STAT-HU

CATEGORICAL DATA ANALYSIS
STAT 3062

Teaching Material

Prepared by

AWOL SEID

Mobile: +2519 21 81 11 54

DEPARTMENT OF STATISTICS
HARAMAYA UNIVERSITY

CATEGORICAL DATA ANALYSIS
STAT 3062

Teaching Material

Prepared by

AWOL SEID

Mobile: +2519 21 81 11 54

DEPARTMENT OF STATISTICS
HARAMAYA UNIVERSITY

Contents

Preface

This teaching material contains lecture notes of a one semester course on "Categorical Data Analysis - Stat 3062" based on the newly harmonized undergraduate statistics curriculum. In order to comply with the goal of the course, in each chapter, it starts with a general introduction to the topics and their applications in different areas.

It is not only intended to statistics students but also other department students and researchers who need to work on applied categorical data analysis. Therefore, every reader is supposed to understand the statistical analysis and models presented in these lecture notes, and know how and when to use them.

This teaching material will never achieve a final version since it is under constant review, and subject to changes and extensions. Therefore, comments and suggestions are always welcome.

Chapter 1

Introduction

1.1 Objective and Learning Outcomes

An important consideration in determining the appropriate analysis of categorical variables is the scale of measurement and their distributions. Hence, the objective of this chapter is to review variable classifications, common discrete probability distributions and significance tests for a binomial proportion.

Upon completion of this chapter, students are expected to:

- Differentiate the different types of categorical variables and understand their corresponding probability distributions.
- Know the three major large sample inferential methods (Wald, Score and Likelihood-ratio tests).
- Determine exact p-values and exact confidence intervals for small sample inferences.

1.2 Categorical Response Data

1.2.1 Categorical versus Continuous Variables

A *categorical* variable is a variable that can take on one of a limited, and usually fixed, number of possible values. That is, it has a measurement scale consisting of a set of categories. Such scales occur frequently in the health sciences (e.g., whether a patient survives an operation: yes, no), social sciences (for measuring attitudes and opinions), behavioral sciences (example, diagnosis of type of mental illness: schizophrenia, depression, neurosis), public health (example, whether awareness of AIDS has led to increased use of condoms: yes, no), zoology (example, alligators' primary food choice: fish, invertebrate, reptile), education (example, examination result: pass, fail) and marketing (example, consumers' preference among brands of a product: Brand A, Brand B, Brand C). They even are pervasive in highly quantitative fields such as engineering sciences and industrial quality control, when items are classified according to whether or not they conform to certain standards.

There are two common kinds of categorical variables: *nominal* and *ordinal*. The first kind, *nominal* variables, have a set of mutually exclusive categories which cannot be ordered. The number of occurrences in each category is referred to as the frequency (count) for that category. When nominal variables have two categories, they are termed as *binary* (dichotomous). For example, gender (male or female) and patient outcomes (dead or alive) are binary variables. A nominal variable which has multiple categories, is referred to a *multinomial*(*polytomous*) variable. For example, blood type (A, B, AB or O), teaching method (lecturing, using slides, discussion or other), favorite Ethiopian music (tizita, ambasel, anchihoye or bati), marital status (single, married, widowed, divorced), preference of soft drink (coca, fanta, sprite, pepsi, mirinda or 7up) and party affiliation (Republican, Democrat, Independent) are all multinomial variables.

The second kind of variables, *ordinal* variables, are where the categories are ordered. For example, clinical stage of a disease (none, mild or severe) and academic qualifications (BSc, MSc or PhD) are ordinal variables. Note that quantitative variables grouped into a small number of categories (example, Age < 18, $18-24$, $25-34$ and ≥ 35 years) are ordinal too. Ordinal variables generally indicate that some subjects are better than others but then, we can not say by how much better, because the intervals between categories are not equal.

In addition to nominal or ordinal variables, categorical data also consists of variables with a finite number of *discrete values* (really, a small number of discrete values). That is, categorical data may arise in a form of simple *counts*, for example, number of children in a family, CD4 counts in an HIV/AIDS patient, $\cdots$.

It must be noted that the distinction between continuous and discrete variables is the number of values they can take. Therefore, since continuous variables can take lots of values, they cannot be considered as categorical.

The reason for distinguishing between variables is that the method of data analysis depends on the scale of measurement and their distribution. Methods designed for ordinal variables cannot be used with nominal variables. Though ordinal variables are qualitative, they are treated in a quantitative manner in a statistical analysis by assigning ordered scores to the categories. Thus, methods designed for ordinal variables utilize the order of the category (low to high or high to low) unlike methods designed for nominal variables. On the contrary, methods designed for nominal variables can be used with ordinal variables as nominal variables are lower in the measurement scale. Since the methods designed for nominal variables do not use the order of the categories, it can result serious loss of power (Agresti, 2007, 2002). Hence, it is a must to apply appropriate methods for the actual scale.

1.2.2 Response versus Explanatory Variables

Based on the role of variables in a statistical analysis, variables can be classified as *dependent* and *independent* variables.

- A *dependent* variable is a variable, that is, of primary interest to be determined as an outcome. For example, the outcome of a certain treatment or the educational achievement level can be considered dependent variables. The terms *outcome*, *response* and *dependent* are used interchangeably.

- An *independent* variable is a variable to be used to determine the value of the dependent variable. It is also called a *factor*, an *exposure*, a *predictor* or a *covariate.*

  There are two types of independent variables: *attribute* (*measured*) and *active* (*manipulated*) variables.

  - An *attribute* independent variable is a variable whose values are *preexisting characteristics* of objects under study. The values of such a variable cannot be systematically changed or manipulated. For example, education, sex, socio-economic status, $\cdots$.
  - An *active* independent variable can be experimentally manipulated. Such an independent variable is a necessary (but not sufficient) condition to make *cause-and-effect* conclusions. For example, a researcher might investigate a new kind of therapy compared to the traditional treatment (the treatment group each person is assigned to). A second example could be a design to evaluate the effect of different fertilizers on crop yields. A third example might be to study the effect of a new teaching method, such as cooperative learning, on student performance. Studies with active independent variables are experimental studies.

  Even though a statistical analysis does not differentiate whether an independent variable is an attribute or active, there is a crucial difference in interpretation. For scientific researches in applied disciplines, the need to demonstrate that a given intervention or treatment causes change in behaviour or performance is extremely important. Only the approaches that have an active independent variable can allow one to infer that the change (difference) in the independent variable caused the change (difference) in the dependent variable. In contrast, a significant difference between or among persons with different values of an attribute independent variable should not lead one to conclude that the attribute independent variable caused the dependent variable to change.

Based on the type and role of variables, the common statistical methods are listed in the following table.

Dependent Variable	Independent Variable	Method
Continuous	Binary	t test
Continuous	Multinomial	ANOVA
Continuous	Continuous	Correlation
Continuous	Quantitative/Categorical/Both	Linear Regression
Categorical	Categorical	χ^2 test
Binary	Quantitative/Categorical/Both	Binary Logistic Regression
Multinomial	Quantitative/Categorical/Both	Multinomial Logistic Regression
Ordinal	Quantitative/Categorical/Both	Ordinal Logistic Regression
Discrete	Quantitative/Categorical/Both	Poisson Regression
Time-to-event	Quantitative/Categorical/Both	Survival Models

Note: For correlation and χ^2 test, there is no need to differentiate variables as dependent and independent.

The subject of this course is the analysis of categorical response variables. It is mainly concerned with those statistical methods which are relevant when there is just one categorical response variable. There can be several explanatory variables which may be either quantitative, categorical or both.

1.3 Probability Distributions for Categorical Data

Inferential statistical analysis requires assumptions about the probability distribution of the response variable. For regression models and analysis of variance, the continuous response variable is assumed to follow normal distribution. For a categorical response, there are three common distributions; binomial, multinomial and poisson.

1.3.1 The Binomial Distribution

A binomial distribution is one of the most frequently used discrete distribution which is very useful in many practical situations involving only two types of outcomes.

Recall that a Bernoulli trial is a trial with only two mutually exclusive and exhaustive outcomes (outcomes that can be reduced to two) which are labeled as ”success” and ”failure”. Let Y denote the number of successes out of n Bernoulli trials.

	Outcome		
	Success	Failure	Total
Frequency	y	$n-y$	n
Probability	π	$1-\pi$	1

Under the assumption of independent and identical trials, Y has the binomial distribution with the number of trials n and probability of success π, $Y \sim Bin(n, \pi)$. Therefore, the probability of y successes out of the n trials is:

$$P(Y = y) = \binom{n}{y} \pi^y (1-\pi)^{n-y}, \; y = 0, 1, 2, \cdots, n$$

The mean μ and variance σ^2 of the number of successes are $E(Y) = \mu = n\pi$ and $V(Y) = \sigma^2 = n\pi(1-\pi)$, respectively.

The binomial distribution is always symmetric when $\pi = 0.50$. For fixed n, it becomes more skewed as π moves toward 0 or 1. Specifically, the distribution is right-skewed when $\pi < 0.5$ and it is left-skewed when $\pi > 0.5$.

For fixed π, it becomes more symmetric as n increases. When n is large, it can be approximated by a normal distribution with $\mu = n\pi$ and $\sigma^2 = n\pi(1-\pi)$. A guideline is that the expected number of both outcomes, $n\pi$ and $n(1-\pi)$, should both be at least 5. For $\pi = 0.50$, it requires only $n \geq 10$. For $\pi = 0.10$ (or $\pi = 0.90$), it requires $n \geq 50$. When π gets nearer to 0 or 1, larger samples are needed to attain normality.

1.3.2 The Multinomial Distribution

The multinomial distribution is an extension of binomial distribution. In this case, each trial has more than two mutually exclusive and exhaustive outcomes. Similar to Bernoulli trials, the trials are independent with the same category probabilities.

Let J denote the number of outcomes in a multinomial experiment and let $Y_i;\ i = 1, 2, \cdots, J$ denote the number of times that the i^{th} outcome occurs among n trials. Let $\pi_i;\ i = 1, 2, \cdots, J$ be the probability that the i^{th} outcome occurs on any trial, where $\pi_1 + \pi_2 + \cdots + \pi_J = 1$.

	Outcome Categories						
	1	2	$\cdots$	j	$\cdots$	J	Total
Frequency	n_1	n_2	$\cdots$	n_j	$\cdots$	n_J	n
Probability	π_1	π_2	$\cdots$	π_j	$\cdots$	π_J	1

Thus, $(Y_1, Y_2, \cdots, Y_J)$ has a multinomial distribution with parameters $n; \pi_1, \pi_2, \cdots, \pi_J$ and write as $(Y_1, Y_2, \cdots, Y_J) \sim Multi(n; \pi_1, \pi_2, \cdots, \pi_J)$. Therefore, the probability of observing n_1 outcome 1's, n_2 outcome 2's, $\cdots$, n_J outcome J's among the n multinomial trials is:

$$P(Y_1 = n_1, Y_2 = n_2, \cdots, Y_J = n_J) = \frac{n!}{n_1! n_2! \cdots n_J!} \pi_1^{n_1} \pi_2^{n_2} \cdots \pi_J^{n_J} = \frac{n!}{\prod_{i=1}^{J} n_i!} \prod_{i=1}^{J} \pi_i^{n_i}$$

where $n_1 + n_2 + \cdots + n_J = n$. For outcome j, $Y_j \sim Bin(n_j, \pi_j)$ with mean $E(Y_j) = \mu_j = n\pi_j$ and variance $V(Y_j) = \sigma_j^2 = n\pi_j(1-\pi_j)$. Also, if $J = 2$, the multinomial distribution reduces to binomial distribution, $(Y_1, Y_2) \sim Multi(n; \pi_1, \pi_2)$.

1.3.3 The Poisson Distribution

Poisson distribution is another theoretical discrete probability distribution, which is useful for modeling the number of successes in a certain time, space, $\cdots$. It differs from binomial distribution in the sense that it is not possible to count the number of failures even though the number of successes is known. For example, in the case of patients coming to hospital for emergency treatment, only the number of patients arriving in a given hour is known but it is not possible to count the number of patients not coming for emergency treatment in that hour.

Accordingly, it is not possible to determine the number of trials (total number of outcomes - successes and failures) and hence binomial distribution cannot be applied as a decision making tool. In such situation the poisson distribution should be used given the average number of successes.

Let Y be the number of successes in a specific time or space. Its probabilities depend on a single parameter, μ which is the average number of successes in a certain time or space. Thus, $Y \sim Poisson(\mu)$. The probability of y successes in that specific time or space is:

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

A key feature of the Poisson distribution is that its variance equals its mean, i.e., $E(Y) = \mu = \text{Var}(Y)$. The counts vary more when their mean is higher. Also the distribution approaches normality as μ increases and it approximates binomial if n is large and π is small, with $\mu = n\pi$.

1.4 Statistical Inference for a Proportion π

Recall a binary variable is a variable having only two categories, for example: patient outcome (cured or dead), development of cancer (yes or no). One of the categories is labeled as success and the other as failure. Mostly, the success outcome is coded by 1 and the failure is coded by 0.

The probability of a success is denoted by π and the probability of a failure is denoted by $1 - \pi$. Then the probability distribution for the number of successes y in n independent and identical trials, is:

$$P(Y = y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}; \ y = 0, 1, 2, \cdots, n.$$

Recall the mean and variance of the number of successes y are $n\pi$ and $n\pi(1 - \pi)$, respectively. If both the expected number of outcomes are at least 5, then a normal distribution with mean $n\pi$ and variance $n\pi(1 - \pi)$ can be used as an approximation for the binomial.

If $Y \sim \text{Bin}(n, \pi)$, then $Y \sim \mathcal{N}(n\pi, n\pi(1-\pi))$. The approximation becomes more precise for large n.

In a random sample of n from a population, if there are y successes, then the sample proportion of successes is $p = \frac{y}{n}$ (alternatively, it can be denoted by $\hat{\pi}$). The point estimator of the binomial parameter π is the sample proportion of successes p (p estimates π). The mean of the sampling distribution of a sample proportion p is $E(p) = \mu_p = \pi$. Also, the variance of the sample proportion of successes is $V(p) = \sigma_p^2 = \pi(1-\pi)/n$. Hence, for large sample size, the sampling distribution of a sample proportion is normal with mean π and variance $\pi(1-\pi)/n$. That is, $p \sim \mathcal{N}[\pi, \pi(1-\pi)/n]$.

Therefore, the standard error of the sample proportion p is $\text{SE}(p) = \sigma_p = \sqrt{\pi(1-\pi)/n}$. Consequently, the estimated standard error of the sample proportion p is $\widehat{\text{SE}}(p) = \hat{\sigma}_p = \sqrt{p(1-p)/n}$.

1.4.1 Maximum Likelihood Estimation

A *likelihood function* is the probability of the observed data, expressed as a function of the parameter. For a binomial distribution, with $y = 0$ successes in $n = 5$ trials, the likelihood function is $\ell(\pi) = (1-\pi)^5$ which is defined for π between 0 and 1. If $\pi = 0.60$ for instance, the probability that $y = 0$ is $\ell(0.60) = (1-0.60)^5 = 0.0102$. Likewise, if $\pi = 0.40$ then $\ell(0.40) = (1-0.40)^5 = 0.0778$, if $\pi = 0.20$ then $\ell(0.20) = (1-0.20)^5 = 0.3277$ and if $\pi = 0.0$ then $\ell(0.0) = (1-0.0)^5 = 1.0$.

The *maximum likelihood estimate* of a parameter is a value at which the likelihood function is maximized. Consider the previous example, the likelihood function $\ell(\pi) = (1-\pi)^5$ is maximized at $\pi = 0.0$. Thus, when $n = 5$ trials have $y = 0$ successes, the maximum likelihood estimate of π equals 0.0. This means that the result $y = 0$ in $n = 5$ trials is more likely to occur when $\pi = 0.00$ than when π equals any other value.

In general, for the binomial outcome of y successes in n trials, the maximum likelihood estimate of π is $\hat{\pi} = p = y/n$. This is the sample proportion of successes for n trials. For observing $y = 3$ successes in $n = 5$ trials, the maximum likelihood estimate of π equals $p = 3/5 = 0.60$. The result $y = 3$ in $n = 5$ trials is more likely to occur when $\pi = 0.60$ than when π equals any other value.

The expected value of the sample proportion p is $E(p) = \pi$ and its variance is $\sigma^2(p) = \pi(1-\pi)/n$.

- Since $E(p) = \pi$, p is an unbiased estimator of π. But unbiasedness is not true for all ML estimators.
- As the number of trials n increases, $\sigma^2(p)$ decreases toward zero; that is, the sample

proportion tends to be closer to the population proportion π. Thus, the estimator p is consistent. Consistency is true for all ML estimators.

- For large n, the sampling distribution of p is approximately normal, that is, $p \sim \mathcal{N}[\pi, \pi(1-\pi)/n]$. This large sample inferential method is also true for all ML estimators.

1.4.2 Wald, Score and Likelihood-Ratio Tests

The interest here is whether the population proportion of success π takes a particular value, say π_0.

The Wald Test

The Wald test uses the sample proportion p for estimating the standard error of the sample proportion p. That is, the estimated standard error is $\widehat{\mathrm{SE}}(p) = \hat{\sigma}_p = \sqrt{p(1-p)/n}$.

Step 1: State both the null and alternative hypotheses. There three options are:

Option 1: $H_0 : \pi = \pi_0$ vs $H_1 : \pi \neq \pi_0$

Option 2: $H_0 : \pi = \pi_0$ vs $H_1 : \pi < \pi_0$

Option 3: $H_0 : \pi = \pi_0$ vs $H_1 : \pi > \pi_0$

Step 2: Specify the level of significance α and obtain the critical value. The critical value is $z_{\alpha/2}$ for a two sided test and z_α for a one sided test.

Step 3: The Wald test statistic defined as:

$$Z = \frac{p - \pi}{\sqrt{p(1-p)/n}} \sim \mathcal{N}(0, 1).$$

Step 4: Decision: H_0 can be rejected if $|z_{cal}| > z_{crt}$ or p-value$< \alpha$.

Step 5: Conclusion.

Example 1.1. Of 1464 HIV/AIDS patients under HAART treatment in Jimma University Specialized Hospital from 2007-2011, 331 defaulted. Did the proportion of defaulter patients different from one fourth?

Solution: Let π denote the proportion of defaulter patients. The sample proportion of defaulters is $p = \frac{331}{1464} = 0.226$. For a sample of size $n = 1464$, the estimated standard error of p is $\widehat{\mathrm{SE}}(P) = \sqrt{0.226(1-0.226)/1464} = 0.011$.

Step 1: Hypothesis:

$H_0 : \pi = 0.25$ The proportion of defaulter patients is not significantly different from 25%.

$H_1 : \pi \neq 0.25$ The proportion of defaulter patients is significantly different from 25%.

Step 2: Assuming $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$

Step 3: The calculated value of the Wald test statistic is:

$$z = \frac{p - \pi}{\sqrt{p(1-p)/n}} = \frac{0.226 - 0.25}{\sqrt{0.226(1 - 0.226)/1464}} = -2.18$$

Step 4: Decision: Since $|z| = 2.18 > 1.96$, H_0 can be rejected. Or it is easy to find the two-sided p-value which is the probability that the absolute value of a standard normal variate exceeds 2.18, that is, $p-\text{value} = 2P(Z > 2.18) = 2(0.0146) = 0.0292$.

Step 5: Conclusion: Since, the one-sided p-value is 0.0146, there is a strong evidence that, $\pi < 0.25$, that is, the proportion of defaulter patients is fewer than a quarter at 5% level of significance.

The Score Test

The *Score* test is an alternative possible test which uses a known standard error. This known standard error is obtained by substituting the assumed value under the null hypothesis π_0. That is, $\hat{\sigma}_P = \sqrt{\pi_0(1-\pi_0)/n}$. Hence, the Score test statistic for a binomial proportion is:

$$Z = \frac{P - \pi}{\sqrt{\pi_0(1-\pi_0)/n}} \sim \mathcal{N}(0, 1).$$

Example 1.2. Recall example 1.1. Test the hypothesis using the Score test.

Solution: Let π denote the proportion of defaulter patients. The sample proportion of defaulters is $p = \frac{331}{1464} = 0.226$. For Score test, the known standard error of P is $\widehat{\text{SE}}(P) = \sqrt{0.25(1 - 0.25)/1464} = 0.0113$.

Step 1: Hypothesis:

$H_0 : \pi = 0.25$ The proportion of defaulter patients is not significantly different from 25%.

$H_1 : \pi \neq 0.25$ The proportion of defaulter patients is significantly different from 25%.

Step 2: Assuming $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$

Step 3: The calculated value of the Score test statistic is:

$$z = \frac{p - \pi}{\sqrt{\pi_0(1-\pi_0)/n}} = \frac{0.226 - 0.25}{\sqrt{0.25(1 - 0.25)/1464}} = -2.12$$

Step 4: Decision: Since $|z| = 2.12 > 1.96$, H_0 should be rejected. Also, the two-sided p-value is $2P(Z > 2.12) = 2(0.0170) = 0.034$ which leads to the rejection of H_0.

Step 5: Conclusion: There is a strong evidence that, $\pi < 0.25$, that is, the proportion of defaulter patients is fewer than a quarter at 5% level of significance..

The Likelihood-Ratio Test

The *likelihood-ratio* test is based on the ratio of two maximizations of the likelihood function. The first is the maximized value of the likelihood function over the possible parameter value(s) that the parameter assumes under the null hypothesis. The second is the maximized value of the likelihood function among all possible parameter values, permitting the null or the alternative hypothesis to be true.

Let ℓ_0 denote the maximized value of the likelihood function under the null hypothesis, and let ℓ_1 denote the maximized value in general. Note that ℓ_1 is always at least as large as ℓ_0.

For a binomial proportion, $\ell_0 = \ell(\pi_0)$ and $\ell_1 = \ell(p)$. Thus, the *likelihood-ratio* test statistic is

$$G^2 = -2\log(\ell_0/\ell_1) = -2(\log \ell_0 - \log \ell_1) \sim \chi^2(1).$$

Note that $G^2 \geq 0$. If ℓ_0 and ℓ_1 are approximately equal, then G^2 will approach to 0. This indicates that there is no sufficient evidence to reject H_0 (in favor of H_0). If ℓ_0 is by far less than ℓ_1, then G^2 will be very large indicating a strong evidence against H_0.

Likelihood-ratio CI: The $(1-\alpha)100\%$ likelihood-ratio confidence interval is obtained by solving $-2\log(\ell_0/\ell_1) \leq \chi^2_\alpha(1)$ for π_0.

Example 1.3. Recall example 1.6. Test $H_0 : \pi = 0.50$ using likelihood-ratio and construct its confidence interval.

Solution: Since $n = 16$ and $y = 0$, the Binomial likelihood function is $\ell = \ell(\pi) = (1-\pi)^{16}$. Under $H_0 : \pi = 0.50$, the binomial probability of the observed result of $y = 0$ successes is $\ell_0 = \ell(0.5) = 0.5^{16}$. The likelihood-ratio test compares this to the value of the likelihood function at the ML estimate of $p = 0$, which is, $\ell_1 = \ell(0) = 1$. Thus, the likelihood-ratio test statistic is $G^2 = -2\log(0.50^{16}) = -32\log(0.50) = 22.18$. Since $G^2 = 22.18 > \chi^2_{0.05}(1) = 3.84$, H_0 should be rejected.

1.4.3 Interval Estimation

Wald CI: The $(1-\alpha)100\%$ (Wald) confidence interval for the population proportion π is given by:

$$\left[p \pm z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} \right].$$

This is a large sample confidence interval for the population proportion π which uses the sample proportion p as the mid-point of the interval.

Example 1.4. Recall example 1.1. Construct the 95% CI for the population proportion of HIV/AIDS patients who were defaulted.

Solution: For $n = 1464$ observations, $p = 0.226$. And $z_{\alpha/2} = z_{0.025} = 1.96$. The 95% confidence interval is $[0.226 \pm 1.96\frac{0.226(1-0.226)}{1464}] = (0.204, 0.248)$. Therefore, the proportion of HIV/AIDS patients who were defaulted is between 0.204 and 0.248 at 0.05 level of significance.

Note: The Wald confidence interval for π is based on a normal approximation to the binomial distribution. The rule is that both $n\pi$ and $n\pi(1-\pi)$ should be at least 5. Unless π is close to 0.50, it does not work well if n is not very large. That is, it works poorly to use the sample proportion as the mid-point of the confidence interval when π is near 0 or 1.

Score CI: The Score confidence interval uses a duality with significance tests. It is constructed by inverting results of a significance test using the null standard error. This confidence interval consists of all values π_0's for the null hypothesis parameter that are 'not rejected' at a given significance level.

For a binomial proportion, given n and p with a critical value $\pm z_{\alpha/2}$, the π_0 solutions for the equation

$$\frac{|p - \pi_0|}{\sqrt{\pi_0(1-\pi_0)/n}} = \pm z_{\alpha/2}$$

are the end points of the Score confidence interval for π. Squaring both sides gives an equation which is quadratic in π_0. This method does not require estimation of π in the standard error, since the standard error in the test statistic uses the null value π_0.

Example 1.5. A clinical trial is conducted to evaluate a new treatment. This experiment has nine successes in the first 10 trials. Construct the 95% Score and Wald CIs.

Solution: The sample proportion of successes $p = 0.90$ based on $n = 10$ trials. The solutions for $n(p-\pi_0)^2 = \pi_0(1-\pi_0)z^2_{\alpha/2}$ are 0.596 and 0.982. Thus, the 95% Score CI is (0.596, 0.982).

By contrast, using the estimated standard error gives confidence interval (0.714, 1.086) in which the upper limit is greater than 1. That is why, it is said Wald CI works poorly when the parameter may fall near the boundary values of 0 or 1.

Example 1.6. Of $n = 16$ students, $y = 0$ answered "yes" for the question "Did you ever smoke cigarette?". Construct the 95% Wald and Score confidence intervals for the population proportion of smoker students.

Solution: Let π be the population proportion of smoker students. Since $y = 0$, $p = \frac{0}{16} = 0$. The 95% Wald CI is given by $(p \pm z_{\alpha/2}\sqrt{p(1-p)/n}) = (0 \pm 1.96\sqrt{0(1-0)/16}) = (0,\ 0)$. As said before when the number of successes is near 0 or near n, Wald methods do not provide sensible results.

The 95% Score confidence interval is obtained by solving $|0 - \pi_0| = \pm 1.96\sqrt{\pi_0(1-\pi_0)/16}$ for π_0. By contrast this provides the interval (0, 0.316) which is sensible than the Wald interval (0, 0).

LR CI: The likelihood-ratio confidence interval is $-2\log(\ell_0/\ell_1) \leq \chi^2_\alpha(1)$. Here, $\ell_0 = \ell(\pi_0) = (1-\pi_0)^{16}$ and $\ell_1 = \ell(0) = 1$. Thus, $-2\log(1-\pi_0)^{16} \leq 3.84$ which implies $\pi_0 \leq 0.113$. Therefore, the 95% likelihood-ratio confidence interval is (0.0, 0.113) which is narrower than the Score CI.

Example 1.7. Recall example 1.5: a clinical trial that has nine successes in the first 10 trials. Test the hypothesis of $H_0 : \pi = 0.5$ using the three methods and construct the corresponding confidence intervals.

Solution: The Wald test is

$$z = \frac{0.90 - 0.50}{\sqrt{0.90(1-0.90)/10}} = 4.22.$$

The corresponding chi-squared statistic is $z^2 = (4.22)^2 = 17.8$ ($df = 1$). Since $z = 4.22 > z_{0.025} = 1.96$ or $z^2 = 17.8 > \chi^2_{0.05}(1) = 3.84$, there is sufficient evidence to reject H_0.

The score test is

$$z = \frac{0.90 - 0.50}{\sqrt{0.5(1-0.5)/10}} = 2.53.$$

Again using the Score test, since $z = 2.53 > z_{0.025} = 1.96$, H_0 should be rejected.

For the likelihood-ratio test, the maximum value of the likelihood function is obtained as $\ell_1 = \binom{10}{9}(0.90)^9(0.10)^1 = 0.3874$. Also, when $H_0 : \pi = 0.50$ is true, the likelihood value is $\ell_0 = \binom{10}{9}(0.50)^9(0.50)^1 = 0.0098$. Thus, the value of the likelihood-ratio test statistic is

$$G^2 = -2\log(\ell_0/\ell_1) = -2\log(0.0098/0.3874) = -2\log(0.0253) = 7.3539.$$

From the chi-squared distribution with $df = 1$ at 5% level of significance, $\chi^2_{0.05} = 3.84$, this statistic has a larger value which results the rejection of H_0.

1.4.4 Small Sample Binomial Inference

When the sample size is small to moderate, the Wald test is the least reliable of the three tests. In other cases, for large samples they have similar behavior when H_0 is true.

For ordinary regression models assuming a normal distribution, the three tests provide identical results. A marked divergence in the values of the three statistics indicates that the distribution of the maximum likelihood estimator may be far from normality. In that case, small sample methods are more appropriate than large sample methods.

Exact p-values

For small samples, it is safer to use the binomial distribution directly (rather than a normal approximation) to calculate the p-values. For $H_0 : \pi = \pi_0$, the p-value is based on the binomial distribution with parameters n and π_0, $Bin(n, \pi_0)$.

For $H_1 : \pi > \pi_0$, the exact one-sided p-value is $P(Y \geq y) = \sum_{x=y}^{n} \binom{n}{x} \pi_0^x (1-\pi_0)^{n-x}$. Similarly, for $H_1 : \pi < \pi_0$, the exact one-sided p-value is $P(Y \leq y) = \sum_{x=0}^{y} \binom{n}{x} \pi_0^x (1-\pi_0)^{n-x}$.

It is easy to calculate a two-sided p-value for a symmetric distribution centered at 0, such as $Z \sim \mathcal{N}(0, 1)$, which is $P(|Z| > z) = 2 \times P(Z \geq |z|)$. In general, if the distribution is symmetric but not necessary centered at 0, then the exact two-sided p-value is $2 \times min[P(Y \geq y), P(Y \leq y)]$

Example 1.8. Recall again example 1.5. Find the exact one sided and two-sided p-values.

Solution: The historical norm for the clinical trial is 50%. So we want to test if the response rate of the new treatment is greater than 50%. For $H_1 : \pi > 0.50$, p-value=$P(Y \geq 9) = P(Y = 9) + P(Y = 10) = 0.0107$. For $H_1 : \pi \neq 0.50$, p-value=$2 \times P(Y \geq 9) = 2 \times [P(Y = 9) + P(Y = 10)] = 2 \times 0.0107 = 0.0214$. In both cases, H_0 should be rejected at 5% level of significance. That is, the treatment is significantly effective.

Exact Confidence Interval

A $(1 - \alpha)100\%$ confidence interval for π is of the form $P(\pi_L \leq \pi \leq \pi_U) = 1 - \alpha$ where π_L and π_U are the lower and upper end points of the interval. Given the level of significance α, observed number of successes y and number of trials n, the endpoints π_L and π_U, respectively, satisfy

$$P(Y \geq y|\pi = \pi_L) = \sum_{x=y}^{n} \binom{n}{x} \pi_L^x (1 - \pi_L)^{n-x} = \alpha/2$$

and

$$P(Y \leq y|\pi = \pi_U) = \sum_{x=y}^{n} \binom{n}{x} \pi_U^x (1 - \pi_U)^{n-x} = \alpha/2$$

except that the lower bound $\pi_L = 0$ when $y = 0$ and the upper bound $\pi_U = 1$ when $y = n$. It can figure out π_L and π_U by plugging different values for π_L and π_U until values that approximate $\alpha/2$ are obtained. In fact, this can be easily implemented using a computer, so there is no need to do it by hand.

Example 1.9. If 4 successes are observed in 5 trials, find the 95% exact confidence interval.

Solution: The lower bound π_L of the exact confidence interval (π_L, π_U) is the value of π_L for which $P(Y \geq 4|\pi = \pi_L) = \sum_{y=4}^{5} \binom{5}{y}\pi_L^y(1-\pi_L)^{5-y}$ approximates 0.025. Similarly, the upper bound π_U of the exact confidence interval (π_L, π_U) is the value of π_U for which $P(Y \leq 4|\pi = \pi_U) = \sum_{y=0}^{4} \binom{5}{y}\pi_U^y(1-\pi_U)^{5-y}$ approximates 0.025. Using trial and error, the values of π_L and π_U can be determined as shown in the following table.

Lower Bound		Upper Bound	
π_L	$P(Y \geq 4\|\pi = \pi_L)$	π_U	$P(Y \leq 4\|\pi = \pi_U)$
0.250	0.0156	0.800	0.6723
0.260	0.0181	0.900	0.4095
0.270	0.0208	0.950	0.2262
0.280	0.0238	0.990	0.0490
0.285	$0.02547 \approx 0.025$	0.995	$0.02475 \approx 0.025$

Thus, the 95% exact confidence interval for π is (0.285,0.995).

1.5 Comparing Two Population Proportions

For comparisons of two population proportions, independent random samples are assumed to be drawn from two binomial populations with parameters π_1 and π_2. If y_1 is the number of successes to be observed for a random sample of size n_1 from population (group) 1 and y_2 is the number of successes to be observed for a random sample of size n_2 from population (group) 2, then the point estimators of π_1 and π_2 are the sample proportions $p_1 = \frac{y_1}{n_1}$ and $p_2 = \frac{y_2}{n_2}$, respectively.

The interest is whether the two population proportions are equal $\pi_1 = \pi_2$, that is, whether the difference between the two population proportions (absolute risk) is zero $\pi_1 - \pi_2 = 0$. The point estimator of the difference of the population proportions $\pi_1 - \pi_2$ is $p_1 - p_2$. The mean of the sampling distribution of the difference of the sample proportions $p_1 - p_2$ is $E(p_1 - p_2) = \mu_{p_1-p_2} = \pi_1 - \pi_2$. The variance of the sampling distribution of the difference of the population proportions $p_1 - p_2$ is also given as $V(p_1 - p_2) = \sigma^2_{p_1-p_2} = \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}$. Thus, $p_1 - p_2 \sim \mathcal{N}\left[\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right]$.

The standard error is $\text{SE}(p_1 - p_2) = \sigma_{p_1-p_2} = \sqrt{\frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}}$. The estimated standard error is $\widehat{\text{SE}}(p_1 - p_2) = \hat{\sigma}_{p_1-p_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$.

1.5.1 Testing for Difference of Two Population Proportions

Step 1: State both the null and alternative hypotheses. There three possible options are:

Option 1: $H_0 : \pi_1 - \pi_2 = 0$ vs $H_1 : \pi_1 - \pi_2 \neq 0$

Option 2: $H_0 : \pi_1 - \pi_2 = 0$ vs $H_1 : \pi_1 - \pi_2 < 0$

Option 3: $H_0 : \pi_1 - \pi_2 = 0$ vs $H_1 : \pi_1 - \pi_2 > 0$

Step 2: Specify the level of significance α and obtain the critical value. The critical value for a two sided test is $z_{\alpha/2}$ whereas the critical value for a one sided test is z_α.

Step 3: Use the z test statistic and obtain its calculated value:

$$Z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_2)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} \sim \mathcal{N}(0, 1).$$

Step 4: Decision: If $|z_{cal}| > z_{tab}$ ($p-$ value $< \alpha$), H_0 can be rejected.

Step 5: Conclusion.

Example 1.10. A study looked at the effects of OC use on heart disease in women 40-44 years of age. The researchers found that among 50 current OC users at baseline, 13 women developed a myocardial infarction (MI) over a 3 year period, whereas among 100 non-OC users, 7 developed an MI over a 3-year period. Assess the statistical significance of the results.

Solution: Let π_1 be the proportion of MI among OC users and π_2 be the proportion of MI among non-OC users. The sample proportion of MI among OC users is $p_1 = \frac{13}{50} = 0.26$ and the sample proportion of MI among non-OC users is $p_2 = \frac{7}{100} = 0.07$.

Step 1: Hypothesis:

$H_0 : \pi_1 - \pi_2 = 0$. The proportions of MI among OC users and non-OC users are not significantly different. That is, OC has not a significant effect.

$H_1 : \pi_1 - \pi_2 \neq 0$. The proportions of MI among OC users and non-OC users are significantly different. That is, OC has a significant effect.

Step 2: Assuming $\alpha = 0.05$, $z_{0.025} = 1.96$.

Step 3: The calculated value of the z test statistic is:

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{\frac{p_1(1-p_2)}{n_1} + \frac{p_2(1-p_2)}{n_2}}} = \frac{(0.26 - 0.07) - 0}{\sqrt{\frac{0.26(1-0.26)}{50} + \frac{0.07(1-0.07)}{100}}} = \frac{0.19}{0.067} = 2.836$$

Step 4: Decision: Since $z_{cal} = 2.836 > z_{0.025} = 1.96$, H_0 can be rejected. Or $p-$value $= 2 \times P[Z > 2.836] = 2 \times 0.0023 = 0.0046 < \alpha = 0.05$.

Step 5: Conclusion. The proportions of MI among OC users and non-OC users are significantly different at 5% level of significance. That is, OC use has a significant positive effect to develop MI at 5% level of significance.

1.5.2 Interval Estimation for $\pi_1 - \pi_2$

The $(1-\alpha)100\%$ confidence interval for the difference of the two population proportions $\pi_1 - \pi_2$ are given by:

$$\left\{ (p_1 - p_2) \pm z_{\alpha/2}\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}} \right\}.$$

Example 1.11. Consider again example 1.10 and construct the 95% confidence interval for the difference in the proportions of MI between OC and non-OC users.

Solution: The 95% confidence interval for the difference in the proportions of MI between OC and non-OC users $\pi_1 - \pi_2$ is:

$$\left\{ (0.26 - 0.07) \pm 1.96\sqrt{\frac{0.26(1-0.26)}{50} + \frac{0.07(1-0.07)}{100}} \right\} = (0.059,\ 0.321).$$

Since the confidence interval is greater than 0, OC use has a significant positive effect to develop MI at 5% level of significance.

Chapter 2

Contingency Tables

2.1 Objective and Learning Outcomes

For a single categorical variable, the data can summarized by counting the number of observations (frequency) in each category. The sample proportions in the categories estimate the category probabilities. For two or more categorical variables, the data is summarized in a tabular form in which the cells of the table contain number of observations (frequencies) in the intersection categories of the variables. Such a table is called contingency table. The objective of this chapter is to discuss statistical methods to be used for contingency table analysis.

Upon completion of this chapter, students are expected to:

- Determine probability structures (joint, marginal and conditional distributions) for contingency tables.
- Use the Pearson's chi-square and likelihood-ratio tests to examine independence of factors.
- Define the difference of proportions, relative risk and odds ratio, and use them to test independence of factors in a multinomial sample.
- Differentiate marginal and conditional associations, marginal and conditional independence in three-way contingency tables.
- Test homogeneity of proportions and check goodness-of-fit of a set of data to a specific probability distribution.

2.2 Two-Way Contingency Table

2.3 Contingency Table Method

Let X and Y denote two categorical variables with I and J categories (levels), respectively. Then, classifications of subjects on both variables have IJ possible combinations and the contingency table is called a *two-way* table or an $I \times J$ (read as I-by-J) table.

Suppose N subjects are classified on both X and Y as shown in Table 2.1. Then N_{ij} represents the number of subjects belonging to the i^{th} category of X and j^{th} category of Y.

Table 2.1: Layout of an $I \times J$ Contingency Table

	Y						
X	1	2	$\cdots$	j	$\cdots$	J	Total
1	N_{11}	N_{12}	$\cdots$	N_{1j}	$\cdots$	N_{1J}	N_{1+}
2	N_{21}	N_{22}	$\cdots$	N_{2j}	$\cdots$	N_{2J}	N_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	N_{i1}	N_{i2}	$\cdots$	N_{ij}	$\cdots$	N_{iJ}	N_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	N_{I1}	N_{I2}	$\cdots$	N_{Ij}	$\cdots$	N_{IJ}	N_{I+}
Total	N_{+1}	N_{+2}	$\cdots$	N_{+j}	$\cdots$	N_{+J}	N

Here, N_{i+} and N_{+j} are the marginal totals representing the number of subjects belonging to the i^{th} category of X and the j^{th} category of Y, respectively. Note that $N_{i+} = \sum_{j=1}^{J} N_{ij}$ and $N_{+j} = \sum_{i=1}^{I} N_{ij}$. Also, the population size $N = \sum_{i=1}^{I} N_{i+} = \sum_{j=1}^{J} N_{+j} = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}$.

2.3.1 Probability Structures for Contingency Tables

The joint probability distribution of the responses (X, Y) of a subject chosen randomly from some population can be determined from the contingency table. This joint distribution determines the relationship between the two categorical variables. Also, from this distribution, the marginal and conditional distributions can be determined.

Joint and Marginal Probabilities

The (true) probability of a subject being in the i^{th} category of X and j^{th} category of Y is

$$P(X = i, Y = j) = \pi_{ij} = \frac{N_{ij}}{N}.$$

The probability distribution $\{\pi_{ij}\}$ is the joint distribution of X and Y shown in Table 2.2. The marginal distribution of each variable is the sum of the joint probabilities over all the categories of the other variable. That is,

$$P(X = i) = \pi_{i+} = \sum_{j=1}^{J} \pi_{ij} = \frac{N_{i+}}{N} \text{ and } P(Y = j) = \pi_{+j} = \sum_{i=1}^{I} \pi_{ij} = \frac{N_{+j}}{N}.$$

Table 2.2: Joint and Marginal Distributions X and Y

	Y						
X	1	2	$\cdots$	j	$\cdots$	J	Total
1	π_{11}	π_{12}	$\cdots$	π_{1j}	$\cdots$	π_{1J}	π_{1+}
2	π_{21}	π_{22}	$\cdots$	π_{2j}	$\cdots$	π_{2J}	π_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	π_{i1}	π_{i2}	$\cdots$	π_{ij}	$\cdots$	π_{iJ}	π_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	π_{I1}	π_{I2}	$\cdots$	π_{Ij}	$\cdots$	π_{IJ}	π_{I+}
Total	π_{+1}	π_{+2}	$\cdots$	π_{+j}	$\cdots$	π_{+J}	1

Thus, $\{\pi_{i+}\}$ is the marginal distribution of X and $\{\pi_{+j}\}$ is the marginal distribution of Y. The marginal distributions provide single-variable information. Note also that $\sum\limits_{i=1}^{I} \pi_{i+} = \sum\limits_{j=1}^{J} \pi_{+j} = \sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J} \pi_{ij} = 1$.

Conditional Probabilities

The joint distribution of X and Y is more useful if both variables are responses. But if one of the variable is explanatory (fixed), the notion of the joint distribution is no longer useful.

If X is fixed, for each category of X, Y has a probability distribution. Hence, it is important to study how the distribution of Y changes as the category of X changes.

Given that a subject is belong to the i^{th} category of X, then

$$P(Y = j|X = i) = \pi_{j|i} = \frac{\pi_{ij}}{\pi_{i+}}$$

denotes the conditional probability of that subject belonging to the j^{th} category of Y. In other words, $\pi_{j|i}$ is the conditional probability of a subject being in the j^{th} category of Y if it is in the i^{th} category of X. Thus, $\{\pi_{j|i};\ j = 1, 2, \cdots, J\}$ is the conditional distribution of Y at the i^{th} category of X. Note also that $\displaystyle\sum_{j=1}^{J} \pi_{j|i} = 1$.

Table 2.3: Conditional Distributions of Y Given X

	Y						
X	1	2	$\cdots$	j	$\cdots$	J	Total
1	$\pi_{1\|1}$	$\pi_{2\|1}$	$\cdots$	$\pi_{j\|1}$	$\cdots$	$\pi_{J\|1}$	1
2	$\pi_{1\|2}$	$\pi_{2\|2}$	$\cdots$	$\pi_{j\|2}$	$\cdots$	$\pi_{J\|2}$	1
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	$\pi_{1\|i}$	$\pi_{2\|i}$	$\cdots$	$\pi_{j\|i}$	$\cdots$	$\pi_{J\|i}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	$\pi_{1\|I}$	$\pi_{2\|I}$	$\cdots$	$\pi_{j\|I}$	$\cdots$	$\pi_{J\|I}$	1

The probabilities $\{\pi_{1|i}, \pi_{2|i}, \cdots, \pi_{j|i}, \cdots, \pi_{J|i}\}$ form the conditional distribution of Y at the i^{th} category of X. A principal aim in many studies is to compare the conditional distribution of Y at various level of X.

Example 2.1. In the HAART Data used by Seid *et al.* (2014), there were 1464 HIV/AIDS patients. Of these 22.6% were defaulters. 63.5% of these patients were females including 189 defaulters.

1. Construct the contingency table.
2. Find the joint and marginal distributions.
3. If a patient is selected at random, what is the probability that the patient is
   (a) a female and defaulter?
   (b) a male?
   (c) defaulter if the patient is female?

Solution:

1. The contingency table is

	Defaulter		
Gender	Yes (1)	No (2)	Total
Female (1)	$N_{11} = 189$	$N_{12} = 741$	$N_{1+} = 930$
Male (2)	$N_{21} = 142$	$N_{22} = 392$	$N_{2+} = 534$
Total	$N_{+1} = 331$	$N_{+2} = 1133$	$N = 1464$

2. The joint and marginal distributions are

	Defaulter		
Gender	Yes (1)	No (2)	Total
Female (1)	$\pi_{11} = 0.129$	$\pi_{12} = 0.506$	$\pi_{1+} = 0.635$
Male (2)	$\pi_{21} = 0.097$	$\pi_{22} = 0.268$	$\pi_{2+} = 0.365$
Total	$\pi_{+1} = 0.226$	$\pi_{+2} = 0.774$	1.000

3. If a patient is selected at random,

   (a) $P(\text{Gender} = 1, \text{Defaulter} = 1) = \frac{N_{11}}{N} = \frac{189}{1464} = 0.1291.$

   (b) $P(\text{Gender} = 2) = \frac{N_{2+}}{N} = \frac{534}{1464} = 0.3648.$

   (c) $P(\text{Defaulter} = 1|\text{Gender} = 1) = \frac{N_{11}}{N_{1+}} = \frac{189}{930} = 0.2032.$

2.3.2 Statistical Independence

Statistical independence is a condition of no relationship between two variables in a population. In probability terms, two categorical variables are defined to be independent if all joint probabilities are the product of their marginal probabilities. That is, if X and Y are independent then $\pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j.

Also, when X and Y are independent, each conditional distribution of Y is identical to the marginal distribution of Y. That is, $\pi_{j|i} = \pi_{+j}$ for all i. Thus, two categorical variables are independent when $\pi_{j|1} = \pi_{j|2} = \cdots = \pi_{j|I}$ for $j = 1, 2, \cdots, J$; that is, the probability of any category of Y is the same in each category of X which is often referred as *homogeneity* of conditional distributions. This is a more better definition of independence than $\pi_{ij} = \pi_{i+}\pi_{+j}$ when one of the variables is explanatory.

Example 2.2. Recall example 2.2. Are the sex of the patient and defaulting statistically independent? The answer is No. Why?

2.3.3 Binomial, Multinomial and Poisson Sampling

The probability distributions introduced in Section 1.3 on page 4 can be extended to cell counts in a contingency table.

Table 2.2 and 2.3 display population notations for joint (and marginal) and conditional distributions for an $I \times J$ table, respectively. For sample data, the notation n_{ij} instead of N_{ij} and p_{ij} instead of π_{ij} are used.

Multinomial Sampling Models

If a sample of n subjects are classified based on two categorical variables (one with I and the other with J categories), there will be IJ possible outcomes (cells). Let Y_{ij} denote the number of outcomes in the i^{th} category of X and j^{th} category of Y, and let π_{ij} be its corresponding probability. Then the probability mass function of the cell counts has the multinomial form

$$P(Y_{11} = n_{11}, Y_{12} = n_{12}, \cdots, Y_{IJ} = n_{IJ}) = \frac{n!}{\prod\limits_{i=1}^{I}\prod\limits_{j=1}^{J} n_{ij}!}\prod_{i=1}^{I}\prod_{j=1}^{J}\pi_{ij}^{n_{ij}}$$

such that $\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J} n_{ij} = n$ and $\sum\limits_{i=1}^{I}\sum\limits_{j=1}^{J} \pi_{ij} = 1$.

Example 2.3. To study the relationship between smoking cigarette (Yes, No) and occurrence of lung cancer (Yes, No), the data can be summarized in a 2 × 2 table format as follows.

	Lung Cancer		
Smoking	Yes	No	Total
Yes	$n_{11} =$	$n_{12} =$	$n_{1+} =$
No	$n_{21} =$	$n_{22} =$	$n_{2+} =$
Total	$n_{+1} =$	$n_{+2} =$	$n =$

If a random sample $n = 300$ individuals is taken and classified according to these two variables (smoking and lung cancer), then the total sample size n is treated as fixed. Hence, the four cells are treated as a multinomial random variables with $n = 300$ trials and unknown joint probabilities $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$. For example, if $\{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\} = \{0.10, 0.20, 0.40, 0.30\}$, then

$$P(n_{11}, n_{12}, n_{21}, n_{22}) = \frac{200!}{n_{11}!n_{12}!n_{21}!n_{22}!}0.10^{n_{11}}0.20^{n_{12}}0.40^{n_{21}}0.30^{n_{22}}.$$

Independent Multinomial (Binomial) Sampling Models

If one of the two variables is explanatory, the observations on the response variable occur separately at each category of the explanatory variable. In such case, the marginal totals of the explanatory variable are treated as fixed. Thus, for the i^{th} category of the explanatory variable, the cell counts $\{Y_{ij}; j = 1, 2, \cdots, J\}$ has a multinomial form with probabilities $\{\pi_{j|i}; j = 1, 2, \cdots, J\}$. That is,

$$P(Y_{i1} = n_{i1}, Y_{i2} = n_{i2}, \cdots, Y_{iJ} = n_{iJ}) = \frac{n_{i+}!}{\prod\limits_{j=1}^{J} n_{ij}!}\prod_{j=1}^{J}\pi_{j|i}^{n_{ij}}$$

provided that $\sum_{j=1}^{J} n_{ij} = n_{i+}$ and $\sum_{j=1}^{J} \pi_{j|i} = 1$. If $J = 2$, it will reduced to binomial distribution.

When samples at different categories of the explanatory variable are independent, the joint probability mass function for the entire cells of the contingency table is the product of the multinomial functions at various categories. That is,

$$P(Y_{11} = n_{11}, Y_{12} = n_{12}, \cdots, Y_{IJ} = n_{IJ}) = \prod_{i=1}^{I} \frac{n_{i+}!}{\prod\limits_{j=1}^{J} n_{ij}!} \prod_{j=1}^{J} \pi_{j|i}^{n_{ij}}.$$

This sampling scheme is called independent (product) multinomial sampling. Again here if $J = 2$, it will be an independent (product) binomial sampling.

Example 2.4. Recall example 2.3. Suppose, instead, random samples of 100 smokers and 200 nonsmokers are taken, and follow up both groups for some years. Finally, each group is classified based on a clinical examination whether they developed lung cancer or not. {It is like a *prospective* design or a *cohort* study '*looking in the future*'. In this case, the marginal totals for smoking status are fixed at $n_{1+} = 100$ and $n_{2+} = 200$ (i.e., the marginal distribution of smoking status is fixed by the sampling design). Such studies provide proportions for the conditional distribution of developing lung cancer, given smoking status.} Thus, for each smoking status, the recoded results will be independent binomial samples.

In another way, if random samples of 100 individuals who have lung cancer and 200 individuals who do not have lung cancer are selected, and classified each sample based on the smoking history of the individuals. Now, the marginal totals for lung cancer are fixed at 100 and 200. {It is a *retrospective* design or a *case-control* study '*looking in the past*'. In this case, the marginal totals for lung cancer status are fixed at $n_{+1} = 100$ and $n_{+2} = 200$. Using this retrospective sample, the probability of lung cancer at each category of smoking habit can not be estimated.} Hence, for each lung cancer outcome, the recoded results are independent binomial samples.

Poisson Sampling Models

A poisson sampling model treats the cell counts as independent poisson random variables with parameters $\{\mu_{ij}\}$. Thus, the joint probability mass function for all outcomes is, therefore, the product of the poisson probabilities for the IJ cells;

$$P(Y_{11} = n_{11}, Y_{12} = n_{12}, \cdots, Y_{IJ} = n_{IJ}) = \prod_{i=1}^{I} \prod_{j=1}^{J} \frac{e^{-\mu_{ij}} \mu_{ij}^{n_{ij}}}{n_{ij}!}.$$

Example 2.5. Recall again example 2.3. If no sample is taken, the total sample size is a random variable. As a result, the number of observations at the four combinations of the two variables are treated as independent poisson random variables with unknown means $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\}$. If, for example, $\{\mu_{11}, \mu_{12}, \mu_{21}, \mu_{22}\} = \{10, 50, 60, 20\}$, we can easily find $P(n_{11}, n_{12}, n_{21}, n_{22})$.

Example 2.6. Given the following data from a political science study concerning opinion in a particular city of a new governmental policy affiliation.

	Policy Opinion			
Party	Favor Policy	Do not Favor Policy	No Opinion	Total
Democrats	200	200	100	500
Republicans	250	175	75	500
Total	450	375	175	1000

1. What are the sampling techniques that could have produced these data?
2. Construct the probability structure.
3. Find the multinomial sampling and independent multinomial sampling models.

Solution:

1. Two distinct sampling procedures can be considered that could have produced the data. In the first, a random sample of 1000 individuals in the city might be selected (the total sample size is fixed at 1000) and each individual is asked his/her party affiliation (democrats or republicans) and his/her opinion concerning the new policy (favor, do not favor or no opinion). This sampling scheme is multinomial sampling which elicits two responses from each individual. Hence, totally there are $2 \times 3 = 6$ response categories.

   In the second sampling scheme, a random sample of 500 democrats was selected from a list of registered democrats in the city and each democrat was asked his or her opinion concerning the new policy (favor, do not favor or no opinion) and a completely analogous procedure was used on 500 republicans (the marginal totals of both political party affiliations are fixed at 500 a priori). This is an independent multinomial sampling scheme which elicits only one response from each individual. Now, there are 3 response categories for each party affiliation.

2. The probability structure is

	Policy Opinion			
Party	Favor Policy	Do not Favor Policy	No Opinion	Total
Democrats	0.200	0.200	0.100	0.500
Republicans	0.250	0.175	0.075	0.500
Total	0.450	0.375	0.175	1.000

3. The multinomial sampling uses the above joint probability structure. For the independent multinomial sampling models, the conditional probability distribution for each party, shown below, is used.

	Policy Opinion			
Party	Favor Policy	Do not Favor Policy	No Opinion	Total
Democrats	0.40	0.40	0.20	1.00
Republicans	0.50	0.35	0.15	1.00

2.4 Chi-squared Tests of Independence

For a multinomial sampling with probabilities π_{ij} in an $I \times J$ contingency table, the null hypothesis of statistical independence is $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$ for all i and j. For independent multinomial samples, independence corresponds to homogeneity of each outcome probability among the categories of the fixed variable. The marginal probabilities then determine the joint probabilities.

Under $H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}$, the expected values of cell counts are $\{\mu_{ij} = n\pi_{i+}\pi_{+j}\}$. That is, μ_{ij} is the expected number of subjects in the i^{th} category of X and j^{th} category of Y. Since $\{\pi_{i+}\}$ and $\{\pi_{+j}\}$ are unknown, their maximum likelihood estimates, respectively, are $\left\{p_{i+} = \frac{n_{i+}}{n}\right\}$ and $\left\{p_{+j} = \frac{n_{+j}}{n}\right\}$. which are the sample marginal proportions. Hence, the estimated expected frequencies are $\left\{\hat{\mu}_{ij} = np_{i+}p_{+j} = \frac{n_{i+}n_{+j}}{n}\right\}$.

Table 2.4: Observed and Expected Frequencies in an $I \times J$ Table

	Y						
X	1	2	$\cdots$	j	$\cdots$	J	Total
1	n_{11} $(\hat{\mu}_{11})$	n_{12} $(\hat{\mu}_{12})$	$\cdots$	n_{1j} $(\hat{\mu}_{1j})$	$\cdots$	n_{1J} $(\hat{\mu}_{1J})$	n_{1+}
2	n_{21} $(\hat{\mu}_{21})$	n_{22} $(\hat{\mu}_{22})$	$\cdots$	n_{2j} $(\hat{\mu}_{2j})$	$\cdots$	n_{2J} $(\hat{\mu}_{2J})$	n_{2+}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1} $(\hat{\mu}_{i1})$	n_{i2} $(\hat{\mu}_{i2})$	$\cdots$	n_{ij} $(\hat{\mu}_{ij})$	$\cdots$	n_{iJ} $(\hat{\mu}_{iJ})$	n_{i+}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
I	n_{I1} $(\hat{\mu}_{I1})$	n_{I2} $(\hat{\mu}_{I2})$	$\cdots$	n_{Ij} $(\hat{\mu}_{Ij})$	$\cdots$	n_{IJ} $(\hat{\mu}_{IJ})$	n_{I+}
Total	n_{+1}	n_{+2}	$\cdots$	n_{+j}	$\cdots$	n_{+J}	n

2.4.1 The Chi-square Test Statistic

The Pearson chi-squared statistic for testing independence of two categorical variables is defined as:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} \sim \chi^2[(I-1)(J-1)].$$

Step 1: Hypothesis:

$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}\ \forall i, j$. The two variables have no significant association.

H_1 : **not** H_0. The variables are significantly associated.

Step 2: Obtain the critical value $\chi^2_\alpha[(I-1)(J-1)]$.

Step 3: The calculated value of the X^2 test statistic is:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}.$$

Step 4: Decision: If $X^2_{cal} > \chi^2_\alpha[(I-1)(J-1)]$, the null hypothesis H_0 of no statistical association can be rejected. Or if $p-\text{value} = P(\chi^2[(I-1)(J-1)] > X^2_{cal})$ is smaller than α, H_0 can be rejected.

Step 5: Conclusion.

2.4.2 The Likelihood-Ratio Test Statistic

The likelihood-ratio test statistic is an alternative test for independence that uses likelihood values. A likelihood-ratio statistic is defined as $G^2 = -2\log(\ell_0/\ell_1)$ where ℓ_0 is the maximized value of the likelihood function under H_0 and ℓ_1 is the maximized value of the likelihood function in general. Therefore, the likelihood-ratio test statistic for independence can be easily derived as

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right) \sim \chi^2[(I-1)(J-1)].$$

When H_0 holds, the Pearson X^2 and likelihood-ratio G^2 statistics both have asymptotic chi-squared distributions with $[(I-1)(J-1)]$ degrees of freedom. For a better approximation, the general rule is that the smallest expected frequency should be at least 5. In general, if more than 20% of the expected frequencies are less than 5, the approximation worsens (that is, the test is not valid).

Example 2.7. The table below shows the distribution of HIV/AIDS patients by the survival outcome (active, dead, transferred to other hospital and lost-to-follow) and gender.

	Survival Outcome				
Gender	Active	Dead	Transferred	Lost-to-follow	Total
Female	741	25	63	101	930
Male	392	20	52	70	534
Total	1133	45	115	171	1464

Test whether or not the survival outcome depends on gender using both the Pearson chi-square and likelihood-ratio tests.

Solution: First let us find the expected cell counts, $\hat{\mu}_{ij} = \dfrac{n_{i+}n_{+j}}{n}$.

	Survival Outcome				
Gender	Active	Dead	Transferred	Lost-to-follow	Total
Female	741 (719.7)	25 (28.6)	63 (73.1)	101 (108.6)	930
Male	392 (413.3)	20 (16.4)	52 (41.9)	70 (62.4)	534
Total	1133	45	115	171	1464

Step 1: Hypothesis:

$H_0 : \pi_{ij} = \pi_{i+}\pi_{+j}\ \forall i, j$. Survival outcome and gender have no significant association.

H_1 : **not** H_0. Survival outcome depends on gender.

Step 2: The critical value $\chi^2_\alpha[(2-1)(4-1)] = \chi^2_{0.05}(3) = 7.8147$.

Step 3: The calculated value of the X^2 and G^2 test statistics, respectively, are:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}} = \frac{(741-719.7)^2}{719.7} + \frac{(25-28.6)^2}{28.6} + \cdots + \frac{(70-62.4)^2}{62.4}$$
$$= 8.2172$$

and

$$G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij}\log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right) = 2\left[741\log\left(\frac{741}{719.7}\right) + 25\log\left(\frac{25}{28.6}\right) + \cdots + 70\log\left(\frac{70}{62.4}\right)\right]$$
$$= 8.0720$$

Step 4: Decision: Since both statistics have larger values than $\chi^2_{0.05}(3) = 7.8147$, the null hypothesis H_0 can be rejected. Also, $p-$ value $= P(\chi^2(3) > 8.0720) = 0.0445$ suggests rejection of no association between the two variables.

Step 5: Conclusion: The survival outcome of patients depends on gender at 5% level of significance.

2.5 Measuring Strength of Association

There are many situations where both the independent and dependent variables have two levels. Let X (explanatory) and Y (response) be binary variables. The data can be displayed in a 2×2 contingency table in which the rows are the levels of X and the columns are the levels of Y. Let us use the generic terms success and failure for the outcome categories of Y.

	Y		
X	Success (1)	Failure (2)	Total
1	N_{11}	N_{12}	N_{1+}
2	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N

For each category i; $i = 1, 2$ of X, $P(Y = j|X = i) = \pi_{j|i}$; $j = 1, 2$. Then, the conditional probability structure is as follows.

	Y		
X	Success (1)	Failure (2)	Total
1	$\pi_{1\|1}$	$\pi_{2\|1}$	1
2	$\pi_{1\|2}$	$\pi_{2\|2}$	1

Here, $\pi_{1|1}$ and $\pi_{1|2}$ are the proportions of successes in category 1 and 2 of X, respectively. From now onwards, let us use π_1 and π_2 are the proportions of successes in category 1 and 2 of X, respectively.

	Y		
X	Success (1)	Failure (2)	Total
1	π_1	π_1'	1
2	π_2	π_2'	1

In chi-square test, the question of interest is whether there is a statistical association between the explanatory (X) and the response (Y) variables. The hypothesis to be tested is

$$H_0 : \pi_1 = \pi_2 \text{ (There is no association between } X \text{ and } Y)$$
$$H_1 : \pi_1 \neq \pi_2 \text{ (There is an association between } X \text{ and } Y)$$

A significant chi-squared test merely tells the existence of the association between the variables. If an association exists, the next task is identifying the category of X which has a larger (smaller) proportion of successes. This can be done by calculating the *difference of proportions*, a *relative risk* and an *odds ratio*.

2.5.1 Difference of Proportions (Absolute Risk)

The difference of proportions (absolute risk) is a simple procedure which compares the probability of success between two groups. It is calculated as $\pi_1 - \pi_2$. It is interesting that the difference in proportions ranges between -1 and +1. If $\pi_1 - \pi_2 \approx 0$, the proportion of successes in both categories of X are almost the same (0 is a baseline for comparison). That is, if $\pi_1 - \pi_2 \approx 0$, categories of X have identical conditional distributions. On the contrary, if $\pi_1 - \pi_2 \approx \pm 1$, the association between X and Y is strong (indicates a high level of association).

Let p_1 and p_2 be the sample proportion of successes in category 1 and 2 of X, respectively. The difference of the sample proportion of successes $p_1 - p_2$ estimates the difference of the population proportion of successes $\pi_1 - \pi_2$. (Details are already discussed in Section 2.5.1).

Example 2.8. An educational researcher designs a study to compare the effectiveness of teaching English to non-English speaking people by a computer software program and by the traditional classroom system. The researcher randomly assigns 35 students from a class of 100 to instruction using the computer. The remaining 65 students are instructed using the traditional method. At the end of a 6-month instructional period, all 100 students are given an examination with the results reported in the following table.

	Examination Result		
Instruction Method	Pass	Fail	Total
Traditional	45	20	65
Computer	32	3	35
Total	77	23	100

Find the difference of the pass proportions and interpret. Also test the significance using the 95% confidence interval.

Solution: The conditional probabilities for each instruction method are shown in the following table.

	Examination Result		
Instruction Method	Pass	Fail	Total
Traditional	$p_1 = 0.692$	$p'_1 = 0.308$	1
Computer	$p_2 = 0.914$	$p'_2 = 0.086$	1

The difference in the sample pass proportions is $p_1 - p_2 = 0.692 - 0.914 = -0.222$. Since the difference is less than 0, computer instruction *seems* to be a better way to improve the academic performance of students in English course. The probability of passing in the traditional instruction method *decreases by* 0.222 as compared to passing in the computer instruction method. Or, the probability of passing in the computer instruction method *increases by* 0.222 as compared to passing in the traditional instruction method.

The 95% confidence interval for the difference in the pass proportions between the traditional and computer instruction methods $\pi_1 - \pi_2$ is

$$\left[(0.692 - 0.914) \pm 1.96\sqrt{\frac{0.692(1-0.692)}{65} + \frac{0.914(1-0.914)}{35}}\right]$$

$$= (-0.222 \pm \sqrt{0.0033 + 0.022}) = (-0.222 \pm 0.0742) = (-0.2962, -0.1478).$$

Thus, since the confidence interval is less than 0, the difference of the pass proportions in the two instruction methods is significantly different (particularly computer instruction is better than traditional instruction). Specifically, the probability of passing in the traditional instruction method *decreases by* between 0.1478 and 0.2962 at 5% significance level as compared to passing in the computer instruction method.

2.5.2 Relative Risk

Relative risk is the ratio of the probability of successes in two groups. That is,

$$r = \frac{\pi_1}{\pi_2} = \frac{N_{11}N_{12}}{N_{1+}N_{2+}}.$$

The value of a relative risk is non-negative, that is, $r \geq 0$. If $r \approx 1$, the proportion of successes in the two categories of X are approximately the same. This corresponds to independence or it is baseline for comparison. On the other hand, values of the relative risk r farther from 1 in a given direction represent stronger association. A relative risk of 4 is farther from independence than a relative risk of 2, and a relative risk of 0.25 is farther from independence than a relative risk of 0.50. Two values for relative risk (for example, 4 and 0.25) represent the same strength of association, but in opposite directions, when one value is the inverse of the other.

The sample relative risk $\hat{r} = \frac{p_1}{p_2}$ estimates the population relative risk r.

Example 2.9. Find the relative risk for the data given on example 2.8 and interpret it.

Solution: The conditional probabilities for each instruction method are shown in the following table.

	Examination Result		
Instruction Method	Pass	Fail	Total
Traditional	$p_1 = 0.692$	$p_1' = 0.308$	1
Computer	$p_2 = 0.914$	$p_2' = 0.086$	1

The estimate of the relative risk is $\hat{r} = \frac{p_1}{p_2} = \frac{0.692}{0.914} = 0.757$. It can be interpreted as follows:

- The proportion of passing in the traditional instruction method is 0.757 *times* the proportion of passing in the computer instruction method.
- The traditional instruction method *reduces* the probability of passing by $(1-\hat{r})100\% = (1-0.757)100\% = 24.3\%$ relative to computer instruction method.
- Or, by inverting, the probability of passing in the computer instruction method is 1.321 *times* the probability of passing in the traditional instruction method.
- This means, computer instruction method (relative to traditional instruction method) *increases* the probability of passing the exam by $(\hat{r}-1)100\% = (1.321-1)100\% = 32.1\%$.

Note: Relative risk is a widely reported measure of association between exposure status and disease state for prospective studies (cohort and randomized clinical trials). In such case, the levels of the explanatory variable are being exposed (E) and being unexposed (E'), and the levels of the response variable are having a disease (D) and not-having a disease (D').

	Disease		
Exposure	Present (D)	Absent (D')	Total
Exposed (E)	n_{11}	n_{12}	n_{1+}
Unexposed (E')	n_{21}	n_{22}	n_{2+}
Total	?	?	n

For this particular case, relative risk is a ratio of the probability of having a disease among those exposed to the probability of having the disease among those unexposed:

$$r = \frac{P(D|E)}{P(D|E')}.$$

- A relative risk of 1.0 implies that the risk of a disease is the same in both exposed and unexposed groups (no association between the exposure and the disease).
- A relative risk greater than 1.0 implies the exposed group have a higher probability of having a disease than the unexposed group (the exposure is a *risk* factor).
- A relative risk less than 1.0 implies that the exposed group has a lower chance of having disease than unexposed group (it is expected in drug efficacy studies, the exposure is a *protective* factor).

Testing for a Relative Risk

To infer about a relative risk r, the sampling distribution of the sample relative risk $\hat{r}$ should be determined. The values of the relative risk are highly skewed to the right. As a result, by taking the logarithm of $\hat{r}$, it turns out that $\log(\hat{r})$ is approximately normally distributed for large values of n. If the probability of successes are approximately equal in the two groups, then $r = 1$ or $\log(r) = 0$ indicating no statistical association between the two variables.

The standard error of $\log(\hat{r})$ is determined to be:

$$\text{SE}[\log(\hat{r})] = \sqrt{\frac{1}{N_{11}} - \frac{1}{N_{1+}} + \frac{1}{N_{21}} - \frac{1}{N_{2+}}}$$

which can be estimated by:

$$\widehat{\text{SE}}[\log(\hat{r})] = \sqrt{\frac{1}{n_{11}} - \frac{1}{n_{1+}} + \frac{1}{n_{21}} - \frac{1}{n_{2+}}}.$$

Step 1: Hypothesis:

$H_0 : \log(r) = 0$ The two variables have no significant association.

$H_1 : \log(r) \neq 0$ The two variables are significantly associated.

Step 2: Obtain the critical value $z_{\alpha/2}$.

Step 3: Under $H_0 : \log(r) = 0$, for large values of n the test statistic is defined as:

$$Z = \frac{\log(\hat{r}) - \log(r)}{\widehat{\text{SE}}[\log(\hat{r})]} \sim \mathcal{N}(0, 1).$$

Step 4: Decision: If $|z_{cal}| > z_{\alpha/2}$, H_0 should be rejected.

Step 5: Conclusion.

Example 2.10. Test the significance of the relative risk for the data given on example 2.8.

Solution: The estimate of the relative risk is $\hat{r} = \frac{p_1}{p_2} = \frac{0.692}{0.914} = 0.757$ which implies $\log(\hat{r}) = \log(0.757) = -0.2784$ and the estimated standard error of $\log(\hat{r})$ is $\widehat{\text{SE}}[\log(\hat{r})] = 0.0975$.

Step 1: Hypothesis:

$H_0 : r = 1 \Rightarrow \log(r) = 0$. Instruction method and exam result have no significant association.

$H_1 : r \neq 1 \Rightarrow \log(r) \neq 0$. Instruction method and exam result have a significant association.

Step 2: Using $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$.

Step 3: The calculated value of the z test statistic is:

$$z = \frac{\log(0.757) - 0}{\sqrt{\frac{1}{45} - \frac{1}{65} + \frac{1}{32} - \frac{1}{35}}} = -2.86.$$

Step 4: Decision: Since $|z_{cal}| = 2.86 > z_{0.025} = 1.96$, H_0 should be rejected.

Step 5: Conclusion: Therefore, the relative risk is significantly different from 1. Instruction method has a significant effect on examination result at 5% significance level. Specifically, the computer instruction method has a positive effect in passing the examination.

Confidence Interval for a Relative Risk

The $(1-\alpha)100\%$ confidence interval for the log of a relative risk $\log(r)$ is given by

$$\{\log(\hat{r}) \pm z_{\alpha/2}\widehat{\text{SE}}[\log(\hat{r})]\}.$$

Taking the exponentials of the end points this confidence interval provides the confidence interval for a relative risk r, that is,

$$\exp\{\log(\hat{r}) \pm z_{\alpha/2}\widehat{\text{SE}}[\log(\hat{r})]\}.$$

Example 2.11. An efficacy study was conducted for the drug pamidronate in patients with Paget's disease of bone. In this randomized clinical trial, patients were assigned at random to receive either pamidronate (E) or placebo (E'). One end point was the occurrence of any skeletal events after 9 cycles of treatment D and non-occurrence D'. The results are given in the following table.

	Skeletal Event		
Exposure	Yes (D)	No (D')	Total
Pamidronate (E)	47	149	196
Placebo (E')	74	107	181
Total	121	256	377

Compute a 95% confidence interval for the relative risk of suffering skeletal events (in a time period of this length) for patients on pamidronate relative to patients not on the drug.

Solution: Let $\pi_1 = P(D|E)$ and $\pi_2 = P(D|E')$. Thus, the estimated probability of patients suffering skeletal events among those receiving the drug, and among those receiving the placebo are $p_1 = \frac{47}{196} = 0.240$ and $p_2 = \frac{74}{181} = 0.409$, respectively.

Then, the estimated relative risk r is $\hat{r} = \frac{0.240}{0.409} = 0.587$ and its log value is $\log(\hat{r}) = -0.533$. The estimated standard error of log of the estimated relative risk $\log(\hat{r})$ is $\widehat{\text{SE}}[\log(\hat{r})] = \sqrt{\frac{1}{47} - \frac{1}{196} + \frac{1}{74} - \frac{1}{181}} = 0.155$.

The 95% confidence interval for the log of the relative risk $\log(r)$ is $-0.533 \pm 1.96(0.155) = (-0.837,\ -0.229)$. Therefore, the 95% confidence interval for the relative risk r is

$$\{\exp(-0.837),\ \exp(-0.229)\} = (0.433,\ 0.795).$$

Thus, the relative risk of suffering a skeletal event (in this time period) for patients on pamidronate (relative to patients not on pamidronate) is between 0.433 and 0.795 at 5% significance level. Since this entire interval is below 1, it can be concluded that pamidronate is effective in reducing the risk of skeletal events. Furthermore, pamidronate reduces the risk of skeletal events by $(1 - \hat{r})100\% = (1 - 0.587)100\% = 41.3\%$.

2.5.3 Odds Ratio

Before defining an odds ratio, let us define what an odds is? An odds (Ω) is the ratio of the probability of success to the probability of failure in a particular group.

$$\Omega = \frac{p(\text{success})}{p(\text{failure})} = \frac{\pi}{1-\pi} = \frac{\text{number of successes}}{\text{number of failures}}$$

Like a relative risk, an odds is a nonnegative number ($0 \leq \Omega < \infty$). If $\Omega = 1$, a successes is as likely as a failure. If $\Omega < 1$, a success is less likely and if $\Omega > 1$, a success is more likely to occur than a failure. Inversely,

$$\pi = \frac{\Omega}{1+\Omega}.$$

Odds ratio is the ratio of two odds. For a 2×2 table, for each group i of X, the odds of successes (instead of failures) is

$$\Omega_i = \frac{\pi_i}{1-\pi_i} = \frac{\pi_i}{\pi_i'};\ i = 1, 2.$$

Thus, the odds ratio is

$$\theta = \frac{\Omega_1}{\Omega_2} = \frac{\pi_1\pi_2'}{\pi_2\pi_1'} = \frac{N_{11}N_{22}}{N_{12}N_{21}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

Like a relative risk and an odds, an odds ratio is also non negative, that is, $\theta \geq 0$. An odds ratio of 1 implies independence of X and Y which is a baseline for comparison. If it larger than 1 ($\Omega_1 > \Omega_2$), a success is more likely to occur in category 1 of X than in category 2. If the odds ratio is near zero ($\Omega_1 < \Omega_2$), then a success is less likely to occur in category 1 than category 2.

Similar to a relative risk, values of an odds ratio θ farther from 1 in a given direction represent stronger association, that is, an odds ratio of 6 is farther from independence than an odds ratio of 2, and an odds ratio of 0.20 is farther from independence than an odds ratio of 0.60. Also, two values for odds ratio, when one value is the inverse of the other (for example, 5 and 0.20) represent the same strength of association, but in opposite directions.

The sample odds ratio $\hat{\theta}$ is used to estimate the population odds ratio θ which is given by

$$\hat{\theta} = \frac{\widehat{\Omega}_1}{\widehat{\Omega}_2} = \frac{n_{11}n_{22}}{n_{12}n_{21}} = \frac{p_{11}p_{22}}{p_{12}p_{21}}.$$

Example 2.12. Again recall example 2.8. Find the odds ratio and interpret.

Solution: The estimated probability of passing in the traditional instruction method is $p_1 = 0.692$. Then, the estimated odds of passing in this group is $\widehat{\Omega}_1 = \frac{0.692}{1-0.692} = 2.247$ which means the probability of passing in the traditional instruction group is 2.247 times the probability of failing in that group.

Similarly, the estimated probability of passing in the computer instruction group is $p_2 = 0.914$. Hence, the estimated odds of passing in this group is $\widehat{\Omega}_2 = \frac{0.914}{1-0.914} = 10.628$ which means the probability of passing in the computer instruction group is 10.627 times the probability of failing.

Therefore, the odds ratio of passing the exam (instead of failing) is the ratio of the odds of passing in the traditional instruction method to the odds of passing in the computer instruction group, that is, $\hat{\theta} = \frac{\widehat{\Omega}_1}{\widehat{\Omega}_2} = \frac{2.247}{10.628} = 0.211$. This value can be interpreted in different ways as follows.

- The odds of passing (instead of failing) the exam in the traditional instruction method is 0.211 *times* the odds of passing in the computer instruction method.
- The odds of passing (instead of failing) in the traditional instruction group *decreases by a factor of* 0.211 relative to the odds of passing in the computer instruction group.
- That is, the odds of passing (instead of failing) in the traditional instruction group is $(1 - \hat{\theta})100\% = (1 - 0.211)100\% = 78.9\%$ *lower than* the odds of passing in the computer instruction group.
- Those in the traditional instruction method group are 0.211 times *less likely* to pass the exam (instead of failing) than those in the computer instruction group.
- Or inversely, the odds of passing (instead of failing) the exam in the computer instruction group is 4.739 *times* the odds of passing in the traditional instruction group.

- The odds of passing (instead of failing) the exam in the computer instruction group *increases by a factor of* 4.739 as compared to those in the traditional instruction group.
- This means, the odds of passing (instead of failing) the exam in the computer instruction method is $(\hat{\theta} - 1)100\% = (4.739 - 1)100\% = 373.9\%$ *higher than* the odds of passing in the traditional instruction group.
- Those in the computer instruction group are 4.739 times *more likely* to pass the exam (instead of failing) than those in the traditional instruction method group.

Example 2.13. Given the following contingency table for the variable "death penalty for crime".

	Race		
Penalty	Blacks	Nonblacks	Total
Death Sentence	28	22	50
Life Imprisonment	45	52	97
Total	73	74	147

Find the odds of receiving a death sentence and interpret. Also, calculate the odds ratio for receiving a death penalty and interpret.

Solution: The estimated probability of receiving a death sentence is $\frac{50}{147} = 0.34$ (34%). Then, the estimated odds of receiving a death sentence (instead of a life imprisonment sentence) is $\frac{50}{97} = 0.516$ (51.6%). Receiving a death sentence is *half as likely as* life imprisonment or receiving a life imprisonment sentence is *twice as likely as* receiving a death penalty.

The odds ratio for receiving a death penalty (instead of life imprisonment) is the ratio of the odds if black to the odds if nonblack. It is estimated as 1.47 which means blacks are 1.47 *times more likely* to receive a death sentence (instead of life imprisonment) than nonblacks. This means, the risk (odds) of death sentence (instead of life imprisonment) for blacks *increases by a factor of* 1.47 as compared to nonblacks. Or the risk (odds) of death sentence for blacks are 47% *higher than* the risk (odds) of a death sentence for nonblacks.

Note: For retrospective (case-control) studies, subjects are identified as cases (D) or controls (D'), and it is observed whether the subjects had been exposed to the risk factor (E) or not (E'). Since the samplings are not from the populations of exposed and unexposed, and observing whether or not disease occurs (as in prospective studies), $P(D|E)$ or $P(D|E')$, cannot be estimated.

	Disease		
Exposure	Present (D)	Absent (D')	Total
Exposed (E)	n_{11}	n_{12}	?
Unexposed (E')	n_{21}	n_{22}	?
Total	n_{+1}	n_{+2}	n

- If the odds ratio is 1.0, the odds (and thus probability) of disease is the same for both groups (no association between an exposure and a disease).
- If the odds ratio is greater than 1.0, the odds (and thus probability) of disease is higher among exposed than unexposed (the exposure is a *risk* factor).
- If the odds ratio is less than 1.0, the odds (and thus probability) of disease is lower among exposed than unexposed (the exposure is a *protective* factor).

Testing for an Odds Ratio

To infer about an odds ratio θ, the sampling distribution of $\log(\hat{\theta})$ is used due to the similar reasons used for a relative risk. If the odds of successes are equal in the two groups, then $\theta = 1$ or $\log(\theta) = 0$ indicating independence (no statistical association).

The standard error of the log of an odds ratio $\log(\hat{\theta})$ can be determined using statistical theory as:

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

which can be estimated by:

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

Step 1: Hypothesis:

H_0 : $\textbf{OR} = 1 \Rightarrow \log(\textbf{OR}) = 0$. The two variables have no significant association.

H_1 : $\textbf{OR} \neq 1 \Rightarrow \log(\textbf{OR}) \neq 0$. The two variables are significantly associated.

Step 2: Obtain the critical value $z_{\alpha/2}$.

Step 3: Under $H_0 : \log(\theta) = 0$, for large values of n the test statistic is defined as:

$$Z = \frac{\log(\hat{\theta}) - \log(\theta)}{\widehat{\text{SE}}[\log(\hat{\theta})]} \sim \mathcal{N}(0, 1).$$

Step 4: Decision: If $|z_{cal}| > z_{\alpha/2}$, H_0 can be rejected.

Step 5: Conclusion.

Example 2.14. Test the significance of the odds ratio for the data given at example 2.13.

Solution: It is easily to calculate that $\hat{\theta} = 1.47$ and $\log(\hat{\theta}) = 0.385$. Also, the standard error of $\log(\hat{\theta})$ is $\widehat{\text{SE}}[\log(\hat{\theta})] = 0.349$.

Step 1: Hypothesis:

$H_0 : \theta = 1 \Rightarrow \log(\theta) = 0$. Death penalty and race have no significant association.

$H_1 : \theta \neq 1 \Rightarrow \log(\theta) \neq 0$. Death penalty and race have a significant association.

Step 2: Using $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$.

Step 3: The calculated value of the Z test statistic is:

$$z = \frac{\log(\hat{\theta}) - 0}{\sqrt{\frac{1}{28} + \frac{1}{22} + \frac{1}{45} + \frac{1}{52}}} = 1.103.$$

Step 4: Decision: Since $|z_{cal}| = 1.103 < z_{0.025} = 1.96$, H_0 cannot be rejected.

Step 5: Conclusion: Therefore, there is not much evidence of association between penalty for crime and race at 5% significance level.

Confidence Interval for an Odd Ratio

The $(1 - \alpha)100\%$ confidence interval for an odds ratio θ is given by

$$\exp\{\log(\hat{\theta}) \pm z_{\alpha/2}\widehat{\text{SE}}[\log(\hat{\theta})]\}.$$

Example 2.15. An epidemiological case-control study was reported, with cases being 537 people diagnosed with lung cancer (D) and controls being made up of 500 people with no lung cancer (D'). One risk factor measured was whether or not the subject had smoked a cigarette (a smoker - E, a non-smoker - E'). The following table gives the numbers of subjects falling in each possible combination.

	Lung Cancer		
Exposure	Yes (D)	No (D')	Total
Smoker (E)	339	149	488
Nonsmoker (E')	198	351	549
Total	537	500	1037

Compute a 95% confidence interval for the population odds ratio, and determine whether or not cigarette smoking is associated with higher (or possibly lower) odds (and probability) of developing lung cancer.

Solution: The estimated odds ratio for developing cancer cancer in smokers and non-smokers is $\hat{\theta} = \frac{339(351)}{149(198)} = 4.03$. This implies $\log(\hat{\theta}) = 1.394$ and its estimated standard error is $\widehat{\text{SE}}\{\log(\hat{\theta})\} = \sqrt{\frac{1}{339} + \frac{1}{149} + \frac{1}{198} + \frac{1}{351}} = 0.133$. Therefore, the 95% confidence interval for the odds ratio θ is

$$\{\exp[1.394 - 1.96(0.133)],\ \exp[1.394 + 1.96(0.133)]\} = (3.110,\ 5.231).$$

That is, the risk of developing lung cancer is between 3.11 and 5.231 times higher among smokers than non-smokers at $\alpha = 0.05$.

Odds Ratios in an $I \times J$ Table

For a 2×2 table, a single number such as an odds ratio can summarize the association. For an $I \times J$ table, it is rarely possible to summarize association by a single number without some loss of information. However, a set of $(I-1)(J-1)$ local odds ratios can describe certain features of the association (the rest odds ratios can be determined from these odds ratios).

Consider category i and $i+1$ of X, and category j and $j+1$ of Y in an $I \times J$ contingency table. Then, the odds ratio:

$$\theta_{ij} = \frac{N_{ij}N_{i+1,j+1}}{N_{i,j+1}N_{i+1,j}} = \frac{\pi_{ij}\pi_{i+1,j+1}}{\pi_{i,j+1}\pi_{i+1,j}};\ i = 1, 2, \cdots, I-1,\ j = 1, 2, \cdots, J-1$$

compares the probability of category j (instead of $j+1$) of Y in category i of X as compared to category $i+1$ of X.

As usual, the estimated odds ratio for comparing category j (instead of $j+1$) of Y between category i and $i+1$ of X is:

$$\hat{\theta}_{ij} = \frac{n_{ij}n_{i+1,j+1}}{n_{i,j+1}n_{i+1,j}} = \frac{p_{ij}p_{i+1,j+1}}{p_{i,j+1}p_{i+1,j}};\ i = 1, 2, \cdots, I-1,\ j = 1, 2, \cdots, J-1.$$

Independence is equivalent to all odds ratios equal to 1 (that is, non-significance of all odds ratios).

Example 2.16. Suppose 980 individuals are classified according to their favorite soft drink preference (Fanta, Coca and Sprite) and gender as shown below.

	Soft Drink			
Gender	Fanta	Coca	Sprite	Total
Females	279	225	73	577
Males	165	191	47	403
Total	444	416	120	980

By looking at the frequencies in the table, guess which gender (male or female) seems more likely to prefer coca? Why? Find all (local) odds ratios and test their significance.

Solution: The association between gender and soft drink preference can be checked using the chi-square or likelihood-ratio tests.

Step 1: Hypothesis:

H_0 : There is no significant association between soft drink preference and gender.

H_1 : Soft drink preference significantly depends on gender.

Step 2: Assuming $\alpha = 0.05$, the critical value is $z_{0.025} = 1.96$.

Step 3: The z test statistic is used for testing each odds ratio:

	Fanta versus Coca	Fanta versus Sprite	Coca versus Sprite
Odds Ratio ($\hat{\theta}_{ij}$)	$\frac{279(191)}{225(165)} = 1.435$	$\frac{279(47)}{73(165)} = 1.089$	$\frac{225(47)}{73(191)} = 0.758$
Log Odds Ratio $\{\log(\hat{\theta}_{ij})\}$	$\log(1.435) = 0.361$	$\log(1.089) = 0.085$	$\log(0.758) = -0.120$
$\widehat{\text{SE}}[\log(\hat{\theta}_{ij})]$	0.139	0.211	0.211
Test Statistic (z)	2.597	0.402	−0.569
Decision	Reject H_0	Do not reject H_0	Do not reject H_0

Step 4: Decision: Since one of the three odds ratios is significant at 5% significance level, the null hypothesis of no significant association is rejected.

Step 5: Conclusion: Therefore, there is a significant difference in the preference of Fanta (instead of Coca) by females as compared to males at 5% level of significance. Hence, from this analysis, it can be concluded that:

- Females are 1.435 times *more likely* to prefer Fanta (instead of Coca) than that of males.
- The odds of preferring Fanta (instead of Coca) by females is 43.5% *higher than* that of males.
- Males are 0.697 times *less likely* to prefer Fanta (instead of Coca) than females.
- The odds of preferring Fanta (instead of Coca) by males is 30.3% *lower than* that of females.

2.6 Exact Inference for Small Samples

The inferential methods of the previous sections are all large sample methods. The Pearson chi-square statistic is only approximated by the chi-square distribution, and that approximation worsens with small expected frequencies. When there are very small expected frequencies, the possible values of the chi-square statistic are quite discrete. For example, for a 2×2 table with only 4 observations in each row and column, the only possible values of chi-square are 8, 2, and 0. It should be clear that a continuous chi-square distribution is not a good match for a discrete distribution having only 3 values. In such cases, when n is small, alternative methods use exact distributions rather than large sample approximations.

In this section, small sample test of independence for 2×2 tables, which is called Fisher's exact inference is discussed. As described in Section 2.3.3, in poisson sampling - the sample size is not fixed unlike multinomial sampling, and in independent multinomial (binomial) sampling only one set of the marginal totals are fixed. In addition, in a 2×2 table, if both sets of the marginal total are fixed, it yields a hypergeometric distribution, that is,

$$P(Y_{11} = n_{11}) = \frac{\binom{n_{1+}}{n_{11}}\binom{n_{2+}}{n_{+1} - n_{11}}}{\binom{n}{n_{+1}}}.$$

Given the marginal totals, n_{11} determines the other three cell counts. The exact p-value is determined using the hypergeometric distribution. The procedure to calculate the p-value for testing $H_0 : \theta = 1$ is as follows. Of the four marginal totals, select the smallest one and create ordered pair of integers with that sum. Next complete the 2×2 table for each of the ordered pair. Then, the two-sided p-value is given by $P(Y_{11} \leq n_{11})$ where n_{11} is the observed frequency in cell $(1, 1)$. For a one sided test, the p-value is found by comparing the observed frequency n_{11} to its expected value $\hat{\mu}_{11}$. If $n_{11} > \hat{\mu}_{11}$, then the onesided (right-sided alternative: $H_1 : \theta > 1$) p-value is $P(Y_{11} \geq \hat{\mu}_{11})$ and if $n_{11} < \hat{\mu}_{11}$, then the onesided (left-sided alternative: $H_1 : \theta < 1$) p-value is $P(Y_{11} \leq n_{11})$.

Example 2.17. Suppose A and B are two small colleges, the results of the beginning Statistics course at each of the two colleges are given below.

	Statistics		
Colleges	Pass	Fail	Total
A	8	14	22
B	1	3	4
Total	9	17	26

Do the data provide sufficient evidence to indicate that the proportion of passing Statistics differs for the two colleges?

Solution: The hypothesis to be tested is, $H_0 : \pi_{1|A} = \pi_{1|B}$, the proportion of passing Statistics do not differ significantly for the two colleges. Since the sample sizes are small, Fisher's exact test will be used. Since $n_{2+} = 4$ is the smallest marginal total, the following ordered pairs for (n_{21}, n_{22}) can be determined: (0, 4), (1, 3), (2, 2), (3, 1) and (4,0). For each pair, the 2×2 table is completed and the corresponding probability is computed using

$$P(Y_{11} = n_{11}) = \frac{n_{1+}!\ n_{2+}!\ n_{+1}!\ n_{+2}!}{n!\ n_{11}!\ n_{12}!\ n_{21}!\ n_{22}!}.$$

For (n_{21}, n_{22})=(0, 4):

$$\begin{array}{|c|c|}\hline 9 & 13 \\ \hline 0 & 4 \\ \hline \end{array} \Rightarrow P(Y_{11} = 9) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 9!\ 13!\ 0!\ 4!} = 0.159197$$

For (n_{21}, n_{22})=(1, 3):

$$\begin{array}{|c|c|}\hline 8 & 14 \\ \hline 1 & 3 \\ \hline \end{array} \Rightarrow P(Y_{11} = 8) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 8!\ 14!\ 1!\ 3!} = 0.409365$$

For (n_{21}, n_{22})=(2, 2):

$$\begin{array}{|c|c|}\hline 7 & 15 \\ \hline 2 & 2 \\ \hline \end{array} \Rightarrow P(Y_{11} = 7) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 7!\ 15!\ 2!\ 2!} = 0.327492$$

For (n_{21}, n_{22})=(3, 1):

$$\begin{array}{|c|c|}\hline 6 & 16 \\ \hline 3 & 1 \\ \hline \end{array} \Rightarrow P(Y_{11} = 6) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 6!\ 16!\ 3!\ 1!} = 0.095518$$

For (n_{21}, n_{22})=(4, 0):

$$\begin{array}{|c|c|}\hline 5 & 17 \\ \hline 4 & 0 \\ \hline \end{array} \Rightarrow P(Y_{11} = 5) = \frac{22!\ 4!\ 9!\ 17!}{26!\ 5!\ 17!\ 4!\ 0!} = 0.008428$$

Since the observed frequency $n_{11} = 8$, the two sided p-value is $P(Y_{11} \leq 8) = P(Y_{11} = 5) + P(Y_{11} = 6) + P(Y_{11} = 7) + P(Y_{11} = 8) = 1$. Hence, there is not enough evidence to conclude that the proportion of passing Statistics differs for the two colleges.

Since the observed frequency $n_{11} = 8 > \hat{\mu}_{11} = 7.6$, the alternative hypothesis is ($H_1 : \pi_{1|A} > \pi_{1|B}$). Then the onesided p-value is $P(Y_{11} \geq 7.6) = P(Y_{11} = 8) + P(Y_{11} = 9) = 0.159197 + 0.409365 = 0.568562$. Again, there is not enough evidence to indicate that the probability of passing Statistics is higher at college A than at college B.

2.7 Measures of Linear Association for Ordinal Variables

In situations where both the explanatory and response variables are ordinal, the X^2 and G^2 tests ignore the fact that the levels of the variables have distinct orderings. When both variables are ordinal, there will be an interest to examine whether individuals with high levels of an explanatory variable tend to have high (low) levels of the corresponding response variable. For instance, suppose that the explanatory variable is dose, with increasing (possibly numeric) levels of amount of drug given to a patient, and the response variable is categorical measuring the degree of improvement. Then, it is essential to determine if as dose increases, the degree of improvement increases.

Many measures have been developed for this type of ordinal variables classification. Most analytical techniques are based on concordant and discordant pairs. A *concordant* pair involves a pair where a subject is higher on both variables than other subject. A *discordant* pair is a pair where a subject is higher on one variable, but lower on the other variable, than other subject. If a pair is said to be *tied* if a subject is in the same category of a variable.

More concordant pairs than discordant pairs indicates a *positive association* between the two variables whereas more discordant pairs than concordant pairs indicates *negative association* between the variables.

Consider the following table

	Income Level		
Education Level	Low	High	Total
High School	N_{11}	N_{12}	
College	N_{21}	N_{22}	
Total			

Looking at the above table, it is easy to observe that income category is ordered by low and high. Similarly education category is ordered, with education ending at high school being the low category and education ending at college being the high category. All N_{11} observations represent individuals in low income and low education category and all N_{22} observations represent individuals in high income and high education category. Thus, there are $C = N_{11}N_{22}$ concordant pairs. On the other hand, all N_{12} observations are higher on the income variable and lower on the education variable, while all N_{21} observations are lower on the income variable and higher on the education variable. Thus, there are $D = N_{12}N_{21}$ discordant pairs.

2.7.1 The Gamma Measure

The strength of the association can be measured by calculating the difference in the proportions of concordant and discordant pairs. This is called the *gamma* (γ) measure which is defined as

$$\gamma = \frac{C}{C+D} - \frac{D}{C+D} = \frac{C-D}{C+D}.$$

Since γ represents the difference in proportions, its value is between -1 and 1. A positive value of gamma indicates a positive association while a negative value of gamma indicates a negative association. A value close to zero indicates weak association.

Let us consider again the above 2×2 table. Let $n_{11} = 25$, $n_{12} = 12$, $n_{21} = 11$ and $n_{22} = 14$. The number of concordant pais is $\widehat{C} = n_{11}n_{22} = 25(14) = 350$; the number of discordant pairs is $\widehat{D} = n_{12}n_{21} = 12(11) = 132$. Therefore, $\widehat{\gamma} = 0.45$ which indicates that the association between education level and income is medium-positive.

For an $I \times J$ table, the number of concordant pairs is $C = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}(\sum_{h=i+1}^{I}\sum_{k=j+1}^{J} N_{hk})$ and the number of discordant pairs is $D = \sum_{i=1}^{I}\sum_{j=1}^{J} N_{ij}(\sum_{h=i+1}^{I}\sum_{k=1}^{j-1} N_{hk})$.

Example 2.18. Find the gamma measure of association for the following cross-classification of HIV/AIDS patients by Clinical Stage and Functional Status.

	Functional Status			
Clinical Stage	Bedridden	Ambulatory	Working	Total
Stage I	0	23	324	347
Stage II	11	96	407	514
Stage III	28	233	235	496
Stage IV	18	52	37	107
Total	57	404	1003	1464

Solution: The total number of concordant pairs is

$$\begin{aligned}\widehat{C} =& 0(96+407+233+235+52+37)+23(407+235+37)\\ &+11(233+235+52+37)+96(235+37)+28(52+37)+233(37)\\ =&58969\end{aligned}$$

The total number of discordant pairs is

$$\begin{aligned}\widehat{D} =& 23(11+28+18)+324(11+96+28+233+18+52)+96(28+18)\\ &+407(28+233+18+52)+233(18)+235(18+52)\\ =&303000\end{aligned}$$

In this example, $\widehat{C} < \widehat{D}$, suggesting a tendency for low clinical stage to occur with high functional status of patients and higher clinical stages with lower functional status.

$$\widehat{\gamma} = \frac{\widehat{C} - \widehat{D}}{\widehat{C} + \widehat{D}} = \frac{58969 - 303000}{58969 + 303000} = -0.674$$

Of the untied pairs, the proportion of concordant pairs is 0.674 lower than the proportion of discordant pairs. This indicates that there is a medium negative linear association between clinical stage and functional status of HIV/AIDS patients. That is, as the clinical stage (severity) of the patient increases, the functional status of the patient decreases and vice versa.

2.7.2 The Kendall's tau-b

Kendall's tau-b, denoted τ_b, is a more sensitive measure of association between two ordinal variables. The formula for calculating Kendall's tau-b τ_b is:

$$\tau_b = \frac{C - D}{0.5\sqrt{\left(N^2 - \sum_{i=1}^{I} N_{i+}^2\right)\left(N^2 - \sum_{j=1}^{J} N_{+j}^2\right)}}.$$

The estimated value of Kendall's tau-b $\hat{\tau}_b$ is also obtained by substituting the sample frequencies in place of the population frequencies as:

$$\hat{\tau}_b = \frac{\widehat{C} - \widehat{D}}{0.5\sqrt{\left(n^2 - \sum_{i=1}^{I} n_{i+}^2\right)\left(n^2 - \sum_{j=1}^{J} n_{+j}^2\right)}}.$$

This measure has the advantage of adjusting for ties. The result of adjusting for ties is that the value of τ_b is always a little closer to 0 than the corresponding value of gamma.

Example 2.19. Find the Kendall's tau-b τ_b for the data given in example 2.18.

Solution:

$$\begin{aligned}\hat{\tau}_b &= \frac{58969 - 303000}{0.5\sqrt{[1464^2 - (57^2 + 404^2 + 1003^2)][1464^2 - (347^2 + 514^2 + 496^2 + 107^2)]}} \\ &= \frac{-244031}{0.5\sqrt{(2143296 - 1172474)(2143296 - 642070)}} \\ &= -0.404\end{aligned}$$

2.8 Association in Three-Way Tables

An important part of most studies, especially observational studies, is the choice of control variables. In studying the effect of X on Y, one should control any covariate that can influence that relationship. This involves using some mechanism to hold the covariate constant. Otherwise, an observed effect of X on Y may actually reflect effects of that covariate on both X and Y. The relationship between X and Y then shows *confounding.* Experimental studies can remove effects of confounding covariates by randomly assigning subjects to different levels of X, but this is not possible with observational studies.

2.8.1 Partial Tables

The variable Z can be controlled by studying the XY relationship at fixed levels of Z. Two-way cross-sectional slices, called *partial* tables, of the three-way contingency table cross-classify X and Y at separate categories of Z. Thsese *partial* tables display the XY relationship while removing the effect of Z by holding its value constant.

The two-way contingency table obtained by combining the partial tables is called the XY *marginal* table. Each cell count in the marginal table is a sum of counts from the same location in the partial tables. The marginal table, rather than controlling Z, ignores it. The marginal table contains no information about Z. It is simply a two-way table relating X and Y but may reflect the effects of Z on X and Y.

The associations in partial tables are called *conditional* associations, because they refer to the effect of X on Y conditional on fixing Z at some level. Conditional associations in partial tables can be quite different from associations in marginal tables.

Example 2.20. Consider the following cross-classification of subjects by gender (Male, Female), smoking (Yes, No) and occurrence of lung cancer (Yes, No).

		Lung Cancer		
Gender	Smoking	Yes	No	Lung Cancer (%)
Male	Yes	45	100	31.0345
	No	13	102	11.3044
Female	Yes	10	402	2.4272
	No	0	12	0.0000
Total	Yes	55	502	9.8743
	No	13	114	10.2362

For each combination of gender and smoking, the above table displays the percentage of subjects who developed lung cancer. These describe the *conditional* associations. When the subjects were male, lung cancer was occurred $31.0345\% - 11.3044\% = 19.7301\%$ more often for smokers than for non-smokers. When the subjects were female, lung cancer was occurred $2.4272\% - 0.0000\% = 2.4272\%$ more often for smokers than for non-smokers. Controlling for subjects' gender by keeping it fixed, lung cancer was occurred more often on smokers than on non-smokers.

Overall, 9.8743% of smokers and 10.2362% of non-smokers developed lung cancer. Ignoring subjects' gender, lung cancer was occurred more often on non-smokers than on smokers. The association reverses direction compared to the partial tables. Therefore, it can be misleading to analyze only marginal tables of a multi-way contingency table as this example illustrates.

The result that a marginal association can have a different direction from each conditional association is called *Simpson's paradox.* It applies to quantitative as well as categorical variables.

2.8.2 Conditional and Marginal Odds Ratios

Odds ratios can describe *marginal* and *conditional* associations. Consider a $2 \times 2 \times K$ tables, where K denotes the number of categories of a control variable, Z. Within a fixed category k of Z, the odds ratio

$$\theta_{11(k)} = \frac{N_{11k}N_{22k}}{N_{12k}N_{21k}} = \frac{\pi_{11k}\pi_{22k}}{\pi_{12k}\pi_{21k}}$$

describes conditional XY association in partial table k. The odds ratios for the K partial tables are called XY *conditional* odds ratios. These can be quite different from marginal odds ratios. The XY *marginal* odds ratio is

$$\theta_{11} = \frac{N_{11+}N_{22+}}{N_{12+}N_{21+}} = \frac{\pi_{11+}\pi_{22+}}{\pi_{12+}\pi_{21+}}.$$

Example 2.21. The conditional odds ratios for males and females in example 2.20 are

$$\hat{\theta}_{(1)} = \frac{45(102)}{13(100)} = 3.53 \text{ and } \hat{\theta}_{(2)} = \frac{10(12)}{0(402)} \approx \infty,$$

respectively. The risk of developing lung cancer for male smokers is 3.53 times higher than that of non-smokers. Yet within each gender category, those odds were smaller for non-smokers. Whereas the marginal odds ratio

$$\hat{\theta} = \frac{55(114)}{13(502)} = 0.96$$

indicates the risk of developing lung cancer is 4% lower for smokers than for non-smokers. This reversal in the association after controlling for gender illustrates Simpson's paradox.

For an $I \times J \times K$ table, in general, the *conditional* odds ratio within a fixed category k of Z is given by

$$\theta_{ij(k)} = \frac{N_{ijk}N_{i+1,j+1,k}}{N_{i,j+1,k}N_{i+1,j,k}} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\pi_{i+1,j,k}};\ i = 1, 2, \cdots, I-1,\ j = 1, 2, \cdots, J-1.$$

2.8.3 Marginal and Conditional Independence

An $I \times J \times K$ table describes the relationship between X and Y, controlling for Z. If X and Y are independent in partial table k, then X and Y are called *conditionally independent* at level k of Z. For a response Y, this means $P(Y = j|X = i, Z = k) = P(Y = j|Z = k)$ for all i and j. More generally, X and Y are said to be *conditionally independent* given Z when they are *conditionally independent* at every level of Z, that is, when the above equation holds for all k. Then, given Z, Y does not depend on X. In other words, *conditional independence* is equivalent to

$$\pi_{ijk} = \frac{\pi_{i+k}\pi_{+jk}}{\pi_{++k}}, \quad \text{for all } i, j \text{ and } k.$$

But, *conditional independence* does not imply *marginal independence.*

An $I \times J \times K$ table has homogeneous XY association when $\theta_{ij(1)} = \theta_{ij(2)} = \cdots = \theta_{ij(K)}$ for all i and j. Then, the effect of X on Y is the same at each category of Z. Conditional independence of X and Y is the special case in which each $\theta_{ij(k)} = 1$.

Under *homogeneous XY* association, homogeneity also holds for the other associations. For instance, the *conditional* odds ratio between two categories of X and two categories of Z is identical at each category of Y. For the odds ratio, *homogeneous association* is a symmetric property. It applies to any pair of variables viewed across the categories of the third. When it occurs, there is said to be no interaction between two variables in their effects on the other variable. When interaction exists, the *conditional* odds ratio for any pair of variables changes across categories of the third.

Example 2.22. For the lung cancer data on example 2.20, $\hat{\theta}_{(1)} = 3.53$ and $\hat{\theta}_{(2)} = \infty$. The values are not close, but the second estimate is unstable because of the zero cell count. Adding 0.1 to each cell count, $\hat{\theta}_{(2)} = 3.04$. Because $\hat{\theta}_{(2)}$ is unstable and because further variation occurs from sampling variability, these partial tables do not necessarily contradict homogeneous association in a population.

2.9 Chi-square Test of Homogeneity

The chi-square test can be used to test the equality (homogeneity) of population proportions for three or more groups. In this case, samples are selected from, say J, different groups and the interest is whether or not the proportion of a certain characteristic is the same for each population. Thus, the null hypothesis to be tested is $H_0 : \pi_1 = \pi_2 = \cdots = \pi_J$. Under H_0, the usual chi-square and likelihood ratio test statistics are used. Rejecting of the null hypothesis means at least one of the proportions is significantly different from the others.

Example 2.23. A researcher took a random sample of 293 students from five departments (53 from department A, 65 from department B, 50 from department C, 65 from department D, and 60 from department E) of a certain university to determine if they passed a statistics course (Yes, No). The cross-tabulated data is as shown below.

	Department					
Pass	A	B	C	D	E	Total
Yes	50	60	49	56	20	235
No	3	5	1	9	40	58
Total	53	65	50	65	60	293

Is there sufficient evidence to reject the hypothesis that the proportion of passing statistics course is the same among the five departments?

Solution: The null hypothesis to be tested hear is $H_0 : \pi_A = \pi_B = \pi_C = \pi_D = \pi_E$. Of the entire 293 students, since 235 of them passed the course, the overall sample proportion of passing students is $235/293 = 0.802$. If $H_0 : \pi_A = \pi_B = \pi_C = \pi_D = \pi_E$, is true, the best estimate of the passing proportion is 0.802. Therefore, the expected number of passing students in department A, B, C, D and E are $\hat{\mu}_A = 53(0.802) = 42.506$, $\hat{\mu}_B = 65(0.802) = 52.130$, $\hat{\mu}_C = 50(0.802) = 40.100$, $\hat{\mu}_D = 65(0.802) = 52.130$ and $\hat{\mu}_E = 60(0.802) = 48.120$, respectively. Note that these expected values are the usual ones.

	Department					
Pass	A	B	C	D	E	Total
Yes	50 (42.506)	60 (52.130)	49 (40.100)	56 (52.130)	20 (48.120)	235
No	3 (10.491)	5 (12.867)	1 (9.898)	9 (12.867)	40 (11.877)	58
Total	53	65	50	65	60	293

Thus, the values of both test statistics are $X^2 = 107.113$ and $G^2 = 94.786$. Since $\chi^2_{0.05}(4) = 2.1318$, the null hypothesis of homogeneous (equal) passing proportions in statistics course among the five departments is rejected. This means the proportion of passing the course in at least one department is significantly different from the others.

2.10 Chi-square Test of Goodness-of-fit

Again, both the chi-square and likelihood-ratio tests are also used for addressing the question of whether a certain data follow any pattern, or fit a specified (assumed) probability distribution such as the binomial or multinomial. In such case, the observed frequencies are compared the expected frequencies of the probability distribution of interest. These tests are called *goodness-of-fit* tests.

Multinomial (Binomial) Probability Distribution

Suppose a sample of n subjects are classified based on a multinomial variable with J categories in which n_j of them are in category j; $j = 1, 2, \cdots, J$ of the variable. Consider the null hypothesis $H_0 : \pi_j = \pi_{j0}$; $j = 1, 2, \cdots, J$ provided that $\sum_{i=1}^{J} \pi_j = 1$. This null hypothesis states that the population follows a multinomial distribution with the specified probabilities π_{10}, π_{20}, $\cdots$, and π_{J0} of the J categories. Then, the alternative one states the population does not follow a multinomial distribution with the specified probabilities.

Under H_0, the expected values of $\{n_j\}$ are $\mu_j = n\pi_{j0}$; $j = 1, 2, \cdots, J$. Thus, the Pearson chi-squared and likelihood-ratio statistics are

$$X^2 = \sum_{j=1}^{J} \frac{(n_j - \mu_j)^2}{\mu_j} \sim \chi^2(J-1) \text{ and } G^2 = 2\sum_{j=1}^{J} n_j \log\left(\frac{n_j}{\mu_j}\right) \sim \chi^2(J-1).$$

For fixed J, as n increases the distribution of Pearson X^2 usually converges to chi-squared more quickly than that of G^2. The chi-squared approximation is usually poor for G^2 when $n/J < 5$.

Example 2.24. Among its many applications, Pearson's test was used in genetics to test Mendel's theories of natural inheritance. Mendel crossed pea plants of pure yellow strain with plants of pure green strain. He predicted that second-generation hybrid seeds would be 75% yellow and 25% green, yellow being the dominant strain. One experiment produced $n = 8023$ seeds, of which $n_1 = 6022$ were yellow and $n_2 = 2001$ were green. Test Mendel's hypothesis using both the Pearson and likelihood-ratio tests.

Solution: The hypothesis to be tested is $H_0 : \pi_{10} = 0.75$, $\pi_{20} = 0.25$ {that is, the population proportions of second-generation hybrid seeds are 75% yellow and 25% green (multinomial); or the population proportion second-generation hybrid seeds are 75% yellow (binomial), or the population proportion second-generation hybrid seeds are 25% green (binomial)}.

Thus, $\mu_1 = n\pi_{10} = 6017.25$ and $\mu_2 = n\pi_{20} = 2005.75$. Both the Pearson X^2 and likelihood-ratio G^2 tests have values 0.015 which is less than $\chi^2_{0.05}(1) = 3.84$. Hence, the experiment does not contradict Mendel's hypothesis which means second-generation hybrid seeds will not be significantly different from 75% yellow and 25% green.

Consider the null hypothesis that cell probabilities in two-way tables equal to certain specified values π_{ij0}; $i = 1, 2, \cdots, I$ and $j = 1, 2, \cdots, J$. For a sample of n observations with cell counts $\{n_{ij}\}$, the expected frequencies are $\{\mu_{ij} = n\pi_{ij0}\}$ when H_0 is true. This notation refers to two-way tables, but similar notions apply to a set of counts for multi-way tables. Consequently, the Pearson chi-squared and likelihood-ratio statistics for a two-way table are:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}} \sim \chi^2(IJ-1) \text{ and } G^2 = 2\sum_{i=1}^{I}\sum_{j=1}^{J} n_{ij} \log\left(\frac{n_{ij}}{\mu_{ij}}\right) \sim \chi^2(IJ-1).$$

respectively. As said previously, the chi-squared approximation is also poor hear for G^2 when $n/IJ < 5$.

Chapter 3

Logistic Regression

3.1 Objective and Learning Outcomes

In a linear regression model, it is implicitly assumed that the response variable is continuous following a normal distribution. There are also cases where the response variable is categorical (binary, multinomial, ordinal or count) in nature. This chapter deals with the case where the response variable is binary with outcomes, say, success and failure. Therefore, a statistical modeling approach used to describe the relationship of such a binary response variable to one or more explanatory variable(s) is called *logistic* regression.

Upon completion of this chapter, students are expected to:

- Understand why the usual linear regression model is not appropriate for a categorical response variable.
- Fit a logistic regression and interpret the parameter estimates in terms of odds ratio.
- Conduct inferences about the overall significance of the model and the individual parameters.
- Construct confidence intervals for the parameters, odds ratio and probability of success.

3.2 Binary Logistic Regression

A binary logistic regression predicts the probability of success in a dichotomous dependent variable, for example, whether a person will develop a disease or whether a certain patient will survive a surgical procedure. There could be one or more independent variables which can be, as usual, either continuous, categorical or both.

3.2.1 Why Not the Linear Regression?

Consider a regression model with a binary response variable $y_i = \alpha + \beta x_i + \varepsilon_i$ where x_i is the study hours per day of a student, and $y_i = 1$ if the student passed Statistics course and $y_i = 0$ if the student failed the course.

Let $\pi(x_i)$ denote the conditional probability that the student will pass Statistics course given the study hours, that is, $P(Y_i = 1|X_i = x_i)$ where $0 \leq \pi(x_i) \leq 1$. Then, the above model can be written as $\pi(x_i) = \alpha + \beta x_i + \varepsilon_i$ which looks like a typical linear regression model. Since the response variable is binary, it is called *linear probability model* (LPM).

It would seem the usual least squares estimation can be applied, but, it poses several problems. Obviously, since the response variable Y_i takes the value 1 with probability $\pi(x_i)$ and 0 with probability $1 - \pi(x_i)$, the basic random variable has a point-binomial or Bernoulli probability distribution, $P(Y_i = y_i) = \pi(x_i)^{y_i}[1 - \pi(x_i)]^{1-y_i}$; $y_i = 0, 1$. Therefore, the assumption of normality for ε_i is not fulfilled for a linear probability model. Because $\varepsilon_i = y_i - \alpha - \beta x_i$, like y_i, the disturbance ε_i also takes only two values; that is, it takes the value $1 - (\alpha + \beta x_i)$ with probability $\pi(x_i)$ and the value $-(\alpha + \beta x_i)$ with probability $1 - \pi(x_i)$. Hence, the errors follow the Bernoulli distribution.

y_i	ε_i	Probability
1	$1 - \alpha - \beta x_i$	$\pi(x_i)$
0	$-\alpha - \beta x_i$	$1 - \pi(x_i)$
Total		1

In fact, the nonfulfillment of the normality assumption may not be so critical because the least squares estimation does not require the disturbances to be normally distributed, the least squares point estimates still remain unbiased. The errors are assumed to be normally distributed for the purpose of statistical inference, but this assumption is not necessary if the objective is point estimation. Besides, as the sample size increases indefinitely, statistical theory shows that the least squares estimators tend to be normally distributed generally.

Another problem of least squares is that the errors are not homoscedastic. This is, however, not surprising. For the distribution of the error term, applying the definition of variance for a Bernoulli distribution, $var(\varepsilon_i) = \pi(x_i)[1 - \pi(x_i)]$. Therefore, the variance of ε_i ultimately depends on the values of x_i. Hence, the error variance is heteroscedastic (not homoscedastic). It is known, in the presence of heteroscedasticity, least squares estimators, although unbiased, they are not efficient, that is, they do not have minimum variance. But the problem of heteroscedasticity, like the problem of nonnormality, is not insurmountable.

The real problem with the least squares estimation of the linear probability model is that it may predict impossible values (negative values or values larger than 1). There is no guarantee that $\pi(x_i)$ will necessarily fulfill the restriction $0 \leq \pi(x_i) \leq 1$. Due to these problems, the linear probability model is not appropriate for modeling a binary response variable.

3.2.2 The Logistic Function

Recall the logistic function is

$$f(z) = \frac{1}{1+\exp(-z)}; \ -\infty < z < \infty.$$

When $z = -\infty$, $f(-\infty) = 0$ and when $z = \infty$, $f(\infty) = 1$. Note also that $f(0) = \frac{1}{2}$.

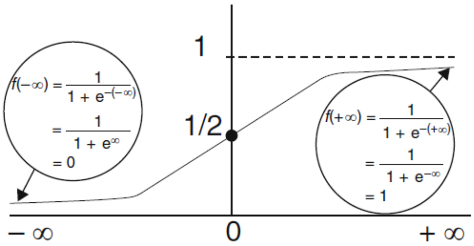

Figure 3.1: Plot of the Logistic Function

Thus, as the figure describes the range of $f(z)$ is between 0 and 1 (that is, $0 \leq f(z) \leq 1$) regardless of the value of z. Therefore, it is suitable for use as a probability model. Hence, to indicate that $f(z)$ is a probability value, the notation $\pi(z)$ can be used instead. That is,

$$\pi(z) = \frac{1}{1+\exp(-z)}; \ -\infty < z < \infty$$

where $\pi(z) = P(Y = 1|Z = z)$.

3.2.3 The Simple Logistic Regression

To begin with the simplest model, consider the case of a binary outcome and a single predictor variable x. Hence, in the logistic function, z is expressed as a function (mostly linear function) of the explanatory variable. That is, $z_i = g(x_i) = \alpha + \beta x_i$. As a result, the simple logistic probability model is:

$$\pi(x_i) = \frac{1}{1+\exp[-(\alpha+\beta x_i)]}$$

where $\pi(x_i) = P(Y_i = 1|X_i = x_i) = 1 - P(Y_i = 0|X_i = x_i)$. It can also be written as

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

As can be seen from this model, the relationship between the response variable (probability of success) and the explanatory variable is not linear. However, it can be linearized by using different transformations of the probability of success and the most common one is called a *logit* or *log-odds* transformation.

The Logit Transformation

In the previous chapter, odds is defined as the ratio of the probability of success to the probability of failure. Hence, the odds of successes at a particular value x_i of the explanatory variable is

$$\Omega(x_i) = \frac{\pi(x_i)}{1 - \pi(x_i)}.$$

Thus, the odds of successes for a simple logistic regression model is $\Omega(x_i) = \exp(\alpha + \beta x_i)$. If $\Omega(x_i) = 1$, then a success is as likely as a failure at the particular value x_i of the explanatory variable. If $\Omega(x_i) > 1$, then $\log \Omega(x_i) > 0$, a success is more likely to occur than a failure. On the other hand, if $\Omega(x_i) < 1$, then $\log \Omega(x_i) < 0$, a success is less likely than a failure.

The *logit* of the probability of success is given by the natural logarithm of the odds of successes. Therefore, the logit of the probability of success is a linear function of the explanatory variable. Thus, the simple logistic model is

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

This is particulary called the *logit* model as it uses the *logit* transformation or the *log-odds scaling* (or *logit link function*) which is a reasonable choice for binary response models.

To clarify the relationship between probabilities, odds, and the natural log of the odds (logit), the following table includes probability values along with their corresponding odds as well as the natural log of the odds, log(odds). The table demonstrates that as the probability gets smaller and approaches 0, the odds also approach 0 while the log odds approach $-\infty$(negative infinity), and as the probability gets larger and approaches 1, the odds also get larger while the log odds approach $+\infty$(positive infinity). Therefore, while probabilities can theoretically vary from 0 to 1 with a midpoint of 0.5, the corresponding odds can theoretically vary from 0 to $+\infty$ with 1 corresponding to the probability midpoint, and the natural log of the odds can theoretically vary from $+\infty$ to $+\infty$ with 0 corresponding to the probability midpoint.

$\pi(x_i)$	$1-\pi(x_i)$	$\Omega(x_i)$	logit $\pi(x_i)$
0.001	0.999	0.001	-6.908
0.010	0.990	0.010	-4.605
0.100	0.900	0.111	-2.198
0.200	0.800	0.250	-1.386
0.300	0.700	0.429	-0.846
0.400	0.600	0.667	-0.405
0.500	0.500	1.000	0.000
0.600	0.400	1.500	0.405
0.700	0.300	2.333	0.847
0.800	0.200	4.000	1.386
0.900	0.100	9.000	2.197
0.990	0.010	99.000	4.595
0.999	0.001	999.000	6.907

Thus, the range of the log(odds) more closely resembles the standard normal distribution in that it is unbounded, has a midpoint of 0, and is symmetric around the midpoint.

There are also other models that are used in practice. The probit model or the complementary log-log model might be appropriate when the logit model does not fit the data well.

Interpretation of the Parameters

The parameters, α and β, are the intercept and slope of the logit model, respectively. Because the predicted value, probability, in logistic regression is different from the predicted value, mean, in linear regression, the interpretations of the intercept, α, and slope, β, are also somewhat different as these must be interpreted in the context of the predicted response.

The logit model is monotone depending on the *sign* of the parameter β. Its sign determines whether the probability of success is increasing or decreasing, as shown in figure 3.2, when the value of the explanatory variable increases. When the parameter β is zero, Y is independent of X. Then, $\pi(x_i) = \frac{\exp(\alpha)}{1+\exp(\alpha)}$ which is identical for all x_i, so the curve becomes a straight (horizontal) line.

The slope parameter of a logit model can be interpreted in terms of an odds ratio. From logit $\pi(x_i) = \alpha + \beta x_i$, an odds is an exponential function of x_i. This provides a basic interpretation for the *magnitude* of the slope parameter β. The odds at x_i is $\Omega(x_i) = \exp(\alpha + \beta x_i)$ and the odds at $x_i + 1$ is $\Omega(x_i + 1) = \exp[\alpha + \beta(x_i + 1)]$. Thus, the odds ratio is

$$\theta = \frac{\Omega(x_i + 1)}{\Omega(x_i)} = \exp(\beta).$$

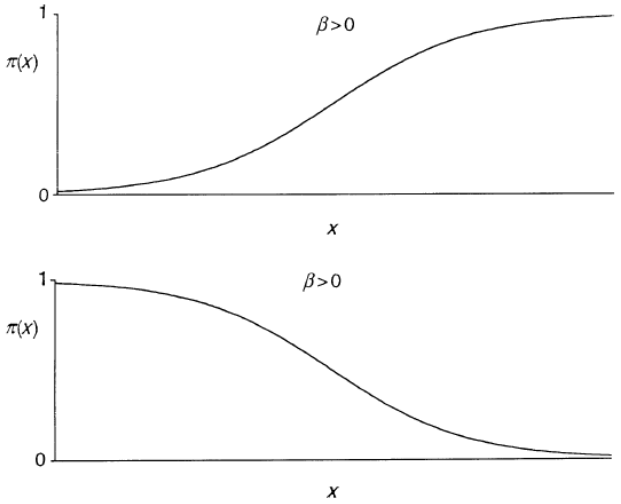

Figure 3.2: Plot of the Logistic Probability

This value is the multiplicative effect of the odds of successes due to a unit change in the explanatory variable. That is, for every one unit increase in x_i, the odds changes by a factor of $\exp(\beta)$. Similarly, for an m units increase in x_i, say $x_i + m$ versus x_i, the corresponding odds ratio becomes $\exp(m\beta)$.

Also, the parameter β determines the *slope* (*rate of change* or *marginal effect*) of the probability of success at a certain value of the explanatory variable. This *rate of change* (*marginal effect*) at a particular x_i value is described by drawing a straight line tangent to the curve at that point. That line will have a *slope* of $\pi(x_i)[1 - \pi(x_i)]\beta$. This is the *rate of change* (*slope* or *marginal effect*) of $\pi(x_i)$ at a particular value of x_i. For example, the line tangent to the curve at x_i for which $\pi(x_i) = 0.5$ has a *slope* $(0.5)(1-0.5)\beta = 0.25\beta$. If $\pi(x_i)$ is 0.9 or 0.1, it has a *marginal effect* 0.09β. As the probability of success approaches either 0 or 1, the *rate of increment* (*decrement*) of the curve approaches to 0. The steepest *slope* of the curve is attained at x_i for which the probability of success is 50%. Thus, solving

$$\frac{1}{1 + \exp[-(\alpha + \beta x_i)]} = 0.5$$

for x_i implies $x_i = -\frac{\alpha}{\beta}$. This x_i value is called *medial effective level* (EL_{50}). At this value, each outcome has a 50% chance of occurring.

The intercept α is, not usually of particular interest, used to obtain the odds (probability) at $x_i = 0$. Also, by centering the explanatory variable at 0 {that is, replacing x_i by $(x_i - \bar{x})$}, α becomes the logit at that mean, and thus $\pi(\bar{x}) = \frac{\exp(\alpha)}{1+\exp(\alpha)}$.

The estimated logistic regression model is written as:

$$\text{logit } \hat{\pi}(x_i) = \log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = \hat{\alpha} + \hat{\beta}x_i.$$

Example 3.1. For studying the effect of age (continuous variable) on the occurrence of hypertension (coded as 1 for presence and 0 for absence), a sample of 13 individuals were examined. The ages (in years) of persons having hypertension are 45, 60, 60, 60, 55, 55, 20 and those who do not have hypertension are 20, 20, 18, 30, 55, 18. For these data, the following parameter estimates were obtained.

Variable	Parameter Estimate
Intercept	-3.4648
Age	0.0931

1. Write the model that allows the prediction of the probability of having hypertension at a given age.
2. What is the estimated probability of having hypertension at the minimum and maximum ages of this study.
3. What is the estimated probability of having hypertension at the age of 35. Also find the odds of having hypertension at this age.
4. Find the estimated probability of success at the sample mean and determine the incremental change (marginal effect) at that point.
5. Write out the estimated logit model.
6. Find the estimated odds ratio of having hypertension and interpret.
7. Determine the estimated median effective level (EL_{50}) and interpret.

Solution: Let Y= hypertension and X= age. Then $\hat{\pi}(x_i) = \widehat{P}(Y=1|x_i)$ is the estimated probability of having hypertension, $Y=1$, given the age x_i of an individual i.

1. The estimated probability of hypertension at a given age is given by:

$$\hat{\pi}(x_i) = \frac{\exp(-3.4648 + 0.0931x_i)}{1+\exp(-3.4648+0.0931x_i)}.$$

2. The estimated probability of having hypertension at the age of 35 years is $\hat{\pi}(35) = 0.4486$ and its estimated odds is $\widehat{\Omega}(35) = 0.8136$.
3. The mean age of the sample is 39.69 years. The estimated probability of having hypertension at this mean age is $\hat{\pi}(39.69) = 0.5573$ and the rate of change (marginal effect) at this mean value is $\hat{\pi}(39.69)[1-\hat{\pi}(39.69)]\hat{\beta} = 0.5573(1-0.5573)(0.0931) = 0.0230$. The probability of having hypertension at the age of 39.69 years increases by 2.30%.

4. The estimated logit model is written as
$$\log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = -3.4648 + 0.0931x_i.$$

5. The estimated odds ratio is $\exp(\hat{\beta}) = \exp(0.0931) = 1.0976$. Hence, the odds (risk) of having hypertension is 1.0976 times larger for every year older an individual is. In other words, as the age of an individual increases by one year, the odds (risk) of developing hypertension increases by a factor of 1.0976. Or the odds (risk) of having hypertension increases by $[\exp(0.0931) - 1] \times 100\% = 9.76\%$ every year.

6. The estimated median effective level, the estimated age in years at which an individual has a 50% chance of having hypertension, is $\widehat{\mathrm{EL}}_{50} = -\hat{\alpha}/\hat{\beta} = -(-3.4648)/0.0931 = 37.2159$.

3.2.4 Logit Models with Categorical Predictors

Like ordinary regression, logistic regression extends to include qualitative explanatory variables, often called *factors*.

Binary Predictors

For simplicity, let us consider a binary predictor, X, representing an exposure which refers to a risk factor such as smoking (smoker, nonsmoker) or patient characteristics like sex (male, female), residence (urban, rural). The simple logit model is
$$\log\left[\frac{\pi(x_i)}{1-\pi(x_i)}\right] = \alpha + \beta x_i \text{ where } x_i = \begin{cases} 1, & \text{exposed group;} \\ 0, & \text{unexposed group.} \end{cases}$$

From this model, the odds in the exposed group is given by $\Omega(1) = \exp(\alpha + \beta)$ and the odds in the unexposed group is $\Omega(0) = \exp(\alpha)$. This implies, $\exp(\beta)$ as the odds ratio associated with an exposure (exposed $x_i = 1$ versus unexposed $x_i = 0$), which is equivalent to the odds ratio in a 2×2 table.

In other words, the estimates of the parameters of a logit model for a 2×2 table can be easily determined from the cell frequencies. Consider the 2×2 table below. Setting $x_i = 0$

	Response		
Exposure	Success (1)	Failure (0)	Total
Exposed (1)	n_{11}	n_{10}	n_{1+}
Unexposed (0)	n_{01}	n_{00}	n_{0+}
Total	n_{+1}	n_{+0}	n

for the unexposed group and then solving for α gives the estimated intercept of the logit model in terms of the natural logarithm of the odds of successes in the unexposed group. That is,

$$\hat{\alpha} = \log\left[\frac{\hat{\pi}(0)}{1-\hat{\pi}(0)}\right] = \log\left(\frac{n_{01}}{n_{00}}\right).$$

Similarly, the estimate of the slope of the logit model is derived as the natural logarithm of the odds ratio associated with an exposure by setting $x_i = 1$ for the exposed group,

$$\hat{\beta} = \log\left[\frac{\hat{\pi}(1)}{1-\hat{\pi}(1)}\right] - \hat{\alpha} = \log\left[\frac{\hat{\pi}(1)}{1-\hat{\pi}(1)}\right] - \left[\frac{\hat{\pi}(0)}{1-\hat{\pi}(0)}\right] = \log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right).$$

As discussed before, the *marginal effect* of a continuous explanatory variable, which is very useful when interpreting a binary logit model, is the partial derivative of the probability of success with respect to that variable.

Similarly, the *discrete change* of a binary explanatory variable is the difference in estimated probabilities when the variable value is 1 and when it is 0. Note that *marginal effects* and *discrete changes* look similar but are not equal in conceptual and numerical senses.

Example 3.2. In a study of cigarette smoking and risk of lung cancer, a logistic regression analysis is used to determine how much greater the odds are finding cases of the diseases among subjects who have ever smoked than among those who have never smoked.

	Lung Cancer		
Smoking	Case (1)	Control (0)	Total
Yes (1)	77	123	200
No (0)	54	171	225
Total	131	294	425

Given the parameter estimates from a statistical software as follows:

Variable	Parameter Estimate
Intercept	-1.1527
Smoking	0.6843

Write out the estimated model and interpret the slope estimate. Also find the discrete change.

Solution: Let Y= lung cancer where

$$y_i = \begin{cases} 1, & \text{if the subject develops lung cancer - Case;} \\ 0, & \text{otherwise (if the subject does not develop lung cancer) - Control.} \end{cases}$$

For the explanatory variable, let X= smoking status where

$$x_i = \begin{cases} 1, & \text{if the subject had ever smoked - Smoker;} \\ 0, & \text{otherwise (if the subject had never smoked) - Nonsmoker.} \end{cases}$$

Thus, $\hat{\pi}(x_i)$ is the estimated probability of developing lung cancer, $Y = 1$, given the smoking status, $x_i = 1$ for smokers and $x_i = 0$ for non-smokers. The parameter estimates can also be obtained manually. The estimates are

$$\hat{\alpha} = \log\left(\frac{n_{01}}{n_{00}}\right) = \log\left(\frac{54}{171}\right) = -1.1527$$

and

$$\hat{\beta} = \log\left(\frac{n_{11}n_{00}}{n_{10}n_{01}}\right) = \log\left[\frac{77(171)}{123(54)}\right] = 0.6843.$$

Thus, the estimated model is

$$\log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = -1.1527 + 0.6843x_i.$$

The estimated odds ratio is $\exp(0.6843) = 1.9824$. Thus, smokers are 1.9824 times (98.24%) more likely to develop lung cancer as compared to nonsmokers. Or the odds (risk) of developing lung cancer is 98.24% higher for smokers than for nonsmokers {the odds (risk) of developing lung cancer among smokers is 98.24% higher than that of among nonsmokers}.

The discrete change is $\hat{\pi}(1) - \hat{\pi}(0) = 0.3850 - 0.2400 = 0.1450$. The probability of developing lung cancer increases by 14.50% for smokers relative to nonsmokers.

Example 3.3. The following table presents the cross-classification of 1464 HIV/AIDS patients involved in Seid *et al.* (2014) study by defaulting (Yes, No) and gender (Female, Male).

	Defaulter		
Gender	Yes (1)	No (0)	Total
Female (1)	189	741	930
Male (0)	142	392	534
Total	331	1133	1464

The parameter estimates are provided in the following table:

Variable	Parameter Estimate
Intercept	-1.0154
Smoking	-0.3508

Write out the estimated model and interpret the estimated slope.

Solution: Let Y= defaulter where $y_i = 1$ if the patient was defaulted from the HAART treatment and $y_i = 0$ otherwise (if the patient was active on the treatment). Let X= gender of the patient where $x_i = 1$ if the patient is female and $x_i = 0$ otherwise (if the patient is male).

Then $\hat{\pi}(x_i)$ is the estimated probability of the patient being defaulted from the HAART treatment. The estimated model is

$$\log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = -1.0154 - 0.3508x_i.$$

The odds ratio is $\exp(-0.3508) = 0.7041$. This means that female patients are 0.7041 times (29.59%) less likely to default from HAART treatment as compared to male patients. Or, the risk of being defaulted is 29.59% lower for female patients than for male patients (the risk of being defaulted for male patients is 42.02% higher than the risk of being defaulted for female patients).

Polytomous Explanatory Variables

If there is a categorical explanatory variable with more than two categories, then it is inappropriate to include it in the model as if it was quantitative. This is because the codes used to represent the various categories are merely identifiers and have no numeric significance. In such case, a set of binary variables, called design (dummy, indicator) variables, should be created to represent such a polytomuous variable.

Suppose, for example, that one of the explanatory variable is marital status with three categories: "Single", "Married", "Separated". In this case, taking one of the categories as a reference (comparison group), two design variables (d_1 and d_2) are required to represent marital status in a regression model. For example, if the category "Single" is taken as a reference, the two design variables, d_1 and d_2 are set to 0; when the subject is "Married", d_1 is set to 1 while d_2 is still 0; when the marital status of the subject is "Separated", $d_1 = 0$ and $d_2 = 1$ are used. The following table shows this example of design variables for marital status:

	Design Variables	
Marital Status	Married (d_1)	Separated (d_2)
Single	0	0
Married	1	0
Separated	0	1

In general, if a polytomuous variable X has m categories, then $m-1$ design variables are needed. The $m-1$ design variables are denoted as d_u and the coefficients of those design variables are denoted as β_u, $u = 1, 2, \cdots, m-1$. Thus, the logit model would be:

$$\text{logit } \hat{\pi}(x_i) = \log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = \alpha + \beta_1 d_{i1} + \beta_2 d_{i2} + \cdots + \beta_{m-1} d_{i,m-1}.$$

Therefore, when there is a binary response variable and a polytomous explanatory variable, the data can be presented using a $2\times m$ table. Taking one of the category of the explanatory variable as a reference, $m-1$ stratified 2×2 tables can be constructed. Then the parameter estimates corresponding to each design variable can be easily determined from each table. If category m is taken as a reference, then $\hat{\alpha} = \log\left(\frac{n_{m1}}{n_{m0}}\right)$ and $\hat{\beta}_u = \log\left(\frac{n_{u1}n_{m0}}{n_{u0}n_{m1}}\right)$; $u = 1, 2, \cdots, m-1$.

Example 3.4. Given the following cross-classified data on race and coronary heart disease for 100 subjects.

	Race				
CHD	White	Black	Hispanic	Other	Total
Present (1)	5	20	15	10	50
Absent (0)	20	10	10	10	50
Total	25	30	25	20	100

Software provides the following parameter estimates.

Variable	Parameter Estimate
Intercept	-1.386
Black (d_1)	2.079
Hispanic (d_2)	1.792
Other (d_3)	1.386

Specify the design variables for race using ”white” as a reference group. Calculate the parameter estimates manually from the cell counts of the contingency table and compare them with the software estimates. Write out the estimated model and interpret.

Solution: Since the variable ”Race” has four categories, three design variables are needed.

	Design Variables		
Race	Black (d_1)	Hispanic (d_2)	Other (d_3)
White	0	0	0
Black	1	0	0
Hispanic	0	1	0
Other	0	0	1

Let $\hat{\pi}(x_i)$ be the estimated probability of developing coronary heart disease given the race of an individual. Thus,

$$\log\left[\frac{\hat{\pi}(x_i)}{1-\hat{\pi}(x_i)}\right] = -1.386 + 2.079d_{i1} + 1.792d_{i2} + 1.386d_{i3}.$$

Blacks are about 8 $\{\exp(2.079) = 7.996\}$ times more likely to develop coronary heart disease as compared to whites. Similarly, the odds (risk) of coronary heart disease for hispanics is about 6 $\{\exp(1.792) = 6.001\}$ times that of whites. The odds (risk) of coronary heart disease for other (neither blacks nor hispanics) races is about 4 $\{\exp(1.386) = 3.999\}$ times that of whites.

3.2.5 Multiple Logistic Regression

Suppose there are k explanatory variables (categorical, continuous or both) to be considered simultaneously. Then, the multiple logit model is written as:

$$\text{logit } \pi(\boldsymbol{x}_i) = \log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}.$$

Similar to the simple logistic regression, $\exp(\beta_j)$ represents the (partial) odds ratio associated with an exposure if X_j is binary (exposed $x_{ij} = 1$ versus unexposed $x_{ij} = 0$); or it is the odds ratio due to a unit increase if X_j is continuous ($x_{ij} = x_{ij}+1$ versus $x_{ij} = x_{ij}$).

If the j^{th} explanatory variable, X_j, has m_j levels, then the multiple logit model with k variables would be

$$\log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{j-1} x_{i,j-1} + \sum_{u=1}^{m_j-1} \beta_{ju} d_{iju} + \beta_{j+1} x_{i,j+1} + \cdots + \beta_k x_{ik}$$

where the d_{ju}'s are the $m_j - 1$ design variables and $\beta_{ju}, u = 1, 2, \cdots, m_j - 1$ are their corresponding parameters.

Note: Odd ratios obtained from a simple logistic regression (one independent variable) are called *crude odds ratios* (COR) and odd ratios obtained from a multiple logistic regression (two or more independent variables) are called *adjusted odds ratios* (AOR).

Example 3.5. To determine the effect of vision status (1=vision problem, 0=no vision problem) and driver education (1=took driver education, 0=did not take driver education) of a driver on car accident (did the subject had an accident in the past year?), the following parameter estimates are obtained from a sample of 210 individuals. Interpret the results.

Variable	Parameter Estimate
Intercept	0.1110
Vision	1.7139
Education	-1.5001

Solution: Let Y= car accident ($y_i = 1$ if a subject had an accident in the past year and $y_i = 0$ if a subject had not an accident in the past year). Let X_1= vision problem ($x_{i1} = 1$ if a subject had a vision problem and $x_{i1} = 0$ if a subject had not a vision problem). Let X_2= driver education ($x_{i2} = 1$ if a subject took driver education, $x_{i2} = 0$ if a subject did not take driver education).

The estimated logit model is $\log\left[\frac{\hat{\pi}(\boldsymbol{x}_i)}{1-\hat{\pi}(\boldsymbol{x}_i)}\right] = 0.1110 + 1.7139x_{i1} - 1.5001x_{i2}$. The estimated odds ratio associated with vision problem is $\exp(1.7139) = 5.551$. The odds of having accident for a person with vision problem is 5.551 times that of a person with no vision problem assuming driver education the same. In other words, drivers who have vision problem are 5.551 times more likely to have an accident as compared to those with no vision problem.

Also, the estimated odds ratio associated with education problem is $\exp(-1.5001) = 0.223$. Drivers who took driving education are 0.223 times less likely to have an accident as compared to those who did not take driving education assuming the same vision status, that is, the risk of having an accident for those who took a driving education is 77.7% lower than those who did not take a driving education.

3.3 Statistical Inference

Recall the binary response probability given the values of the explanatory variables is

$$\pi(\boldsymbol{x}_i) = \frac{\exp(\sum\limits_{j=0}^{k} \beta_j x_{ij})}{1 + \exp(\sum\limits_{j=0}^{k} \beta_j x_{ij})} \tag{3.1}$$

where $x_{i0} = 1$ for all $i = 1, 2, \cdots, n$. Equivalently using the logit transformation, it can be written as

$$\log\left[\frac{\pi(\boldsymbol{x}_i)}{1 - \pi(\boldsymbol{x}_i)}\right] = \sum_{j=0}^{k} \beta_j x_{ij}. \tag{3.2}$$

3.3.1 Parameter Estimation

The goal of logistic regression model is to estimate the $k + 1$ unknown parameters of the model. This is done with maximum likelihood estimation which entails finding the set of parameters for which the probability of the observed data is largest.

Given a data set with n independent observations. Suppose these responses are grouped into m unique covariate patterns (called populations). Then each binary response Y_i; $i = 1, 2, \cdots, m$ has an independent Binomial distribution with parameter n_i and $\pi(\boldsymbol{x}_i)$, that is,

$$P(Y_i = y_i) = \binom{n_i}{y_i} \pi(\boldsymbol{x}_i)^{y_i} [1 - \pi(\boldsymbol{x}_i)]^{n_i - y_i};\ y_i = 0, 1, 2, \cdots, n_i$$

where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$ for population i and $\sum\limits_{i=1}^{m} n_i = n$. Then, the joint probability mass function of the vector of m Binomial random variables, $\boldsymbol{Y}^t = (Y_1, Y_2, \cdots, Y_m)$, is the product of the m Binomial distributions

$$P(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{m} \binom{n_i}{y_i} \pi(\boldsymbol{x}_i)^{y_i} [1 - \pi(\boldsymbol{x}_i)]^{n_i - y_i}. \tag{3.3}$$

The joint probability mass function in equation (3.3) expresses the values of $\boldsymbol{y}$ as a function of known, fixed values for $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots, \beta_k)^t$. The likelihood function has the same form as the probability mass function, except that it expresses the values of $\boldsymbol{\beta}$ in terms of known, fixed values for $\boldsymbol{y}$. Thus,

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{m} \binom{n_i}{y_i} \pi(\boldsymbol{x}_i)^{y_i}[1-\pi(\boldsymbol{x}_i)]^{n_i-y_i} \tag{3.4}$$

Note that the combination term does not contain any of the $\pi(\boldsymbol{x}_i)$. As a result, it is essentially constant that can be ignored: maximizing the equation without the combination term will come to the same result as if it was included. Therefore, equation (3.4) can be written as:

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{m} \pi(\boldsymbol{x}_i)^{y_i}[1-\pi(\boldsymbol{x}_i)]^{n_i-y_i} \tag{3.5}$$

and it can be re-arranged as:

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{m} \left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right]^{y_i} [1-\pi(\boldsymbol{x}_i)]^{n_i} \tag{3.6}$$

By substituting the odds of successes and probability of failure in equation (3.6), the likelihood function becomes

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{m} \left[\exp\left(y_i \sum_{j=0}^{k} \beta_j x_{ij}\right)\right] \left[1+\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)\right]^{-n_i} \tag{3.7}$$

Since the logarithm is a monotonic function, any maximum of the likelihood function will also be a maximum of the log-likelihood function and vice versa. Thus, taking the natural logarithm of equation (3.7) gives the log-likelihood function:

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{m} \left\{ y_i \sum_{j=0}^{k} \beta_j x_{ij} - n_i \log\left[1+\exp\left(\sum_{j=0}^{k} \beta_j x_{ij}\right)\right]\right\} \tag{3.8}$$

To find the critical points of the log-likelihood function, first, equation (3.8) should be partially differentiated with respect to each $\beta_j; j = 0, 1, \cdots, k$ which results in a system of $k+1$ nonlinear equations with the $k+1$ unknown parameters as shown in equation (3.9) below:

$$\frac{\partial L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{m} [y_i x_{ij} - n_i \pi(\boldsymbol{x}_i) x_{ij}] = \sum_{i=1}^{m} [y_i - n_i \pi(\boldsymbol{x}_i)] x_{ij}; \quad j = 0, 1, 2, \cdots, k. \tag{3.9}$$

The maximum likelihood estimates for $\boldsymbol{\beta}$ can be, then, found by setting each of the $k+1$ equation equal to zero and solving for each β_j. Since the second partial derivatives of the log-likelihood function:

$$\frac{\partial^2 L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \beta_j \partial \beta_h} = -\sum_{i=1}^{m} n_i \pi(\boldsymbol{x}_i)[1-\pi(\boldsymbol{x}_i)] x_{ij} x_{ih}; \quad j, h = 0, 1, 2, \cdots, k \tag{3.10}$$

is negative semidefinite, the log-likelihood is a concave function of the parameter $\boldsymbol{\beta}$. In addition, equation (3.10) represents the variance-covariance matrix of the parameter estimates which is a function of $var(Y_i) = n_i\pi(\boldsymbol{x}_i)[1 - \pi(\boldsymbol{x}_i)]$.

These equations do not have a closed form solution. Several optimization techniques are available for finding the maximizing estimates of the parameters. Of these, the Newton-Raphson method is the one which is commonly used.

3.3.2 Overall Significance of the Model

Once a logistic regression model is estimated, the next task is to answer the question "Does the entire set of explanatory variables contribute significantly to the prediction of the response?". In this case, two models are to be fitted; one with all explanatory variables (full model) and the other with no explanatory variable (null model).

Likelihood-Ratio/Deviance Test

If the model has k explanatory variables (either binary or continuous), the null hypothesis of no contribution of all the k explanatory variables is $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$. Let ℓ_0 denote the maximized value of the likelihood function of the null model which has only one parameter, that is, the intercept. That is, $\ell_0 = \ell(\hat{\beta}_0)$. Also let ℓ_M denote the maximized value of the likelihood function of the model M with all explanatory variables (having $k+1$ parameters). Here, $\ell_M = \ell(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)$.

Then, the likelihood-ratio test statistic is $G^2 = -2\log(\ell_0/\ell_M) = -2(\log\ell_0 - \log\ell_M) \sim \chi^2(k)$. Deviance is -2 times the log-likelihood value of a model. Thus, $G^2 = D_0 - D_M \sim \chi^2(k)$.

Rejection of the null hypothesis, has an interpretation analogous to that in multiple linear regression using F test, indicates at least one of the k parameters is significantly different from zero.

Example 3.6. Suppose, a study was conducted with the objective of identifying the risk factors associated with HIV/AIDS HAART treatment defaulter patients. Of 1464 patients, 331 were defaulted and the remaining 1133 were actively following the treatment. Five variables which were considered as explanatory variables are age in years (Age), weight in kilograms (Weight), Gender (0=Female, 1=Male), Functional Status (0=Working, 1=Ambulatory, 2=Bedridden) and number of baseline CD4 counts (CD4). The parameter estimates and their corresponding standard errors are presented in the following table.

Variable	Parameter Estimate	Standard Error
Intercept	-0.3120	0.4299
Age	-0.0282	0.0080
Weight	-0.0051	0.0071
Gender	0.5372	0.1438
Ambulatory	0.4959	0.1448
Bedridden	1.2610	0.2882
Working	Ref.	
CD4	-0.0007	0.0004

The log-likelihood value of the null model is -782.5257 and the log-likelihood value of the full model is -753.2892. Test the significance of the entire five variables altogether.

Solution: The response variable takes the value $y_i = 1$ if the patient was defaulted and $y_i = 0$ otherwise (if the patient was on the treatment).

The design variables for Functional Status are:

	Design Variables	
Functional Status	Ambulatory (d_{41})	Bedridden (d_{42})
Working	0	0
Ambulatory	1	0
Bedridden	0	1

Now the model can be written as

$$\log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] = \beta_0 + \beta_1 \text{ Age}_i + \beta_2 \text{ Weight}_i + \beta_3 \text{ Gender}_i$$
$$+ \beta_{41} \text{ Ambulatory}_i + \beta_{42} \text{ Bedridden}_i + \beta_5 \text{ CD4}_i$$

The null hypothesis to be tested is $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_{41} = \beta_{42} = \beta_5 = 0$. The test statistic value is $G^2 = -2(\log \ell_0 - \log \ell_M) = -2[-782.5257 - (-753.2892)] = 58.473$ which is greater than $\chi^2_{0.05}(6) = 12.592$. Therefore, H_0 should be rejected. At least one of the parameter is significantly different from zero.

3.3.3 Significance Test for Parameters

Once the null hypothesis of no contribution of all the explanatory variables to the model is rejected, there is a need to look at which of the variables are significant and which are not. The Wald test is used to identify the statistical significance of each coefficient (β_j) of the logit model. That is, it is used to test the null hypothesis $H_0 : \beta_j = 0$ which states that factor X_j does not have significant value added to the prediction of the response given that other factors are already included in the model. The test statistic for large sample size is, therefore,

$$Z_j = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \sim N(0,1).$$

Example 3.7. Recall example 3.6. Write out the estimated model and identify the significant explanatory variables using Wald test, and interpret the results.

Solution: We have that the estimated model is:

$$\begin{aligned}\log\left[\frac{\pi(\boldsymbol{x}_i)}{1-\pi(\boldsymbol{x}_i)}\right] =& -0.3120 - 0.0282 \text{ Age}_i - 0.0051 \text{ Weight}_i + 0.5372 \text{ Gender}_i \\ &+ 0.4959 \text{ Ambulatory}_i + 1.2610 \text{ Bedridden}_i - 0.0007 \text{ CD4}_i\end{aligned}$$

The Wald test help us to identify those parameters which are responsible for rejection of the null hypothesis of all the parameters are zero. The value of the Wald test for each parameter which is obtained by dividing each parameter estimate by the corresponding standard error estimate is given in the following table.

Variable	Parameter Estimate	Standard Error	Wald Test
Intercept	-0.3120	0.4299	-0.7258
Age	-0.0282	0.0080	-3.5250*
Weight	-0.0051	0.0071	-0.7183
Gender	0.5372	0.1438	3.7357*
Ambulatory	0.4959	0.1448	3.4247*
Bedridden	1.2610	0.2882	4.3754*
Working	Ref.		
CD4	-0.0007	0.0004	-1.7500

As it can be seen from this table, age, gender and functional status (since both of the design variables are significant) are significant at 5% level of significance. When the age of the patient increases by one year, the odds of being defaulted decreases by a factor of $\exp(-0.0282) = 0.9723$ assuming all other variables are same. Also, males are $\exp(0.5372) = 1.7112$ times more likely to default than females, that is, the odds of being defaulted for males is 71.12% higher than that of females assuming the other variables constant. Again, assuming all other variables constant, ambulatory and bedridden patients are 1.6420 and 3.5290 times more likely to be defaulted than working patients, respectively.

Significance of a Polytomous Predictor

The Wald test considered above is used to identify the statistical significance of a binary or continuous explanatory variable. Whenever a multinomial explanatory variable is included (excluded) in (from) the model, all of its design variables should be included (excluded); to do otherwise implies the variables are recorded. By just looking at the Wald statistics of the design variables, the contribution of the variable could not be determined. Hence, the Wald test can be not used to check the significance of such a variable, rather the likelihood-ratio test should be used.

If X_j has m categories, then the null hypothesis of no contribution of this multinomial variable is $H_0 : \beta_{j1} = \beta_{j2} = \cdots = \beta_{j,m-1} = 0$. The likelihood-ratio test statistic is $G^2 = -2(\log \ell_R - \log \ell_M) \sim \chi^2(m-1)$ where ℓ_R is the maximized likelihood value under H_0 (excluding the multinomial variable X_j) and ℓ_M is the maximized likelihood value of the full model.

Example 3.8. Again recall example 3.6. Test the significance of functional status.

Solution: Since functional status is a multinomial variable with $m = 3$ categories, wald test cannot be used for checking its significance. The null hypothesis is $H_0 : \beta_{41} = \beta_{42} = 0$. Here, β_{41} and β_{42} are the parameters associated with the two design variables of functional status; ambulatory and bedridden, respectively. Therefore, the model in example 3.6 is re-fitted without the two design variables of marital status. When fitted, the log-likelihood value becomes -765.7410.

The likelihood-ratio test statistic is $G^2 = -2(\log \ell_R - \log \ell_M) = -2[-765.7410 - (-753.2892)] = 24.9036$. Since this value is greater than $\chi^2_{0.05}(2) = 5.9915$, functional status has a significant contribution to the model.

3.3.4 Confidence Intervals

Confidence Intervals for Parameters

Confidence intervals are more informative than tests. A confidence interval for β_j results from inverting a test of $H_0 : \beta_j = \beta_{j0}$. The interval is the set of β_{j0}'s for which the z test statistic is not greater than $z_{\alpha/2}$. This means $|\hat{\beta}_j - \beta_{j0}| \leq z_{\alpha/2}|\widehat{\mathrm{SE}}(\hat{\beta}_j)|$. This yields the confidence interval

$$\left[\hat{\beta}_j \pm z_{\alpha/2}\widehat{\mathrm{SE}}(\hat{\beta}_j)\right]$$

for β_j; $j = 1, 2, \cdots, k$. As the point estimate of the odds ratio associated to X_j is $\exp(\hat{\beta}_j)$ and its confidence interval is

$$\left\{\exp\left[\hat{\beta}_j \pm z_{\alpha/2}\widehat{\mathrm{SE}}(\hat{\beta}_j)\right]\right\}.$$

Example 3.9. Recall example 3.6 and construct the 95% confidence interval for each parameter and the corresponding odds ratio.

Solution: The critical value $z_{0.025} = 1.96$

Variable	$\hat{\beta}_j$	$\widehat{\text{SE}}(\hat{\beta}_j)$	95% CI for β_j	95% CI for $\text{OR}_j = \exp(\beta_j)$
Intercept	-0.3120	0.4299		
Age	-0.0282	0.0080	(-0.0439, -0.0125)*	(0.9570, 0.9876)*
Weight	-0.0051	0.0071	(-0.0190, 0.0088)	(0.9812, 1.0088)
Gender	0.5372	0.1438	(0.2554, 0.8190)*	(1.2910, 2.2682)*
Ambulatory	0.4959	0.1448	(0.2121, 0.7797)*	(1.2363, 2.1808)*
Bedridden	1.2610	0.2882	(0.6961, 1.8259)*	(2.0059, 6.2084)*
Working	Ref.			
CD4	-0.0007	0.0004	(-0.0015, 0.0001)	(0.9985, 1.0001)

Confidence Intervals for Predicted Probabilities

For summarizing the relationship, other characteristics may have greater importance such as $\pi(\boldsymbol{x}_i)$ at various $\boldsymbol{x}_i$ values. Consider the simple logistic model, logit $\hat{\pi}(x_i) = \hat{\alpha} + \hat{\beta} x_i$. For a fixed $x_i = x_0$, logit $\hat{\pi}(x_0) = \hat{\alpha} + \hat{\beta} x_0$ has a large standard error given by

$$\sqrt{\text{var}(\hat{\alpha}) + x_0^2 \text{ var}(\hat{\beta}) + 2x_0 \text{ cov}(\hat{\alpha}, \hat{\beta})}.$$

A $(1-\alpha)100\%$ confidence interval for logit $\pi(x_0)$ is

$$\left[(\hat{\alpha} + \hat{\beta} x_0) \pm z_{\alpha/2}\sqrt{\text{var}(\hat{\alpha} + \hat{\beta} x_0)}\right].$$

Substituting each end point into the inverse transformation

$$\pi(x_0) = \frac{\exp\{\text{logit}[\hat{\pi}(x_0)]\}}{1 + \exp\{\text{logit}[\hat{\pi}(x_0)]\}}$$

gives the corresponding interval for $\pi(x_0)$.

Example 3.10. Recall example 3.6, in which the estimated model is logit $\hat{\pi}(x_i) = -3.4648 + 0.0931 x_i$. The variance-covariance matrix of the estimated parameters is:

$$\begin{pmatrix} 3.4037 & -0.0744 \\ & 0.0019 \end{pmatrix}$$

Find the 95% confidence interval for the odds ratio and for the probability of success at the age of 39.6923 years ($x_i = 39.6923$).

Solution: $\hat{\beta} = 0.0931$, $\widehat{\text{var}}(\hat{\alpha}) = 3.4037$, $\widehat{\text{var}}(\hat{\beta}) = 0.0019$ and $\widehat{\text{cov}}(\hat{\alpha}, \hat{\beta}) = -0.0744$.

The 95% confidence interval for β is

$$\left[\hat{\beta} \pm z_{\alpha/2}\sqrt{\widehat{\text{var}}(\hat{\beta})}\right] = \left(0.0931 \pm 1.96\sqrt{0.0019}\right) = (0.0077, 0.1785).$$

This implies, the confidence interval for the odds ratio is

$$[\exp(0.0077, 0.1785)] = [\exp(0.0077), \exp(0.1785)] = (1.0077, 1.1954).$$

Also, to construct the confidence interval for the proportion of having hypertension at the age of 39.6923 years, the estimated probability of having hypertension at the age of 39.6923 years is logit $\hat{\pi}(39.6923) = -3.4648 + 0.0931(39.6923) = 0.2306$ and its estimated variance is

$$\begin{aligned}\widehat{\text{var}}\{\text{logit }[\hat{\pi}(39.6923)]\} &= \widehat{\text{var}}(\hat{\alpha}) + 39.6923^2\,\widehat{\text{var}}(\hat{\beta}) + 2(39.6923)\,\widehat{\text{cov}}(\hat{\alpha}, \hat{\beta})\\ &= 3.4037 + 39.6923^2(0.0019) + 2(39.6923)(-0.0744)\\ &= 0.4909\end{aligned}$$

The 95% confidence interval for logit $\pi(39.6923)$ is $(0.2306 \pm 1.96\sqrt{0.4909}) = (-1.1427, 1.6039)$. Thus, the 95% confidence interval for the probability of hypertension at the age of 39.6923 years is

$$\left[\frac{\exp(-1.1427)}{1+\exp(-1.1427)}, \frac{\exp(1.6039)}{1+\exp(1.6039)}\right] = (0.2418, 0.8326).$$

This confidence interval is very wide which may be due to the small sample size, $n = 13$.

Testing for a Set of Predictors

Sometimes, determining the contribution of a group of variables may be an interest. As usual, two models; one with all explanatory variables (full model) and the other without the explanatory variables to be tested (reduced model) are to be fitted. Thus, the reduced model is a special case of the full model.

Let ℓ_M denote the maximized value of the likelihood function for the model of interest M with $p_M = k+1$ parameters and let ℓ_R denote the maximized value of the likelihood function for the reduced model R with $p_R = k + 1 - q$ parameters. Note that model R is nested under model M. Thus, the null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_q = 0$ of no contribution of all the q predictors in model M (according to the alternative, at least one of the extra parameters in the full model is nonzero) is tested using $G^2 = -2(\log \ell_R - \log \ell_M) \sim \chi^2(q)$ where $\ell_M = \ell(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_k)$ and $\ell_R = \ell(\hat{\beta}_0, \hat{\beta}_{q+1}, \hat{\beta}_{q+2}, \cdots, \hat{\beta}_k)$.

Example 3.11. Recall example 3.6. Obtain the best fitting model.

Solution: Considering that the over all goal is to obtain the best fitting model, the logical step is to fit a reduced model containing only those significant variables and compare it to the model containing all the variables.

For our case, the model is of the form

$$\begin{aligned}\text{logit }\pi(\boldsymbol{x}_i) = \beta_0 &+ \beta_1 \text{ Age}_i + \beta_2 \text{ Weight}_i + \beta_3 \text{ Gender}_i\\ &+ \beta_{41} \text{ Ambulatory}_i + \beta_{42} \text{ Bedridden}_i + \beta_5 \text{ CD4}_i.\end{aligned}$$

Note that the variables Weight and CD4 are not significant. As a result, a new model is fitted excluding these insignificant variables. This new model has a log-likelihood value of -754.9283, and the parameter estimates and standard errors are in the following table.

Variable	Parameter Estimate	Standard Error	Wald Test
Intercept	-0.6858	0.2623	-2.6146*
Age	-0.0295	0.0079	-3.7342*
Gender	0.5305	0.1372	3.8666*
Ambulatory	0.5679	0.1375	4.1302*
Bedridden	1.3571	0.2827	4.8005*

The difference in this model is the exclusion of the Weight and CD4 variables. Thus, this reduced model is

$$\begin{aligned}\text{logit } \pi(\boldsymbol{x}_i) = \beta_0 + \beta_1 \text{ Age}_i + \beta_3 \text{ Gender}_i \\ + \beta_{41} \text{ Ambulatory}_i + \beta_{42} \text{ Bedridden}_i.\end{aligned}$$

Therefore, to determine whether the two variables should be included or not, the null hypothesis is $H_0 : \beta_2 = \beta_5 = 0$. The likelihood-ratio test statistic value is $G^2 = -2(\log \ell_R - \log \ell_M) = -2[-754.9283 - (-753.2892)] = 3.2782$ which is less than $\chi^2_{0.05}(2) = 5.9915$. Hence, there is no advantage of including both the Weight and CD4 variables in the model. Thus, the best fitting model is

$$\begin{aligned}\text{logit } \hat{\pi}(\boldsymbol{x}_i) = -0.6858 - 0.0295 \text{ Age}_i + 0.5305 \text{ Gender}_i \\ + 0.5679 \text{ Ambulatory}_i + 1.3571 \text{ Bedridden}_i.\end{aligned}$$

However, CD4 is known to be a "biologically important" variable. In this case, the decision to include or exclude the CD4 variable should be made in conjunction with subject matter experts.

Chapter 4

Model Building and Diagnostics

4.1 Objective and Learning Outcomes

In the previous chapter, a logistic regression model is fitted with a fixed set of explanatory variables and explored techniques of inference assuming that the model and the chosen variables were correct. Generally, every probability model is an assumption that may or may not be satisfied by the data. Also, in practice there is often uncertainty regarding which explanatory variables have to be included in a model. Therefore, the objective of this chapter is to describe the common model building procedures and diagnostics methods in fitting multiple logistic regression.

Upon completion of this chapter, students are expected to:

- Compare nested models using the likelihood-ratio (deviance) test and nonnested models using information criteria.
- Calculate measures of the predictive power (pseudo R^2s) from the likelihood values of a logistic model.
- Determine and interpret the overall proportion of correct classifications.
- Check the adequacy of a fitted model using the Pearson, deviance and Hosmer-Lemeshow gooodness-of-fit tests.

4.2 Model Selection

Model selection consists of identifying an appropriate probability model and choosing a set of explanatory variables to be used in the model. With several explanatory variables, there are many potential models. The model selection process becomes harder as the number of explanatory variables increases, because of the rapid increase in possible effects and interactions. There are two competing goals of model selection. The first is the model should be complex enough to fit the data well. A more complex model might contain a nonlinear effect, such as a quadratic term to allow the effect of a predictor to change directions as its value increases. Models with multiple predictors would consider interaction terms. On the other hand, it should be simple to interpret, smoothing rather than over fitting the data. Then, a search among many models may provide clues about which explanatory variables are associated with the response.

Variable selection is the process of reducing the size of the model from a potentially large number of variables to a more manageable and interpretable set. There are many approaches to selecting variables. The three most common ones are described below.

- *Forward* Selection:
  1. Fit a simple logistic regression model to each factor, one at a time.
  2. Select the most important factor according to a certain predetermined criterion.
  3. Test for the significance of the factor selected in step 2 and determine according to a certain predetermined criterion, whether or not to add this factor to the model.
  4. Repeat step 2 and 3 for those variables not yet in the model. At any subsequent step, if none meets the criterion in step 3 no more variables are included in the model and the process is terminated.
- *Backward* Elimination:
  1. Fit multiple logistic regression model containing all available explanatory variables.
  2. Select the least important variable according to a certain predetermined criterion; this is done by considering one factor at a time.
  3. Test for the significance of the factor selected in step 2 and determine according to the predetermined criterion, whether or not to delete this factor from the model.
  4. Repeat step 2 and 3 for those variables still in the model. At any subsequent step, if none meets the criterion in step 3, no more variables are removed from the model and the process is terminated.
- *Stepwise* Selection: It is a modified version of forward selection that permits re-examination, at every step, of the variables incorporated in the model in the previous steps. A variable entered at an early stage may become superfluous at a later stage because of its relationship with other variables now in the model; the information it provides becomes redundant. That variable may be removed if meeting the elimination criterion and the model is re-fitted with the remaining variables, and the forward process goes on. The entire process, one step forward followed by one step backward, continues until more variables can be added or removed.

4.3 Measures of Predictive Power

4.3.1 Pseudo R^2 Measures

In ordinary regression, the coefficient of determination R^2 and the multiple correlation R describe the power of the explanatory variables to predict the response, with $R \approx 1$ for best prediction. Despite the various attempts to define analogs for categorical response models, there is no proposed measure as widely useful as R and R^2. Some of the proposed measures which directly use the likelihood function are presented here.

Let the maximized likelihood be denoted by ℓ_M for a given model, ℓ_S for the saturated model and ℓ_0 for the null model containing only an intercept term. These probabilities are not greater than 1, thus log-likelihoods are nonpositive. As the model complexity increases, the parameter space expands, so the maximized log-likelihood increases. Thus, $\ell_0 \leq \ell_M \leq \ell_S \leq 1$ or $\log \ell_0 \leq \log \ell_M \leq \log \ell_S \leq 0$. The measure

$$R^2 = \frac{\log \ell_M - \log \ell_0}{\log \ell_S - \log \ell_0}$$

lies in between 0 and 1. It is zero when the model provides no improvement in fit over the null model and it will be 1 when the model fits as well as the saturated model.

The McFadden R^2

Since the saturated model has a parameter for each subject, the $\log \ell_S$ approaches to zero. Thus, $\log \ell_S = 0$ simplifies $R^2_{\text{McFadden}} = 1 - (\log \ell_M / \log \ell_0)$.

The Cox & Snell R^2

The Cox & Snell modified R^2 is $R^2_{\text{Cox-Snell}} = 1 - (\ell_0/\ell_M)^{2/n} = 1 - [\exp(\log \ell_0 - \log \ell_M)]^{2/n}$.

The Nagelkerke R^2

Because the $R^2_{\text{Cox-Snell}}$ value cannot reach 1, Nagelkerke modified it. The correction increases the Cox & Snell version to make 1 a possible value for R^2.

$$R^2_{\text{Nagelkerke}} = \frac{1 - (\ell_0/\ell_M)^{2/n}}{1 - (\ell_0)^{2/n}} = \frac{1 - [\exp(\log \ell_0 - \log \ell_M)]^{2/n}}{1 - [\exp(\log \ell_0)]^{2/n}}$$

Example 4.1. Obtain the McFadden, Cox & Snell, and Nagelkerke pseudo R^2s for the model fitted on example 3.11.

Solution: Note that $\log \ell_M = -754.9283$ which is given on example 3.11. Also $\log \ell_0 =$

-782.5257 and $n = 1464$ as given on example 3.6. Therefore,

$$\begin{aligned}
R^2_{\text{McFadden}} &= 1 - (-754.9283/-782.5257) = 0.0352\\
R^2_{\text{Cox-Snell}} &= 1 - [\exp(-782.5257 + 754.9283)]^{2/1464} = 0.0370\\
R^2_{\text{Nagelkerke}} &= \frac{1 - [\exp(-782.5257 + 754.9283]^{2/1464}}{1 - [\exp(-782.5257)]^{2/1464}} = 0.056.
\end{aligned}$$

4.3.2 Classification Tables

A classification table is also useful to summarize the predictive power of a binary logistic model. The table cross-classifies the binary response with a prediction of whether $y = 0$ or $y = 1$. The prediction is $\hat{y} = 1$ when $\hat{\pi} > \pi_0$ and $\hat{y} = 0$ when $\hat{\pi} \leq \pi_0$, for some cutoff π_0. Most classification tables use $\pi_0 = 0.5$. However, if a low (high) proportion of observations have $y = 1$, the model fit may never (always) have $\hat{\pi} > 0.50$, in which case one never (always) predicts $\hat{y} = 1$. Another possibility takes π_0 as the sample proportion of successes, which is $\hat{\pi}$ for the model containing only an intercept term.

Summary of prediction power from the classification table is the overall proportion of correct classifications. This estimates

$$\begin{aligned}
P(\text{correct classification}) &= P(y = 1 \text{ and } \hat{y} = 1) + P(y = 0 \text{ and } \hat{y} = 0)\\
&= P(y = 1) \cdot P(\hat{y} = 1|y = 1) + P(y = 0) \cdot P(\hat{y} = 0|y = 0).
\end{aligned}$$

Limitations of this table are that it collapses continuous predictive values $\hat{\pi}$ into binary ones, the choice of π_0 is arbitrary, and it is highly sensitive to the relative numbers of times $y = 1$ and $y = 0$.

Example 4.2. Recall example 3.1. The fitted probabilities of having hypertension for each individual is given in the following table. Using the cutoff value $\hat{\pi}_0 = 0.50$, find the proportion of correct classification.

Age (x_i)	Hypertension (y_i)	Probability [$\hat{\pi}(x_i)$]
20	1	0.1676
55	1	0.8397
55	1	0.8397
60	1	0.8929
60	1	0.8929
60	1	0.8929
45	1	0.6736
18	0	0.1432
55	0	0.8397
30	0	0.3381
18	0	0.1432
20	0	0.1676
20	0	0.1676

Solution: Based of the cutoff value 0.5, the probabilities above 0.5 are taken as 1 and those probabilities less than or equal to 0.5 are taken as 0. Hence, $\hat{y}_i = (0, 1, 1, 1, 1, 1, 1, 0, 1, 0, 0, 0, 0)$. Thus:

$$\begin{aligned} P(\text{correct classification}) &= P(y = 1) \cdot P(\hat{y} = 1|y = 1) + P(y = 0) \cdot P(\hat{y} = 0|y = 0) \\ &= \frac{7}{13} \cdot \frac{6}{7} + \frac{6}{13} \cdot \frac{5}{6} = 0.8462 \end{aligned}$$

Therefore, P(correct classification) is 84.62%.

Information Criteria (AIC and BIC)

Deviance or likelihood-ratio tests are used for comparing nested models. When there are non nested models, information criteria can help to select the good model. The best known ones are the Akaike Information Criterion (AIC) and Bayesian Information Criteria (BIC). Both judge a model by how close its fitted values tend to be to the true expected values. Also, both are calculated based on the likelihood value of a particular model M as $AIC = -2\log \ell_M + 2p_M$ and $BIC = -2\log \ell_M + p_M \log(n)$ where p_M is number of parameters in the model and n is the sample size. A model having smaller AIC or BIC is better.

Example 4.3. Find the AIC and BIC values for the model given on example 3.11.

Solution: It is already given $\log \ell_M = -754.9283$, $p_M = 5$ and $n = 1464$. This implies the $AIC = 1519.8566$ and $BIC = 1546.3012$.

4.4 Model Checking

Once the variable selection process is addressed, then the selected model should be explored for assessing whether the assumptions of the probability model are satisfied. The diagnostic methods for logistic regression, like that of linear regression, mostly rely residuals which compare observed and predicted values. Goodness-of-fit statistics are often computed as an objective measures of the overall fit of a model. A model checked and if it is found lacking the fit, a new model is proposed - fitted and then checked. And this process is repeated until a satisfactory model is found.

Similar to grouping the observations by the unique covariate patterns for the purpose of estimating the parameters, again here for the purpose of checking the goodness-of-fit of a model, the n independent responses are grouped into m unique covariate patterns (populations) each with $n_i; i = 1, 2, \cdots, m$ observations where $\sum\limits_{i=1}^{m} n_i = n$. Of the n_i observations in each covariate pattern, if n_{1i} successes are observed, then $n_{0i} = n_i - n_{1i}$ of them are failures. Thus, the raw residual is the difference between the observed number of successes n_{1i} and expected number of successes $\hat{\mu}(\boldsymbol{x}_i) = n_i\hat{\pi}(\boldsymbol{x}_i)$ for each value of the covariate $\boldsymbol{x}_i$.

4.4.1 The Pearson Chi-squared Goodness-of-fit Statistic

The Pearson residual is the standardized difference between the observed and expected number of successes. That is,

$$r_i = \frac{n_{1i} - n_i\hat{\pi}(\boldsymbol{x}_i)}{\sqrt{n_i\hat{\pi}(\boldsymbol{x}_i)[1-\hat{\pi}(\boldsymbol{x}_i)]}}; \quad i = 1, 2, \cdots, m.$$

Thus, the Pearson chi-squared statistic is the sum of the square of standardized residuals:

$$X^2 = \sum_{i=1}^{m} \frac{[n_{1i} - n_i\hat{\pi}(\boldsymbol{x}_i)]^2}{n_i\hat{\pi}(\boldsymbol{x}_i)[1-\hat{\pi}(\boldsymbol{x}_i)]} \sim \chi^2(m-k).$$

When this statistic is close to zero, it indicates a good model fit to the data. When it is large, it is an indication of lack of fit. Often the Pearson residuals r_i are used to determine exactly where the lack of fit occurs.

Example 4.4. Recall again example 3.1. Test the adequacy of the model using the Pearson chi-squared test.

Solution: The fitted probabilities are obtained from the fitted model. Note here the number of populations (aggregate values of the explanatory variable) is $m = 6$. Thus,

$$r_i = \frac{n_{1i} - n_i\hat{\pi}(\boldsymbol{x}_i)}{\sqrt{n_i\hat{\pi}(\boldsymbol{x}_i)[1-\hat{\pi}(\boldsymbol{x}_i)]}}; \quad i = 1, 2, \cdots, 6$$

Group (x_i)	Frequency (n_i)	Successes (n_{1i})	Probability [$\hat{\pi}(x_i)$]	r_i	r_i^2
18	2	0	0.1432	-0.5782	0.3343
20	3	1	0.1676	0.7685	0.5906
30	1	0	0.3381	-0.7147	0.5108
45	1	1	0.6736	0.6961	0.4846
55	3	2	0.8397	-0.8169	0.6673
60	3	3	0.8929	0.5999	0.3599
Total	13	7			2.9475

The Pearson chi-squared test statistic becomes $X^2 = \sum_{i=1}^{6} r_i^2 = 2.9475$ which is smaller than $\chi^2_{0.05}(6-2) = \chi^2_{0.05}(4) = 9.4877$, indicating that the model is a good fit to the data.

4.4.2 The Deviance Statistic

The deviance, like the Pearson chi-squared, is used to test the adequacy of the logistic model. As shown before, the maximum likelihood estimates of the parameters of the logistic regression are estimated iteratively by maximizing the Binomial likelihood function. Maximizing the likelihood function is equivalent to minimizing the deviance function. The choices for $\hat{\beta}_j$; $j = 0, 1, \cdots, k$ that minimize the deviance are the parameter values that make the observed and fitted proportions as close together as possible in a 'likelihood sense'. The deviance is given by:

$$D = 2\sum_{i=1}^{m}\left\{n_{1i}\log\left[\frac{n_{1i}}{n_i\hat{\pi}(\boldsymbol{x}_i)}\right] + (n_i - n_{1i})\log\left[\frac{n_i - n_{1i}}{n_i[1-\hat{\pi}(\boldsymbol{x}_i)]}\right]\right\} \sim \chi^2(m-k)$$

where the fitted probabilities $\hat{\pi}(\boldsymbol{x}_i)$ satisfy logit $\hat{\pi}(\boldsymbol{x}_i) = \sum_{j=0}^{k}\hat{\beta}_j x_{ij}$ and $x_{i0} = 1$. The deviance is small when the model fits the data, that is, when the observed and fitted proportions are close together. Large values of D (small p-values) indicate that the observed and fitted proportions are far apart, which suggests that the model is not good.

4.4.3 The Hosmer-Lemeshow Test Statistic

The Pearson chi-squared goodness-of-fit test cannot be readily applied if there are only one or a few observations for each possible value (combination of values) of the explanatory variable(s). Consequently, the Hosmer-Lemeshow statistic, the best goodness-of-fit test with continuous explanatory variables, was developed to address this problem. The idea is to aggregate similar observations into (mostly 10 - decile) groups that have large enough samples so that a Pearson statistic is computed on the observed and predicted counts from the groups. That is,

$$HL = \sum_{i=1}^{m}\frac{[n_{1i} - n_i\hat{\pi}(\boldsymbol{x}_i)]^2}{n_i\hat{\pi}(\boldsymbol{x}_i)[1-\hat{\pi}(\boldsymbol{x}_i)]} \sim \chi^2(m-2).$$

Chapter 5

Multicategory Logit Models

5.1 Objective and Learning Outcomes

The objective of this chapter is to extend the standard logistic regression model to handle outcome variables that have more than two categories. Multinomial logistic regression is used when the categories of the outcome variable are nominal, that is, they do not have any natural order. When the categories of the outcome variable do have a natural order, ordinal logistic regression may also be appropriate.

Upon completion of this chapter, students are expected to:

- Fit and interpret multinomial and ordinal logistic regression models for a multinomial response variable.
- Calculate the probability of each category of a multinomial and ordinal response variable given the values of the explanatory variables.
- Differentiate proportional and nonproportional odds models.

5.2 Logit Models for Nominal Responses

Multinomial logistic regression is used to predict a nominal dependent variable given one or more independent variables. It is an extension of binomial logistic regression to allow for a dependent variable with more than two categories.

Let Y be a categorical response with J categories. Let $P(Y = j|\boldsymbol{x}_i) = \pi_j(\boldsymbol{x}_i)$ at a fixed setting $\boldsymbol{x}_i$ for explanatory variables with $\sum\limits_{j=1}^{J} \pi_j(\boldsymbol{x}_i) = 1$. Thus, Y has a multinomial distribution with probabilities $\{\pi_1(\boldsymbol{x}_i), \pi_2(\boldsymbol{x}_i), \cdots, \pi_J(\boldsymbol{x}_i)\}$.

Multinomial (also called polytomous) logit models for nominal response variables simultaneously describe log odds for all $\binom{J}{2}$ pairs of categories. Of these, a certain choice of $J-1$ are enough to determine all, the rest are redundant. An odds for a multinomial response can be defined to be a comparison of *any pair* of response categories. For example, the odds of category 1 relative to category 3 is simply the ratio π_1/π_3.

5.2.1 Baseline Category Logit Models

Logit models for multinomial responses are developed by selecting one response category, often the first (last) category or the most common one, as a baseline (reference) and forming the odds of the remaining $J-1$ categories against this category. For example, the *multinomial logit* model (also called *baseline category logit* model) pairing each response category with the last category,

$$\log\left[\frac{\pi_j(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)}\right] = \beta_{j0} + \beta_{j1}x_{i1} + \beta_{j2}x_{i2} + \cdots + \beta_{jk}x_{ik};\ j = 1, 2, \cdots, J-1$$

simultaneously describes the effects of the explanatory variables on the $J-1$ logit models (if $J = 2$, it simplifies to binary logistic regression model). The intercepts and effects vary according to the response paired with the baseline. That is, each model has its own intercept and slope. Also note that for the reference category, $\beta_{J0} = \beta_{J1} = \beta_{J2} = \cdots = \beta_{Jk} = 0$.

The $J-1$ equations also determine parameters for logit models with other pairs of response categories, since

$$\log\left[\frac{\pi_1(\boldsymbol{x}_i)}{\pi_2(\boldsymbol{x}_i)}\right] = \log\left[\frac{\pi_1(\boldsymbol{x}_i)/\pi_J(\boldsymbol{x}_i)}{\pi_2(\boldsymbol{x}_i)/\pi_J(\boldsymbol{x}_i)}\right] = \log\left[\frac{\pi_1(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)}\right] - \log\left[\frac{\pi_2(\boldsymbol{x}_i)}{\pi_J(\boldsymbol{x}_i)}\right].$$

Example 5.1. Based on the survival outcome of HAART treatment, HIV/AIDS patients were classified into four categories (0= Active, 1= Dead, 2= Transferred to other hospital, 3= Lost-to-follow). To identify factors associated with these survival outcomes, a multinomial logit model was fitted. Three explanatory variables that were considered are Age, Gender (0= Female, 1= Male) and Functional Status (0= Working, 1= Ambulatory, 2= Bedridden). The parameter estimates are presented as follows (values in brackets are standard errors).

				Functional Status	
logit	Intercept	Age	Gender	Ambulatory	Bedridden
$\log(\hat{\pi}_D/\hat{\pi}_A)$	-3.271 (0.624)	-0.020 (0.018)	0.564 (0.325)	0.940 (0.333)	2.280 (0.479)
$\log(\hat{\pi}_T/\hat{\pi}_A)$	-1.882 (0.413)	-0.030 (0.012)	0.635 (0.211)	0.833 (0.209)	1.584 (0.393)
$\log(\hat{\pi}_L/\hat{\pi}_A)$	-1.116 (0.343)	-0.031 (0.010)	0.455 (0.178)	0.292 (0.183)	0.828 (0.395)

Write the estimated multinomial logit models and interpret. Also, find the estimated logit model for the log odds of dead instead of transferred to other hospital.

Solution: Let Y = survival outcome, X_1 = age of the patient, X_2 = gender and X_3= functional status.

Each model is written as:

$$\log\left[\frac{\hat{\pi}_j(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] = \hat{\beta}_{j0} + \hat{\beta}_{j1}x_{i1} + \hat{\beta}_{j2}x_{i2} + \hat{\beta}_{j31}d_{i31} + \hat{\beta}_{j32}d_{i32};\ j = D, T, L.$$

For example, the estimated model for the log odds of being dead instead of active is

$$\log\left[\frac{\hat{\pi}_D(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] = -3.271 - 0.020x_{i1} + 0.564x_{i2} + 0.940d_{i31} + 2.280d_{i32}.$$

An increase in the age of a patient by one year decreases the odds of being dead (instead of active) by 2% (a factor of $\exp(-0.020) = 0.98$). The odds that male patients being dead (instead of active) is $\exp(0.565) = 1.759$ times that of females, or the odds of being dead (instead of active) is 75.9% higher for males than for females. In other words, relative to female patients, male patients are 1.759 times (75.9%) more likely to be dead (instead of active). Also, ambulatory patients are $\exp(0.941) = 2.563$ times more likely to be dead (instead of active) as compared to working patients. Similarly, bedridden patients are $\exp(2.280) = 9.777$ times more likely to be dead (instead of active) relative to working patients. The functional status effects indicate that the odds of being dead (instead of active) are relatively higher for bedridden patients relative to ambulatory patients.

The estimated model for the log odds of being transferred instead of active is

$$\log\left[\frac{\hat{\pi}_T(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] = -1.882 - 0.030x_{i1} + 0.635x_{i2} + 0.833d_{i31} + 1.584d_{i32}.$$

An increase in the age of a patient by a year decreases the odds of being transferred to other hospital (instead of active) by 3% (a factor of $\exp(-0.030) = 0.970$). The odds that male patients being transferred to other hospital (instead of active) is $\exp(0.635) = 1.887$ times that of females, or the odds of being transferred to other hospital (instead of active) is 88.7% higher for males than for females. In other words, male patients are 1.887 times (88.7%) more likely to be transferred to other hospital (instead of active) as compared to female patients. Also, relative to working patients, ambulatory patients are $\exp(0.833) = 2.300$ times more likely to be transferred to other hospital (instead of active). Similarly, bedridden patients are $\exp(1.584) = 4.874$ times more likely to be transferred to other hospital (instead of active) as compared to working patients.

Also, the estimated model for the log odds of being lost-to-follow instead of active is

$$\log\left[\frac{\hat{\pi}_L(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] = -1.116 - 0.031x_{i1} + 0.455x_{i2} + 0.292d_{i31} + 0.828d_{i32}.$$

The odds of being lost-to-follow (instead of active) decreases by 3.1% (a factor of $\exp(-0.031) = 0.969$) every year older an individual is. Male patients are $\exp(0.455) = 1.576$ times (57.6%)

more likely to be lost-to-follow (instead of active) relative to female patients. As compared to working patients, ambulatory patients are exp(0.292) = 1.339 times (33.9%) more likely to be lost-to-follow (instead of active). Similarly, bedridden patients are exp(0.828) = 2.289 times more likely to be lost-to-follow (instead of active) as compared to working patients.

The estimated model for being dead instead of transferred to other hospital is

$$\begin{aligned}\log\left[\frac{\hat{\pi}_D(\boldsymbol{x}_i)}{\hat{\pi}_T(\boldsymbol{x}_i)}\right] =& \log\left[\frac{\hat{\pi}_D(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right] - \log\left[\frac{\hat{\pi}_T(\boldsymbol{x}_i)}{\hat{\pi}_A(\boldsymbol{x}_i)}\right]\\ =& -3.271 - 0.020x_{i1} + 0.564x_{i2} + 0.940d_{i31} + 2.280d_{i32}\\ & -(-1.882 - 0.030x_{i1} + 0.635x_{i2} + 0.833d_{i31} + 1.584d_{i32})\\ =& -1.389 + 0.010x_{i1} - 0.071x_{i2} + 0.107d_{i31} + 0.696d_{i32}.\end{aligned}$$

Therefore, the estimated model for the log odds of dead instead of transferred to other hospital is

$$\log\left[\frac{\hat{\pi}_D(\boldsymbol{x}_i)}{\hat{\pi}_T(\boldsymbol{x}_i)}\right] = -1.389 + 0.010x_{i1} - 0.071x_{i2} + 0.107d_{i31} + 0.696d_{i32}.$$

5.2.2 Multinomial Response Probabilities

The probabilities for each category of the multinomial response also can be found in terms of the model. Using the properties of logarithms, the logit models for a multinomial responses can be re-written as $\pi_j(\boldsymbol{x}_i) = \pi_J(\boldsymbol{x}_i)\exp(\beta_{j0} + \beta_{j1}x_{i1} + \beta_{j2}x_{i2} + \cdots + \beta_{jk}x_{ik})$ for $j = 1, 2, \cdots, J-1$.

Since $\sum_{h=1}^{J}\pi_h(\boldsymbol{x}_i) = 1$, $\sum_{h=1}^{J-1}\pi_J(\boldsymbol{x}_i)\exp(\beta_{h0} + \beta_{h1}x_{i1} + \beta_{h2}x_{i2} + \cdots + \beta_{hk}x_{ik}) + \pi_J(\boldsymbol{x}_i) = 1$. By factoring out the common term $\pi_J(\boldsymbol{x}_i)$, the probability of the reference category is

$$\pi_J(\boldsymbol{x}_i) = \frac{1}{1 + \sum_{h=1}^{J-1}\exp(\beta_{h0} + \beta_{h1}x_{i1} + \beta_{h2}x_{i2} + \cdots + \beta_{hk}x_{ik})}.$$

Hence, the equation that expresses multinomial logit models directly in terms of response probabilities $\{\pi_j(\boldsymbol{x}_i)\}$ is

$$\pi_j(\boldsymbol{x}_i) = \frac{\exp(\beta_{j0} + \beta_{j1}x_{i1} + \beta_{j2}x_{i2} + \cdots + \beta_{jk}x_{ik})}{1 + \sum_{h=1}^{J-1}\exp(\beta_{h0} + \beta_{h1}x_{i1} + \beta_{h2}x_{i2} + \cdots + \beta_{hk}x_{ik})};\ j = 1, 2, \cdots, J-1.$$

Or, in general for all the response categories, it can be written as

$$\pi_j(\boldsymbol{x}_i) = \frac{\exp(\beta_{j0} + \beta_{j1}x_{i1} + \beta_{j2}x_{i2} + \cdots + \beta_{jk}x_{ik})}{\sum_{h=1}^{J}\exp(\beta_{h0} + \beta_{h1}x_{i1} + \beta_{h2}x_{i2} + \cdots + \beta_{hk}x_{ik})};\ j = 1, 2, \cdots, J$$

where $\beta_{Jp} = 0$ for $p = 1, 2, \cdots, k$.

Example 5.2. Consider the previous example. Find the estimated probability of each outcome for a 40 years old female patient who were working.

Solution: The estimated probability of each outcome with $x_{i1} = 40$, $x_{i2} = 0$ and $d_{i31} = d_{i32} = 0$

$$\begin{aligned}
\hat{\pi}_D(\boldsymbol{x}_i) &= \frac{\exp[-3.271 - 0.020(40)]}{1 + \exp[-3.271 - 0.020(40)] + \exp[-1.882 - 0.030(40)] + \exp[-1.116 - 0.031(40)]} \\
&= 0.0147 \\
\hat{\pi}_T(\boldsymbol{x}_i) &= \frac{\exp[-1.882 - 0.030(40)]}{1 + \exp[-3.271 - 0.020(40)] + \exp[-1.882 - 0.030(40)] + \exp[-1.116 - 0.031(40)]} \\
&= 0.0396 \\
\hat{\pi}_L(\boldsymbol{x}_i) &= \frac{\exp[-1.116 - 0.031(40)]}{1 + \exp[-3.271 - 0.020(40)] + \exp[-1.882 - 0.030(40)] + \exp[-1.116 - 0.031(40)]} \\
&= 0.0819 \\
\hat{\pi}_A(\boldsymbol{x}_i) &= \frac{1}{1 + \exp[-3.271 - 0.020(40)] + \exp[-1.882 - 0.030(40)] + \exp[-1.116 - 0.031(40)]} \\
&= 0.8638
\end{aligned}$$

The value 1 in each denominator and in the numerator of $\hat{\pi}_A(\boldsymbol{x}_i)$ represents exp(0) for which $\hat{\beta}_0 = \hat{\beta}_1 = \cdots = \hat{\beta}_k = 0$ with the baseline category.

5.3 Cumulative Logit Models for Ordinal Responses

Many categorical response variables have a natural ordering to their categories or called *levels*. For example, a response variable (like amount of agreement) may be measured using a Likert scale with categories 'strongly disagree', 'disagree', 'neutral', 'agree' or 'strongly agree'. Ordinal logistic regression is used to predict such an ordinal dependent variable given one or more independent variables.

5.3.1 Cumulative Logits

Let Y is an ordinal response with J categories. Then there are $J - 1$ ways to dichotomize these outcomes. These are $Y_i \leq 1$ $(Y_i = 1)$ versus $Y_i > 1$, $Y_i \leq 2$ versus $Y_i > 2$, $\cdots$, $Y_i \leq J - 1$ versus $Y_i > J - 1$ $(Y_i = J)$. With this categorization of Y_i, $P(Y_i \leq j)$ is the cumulative probability that Y_i falls at or below category j. That is, for outcome j, the cumulative probability is

$$P(Y_i \leq j|\boldsymbol{x}_i) = \pi_1(\boldsymbol{x}_i) + \pi_2(\boldsymbol{x}_i) + \cdots + \pi_j(\boldsymbol{x}_i);\ j = 1, 2, \cdots, J$$

where $P(Y_i \leq j|\boldsymbol{x}_i) = 1$. Each cumulative logit model uses all the J response levels. A model for logit $[P(Y \leq j|\boldsymbol{x}_i)]$ alone is the usual logit model for a binary response in which categories from 1 to j form one outcome and categories from $j+1$ to J form the second. That is,

$$\begin{aligned}\text{logit } P(Y_i \leq j) &= \log\left[\frac{P(Y_i \leq j)}{1 - P(Y_i \leq j)}\right] \\ &= \log\left[\frac{P(Y_i \leq j)}{P(Y_i > j)}\right] \\ &= \log\left[\frac{\pi_1(\boldsymbol{x}_i) + \pi_2(\boldsymbol{x}_i) + \cdots + \pi_j(\boldsymbol{x}_i)}{\pi_{j+1}(\boldsymbol{x}_i) + \pi_{j+2}(\boldsymbol{x}_i) + \cdots + \pi_J(\boldsymbol{x}_i)}\right];\ j = 1, 2, \cdots, J-1.\end{aligned}$$

5.3.2 Proportional Odds Model

Formally, a model that simultaneously uses all cumulative logits assuming linear relationship with the explanatory variables is

$$\text{logit } P(Y_i \leq j|\boldsymbol{x}_i) = \beta_{j0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik};\ j = 1, 2, \cdots, J-1.$$

Each cumulative logit has its own intercept which usually are not of interest except for computing response probabilities. Since logit $[P(Y_i \leq j|\boldsymbol{x}_i)]$ increases in j for a fixed $\boldsymbol{x}_i$ and the logit is an increasing function of this probability, each intercept increases in j.

But, the model assumes the *same slope* (its associated odds ratio called *cumulative odds ratio*) regardless of the category j. This is called *proportional odds* assumption which means the distance between each category is equivalent (proportional odds). That is, each model has the same effect associated with each explanatory variable (the effects of the explanatory variables are the same regardless of which cumulative probabilities are used).

The slope parameters can be interpreted in the same way as a binary logistic regression parameters - except in this case, there are three transitions estimated instead of one transition - as there would be with a dichotomous dependent variable. A positive parameter indicates an increased chance that a subject with a higher score on the independent variable will be observed in a higher category. A negative parameter indicates that the chances that a subject with a higher score on the independent variable will be observed in a lower category.

The intercepts can be used to calculate predicted probabilities for a person with a given set of characteristics of being in a particular category.

Example 5.3. To determine the effect of Age and Gender (0= Female, 1=Male) on the Clinical Stage of HIV/AIDS patients (1= Stage I, 2= Stage II, 3= Stage III and 4= Stage IV), the following parameter estimates of ordinal logistic regression are obtained. The loglikelihood values of the null and the full models are -1854.3173 and -1852.1351, respectively.

Variable	Parameter Estimate	Standard Error
Intercept 1	-0.9905	0.1884
Intercept 2	0.5383	0.1870
Intercept 3	2.7246	0.2066
Age	0.0034	0.0055
Gender	0.1789	0.1028

Obtain the cumulative logit model and interpret.

Solution: Let Y= Clinical Stage of patients (1= Stage I, 2= Stage II, 3= Stage III and 4= Stage IV), X_1= Age and X_2= Gender (0= Female, 1=Male).

Hence, the model has the form logit $\widehat{P}(Y_i \leq j|\boldsymbol{x}_i) = \hat{\beta}_{j0} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2};\ j = 1, 2, 3$. With $J = 4$ categories, the model has three cumulative logits. These are:

$$\begin{aligned}
\text{logit } \widehat{P}(Y_i \leq 1|\boldsymbol{x}_i) &= -0.9905 + 0.0034x_{i1} + 0.1789x_{i2} \\
\text{logit } \widehat{P}(Y_i \leq 2|\boldsymbol{x}_i) &= 0.5383 + 0.0034x_{i1} + 0.1789x_{i2} \\
\text{logit } \widehat{P}(Y_i \leq 3|\boldsymbol{x}_i) &= 2.7246 + 0.0034x_{i1} + 0.1789x_{i2}.
\end{aligned}$$

The cumulative estimate $\hat{\beta}_1 = 0.0034$ suggests an increase in the age of the patient leads to be in higher clinical stages given the gender. Being in smaller ages reduces the likelihood of being in a higher clinical stage category. Also, the estimate $\hat{\beta}_2 = 0.1789$ males are more likely to be in higher clinical stages as compared to females given the age of the patient. That is, being male increases the likelihood of being in a higher clinical stage category.

5.3.3 Cumulative Response Probabilities

The response probabilities $P(Y_i = j|\boldsymbol{x}_i)$ of an ordinal logit model is determined as $P(Y_i = j|\boldsymbol{x}_i) = P(Y_i \leq j|\boldsymbol{x}_i) - P(Y_i \leq j-1|\boldsymbol{x}_i)$ where the cumulative response probabilities $P(Y_i \leq j|\boldsymbol{x}_i)$ are given by

$$P(Y_i \leq j|\boldsymbol{x}_i) = \frac{\exp(\beta_{j0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})}{1 + \exp(\beta_{j0} + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik})};\ j = 1, 2, \cdots, J-1.$$

Hence, an ordinal logit model estimates the cumulative probability of being in one category versus all lower or higher categories.

Example 5.4. Recall example 5.3. Find the estimated probabilities of each clinical stage for a female patient at the mean age 34.01 years.

Solution: The estimated probability of response clinical stage j or below is:

$$\widehat{P}(Y_i \leq j|\boldsymbol{x}_i) = \frac{\exp(\hat{\beta}_{j0} + 0.0034x_{i1} + 0.1789x_{i2})}{1 + \exp(\hat{\beta}_{j0} + 0.0034x_{i1} + 0.1789x_{i2})};\ j = 1, 2, 3.$$

The cumulative response probability of a female patient at the age of 34.01 years being in clinical stage I, clinical stages I or II, clinical stages I, II or III, respectively, are:

$$\begin{aligned}
\widehat{P}(Y_i \leq 1|\boldsymbol{x}_i) &= \frac{\exp[-0.9905 + 0.0034(34.01) + 0.1789(0)]}{1 + \exp[-0.9905 + 0.0034(34.01) + 0.1789(0)]} \\
&= 0.2942 \\
\widehat{P}(Y_i \leq 2|\boldsymbol{x}_i) &= \frac{\exp[0.5383 + 0.0034(34.01) + 0.1789(0)]}{1 + \exp[0.5383 + 0.0034(34.01) + 0.1789(0)]} \\
&= 0.6579 \\
\widehat{P}(Y_i \leq 3|\boldsymbol{x}_i) &= \frac{\exp[2.7246 + 0.0034(34.01) + 0.1789(0)]}{1 + \exp[2.7246 + 0.0034(34.01) + 0.1789(0)]} \\
&= 0.9448
\end{aligned}$$

Note also that $\widehat{P}(Y_i \leq 4|\boldsymbol{x}_i) = 1$. Thus, the actual response probability of a female patient of 34.01 years old at each clinical stage is calculated as

$$\begin{aligned}
\widehat{P}(Y_i = 1|\boldsymbol{x}_i) &= \widehat{P}(Y_i \leq 1|\boldsymbol{x}_i) \\
&= 0.2942 \\
\widehat{P}(Y_i = 2|\boldsymbol{x}_i) &= \widehat{P}(Y_i \leq 2|\boldsymbol{x}_i) - \widehat{P}(Y_i = 1|\boldsymbol{x}_i) \\
&= 0.6579 - 0.2942 \\
&= 0.3637 \\
\widehat{P}(Y_i = 3|\boldsymbol{x}_i) &= \widehat{P}(Y_i \leq 3|\boldsymbol{x}_i) - \widehat{P}(Y_i \leq 2|\boldsymbol{x}_i) \\
&= 0.9448 - 0.6579 \\
&= 0.2869 \\
\widehat{P}(Y_i = 4|\boldsymbol{x}_i) &= 1 - \widehat{P}(Y_i = 3|\boldsymbol{x}_i) \\
&= 1 - 0.9448 \\
&= 0.0552
\end{aligned}$$

5.3.4 Nonproportional Odds Model

A proportional odds model is one of the preferred ways to account for an ordered response, because the slope regression parameters are constant over the response categories. While this can greatly simplify the model, it imposes the assumption that association affects the logit of cumulative probabilities the same way for all $j = 1, 2, \cdots, J-1$. This assumption may not hold in all situations. An alternative model that relaxes this assumption is a *nonproportional (partial proportional) odds* model which is written as

$$\text{logit } P(Y_i \leq j|\boldsymbol{x}_i) = \beta_{j0} + \beta_{j1}x_{i1} + \beta_{j2}x_{i2} + \cdots + \beta_{jk}x_{ik};\ j = 1, 2, \cdots, J-1.$$

Notice that all the slope parameters are now allowed to vary across the levels of the ordinal response.

Because the proportional odds model is a special case of nonproportional odds model, the proportional odds assumption can be tested through the hypothesis $H_0 : \beta_{1p} = \beta_{2p} = \cdots = \beta_{J-1,p}$ for $p = 1, 2, \cdots , k$. The test is conducted as a likelihood-ratio test where the degrees of freedom for the χ^2 distribution is the difference in the number of parameters between the two models, $(k + 1)(J - 1) - (p + J - 1) = (J - 2)p$. Rejecting the proportional odds assumption suggests that the nonproportional odds model may be preferred. But failing to reject the proportional odds hypothesis is not a proof that the assumption holds. However, it offers some assurance that a proportional odds model provides a reasonable approximation to true relationships between the ordinal response and the explanatory variables.

Example 5.5. Recalling example 5.3, the parameter estimates of a nonproportional odds model, with a loglikelihood value of -1850.1355, are given as follows.

Variable	Parameter Estimate	Standard Error
Intercept 1	1.0736	0.2376
Intercept 2	-0.5840	0.2062
Intercept 3	-2.6955	0.3851
Age 1	-0.0007	0.0071
Age 2	0.0057	0.0061
Age 3	0.0033	0.0112
Gender 1	0.3509	0.1376
Gender 2	0.0928	0.1144
Gender 3	0.1096	0.2126

Write out the estimated models.

Solution: The model has the usual form logit $\widehat{P}(Y_i \leq j|\boldsymbol{x}_i) = \hat{\beta}_{j0} + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2}$; $j = 1, 2, 3$. Three cumulative logits are

$$\begin{aligned}
\text{logit } \widehat{P}(Y_i \leq 1|\boldsymbol{x}_i) &= \quad 1.0736 - 0.0007x_{i1} + 0.3509x_{i2} \\
\text{logit } \widehat{P}(Y_i \leq 2|\boldsymbol{x}_i) &= -0.5840 + 0.0057x_{i1} + 0.0928x_{i2} \\
\text{logit } \widehat{P}(Y_i \leq 3|\boldsymbol{x}_i) &= -2.6955 + 0.0033x_{i1} + 0.1096x_{i2}.
\end{aligned}$$

Chapter 6

Count Regression Models

6.1 Objective and Learning Outcomes

The objective of this chapter is to introduce the basics of poisson regression model which is a statistical modeling scenario where the responses are counts or frequencies, that is, non-negative integers. Upon completion of this chapter, students are expected to know when and how to apply poisson and negative binomial regression models. Count regressions such as poisson and negative-binomial models are used for modelling *count (discrete)* response variables: for example, the number of hospital admissions or the number of accidents over some period of time. The unit of analysis may be a person (e.g., number of infections per patient per year), an institution (e.g., number of admissions per hospital per month) or a place (e.g., number of car accidents per city per day). As a first pass, such a dependent variable could be analyzed as a continuous outcome. However, unlike a continuous variable, with counts there cannot be negative numbers. Also, the distribution of counts often tend to be skewed to the right and does not fit a normal distribution.

Count regression models are also used to model *incidence rate* or incidence of rare diseases. Incidence rate measures the rate at which a group of people develops a disease or condition. Often it is of interest to compare incidence rates. For example, is the incidence of diabetes higher in one city than another or is higher among men than women. As is true of counts, incidence rates cannot be negative. As a result, in situations such as these, analyzing the data with a technique such as linear regression is not appropriate.

6.2 The Exponential Function

Count regression models are modeled based on the exponential function. For any real number z, the exponential function is $f(z) = \exp(z)$. This function is nonnegative for all values of z. That is, if $z = -\infty$, then $f(-\infty) = 0$, if $z = 0$, then $f(0) = 1$ and if $z = \infty$, then $f(\infty) = \infty$.

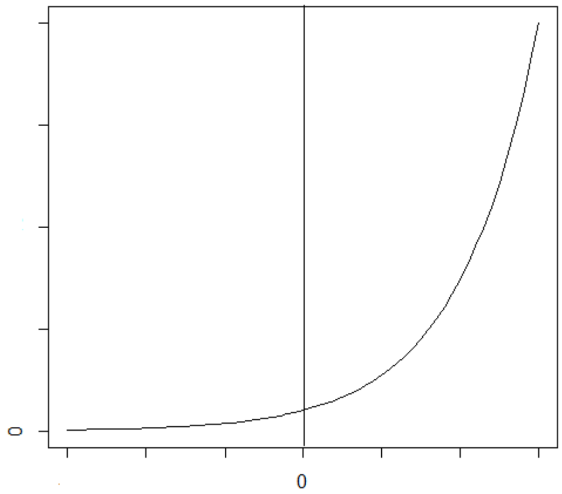

Figure 6.1: Plot of the Exponential Function

The figure also shows that the range of f is in between 0 and ∞ for every real number z. Therefore, $0 \leq f(z) < \infty$.

6.3 The Poisson Regression Model

To obtain the poisson regression model from the exponential function, z should be expressed as a function (mostly linear function) of the explanatory variable(s). That is, $z_i = g(x_i) = \alpha + \beta x_i$ for a single explanatory variable X. As a result, the simple poisson regression model can be written as $f(x_i) = \exp(\alpha + \beta x_i)$. Here, since $f(x_i)$ represents the mean response, let us use the notation $\mu(x_i)$. That is, $\mu(x_i) = \exp(\alpha + \beta x_i)$. This model can be linearized using the natural logarithm transformation as:

$$\log \mu(x_i) = \alpha + \beta x_i.$$

Here α and β are the intercept and slope parameters of the log-linear model. The slope parameter is commonly interpreted in terms of an incidence rate ratio (IRR). A one unit increase in x_i has a multiplicative impact of $\exp(\beta)$ on the mean response, that is, the mean of Y_i at $x_i + 1$ is the mean of Y_i at x_i multiplied by $\exp(\beta)$. If $\beta = 0$, then the multiplicative factor is 1, the mean of Y_i does not change as x_i changes. If $\beta > 0$, then $\exp(\beta) > 1$ and the mean of Y_i increases as x_i increases. If $\beta < 0$, the mean decreases as x_i increases.

Similarly, if there are k explanatory variables, the multiple poisson regression model is written as:

$$\log \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} = \sum_{j=0}^{k} \beta_j x_{ij} \tag{6.1}$$

where $x_{i0} = 1$ for all $i = 1, 2, \cdots, n$. Here, $\mu(\boldsymbol{x}_i)$ is the conditional mean of Y_i given $\boldsymbol{x}_i$ where $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{ik})$.

The sample poisson regression model is:

$$\log \hat{\mu}(\boldsymbol{x}_i) = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_k x_{ik} = \sum_{j=0}^{k} \hat{\beta}_j x_{ij} \tag{6.2}$$

- $\hat{\mu}(\boldsymbol{x}_i)$ is the estimated mean response.
- $\hat{\beta}_0$ is the estimated intercept of the log-linear model.
- $\hat{\beta}_j; j = 1, 2, \cdots, k$ is the j^{th} estimated (partial) slope associated with the j^{th} independent variable.

Example 6.1. Suppose a study is conducted in identifying factors associated with CD4 counts of 1464 HIV/AIDS patients at the start of HAART treatment. Here the response variable is CD4 count of a patient and the explanatory variables were Age in years (Age), Gender (0=Female, 1=Male) and Functional Status (0=Working, 1=Ambulatory, 2=Bedridden). The parameter estimates and their corresponding standard errors of the poisson regression model are given in the following table.

Variable	Parameter Estimate	Standard Error
Intercept	5.4625	0.0079
Age	0.0060	0.0002
Gender	-0.1982	0.0041
Ambulatory	-0.3783	0.0046
Bedridden	-0.6296	0.0123

Obtain the estimated model and interpret the estimates.

Solution: Let Y= CD4 count, X_1= Age, X_2= Gender (0=Female, 1=Male) and X_3= Functional Status (0=Working, 1=Ambulatory, 2=Bedridden). The estimated model is:

$$\log \hat{\mu}(\boldsymbol{x}_i) = 5.4625 + 0.0060 x_{i1} - 0.1982 x_{i2} - 0.3783 d_{i31} - 0.6296 d_{i32}.$$

As the age of the patient increases by one year, the mean CD4 count increases by 0.60% [$\exp(0.0060) - 1 = 0.60\%$]. The mean CD4 count of male patients decreases by 17.98% [$1 - \exp(-0.1982) = 17.98\%$] than female patients. Similarly the mean CD4 counts of ambulatory and bedridden patients decreases by 31.50% and 46.72% than working patients, respectively.

6.3.1 Estimation

Inference on the model and its parameters follows exactly the same approach as used for logistic regression. Like other regression modeling, the goal of poisson regression is to estimate the $k+1$ unknown parameters of the model. The method of maximum likelihood is used to estimate the parameters which follows closely the approach used for logistic regression.

Consider a random variable Y that can take on a set of count values. Given a dataset with a sample size of n where each observation is independent. Thus, $\boldsymbol{Y}$ can be considered as a vector of n poisson random variables. That is, each individual count response Y_i; $i = 1, 2, \cdots, n$ has an independent poisson distribution with parameter $\mu(\boldsymbol{x}_i)$, that is,

$$P(Y_i = y_i) = \frac{\mu(\boldsymbol{x}_i)^{y_i} \exp[-\mu(\boldsymbol{x}_i)]}{y_i!};\ y_i = 0, 1, 2, \cdots .$$

Then, the joint probability mass function of $\boldsymbol{Y}^t = (Y_1, Y_2, \cdots, Y_n)$ is the product of the n poisson distributions. Thus, the likelihood function is:

$$\ell(\boldsymbol{\beta}|\boldsymbol{y}) = \prod_{i=1}^{n} \frac{\mu(\boldsymbol{x}_i)^{y_i} \exp[-\mu(\boldsymbol{x}_i)]}{y_i!} \tag{6.3}$$

where $\mu(\boldsymbol{x}_i) = \exp(\sum\limits_{j=0}^{k} \beta_j x_{ij})$. Also, the log-likelihood function becomes:

$$L(\boldsymbol{\beta}|\boldsymbol{y}) = \sum_{i=1}^{n} y_i \log\ [\mu(\boldsymbol{x}_i)] - \sum_{i=1}^{n} \mu(\boldsymbol{x}_i) - \sum_{i=1}^{n} \log\ (y_i!). \tag{6.4}$$

Then, partially differentiating the log-likelihood with respect to β_j; $j = 0, 1, 2, \cdots, k$ and setting it equal to zero results $k+1$ equations with $k+1$ unknown parameters. That is,

$$\frac{\partial L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \beta_j} = \sum_{i=1}^{n} [y_i - \mu(\boldsymbol{x}_i)] x_{ij} = 0;\ j = 0, 1, 2, \cdots, k. \tag{6.5}$$

which is usually solved with some numerical method like the Newton-Raphson algorithm.

Also, the second partial derivative of the log-likelihood function yields the variance-covarince matrix of the estimated parameters:

$$\frac{\partial^2 L(\boldsymbol{\beta}|\boldsymbol{y})}{\partial \beta_j \beta_h} = -\sum_{i=1}^{n} \mu(\boldsymbol{x}_i) x_{ij} x_{ih};\quad j = h = 0, 1, 2, \cdots, k. \tag{6.6}$$

6.3.2 Significance Tests

Let ℓ_M denote the maximized value of the likelihood function for the fitted model M with all the k explanatory variables. Let ℓ_0 denote the maximized value of the likelihood function for the fitted model with no explanatory variables (having only one parameter, that is, the intercept). The likelihood-ratio test statistic is $G^2 = -2(\log \ell_0 - \log \ell_M) = D_0 - D_M \sim \chi^2(k)$. Rejection of the null hypothesis implies at least one of the parameter is significantly different from zero. Then, Wald test can be used to look at the significance of each variable ($H_0 : \beta_j = 0$) using a Z statistic in which

$$Z_j = \frac{\hat{\beta}_j}{\widehat{\text{SE}}(\hat{\beta}_j)} \sim N(0,1)$$

for large sample size.

Example 6.2. The log-likelihood value of the model given in example 6.1 is -85956.40 and the corresponding null model is -92061.31. Test the overall significance of the model and also identify the significant variables using wald test.

Solution: The model is of the form $\log \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{31} d_{i31} + \beta_{32} d_{i32}$. For testing the significance of the model, the hypothesis to be tested is $H_0 : \beta_1 = \beta_2 = \beta_{31} = \beta_{32} = 0$. Thus, the likelihood-ratio statistic is $G^2 = -2(\log \ell_0 - \log \ell_M) = -2[-92061.31-(-85956.40)] = 12209.82$ which is very larger than $\chi^2_{0.05}(4) = 1.145$. Therefore, at least one of the explanatory variable is significant.

To identify the significant explanatory variables one by one, the Wald statistics are calculated as shown in the following table.

Variable	t Statistic	95% CI for β	$\widehat{\text{IRR}}$	95% CI for IRR
Intercept	691.46*	(5.4470, 5.4780)*		
Age	30.00*	(0.0056, 0.0064)*	1.0060	(1.0056, 1.0064)*
Gender	-48.34*	(-0.2062, -0.1902)*	0.8202	(0.8137, 0.8268)*
Ambulatory	-82.33*	(-0.3877, -0.3697)*	0.6848	(0.6786, 0.6909)*
Bedridden	-30.76*	(-0.4024, -0.3542)*	0.6850	(0.6687, 0.7017)*

As can be seen, all the three explanatory variables are significantly associated with the CD4 counts of HIV/AIDS patients.

6.3.3 Model Diagnostics

Just as in any model fitting procedure, analysis of residuals is important in fitting poisson regression. Residuals can provide guidance concerning the overall adequacy of the model, assist in verifying assumptions, and can give an indication concerning the appropriateness of the selected link function.

The ordinary or raw residuals are just the differences between the observations and the fitted values, $e_i = y_i - \mu(\boldsymbol{x}_i)$, which have limited usefulness. The Pearson residuals are the standardized differences

$$r_i = \frac{y_i - \mu(\boldsymbol{x}_i)}{\sqrt{\mu(\boldsymbol{x}_i)}}.$$

These residuals fluctuate around zero, following approximately a normal distribution when $\mu(\boldsymbol{x}_i)$ is large. When the model holds, these residuals are less variable than standard normal, however, because the numerator must use the fitted value $\hat{\mu}(\boldsymbol{x}_i)$ rather than the true mean $\mu(\boldsymbol{x}_i)$. Since the sample data determine the fitted value, $[y_i - \hat{\mu}(\boldsymbol{x}_i)]$ tends to be smaller than $[y_i - \mu(\boldsymbol{x}_i)]$.

Since, the standardized residual takes $[y_i - \hat{\mu}(\boldsymbol{x}_i)]$ and divides it by its estimated standard error $\sqrt{\hat{\mu}(\boldsymbol{x}_i)}$, it does have an approximate standard normal distribution when $\mu(\boldsymbol{x}_i)$ is large. With standardized residuals, it is easier to tell when a deviation $[y_i - \hat{\mu}(\boldsymbol{x}_i)]$ is "large".

Components of the deviance are alternative measures of lack of fit. The deviance residuals are $d_i = \pm\sqrt{y_i \log\left[y_i/\hat{\mu}(\boldsymbol{x}_i)\right] - [y_i - \hat{\mu}(\boldsymbol{x}_i)]}; i = 1, 2, \cdots, n$ where the sign is the sign of the ordinary residual. The deviance residuals approach zero when the observed values of the response and the fitted values are closer to each other.

6.4 The Negative-Binomial Regression Model

For a poisson distribution, the variance and the mean are equal. Often count data vary more than the expected. The phenomenon of the data having greater variability than expected is called *over-dispersion*. But, over-dispersion is not an issue in ordinary regression models assuming normally distributed response, because the normal distribution has a separate parameter to describe the variability.

In the presence of over-dispersion, a negative binomial model is should be applied. Like a poisson model, a negative binomial model expresses the log mean response in terms of the explanatory variables. But a negative binomial model has an additional parameter called a *dispersion parameter*. That is, because, the negative binomial distribution has mean $E(Y) = \mu$ and variance $\text{Var}(Y) = \mu + \psi\mu^2$ where $\psi > 0$. The index ψ is a dispersion parameter. As ψ approaches 0, $Var(Y)$ goes to μ and the negative binomial distribution converges to the poisson distribution. The farther ψ falls above 0, the greater the over-dispersion relative to poisson variability.

Example 6.3. Consider example 6.1. The parameter estimates and their corresponding standard errors of the negative binomial regression are given below.

Variable	Parameter Estimate	Standard Error
Intercept	5.4202	0.0867
Age	0.0067	0.0023
Gender	-0.1841	0.0443
Ambulatory	-0.3743	0.0460
Bedridden	-0.6332	0.1066
$\hat{\psi}$	0.6022 [CI: (0.5628,0.6443)]	0.0208

The log-likelihood value of this model is -9083.73 and that of the null model is -9135.30. Compare and contrast the parameter estimates with that of the poisson regression. In addition, compare both models by finding their corresponding AIC values.

Solution: As the dispersion parameter ψ is significantly larger than 0, it assures that the negative binomial regression model is more appropriate than the poisson regression model.

Chapter 7

Loglinear Models for Contingency Tables

A loglinear model is another modeling method, in addition to a logistic regression model, to be used for analyzing categorical data. It is useful to describe association patterns among a set of categorical response variables. The choice of the (logistic regression or loglinear) models depends on the characteristics of the explanatory variables. If the explanatory variables are categorical and/or continuous data, the logistic regression model should be used. If the explanatory variables are categorical data, the loglinear model should be used. Loglinear models are mostly used when at least two variables in a contingency table are response variables.

There are three views of loglinear models. The first is to examine the joint frequency distribution of two or more categorical variables in which results are expressed in terms of a distribution type that the variables jointly display. The second view is to assess the possible dependence of among the variables in which results are expressed in terms of conditional probabilities of states of one variables given other variable(s) levels. The last one is studying association patterns of response variables in which results are expressed in terms of interactions among variables.

For example, suppose we are interested in relationships among gender (Female, Male), smoking (Yes, No), tea drinking (Yes, No) and coffee drinking (Yes, No). From the $2 \times 2 \times 2 \times 2$ contingency table shown in Table 7.1, we can describe $4C_2 = 6$ two-way associations ($C \times T$, $C \times S$, $C \times G$, $T \times S$, $T \times G$, $S \times G$), $4C_3 = 6$ three-way associations ($C \times T \times S$, $C \times T \times G$, $T \times S \times G$, $C \times S \times G$) as well as four main effects. The loglinear model tests whether each association is significant in the model.

In general, when there are sets of categorical response variables and there is no distinction between response and explanatory variables, the loglinear model provides a good statistical analysis for testing associations and interactions among sets of categorical response variables.

Table 7.1: Cross-Classification of Subjects by Gender, Smoking, Tea and Coffee Drinking

			Coffee	
Gender	Smoking	Tea	Yes	No
Female	Yes	Yes	15	5
		No	30	14
	No	Yes	17	8
		No	14	2
Male	Yes	Yes	23	6
		No	15	11
	No	Yes	18	7
		No	5	1

7.1 Loglinear Models for Two-way Tables

Consider an $I \times J$ contingency table that cross-classifies a multinomial sample of n subjects on two categorical responses, X and Y. The cell probabilities are $\{P(X = i, Y = j) = \pi_{ij}\}$ and the observed frequencies are $\{n_{ij}\}$ provided that

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{ij} = 1 \text{ and } \sum_{i=1}^{I}\sum_{j=1}^{J}n_{ij} = n.$$

Thus, the expected frequencies are $\{\mu_{ij} = n\pi_{ij}\}$. Loglinear model uses $\{\mu_{ij}\}$ rather than $\{\pi_{ij}\}$, so they can also apply with poisson sampling for $N = IJ$ independent cell counts $\{Y_{ij}\}$ having $\{\mu_{ij} = E(Y_{ij})\}$.

7.1.1 The Independence Model

Under the assumption of statistical independence of two response variables, $\pi_{ij} = \pi_{i+}\pi_{+j}$. As a result, the expected frequencies are $\mu_{ij} = n\pi_{i+}\pi_{+j}$. Hence, the loglinear model of independence is:

$$\log \mu_{ij} = \log n + \log \pi_{i+} + \log \pi_{+j}.$$

It can be expressed as:

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

where λ is the grand mean of the logarithms of the expected cell frequencies or more specifically,

$$\lambda = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\log \mu_{ij}$$

where I and J indicate the numbers of categories of X and Y. The parameters λ_i^X and λ_j^Y are the main effects of variable X and Y, respectively, which can be recalculated as

$$\lambda_i^X = \frac{1}{J}\sum_{j=1}^{J} \log \mu_{ij} - \lambda$$

and

$$\lambda_j^Y = \frac{1}{I}\sum_{i=1}^{I} \log \mu_{ij} - \lambda.$$

Since the parameters λ_i^X and λ_j^Y are expressed interms of differences from the grand mean λ, the following equation holds:

$$\sum_{i=1}^{I} \lambda_i^X = \sum_{j=1}^{J} \lambda_j^Y = 0.$$

With these constraints, λ_i^X and λ_j^Y are coefficients of dummy variables for the first $(I-1)$ categories of X and $(J-1)$ categories of Y, respectively.

Interpretation of Parameters

The model $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y$ does not distinguish between response and explanatory variables. It treats both jointly as responses, modeling $\{\mu_{ij}\}$ for combinations of their levels. To interpret parameters, however, it is helpful to treat variables asymmetrically.

Consider an $I \times 2$ tables. In category i of X, the logit model equals:

$$\text{logit}[P(Y=1|X=i)] = \log\left[\frac{P(Y=1|X=i)}{P(Y=2|X=i)}\right] = \log\left(\frac{\mu_{i1}/\mu_{i+}}{\mu_{i2}/\mu_{i+}}\right) = \log\left(\frac{\mu_{i1}}{\mu_{i2}}\right)$$

This implies,

$$\text{logit}[P(Y=1|X=i)] = \log \mu_{i1} - \log \mu_{i2} = (\lambda + \lambda_i^X + \lambda_1^Y) - (\lambda + \lambda_i^X + \lambda_2^Y) = \lambda_1^X - \lambda_2^Y.$$

The final term, $\lambda_1^X - \lambda_2^Y$, does not depend on i, that is, $\text{logit}[P(Y=1|X=i)]$ is identical at each level of X. That is, in each category of X, the odds of response 1 of Y is $\exp(\lambda_1^X - \lambda_2^Y)$.

An analogous property holds when $J > 2$. Of course, with a single response variable, logit models apply directly and loglinear models are not needed.

7.1.2 The Saturated Model

The loglinear model discussed before contains only main effect terms. And it rarely fits. Therefore, interaction terms often are necessary to obtain estimates for cell frequencies that are close enough to the observed values.

Models that include all possible main effects and interactions are called saturated models. For example, the saturated loglinear model for two-way contingency tables is:

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

The $\{\lambda_{ij}^{XY}\}$ are association terms that reflect deviations from independence. These represent interactions between X and Y, where by the effect of one variable on μ_{ij} depends on the level of the other. The parameters for interactions can be expressed as differences from the grand mean.

$$\sum_{i=1}^{I} \lambda_{ij}^{XY} = \sum_{j=1}^{J} \lambda_{ij}^{XY} = 0$$

Thus, λ_{ij}^{XY} is the coefficient of the product of dummy variables for λ_i^X and λ_j^Y. The independence model results when all $\lambda_{ij}^{XY} = 0$. The saturated model is the most general model for twoway contingency tables. For it, direct relationships exist between log odds ratios and $\{\lambda_{ij}^{XY}\}$. For example, for 2 × 2 tables,

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

Thus, $\{\lambda_{ij}^{XY}\}$ determine the association.

In practice, unsaturated models are preferable, since their fit smooths the sample data and has simpler interpretations. For tables with at least three variables, unsaturated models can include association terms. Then, loglinear models are more commonly used to describe associations (through two-factor terms) than to describe odds (through single-factor terms).

7.2 Loglinear Models for Three-way Tables

Loglinear models for three-way tables describe their independence and association patterns. A three-way $I \times J \times K$ cross-classification of response variables X, Y and Z has several potential types of independence. The cell probabilities are $\{P(X = i, Y = j, Z = k) = \pi_{ijk}\}$ and the observed frequencies are $\{n_{ijk}\}$ provided that

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} \pi_{ijk} = 1 \text{ and } \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk} = n.$$

Thus, the expected frequencies are $\{\mu_{ijk} = n\pi_{ijk}\}$. Hence, the model also applies to poisson sampling with means $\{\mu_{ijk}\}$.

The general loglinear model for a three-way table is

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

where λ is the grand mean of the logarithm of the expected cell frequencies,

$$\lambda = \frac{1}{IJK}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\log \mu_{ijk}$$

The main effect parameters λ_i^X, λ_j^Y and λ_k^Z can be recalculated as $\lambda_i^X = \frac{1}{JK}\sum_{j=1}^{J}\sum_{k=1}^{K}\log \mu_{ijk} - \lambda$, $\lambda_j^Y = \frac{1}{IK}\sum_{i=1}^{I}\sum_{k=1}^{K}\log \mu_{ijk} - \lambda$ and $\lambda_k^Z = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\log \mu_{ijk} - \lambda$.

Since the parameters λ_i^X, λ_j^Y and λ_k^Z are expressed interms of differences from the grand mean λ, the following equation holds.

$$\sum_{i=1}^{I}\lambda_i^X = \sum_{j=1}^{J}\lambda_j^Y = \sum_{k=1}^{K}\lambda_k^Z = 0.$$

Similarly, the two-way interaction parameters λ_{ij}^{XY}, λ_{ik}^{XZ} and λ_{jk}^{YZ} are calculated as: $\lambda_{ij}^{XY} = \frac{1}{K}\sum_{k=1}^{K}\log \mu_{ijk} - \lambda$, $\lambda_{ik}^{XZ} = \frac{1}{J}\sum_{j=1}^{J}\log \mu_{ijk} - \lambda$ and $\lambda_{jk}^{YZ} = \frac{1}{I}\sum_{i=1}^{I}\log \mu_{ijk} - \lambda$.
As a result,

$$\sum_{i=1}^{I}\lambda_{ij}^{XY} = \sum_{j=1}^{J}\lambda_{ij}^{XY} = \sum_{i=1}^{I}\lambda_{ik}^{XZ} = \sum_{k=1}^{K}\lambda_{ik}^{XZ} = \sum_{j=1}^{J}\lambda_{jk}^{YZ} = \sum_{k=1}^{K}\lambda_{jk}^{YZ} = 0.$$

Also, for the three-way interaction,

$$\sum_{i=1}^{I}\lambda_{ijk}^{XYZ} = \sum_{j=1}^{J}\lambda_{ijk}^{XYZ} = \sum_{k=1}^{K}\lambda_{ijk}^{XYZ} = 0.$$

With dummy variables, λ_{ijk}^{XYZ} is the coefficient of the product of the i^{th} dummy variable for X, j^{th} dummy variable for Y, and k^{th} dummy variable for Z.

The above model includes all possible main effects and interactions which is a saturated model for three-way tables. Saturated models exactly reproduce the observed frequency distribution. There is no degree of freedom left. There is really no business in statistically testing the fit of a saturated model because it does not provide any reduction of data. Saturated models are, however, often used to generate hints as to what parameters might be strong. These hints give only first insights that may have to be revised. Parameter estimates often are correlated and depend partly on the presence or absence of other parameters in the equation.

In general, for $d > 1$ variables in a cross-classification, there are $dC_1 = d$ main effects, $dC_2 = \frac{1}{2}d(d-1)$ two-way interaction terms, $\cdots$, $dC_d = 1$ d-way interaction terms in the saturated model.

7.2.1 Types of Independence

Setting certain parameters equal to zero in the general loglinear model yields different models to be introduced next.

- The three variables are mutually independent when $\pi_{ijk} = \pi_{i++}\pi_{+j+}\pi_{++k}$ for all i, j and k. Mutual independence has loglinear form

  $$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$$

- Variable Y is jointly independent of X and Z when $\pi_{ijk} = \pi_{i+k}\pi_{+j+}$ for all i, j and k. This is ordinary two-way independence between Y and a variable composed of the IK combinations of levels of X and Z. The loglinear model is

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

  Similarly, X could be jointly independent of Y and Z, or Z could be jointly independent of X and Y. Mutual independence implies joint independence of any one variable from the others.

- Variables X and Y are conditionally independent given Z when $\pi_{ij|k} = \pi_{i+|k}\pi_{+j|k}$ or $\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}$ for all i, j and k. Conditional independence of X and Y, given Z, is the loglinear model is

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

  This is a weaker condition than mutual or joint independence. Mutual independence implies that Y is jointly independent of X and Z, which itself implies that X and Y are conditionally independent.

A model that permits all three pairs to be conditional dependent is

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

For this model, the conditional odds ratios between any two variables are identical at each category of the third variable which is shown in the next section. That is, each pair has a homogenous association and the model is called the loglinear model of homogenous association or of no three factor interaction.

To ease referring to the above models discussed, the following table assigns to each model a symbol that lists the highest-order terms for each variable. For instance, the model of conditional independence between X and Y has symbol (XY, YZ), since its highest-order terms are λ_{ik}^{XZ} and λ_{jk}^{YZ}.

Table 7.2: Loglinear Models for Three-Way Contingency Tables

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

7.2.2 The Hierarchical Model

Hierarchical models are models which include all lower-order terms composed from variables contained in a higher-order model term. The models in Table 7.2 are all hierarchical. When the model contains λ_{ij}^{XY}, it also contains λ_i^X and λ_j^Y.

A reason for including lower-order terms is that, otherwise, the statistical significance and the interpretation of a higher-order term depends on how variables are coded. This is undesirable, and with hierarchical models the same results occur no matter how variables are coded. For example, the model $\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_{ij}^{XY}$, it is not hierarchical. It permits association but forces unnatural behavior of expected frequencies, with the pattern depending on constraints used for parameters.

7.2.3 Interpreting Model Parameters

Interpretations of loglinear model parameters use their highest-order terms. For instance, interpretations for the loglinear model of homogenous associations use the two-factor terms to describe conditional odds ratios. At a fixed level k of Z, the conditional association between X and Y uses $(I-1)(J-1)$ odds ratios, such as the local odds ratios

$$\theta_{ij(k)} = \frac{\pi_{ijk}\pi_{i+1,j+1,k}}{\pi_{i,j+1,k}\pi_{i+1,j,k}}, \quad i = 1, 2, \cdots, I-1, \quad j = 1, 2, \cdots, J-1.$$

Similarly, $(I-1)(J-1)$ odds ratios $\{\theta_{i(j)k}\}$ describe XZ conditional association, and $(J-1)(K-1)$ odds ratios $\{\theta_{(i)jk}\}$ describe YZ conditional association. Loglinear models have characterizations using constraints on conditional odds ratios. For instance, conditional independence of X and Y is equivalent to $\{\theta_{ij(k)} = 1, \quad i = 1, 2, \cdots, I-1, \quad j = 1, 2, \cdots, J-1, \quad k = 1, 2, \cdots, K\}$.

The two-factor parameters relate directly to the conditional odds ratios. Thus, $\log \theta_{ij(k)}$ yields

$$\log \theta_{ij(k)} = \log\left(\frac{\mu_{ijk}\mu_{i+1,j+1,k}}{\mu_{i+1,j,k}\mu_{1,j+1,k}}\right) = \lambda^{XY}_{ij} + \lambda^{XY}_{i+1,j+1} - \lambda^{XY}_{i,j+1} - \lambda^{XY}_{i+1,j}.$$

Since the right-hand side is the same for all k, an absence of three-factor interaction is equivalent to

$$\theta_{ij(1)} = \theta_{ij(2)} = \cdots = \theta_{ij(K)} \quad \text{for all } i \text{ and } j.$$

The same argument for the other conditional odds ratios shows that model (XY, XZ, YZ) is also equivalent to

$$\theta_{i(1)k} = \theta_{i(2)k} = \cdots = \theta_{i(J)k} \quad \text{for all } i \text{ and } k$$

and to

$$\theta_{(1)jk} = \theta_{(2)jk} = \cdots = \theta_{(I)jk} \quad \text{for all } j \text{ and } k.$$

Any model not having the three-factor interaction term has a homogeneous association for each pair of variables.

The λ^{XYZ}_{ijk} term in the general model refers to three-factor interaction. It describes how the odds ratio between two variables changes across categories of the third. Consider a $2 \times 2 \times 2$ tables. By direct substitution of the general model formula,

$$\begin{aligned}\log\left(\frac{\theta_{11(1)}}{\theta_{11(2)}}\right) =& \log\left[\frac{(\mu_{111}\mu_{221})/(\mu_{121}\mu_{211})}{(\mu_{112}\mu_{222})/(\mu_{122}\mu_{212})}\right] \\ =& (\lambda^{XYZ}_{111} + \lambda^{XYZ}_{221} - \lambda^{XYZ}_{121} - \lambda^{XYZ}_{211}) - (\lambda^{XYZ}_{112} + \lambda^{XYZ}_{222} - \lambda^{XYZ}_{122} - \lambda^{XYZ}_{212})\end{aligned}$$

For constraints setting the second-category parameters equal to 0, this log ratio of odds ratios equals λ^{XYZ}_{111}. When $\lambda^{XYZ}_{111} = 0$, $\theta_{11(1)} = \theta_{11(2)}$, giving homogeneous XY association.

7.3 Fitting Loglinear Models

When fitting loglinear models, the first goal is to find a model that describes the cross-classified data such that there are, statistically, only random discrepancies between the observed and expected frequencies. Another goal is to get more parsimonious models than the saturated model which does not provide any data reduction and has no more use the raw data them selves. Parsimonious models contain as few parameters as possible and interactions of the lowest possible order.

7.3.1 Specification and Estimation

The maximum number of independent parameters for the interaction between two variables, say X with I categories and Y with J categories, is $(I-1)(J-1)$. The maximum number of independent parameters for the interaction between three variables, say X with I categories, Y with J categories and Z with K categories, is $(I-1)(J-1)(K-1)$, and so forth.

Using the design matrices, the general loglinear model can be formulated in a fashion analogous to the general linear model, $\log \boldsymbol{\mu} = \boldsymbol{X\beta}$ where $\boldsymbol{\mu}$ is the vector of the expected cell frequencies, $\boldsymbol{X}$ is the design matrix that contains one vector per main effect or interaction parameter and $\boldsymbol{\beta}$ is a vector of parameters.

Consider the cross-classification by three dichotomous variables: Gender (G), Coffee Drinking (C) and Smoking (S). Table 7.3 contains the design matrix for the $G \times C \times S$ cross-classification. This table contains four blocks of vectors. The first block contains the constant vector. In the general linear model, a constant vector of ones has the effect that the intercept parameter estimate is equal to the arithmetic mean of the response variable. In the general loglinear model, this constant vector yields, for parameter λ, the arithmetic mean of the logarithms of the expected cell frequencies.

Table 7.3: Design Matrix for 2 × 2 × 2 Cross-Classification (Effect Coding)

	Vectors in Design Matrix							
		Main Effects			Two-Way Interactions			Three-Way Interaction
μ_{ijk}	Constant	G	C	S	$G \times C$	$G \times S$	$C \times S$	$G \times C \times S$
μ_{111}	1	1	1	1	1	1	1	1
μ_{112}	1	1	1	-1	1	-1	-1	-1
μ_{121}	1	1	-1	1	-1	1	-1	-1
μ_{122}	1	1	-1	-1	-1	-1	1	1
μ_{211}	1	-1	1	1	-1	-1	1	-1
μ_{212}	1	-1	1	-1	-1	1	-1	1
μ_{221}	1	-1	-1	1	1	-1	-1	1
μ_{222}	1	-1	-1	-1	1	1	1	-1

The second block contains the vectors for the main effects of the variables: G, C and S. Here, effect coding is used to generate the vectors for variable main effects. Alternatives include dummy coding which are equivalent in the sense that they allow one to test exactly the same hypothesis. For didactical purposes, however, the effect coding is preferred which makes it easier to identify group of contrasted cells.

The third block of the table contains the vectors for the two-way interactions. Just as with effect coding in the general linear model, these vectors result from element-wise multiplication of main effect vectors of the interacting variables.

The fourth block of coding vectors contains the effect of coding vector for the three-way interaction of the variables: G, C and S.

Interpretation of the vectors proceeds as follows. Interactions modify main effects under consideration of categories of other variables. For instance, the first interaction vector in the block of two-way interaction vectors in Table 7.3 describes the interaction between variables G and C. The main effect of variable C contrasts states 1 and 2. The levels of variable G are not considered. In other words, it is assumed that this contrast is the same across the two levels of variable E. The interaction term $G \times C$ repeats the main effect statement for C only for the first category of G, that is, in the upper half of the vector. In the lower half, this contrast takes the opposite direction.

The $G \times S$ and $C \times S$ interactions can be interpreted in an analogous fashion. Accordingly, the $G \times C \times S$ interaction is a modification of the $G \times C$ interaction that considers the categories of variable S. Or it can be seen as a modification of $G \times S$ under consideration of C, or, as a modification of $C \times S$ under consideration of G.

One the models to be estimated are specified, then the method of maximum likelihood estimation technique can be used. The estimation process for loglinear models involves two steps. The first step is the calculation of estimates for the values of the λ parameters. The second step is the calculation of estimated expected frequencies using the λ parameter estimates.

7.3.2 Statistical Significance Tests

In explanatory research, theories allow to derive models that reflect propositions of these theories. These propositions are then translated into patterns of variable relationships and tested using loglinear models that include these patterns. In many instances, theories allow to derive more than one plausible model. Differences between these models may concern parsimony and type of association pattern, discrepancy pattern or sampling distribution. There are many ways to compare competing models. If competing models operate at different hierarchical levels, differences between these models can be statistically tested.

Significance tests in loglinear models are used to determine whether the model fits, whether the parameters are statistically significant, and whether there is a difference between the observed and expected frequency of each cell (that is, residual analysis).

- **Goodness-of-fit Tests:** The goodness-of-fit test of a model determines whether a model adequately describes the observed frequency distribution. For this purpose, there are two commonly used tests; the likelihood ratio (G^2) test and the Pearson chi-square (G^2) test. They are described, respectively,

$$G^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} n_{ijk}\log\left(\frac{n_{ijk}}{\hat{\mu}_{ijk}}\right) \text{ and } \chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\frac{(n_{ijk}-\hat{\mu}_{ijk})^2}{\hat{\mu}_{ijk}}.$$

Both of these tests provide good χ^2 approximations and are asymptotically equivalent. The question is how to estimate the degrees of freedom in the design matrix approach to loglinear modeling. In fact, it is simple. Let t be the number of cells in the cross-classification and let v be the number of columns in the design matrix, including the constant vector, with $v \leq t$. Then, the degrees of freedom for a given model is calculated as $df = t - v$.

If the values of G^2 and χ^2 are smaller than the critical value, then it indicates that the model adequately fits. But, it should be noted that, unlike many applications of general linear model, in the applications of loglinear models the emphasis is on both the overall model fit and significance of effect parameters. Whereas regression parameter estimates typically are interpreted when they are statistically significant, regardless of model fit, loglinear model parameter estimates are interpreted only if the model provides acceptable fit.

- **Significance of Parameters:** If the model adequately fits, next statistically significant parameters of the target model are to be interpreted. For each vector in the design matrix, the parameter λ is estimated as $\hat{\lambda}$ with an estimated standard error $\widehat{SE}(\hat{\lambda})$. Then, the value of test statistic

$$z = \frac{\hat{\lambda}}{\widehat{SE}(\hat{\lambda})}$$

can be compared with the critical value of a standard normal distribution. The parameter is significant if the value of $|z|$ is greater than the critical value.

- **Residual Analysis:** The third component of statistical significance testing in loglinear modeling involves residual analysis. The raw residuals are $(n_{ijk} - \hat{\mu}_{ijk})$. More frequently, standardized residuals defined as

$$\frac{n_{ijk} - \hat{\mu}_{ijk}}{\sqrt{\hat{\mu}_{ijk}}}$$

are used. If the model fits, each standardized residual is approximately normally distributed with mean 0 and variance 1.

In general, the loglinear model fitting process involves four steps: specifications of models to be tested, estimation of the models, significance tests and finally interpretation of results. Interpretation of results reflects the goal of the analysis. If it is a goal to fit a model, the overall goodness of fit results for all fitting models are evaluated with respect to substantive assumptions and such desiderates as parsimony. If special hypotheses are tested, the parameters for these hypothesis, their meaning and statistical significance are important.

Example 7.1. A sample of $n = 516$ adults are cross-classified according to three variables: Marital Status (M; 1=Married, 2=Single), Gender (G; 1=Male, 2=Female), and Size of Social Network (S; 1=Small Social Network, 2=Large Social Network). The cross-classification provides a $2 \times 2 \times 2$ table. We will analyze this data under the assumption that Marital Status and Gender interact such that older women are less likely to be married than older men, and that large networks are more likely among married people.

Table 7.4: Cross-Classification of Subjects by Marital Status, Gender and Size of Social Network

		Social Network (S)	
Marital Status (M)	Gender (G)	Small (1)	Large (2)
Married (1)	Male (1)	48	87
	Female (2)	5	14
Single (2)	Male (1)	78	45
	Female (2)	130	109

The goal is to find a fitting and parsimonious model that reflect our assumptions. To achieve this, the four steps of fitting and testing in loglinear modeling are to be followed.

- **Step 1: Specification of the Loglinear Models.** The current analysis is explanatory in the sense that we have explicit assumptions. These assumptions are explicit enough so that they can be rejected by data. Specifically, the assumptions that older women are less frequently married than older men, and that larger networks are more frequent that smaller networks among married people fail to be confirmed. To test these assumptions, the following models are specified.

  1. The null model postulates that variables do not interact. That is, the main effects are sufficient to explain the frequency distribution. The null model is tested for two reasons. First, by fitting the null model, we have a model at a relatively low hierarchy level that can be used for comparisons with the finally accepted model. It is one of the requirement that the finally accepted model not only fit by itself, but that it also be statistically significantly better than the null model. Second, if the null model provides such a good fit that statistically significant improvements are not possible, then there is no need to test more complex, less parsimonious models.

2. The target model is the model that is closest to our assumptions. It provides vectors in the design matrix for each of the main effects and interactions involved. Table 7.5 contains the design matrices for both the null and the target model.

The first two columns of Table 7.5 contain the configuration indexes and the observed frequencies. Column 3, 4 and 5 contain vectors for the main effects of variable M, G and S. The elements of the interaction vector result from element-wise multiplication of the first and the second vectors of the design matrix. The last vector is the translation of the assumption that among married people, large networks are more likely than small networks. This vector contrasts two groups of individuals, both composed of married people. The first group involved in the contrast, marked by -1, involves men and women with small networks. The second group, marked by 1, involves men and women with large networks. The single members of this sample are not involved in this contrast and are in cells marked by 0's. Table 7.5 contains the vectors necessary for testing both the null model and the target model. The only exception is the constant vector of 1's, which is implied. The first three vectors are the only ones needed for the null model. All the five vectors are needed for the target model.

Table 7.5: Design Matrices for $M \times G \times S$ Cross-Classification in Table 7.4

		Design Matrix Vectors				
		Main Effects			Interaction	
Configuration (MGS)	n_{ijk}	M	G	S	$M \times G$	Special Contrast
111	48	1	1	1	1	-1
112	87	1	1	-1	1	1
121	5	1	-1	1	-1	-1
122	14	1	-1	-1	-1	1
211	78	-1	1	1	-1	0
212	45	-1	1	-1	-1	0
221	130	-1	-1	1	1	0
222	109	-1	-1	-1	1	0

- **Step 2: Estimation of the Null and Target Models.** As discussed before, estimation of loglinear models involves two main steps. The first step involves parameter estimation while the second step involves estimation of expected cell frequencies and residuals.

  Table 7.6 and Table 7.7 present results from the first and second step for both models, respectively.

Table 7.6: Parameter Estimates for the Null Model and Target Model Specified in Table 7.5

	Null Model			Target Model		
Parameter	$\hat{\lambda}$	$\widehat{SE}(\hat{\lambda})$	z	$\hat{\lambda}$	$\widehat{SE}(\hat{\lambda})$	z
Main Effect M	-0.43	0.05	-8.60*	-0.63	0.07	-9.00*
Main Effect G	0.00	0.04	0.00	0.32	0.07	4.57*
Main Effect S	0.01	0.04	0.25	0.15	0.05	3.00*
Interaction $M \times G$	-	-	-	0.66	0.07	9.43*
Special Contrast	-	-	-	0.47	0.10	4.70*

*statistically significantly different from 0.

Table 7.7: Expected Frequencies and Standardized Residuals from the Null and Target Models for Data in Table 7.4

		Null Model		Target Model	
(MGS)	n_{ijk}	$\hat{\mu}$	Standardized Residual	$\hat{\mu}$	Standardized Residual
111	48	38.95	1.45	46.46	0.23
112	87	38.05	7.94*	88.54	-0.16
121	5	38.05	-5.44*	6.54	-0.60
122	14	38.05	-3.90*	12.46	0.44
211	78	91.55	-1.42	70.67	0.87
212	45	89.45	-4.70*	52.33	-1.01
221	130	91.55	4.02*	137.33	-0.63
222	109	89.45	2.07*	101.67	0.73

*indicates statistically significant deviation of residual from zero.

- **Step 3: Significance Tests.**
  - **Goodness-of-fit Tests:** To evaluate the overall model fit, the likelihood ratio test G^2 and the Pearson chi-square test χ^2 are used. That is,

$$G^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} n_{ijk} \log\left(\frac{n_{ijk}}{\hat{\mu}_{ijk}}\right) \text{ and } \chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2} \frac{(n_{ijk}-\hat{\mu})^2}{\hat{\mu}_{ijk}}.$$

As a result, for the null model, $G^2 = 162.864$ and $\chi^2 = 154.372$, and the critical value is $\chi^2_{0.05}(8-4) = \chi^2_{0.05}(4) = 9.4877$. This implies, the deviations of the observed frequencies from the expected frequencies are not random and thus reject the null model. Therefore, the three parameters for the null model cannot be interpreted.

For the target model, $G^2 = 3.380$ and $\chi^2 = 3.334$, and the critical value is $\chi^2_{0.05}(8-6) = \chi^2_{0.05}(2) = 5.991$. Hence, the model that involves three main effect parameters, one interaction and one special contrast fits very well. In other words, the target model provides a very good rendering of the observed frequency distribution.

The parameters estimated for the two loglinear models, their standard errors, and their z statistics are appear in Table 7.6. Before looking at the significant of parameters in the target model, it must be checked whether the target model shows a significant improvement over the null model. The difference in the likelihood ratio is $\Delta G^2 = 162.864 - 3.380 = 159.484$ and the difference in the degrees of freedom is $\Delta df = 4 - 2 = 2$. The ΔG^2 is statistically significant and we can conclude that the target model not only fits but also provides a significant improvement over the null model.

- **Significance of Parameters:** Now, we can move to significance tests concerning the parameter estimates. As can be seen from Table 7.6, only the main effect M parameter for the null model and all the five parameters of the target model are statistically significant.
- **Significance of Residuals:** The null model in Table 7.7 suggests that six out of eight estimated expected frequencies deviate significantly from the observed frequency. This is another indicator implying poor fit of the main effects model. The target model in the same table suggests none of the standardized residuals is significant. Thus, the target model fits very well even at the level of residuals.

- **Step 4: Interpretation of Results.** Estimating the three main effect parameters first makes sure that the marginal totals are reproduced. These marginal totals are Married = 154, Single = 362, Male = 258, Female = 258, Small Networks = 261 and Large Networks = 255. Parameter interpretation must always consider the other variables in the equation. The interaction suggests that the ratio of married men to married women is not the same as the ratio of single men to single women in the population under study. The special contrast indicates that, among married people, large networks are more likely than small networks, considering that these two statements account for statistically significant portions of the variation in the frequency table.

Note, however, the main effect M parameter for the null model is significant, but its estimate cannot be interpreted because the model does not fit.

Bibliography

Agresti, A. (2007). An Introduction to Categorical Data Analysis. 2nd ed. *Wiley Series in Probability and Statistics.*

Agresti, A. (2002). Categorical Data Analysis. 2nd ed. *Wiley Series in Probability and Statistics.*

David G.K. and Mitchel K. (2010) Logistic Regression: A Self-Learning Text, 3rd ed. *Springer Science+Business Media*

Eye, A. V., and Niedermeier, K. E. (1999) Statistical Analysis of Longitudinal Categorical Data in the Scocial and Behavioural Sciences - An Introduction with Computer Illustrations. *Lawrence Erlbaum Associates, Inc.*

Hosmer, D.W., and Lemeshow, S. (2000) Applied Logistic Regression, 2nd ed. *New York: Wiley.*

Kleinbaum, D.G., Kupper, L.L., Nizam, A., and Muller, K.E., (2008) Applied Regression Analysis and Other Multivariable Methods, 4th ed. *Duxbury Press/Cengage Learning.*

Ronald, P.C., and Jeffrey, K.S., (1997) Applied Statistics and the SAS Programming Language, 4th ed. *Prentice Hall.*

Seid, A., (2015). Multilevel Modeling of the Progression of HIV/AIDS Disease Among Patients Under HAART Treatment. *Ann. Data. Sci.* Springer-Verlag Berlin Heidelberg. 02(02):217-230.

Seid, A., Muluye, G., Belay, B. and Yehenew, G., (2014). Joint modeling of longitudinal CD4 counts and time-to-default from HAART treatment: a comparison of separate and joint models. *Electron. J. Appl. Stat. Anal.* Salento Univeristy. 07(02):292-314.