Low Birth Weight

Jack Tubbs

September 2021

Contents

Discussion of the Problem

The data contain information about infant mortality in 2003 and were obtained from the US National Center for Health Statistics. A random sample of 2,500+ observations is used in this example. This data are observational, in which case, meaningful inference is limited. The description below is for a causal inference example, which is beyond the scope of this course, given in SAS.

Our approach is to investigate this problem using the material given in the first part of Chapter 3 in the methods lecture notes.

The main variables in the analysis are as follows:

- The treatment variable is **Smoking**. It is an indicator of maternal smoking behavior, with values Yes and No.
- The outcome variable is **Death**. It is an indicator of infant death within one year of birth, with values Yes and No.

- The mediator variable is **LowBirthWgt**. It is an indicator of low birth weight (less than 2,500 grams), with values Yes and No.

The analysis also includes five confounding covariates:

- **AgeGroup** represents maternal ages of less than 20, between 20 and 35, and greater than 35, with values 1, 2, and 3, respectively.
- **Drinking** is an indicator of maternal drinking during pregnancy, with values Yes and No.
- **Married** is an indicator of marital status, with values Yes and No.
- **Race** is an indicator of race, with values Asian, Black, Hispanic, Native (native American), and White.
- **SomeCollege** is an indicator of whether the mother has 12 or more years of education, with values Yes and No.

R

Needed Packages

```
if(!require(FSA)){install.packages("FSA")}
if(!require(ggplot2)){install.packages("ggplot2")}
if (!require("mosaic")) install.packages("mosaic", dep=FALSE)
if (!require("nortest")) install.packages("nortest", dep=TRUE)
if (!require("epitools")) install.packages("epitools", dep=TRUE)
if (!require("prettyR")) install.packages("prettyR", dep=TRUE)
if (!require("rms")) install.packages("rms", dep=TRUE)
# add other as needed
```

Problem – Infant Birth Weight Data

Read data from SAS input file

```
# this data came from SASHELP.BWEIGHT
bw = read.csv('bwgt.csv', header = TRUE)
bw = data.frame(bw)
#summary(bw)
bw = transform(bw, AgeGroup.f = as.factor(AgeGroup))
bw = transform(bw, Race.f = as.factor(Race))
bw = transform(bw, Drinking.f = as.factor(Drinking))
bw = transform(bw, Death.f = as.factor(Death))
bw = transform(bw, Smoking.f = as.factor(Smoking))
bw = transform(bw, SomeCollege.f = as.factor(SomeCollege))
bw = transform(bw, LowBirthWgt.f = as.factor(LowBirthWgt))
```

```
tally(~ AgeGroup + Race.f, data=bw)
```

```
##         Race.f
## AgeGroup Asian Black Hispanic Native White
##        1     8    91       83      6   169
##        2   101   375      475     22  1337
##        3    36    52       66      4   264
```

```
tally(~ Race.f | AgeGroup.f, data=bw)
```

```
##           AgeGroup.f
## Race.f        1    2    3
##   Asian       8  101   36
```

```
##   Black      91  375   52
##   Hispanic   83  475   66
##   Native      6   22    4
##   White     169 1337  264
```

```
library(mosaic)
mytab = tally(~ Race.f | AgeGroup.f, data=bw)
addmargins(mytab)
```

```
##           AgeGroup.f
## Race.f        1    2    3  Sum
##   Asian       8  101   36  145
##   Black      91  375   52  518
##   Hispanic   83  475   66  624
##   Native      6   22    4   32
##   White     169 1337  264 1770
##   Sum       357 2310  422 3089
```

```
prop.table(mytab, 1)
```

```
##           AgeGroup.f
## Race.f              1          2          3
##   Asian    0.05517241 0.69655172 0.24827586
##   Black    0.17567568 0.72393822 0.10038610
##   Hispanic 0.13301282 0.76121795 0.10576923
##   Native   0.18750000 0.68750000 0.12500000
##   White    0.09548023 0.75536723 0.14915254
```

```
library(epitools)
attach(bw)
mytab = tally(~ LowBirthWgt.f | Death.f, data=bw)
addmargins(mytab)
```

```
##              Death.f
## LowBirthWgt.f   No  Yes  Sum
##           No  2278  198 2476
##           Yes  205  408  613
##           Sum 2483  606 3089
```

```
prop.table(mytab, 1)
```

```
##              Death.f
## LowBirthWgt.f         No        Yes
##           No  0.92003231 0.07996769
##           Yes 0.33442088 0.66557912
```

```
riskratio(x=Smoking.f, y=Death.f)
```

```
## $data
##          Outcome
## Predictor   No Yes Total
##            156  41   197
##     No    1786 405  2191
##     Yes    541 160   701
##     Total 2483 606  3089
##
## $measure
##          risk ratio with 95% C.I.
```

```
## Predictor  estimate     lower    upper
##           1.0000000        NA       NA
##        No  0.8881678 0.6670955 1.182502
##       Yes 1.0966911 0.8087967 1.487063
##
## $p.value
##          two-sided
## Predictor midp.exact fisher.exact chi.square
##                  NA           NA         NA
##        No   0.4206715    0.4449380  0.4220298
##       Yes   0.5556210    0.6286442  0.5493629
##
## $correction
## [1] FALSE
##
## attr(,"method")
## [1] "Unconditional MLE & normal approximation (Wald) CI"
```

SAS

Code 1

```
/*
The Sashelp.BirthWgt data set contains 100,000 random observations about
infant mortality in 2003 from the US National Center for Health Statistics.
Each observation records infant death  within one year of birth, birth weight,
maternal smoking and drinking behavior, and other background characteristics
of the mother.
*/

title "Sashelp.bweight --- Infant Birth Weight";
data birthwgt; set sashelp.birthwgt;
run;

proc contents data=birthwgt varnum;
ods select position;
run;

title "The First Five Observations Out of 100,000";
proc print data=birthwgt(obs=10);
run;
```

Sashelp.bweight — Infant Birth Weight

The CONTENTS Procedure

Variables in Creation Order			
#	Variable	Type	Len
1	LowBirthWgt	Char	3
2	Married	Char	3
3	AgeGroup	Num	8
4	Race	Char	9
5	Drinking	Char	3

Variables in Creation Order			
#	Variable	Type	Len
6	Death	Char	3
7	Smoking	Char	3
8	SomeCollege	Char	3

The First Five Observations Out of 100,000

Obs	LowBirthWgt	Married	AgeGroup	Race	Drinking	Death	Smoking	SomeCollege
1	No	No	3	Asian	No	No	No	Yes
2	No	No	2	White	No	No	No	No
3	Yes	Yes	2	Native	No	Yes	No	No
4	No	No	2	White	No	No	No	No
5	No	No	2	White	No	No	No	Yes
6	No	No	2	White	No	No	No	
7	No	No	2	Asian	No	No	No	Yes
8	No	No	3	White	No	No	No	Yes
9	No	Yes	1	Black	No	No	No	No
10	No	No	2	Native	No	No	No	Yes

Code 2

I have changed 'Yes' responses to 'Affirm' as SAS orders the variables in the tables using an alphabetical ordering. This new order allows one to have a better interpretation of results.

```
*Create a new smaller data set;
title 'New Sample of Size 2,500';
proc surveyselect data=birthwgt out=new2 method=srs n=2500
                  seed=2021;
run;

/* I needed more death records than the srs gave me */
data new; set birthwgt; if death = 'Yes';
run;

/*merge the two files into one */
data new_bwgt; set new new2;
run;

data new_bwgt; set new_bwgt;
if LowBirthWgt = 'Yes' then LowBirthWgt = 'Affirm';
if Death = 'Yes' then Death = 'Affirm';
if Smoking = 'Yes' then Smoking = 'Affirm';
if Drinking = 'Yes' then Drinking = 'Affirm';


title 'Test for Association between Low Birth Weight and Smoking';
proc freq data=new_bwgt;* order=freq;
tables smoking*LowBirthWgt/norow nopercent chisq relrisk riskdiff;
run;
```

```
title 'Test for Association between Low Birth Weight and drinking';
proc freq data=new_bwgt;* order=freq;
tables drinking*LowBirthWgt/norow nopercent chisq relrisk riskdiff;
run;
```

New Sample of Size 2,500

The SURVEYSELECT Procedure

Selection Method	Simple Random Sampling

Input Data Set	BIRTHWGT
Random Number Seed	2021
Sample Size	2500
Selection Probability	0.025
Sampling Weight	40
Output Data Set	NEW2

Test for Association between Low Birth Weight and Smoking

The FREQ Procedure

Table of Smoking by LowBirthWgt			
Smoking	*LowBirthWgt*		
	Aff	*No*	*Total*
Aff	155 26.96	546 23.56	701
No	420 73.04	1771 76.44	2191
Total	575	2317	2892
Frequency Missing = 197			

Note	Statistics for Table of Smoking by LowBirthWgt

In the following table there is not a significant association at the .05 level between Low Birth Weight and Smoking. This is seen in the chi-square statistic and the relative risk and odds ratio.

Statistic	*DF*	*Value*	*Prob*
Chi-Square	1	2.8856	0.0894
Likelihood Ratio Chi-Square	1	2.8341	0.0923
Continuity Adj. Chi-Square	1	2.7038	0.1001
Mantel-Haenszel Chi-Square	1	2.8846	0.0894
Phi Coefficient		0.0316	
Contingency Coefficient		0.0316	
Cramer's V		0.0316	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	155
Left-sided Pr <= F	0.9593
Right-sided Pr >= F	0.0510
Table Probability (P)	0.0102
Two-sided Pr <= P	0.0921

Column 1 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.2211	0.0157	0.1904	0.2518	0.1909	0.2537
Row 2	0.1917	0.0084	0.1752	0.2082	0.1754	0.2088
Total	0.1988	0.0074	0.1843	0.2134	0.1844	0.2138
Difference	0.0294	0.0178	−0.0054	0.0643		
Difference is (Row 1 - Row 2)						

Column 2 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.7789	0.0157	0.7482	0.8096	0.7463	0.8091
Row 2	0.8083	0.0084	0.7918	0.8248	0.7912	0.8246
Total	0.8012	0.0074	0.7866	0.8157	0.7862	0.8156
Difference	−0.0294	0.0178	−0.0643	0.0054		
Difference is (Row 1 - Row 2)						

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	1.1970	0.9725	1.4734
Relative Risk (Column 1)	1.1535	0.9796	1.3582
Relative Risk (Column 2)	0.9636	0.9218	1.0074

Note	Sample Size = 2892 Frequency Missing = 197

Test for Association between Low Birth Weight and drinking

In the following table there is not a significant association at the .05 level between Low Birth Weight and Drinking. This is seen in the chi-square statistic and the relative risk and odds ratio.

The FREQ Procedure

Table of Drinking by LowBirthWgt			
Drinking	*LowBirthWgt*		
	Aff	*No*	*Total*
Aff	74 12.87	325 14.03	399
No	501 87.13	1992 85.97	2493
Total	575	2317	2892
Frequency Missing = 197			

Note	Statistics for Table of Drinking by LowBirthWgt

Statistic	*DF*	*Value*	*Prob*
Chi-Square	1	0.5187	0.4714
Likelihood Ratio Chi-Square	1	0.5263	0.4682
Continuity Adj. Chi-Square	1	0.4260	0.5140
Mantel-Haenszel Chi-Square	1	0.5185	0.4715
Phi Coefficient		−0.0134	
Contingency Coefficient		0.0134	
Cramer's V		−0.0134	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	74
Left-sided Pr <= F	0.2588
Right-sided Pr >= F	0.7835
Table Probability (P)	0.0423
Two-sided Pr <= P	0.4998

Column 1 Risk Estimates						
	Risk	*ASE*	*95% Confidence Limits*		*Exact 95% Confidence Limits*	
Row 1	0.1855	0.0195	0.1473	0.2236	0.1485	0.2271
Row 2	0.2010	0.0080	0.1852	0.2167	0.1854	0.2172
Total	0.1988	0.0074	0.1843	0.2134	0.1844	0.2138
Difference	−0.0155	0.0210	−0.0568	0.0258		
Difference is (Row 1 - Row 2)						

Column 2 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.8145	0.0195	0.7764	0.8527	0.7729	0.8515
Row 2	0.7990	0.0080	0.7833	0.8148	0.7828	0.8146
Total	0.8012	0.0074	0.7866	0.8157	0.7862	0.8156
Difference	0.0155	0.0210	−0.0258	0.0568		
Difference is (Row 1 - Row 2)						

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	0.9053	0.6906	1.1869
Relative Risk (Column 1)	0.9229	0.7406	1.1500
Relative Risk (Column 2)	1.0194	0.9689	1.0725

Note	Sample Size = 2892 Frequency Missing = 197

Code 3

```
title 'Test for Association between Low Birth Weight and Smoking';
title2 'Controlling for Death';
proc freq data=new_bwgt;* order=freq;
tables death*smoking*LowBirthWgt /nopercent norow chisq cmh;
run;

title 'Test for Association between Low Birth Weight and Drinking';
title2 'Controlling for Death';
proc freq data=new_bwgt;* order=freq;
tables death*drinking*LowBirthWgt /nopercent norow chisq cmh;
run;
```

Test for Association between Low Birth Weight and Smoking

Controlling for Death

The FREQ Procedure

Table 1 of Smoking by LowBirthWgt			
Controlling for Death=Aff			
Smoking	LowBirthWgt		
	Aff	No	Total
Aff	102 26.63	58 31.87	160
No	281 73.37	124 68.13	405
Total	383	182	565
Frequency Missing = 41			

Note	Statistics for Table 1 of Smoking by LowBirthWgt Controlling for Death=Aff

Statistic	DF	Value	Prob
Chi-Square	1	1.6664	0.1967
Likelihood Ratio Chi-Square	1	1.6467	0.1994
Continuity Adj. Chi-Square	1	1.4185	0.2337
Mantel-Haenszel Chi-Square	1	1.6635	0.1971
Phi Coefficient		−0.0543	
Contingency Coefficient		0.0542	
Cramer's V		−0.0543	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	102
Left-sided Pr <= F	0.1172
Right-sided Pr >= F	0.9172

Fisher's Exact Test	
Table Probability (P)	0.0344
Two-sided Pr <= P	0.2304

Note	Sample Size = 565 Frequency Missing = 41

Table 2 of Smoking by LowBirthWgt			
Controlling for Death=No			
Smoking	LowBirthWgt		
	Aff	No	Total
Aff	53 27.60	488 22.86	541
No	139 72.40	1647 77.14	1786
Total	192	2135	2327
Frequency Missing = 156			

Note	Statistics for Table 2 of Smoking by LowBirthWgt Controlling for Death=No

Statistic	DF	Value	Prob
Chi-Square	1	2.2246	0.1358
Likelihood Ratio Chi-Square	1	2.1450	0.1430
Continuity Adj. Chi-Square	1	1.9666	0.1608
Mantel-Haenszel Chi-Square	1	2.2237	0.1359
Phi Coefficient		0.0309	
Contingency Coefficient		0.0309	
Cramer's V		0.0309	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	53
Left-sided Pr <= F	0.9410
Right-sided Pr >= F	0.0821
Table Probability (P)	0.0231
Two-sided Pr <= P	0.1532

Note	Sample Size = 2327 Frequency Missing = 156

Test for Association between Low Birth Weight and Smoking

Controlling for Death

The FREQ Procedure

Note	Summary Statistics for Smoking by LowBirthWgt Controlling for Death

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	*Alternative Hypothesis*	*DF*	*Value*	*Prob*
1	*Nonzero Correlation*	1	0.0640	0.8003
2	*Row Mean Scores Differ*	1	0.0640	0.8003
3	*General Association*	1	0.0640	0.8003

Common Odds Ratio and Relative Risks				
Statistic	*Method*	*Value*	*95% Confidence Limits*	
Odds Ratio	Mantel−Haenszel	1.0328	0.8004	1.3327
	Logit	1.0374	0.8066	1.3341
Relative Risk (Column 1)	Mantel−Haenszel	1.0170	0.8909	1.1610
	Logit	0.9675	0.8563	1.0932
Relative Risk (Column 2)	Mantel−Haenszel	0.9954	0.9599	1.0323
	Logit	0.9809	0.9513	1.0114

Breslow-Day Test for Homogeneity of Odds Ratios	
Chi-Square	3.8091
DF	1
Pr > ChiSq	0.0510

Note	Sample Size = 2892 Frequency Missing = 197

Test for Association between Low Birth Weight and Drinking

Controlling for Death

The FREQ Procedure

Table 1 of Drinking by LowBirthWgt			
Controlling for Death=Aff			
Drinking	LowBirthWgt		
	Aff	No	Total
Aff	45 11.75	25 13.74	70
No	338 88.25	157 86.26	495
Total	383	182	565
Frequency Missing = 41			

Note	Statistics for Table 1 of Drinking by LowBirthWgt Controlling for Death=Aff

Statistic	DF	Value	Prob
Chi-Square	1	0.4487	0.5029
Likelihood Ratio Chi-Square	1	0.4419	0.5062
Continuity Adj. Chi-Square	1	0.2843	0.5939
Mantel-Haenszel Chi-Square	1	0.4479	0.5033
Phi Coefficient		−0.0282	
Contingency Coefficient		0.0282	
Cramer's V		−0.0282	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	45
Left-sided Pr <= F	0.2940
Right-sided Pr >= F	0.7912
Table Probability (P)	0.0852
Two-sided Pr <= P	0.4976

Note	Sample Size = 565 Frequency Missing = 41

Table 2 of Drinking by LowBirthWgt			
Controlling for Death=No			
Drinking	LowBirthWgt		
	Aff	No	Total
Aff	29 15.10	300 14.05	329
No	163 84.90	1835 85.95	1998
Total	192	2135	2327
Frequency Missing = 156			

Note	Statistics for Table 2 of Drinking by LowBirthWgt Controlling for Death=No

Statistic	DF	Value	Prob
Chi-Square	1	0.1608	0.6884
Likelihood Ratio Chi-Square	1	0.1581	0.6909
Continuity Adj. Chi-Square	1	0.0858	0.7696
Mantel-Haenszel Chi-Square	1	0.1607	0.6885
Phi Coefficient		0.0083	
Contingency Coefficient		0.0083	
Cramer's V		0.0083	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	29
Left-sided Pr <= F	0.7002
Right-sided Pr >= F	0.3773
Table Probability (P)	0.0775
Two-sided Pr <= P	0.6660

Note	Sample Size = 2327 Frequency Missing = 156

Test for Association between Low Birth Weight and Drinking

Controlling for Death

The FREQ Procedure

Note	Summary Statistics for Drinking by LowBirthWgt Controlling for Death

Cochran-Mantel-Haenszel Statistics (Based on Table Scores)				
Statistic	*Alternative Hypothesis*	*DF*	*Value*	*Prob*
1	*Nonzero Correlation*	1	0.0102	0.9194
2	*Row Mean Scores Differ*	1	0.0102	0.9194
3	*General Association*	1	0.0102	0.9194

Common Odds Ratio and Relative Risks				
Statistic	*Method*	*Value*	*95% Confidence Limits*	
Odds Ratio	Mantel−Haenszel	0.9834	0.7080	1.3659
	Logit	0.9836	0.7110	1.3609
Relative Risk (Column 1)	Mantel−Haenszel	0.9908	0.8261	1.1883
	Logit	0.9668	0.8191	1.1413
Relative Risk (Column 2)	Mantel−Haenszel	1.0021	0.9611	1.0449
	Logit	0.9942	0.9592	1.0305

Breslow-Day Test for Homogeneity of Odds Ratios	
Chi-Square	0.5994
DF	1
Pr > ChiSq	0.4388

Note	Sample Size = 2892 Frequency Missing = 197

Code 4

```
title 'Test for Association between Low Birth Weight and Death';
title2 '';
proc freq data=new_bwgt;* order=freq;
tables LowBirthWgt*death/norow nopercent chisq relrisk riskdiff;
run;

ods latex close;
```

Test for Association between Low Birth Weight and Death

In the following table there is a significant association at the .05 level between Death and Low Birth Weight. This is seen in the chi-square statistic and the relative risk and odds ratio.

The FREQ Procedure

Table of LowBirthWgt by Death			
LowBirthWgt	*Death*		
	Aff	*No*	*Total*
Aff	408 67.33	205 8.26	613
No	198 32.67	2278 91.74	2476
Total	606	2483	3089

Note	Statistics for Table of LowBirthWgt by Death

Statistic	*DF*	*Value*	*Prob*
Chi-Square	1	1068.5596	<.0001
Likelihood Ratio Chi-Square	1	897.1241	<.0001
Continuity Adj. Chi-Square	1	1064.8493	<.0001
Mantel-Haenszel Chi-Square	1	1068.2137	<.0001
Phi Coefficient		0.5882	
Contingency Coefficient		0.5070	
Cramer's V		0.5882	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	408
Left-sided Pr <= F	1.0000
Right-sided Pr >= F	<.0001
Table Probability (P)	<.0001
Two-sided Pr <= P	<.0001

Column 1 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.6656	0.0191	0.6282	0.7029	0.6267	0.7029
Row 2	0.0800	0.0055	0.0693	0.0907	0.0696	0.0914
Total	0.1962	0.0071	0.1822	0.2102	0.1823	0.2106
Difference	0.5856	0.0198	0.5468	0.6245		
Difference is (Row 1 - Row 2)						

Column 2 Risk Estimates						
	Risk	ASE	95% Confidence Limits		Exact 95% Confidence Limits	
Row 1	0.3344	0.0191	0.2971	0.3718	0.2971	0.3733
Row 2	0.9200	0.0055	0.9093	0.9307	0.9086	0.9304
Total	0.8038	0.0071	0.7898	0.8178	0.7894	0.8177
Difference	−0.5856	0.0198	−0.6245	−0.5468		
Difference is (Row 1 - Row 2)						

Note all the confidence intervals do not contain one, indicating a strong association between infant birth weight and survival.

Odds Ratio and Relative Risks			
Statistic	Value	95% Confidence Limits	
Odds Ratio	22.8979	18.3410	28.5869
Relative Risk (Column 1)	8.3231	7.2003	9.6210
Relative Risk (Column 2)	0.3635	0.3249	0.4067

Note	Sample Size = 3089

Loglinear Models for Contingency Tables

I do not normally cover this material until the spring semester. It is covered in greater detail in our graduate course on Categorical Models. Yet, this example is ideally suited for this approach. Agresti covers this material in Chapter 8 and SDK covers the material in chapter 16.

Two-Way Tables

Loglinear Models for the 2×2 Table

Suppose that one has the 2×2 table given by

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

With cell probabilities given by

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

One of the foundational issues in analysis of these tables is to test for independence of the two random variables X and Y. Independence implies that,

$$\frac{\pi_{11}}{\pi_{+1}} = \frac{\pi_{12}}{\pi_{+2}} = \pi_{1+}$$

and

$$\pi_{11} = \pi_{1+}\pi_{+1}.$$

From which it follows that

$$\pi_{ij} = \pi_{i+}\pi_{+j} \quad i, j = 1, 2.$$

If X and Y are independent then it follows that the odds ratio, ψ, is

$$\psi = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}} = 1.$$

Taking the log of both sides leads to,

$$log\ \psi = log\ \pi_{11} + log\ \pi_{22} - log\ \pi_{12} - log\ \pi_{21} = 0.$$

Now consider the log transformation of the expected counts for the ij^{th} cell given by m_{ij}. For which one has,

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

for $i, j = 1, 2$ where $m_{ij} = n\pi_{ij}$. This equation is called the **saturated loglinear model** for the 2×2 table.

Since there are $1 + 2 + 2 + 4 = 9$ parameters in this model and only four observations (cell frequencies) it is necessary to define the following constraints on the model,

$$\sum_i \lambda_i^X = 0 \quad \sum_j \lambda_j^Y = 0 \quad \sum_i \lambda_{ij}^{XY} = \sum_j \lambda_{ij}^{XY} = 0.$$

The loglinear model expected cell counts can be written as,

X	Y=1	Y=2
1	$exp(\mu + \lambda_1^X + \lambda_1^Y + \lambda_{11}^{XY})$	$exp(\mu + \lambda_1^X - \lambda_1^Y - \lambda_{11}^{XY})$
2	$exp(\mu - \lambda_1^X + \lambda_1^Y - \lambda_{11}^{XY})$	$exp(\mu - \lambda_1^X - \lambda_1^Y + \lambda_{11}^{XY})$

Using the above loglinear model, the odds ratio can be written as

$$\psi = \frac{m_{11}m_{22}}{m_{12}m_{21}}$$

so that

$$log\ \psi = log\ m_{11} + log\ m_{22} - log\ m_{12} - log\ m_{21} = 4\lambda_{11}^{XY}.$$

In which case, the hypothesis of independence of X and Y is equivalent to $H_0 : \lambda_{11}^{XY} = 0$. Thus, the loglinear model when the assumption of independence holds becomes,

$$log\ (m_{ij}) = \mu + \lambda_i^X + \lambda_j^Y \quad i, j = 1, 2.$$

$R \times C$ Tables

Agresti considers the model given by,

$$\mu_{ij} = \mu\alpha_i\beta_j.$$

From which one has the *saturated model*

$$log\ \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_{ij}^{XY} \tag{1}$$

and the *independent model*,

$$log\ \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y \tag{2}$$

for $i = 1, 2, \ldots, R, j = 1, 2, \ldots, C$.

Three-Way Tables

Types of Independence

- The **saturated model** has loglinear form,

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

- **Mutual Independence** X, Y, and Z are mutually independent when

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

  for all i, j and k. Mutual independence implies that the expected frequencies $\{\mu_{ijk}\}$ have the loglinear form given by,

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

  in which case

$$\lambda_{ij}^{XY} = \lambda_{ik}^{XZ} = \lambda_{jk}^{YZ} = \lambda_{ijk}^{XYZ} = 0.$$

- **Joint Independence** Variable Y is jointly independent of X and Z when

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

  for all i, j and k. The loglinear form can be written as,

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

  in which case

$$\lambda_{ij}^{XY} = \lambda_{jk}^{YZ} = \lambda_{ijk}^{XYZ} = 0.$$

  Mututal independence implies joint independence of any one variable from the others.

- **Conditional Independence** X and Y are conditionally independent, given Z when

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

for all i, j and k. This holds for joint probabilities over the entire table, hence,

$$\pi_{ijk} = \pi_{i+k}\pi_{+jk}/\pi_{++k}.$$

The loglinear form can be written as,

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

in which case

$$\lambda_{ij}^{XY} = \lambda_{ijk}^{XYZ} = 0.$$

SAS

Code 5

```
title 'Three Way Model';
title2 'Death, Low Birth Weight and Smoking';
proc catmod data=new_bwgt;
   model death*smoking*lowbirthwgt=_response_
         / noparm pred=freq;
   loglin death|smoking|lowbirthwgt @ 2;
run;

title2 'Final Model';
proc catmod data=new_bwgt;
   model death*smoking*lowbirthwgt=_response_
         / noparm pred=freq;
   loglin death|lowbirthwgt death|smoking;
run;
```

Three Way Model

Death, Low Birth Weight and Smoking

The CATMOD Procedure

Data Summary			
Response	Death*Smoking*LowBirthWg	*Response Levels*	8
Weight Variable	None	*Populations*	1
Data Set	NEW_BWGT	*Total Frequency*	2892
Frequency Missing	197	*Observations*	2892

Population Profiles	
Sample	*Sample Size*
1	2892

Response Profiles			
Response	Death	Smoking	LowBirthWgt
1	Aff	Aff	Aff
2	Aff	Aff	No
3	Aff	No	Aff
4	Aff	No	No
5	No	Aff	Aff
6	No	Aff	No
7	No	No	Aff
8	No	No	No

Maximum Likelihood Analysis
Maximum likelihood computations converged.

Maximum Likelihood Analysis of Variance			
Source	DF	Chi-Square	Pr > ChiSq
Death	1	154.35	<.0001
Smoking	1	369.69	<.0001
Death*Smoking	1	3.46	0.0627
LowBirthWgt	1	152.09	<.0001
Death*LowBirthWgt	1	718.75	<.0001
Smoking*LowBirthWgt	1	0.06	0.8002
Likelihood Ratio	1	3.73	0.0535

Maximum Likelihood Predicted Values for Response Functions					
Function Number	Observed		Predicted		Residual
	Function	Standard Error	Function	Standard Error	
1	−2.78174	0.102035	−2.70818	0.091798	−0.07355
2	−3.34627	0.133598	−3.47632	0.123381	0.130054
3	−1.76836	0.064544	−1.79023	0.064295	0.021877
4	−2.58643	0.093122	−2.52479	0.08532	−0.06164
5	−3.43642	0.139553	−3.58015	0.127129	0.143728
6	−1.2164	0.05154	−1.19711	0.050321	−0.01929
7	−2.47224	0.088326	−2.41662	0.081669	−0.05562

Maximum Likelihood Predicted Values for Frequencies							
Death	*Smoking*	*LowBirthWgt*	*Observed*		*Predicted*		*Residual*
			Frequency	*Standard Error*	*Frequency*	*Standard Error*	
Aff	*Aff*	*Aff*	102	9.919803	109.2989	9.548808	−7.2989
Aff	*Aff*	*No*	58	7.539018	50.7011	5.985303	7.298899
Aff	*No*	*Aff*	281	15.92786	273.7011	15.29083	7.298899
Aff	*No*	*No*	124	10.89418	131.2989	10.55222	−7.2989
No	*Aff*	*Aff*	53	7.213092	45.7011	5.566915	7.298899
No	*Aff*	*No*	488	20.14086	495.2989	19.91188	−7.2989
No	*No*	*Aff*	139	11.50301	146.2989	11.1763	−7.2989
No	*No*	*No*	1647	26.62762	1639.701	26.38253	7.298899

Three Way Model

Final Model

The CATMOD Procedure

Data Summary			
Response	Death*Smoking*LowBirthWg	*Response Levels*	8
Weight Variable	None	*Populations*	1
Data Set	NEW_BWGT	*Total Frequency*	2892
Frequency Missing	197	*Observations*	2892

Population Profiles	
Sample	*Sample Size*
1	2892

Response Profiles			
Response	*Death*	*Smoking*	*LowBirthWgt*
1	Aff	Aff	Aff
2	Aff	Aff	No
3	Aff	No	Aff
4	Aff	No	No
5	No	Aff	Aff
6	No	Aff	No
7	No	No	Aff
8	No	No	No

Maximum Likelihood Analysis
Maximum likelihood computations converged.

Maximum Likelihood Analysis of Variance			
Source	*DF*	*Chi-Square*	*Pr > ChiSq*
Death	1	167.50	<.0001
LowBirthWgt	1	201.07	<.0001
*Death*LowBirthWgt*	1	721.21	<.0001
Smoking	1	405.06	<.0001
*Death*Smoking*	1	6.34	0.0118
Likelihood Ratio	2	3.79	0.1502

All the above terms are significant, including the association between smoking and death!

Maximum Likelihood Predicted Values for Response Functions					
Function Number	*Observed*		*Predicted*		*Residual*
	Function	*Standard Error*	*Function*	*Standard Error*	
1	−2.78174	0.102035	−2.71524	0.08769	−0.0665
2	−3.34627	0.133598	−3.45927	0.102826	0.112998
3	−1.76836	0.064544	−1.78652	0.06252	0.018168
4	−2.58643	0.093122	−2.53055	0.082415	−0.05588
5	−3.43642	0.139553	−3.60304	0.089917	0.166622
6	−1.2164	0.05154	−1.19431	0.049075	−0.02208
7	−2.47224	0.088326	−2.40873	0.075344	−0.06351

Maximum Likelihood Predicted Values for Frequencies							
Death	*Smoking*	*LowBirthWgt*	*Observed*		*Predicted*		*Residual*
			Frequency	*Standard Error*	*Frequency*	*Standard Error*	
Aff	*Aff*	*Aff*	102	9.919803	108.4602	8.907781	−6.46018
Aff	*Aff*	*No*	58	7.539018	51.53982	5.057426	6.460177
Aff	*No*	*Aff*	281	15.92786	274.5398	14.94761	6.460177
Aff	*No*	*No*	124	10.89418	130.4602	9.976526	−6.46018
No	*Aff*	*Aff*	53	7.213092	44.63773	3.537735	8.362269
No	*Aff*	*No*	488	20.14086	496.3623	19.48686	−8.36227
No	*No*	*Aff*	139	11.50301	147.3623	10.41251	−8.36227
No	*No*	*No*	1647	26.62762	1638.638	26.05253	8.362269

Code 6

```
title 'Three Way Model';
title2 'Death, Low Birth Weight and Smoking';
proc logistic data=new_bwgt plots=(oddsratio effect) ;*where race ne 'Native';
class AgeGroup Death Drinking LowBirthWgt Married Smoking somecollege;
model death(event='Aff') =
             Drinking LowBirthWgt Smoking /expb;
run;
```

Logistic Models for Binary Data

Let Y be a binary response variable where $\Pr[Y = 1 \mid \mathbf{x}] = \pi(\mathbf{x})$ and $\Pr[Y = 0 \mid \mathbf{x}] = 1 - \pi(\mathbf{x})$ with covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)$. There are several potential approaches to this modeling problem. One could use the ordinary least squares approach[1], called the **Linear Probability model**, given as,

$$\pi(\mathbf{x}) = \alpha + \beta'\mathbf{x}.$$

This model has a structural defect since $\pi(x)$ is not restricted to the interval $[0, 1]$ for all x. A better model is the **Logistic Regression Model**[2] given as,

$$y = log\left[\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right] = (\alpha + \beta'\mathbf{x}),$$

where y is the log odds and $\pi(\mathbf{x})$ is the probability of the event of interest for the covariate $\mathbf{x}$. It follows that,

$$\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})} = exp(\alpha + \beta'\mathbf{x}),$$

and

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

Logistic Regression with Categorical Predictors

SDK considered using the Logistic model with categorical predictors. The SAS file is

```
title 'Agresti Coronary Example 1';

data coronary;
  input sex ecg ca count @@;
datalines;
0 0 0  11 0 0 1  4
0 1 0  10 0 1 1  8
1 0 0   9 1 0 1  9
1 1 0   6 1 1 1 21
;
run;
title2 'LOGIT Link';
proc logistic data=coronary desc plots=(oddsratio);
class sex ecg;
      freq count;
      model ca=sex ecg / expb scale=none aggregate;
   output out=predict pred=prob;
run;
```

[1]GLM with normal data and the identity link function.
[2]GLM with binary data and the canonical logit link.

```
   data predict; set predict; prob = 1 - prob; drop _level_;

proc print data=predict;
run;
```

Coronary Example Output

The model for the four discrete events is given by

Sex	ECG	$\Pr[CA\ Disease] = \theta_{hi}$	Odds of CA Disease
Females	< 0.1	$e^{\alpha}/(1+e^{\alpha})$	e^{α}
Females	≥ 0.1	$e^{\alpha+\beta_2}/(1+e^{\alpha+\beta_2})$	$e^{\alpha+\beta_2}$
Males	< 0.1	$e^{\alpha+\beta_1}/(1+e^{\alpha+\beta_1})$	$e^{\alpha+\beta_1}$
Males	≥ 0.1	$e^{\alpha+\beta_1+\beta_2}/(1+e^{\alpha+\beta_1+\beta_2})$	$e^{\alpha+\beta_1+\beta_2}$

Parameter	Estimate	SE	Interpretation
α	-1.17	0.485	log odds of coronary disease
			for females with ECG < 0.1
β_1	1.28	0.498	increment to log odds for males
β_2	1.05	0.498	increment to log odds for high ECG

Sex	ECG	Logit	Odds of CA Disease
Female	< 0.1	$\hat{\alpha} = -1.17$	$e^{\hat{\alpha}} = e^{-1.17} = 0.3089$
Female	≥ 0.1	$\hat{\alpha} + \hat{\beta}_2 = -0.12$	$e^{\hat{\alpha}+\hat{\beta}_2} = e^{-0.12} = 0.8867$
Male	< 0.1	$\hat{\alpha} + \hat{\beta}_1 = 0.10$	$e^{\hat{\alpha}+\hat{\beta}_1} = e^{0.10} = 1.11$
Male	≥ 0.1	$\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 = 1.157$	$e^{\hat{\alpha}+\hat{\beta}_1+\hat{\beta}_2} = e^{1.157} = 3.18$

Logistic Model - Low Birth weight

The LOGISTIC Procedure

Model Information	
Data Set	WORK.NEW_BWGT
Response Variable	Death
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	4589
Number of Observations Used	4242

Response Profile		
Ordered Value	*Death*	*Total Frequency*
1	Aff	545
2	No	3697

Note	Probability modeled is Death='Aff'.

Note	347 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information		
Class	*Value*	*Design Variables*
Drinking	*Aff*	1
	No	−1
LowBirthWgt	*Aff*	1
	No	−1
Smoking	*Aff*	1
	No	−1

Model Convergence Status
Convergence criterion (GCONV=1E−8) satisfied.

Model Fit Statistics		
Criterion	*Intercept Only*	*Intercept and Covariates*
AIC	3255.457	2299.268
SC	3261.810	2324.679

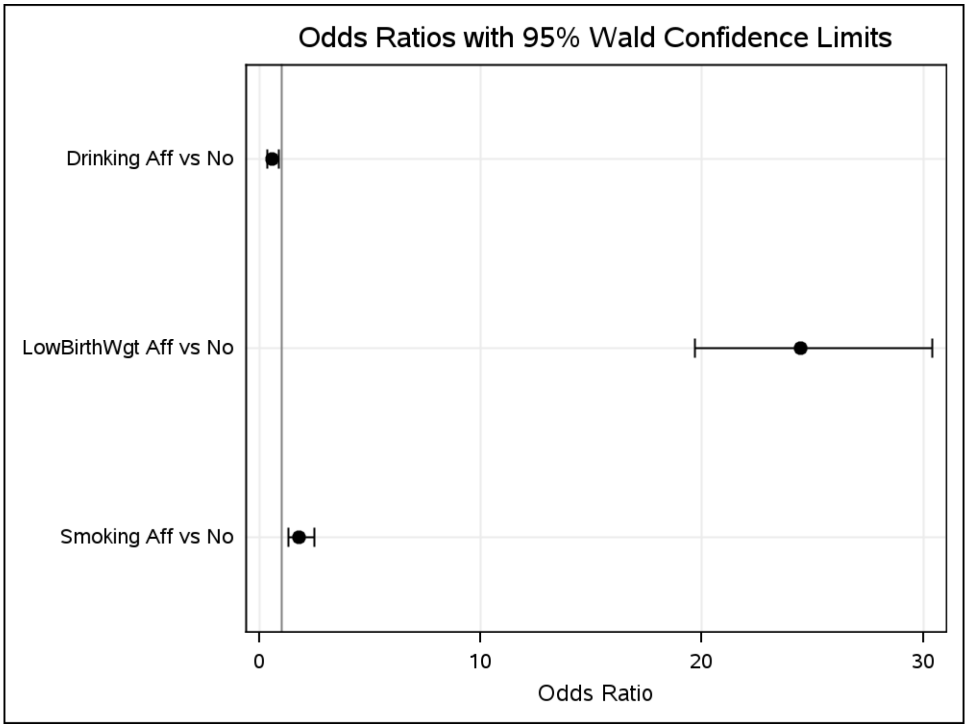
Model Fit Statistics		
Criterion	*Intercept Only*	*Intercept and Covariates*
-2 Log L	3253.457	2291.268

Testing Global Null Hypothesis: BETA=0			
Test	*Chi-Square*	*DF*	*Pr > ChiSq*
Likelihood Ratio	962.1896	3	<.0001
Score	1320.6291	3	<.0001
Wald	848.1561	3	<.0001

Type 3 Analysis of Effects			
Effect	*DF*	*Wald Chi-Square*	*Pr > ChiSq*
Drinking	1	6.7534	0.0094
LowBirthWgt	1	831.3778	<.0001
Smoking	1	12.7496	0.0004

Analysis of Maximum Likelihood Estimates							
Parameter		*DF*	*Estimate*	*Standard Error*	*Wald Chi-Square*	*Pr > ChiSq*	*Exp(Est)*
Intercept		1	−1.4015	0.0839	278.7291	<.0001	0.246
Drinking	*Aff*	1	−0.2814	0.1083	6.7534	0.0094	0.755
LowBirthWgt	*Aff*	1	1.5986	0.0554	831.3778	<.0001	4.946
Smoking	*Aff*	1	0.2966	0.0831	12.7496	0.0004	1.345

Odds Ratio Estimates			
Effect	*Point Estimate*	*95% Wald Confidence Limits*	
Drinking Aff vs No	0.570	0.373	0.871
LowBirthWgt Aff vs No	24.462	19.684	30.400
Smoking Aff vs No	1.810	1.307	2.506



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	71.8	*Somers' D*	0.634
Percent Discordant	8.5	*Gamma*	0.789
Percent Tied	19.7	*Tau-a*	0.142
Pairs	2014865	*c*	0.817