

SAS Program Notes

Biostatistics: A Guide to Design, Analysis, and Discovery

Chapter 14: Logistic and Proportional Hazards Regression

Before we begin, we would like to mention that there are several good books on logistic regression. Some of the most popular are:

Allison PD. **Logistic Regression Using the SAS System Theory and Application**, SAS Institute Inc. 1999

Collett D. **Modeling Binary Data**, Chapman & Hall 1991

Hosmer DW and Lemeshow S. **Applied Logistic Regression**. John Wiley & Sons 2000

Kleinbaum DG, Klein M. **Logistic Regression: A Self-Learning Text 2nd**, Springer 2002

Note 14.1 - Logistic Regression Analysis

PROC LOGISTIC is the SAS procedure used to conduct logistic regression analyses. As an example, we use the pulmonary function test (**PFT**) data presented in Table 14.1. In the data set, the variable **PFT** equals '1' to represent a normal pulmonary function test result and '0' to indicate a result that is not normal. The variable **LEAD** equals '1' to represent low levels of lead and '0' to indicate high lead levels. In this case, we are considering the high lead levels as the reference level. We can use the following SAS code to create the data presented in Table 14.1.

SAS commands:

```
DATA PFT;
  INPUT PFT LEAD COUNT;
DATALINES;
1 0 82
1 1 368
0 0 10
0 1 19
;
PROC LOGISTIC DESCENDING;
  WEIGHT COUNT;
  MODEL PFT = LEAD;
RUN;
```

SAS output:

The SAS System
The LOGISTIC Procedure
Model Information

Data Set

WORK.PFT

Response Variable PFT
 Number of Response Levels 2
 Weight Variable COUNT
 Model binary logit
 Optimization Technique Fisher's scoring

Number of Observations Read 4
 Number of Observations Used 4
 Sum of Weights Read 479
 Sum of Weights Used 479

Response Profile

Ordered Value	PFT	Total Frequency	Total Weight
1	1	2	450.00000
2	0	2	29.00000

Probability modeled is PFT=1.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	220.863	218.838
SC	220.249	217.611
-2 Log L	218.863	214.838

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.0248	1	0.0448
Score	4.6421	1	0.0312
Wald	4.4105	1	0.0357

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.1040	0.3349	39.4608	<.0001
LEAD	1	0.8596	0.4093	4.4105	0.0357

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
LEAD	2.362	1.059 5.269

Association of Predicted Probabilities and Observed Responses

	Percent Concordant	Percent Discordant	Percent Tied	Pairs	Somers' D	Gamma	Tau-a	c
	25.0	25.0	50.0	4	0.000	0.000	0.000	0.500

The values for the Wald Statistic shown in the text in Table 14.3 on page 391 are the square roots of the Wald Statistic shown in the above output. In the text, under the null hypothesis, the test statistic shown in the text would follow the standard normal distribution. Its square, the statistic shown above, would follow the chi-square distribution with one degree of freedom.

In Example 14.3, we can use **PROC LOGISTIC** to explore the relationship between diabetes and body mass index (BMI) using individuals from the DIG200 data set. In the example, body mass index was rounded to the nearest whole number. In the SAS commands below, we create the new variable **ROUNDBMI** by rounding the variable BMI to the nearest whole number.

SAS commands:

```
PROC IMPORT FILE= 'C:\DIG200.XLS' OUT=DIG200_DATA REPLACE;
RUN;

DATA DIG200;
  SET DIG200_DATA;
  ROUNDBMI = ROUND(BMI, 1);
PROC LOGISTIC DESCENDING;
  MODEL DIABETES = ROUNDBMI;
RUN;
```

SAS output:

```

                                The SAS System
                                The LOGISTIC Procedure
                                Model Information

Data Set                        WORK.DIG200
Response Variable                diabetes          diabetes
Number of Response Levels       2
Model                            binary logit
Optimization Technique           Fisher's scoring

                                Number of Observations Read      200
                                Number of Observations Used      200

                                Response Profile

Ordered Value      diabetes      Total
Frequency
1                   1             54
2                   0             146

Probability modeled is diabetes=1.

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

                                Model Fit Statistics

Criterion            Intercept Only      Intercept and
Covariates
AIC                  235.304          231.701
SC                   238.602          238.298
```

		-2 Log L	233.304	227.701		
Testing Global Null Hypothesis: BETA=0						
Test		Chi-Square	DF	Pr >	ChiSq	
Likelihood Ratio		5.6025	1	0.0179		
Score		5.7875	1	0.0161		
Wald		5.5190	1	0.0188		
Analysis of Maximum Likelihood Estimates						
Parameter	DF	Estimate	Standard Error	Chi-Square	Wald	Pr > ChiSq
Intercept	1	-3.0342	0.8933	11.5364		0.0007
ROUNDBMI	1	0.0755	0.0321	5.5190		0.0188
Odds Ratio Estimates						
Effect	Point Estimate	95% Wald Confidence Limits				
ROUNDBMI	1.078	1.013 1.148				
Association of Predicted Probabilities and Observed Responses						
Percent Concordant	57.2	Somers' D	0.208			
Percent Discordant	36.4	Gamma	0.222			
Percent Tied	6.5	Tau-a	0.082			
Pairs	7884	c	0.604			

The values for the Wald Statistic shown in the text in Table 14.4 on page 392 again are the square roots of the Wald Statistic shown in the above output.

To consider a 5 kg/m² increase in BMI, we can use the **UNIT** statement followed by the variable of interest which is **ROUNDBMI** set equal to 5. The **UNIT** statement in SAS Version 9.1 is displayed in red as shown in the SAS code below. In the last portion of output, **Adjusted Odds Ratios** are displayed that produce the odds of diabetes for a 5 kg/m² increase in body mass index.

SAS commands:

```
PROC IMPORT FILE= 'C:\DIG200.XLS' OUT=DIG200_DATA REPLACE;
RUN;

DATA DIG200;
  SET DIG200_DATA;
  ROUNDBMI = ROUND(BMI, 1);
PROC LOGISTIC DESCENDING;
  MODEL DIABETES = ROUNDBMI;
  UNIT ROUNDBMI = 5;
RUN;
```

SAS output:

The SAS System
The LOGISTIC Procedure
Model Information

Data Set WORK.DIG200
 Response Variable diabetes diabetes
 Number of Response Levels 2
 Model binary logit
 Optimization Technique Fisher's scoring

Number of Observations Read 200
 Number of Observations Used 200

Response Profile

Ordered Value	diabetes	Total Frequency
1	1	54
2	0	146

Probability modeled is diabetes=1.

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	235.304	231.701
SC	238.602	238.298
-2 Log L	233.304	227.701

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5.6025	1	0.0179
Score	5.7875	1	0.0161
Wald	5.5190	1	0.0188

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.0342	0.8933	11.5364	0.0007
ROUNDBMI	1	0.0755	0.0321	5.5190	0.0188

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
ROUNDBMI	1.078	1.013 1.148

Association of Predicted Probabilities and Observed Responses

Percent Concordant	57.2	Somers' D	0.208
Percent Discordant	36.4	Gamma	0.222
Percent Tied	6.5	Tau-a	0.082
Pairs	7884	c	0.604

Adjusted Odds Ratios

Effect	Unit	Estimate
ROUNDBMI	5.0000	1.458

Note 14.2 - Logistic Regression Analysis continued

In Example 14.6, we fitted a multivariate logistic regression model where the outcome is pulmonary function test (PFT) and the independent variables are lead (LEAD- equals '1' to represent low levels of lead and '0' to indicate high lead levels) and smoking (SMOKING- equals '0' to represent a heavy smoker, '1' for a light smoker, '2' for a former smoker, and '3' for someone who has never smoked). A way of introducing the smoking variable into the model is to create dummy variables out of the variable SMOKING. In the SAS commands below, we use IF-THEN statements to create the three dummy variables for smoking status. We begin by setting the three dummy variables SMOK1, SMOK2, and SMOK3 equal to '0'. Then variable SMOK1 will have the value of '1' if SMOKING is '1' and '0' otherwise; the variable SMOK2 will have the value of '1' if SMOKING is '2' and '0' otherwise; and the variable SMOK3 will have the value of '1' if SMOKING is '3' and '0' otherwise. This means that the heavy level of smoking is the reference level. See the figure below and refer to the text for more information on creating dummy variables.

SMOKING	SMOK1	SMOK2	SMOK3
0-heavy	0	0	0
1-light	1	0	0
2-former	0	1	0
3-never	0	0	1

SAS commands:

```
DATA PFT;
  INPUT PFT LEAD SMOKING COUNT;
  SMOK1=0; SMOK2=0; SMOK3=0;
  IF SMOKING = 1 THEN SMOK1=1;
  IF SMOKING = 2 THEN SMOK2=1;
  IF SMOKING = 3 THEN SMOK3=1;
  OUTPUT;
DATALINES;
1 1 0 84
1 1 1 75
1 1 2 49
1 1 3 160
1 0 0 16
1 0 1 21
1 0 2 12
1 0 3 33
0 1 0 3
0 1 1 6
0 1 2 6
0 1 3 4
0 0 0 3
0 0 1 2
0 0 2 2
0 0 3 3
;
PROC LOGISTIC DESCENDING;
```

```

WEIGHT COUNT;
MODEL PFT = LEAD SMOK1 SMOK2 SMOK3;
RUN;

```

SAS output:

The SAS System
The LOGISTIC Procedure
Model Information

Data Set	WORK.PFT
Response Variable	PFT
Number of Response Levels	2
Weight Variable	COUNT
Model	binary logit
Optimization Technique	Fisher's scoring
Number of Observations Read	16
Number of Observations Used	16
Sum of Weights Read	479
Sum of Weights Used	479

Response Profile

Ordered Value	PFT	Total Frequency	Total Weight
1	1	8	450.00000
2	0	8	29.00000

Probability modeled is PFT=1.

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	220.863	218.949
SC	221.636	222.812
-2 Log L	218.863	208.949

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.9143	4	0.0419
Score	10.8622	4	0.0282
Wald	9.9943	4	0.0405

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	2.1778	0.5096	18.2643	<.0001
LEAD	1	0.8374	0.4135	4.1024	0.0428
SMOK1	1	-0.2892	0.5618	0.2650	0.6067
SMOK2	1	-0.7673	0.5673	1.8292	0.1762
SMOK3	1	0.5076	0.5719	0.7879	0.3747

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
LEAD	2.310	1.027	5.195
SMOK1	0.749	0.249	2.252
SMOK2	0.464	0.153	1.411
SMOK3	1.661	0.542	5.096

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	43.8	Somers' D	0.000
Percent Discordant	43.8	Gamma	0.000
Percent Tied	12.5	Tau-a	0.000
Pairs	64	c	0.500

Again the Wald statistics shown in the above table are the squares of the test statistics shown in Table 14.7 on page 396 in the text.

The following SAS commands produce the same output. However, instead of creating the dummy variables, we use the **CLASS** statement followed by the categorical variable **SMOKING**. In parentheses, we use the statement **REF='0'** to specify that the reference category is the value '0' of the **SMOKING** variable referring to heavy smokers.

SAS commands:

```

DATA PFT;
  INPUT PFT LEAD SMOKING COUNT;
DATALINES;
1 1 0 84
1 1 1 75
1 1 2 49
1 1 3 160
1 0 0 16
1 0 1 21
1 0 2 12
1 0 3 33
0 1 0 3
0 1 1 6
0 1 2 6
0 1 3 4
0 0 0 3
0 0 1 2
0 0 2 2
0 0 3 3
;
PROC LOGISTIC DESCENDING;
  WEIGHT COUNT;
  CLASS SMOKING (REF='0');
  MODEL PFT = LEAD SMOKING;
RUN;

```

To obtain the estimated probability also referred to as the predicted probability of having a normal PFT result, we can use the **OUTPUT** statement along with **OUT = NEW** which creates a

temporary data set named NEW containing statistics relevant to the logistic regression model. In this case we use the keyword `PRED=PHAT` to obtain the estimated probabilities stored under the variable name PHAT. In Table 14.10, we display the list of covariate patterns along with the estimated probability of having the outcome of interest obtained from the logistic regression model. Using the following SAS commands, notice that we end up printing 16 covariate patterns; however, there are really only eight distinct covariate patterns. SAS displays 16 covariate patterns because the outcome variable `PFT` is also included in the output creating the duplication of covariate patterns. We have highlighted the first eight covariate patterns to match the predicted probabilities in Table 14.10.

SAS commands:

```

DATA PFT;
  INPUT PFT LEAD SMOKING COUNT;
  SMOK1=0; SMOK2=0; SMOK3=0;
  IF SMOKING = 1 THEN SMOK1=1;
  IF SMOKING = 2 THEN SMOK2=1;
  IF SMOKING = 3 THEN SMOK3=1;
  OUTPUT;
DATALINES;
1 1 0 84
1 1 1 75
1 1 2 49
1 1 3 160
1 0 0 16
1 0 1 21
1 0 2 12
1 0 3 33
0 1 0 3
0 1 1 6
0 1 2 6
0 1 3 4
0 0 0 3
0 0 1 2
0 0 2 2
0 0 3 3
;
PROC LOGISTIC DATA=PFT DESCENDING NOPRINT;
  WEIGHT COUNT;
  MODEL PFT = LEAD SMOK1 SMOK2 SMOK3;
  OUTPUT OUT= NEW PRED=PHAT;
RUN;

DATA PROB;
  SET NEW;
PROC PRINT;
RUN;

```

SAS output:

The SAS System									
Obs	PFT	LEAD	SMOKING	COUNT	SMOK1	SMOK2	SMOK3	_LEVEL_	PHAT
1	1	1	1	84	0	0	0	1	0.95326

2	1	1	2	75	1	0	0	1	0.93855
3	1	1	3	49	0	1	0	1	0.90447
4	1	1	4	160	0	0	1	1	0.97133
5	1	0	1	16	0	0	0	1	0.89824
6	1	0	2	21	1	0	0	1	0.86860
7	1	0	3	12	0	1	0	1	0.80385
8	1	0	4	33	0	0	1	1	0.93616
9	0	1	1	3	0	0	0	1	0.95326
10	0	1	2	6	1	0	0	1	0.93855
11	0	1	3	6	0	1	0	1	0.90447
12	0	1	4	4	0	0	1	1	0.97133
13	0	0	1	3	0	0	0	1	0.89824
14	0	0	2	2	1	0	0	1	0.86860
15	0	0	3	2	0	1	0	1	0.80385
16	0	0	4	3	0	0	1	1	0.93616

Using the following SAS commands which include **PROC SORT** and **PROC FREQ**, we are able to create a contingency table of the predicted probabilities versus **PFT** status.

SAS commands:

```
PROC SORT DATA=PROB;
  BY PHAT;
PROC FREQ DATA=PROB;
  TABLES PHAT*PFT;
  WEIGHT COUNT;
RUN;
```

SAS output:

The SAS System
The FREQ Procedure

Table of PHAT by PFT

PHAT(Estimated Probability)		PFT		Total
Frequency	Percent	0	1	
0.8038504689		2	12	14
		0.42	2.51	2.92
		14.29	85.71	
		6.90	2.67	
0.8685983048		2	21	23
		0.42	4.38	4.80
		8.70	91.30	
		6.90	4.67	
0.8982397962		3	16	19
		0.63	3.34	3.97

	15.79 10.34	84.21 3.56	
0.904474418	6 1.25 10.91 20.69	49 10.23 89.09 10.89	55 11.48
0.9361604396	3 0.63 8.33 10.34	33 6.89 91.67 7.33	36 7.52
0.9385461597	6 1.25 7.41 20.69	75 15.66 92.59 16.67	81 16.91
0.9532579751	3 0.63 3.45 10.34	84 17.54 96.55 18.67	87 18.16
0.9713306341	4 0.84 2.44 13.79	160 33.40 97.56 35.56	164 34.24
Total	29 6.05	450 93.95	479 100.00

We can create an ROC curve from the data provided in Example 14.9 with the SAS commands below. Note we use variables **X** and **Y** just to create a reference line.

SAS commands:

```

DATA ROCCURVE;
  INPUT SENS SPEC X Y;
  ONEMSPEC=1-SPEC;
  DATALINES;
1.000 0.000 0.000 0.000
0.973 0.069 0.100 0.100
0.927 0.138 0.200 0.200
0.891 0.241 0.300 0.300
0.782 0.448 0.400 0.400
0.709 0.552 0.600 0.600
0.542 0.759 0.800 0.800
0.356 0.862 0.900 0.900
0.000 1.000 1.000 1.000
;
GOPTIONS DEVICE= GIF VPOS= 24 HPOS= 75 VSIZE= 5 HSIZE= 6 FTEXT=COMPLEX;
ODS HTML;
ODS GRAPHICS ON;
ODS NOPTITLE;

```

```

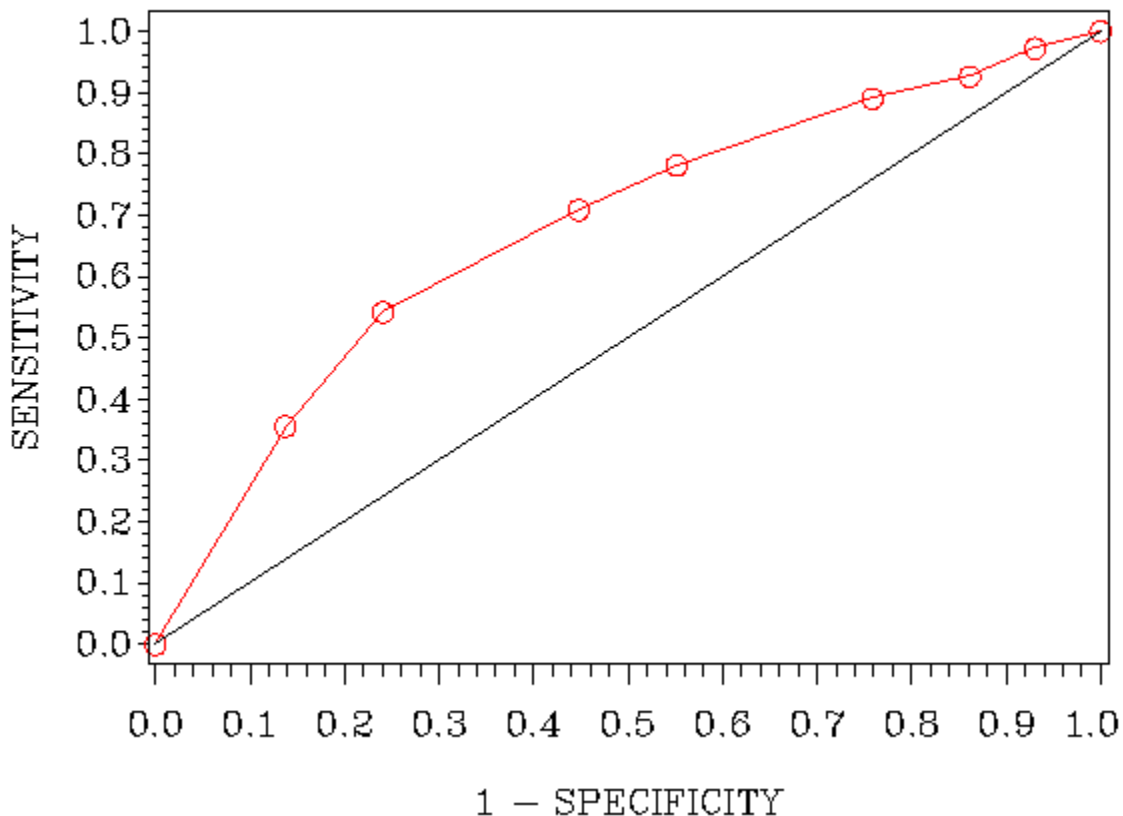
SYMBOL1 V=CIRCLE I=JOIN C=RED;
SYMBOL2 V=NONE I=JOIN C=BLACK;

PROC Gplot;
  PLOT SENS*ONEMSPEC=1 Y*X=2/OVERLAY HAXIS=AXIS1 VAXIS=AXIS2;
  AXIS1 LABEL=('1 - SPECIFICITY');
  AXIS2 LABEL=(A=90 R=0 'SENSITIVITY');
RUN;
QUIT;

ODS GRAPHICS OFF;
ODS HTML CLOSE;

```

SAS output:



Note 14.3 - Pearson and deviance residuals

For simplicity we create a new data set that contains the predicted probability of a normal PFT, the number of individuals with a unique covariate pattern, and the number of individuals with a normal PFT among all those with the same covariate pattern. We will compute the Pearson and deviance residuals and store their values in `P_RESID` and `D_RESID` as shown in the SAS commands below. Because the data are sorted by `PHAT`, the values presented in Table 14.11 are

arranged differently from the output here; however, this is only a matter of rearranging observations.

SAS commands:

```
DATA RESIDUALS;
  INPUT PHAT N Y;
  P_RESID = (Y-N*PHAT)/SQRT(N*PHAT*(1-PHAT));
  D_RESID = SIGN(Y-N*PHAT)*SQRT(2*Y*LOG(Y/(N*PHAT))+2*(N-Y)*LOG((N-Y)/(N*(1-
  PHAT))));
DATALINES;
0.8038505 14 12
0.8685983 23 21
0.8982398 19 16
0.9044744 55 49
0.9361604 36 33
0.9385462 81 75
0.953258 87 84
0.9713306 164 160
;
PROC PRINT;
RUN;
```

SAS output:

The SAS System

Obs	PHAT	N	Y	P_RESID	D_RESID
1	0.80385	14	12	0.50217	0.52198
2	0.86860	23	21	0.63093	0.66672
3	0.89824	19	16	-0.80932	-0.75553
4	0.90447	55	49	-0.34226	-0.33542
5	0.93616	36	33	-0.47844	-0.45828
6	0.93855	81	75	-0.47294	-0.45900
7	0.95326	87	84	0.54171	0.56715
8	0.97133	164	160	0.32839	0.33686

To obtain the sum of squares, we can simply create another temporary data set we called RESID2 along with setting the data set RESIDUALS, and then we create the variables **P_RESID2** and **D_RESID2** which contain the squared values for the Pearson and Deviance residuals. Finally, we use **PROC MEANS** with the option **SUM** to display the sum of squares for both residuals as shown in the SAS commands below. Note that these values are displayed at the bottom of Table 14.11.

SAS commands:

```
DATA RESID2;
  SET RESIDUALS;
  P_RESID2= P_RESID**2;
  D_RESID2= D_RESID**2;
  LABEL P_RESID2 = 'PEARSON RESIDUALS' D_RESID2 = 'DEVIANCE RESIDUALS';
PROC MEANS SUM;
```

```
VAR P_RESID2 D_RESID2;  
RUN;
```

SAS output:

```
The SAS System  
The MEANS Procedure
```

Variable	Label	Sum
P_RESID2	PEARSON RESIDUALS	2.2762465
D_RESID2	DEVIANCE RESIDUALS	2.2561520

Recall from the textbook that the sum of squares of the residuals creates a test statistic that follows a chi-square distribution with degrees of freedom equal to the number of covariate patterns minus the number of parameters in the model plus one (i.e. $8 - (4 + 1) = 8 - 5 = 3$). To calculate the critical value, we can use the SAS function **CINV(p,df)** which allows one to calculate the **p**-th percentile of a chi-square distribution with **df** degrees of freedom as shown in the SAS code below.

SAS commands:

```
DATA CHI_SIGNIF;  
INVCHI= CINV(0.95, 3);  
PROC PRINT;  
RUN;
```

SAS output:

```
The SAS System  
Obs      INVCHI  
  
1         7.81473
```

We could have obtained the [Deviance and Pearson Goodness-of-Fit Statistics](#) by simply using the [AGGREGATE](#) and [SCALE = NONE](#) option. These options are invoked when the number of covariate patterns is small relative to the sample size. Read [14.2.3 Goodness-of-Fit Statistics](#) in the text where we describe using other techniques in cases where you have a continuous covariate in model and the number of covariate patterns is approximately equal to the sample size.

SAS commands:

```
DATA PFT;  
INPUT PFT LEAD SMOKING COUNT;  
SMOK1=0; SMOK2=0; SMOK3=0;  
IF SMOKING = 2 THEN SMOK1=1;  
IF SMOKING = 3 THEN SMOK2=1;  
IF SMOKING = 4 THEN SMOK3=1;  
OUTPUT;
```

```

DATALINES;
1 1 1 84
1 1 2 75
1 1 3 49
1 1 4 160
1 0 1 16
1 0 2 21
1 0 3 12
1 0 4 33
0 1 1 3
0 1 2 6
0 1 3 6
0 1 4 4
0 0 1 3
0 0 2 2
0 0 3 2
0 0 4 3
;
PROC LOGISTIC DATA=PFT DESCENDING;
  WEIGHT COUNT;
  MODEL PFT = LEAD SMOK1 SMOK2 SMOK3/AGGREGATE SCALE=NONE;
RUN;

```

SAS output:

The SAS System
The LOGISTIC Procedure

Model Information

Data Set	WORK.PFT
Response Variable	PFT
Number of Response Levels	2
Weight Variable	COUNT
Model	binary logit
Optimization Technique	Fisher's scoring
Number of Observations Read	16
Number of Observations Used	16
Sum of Weights Read	479
Sum of Weights Used	479

Response Profile

Ordered Value	PFT	Total Frequency	Total Weight
1	1	8	450.00000
2	0	8	29.00000

Probability modeled is PFT=1.

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

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
-----------	-------	----	----------	------------

Deviance	2.2562	3	0.7521	0.5210
Pearson	2.2762	3	0.7587	0.5171

Number of unique profiles: 8

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	220.863	218.949
SC	221.636	222.812
-2 Log L	218.863	208.949

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.9143	4	0.0419
Score	10.8622	4	0.0282
Wald	9.9943	4	0.0405

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.1778	0.5096	18.2643	<.0001
LEAD	1	0.8374	0.4135	4.1024	0.0428
SMOK1	1	-0.2892	0.5618	0.2650	0.6067
SMOK2	1	-0.7673	0.5673	1.8292	0.1762
SMOK3	1	0.5076	0.5719	0.7879	0.3747

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
LEAD	2.310	1.027 5.195
SMOK1	0.749	0.249 2.252
SMOK2	0.464	0.153 1.411
SMOK3	1.661	0.542 5.096

Association of Predicted Probabilities and Observed Responses

Percent Concordant	43.8	Somers' D	0.000
Percent Discordant	43.8	Gamma	0.000
Percent Tied Pairs	12.5	Tau-a	0.000
	64	c	0.500

In Example 14.7, we developed a logistic regression model to predict diabetes (**DIABETES**- '0' if the individual does not have diabetes and '1' if the individual does) using BMI (**ROUNDBMI**- a continuous variable rounded to the nearest whole number), treatment (**TRTMT**- '0' if the individual was on placebo and '1' if the individual was on treatment), and race (**NEWRACE**- '0' for whites and '1' for non-whites) from the DIG200 dataset. The multivariate analysis can be conducted using the SAS commands present below.

SAS commands:

```
PROC IMPORT FILE='C:\DIG200.XLS' OUT=DIG200DATA REPLACE;
RUN;
```



```

DATA DIG200;
  SET DIG200DATA;
  ROUNDBMI = ROUND(BMI,1);
  IF RACE = 1 THEN NEWRACE = 0;
  IF RACE = 2 THEN NEWRACE = 1;
  PROC LOGISTIC DESCENDING;
    MODEL DIABETES = ROUNDBMI TRTMT NEWRACE;
  RUN;

```

SAS output:

The SAS System Model Information

Data Set	WORK.DIG200	
Response Variable	diabetes	diabetes
Number of Response Levels	2	
Model	binary logit	
Optimization Technique	Fisher's scoring	

Number of Observations Read	200
Number of Observations Used	200

Response Profile

Ordered Value	diabetes	Total Frequency
1	1	54
2	0	146

Probability modeled is diabetes=1.

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	235.304	225.833
SC	238.602	239.026
-2 Log L	233.304	217.833

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	15.4708	3	0.0015
Score	15.5953	3	0.0014
Wald	14.2213	3	0.0026

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.9476	0.9142	10.3960	0.0013
ROUNDBMI	1	0.0805	0.0329	5.9868	0.0144

trtmt	1	-0.7963	0.3386	5.5322	0.0187
NEWRACE	1	0.9036	0.4403	4.2124	0.0401

Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits
ROUNDBMI	1.084	1.016 1.156
trtmt	0.451	0.232 0.876
NEWRACE	2.468	1.042 5.850

Association of Predicted Probabilities and Observed Responses

Percent Concordant	65.4	Somers' D	0.332
Percent Discordant	32.2	Gamma	0.340
Percent Tied	2.4	Tau-a	0.132
Pairs	7884	c	0.666

Again the Wald statistics shown here are the squares of those shown in Table 14.9 on page 397 in the text.

Note 14.4 - Ordered Logistic Regression

In Example 14.10 we presented a subsample of data from the NHANES III, Phase II adult study. The subsample contains 1,000 individual observations. First, we use **PROC FREQ** to examine the distribution of the variables: **vit**, **health**, and **educat**. The variable **vit** equals '1' to represent vitamin use and '2' to represent no use. The variable **health** equals '1' to represent excellent health, '2' to represent very good health, '3' to represent good health, '4' to represent fair health, and '5' to represent poor health. The variable **educat** represents years of education.

SAS commands:

```
PROC IMPORT FILE='C:\NHANES_III_PH_II.XLS' OUT=NHANESDATA REPLACE;
RUN;

DATA NHANES3;
  SET NHANESDATA;
PROC FREQ;
  TABLES VIT HEALTH EDUCAT;
RUN;
```

SAS output:

The SAS System

The FREQ Procedure

vit

vit	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	430	43.00	430	43.00

2	570	57.00	1000	100.00
---	-----	-------	------	--------

health

health	Frequency	Percent	Cumulative Frequency	Cumulative Percent
1	229	22.90	229	22.90
2	299	29.90	528	52.80
3	314	31.40	842	84.20
4	129	12.90	971	97.10
5	29	2.90	1000	100.00

educat

educat	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	5	0.50	5	0.50
1	1	0.10	6	0.60
2	6	0.60	12	1.20
3	8	0.80	20	2.00
4	7	0.70	27	2.70
5	8	0.80	35	3.50
6	18	1.80	53	5.30
7	14	1.40	67	6.70
8	43	4.30	110	11.00
9	43	4.30	153	15.30
10	56	5.60	209	20.90
11	45	4.50	254	25.40
12	345	34.50	599	59.90
13	69	6.90	668	66.80
14	84	8.40	752	75.20
15	42	4.20	794	79.40
16	100	10.00	894	89.40
17	94	9.40	988	98.80
88	7	0.70	995	99.50
99	5	0.50	1000	100.00

In Table 14.12, there are only four categories for perceived health because categories 4 and 5 were collapsed. For the education variable missing or nonresponses were eliminated. Next we provide the SAS commands that produce the output for Model I in Example 14.10.

SAS commands:

```

PROC IMPORT FILE='C:\NHANES_III_PH_II.XLS' OUT=NHANESDATA REPLACE;
RUN;

DATA NHANES3;
  SET NHANESDATA;
  NEWVIT = 0;
  IF VIT = 1 THEN NEWVIT = 1;
  NEWHEALTH = HEALTH;
  IF HEALTH = 5 THEN NEWHEALTH = 4;

```

```

IF EDUCAT = 88 OR EDUCAT = 99 THEN DELETE;
PROC LOGISTIC;
  MODEL NEWHEALTH = NEWVIT;
RUN;

```

SAS output:

The SAS System
The LOGISTIC Procedure

Model Information

Data Set	WORK.NHANES3
Response Variable	NEWHEALTH
Number of Response Levels	4
Model	cumulative logit
Optimization Technique	Fisher's scoring
Number of Observations Read	988
Number of Observations Used	988

Response Profile

Ordered Value	NEWHEALTH	Total Frequency
1	1	227
2	2	299
3	3	309
4	4	153

Probabilities modeled are cumulated over the lower Ordered Values.

Model Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Score Test for the Proportional Odds Assumption

Chi-Square	DF	Pr > ChiSq
1.3486	2	0.5095

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	2677.558	2673.553
SC	2692.245	2693.136
-2 Log L	2671.558	2665.553

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	6.0043	1	0.0143

Score	6.0159	1	0.0142
Wald	5.9641	1	0.0146

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept 1	1	-1.3384	0.0923	210.0621	<.0001
Intercept 2	1	0.00634	0.0808	0.0062	0.9374
Intercept 3	1	1.5808	0.0993	253.6021	<.0001
NEWVIT	1	0.2834	0.1160	5.9641	0.0146

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
NEWVIT	1.328	1.058	1.667

Association of Predicted Probabilities and Observed Responses

Percent Concordant	27.4	Somers' D	0.058
Percent Discordant	21.6	Gamma	0.119
Percent Tied	51.0	Tau-a	0.043
Pairs	358162	c	0.529

Note 14.5 - Conditional Logistic Regression

In Example 14.11 we presented data on twenty-seven cases and controls matched on age, sex, and race. We can use **PROC PHREG** to conduct conditional logistic regression. Note that **PROC PHREG** does not have a descending option like **PROC LOGISTIC**, and the smaller value is considered the event and the larger value is a non-event thus we have created the variable **NEWCVD** out of **CVD** in the SAS commands below. Next to **STRATA** is where the variable **SET** that pertains to the matching is located. The SAS out refers to Model I in Example 14.11.

SAS commands:

```
PROC IMPORT FILE='C:\TABLE14-14.XLS' OUT=MATCHDATA REPLACE;
RUN;

DATA MATCH;
  SET MATCHDATA;
NEWCVD = 0;
IF CVD = 0 THEN NEWCVD = 1;
PROC PHREG;
  MODEL NEWCVD=PREVMI;
  STRATA SET;
RUN;
```

SAS output:

The SAS System
The PHREG Procedure

Model Information

Data Set WORK.MATCH
Dependent Variable NEWCVD
Ties Handling BRESLOW

Number of Observations Read 54
Number of Observations Used 54

Summary of the Number of Event and Censored Values

Stratum	set	Total	Event	Censored	Percent Censored
1	1	2	2	0	0.00
2	2	2	2	0	0.00
3	3	2	2	0	0.00
4	4	2	2	0	0.00
5	5	2	2	0	0.00
6	6	2	2	0	0.00
7	7	2	2	0	0.00
8	8	2	2	0	0.00
9	9	2	2	0	0.00
10	10	2	2	0	0.00
11	11	2	2	0	0.00
12	12	2	2	0	0.00
13	13	2	2	0	0.00
14	14	2	2	0	0.00
15	15	2	2	0	0.00
16	16	2	2	0	0.00
17	17	2	2	0	0.00
18	18	2	2	0	0.00
19	19	2	2	0	0.00
20	20	2	2	0	0.00
21	21	2	2	0	0.00
22	22	2	2	0	0.00
23	23	2	2	0	0.00
24	24	2	2	0	0.00
25	25	2	2	0	0.00
26	26	2	2	0	0.00
27	27	2	2	0	0.00

Total		54	54	0	0.00

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	37.430	36.419
AIC	37.430	38.419
SBC	37.430	40.408

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	1.0107	1	0.3147
Score	1.0000	1	0.3173
Wald	0.9785	1	0.3226

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
prevmi	1	-0.51081	0.51640	0.9785	0.3226	0.600	prevmi

Note 14.6 - Cox Proportional Hazards Regression

In the DIG200 data set, the variable **TRTMT** equals '0' to represent placebo and '1' to represent the treatment. The variable **AGE** represents age in years and **BMI** represents body mass index in kg/m². The variable **DEATH** equals '0' to represent a censored value and '1' to represent a death, and the variable **DEATHDAYS** represents the number of days from randomization to the occurrence of a censor or a death.

To perform Cox's proportional hazards regression, we use the SAS procedure **PROC PHREG**. In the **MODEL** statement, the variable containing information on time to event (**DEATHDAYS**) is specified followed by an * and then the censoring indicator (**DEATH**). The independent variable (**TRTMT**) is specified after the equal sign. The SAS commands below provide the output found for Model I in Table 14.18.

SAS commands:

```
PROC IMPORT FILE='C:\DIG200.XLS' OUT=DIG200DATA REPLACE;
RUN;

DATA DIG200;
  SET DIG200DATA;
PROC PHREG;
  MODEL DEATHDAY*DEATH(0)=TRTMT;
RUN;
```

SAS output:

The SAS System Model Information

Data Set	WORK.DIG200	
Dependent Variable	deathday	deathday
Censoring Variable	death	death
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Number of Observations Read	200
Number of Observations Used	200

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
200	72	128	64.00

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	708.339	707.622
AIC	708.339	709.622
SBC	708.339	711.899

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.7169	1	0.3972
Score	0.7151	1	0.3977
Wald	0.7128	1	0.3985

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
trtmt	1	-0.20071	0.23773	0.7128	0.3985	0.818	trtmt

For Model II, we added AGE in the model along with TRTMT.

SAS commands:


```

PROC IMPORT FILE='C:\DIG200.XLS' OUT=DIG200DATA REPLACE;
RUN;

DATA DIG200;
  SET DIG200DATA;
PROC PHREG;
  MODEL DEATHDAY*DEATH(0)=TRTMT AGE;
RUN;

```

SAS output:

The SAS System
Model Information

Data Set	WORK.DIG200	
Dependent Variable	deathday	deathday
Censoring Variable	death	death
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Number of Observations Read	200
Number of Observations Used	200

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
200	72	128	64.00

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	708.339	707.524
AIC	708.339	711.524
SBC	708.339	716.078

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.8151	2	0.6653
Score	0.8143	2	0.6655
Wald	0.8119	2	0.6664

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
trtmt	1	-0.20147	0.23777	0.7180	0.3968	0.818	trtmt
age	1	-0.00329	0.01048	0.0985	0.7536	0.997	age

For Model III, we added BMI along with TRTMT and AGE.

SAS commands:

```
PROC IMPORT FILE='C:\DIG200.XLS' OUT=DIG200DATA REPLACE;
RUN;
```

```
DATA DIG200;
  SET DIG200DATA;
PROC PHREG;
  MODEL DEATHDAY*DEATH(0)=TRTMT AGE BMI;
RUN;
```

SAS output:

The SAS System Model Information

Data Set	WORK.DIG200	
Dependent Variable	deathday	deathday
Censoring Variable	death	death
Censoring Value(s)	0	
Ties Handling	BRESLOW	

Number of Observations Read	200
Number of Observations Used	200

Summary of the Number of Event and Censored Values

Total	Event	Censored	Percent Censored
200	72	128	64.00

Convergence Status

Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics

Criterion	Without Covariates	With Covariates
-2 LOG L	708.339	707.408
AIC	708.339	713.408
SBC	708.339	720.238

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.9314	3	0.8178
Score	0.9276	3	0.8188
Wald	0.9252	3	0.8193

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Variable Label
trtmt	1	-0.19799	0.23800	0.6920	0.4055	0.820	trtmt
age	1	-0.00308	0.01056	0.0850	0.7707	0.997	age
bmi	1	-0.00846	0.02495	0.1149	0.7346	0.992	bmi