

Models for Binary Outcomes

Introduction

The simple or binary response (for example, success or failure) analysis models the relationship between a binary response variable and one or more explanatory variables. For a binary response variable Y , it assumes:

$$g(p) = \beta'x$$

where p is $\text{Prob}(Y=y_1)$ for y_1 as one of two ordered levels of Y , β is the parameter vector, x is the vector of explanatory variables, and g is a function of which p is assumed to be linearly related to the explanatory variables.

The binary response model shares a common feature with a more general class of linear models that a function $g=g(\mu)$ of the mean μ of the dependent variable is assumed to be linearly related to the explanatory variables. The function $g(\mu)$, often referred as the link function, provides the link between the random or stochastic component and the systematic or deterministic component of the response variable. For the binary response model, logistic and probit regression techniques are often employed among all others.

Logistic Regression

For a binary response variable Y , the logistic regression has the form:

$$\text{logit}(p) \equiv \log \frac{p}{1-p} = \beta'x$$

or equivalently,

$$p = \frac{\exp(\beta'x)}{1+\exp(\beta'x)}$$

The logistic regression models the logit transformation of the i th observation's event probability, p_i , as a linear function of the explanatory variables in the vector x_i . The logistic regression model uses the logit as the link function.

Logistic Regression with SAS

LOGISTIC Procedure

Suppose the response variable Y is 0 or 1 binary (This is not a limitation. The values can be either numeric or character as long as they are dichotomous), and X_1 and X_2 are two regressors of interest. To fit a logistic regression, you can use:

```
proc logistic; model y=x1 x2; run;
```

SAS PROC LOGISTIC models the probability of $Y=0$ by default. In other words, SAS chooses the smaller value to estimate its probability. One way to change the default setting in order to model the probability of $Y=1$ in SAS is to specify the DESCENDING option on the PROC LOGISTIC statement. That is, use:

```
proc logistic descending;
```

Example 1: SAS Logistic Regression in PROC LOGISTIC (individual data)

The following data are from Cox (Cox, D. R., 1970. *The Analysis of Binary Data*, London, Methuen, p. 86). At the specified time (T) of heating, a number of ingots are tested for some temperature settings and whether an ingot is ready or not (S) for rolling is recorded. S=0 means not ready and S=1 means ready. You want to know if the time of heating affects whether an ingot is ready or not for rolling.

	T	S
1	7	1
2	7	1
.	.	.
55	7	1
1	14	0
2	14	0
3	14	1
4	14	1
.	.	.
157	14	1
1	27	0
2	27	0
.	.	.
7	27	0
8	27	1
9	27	1
.	.	.
159	27	1
1	51	0
2	51	0
3	51	0
4	51	1
.	.	.
16	51	1

With this data set INGOT, you can use:

```
proc logistic data=ingot;
model s=t;
run;
```

As a result, you will have the following SAS output:

Sample Program: Logistic Regression

The LOGISTIC Procedure

```
Data Set: WORK.INGOT
Response Variable: S
Response Levels: 2
Number of Observations: 387
```

Link Function: Logit

Response Profile

Ordered Value	S	Count
1	0	12
2	1	375

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	108.988	99.375	.
SC	112.947	107.291	.
-2 LOG L Score	106.988	95.375	11.614 with 1 DF (p=0.0007) 15.100 with 1 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-5.4152	0.7275	55.4005	0.0001	.	.
T	1	0.0807	0.0224	13.0290	0.0003	0.442056	1.084

Association of Predicted Probabilities and Observed Responses

Concordant = 59.2%	Somers' D = 0.499
Discordant = 9.4%	Gamma = 0.727
Tied = 31.4%	Tau-a = 0.030
(4500 pairs)	c = 0.749

The result shows that the estimated logit is

$$\log \frac{p}{1-p} = -5.4152 + 0.0807 * T$$

where p is the probability of having an ingot not ready for rolling. The slope coefficient 0.0807 represents the change in log odds for a one unit increase in T (time of heating). Its odds ratio 1.084 is the ratio of odds for a one unit change in T. The odds ratio can be computed by exponentiating the log odds, i.e., $\exp(\log \text{odds})$, which is $\exp(0.0807)=1.084$ in this example.

If you had used the DESCENDING option:

```
proc logistic descending;
model s=t;
```

```
run;
```

it would have yielded the following estimated logit:

$$\log \frac{p}{1-p} = 5.4152 - 0.0807 * T \text{ with } T\text{'s odds ratio } 0.922$$

where p is the probability of having an ingot ready for rolling.

You may have the same data set arranged in the following frequency format:

T	S	F
7	1	55
14	0	2
14	1	155
27	0	7
27	1	152
51	0	3
51	1	13

In this case, to have the same output as above, you can use the syntax:

```
proc logistic;
freq f;
model s=t;
run;
```

The LOGISTIC procedure also allows the input of binary response data that are grouped so that you can use:

```
proc logistic;
model r/n=x1 x2;
run;
```

where N represents the number of trials and R represents the number of events.

Example 2: SAS Logistic Regression in PROC LOGISTIC (grouped data)

The data set described in the previous example can be arranged in a different way. At the specified time(T) of heating, the number of ingots (N) tested and the number (R) not ready for rolling can be recorded. Now you have:

T	R	N
7	0	55
14	2	157
27	7	159
51	3	16

With this data set INGOT2, you can use:

```
proc logistic data=ingot2;
model r/n=t;
run;
```

The SAS output will be:

Sample Program: Logistic Regression

The LOGISTIC Procedure

Data Set: WORK.INGOT2
 Response Variable (Events): R
 Response Variable (Trials): N
 Number of Observations: 4
 Link Function: Logit

Response Profile

Ordered Value	Binary Outcome	Count
1	EVENT	12
2	NO EVENT	375

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	108.988	99.375	.
SC	112.947	107.291	.
-2 LOG L Score	106.988	95.375	11.614 with 1 DF (p=0.0007)
	.	.	15.100 with 1 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-5.4152	0.7275	55.4005	0.0001	.	.
T	1	0.0807	0.0224	13.0290	0.0003	0.442056	1.084

Association of Predicted Probabilities and Observed Responses

Concordant = 59.2%	Somers' D = 0.499
Discordant = 9.4%	Gamma = 0.727
Tied = 31.4%	Tau-a = 0.030
(4500 pairs)	c = 0.749

Sometimes you may be interested in the change in log odds, and thus the corresponding change in odds ratio for some amount other than one unit change in the explanatory variable. In this case, you can customize your own odds calculation. You can use the UNITS option:

```
proc logistic;
model y=x1 x2;
units x1=list;
run;
```

where *list* represents a list of units in change that are of interest for the variable X1. Each unit of change in a list has

one of the following forms:

```
number
SD or -SD
number*SD
```

where *number* is any non-zero number and SD is the sample standard deviation of the corresponding independent variable X1.

Example 3: Customized Odds Computation

Using the same data set in Example 2, if you use:

```
proc logistic data=ingot2;
model r/n=t;
units t=10 -10 sd 2*sd;
run;
```

you will have the following result in addition to the output in Example 2:

Conditional Odds Ratio

Variable	Unit	Odds Ratio
T	10.0000	2.241
T	-10.0000	0.446
T	9.9361	2.230
T	19.8721	4.971

In this example, you calculated four different odd ratio, each corresponding to change in 10 unit increase, 10 unit decrease, 1 standard deviation increase, and 2 standard deviation increase in T, respectively.

From the SAS PROC LOGISTIC output, you can also obtain predicted probability values. Suppose you want to know the predicted probabilities of having an ingot not ready for rolling ($Y=0$) at each level of time of heating in the data set from Example 2. The predicted probability, p , can be computed from the formula:

$$p = \frac{\exp(\beta'x)}{1 + \exp(\beta'x)}$$

Thus, for example, at $T=7$,

$$p = \frac{\exp(-5.4152 + 0.0807*7)}{1 + \exp(-5.4152 + 0.0807*7)} = 0.00777$$

This computation can be easily obtained as a part of the SAS output by using the OUTPUT statement and PRINT procedure:

```
proc logistic;
model r/n=x1 x2;
output out=filename predicted=varname;
run;
proc print data=filename;
run;
```

where *filename* is the output data set name and *varname* is the variable name for predicted probabilities. The SAS output will show all the predicted probabilities for all observation points.

However, if you need to know the predicted probabilities at some levels of explanatory variables other than levels the data set provides, you need to do something different. You need to create a new SAS data set with missing values for the response variable. Then you merge the new data with the original data and run the logistic regression using the merged data set. Because the new data set has missing values for the response variable, they do not affect the model fit. But the predicted probabilities will be also calculated for the new observations.

Example 4: Predicted Probability Computation

Using the data in Example 2, if you use:

```
proc logistic data=ingot2;
model r/n=t;
output out=prob predicted=phat;
run;
proc print data=prob;
run;
```

you will have the following additional result to the output in Example 2:

Sample Program: Logistic Regression

OBS	T	R	N	PHAT
1	7	0	55	0.00777
2	14	2	157	0.01358
3	27	7	159	0.03782
4	51	3	16	0.21422

Now suppose you want to compute the predicted probabilities at T=10,20,30,40,50, and 60. You can use the following syntax:

```
data ingot2;
input t r n;
cards;
  7 0 55
 14 2 157
 27 7 159
 51 3 16
;
data new;
input t @@;
r=.;
n=.;
cards;
10 20 30 40 50 60
;
data merged;
set ingot2 new;
run;
proc logistic data=merged;
```

```

model r/n=t;
output out=prob predicted=phat;
run;
proc print data=prob;
run;

```

You will have the following additional output to show the predicted probability at each level of T of interest:

Sample Program: Logistic Regression

OBS	T	R	N	PHAT
1	7	0	55	0.00777
2	14	2	157	0.01358
3	27	7	159	0.03782
4	51	3	16	0.21422
5	10	.	.	0.00987
6	20	.	.	0.02185
7	30	.	.	0.04768
8	40	.	.	0.10089
9	50	.	.	0.20095
10	60	.	.	0.36045

PROBIT Procedure

You can even use the PROC PROBIT to fit a logistic regression by specifying LOGISTIC as the cumulative distribution type in the MODEL statement. To fit a logistic regression model, use:

```

proc probit;
class y;
model y=x1 x2 / d=logistic;
run;

```

or

```

proc probit;
model r/n=x1 x2 / d=logistic;
run;

```

depending on your data set. If a single response variable is given in the MODEL statement, it must be listed in a CLASS statement. Unlike the PROC LOGISTIC, the PROC PROBIT is capable of dealing with categorical variables as regressors as shown in the following syntax:

```

proc probit;
class x2;
model r/n=x1 x2 / d=logistic;
run;

```

where X2 is a categorical regressor.

Example 5: SAS Logistic Regression in PROC PROBIT

Using the data in Example 2, you may use:

```
proc probit data=ingot2;
model r/n=t / d=logistic;
run;
```

The resulting SAS output will be:

Sample Program: Logistic Regression

Probit Procedure

```
Data Set           =WORK.INGOT2
Dependent Variable=R
Dependent Variable=N
Number of Observations=   4
Number of Events      =   12      Number of Trials =   387
```

Log Likelihood for LOGISTIC -47.68727905

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-5.4151721	0.727541	55.40004	0.0001	Intercept
T	1	0.08069587	0.022356	13.02885	0.0003	

Probit Model in Terms of Tolerance Distribution

	MU	SIGMA
	67.10594	12.39221

Estimated Covariance Matrix for Tolerance Parameters

	MU	SIGMA
MU	121.813302	35.655509
SIGMA	35.655509	11.786672

GENMOD Procedure

The GENMOD procedure fits generalized linear models (Nelder and Wedderburn, 1972, "Generalized Linear Models," *Journal of the Royal Statistical Society A*, 135, pp. 370-384). Logistic regression can be modeled as a class of generalized linear model where the response probability distribution function is binomial and the link function is logit. To use PROC GENMOD for a logistic regression, you can use:

```
proc genmod;
model y=x1 x2 / dist=binomial link=logit;
run;
```

or

```
proc genmod;
model r/n=x1 x2 / dist=binomial link=logit;
run;
```

Example 6: SAS Logistic Regression in PROC GENMOD

Using the data in Example 2, you may use:

```
proc genmod data=ingot2;
model r/n=t / dist=binomial link=logit;
run;
```

You will have the following SAS output:

Sample Program: Logistic Regression

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.INGOT2
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	R
Dependent Variable	N
Observations Used	4
Number Of Events	12
Number Of Trials	387

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	1.0962	0.5481
Scaled Deviance	2	1.0962	0.5481
Pearson Chi-Square	2	0.6749	0.3374
Scaled Pearson X2	2	0.6749	0.3374
Log Likelihood	.	-47.6873	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-5.4152	0.7275	55.4000	0.0001
T	1	0.0807	0.0224	13.0289	0.0003
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

PROC GENMOD is especially convenient when you need to use categorical or class variables as regressors. In this case, you can use:

```
proc genmod;
class x2;
model y=x1 x2 / dist=binomial link=logit;
run;
```

where X2 is a categorical regressor.

Example 7: SAS Logistic Regression in PROC GENMOD (categorical regressors)

This example is excerpted from a SAS manual (SAS, 1996, *SAS/STAT Software Changes and Enhancements through Release 6.11*, pp. 279-284). In an experiment comparing the effects of five different drugs, each drug was tested on a number of different's subjects. The outcome of each experiment was the presence or absence of a positive response in a subject. The following data represent the number of responses R in the N subjects for the five different drugs, labeled A through E. The response is measured for different levels of a continuous covariate X for each drug. The drug type and the covariate X are explanatory variables in this experiment. The number of response R is modeled as a binomial random variable for each combination of the explanatory variable values, with the binomial number of trials parameter equal to the number of subjects N and the binomial probability equal to the probability of a response. The following DATA step creates the data set DRUG:

```
data drug;
input drug$ x r n;
cards;
A .1 1 10
A .23 2 12
A .67 1 9
B .2 3 13
B .3 4 15
B .45 5 16
B .78 5 13
C .04 0 10
C .15 0 11
C .56 1 12
C .7 2 12
D .34 5 10
D .6 5 9
D .7 8 10
E .2 12 20
E .34 15 20
E .56 13 15
E .8 17 20
;
```

A logistic regression for these data is a generalized linear model with response equal to the binomial proportion R/N. PROC GENMOD can be used as follows:

```
proc genmod data=drug;
class drug;
```

```

model r/n=x drug / dist=binomial link=logit;
run;

```

You will have the SAS output:

Sample Program: Logistic Regression

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.DRUG
Distribution	BINOMIAL
Link Function	LOGIT
Dependent Variable	R
Dependent Variable	N
Observations Used	18
Number Of Events	99
Number Of Trials	237

Class Level Information

Class	Levels	Values
DRUG	5	A B C D E

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	12	5.2751	0.4396
Scaled Deviance	12	5.2751	0.4396
Pearson Chi-Square	12	4.5133	0.3761
Scaled Pearson X2	12	4.5133	0.3761
Log Likelihood	.	-114.7732	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	0.2792	0.4196	0.4430	0.5057
X	1	1.9794	0.7660	6.6770	0.0098
DRUG	A	-2.8955	0.6092	22.5894	0.0001
DRUG	B	-2.0162	0.4052	24.7628	0.0001
DRUG	C	-3.7952	0.6655	32.5258	0.0001
DRUG	D	-0.8548	0.4838	3.1218	0.0773
DRUG	E	0.0000	0.0000	.	.

SCALE	0	1.0000	0.0000	.	.
-------	---	--------	--------	---	---

NOTE: The scale parameter was held fixed.

In this example, PROC GENMOD automatically generates five dummy variables for each value of the class variable DRUG. Therefore, the same result could be obtained without using PROC GENMOD, but employing PROC LOGISTIC:

```
if drug='A' then drugdum1=1; else drugdum1=0;
if drug='B' then drugdum2=1; else drugdum2=0;
if drug='C' then drugdum3=1; else drugdum3=0;
if drug='D' then drugdum4=1; else drugdum4=0;
if drug='E' then drugdum5=1; else drugdum5=0;
proc logistic data=drug2;
model r/n=x drugdum1 drugdum2 drugdum3 drugdum4 drugdum5;
run;
```

where the first five lines must be included in the DATA step to create a new data set DRUG2. Notice that one of the five dummy variables is redundant.

The resulting output will be:

Sample Program: Logistic Regression

The LOGISTIC Procedure

Data Set: WORK.DRUG2
 Response Variable (Events): R
 Response Variable (Trials): N
 Number of Observations: 18
 Link Function: Logit

Response Profile

Ordered Value	Binary Outcome	Count
1	EVENT	99
2	NO EVENT	138

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	324.105	241.546	.
SC	327.573	262.355	.
-2 LOG L Score	322.105	229.546	92.558 with 5 DF (p=0.0001)
	.	.	82.029 with 5 DF (p=0.0001)

NOTE: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

$$\text{DRUGDUM5} = 1 * \text{INTERCPT} - 1 * \text{DRUGDUM1} - 1 * \text{DRUGDUM2} - 1 * \text{DRUGDUM3} - 1 * \text{DRUGDUM4}$$

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	0.2792	0.4196	0.4430	0.5057	.	.
X	1	1.9794	0.7660	6.6772	0.0098	0.259740	7.238
DRUGDUM1	1	-2.8955	0.6092	22.5895	0.0001	-0.539417	0.055
DRUGDUM2	1	-2.0162	0.4052	24.7628	0.0001	-0.476082	0.133
DRUGDUM3	1	-3.7952	0.6654	32.5336	0.0001	-0.822382	0.022
DRUGDUM4	1	-0.8548	0.4838	3.1218	0.0773	-0.154773	0.425
DRUGDUM5	0	0

Association of Predicted Probabilities and Observed Responses

Concordant = 82.3%	Somers' D = 0.686
Discordant = 13.7%	Gamma = 0.714
Tied = 4.0%	Tau-a = 0.335
(13662 pairs)	c = 0.843

CATMOD Procedure

SAS CATMOD (CATEGorical data MODELing) procedure fits linear models to functions of response frequencies and can be used for logistic regression. The basic syntax is:

```
proc catmod;
direct x1;
response logits;
model y=x1 x2;
run;
```

where X1 is a continuous quantitative variable and X2 is a categorical variable. You must specify your continuous regressors in the DIRECT statement. Because the CATMOD procedure is mainly designed for the analysis of categorical data, it is not recommended for use with a continuous regressor with a large number of unique values.

Example 8: SAS Logistic Regression in PROC CATMOD

Using the data in Example 1, if you use:

```
proc catmod data=ingot;
direct t;
response logits;
model s=t;
run;
```

you will see the result:

Sample Program: Logistic Regression

CATMOD PROCEDURE

Response: S	Response Levels (R)=	2
Weight Variable: None	Populations (S)=	4
Data Set: INGOT	Total Frequency (N)=	387
Frequency Missing: 0	Observations (Obs)=	387

POPULATION PROFILES

Sample	T	Sample Size
1	7	55
2	14	157
3	27	159
4	51	16

RESPONSE PROFILES

Response	S
1	0
2	1

MAXIMUM-LIKELIHOOD ANALYSIS

Iteration	Sub Iteration	-2 Log Likelihood	Convergence Criterion	Parameter Estimates	
				1	2
0	0	536.49592	1.0000	0	0
1	0	152.59147	0.7156	-2.1503	0.0138
2	0	106.76794	0.3003	-3.5040	0.0361
3	0	96.711696	0.0942	-4.6746	0.0633
4	0	95.411914	0.0134	-5.2884	0.0779
5	0	95.374601	0.000391	-5.4109	0.0806
6	0	95.374558	4.5308E-7	-5.4152	0.0807
7	0	95.374558	6.605E-13	-5.4152	0.0807

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
-----	-----	-----	-----
INTERCEPT	1	55.40	0.0000
T	1	13.03	0.0003
LIKELIHOOD RATIO	2	1.10	0.5781

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-5.4152	0.7275	55.40	0.0000
T	2	0.0807	0.0224	13.03	0.0003

LOGISTIC REGRESSION Procedure

Unlike in SAS, the SPSS procedure LOGISTIC REGRESSION models the probability of $Y=1$ or Y 's higher sorted value. Suppose the response variable Y is 0 or 1 binary (This is not a limitation for SPSS either. The values can be either numeric or character as long as they are dichotomous), and $X1$ and $X2$ are two regressors of interest. To run a logistic regression, use:

```
logistic regression var=y with x1 x2.
```

Example 9: SPSS Logistic Regression in LOGISTIC REGRESSION procedure (individual data)

Using the data in Example 1, you can use:

```
logistic regression var=s with t.
```

You will have the SPSS output:

L O G I S T I C R E G R E S S I O N

```
Total number of cases:      387 (Unweighted)
Number of selected cases:    387
Number of unselected cases:  0

Number of selected cases:      387
Number rejected because of missing data:  0
Number of cases included in the analysis: 387
```

Dependent Variable Encoding:

Original Value	Internal Value
.00	0
1.00	1

Dependent Variable.. S

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 106.98843

* Constant is included in the model.

Beginning Block Number 1. Method: Enter

Variable(s) Entered on Step Number

1.. T

Estimation terminated at iteration number 6 because Log Likelihood decreased by less than .01 percent.

-2 Log Likelihood 95.375
 Goodness of Fit 346.446

	Chi-Square	df	Significance
Model Chi-Square	11.614	1	.0007
Improvement	11.614	1	.0007

Classification Table for S

Observed		Predicted		Percent Correct
		.00	1.00	
		0	1	
.00	0	0	12	.00%
1.00	1	0	375	100.00%
Overall				96.90%

----- Variables in the Equation -----

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
T	-.0807	.0224	13.0289	1	.0003	-.3211	.9225
Constant	5.4152	.7275	55.4000	1	.0000		

The output shows that the estimated logit is

$$\log \frac{p}{1-p} = 5.4152 - 0.0807 * T$$

where p is the probability of having an ingot ready for rolling. This is the same result as with the use of the DESCENDING option in SAS PROC LOGISTIC.

Probit Regression

Probit regression can be employed as an alternative to the logistic regression in binary response models. For a binary response variable Y, the probit regression model has the form:

$$\Phi^{-1}(p) = \beta'x$$

or equivalently,

$$p = \Phi(\beta'x)$$

where Φ^{-1} is the inverse of the cumulative standard normal distribution function, often referred as probit or normit, and Φ is the cumulative standard normal distribution function.

The probit regression model can be viewed also as a special case of the generalized linear model whose link function is probit.

Probit Regression with SAS

LOGISTIC Procedure

The SAS LOGISTIC procedure can be also used for a probit regression. To fit a probit regression use the LINK=NORMIT (or PROBIT) option:

```
proc logistic;
model y=x1 x2 / link=normit;
run;
```

Example 11: SAS Probit Regression in PROC LOGISTIC

Using the data in Example 1, you can use:

```
proc logistic data=ingot;
model s=t / link=normit;
run;
```

You will have the SAS output:

Sample Program: Probit Regression

The LOGISTIC Procedure

```
Data Set: WORK.INGOT
Response Variable: S
Response Levels: 2
Number of Observations: 387
Link Function: Normit
```

Response Profile

Ordered Value	S	Count
1	0	12
2	1	375

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	108.988	99.018	.
SC	112.947	106.934	.

Categorical Analysis - Part 1

-2 LOG L	106.988	95.018	11.971 with 1 DF (p=0.0005)
Score	.	.	15.100 with 1 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCPT	1	-2.8004	0.3284	72.7050	0.0001	.
T	1	0.0391	0.0113	11.9525	0.0005	0.388259

Association of Predicted Probabilities and Observed Responses

Concordant = 59.2%	Somers' D = 0.499
Discordant = 9.4%	Gamma = 0.727
Tied = 31.4%	Tau-a = 0.030
(4500 pairs)	c = 0.749

PROBIT Procedure

You can use the PROC PROBIT to fit a probit model. The basic syntax you can use is:

```
proc probit;  
class y;  
model y=x1 x2;  
run;
```

or

```
proc probit;  
model r/n=x1 x2;  
run;
```

or

```
proc probit;  
class x2;  
model r/n=x1 x2;  
run;
```

depending on the nature of the data set.

Example 12: SAS Probit Regression in PROC PROBIT

Using the data in Example 1, you can use:

```
proc probit data=ingot;  
class s;  
model s=t;  
run;
```

You will have the following SAS output:

Sample Program: Probit Regression

Probit Procedure
Class Level Information

Class	Levels	Values
S	2	0 1

Number of observations used = 387

Probit Procedure

Data Set =WORK.INGOT
Dependent Variable=S

Weighted Frequency Counts for the Ordered Response Categories

Level	Count
0	12
1	375

Log Likelihood for NORMAL -47.5087804

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-2.8003508	0.331621	71.30839	0.0001	Intercept
T	1	0.0390757	0.011425	11.69807	0.0006	

Probit Model in Terms of Tolerance Distribution

MU	SIGMA
71.66476	25.59135

Estimated Covariance Matrix for Tolerance Parameters

	MU	SIGMA
MU	186.336614	98.799500
SIGMA	98.799500	55.985053

Example 13: SAS Probit Regression in PROC PROBIT (categorical regressors)

Using the data in Example 7, if you use:

```
proc probit data=drug;
class drug;
model r/n=x drug;
run;
```

you will have the result:

Sample Program: Probit Regression

Probit Procedure
Class Level Information

Class	Levels	Values
DRUG	5	A B C D E

Number of observations used = 18

Probit Procedure

Data Set =WORK.DRUG
 Dependent Variable=R
 Dependent Variable=N
 Number of Observations= 18
 Number of Events = 99 Number of Trials = 237

Log Likelihood for NORMAL -114.6516555

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	0.19031335	0.24926	0.582954	0.4452	Intercept
X	1	1.15885442	0.438333	6.989539	0.0082	
DRUG	4			64.33502	0.0001	
	1	-1.7087998	0.331686	26.5416	0.0001	A
	1	-1.2286831	0.239099	26.40741	0.0001	B
	1	-2.2309708	0.343196	42.2574	0.0001	C
	1	-0.5079719	0.291889	3.028612	0.0818	D
	0	0	0	.	.	E

SAS PROC PROBIT models the probability of $Y=0$ or of Y 's lower sorted value by default. This default can be altered by using the ORDER option in the PROC PROBIT statement. For example,

```
proc probit order=freq;
```

specifies the sorting order for the levels of the classification variables (specified in the CLASS statement) in a descending frequency count; levels with the most observations come first in the order.

Example 14: Altering Order

You may need to model the probability of the value with the higher count. In Example 12, $Y=1$ has the count 375 and $Y=0$ has 12. If you use:

```
proc probit order=freq data=ingot;
class s;
```

```
model s=t;
run;
```

you will have the following output:

Sample Program: Probit Regression

Probit Procedure
Class Level Information

Class	Levels	Values
S	2	1 0

Number of observations used = 387

Probit Procedure

Data Set =WORK.INGOT
Dependent Variable=S

Weighted Frequency Counts for the Ordered Response Categories

Level	Count
1	375
0	12

Log Likelihood for NORMAL -47.5087804

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	2.80035085	0.331621	71.30839	0.0001	Intercept
T	1	-0.0390757	0.011425	11.69807	0.0006	

Probit Model in Terms of Tolerance Distribution

MU	SIGMA
71.66476	25.59135

Estimated Covariance Matrix for Tolerance Parameters

	MU	SIGMA
MU	186.336614	98.799500
SIGMA	98.799500	55.985053

Sometimes you will need to know the predicted probability values. For example, if you need to know the probability of having an ingot not ready for rolling ($Y=0$) at $T=7$ from Example 12, you can compute the probability using the formula:

$$p = \Phi(\beta'x) = \Phi(-2.8003508 + 0.0390757*7) = 0.006$$

from the standard normal probability distribution table. You can obtain this kind of computation using the OUTPUT statement and the PRINT procedure:

```
proc probit;
model r/n=x1 x2;
output out=filename prob=varname;
run;
proc print data=filename;
run;
```

where filename is the output data set name and varname is the variable name for predicted probabilities. The SAS output will show all the predicted probabilities for all observation points.

Example 15: Predicted Probability Computation

Using the data in Example 2, if you use:

```
proc probit data=ingot2;
model r/n=t;
output out=prob2 prob=phat;
run;
proc print data=prob2;
run;
```

you will have the following additional result:

Sample Program: Probit Regression

OBS	T	S	N	PHAT
1	7	0	55	0.00576
2	14	2	157	0.01212
3	27	7	159	0.04047
4	51	3	16	0.20969

GENMOD Procedure

Probit regression can be modeled as a class of generalized linear models in which the response probability function is binomial and the link function is probit. Therefore you can use the PROC GENMOD to fit a probit model:

```
proc genmod;
model r/n=x1 x2 / dist=binomial link=probit;
run;
```

Example 16: SAS Probit Regression in PROC GENMOD

Using the data as in Example 2, you may use:

```
proc genmod data=ingot2;
model r/n=t / dist=binomial link=probit;
run;
```

You will have the following SAS output:

Sample Program: Probit Regression

The GENMOD Procedure

Model Information

Description	Value
Data Set	WORK.INGOT2
Distribution	BINOMIAL
Link Function	PROBIT
Dependent Variable	R
Dependent Variable	N
Observations Used	4
Number Of Events	12
Number Of Trials	387

Criteria For Assessing Goodness Of Fit

Criterion	DF	Value	Value/DF
Deviance	2	0.7392	0.3696
Scaled Deviance	2	0.7392	0.3696
Pearson Chi-Square	2	0.4228	0.2114
Scaled Pearson X2	2	0.4228	0.2114
Log Likelihood	.	-47.5088	.

Analysis Of Parameter Estimates

Parameter	DF	Estimate	Std Err	ChiSquare	Pr>Chi
INTERCEPT	1	-2.8004	0.3316	71.3084	0.0001
T	1	0.0391	0.0114	11.6981	0.0006
SCALE	0	1.0000	0.0000	.	.

NOTE: The scale parameter was held fixed.

Probit Regression with SPSS

Again, unlike in SAS, SPSS models the probability of Y=1 or of Y's higher sorted value. To fit a probit regression, use:

```
probit r of n with x1 x2
/model probit.
```

Example 17: SPSS Probit Regression in PROBIT procedure

Using the data in Example 1, you can use:

```
compute n = 1.
execute.
```

```
probit r of n with t
/model probit
/print none.
```

The resulting SPSS output will be:

```
* * * * * P R O B I T   A N A L Y S I S * * * * *
```

Parameter estimates converged after 13 iterations.

Optimal solution found.

Parameter Estimates (PROBIT model: (PROBIT(p)) = Intercept + BX):

	Regression Coeff.	Standard Error	Coeff./S.E.
T	-.03908	.01142	-3.42024

	Intercept	Standard Error	Intercept/S.E.
	2.80035	.33162	8.44443

Pearson Goodness-of-Fit Chi Square = 352.383 DF = 385 P = .882

Since Goodness-of-Fit Chi square is NOT significant, no heterogeneity factor is used in the calculation of confidence limits.

SPSS PROBIT procedure supports the inclusion of categorical variables as explanatory variables. However, it only accepts numerically coded categorical variables. If your categorical variable is string, you need to reassign string values to numerical values in a new variable before running PROBIT.

Example 18: SPSS Probit Regression in PROBIT procedure (categorical regressors)

Using the data in Example 7, you can use:

```
autorecode variables = drug / into drug2.
probit r of n by drug2(1 5) with x
/model probit
/print none
/criteria iterate(20) steplimit(.1).
```

where AUTORECORD reassigns the string values A, B, C, D, and E of the variable DRUG to the consecutive integers 1,2,3,4, and 5 in the new variable DRUG2.

The resulting SPSS output will be:

```
* * * * * P R O B I T   A N A L Y S I S * * * * *
```

DATA Information

18 unweighted cases accepted.

0 cases rejected because of out-of-range group values.

0 cases rejected because of missing data.

0 cases are in the control group.

Group Information

DRUG2	Level	N of Cases	Label
	1	3	A
	2	4	B
	3	4	C
	4	3	D
	5	4	E

MODEL Information

ONLY Normal Sigmoid is requested.

Parameter estimates converged after 18 iterations.

Optimal solution found.

Parameter Estimates (PROBIT model: (PROBIT(p)) = Intercept + BX):

	Regression Coeff.	Standard Error	Coeff./S.E.
X	1.15886	.43833	2.64378

	Intercept	Standard Error	Intercept/S.E.	DRUG2
	-1.51849	.32692	-4.64487	A
	-1.03837	.26286	-3.95027	B
	-2.04067	.37455	-5.44829	C
	-.31766	.33551	-.94678	D
	.19031	.24926	.76350	E

Pearson Goodness-of-Fit Chi Square = 4.383 DF = 12 P = .975

Since Goodness-of-Fit Chi square is NOT significant, no heterogeneity factor is used in the calculation of confidence limits.

Models for Multiple Outcomes

Introduction

Analysis for multiple outcomes or choices models the relationship between a multiple response variable and one or more explanatory variables. There are two broad types of outcome sets, ordered (or ordinal) and unordered. The choice of travel mode (by car, bus, or train) is unordered. Bond ratings, taste tests (from strong dislike to excellent taste), levels of insurance coverage (none, part, or full) are ordered by design. Two different types of approaches are

employed for the two types of models.

Models for Ordered Multiple Choices

The ordered multiple choice model assumes the relationship:

$$g(\text{Prob}(Y \leq j)) = \alpha_j + \beta'x \text{ for } j = 1, \dots, k$$

where the response of the variable Y is measured in one of $k+1$ different categories, α_j are k intercept parameters, and b is the slope parameter vector not including the intercept term. By construction, $\alpha_1 < \alpha_2 < \dots < \alpha_{k-1} < \alpha_k$ holds. This model fits a common slopes cumulative model, which is a parallel lines regression model based on the cumulative probabilities of the response categories. Ordered logit and ordered probit models are most commonly used.

Ordered Logit Regression

Ordered logit model has the form:

$$\text{logit}(p_1) \equiv \log \frac{p_1}{1 - p_1} = \alpha_1 + \beta'x$$

$$\text{logit}(p_1 + p_2) \equiv \log \frac{p_1 + p_2}{1 - p_1 - p_2} = \alpha_2 + \beta'x$$

⋮

$$\text{logit}(p_1 + p_2 + \dots + p_k) \equiv \log \frac{p_1 + p_2 + \dots + p_k}{1 - p_1 - p_2 - \dots - p_k} = \alpha_k + \beta'x$$

$$\text{and } p_1 + p_2 + \dots + p_{k+1} = 1$$

This model is known as the proportional-odds model because the odds ratio of the event $Y \leq j$ is independent of the category j . The odds ratio is assumed to be constant for all categories.

Ordered Logit Regression with SAS

LOGISTIC Procedure

SAS PROC LOGISTIC is a procedure you can use for an ordered multiple outcome model as well as for a binary model. All previous discussions about the binary logistic regression estimation in PROC LOGISTIC are also valid for ordered logit model. To fit an ordered logit model, you can use:

```
proc logistic;
model y=x1 x2;
run;
```

where Y is the ordinally scaled multiple response variable, and $X1$ and $X2$ are two regressors of interest.

Example 19: SAS Ordered Logit Regression in PROC LOGISTIC

The following data are from McCullagh and Nelder (McCullagh and Nelder, 1989, *Generalized Linear Models*, London, Chapman Hall, p. 175) and used in a SAS manual (SAS, 1996, SAS/STAT Software Changes and Enhancements through Release 6.11, pp. 435-438). Consider a study of the effects on taste of various cheese additives. Researchers tested four cheese additives and obtained 52 response ratings for each additive. Each response was

measured on a scale of nine values ranging from strong dislike (1) to excellent taste (9). The data set CHEESE has five variables Y, X1, X2, X3, and F. The variable Y contains the response rating and the variables X1, X2, and X3 are dummy variables, representing the first, second, and third additive, respectively; for the fourth additive, X1=X2=X3=0. F gives the frequency of occurrence of the observation. The following DATA step creates the data set CHEESE:

```
data cheese;
input x1 x2 x3 y f;
cards;
1 0 0 1 0
1 0 0 2 0
1 0 0 3 1
1 0 0 4 7
1 0 0 5 8
1 0 0 6 8
1 0 0 7 19
1 0 0 8 8
1 0 0 9 1
0 1 0 1 6
0 1 0 2 9
0 1 0 3 12
0 1 0 4 11
0 1 0 5 7
0 1 0 6 6
0 1 0 7 1
0 1 0 8 0
0 1 0 9 0
0 0 1 1 1
0 0 1 2 1
0 0 1 3 6
0 0 1 4 8
0 0 1 5 23
0 0 1 6 7
0 0 1 7 5
0 0 1 8 1
0 0 1 9 0
0 0 0 1 0
0 0 0 2 0
0 0 0 3 0
0 0 0 4 1
0 0 0 5 3
0 0 0 6 7
0 0 0 7 14
0 0 0 8 16
0 0 0 9 11
;
```

Because the response variable Y is ordinally scaled, you can estimate an ordered logit model. You can use:

```
proc logistic data=cheese;
freq f;
model y=x1-x3;
run;
```

You will have the following SAS output:

Sample Program: Ordered Logit Regression

The LOGISTIC Procedure

Data Set: WORK.CHEESE
 Response Variable: Y
 Response Levels: 9
 Number of Observations: 28
 Frequency Variable: F
 Link Function: Logit

Response Profile

Ordered Value	Y	Count
1	1	7
2	2	10
3	3	19
4	4	27
5	5	41
6	6	28
7	7	39
8	8	25
9	9	12

NOTE: 8 observation(s) having zero frequencies or weights were excluded since they do not contribute to the analysis.

Score Test for the Proportional Odds Assumption

Chi-Square = 17.2868 with 21 DF (p=0.6936)

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	875.802	733.348	.
SC	902.502	770.061	.
-2 LOG L	859.802	711.348	148.454 with 3 DF (p=0.0001)
Score	.	.	111.267 with 3 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCP1	1	-7.0802	0.5624	158.4865	0.0001	.	.
INTERCP2	1	-6.0250	0.4755	160.5507	0.0001	.	.
INTERCP3	1	-4.9254	0.4272	132.9477	0.0001	.	.
INTERCP4	1	-3.8568	0.3902	97.7086	0.0001	.	.
INTERCP5	1	-2.5206	0.3431	53.9713	0.0001	.	.
INTERCP6	1	-1.5685	0.3086	25.8379	0.0001	.	.
INTERCP7	1	-0.0669	0.2658	0.0633	0.8013	.	.
INTERCP8	1	1.4930	0.3310	20.3443	0.0001	.	.
X1	1	1.6128	0.3778	18.2258	0.0001	0.385954	5.017
X2	1	4.9646	0.4741	109.6453	0.0001	1.188080	143.257
X3	1	3.3227	0.4251	61.0936	0.0001	0.795146	27.735

The LOGISTIC Procedure

Association of Predicted Probabilities and Observed Responses

Concordant = 67.6%	Somers' D = 0.578
Discordant = 9.8%	Gamma = 0.746
Tied = 22.6%	Tau-a = 0.500
(18635 pairs)	c = 0.789

This result shows eight fitted regression lines as follows:

$$\text{logit}(p_1) = -7.0802 + 1.6128*X1 + 4.9646*X2 + 3.3227*X3$$

$$\text{logit}(p_1 + p_2) = -6.0250 + 1.6128*X1 + 4.9646*X2 + 3.3227*X3$$

$$\text{logit}(p_1 + p_2 + \dots + p_8) = 1.4930 + 1.6128*X1 + 4.9646*X2 + 3.3227*X3$$

where p_1 is the probability of being strongly disliked, i.e, the probability of $Y=1$, and so on. Positive coefficients of $X1$, $X2$ and $X3$ indicate that adding those additives is associated with increased probability of the cheese being disliked. The estimated odds are reported 5.017, 143.257 and 27.735 for $X1$, $X2$ and $X3$ respectively. Each odd is constant for all categories.

Example 20: Predicted Probability Computation

You can compute the predicted probability at a certain level of independent variables. For example, you can use the following formula to compute the predicted probabilities at $X1=1$, $X2=0$ and $X3=0$ for the model in Example 19:

$$p_1 = \frac{\exp(-7.0802+1.6128)}{1+\exp(-7.0802+1.6128)} = 0.000420$$

$$p_1+p_2 = \frac{\exp(-6.0250+1.6128)}{1+\exp(-6.0250+1.6128)} = 0.01198 \text{ and thus } p_2=0.01198-0.000420=0.00778$$

and so on. However, this computation can be easily obtained for each combination of additives by using:

```
proc logistic data=cheese;
```

```

freq f;
model y=x1-x3;
output out=prob predicted=phat;
run;
proc print data=prob;
run;

```

You will have the following additional output:

Sample Program: Ordered Logit Regression

OBS	X1	X2	X3	Y	F	_LEVEL_	PHAT
1	1	0	0	1	0	1	0.00420
2	1	0	0	1	0	2	0.01198
3	1	0	0	1	0	3	0.03514
4	1	0	0	1	0	4	0.09587
5	1	0	0	1	0	5	0.28746
6	1	0	0	1	0	6	0.51106
7	1	0	0	1	0	7	0.82432
8	1	0	0	1	0	8	0.95713
				.			
				.			
281	0	0	0	9	11	1	0.00084
282	0	0	0	9	11	2	0.00241
283	0	0	0	9	11	3	0.00721
284	0	0	0	9	11	4	0.02070
285	0	0	0	9	11	5	0.07443
286	0	0	0	9	11	6	0.17242
287	0	0	0	9	11	7	0.48329
288	0	0	0	9	11	8	0.81652

In this output you have 8 observations for each additive-response combination. The observation with `_LEVEL_=1` shows the predicted probability of `Y=1`, the observation with `_LEVEL_=2` shows the predicted probability of `Y=1` or `2`, and so on.

PROBIT Procedure

To use the SAS PROC PROBIT to fit an ordered logit model, use the syntax:

```

proc probit;
class y;
model y = x1 x2 / d=logistic;
run;

```

where `Y` is the ordinally scaled multiple response variable, and `X1` and `X2` are two regressors of interest.

Example 21: SAS Ordered Logit Regression in PROC PROBIT

In this example, you are using the same data set as in Example 19. However, SAS PROC PROBIT does not accept a data set in a frequency format. You need to have the same data set in an individual data format; i.e., you need to have:

```

X1    X2    X3    Y

```

Categorical Analysis - Part 1

```

1    0    0    3
1    0    0    4
1    0    0    4
      .
      .
1    0    0    4 (7 rows of the same data)
1    0    0    5
1    0    0    5
      .
      .
1    0    0    5 (8 rows of the same data)
      .
      .
      .

```

With this new data set, CHEESE2, you can use:

```

proc probit data=cheese2;
class y;
model y = x1-x3 / d=logistic;
run;

```

The resulting SAS output will be:

Sample Program: Ordered Logit Regression

Probit Procedure
Class Level Information

Class	Levels	Values
Y	9	1 2 3 4 5 6 7 8 9

Number of observations used = 208

Probit Procedure

Data Set =WORK.CHEESE2
Dependent Variable=Y

Weighted Frequency Counts for the Ordered Response Categories

Level	Count
1	7
2	10
3	19
4	27
5	41
6	28
7	39
8	25

Log Likelihood for LOGISTIC -355.6739524

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-7.0801649	0.56401	157.5844	0.0001	Intercept
X1	1	1.6127909	0.380544	17.96169	0.0001	
X2	1	4.96463991	0.476721	108.4546	0.0001	
X3	1	3.32268278	0.42183	62.04439	0.0001	
INTER.2	1	1.0551848	0.324654			2
INTER.3	1	2.15474934	0.387165			3
INTER.4	1	3.22336352	0.420573			4
INTER.5	1	4.55961327	0.454216			5
INTER.6	1	5.5116267	0.479248			6
INTER.7	1	7.01328969	0.520899			7
INTER.8	1	8.57313924	0.587685			8

Notice that intercept parameter estimates are computed as:

$$\alpha_1 = -7.0802, \alpha_2 = -7.0802 + 1.0552, \alpha_3 = -7.0802 + 2.1547, \alpha_4 = -7.0802 + 3.2234,$$

$$\alpha_5 = -7.0802 + 4.5596, \alpha_6 = -7.0802 + 5.5116, \alpha_7 = -7.0802 + 7.0133, \alpha_8 = -7.0802 + 8.5731$$

Ordered Probit Regression

The ordered probit model has the form:

$$\Phi^{-1}(p_1) = \alpha_1 + \beta'x$$

$$\Phi^{-1}(p_1 + p_2) = \alpha_2 + \beta'x$$

$$\vdots$$

$$\Phi^{-1}(p_1 + p_2 + \dots + p_k) = \alpha_k + \beta'x$$

and $p_1 + p_2 + \dots + p_{k+1} = 1$

or equivalently,

$$p_1 = \Phi(\alpha_1 + \beta'x)$$

$$p_2 = \Phi(\alpha_2 + \beta'x) - \Phi(\alpha_1 + \beta'x)$$

$$\vdots$$

$$p_k = \Phi(\alpha_k + \beta'x) - \Phi(\alpha_{k-1} + \beta'x)$$

$$p_{k+1} = 1 - \Phi(\alpha_k + \beta'x)$$

where Φ^{-1}

is the inverse of the cumulative standard normal distribution function, often referred to as probit or normit, and Φ

denotes the cumulative standard normal distribution function.

Ordered Probit Regression with SAS

LOGISTIC Procedure

To fit an ordered probit model in PROC LOGISTIC, use the LINK=NORMIT (or PROBIT) option as:

```
proc logistic;
model y=x1 x2 / link=normit;
run;
```

Example 22: SAS Ordered Probit Regression in PROC LOGISTIC

Using the data set CHEESE in Example 19, if you use:

```
proc logistic data=cheese;
freq f;
model y=x1-x3 / link=normit;
run;
```

you will have:

Sample Program: Ordered Probit Regression

The LOGISTIC Procedure

```
Data Set: WORK.CHEESE
Response Variable: Y
Response Levels: 9
Number of Observations: 28
Frequency Variable: F
Link Function: Normit
```

Response Profile

Ordered Value	Y	Count
1	1	7
2	2	10
3	3	19
4	4	27
5	5	41
6	6	28
7	7	39
8	8	25
9	9	12

NOTE: 8 observation(s) having zero frequencies or weights were excluded since they do not contribute to the analysis.

Score Test for the Equal Slopes Assumption

Chi-Square = 15.0251 with 21 DF (p=0.8217)

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	875.802	729.391	.
SC	902.502	766.104	.
-2 LOG L	859.802	707.391	152.411 with 3 DF (p=0.0001)
Score	.	.	108.491 with 3 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate
INTERCP1	1	-4.0762	0.2867	202.1202	0.0001	.
INTERCP2	1	-3.5087	0.2496	197.6200	0.0001	.
INTERCP3	1	-2.8628	0.2248	162.2226	0.0001	.
INTERCP4	1	-2.2356	0.2067	117.0124	0.0001	.
INTERCP5	1	-1.4641	0.1858	62.0947	0.0001	.
INTERCP6	1	-0.9155	0.1730	28.0144	0.0001	.
INTERCP7	1	-0.0276	0.1607	0.0296	0.8634	.
INTERCP8	1	0.8779	0.1841	22.7341	0.0001	.
X1	1	0.9643	0.2119	20.7122	0.0001	0.418540
X2	1	2.8618	0.2508	130.2471	0.0001	1.242206
X3	1	1.9408	0.2296	71.4236	0.0001	0.842421

The LOGISTIC Procedure

Association of Predicted Probabilities and Observed Responses

Concordant = 58.4%	Somers' D = 0.512
Discordant = 7.2%	Gamma = 0.781
Tied = 34.4%	Tau-a = 0.443
(18635 pairs)	c = 0.756

This result shows eight fitted regression lines as:

$$\Phi^1(p_1) = -4.0762 + 0.9643*X1 + 2.8618*X2 + 1.9408*X3$$

$$\Phi^1(p_1 + p_2) = -3.5087 + 0.9643*X1 + 2.8618*X2 + 1.9408*X3$$

.

.

$$\Phi^1(p_1 + p_2 + \dots + p_k) = 0.8779 + 0.9643*X1 + 2.8618*X2 + 1.9408*X3$$

PROBIT Procedure

You can use the SAS PROC PROBIT to fit an ordered probit model:

```
proc probit;
class y;
model y = x1 x2;
run;
```

Example 23: SAS Ordered Probit Regression in PROC PROBIT

Using the data set CHEESE2 in Example 21, you can use:

```
proc probit data=cheese2;
class y;
model y = x1-x3;
run;
```

Your SAS output will be:

Sample Program: Ordered Probit Regression

Probit Procedure
Class Level Information

Class	Levels	Values
Y	9	1 2 3 4 5 6 7 8 9

Number of observations used = 208

Probit Procedure

Data Set =WORK.CHEESE2
Dependent Variable=Y

Weighted Frequency Counts for the Ordered Response Categories

Level	Count
1	7
2	10
3	19
4	27
5	41

6	28
7	39
8	25
9	12

Log Likelihood for NORMAL -353.6953428

Probit Procedure

Variable	DF	Estimate	Std Err	ChiSquare	Pr>Chi	Label/Value
INTERCPT	1	-4.0761911	0.2868	202	0.0001	Intercept
X1	1	0.9642503	0.211624	20.761	0.0001	
X2	1	2.86184814	0.24956	131.5057	0.0001	
X3	1	1.94080682	0.230248	71.05121	0.0001	
INTER.2	1	0.56749271	0.166663			2
INTER.3	1	1.21337982	0.200722			3
INTER.4	1	1.84054882	0.216171			4
INTER.5	1	2.61204661	0.229904			5
INTER.6	1	3.16070252	0.241469			6
INTER.7	1	4.04855742	0.263792			7
INTER.8	1	4.95412966	0.298786			8

This result shows eight fitted regression lines as:

$$\Phi^1(p_1) = -4.0762 + 0.9643*X1 + 2.8618*X2 + 1.9408*X3$$

$$\Phi^1(p_1 + p_2) = -4.0762 + 0.5675 + 0.9643*X1 + 2.8618*X2 + 1.9408*X3$$

.

$$\Phi^1(p_1 + p_2 + \dots + p_k) = -4.0762 + 4.9541 + 0.9643*X1 + 2.8618*X2 + 1.9408*X3$$

which are the same results as we obtained in Example 22.

Example 24: Predicted Probability Computation

Predicted probability computation can be easily obtained using:

```
proc probit data=cheese2;
class y;
model y = x1-x3;
output out=prob2 prob=phat;
run;
proc print data=prob2;
run;
```

As a result, you will have:

Sample Program: Ordered Probit Regression

OBS	X1	X2	X3	Y	_LEVEL_	PHAT
1	1	0	0	3	1	0.00093

2	1	0	0	3	2	0.00547
3	1	0	0	3	3	0.02881
4	1	0	0	3	4	0.10179
5	1	0	0	3	5	0.30857
6	1	0	0	3	6	0.51945
7	1	0	0	3	7	0.82552
8	1	0	0	3	8	0.96728
				.		
				.		
1657	0	0	0	9	1	0.00002
1658	0	0	0	9	2	0.00023
1659	0	0	0	9	3	0.00210
1660	0	0	0	9	4	0.01269
1661	0	0	0	9	5	0.07158
1662	0	0	0	9	6	0.17997
1663	0	0	0	9	7	0.48898
1664	0	0	0	9	8	0.81001

Models for Unordered Multiple Choices

The unordered multiple choice model assumes the relationship:

$$g(\text{Prob}(Y=j)) = \beta_j'x \text{ for } j = 1, \dots, k+1$$

where the response of the variable Y is measured in one of $k+1$ different categories, and β_j is the parameter vector for each j . This model is made operational by a particular choice of the distributional form of g . Although two models, logit and probit could be considered as before, the probit model is practically hard to employ. Two different logit models are commonly used; one is multinomial logit or generalized logit model and the other is conditional logit (McFadden, 1974, "Conditional Logit Analysis of Qualitative Choice Behavior," *Frontiers in Econometrics*, Zarembka ed., New York, Academic Press, pp. 105-142) or discrete choice model (this is also often referred as multinomial logit model, resulting in a conflict in terminology). The major difference between the two models is found in the characteristics of the vector x . The multinomial logit model is typically (but not necessarily) used for the data in which x variables are the characteristics of individuals, not the characteristics of the choices. The conditional logit model is typically (but not necessarily) employed in the case where x variables are the characteristics of the choices, often called attributes of the choices.

Multinomial Logit Regression

The multinomial logit model has the form:

$$p_j = \frac{\exp(\beta_j'x)}{\sum_j \exp(\beta_j'x)} \text{ for } j = 1, \dots, k+1$$

β_{k+1} can be set to 0 (zero vector) as a normalization and thus:

$$p_{k+1} = \frac{1}{\sum_j \exp(\beta_j'x)}$$

As a result, the j logit has the form:

$$\log \frac{p_j}{p_{k+1}} = \beta_j'x \text{ for } j = 1, \dots, k$$

Multinomial Logit Regression with SAS

CATMOD Procedure

The SAS CATMOD procedure is capable of dealing with various types of the multinomial logit model. The basic syntax to fit a multinomial logit model is:

```
proc catmod;
  direct x1;
  response logits;
  model y=x1 x2;
run;
```

where X1 is a continuous quantitative variable and X2 is a categorical variable. You must specify your continuous regressors in the DIRECT statement.

The RESPONSE statement specifies the functions of response probabilities used to model the response functions as a linear combination of the parameters. Depending on your model, you can specify other types of responses beside the LOGITS. For example, among all others,

CLOGITS (cumulative logits): $\log \frac{p_1}{1 - p_1}, \log \frac{p_1 + p_2}{1 - p_1 - p_2}, \dots, \log \frac{p_1 + p_2 + \dots + p_k}{1 - p_1 - p_2 - \dots - p_k}$

ALOGITS (adjacent logits): $\log \frac{p_2}{p_1}, \log \frac{p_3}{p_2}, \dots, \log \frac{p_k}{p_{k-1}}$

The default is LOGITS (generalized logits) and it models:

$$\log \frac{p_j}{p_{k+1}} = \beta_j x \text{ for } j = 1, \dots, k$$

Example 25: SAS Multinomial Logit Regression in PROC CATMOD

In this example, you are using a modified version of the data set CHEESE in Example 19. Zero frequency for some observations causes a sparseness of the data and thus you may have problems in fitting the multinomial logit model. In order to avoid the zero frequency, the following is being tried:

X1	X2	X3	Y	F
1	0	0	1	1
1	0	0	2	1
1	0	0	3	1
1	0	0	4	5
1	0	0	5	8
1	0	0	6	8
1	0	0	7	19
1	0	0	8	8
1	0	0	9	1
0	1	0	1	6
0	1	0	2	9
0	1	0	3	12
0	1	0	4	11

Categorical Analysis - Part 1

```

0 1 0 5 7
0 1 0 6 4
0 1 0 7 1
0 1 0 8 1
0 1 0 9 1
0 0 1 1 1
0 0 1 2 1
0 0 1 3 6
0 0 1 4 8
0 0 1 5 23
0 0 1 6 7
0 0 1 7 4
0 0 1 8 1
0 0 1 9 1
0 0 0 1 1
0 0 0 2 1
0 0 0 3 1
0 0 0 4 1
0 0 0 5 1
0 0 0 6 6
0 0 0 7 14
0 0 0 8 16
0 0 0 9 11

```

You are supposed to rearrange this data in the way you have used in Example 21. Furthermore, you are collapsing the nine response categories into three for a simpler illustration. That is, you will create a new response variable YLESS such that

```

if y=1 or y=2 or y=3 then yless=1;
else if y=4 or y=5 or y=6 then yless=2;
else yless=3;

```

With this new data set CHEESE3, if you use:

```

proc catmod data=cheese3;
direct x1-x4;
response logits;
model yless=x1-x4 / noiter freq;
run;

```

the resulting output will be:

Sample Program: Multinomial Logit Regression

CATMOD PROCEDURE

Response: YLESS	Response Levels (R)=	3
Weight Variable: None	Populations (S)=	4
Data Set: CHEESE3	Total Frequency (N)=	208
Frequency Missing: 0	Observations (Obs)=	208

POPULATION PROFILES

Sample	X1	X2	X3	X4	Sample Size
1	0	0	0	1	52
2	0	0	1	0	52
3	0	1	0	0	52
4	1	0	0	0	52

RESPONSE PROFILES

Response	YLESS
1	1
2	2
3	3

RESPONSE FREQUENCIES

Sample	Response Number			
	1	2	3	
1	3	8		41
2	8	38		6
3	27	22		3
4	3	21		28

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	2	33.46	0.0000
X1	2	7.79	0.0203
X2	2	36.33	0.0000
X3	2	37.07	0.0000
X4	0*	.	.
LIKELIHOOD RATIO	0	.	.

NOTE: Effects marked with '*' contain one or more redundant or restricted parameters.

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-2.6150	0.5981	19.12	0.0000
	2	-1.6341	0.3865	17.88	0.0000
X1	3	0.3814	0.8525	0.20	0.6546

	4	1.3464	0.4824	7.79	0.0053
X2	5	4.8122	0.8532	31.81	0.0000
	6	3.6266	0.7267	24.90	0.0000
X3	7	2.9026	0.8058	12.97	0.0003
	8	3.4800	0.5851	35.37	0.0000
X4	9
	10

The estimation result shows two regression lines:

$$\log \frac{p_1}{p_3} = -2.6150 + 0.3814*X1 + 4.8122*X2 + 2.9026*X3$$

$$\log \frac{p_2}{p_3} = -1.6341 + 1.3464*X1 + 3.6266*X2 + 3.4800*X3$$

Thus, the estimated logit at each combination of X's is

$$\text{at } X1=1, X2=0, X3=0 \text{ and } X4=0, \log \frac{p_1}{p_3} = -2.6150 + 0.3814 = -2.2336$$

$$\log \frac{p_2}{p_3} = -1.6341 + 1.3464 = -0.2877$$

$$\text{at } X1=0, X2=1, X3=0 \text{ and } X4=0, \log \frac{p_1}{p_3} = -2.6150 + 4.8122 = 2.1972$$

$$\log \frac{p_2}{p_3} = -1.6341 + 3.6266 = 1.9925$$

$$\text{at } X1=0, X2=0, X3=1 \text{ and } X4=0, \log \frac{p_1}{p_3} = -2.6150 + 2.9026 = 0.2876$$

$$\log \frac{p_2}{p_3} = -1.6341 + 3.4800 = 1.8459$$

$$\text{at } X1=0, X2=0, X3=0 \text{ and } X4=1, \log \frac{p_1}{p_3} = -2.6150$$

$$\log \frac{p_2}{p_3} = -1.6341$$

If you need to know the predicted probabilities you can compute them by applying the formula. For example, at $X1=1$, $X2=0$ and $X3=0$,

$$p_1 = \frac{\exp(-2.2336)}{1 + \exp(-2.2336) + \exp(-0.2877)} = 0.05769$$

$$p_2 = \frac{\exp(-0.2877)}{1 + \exp(-2.2336) + \exp(-0.2877)} = 0.40384$$

This computation can be easily obtained by including the PROB statement in the MODEL command as:

```
proc catmod data=cheese3;
  direct x1-x4;
  response logits;
```

```
model yless=x1-x4 / noiter freq prob;
run;
```

In addition to the output above, you will get the following:

RESPONSE PROBABILITIES

Sample	Response Number		
	1	2	3
1	0.05769	0.15385	0.78846
2	0.15385	0.73077	0.11538
3	0.51923	0.42308	0.05769
4	0.05769	0.40385	0.53846

Example 26: SAS Multinomial Logit Regression in PROC CATMOD (categorical regressors)

In Example 25, X1, X2, X3, and X4 are dummy variables to denote each type of additive. The same result can be obtained by employing a categorical variable to represent types of additives. To do this, you can create a new variable X such that

```
if x1=1 then x=1;
else if x2=1 then x=2;
else if x3=1 then x=3;
else x=4;
```

Now if you use:

```
proc catmod data=cheese3;
response logits;
model yless = x / noiter freq prob;
run;
```

the resulting SAS output will be:

Sample Program: Multinomial Logit Regression

CATMOD PROCEDURE

Response: YLESS	Response Levels (R)=	3
Weight Variable: None	Populations (S)=	4
Data Set: CHEESE3	Total Frequency (N)=	208
Frequency Missing: 0	Observations (Obs)=	208

POPULATION PROFILES

Sample	X	Sample Size
1	1	52
2	2	52
3	3	52
4	4	52

RESPONSE PROFILES

Response YLESS

1	1
2	2
3	3

RESPONSE FREQUENCIES

Sample	Response Number		
	1	2	3
1	3	21	28
2	27	22	3
3	8	38	6
4	3	8	41

RESPONSE PROBABILITIES

Sample	Response Number		
	1	2	3
1	0.05769	0.40385	0.53846
2	0.51923	0.42308	0.05769
3	0.15385	0.73077	0.11538
4	0.05769	0.15385	0.78846

MAXIMUM-LIKELIHOOD ANALYSIS-OF-VARIANCE TABLE

Source	DF	Chi-Square	Prob
INTERCEPT	2	18.26	0.0001
X	6	78.70	0.0000
LIKELIHOOD RATIO	0	.	.

ANALYSIS OF MAXIMUM-LIKELIHOOD ESTIMATES

Effect	Parameter	Estimate	Standard Error	Chi-Square	Prob
INTERCEPT	1	-0.5909	0.2946	4.02	0.0449
	2	0.4791	0.2242	4.57	0.0326
X	3	-1.6427	0.5209	9.95	0.0016
	4	-0.7668	0.3032	6.39	0.0114
	5	2.7881	0.5215	28.59	0.0000
	6	1.5133	0.4895	9.56	0.0020
	7	0.8786	0.4823	3.32	0.0685
	8	1.3667	0.3831	12.73	0.0004

The result shows each estimated logit can be calculated as:

$$\text{at } X=1, \log \frac{p_1}{p_3} = -0.5909 - 1.6427 = -2.2336 \quad \text{and} \quad \log \frac{p_2}{p_3} = 0.4791 - 0.7668 = -0.2877$$

$$\text{at } X=2, \log \frac{p_1}{p_3} = -0.5909 + 2.7881 = 2.1972 \quad \text{and} \quad \log \frac{p_2}{p_3} = 0.4791 + 1.5133 = 1.9924$$

$$\text{at } X=3, \log \frac{p_1}{p_3} = -0.5909 + 0.8786 = 0.2877 \quad \text{and} \quad \log \frac{p_2}{p_3} = 0.4791 + 1.3667 = 1.8458$$

$$\text{at } X=4, \log \frac{p_1}{p_3} = -0.5909 - (-1.6427 + 2.7881 + 0.8786) = -2.6149$$

$$\text{and} \quad \log \frac{p_2}{p_3} = 0.4791 - (-0.7668 + 1.5133 + 1.3667) = -1.6341$$

This is exactly the same logit computation as in the previous example.

Multinomial Logit Regression with SPSS

GENLOG Procedure

The SPSS GENLOG procedure conducts the general loglinear analysis and the logit model can be treated as a special class of loglinear models. However, SPSS is only capable of dealing with a multinomial logit model with categorical independent variables. The basic syntax to fit a multinomial logit model is:

```
genlog y by x1 x2
  /model=multinomial
  /design y y*x1 y*x2 y*x1*x2.
```

where Y is response variable and X1 and X2 are categorical regressors.

Example 27: SPSS Multinomial Logit Regression in GENLOG

You are using the same data in Example 25. You can use:

```
genlog yless by x
  /model=multinomial
  /print freq estim
  /plot none
  /criteria=cin(95) iteration(20) converge(.001) delta(0)
  /design yless yless*x.
```

The resulting SPSS output will be:

```
-----
                          GENERALIZED LOGLINEAR ANALYSIS
-----
```

Data Information

```
208 cases are accepted.
  0 cases are rejected because of missing data.
208 weighted cases will be used in the analysis.
12 cells are defined.
  0 structural zeros are imposed by design.
  0 sampling zeros are encountered.
```

Variable Information

Factor	Levels	Value
YLESS	3	1.00
		2.00
		3.00
X	4	1.00
		2.00
		3.00
		4.00

Model and Design Information

Model: Multinomial Logit
 Design: Constant + YLESS + YLESS*X

Note: There is a separate constant term for each combination of levels of the independent factors.

Correspondence Between Parameters and Terms of the Design

Parameter	Aliased	Term
1		Constant for [X = 1.00]
2		Constant for [X = 2.00]
3		Constant for [X = 3.00]
4		Constant for [X = 4.00]
5		[YLESS = 1.00]
6		[YLESS = 2.00]
7	x	[YLESS = 3.00]
8		[YLESS = 1.00]*[X = 1.00]
9		[YLESS = 1.00]*[X = 2.00]
10		[YLESS = 1.00]*[X = 3.00]
11	x	[YLESS = 1.00]*[X = 4.00]
12		[YLESS = 2.00]*[X = 1.00]
13		[YLESS = 2.00]*[X = 2.00]
14		[YLESS = 2.00]*[X = 3.00]
15	x	[YLESS = 2.00]*[X = 4.00]
16	x	[YLESS = 3.00]*[X = 1.00]
17	x	[YLESS = 3.00]*[X = 2.00]
18	x	[YLESS = 3.00]*[X = 3.00]

19 x [YLESS = 3.00]*[X = 4.00]

Note: 'x' indicates an aliased (or a redundant) parameter.
 These parameters are set to zero.

 Convergence Information

Maximum number of iterations: 20
 Relative difference tolerance: .001
 Final relative difference: 2.92779E-14

Maximum likelihood estimation converged at iteration 1.

Table Information

Factor	Value	Observed Count	%	Expected Count	%
X	1.00				
YLESS	1.00	3.00 (5.77)		3.00 (5.77)	
YLESS	2.00	21.00 (40.38)		21.00 (40.38)	
YLESS	3.00	28.00 (53.85)		28.00 (53.85)	
X	2.00				
YLESS	1.00	27.00 (51.92)		27.00 (51.92)	
YLESS	2.00	22.00 (42.31)		22.00 (42.31)	
YLESS	3.00	3.00 (5.77)		3.00 (5.77)	
X	3.00				
YLESS	1.00	8.00 (15.38)		8.00 (15.38)	
YLESS	2.00	38.00 (73.08)		38.00 (73.08)	
YLESS	3.00	6.00 (11.54)		6.00 (11.54)	
X	4.00				
YLESS	1.00	3.00 (5.77)		3.00 (5.77)	
YLESS	2.00	8.00 (15.38)		8.00 (15.38)	
YLESS	3.00	41.00 (78.85)		41.00 (78.85)	

 Goodness-of-fit Statistics

	Chi-Square	DF	Sig.
Likelihood Ratio	.0000	0	.
Pearson	.0000	0	.

Analysis of Dispersion

Source of Dispersion	Entropy	Concentration	DF
Due to Model	55.4019	35.2404	12
Due to Residual	163.2376	97.3462	402
Total	218.6395	132.5865	414

Measures of Association

Entropy = .2534
 Concentration = .2658

Parameter Estimates

Constant	Estimate
1	3.3322
2	1.0986
3	1.7918
4	3.7136

Note: Constants are not parameters under multinomial assumption.
 Therefore, standard errors are not calculated.

Parameter	Estimate	SE	Z-value	Asymptotic 95% CI	
				Lower	Upper
5	-2.6150	.5981	-4.37	-3.79	-1.44
6	-1.6341	.3865	-4.23	-2.39	-.88
7	.0000
8	.3814	.8525	.45	-1.29	2.05
9	4.8122	.8533	5.64	3.14	6.48
10	2.9026	.8058	3.60	1.32	4.48
11	.0000
12	1.3464	.4824	2.79	.40	2.29
13	3.6266	.7268	4.99	2.20	5.05
14	3.4800	.5851	5.95	2.33	4.63
15	.0000
16	.0000
17	.0000
18	.0000
19	.0000

SPSS output shows the following model structure:

Assuming YLESS=3 is the reference category, the estimated logit of YLESS=1 at X=1 is

$$\log \frac{m_{11}}{m_{31}} = \lambda^{\text{YLESS}=1} + \lambda^{\text{YLESS}=1 \text{ and } X=1}$$

where m_{11} is the predicted count for YLESS=1 at X=1 and m_{31} is the predicted count for YLESS=3 at X=1. Similarly, the estimated logit of YLESS=2 at X=1 is

$$\log \frac{m_{21}}{m_{31}} = \lambda^{\text{YLESS}=2} + \lambda^{\text{YLESS}=2 \text{ and } X=1}$$

where m_{21} is the predicted count for YLESS=2 at X=1. Other possible logit computations at X=2,3 and 4 can be derived in the same manner.

The output provides the following estimation results:

$$\log \frac{m_{11}}{m_{31}} = -2.6150 + 0.3814 = -2.2336$$

$$\log \frac{m_{21}}{m_{31}} = -1.6341 + 1.3464 = -0.2877$$

$$\log \frac{m_{12}}{m_{32}} = -2.6150 + 4.8122 = 2.1972$$

$$\log \frac{m_{22}}{m_{32}} = -1.6341 + 3.6266 = 1.9925$$

$$\log \frac{m_{13}}{m_{33}} = -2.6150 + 2.9026 = 0.2876$$

$$\log \frac{m_{23}}{m_{33}} = -1.6341 + 3.4800 = 1.8459$$

$$\log \frac{m_{14}}{m_{34}} = -2.6150 + 0.0000 = -2.6150$$

$$\log \frac{m_{24}}{m_{34}} = -1.6341 + 0.0000 = -1.6341$$

Conditional Logit Regression

The conditional logit model has the form:

$$p_j = \frac{\exp(\beta'x_j)}{\sum_j \exp(\beta'x_j)} \text{ for } j = 1, \dots, k+1$$

In this model, subjects are presented with choice alternatives and asked to choose the most preferred alternative. The set of alternatives is typically the same for all subjects and the explanatory variables are all choice specific. Unlike in the multinomial logit model, the parameters are not specific to the choice.

Conditional Logit Regression with SAS

PHREG Procedure

The SAS PHREG procedure performs regression analysis of survival data based on the Cox proportional hazards model. Its likelihood function is similar to that of the conditional logit model.

To fit a conditional logit model with PROC PHREG, you need to rearrange the data set in a form that is consistent with survival analysis data. The most preferred choice is said to occur at time 1 and all other choices are said to occur at later times or to be censored. You also need to create a status variable to denote whether the observation was censored or not, i.e., whether the alternative was chosen or not. The censoring indicator variable has the value of 0 if the alternative was censored (not chosen) and 1 if not censored (chosen). The basic syntax is:

```
proc phreg;
strata strata_varname;
model time_varname*status_varname(0) = x1 x2;
run;
```

where *strata_varname* is the name of variable to specify the variable that determines the stratification, *time_varname* is the name of failure time variable (the smaller value means the alternative was chosen), *status_varname* is the name of the censoring indicator variable, of which 0 is the value to indicate censoring, and X1 and X2 are explanatory variables.

Example 28: SAS Conditional Logit Regression in PROC PHREG

This example is from SAS (SAS, 1995, *Logistic Regression Examples Using the SAS System*, pp. 2-3). Chocolate candy data are generated in which 10 subjects are presented with eight different chocolate candies. The subjects choose one preferred candy from among the eight types. The eight candies consist of eight combinations of dark(1) or milk(0) chocolate, soft(1) or hard(0) center, and nuts(1) or no nuts(0). The following data step creates the data set CHOCO:

```
data choco;
input subject choose dark soft nuts @@;
t=2-choose;
cards;
 1 0 0 0 0   1 0 0 0 1   1 0 0 1 0   1 0 0 1 1
 1 1 1 0 0   1 0 1 0 1   1 0 1 1 0   1 0 1 1 1
 2 0 0 0 0   2 0 0 0 1   2 0 0 1 0   2 0 0 1 1
 2 0 1 0 0   2 1 1 0 1   2 0 1 1 0   2 0 1 1 1
 3 0 0 0 0   3 0 0 0 1   3 0 0 1 0   3 0 0 1 1
 3 0 1 0 0   3 0 1 0 1   3 1 1 1 0   3 0 1 1 1
 4 0 0 0 0   4 0 0 0 1   4 0 0 1 0   4 0 0 1 1
 4 1 1 0 0   4 0 1 0 1   4 0 1 1 0   4 0 1 1 1
 5 0 0 0 0   5 1 0 0 1   5 0 0 1 0   5 0 0 1 1
 5 0 1 0 0   5 0 1 0 1   5 0 1 1 0   5 0 1 1 1
 6 0 0 0 0   6 0 0 0 1   6 0 0 1 0   6 0 0 1 1
 6 0 1 0 0   6 1 1 0 1   6 0 1 1 0   6 0 1 1 1
 7 0 0 0 0   7 1 0 0 1   7 0 0 1 0   7 0 0 1 1
 7 0 1 0 0   7 0 1 0 1   7 0 1 1 0   7 0 1 1 1
 8 0 0 0 0   8 0 0 0 1   8 0 0 1 0   8 0 0 1 1
 8 0 1 0 0   8 1 1 0 1   8 0 1 1 0   8 0 1 1 1
 9 0 0 0 0   9 0 0 0 1   9 0 0 1 0   9 0 0 1 1
 9 0 1 0 0   9 1 1 0 1   9 0 1 1 0   9 0 1 1 1
10 0 0 0 0   10 0 0 0 1   10 0 0 1 0   10 0 0 1 1
10 0 1 0 0   10 1 1 0 1   10 0 1 1 0   10 0 1 1 1
;
```

where SUBJECT is the subject number, CHOOSE is the status variable, and T is the time variable. Because this data set is arranged in a survival analysis form you can use the PROC PHREG. You can use the syntax:

```
proc phreg data=choco;
strata subject;
```

```

model t*choose(0)=dark soft nuts;
run;

```

As a result, you will have:

Sample Program: Conditional Logit Regression

The PHREG Procedure

```

Data Set: WORK.CHOCO
Dependent Variable: T
Censoring Variable: CHOOSE
Censoring Value(s): 0
Ties Handling: BRESLOW

```

Summary of the Number of Event and Censored Values

Stratum	SUBJECT	Total	Event	Censored	Percent Censored
1	1	8	1	7	87.50
2	2	8	1	7	87.50
3	3	8	1	7	87.50
4	4	8	1	7	87.50
5	5	8	1	7	87.50
6	6	8	1	7	87.50
7	7	8	1	7	87.50
8	8	8	1	7	87.50
9	9	8	1	7	87.50
10	10	8	1	7	87.50
Total		80	10	70	87.50

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L Score	41.589	28.727	12.862 with 3 DF (p=0.0049)
Wald	.	.	11.600 with 3 DF (p=0.0089)
	.	.	8.928 with 3 DF (p=0.0303)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Risk Ratio
DARK	1	1.386294	0.79057	3.07490	0.0795	4.000
SOFT	1	-2.197225	1.05409	4.34502	0.0371	0.111
NUTS	1	0.847298	0.69007	1.50762	0.2195	2.333

The result shows the estimation result as:

$$P_j = \frac{\exp(1.386294 * DARK_j - 2.197225 * SOFT_j + 0.847298 * NUTS_j)}{\sum_j \exp(1.386294 * DARK_j - 2.197225 * SOFT_j + 0.847298 * NUTS_j)} \text{ for } j = 1, \dots, 8$$

The positive parameter estimates of DARK and NUTS mean that dark and nuts each increases the preference. The negative parameter estimate of SOFT denotes soft center decreases the preference.

For each of eight types of candies, the predicted probabilities can be computed as follows:

Choice	DARK	SCFT	NUTS	$\exp(\beta x_j)$	Predicted Probability
1	0	0	0	1.000	0.054
2	0	0	1	2.333	0.126
3	0	1	0	0.111	0.006
4	0	1	1	0.259	0.014
5	1	0	0	4.000	0.216
6	1	0	1	9.333	0.504
7	1	1	0	0.444	0.024
8	1	1	1	1.037	0.056
			sum	18.518	1.000

This shows that the most preferred type of candy is the dark chocolate with a hard center and nuts.

Conditional Logit Regression with SPSS

COXREG Procedure

With SPSS, you can use the COXREG procedure to fit a conditional logit model. The basic syntax is:

```
coxreg time_varname with X1 X2
  /status=status_varname(1)
  /strata=strata_varname.
```

where *time_varname* is the name of the failure time variable (the smaller value means the alternative was chosen), *status_varname* is the name of the censoring indicator variable, of which 1 is the value to indicate the event has occurred (not censored), *strata_varname* is the name of variable to specify the variable that determines the stratification, and X1 and X2 are explanatory variables.

Example 29: SPSS Conditional Logit Regression in COXREG procedure

Using the data in Example 28, if you use:

```
coxreg t with dark soft nuts
  /status=choose(1)
  /strata=subject.
```

you will have the following SPSS output:

C O X R E G R E S S I O N

Categorical Analysis - Part 1

80 Total cases read
 0 Cases with missing values
 0 Valid cases with non-positive times
 0 Censored cases before the earliest event in a stratum
 0 Total cases dropped
 80 Cases available for the analysis

Dependent Variable: T

SUBJECT	Events	Censored
1.00	1	7 (87.5%)
2.00	1	7 (87.5%)
3.00	1	7 (87.5%)
4.00	1	7 (87.5%)
5.00	1	7 (87.5%)
6.00	1	7 (87.5%)
7.00	1	7 (87.5%)
8.00	1	7 (87.5%)
9.00	1	7 (87.5%)
10.00	1	7 (87.5%)
Total	10	70 (87.5%)

Beginning Block Number 0. Initial Log Likelihood Function

-2 Log Likelihood 41.589

Beginning Block Number 1. Method: Enter

Variable(s) Entered at Step Number 1..

DARK
 NUTS
 SOFT

Coefficients converged after 5 iterations.

-2 Log Likelihood 28.727

	Chi-Square	df	Sig
Overall (score)	11.600	3	.0089
Change (-2LL) from			
Previous Block	12.862	3	.0049
Previous Step	12.862	3	.0049

----- Variables in the Equation -----

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
DARK	1.3863	.7906	3.0749	1	.0795	.1608	4.0000
NUTS	.8473	.6901	1.5076	1	.2195	.0000	2.3333

SOFT	-2.1972	1.0541	4.3450	1	.0371	-.2375	.1111
------	---------	--------	--------	---	-------	--------	-------

Covariate Means

Variable	Mean
DARK	.5000
NUTS	.5000
SOFT	.5000

The estimation result is exactly the same as what you obtained with SAS.

Permission to use this document is granted so long as the author is acknowledged and notified.

Please send comments and suggestions to: statmath@indiana.edu

Copyright 1995-1999, [Indiana University](http://www.indiana.edu).

Last modified: Thursday, 24-Feb-2000 15:28:59 EST

URL /~statmath/stat/all/cat/printable.html