Chapter 8 Logistic Regression I: Dichotomous Response

8.1 Introduction

- This chapter focuses on statistical models. Statistical modeling allows you to address questions about association in terms of hypotheses concerning model parameters.

- If sampling assumptions are plausible, statistical models can be used to make inferences from study population to larger target population.

- Logistic regression is a form of statistical modeling often appropriate for categorical outcome variables: describes relationship between categorical response and set of explanatory variables.

- Response variable can be dichotomous or polytomous:

  Chapter 8: Dichotomous response.
  Chapter 9: Polytomous response (nominal or ordered).

- Explanatory variables can be categorical or continuous.

- Logistic regression has applications in fields such as medical research, epidemiology, social research, banking, and market research. One advantage is model interpretation through odds ratios.

- Procedures used to perform logistic regression:

LOGISTIC: designed for logistic regression, provides odds ratio estimates and model diagnostics

GENMOD: procedure for analyzing generalized linear models, of which logistic regression is a simple case.

Logistic Regression for relationship between dichotomous response and one quantitative predictor:

$$y = \begin{cases} 1 \text{ if target attribute} \\ 0 \text{ if else} \end{cases}$$

x = quantitative predictor

$$\Pr\{y = 1\} = \Pi(x) = \exp(\alpha + \beta x) / (1 + \exp(\alpha + \beta x))$$

Population for predictor	$\Pr\{y=1\}$	$\Pr\{y=0\}$	$\frac{\Pr\{y=1\}}{\Pr\{y=0\}}$
$(x+1)$	$\frac{\exp(\alpha+\beta(x+1))}{1+\exp(\alpha+\beta(x+1))}$	$\frac{1}{\{1+\exp(\alpha+\beta(x+1))\}}$	$\exp(\alpha+\beta(x+1))$
x	$\frac{\exp(\alpha+\beta x)}{\{1+\exp(\alpha+\beta x)\}}$	$\frac{1}{\{1+\exp(\alpha+\beta x)\}}$	$\exp(\alpha+\beta x)$
$(x+1)=1$	$\frac{\exp(\alpha+\beta)}{\{1+\exp(\alpha+\beta)\}}$	$\frac{1}{\{1+\exp(\alpha+\beta)\}}$	$\exp(\alpha+\beta)$
$x=0$	$\frac{\exp(\alpha)}{\{1+\exp(\alpha)\}}$	$\frac{1}{\{1+\exp(\alpha)\}}$	$\exp(\alpha)$
$\left[\frac{\Pr\{y=1\mid(x+1)\}}{\Pr\{y=0\mid(x+1)\}}\right]\Big/\left[\frac{\Pr\{y=1\mid x\}}{\Pr\{y=0\mid x\}}\right]$			$\exp(\beta)$

8.2 Dichotomous Explanatory Variables

8.2.1 Logistic Model

- Following table displays coronary artery disease data from Chapter 3, where Mantel-Haenszel methods were used to analyze the data. There, we found ECG to be clearly associated with disease status, adjusted for gender.

Gender	ECG	Disease	No Disease	Total
Female	< 0.1 ST segment depression	4	11	15
Female	≥ 0.1 ST segment depression	8	10	18
Male	< 0.1 ST segment depression	9	9	18
Male	≥ 0.1 ST segment depression	21	6	27

- Assume the data arise from a stratified simple random sample so that presence of coronary artery disease is distributed binomially for each GENDER × ECG combination:

$$\Pr\{n_{hij}\} = \prod_{h=1}^{2}\prod_{i=1}^{2} \frac{n_{hi+}!}{n_{hi1}!n_{hi2}!}\theta_{hi}^{n_{hi1}}(1-\theta_{hi})^{n_{hi2}},$$

where θ_{hi} is probability that person of *h*th gender with *i*th ECG status has coronary artery disease, and n_{hi1} and n_{hi2} are numbers of persons of *h*th gender and *i*th ECG with and without coronary artery disease, respectively. A logistic model can then be applied to describe variation among the $\{\theta_{hi}\}$

- θ_{hi} can be expressed in the following forms:

$$\theta_{hi} = \frac{1}{1+\exp\left\{-\left(\alpha + \sum_{k=1}^{t} \beta_k x_{hik}\right)\right\}}$$

$$= \frac{\exp\left\{\alpha + \sum_{k=1}^{t} \beta_k x_{hik}\right\}}{1+\exp\left\{\alpha + \sum_{k=1}^{t} \beta_k x_{hik}\right\}},$$

where α is intercept parameter, $\{x_{hik}\}$ are t explanatory variables for hth gender and ith ECG; $k = 1,\ldots,t$; and $\{\beta_k\}$ are t regression parameters

- Matrix form of equation:

$$\theta_{hi} = \frac{\exp(\boldsymbol{x}'_{hi}\boldsymbol{\beta})}{1+\exp(\boldsymbol{x}'_{hi}\boldsymbol{\beta})}$$

- Odds of CA disease for *hi*th group is expressed as:

$$\frac{\theta_{hi}}{1-\theta_{hi}} = \exp\left\{\alpha + \sum_{k=1}^{t} \beta_k x_{hik}\right\}$$

- Taking log of both sides produces linear model for *logit*:

$$\log\left\{\frac{\theta_{hi}}{1-\theta_{hi}}\right\} = \alpha + \sum_{k=1}^{t} \beta_k x_{hik}$$

- Model is for the log odds of coronary artery disease vs. no coronary artery disease for the *hi*th group
- Model-predicted odds ratios are obtained by exponentiating the model parameter estimates
- Maximum likelihood methods are used to estimate α and β. LOGISTIC uses the Fisher scoring method (equivalent to iteratively weighted least squares), while GENMOD uses Newton-Raphson algorithms
- Note that estimated coefficients are approximately normal, and that estimated standard errors are provided

- Marginal tables for each main effect (all counts ≥ 5)

Gender

	Disease	No Disease	Total
Female	12	21	33
Male	30	15	45
Total	42	36	78

ECG

	Disease	No Disease	Total
ECG < 0.1	13	20	33
ECG $\geq$ 0.1	29	16	45
Total	42	36	78

8.2.2 Model Fitting

- First consider model for coronary disease data with main effects for gender and ECG:

$$\begin{bmatrix} \text{logit}(\theta_{11}) \\ \text{logit}(\theta_{12}) \\ \text{logit}(\theta_{21}) \\ \text{logit}(\theta_{22}) \end{bmatrix} = \begin{bmatrix} \alpha & & \\ \alpha & & +\beta_2 \\ \alpha & +\beta_1 & \\ \alpha & +\beta_1 & +\beta_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

where α = log odds of coronary artery disease for females with ECG < 0.1

β_1 = increment in log odds for males

β_2 = increment in log odds for ECG $\geq$ 0.1

- Formulas for cell probabilities and odds predicted by model:

Gender	ECG	Pr {CA Disease} = θ_{hi}	Odds of CA Disease
Female	< 0.1	$e^{\alpha}/(1+e^{\alpha})$	e^{α}
Female	≥ 0.1	$e^{\alpha+\beta_2}/\left(1+e^{\alpha+\beta_2}\right)$	$e^{\alpha+\beta_2}$
Male	< 0.1	$e^{\alpha+\beta_1}/\left(1+e^{\alpha+\beta_1}\right)$	$e^{\alpha+\beta_1}$
Male	≥ 0.1	$e^{\alpha+\beta_1+\beta_2}/\left(1+e^{\alpha+\beta_1+\beta_2}\right)$	$e^{\alpha+\beta_1+\beta_2}$

- Odds ratio for males vs. females for either low or high ECG is:

$$\frac{e^{\alpha+\beta_1}}{e^{a}} = e^{\beta_1} \text{ or } \frac{e^{\alpha+\beta_1+\beta_2}}{e^{\alpha+\beta_2}} = e^{\beta_1}$$

- Similarly, odds ratio for high ECG vs. low ECG is determined by forming ratio of odds of CA disease for either gender:

$$\frac{e^{\alpha+\beta_1+\beta_2}}{e^{a+\beta_1}} = e^{\beta_2} \text{ or } \frac{e^{\alpha+\beta_2}}{e^{\alpha}} = e^{\beta_2}$$

- Unlike odds ratios calculated from individual 2 × 2 tables, these odds ratios have been adjusted for all other explanatory variables in model

Logistic Regression for relationship between dichotomous response and two quantitative predictors

$$\Pr\{y=1|x_1,x_2\} = \Pi(x_1,x_2) = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2)}{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2)}$$

$$\frac{\Pr\{y=1|x_1,x_2\}}{\Pr\{y=0|x_1,x_2\}} = \phi(x_1,x_2) = \exp(\alpha+\beta_1 x_1+\beta_2 x_2)$$

$$\frac{\phi(x_1+1,x_2)}{\phi(x_1,x_2)} = \frac{\exp(\alpha+\beta_1(x_1+1)+\beta_2 x_2)}{\exp(\alpha+\beta_1 x_1+\beta_2 x_2)} = \exp(\beta_1)$$

$$\frac{\phi(x_1,x_2+1)}{\phi(x_1,x_2)} = \frac{\exp(\alpha+\beta_1 x_1+\beta_2(x_2+1))}{\exp(\alpha+\beta_1 x_1+\beta_2 x_2)} = \exp(\beta_2)$$

$$\frac{\phi(x_1+1,x_2+1)}{\phi(x_1,x_2)} = \exp(\beta_1+\beta_2)$$

$$\frac{\phi(x_1+1,x_2+1)}{\phi(x_1,x_2+1)} = \exp(\beta_1)$$

$$\frac{\phi(x_1+1,x_2+1)}{\phi(x_1+1,x_2)} = \exp(\beta_2)$$

Note that dichotomous predictors have $x_1 = x_2 = 0$ here as reference population

8.2.3 Goodness of Fit:

- Need to assess how close model-predicted values are to corresponding observed values
- Test statistics to assess fit of model in this manner are known as *goodness of fit (GOF) statistics*
- GOF statistics have approximate chi-square distributions. If they are larger than a tolerable value, then model is oversimplified and we need to identify other factors
- Two traditional goodness-of-fit tests are Pearson chi-square, Q_P, and likelihood ratio chi-square, Q_L, also known as *deviance*:

$$Q_P = \sum_{h=1}^{2}\sum_{i=1}^{2}\sum_{j=1}^{2}\left(n_{hij} - m_{hij}\right)^2 / m_{hij}$$

$$Q_L = \sum_{h=1}^{2}\sum_{i=1}^{2}\sum_{j=1}^{2} 2n_{hij} \log\left(\frac{n_{hij}}{m_{hij}}\right),$$

where m_{hij} are model-predicted counts defined as:

$$m_{hij} = \begin{cases} n_{hi+}\hat{\theta}_{hi} & \text{for } j = 1 \\ n_{hi+}(1-\hat{\theta}_{hi}) & \text{for } j = 2 \end{cases}$$

- If model fits, both Q_P and Q_L are distributed as χ^2 with d.f. equal to number of rows in table minus number of parameters
- Sample size guidelines for these statistics to be approximately chi-square include:
    1. At least 10 subjects in each group ($n_{hi+} \geq 10$)
    2. 80% of predicted counts (m_{hij}) are ≥ 5
    3. All other expected counts > 2, with no 0 counts

8.2.4 Using PROC LOGISTIC

- Specify response variable and explanatory variables in the MODEL statement
- LOGISTIC fits model via maximum likelihood estimation. Parameter estimates, standard errors, and statistics to assess model fit are produced
- Provides several model selection methods, puts predicted values and other statistics into output data sets, includes a number of options for controlling model-fitting process
- Example of analysis data set:

```
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;
```

- CA is response variable: 1 if CA disease is present,
  0 otherwise

  Response variable is ordered alphanumerically, which means that logistic models Pr {CA = 0}. Because modeling Pr {CA = 1} is most likely of interest, can alter default by using EVENT='1' or DESCENDING option. Difference in models is that sign of parameter estimates is changed
- SEX (0 for females, 1 for males) and ECG (0 for lower ST segment depression, 1 for higher) are explanatory variables, and provide values for model matrix
- When data are in count form, use FREQ statement
- SCALE = NONE and AGGREGATE are used to request goodness of fit statistics

```
proc logistic data=coronary;
freq count;
        model ca(event='1') = sex ecg / scale = none aggregate;

run;
```

Response Profile

```
                    Response Profile

                Ordered
                  Value       CA      Count

                      1        0         36
                      2        1         42
                Probability modeled is ca=1.
```

Goodness of Fit Statistics

```
        Deviance and Pearson Goodness-of-Fit Statistics
                                                          Pr >
   Criterion        DF        Value     Value/DF    Chi-Square

   Deviance          1       0.2141       0.2141        0.6436
   Pearson           1       0.2155       0.2155        0.6425
              Number of unique profiles: 4
```

Testing Joint Significance of the Explanatory Variables

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	109.669	101.900
SC	112.026	108.970
-2 Log L	107.669	95.900

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	11.7694	2	0.0028
Score	11.2410	2	0.0036
Wald	10.0644	2	0.0065

Main Effects Model: ANOVA Table

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.1747	0.4854	5.8571	0.0155
sex	1	1.2770	0.4980	6.5750	0.0103
ecg	1	1.0545	0.4980	4.4844	0.0342

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex	3.586	1.351	9.516
ecg	2.871	1.082	7.618

8.2.5 Interpretation of Main Effects Model

- In a strict sense, results apply only to population consisting of those persons who visited this medical clinic and required catheterization
- Wald statistics take form of squared ratio of estimate to its standard error. They are easy to compute and based on normal theory. Statistical properties of likelihood ratio statistics are more optimal for small samples
- Model equation can be written as:

$$\text{logit}(\theta_{hi}) = -1.1747 + 1.2770\ \text{SEX} + 1.0545\ \text{ECG}$$

- Interpretation of parameters:

Parameter	Estimate (SE)	Interpretation
α	-1.1747 (0.485)	Log odds of coronary disease for females with low ECG
β_1	1.2770 (0.498)	Increment to log odds for males
β_2	1.0545 (0.498)	Increment to log odds for high ECG

- Odds ratio of CA disease for males vs. females:

$$e^{\hat{\beta}_1} = e^{1.2770} = 3.586$$

- Odds ratio of CA disease for high ECG vs. low ECG:

$$e^{\hat{\beta}_2} = e^{1.0545} = 2.871$$

- Predicted values can be produced with the following additional code:

```
proc logistic data=coronary descending;
     freq count;
     model ca = sex ecg;
     output out=predict pred=prob;
run;
proc print data=predict;
run;
```

Predicted Values Output Data Set

Obs	sex	ecg	ca	count	_LEVEL_	prob
1	0	0	0	11	1	0.23601
2	0	0	1	4	1	0.23601
3	0	1	0	10	1	0.46999
4	0	1	1	8	1	0.46999
5	1	0	0	9	1	0.52555
6	1	0	1	9	1	0.52555
7	1	1	0	6	1	0.76075
8	1	1	1	21	1	0.76075

- Model-predicted logits and odds of CA disease

Sex	ECG	Logit	Odds of CA disease
Female	< 0.1	$\hat{\alpha} = -1.1747$	$e^{\hat{\alpha}} = e^{-1.1747} = 0.3089$
Female	≥ 0.1	$\hat{\alpha} + \hat{\beta}_2 = -0.1202$	$e^{\hat{\alpha}+\hat{\beta}_2} = e^{-0.1202} = 0.8867$
Male	< 0.1	$\hat{\alpha} + \hat{\beta}_1 = 0.1023$	$e^{\hat{\alpha}+\hat{\beta}_1} = e^{0.1023} = 1.1077$
Male	≥ 0.1	$\hat{\alpha} + \hat{\beta}_1 + \hat{\beta}_2 = 1.1568$	$e^{\hat{\alpha}+\hat{\beta}_1+\hat{\beta}_2} = e^{1.1568} = 3.1797$

- Model-predicted probabilities of CA disease:

Sex	ECG	Probability of CA disease
Female	< 0.1	$\frac{e^{\hat{\alpha}}}{1+e^{\hat{\alpha}}} = \frac{e^{-1.1747}}{1+e^{-1.1747}} = 0.236$
Female	≥ 0.1	$\frac{e^{\hat{\alpha}+\hat{\beta}_2}}{1+e^{\hat{\alpha}+\hat{\beta}_2}} = \frac{e^{-0.1202}}{1+e^{-0.1202}} = 0.470$
Male	< 0.1	$\frac{e^{\hat{\alpha}+\hat{\beta}_1}}{1+e^{\hat{\alpha}+\hat{\beta}_1}} = \frac{e^{0.1023}}{1+e^{0.1023}} = 0.526$
Male	≥ 0.1	$\frac{e^{\hat{\alpha}+\hat{\beta}_1+\hat{\beta}_2}}{1+e^{\hat{\alpha}+\hat{\beta}_1+\hat{\beta}_2}} = \frac{e^{1.1568}}{1+e^{1.1568}} = 0.761$

- To use the CLASS statement to replicate this analysis:

```
proc logistic data=coronary descending;
    freq count;
    class sex(ref=0) ecg(ref=0) / param=ref;
    model ca = sex ecg;
run;
```

Class Level Information

Class	Value	Design Variables 1
sex	0	0
	1	1
ecg	0	0
	1	1

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-1.1747	0.4854	5.8571	0.0155
sex	1	1	1.2770	0.4980	6.5750	0.0103
ecg	1	1	1.0545	0.4980	4.4844	0.0342

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
sex 1 vs 0	3.586	1.351	9.516
ecg 1 vs 0	2.871	1.082	7.618

8.2.6 Alternative Methods of Assessing Goodness of Fit

- Other strategies are based on fitting an appropriate expanded model, and testing whether contribution of additional terms is nonsignificant. If so, then original model has adequate fit
- Compute likelihood ratio tests for significance of additional terms by taking difference in –2 LOG L's of reduced and expanded models. Difference is $\approx$ chi-square with df equal to the difference in the number of parameters in the two models
- Can also examine Wald statistic for additional parameters in order to assess goodness of fit
- Expanded model contains main effects for sex and ECG, and their interaction. Likelihood ratio statistic tests significance of interaction term and serves as goodness-of-fit test

- Saturated model can be written as:

$$\begin{bmatrix} \text{logit}(\theta_{11}) \\ \text{logit}(\theta_{12}) \\ \text{logit}(\theta_{21}) \\ \text{logit}(\theta_{22}) \end{bmatrix} = \begin{bmatrix} \alpha \\ \alpha + \beta_2 \\ \alpha + \beta_1 \\ \alpha + \beta_1 + \beta_2 + \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

- An interaction between `sex` and `ecg` can be added by specifying `sex*ecg`:

```
ods select FitStatistics ParameterEstimates;
proc logistic data=coronary descending;
    freq count;
    class sex(ref=0) ecg(ref=0) / param=ref;
    model ca = sex  ecg  sex*ecg;
run;
```

- –2 LOG L for saturated model: 95.686

  –2 LOG L for main effects model: 95.900

  Difference is 0.214 with 1 df, adequacy of model is supported

- Can always compute likelihood ratio test in this manner for contribution of a particular model term or set of model terms

- Note likelihood ratio test value is same as deviance reported for main effects model

- Note value of Wald statistic is 0.215 for interaction

Results for Saturated Model

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	109.669	103.686
SC	112.026	113.112
-2 Log L	107.669	95.686

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.0116	0.5839	3.0018	0.0832
sex 1	1	1.0116	0.7504	1.8172	0.1776
ecg 1	1	0.7885	0.7523	1.0985	0.2946
sex*ecg 1 1	1	0.4643	1.0012	0.2151	0.6428

For dichotomous predictors with $x_1 = x_2 = 0$ as the reference population

Population x_1	x_2	$\Pr\{y=1 \mid x_1, x_2\}$	$\phi(x_1, x_2)$
0	0	$e^{\alpha}/(1+e^{\alpha})$	e^{α}
0	1	$e^{\alpha+\beta_2}/(1+e^{\alpha+\beta_2})$	$e^{\alpha+\beta_2}$
1	0	$e^{\alpha+\beta_1}/(1+e^{\alpha+\beta_1})$	$e^{\alpha+\beta_1}$
1	1	$e^{\alpha+\beta_1+\beta_2+\beta_3}/(1+e^{\alpha+\beta_1+\beta_2+\beta_3})$	$e^{\alpha+\beta_1+\beta_2+\beta_3}$

$$\phi(1,0)/\phi(0,0) = e^{\beta_1}$$

$$\phi(0,1)/\phi(0,0) = e^{\beta_2}$$

$$\phi(1,1)/\phi(0,0) = e^{\beta_1+\beta_2+\beta_3}$$

$$\phi(1,1)/\phi(0,1) = e^{\beta_1+\beta_3}$$

$$\phi(1,1)/\phi(1,0) = e^{\beta_2+\beta_3}$$

$$\left\{\frac{\phi(1,1)}{\phi(0,1)} \Big/ \frac{\phi(1,0)}{\phi(0,0)}\right\} = e^{\beta_3} = \left\{\frac{\phi(1,1)}{\phi(1,0)} \Big/ \frac{\phi(0,1)}{\phi(0,0)}\right\}$$

Logistic Regression for relationship between dichotomous response and two quantitative predictors and their interaction

$$\Pr\{y=1 \mid x_1, x_2\} = \prod(x_1, x_2) = \frac{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2)}{\{1+\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2)\}}$$

$$\frac{\Pr\{y=1|x_1,x_2\}}{\Pr\{y=0|x_1,x_2\}} = \phi(x_1, x_2) = \exp(\alpha + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2)$$

$$\frac{\phi(x_1+1,x_2)}{\phi(x_1,x_2)} = \frac{\exp(\alpha+\beta_1(x_1+1)+\beta_2 x_2+\beta_3(x_1+1)x_2)}{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2)} = \exp(\beta_1 + \beta_3 x_2)$$

$$\frac{\phi(x_1,x_2+1)}{\phi(x_1,x_2)} = \frac{\exp(\alpha+\beta_1 x_1+\beta_2(x_2+1)+\beta_3 x_1(x_2+1))}{\exp(\alpha+\beta_1 x_1+\beta_2 x_2+\beta_3 x_1 x_2)} = \exp(\beta_2 + \beta_3 x_1)$$

$$\frac{\phi(x_1+1,x_2+1)}{\phi(x_1,x_2)} = \exp(\beta_1 + \beta_2 + \beta_3(x_1 + x_2 + 1))$$

$$\frac{\phi(x_1+1,x_2+1)}{\phi(x_1,x_2+1)} = \exp(\beta_1 + \beta_3(x_2 + 1))$$

$$\frac{\phi(x_1+1,x_2+1)}{\phi(x_1+1,x_2)} = \exp(\beta_2 + \beta_3(x_1 + 1))$$

8.4 Qualitative Explanatory Variables

- Previous examples have dealt with dichotomous outcomes when explanatory variables were also dichotomous
- Logistic regression allows for combinations of dichotomous, nominal, ordinal or continuous explanatory variables
- This section is concerned with handling explanatory variables that are qualitative and contain three or more levels
- Following data come from study on urinary tract infections (Koch, Imrey, *et al*) Investigators were interested in whether the pattern of treatment differences are the same across diagnoses, i.e., is there a treatment × diagnosis interaction?

Diagnosis	Treatment	Cured	Not Cured	Prop. Cured
Complicated	A	78	28	0.736
Complicated	B	101	11	0.902
Complicated	C	68	46	0.596
Uncomplicated	A	40	5	0.889
Uncomplicated	B	54	5	0.915
Uncomplicated	C	34	6	0.850

- Assume data arose from stratified simple random sample so that response is distributed binomially for each diagnosis × treatment combination:

$$\Pr\{n_{hij}\} = \prod_{h=1}^{2}\prod_{i=1}^{3}\frac{n_{hi+}!}{n_{hi1}!\,n_{hi2}!}\theta_{hi}^{n_{hi1}}(1-\theta_{hi})^{n_{hi2}},$$

where θ_{hi} is probability that person with hth diagnosis receiving ith treatment is cured. n_{hi1} and n_{hi2} are number of patients of hth diagnosis and ith treatment who were and were not cured, respectively

8.4.1 Model Fitting

- Since there is interest in interaction term, preliminary model includes main effects and their interaction. Parameter α is intercept, β_1 is incremental for complicated diagnosis, β_2 is incremental effect for treatment A, β_3 is incremental effect for treatment B, and β_4 and β_5 represent interaction terms

$$\begin{bmatrix} \log it(\theta_{11}) \\ \text{logit}(\theta_{12}) \\ \text{logit}(\theta_{13}) \\ \text{logit}(\theta_{21}) \\ \text{logit}(\theta_{22}) \\ \text{logit}(\theta_{23}) \end{bmatrix} = \begin{bmatrix} \alpha+\beta_1+\beta_2 \quad +\beta_4 \\ \alpha+\beta_1 \quad +\beta_3 \quad +\beta_5 \\ \alpha+\beta_1 \\ \alpha \quad +\beta_2 \\ \alpha \quad +\beta_3 \\ \alpha \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \end{bmatrix}$$

8.4.2 PROC LOGISTIC for Nominal Effects

```
data uti;
     input diagnosis : $13. treatment $ response $ count @@;
     cards;
complicated     A    cured    78   complicated     A     not   28
complicated     B    cured    101  complicated     B     not   11
complicated     C    cured    68   complicated     C     not   46
uncomplicated   A    cured    40   uncomplicated   A     not   5
uncomplicated   B    cured    54   uncomplicated   B     not   5
uncomplicated   C    cured    34   uncomplicated   C     not   6
;
```

- Example of reference cell coding using the CLASS statement:

```
proc logistic data=uti;
   freq count;
   class diagnosis treatment / param=ref;
   model response = diagnosis treatment;
run;
```

Class Level Information

Class	Value	Design Variables 1	2
diagnosis	complicated	1	
	uncomplicated	0	
treatment	A	1	0
	B	0	1
	C	0	0

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
diagnosis	1	10.2885	0.0013
treatment	2	24.6219	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	1.4184	0.2987	22.5505	<.0001
diagnosis complicated	1	-0.9616	0.2998	10.2885	0.0013
treatment A	1	0.5847	0.2641	4.9020	0.0268
treatment B	1	1.5608	0.3160	24.4010	<.0001

- Odds ratios for TREATMENT indicate the odds of being cured with treatment A are 1.8 times as high as those with treatment C, and the odds of being cured with treatment B are 4.8 times as high as those with treatment C

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
diagnosis complicated vs uncomplicated	0.382	0.212	0.688
treatment A vs C	1.795	1.069	3.011
treatment B vs C	4.762	2.564	8.847

- Model-predicted probabilities and odds from main effects model:

Diagnosis	Trt	Pr {Cured}= θ_{hi}	Odds of Cured
Complicated	A	$e^{\alpha+\beta_1+\beta_2}/(1+e^{\alpha+\beta_1+\beta_2})$	$e^{\alpha+\beta_1+\beta_2}$
Complicated	B	$e^{\alpha+\beta_1+\beta 3}/(1+e^{\alpha+\beta_1+\beta_3})$	$e^{\alpha+\beta_1+\beta_3}$
Complicated	C	$e^{\alpha+\beta_1}/(1+e^{\alpha+\beta_1})$	$e^{\alpha+\beta_1}$
Uncomplicated	A	$e^{\alpha+\beta_2}/(1+e^{\alpha+\beta_2})$	$e^{\alpha+\beta_2}$
Uncomplicated	B	$e^{\alpha+\beta_3}/(1+e^{\alpha+\beta_3})$	$e^{\alpha+\beta_3}$
Uncomplicated	C	$e^{\alpha}/(1+e^{\alpha})$	e^{α}

- By default, SAS uses the last level of a CLASS variable as its reference. If we wish to have a user-specified reference population, we can do so without any recoding.

- For example, if we wanted to have those on Treatment A with a complicated diagnosis as the reference population, we could invoke the following code:

```
proc logistic data=uti;
      freq count;
      class diagnosis(ref='complicated') treatment(ref='A')
       / param=ref;
      model response = diagnosis treatment;
run;
```

Class Level Information

Class	Value	Design Variables 1	2
diagnosis	complicated	0	
	uncomplicated	1	
treatment	A	0	0
	B	1	0
	C	0	1

- Alternatively, we could use the REF=FIRST syntax to indicate that we would like the first ordered value to be the reference.

```
proc logistic data=uti;
     freq count;
     class diagnosis(ref=first) treatment(ref=first)
      / param=ref;
     model response = diagnosis treatment;
run;
```

Using the CLASS Statement in PROC LOGISTIC

- Using the CLASS statement in PROC LOGISTIC allows you to specify reference cell coding. However, the default coding scheme is effect, or deviation from the mean, coding.
- Example of effect coding using data set from 8.4:

```
proc logistic data=uti;
    freq count;
    class diagnosis treatment;
    model response = diagnosis treatment;
run;
```

Class Level Information

Class	Value	Design Variables 1	2
diagnosis	complicated	1	
	uncomplicated	-1	
treatment	A	1	0
	B	0	1
	C	-1	-1

- The CLASS statement also allows for convenient specification of interaction terms:

```
proc logistic data=uti;
  freq count;
  class diagnosis treatment;
  model response = diagnosis treatment
                   diagnosis*treatment;
run;
```

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
diagnosis	1	7.9448	0.0048
treatment	2	11.0731	0.0039
diagnosis*treatment	2	2.6384	0.2674

Analysis of Maximum Likelihood Estimates

Parameter			DF	Estimate	Standard Error	Chi-Square
Intercept			1	1.6376	0.1514	116.9795
diagnosis	complicated		1	-0.4268	0.1514	7.9448
treatment	A		1	-0.0856	0.2138	0.1604
treatment	B		1	0.6605	0.2225	8.8082
diagnosis*treatment	complicated	A	1	-0.1007	0.2138	0.2217
diagnosis*treatment	complicated	B	1	0.3458	0.2225	2.4141

Analysis of Maximum Likelihood Estimates

Parameter			Pr > ChiSq
Intercept			<.0001
diagnosis	complicated		0.0048
treatment	A		0.6888
treatment	B		0.0030
diagnosis*treatment	complicated	A	0.6377
diagnosis*treatment	complicated	B	0.1202

- To determine if interaction is meaningful, fit full and reduced model and take difference in likelihoods

Full: `model response = diagnosis treatment diagnosis*treatment;`

(–2 log L = 447.56)

Reduced: `model response = diagnosis treatment;`

(–2 log L = 450.07)

- Difference in number of parameters in models is 2, so compare difference in likelihoods to chi-square distribution with 2 df. This indicates that interaction term is not significant, and goodness-of-fit of main effects model (reduced model) is supported

- Differential effect main effects model is written as:

$$\begin{bmatrix} \log it(\theta_{11}) \\ \log it(\theta_{12}) \\ \log it(\theta_{13}) \\ \log it(\theta_{21}) \\ \log it(\theta_{22}) \\ \log it(\theta_{23}) \end{bmatrix} = \begin{bmatrix} \alpha + \beta_1 + \beta_2 \\ \alpha + \beta_1 + \beta_3 \\ \alpha + \beta_1 - \beta_2 - \beta_3 \\ \alpha - \beta_1 + \beta_2 \\ \alpha - \beta_1 + \beta_3 \\ \alpha - \beta_1 - \beta_2 - \beta_3 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & 0 \\ 1 & -1 & 0 & 1 \\ 1 & -1 & -1 & -1 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

α = average log odds of cure

β_1 = differential change in log odds for complicated diagnosis

β_2 = differential change in log odds for treatment A

β_3 = differential change in log odds for treatment B

- Formulas for model-predicted probabilities and odds:

Diagnosis	Treatment	Pr{Cured}	Odds of Cured
Complicated	A	$e^{\alpha+\beta_1+\beta_2}/(1+e^{\alpha+\beta_1+\beta_2})$	$e^{\alpha+\beta_1+\beta_2}$
Complicated	B	$e^{\alpha+\beta_1+\beta_3}/(1+e^{\alpha+\beta_1+\beta_3})$	$e^{\alpha+\beta_1+\beta_3}$
Complicated	C	$e^{\alpha+\beta_1-\beta_2-\beta_3}/(1+e^{\alpha+\beta_1-\beta_2-\beta_3})$	$e^{\alpha+\beta_1-\beta_2-\beta_3}$
Uncomplicated	A	$e^{\alpha-\beta_1+\beta_2}/(1+e^{\alpha-\beta_1+\beta_2})$	$e^{\alpha-\beta_1+\beta_2}$
Uncomplicated	B	$e^{\alpha-\beta_1+\beta_3}/(1+e^{\alpha-\beta_1+\beta_3})$	$e^{\alpha-\beta_1+\beta_3}$
Uncomplicated	C	$e^{\alpha-\beta_1-\beta_2-\beta_3}/(1+e^{\alpha-\beta_1-\beta_2-\beta_3})$	$e^{\alpha-\beta_1-\beta_2-\beta_3}$

- Odds of being cured for complicated diagnosis vs. uncomplicated diagnosis (using Treatment A) is:

$$\frac{e^{\alpha+\beta_1+\beta_2}}{e^{\alpha-\beta_1+\beta_2}} = e^{2\beta_1}$$

- Odds of being cured for Treatment A vs. Treatment B (using complicated diagnosis) is:

$$\frac{e^{\alpha+\beta_1+\beta_2}}{e^{\alpha+\beta_1+\beta_3}} = e^{\beta_2-\beta_3}$$

- Odds of being cured for Treatment A vs. Treatment C (using complicated diagnosis) is:

$$\frac{e^{\alpha+\beta_1+\beta_2}}{e^{\alpha+\beta_1-\beta_2-\beta_3}} = e^{2\beta_2+\beta_3}$$

- Odds of being cured for Treatment B vs. Treatment C (using complicated diagnosis) is:

$$\frac{e^{\alpha+\beta_1+\beta_3}}{e^{\alpha+\beta_1-\beta_2-\beta_3}} = e^{\beta_2+2\beta_3}$$

Type III Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
diagnosis	1	10.2885	0.0013
treatment	2	24.6219	<.0001

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	1.6528	0.1557	112.7189	<.0001
diagnosis complicated	1	-0.4808	0.1499	10.2885	0.0013
treatment A	1	-0.1304	0.1696	0.5914	0.4419
treatment B	1	0.8456	0.1970	18.4336	<.0001

- Plot odds ratio and predicted probabilities using PLOTS option
- Use ODDSRATIO statement with CL=BOTH option to obtain Wald and profile likelihood confidence intervals for the ORs

```
ods graphics on;
proc logistic plots(only)=(effect(clband yrange=(.5,1)
        x=treatment*diagnosis) oddsratio(logbase=2));
     freq count;
     class diagnosis treatment;
     model response = diagnosis treatment /
        scale=none aggregate;
     oddsratio treatment / cl=both;
     oddsratio diagnosis / cl=both;
run;
ods graphics off;
```

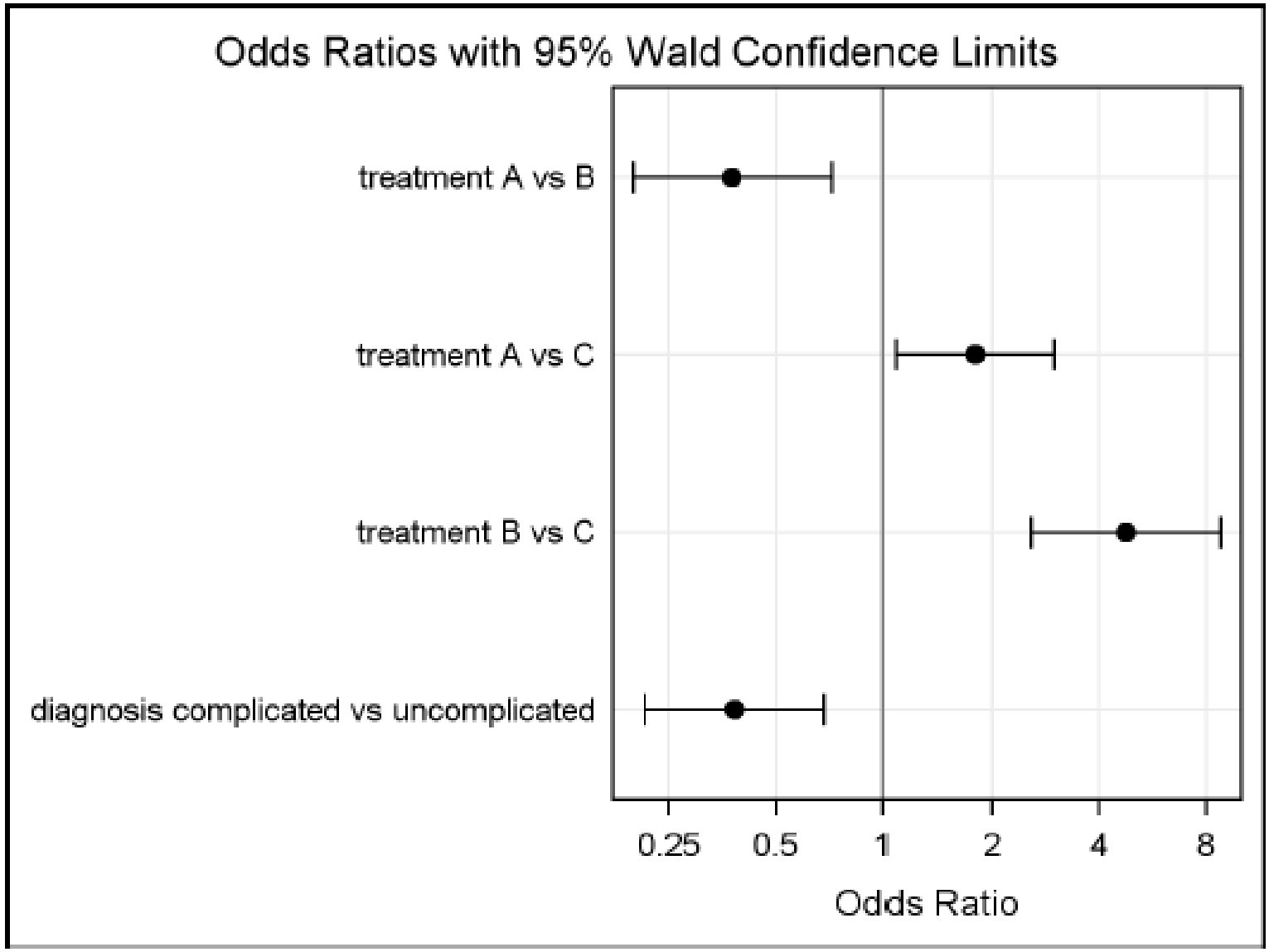
The LOGISTIC Procedure

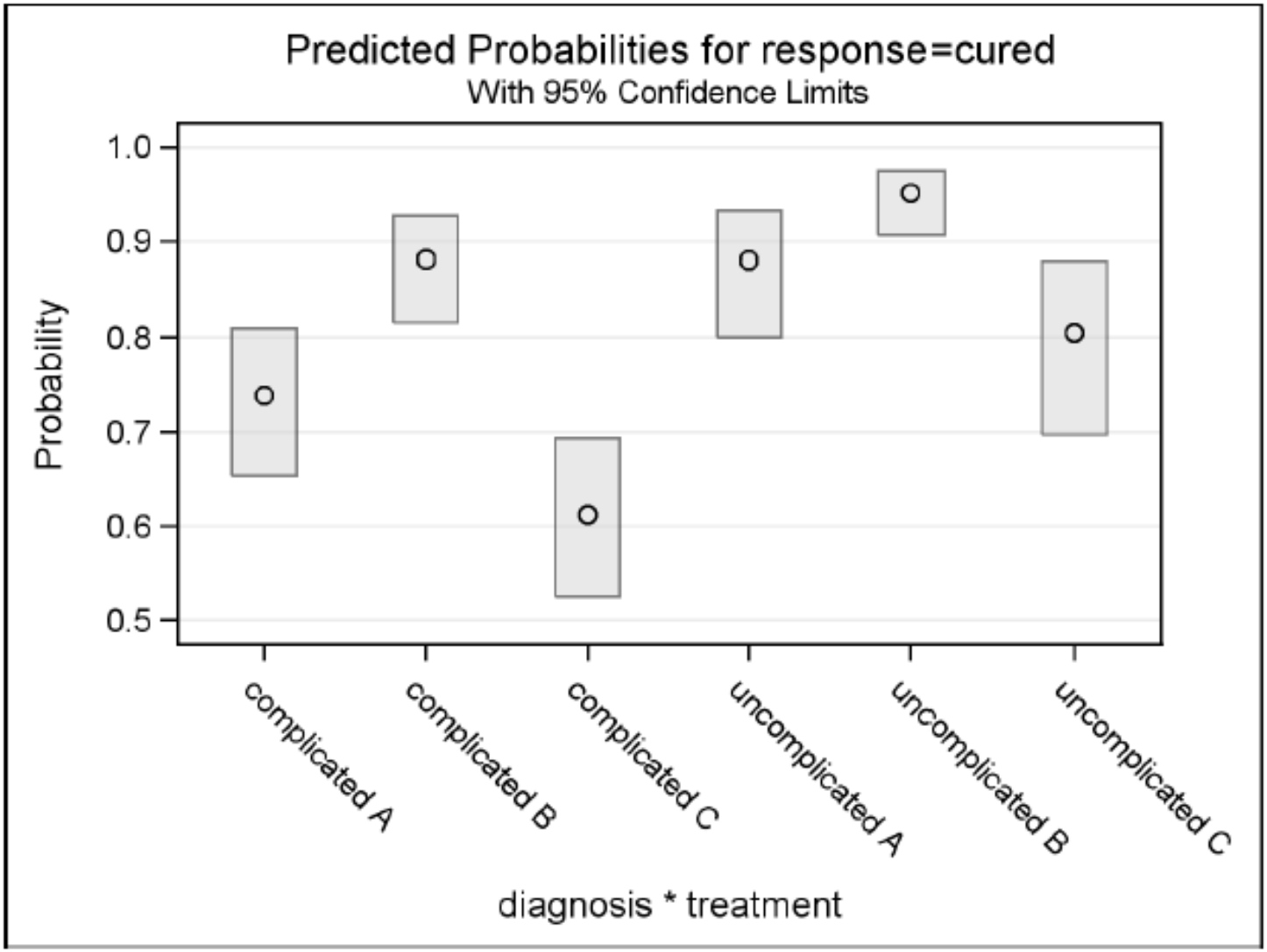
Odds Ratio Estimates and Wald Confidence Intervals

Label	Estimate	95% Confidence Limits	
treatment A vs B	0.377	0.197	0.721
treatment A vs C	1.795	1.069	3.011
treatment B vs C	4.762	2.564	8.847
Diagnosis complicated vs uncomplicated	0.382	0.212	0.688

Odds Ratio Estimates and Profile-Likelihood Confidence Intervals

Label	Estimate	95% Confidence Limits	
treatment A vs B	0.377	0.193	0.711
treatment A vs C	1.795	1.074	3.031
treatment B vs C	4.762	2.615	9.085
Diagnosis complicated vs uncomplicated	0.382	0.206	0.672


Odds Ratios with 95% Wald Confidence Limits
treatment A vs B
treatment A vs C
treatment B vs C
diagnosis complicated vs uncomplicated
0.25
0.5
1
2
4
8
Odds Ratio


Predicted Probabilities for response=cured
With 95% Confidence Limits
Probability
1.0
0.9
0.8
0.7
0.6
0.5
complicated A
complicated B
complicated C
uncomplicated A
uncomplicated B
uncomplicated C
diagnosis * treatment

8.4.3 Testing Hypotheses about the Parameters

- In previous analysis, both effects for treatment were significant. Can generate overall effect for treatment by computing likelihood ratio test for main effects model compared to model with diagnosis effect only, and compute difference (2 df test)
- May be interested in comparing treatment A vs B, or treatment B vs C. Can generate these tests using CONTRAST statements in LOGISTIC
- Create linear combinations of parameters and test if they are significantly different from zero:

$$H_0: \boldsymbol{L\beta} = \mathbf{0}$$

Wald statistic for given linear combination $\boldsymbol{L}$ is computed as:

$$Q_w = (\boldsymbol{L}\hat{\boldsymbol{\beta}})'(\boldsymbol{L}\boldsymbol{V}(\hat{\boldsymbol{\beta}})\boldsymbol{L}')^{-1}(\boldsymbol{L}\hat{\boldsymbol{\beta}})$$

where Q_W follows chi-square distribution with df = number of linearly independent rows of $\boldsymbol{L}$

- Test for A vs. B: H_0: $\beta_2 - \beta_3 = 0$
  Test for A vs. C: H_0: $\beta_2 = 0$
- Joint test of equality of treatments A, B and C:

  H_0: $\beta_2 = \beta_3 = 0$
- To implement hypothesis tests in PROC LOGISTIC:

```
proc logistic;
  freq count;
  class diagnosis treatment / param=ref;
  model response = diagnosis treatment;
  contrast ‘A vs B' treatment 1 -1 / estimate=exp;
  contrast 'A' treatment 1 0;
  contrast 'joint test' treatment 1 0,
                        treatment 0 1;
run;
```

- Results of main effects model contrasts

Contrast Test Results

Contrast	DF	Wald Chi-Square	Pr > ChiSq
A vs B	1	8.6919	0.0032
A	1	4.9020	0.0268
joint test	2	24.6219	<.0001

Contrast Estimation and Testing Results by Row

Contrast	Type	Row	Estimate	Standard Error	Alpha	Lower Limit	Upper Limit
A vs B	EXP	1	0.3768	0.1247	0.05	0.1969	0.7210

Contrast Estimation and Testing Results by Row

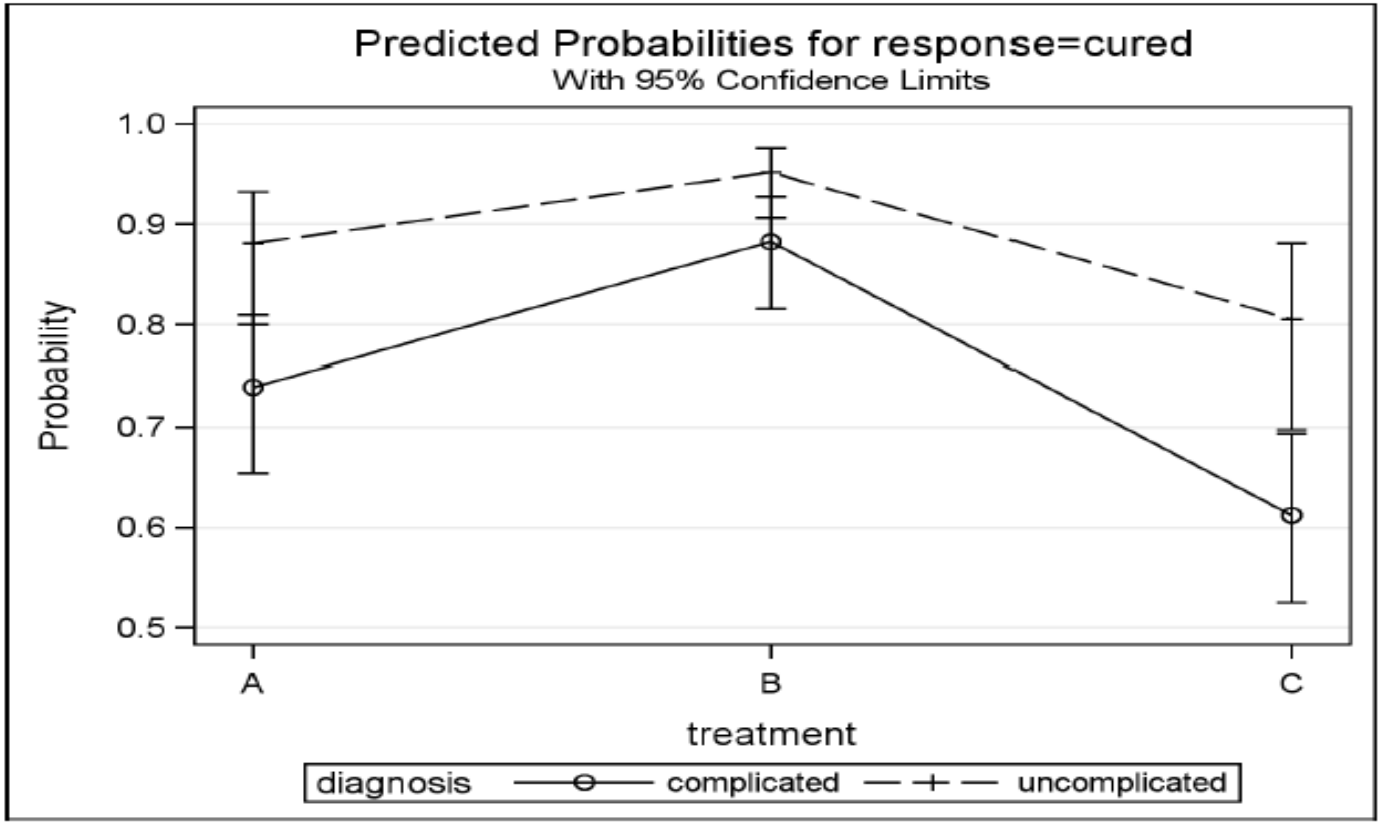
Contrast	Type	Row	Wald Chi-Square	Pr > ChiSq
A vs B	EXP	1	8.6919	0.0032

8.4.4 Additional Graphics

- The PLOTS= option in the PROC LOGISTIC statement provides a number of graphs, including an EFFECT plot to summarize the results of the analysis

```
ods graphics on;
proc logistic plots(only)=(effect(x=treatment sliceby=diagnosis
                            clbar connect yrange=(0.5));
    freq count;
    class diagnosis treatment / param=ref;
    model response = diagnosis treatment;
    oddsratio treatment / cl=pl;
    oddsratio treatment / cl=pl;
run;
ods graphics off;
```

- Persons with uncomplicated diagnosis do better than those with complicated diagnosis for all treatments. Persons who receive treatment B did best, and persons receiving treatment A did better than C



8.5 Continuous and Ordinal Explanatory Variables

8.5.1 Goodness of Fit

- Logistic regression analysis can involve continuous variables. Analysis strategies are same as previously described, except evaluating goodness of fit

- Following data are from study on coronary artery disease; AGE is a continuous explanatory variable, ECG is ordinal (0,1,2):

```
data coronary;
      input sex ecg age ca @@;
      cards;
      0 0 28 0   1 0 42 1   0 1 46 0   1 1 45 0
      0 0 34 0   1 0 44 1   0 1 48 1   1 1 45 1
      0 0 38 0   1 0 45 0   0 1 49 0   1 1 46 1
      0 0 41 1   1 0 46 0   0 1 49 0   1 1 48 1
      ⋮
      ;
run;
```

- SEX by ECG by AGE cross-classification produces 68 unique groups from the 78 observations. Therefore sample size requirements for Pearson χ^2 and likelihood ratio goodness-of-fit tests (each predicted cell count ≥ 5) not met
- Alternative 1: Fit desired model. Fit expanded model with additional explanatory variables. Evaluate difference in log likelihood ratio statistics. Difference is distributed χ^2 with df equal to difference in df in the two models
- Alternative 2: Fit desired model, including SELECTION = FORWARD to potentially add variables to model. Examine residual score statistic Q_{RS} that assesses joint contribution of remaining effects not yet incorporated in model. If there is an association, these variables should also be included in model
- Alternative 3: Fit desired model, specifying LACKFIT option in MODEL statement. This produces Hosmer and Lemeshow goodness-of-fit statistic which is compared to χ^2 distribution with t df (t is number of decile groups minus 2)

8.5.2 Fitting a Main Effects Model

- Main effects model includes: SEX, ECG and AGE
- To determine number of factors to include in expanded model: Need ≥ 5 observations for the rarer outcome level per parameter. Therefore, since 37 observations have no coronary artery disease, and 41 observations have coronary artery disease: model can only support 37/5 = 7 or 8 parameters
- Expanded model includes: main effects, squared terms for ECG and AGE, and all pairwise interactions
- To fit main effects model and compute score test:

```
proc logistic descending;
     model ca = sex ecg age
        ecg*ecg age*age sex*ecg sex*age ecg*age /
        selection=forward include=3 details lackfit;
run;
```

- Residual Chi-Square Q_{RS} = 2.3277 with df = 5 (since difference between number of parameters of models is 9 – 4 = 5). *p*-value = 0.8022, supporting goodness of fit of the main effects model

Assessing Fit

Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
2.3277	5	0.8022

Analysis of Effects Not in the Model

Effect	DF	Score Chi-Square	Pr > ChiSq
ecg*ecg	1	0.3766	0.5394
age*age	1	0.7712	0.3798
sex*ecg	1	0.0352	0.8513
sex*age	1	0.0290	0.8647
ecg*age	1	0.8825	0.3475

The likelihood ratio test could be formulated by comparing the –2 Log Likelihood from each model:

Model with quadratic and interaction terms:	84.379
Main effects model:	86.811

The difference is 2.432 with df = 5, producing $p = 0.787$, which agrees closely with the residual score test.

- DETAILS option causes printing of "Analysis of Variables Not in the Model". Each test is not significant, indicating adequate fit of the model without these factors
- Hosmer and Lemeshow statistic has value of 4.7766 with 8 df and p-value = 0.7812. This measure also supports adequacy of main effects model (Hosmer and Lemeshow statistic can also be used when all explanatory variables are qualitative)

- Estimated equation for log odds from main effects model:

$$\log it(\theta_{hi}) = -5.6418 + 1.3564SEX + 0.8732ECG + 0.0929AGE$$

- Coronary artery disease is positively associated with age and ST segment depression, and is more likely for males in this population
- Odds ratio for AGE is 1.097: extent to which odds increase each year. More desirable statistic is extent to which odds increase per 10 years of age: $e^{10\times 0.0929} = 2.53$. Can compute this using UNITS statement:

```
proc logistic descending;
     model ca = sex ecg age;
     units age = 10;
run;
```

Adjusted Odds Ratios

Effect	Unit	Estimate
age	10.0000	2.531

8.4.5 Fitting Models with Interactions

- Example: Examining association between occupational environment and prevalence of respiratory ailments associated with the disease byssinosis
- See page 226 for SAS data input

Workplace Condition	Years Employment	Smoking	Complaints	
			Yes	No
Dusty	< 10	Yes	30	203
Dusty	< 10	No	7	119
Dusty	≥ 10	Yes	57	161
Dusty	≥ 10	No	11	81
Not Dusty	< 10	Yes	14	1340
Not Dusty	< 10	No	12	1004
Not Dusty	≥ 10	Yes	24	1360
Not Dusty	≥ 10	No	10	986

```
proc logistic data=byss;
   freq count;
   class work years(ref=first) smoke(ref=first) /param=ref;
   model status(event=last) = work|years|smoke@2 /
       scale=none aggregate;
run;
```

Response Profile

Ordered Value	status	Total Frequency
1	no	5254
2	yes	165

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	0.6943	1	0.6943	0.4047
Pearson	0.6905	1	0.6905	0.4060

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
work	1	23.9781	<.0001
years	1	0.0085	0.9267
work*years	1	2.3264	0.1272
smoke	1	0.0100	0.9202
work*smoke	1	3.2242	0.0726
years*smoke	1	0.9101	0.3401

- YEARS*SMOKE interaction can be removed

```
ods graphics on;
proc logistic plots(only)=oddsratio(logbase=2));
   freq count;
   class work years(ref=first) smoke(ref=first) /param=ref;
   model status(event=last) = work years smoke
                       work*years work*smoke
                      /scale=none aggregate;
   effectplot interaction (x=work)/at(smoke=all years=all) link noobs;
   oddsratios work;
run;
ods graphics off;
```

Deviance and Pearson Goodness-of-Fit Statistics

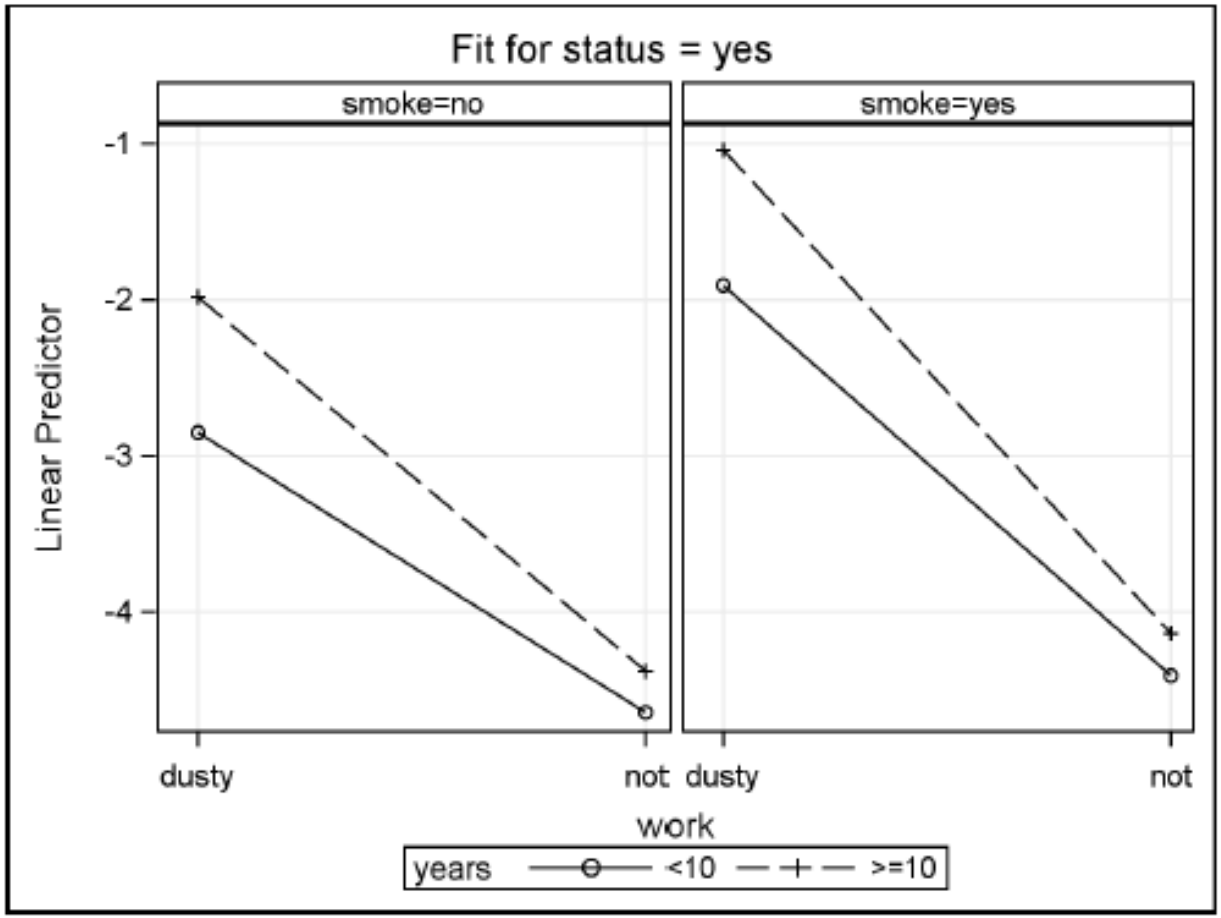
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	1.6016	2	0.8008	0.4490
Pearson	1.6027	2	0.8013	0.4487

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-4.6446	0.2598	319.6394	<.0001
work	dusty	1	1.7936	0.3833	21.9032	<.0001
years	>=10	1	0.2651	0.2622	1.0221	0.3120
smoke	yes	1	0.2387	0.2696	0.7843	0.3758
work*years	dusty >=10	1	0.6014	0.3444	3.0491	0.0808
work*smoke	dusty yes	1	0.7047	0.3857	3.3387	0.0677

- Pairwise interactions indicate that one variable's effect depends on the level of a second variable

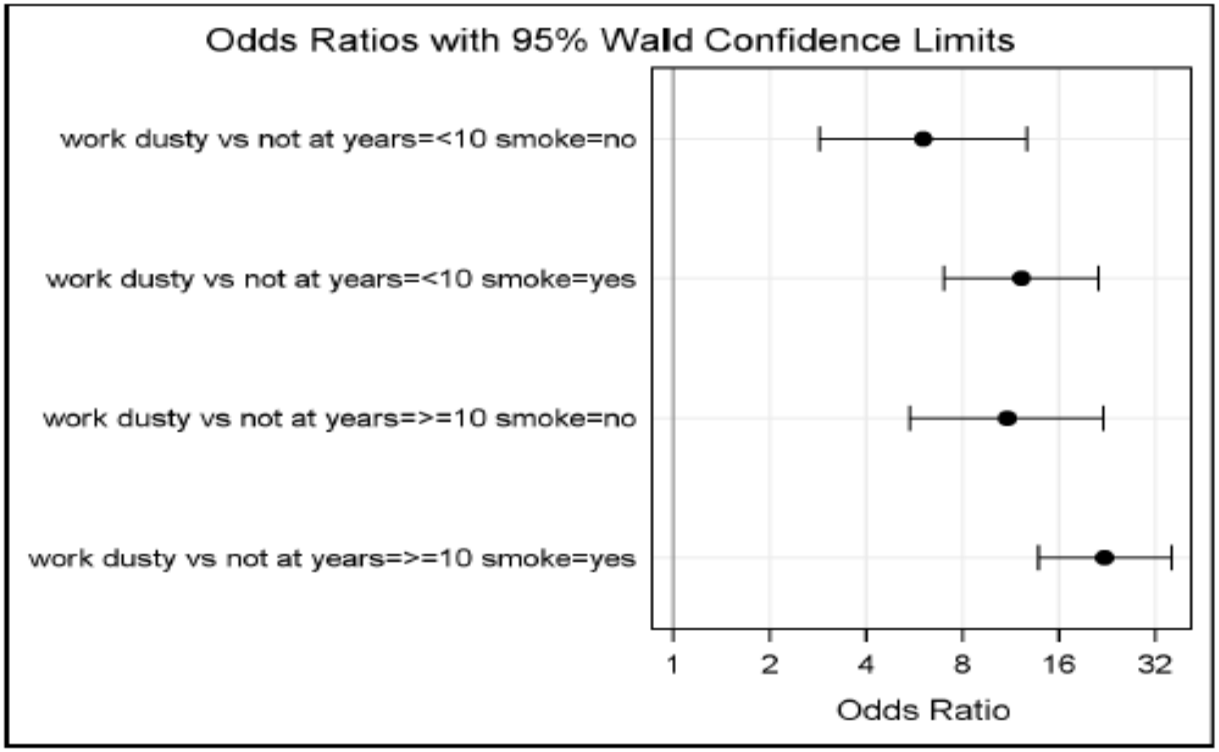
- Changes interpretations of the main effects



- The y-axis is the log odds of byssinosis symptoms. Lack of parallel lines indicates the effect of the interaction terms.

Odds Ratio Estimates and Wald Confidence Intervals

Label	Estimate	95% Confidence Limits	
work dusty vs not at years=<10 smoke=no	6.011	2.836	12.741
work dusty vs not at years=<10 smoke=yes	12.163	6.926	21.359
work dusty vs not at years=>=10 smoke=no	10.968	5.430	22.156
work dusty vs not at years=>=10 smoke=yes	22.192	13.652	36.073



PROC LOGISTIC: Plot Options

- `plots(only)=`
  - Suppresses default plots and only outputs requested plots
- `plots=(effect(clband yrange=(.5,1) x=treatment*diagnosis))`
  - Produces an effect plot with confidence limit bands
  - `yrange=(.5,1)` specifies range of y-axis. Default is (0,1).
  - `x=treatment*diagnosis` requests an effect be plotted (on the x-axis) for every combination of treatment and diagnosis
- `plots=(effect(x=treatment sliceby=diagnsosis clbar connect))`
  - Produces an effect plot with error bars
  - `x=treatment` specifies the x-axis
  - `sliceby=diagnosis` requests plots at each level of diagnosis
  - `connect` requests plotted estimates to be connect by lines

PROC LOGISTIC: Plot Options

- `plots=(oddsratio(logbase2))`
  - Plots odds ratios with 95% Wald confidence intervals
  - Creates log base 2 scale for the x-axis

- `effectplot interaction (x=work) / at(smoke=all years=all) link noobs;`
  - Creates effect plots displaying effects of interactions, with workplace as x-axis
  - `link` option requests y-axis to be on scale of linear predictor (log odds)

Appendix A Statistical Methodology for Dichotomous Logistic Regression

- Consideration of relationship between dichotomous outcome variable and set of explanatory variables arises from:

  i) clinical trials where explanatory variables are treatment, stratification variables, and background covariables

  ii) observational studies where explanatory variables represent factors for evaluation and background variables

- Assume the data arise from a stratified simple random sample so that a dichotomous outcome for the respective strata has a product binomial distribution:

$$\Pr\left\{n_{ij}\right\} = \prod_{i=1}^{s} \frac{n_{i+}!}{n_{i1}!n_{i2}!} \theta_i^{n_{i1}} \left(1-\theta_i\right)^{n_{i2}}$$

where θ_i is the probability that a randomly selected subject from the i-th stratum has outcome $j = 1$, and n_{i1} and n_{i2} are the numbers of subjects from the ith stratum with the $j = 1$ and $j = 2$ outcomes, $n_{i+} = (n_{i1} + n_{i2})$ and $i = 1,2, \ldots, s$.

- Model for θ specified as:

$$\theta = \frac{\exp\left(\alpha + \sum_{k=1}^{t} \beta_k x_k\right)}{1+\exp\left(\alpha + \sum_{k=1}^{t} \beta_k x_k\right)}$$

where $x_i, \ldots, x_t$ are explanatory variables

- Odds are written as :

$$\frac{\theta}{1-\theta} = \exp\left(\alpha + \sum_{k=1}^{t} \beta_k x_k\right)$$

- Model for logit is linear:

$$\log\left(\frac{\theta}{1-\theta}\right) = \alpha + \sum_{k=1}^{t} \beta_k x_k$$

- $\exp(\beta_k)$ = odds ratios for unit changes in x_k

- $\exp(\alpha)$ = odds when $x_1 = \ldots = x_t = 0$

- When data are from sampling process equivalent to stratified simple random sampling from subpopulations according to explanatory variables so that a product of binomial distributions applies, maximum likelihood estimates are obtained by solving:

$$\sum_{i=1}^{s} n_{i+}\hat{\theta}_i(1, x_{i1}, \ldots, x_{it}) = \sum_{i=1}^{s} n_{i1}(1, x_{i1}, \ldots, x_{it})$$

- Model-predicted value for θ_i:

$$\hat{\theta}_i = \frac{\exp\left(\hat{\alpha} + \sum_{k=1}^{t} \hat{\beta}_k x_k\right)}{1 + \exp\left(\hat{\alpha} + \sum_{k=1}^{t} \hat{\beta}_k x_k\right)},$$

where $\hat{\alpha}$ and $\hat{\beta}_k$ have approximate multivariate normal distributions

- The estimated covariance matrix for $\hat{\boldsymbol{\beta}}$ is $\left(\boldsymbol{X}_A'\boldsymbol{D}_v\boldsymbol{X}_A\right)^{-1}$

  where $\boldsymbol{D}_v = Diag\left\{n_{i+}\hat{\theta}_i(1-\hat{\theta}_i)\right\}$ and $\boldsymbol{X}_A = [\mathbf{1}, \boldsymbol{X}]$

- Goodness of fit can be assessed with Pearson chi-square statistics when sample sizes are adequate (80% of $\{n_{i1}\}$ and $\{n_i - n_{i1}\}$ are ≥ 5 and all others are ≥ 2):

$$Q_P = \sum_{i=1}^{s} \frac{(n_{i1} - n_{i+}\hat{\theta}_i)^2}{n_{i+}\hat{\theta}_i(1-\hat{\theta}_i)},$$

which is $\approx \chi^2(s-1-t)$

- Can also use log-likelihood ratio statistic (deviance) to evaluate goodness of fit

- In situations with continuous explanatory variables, Q_P is not appropriate. Instead, we need to fit expanded and reduced models. Let the original model matrix $\boldsymbol{X}$ have rank t, and expanded model $[\boldsymbol{X}, \boldsymbol{W}]$ have rank $t + w$. Evaluate significance of $\boldsymbol{W}$ by taking difference of log-likelihood statistics:

$$Q_{LR} = \sum_{i=1}^{s} \sum_{j=1}^{2} 2n_{ij} \log\left(\frac{m_{ij,w}}{m_{ij}}\right),$$

which is $\approx \chi^2(w)$; here $m_{i1} = n_{i+}\hat{\theta}_i$ and $m_{i2} = (n_{i+} - m_{i1})$ and similarly for $m_{ij,w}$.

- Another approach to goodness of fit that does not involve fitting expanded model is the score statistic to assess the association of residuals with $\boldsymbol{W}$:

$$Q_S = \boldsymbol{g}'\{\boldsymbol{W}'[\boldsymbol{D}_v - \boldsymbol{D}_v \boldsymbol{X}_A (\boldsymbol{X}_A' \boldsymbol{D}_v \boldsymbol{X}_A)^{-1} \boldsymbol{X}_A' \boldsymbol{D}_v]\boldsymbol{W}\}^{-1} \boldsymbol{g},$$

where the residuals $\boldsymbol{g} = \boldsymbol{W}'(\boldsymbol{n}_{*1} - \boldsymbol{m}_{*1})$

$$Q_S \approx \chi^2(w)$$

here, $\boldsymbol{D}_v = Diag\left\{n_i \hat{\theta}_i (1 - \hat{\theta}_i)\right\}$ and $\boldsymbol{X}_A = [\mathbf{1}, \boldsymbol{X}]$

8.8 Using GENMOD Procedure for Logistic Regression

<u>Generalized Linear Models</u>

- GENMOD fits generalized linear models: including not only classical linear models but logistic and probit (binary data), loglinear models (multinomial data) and Poisson regression (Poisson data)

- Generalized linear model has three components:
  i) response variable $\{y_i\}$ with probability distribution
  ii) set of explanatory variables $\boldsymbol{x}_i$ and parameters $\boldsymbol{\beta}$
  iii) monotonic link function g that describes how expected value of y_i (denoted μ_i) is related to $\boldsymbol{x}'_i\boldsymbol{\beta}$:

$$g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$$

- Generalized linear model is constructed by choosing appropriate link function and response probability distribution

Model	Probability Distribution	Link Function
Classical linear	Normal	$g(\mu) = \mu$
Logistic	Binomial	$g(\mu) = \log\left(\frac{\mu}{1-\mu}\right)$
Poisson	Poisson	$g(\mu) = \log(\mu)$

Fitting Logistic Regression Models with PROC GENMOD

- An attractive feature for using GENMOD is ease in handling qualitative variables with CLASS statement
- Reference cell (or incremental effects) coding is used (similar to LOGISTIC with CLASS statement and PARAM=REF or PARAM=GLM)
- GENMOD allows specification of a single response variable (using FREQ statement if data contain frequency counts) or outcome in *events/trials* form.

- Ex: Use urinary tract data in events/trial form to perform logistic regression in PROC GENMOD

```
data uti2;
     input diagnosis:$13. treatment $ events trials;
     datalines;
complicated A       78      106
     ⋮              ⋮        ⋮       ⋮
uncomplicated       C       34      40
;
run;

proc genmod;
     class diagnosis treatment;
     model events/trials = diagnosis treatment /
           link=logit  dist=binomial type3 aggregate;
run;
```

- To assess whether any treatments are similar, test linear combinations of parameters:

$$H_0: \boldsymbol{L\beta} = \mathbf{0}$$

- Likelihood ratio test is computed by default, Wald test can be produced by request

- Test for whether treatment A is equivalent to treatment B:

$$H_0: \beta_A = \beta_B$$

- Test for whether treatment A is equivalent to treatment C:

$$H_0: \beta_A = \beta_C$$

- Tests are requested with CONTRAST statement:

```
proc genmod;
    class diagnosis treatment;
    model events/trials = diagnosis treatment
      / link=logit  dist=binomial;
    contrast 'treatment'  treatment 1  0 -1,
                                    0  1 -1;
    contrast 'A-B'  treatment  1 -1   0;
    contrast 'A-C'  treatment  1  0  -1;
run;
```

- The ESTIMATE statement may be used to request an estimate and confidence interval for exp($\boldsymbol{L\beta}$).

Goodness of Fit

Class	Levels	Values
DIAGNOSIS	2	complicated uncomplicated
TREATMENT	3	A B C

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	2.5147	1.2573
Scaled Deviance	2	2.5147	1.2573
Pearson Chi-Square	2	2.7574	1.3787
Scaled Pearson X2	2	2.7574	1.3787
Log Likelihood		-225.0355	
Full Log Likelihood		-13.4690	
AIC (smaller is better)		34.9379	
AICC (smaller is better)		35.0228	
BIC (smaller is better)		51.5996	

Parameter Estimates

Analysis Of Parameter Estimates

Parameter		DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT		1	1.4184	0.2987	22.5505	0.0001
DIAGNOSIS	complicated	1	-0.9616	0.2998	10.2885	0.0013
DIAGNOSIS	uncomplicated	0	0.0000	0.0000	.	.
TREATMENT	A	1	0.5847	0.2641	4.9020	0.0268
TREATMENT	B	1	1.5608	0.3160	24.4010	0.0001
TREATMENT	C	0	0.0000	0.0000	.	.
SCALE		0	1.0000	0.0000	.	.

LR Statistics For Type 3 Analysis

Source	DF	ChiSquare	Pr>Chi
DIAGNOSIS	1	11.72	0.0006
TREATMENT	2	28.11	0.0001

Contrasts

CONTRAST Statement Results

Contrast	DF	ChiSquare	Pr>Chi	Type
treatment	2	28.1137	0.0001	LR
A-B	1	9.2218	0.0024	LR
A-C	1	4.9883	0.0255	LR

8.6 A Note on Diagnostics

Pearson Residuals: $$r_i = \frac{y_i - n_i\hat{\theta}_i}{\sqrt{n_i\hat{\theta}_i(1-\hat{\theta}_i)}}$$

- Used to compare differences between observed counts and their predicted values, scaled by the observed count's standard deviation.

- You can examine the r_i's to determine how well the model fits the invidual groups.

- Residual values exceeding 2 are considered to be indicative of lack of fit.

- The sums of the squares of the r_i's is Q_P

Deviance Residuals:

$$d_i = \text{sgn}(y_i - \hat{y}_i)\left[2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)\right]^{1/2}$$

where $\hat{y}_i = n\hat{\theta}_i$

- Both Pearson and deviance residuals can be standardized to have approximately unit variances

- The likelihood residual is another option, and is a weighted combination of the standardized Pearson and deviance residuals

- Standardized deviance residuals and likelihood residuals are recommended as they rank extreme observations well and are reasonably well approximated by a standard normal distribution when the numbers in each group are large enough

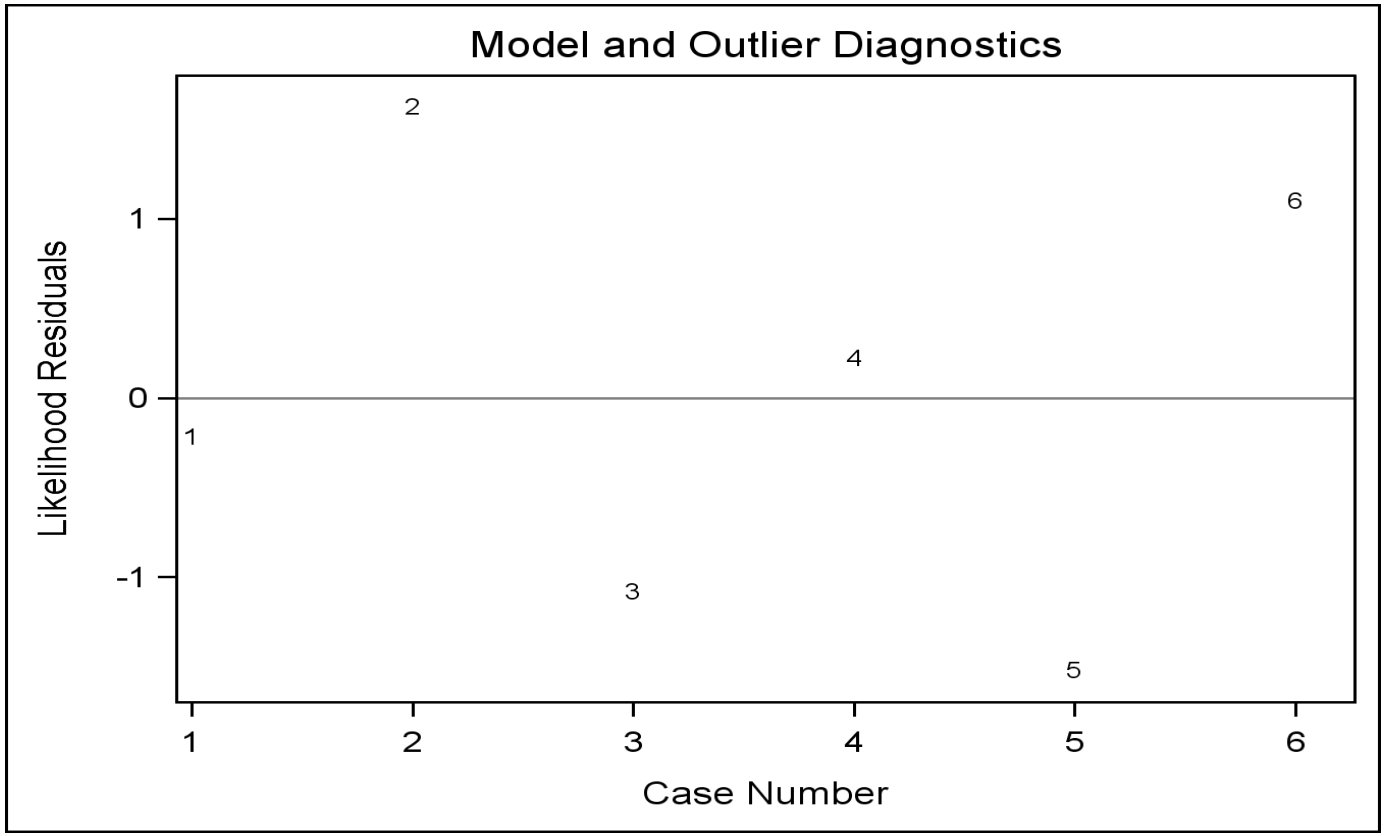
- Residuals can be examined in an index plot, in which residuals are plotted against the corresponding observation number.

- The INFLUENCE option requests that PROC LOGISTIC provide regression diagnostics.

- Data must be in *events/trial* form. Otherwise, when you compute residuals, they are calculated using a group size of 1.

```
data uti2;
   input diagnosis : $13.  treatment $ response trials;
datalines;
   complicated    A  78   106
   complicated    B  101  112
   complicated    C  68   114
   uncomplicated  A  40    45
   uncomplicated  B  54    59
   uncomplicated  C  34    40
;
run;
```

```
ods graphics on;
proc logistic data=uti2 plots(label)=influence(unpack stdres);
   class diagnosis treatment / param=ref;
   model response/trials = diagnosis treatment;
run;
ods graphics off;
```



Byssinosis Example – Main Effects Model

The following code can be used to create a main effects model for the byssinosis data in event/trials syntax, wherein some of the pairwise interactions had previously been found to be important.

```
data byss2;
  input work $ years $ smoke $ count trials @@;
  datalines;
dusty <10 yes 30 233  dusty <10 no 7  126  dusty >=10 yes 57 218  dusty >=10 no 11 92
not   <10 yes 14 1354 not   <10 no 12 1016 not   >=10 yes 24 1384 not   >=10 no 10 996
  ;
run;

ods graphics on;
proc logistic data=byss2 plots(label)=influence(unpack stdres);
   class work years(ref=first) smoke(ref=first) / param=ref;
   model count/trials = work years smoke work*years work*smoke years*smoke
   / selection=forward include=3 details scale=none aggregate;
run;
ods graphics off;
```

Byssinosis Example – Main Effects Model (continued)

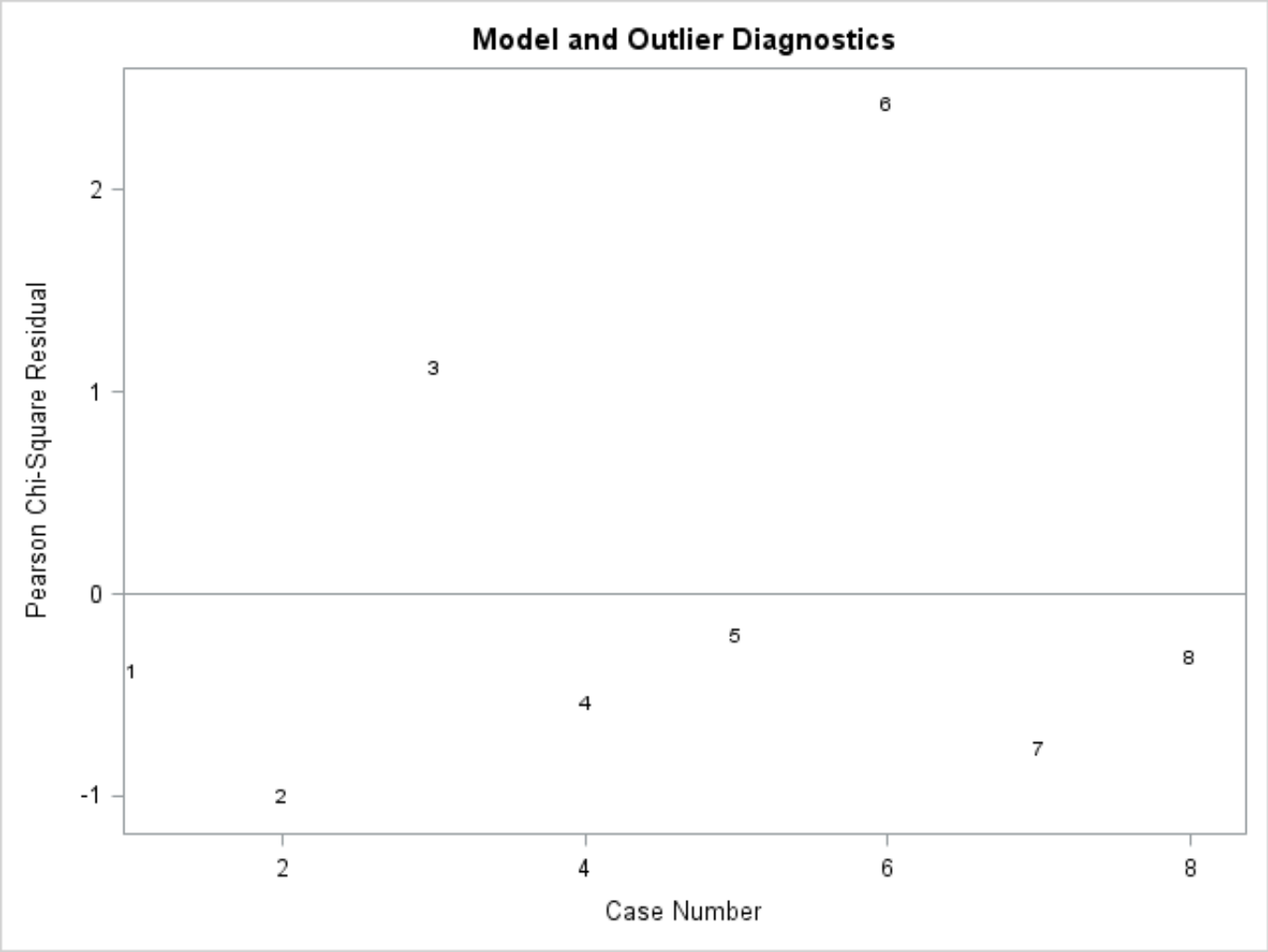
Residual Chi-Square Test

Chi-Square	DF	Pr > ChiSq
7.9772	3	0.0465

Analysis of Effects Eligible for Entry

Effect	DF	Score Chi-Square	Pr > ChiSq
work*years	1	3.1737	0.0748
work*smoke	1	3.4847	0.0619
years*smoke	1	1.2449	0.2645

Byssinosis Example – Main Effects Model (continued)



8.7.1 Examples of Non-Convergence

- Consider following table:

Factor	Response = Yes	Response = No
Factor 1	15	0
Factor 2	0	34

Computing odds ratio results in: $\frac{15\times34}{0\times0}$, which is infinite.
Since odds ratio is e^{β}, then β is also infinite
Warning message for complete separation is produced in output, ML estimate does not exist

- If convergence is not attained within 8 iterations, PROC LOGISTIC computes *Pr*(allocation each observation to correct response group):
  i) if prob = 1 for all observations, there is *complete separation* of data points and process is halted
  ii) if prob = 1 for nearly all observations, there is *quasicomplete separation* and process is halted
  iii) if neither of these exists, there is *overlapping* → ML estimates exist and are unique

- Problems of complete and quasi-complete separation generally occur for small sample sizes; usually quasi-complete does not occur if you have continuous explanatory variables; complete can always occur
- Following example has several zero frequencies for response distribution within groups:

Treatment Group	Yes	No
Control	0	0
Treatment A	8	0
Treatment B	0	2
Treatment A + B	21	6

```
proc logistic;
      freq count;
      model response = treatA treatB;
run;
```

Warning message for quasicomplete separation is produced in output, ML estimate may not exist

Log:

```
WARNING: There is possibly a quasi-complete separation of data points. The
         maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning.
         Results shown are based on the last maximum likelihood iteration.
         Validity of the model fit is questionable.
```

Output:

```
                              Model Convergence Status

                 Quasi-complete separation of data points detected.

WARNING: The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning.
         Results shown are based on the last maximum likelihood
iteration.
         Validity of the model fit is questionable.
```

- Following example includes two dichotomous explanatory variables and zero counts resulting in complete separation:

Gender	Region	Yes	No
Female	I	0	5
Female	II	1	0
Male	I	0	175
Male	II	53	0

```
proc logistic;
     model response = gender region;
run;
```

Warning message for complete separation is produced in output, ML estimate does not exist

Exact Methods in Logistic Regression

- It is now possible to compute parameter estimates, confidence intervals, and *p*-values using methodology based on exact distributions
- SAS provides this methodology beginning in version 8.1; exact estimates and confidence intervals are also available using LogXact from CYTEL Software Corporation
- Exact methods are particularly useful when there is non-convergence (or non-availability) for estimates from approximate methods.

Exact methods for the examples of non-convergence:

Factor	Response = Yes	Response = No
Factor 1	15	0
Factor 2	0	34

```
data factor;
  input factor response count @@;
  datalines;
1 0 0    1 1 15
2 0 34   2 1 0
;
proc logistic descending;
   freq count;
   class factor / param=ref;
   model response=factor;
   exact factor / estimate=odds;
run;
```

```
                          Model Convergence Status

                   Complete separation of data points detected.

WARNING: The maximum likelihood estimate does not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning.
         Results shown are based on the last maximum likelihood iteration.
         Validity of the model fit is questionable.
```

```
                    Testing Global Null Hypothesis: BETA=0

            Test                  Chi-Square       DF     Pr > ChiSq

            Likelihood Ratio         60.3568        1         <.0001
            Score                    49.0000        1         <.0001
            Wald                      0.2720        1         0.6020
```

Note the conflicting nature of the *p*-values. None is trustworthy.

```
WARNING: The validity of the model fit is questionable.

                    Analysis of Maximum Likelihood Estimates

                                          Standard
     Parameter       DF    Estimate        Error    Chi-Square    Pr > ChiSq

     Intercept        1     -9.4177      19.0251        0.2450        0.6206
     factor     1     1     18.9980      36.4303        0.2720        0.6020


                              Odds Ratio Estimates

                                  Point          95% Wald
               Effect           Estimate      Confidence Limits

               factor 1 vs 2    >999.999      <0.001     >999.999
```

Note the almost impossibly large estimates for the log odds ratio and its standard error. Neither is trustworthy.

Exact Results

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-Value --- Exact	Mid
factor	Score	48.0000	<.0001	<.0001
	Probability	6.35E-13	<.0001	<.0001

Exact Odds Ratios

Parameter		Estimate	95% Confidence Limits		p-Value
factor	1	613.522*	62.864	Infinity	<.0001

NOTE: * indicates a median unbiased estimate.

```
data quasi;
   input treatA treatB response count @@;
   datalines;
   0 0 0 0  0 0 1  0
   0 1 0 2  0 1 1  0
   1 0 0 0  1 0 1  8
   1 1 0 6  1 1 1 21
   ;

proc logistic descending;
   freq count;
   model response= TreatA TreatB;
   exact TreatA TreatB / estimate=both;
run;
```

Model Convergence Status

Quasi-complete separation of data points detected.

```
WARNING: The maximum likelihood estimate may not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning.
         Results shown are based on the last maximum likelihood iteration.
         Validity of the model fit is questionable.
```

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	10.0294	2	0.0066
Score	9.4626	2	0.0088
Wald	0.0089	2	0.9956

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-1.9635	367.8	0.0000	0.9957
treatA	1	13.6453	347.1	0.0015	0.9686
treatB	1	-10.4290	121.7	0.0073	0.9317

Exact Results

Exact Conditional Analysis

Conditional Exact Tests

Effect	Test	Statistic	--- p-Value --- Exact	Mid
treatA	Score	5.4444	0.0690	0.0345
	Probability	0.0690	0.0690	0.0345
treatB	Score	2.0843	0.2994	0.2082
	Probability	0.1824	0.2994	0.2082

Exact Odds Ratios

Parameter	Estimate	95% Confidence Limits		p-Value
treatA	7.062*	0.522	Infinity	0.1379
treatB	0.352*	0	2.850	0.3647

NOTE: * indicates a median unbiased estimate.

Gender	Region	Yes	No
Female	I	0	5
Female	II	1	0
Male	I	0	175
Male	II	53	0

```
data complete;
   input gender region count response @@;
   datalines;
   0 0 0  1  0 0 5   0
   0 1 1  1  0 1 0   0
   1 0 0  1  1 0 175 0
   1 1 53 1  1 1 0   0
   ;
proc logistic descending;
   freq count;
   model response = gender region;
   exact gender region / estimate=both;
run;
```

Model Convergence Status

Complete separation of data points detected.

WARNING: The maximum likelihood estimate does not exist.
WARNING: The LOGISTIC procedure continues in spite of the above warning.
Results shown are based on the last maximum likelihood iteration.
Validity of the model fit is questionable.

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	252.7878	2	<.0001
Score	234.0000	2	<.0001
Wald	0.6762	2	0.7131

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-9.6290	52.3141	0.0339	0.8540
gender	1	-1E-14	52.9721	0.0000	1.0000
region	1	19.8262	24.1158	0.6759	0.4110

WARNING: The validity of the model fit is questionable.

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
gender	1.000	<0.001	>999.999
region	>999.999	<0.001	>999.999

Even though the exact methods are appropriate for these data, SAS is unable to perform the necessary computation due to the large cell count of 175, and no exact results are provided.

Log:

```
WARNING: There is a complete separation of data points. The maximum
         likelihood estimate does not exist.
WARNING: The LOGISTIC procedure continues in spite of the above
         warning. Results shown are based on the last maximum
         likelihood iteration. Validity of the model fit is
         questionable.
WARNING: The permutation distribution contains frequencies larger
         than 9.0071993E15; accuracy was lost.
```

- Example: the following data are from a study on liver function outcomes for high risk overdose patients in which antidote and historical control groups are compared

Time to Hospital	Antidote		Control	
	Severe	Not Severe	Severe	Not Severe
Early	6	12	6	2
Delayed	3	4	3	0
Late	5	1	6	0

- These data do not present a complete or quasicomplete separation problem. However, due to the small cell counts, exact logistic regression is the appropriate method.

```
data liver;
   input time $ group $ status $ count @@;
datalines;
early   antidote severe 6 early   antidote not 12
early   control  severe 6 early   control  not  2
delayed antidote severe 3 delayed antidote not 4
delayed control  severe 3 delayed control  not 0
late    antidote severe 5 late    antidote not 1
late    control  severe 6 late    control  not 0
;
run;
```

```
proc logistic descending;
   freq count;
   class time (ref=‘early’) group(ref=‘control’) / param=ref;
   model status = time group / clparm=wald;
run;
```

Global Fit Statistics

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	16.3913	3	0.0009
Score	13.4256	3	0.0038
Wald	10.2488	3	0.0166

MLE Estimates

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1.4132	0.7970	3.1439	0.0762
time delayed	1	0.7024	0.8344	0.7087	0.3999
time late	1	2.5533	1.1667	4.7893	0.0286
group antidote	1	-2.2170	0.8799	6.3480	0.0118

Odds Ratio Estimates

Odds Ratio Estimates

Effect		Point Estimate	95% Wald Confidence Limits	
time	delayed vs early	2.019	0.393	10.359
time	late vs early	12.849	1.305	126.471
group	antidote vs control	0.109	0.019	0.611

- However, we would not report the maximum likelihood estimates and corresponding odds ratios due to sample size concerns.

- The following statements request an exact analysis:

```
proc logistic descending;
   freq count;
   class time (ref=‘early’) group(ref=‘control’) / param=ref;
   model status = time group / scale=none aggregate clparm=wald;
   exact ‘Model 1’ intercept time group / estimate=both;
   exact ‘Joint Test’ time group / joint;
run;
```

Exact Results

Exact Conditional Analysis

Exact Conditional Tests for Model 1

Effect	Test	Statistic	--- p-Value --- Exact	Mid
Intercept	Score	3.4724	0.1150	0.0922
	Probability	0.0457	0.1150	0.0922
time	Score	6.0734	0.0442	0.0418
	Probability	0.00471	0.0442	0.0418
group	Score	7.1656	0.0085	0.0050
	Probability	0.00698	0.0085	0.0050

Exact Conditional Tests for Joint Test

Effect	Test	Statistic	--- p-Value --- Exact	Mid
Joint	Score	13.1459	0.0027	0.0027
	Probability	0.000015	0.0015	0.0015
time	Score	6.0734	0.0442	0.0418
	Probability	0.00471	0.0442	0.0418
group	Score	7.1656	0.0085	0.0050
	Probability	0.00698	0.0085	0.0050

Exact Parameter Estimates for Model 1

Parameter		Estimate	Standard Error	95% Confidence Limits		Two-Sided p-value
Intercept		1.3695	0.7903	-0.2361	3.6386	0.1140
time	delayed	0.6675	0.8141	-1.2071	2.6444	0.6667
time	late	2.4388	1.1425	0.1364	6.4078	0.0331
group	antidote	-2.0992	0.8590	-4.5225	-0.3121	0.0154

Exact Odds Ratios for Model 1

Parameter		Estimate	95% Confidence Limits		p-Value
Intercept		3.934	0.790	38.037	0.1140
time	delayed	1.949	0.299	14.075	0.6667
time	late	11.460	1.146	606.546	0.0331
group	antidote	0.123	0.011	0.732	0.0154

Firth Bias Reduction Method

- An alternative strategy to exact methods is Firth's penalized likelihood method. This is a bias reduction method that adds a term to the usual log-likelihood function. When the resulting penalized likelihood method is maximized, it shrinks the estimates towards zero.

- Firth's method is especially useful when you are dealing with continuous explanatory variables and exact methods may not be applicable. It always produces parameter estimates when the issue is complete or quasi-complete separation.

- Request Firth's method using the FIRTH option in the MODEL statement of PROC LOGISTIC

    - Should always use CLPARM=PL option with Firth's method since the profile likelihood based confidence limits will be based on the penalized likelihood

```
proc logistic data=liver;
   freq count;
   class time (ref=‘early’) group(ref=‘control’) / param=ref;
   model status = time group / firth clparm=pl;
run;
```

Parameter Estimates and Profile-Likelihood Confidence Intervals

Parameter		Estimate	95% Confidence Limits	
Intercept		1.2077	-0.0769	2.8718
time	delayed	0.6374	-0.9007	2.2523
time	late	2.1543	0.4031	4.5421
group	antidote	-1.9526	-3.7557	-0.5053

In general, exact tests are recommended for small sample situations, but the Firth penalized likelihood approach is a useful alternative, especially when exact methods are computationally infeasible

Firth's method applied to previous example of completely separated data:

Gender	Region	Yes	No
Female	I	0	5
Female	II	1	0
Male	I	0	175
Male	II	53	0

```
proc logistic data=complete descending;
   freq count;
   model response = gender region / firth clparm=pl
   exact gender region;
run;
```

The exact results were non-conclusive because the computations ran into a degenerate distribution. The Firth method, however, does produce estimates.

Penalized Parameter Estimates

Analysis of Penalized Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.4001	1.6189	2.1978	0.1382
gender	1	-3.4599	2.1523	2.5843	0.1079
Region	1	10.5320	2.0164	27.2817	<.0001

Parameter Estimates and Profile-Likelihood Confidence Intervals

Parameter	Estimate	95% Confidence Limits	
Intercept	-2.4001	.	-0.2218
gender	-3.4599	-8.7265	.
region	10.5320	7.5460	16.2653

These estimates should be used cautiously. However, the confidence interval for region conveys the impression that region is an important effect.

One way to evaluate the parameter estimates is to collapse the two tables into one 2×2 table and add 0.5 to each of the counts. Collapsing over gender is justified since gender appears to have no effect:

Region	Yes	No
I	0.5	180.5
II	54.5	0.5

If you compute the odds ratio for this table, you obtain (0.5)(0.5)/(54.5)(180.5) = 0.00003, which is about the same as the exponentiated parameter for region. Thus, these estimates appear to be reasonable.

8.7.4 Exact Confidence Limits for Common Odds Ratios for 2x2 Tables

When you have multiple 2 × 2 tables, you may be interested in computing exact confidence limits for the average odds ratio among the set.

To do so, formulate the analysis as a regression where the column variable is the response variable and the row and stratification variables are the explanatory variables.

Then, condition on the stratification variable and estimate the odds ratio for the row variable. This odds ratio will be an average odds ratio.

Example: Association of office exercise program and test results, stratified by location.

Cardiovascular Test Outcomes

Location	Program	Good	Not Good	Total
Downtown	Office	12	5	17
Downtown	Home	3	5	8
	Total	15	10	26
Satellite	Office	6	1	7
Satellite	Home	1	3	4
	Total	7	4	11

See page 250 for SAS code to input data

```
proc logistic;
   freq count;
   class location program(ref=first) / param=ref;
   model outcome = location program;
   exact program / estimate=both;
run;
```

Exact Test Results

Exact Conditional Tests

Effect	Test	Statistic	p-value Exact	p-value Mid
program	Score	5.5739	0.0307	0.0215
	Probability	0.0183	0.0307	0.0215

Exact Odds Ratio

Parameter		Estimate	95% Confidence Limits		Two-Sided p-value
program	office	5.413	1.049	33.312	0.0424

Asymptotic Odds Ratio

Effect	Point Estimate	95% Wald Confidence Limits	
program office vs home	6.111	1.331	28.062

Using the exact method provides a more accurate picture in this example than the inappropriate asymptotic method.