$$E_{ij} = np_{i+}p_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}.$$

Hence,

$$E_{ij} = \frac{[i^{th}\text{row total}][j^{th}\text{column total}]}{[\text{total sample size}]}$$

Example - Aspirin use and heart attack

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10845	11034
Aspirin	104	10933	11037
Total	293	21778	22071

Estimated Expected Cell Counts

If you work through the (2×2) table, you will see

$$E_{11} = \frac{[1^{st}\text{row total}].[1^{st}\text{column total}]}{[\text{total sample size}]}$$

$$= \frac{(11034)(293)}{22071}$$

$$= 146.48$$

$$E_{12} = \frac{[1^{st}\text{row total}].[2^{nd}\text{column total}]}{[\text{total sample size}]}$$

$$= \frac{(11034)(21778)}{22071}$$

$$= 10888$$

$$E_{21} = \frac{[2^{nd}\text{row total}].[1^{st}\text{column total}]}{[\text{total sample size}]}$$

$$= \frac{(11037)(293)}{22071}$$

$$= 146.52$$

SAS Program

```
data m.aspirin;
input trt $ outcome $ count;
cards;
1(P) HA 189
1(P) NHA 10845
2(A) HA 104
2(A) NHA 10933
;

proc freq data=m.aspirin;
table trt*outcome/expected chisq measures;
weight count; /* tells SAS how many observations */
              /* in each cell of 2x2 table       */
run;
```

SAS output

TABLE OF TRT BY OUTCOME

TRT	OUTCOME		
Frequency			
Expected			
Percent			
Row Pct			
Col Pct	HA	NHA	Total
1(P)	189	10845	11034
	146.48	10888	
	0.86	49.1	49.99
	1.71	98.29	
	64.51	.80	
2(A)	104	10933	11037
	146.52	10890	
	0.47	49.54	50.01
	0.94	99.06	
	35.49	50.20	
Total	293	21778	22071
	1.33	98.67	100.00

SAS Output: Continued

STATISTICS FOR TABLE OF TRT BY OUTCOME

Statistic	DF	Value	Prob
Chi-Square	1	25.014	0.001<=Pearson
Likelihood Ratio Chi-Square	1	25.372	0.001<=LR STAT
Continuity Adj. Chi-Square	1	24.429	0.001
Mantel-Haenszel Chi-Square	1	25.013	0.001
Fisher's Exact Test (Left)			1.000
(Right)			3.25E-07
(2-Tail)			5.03E-07

SAS Output: Continued

Estimates of the Relative Risk (Row1/Row2)

Type of Study	Value	95% Confidence Bounds	
Case-Control	1.832	1.440	2.331<=(OR, using logOR)
Cohort (Col1 Risk)	1.818	1.433	2.306<=(RR, using logRR)
Cohort (Col2 Risk)	0.992	0.989	0.995

Comparing Test Statistics

- We want to compare test statistics for

$$H_0 : p_1 = p_2 = p \text{ versus } H_1 : p_1 \neq p_2$$

- Looking at the (square of the) Wald statistics from earlier, as well as the Likelihood Ratio and Pearson's Chisquare,

STATISTIC	VALUE
Wald-Risk Diff	25.00
Wald-Log(RR)	24.34
Wald-Log(OR)	24.28
LR	25.37
Pearson's	25.01

Comparing Test Statistics

- All of the test statistics are approximately χ^2_1 under the null.
- Note, the likelihood ratio and Pearson's chi-squared statistic just depend on the estimated expected cell counts (predicted probabilities) and not how we measure the treatment difference.
- However, the WALD statistic does depend on what treatment difference (Risk Difference, log OR, log RR) we use in the test statistic.
- In other words, the WALD test statistics using the Risk Difference, log OR, and log RR will usually be slightly different (as we see in the example).

Chi-squared Tests of Independence: Summary

The null hypothesis of statistical independence in a general $I \times J$ 2-way table

	Column			
Row	1	2	...	J
1	n_{11}	n_{12}	...	n_{1J}
2	n_{21}	n_{22}	...	n_{2J}
...				
I	n_{I1}	n_{I2}	...	n_{IJ}

has following form:

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

for all i and j.

How do we proceed?

- We estimate the expected cell frequencies under H_0:

$$\hat{\mu}_{ij} = np_{i+}p_{+j} = n\frac{n_{i+}}{n}\frac{n_{+j}}{n} = \frac{n_{i+}n_{+j}}{n}.$$

- The Pearson and likelihood-ratio statistics equal:

$$X^2 = \sum \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}, \qquad G^2 = 2\sum n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right).$$

- The large-sample distributions of these statistics are chi-squared distributions with $df = (I-1)(J-1)$.
- The value of df is the difference between the number of parameters under the alternative and null hypotheses, or:

$$(I-1)(J-1).$$

Gender Gap Example

Cross classification of Party Identification by Gender

	Party Identification			
Gender	Democrat	Independent	Republicans	Total
Females	279	73	225	577
	(261.4)	(70.7)	(244.9)	
Males	165	47	191	403
	(182.6)	(49.3)	(171.1)	
Total	444	120	416	980

Estimated expected frequencies for hypothesis of independence in parentheses.

- $X^2 = 7.01$ (check!)
- $G^2 = 7.00$ (check!)
- $df = (2-1)(3-1) = 2$
- For each statistic, p-value=0.03
- This suggests that party identification and gender are associated!

SAS Program

```
data m.party;
input gender $ party $ count;
cards;
F D 279
F I 73
F R 225
M D 165
M I 47
M R 191
;

proc freq data=m.party;
table gender*party/expected chisq;
weight count;
run;
```

SAS Output

TABLE OF GENDER BY PARTY

GENDER	PARTY			
Frequency				
Expected				
Percent				
Row Pct				
Col Pct	D	I	R	Total
F	279	73	225	577
	261.42	70.653	244.93	
	28.47	7.45	22.96	58.88
	48.35	12.65	38.99	
	62.84	60.83	54.09	
M	165	47	191	403
	182.58	49.347	171.07	
	16.84	4.80	19.49	41.12
	40.94	11.66	47.39	
	37.16	39.17	45.91	
Total	444	120	416	980
	45.31	12.24	42.45	100.00

SAS Output (Continued)

STATISTICS FOR TABLE OF GENDER BY PARTY

Statistic	DF	Value	Prob
Chi-Square	2	7.010	0.030<= Pearson's
Likelihood Ratio Chi-Square	2	7.003	0.030<= LR STAT
Mantel-Haenszel Chi-Square	1	6.758	0.009
Phi Coefficient		0.085	
Contingency Coefficient		0.084	
Cramer's V		0.085	

Residuals

- The test statistics and their associated p-values, described in the previous section, describe the evidence against the null hypothesis.
- **How can we describe/understand the *nature* of this evidence?**
- Adjusted Residuals:

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

  - The sample distribution of each residual under H_0 is a large-sample standard normal distribution.
  - In practice this means that a residual, which exceeds 2 or 3 in absolute value, indicates lack of fit of H_0 *in that cell*.

Gender Gap Example

Cross classification of Party Identification by Gender

	Party Identification			
Gender	Democrat	Independent	Republicans	Total
Females	279	73	225	577
	(2.29)	(0.46)	(-2.62)	
Males	165	47	191	403
	(-2.29)	(-0.46)	(2.62)	
Total	444	120	416	980

Adjusted Residuals (in parentheses).

Calculating the Adjusted Residuals

First Cell:

$$\begin{aligned} n_{11} &= 279 \\ \hat{\mu}_{11} &= 261.4 \\ p_{1+} &= 577/980 = 0.589 \\ p_{+1} &= 444/980 = 0.453 \end{aligned}$$

Therefore, the adjusted residual for the first cell equals

$$\frac{279 - 261.4}{\sqrt{261.4(1 - 0.589)(1 - 0.453)}} = 2.29$$

Interpretation

- Large (> 2) positive residuals for female democrats and male republicans
- Large (< -2) negative residuals for male democrats and female republicans
- Thus, significantly more female democrats and male republicans and fewer female republicans and male democrats than expected from the hypothesis of independence.

How to describe evidence of gender gap?

- Odds ratio of 2×2 table of Democrat and Republican identifiers.
- OR=279*191/225*165=1.44
- The odds of identifying with the democrats, rather than with the republicans is 44% higher for females than for males.

Question

The residual for females in a given party identification class is the negative of the one for males. Why?

SAS Program

```
data m.party;
input gender $ party $ count;
cards;
F D 279
F I 73
F R 225
M D 165
M I 47
M R 191
;

proc genmod data=m.party;
class gender party;
model count=gender party/dist=poi link=log obstats residuals;
run;
```

SAS Output

Observation Statistics

COUNT	Pred	Xbeta	Std	HessWgt	Lower	Upper
279	261.4163	5.5661	0.0545	261.4163	234.9544	290.8585
73	70.6531	4.2578	0.0951	70.6531	58.6371	85.1313
225	244.9306	5.5010	0.0558	244.9306	219.5452	273.2513
165	182.5837	5.2072	0.0609	182.5837	162.0290	205.7458
47	49.3469	3.8989	0.0990	49.3469	40.6462	59.9102
191	171.0694	5.1421	0.0622	171.0694	151.4450	193.2367

Observation Statistics

Resraw	Reschi	Resdev	StResdev	StReschi	Reslik
17.5837	1.0875	1.0757	2.2681	2.2932	2.2876
2.3469	0.2792	0.2777	0.4623	0.4648	0.4639
-19.9306	-1.2735	-1.2914	-2.6545	-2.6178	-2.6265
-17.5837	-1.3013	-1.3231	-2.3315	-2.2932	-2.3056
-2.3469	-0.3341	-0.3368	-0.4686	-0.4648	-0.4667
19.9306	1.5238	1.4956	2.5693	2.6178	2.6014

Limitations of χ^2 Tests

- Chi-squared tests only indicate the degree of evidence for an association.
- They do not study the nature of the association
  - Study residuals
  - Estimate odds ratio's, ...
- X^2 and G^2 chi-squared require large samples
  - Special techniques required for small samples
- X^2 and G^2 do not depend on the order in which rows and columns are listed.
  - When at least one variable is ordinal, more powerful tests of independence can usually be applied.

Testing Independence for Ordinal Data

- The chi-squared test of independence using X^2 and G^2 treats both classifications as *nominal*.
- When the rows and/or columns are *ordinal*, test statistics that utilize the ordinality are usually more appropriate (powerful).
- In that case, the investigation of "trend" association is quite common:

  > As the level of X increases, responses on Y tend to increase toward higher levels, or decrease toward lower levels.

- We look for a single parameter that can describe such an ordinal trend association.

Intuitive Outline

- Assign scores to categories
- Measure the degree of *linear trend* or correlation.

Linear Trend Alternative to Independence

- Let $u_1 \leq u_2 \leq u_I$ denote scores for the rows.
- Let $v_1 \leq v_2 \leq v_J$ denote scores for the columns.
- Scores reflect distances between categories.
- A statistic for testing the null hypothesis of independence against the two-sided alternative hypothesis of nonzero true correlation is:

$$M^2 = (n-1)r^2,$$

where r is the *Pearson product-moment* correlation:

$$r = \frac{\sum_{i,j} u_i v_j n_{ij} - (\sum_i u_i n_{i+})(\sum_j v_j n_{+j})/n}{\sqrt{\left[\sum_i u_i^2 n_{i+} - \frac{(\sum_i u_i n_{i+})^2}{n}\right]\left[\sum_j v_j^2 n_{+j} - \frac{(\sum_j v_j n_{+j})^2}{n}\right]}}.$$

  - r can be computed using standard software
  - r falls between -1 and $+1$
  - Independence implies $r = 0$

Linear Trend Alternative to Independence

- M^2 has a large-sample chi-square distribution with only 1 degree of freedom (1 single extra parameter: the correlation r).
- $$M = \sqrt{n-1}r$$

  has a standard normal distribution
- M applies to "directional" alternatives

Advantages over X^2 and G^2

- Power advantage, if association truly has positive or negative trend

"*Easier to detect a significant difference*"

  - a large M^2 value based on $df = 1$ falls farther out in the right hand tail than a comparable value of X^2 or G^2.
  - thus, P-value is smaller
  - thus, easier to detect significant positive or negative trend!!

- When several cell counts are small, the chi-squared approximations for X^2 and G^2 are worse than for M^2.
  - for small sample sizes the sampling distributions tend to be closer to chi-squared when df is smaller.

Choice of Scores

- If the variable is ordinal, but not numerical, such as

  (NONE, MILD, SEVERE)

  investigators often just use (1,2,3).
- If the variable is a crude grouping of an underlying continuous variable, such as AGE with levels [20,30), [30,40), [40,50), investigators often use the *midpoints* of the intervals
- For some variables, such as "dose", investigators mostly use the *actual numerical value*.
- An alternative approach is the use of "midranks", however this approach is often not appropriate, since it may not reflect distances between categories.

Alcohol and Infant Malformation Example

Prospective study of maternal drinking and infant malformations.

Alcohol Consumption	Malformation		
	Absent	Present	Total
0	17066	48	17114
< 1	14464	38	14502
$1 - 2$	788	5	793
$3 - 5$	126	1	127
≥ 6	37	1	38

Problem

Conduct a test for independence, using

(a) X^2

(b) G^2

Do you think the sampling distributions for these test statistics are really chi-squared distributions? Why? Why not?

Solution

- $df = 4$
- $X^2 = 12.1$ (p-value=0.02)
- $G^2 = 6.2$ (p-value=0.19)

Mixed messages!!

In both cases, we have ignored the ordinality of alcohol consumption!

Cell counts are small, moderate, extremely large, ... Hence, sampling distributions of X^2 or G^2 may not be chi-squared.

Example (Continued)

Alcohol Consumption	Malformation Absent	Malformation Present	Total	Percentage Present	Adjusted Residual
0	17066	48	17114	0.28	-0.18
< 1	14464	38	14502	0.26	-0.71
$1 - 2$	788	5	793	0.63	1.84
$3 - 5$	126	1	127	0.79	1.06
≥ 6	37	1	38	2.63	2.71

- The percentage of malformation cases increases at each increase in level of alcohol consumption.
- The adjusted residuals for malformation cases
  - are negative for low levels of alcohol consumption
  - are positive for high levels of alcohol consumption
  - change substantially with slight changes in alcohol consumption

suggests a tendency for malformations to be more likely at higher levels of alcohol consumption

Example (Continued)

Applying the Trend Test...

- Requires scores for levels of alcohol
- Use midpoints of categories:

$$v_1 = 0, v_2 = 0.5, v_3 = 1.5, v_4 = 4.0, v_5 = 7.0$$

- Last score arbitrary
- Calculate r and M^2 using software (e.g. SAS, proc CORR and PROC FREQ).

SAS Program

```
data m.infants;
input alcohol malform counts;
cards;
0 0 17066
0 1 48
0.5 0 14464
0.5 1 38
1.5 0 788
1.5 1 5
4 0 126
4 1 1
7 0 37
7 1 1
;

proc corr data=m.infants;
var alcohol malform;
freq counts;
run;

proc freq data=m.infants;
tables alcohol*malform/chisq cmh1;
weight counts;
run;
```

SAS Output

```
Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 32574
/ FREQ Var = COUNTS

                          ALCOHOL          MALFORM

          ALCOHOL         1.00000          0.01420
                           0.0              0.0104

          MALFORM         0.01420          1.00000
                           0.0104           0.0


   STATISTICS FOR TABLE OF ALCOHOL BY MALFORM

Statistic                     DF      Value       Prob

Chi-Square                     4     12.082      0.017
Likelihood Ratio Chi-Square    4      6.202      0.185
Mantel-Haenszel Chi-Square     1      6.570      0.010
Phi Coefficient                       0.019
Contingency Coefficient               0.019
Cramer's V                            0.019

Sample Size = 32574
WARNING: 30% of the cells have expected counts less
         than 5. Chi-Square may not be a valid test.


Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

 Statistic   Alternative Hypothesis    DF       Value      Prob

    1        Nonzero Correlation        1       6.570     0.010
```

Results

- $r = 0.014$
- $M^2 = (32573)(0.014)^2 = 6.6$
- p-value=0.01 => strong evidence of non-zero correlation!
- $M = 2.56$ with p-value=0.001 => strong evidence of a *positive* correlation

FUTURE MUSIC:

How to build models in which malformation probabilities change linearly according to alcohol consumption?

Trend Test for $I \times 2$ and $2 \times J$ Tables

$I \times 2$ Table

- Binary response variable
- How does proportion of "successes" (or "failures") varies across the levels of X?
- M^2 tries to detect a linear trend.
- Small p-values indicate strong evidence of such trend.
- This $I \times 2$ version of the trend test is called the *Cochran-Armitage Trend Test*

$2 \times J$ Table

- Is there a difference between the two row means of the scores on Y?
- Small p-values suggest that the true difference in row means is nonzero.
- When we use midrank scores for Y, the test is called the *Wilcoxon or Mann-Whitney* test (cfr. nonparametrics course).

Nominal-Ordinal Tables

- M^2 treats both classifications as ordinal.
- When one variable is nominal but has only 2 categories, it still applies.
- When one variable is nominal with more than 2 categories, a more complex test statistic is needed (not considered here).

Exact Inference for Small Samples

The confidence intervals and tests presented so far in this chapter are large-sample methods.
As the sample size grows, the cell counts grow, and "chi-squared" statistics such as X^2, G^2 and M^2 have distributions that are more nearly chi-squared.
When the sample size is small, one can perform inference using *exact* distributions rather than large-sample approximations.
Fisher (1935) criticized large sample methods:

> *"Not only does it take a cannon to shoot a sparrow, but it misses the sparrow . . . . Only by tackling small sample problems on their merits does it seem possible to apply accurate tests."*

In this section we will restrict attention to 2×2 tables only.

Fisher's Exact Test: One-Sided Test

- The null hypothesis of independence corresponds to an odds ratio of $\theta = 1$:

$$H_0 : \theta = 1$$

- The alternative hypothesis is

$$H_a : \theta > 1$$

or

$$H_a : \theta < 1$$

Intuitive Outline

- The central idea is to enumerate all possible tables consistent with a given set of marginal totals and add up the probabilities of those tables "more extreme" than the one observed.
- For *given* row and column marginal totals, the value for n_{11} determines the other three cell counts.
- Hence, conditional on the margins, a 2×2 table is a one-dimensional random variable (n_{11}) with a *hypergeometric* distribution.
- When $\theta = 1$, the probability of a particular value n_{11} equals:

$$P(n_{11}) = \frac{\begin{pmatrix} n_{1+} \\ n_{11} \end{pmatrix} \begin{pmatrix} n_{2+} \\ n_{+1} - n_{11} \end{pmatrix}}{\begin{pmatrix} n \\ n_{+1} \end{pmatrix}}.$$

How do we proceed?

The process for doing hand calculations (in the case $H_a : \theta < 1$) would be as follows:

1. Rearrange the rows and columns of the observed table so that the smaller total is in the first row and the smaller column total is in the first column.
2. Start with the table having 0 in the (1,1) cell (top-left cell). The other cells in this table are determined automatically from the fixed row margins and column margins
3. Construct the next table by increasing the (1,1) cell from 0 to 1 and decreasing all other cells accordingly.
4. Calculate and add up the probabilities of those tables with cell (1,1) having values "more extreme" than the observed frequency.

The case $H_a : \theta > 1$ can be done similarly.

Interpretation

Given the marginal totals, tables having larger (smaller) n_{11} values also have larger (smaller) sample odds ratios $\hat{\theta} = (n_{11}n_{22})/(n_{12}n_{21})$.
Hence, they provide stronger evidence in favor of the alternative $H_a : \theta > 1 (< 1)$.
The p-value equals the right-tail (left-tail) hypergeometric probability that n_{11} is at least (at the most) as large as the observed value.

Fisher's Tea Taster Experiment

A colleague of Fisher's at Rothamsted Experiment Station near London claimed that, when drinking tea, she could distinguish whether milk or tea was added to the cup first. To test her claim, Fisher designed an experiment in which she tasted eight cups of tea. Four cups had milk added first, and the other four had tea added first. She was told there were four cups of each type, so that she should try to select the four that had milk added first. The cups were presented to her in random order. The results of the experiment are given in the following table.

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	8

Fisher's Tea Taster (Continued)

- We conduct Fisher's exact test of

$$H_0 : \theta = 1$$

against

$$H_a : \theta > 1.$$

  - H_0 states that Fisher's colleague's guess was independent of the order of pouring
  - H_a reflects Fisher's colleague's claim: a positive association between true order of pouring and her guess.
- The design fixes both margins naturally.
- Under H_0, the distribution of n_{11} is the hypergeometric distribution defined for all 2×2 tables having row and column margins (4,4).
- The observed table has null probability

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

Fisher's Tea Taster (Continued)

- The only table that is more extreme, for the alternative $H_a : \theta > 1$ consists of four correct guesses. It has $n_{11} = n_{22} = 4$ and $n_{12} = n_{21} = 0$. Its probability is

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

- For a right one-sided alternative, the p-value equals the right-tail probability that n_{11} is at least as large as observed:

$$p\text{-value} = \frac{16}{70} + \frac{1}{70} = 0.243.$$

Conclusion

The experiment did not establish an association between the actual order of pouring and the guess.

SAS Program

```
data m.tea;
input poured guess count;
cards;
1 1 3
2 1 1
1 2 1
2 2 3
;

proc freq data=m.tea;
weight count;
tables poured*guess/exact;
run;
```

SAS Output

STATISTICS FOR TABLE OF POURED BY GUESS

Statistic	DF	Value	Prob
Chi-Square	1	2.000	0.157
Likelihood Ratio Chi-Square	1	2.093	0.148
Continuity Adj. Chi-Square	1	0.500	0.480
Mantel-Haenszel Chi-Square	1	1.750	0.186
Fisher's Exact Test (Left)			0.986
(Right)			0.243
(2-Tail)			0.486
Phi Coefficient		0.500	
Contingency Coefficient		0.447	
Cramer's V		0.500	

Sample Size = 8
WARNING: 100% of the cells have expected counts less
than 5. Chi-Square may not be a valid test.

R Code

```
TeaTasting <-matrix(c(3, 1, 1, 3),nrow = 2,
                    dimnames = list(Guess = c("Milk", "Tea"),
                                    Truth = c("Milk", "Tea")))
fisher.test(TeaTasting, alternative = "greater")
```

R Output

```
 Fisher's Exact Test for Count Data

data:  TeaTasting
p-value = 0.2429
alternative hypothesis: true odds ratio is greater than 1
95 percent confidence interval:
 0.3135693       Inf
sample estimates:
odds ratio
  6.408309
```

Fisher's Exact Test: Two-Sided Test

- Suppose we want to test

$$H_0 : \theta = 1$$

versus

$$H_a : \theta \neq 1$$

.

- The exact p-value here is:

$$\Pr\left[\begin{array}{l}\text{seeing a result as likely or less likely than} \\ \text{the observed result in either direction}\end{array} \middle| H_0 : \theta = 1\right].$$

How to calculate the 2-sided p-value?

1. Calculate the probability of the observed result under the null (π).
2. Calculate the probabilities of all possible values of n_{11}:

$$\pi_l = Pr[n_{11} = \ell | h_0 : \theta = 1].$$

3. Sum the probabilities that are less than or equal to the observed probability π.

Example: Fisher's Tea Taster

n_{11}	Probability
0	0.014
1	0.229
2	0.514
3	0.229
4	0.014

The exact two-sided p-value is obtained by summing all probabilities who are no greater than $P(3) = 0.229$.
This yields:

$$P(0) + P(1) + P(3) + P(4) = 0.486$$

(cfr. SAS Output)
Note:
When the row or column marginal totals are equal, the hypergeometric distribution is symmetric, and the two-sided p-value doubles the one-sided one.

Small Sample Confidence Interval for Odds Ratios

- Generalization of Fisher's Exact Test
- A 95% confidence interval contains all values of θ_0 for which the exact test of $H_0 : \theta = \theta_0$ yields $p > 0.05$.
- Computations for these types of confidence intervals are complex and require specialized software (StatXact, Cytel Software, Cambridge, MA).

Example: Fisher's Tea Taster

The "exact" 95% confidence interval for the true odds ratio equals

$$(0.21, 626.17)$$

The interval is so wide, because the sample size is so small.

R Code

```
TeaTasting <-matrix(c(3, 1, 1, 3),nrow = 2,
                    dimnames = list(Guess = c("Milk", "Tea"),
                                    Truth = c("Milk", "Tea")))
fisher.test(TeaTasting, alternative = "two.sided")
```

R Code

```
 Fisher's Exact Test for Count Data

data:  TeaTasting
p-value = 0.4857
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
   0.2117329 621.9337505
sample estimates:
odds ratio
  6.408309
```

Chapter 3: Three-way Contingency Tables

This chapter generalizes the methods of Chapter 2 regarding two-way contingency tables to multi-way tables.

The main topic is analyzing the association between two categorical variables X and Y, while controlling for effects of a possibly confounding variable Z.

Example

Is passive smoking associated with lung cancer?

A valid analysis should account for several potential "confounders" such as age, socio-economic status, etc.:

For example,suppose spouses of nonsmokers tend to be younger than spouses of smokers, and suppose younger people are less likely to have lung cancer. Then, a lower proportion of lung cancer cases among spouses of nonsmokers may simply reflect their lower average age.

Partial Association

Let us first introduce some notation:

Y: categorical response variable
X: categorical explanatory variable
Z: categorical control variable

Partial Tables

Partial table:
a two-way cross-sectional slice of the three-way table where X and Y are cross classified at separate levels of the control variable Z.

- Such a table displays the $X - Y$ relationship at fixed levels of Z.
- Hence, it shows the effect of X on Y while controlling for Z.

Marginal table:
a two-way contingency table obtained by combining the partial tables.

- Each cell count in the marginal table is a sum of counts from the same cell locations in the partial tables.
- The marginal table *ignores* Z rather than controlling for it.

Example

- 715 births, cross classified by clinic (Z), prenatal care (X) and outcome (Y)
- prevalence study

CLINIC	PRENATAL CARE	OUTCOME died	OUTCOME lived
1	less	3	176
	more	4	293
2	less	17	197
	more	2	23

Construct the corresponding partial tables and the marginal table.

Solution

Partial tables

Clinic=1

CARE	OUTCOME	
less	3	176
more	4	293

Clinic=2

CARE	OUTCOME	
less	17	197
more	2	23

Marginal table

CARE	OUTCOME	
less	20	373
more	6	316

Conditional Associations versus Marginal Associations

Conditional Association: association in a partial table (because they refer to the effect of X on Y, conditional on the level of Z)

Marginal Association: association in a marginal table

Remarks:

- Conditional associations in partial tables can be very different from associations in marginal tables
- It can even be misleading to analyze only a marginal table of a multi-way contingency table (Simpson's Paradox: see further down)

Example (continued)

Marginal and partial odds ratios for CLINIC (Z), CARE (X) and OUTCOME (Y).

Association	(Z,X)	(Z,Y)	(X,Y)
MARGINAL	0.070*	0.173*	2.824*
PARTIAL (1)	0.088*	0.198*	1.249
PARTIAL (2)	0.070*	0.157*	0.992

* p $<$ 0.05 for OR=1

Example (Continued)

- The marginal and partial odds ratios for (CLINIC,CARE) are similar.
- The marginal and partial odds ratios for (CLINIC,OUTCOME) are similar.
- On the other hand, we see that the marginal odds ratio for (CARE,OUTCOME) is 2.8, meaning "the odds of death is 2.8 times greater for less care than more care".
- However, controlling for the clinic, CARE and OUTCOME appear to be independent (partial OR's about 1, and non-significant).

Confounding

- When the partial and marginal associations are different, there is said to be CONFOUNDING.
- Confounding occurs when two variables are associated with a third in a way to obscure their relationship.

Example (Continued)

CLINIC (Z) can confound the relationship between CARE (X) and OUTCOME (Y) when Z is related to both X and Y:

1. The CLINIC (Z) and CARE (X) relationship

- The marginal and partial OR ≈ 0.07.
- In particular, for clinic 1, a majority of infants received prenatal care (60%), regardless of outcome, whereas in clinic 2, only about 10% of the infants received prenatal care.

2. The CLINIC (Z) and OUTCOME (Y) relationship

- The marginal and partial OR ≈ 0.2.
- There is another important difference between the 2 clinics; the outcome differs by clinic; the death rate in clinic 1 is 1.5% and in clinic 2 it is 8%.

Discussion

- Thus, infants in clinic 2 tend to receive "less" prenatal care, and infants in clinic 2 tend to die more, making it appear, when just looking at CARE (X) and OUTCOME (Y), as if infants who get less prenatal care tend to die more.
- In particular, the marginal OR=2.8.
- However, within clinic, there appears to be no relationship between CARE and OUTCOME. The partial OR ≈ 1.

Death Penalty Example: Simpson's Paradox

This example (Agresti 1996, 2002) originally comes from an article that studied effects of racial characteristics on whether individuals convicted of homicide receive the death penalty.

Victim's Race	Defendant's Race	Death Penalty Yes	Death Penalty 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

Death Penalty Example: Simpson's Paradox

- The 674 subjects classified in the table were the defendants in indictments involving cases with multiple murders in Florida between 1976 and 1987.
- The variables are:
  - $Y =$ "death penalty verdict" (Yes/No)
  - $X =$ "race of defendant" (White/Black)
  - $Z =$ "race of victim" (White/Black)

Death Penalty Example: Simpson's Paradox

Goal

Study the effect of defendant's race (X) on death penalty verdict (Y), treating victim's race as control variable (Z).

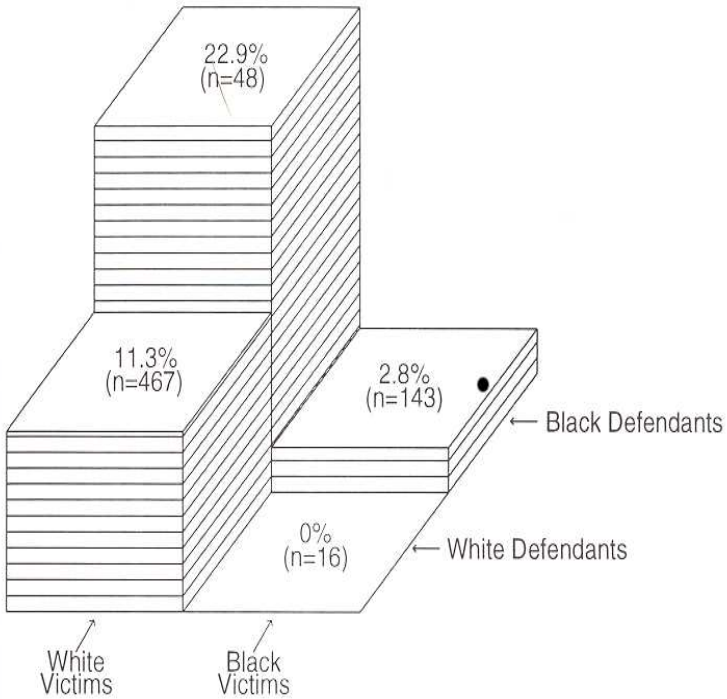


Figure Percent receiving death penalty.

Above figure displays the percentage of defendants who received death penalty.

Death Penalty Example: Simpson's Paradox

Results for Partial Tables

- When the victims were white, the death penalty was imposed

$$22.9\% - 11.3\% = 11.6\%$$

more often for black defendants than for white defendants.

- When the victims were black, the death penalty was imposed

$$2.8\% - 0\% = 2.8\%$$

more often for black defendants than for white defendants.

Thus, *controlling* for victim' race (by keeping it fixed), the percentage of "yes" death penalty verdicts was higher for black defendants than for white defendants.

Death Penalty Example: Simpson's Paradox

Results for Marginal Table

The marginal table for defendant's race and death penalty verdict is obtained by summing the cell counts over the two levels of victim's race:

Defendant's Race	Death Penalty Yes	Death Penalty No	Percentage Yes
White	53	430	11.0
Black	15	176	7.9

- *Ignoring* victim's race, the percentage of "yes" death penalty verdicts was lower for black defendants than for white defendants

The association reverses direction compared to the partial tables!

Death Penalty Example: Simpson's Paradox

Question

Why does the association between death penalty verdict and defendant's race differ so much when we ignore versus control victim's race?

Let us therefore study the relationship between the control variable (victim's race) and each of the other variables.

Death Penalty Example: Simpson's Paradox

Association between victim's race and defendant's race

- Extremely strong association!
- Indeed, the OR can be calculated as:

$$\frac{467 \times 143}{48 \times 16} = 87.0 \quad (p < 0.001)$$

- Hence, the odds that a white defendant had a white victim (versus a black victim) are estimated to be 87 times the odds that a black defendant had a white victim (versus a black victim).

Death Penalty Example: Simpson's Paradox

Association between victim's race and death penalty

- The OR can be calculated as:

$$\frac{64 \times 155}{451 \times 4} = 5.5 \quad (p < 0.001)$$

- Hence, the odds that of obtaining a death penalty (versus not) for a white victim is 5.5 times the odds of obtaining a death penalty (versus not) for a black victim.

Death Penalty Example: Simpson's Paradox

Summary

- Whites tend to kill whites and killing whites is more likely to result in the death penalty.
- This suggests that the marginal association (ignoring victim's race) should show a greater tendency for white defendants to receive the death penalty than do the conditional associations.

Simpson's Paradox

The result that a marginal association can have different direction from the conditional associations is called *Simpson's Paradox*. This result applies to quantitative as well as categorical variables.

Conditional and Marginal Odds Ratios: Definition and Notation

One can describe marginal and conditional associations using odds ratios. We illustrate for $2 \times 2 \times K$ tables, where K denotes the number of levels of a control variable Z.

- Let n_{ijk}=number of subjects with $X = i, Y = j, Z = k$.
- Denote the expected frequency by $\mu_{ijk} = E(Y_{ijk})$.

Conditional odds ratio

The conditional $X - Y$ odds ratio within a fixed level k of Z is:

$$\begin{aligned} \theta_{XY(k)} &= \frac{\mu_{11k}\mu_{22k}}{\mu_{12k}\mu_{21k}} \\ \hat{\theta}_{XY(k)} &= \frac{n_{11k}n_{22k}}{n_{12k}n_{21k}} \end{aligned}$$

Marginal odds ratio

The marginal $X - Y$ odds ratio is:

$$\begin{aligned} \theta_{XY} &= \frac{\mu_{11+}\mu_{22+}}{\mu_{12+}\mu_{21+}} \\ \hat{\theta}_{XY} &= \frac{n_{11+}n_{22+}}{n_{12+}n_{21+}}, \end{aligned}$$

Death Penalty Example

Calculate

1. the conditional odds ratio for the association between defendant's race and death penalty when victim's race is white.
2. the conditional odds ratio for the association between defendant's race and death penalty when victim's race is black.
3. the marginal odds ratio for defendant's race and death penalty.

Death Penalty Example: Solution

1. $\hat{\theta}_{XY(w)} = (53 \times 37)/(414 \times 11) = 0.43$
   (odds for white defendants to receive death penalty is 43% of the odds for black defendants to receive death penalty)
2. $\hat{\theta}_{XY(b)} = (0 \times 139)/(4 \times 16) = 0.00$
   (since death penalty was never given to white defendants having black victims)
3. $\hat{\theta}_{XY} = (53 \times 176)/(15 \times 430) = 1.45$
   (odds of death penalty were 45% higher for white defendants than for black defendants)

Note:
This reversal in the association when we control for victim's race illustrates Simpson's paradox.

Marginal versus Conditional Independence

Conditional Independence

- X and Y are *conditionally independent*, *given* Z, if X and Y are independent in EACH partial table.
- All conditional odds ratios between X and Y then equal 1.

Marginal versus Conditional Independence

Marginal versus Conditional Independence

- Conditional independence of X and Y does not imply marginal independence.
- If Z is partially related to BOTH X and Y, then it confounds the relationship between X and Y.
- Alternatively, if either
  1. Z and X are conditionally independent given Y, i.e.

$$\theta_{ZX(y)} = 1(y = 1, 2)$$

  or

  2. Z and Y are conditionally independent given X, i.e.

$$\theta_{ZY(x)} = 1(x = 1, 2)$$

  then Z cannot be a confounder.

Marginal versus Conditional Independence

Marginal versus Conditional Independence

- Because of the simplicity of looking at a single (2×2) table to study the relationship between X and Y instead of having to look in each partial table given Z, it is important to know when Z is NOT a confounder, so that we can "collapse" over Z to study the relationship between X and Y in a simplified manner.

Example

The joint probabilities in following table show a hypothetical relationship among three variables for new graduates of a university.

		Income	
Major	Gender	Low	High
Liberal Arts	Female	0.18	0.12
	Male	0.12	0.08
Science	Female	0.02	0.08
	Male	0.08	0.32
Total	Female	0.20	0.20
	Male	0.20	0.40

- Income (Y) and Gender (X) are conditionally independent given major discipline (Z) (Check!)
- However, the odds ratio for the (income,gender) marginal table equals 2.0 (not independent when we ignore major) (Check!)

Problem

Why are the odds of a high income twice as high for males as for females?

Answer

- The conditional odds X-Z and Y-Z odds ratios equal 6 (CHECK!).
- Hence, the conditional odds (given income) of majoring in science are six times higher for males than for females.
- And, the conditional odds (given gender) of having a high income are 6 times higher for those majoring in science.

Homogeneous Association

There is *homogeneous* X-Y association in a $2 \times 2 \times K$ table when

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

- Hence, any conditional odds ratio formed using two levels of X and two levels of Y is the same at each level of Z.
- A single number describes the $X - Y$ conditional associations.

Example

For

- X=smoking (yes,no),
- Y=lung cancer (yes,no),
- Z=age $(< 45, 45 - 65, > 65)$,

suppose

- $\hat{\theta}_{XY(1)} = 1.2$,
- $\hat{\theta}_{XY(2)} = 2.8$,
- $\hat{\theta}_{XY(3)} = 6.2$.

Hence, the conditional odds ratio seems to change across levels of the third variable (homogeneous association does not exist?): smoking seems to have a weak effect on lung cancer for young people, but the effect strengthens considerably with age.

Death Penalty Example

- $\hat{\theta}_{XY(1)} = 0.43$
- $\hat{\theta}_{XY(2)} = 0.00$

Notes:

- The values are not close but the second estimate is unstable because of the zero cell count.
- If we add $1/2$ to each cell count, we obtain $\hat{\theta}_{XY(2)} = 0.94$.
- Because the second estimate is so unstable and because further variation can occur from sampling variability, these values do not necessarily contradict homogeneous association.

How to check whether sample data are consistent with homogeneous association or conditional independence?

Answer follows...

Cochran-Mantel-Haenszel Methods

In this section we present

- a test for conditional independence,
- a test of homogeneous association for the K conditional odds ratios in a $2 \times 2 \times K$ tables,
- and show how to combine the sample odds ratios from the K partial tables into a single measurement of partial association.

Example: Chinese Smoking and Lung Cancer Study, with Information Relevant to Cochran-Mantel-Haenszel Test

		Lung Cancer				
City	Smoking	Yes	No	Odds Ratio	μ_{11k}	$Var(n_{11k})$
Beijing	Smokers	126	100	2.20	113.0	16.9
	Nonsmokers	35	61			
Shangai	Smokers	908	688	2.14	773.2	179.3
	Nonsmokers	497	807			
Shenyang	Smokers	913	747	2.18	799.3	149.3
	Nonsmokers	336	598			
Nanjing	Smokers	235	172	2.85	203.5	31.1
	Nonsmokers	58	121			
Harbin	Smokers	402	308	2.32	355.0	57.1
	Nonsmokers	121	215			
Zhengzhou	Smokers	182	156	1.59	169.0	28.3
	Nonsmokers	72	98			
Taiyuan	Smokers	60	99	2.37	53.0	9.0
	Nonsmokers	11	43			
Nanchang	Smokers	104	89	2.00	96.5	11.0
	Nonsmokers	21	36			

Cochran-Mantel-Haenszel Test for Conditional Independence

H_0: X and Y are conditionally independent, given Z,
or
$H_0 : \theta_{XY(k)} = 1$ for all partial tables.

Cochran-Mantel-Haenszel Test for Conditional Independence

Background

- Standard sampling models treat the cell counts as
  1. independent Poisson variates
  2. multinomial counts with fixed overall sample size
  3. multinomial counts with fixed sample size for each partial table (with counts in different partial tables bein independent)
  4. independent binomial samples within each partial table with row totals fixed
- In partial table k, the row totals are $\{n_{1+k}, n_{2+k}\}$ and the column totals are $\{n_{+1k}, n_{+2k}\}$.
- Given both these totals, all these sampling schemes yield a hypergeometric distribution for the count n_{11k} in the first row and the first column.
- That cell count determines all other counts in the partial table.

Cochran-Mantel-Haenszel Test for Conditional Independence

Construction of Test Statistics

- Under H_0, the mean and variance of n_{11k} are

$$\mu_{11k} = E(n_{11k}) = \frac{n_{1+k}n_{+1k}}{n_{++k}}$$

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

- The *Cochran-Mantel-Haenszel* (CMH) statistic is then defined by:

$$CMH = \frac{[\sum_k (n_{11k} - \mu_{11k})]^2}{\sum_k Var(n_{11k})}.$$

  - When the true odds ratio exceeds 1 in partial table k, we expect to observe $(n_{11k} - \mu_{11k}) > 0$.
  - The test statistic combines these differences across all K tables.
  - CMH takes larger values when $(n_{11k} - \mu_{11k})$ is consistently positive or consistently negative for all tables.

- CMH has a large-sample chi-squared distribution with $df = 1$.

Cochran-Mantel-Haenszel Test for Conditional Independence

Remarks

- Valid χ^2 approximation if X-Y marginal totals are large (like for a large number K of sparse partial tables)
- CMH test works best when X-Y association is similar in each partial table.
- CMH test is inappropriate when X-Y association varies dramatically among the partial tables.

Example: Chinese smoking and Lung Cancer Study

Problem:
Test the hypothesis of conditional independence which states that the true odds ratio between smoking and lung cancer equals 1 for each city.

Solution

- $\sum_k n_{11k} = 2930$
- $\sum_k \mu_{11k} = 2562.5$
- $\sum_k Var(n_{11k}) = 482.1$
- $CMH = (2930 - 2562.5)^2/482.1 = 280.1$

Hence, there is very strong evidence against conditional independence!

Remark:
A statistical analysis that combines information from several studies is called a *meta-analysis*. The meta-analysis provides stronger evidence of an association than any single partial table by itself.

SAS Program

```
libname m 'c:\tex\ddacourse';

data m.chinese;
input center smoke cancer count @@;
cards;
1 1 1 126 1 1 2 100 1 2 1 35 1 2 2 61
2 1 1 908 2 1 2 688 2 2 1 497 2 2 2 807
3 1 1 913 3 1 2 747 3 2 1 336 3 2 2 598
4 1 1 235 4 1 2 172 4 2 1 58 4 2 2 121
5 1 1 402 5 1 2 308 5 2 1 121 5 2 2 215
6 1 1 182 6 1 2 156 6 2 1 72 6 2 2 98
7 1 1 60 7 1 2 99 7 2 1 11 7 2 2 43
8 1 1 104 8 1 2 89 8 2 1 21 8 2 2 36
;

proc freq data=m.chinese;
weight count;
tables center*smoke*cancer/cmh1 chisq;
run;
```

SAS Output

```
Cochran-Mantel-Haenszel Statistics (Based on Table Scores)

Statistic    Alternative Hypothesis    DF       Value      Prob
   1         Nonzero Correlation        1     280.138     0.001
```

Estimation of Common Odds Ratio

When the association seems stable across partial tables, we can estimate an assumed common value of the K true odds ratios.
The *Mantel-Haenszel estimator* of that common value equals:

$$\hat{\theta}_{MH} = \frac{\sum_k (n_{11k} n_{22k})/n_{++k}}{\sum_k (n_{12k} n_{21k})/n_{++k}}.$$

The standard error for $\log(\theta_{MH})$ has a very complicated formula (Agresti, 2002 p. 234). We will not report it here.
Confidence intervals (and thus the standard error) can be obtained, however, from SAS PROC FREQ!

Example: Chinese smoking and lung cancer

- $\hat{\theta}_{MH} = \frac{126*61/322+\cdots+104*36/250}{35*100/322+\cdots+21*89/250} = 2.17$

SAS Output

Estimates of the Common Relative Risk (Row1/Row2)

Type of Study	Method	Value	95% Confidence Bounds	
Case-Control	Mantel-Haenszel	2.174	1.985	2.382
(Odds Ratio)	Logit	2.173	1.983	2.382
Cohort	Mantel-Haenszel	1.519	1.447	1.595
(Col1 Risk)	Logit	1.513	1.436	1.594
Cohort	Mantel-Haenszel	0.700	0.671	0.730
(Col2 Risk)	Logit	0.701	0.673	0.730

The confidence bounds for the M-H estimates are test-based.

Interpretation

The odds of lung cancer for smokers equal about twice the odds for non-smokers.

Testing Homogeneity of Odds Ratios

Suppose we want to test:

$$H_0 : \theta_{XY(1)} = \theta_{XY(2)} = \cdots = \theta_{XY(K)}.$$

Then, we can use the following test statistic (Breslow-Day):

$$\sum_{i,j,k} \frac{(n_{ijk} - \hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}},$$

where $\hat{\mu}_{ijk}$ denote estimated expected frequencies in the kth partial table that have following properties:

Testing Homogeneity of Odds Ratios

- the same marginal totals as the observed data,
- odds ratio equal to the Mantel-Haenszel estimate $\hat{\theta}_{MH}$ of a common odds ratio, by solving

$$\frac{\hat{\mu}_{11k}(n_{++k} - n_{1+k} - n_{+1k} + \hat{\mu}_{11k})}{(n_{1+k} - \hat{\mu}_{11k})(n_{+1k} - \hat{\mu}_{11k})} = \hat{\theta}_{MH}$$

- The closer the cell counts fall to the values having a common odds ratio, the smaller the statistic and the less evidence against H_0.
- Calculation of the $\hat{\mu}_{ijk}$ satisfying a common odds ratio is complex and not discussed here. Standard software (e.g. SAS PROC FREQ) report this statistic.
- The Breslow-Day statistic has a large-sample ($\hat{\mu}_{ijk} \geq 5$ in at least about 80% of the cells) chi-squared statistic with $df = K - 1$.

Chinese smoking and lung cancer

```
      Breslow-Day Test for Homogeneity of the Odds Ratios

Chi-Square =    5.200           DF =    7           Prob = 0.636
```

Chapter 4:
Generalized Linear Models Previous chapters presented methods for analyzing associations in two-way and three-way contingency tables.

These methods help us investigate effects of explanatory variables on *categorical* response variables.

Even if we are interested only in the relationship between a response and an explanatory variable, we may still have to control for at least one confounder that can influence the relationship under investigation.

We then end up studying at least three factors simultaneously.

Model Building

In this chapter we will use *models* as the basis of such analyses.

The goal is to find the best fitting and most parsimonious, yet biologically reasonable model to describe the relationship between an outcome (dependent or response variable) and a set of independent (predictor or explanatory) variables.

The most common example of modeling is the usual *linear regression model* where the outcome variable is continuous.

In a lot of cases however, the outcome variable is discrete. In that case *generalized linear models* are often used.

Model Building

A good-fitting model has several benefits.

1. inferences for model parameters help us evaluate which explanatory variables affect the response, while controlling effects of possible confounding variables.
2. estimation of parameters is more informative than mere significance testing (sizes of estimated model parameters determine the strength and importance of the effects).
3. model based predicted values can be obtained.
4. models can handle more complicated situations than those in chapters 2 and 3 (e.g. analyzing simultaneously the effects of several explanatory variables).

Generalized Linear Models

This is a broad class of models that includes ordinary regression and ANOVA models for continuous response variables as well as models for categorical response variables.

A generalized linear model (GLM) is a unifying framework for a wide range of settings:

- normal: linear models: multiple regression, ANOVA
- binary: probit and logit (logistic) regression
- counts: Poisson regression
- categorical data: log-linear modelling
- ...

Components of a GLM

All GLMs have three components:

1. the **random component** identifies the response variable Y and assumes a probability distribution for it:
   - Y=continuous response
     $\Rightarrow$ *normal* distribution
   - Y=number of "successes" out of a certain fixed number of trials
     $\Rightarrow$ *binomial* distribution
   - Y=nonnegative count (such as a cell count in a contingency table)
     $\Rightarrow$ *poisson* distribution

Components of a GLM

2. the **systematic component** specifies the explanatory variables used as predictors in the model:

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- This linear combination of explanatory variables is the *linear predictor*
- Some $\{x_j\}$ may be based on others in the model, e.g.
  - ★ $x_3 = x_1 x_2$ (to allow for interaction)
  - ★ $x_3 = x_2^2$ (to allow for quadratic effect of x_2)

Components of a GLM

3. the **link** describes the functional relationship between the systematic component and the expected value (mean) ($\mu = E(Y)$) of the random component. For a general link function $g(.)$, we have:

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Some examples of link functions

- *identity link* (ordinary regression model for continuous responses)

$$\mu = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- *log link* (loglinear model for nonnegative μ, such as count data)

$$\log(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- *logit link* (logit model when μ is between 0 and 1, such as a probability)

$$\text{logit}(\mu) = \log\left(\frac{\mu}{1-\mu}\right) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

- *probit link* (probit model when μ is between 0 and 1, such as a probability)

$$\Phi^{-1}(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

Normal GLM

Ordinary regression and ANOVA models for continuous variables are special cases of GLMs:

- Assume a normal distribution for the random component
- Model the mean directly using the identity link $g(\mu) = \mu$.

In general, a GLM generalizes ordinary regression models in two ways:

1. It allows the random component to have a distribution other than the normal.
2. It allows modelling some function of the mean.

GLM for Binary Data: Logistic/Probit Regression

Many categorical response variables have only two categories, e.g.

- a vote in an election (democrat, republican)
- choice of automobile (domestic, foreign)
- diagonsis whether woman has breast cancer (yes,no)

A binary respons variable Y is called a *Bernoulli* variable and its distribution is specified by the probability of "success",

$$\pi = P(Y = 1) = E(Y).$$

In this section we introduce GLMs for binary response data.

Relationship between Age and Coronary Heart Disease Status (CHD) of 100 subjects

Data are available on a study with 100 participants that investigates age as a possible risk factor for heart disease. The table contains following variables:

- ID: identifier variable
- AGRP: age group
- AGE: age in years
- CHD: presence (1) or absence (0) of evidence of coronary heart disease.

Relationship between Age and Coronary Heart Disease Status (CHD) of 100 subjects

Age and Coronary Heart Disease Status (CHD) of 100 Subjects.

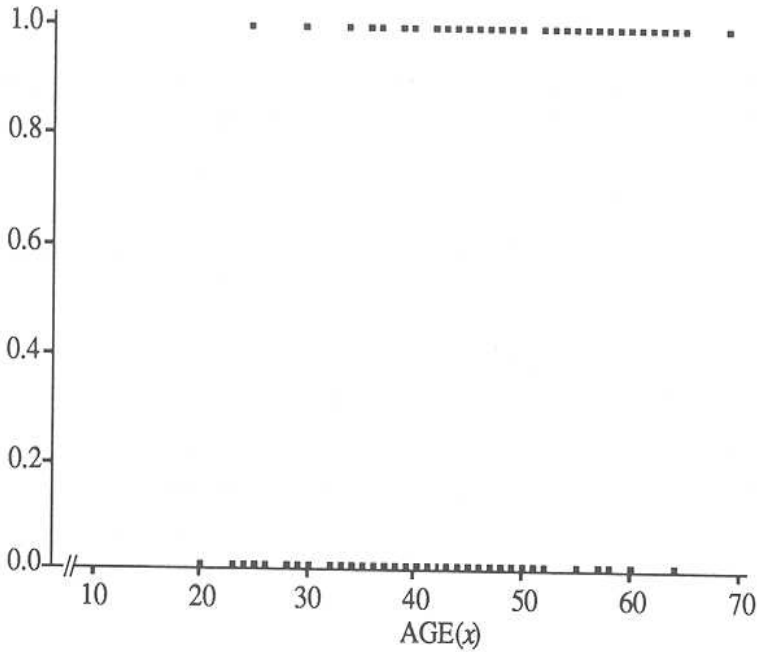
ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD
1	1	20	0	35	3	38	0	68	6	51	0
2	1	23	0	36	3	39	0	69	6	52	0
3	1	24	0	37	3	39	1	70	6	52	1
4	1	25	0	38	4	40	0	71	6	53	1
5	1	25	1	39	4	40	1	72	6	53	1
6	1	26	0	40	4	41	0	73	6	54	1
7	1	26	0	41	4	41	0	74	7	55	0
8	1	28	0	42	4	42	0	75	7	55	1
9	1	28	0	43	4	42	0	76	7	55	1
10	1	29	0	44	4	42	0	77	7	56	1
11	2	30	0	45	4	42	1	78	7	56	1
12	2	30	0	46	4	43	0	79	7	56	1
13	2	30	0	47	4	43	0	80	7	57	0
14	2	30	0	48	4	43	1	81	7	57	0
15	2	30	0	49	4	44	0	82	7	57	1
16	2	30	1	50	4	44	0	83	7	57	1
17	2	32	0	51	4	44	1	84	7	57	1
18	2	32	0	52	4	44	1	85	7	57	1
19	2	33	0	53	5	45	0	86	7	58	0
20	2	33	0	54	5	45	1	87	7	58	1
21	2	34	0	55	5	46	0	88	7	58	1
22	2	34	0	56	5	46	1	89	7	59	1
23	2	34	1	57	5	47	0	90	7	59	1
24	2	34	0	58	5	47	0	91	8	60	0
25	2	34	0	59	5	47	1	92	8	60	1
26	3	35	0	60	5	48	0	93	8	61	1
27	3	35	0	61	5	48	1	94	8	62	1
28	3	36	0	62	5	48	1	95	8	62	1
29	3	36	1	63	5	49	0	96	8	63	1
30	3	36	0	64	5	49	0	97	8	64	0
31	3	37	0	65	5	49	1	98	8	64	1
32	3	37	1	66	6	50	0	99	8	65	1
33	3	37	0	67	6	50	1	100	8	69	1
34	3	38	0								

Goal

Investigate the relationship between age and presence or absence of CHD.

Had the response been continuous rather than binary, we probably would begin by forming a scatterplot of the outcome versus the independent variable. We would use this scatterplot to provide an impression of the nature and strength of any relationship between the outcome and the independent variable.

Scatter Plot



Plot of CHD by Age.

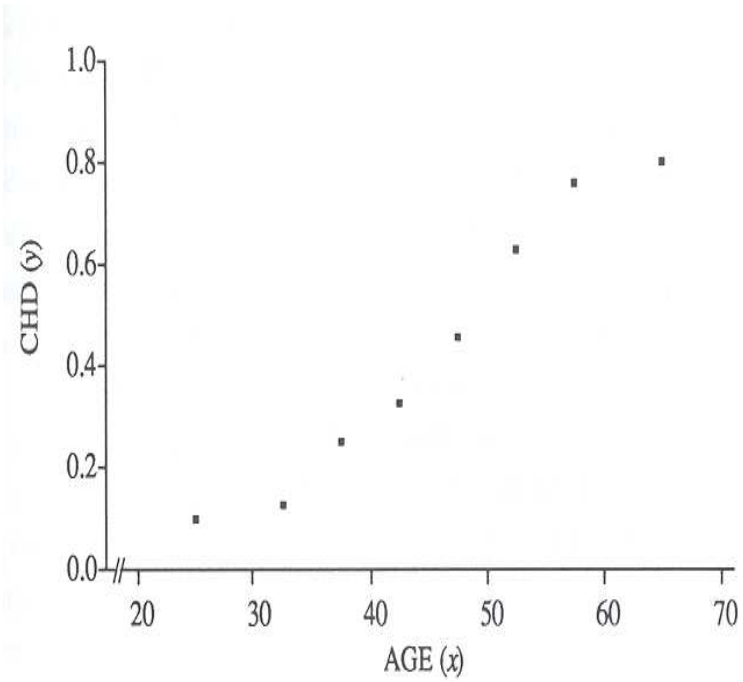
- No clear picture of the nature of the relationship between CHD and age.
- Some tendency for subjects with no CHD to be younger than subjects with CHD.

Alternative Data Representation

- Create intervals for age (AGRP).
- Calculate the mean (or proportion), π, of "successes" (subjects with CHD) in each age group.
- We assume that the number of "successes" in a certain age group follows a binomial distribution (random component of a GLM for binary responses).
- The value of π can vary as the value x of the covariate changes. We replace the π notation by $\pi(x)$ to reflect its dependence on that value.

		CHD		
AGRP	n	Absent	Present	Proportion (Present)
20-29	10	9	1	0.10
30-34	15	13	2	0.13
35-39	12	9	3	0.25
40-44	15	10	5	0.33
45-49	13	7	6	0.46
50-54	8	3	5	0.63
55-59	17	4	13	0.76
60-69	10	2	8	0.80
Total	100	57	43	0.43

Plot of the proportion of people with CHD versus the midpoint of each age interval:



Plot of the Mean of CHD in Each Age Group.

- S-shaped curve with a threshold: typical form for a risk curve!

How can we fit a model (curve) to these data?

Linear Probability Model

One approach to modelling the effect of a covariate X uses the form of ordinary regression:

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

This model is called a *linear probability model*:

- The probability of success changes linearly in x.
- The parameter β represents the change in the probability per unit change in x.
- This model is a GLM with *binomial* random component and *identity* link function.

Major structural defect of the linear probability model

- Probabilities fall between 0 and 1, whereas linear functions may take values over the entire real line.
- The linear probability model may predict $\pi(x) < 0$ and $\pi(x) > 1$ for sufficiently large or small x values.
- The model can be valid over a finite range of x values. However, most applications require a more complex model form.

Logistic Regression Model

The *logistic regression* or *logit* model takes following form:

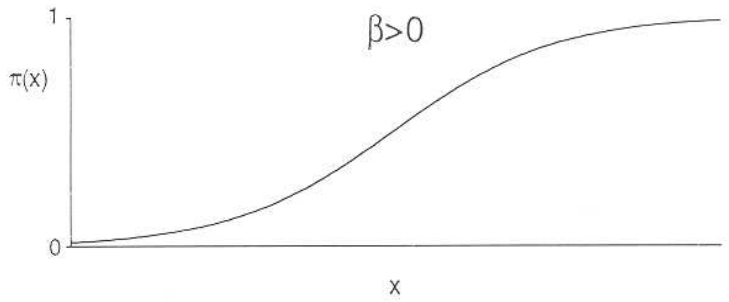
$$\log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x,$$

and thus

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

This function represents S-shaped curves (often realistic shapes for the relationship), with $\pi(x)$ increasing continuously as x increases, or $\pi(x)$ decreasing continuously as x increases:

Logistic Regression Model



Logistic regression functions.

Logit model: a special case of GLMs

- The random component for the (succes, failure) determination is binomial.
- The link function is the *logit* transformation:

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

Important Property

- π is restricted to 0 and 1,
- but $\text{logit}(\pi)$ can take any real value.
- The linear predictors (such as $\alpha + \beta x$) that form the systematic component of a GLM also take any real value.
- Hence, this model does not have the structural problem that the linear probability model has.

Parameter Interpretation

The parameter β determines the rate of increase or decrease of the curve.

- When $\beta > 0$, $\pi(x)$ increases as x increases (see previous figure (a)).
- When $\beta < 0$, $\pi(x)$ decreases as x increases (see previous figure (b)).
- When $\beta = 0$, the curve flattens to a horizontal line.

Probit Model

The probit model has expression:

$$\text{probit}(\pi(x)) = \Phi^{-1}(\pi(x)) = \alpha + \beta x,$$

where $\Phi^{-1}(.)$ is the inverse standard normal cumulative distribution. This model is an alternative to the logit model, also not showing the structural problem such as the linear probability model.

Comparison of Probit and Logit Model

- For practical purposes, probit and logistic regression curves look the same.
- One seldom encounters data for which a logistic model fits well but the probit model fits poorly, or vice versa.
- There are two important reasons that make logistic and probit models popular:
  - The range of the logit/probit function is between 0 and 1; that makes it suitable for use as a probability model, representing individual risk.
  - The logistic/probit curves have an increasing S-shape with a threshold; that makes it suitable for use as a biological model, representing risk (due to some exposure).

Comparison of Probit and Logit Model

- Parameter estimates differ for the two models, since their links have different scales.
- The probit model was introduced in 1934 for models in toxicology.
- The logistic regression model was not studied until a decade later, but is now much more popular than the probit model.
- One **advantage** of the logistic regression model over the probit model is that the logistic regression effects can also be interpreted using odds ratios (see later).

Logistic regression models will be studied in detail in Chapter 5.

GLM for Count Data: Poisson Regression

Many discrete response variables have counts as possible outcomes, e.g.

- the number of automobile thefts in a sample of cities worldwide in 1995.
- the number of viruses in a solution
- the number of defective teeth in an individual
- the number of suicides in New York in 1999

In this section we introduce GLMs for count data.

These GLMs assume a Poisson distribution for the random component. Like counts, Poisson variates can take any nonnegative integer value.

Remember, the Poisson distribution is characterized by a parameter μ:

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

where

$$\mu = E(Y)$$

Poisson Regression Model

The Poisson distribution has *positive* mean.

Therefore, the Poisson mean in GLMs is commonly modelled using a *log-link*:

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

For this model, the mean satisfies the exponential relationship:

$$\mu = \exp(\alpha + \beta x) = \exp(\alpha)\exp(\beta)^x$$

Interpretation

- The mean of Y at $x+1$ equals the mean of Y at x multiplied by $\exp \beta$.
- If $\beta = 0$, then the mean of Y does not change as X changes.
- If $\beta > 0$ then the mean of Y increases as X increases.
- If $\beta < 0$ then the mean of Y decreases as X increases.

More about Poisson regression models in Chapter 8.

Chapter 5:
Simple Logistic Regression

Logistic Regression Models

Let us now take a closer look at the statistical modeling of binary response variables, for which the response measurement for each subject is a "success" or "failure". Binary data are perhaps the most common form of categorical data, and the methods of this chapter are of fundamental importance. The most popular model for binary data is *logistic regression*. This chapter studies the application of logistic regression in greater detail.

The Logistic Regression Model (Recapitulation)

The logistic regression model has linear form for the logit of the succes probability $\pi(x)$ when X takes value x:

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

The relationship between $\pi(x)$ and the x is then described by the *logistic function*:

$$\pi(x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}. \tag{2}$$

There are two important reasons that make logistic regression popular:

1. The range of the logistic function is between 0 and 1; that makes it suitable for use as a probability model, representing individual risk.
2. The logistic curve has an increasing S-shape with a threshold; that makes it suitable for use as a biological model, representing risk due to exposure.

Interpreting the Logistic Regression Model

Classic Way

There are several ways of interpreting the model formulas (1) and (2)
The parameter β determines the rate of increase or decrease of the curve.

- When $\beta > 0$, $\pi(x)$ increases as x increases.
- When $\beta < 0$, $\pi(x)$ decreases as x increases.
- When $\beta = 0$, the curve flattens to a horizontal line.

The steepest slope of the curve (2) occurs at x for which $\pi(x) = 0.5$; that value is

$$x = -\alpha/\beta.$$

This x value is called the *median effective level* and is denoted by EL_{50}.
It represents the level at which each outcome has a 50% chance.

Interpreting the Logistic Regression Model

Classic Way

Example (revisited): Relationship between Age and Coronary Heart Disease Status (CHD) of 100 subjects

Data are available on a study with 100 participants that investigates age as a possible risk factor for heart disease. The table contains following variables:

- ID: identifier variable
- AGRP: age group
- AGE: age in years
- CHD: presence (1) or absence (0) of evidence of coronary heart disease.

Interpreting the Logistic Regression Model

Classic Way

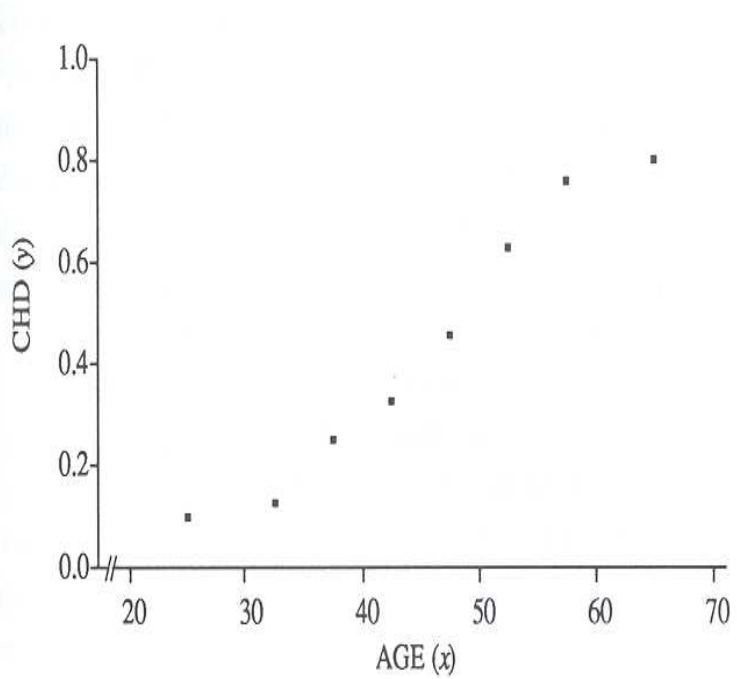
Age and Coronary Heart Disease Status (CHD) of 100 Subjects.

ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD	ID	AGRP	AGE	CHD
1	1	20	0	35	3	38	0	68	6	51	0
2	1	23	0	36	3	39	0	69	6	52	0
3	1	24	0	37	3	39	1	70	6	52	1
4	1	25	0	38	4	40	0	71	6	53	1
5	1	25	1	39	4	40	1	72	6	53	1
6	1	26	0	40	4	41	0	73	6	54	1
7	1	26	0	41	4	41	0	74	7	55	0
8	1	28	0	42	4	42	0	75	7	55	1
9	1	28	0	43	4	42	0	76	7	55	1
10	1	29	0	44	4	42	0	77	7	56	1
11	2	30	0	45	4	42	1	78	7	56	1
12	2	30	0	46	4	43	0	79	7	56	1
13	2	30	0	47	4	43	0	80	7	57	0
14	2	30	0	48	4	43	1	81	7	57	0
15	2	30	0	49	4	44	0	82	7	57	1
16	2	30	1	50	4	44	0	83	7	57	1
17	2	32	0	51	4	44	1	84	7	57	1
18	2	32	0	52	4	44	1	85	7	57	1
19	2	33	0	53	5	45	0	86	7	58	0
20	2	33	0	54	5	45	1	87	7	58	1
21	2	34	0	55	5	46	0	88	7	58	1
22	2	34	0	56	5	46	1	89	7	59	1
23	2	34	1	57	5	47	0	90	7	59	1
24	2	34	0	58	5	47	0	91	8	60	0
25	2	34	0	59	5	47	1	92	8	60	1
26	3	35	0	60	5	48	0	93	8	61	1
27	3	35	0	61	5	48	1	94	8	62	1
28	3	36	0	62	5	48	1	95	8	62	1
29	3	36	1	63	5	49	0	96	8	63	1
30	3	36	0	64	5	49	0	97	8	64	0
31	3	37	0	65	5	49	1	98	8	64	1
32	3	37	1	66	6	50	0	99	8	65	1
33	3	37	0	67	6	50	1	100	8	69	1
34	3	38	0								

Interpreting the Logistic Regression Model

Classic Way

Plot of the proportion of people with CHD versus the midpoint of each age interval:



Plot of the Mean of CHD in Each Age Group.

- S-shaped curve with a threshold: typical form for a risk curve!

Interpreting the Logistic Regression Model

Classic Way

Results of Fitting the Logistic Regression Model to the Data
Using e.g. PROC GENMOD or PROC LOGISTIC in SAS we obtain the ML parameter estimates for the logistic regression models:

Variable	Estimated Coefficient	Standard Error
Constant	-5.310	1.134
AGE	0.111	0.024

The predicted value for CHD as a function of AGE is then:

$$\hat{\pi}(x) = \frac{\exp(-5.31 + 0.111 \times AGE)}{1 + \exp(-5.31 + 0.111 \times AGE)}$$

Since $\beta > 0$, the predicted value is higher at higher values of AGE.

Interpreting the Logistic Regression Model

Classic Way

Calculate the predicted probability at the minimum AGE level in this study.
Calculate the predicted probability at the maximum AGE level in this study.
Calculate the median effective level.

Interpreting the Logistic Regression Model

Classic Way

Solution:

At the minimum AGE level in this study (i.e. 20), the predicted probability is $\frac{\exp(-5.31+0.111\times 20)}{1+\exp(-5.31+0.111\times 20)} = 4.4\%$

At the maximum AGE level, the predicted probability is 91.2%.

The median effective level is the age level at which the predicted probability equals 50%, which is $5.310/0.111 = 47.8$

Odds Ratio Interpretation

Another interpretation of the logistic regression model uses the *odds* and the *odds ratio*.

One important objective of regression analyses such as the one in (1) is "measuring" the strength of a statistical relationship between the binary dependent variable and each independent variable or covariate measured from patients; findings may lead to important decisions inpatient management (or public health interventions in other examples). In epidemiologic studies, such effects are usually measured by the *relative risk* or *odds ratio*.

When the logistic model is used, the measure is *odds ratio*!

Binary Covariate

We first consider the case of a *binary covariate*, for example:

$$\begin{aligned} X &= 0 \text{ if the patient is not exposed} \\ X &= 1 \text{ if the patient is exposed} \end{aligned}$$

Here, the term "exposed" may refer to any risk factor such as smoking, or a patient's characteristic such as race (white/nonwhite) or sex (male/female).
Clearly,

$$\log(Odds; nonexposed) = \alpha$$
$$\log(Odds; exposed) = \alpha + \beta.$$

Hence,

$$\frac{Odds; exposed}{Odds, nonexposed} = \exp \beta$$

In other words, the primary regression coefficient β represents the log odds ratio associated with the exposure, exposed versus nonexposed.

Continuous Covariate

Similarly, we have for a continuous covariate X and any value x of X,

$$\log(Odds; X = x) = \alpha + \beta(x)$$

$$\log(Odds; X = x + 1) = \alpha + \beta(x + 1).$$

Hence,

$$\frac{Odds; X = x + 1}{Odds, X = x} = \exp \beta$$

In other words, the primary regression coefficient β represents the log odds ratio associated with *one unit increase* in the value of X, $X = x + 1$ versus $X = x$.

Continuous Covariate

The primary regression coefficient β can be estimated iteratively using computer packages such as SAS. From the results we obtain a point estimate

$$\widehat{OR} = \exp(\hat{\beta})$$

and its 95% confidence interval

$$\exp\left[\hat{\beta} - 1.96SE(\hat{\beta}), \hat{\beta} + 1.96SE(\hat{\beta})\right].$$

Example (continued): Age–CHD Relationship

The odds ratio of CHD versus NO CHD for a unit increase in age equals

$$\exp(0.111) = 1.12$$

Inference for Logistic Regression

We have studied how the fit of a logistic regression model helps us describe the effects of a predictor on a binary response variable;
We next present statistical inference for the model parameters, to help judge the significance and size of the effects.
Widely available software reports the parameter estimates and their standard errors.

Confidence Intervals for Effects

A large-sample confidence interval for the parameter β in the logistic regression model $\text{logit}(\pi(x)) = \alpha + \beta x$, is

$$\hat{\beta} \stackrel{+}{-} z_{\alpha/2}(ASE)$$

with ASE the asymptotic standard error of β.
While we have not (yet) formally discussed how the estimates of the standard errors of the estimated parameters are obtained, they are routinely printed out by computer software.
Exponentiating the endpoints of this interval yields one for $\exp \beta$, the multiplicative effect on the odds of a 1–unit increase in X.

Example: Relationship AGE and CHD

To illustrate, we continue our logistic regression of the CHD data.
The estimated effect of age in the fitted equation for the probability of a CHD was:

$$\hat{\beta} = 0.111 \text{ with } ASE = 0.024.$$

A 95% confidence interval for the effect of age is

$$0.111 \pm 1.96 \times 0.024, \text{ or } (0.064, 0.158).$$

The confidence interval for $\exp \beta$, the effect on the odds per year equals

$$(\exp 0.064, \exp 0.158) = (1.066, 1.171).$$

Interpretation

We can thus conclude that a 1 year increase in age results in an increase of at least 6.6% and at the most 17% in the odds of having a CHD.

Significance Testing

How can we test for the significance of the effect of X on the binary response?

$$H_0 : \beta = 0$$

Significance Testing

- The Wald test is obtained by comparing the maximum likelihood estimate of the slope parameter β to an estimate of its standard error.
- For large samples and under H_0, the test statistic

$$Z = \frac{\hat{\beta}}{ASE}$$

has a standard normal distribution. This test statistic can be used with a 1-sided or a 2-sided alternative.

- Equivalently, for a 2-sided alternative, we can use the test statistic

$$Z^2 = \frac{\hat{\beta}^2}{ASE^2},$$

having a large-sample χ^2_1 distribution

Likelihood Ratio Statistic

- Though the Wald tests work well for large samples, the likelihood-ratio test is more powerful and reliable for sample sizes used in practice.
- The test statistic compares the maximum L_0 of the log-likelihood function when $\beta = 0$ (i.e. when $\pi(x)$ is forced to be identical at all x values) to the maximum L_1 of the log-likelihood function for unrestricted β.
- The test statistic

$$G^2 = -2(L_0 - L_1)$$

  has a large-sample χ^2_1 distribution.
- Most software for logistic regression reports the maximized log-likelihoods L_0 and L_1 and the likelihood-ratio statistic derived from those maxima.

Example: Relationship Age–CHD (Continued)

SAS Program

```
proc logistic descending;
model chd=age;
run;

/*Full Model*/
proc genmod data=m.agechd;
model chd=age/link=logit dist=bin;
run;

/*Reduced Model*/
proc genmod data=m.agechd;
model chd=/link=logit dist=bin;
run;
```

SAS Output

Proc Logistic

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
-2 LOG L	136.663	107.353	29.310 with 1 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCPT	1	-5.3095	1.1337	21.9350	0.0001	.
AGE	1	0.1109	0.0241	21.2541	0.0001	1.117

SAS Output

Full Model

Criteria For Assessing Goodness Of Fit

	Criterion	DF	Value	Value/DF
	Deviance	98	107.3531	1.0954
=>	Log Likelihood	.	-53.6765	.

Analysis Of Parameter Estimates

	Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
	INTERCEPT	1	-5.3095	1.1337	21.9350	0.0001
=>	AGE	1	0.1109	0.0241	21.2541	0.0001

SAS Output

Reduced Model

```
            Criteria For Assessing Goodness Of Fit

            Criterion              DF          Value      Value/DF

=>          Log Likelihood          .       -68.3315             .
```

Summary of Results

1. Wald statistics
   - $z = \hat{\beta}/ASE = 0.111/0.024 = 4.610$
     The 2-tailed p-value (based on a standard normal distribution) is 0.0001.
   - $z^2 = \hat{\beta}^2/ASE^2 = 0.111^2/0.024^2 = 21.252$
     The 2-tailed p-value (based on a χ^2_1 distribution) is 0.0001.

   Both statistics show a strong evidence of a positive effect of AGE on coronary heart disease.

Summary of Results

2 Likelihood ratio statistic

- the log-likelihood for the model containing only a constant term is $L_0 = -68.332$.
- the log-likelihood for the model containing the independent variable AGE, along with the constant term is $L_1 = -53.667$.
- The value for the likelihood ratio test statistic is thus

$$-2[-68.332 - (-53.667)] = 29.31.$$

The p-value (based on a χ^2_1 distribution) is 0.0001.

This provides even stronger evidence than the Wald statistic of an AGE effect.

Distribution of Probability Estimates

The estimated probability that $Y = 1$ at a fixed setting x of X equals

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

How can we construct a confidence interval for the true predicted probability?

Distribution of Probability Estimates

Outline

- Start from the linear predictor $\alpha + \beta x$
- The estimated linear predictor has large-sample ASE given by the estimated square root of

$$\text{Var}(\hat{\alpha} + \hat{\beta}x) = \text{Var}(\hat{\alpha}) + 2x\text{Cov}(\hat{\alpha}, \hat{\beta}) + x^2\text{Var}(\hat{\beta}).$$

- This can be calculated using the covariance matrix of the model parameters (reported by standard software packages).
- A 95% confidence interval for the true logit is thus

$$(\hat{\alpha} + \hat{\beta}x) \stackrel{+}{-} 1.96ASE$$

- Substituting the endpoints of this interval in the exponents of

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

gives a corresponding interval for the predicted probability.

Problem

For the CHD example, calculate the predicted probability of having a coronary heart disease at age 67, together with the 95% confidence interval for the true probability.

SAS Program

```
proc genmod data=m.agechd;
model chd=age/link=logit dist=bin covb;
run;
```

- The COVB option requests that an estimate of the parameter estimate covariance matrix be printed.

SAS Output

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-5.3095	1.1337	21.9350	0.0001
AGE	1	0.1109	0.0241	21.2541	0.0001

Estimated Covariance Matrix

Parameter Number	PRM1	PRM2
PRM1	1.28517	-0.02668
PRM2	-0.02668	0.0005789

Solution

- The logistic regression fit yields the following predicted probability:

$$\hat{\pi}(67) = \frac{\exp(-5.3095 + 0.1109 * 67)}{1 + \exp(-5.3095 + 0.1109 * 67)} = 0.893.$$

- The predicted logit is $-5.3095 + 0.1109 * 67 = 2.121$.
- SAS PROC GENMOD reports

$$\begin{aligned} \widehat{\text{Var}}(\hat{\alpha}) &= 1.285 \\ \widehat{\text{Var}}(\hat{\beta}) &= 0.000579 \\ \widehat{\text{Cov}}(\hat{\alpha}, \hat{\beta}) &= -0.0267 \end{aligned}$$

from which the estimated variance of the predicted logit equals:

$$1.285 + 2(67)(-0.0267) + 67^2(0.000579) = 0.306$$

Solution

- The 95% confidence interval for the true logit equals:

$$(1.036, 3.205)$$

- This translates to the following interval for the CHD probability at age 67:

$$(0.738, 0.961)$$

Assessing the Fit of the Model

So far, we have used logistic regression for description and inference about the effects of predictors on binary responses.
There is no guarantee, however, that a particular model of this form is appropriate or that it provides a good fit to the data.
This section discusses ways of checking the model fit.

Goodness of Fit for Models with a Fixed Number of Settings of the Explanatory Variables

Suppose we have R settings for the Explanatory Variables, indexed by j and denote by

- y_j: # successes in setting j
- n_j: # number of observations in setting j
- $\hat{\pi}_j$: estimated success probability under a given model M
- y_j/n_j: estimated success probability under a "saturated" model.

Goodness of Fit for Models with a Fixed Number of Settings of the Explanatory Variables

Outline: Compare fitted and observed counts using

1. Deviance D^2
2. Pearson's χ^2

- These statistics have approximate χ^2 distributions.
- Large χ^2 or D^2 values provide evidence of lack of fit.
- When the fit is poor, residuals (and other diagnostic measures) describe the influence of individual observations on the model fit and highlight reasons for inadequacy.

Deviance

The likelihood ratio statistic for a given model M versus a "saturated" model is often called the deviance, denoted by D^2:

$$\begin{aligned} D^2(M) &= -2[\log[L(M)] - \log[L(Sat)]] \\ &= 2\sum_{j=1}^{R}\left[y_j \log\left(\frac{y_j}{n_j\hat{\pi}_j}\right) + (n_j - y_j)\log\left(\frac{n_j - y_j}{n_j(1-\hat{\pi}_j)}\right)\right] \\ &= 2\sum_{j=1}^{R}\sum_{k=1}^{2} O_{jk}\log\left(\frac{O_{jk}}{E_{jk}}\right) \\ &\sim \chi^2_{df} \end{aligned}$$

under the null and for large n_j.
As a rough rule–of–thumb, we should have that

$$75\% \text{ of the } n_j \geq 10$$

for the chi-square approximation to hold.

Deviance

The degrees of freedom, called *residual df* for the model, equal:

$$df = \#\text{ parameters in sat. model} - \#\text{ parameters in } M$$

The deviance D^2 is often used as a measure of overall goodness-of-fit of the model, and is a test statistic for terms **left out** of the model.

Pearson's Chi-square

Alternatively, the Pearson's chi-square is used to look at the goodness of fit for a given model versus the saturated model:

$$\begin{aligned} X^2 &= \sum_{j=1}^{R} \left(\frac{[y_j - n_j\hat{\pi}_j]^2}{n_j\hat{\pi}_j} + \frac{[(n_j - y_j) - n_j(1-\hat{\pi}_j)]^2}{n_j(1-\hat{\pi}_j)} \right) \\ &= \sum_{j=1}^{R}\sum_{k=1}^{2} \frac{(O_{jk} - E_{jk})^2}{E_{jk}} \end{aligned}$$

Again, the degrees of freedom equal:

$$df = \#\text{ parameters in sat. model } - \#\text{ parameters in } M$$

Relationship with Likelihood Ratio Statistic for Nested Models

Previously we have introduced the likelihood ratio test statistic G^2 to test for parameters equalling 0.
Suppose you want to test whether model M_1 holds when the alternative is M_2, where M_1 is nested in M_2, e.g.
The likelihood ratio statistic G^2 was then introduced as:

$$\begin{aligned} G^2 &= -2\{\log[L(M_1)] - \log[(M_2)]\} \\ &= -2\{\log[L(M_1)] - \log[L(Sat)]\} + 2\{\log[L(M_2)] - \log[L(Sat)]\} \\ &= D^2(M_1) - D^2(M_2) \end{aligned}$$

Under the null and for large samples this follows a χ^2 distribution with

$$\begin{aligned} df &= (\#\ \text{pars}(Sat.) - \#\ \text{pars}(M_1)) - (\#\ \text{pars}(Sat.) - \#\ \text{pars}(M_2)) \\ &= \#\ \text{pars}(M_2) - \#\ \text{pars}(M_1) \end{aligned}$$

Relationship with Likelihood Ratio Statistic for Nested Models

- G^2 is a test statistic for whether parameters **in the model** are 0
- D^2 is a test statistic for whether parameters **not in the model** are 0

Example: Toxicology Study

Consider the following toxicology study which concerns the effects in rats of a certain toxic compound. The doses selected for the study were 0, 10, 100, 1000 mg. A number of 100 different rats is assigned to each dose group and some time after, evidence for toxicity is recorded (yes/no). The data are:

	Toxicity	
Dose	Yes	No
1	8	92
10	15	85
100	22	78
1000	26	74

Example: Toxicology Study

Question of Interest:
Does the probability (π) on a toxic event vary with dose (d)?

The logistic regression model is:

$$\text{logit}(\pi) = \beta_0 + \beta_1 d$$

SAS Program

```
data m.toxic2;
input dose y count;
cards;
1       8       100
10      15      100
100     22      100
1000    26      100
;
run;

proc genmod data=m.toxic2;
model y/count=dose/dist=bin link=logit;
run;
```

SAS Output

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	6.9618	3.4809
Scaled Deviance	2	6.9618	3.4809
Pearson Chi-Square	2	6.7383	3.3692
Scaled Pearson X2	2	6.7383	3.3692
Log Likelihood	.	-183.6252	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-1.7808	0.1692	110.8291	0.0001
DOSE	1	0.0008	0.0003	7.0915	0.0077
SCALE	0	1.0000	0.0000	.	.

Results

- Significant dose effect
- For a 990 unit increase in dose (from 10 to 1000),

$$OR(1000:10) = e^{\hat{\beta}_1(1000-10)} = 2.21$$

  the odds of some toxicity doubles.
- The observed OR for these two rows is

$$\frac{26.85}{15.74} = 1.99,$$

  so the model overestimates this odds ratio by a little.
- Also, based on X^2 and D^2, the fit is not spectacular

Other Model

Other possible models could include squared terms, cubic terms, etc.
For example, the model including squared terms is:

$$\text{logit}(\pi) = \beta_0 + \beta_1 d + \beta_2 d^2$$

SAS Program

```
data m.toxic2;
input dose y count;
dose2=dose*dose;
cards;
1       8       100
10      15      100
100     22      100
1000    26      100
;
run;

proc genmod data=m.toxic2;
model y/count=dose dose2/dist=bin link=logit;
run;
```

SAS PROC GENMOD Output

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	1.8988	1.8988
Scaled Deviance	1	1.8988	1.8988
Pearson Chi-Square	1	1.8899	1.8899
Scaled Pearson X2	1	1.8899	1.8899
Log Likelihood	.	-181.0937	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-2.1088	0.2378	78.6215	0.0001
DOSE	1	0.0095	0.0039	6.0320	0.0140
DOSE2	1	-0.0000	0.0000	5.1729	0.0229
SCALE	0	1.0000	0.0000	.	.

Summary

Model	Pars IN model $df(G^2)$	G^2	pars NOT in model $df(D^2)$	D^2 (p-value)
Null (β_0)	0	0	3	13.78 (.0032)
x	1	6.82	2	6.96 (.0308)
x, x^2	2	11.88	1	1.9 (.1682)
SATURATED	3	13.78	0	0 (.)

- Overall, the model with linear and quadratic terms appears to be the best fit.
- Note, the parameter estimate for the coefficient of x^2 is very small, but that is because x^2 is large, especially when $x = 1000$. Maybe I should have chosen $\log(x)$ as covariate.

Residuals for Logit Models

- Goodness-of-fit statistics such as D^2 and χ^2 are summary indicators of the overall quality of fit.
- To see where the model does not fit well, you can sometimes look at residuals.

Standard (unadjusted) Pearson Residuals

For a GLM with binomial random component, the Pearson residual for the fit at setting j is:

$$e_j = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{[n_j\hat{\pi}_j(1-\hat{\pi}_j)]}}$$

- If the model fits, then, asymptotically,

$$e_j \sim N(0,1)$$

  if n_j large compared to the number of parameters!
- Hence, absolute values larger than 2 indicate possible lack of fit.

Note that the Pearson statistic for testing the model fit satisfies:

$$X^2 = \sum e_j^2.$$

Adjusted Pearson Residuals

If n_j not large compared to number of parameters, then e_j is approximately normally distributed BUT with a variance that is smaller than the variance of a standard normal.
In that case an adjusted residual can be constructed

$$e_j = \frac{y_j - n_j\hat{\pi}_j}{\sqrt{[n_j\hat{\pi}_j(1-\hat{\pi}_j)(1-h_j)]}}$$

where h_j is a so-called "leverage" value.
Leverage values:

- $0 \leq h_j \leq 1$
- h_j indicates that obsevation j is distant from the center of all observations.
- The greater an observation's leverage value the greater its potential influence.
- h_j is considered "large" if it is more than twice the mean value of all leverage values ($\overline{h}$)

Deviance Residuals

The deviance residual for the fit at setting j is defined as:

$$d_j = \overset{+}{-} \sqrt{\left[y_j \log\left(\frac{y_j}{n_j \hat{p}_j}\right) + (n_j - y_j) \log\left(\frac{n_j - y_j}{n_j(1-\hat{p}_j)},\right)\right]}$$

where the sign ($+$ or $-$) is the same as $(y_j - n_j\hat{p}_j)$
When $y_j = 0$ or $y_j = n_j$, the deviance residual is defined as:

$$d_j = \begin{cases} -\sqrt{2n_j|\log(1-\hat{p}_j)|} & \text{if } y_j = 0 \\ \sqrt{2n_j|\log(\hat{p}_j)|} & \text{if } y_j = n_j \end{cases}$$

When none of the y_j equal 0 or n_j, then

$$D^2 = \sum_{j=1}^{R} d_j^2$$

Remarks

Graphical displays are also useful for showing lack of fit, e.g.

- plot of observed versus fitted proportions
- plot of observed and fitted proportions versus explanatory variables

Using SAS to obtain the residuals

Toxicology Example

```
proc logistic;
model y/count=dose dose2/influence;
run;
```

Regression Diagnostics

Case Number	Covariates DOSE	DOSE2	Pearson Residual Value	(1 unit = 0.13) -8 -4 0 2 4 6 8
1	1.0000	1.0000	-0.9351	* | |
2	10.0000	100.0	1.0047	| | *|
3	100.0	10000.0	-0.0777	| *| |
4	1000.0	1000000	0.000665	| * |

Regression Diagnostics

Case Number	Deviance Residual Value	(1 unit = 0.12) -8 -4 0 2 4 6 8	Hat Matrix Diagonal Value	(1 unit = 0.06) 0 2 4 6 8 12 16
1	-0.9767	|* | |	0.5373	| * |
2	0.9689	| | *|	0.4659	| * |
3	-0.0779	| *| |	0.9968	| *|
4	0.000665	| * |	1.0000	| *|

Results

- Pearson's chi-square for the model (x, x^2) versus the saturated model is the sum of squares of the Pearson residuals and equals

$$X^2 = 1.89 \qquad (1df) \qquad p - \text{value} = 0.1692.$$

- This is similar to the deviance for the model (x, x^2) versus the saturated model,

$$D^2 = 1.90 \qquad (1df) \qquad p - \text{value} = 0.1682.$$

- The model (x, x^2) seems to fit OK.
- No leverage points.

Diagnostic Measures of Influence

Some observations may have much influence in determining parameter estimates. The fit could be quite different if they were deleted. It may therefore be informative to report the fit of the model after deleting one or two observations, if the fit with them seems misleading
Formulas for leverages and diagnostic measures of influence are complex, so we do not reproduce them here. Most software for logistic regression produces these diagnostics.

Diagnostic Measures of Influence

Influence measures for each observation include the following:

1. For each parameter in the model, the change in the parameter estimate when the observation is deleted. This change, divided by its standard error, is called *dfbeta*.
2. A measure of the change in a joint confidence interval for the parameters produced by deleting the observation. This confidence interval displacement diagnosted is denoted by c.
3. The change in X^2 or G^2 goodness-of-fit statistics when the observation is deleted.

For each measure, the larger the value, the greater the observation's influence.

Further reading: Hosmer and Lemeshow (2000) "Applied Logistic Regression" Wiley Series in Probability and Statistics

Goodness of Fit for Models with Continuous Predictors

As stated earlier, the Deviance and Pearson's chi-square will be approximately chi-square if

$$75\% \text{ of the } n_j \geq 10.$$

However, n_j is often small. For example,

- the covariates may be continuous,
- the model may have a lot of covariates (so that very few individuals have the same pattern),

In these cases, alternative goodness of fit statistics are needed such as:

1. Use D^2 and X^2 on grouped data
2. Hosmer-Lemeshow goodness of fit statistic

Use of AGGREGATE option in SAS

Example: AGE–CHD Relationship
SAS Program

```
proc logistic data=m.agechd descending;
model chd=age/aggregate scale=none;
run;
```

SAS Output

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > Chi-Square
Deviance	41	23.7543	0.5794	0.9857
Pearson	41	21.3113	0.5198	0.9953

Number of unique profiles: 43

SAS Output

- The SCALE=none option tells LOGISTIC not to adjust the GOF statistics for overdispersion (topic will be pursued later)
- The AGGREGATE option forms the cross-classification of all levels of the explanatory variable(s).
- That way we get 43 unique profiles.
- For each profile, the observed and expected frequencies for each of the two outcomes of the dependent variable are calculated based on the fitted model.
- D^2 and X^2 can then be computed.
- Where do the 41 degrees of freedom come from?

Question

Do you think the AGGREGATE option is really useful here?

Answer

AGGREGATE is often useful but does not help when there are

- many explanatory variables (Chapter 6)
- variables measured on a continuum

In these cases there will be nearly as many profiles as original observations and nothing is accomplished by aggregating.

Alternative Way of Grouping

SAS Program

```
data m.agegroup;
input agegrp $ age absent present;
total=absent+present;
cards;
20-29   24.5    9       1
30-34   32      13      2
35-39   37      9       3
40-44   42      10      5
45-49   47      7       6
50-54   52      3       5
55-59   57      4       13
60-69   64.5    2       8
;
run;

proc logistic data=m.agegroup;
model present/total=age / scale=none aggregate;
run;
```

SAS Output

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > Chi-Square
Deviance	6	0.5242	0.0874	0.9975
Pearson	6	0.5355	0.0893	0.9974

Number of unique profiles: 8

Hosmer-Lemeshow goodness of fit statistic

Outline:

- Suppose, for example, there is **one continuous covariate** x_j and we want to determine if the model

$$\text{logit}(\pi_j) = \alpha + \beta x_j$$

  provides a good fit to the data (where there are $j = 1, \ldots, n$ observations in the dataset).
- One possible way is to fit a broader model (with interactions, squared, cubic, etc. terms) and see if those "extra" terms are significant.

Hosmer-Lemeshow goodness of fit statistic

- Hosmer and Lemeshow suggest forming G (usually 10) extra terms based on deciles of the covariate x, i.e.

  We form 10 groups of approximately equal size. The first group contains the n/10 subjects with the smallest values of x, the second group contains the $n/10$ subjects with the next smallest values of x, etc. ... the last group contains the $n/10$ subjects with the highest values of x.

  (Of course, in practice it is not possible to form groups of exactly equal size.)
- Suppose we define the $G-1$ indicators (the last one is redundant)

$$I_{jg} = \begin{cases} 1 & \text{if individual j is in group g} \\ 0 & otherwise \end{cases}$$

Hosmer-Lemeshow goodness of fit statistic

- Then, to test goodness-of-fit we consider the alternative model

$$\text{logit}(\pi_j) = \alpha + \beta x_j + \gamma_1 I_{j1} + \cdots + \gamma_9 I_{j9}$$

- If model

$$\text{logit}(\pi_j) = \alpha + \beta x_j$$

  is appropriate, then

$$\gamma_1 = \cdots = \gamma_9 = 0.$$

The Hosmer-Lemeshow statistic

- Hosmer and Lemeshow proposed a "Pearson-like" statistic (X^2_{HL}) to test for
$$H_0 : \gamma_1 = \cdots = \gamma_9 = 0,$$
- Essentially one can show (not in this course) that
$$X^2 = X^2_{HL} + \text{ pos. other stuff.}$$
- Hosmer and Lemeshow's statistic is approximately distributed as
$$X^2_{HL} \sim \chi^2_{G-2}$$
- The Hosmer-Lemeshow statistic can be obtained in SAS Proc Logistic.

Age–Chd Relationship Example

```
proc logistic data=m.agechd descending;
model chd=age/lackfit;
run;
```

Selected Output:

Hosmer and Lemeshow Goodness-of-Fit Test

		CHD = 1		CHD = 0	
Group	Total	Observed	Expected	Observed	Expected
1	10	1	0.79	9	9.21
2	10	1	1.34	9	8.66
3	10	2	1.91	8	8.09
4	11	3	2.96	8	8.04
5	11	4	4.05	7	6.95
6	10	5	4.68	5	5.32
7	10	5	5.80	5	4.20
8	13	10	9.26	3	3.74
9	15	12	12.21	3	2.79

Goodness-of-fit Statistic = 0.6978 with 7 DF (p=0.9984)

Overdispersion

When estimating logit models with grouped data, it often happens that the model doesn't fit–the deviance and Pearson chi-square are large, relative to the degrees of freedom.

Lack of fit is then sometimes described as *overdispersion*, namely if the true variance is greater than $\pi(1-\pi)$ (the normal variance for a Bernoulli distribution).

Overdispersion has several possible causes:

- incorrectly specified model
  - omitting important covariates in the model
  - the need to transform some explanatory factors
  - ...
- lack of independence in the observations

Concerns

Overdispersion is a common phenomenon in practice and causes a lot of concern because the implications are serious:

- the analysis which assumes the logistic model underestimates the standard error(s) and thus,
- wrongly inflates the test statistics, and
- the level of significance.

Measuring and Monitoring Dispersion

Suppose that data are with replications consisting of m subgroups (with identical covariate values).
Dispersion can then measured by

- the scaled deviance, i.e.

$$\chi^2_D/df$$

- the scaled Pearson chi-square, i.e.

$$\chi^2_P/df$$

When the values of these statistics are much larger than one, the assumption of binomial variability may not be valid and the data are said to exhibit overdispersion.

Fitting an Overdispersed Logistic Model

One way of correcting overdispersion is to multiply the covariance matrix of the parameters by the value of the overdispersion parameter ϕ, i.e.

- the scaled Pearson chi-square, or
- the scaled deviance.

In this correction process, the parameter estimates are not changed. However, their standard errors are adjusted (increased), affecting their significance levels (reduced).

Example

The data investigate the toxicity of a certain chemical compound.
Five groups of 20 rats each were fed for four weeks with a diet mixed with that compound at five different doses. At the end of the study, their lungs were harvested and subjected to histopathological examinations to obseve for signs of toxicity (yes=1, no=0). The results were:

Group	Dose (mg)	Number of Rats	Nr of Rats with Toxicity
1	5	20	1
2	10	20	3
3	15	20	7
4	20	20	14
5	30	20	10

Program

```
libname m 'c:\tex\ddacourse';

data m.toxic;
input group dose n toxic;
cards;
1 5 20 1
2 10 20 3
3 15 20 7
4 20 20 14
5 30 20 10
;
run;

/*Routine Fit of the logistic model*/
proc logistic descending;
model toxic/n=dose/scale=none aggregate;
run;

/*Fitting an overdispersed model, controlling for the scaled deviance*/
proc logistic descending;
model toxic/n=dose/scale=D aggregate;
run;

/*Fitting an overdispersed model, controlling for the scaled Pearson*/
proc logistic descending;
model toxic/n=dose/scale=P aggregate;
run;
```

- The option DESCENDING is needed because PROC LOGISTIC models P(Y=0) instead of P(Y=1).

SAS Output (Routine Fit)

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > Chi-Square
Deviance	3	10.9919	3.6640	0.0118
Pearson	3	10.7863	3.5954	0.0129

Number of events/trials observations: 5

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-2.3407	0.5380	18.9260	0.0001
DOSE	1	0.1017	0.0277	13.5138	0.0002

- Obvious sign of overdispersion
- Dose effect highly significant

Output (after controlling for scaled deviance)

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > Chi-Square
Deviance	3	10.9919	3.6640	0.0118
Pearson	3	10.7863	3.5954	0.0129

Number of events/trials observations: 5

NOTE: The covariance matrix has been multiplied by the heterogeneity factor 3.66396.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square
INTERCPT	1	-2.3407	1.0299	5.1654	0.0230
DOSE	1	0.1017	0.0530	3.6883	0.0548

- Point estimates remain the same
- Standard errors are larger
- Dose effect no longer significant at 5% level

SAS Output (after controlling for scaled Pearson)

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	DF	Value	Value/DF	Pr > Chi-Square
Deviance	3	10.9919	3.6640	0.0118
Pearson	3	10.7863	3.5954	0.0129

Number of events/trials observations: 5

NOTE: The covariance matrix has been multiplied by the heterogeneity factor 3.59544.

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-2.3407	1.0202	5.2639	0.0218	.	.
DOSE	1	0.1017	0.0525	3.7586	0.0525	0.484944	1.107

- Point estimates remain the same
- Standard errors are larger
- Dose effect no longer significant at 5% level

Chapter 6:
Multiple Logistic Regression

Multiple Logistic Regression

Introduction

Logistic regression, like ordinary regression extends to models incorporating multiple explanatory variables.
The predictors can be quantitative, qualitative, or of both types.
We will first show the use of dummy variables for including qualitative predictors (often called *factors*).
Afterwards, we present the general form of multiple logistic regression models.

Logit Models for Qualitative Predictors

Dummy variables in Logit Models

Suppose that a binary response variable Y has 1 nominal predictor X with R different categories.
Then we can construct $R-1$ (!) *dummy variables* that indicate the different categories.
One category is chosen as the baseline category.

Example

Consider, for example a model with $R = 3$, where

- X=Drug A
- X=Drug B
- X=Placebo

and $Y = 1$ is a succesful response.

We can then define the following 2 dummy variables:

$$\begin{aligned} X_1 &= 1 \text{ if Drug A} \\ &= 0 \text{ otherwise} \end{aligned}$$

$$\begin{aligned} X_2 &= 1 \text{ if Drug B} \\ &= 0 \text{ otherwise} \end{aligned}$$

We then fit the model:

$$\text{logit}\pi = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

Interpretation

- For an individual on placebo:

$$\text{logit}\pi = \beta_0$$

(Placebo is the baseline category)

- For an individual on Drug A:

$$\text{logit}\pi = \beta_0 + \beta_1$$

- For an individual on Drug B:

$$\text{logit}\pi = \beta_0 + \beta_2$$

Hence,

- $\exp(\beta_1)$ is the odds ratio for drug A relative to placebo
- $\exp(\beta_2)$ is the odds ratio for drug B relative to placebo
- $\exp(\beta_1 - \beta_2)$ is the odds ratio drug A relative to drug B

Extension to Several Qualitative Predictors

The previous discussion can be easily extended to a situation where Y has several nominal predictors X_i, each with R_i different categories.
For *each* predictor we then construct $R_i - 1$ dummies.
A potential "danger" with several predictors lies in the presence of interactions (i.e. one predictor modifies the effect of another).
Cross–product terms between the different predictors will help in the investigation of possible interactions.

Special Case

Suppose that a binary response Y has two nominal predictors X and Z, each with 2 levels (e.g. X=gender (male/female), Z=race (black/white)).
For each variable we can construct 1 dummy (0/1) variable.
Consider the following model for the probability π that $Y = 1$ in the $2 \times 2 \times 2$ contingency table:

$$\text{logit}(\pi) = \alpha + \beta_1 x + \beta_2 z.$$

The model specified above assumes an absence of interaction:
the effect of one factor is the same at each level of the other factor.

Model Interpretation

- At a fixed level z of Z, the effect on the logit of changing from $x = 0$ to $x = 1$ (difference of log odds) equals:

$$[\alpha + \beta_1(1) + \beta_2 z] - [\alpha + \beta_1(0) + \beta_2 z] = \beta_1$$

- Hence, $\exp \beta_1$ reflects the conditional odds ratio between X and Y at a fixed level of Z.
- In other words, controlling for Z, the odds of success at $x = 1$ equals $\exp \beta_1$ times the odds of success at $x = 0$.
- This conditional odds ratio is the same at each level z of Z.
- The model satisfies *homogeneous association* (common value of the odds ratio for the partial tables at the two levels of Z).

Model Interpretation

- Conditional independence exists between X and Y, controlling for Z, if $\beta_1 = 0$ (in which case the common odds ratio equals 1).
- In that case, the simpler model

$$\text{logit}(\pi) = \alpha + \beta_2 z$$

  applies to the three-way table.
- One can test $H_0 : \beta_1 = 0$ using a Wald or likelihood-ratio statistic comparing the two models.

AZT and AIDS Example

338 veterans whose immune systems were beginning to falter after infection with the AIDS virus were randomly assigned either to receive AZT immediately or to wait unitl their T cells showed severe immune weakness. Subjects were cross classified according to race (Z), whether they received AZT immediately (X), and whether they developed AIDS symptoms during the three-year study (Y).

		Symptoms	
Race	AZT Use	Yes	No
White	Yes	14	93
	No	32	81
Black	Yes	11	52
	No	12	43

SAS Program

```
libname m 'c:\tex\ddacourse';

data m.azt;
input race $ azt $ yes no;
cases=yes+no;
cards;
white yes 14 93
white no  32 81
black yes 11 52
black no  12 43
;
run;

proc genmod data=m.azt order=data;
class azt race;
model yes/cases=race azt/link=logit dist=bin type3;
run;
```

- The CLASS statement declares race and azt as dummy variables.
  The parameter estimate for the last level of each factor equals 0
- The TYPE3 option provides likelihood-ratio tests for testing the significance of each individual predictor in the model.

SAS Output

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	1	1.3835	1.3835
Scaled Deviance	1	1.3835	1.3835
Pearson Chi-Square	1	1.3910	1.3910
Scaled Pearson X2	1	1.3910	1.3910
Log Likelihood	.	-167.5756	.

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-1.0736	0.2629	16.6705	0.0001
RACE	white	1	0.0555	0.2886	0.0370	0.8476
RACE	black	0	0.0000	0.0000	.	.
AZT	yes	1	-0.7195	0.2790	6.6507	0.0099
AZT	no	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
RACE	1	0.0371	0.8473
AZT	1	6.8709	0.0088

Summary of Results

- The ML estimate of the effect of AZT is -0.72
- The estimated odds ratio between immediate AZT use and development of AIDS symptoms is $\exp(-0.72) = 0.49$
- Hence, the estimated odds of developing symptoms are half as high for those who took AZT immediately.
- Both the likelihood-ratio statistic and the Wald statistic show evidence of conditional association between AZT treatment and development of AIDS symptoms.
- The effect of race is not significant.

ANOVA-Type representation of Factors

An alternative representation of factors in logistic regression models resembles the way ANOVA models ordinarily express them:

$$\text{logit}(\pi) = \alpha + \beta_i^X + \beta_k^Z \tag{3}$$

- The parameters $\{\beta_i^X\}$ represent the effects of X
- The parameters $\{\beta_k^Z\}$ represent the effects of Z.

(The X and Z superscripts are simply labels, and do not represent powers.)

ANOVA-Type representation of Factors

With this notation conditional independence between X and Y, given Z, corresponds to:

$$\beta_1^X = \beta_2^X = \cdots = \beta_I^X$$

Each factor has as many parameters as it has levels, but one is redundant
There exist several ways to account for redundancies in parameters:

- setting parameter for last category equal to zero
- setting parameter for first category equal to zero
- setting sum of factor's parameters equal to zero

Different parameter coding schemes yield the same estimated probabilities.

AZT and AIDS Example

	Definition of Parameters		
Parameter	Last-zero	First=zero	Sum=zero
Intercept	-1.074	-1.738	-1.406
AZT-Yes	-0.720	0.000	-0.360
AZT-No	0.000	0.720	0.360
Race-W	0.055	0.000	0.028
Race-B	0.000	-0.055	-0.028

- For any coding scheme, the differences $\beta_1^X - \beta_2^X$ and $\beta_1^Z - \beta_2^Z$ are the same. These represent the conditional log odds ratios of X and Z with the response, given the other is the same.
- Different coding schemes yield the same estimated probabilities.

Logit Models for $2 \times 2 \times K$ Contingency Tables

$$\text{logit}(\pi) = \alpha + \beta_i^X + \beta_k^Z \tag{4}$$

An important special case of logit models with qualitative predictors occurs when

- X is a binary classification of two groups (e.g. 2 experimental treatments)
- Z is a control variable with K levels (e.g. several locations for conducting the experiment)

In this model, conditional independence exists between X and Y, controlling for Z, **if** (has to be formally tested)

$$\beta_1^X = \beta_2^X$$

Logit Models for $2 \times 2 \times K$ Contingency Tables

In that case the common $X - Y$ odds ratio,

$$\exp(\beta_1^X - \beta_2^X) = 1$$

Upon absorbing the common value of β_i^X into the α term, this yields the simpler model

$$\text{logit}(\pi) = \alpha + \beta_k^Z \tag{5}$$

How can we test for conditional independence?

- Likelihood-ratio statistic

$$-2(L_0 - L_1),$$

  comparing the full model (4) with the simpler model (5).
- Wald statistic
- (Chapter 3: Cochran-Mantel-Haenszel test),
  performs well when the association between X and Y is similar in each partial table (analogy!)

Hence, the likelihood-ratio test is an alternative to the CMH procedure for testing conditional independence in $2 \times 2 \times K$ tables!

Estimation of Common Odds Ratio for the K partial tables

- The ML estimate

$$\exp(\beta_1^X - \beta_2^X)$$

Note:
depending on the coding scheme that we use (last=zero, first=zero), this may reduce to:

  - $\exp(\beta_1)$
  - $\exp(-\beta_2)$

- (Chapter 3: Mantel-Haenszel estimate)

Testing Homogeneity of Odds Ratios

- For model (4) with binary X, the odds ratio between X and Y is assumed to be the same at each level of Z (additive model).
  - Tests for GOF provide tests of homogeneous odds ratios between X and Y at the K levels of Z
  - For example χ^2 and D^2 at $K-1$ degrees of freedom
- (Chapter 3: Breslow-Day Statistic)

AZT and AIDS Example

Test for conditional independence

likelihood–ratio	6.9 (p-value=.009, df=1)
Wald statistic	6.7 (p-value=.009, df=1)
CMH statistic	6.8 (p-value=.009, df=1)

Estimation of common odds ratio

ML estimate	0.49
Mantel–Haenszel estimate	0.49

Test for homogeneity

χ^2	1.39 (p-value=.238 ,df=1)
D^2	1.38 (p-value=.240 ,df=1)
Breslow-Day statistic	1.39 (p-value=.238, df=1)

When the sample size is large relative to the number of strata: similarity of results!

Multiple Logistic Regression

Multiple logistic regression involves a linear combination of explanatory variables:

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

- The explanatory variables can be
  - quantitative
  - qualitative
  - both types
- The parameter β_i reflects the effect of X_i on the log odds that $Y = 1$, controlling for the other Xs.

Questions of interest:

1. Which factors are most closely related to the dependent variable Y?
2. Is the effect of one factor on the response variable influenced by the presence of other factors? (interactions?)

Horseshoe Crab Example

Data come from a study of nesting horseshoe crabs (Agresti 1996, 2002). Each female horseshoe crab in the study had a male crab attaced to her in her nest. The study investigated factors that affect whether the female crab had any other males, called *satellites*, residing nearby her. Explanatory variables thought possibly to affect this included the female crab's

- color
- spine condition
- weight
- carapace width

The response outcome (Y) for each female crab is a binary indicator of whether she has satellites (yes=1, no=0).

Horseshoe Crab Example

OBS	COLOR	SPINE	WIDTH	SATELL	WEIGHT	Y
1	3	3	28.3	8	3.05	1
2	4	3	22.5	0	1.55	0
3	2	1	26.0	9	2.30	1
4	4	3	24.8	0	2.10	0
5	4	3	26.0	4	2.60	1
6	3	3	23.8	0	2.10	0
...						
170	4	3	29.0	4	3.275	1
171	2	1	28.0	0	2.625	0
172	5	3	27.0	0	2.625	0
173	3	2	24.5	0	2.000	0

Horseshoe Crab Example Using Color and Width Predictors

Question:
What is the effect of a female crab's color and width on her likeliness to have satellites?
Color has five categories:

1. light
2. medium light
3. medium
4. medium dark
5. dark

Horseshoe Crab Example Using Color and Width Predictors

Color is a surrogate for age, older crabs tending to be darker.
The sample contains no light crabs, so our models will use only the other four categories, whcih can be represented using three dummy variables:

$$\begin{aligned} c_1 &= 1 \text{ for medium light color, and } 0 \text{ otherwise,} \\ c_2 &= 1 \text{ for medium color, and } 0 \text{ otherwise,} \\ c_3 &= 1 \text{ for medium dark color, and } 0 \text{ otherwise.} \end{aligned}$$

The crab's color is dark (category 4) when

$$c_1 = c_2 = c_3 = 0$$

Model

We then consider the following model:

$$\text{logit}\pi = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x,$$

where x denotes the crab's weight.
This model assumes a lack of interaction between color and width in their effects on the response.

SAS Program

```
libname m 'c:\tex\ddacourse';

data m.crab;
input  color spine  width  satell  weight;
weight=weight/1000;
color=color-1;
if satell>0 then y=1; if satell=0 then y=0;
cards;
3  3  28.3  8  3050

4  3  22.5  0  1550

2  1  26.0  9  2300

.....

5  3  27.0  0  2625

3  2  24.5  0  2000
;
run;

/*Model using proc genmod: last category is baseline category*/
proc genmod data=m.crab2;
class color;
model y = color width / dist=bin link=logit;
run;

/*Alternative implementation using proc logistic*/
proc logistic data=m.crab2 descending;
class color/param=ref;  /*tell SAS to use the last category as reference category (not standard!)*/
model y= color width;
run;
```

SAS Output

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	168	187.4570	1.1158
Scaled Deviance	168	187.4570	1.1158
Pearson Chi-Square	168	168.6590	1.0039
Scaled Pearson X2	168	168.6590	1.0039
Log Likelihood		-93.7285	

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Odds Ratio
INTERCPT	1	-12.7151	2.7618	21.1965	0.0001	.
C1	1	1.3299	0.8525	2.4335	0.1188	3.781
C2	1	1.4023	0.5484	6.5380	0.0106	4.065
C3	1	1.1061	0.5921	3.4901	0.0617	3.023
WIDTH	1	0.4680	0.1055	19.6573	0.0001	1.597

Question

- Calculate the predicted probability for a medium-light crab of average width (26.3 cm)
- Calculate the predicted probability for a dark crab of average width (26.3 cm)

Prediction Equations

- Prediction equation for medium light crabs:

$$\text{logit}\hat{\pi} = -12.7151 + 1.3299 + 0.4680x = -11.385 + 0.468x$$

- Prediction equation for dark crabs:

$$\text{logit}\hat{\pi} = -12.7151 + 0.4680x$$

$(c_1 = c_2 = c_3 = 0)$

Predicted Probabilities

Using the prediction equations we can use predicted probabilities, e.g.

- The predicted probability for a medium-light crab of average width (26.3cm) equals .715 (check!)
- The predicted probability for a dark crab of average width (26.3cm) equals .399 (check!)

Model Interpretation

The model asumes a lack of interaction between color and width in their effects on the response:

- Width has the same effect on the response for all colors, i.e.
- a multiplicative effect of $\exp(0.468) = 1.60$ on the odds that $Y = 1$.
- This implies that two curves never cross.
- For example, at all width values, dark crabs (color 4) have a lower predicted probability than the other colors.

Model Interpretation

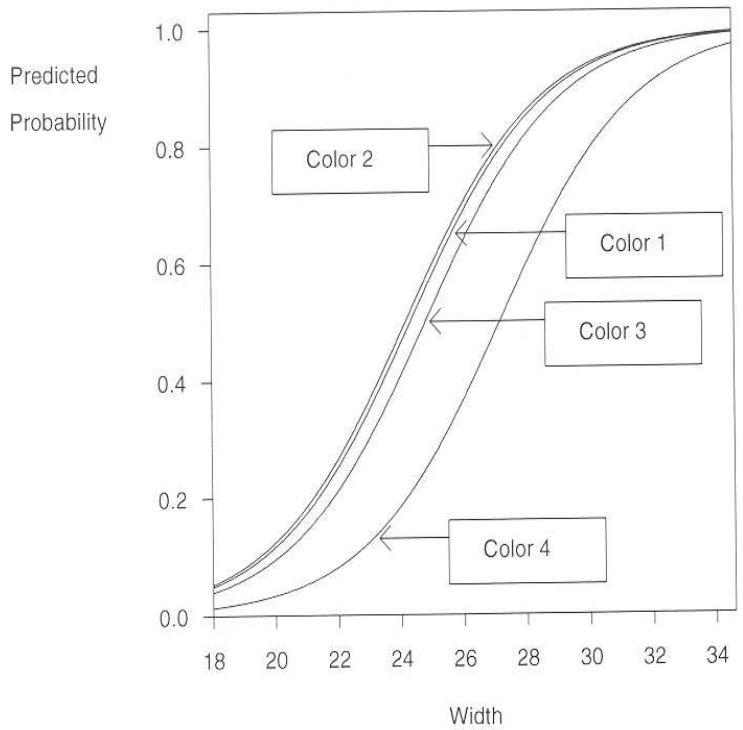


Figure Logistic regression model using width and color predictors.

Model Interpretation

- The difference in color parameter estimates between medium-light crabs and dark crabs equals 1.33.
- At any width, the estimated odds that a medium light crab has a satellite is $\exp(1.33) = 3.8$ times higher than the estimated odds for a dark crab.
- We can check this using the probabilities calculated at width 26.3:
  - the odds for a medium-light crab equals $.751/.285 = 2.51$,
  - the odds for a dark crab equals $.399/.601 = 0.66$,
  - yielding an odds ratio of 2.51/0.66=3.8!

Conclusion:

The color estimates indicate that dark (older) crabs are less likely than crabs of other colors to have satellites.

Model Comparison

One can use the likelihood-ratio method to test hypotheses about parameters in multiple logistic regression models.
For example, for the horseshoe crab data,

1. controlling for width, test whether the probability of a satellite is independent of color.
2. test whether we can safely assume a lack of interaction between color and width.

1. Controlling for width, test whether the probability of a satellite is independent of color.

$$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$$

Outline:

- Calculate the maximized log-likelihood L_1 for the full model
- Calculate the maximized log-likelihood L_0 for the reduced model
- Use test statistic $-2(L_0 - L_1)$ with a large sample χ^2 distribution with

$$df = \#\text{pars(full)} - \#\text{pars(reduced)}$$

.

SAS Program

```
proc logistic data=m.crab2 descending;
model y=c1 c2 c3 width;
run;

proc logistic data=m.crab2 descending;
model y=width;
run;
```

Results

	Value	df	p-value
L_1	-93.73		
L_0	-97.23		
$-2(L_0 - L_1)$	7.00	3	.07

No strong evidence of a color effect!

Alternative SAS Program

```
proc genmod data=m.crab2;
class color;
model y = color width / dist=bin link=logit type3;
run;
```

SAS Output

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
COLOR	3	6.9956	0.0720
WIDTH	1	24.6038	0.0001

2. Test whether we can safely assume a lack of interaction between color and width.

More generally, one can compare maximized log-likelihoods for any pair of "nested" models, e.g. with and without interaction terms.
The full model now has three additional terms, the cross products of width with the color dummy variables:

$$\text{logit}(\pi) = \alpha + \beta_1 c_1 + \beta_2 c_2 + \beta_3 c_3 + \beta_4 x + \beta_5 c_1 x + \beta_6 c_2 x + \beta_7 c_3 x$$

SAS Program

```
proc genmod data=m.crab2;
class color;
model y = color width color*width/ dist=bin link=logit type3;
run;
```

Results

	Value	df	p-value
L_1	-91.54		
L_0	-93.73		
$-2(L_0 - L_1)$	4.38	3	.22

No evidence of interaction!

Results

The reduced model has the advantage of simpler interpretations.
In fact this model fits adequately according to formal GOF tests. For example, the Hosmer-Lemeshow test with ten groups of predicted probilities yields a statistic of 3.7, based on $df = 8$ (p-value=.883).

Quantitative Treatment of Ordinal Predictor

Color can also be treated in a quantitative manner:

- It has a natural ordering of categories
- Therefore it may have a linear effect for a set of monotone scores assigned to its categories, e.g. $c = (1, 2, 3, 4)$

To test this we fit the model

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

SAS Program

```
proc genmod data=m.crab2;
model y/n = color width/ dist=bin link=logit;
run;
```

SAS Output

```
Log Likelihood          .       -94.5606             .
```

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-10.0708	2.8069	12.8733	0.0003
COLOR	1	-0.5090	0.2237	5.1791	0.0229
WIDTH	1	0.4583	0.1040	19.4129	0.0001

- Strong evidence of effects of color and width
- At a given width, for every one category increase in color darkness, the estimated odds of a satellite multiplies by $\exp^{-0.509} = 0.6$
- For instance, the estimated odds of a satellite for medium colored crabs are 60% of those for medium–light crabs.

How can we test for a linear trend in color?

outline:

- Fit a model that has a separate parameter for each color (full model)
- Fit a simpler model where color is treated in a quantitative manner
- Compare both models using a likelihood–ratio test.

Results

	Value	df	p-value
L_1	-93.73		
L_0	-94.56		
$-2(L_0 - L_1)$	1.66	2	.44

Simplification seems permissible!

Model Selection with Several Predictors

The horseshoe crab dataset has four predictors:

1. color (4 categories)
2. spine condition (3 categories)
3. weight
4. width of the carapace shell

We can fit a logistic regression model using all these predictors.
A potential drawback with such models is the problem of *multicollinearity*:

- Strong correlations among predictors make it seem that no one variable is important when all the others are in the model.
- A variable may seem to have little effect simply because it "overlaps" with other predictors in the model.

Horseshoe Crab Example

```
proc genmod data=m.crab2;
class color spine;
model y/n = color spine width weight/ dist=bin link=logit;
run;

proc genmod data=m.crab2;
class color spine;
model y/n = / dist=bin link=logit;
run;
```

Results

	Value	df	p-value
L_1	-92.60		
L_0	-112.88		
$-2(L_0 - L_1)$	40.56	7	$<.0001$

Extremely strong evidence that at least one predictor has an effect

SAS Output

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-9.2734	3.8378	5.8386	0.0157
COLOR	1	1	1.6087	0.9355	2.9567	0.0855
COLOR	2	1	1.5058	0.5667	7.0607	0.0079
COLOR	3	1	1.1198	0.5933	3.5624	0.0591
COLOR	4	0	0.0000	0.0000	.	.
SPINE	1	1	-0.4003	0.5027	0.6340	0.4259
SPINE	2	1	-0.4963	0.6292	0.6222	0.4302
SPINE	3	0	0.0000	0.0000	.	.
WIDTH		1	0.2631	0.1953	1.8152	0.1779
WEIGHT		1	0.8258	0.7038	1.3765	0.2407

- Lack of significance (compared with overall test) is a warning signal of potential multicollinearity

Example (continued)

```
proc corr data=m.crab2 noprob;
var color spine weight width;
run;
```

Pearson Correlation Coefficients / N = 173

	COLOR	SPINE	WEIGHT	WIDTH
COLOR	1.00000	0.37850	-0.25078	-0.26439
SPINE	0.37850	1.00000	-0.16648	-0.12189
WEIGHT	-0.25078	-0.16648	1.00000	0.88687
WIDTH	-0.26439	-0.12189	0.88687	1.00000

High sample correlation between width and weight:

- Width and weight serve equally well as predictors but it is redundant to use them both
- The effect of width while controlling for weight is almost zero, since weight naturally increases as width does
- In further analysis, we use width alone together with color and spine as predictors

Example (continued)

```
proc genmod data=m.crab2;
class color spine;
model y/n = color spine width/ dist=bin link=logit;
run;
```

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	-12.3908	2.8194	19.3154	0.0001
COLOR	1	1	1.6683	0.9329	3.1985	0.0737
COLOR	2	1	1.5249	0.5672	7.2285	0.0072
COLOR	3	1	1.1443	0.5933	3.7199	0.0538
COLOR	4	0	0.0000	0.0000	.	.
SPINE	1	1	-0.3770	0.5019	0.5643	0.4525
SPINE	2	1	-0.4348	0.6254	0.4834	0.4869
SPINE	3	0	0.0000	0.0000	.	.
WIDTH		1	0.4562	0.1078	17.9141	0.0001

- For color, the largest difference is between the first and fourth level.
- For spine condition, the largest difference is between the second and the third level.
- There is evidence of a strong width effect on the presence of satellites, controlling for color and spine condition.

Backward Elimination of Predictors

In many applications our major interest is to identify important predictors.
In other words, we wish to identify from many available factors a small subset that relate significantly to the outcome.
In that identification procedure we wish to avoid a large Type I error, which corresponds to including a predictor that has no real relationship to the outcome.
Such an inclusion can greatly confuse the interpretation of the regression results.
A commonly encountered strategy for selecting variables is the *backward selection procedure.*

Backward Selection Procedure

This procedure is based on a sequence of model comparisons.
The *deviance* of a model is the G^2 test of goodness–of–fit based on comparing the model to the saturated model.
The difference of deviances between two models is the likelihood–ratio statistic $-2(L_0 - L_1)$ for comparing them.

1: Fit the most complex multiple logistic regression model that you are willing to consider.

2: Eliminate the term in the model that has the largest p-value when we test that its parameter equals zero (using likelihood–ratio)

3: Repeat previous steps for the variables still in the model. If no more variables can be removed, the process is terminated.

Remark: We test only the highest–order terms for each variable!
It is inappropriate for instance to remove a main effect term if the model contains higher–order interactions involving that term.

Results of Fitting Several Logistic Regression Models to Horseshoe Crab Example

Model	Predictors	Deviance	df	AIC	Models Compared	Difference	p-value
(1)	c*s*w	170.44	149	212.4	–	–	–
(2)	c*s+c*w+s*w	173.68	155	209.7	(2)-(1)	3.2(df=6)	.36
(3a)	c*s+s*w	177.34	158	207.3	(3a)-(2)	3.7(df=3)	.30
(3b)	c*w+s*w	181.56	161	205.6	(3b)-(2)	7.9(df=6)	.25
(3c)	c*s+c*w	173.69	157	205.7	(3c)-(2)	0.0(df=2)	1.0
(4a)	s+c*w	181.64	163	201.6	(4a)-(3c)	8.0(df=6)	.24
(4b)	w+c*s	177.61	160	203.6	(4b)-(3c)	3.9(df=3)	.27
(5)	c+s+w	186.61	166	200.6	(5)-(4b)	9.0(df=6)	.17
(6a)	c+s	208.83	167	220.8	(6a)-(5)	22.2(df=1)	.00
(6b)	s+w	194.42	169	202.4	(6b)-(5)	7.8(df=3)	.05
(6c)	c+w	187.46	168	197.5	(6c)-(5)	0.8(df=2)	.67
(7a)	c	212.06	169	220.1	(7a)-(6c)	24.5(df=1)	.00
(7b)	w	194.45	171	198.5	(7b)-(6c)	7.0(df=3)	.07
(8)	(c=dark+w)	187.96	170	194.0	(8)-(6c)	0.5 (df=2)	
(9)	none	225.76	1b72	227.8	(9)-(8)	37.8(df=2)	.00

As a final step, one can use goodness-of-fit statistics to further check the fit of this model.

Remarks

- Computerized variable selection procedures should be used with caution, e.g. the backward option in SAS PROC LOGISTIC

```
proc logistic data=m.crab4 descending;
model y=color spine width cs cw sw csw/selection=backward;
run;
```

Summary of Backward Elimination Procedure

Step	Variable Removed	Number In	Wald Chi-Square	Pr > Chi-Square
1	WIDTH	6	1.2008	0.2732
2	SPINE	5	0.7967	0.3721
3	CS	4	0.5595	0.4544
4	CSW	3	0.2467	0.6194
5	SW	2	1.1284	0.2881

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	1.8197	0.5914	9.4668	0.0021	.	.
COLOR	1	-5.1225	1.0666	23.0677	0.0001	-2.264813	0.006
CW	1	0.1783	0.0414	18.5209	0.0001	1.977040	1.195

Remarks

- In a model with a large number of terms, 1 or 2 that are not really important may look impressive due to chance
- In addition, it often makes sense to include certain variables of special interest in a model and report their estimated effects even if they are not statistically significant at some level.

Akaike's Information Criterion (AIC)

Other criteria besides significance tests can help select a good model in terms of estimating quantities of interest. The best known is the Akaike Information Criterion (AIC):

$$\text{AIC} = -2(\text{maximized log likelihood - number of parameters in model}).$$

This penalizes a model for having many parameters
(cfr. Agresti (2002), page 216)

Collinearity and Confounding in Logistic Regression: Some Practical Guidelines

Confounding

- Note, confounding can be a problem in logistic regression with many covariates.
- One covariate confounds the relationship between the response and another covariate if it is related to **both** the response and covariate.
- Suppose you want to control for a possible confounding factor that is not of interest itself.
- If, in the fitted model, the confounding factor is not significant, but it changes the significance and estimated odds ratios for the covariates of interest, then you should always keep the confounding factor in the model.

Collinearity and Confounding in Logistic Regression: Some Practical Guidelines

Collinearity

- Strictly speaking, collinearity refers to correlation among the covariates.
- Thus, if there is collinearity in a dataset, there will very often also be confounding, since many of the covariates will be related to both the response and other covariates.
- Sometimes, collinearity is unavoidable if we have both x and x^2 as a covariate, which are usually highly correlated.
- In extreme cases, where two covariates are very highly correlated, we get unstable fitted equations, symptomatic of collinearity among the covariates. When a pair of covariates are collinear, estimated coefficients may even change signs when the covariates are in the model together versus in the model separately.
- When there is collinearity, one should consider which of the collinear variables is most important.

Model Building Strategies: Some Helpful Steps

- Building logistic regression models when there are many possible covariates can be a bewildering experience.
- There can be many interactions to consider, categorical versus continuous covariates, data transformations, etc.
- It is often useful to work hierarchically, looking at increasingly more complex structures of nested models, using test statistics (likelihood ratio, Wald, ...) in deciding which covariates are important or not important in predicting response.

Univariate Relationships

- Often, you first look at the relationship between the response and each covariate separately:
  1. e.g. with Pearson's chi-square (or Fisher's exact test) for a categorical covariate.
  2. if the covariate is ordered or continuous, you often look at the simple logistic regression

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

     and test for significance of β in the regression.
- These relationships are sometimes called "univariate" relationships

Multivariate Models

- After carefully looking at the univariate analyses as a screening tool, you can run a "multivariate" analysis including all covariates thought to be important from the univariate analyses.
- Hosmer and Lemeshow recommend including any covariate in the multivariate analyses which had p-value less than .25 in a univariate analysis.
- Hosmer and Lemeshow also recommend including other variables known to be important (treatment, exposure variables, possible confounding variables, etc.) that may not be significant.
- Consider the importance of each covariate in the model, and look to see whether some could be deleted or others need to be added (via test statistics, usually).

Multivariate Models

- Once you feel "close" to a final model, look more carefully for possible interactions, recording of variables that might be helpful, addition of quadratic terms, or other "transformations", etc.
- If possible, meet with an investigator on the subject matter to see if the model is biologically plausible.

No Unique Final Model

- There is more than ONE "final model".
- In complex datasets, we often will present the results of several related models.
- What is important is that you write up your model building strategy in a fair and descriptive way.
- "Statistical significance" is not the only reason to keep a covariate in the model. If a covariate is thought (and shown by others) to be a confounding variable, for the association between the exposure and disease, then it should be kept in!
- You may also be interested in p-values for covariates which are not significant, so you often leave them in the model to show they are not significant.

Stepwise Regression

- As in linear regression, there is step-up, setp-down, and stepwise regression.
- We have previously concentrated on step-down regression.
- In the **Step-up Procedure:**
  1. Fit the intercept only model, or some other relatively simple model that includes only important covariates.
  2. Fit all models that add an additional covariate to the model in 1.
  3. Choose the model with the best fit out of these. If this new model fits significantly better (say, if the p-value for the extra covariate is less than .05), keep this covariate in. If not, you might stop with the current model.
- In the **Step-down Procedure:**

A step-down procedure starts with a very complex model and then tries to delete covariates that help the least.

Stepwise Regression

- Hybrid **Stepwise Methos:**

  Often, people use a hybrid method, trying to step-up or step down in tandem at each step.
- Although it is done often, letting the computer select your final model by one of these stepwise procedures can lead to a biological implausible model just by chance.
- You are doing so many tests in the stepwise procedures that you may blow up your α-level out of the water.

Goodness-of-Fit

- Once you have come up with your "best" model, you want to consider the goodness-of-fit of the model you have selected.
- The goodness-of-fit statistics are the same as those discussed earlier;
  1. You can fit a general model (with interaction terms and quadratics), and see if the extra terms are significant.
  2. The Deviance or Pearson's chi-square can be used if the strata s are sufficient ($\geq 75\%$ of the $n_j \geq 10$).
  3. Hosmer and Lemeshow's statistic can be used if the strata sample sizes are small ($\geq 25\%$ of the $n_j < 10$).

Other Links

- Although logistic regression is by far the most popular way to model Bernoulli data, we can also use other link functions, such as for example the probit link.