Advanced Topics from "And Introduction to Categorical Data Analysis" Chapters 7, 8, 9.

By: Tyler Diestel

A Statistics Senior Project

Cal Poly San Luis Obispo

Winter 2011

Introduction

After finishing Stat 418: An Introduction to Categorical Analysis in winter 2010, I was curious about the many methods of categorical analysis that we were unable to cover. This was nothing more than me just being curious, until I began some of my classes in the spring. Twice throughout my spring quarter there were opportunities to use some of these advanced methods, yet I did not know how to use them properly. It first occurred in my Stat 465: Statistical Consulting class, when one of our clients had some before and after surveys from a camp. These surveys could have been analyzed using matched pair categorical analysis, yet I did not know the methods well enough at the time to run them. An opportunity again arose in my Stat 427: Probability Theory class to better understand McNemar's test, but still I did not have any previous knowledge to bring to the table. At last, I finally decided that I should learn these advanced categorical analysis methods and write a tutorial for anyone else who might be in my shoes in the future.

In my Stat 418 class, by the end of the year, we reach Section 7.1.3 in *An Introduction to Categorical Data Analysis*, (Second Edition), by Alan Agresti. For my senior project, I investigated selected statistical methods from where Stat 418 ended to Section 9.2 of Agresti's text. I have now compiled all of my notes, SAS code, and proofs into this tutorial. To be faithful to Agresti's text, I have kept all of the same chapter, section, subsection, and table headers, so the reader can easily reference the text. At times, I will reference direct models that Agresti used. To indicate that I am directly referring to *An Introduction to Categorical Data Analysis*, I will use the abbreviation 'ICDA.'

This paper begins in Chapter 7 which deals with loglinear models for contingency tables. I will start on Section 7.1.4, which explains how to create loglinear models for three-way contingency tables. Section 7.1.5 will then show how a model with two-factor parameters describes conditional association. Then in Section 7.1.6, we will go over an example to use all of the tools we have learned thus far when analyzing data about alcohol, cigarette and marijuana use.

Section 7.2 describes how we can make inferences from these loglinear models. The first subsection deals with model fit. Then in Section 7.2.2 we start to analyze the cell residuals of these models. In Section 7.2.3, we construct a test to determine if there is conditional association. We will then create conditional odds ratios and confidence intervals for them. Section 7.2.5, applies everything that we have learned to models that are larger than three-way tables. We will then follow this up with an example that deals with automobile accidents in Section 7.2.6. In the next section, we will dive more into the understanding of the three-factor interaction. Lastly, in Section 7.2.8 we will discuss the difference between statistical significance and practical significance.

Section 7.3 ties Chapter 7 with Chapter 4 as it relates loglinear models to logistic models. Section 7.3.1 uses logistic models to interpret loglinear models. Then in the next section we revisit the example in Section 7.2.6 and apply this new perspective. Then in Section 7.3.3 and Section 7.3.4 we discuss when we would use loglinear models over logistic models and vice versa.

We then jump to Chapter 8 where we create models for matched pairs. Section 8.1 compares dependent proportions. It begins by explaining the McNemar Test in Section 8.1.1 and then finishes by estimating the difference in marginal proportions.

In Section 8.2 we construct logistic models to analyze matched pairs. We begin by creating marginal models for marginal proportions. Then we are introduced to subjected-specific tables and population-average tables. In Section 8.2.3 we examine how one can use conditional logistic regression for matched paired data. Section 8.2.4 is very similar to the previous section but its data deals with case-controlled studies. Then in Section 8.2.5 we make a connection between the McNemar Test and the Cochran-Mantel-Haenszel Test.

Section 8.3 discusses how to interpret margins of square contingency tables. The first section explains marginal homogeneity and nominal classification. This is then followed up by an example which deals with different brands of coffee. In Section 8.3.3 we discuss how to analyze ordinal paired data. Then we use this new method to analyze data in an example in Section 8.3.4.

Section 8.4 deals with symmetry and quasi-symmetry models. In the first subsection we are introduced to the symmetry model. Then in Section 8.4.2 we create the quasi-symmetry model, which is a more realistic version of the symmetry model.

We then skip to Section 8.6 where we learn the Bradley-Terry model that ranks subjects in paired situations. In Section 8.6.2, we are able to run this model on a data set of men's tennis players. Next in Section 8.6.2.a I have added my own example which analyzes data from the 2010 Major League Baseball season.

The last chapter we will look at will be Chapter 9, which creates models for correlated and clustered responses. Section 9.1 discusses marginal and conditional models. This has three subsections. The first explains how marginal models can be used

for data with a clustered binary response. We will then run an example in the next section, and lastly, introduce conditional models for a repeated response.

Section 9.2 will be our last section. Here we will first explain the quasi-likelihood approach, and then relate that to the General Estimating Equation (GEE) methodology in Section 9.2.2. The next two sections will be examples to help us understand GEE. Lastly, we will find out the limitations that the GEE has compared to the maximum likelihood method.

Table of Contents

7.1.4 Loglinear Models for Three-Way Tables

A three-way table can be interpreted much like a two-way table in which we would like to test independence. To examine independence, we are going to look at several models used to find expected cell counts.

The first model we are going to look at is the *mutual independence model*:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z.$$

This model is like the two-way model of independence equation 7.1 in ICDA. It treats each variable as conditionally independent and marginally independent. For review, *conditional independence* is when the conditional odds ratio is equal to 1, and *marginal independence* is when the marginal odds ratio is equal to 1. This model is best used when each of the X, Y, and Z variables is unaffected by the other variables, thus causing them to be independent of each other. Below is a proof of why this is the model used for the independence model:

$$Under\ the\ assumption\ of\ independence,$$

$$\pi_{ijk} = (\pi_{i++})(\pi_{+j+})(\pi_{++k})$$

$$\mu_{ijk} = n * \pi_{ijk} = n(\pi_{i++})(\pi_{+j+})(\pi_{++k})$$

$$\log(\mu_{ijk}) = \log\Big(n(\pi_{i++})(\pi_{+j+})(\pi_{++k})\Big)$$

$$\log(\mu_{ijk}) = \log(n) + log(\pi_{i++}) + log(\pi_{+j+}) + log(\pi_{++k})$$

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

We use (X,Y,Z) to denote this model.

The next model allows for independence between only two of the factors instead of all three:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad (7.4).$$

This model allows there to be an association between X and Z controlling for Y, which can be seen in the λ_{ik}^{XZ} term. Likewise, this model allows there to be an association between Y and Z controlling for X, which can be seen in the λ_{jk}^{YZ} term. This model also shows conditional independence between X and Y, controlling for Z. This is why there is no term for the X and Y interaction because $\lambda_{ij}^{XY} = 0$ in this case. We use (XZ, YZ) to denote this model.

To describe an interaction between all three pairs of variables we would use the homogeneous association model:

$$\log(\mu_{ijk}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ} \quad (7.5).$$

This allows all three pair of variables to have conditional association, which means that all odds ratios for any two variables are equal at all levels of the third variable. This is known as *homogeneous association*. To denote this model we will use (XY, XZ, YZ).

The last model contains a term for each variable, each pair of variables, and a term that explains an interaction between all three variables. This is known as the saturated model:

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

The model is the most general model and has a perfect fit because it describes all of the possible interactions. Below is proof of why this must be the saturated model:

$$In\ this\ saturated\ model, there\ is\ a\ single\ constant\ parameter\ (\lambda),$$

$$(I-1) nonredundant\ \lambda_i^X\ parameters, (J-1) nonredundant\ \lambda_j^Y\ parameters,$$

$$(K-1) nonredundant\ \lambda_k^Z\ parameters,$$

$$(I-1)(J-1) nonredundant\ \lambda_{ij}^{XY}\ parameters,$$

$$(I-1)(K-1) nonredundant\ \lambda_{ik}^{XZ}\ parameters,$$

$$(K-1)(J-1) nonredundant\ \lambda_{jk}^{YZ}\ parameters, and$$

$$(I-1)(J-1)(K-1) nonredundant\ \lambda_{ijk}^{XYZ}\ parameters.$$

$$So\ the\ total\ number\ of\ parameters\ equals:$$

$$1+(I-1)+(J-1)+(K-1)+(I-1)(J-1)+(I-1)(K-1)+(J-1)(Z-1)$$

$$+(I-1)(J-1)(K-1)$$

$$\rightarrow IJ+K-1+IK-I-K+1+JK-J-K+1+(IJ-I-J+1)(k-1)$$

$$\rightarrow 1-K-I-J+IJ+IK+JK+IJK-IK-JK+K-IJ+I+J-1$$

$$\rightarrow IJK$$

$$So\ this\ model\ has\ as\ many\ parameters\ as\ observed\ cell\ counts. It\ is\ the$$

$$saturated\ loglinear\ model, having\ the\ maximum\ possible\ number\ of\ parameters$$

$$uniquely\ estimated\ by\ the\ data.$$

7.1.5 Two-Factor Parameters Describe Conditional Association

This next section dives further into the homogeneous association model (Equation 7.5 ICDA)

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

A proof below has been provided to show that this model has the characteristic that the conditional odds ratios between any two variables are the same at each level of the third variable.

$$Let\ z = k\ so\ that\ z\ is\ fixed,$$

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

$$= (\lambda + \lambda_1^X + \lambda_1^Y + \lambda_k^Z + \lambda_{11}^{XY} + \lambda_{1k}^{XZ} + \lambda_{1k}^{YZ}) + (\lambda + \lambda_2^X + \lambda_2^Y + \lambda_k^Z + \lambda_{22}^{XY} + \lambda_{2k}^{XZ} + \lambda_{2k}^{YZ})$$

$$-(\lambda + \lambda_1^X + \lambda_2^Y + \lambda_k^Z + \lambda_{12}^{XY} + \lambda_{1k}^{XZ} + \lambda_{2k}^{YZ}) - (\lambda + \lambda_2^X + \lambda_1^Y + \lambda_k^Z + \lambda_{21}^{XY} + \lambda_{2k}^{XZ} + \lambda_{1k}^{YZ})$$

$$= \lambda_{11}^{XY} + \lambda_{22}^{XY} - \lambda_{12}^{XY} - \lambda_{21}^{XY} \qquad (7.6\ ICDA)$$

The final result has no 'Z' terms contained in it, so it does not depend on *k.* This means that the model has equal odds ratios at all levels of 'Z.' We can similarly prove the same results for all XZ odds ratios at different levels of Y and all YZ odds ratios at different levels of X. Models that do not have a three factor term λ_{ijk}^{XYZ} have homogeneous association.

7.1.6 Example: Alcohol, Cigarette and Marijuana Use

In a survey at Wright State University School of Medicine and the United Health Services in Dayton, Ohio, high school seniors were asked if they have ever tried alcohol, cigarettes, and/or marijuana. The data is below in Table 7.3

Table 7.3: Alcohol, Cigarette, and Marijuana Use

		Marijuana Use	
Alcohol Use	**Cigarette Use**	**Yes**	**No**
Yes	**Yes**	911	538
	No	44	456
No	**Yes**	3	43
	No	2	279

In order to find the fitted values for this data, we will have to put our data into SAS. To do this, we will need to insert three categorical columns to represent each of the three variables and one numeric column to represent the counts.

```
data Alcohol;
input alcohol $ cigarettes $ marijuana $ count @@;
cards;
```

```
y y y 911 y y n 538
y n y 44  y n n 456
n y y 3   n y n 43
n n y 2   n n n 279
;
run;
```

From here we will produce many models using `proc genmod`.

```
/* this will generate results for our (A,C,M) model */
proc genmod data=Alcohol order=data;
class alcohol cigarettes marijuana;
model count = alcohol cigarettes marijuana / dist=poi link=log;
run;

/* this will generate results for our (AC,M) model */
proc genmod data=Alcohol order=data;
class alcohol cigarettes marijuana;
model count = alcohol cigarettes marijuana alcohol*cigarettes/ dist=poi
link=log;
run;

/* this will generate results for our (AM,CM) model */
proc genmod data=Alcohol order=data;
class alcohol cigarettes marijuana;
model count = alcohol cigarettes marijuana
      alcohol*marijuana cigarettes*marijuana/ dist=poi link=log;
run;

/* this will generate results for our (AC,AM,CM) model */
proc genmod data=Alcohol order=data;
class alcohol cigarettes marijuana;
model count = alcohol cigarettes marijuana alcohol*cigarettes
      alcohol*marijuana cigarettes*marijuana/ dist=poi link=log;
run;

/* this will generate results for our (ACM) model */
proc genmod data=Alcohol order=data;
class alcohol cigarettes marijuana;
model count = alcohol cigarettes marijuana alcohol*cigarettes
      alcohol*marijuana cigarettes*marijuana
alcohol*cigarettes*marijuana / dist=poi link=log;
run;
```

Each one of these `procs` represents a different model. To further explain what each of these does and what it outputs, I will pick only one model to examine.

The model that we will examine is the (AM, CM) model, which can be written out $\log(\mu_{ijk}) = \lambda + \lambda_i^A + \lambda_j^C + \lambda_k^M + \lambda_{ik}^{AM} + \lambda_{jk}^{CM}$. The proc that produces the output we would like for this model and the actual output is below:

```
/* this will generate results for our (AM,CM) model */
proc genmod data=Alcohol order=data;
class alcohol cigarettes marijuana;
model count = alcohol cigarettes marijuana
      alcohol*marijuana cigarettes*marijuana/ dist=poi link=log;
run;
```

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Intercept			1	5.1921	0.0609	5.0727	5.3114	7273.54
alcohol	y		1	**1.1272**	0.0641	1.0015	1.2529	309.01
alcohol	n		0	0.0000	0.0000	0.0000	0.0000	.
cigarettes	y		1	**-0.2351**	0.0555	-0.3439	-0.1263	17.94
cigarettes	n		0	0.0000	0.0000	0.0000	0.0000	.
marijuana	y		1	**-6.6209**	0.4737	-7.5493	-5.6924	195.35
marijuana	n		0	0.0000	0.0000	0.0000	0.0000	.
alcohol*marijuana	y	y	1	**4.1251**	0.4529	3.2373	5.0128	82.94
alcohol*marijuana	y	n	0	0.0000	0.0000	0.0000	0.0000	.
alcohol*marijuana	n	y	0	0.0000	0.0000	0.0000	0.0000	.
alcohol*marijuana	n	n	0	0.0000	0.0000	0.0000	0.0000	.
cigarettes*marijuana	y	y	1	**3.2243**	0.1610	2.9088	3.5398	401.17
cigarettes*marijuana	y	n	0	0.0000	0.0000	0.0000	0.0000	.
cigarettes*marijuana	n	y	0	0.0000	0.0000	0.0000	0.0000	.
cigarettes*marijuana	n	n	0	0.0000	0.0000	0.0000	0.0000	.
Scale			0	1.0000	0.0000	1.0000	1.0000	

NOTE: The scale parameter was held fixed.

Next, we have to find the fitted values for this model. We will start by fitting the value for the amount of students who have used all three substances. To do this look above in the output and find all of the values where there is a 'y' and 'y y'. These values are all the ones that represent a person has tried that item. We then take the sum of these numbers including the intercept and exponentiate that value:

$$e^{5.1921 + 1.1272 - 0.2351 - 6.6209 + 4.1251 + 3.2243} \approx 909.24$$

The number 909.24 is our expected count for the number of people who answered that they have tried alcohol, cigarettes, and marijuana. We will do this for all of the different combinations of the three variables and produce Table 7.4:

Table 7.4: Fitted Values for Loglinear Model (AM, CM)

Alcohol Use	Cigarette Use	Marijuana Use: Yes	Marijuana Use: No
Yes	Yes	909.24	438.84
	No	45.76	555.16
No	Yes	4.76	142.16
	No	.24	179.84

With these values we are able to find the estimated conditional and marginal odds ratios. For example, the estimated AM conditional odds ratio can be calculated as such:

$$\frac{\hat{\mu}_{yyy}*\hat{\mu}_{nny}}{\hat{\mu}_{yny}*\hat{\mu}_{nyy}} = \frac{909.24*142.16}{438.84*4.76} = \frac{\hat{\mu}_{yyn}*\hat{\mu}_{nnn}}{\hat{\mu}_{ynn}*\hat{\mu}_{nyn}} = \frac{45.76*179.84}{555.16*.24} = 61.9,$$

and the AM marginal odds ratio can be calculated:

$$\frac{\hat{\mu}_{yy+} * \hat{\mu}_{nn+}}{\hat{\mu}_{yn+} * \hat{\mu}_{ny+}} = \frac{(909.24 + 45.76) * (142.16 + 179.84)}{(438.84 + 555.16) * (4.76 + .24)} = 61.9$$

An interpretation of the AM conditional odds ratio is that for each level of C, students who have tried marijuana have estimated odds of having drunk alcohol that are 61.9 times the estimated odds for students who have not tried marijuana. The difference between the interpretation of the conditional odds ratio and the marginal odds ratio is that the condition odds ratio controls for C, and the marginal odds ratio does not. All of the conditional and marginal odds ratios for this model and a few others can be found in Table 7.5 of ICDA.

Another way to calculate the estimated conditional odds ratio would be to use the Equation 7.6 ICDA. In our model (AM, CM), we can use this equation to find the estimated conditional odds ratio for AM:

$$e^{\hat{\lambda}_{yy}^{AM} + \hat{\lambda}_{nn}^{AM} - \hat{\lambda}_{yn}^{AM} - \hat{\lambda}_{ny}^{AM}}$$

$$= e^{4.1251+0-0-0} = e^{4.1251} = 61.9.$$

7.2 Inference for Loglinear Models

Now that we have learned about the many different models we can make, we are going to learn about which model we want to select in a given situation. We want to choose the best model because we will be able to make better estimations, which will lead to more accurate inferences. In the following section, we will discover how to do this. Some of this material might seem like review from Section 3.4 *Statistical Inference and Model Checking,* but it is slightly different because it is expanded to three-way log linear models.

7.2.1 Chi-Squared Goodness-of-Fit Tests

A couple statistics that we use to see if a model works well or not are the likelihood-ratio and the Pearson statistics:

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

In example 7.1.6, when we used our `proc genmod` procedure to examine the model (AM, CM), we obtained the following output:

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	**2**	**187.7543**	93.8772
Scaled Deviance	2	187.7543	93.8772
Pearson Chi-Square	**2**	**177.6144**	88.8072
Scaled Pearson X2	2	177.6144	88.8072
Log Likelihood		11916.9222	
Full Log Likelihood		-118.3989	
AIC (smaller is better)		248.7977	
AICC (smaller is better)		332.7977	
BIC (smaller is better)		249.2744	

The `Deviance` in this output is the G^2 test statistic, and the `Pearson Chi-Square` is our X^2 statistic. The degrees of freedom (`DF`) equals the number of parameters in the

saturated model minus the number of parameters in our specified model of interest. For instance, suppose the saturated model has 8 parameters, and our model has 6 parameters. Thus, the degrees of freedom for the G^2 and X^2 statistics would be 2. In order to find if this model fits well, we will use either test statistic, its corresponding degrees of freedom, and find the p-value. The p-value of these statistics is less than .001, which indicates a poor fit with the data for model (AM, CM). If we were to run the `proc genmod` procedure with the model (AC, AM, CM), we would obtain $G^2 = X^2 =$ .4 with 1 degree of freedom. This would give us a p-value of .54 which is quite large, so the homogenous model fits very well.

7.2.2 Loglinear Cell Residuals

Cell residuals are another way we can examine whether or not a model is fitting the data well. But instead of accessing the model as a whole, they show how well the model is fitting for each particular cell. The cell residuals can be calculated by taking the difference of the observed count minus the fitted count and then divided by a standard error. These residuals are approximately normal. When examining these residuals we are looking to see if their absolute value is greater than 3 for many cells or greater than 2 for few cells. If the residuals are larger than these numbers then we can conclude that the model does not fit that particular cell well. To show an example of this, we will again look back at example 7.1.6 and compare our fitted values with our observed values for model (AM, CM). Table 7.8 displays these values.

Table7.8: Standardized Residuals for Model (AM, CM)

Drug Use					
A	C	M	Observed Counts	Fitted Counts	Standardized Residuals
Yes	Yes	Yes	911	909.2	3.7
		No	538	438.8	12.8
	No	Yes	44	45.8	-3.7
		No	456	555.2	-12.8
No	Yes	Yes	3	4.8	-3.7
		No	43	142.2	-12.8
	No	Yes	2	0.2	3.7
		No	279	179.8	12.8

These residuals are all too large, which brings us to the same conclusion as the chi-squared test, indicating that there is a lack of fit.

7.2.3 Test about Conditional Associations

To test if the addition of a parameter to a model would improve it, we compare the model with the new parameter to a simpler model without the new one. This is referred as testing a conditional association in a model. This can be easily explained through an example. Let us say that we would like to test if $\lambda^{AC} = 0$. To do this we will compare the model with the new term (AC, AM, CM) to the model without it (AM, CM). Next, we will obtain a likelihood-ratio statistic by taking the simpler model's deviance minus the fuller model's deviance:

$$G^2(AM, CM) - G^2(AC, AM, CM) = 187.8 - .4 = 187.4$$

Then to find the degree of freedom for this statistic, we will take the difference between the two models' *df*. There is only a 1 parameter difference between these models, so *df*=1. This gives us a p-value less than .0001, which indicates that there is strong evidence of an

AC conditional association. Or in other words, the λ^{AC} term is significant in the model. This agrees with our past comparisons between these two models that showed model (AC, AM, CM) is better than model (AM, CM).

7.2.4 Confidence Intervals for Conditional Odds Ratios

Since the ML estimators of loglinear model parameters have approximately a normal distribution for large sample sizes, we can use this fact in generating confidence intervals to estimate the true log odds ratios. This is very similar to what we did in Section 3.4.1 of ICDA. After we obtain the estimates, we can exponentiate these values to come up with a confidence interval for the conditional odds ratios.

Let's refer again back to model (AM, CM), and find a confidence interval for the estimated log odds ratio of λ^{AM}. Our $\hat{\lambda}^{AM}$ equals 4.1251 and its SE equals .4529, thus a 95% confidence interval for the estimated log odds ratio is:

$$4.1251 \pm 1.96(.4529) = (3.24, 5.01)$$

We are able to use the 1.96 to represent the 95% because we are assuming a normal distribution. Next, we'll need to exponentiate that interval to find an interval for the conditional odds ratio:

$$(e^{3.24}, e^{5.01}) = (25.53, 149.90)$$

These large positive numbers indicate a large positive association between users of alcohol and marijuana, regardless of if they have tried cigarettes or not.

7.2.5 Loglinear Models for Higher Dimensions

So far, Section 7.2 has dealt with data for three-way tables. All of these new ideas and properties can be seen in multiway tables.

To explain this we can look at a four-way table and explain its models. A four-way table has a model of mutual independence that is $\log(\mu_{ijkl}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W$, where each variable is independent of the next. There is also a homogeneous model that looks like $\log(\mu_{ijkl}) = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_l^W + \lambda_{ij}^{XY} + \lambda_{ik}^{XZ} + + \lambda_{il}^{XW} + \lambda_{jk}^{YZ} + \lambda_{jl}^{YW} + \lambda_{kl}^{ZW}$. This is denoted by (XY, XZ, XW, YZ, YW, ZW), and allows for there to be an association between each pair of variables at any levels of the other two variable. There can also be models where one of the pairs does not have an association, (XZ, XW, YZ, YW, ZW). This model shows that X and Y are conditionally independent at each level of the other two variables.

The next two models have three-factor terms. A three factor term can be interpreted very similarly to a two factor term. An interpretation of an XYZ term is that there is an association between any pair of X, Y, and Z varying at the third variable, at every fixed level of the W variable. A model that could contain a three-factor term may look like (XYZ, XYW, XZW, YZW). The last model would be the saturated model which would be denoted (XYZW).

7.2.6 Example: Automobile Accidents and Seat Belts

Data was collected in Maine for 68,694 passengers who were in an accident. Their gender (G), location of accident (L), seat-belt use (S), and injury (I) were recorded and can be seen in the Table 7.9 below.

Table 7.9: Injury (I) by Gender (G), Location (L), and Seat Belt Use (S), with Fit of Models (GI, GL, GS, IL, IS, LS) and (GLS, GI, IL, IS)

			Injury		(GI, GL, GS, IL, IS, LS)		(GLS, GI, IL, IS)		
Gender	Location	Seat Belt	No	Yes	No	Yes	No	Yes	Sample Proportion Yes
Female	Urban	No	7,287.00	996.00	7,166.40	993.00	7,273.20	1,009.80	0.12
		Yes	11,587.00	759.00	11,748.30	721.30	11,632.60	713.40	0.06
	Rural	No	3,246.00	973.00	3,353.80	988.80	3,254.70	964.30	0.23
		Yes	6,134.00	757.00	5,985.50	781.90	6,093.50	797.50	0.11
Male	Urban	No	10,381.00	812.00	10,471.50	845.10	10,358.90	834.10	0.07
		Yes	10,969.00	380.00	10,837.80	387.60	10,959.20	389.80	0.03
	Rural	No	6,123.00	1,084.00	6,045.30	1,038.10	6,150.20	1,056.80	0.15
		Yes	6,693.00	513.00	6,811.40	518.20	6,697.60	508.40	0.07

To examine this data, we will use SAS to find the G^2 values for a several models.

```
data CarCrash;
input Gender $ Location $ SeatBelt $ Injury $ count @@;
cards;
f u n n 7287  f u n y 996
f u y n 11587 f u y y 759
f r n n 3246  f r n y 973
f r y n 6134  f r y y 757
m u n n 10381 m u n y 812
m u y n 10969 m u y y 380
m r n n 6123  m r n y 1084
m r y n 6693  m r y y 513
;

/* this will generate results for our (G,L,S,I) model */
proc genmod data=CarCrash order=data;
class Gender Location SeatBelt Injury;
model count = Gender Location SeatBelt Injury / dist=poi link=log;
run;

/* this will generate results for our (GI,GL,GS,IL,IS,LS) model */
proc genmod data=CarCrash order=data;
class Gender Location SeatBelt Injury;
model count = Gender Location SeatBelt Injury
              Gender*Injury Gender*Location
              Gender*SeatBelt Injury*Location
              Injury*SeatBelt Location*SeatBelt / dist=poi link=log;
run;


/* this will generate results for our (GIL,GIS,GLS,ILS) model */
proc genmod data=CarCrash order=data;
class Gender Location SeatBelt Injury;
model count = Gender Location SeatBelt Injury Gender*Injury
```

```
              Injury*Location Injury*SeatBelt Gender*Location*SeatBelt
              Gender*Location*Injury Gender*SeatBelt*Injury
              Injury*Location*SeatBelt/ dist=poi link=log;
run;


/* this will generate results for our (GLS,GI,IL,IS) model */
proc genmod data=CarCrash order=data;
class Gender Location SeatBelt Injury;
model count = Gender Location SeatBelt Injury Gender*Injury
              Injury*Location Injury*SeatBelt Gender*Location*SeatBelt
/ dist=poi link=log;
run;
```

Table 7.10: Goodness-of fit Test for Loglinear Models

Model	G^2	*df*	P-value
(G, I, L, S)	2792.8	11	<0.0001
(GI,GL, GS, IL, IS, LS)	23.4	5	<0.001
(GIL, GIS, GLS, ILS)	1.3	1	0.25
(GLS, GI, IL, IS)	7.5	4	0.11

To find out which model works best, we will examine the simpler models and then work our way up until we find a model that fits well. In Table 7.10, we can see the G^2 statistics for many models. The simplest model is the model of mutual independence (G, I, L, S). This model has an extremely high G^2 statistic, which corresponds with its very low p-value indicating a lack of fit. The next model in complexity is the homogeneous model (GI, GL, GS, IL, IS, LS). This model too appears that it lacks fit with its low p-value, so again we move up in complexity to the next model (GIL, GIS, GLS, ILS), which has a $G^2 = 1.3$ with a p-value of 0.25. This model fits very well with the data, but it is somewhat hard to interpret, so we would like to see if we can find a model that is in between this complex model and the homogeneous model. We will examine this problem in Section 7.2.7.

Just like with the three-way table problem in Sections 7.1.5 and 7.1.6, we can compute the estimated log odds ratios and then exponentiate those estimations to get the estimated conditional odds ratios. Table 7.11 contains these estimates.

Table 7.11: Estimated Conditional Odds Ratios for Two Loglinear Models

	Loglinear Model	
Odds Ratio	(GI,GL, GS, IL, IS, LS)	(GLS, GI, IL, IS)
GI	0.58	0.58
IL	2.13	2.13
IS	0.44	0.44
GL (S = no)	1.23	1.33
GL (S = yes)	1.23	1.17
GS (L = urban)	0.63	0.66
GS (L = rural)	0.63	0.58
LS (G = female)	1.09	1.17
LS (G = male)	1.09	1.03

In Table 7.11, the reason why numbers start to duplicate at the bottom for the model (GI, GL, GS, IL, IS, LS) is because this model is the homogeneous model. This means that each pair of variables has an identical association at every level of the other variables, so we would expect there to be no difference between the odds ratio for GL (S = no) and GL (S = yes). Because the model (GLS, GI, IL, IS) has this three-factor term, there is a difference between GL (S = no) and GL (S = yes) and the other two factor terms.

As in one of the previous sections, we are able to construct confidence intervals for these odds ratios, but first we need to start by computing an interval for the log odds ratios. For example take the estimated log odds ratio for GI in the model (GI, GL, GS, IL,

IS, LS) and use its corresponding SE to create the interval for the log odds ratio. We can look at our SAS output for these values.

Analysis Of Maximum Likelihood Parameter Estimates

Parameter			DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square
Gender*Injury	f	n	1	**-0.5405**	**0.0272**	-0.5939	-0.4872	394.36

A 95% confidence interval for the true log odds ratio for GI is -0.5405± 1.96(0.0272) = (-0.594, -0.487). Then we exponentiate this interval to get (.552, .614), which is a 95% confidence interval for the true odds ratio. The odds of injury for passengers who were male are a little more than half the odds for passengers who were female, for each location-seat belt combination.

7.2.7 Three-Factor Interaction

It is often very difficult to interpret a three-factor term. In Table 7.11 of our previous example, we can see all of the estimated odds ratios for two models. For model (GLS, GI, IL, IS), the table is straight foward until we reach values for GL, where there are two values. One is for the GL odds ratio when a person was wearing a seat belt and one is for when a person was not wearing a seat belt. This is because of the three-factor term GLS which makes these odds ratios for GL differ for each level of S. Since 'I' is not in this three factor term, all the two-factor terms with 'I' have odds ratios that are equal at each pair of levels of the other two variables. This is why the first few lines appear as they have in the past.

In order to find the odds ratios for the pairs that have two of the three variables that are in the three-factor term, we need to calculate fitted odds ratios between two variables at each level of the third. For calculating the GS odds ratio for an urban location we will need four fitted values either from injury-yes or from injury-no that have an urban location. From here we find the estimated odds ratio

$$\frac{7273.2 * 10959.2}{11632.6 * 10358.9} = .66$$

This is means that the estimated odds that males used seatbelts in an urban location are only .66 times the estimated odds for females in an urban location.

7.2.8 Large Samples and Statistical Versus Practical Significance

Whether the sample size is small or large, sample sizes can cause problems for selecting the best model. If the sample size is too small, models that are simpler may seem like they are significant when they actually wouldn't be if the sample size was a bit larger. Also if sample sizes are too large, models that are more complex may seem to better fit the data when in actuality they do a similar job as a more simplistic model. For example, in Table 7.10 it appears that model (GLS, GI, IL, IS) does a better job than model (GI, GL, GS, IL, IS, LS) because it has a smaller G^2 statistic and a larger p-value. But looking at Table 7.11, the models appear to produce relatively the same odds ratios. Thus, it may be better to go with the simpler model because it is easier to interpret and produce almost the same results as the more complex model. This, again, is because there was such a large sample size in this problem, so differences appeared to be significant by

the goodness-of-fit test. There is another way to assess goodness of fit for a model of large sample sizes through the *dissimilarity index*.

The *dissimilarity index,* denoted by D, is a measure that does not depend on sample size.

$$D = \sum \frac{|n_i - \hat{\mu}_i|}{2n} = \sum \frac{|p_i - \hat{\pi}_i|}{2}$$

where *D* is between 0 and 1. A proof of this can be seen below.

$$We\ want\ to\ prove\ \sum \frac{|p_i - \hat{\pi}_i|}{2} \leq 1$$

$$or\ \sum_{i=1}^{k} |p_i - \hat{\pi}_i| \leq 2. Suppose\ the\ set\ P\ implies\ p_i \geq \hat{\pi}_i$$

$$and\ the\ set\ N\ implies\ p_i \leq \hat{\pi}_i$$

$$Thus, \sum_{i=1}^{k} |p_i - \hat{\pi}_i| = \sum_{P} (p_i - \hat{\pi}_i) + \sum_{N} (\hat{\pi}_i - p_i)$$

$$= \sum_{P} (p_i) - \sum_{P} (\hat{\pi}_i) + \sum_{N} (\hat{\pi}_i) - \sum_{N} (p_i)$$

$$= \sum_{P} (p_i) - \sum_{N} (p_i) + \sum_{N} (\hat{\pi}_i) - \sum_{P} (\hat{\pi}_i)$$

$$= \underbrace{(p_1 + p_2 + \cdots + p_j - p_{j+1} - p_{j+2} - \ \ldots - p_k)}_{\substack{\leq 1 \text{ because} \\ \sum_{n=1}^{k} p_n = 1}} + \underbrace{(\hat{\pi}_1 + \hat{\pi}_2 + \cdots + \hat{\pi}_j - \hat{\pi}_{j+1} - \hat{\pi}_{j+2} - \ \ldots - \hat{\pi}_k)}_{\substack{\leq 1 \text{ because} \\ \sum_{n=1}^{k} \pi_n = 1}}$$

$$\leq 2 \ \blacksquare$$

Smaller values of *D* mean that the model fits better. A small *D* show that the model is doing a pretty good job representing the data even though the model might not appear to fit the best.

Now that we have defined *D*, we are now able to apply it to the models in Table 7.9. The model (GI, GL, GS, IL, IS, LS) has a $D = 0.008$ and model (GLS, GI, IL, IS) has $D=0.003$. Both of these small values for *D* indicate that the models are both doing a good job fitting the data. If we were to choose between the two models, I would choose model (GI, GL, GS, IL, IS, LS) because it is easier to interpret. Even though this model had a high G^2 value, this dissimilarity index shows that it still has a pretty good fit.

7.3 The Loglinear-Logistic Connection

In the following sections, we learn when it is appropriate to use a loglinear model versus a logistic model. In particular, loglinear regression models tell us about the associations between categorical responses and logistic regression models tell us how a categorical response relies on a group of explanatory variable.

7.3.1 Using Logistic Models to Interpret Loglinear Models

Like in Section 4.3.3 AIDCA, we can create a logistic from a loglinear model if one of the variables is binary. Our loglinear model will be:

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

Now, let *Y* be our binary response variable, thus making our explanatory variables X and Z. Next, let X be at level *i* and Z be at level *k*.

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

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

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

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

$$= \quad \alpha \quad + \quad \beta_i^X \quad + \quad \beta_k^Z$$

$$logit[P(Y=1)] = \alpha + \beta_i^X + \beta_k^Z$$

By the end of these steps, we have converted what appeared to be the difference of two loglinear models into one logistic model. Thus, we have established a link between the loglinear model and the logistic model.

7.3.2 Example: Auto Accident Data Revisited

In Section 7.2.6, we tested several models and found out that the loglinear model (GLS, GI, IL, IS) fits the data well by inspecting G^2 statistics. Now let's turn this loglinear model into a logistic model by having 'I' be our binary response variable.

$$logit[P(I=1)] = \log\left[\frac{P(I=1)}{1-P(I=1)}\right] = \log\left[\frac{P(I=1|G=g, L=l, S=s)}{P(I=2|G=g, L=l, S=s)}\right]$$

$$= \log\left(\frac{\mu_{g1ls}}{\mu_{g2ls}}\right) = \log(\mu_{g1ls}) - \log(\mu_{g2ls})$$

$$= \left(\lambda + \lambda_g^G + \lambda_1^I + \lambda_l^L + \lambda_s^S + \lambda_{g1}^{GI} + \lambda_{gl}^{GL} + \lambda_{gs}^{GS} + \lambda_{1l}^{IL} + \lambda_{1s}^{IS} + \lambda_{ls}^{LS} + \lambda_{gls}^{GLS}\right) - \ (\lambda + \lambda_g^G$$

$$+ \lambda_2^I + \lambda_l^L + \lambda_s^S + \lambda_{g2}^{GI} + \lambda_{gl}^{GL} + \lambda_{gs}^{GS} + \lambda_{2l}^{IL} + \lambda_{2s}^{IS} + \lambda_{ls}^{LS} + \lambda_{gls}^{GLS}$$

$$= (\lambda_1^I - \lambda_2^I) + \left(\lambda_{g1}^{GI} - \lambda_{g2}^{GI}\right) + (\lambda_{1l}^{IL} - \lambda_{2l}^{IL}) + (\lambda_{1s}^{IS} - \lambda_{2s}^{IS})$$

$$logit[P(I=1)] = \alpha + \beta_g^G + \beta_l^L + \beta_s^S$$

7.3.3 Correspondence between Loglinear and Logistic Models

One of the down sides to the logistic model is that it does not describe the relationship between two of the explanatory variables. To better show this return to Section 7.3.1 where we transformed a loglinear model (XY, XZ, YZ) into a logistic model. In the final step of this transformation the λ_{ik}^{XZ} term cancels out. Now, let's suppose we want to transform the loglinear model (XY, YZ) into a logistic model.

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

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

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

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

It turns out that we have come up with the exact same logistic equation as we did in Section 7.3.1. This occurred because logistic regression neglects this explanatory association. Table 7.12 shows all the different types of logistic equations for a three-way model with Y as a binary response variable and X and Z as the explanatory variables.

Table 7.12: Equivalent Loglinear and Logistic Models for a Three-Way Table With Binary Response Variable Y

Loglinear Symbol	Logistic Model	Logistic Symbol
(Y,XZ)	α	(-)
(XY,XZ)	$\alpha+\beta_i^X$	(X)
(YZ, XZ)	$\alpha+\beta_k^Z$	(Z)
(XY,XZ, YZ)	$\alpha+\beta_i^X + \beta_k^Z$	(X+Z)
(XYZ)	$\alpha+\beta_i^X + \beta_k^Z + \beta_{ik}^{XZ}$	(X*Z)

7.3.4 Strategies in Model Selection

Generally, when there is one binary response variable we should usually use the logistic model because it is simpler than the loglinear model. To explain that it is simpler look at the following two equations,

$$logit[P(I=1)] = \alpha + \beta_g^G + \beta_l^L + \beta_s^S$$

$$\log(\mu_{gils}) = \lambda + \lambda_g^G + \lambda_i^I + \lambda_l^L + \lambda_s^S + \lambda_{gi}^{GI} + \lambda_{gl}^{GL} + \lambda_{gs}^{GS} + \lambda_{il}^{IL} + \lambda_{is}^{IS} + \lambda_{ls}^{LS} + \lambda_{gls}^{GLS}$$

The logistic equation has fewer parameters and is very simple. The problem with loglinear regression is that the equations get complicated with the addition of more variables. We should use loglinear regression if there is more than one response variable or if we want to find out about associations between all the variables. For everything else, we should use logistic regression.

Chapter 8 Models for Matched Pairs

In this chapter, we are going to be dealing with data that has categorical responses for two samples in which there is an obvious pairing between the two samples. This pairing is done because there are several characteristics between particular subjects in one group that are the same in the second group. The responses for each pair of subjects are known as matched pairs. This matching causes the two samples to be statistically dependent.

A pairing between two subjects can simply happen when a group of people is asked two questions thus having a single person's responses be a matched pair. Another common scenario would be if we wanted to test a drug to see if it works. What we might do is have two groups where a subject in one group has an almost identical subject in the second group. Then we would assign one group to be the control group and the other group to be the treatment group. This would mean that the only difference between the two groups would be the drug effect. Both these kind of models are best summarized in square contingence tables.

Section 8.1 explains how to compare dependent proportions. Section 8.2 applies logistic regression to matched pairs. Section 8.3 examines the margins of square contingency tables. We will then briefly introduce symmetry and quasi-symmetry models for square tables in Section 8.4, and lastly Section 8.6 discusses the Bradley-Terry model for paired preferences.

8.1 Comparing Dependent Proportions

Our first example of matched pairs comes from the 2000 General Social survey, in which people were asked two questions: "Would you be willing to pay higher taxes to

help improve the environment?" and "Would you be willing to cut your living standards to help improve the environment?" The results of the responses are summarized in Table 8.1.

Table 8.1 Opinions Relating to Environment

	Cut Living Standards		
Pay Higher Taxes	Yes	No	Total
Yes	227	132	359
No	107	678	785
Total	334	810	1144

To compare the probabilities of a "yes" outcome for each of these questions, we will first look at the marginal proportions. Before we do this, let's go over some notation to be clear about which cells we are talking about. Cell n_{ij}, refers to the number of subjects that answered i to the first question and j to the second question. So in our example, the subjects who answered "yes" to increasing taxes are $n_{11} + n_{12} = n_{1+} = 359$, and the subjects who answered "yes" to cutting standards are $n_{11} + n_{21} = n_{+1} = 334$. These have sample proportions of 359/1144 = 0.31 and 334/1144 = 0.29 respectively. These proportions are known as marginal proportions.

Because most of the subjects in this survey answered either "yes" to both questions or "no" to both questions, we can say that the marginal proportions are correlated. Another way to see that these proportions are strongly correlated is to examine the sample odds ratio:

$$\frac{227 * 678}{132 * 107} = 10.9$$

This means that if someone answered "yes" to one of the questions they are almost 10 times more likely to answer "yes" to the second question.

To help us understand Table 8.1 further, we are going to say that the probability of outcome *i* for the first question and outcome *j* for the second question is π_{ij}. This notation can also be used for the marginal proportions, such as the probability of answering "yes" to the first question is π_{1+} and answering "yes" to the second question is π_{+1}. If these two marginal proportions equal each other then π_{2+} has to equal π_{+2}, since $\pi_{1+} + \pi_{2+} = 1$ has to be true and $\pi_{+1} + \pi_{+2} = 1$ has to be true. When this happens it is said to have *marginal homogeneity*, which simply means that the margins are equal. Marginal homogeneity in a two-way table also implies that $\pi_{12} = \pi_{21}$ because

$$\pi_{1+} = \pi_{+1} \Rightarrow 0 = \pi_{1+} - \pi_{+1} = (\pi_{11} + \pi_{12}) - (\pi_{11} + \pi_{21}) = \pi_{12} - \pi_{21}$$

$$\Rightarrow \pi_{12} = \pi_{21}$$

8.1.1 McNemar Test Comparing Marginal Proportions

When matched pair data can be arranged in a two-way table, as in Table 8.1, we refer to this type of response as a binary response. With data like this we can run a test of marginal homogeneity with a null hypothesis:

$$H_o: \pi_{1+} = \pi_{+1} \; or \; H_o: \pi_{12} = \pi_{21}$$

To come up with a test statistic for this hypothesis, let's first denote $n_{12} + n_{21}$ as n^*. Assuming that the null hypothesis is true, we would then expect that $n_{12} = n_{21}$, which also would equal $\frac{1}{2}n^*$. This means that the probability of one of the subjects contributing to n^* from n_{12} is .5, which is the same probability as being from n_{21}. This allows us to say n_{21} and n_{21} are the number of 'successes' and 'failures' for a binomial distribution having n^* trials and success probability $\frac{1}{2}$.

This binomial distribution has an approximately normal distribution with a mean of $.5n^*$and a standard deviation$\sqrt{n^* * .5 * .5}$, when $n^*$is large enough (>10). This yields a standard normal test statistic,

$$z = \frac{n_{12} - (.5)n^*}{\sqrt{n^* * .5 * .5}} = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} \qquad (8.1)$$

When we square this, we obtain a chi-squared statistic with degrees of freedom equal to 1. This is called the McNemar test, which tests for a comparison between two dependent proportions.

Referring back to Table 8.1, we can generate a McNemar test statistic. First, we will find the standard normal test statistic using Equation 8.1,

$$z = \frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}} = \frac{132 - 107}{\sqrt{132 - 107}} = 1.62$$

Squaring, we get a chi squared statistic of 2.62 with 1 degree of freedom, which gives us a p-value of 0.106. This tells us that there is weak evidence that the probability of a person answering 'yes' was greater for higher taxes than for cutting the standard of living.

If these two samples were actually independent, rather than matched pairs we would set up our two-way table differently. We would have had the two questions be our explanatory row variables and have the binary response be our column variables. Then we would run a test of independence as we did in Section 2.4. Then we would see if the probability of 'yes' was the same for each question. We do not want to run this analysis on our matched pair data because we assume that there is an association between the row and column classifications since the two samples are dependent.

8.1.2 Estimating Differences of Proportions

Another way of testing whether the marginal distributions differ can be done by creating a confidence interval for the true difference of proportions. Thus, we need to find some statistics to reflect the parameter $\pi_{1+} - \pi_{+1}$. Let $p_{ij} = \frac{n_{ij}}{n}$ denote the sample cell proportion. Our statistic of the difference of proportions is $p_{1+} - p_{+1}$. The difference has an estimated variance of

$$\frac{[p_{1+}(1-p_{1+}) + p_{+1}(1-p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21})]}{n} \qquad (8.2\ )$$

To obtain the SE, just take the square root of this, and after some algebra, we can simplify this equation to

$$SE = \sqrt{(n_{12}+n_{21}) - [(n_{12}+n_{21})^2/n]}/n$$

How the SE is derived from Equation 8.2 is shown below.

$$\frac{[p_{1+}(1-p_{1+}) + p_{+1}(1-p_{+1}) - 2(p_{11}p_{22} - p_{12}p_{21})]}{n}$$

$$\Rightarrow \frac{\left[(p_{11}+p_{12})\big(1-(p_{11}+p_{12})\big) + (p_{11}+p_{21})\big(1-(p_{11}+p_{21})\big) - 2(p_{11}p_{22} - p_{12}p_{21})\right]}{n}$$

$$\Rightarrow \frac{[(p_{11}+p_{12}) - (p_{11}+p_{12})^2 + (p_{11}+p_{21}) - (p_{11}+p_{21})^2 - 2p_{11}p_{22} + 2p_{12}p_{21}]}{n}$$

$$\Rightarrow \frac{[(p_{12}+p_{21}) + 2p_{11} + 2p_{12}p_{21} - ({p_{11}}^2 + 2p_{11}p_{12} + {p_{12}}^2) - ({p_{11}}^2 + 2p_{11}p_{21} + {p_{21}}^2) - 2p_{11}p_{22}]}{n}$$

$$\Rightarrow \frac{[(p_{12}+p_{21}) + 2p_{11} + 2p_{12}p_{21} - {p_{11}}^2 - {p_{12}}^2 - {p_{11}}^2 - {p_{21}}^2 - 2p_{11}p_{12} - 2p_{11}p_{21} - 2p_{11}p_{22}]}{n}$$

$$\Rightarrow \frac{[(p_{12}+p_{21}) - ({p_{12}}^2 - 2p_{12}p_{21} + {p_{21}}^2) + 2p_{11} - 2{p_{11}}^2 - 2p_{11}p_{12} - 2p_{11}p_{21} - 2p_{11}p_{22}]}{n}$$

$$\Rightarrow \frac{[(p_{12}+p_{21}) - (p_{12}-p_{21})^2 + 2p_{11}(1-p_{11}-p_{12}-p_{21}-p_{22})]}{n}$$

$$\Rightarrow \frac{[(p_{12}+p_{21})-(p_{12}-p_{21})^2]}{n}$$

$$\Rightarrow \frac{\left[\frac{1}{n}(n_{12}+n_{21})-\frac{1}{n^2}(n_{12}-n_{21})^2\right]}{n}$$

$$\Rightarrow \frac{\left[(n_{12}+n_{21})-\frac{1}{n}(n_{12}-n_{21})^2\right]}{n^2} \; Taking\ the\ square\ root\ we\ have$$

$$\sqrt{\frac{\left[(n_{12}+n_{21})-\frac{1}{n}(n_{12}-n_{21})^2\right]}{n^2}}$$

$$\Rightarrow \sqrt{(n_{12}+n_{21})-[(n_{12}+n_{21})^2/n]}/n$$

With this shown, now we can produce a 95% confidence interval for the difference of proportions, $(p_{1+} - p_{+1}) \pm 1.96\frac{\sqrt{(n_{12}+n_{21})-[(n_{12}+n_{21})^2/n]}}{n}$. In our previous example, a 95% confidence interval would equal,

$$(0.314 - .292) \pm 1.96\left(\frac{\sqrt{(132+107)-[(132+107)^2/1144]}}{1144}\right) = (-0.004, 0.048)$$

Our interpretation would be that we are 95% confident that the probability of a 'yes' response was between 0.004 less and 0.048 higher for paying higher taxes than for accepting a cut in living standards. Because zero is in this interval, we can conclude that there is no difference between the two probabilities.

8.2 Logistic Regression for Matched Pairs

This next section shows how we can use logistic regression to analyze matched pairs.

8.2.1 Marginal Models for Marginal Proportions

To begin applying models to matched pairs, we will start by using the example which asked the two questions about a tax increase and cutting living standards. Let (Y_1, Y_2) represent the response to the two questions, where '1' means they answered 'yes' and '0' means they answered 'no.' This tells us that the marginal probabilities can be written as $P(Y_1=1) = \pi_{1+}$ and $P(Y_2=1) = \pi_{+1}$. Their corresponding statistics are $p_{1+} = \frac{359}{1144} = 0.31$ and $p_{+1} = \frac{334}{1144} = 0.29$ respectively.

Using the identity link function, we can obtain the model

$$P(Y_1 = 1) = \alpha + \delta\,, \qquad P(Y_2 = 1) = \alpha$$

Where δ is the difference between the marginal probabilities. The ML estimate for δ is the differences between the sample marginal proportions, $p_{1+} - p_{+1} = \hat{\delta} = .31 - .29 = 0.02$. The hypothesis test for this model is similar to the hypothesis for the McNemar test but in relation to δ, H_o: $\delta = 0$.

Another model we could use would involve using the logit link,

$$logit[P(Y_1 = 1)] = \alpha + \beta\,, \qquad logit[P(Y_2 = 1)] = \alpha \qquad (8.3)$$

which equals

$$logit[P(Y_t = 1)] = \alpha + \beta x_t$$

This model uses x_t as an indicator variable that is 1 when $t = 1$ and is 0 when $t = 2$. As before e^{β} is an odds ratio, but this odds ratio is comparing the marginal distributions. We

can use the odds ratio for the sample marginal distributions to find the ML estimate. For example, Table 8.1 has $e^{\widehat{\beta}} = \left[\frac{359*810}{785*334}\right] = 1.11$. This can be interpreted as the population odds of answering 'yes' to pay higher taxes are estimated to be 11% higher than the population odds of answering 'yes' to accept cuts in living standards.

Because these models are referring to the marginal distributions, these models are the *marginal models.*

8.2.2 Subject-Specific and Population-Averaged Tables

Table 8.1 shows a two-way table that summarizes how everyone voted by tallying the totals of those who responded 'yes' to both questions, 'no' to both questions, 'yes' to the first and 'no' to the other, and 'no' to the first and 'yes' to the second. Another way to examine this would be to have a model that is particular for each subject. In Table 8.2, we can observe what one of these tables would look like for a single person, who answered 'yes' to both questions.

Table 8.2: Representation of Matched Pair Contributing to Count n_{11} in Table 8.1

	Response	
Issue	Yes	No
Pay Higher Taxes	1	0
Cut Living Standards	1	0

Table 8.2 is just one of the 1144 partial tables that make up a full three-way table that relates to Table 8.1. The entire three-way table has the form 2 x 2 x 1144, where there are 227 tables that look like the table above, 132 that have 'yes' to the first and 'no' to the second, 107 that have 'no' to the first and 'yes' to the second, and 678 that have

'no' as an answer to both. It we were to combine all of these partial tables, we would have Table 8.2.a. This actually shows all the marginal counts to Table 8.1.

Table 8.2.a: Collapsed Partial Tables

	Response	
Issue	Yes	No
Pay Higher Taxes	359	785
Cut Living Standards	334	810

The three-way table that Table 8.2 is part of is known as a *subject-specific table*. These tables are in the form of 2 x 2 x *n* where there are *n* partial tables. Models that are used to analyze these are called *conditional* models, since the effect comparing the responses is *conditional* on the subject. Table 8.1 is an example of a *population-average table,* which essentially is a 2 x 2 cross-classified table of the two responses for all the subjects. Models that explain tables like Table 8.2.a are called *marginal models* and are explained more in Chapter 9.

8.2.3 Conditional Logistic Regression for Matched-Pairs

Equation 8.3 refers to a marginal model that uses the logit link function. We similarly can extend this model to create conditional models where Y_{it} represents observation *t* for subject *i,* and $y_{it} = 1$ means a success. The conditional model would then look like this

$$logit[P(Y_{i1} = 1)] = \alpha_i + \beta, \quad logit[P(Y_{i2} = 1)] = \alpha_i \qquad (8.4)$$

Or as before

$$logit[P(Y_{it} = 1)] = \alpha_i + \beta x_{it}$$

where x_{it} is an indicator variable with $x_{i1} = 1$ and $x_{i2} = 0$.

This α_i parameter allows variability between subjects. To describe this parameter we will examine what happens for certain values of it when referring to our example that dealt with the two environmental improvement questions. Larger positive values for this parameter show that a person will answer 'yes' for both questions. Larger negative values for this term show that a person will answer 'no' for both questions. The greater the magnitude of the parameter the greater the association between observations is. Model 8.4 implies that, for each subject, the odds of answering 'yes' to the first question are e^{β} times the odds of answering 'yes' to the second question. This conditional association is a *subject-specific effect* because it explains what is happening to one particular subject. If marginal homogeneity occurs, then the β term will equal 0, which will mean that the probability of answering 'yes' to the first question is the same as answering 'yes' to the second question.

In order to make an inference about β, we compare the distribution for $t = 1$ and $t = 2$. The drawback to this is that there are as many subject parameters α_i as there are subjects, which makes it extremely hard to fit these models. To fix this we will use the *conditional maximum likelihood.* This maximizes the likelihood function and finds $\hat{\beta}$ for a conditional distribution, which eliminates the subject specific parameters. Table 8.1 has a conditional ML estimate of the odds ratio e^{β} for model (8.4), which equals $\frac{n_{12}}{n_{21}} = \frac{132}{107} = 1.23$. This means that a subject's estimated odds of answering 'yes' are 23% higher for increasing taxes than for cutting their living standards.

This odds ratio is different than the one we found in Section 8.2.1, which was produced using the marginal model. This difference is just another example of how conditional odds ratios may differ from marginal odds ratios.

8.2.4 Logistic Regression for Matched Case-Control Studies

Case-control studies require matched pair analysis. As mentioned earlier in this chapter, a case-control study will have two groups: one that is undergoing a treatment (case) and one that is not (control). Because it is impossible to have a single person be in each group, for the binary response Y, the person conducting the study will match a person from the case group (Y=1) with a person to be in the control group (Y =0) based on many demographics, which will try to mimic that person being in both groups. The study will then observe both groups on the predictor variable *X* and analyzes the *XY* association.

A study of acute myocardial infarction (MI) among Navajo Indians match 144 victims of MI according to age and gender with 144 subjects who did not have this disease. All the individuals were then asked if they had diabetes or not (x=1 if yes or x=0 if no). The data for this study can be found in Table 8.3. This table is the population-average table for this study.

Table 8.3: Previous Diagnoses of Diabetes for Myocardial Infarction Case-Control Pairs

	MI Cases		
MI Controls	Diabetes	No Diabetes	Total
Diabetes	9	16	25
No Diabetes	37	82	119
Total	46	98	144

Table 8.3 can be further separated into 144 partial tables to create a 2 x 2 x 144 subject-specific table. There will be four unique partial tables, which can be seen in Table 8.4. In Table 8.4, the numbers in parenthesis explain how many partial tables there are

that look like that type. For example, there are 9 partial tables that look like the partial table A.

Table 8.4: Possible Case-Control Pairs for Table 8.3

	A(9)		B(16)		C(37)		D(82)	
Diabetes	Case	Control	Case	Control	Case	Control	Case	Control
Yes	1	1	0	1	1	0	0	0
No	0	0	1	0	0	1	1	1

Let the model,

$$logit[P(Y_i = 1)] = \alpha_i + \beta_x$$

represent subject *i.* If we wanted to estimate the odds ratio for XY, we could just use the conditional ML estimate of the odds ratio e^{β} for Table 8.3, which is $\frac{n_{21}}{n_{12}} = \frac{37}{16} = 2.3$. This means that the odds of having diabetes if you have MI are 130% higher than those who do not have MI.

8.2.5 Connection between McNemar and Cochran-Mantel-Haenszel Test

In Section 4.9 we were introduced to the Cochran-Mantel-Haenszel (CMH) chi-squared statistic:

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

If this statistic were to be applied to a 2 x 2 x *n* subject-specific table that relates the response to the observation, we would obtain a statistic that is equivalent to the McNemar statistic.

$$\frac{n_{12} - n_{21}}{\sqrt{n_{12} + n_{21}}}$$

Thus, we can see that the McNemar test is just a special case of the CMH, in which there are *n* binary responses which can be represented in *n* partial tables.

8.3 Comparing Margins of Square Contingency Tables

So far in Chapter 8 we have been dealing with matched paired data that can be displayed in a 2 x 2 table. Now we are going to extend our new methods to any squared table. Let (Y_1, Y_2) represent the observations for a randomly selected subject. We can then create a I x I table that has cell counts in the form of $n_{ij}$for (Y_1, Y_2). These (Y_1, Y_2) will commonly represent our two matched samples and n_{ij} will be their responses.

Let $\pi_{ij} = P(Y_1 = i, Y_2 = j)$. Marginal homogeneity is

$$P(Y_1 = i) = P(Y_2 = j) \quad for\ i = 1, \dots, I$$

This shows that the marginal probability for each row equals the marginal probability for its corresponding column.

8.3.1 Marginal Homogeneity and Nominal Classifications

The way that we will conduct test of marginal homogeneity is by comparing the ML fitted values $\{\hat{\mu}_{ij}\}$ that satisfy marginal homogeneity to the observed counts $\{n_{ij}\}$ using the G^2 or X^2statistics with d.f. = I-1.

8.3.2 Example: Coffee Brand Market Share

A survey was done in which a sample of buyers of instant decaffeinated coffee were asked, "what was the brand of coffee that they bought?" At a later coffee purchase by the same people, the brand of their choice was again recorded. Table 8.5 shows the results of this observational study. From this table, we are able to tell that most buyers did not change their brand preference. We can tell this because if you look along the main diagonal (from $n_{11}$to n_{55}) we see that these are the larger numbers of the table.

Table 8.5: Choice of Decaffeinated Coffee at Two Purchase Dates, with ML Fit Satisfying Marginal Homogeneity in Parentheses

	Second Purchase				
First Purchase	High Point	Taster's Choice	Sanka	Nescafe	Brim
High Point	93(93)	17(13.2)	44(32.5)	7(6.1)	10(7.8)
Taster's Choice	9(12.7)	46(46)	11(10.5)	0(0.0)	9(9.1)
Sanka	17(26.0)	11(11.6)	155(155)	9(11.3)	12(12.8)
Nescafe	6(7.0)	4(3.5)	9(7.5)	15(15)	2(1.8)
Brim	10(14.0)	4(4.0)	12(11.3)	2(2.3)	27(27)

The numbers in the parentheses of Table 8.5 are the ML fitted values that satisfy marginal homogeneity. To obtain these numbers we can use SAS.

```
Data coffee;
input first $ second $ count m11 m12 m13 m14 m21 m22 m23 m24
   m31 m32 m33 m34 m41 m42 m43 m44 m55 m1 m2 m3 m4;
cards;
/* The 1st variable represents the first purchase*/
/* The 2nd variable represents the second purchase*/
/* The 3rd variable represents the counts*/
/* from here was want to place a 1 next to the value that correlates to
   the suffix of the variable. Example m11 has a suffix of 11. So we
   mark a 1 next to that first and second purchases that are in the
   first row and first column. Then since this is a squared table with
   (I-1)2 parameters we need to mark a -1 for any purchase that has brim
   In it unless it is with brim itself. These values will be implied.
   */
high high  93  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
high task  17  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
high sank  44  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
high nesc   7  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
high brim  10 -1 -1 -1 -1  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0
task high   9  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
task task  46  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
task sank  11  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0  0
task nesc   0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0  0
task brim   9  0  0  0  0 -1 -1 -1 -1  0  0  0  0  0  0  0  0  0  0  1  0  0
sank high  17  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0  0
sank task  11  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0  0
sank sank 155  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0  0
sank nesc   9  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0  0
sank brim  12  0  0  0  0  0  0  0  0 -1 -1 -1 -1  0  0  0  0  0  0  0  1  0
nesc high   6  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0  0
nesc task   4  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0  0
nesc sank   9  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0  0
nesc nesc  15  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0  0
nesc brim   2  0  0  0  0  0  0  0  0  0  0  0  0 -1 -1 -1 -1  0  0  0  0  1
brim high  10 -1  0  0  0 -1  0  0  0 -1  0  0  0 -1  0  0  0  0  1  0  0  0
brim task   4  0 -1  0  0  0 -1  0  0  0 -1  0  0  0 -1  0  0  0  0  1  0  0
brim sank  12  0  0 -1  0  0  0 -1  0  0  0 -1  0  0  0 -1  0  0  0  0  1  0
brim nesc   2  0  0  0 -1  0  0  0 -1  0  0  0 -1  0  0  0 -1  0  0  0  0  1
brim brim  27  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  1  0  0  0  0
;
```

```
run;

/* In this model m11 is the expected frequency μ11, m1 denotes μ1+= μ+1.
This notation uses the formulas like μ15= μ1+-μ11-μ12-μ13-μ14 for the terms
in the last column or the last row.
proc genmod;
model count = m11 m12 m13 m14 m21 m22 m23 m24
   m31 m32 m33 m34 m41 m42 m43 m44 m55 m1 m2 m3 m4
   / dist = poi link= identity obstats residuals;
run;
```

When we compare our fitted values to the observed cell counts we get a $G^2 = 12.6$ and a $X^2 = 12.4$. These with 4 d.f. yield a p-value of 0.015, which rejects our null hypothesis of marginal homogeneity.

The sample marginal proportions can easily be obtained and are listed in Table 8.5.a.

Table 8.5.a: Sample Marginal Proportions for Coffee Brands

	Purchase	
	First	Second
High Point	0.32	0.25
Taster's Choice	0.14	0.15
Sanka	0.38	0.43
Nescafe	0.07	0.06
Brim	0.1	0.11

To estimate the change for a given brand, we can compare one brand versus all the other brands to create a two-way table and then analyze it as we have done previously in this chapter. For example, we will compare High Point to all other brands. Table 8.5.b displays this data.

Table 8.5.b: High Point vs Others

	Second Purchase	
First Purchase	High Point	Other
High Point	93	78
Other	42	328

When we run McNemar's test, we obtain a statistic of $\frac{78-12}{\sqrt{78-42}} = 3.3$, which has a p-value equal to 0.001. This tells us that we have strong evidence of a change in population proportions. Another way to examine this would be to take the difference of proportion and create a confidence interval for it to see if the difference is significant. The estimated difference is $0.32 - 0.25 = 0.07$. Then we can calculate the standard error to create a confidence interval:

$$\frac{\sqrt{(78+42) - \frac{(78-42)^2}{541}}}{541} = .02$$

The 95% confidence interval is 0.07±1.96(.02) = (0.03,0.11), thus there is a difference between the first and the second purchase for the brand High Point. The reason why the test for marginal homogeneity was rejected was mainly due to the decrease in the proportion choosing High Point.

8.3.3 Marginal Homogeneity and Order Categories

The test of marginal homogeneity, having d.f.= *I*-1, that we have looked at thus far is meant to find any differences between the marginal proportions. It assumes that there is no order among the different categories. But a more powerful test could be run if there was such an order. For example, if our categories were on a scale from 1 to 10, where 10 denoted the best, and we found out that our data does not have marginal

homogeneity, we would like to know if the data favors the high end of the scale or the low end of the scale. This kind of variable is called ordinal. We can run an ordinal test, which has a d.f. = 1, and it is usually a more powerful than just a test for marginal homogeneity. Also, since the degree of freedom is 1, when *I* is large the association between classifications is strong.

We can create an ordinal logistic model for comparison of the margins:

$$logit[P(Y_{i1} \leq j)] = \alpha_{ij} + \beta, \qquad logit[P(Y_{i2} \leq j)] = \alpha_{ij}$$

This model is a generalization of our binary model (8.4 ICDA) that represents each cumulative logit in terms of subject effects and a margin effect. In these models the β term is assumed to be held constant for each cumulative probability. The interpretation of this model is that, for each pair, the odds that observation 1 falls in category *j* or below are e^{β} times the odds for observation 2.

Here is how we can estimate this β term in our model,

$$\hat{\beta} = \log\left(\frac{\sum\sum_{i<j}(j-i)n_{ij}}{\sum\sum_{i>j}(i-j)n_{ij}}\right) \qquad (8.7)$$

The numerator refers to all the cell counts above the main diagonal, and the denominator refers to all the cell counts below the main diagonal. The ordinal test of marginal homogeneity is testing if β = 0. $\hat{\beta}$ has a SE that equals:

$$SE = \sqrt{\frac{\sum\sum_{i<j}(j-i)^2 n_{ij}}{[\sum\sum_{i<j}(j-i)n_{ij}]^2} + \frac{\sum\sum_{i>j}(i-j)^2 n_{ij}}{[\sum\sum_{i>j}(i-j)n_{ij}]^2}}$$

Our test statistic will be $\hat{\beta}/_{SE}$, which is approximately standard normal.

This method can easily become tedious to do by hand, and there is a simple alternative method that compares the sample means for the two margins, for ordered

category $\{\mu_i\}$. Let $\bar{x} = \sum_i \mu_i p_{i+}$ be the sample mean for the rows, and let $\bar{y} = \sum_i \mu_i p_{+i}$ be the sample mean for the columns. Then our estimated difference of means would be $(\bar{x} - \bar{y})$, with a standard error of $\sqrt{\frac{1}{n}[\sum_i \sum_j (\mu_i - \mu_j)^2 p_{ij}]}$. The ratio of the difference in means and the standard error has an approximate standard normal distribution. This test is used to find the true difference between true marginal means.

8.3.4 Example: Recycle or Drive Less to Help Environment

A General Social Survey asked people "How often do you cut back on driving a car for environmental reasons? and "How often do you make a special effort to sort glass or cans or plastic or paper and so on for recycling?" The results from this survey are in Table 8.6.

Table 8.6: Behaviors on Recycling and Driving Less to Help Environment

	Drive Less			
Recycle	Always	Often	Sometimes	Never
Always	12	43	163	233
Often	4	21	99	185
Sometimes	4	8	77	230
Never	0	1	18	132

In order to find our $\hat{\beta}$ for Equation 8.7, we are going to first have to find the numerator and then find the denominator.

$$numerator = 1(43 + 99 + 230) + 2(163 + 185) + 3(233) = 1767$$

$$denominator = 1(4 + 8 + 18) + 2(4 + 1) + 3(0) = 40$$

This gives us $\hat{\beta} = \log\left(\frac{1767}{40}\right) = 3.79$, which implies that our estimated odds ratio is $e^{3.79} = 44.2$. This means that for each subject the estimated odds of response 'always' on recycling are 44.2 times the estimated odds of the response for driving less. This provides a lot of evidence that people would much rather recycle than drive less.

Now let's look at the alternative method which deals with the difference of sample means. First, we are going to denote the response {Always, Often, Sometimes, Never} by having a score that corresponds to {1,2,3,4}. This allows us to find the mean for driving less ($\bar{y}$) and our mean for recycling ($\bar{x}$).

$$\bar{x} = \frac{[451 + 2 * 309 + 3 * 319 + 4 * 151]}{1230} = 2.14$$

$$\bar{y} = \frac{[20 + 2 * 73 + 3 * 357 + 4 * 780]}{1230} = 3.54$$

Next, we need to calculate our standard error.

$$SE = \sqrt{\frac{1}{n}\left(\left(\frac{12}{1230}\right)(1-1)^2 + \left(\frac{43}{1230}\right)(1-2)^2 \ldots\right)}$$

$$= .05084$$

Thus, our z statistic is $z = \frac{2.14-3.54}{.05084} = -27.6$. This provides significantly enough evidence to reject the null hypothesis that there was marginal homogeneity. It also shows that the responses tended to be considerably more towards the 'Always' end of the response scale on recycling than on the driving less.

8.4 Symmetry and Quasi-Symmetry Models for Square Tables

A square table has proportions that have the property of *symmetry* if

$$\pi_{ij} = \pi_{ji} \qquad (8.8)$$

for all pairs of cells. In other words, the probabilities on one side of the main diagonal are mirror images of the probabilities on the other side of the main diagonal. Also, if a table has the property of symmetry than it also has marginal homogeneity, but for tables where $I > 2$ marginal homogeneity does not necessarily mean that the table is symmetric.

8.4.1 Symmetry as a Logistic Model

The logistic model for symmetry is

$$\log\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = 0 \quad for\ all\ i\ and\ j$$

The ML fit for this model has expected frequency estimates

$$\hat{\mu}_{ij} = (n_{ij} + n_{ji})/2$$

The expected frequencies have a couple special characteristics, $\hat{\mu}_{ij} = \hat{\mu}_{ji}$ and $\hat{\mu}_{ii} = n_{ii}$. These characteristics are all due to the fact of symmetry.

The symmetry model has standard residuals that equal

$$r_{ij} = (n_{ij} + n_{ji})/\sqrt{(n_{ij} + n_{ji})}$$

Having two residuals for each pairing of categories is redundant because $r_{ij} = -r_{ji}$. To test the goodness of fit, we can use the sum of squared residuals, one for each pairing of categories, and run a X^2 test.

8.4.2 Quasi-Symmetry

The symmetry model can easily fit the data poorly because the model is so simple. If there is the slightest difference between the marginal distributions, the model will not fit well. To compensate for marginal heterogeneity, we can use the *quasi-symmetry model,*

$$\log\left(\frac{\pi_{ij}}{\pi_{ji}}\right) = \beta_i - \beta_j \quad for\ all\ i\ and\ j \qquad (8.10)$$

The symmetry model is just a special case of this model, where all the $\beta_i's$ equal zero. To use the quasi-symmetry model you will need to use software. To use the software, we are going to want to ignore the main diagonal values where *i=j*, and treat each pair of cell counts (n_{ij}, n_{ji}) as an independent binomial variate. Next, we want to set up *I* dummy explanatory variables that correspond to the coefficients of the β_i parameters. Then for the logit $\log\left(\frac{\pi_{ij}}{\pi_{ji}}\right)$ for a given pair of categories, the variable $\beta_i = 1$, the variable $\beta_j = -1$, and the variables for all other parameters equal 0. This is much like the SAS code that we produced in Section 8.3.2. We are going to have one explanatory variable be redundant, so we will leave it out of the model because it will be implied.

8.6 Bradley-Terry Model for Paired Preferences

In this section, we are going to discuss a model that provides rankings between pair wise comparisons, which will help us to decide which category is better than another. This model will be easy to apply to comparisons between sports teams or products to find out which one would be better. This model also estimates the probabilities that one team, person, or product will win or lose over another.

To put this model into context right away, we are going to introduce an example. Table 8.9 below displays the results between five professional tennis players for the 2004-2005 year.

Table 8.9: Results of 2004-2005 Tennis Matches for Men Players

	Loser				
Winner	Agassi	Federer	Henman	Hewitt	Roddick
Agassi	-	0	0	1	1
Federer	6	-	3	9	5
Henman	0	1	-	0	1
Hewitt	0	0	2	-	3
Roddick	0	0	1	2	-

8.6.1 The Bradley-Terry Model

The Bradley-Terry model is a logistic model for paired preference data. This models deals with the probability that one player will defeat another. In Table 8.9, we will let Π_{ij} be the probability that player i wins over player j. The probability that player j wins over player i is $\Pi_{ji} = 1 - \Pi_{ij}$. For example, $\Pi_{23}$is the probability that Federer will defeat Henman, and $\Pi_{32}$is the probability that Henman will win over Federer.

The Bradley-Terry model has player parameters $\{\ \beta_i\}$ such that

$$logit(\Pi_{ij}) = \log\left(\frac{\Pi_{ij}}{\Pi_{ji}}\right) = \beta_i - \beta_j \qquad (8.14)$$

When $\beta_i = \beta_j$, $\Pi_{ij} = \Pi_{ji}$ is implied, which means that each player in this match up has a probability of winning equal to .5, and when the probability is greater than .5 then $\beta_i > \beta_j$. One downside to this model is that the data that it is analyzing cannot have any ties between the players, but it is still very useful.

This model is the same as the quasi-symmetry model 8.10. We can estimate our probabilities with this equation:

$$\widehat{\Pi}_{ij} = e^{\widehat{\beta}_i - \widehat{\beta}_j} \Big/ 1 + e^{\widehat{\beta}_i - \widehat{\beta}_j}$$

8.6.2 Example: Ranking Men Tennis Players

Let's run the Bradley-Terry model for Table 8.9, which has data on men's tennis players. Our SAS code is:

```
data tennis;

/*To run the Bradley-Terry model we would like to input our data so
  that we have a wins column, a total number of matches n, and a column
  for wach player */

input win n agassi federer henman hewitt roddick;
cards;

/* each one of these 'blocks' represents how one player did. Each row
   compares that player to one other. The first column tells us how
   many times that specific player beat the other player. The next
   column then shows the number of times those two players met. The
   next five columns indicate, which players we are talking about. For
   each block we put a '1' in the column of the player we are talking
   about, and then we put a '-1' in the place of the player who that
   person is playing against. */

/*Agassi's block*/
0 6  1 -1  0  0  0
0 0  1  0 -1  0  0
1 1  1  0  0 -1  0
1 1  1  0  0  0 -1
```

```
/* Federer's block*/
6 6 -1  1  0  0  0
3 4  0  1 -1  0  0
9 9  0  1  0 -1  0
5 5  0  1  0  0 -1

/*Henman's block*/
0 0 -1  0  1  0  0
1 4  0 -1  1  0  0
0 2  0  0  1 -1  0
1 2  0  0  1  0 -1

/*Hewitt's Block*/
0 1 -1  0  0  1  0
0 9  0 -1  0  1  0
2 2  0  0 -1  1  0
3 5  0  0  0  1 -1

/*Roddicks block*/
0 1 -1  0  0  0  1
0 5  0 -1  0  0  1
1 2  0  0 -1  0  1
2 5  0  0  0 -1  1
;
run;

/* In our proc genmod statement there are a couple new options. The
   first is the 'noint' option. This allows for the intercept to be 0,
   which is what we want in a Bradley Terry model. */

/* The next new option is the covb, which creates an estimated
   covariance matrix. We will need this matrix when we compute
   confidence intervals for the difference in β values */

proc genmod;
model win / n = agassi federer henman hewitt roddick / dist=bin
link=logit noint covb;
run;
```

This code will give us the following output (note that I have removed some output to emphasize the parts that we will be focusing on):

Estimated Covariance Matrix

	Prm2	Prm3	Prm4	Prm5
Prm2	0.96546	0.53327	0.13702	0.20008
Prm3	0.53327	0.86670	0.17268	**0.21386**
Prm4	0.13702	0.17268	0.55449	0.16222
Prm5	0.20008	**0.21386**	0.16222	0.31893

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error
Intercept	0	0.0000	0.0000
agassi	1	1.4489	0.9826
federer	1	**3.8815**	**0.9310**
henman	1	0.1875	0.7446
hewitt	1	**0.5734**	**0.5647**
roddick	0	0.0000	0.0000

Since Roddick was our last player put into SAS, SAS has set $\hat{\beta}_5 = 0$, which tells us that Roddick will be our base player for this data. The $\hat{\beta}'s$ for the other players are $\hat{\beta}_1 = 1.449$ for Agassi, $\hat{\beta}_2 = 3.882$ for Federer, $\hat{\beta}_3 = 0.188$ for Henman, and $\hat{\beta}_4 = 0.573$ for Hewitt. From these $\hat{\beta}'s$ alone, we can tell that Federer is ranked the highest and Roddick is ranked the lowest, but these rankings are only the beginning to the Bradley-Terry model.

Now that we have these $\hat{\beta}'s$, we can compute the probability that any player will win over another. For example, let's compare Federer to Hewitt and discover the probability that Federer would defeat Hewitt. To do this, we will use Model 8.14

$$\hat{\Pi}_{24} = e^{\hat{\beta}_2 - \hat{\beta}_4} \Big/ 1 + e^{\hat{\beta}_2 - \hat{\beta}_4} = e^{3.309} \Big/ 1 + e^{3.309} = .9647$$

This probability is extremely high, yet it is not equal to 1 even though Federer beat Hewitt 9 out of 9 times. This is good because it is more realistic, since there is always going to be a small chance of an upset.

The next thing we can do is create a confidence interval for this number. To do this we are going to use this formula as our standard error:

$$SE = \sqrt{var(\hat{\beta}_i) + var(\hat{\beta}_j) - 2cov(\hat{\beta}_i \hat{\beta}_j)}$$

Our standard error for our example of Federer and Hewitt is

$$SE = \sqrt{(.9310^2 + .5647^2 - 2 * .2139)} = .8705$$

Thus, a 95% confidence interval for $\hat{\beta}_2 - \hat{\beta}_4$ is 3.309 ± 1.96*.8705, which is (1.603, 5.015). This leads us to a 95% confidence interval for the probability that Federer will win (.832, .993).

$$\hat{\Pi}_{24} = {e^{1.603}}/{1 + e^{1.603}} = .832 \quad and \quad \hat{\Pi}_{24} = {e^{5.015}}/{1 + e^{5.015}} = .993$$

This confidence interval tells us that Federer is extremely likely to win because the confidence interval is entirely above .5.

Another downside to this model is that it assumes that each event is independent and identical. This assumption may be false because of some confounding variables such as which types of courts the matches were played on, or what city each match was played in. Some players may have an advantage under certain circumstances, and the Bradley-Terry model does not take this into account.

8.6.2.a Example: MLB-National League West

The Bradley-Terry model is very practical when analyzing the world of sports. In October 2010, the San Francisco Giants and the San Diego Padres were to meet for the last game of the Major League Baseball (MLB) season. If the Giants won they would be able to advance into the playoffs. If the Padres won, they would force a tie breaker between the Giants and the Padres to see who would be the division champions and

advance to the playoffs. Let's use the Bradley-Terry model and find out what the odds would be for the Giants to win this last game of the season against the Padres.

Before we dive into this let's look at the Bradley-Terry model and how it can be used in baseball. First, baseball is perfect for the model because the model only works for untied data. Since there are extra innings in baseball, we have no ties. For this example, we are only going to be comparing teams that are in the National League (NL) West Division. We want to look at the teams in a division because they play each other team in the division multiple times and about the same amount. Table 8.9.a summarizes how the 2010 MLB season went in the NL West minus the last game between the Giants and Padres.

Table 8.9.a: NL West 2010 Outcomes

	Loser				
Winner	Giants	Padres	Dodgers	Rockies	D-Backs
Giants, SF	-	5	10	9	14
Padres, SD	12	-	10	6	10
Dodgers, LA	8	8	-	11	13
Rockies, Col	9	11	7	-	9
D-Backs, AZ	5	8	5	9	-

The SAS code to analyze this table is:

```
data nlwest;
input win n SF SD LA COL AZ;
cards;

/*Giants Block*/
5  17  1 -1  0  0  0
10 18  1  0 -1  0  0
9  18  1  0  0 -1  0
14 19  1  0  0  0 -1

/*Padres Block*/
12 17 -1  1  0  0  0
10 18  0  1 -1  0  0
6  18  0  1  0 -1  0
10 18  0  1  0  0 -1

/*Dodgers Block*/
```

```
8   18 -1   0   1   0   0
8   18   0 -1   1   0   0
11  18   0   0   1 -1   0
13  18   0   0   1   0 -1


/*Rockies Block*/
9   18 -1   0   0   1   0
12  18   0 -1   0   1   0
7   18   0   0 -1   1   0
9   18   0   0   0   1 -1

/*D-Backs Block*/
5   19 -1   0   0   0   1
8   18   0 -1   0   0   1
5   18   0   0 -1   0   1
9   18   0   0   0 -1   1
;
run;
proc genmod;
model win / n = SF SD LA COL AZ / dist=bin link=logit noint covb;
run;
```

With this code we are able to produce the following covariance matrix and the following coefficients:

Estimated Covariance Matrix

	Prm2	Prm3	Prm4	Prm5
Prm2	0.04590	**0.02339**	0.02372	0.02357
Prm3	**0.02339**	0.04674	0.02401	0.02385
Prm4	0.02372	0.02401	0.04675	0.02392
Prm5	0.02357	0.02385	0.02392	0.04627

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Estimate	Standard Error
Intercept	0	0.0000	0.0000
SF	1	**0.5119**	**0.2142**
SD	1	**0.5407**	**0.2162**
LA	1	0.6082	0.2162
COL	1	0.4729	0.2151
AZ	0	0.0000	0.0000

This model has the Arizona Diamond Backs (D-Backs) as its base, so $\hat{\beta}_5 = 0$. The $\hat{\beta}'s$ for the other teams are $\hat{\beta}_1 = 0.512$ for the San Francisco Giants, $\hat{\beta}_2 = 0.541$ for the San Diego Padres, $\hat{\beta}_3 = 0.608$ for the Los Angeles Dodgers, and $\hat{\beta}_4 = 0.473$ for the Colorado Rockies. This model ranks the Dodgers as the best team and the D-Backs as the worst team.

Let's now compute the probability that the Giants will defeat the Padres in the last game of the season.

$$\hat{\Pi}_{12} = e^{\hat{\beta}_1 - \hat{\beta}_2} \Big/ 1 + e^{\hat{\beta}_1 - \hat{\beta}_2} = e^{-0.029} \Big/ 1 + e^{-0.029} = .4928$$

This probability is extremely close to .5 meaning that the chance of the Giants defeating the Padres is pretty much a fifty-fifty chance. We can compute the standard error to find a confidence interval and see if it contains 0.50.

$$SE = \sqrt{(.9310^2 + .5647^2 - 2 * .2139)} = .2141$$

A 95% confidence interval of the difference between the β coefficients is $-0.029 \pm 1.96(0.2141) = (-.4486,\ .3906)$. From this, we can then create a 95% confidence interval for the probability that San Francisco will win against the Padres, which is (.390, .596).

$$\hat{\Pi}_{24} = e^{-0.4486} \Big/ 1 + e^{-0.4486} = .390 \quad and \quad \hat{\Pi}_{24} = e^{0.3906} \Big/ 1 + e^{0.3906} = .596$$

This interval confirms that each team, the Giants and the Padres, were well matched for their next game.

In case you were wondering, the Giants won that last game of the 2010 season and went on to win the World Series, but it all started with this one win against an equally matched team.

Chapter 9 Modeling Correlated, Clustered Responses

Often when we conduct an experiment or do an observational study, we discover that there are subgroups in our sample that have similar traits, which could be affecting our results. For example, if we were to sample kids from a grammar school and observe their GPA, we might see that kids that belong to the same family tend to achieve the same GPA. We would call each family a *cluster*. A *cluster* is a set of observations that have similar traits. They can range from being the results of individuals belonging to a particular family/litter to the results of one particular person over time.

These results that we obtain from a cluster are often correlated with one another. To account for this correlation, we will need to analyze the data differently then we have in the past. In this chapter, we will discuss the marginal models and introduce the conditional models that describe this kind of data, and we will introduce the *generalized estimating equation* in Section 9.2.

9.1 Marginal Models versus Conditional Models

Like our previous models, the purpose of a clustered observations model is to find the probability of a response based on the explanatory variables. These types of models are most commonly used for longitudinal studies. For example, a longitudinal study might try to predict the probability of having a disease based on the drug treatment and the amount of time that has passed.

9.1.1 Marginal Models for a Clustered Binary Response

Before we begin creating our models, we have to denote a few things. Let the number of observations from a particular cluster be *T*. We need a notation for this because clusters tend to vary in size. For example, the number a kids a family has in a school could vary from having 1 to 5 kids. We will then denote each observation in a cluster as $(Y_1, Y_2, \dots, Y_T)$.

When the response is binary (Success or Failure), the *T* success probabilities $\{P(Y_1 = 1), P(Y_2 = 1), \dots, P(Y_T = 1)\}$ are marginal probabilities of a *T*-dimensional contingency table that cross classifies the *T* observations. If we were to take the logit of these marginal probabilities $\{logit[P(Y_1 = 1)]\}$, we would find a way to describe how the marginal probabilities depend on the explanatory variables. These models will be further analyzed in Section 9.2.

9.1.2 Example: Longitudinal Study of Treatments for Depression

A longitudinal study was done that compared a new drug with a standard drug for treating subjects suffering mental depression. Subjects were separated in to two groups: one was for those with a mild case of depression and the other was for those with a severe case of depression. Then in each group, subjects were randomly assigned to either taking the new drug or taking the standard drug. Following 1 week, 2 weeks, and 4 weeks after the initial treatment, the subjects were classified normal (N) or abnormal (A) to describe how their suffering of mental depression was going. The data for this study can be seen in Table 9.1.

Table 9.1: Cross-Classifcation of Responses on Depression ar Three Times (N = Normal, A = Abnormal) by Treatment and Diagnosis Severity

		Response at Three Times							
Diagnosis Severity	Treatment	NNN	NNA	NAN	NAA	ANN	ANA	AAN	AAA
Mild	Standard	16	13	9	3	14	4	15	6
Mild	New Drug	31	0	6	0	22	2	9	0
Severe	Standard	2	2	8	9	9	15	27	28
Severe	New Drug	7	2	5	2	31	5	32	6

In this experiment, a single subject has a binary response of normal or abnormal, and he or she responds three times. Thus, our clusters are the individual person's responses where $T = 3$. These three depression assessments form a multivariate response with three parts, with $Y_t = 1$ for normal and 0 for abnormal at time *t.* Table 9.1 shows a 2 x 2 x 2 table for every possible combination of responses. This table essentially shows 12 marginal distributions.

We can tell a little more about the data if we were to look at a table that displayed the sample proportions of normal responses for the 12 marginal distributions over the three time periods. From this table, we would be able to tell if there is a time effect and/ or a drug effect. Table 9.2 does exactly this.

Table 9.2: Sample Marginal Proportions of Normal Response for Depression Data of Table 9.1

		Sample Proportion		
Diagnosis Severity	Treatment	Week 1	Week 2	Week 4
Mild	Standard	0.51	0.59	0.68
	New Drug	0.53	0.79	0.97
Severe	Standard	0.21	0.28	0.46
	New Drug	0.18	0.5	0.83

To obtain these values, we will examine the cell for mild depression, with a standard treatment for week 1. We calculated 0.51 by looking back at Table 9.1 and observed the

entire first row, which is the data for mild depression, with a standard treatment. Then, because we want the proportion of all those with a normal response for the first week, we counted all the values where the response of the first week was normal (N--). Then we divide that number by the total number of observations with mild depression and a standard treatment.

$$\frac{16 + 13 + 9 + 3}{16 + 13 + 9 + 3 + 14 + 4 + 15 + 6} = .51$$

From Table 9.2 we want to describe what the table tells us about the diagnosis effect, the drug effect, and the time effect. Looking at the table we can tell that between the two diagnoses 'Mild' has a higher sample proportion of normal responses, between the two drugs the 'new drug' has a higher sample proportion of normal responses, and that over time the sample proportion of normal responses increases.

To construct a main effects model for this data, we will let *s* be the initial severity of depression (*s*=1 for severe and 0 for mild). Let *d* denote the type of drug the subject is using (*d* =1 for new drug and 0 for standard drug), and let *t* denote the time of measurement. We will use scores (0, 1, 2), the logs to base 2 of the week numbers 1,2 and 4 because a logit scale usually has an approximate linear effect for the logarithm of time. Next, we will have $P(Y_t = 1)$ denote the probability of a normal response at time *t* for a randomly selected subject. Our model, which shows how our response depends on the explanatory variables (*s, d, t*), would look like:

$$logit[P(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t$$

This model implies that the time effect β_3 is linear and the same for each group. Unfortunately, when we look at Table 9.2, we see that the time effect is stronger for the

new drug than for the standard drug. Thus, we can create a model that has an interaction term for drug-by-time,

$$logit[P(Y_t = 1)] = \alpha + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (d \times t)$$

In this model, the time effect to describe when the standard drug is used is just β_3, but the time effect to describe when the new drug is used would now be $\beta_3 + \beta_4$.

In Section 9.2, we will analyze this model further.

9.1.3 Conditional Models for a Repeated Response

The effects of the marginal models that we described in the previous section are *population-average,* because they average over the entire population rather than looking at each subject specifically. As in Section 8.2.3, we can create a model that is defined at each subject level.

Let Y_{it} be the response for subject *i* at time *t*. Relating back to the previous example dealing with depression, our *subject-specific* model looks like this:

$$logit[P(Y_{it} = 1)] = \alpha_i + \beta_1 s + \beta_2 d + \beta_3 t + \beta_4 (d \times t)$$

In this subject-specific model, the variation between each subject at a particular *s, d,* and *t* is described by the α_i term. This model is known as the *conditional model,* because the effects are conditional on the subject.

The models that we will be talking about for the rest of the chapter describe a *population-average* effect. These models are the marginal models.

9.2 Marginal Modeling: The Generalized Estimating Equations (GEE) Approach

Because ML fitting marginal logit models is so difficult, we are going to introduce a new method for fitting these marginal models.

9.2.1 Quasi-Likelihood Methods

When we use the GLM, we have to first specify a distribution for *Y*. With this distribution, we can then figure out a formula for how its mean $E(Y) = \mu$ depends on the explanatory variables by using a link function to connect the mean to a linear predictor. The distribution of *Y* explains how μ and the variance of μ are related. For example, when the data is binary with a probability of success equal to π, an observation *Y* has $E(Y) = \pi$ and $Var(Y) = \pi(1 - \pi)$. When we use count data with a Poison distribution, the $Var(Y) = \mu$.

When we use the ML method, we have to assume a particular type of distribution for *Y* in order to find a formula which displays how μ depends on the explanatory variables. Another method makes an assumption about the relationship between μ and the $Var(Y)$. This is the *quasi-likelihood method* that we learned in Section 8.4.2. This method does not assume overdispersion for correlated data or unobserved explanatory variables. *Overdispersion* is the presence of greater variability in a data set than would be expected based on a given statistical method. Or simply put, it is having a larger variance than expected. It does this by multiplying the typical variance formula by a constant that is estimated by the data.

For example, let our data be clustered binary data, with *n* subjects in a cluster. The subjects within a cluster are most likely correlated because they share common traits,

which is why they are in a cluster in the first place. The variance of the number of successes in a cluster is probably different from the variance $n\pi(1-\pi)$ for a binomial distribution because a binomial distribution assumes the trials to be uncorrelated or independent. The quasi-likelihood method states that the variance of the number of successes is a multiple, φ, of the usual variance, so our variance is $\varphi n\pi(1-\pi)$, where φ is estimated based on the observed variance of the sample data. When φ > 1, we have overdispersion.

9.2.2 Generalized Estimating Equation Methodology: Basic Ideas

As we have mentioned before, the ML method for clustered categorical data can easily be complicated, but a mathematically simpler alternative is a multivariate generalization of the quasi-likelihood method. This generalization links each marginal mean to a linear predictor and provides a guess for the variance-covariance structure of $(Y_1, \dots, Y_t)$ as opposed to assuming a distribution for $(Y_1, \dots, Y_t)$. This method, like the quasi-likelihood method, uses the variability of the data to create standard errors. This method is called the *GEE method* because the estimates are solutions of *generalized estimating equations.*

After we have declared a marginal model for each Y_t, we must do 2 things for the *GEE* method:

1. We must assume a distribution for each Y_t. This will help us find the relationship between $Var(Y_t)$ and $E(Y_t)$.

2. We must make a guess of what the correlation structure among the $\{Y_t\}$ might look like. This is called the *working correlation* matrix.

There are four types of structures a working correlation matrix might have. Each one assumes something different about the data. The first assumes that $\rho = Corr(Y_s, Y_t)$ is the same for all pairs of *s* and *t.* This is known as the *exchangeable* structure. The next structure assumes that observations that are further apart in time are less correlated. This is called the *autoregressive structure* and is commonly used in time series analysis. It has the form $Corr(Y_s, Y_t) = \rho^{t-s}$. The next type of structure treats all observations as if they were uncorrelated. This is the independence structure and has a working correlation matrix in the form of the identity matrix. It has a $Corr(Y_s, Y_t) = 0$. The last structure is the *unstructured* working correlation matrix, which allows $Corr(Y_s, Y_t)$ to differ between every pair of *s* and *t.*

The working correlation matrix is just a starting point for the GEE. Even if we do not choose the best structure for the working correlation matrix, the GEE will still produce robust standard errors because the information that the sample data provides about dependence will update our initial structure. But, slightly more efficient estimates can be achieved by choosing the right structure. So if you are unsure about which type of structure to use, use the exchangeable structure.

9.2.3 GEE for Binary Data: Depression Study

Let's revisit Table 9.1 and try to analyze the data using this new GEE method. The SAS code can be seen below.

```
data depress;
/* I have not included all of the data in the SAS code so that it would
   be easier to read and duplicate with another study*/
/* the case denotes the subject that is being tested three times*/
/* Diagnose is a variable that is 0 if the person had a mild depression
   And 1 if the person had severe depression. */
/* Treat is a variable that is 0 if the person took the standard drug
   or 1 if the person took the new drug */
/* Time is either 0,1,2 representing if the time was the 1 week, 2 week
   Or 4 week. */
/* Outcome equals 1 if the response was normal or 0 if abnormal*/


input case diagnose treat time outcome ; * outcome=1 is normal;
datalines;


  1  0  0  0  1
  1  0  0  1  1
  1  0  0  2  1
  2  0  0  0  1
  2  0  0  1  1
  2  0  0  2  1
  3  0  0  0  1
  3  0  0  1  1
  3  0  0  2  1
  4  0  0  0  1
  4  0  0  1  1
  4  0  0  2  1
  5  0  0  0  1
  5  0  0  1  1
  5  0  0  2  1
  6  0  0  0  1
  6  0  0  1  1
  6  0  0  2  1
  7  0  0  0  1
  7  0  0  1  1
  7  0  0  2  1
  8  0  0  0  1
  8  0  0  1  1
  8  0  0  2  1
  9  0  0  0  1
  9  0  0  1  1
  9  0  0  2  1
.
.
.
.
330  1  1  2  0
331  1  1  0  0
331  1  1  1  0
331  1  1  2  0
332  1  1  0  0
332  1  1  1  0
332  1  1  2  0
333  1  1  0  0
333  1  1  1  0
```

```
333  1  1  2  0
334  1  1  0  0
334  1  1  1  0
334  1  1  2  0
335  1  1  0  0
335  1  1  1  0
335  1  1  2  0

;
run;


/* In the proc genmod statement there are three new options*/
/* The repeated subject option shows which variable is the cluster. In
   this situation the cluster was the person*/
/* The type option chooses a working correlation matrix. Here we chose
   the exchangeable structure. This could also be type=AR for
   autoregressive, type=INDEP for independence, and type=UNSTR for the
   unstructured correlation matrix. */
/* The last new command is the corrw, which displays the working
   correlation matrix */
proc genmod descending; class case;
model outcome = diagnose treat time treat*time / dist=bin link=logit
type3;
repeated subject=case / type=exch corrw;
run;
```

This produces the working correlation matrix:

Working Correlation Matrix

	Col1	Col2	Col3
Row1	1.0000	-0.0034	-0.0034
Row2	-0.0034	1.0000	-0.0034
Row3	-0.0034	-0.0034	1.0000

Exchangeable Working Correlation

Correlation	-0.003432732

Because the exchangeable correlation is -0.0034, we might think that the correct working correlation matrix for this model may be the one for independence. This is because -0.0034 is extremely close to 0. This is actually unusual behavior for repeated measurement data, but we will now rerun our SAS code with an independence structure.

```
proc genmod descending; class case;
model outcome = diagnose treat time treat*time / dist=bin link=logit
type3;
/*Here we changed the type to have an independent structure*/
repeated subject=case / type=INDEP corrw;
run;

proc genmod descending; class case;
model outcome = diagnose treat time treat*time / dist=bin link=logit
type3;
/* If we were to delete our new statements we would use ML to find our
   fitted values*/
run;
```

Table 9.3: Output from GEE to Fit Logistic Model to Table 9.1

Analysis Of Maximum Likelihood Parameter Estimates

Parameter	DF	Standard Estimate	Standard Error
Intercept	1	-0.0280	0.1639
diagnose	1	-1.3139	0.1464
treat	1	-0.0596	0.2222
time	1	0.4824	0.1148
treat*time	1	1.0174	0.1888

Analysis Of GEE Parameter Estimates Empirical Standard Error Estimates

Parameter	Standard Estimate	Standard Error
Intercept	-0.0280	0.1742
diagnose	-1.3139	0.1460
treat	-0.0596	0.2285
time	0.4824	0.1199
treat*time	1.0174	0.1877

Working CorrelationMatrix

	Col1	Col2	Col3
Row1	1.0000	0.0000	0.0000
Row2	0.0000	1.0000	0.0000
Row3	0.0000	0.0000	1.0000

Table 9.3 displays our GEE estimates based on our working correlation matrix with an independence structure. Next to the GEE output is the output obtained from using ML, which treats all observations as independent. The empirical standard errors use the sample dependence to adjust the independence-based standard errors.

$\hat{\beta}_3 = 0.482$ is the estimated time effect for the standard drug (d=0), and the estimated time effect for the new drug is $\hat{\beta}_3 + \hat{\beta}_4 = 1.50$. If we wanted to test if the interaction term was significant we would just run a hypothesis test with H_o: $\beta_4 = 0$. Our

test statistic would be $z = \frac{1.017-0}{0.188} = 5.4$, which yields a p-value < 0.0001. Thus, there is strong evidence that there is faster improvement with the new drug.

Now, let's find out more information about the other parameters. Holding drug and time constant, the estimated odds of a normal response when the initial diagnosis was severe depression equal $e^{-1.314} = 0.27$ times the estimated odds when the initial diagnosis was *mild* depression. This indicates that a normal response is more likely to occur for a diagnosis *mild*. The estimated drug effect $\hat{\beta}_2 = -0.060$ only applies when t = 0, which has an insignificant effect on the response after 1 week. But after 1 week, we start adding the interaction term to the drug effect and discover that the drug causes more normal responses.

All of these parameters (the severity, the drug treatment, and the time) have significant effects on whether or not the subject had a normal response.

9.2.4 Example: Teratology Overdispersion

In an experiment, female rats with low iron diets were assigned to 4 groups. The 1st group received placebo drugs, the 2nd group received iron supplement injections on days 7 and 10, the 3rd group received iron supplement injections on days 0 and 7, and the 4th group received iron injections weekly. The rats were then impregnated, and then killed. For each fetus in each rat's litter, the response was whether the fetus was dead. The data for this experiment can be found in Table 9.4.

Table 9.4: Response Counts of (litter Size, Number Dead) for 58 Litters of Rats in a Low-Iron Teratology Study

Group 1: Untreated (low iron)
(10,1)(11,4)(12,9)(4,4)(10,10)(11,9)(9,9)(11,11)(10,10)(10,7)(12,12)
(10,9)(8,8)(11,9)(6,4)(9,7)(14,14)(12,7)(11,9)(13,8)(14,5)(10,10)
(12,10)(13,8)(10,10)(14,3)(13,13)(4,3)(8,8)(13,5)(12,12)
Group 2: injections days 7 and 10
(10,1)(3,1)(13,1)(12,0)(14,4)(9,2)(13,2)(16,1)(11,0)(4,0)(1,0)(12,0)
Group 3: injections days 0 and 7
(8,0)(11,1)(14,0)(14,1)(11,0)
Group 4: injections weekly
(3,0)(13,0)(9,2)(17,2)(15,0)(2,0)(14,1)(8,0)(6,0)(17,0)

We are going to observe each fetus and claim that the litter it is from is a single cluster. Let y_i stand for the number of dead fetuses for the T_i fetuses in litter i. Then the probability of death for fetus t in litter i is π_{it}. Let $z_{ig} = 1$ if litter i is in group g and 0 if it is not.

For now, let's suppose that there is no clustering and that y_i is a $bin(T_i, \pi_{it})$ variate. Our model is

$$logit(\pi_{it}) = \alpha + \beta_2 z_{i2} + \beta_3 z_{i3} + \beta_4 z_{i4}$$

This model treats all litters in group g as having the same probability of death, which would be $\frac{e^{\alpha+\beta_g}}{1+e^{\alpha+\beta_g}}$ where $\beta_1 = 0$. In this model, we are comparing all the groups to the placebo group 1. Table 9.5 shows us the ML estimated β coefficients with their standard errors. There is enough evidence to conclude that the probability of death is lower for the treatment groups (groups 2, 3, and 4) than the placebo group (group1). To produce Table 9.5, we can use SAS.

```
data teratology;
/* I have not included all of the data in the SAS code so that it would
   be easier to read and duplicate with another study*/
/* the group1 variable is either 1 if you are in group 1 or 0
   otherwise */
/* the group2 variable is either 1 if you are in group 2 or 0
   otherwise */
/* the group3 variable is either 1 if you are in group 3 or 0
   otherwise */
/* the group4 variable is either 1 if you are in group 4 or 0
   otherwise */
/* Litter variable designates what litter that fetus belonged to */
/* dead is 1 if the fetus died or 0 if the fetus lived*/

input group1 group2 group3 group4 litter dead @@;
cards;
1 0 0 0 1 0  1 0 0 0 1 0  1 0 0 0 1 0  1 0 0 0 1 0  1 0 0 0 1 0  1 0 0
…
1 0 0 0 2 0  1 0 0 0 2 0  1 0 0 0 2 0  1 0 0 0 2 0  1 0 0 0 2 0  1 0 0
…
1 0 0 0 3 0  1 0 0 0 3 0  1 0 0 0 3 0  1 0 0 0 3 1  1 0 0 0 3 1  1 0 0
…
1 0 0 0 3 1
1 0 0 0 4 1  1 0 0 0 4 1  1 0 0 0 4 1  1 0 0 0 4 1
1 0 0 0 5 1  1 0 0 0 5 1  1 0 0 0 5 1  1 0 0 0 5 1  1 0 0 0 5 1  1 0 0
…
1 0 0 0 6 0  1 0 0 0 6 0  1 0 0 0 6 1  1 0 0 0 6 1  1 0 0 0 6 1  1 0 0
…
1 0 0 0 7 1  1 0 0 0 7 1  1 0 0 0 7 1  1 0 0 0 7 1  1 0 0 0 7 1  1 0 0
…
1 0 0 0 8 1  1 0 0 0 8 1  1 0 0 0 8 1  1 0 0 0 8 1  1 0 0 0 8 1  1 0 0
…
1 0 0 0 9 1  1 0 0 0 9 1  1 0 0 0 9 1  1 0 0 0 9 1  1 0 0 0 9 1  1 0 0
…
1 0 0 0 10 0  1 0 0 0 10 0  1 0 0 0 10 0  1 0 0 0 10 1  1 0 0 0 10 1  1
…
1 0 0 0 11 1  1 0 0 0 11 1  1 0 0 0 11 1  1 0 0 0 11 1  1 0 0 0 11 1  1
…
1 0 0 0 12 0  1 0 0 0 12 1  1 0 0 0 12 1  1 0 0 0 12 1  1 0 0 0 12 1  1
…
1 0 0 0 13 1  1 0 0 0 13 1  1 0 0 0 13 1  1 0 0 0 13 1  1 0 0 0 13 1  1
…
1 0 0 0 14 0  1 0 0 0 14 0  1 0 0 0 14 1  1 0 0 0 14 1  1 0 0 0 14 1  1
…
1 0 0 0 15 0  1 0 0 0 15 0  1 0 0 0 15 1  1 0 0 0 15 1  1 0 0 0 15 1  1
…
.
.
.
.
0 0 0 1 56 0  0 0 0 1 56 0  0 0 0 1 56 0  0 0 0 1 56 0  0 0 0 1 56 0  0
..
0 0 0 1 57 0  0 0 0 1 57 0  0 0 0 1 57 0  0 0 0 1 57 0  0 0 0 1 57 0  0
..
0 0 0 1 58 0  0 0 0 1 58 0  0 0 0 1 58 0  0 0 0 1 58 0  0 0 0 1 58 0  0
..
;
```

```
run;

/*This proc produced the GEE method*/

proc genmod descending ; class litter;
model dead = group2 group3 group4 group1 / dist = bin link = logit
type3;
repeated subject = litter / type=exch corrw;
run;

/*This proc produces the Binomial ML model*/
proc genmod descending ; class litter;
model dead = group2 group3 group4 group1 / dist = bin link = logit
type3;

run;
```

Table 9.5: Estimates and Standard Errors (in Parentheses) for Logistic Models Fitted to Teratology Data of Table 9.4

	Type of Logistic Model Fitting	
Parameter	Binomial ML	GEE
Intercept	1.14(0.13)	1.21(0.27)
Group 2	-3.32(0.33)	-3.37(0.43)
Group 3	-4.48(0.73)	-4.58(0.62)
Group 4	-4.13(0.48)	-4.25(0.60)
Overdispersion	None	$\hat{\rho} = 0.19$

The only problem with the ML model is that it assumes there is no correlation between all the observations, but we instinctively believe there should be. As you can see in Table 9.5, when we use the GEE method, the estimated within-cluster correlation is 0.19. This tells us that each observation is not independent from the next within a cluster, thus we should not use the ML model.

9.2.5 Limitations of GEE Compared with ML

The GEE does not specify the complete multivariate distribution, so it does not have a likelihood function. This means that its estimates are not ML estimates. The GEE only specifies the marginal distributions and the correlation structure.

Even though the GEE method is simpler for clustered data than ML, it has some draw backs because it does not have a likelihood function. Some of these include not being able to compare models, check the model fit, and conduct inference about parameters all because we cannot use the likelihood-ratio methods. But, when we do have large samples, we are able to use statistics, such as the Wald statistic, to make inferences because of their approximate normality of estimators along with their estimated covariance matrix. If there is not a large sample, then our empirical based errors usually are lower than the true standard errors.

But, overall the GEE method does take into account the within cluster correlation and is a lot less difficult to compute than the ML.