

# The QLIM Procedure

## Table of Contents

---

<b>OVERVIEW</b> . . . . .	3
<b>GETTING STARTED</b> . . . . .	4
<b>SYNTAX</b> . . . . .	9
Functional Summary . . . . .	9
PROC QLIM Statement . . . . .	10
BY Statement . . . . .	10
MODEL Statement . . . . .	11
ENDOGENOUS Statement . . . . .	15
HETERO Statement . . . . .	16
FREQ Statement . . . . .	17
RESTRICT Statement . . . . .	17
<b>DETAILS</b> . . . . .	18
Box-Cox Transformation and Heteroscedasticity . . . . .	18
Binary Discrete Choice Modeling . . . . .	20
Multinomial Discrete Choice Modeling . . . . .	21
Goodness-of-Fit Measures . . . . .	24
Limited Dependent Variable Models . . . . .	25
<b>EXAMPLES</b> . . . . .	29
Example 1 Ordered Data Modeling . . . . .	29
<b>REFERENCES</b> . . . . .	34



# The QLIM Procedure

---

## Overview

The QLIM (Qualitative and Limited dependent variable Model) procedure analyzes univariate and multivariate limited dependent variable models where dependent variables take discrete values or dependent variables are observed only in a limited range of values. This procedure includes logit, probit, tobit, and general simultaneous equations models. The simultaneous equations model can contain discrete choice and limited endogenous variables as well as continuous endogenous variables.

The QLIM procedure mainly uses the maximum likelihood (ML) method for the single equation model or reduced form equations of the simultaneous equations model. The structural parameters are estimated in the second stage using the least squares method.

The experimental QLIM procedure currently supports the following models:

- linear regression model with heteroscedasticity
- Box-Cox regression with heteroscedasticity
- binary probit and logit with heteroscedasticity
- ordinal probit and logit with heteroscedasticity
- simple multinomial logit
- conditional logit
- tobit (censored and truncated) with heteroscedasticity

The Box-Cox transformation of explanatory variables can be used for discrete choice and limited dependent variable models: binary logit/probit, ordinal logit/probit, and tobit. The multivariate and simultaneous equations models will be supported in a future release.

The MDC procedure supports unordered multinomial logit models, and the COUNTREG procedure estimates count data regression models.

---

## Getting Started

The QLIM procedure is similar in use to the other regression or simultaneous equations model procedures in the SAS<sup>®</sup> System. For example, the following statements are used to estimate a binary choice model using the logistic probability function:

```
proc qlim;
  model y = x1 / type=blogit;
  endogenous discrete=(y 0 1);
run;
```

The response variable, *y*, is numeric and has discrete values. PROC QLIM enables you to specify these binary values in the ENDOGENOUS statement. The ENDOGENOUS statement is not required for univariate discrete choice modeling. You can specify the binary probit model as follows:

```
model y = x1 / type=bprobit;
```

Multiple endogenous variables can be specified with one MODEL statement in the QLIM procedure when two models have the same exogenous variables:

```
model y1 y2 = x1 x2 / type=bprobit;
```

The preceding specification is equivalent to

```
proc qlim type=bprobit;
  model y1 = x1 x2;
  model y2 = x1 x2;
run;
```

When you estimate the conditional logit model, contrary to simple multinomial logit, the data must be arranged by choice. That is, each individual decision maker has an observation for each choice. An indicator variable is needed to identify the actual choice. Each individual is allowed to have a different number of choices. See the “Multinomial Discrete Choice Modeling” section for more details on multinomial choice models. For example, the conditional logit model can be specified using an identification variable, *id*, and a choice variable, *choose*. The indicator variable, *y*, is specified as a dependent variable. Note that for a conditional logit model, data set values for the dependent variable in the MODEL statement are binary and indicate which alternative is chosen among multiple choices. The CHOICE= option identifies the variable that contains all possible choices for each individual or subject:

```
model y = x1 x2 / type=clogit id=(id) choice=(choose);
```

The standard tobit model is estimated with the TYPE=TOBIT option. However, you must specify variables that contain limits of the dependent variable in the ENDOGENOUS statement when the data is limited by specific values. For example, the two-limit censored model requires two variables that contain the lower (*bottom*) and upper (*top*) bound.

```

proc qlim data=a type=tobit;
  model y = x1 x2 x3;
  endogenous censored=(lb=bottom ub=top y);
run;

```

The following example illustrates the use of PROC QLIM. The data are taken from Mroz (1987). This data set is based on a sample of 753 married white women. The dependent variable is a discrete variable of labor force participation (*lfp*). Explanatory variables are the number of children ages 5 or younger (*k5*), the number of children ages 6 to 18 (*k618*), the woman's age (*age*), a dummy variable for the wife's college education (*wc*), a dummy variable for the husband's college education (*hc*), the wife's wage estimate (*lwg*), and the family income excluding the wife's wage (*inc*).

```

data mroz;
  input lfp k5 k618 age wc hc lwg inc;
datalines;
  ... data lines are omitted ...
;
run;

proc qlim data=mroz;
  model lfp = k5 k618 age wc hc lwg inc / type=blogit;
run;

```

Results of this analysis are shown in the following four figures. PROC QLIM first lists the estimation summary table shown in Figure 1. Included are the dependent variable, the number of observations, the log-likelihood function value, the maximum absolute gradient, the number of iterations, the optimization method, AIC, and Schwarz criterion. By default, the QLIM procedure uses the Newton-Raphson optimization technique.

The QLIM Procedure	
Binary Logit Estimates	
Model Fit Summary	
Dependent Variable	lfp
Number of Observations	753
Log Likelihood	-452.63296
Maximum Absolute Gradient	2.09602E-6
Number of Iterations	5
Optimization Method	Newton-Raphson
AIC	921.26591
Schwarz Criterion	958.25844

**Figure 1.** Fit Summary Table of Binary Logit

In the second table, shown in Figure 2, PROC QLIM provides frequency information on each choice. In this example, 428 women participate in the labor force (*lfp*=0).

The QLIM Procedure			
Binary Logit Estimates			
Discrete Response Profile			
Index	lfp	Frequency	Percent
0	0	325	43.16
1	1	428	56.84

**Figure 2.** Choice Frequency Summary

Goodness-of-fit measures are displayed in Figure 3. All measures except McKelvey-Zavoina's definition are based on the log-likelihood function value. The likelihood ratio test statistic has chi-square distribution conditional on the null hypothesis that all slope coefficients are zero. In this example, the likelihood ratio statistic is used to test the hypothesis that  $k_5 = k_{618} = \text{age} = \text{age} = \text{wc} = \text{hc} = \text{lwg} = \text{inc} = 0$

The QLIM Procedure		
Binary Logit Estimates		
Goodness-of-Fit Measures for Discrete Choice Models		
Measure	Value	Formula
Likelihood Ratio (R)	124.48	$2 * (\text{LogL} - \text{LogL0})$
Upper Bound of R (U)	1029.7	$- 2 * \text{LogL0}$
Aldrich-Nelson	0.1419	$R / (R+N)$
Cragg-Uhler 1	0.1524	$1 - \exp(-R/N)$
Cragg-Uhler 2	0.2045	$(1 - \exp(-R/N)) / (1 - \exp(-U/N))$
Estrella	0.1615	$1 - (1 - R/U)^{(U/N)}$
Adjusted Estrella	0.1412	$1 - ((\text{LogL} - K) / \text{LogL0})^{(-2/N * \text{LogL0})}$
McFadden's LRI	0.1209	$R / U$
Veall-Zimmermann	0.2456	$(R * (U+N)) / (U * (R+N))$
McKelvey-Zavoina	0.4769	

N = # of observations, K = # of regressors

**Figure 3.** Likelihood Ratio Test and  $R^2$  Measures

Finally, the parameter estimates and standard errors are shown in Figure 4. All gradients are very small in magnitude, which means that the optimization algorithm is converged to the maximum likelihood value. Note that the log-likelihood function of the binary logit and probit models has a unique maximum.

The QLIM Procedure						
Binary Logit Estimates						
Parameter Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t	Gradient
Intercept	1	3.1821	0.6444	4.94	<.0001	6.044E-8
k5	1	-1.4629	0.1970	-7.43	<.0001	-1.39E-7
k618	1	-0.0646	0.0680	-0.95	0.3423	-4.01E-8
age	1	-0.0629	0.0128	-4.92	<.0001	2.096E-6
wc	1	0.8073	0.2300	3.51	0.0004	9.908E-8
hc	1	0.1117	0.2060	0.54	0.5876	6.605E-8
lwg	1	0.6047	0.1508	4.01	<.0001	1.884E-7
inc	1	-0.0344	0.008208	-4.20	<.0001	-1.13E-6

**Figure 4.** Parameter Estimates of Binary Logit

When the error term has a standard normal distribution, the binary probit models are estimated. The estimated parameters are shown in Figure 5. Note that parameter estimates are not equivalent to logit estimates, since the error variance of the logit model is different from that of probit.

The QLIM Procedure						
Binary Probit Estimates						
Parameter Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t	Gradient
Intercept	1	1.9184	0.3807	5.04	<.0001	-1.25E-7
k5	1	-0.8747	0.1136	-7.70	<.0001	8.245E-8
k618	1	-0.0386	0.0405	-0.95	0.3405	-7.76E-8
age	1	-0.0378	0.007609	-4.97	<.0001	-4.63E-6
wc	1	0.4883	0.1355	3.60	0.0003	-1.22E-7
hc	1	0.0572	0.1240	0.46	0.6448	-8.7E-8
lwg	1	0.3656	0.0878	4.17	<.0001	-2.62E-7
inc	1	-0.0205	0.004777	-4.30	<.0001	-6.15E-7

**Figure 5.** Parameter Estimates of Binary Probit

The heteroscedastic logit model can be estimated using the HETERO statement. If the variance of the logit model is a function of the family income level, the variance can be specified as

$$Var(\epsilon_i) = \exp(\theta inc_i)$$

The following SAS statements estimate the heteroscedastic logit model:

```
proc qlim data=mroz;
  model lfp = k5 k618 age wc hc lwg inc / type=blogit;
  hetero inc;
run;
```

The parameter estimate ( $\theta$ ) of the heteroscedasticity variable is listed as HET1; see Figure 6.

The QLIM Procedure						
Binary Logit Estimates with Heteroscedasticity						
Parameter Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t	Gradient
Intercept	1	4.3193	1.1190	3.86	0.0001	-3.09E-6
k5	1	-1.9313	0.3957	-4.88	<.0001	-4.61E-6
k618	1	-0.0844	0.0903	-0.94	0.3497	1.88E-6
age	1	-0.0824	0.0212	-3.89	0.0001	-0.00006
wc	1	1.2208	0.4317	2.83	0.0047	4.554E-6
hc	1	0.2060	0.2871	0.72	0.4731	4.563E-6
lwg	1	0.7968	0.2363	3.37	0.0007	3.901E-6
inc	1	-0.0580	0.0201	-2.89	0.0038	-0.00009
HET1	1	0.0302	0.0185	1.63	0.1026	-0.00035

**Figure 6.** Parameter Estimates of Binary Logit with Heteroscedasticity

---

## Syntax

The QLIM procedure is controlled by the following statements:

```

PROC QLIM options ;
  BY variables ;
  MODEL dependent variables = regressors / options ;
  FREQ variable ;
  ENDOGENOUS variables <DISCRETE=> <CENSORED=> <TRUN-
  CATED=> ;
  HETERO variables / options ;
  RESTRICT options ;
  OUTPUT options ;

```

---

## Functional Summary

The statements and options used with the QLIM procedure are summarized in the following table:

Description	Statement	Option
<b>Data Set Options</b>		
specify the input data set	QLIM	DATA=
write parameter estimates to an output data set	QLIM	OUTEST=
write predictions to an output data set	OUTPUT	OUT=
<b>Declaring the Role of Variables</b>		
specify BY-group processing	BY	
<b>Group Frequency Information</b>		
specify a frequency variable for grouped data	FREQ	
<b>Printing Control Options</b>		
request all printing options	MODEL	ALL
print correlation matrix of the estimates	MODEL	CORRB
print covariance matrix of the estimates	MODEL	COVB
<b>Model Estimation Options</b>		
specify options specific to Box-Cox transformation	MODEL	BOXCOX=()
specify a choice variable for conditional logit	MODEL	CHOICE=()
specify the type of covariance matrix	MODEL	COVEST=
specify the ID variable	MODEL	ID=()
set the initial values of parameters used by the iterative optimization algorithm	MODEL	INITIAL=()
specify a restriction on the first threshold parameter of the ordinal probit model	MODEL	LIMIT1=

Description	Statement	Option
specify maximum number of iterations	MODEL	MAXITER=
specify the estimation method	MODEL	METHOD=
specify number of choices for each person	MODEL	NCHOICE=
suppress the intercept parameter	MODEL	NOINT
specify the optimization technique	MODEL	OPTMETHOD=
specify that initial values are generated using random numbers	MODEL	RANDOMINIT
specify that the dependent variable contains rank data	MODEL	RANK
specify options for restarting optimization process	MODEL	RESTART=
specify a seed for pseudo-random number generation	MODEL	SEED=
specify the type of the model	MODEL	TYPE=
<b>Heteroscedasticity Model Options</b>		
estimate heteroscedasticity models	HETERO	LINK=
<b>Output Control Options</b>		
output predicted values	OUTPUT	P=

---

## PROC QLIM Statement

### **PROC QLIM** *options* ;

The following options can be used in the PROC QLIM statement:

#### **DATA= SAS-data-set**

specifies the input SAS data set. If the DATA= option is not specified, PROC QLIM uses the most recently created SAS data set.

#### **OUTEST= SAS-data-set**

writes the parameter estimates to an output data set.

In addition, any of the following MODEL statement options can be specified in the PROC QLIM statement, which is equivalent to specifying the option for every MODEL statement: ALL, CORRB, COVB, COVEST=, ID=, ITPRINT, MAXITER=, NOINT, NOPRINT, OPTMETHOD=, RANDOMINIT=, RANK, RESTART=, SEED=, and TYPE=.

---

## BY Statement

### **BY** *variables* ;

A BY statement can be used with PROC QLIM to obtain separate analyses on observations in groups defined by the BY variables.

---

## MODEL Statement

**MODEL** *dependent = regressors / options ;*

The MODEL statement specifies the dependent variable and independent regressor variables for the regression model.

The following options can be used in the MODEL statement after a slash (/).

**CHOICE=(*variable*)**

specifies the variable that contains possible choices for each individual when the conditional logit model is estimated.

**ID=(*variable*)**

specifies the identification variable when there are multiple choice-specific observations.

**LIMIT1=*value***

specifies the restriction of the threshold value of the first category when the ordinal probit or logit model is estimated. LIMIT1=ZERO is the default option. When LIMIT1=VARYING is specified, the threshold value is estimated.

**NCHOICE=*number***

specifies the number of choices for the conditional choice model when all individuals have the same choice set. The NCHOICE= and CHOICE= options must not be used simultaneously.

**NOINT**

suppresses the intercept parameter.

**RANDOMINIT**

**RANDOMINIT=*number***

specifies that initial parameter values are perturbed by uniform pseudo-random numbers for numerical optimization of the objective function. The default is  $U(-1, 1)$ . When the RANDINIT= $r$  option is specified,  $U(-r, r)$  pseudo-random numbers are generated. The value  $r$  should be positive. With a RANDINIT or RANDINIT= option, there are pure random searches for a given number of trials (1000) to get a maximum (or minimum) value of the objective function. For example, when there is a parameter estimate with an initial value of 1, the RANDINIT option will add a generated random number  $u$  to the initial value and compute an objective function value using  $1 + u$ . This option is helpful in finding the initial value automatically if there is no guidance in setting the initial estimate.

**RANK**

specifies that the dependent variable contains ranks. The numbers must be positive integers starting from 1. When the dependent variable has a value of 1, the corresponding alternative is chosen.

**TYPE= *value***

specifies a type of model to be analyzed. The supported model types are

LOGIT	specifies a general logit model
PROBIT	specifies a general probit model
BINOMLOGIT   BLOGIT	specifies a binomial logit model
BINOMPROBIT   BPROBIT	specifies a binomial probit model
ORDINALPROBIT   OPROBIT	specifies an ordinal probit model
MULTINOMLOGIT   MLOGIT	specifies a simple multinomial logit model
CONDITIONLOGIT   CLOGIT	specifies a conditional logit model
TOBIT	specifies a Tobit model

### **BOXCOX Estimation Options**

#### **BOXCOX= (option-list)**

specifies options that are used for Box-Cox regression or regressor transformation. The Box-Cox regression with heteroscedasticity is specified as

```
model y = x1 x2 / boxcox=(bcxparm(1)=y,bcxparm(2)=x1 x2)
hetero z1 z2 / link=exp;
```

PROC QLIM estimates the following Box-Cox regression model:

$$y_i^{(\lambda_1)} = \beta_0 + \beta_1 x_{1i}^{(\lambda_2)} + \beta_2 x_{2i}^{(\lambda_2)} + \epsilon_i$$

$$V(\epsilon_i) = \sigma^2 \exp(\gamma_1 z_{1i} + \gamma_2 z_{2i})$$

When the specific Box-Cox parameter is set equal to a constant, you can use the BCXCONSTANT(=) option. For example, you may want to set the transformation parameter of the dependent variable to 0 instead of estimating it in the model.

$$\log(y_i) = \beta_0 + \beta_1 x_{1i}^{(\lambda_2)} + \beta_2 x_{2i}^{(\lambda_2)} + \epsilon_i$$

```
model y = x1 x2 / boxcox=(bcxconstant(0)=y,bcxparm(2)=x1 x2)
```

The estimate of the Box-Cox parameter,  $\lambda_2$ , is listed as BCX2 in the output. If you want to name it BCX3, you must specify

```
model y = x1 x2 / boxcox=(bcxconstant(0)=y,bcxparm(3)=x1 x2)
```

Interaction terms can also be specified as follows:

$$y_i = \beta_0 + \beta_1 x_{1i}^{(\lambda_2)} + \beta_2 x_{2i}^{(\lambda_2)} + \beta_3 x_{4i}^{(\lambda_3)} x_{2i} \epsilon_i$$

```
model y = x1 x2 /
boxcox=(bcxparm(2)=x1 x2,bcxparm(3)=multf(x4,x2))
```

**BCXCONSTANT**(*number*)= (*variables*)

**BCXCONST**(*number*)= (*variables*)

specifies the value of the fixed Box-Cox parameter and relevant variables.

**BCXLIMIT**=(*value1 value2*)

specifies lower and upper bounds of Box-Cox transformation parameter estimates.

The magnitude of VALUE1 and VALUE2 must be chosen carefully to avoid numerical errors. It would be better to re-scale the variable that contains extreme values.

**BCXPARAMETER**(*number*)= (*variables*)

**BCXPARM**(*number*)= (*variables*)

specifies Box-Cox parameter index and relevant variables.

The interaction terms are specified as

MULTF(*variable, variable*) specifies an interaction term with the first variable transformed

MULTS(*variable, variable*) specifies an interaction term with the second variable transformed

MULTB(*variable, variable*) specifies an interaction term with both variables transformed

At least one variable must be transformed. When both variables are transformed, the same transformation parameter will be used. For example, you can add a new interaction term to your regressors in the following ordinal probit model:

$$y_i^* = \beta_0 + \beta_1 x_{1i}^{(\lambda_2)} + \beta_2 x_{2i}^{(\lambda_2)} + \beta_3 x_{3i}^{(\lambda_3)} x_{4i}^{(\lambda_3)} + \epsilon_i$$

where

$$y_i = j \text{ if } \mu_{j-1} < y_i^* \leq \mu_j, \quad \epsilon_i \sim N(0, 1)$$

See the “Ordinal Probit/Logit” section for more details on ordinal response models. To estimate this model, you need to specify the following SAS statement:

```
model y = x1 x2 / type=oprobit
      boxcox=(bcxparm(2)=x1 x2,bcxparm(3)=multb(x3,x4));
```

### Restart Options

**RESTART**=(*option-list*)

specifies options that are used for the reiteration of the optimization routine. It would be better to get an optimum solution using perturbations once you reach the optimum point. When the ADDRANDOM option is specified, the initial value of reiteration is computed using random grid searches around the initial solution.

```
model y = x1 x2 / type=oprobit
      restart=(addvalue=(.01 .01 .01 .001 .001));
hetero z1 z2 / link=exp;
```

The preceding SAS statement re-estimates a heteroscedastic ordinal probit model by adding ADDVALUE= values. If the ADDVALUE= option contains missing values, the restart option uses the corresponding estimate in the initial stage. If both the ADDVALUE= and ADDRANDOM= options are specified, ADDVALUE= is ignored.

The following options can be used in the RESTART=( ) option. The options are listed within parentheses and separated by commas.

**ADDMAXIT=number**

specifies the number of maximum iterations for the second stage of estimation.

**ADDRANDOM**

**ADDRANDOM=value**

specifies random added values to the estimates in the initial stage. With the ADDRANDOM option,  $U(-1, 1)$  random numbers are created and added to the estimate obtained in the initial stage. When the ADDRANDOM= $r$  option is specified, the uniform random numbers,  $U(-r, r)$ , are generated. The restart initial value is determined based on the given number of random searches.

**ADDVALUE=(value-list)**

specifies added values to the estimates in the initial stage. The missing element is considered as a non-added value for the corresponding estimate. When the ADDVALUE= option is not specified, default values are added.

**Printing Options**

**ALL**

requests all printing options.

**CORRB**

prints the estimated correlation matrix of the parameter estimates.

**COVB**

prints the estimated covariance matrix of the parameter estimates.

**COVEST=value**

The COVEST= option specifies the type of covariance matrix. When COVEST=OP is specified, the outer product matrix is used to compute the covariance matrix of the parameter estimates. The COVEST=HESSIAN option produces the covariance matrix using the inverse Hessian matrix. The quasi-maximum likelihood estimates are computed with COVEST=QML. The default is COVEST=HESSIAN when the Newton-Raphson method is used. COVEST=OP is the default when the OPTMETHOD=QN option is specified.

**ITPRINT**

prints the objective function and parameter estimates at each iteration. The objective function is the full log likelihood function for the maximum likelihood method.

**NOPRINT**

suppresses all printed output.

**Estimation Control Options**

**INITIAL= ( initial-values )**

**START=** (*initial-values*)

specifies initial values for some or all of the parameter estimates. The values specified are assigned to model parameters in the same order as the parameter estimates are printed in the QLIM procedure output. The order of values in the INITIAL=option is: the intercept, the regressor coefficients, and additional parameters.

When you use the INITIAL= option, the initial values in the INITIAL= option should satisfy the restrictions specified for the parameter estimates. If they do not, the initial values you specify are adjusted to satisfy the restrictions.

**MAXITER=** *number*

sets the maximum number of iterations allowed. The default is MAXITER=100.

**OPTMETHOD=** *value*

The OPTMETHOD= option specifies the optimization technique when the estimation method uses non-linear optimization. The OPTMETHOD=QN option specifies the quasi-Newton method. The OPTMETHOD=NR option specifies the Newton-Raphson method. The OPTMETHOD=TR option specifies the trust region method. The default is OPTMETHOD=NR.

---

## ENDOGENOUS Statement

**ENDOGENOUS** *variables DISCRETE=(options) CENSORED=(options) TRUNCATED=(options) ;*

The ENDOGENOUS statement specifies types of endogenous variables. When the SYSTEM option is specified, the ENDOGENOUS statement must be provided. It is required that all LHS variables in the MODEL statement should be listed in the ENDOGENOUS statement. Continuous variables can also be listed in the ENDOGENOUS statement.

**CENSORED=(variables)****CENSORED=(LB=variable UB=variable)**

specifies censored variables. The LB= option specifies the variable that contains the left or lower censoring point, and the UB= option specifies the right or upper censoring point. When neither the LB= nor UB= option is specified, the default censoring point ( $y_i^* > 0$ ) is used.

**DISCRETE=(variables)****DISCRETE=(variable value)**

specifies discrete choice variables with their choice values. However, the choice values can be omitted.

**TRUNCATED=(variables)****TRUNCATED=(LB= UB= variable)**

specifies truncated variables. The LB= option specifies the variable that contains the left or lower truncation point, and the UB= option specifies the right or upper truncation point. When neither the LB= nor UB= option is specified, the default truncation point ( $y_i^* > 0$ ) is used.

---

## HETERO Statement

**HETERO** *variables* < / link= >;

The HETERO statement specifies variables that are related to the heteroscedasticity of the residuals and the way these variables are used to model the error variance. The heteroscedastic regression model supported by PROC QLIM is

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma_i^2)$$

See the “Heteroscedasticity” section for more details on the specification of functional forms.

### LINK=(value)

The functional form can be specified using the LINK= option. The following option values are allowed:

EXP	specifies exponential link function
LINEAR	specifies linear link function

When the LINK= option is not specified, the exponential link function is specified as follows:

$$\sigma_i^2 = \sigma^2 \exp(\mathbf{z}_i' \boldsymbol{\gamma})$$

### SQUARE

estimates the model using the square of exponential or linear heteroscedasticity function. For example, you can specify the following heteroscedasticity function:

$$\sigma_i^2 = \sigma^2 (\exp(\mathbf{z}_i' \boldsymbol{\gamma}))^2$$

```
model y = x1 x2 / type=blogit;
hetero z1 / link=exp square;
```

When the dependent variable is continuous, the HETERO statement estimates the regression model with heteroscedasticity using the maximum likelihood method. For example, the heteroscedastic logit model can be estimated using the following statement:

```
model y = x1 x2 / type=blogit;
hetero z1;
```

---

## FREQ Statement

**FREQ** *variable* ;

The variable in the FREQ statement identifies a variable that contains the frequency of occurrence of each observation. PROC QLIM treats each observation as if it appears  $n$  times, where  $n$  is the value of the FREQ variable for the observation. If it is not an integer, the frequency value is truncated to an integer. If the frequency value is less than 1 or missing, the observation is not used in the model fitting. When the FREQ statement is not specified, each observation is assigned a frequency of 1.

---

## RESTRICT Statement

**RESTRICT** *option* ;

The RESTRICT statement specifies simple parameter restrictions. The sequence of elements in the FIXEDPARAM=, LBOUND=, and UBOUND= option must correspond to the printed sequence of parameter estimates. A RESTRICT statement can be specified for each MODEL statement.

**FIXEDPARAMETER=** (*value-list*)

**FIXEDPARM=** (*value-list*)

specifies the fixed value of parameters. When the LBOUND= or UBOUND= option is specified, the values specified in the FIXEDPARM= option must satisfy the specified boundary conditions.

**LOWERBOUND=** (*value-list*)

**LBOUND=** (*value-list*)

specifies the lower bound of parameters. When there is a FIXEDPARM= option present and the corresponding element in the FIXEDPARM= option does not have a missing value, the relevant element of the LBOUND= option is ignored.

**UPPERBOUND=** (*value-list*)

**UBOUND=** (*value-list*)

specifies the upper bound of parameters. When there is a FIXEDPARM= option present and the corresponding element in the FIXEDPARM= option does not have a missing value, the relevant element of the UBOUND= option is ignored.

**ALL**

specifies that the single element of the FIXEDPARM=, LBOUND=, and UBOUND= options is expanded to all parameters. For example, the model with four parameters can have non-negative boundary constraints if the following RESTRICT statement is specified:

```
restrict lbound=(0) / all;
```

However, only the first parameter is bounded below by 0 if the ALL option is not specified.

---

## Details

---

### Box-Cox Transformation and Heteroscedasticity

#### Heteroscedasticity

If the variance of regression disturbance ( $\epsilon_i$ ) is heteroscedastic, the variance can be specified as a function of variables

$$E(\epsilon_i^2) = \sigma_i^2 = f(\mathbf{z}_i' \boldsymbol{\gamma})$$

The functional form of heteroscedasticity is modeled using one of the following specifications:

$$f(\mathbf{z}_i' \boldsymbol{\gamma}) = \sigma^2 \exp(\mathbf{z}_i' \boldsymbol{\gamma})$$

$$f(\mathbf{z}_i' \boldsymbol{\gamma}) = \sigma^2 \exp(\mathbf{z}_i' \boldsymbol{\gamma})^2$$

$$f(\mathbf{z}_i' \boldsymbol{\gamma}) = \sigma^2 \left(1 + \sum_{l=1}^L \gamma_l^2 z_{li}\right)$$

$$f(\mathbf{z}_i' \boldsymbol{\gamma}) = \sigma^2 \left(1 + \sum_{l=1}^L \gamma_l^2 z_{li}\right)^2$$

However,  $\sigma^2$  is normalized ( $\sigma^2 = 1$ ) for discrete choice models since this parameter is not identified.

The heteroscedastic regression model is estimated using the following log-likelihood function:

$$\ell = -\frac{N}{2} \ln(2\pi) - \sum_{i=1}^N \frac{1}{2} \ln(\sigma_i^2) - \frac{1}{2} \sum_{i=1}^N \left(\frac{e_i}{\sigma_i}\right)^2$$

where  $e_i = y_i - \mathbf{x}_i' \boldsymbol{\beta}$ .

#### Box-Cox Modeling

Let a transformation function  $T(\cdot)$  be defined as follows:

$$T(\lambda, \kappa, x_1, x_2) = \begin{cases} x_1^{(\lambda)} \cdot x_2 & \text{if } \kappa = F \\ x_1 \cdot x_2^{(\lambda)} & \text{if } \kappa = S \\ x_1^{(\lambda)} \cdot x_2^{(\lambda)} & \text{if } \kappa = B \end{cases}$$

where

$$x^{(\lambda)} = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \ln(x) & \text{if } \lambda = 0 \end{cases}$$

Note that

$$x^{(\lambda)} = \ln(x) + \frac{\lambda}{2!} \ln(x)^2 + \frac{\lambda^2}{3!} \ln(x)^3 + \dots$$

Therefore, it can be shown that  $x^{(0)} = \ln(x)$ .

The Box-Cox regression model with interaction terms and heteroscedasticity is written

$$\begin{aligned} y_i^{(\lambda_0)} &= \beta_0 + \sum_{k=1}^K \beta_k x_{ki}^{(\lambda_k)} + \sum_{m=1}^M \gamma_m T(\theta_m, \kappa_m, w_{mi}, z_{mi}) + \epsilon_i \\ &= \mu_i + \epsilon_i \end{aligned}$$

where  $\epsilon \sim N(0, \sigma_i^2)$  and transformed variables must be positive. The variables ( $w_{mi}$  and  $z_{mi}$ ) in the interaction terms can be model regressors ( $x_{ki}$ ). In practice, too many transformation parameters cause numerical problems in model fitting. It would be desirable for the magnitude of transformed variables to be in the tolerable range if the corresponding transformation parameters are  $|\lambda| > 1$ .

The log-likelihood function of the Box-Cox regression model is written

$$\ell = -\frac{N}{2} \ln(2\pi) - \sum_{i=1}^N \ln(\sigma_i) - \frac{1}{2\sigma_i^2} \sum_{i=1}^N e_i^2 + (\lambda_0 - 1) \sum_{i=1}^N \ln(y_i)$$

where  $e_i = y_i^{(\lambda_0)} - \mu_i$ .

When the dependent variable is transformed, the original dependent variable must be truncated so that the Box-Cox transformation is well-defined. Therefore, the transformed variable is also truncated:

$$L < y_i^{(\lambda_0)} < R$$

where  $L = -\infty$  and  $R = -1/\lambda_0$  if  $\lambda_0 < 0$ ;  $L = -1/\lambda_0$  and  $R = \infty$  if  $\lambda_0 > 0$ . The correct log-likelihood function that satisfies regularity condition is

$$\ell^c = \ell - \sum_{i=1}^N \ln[\Phi(R_i) - \Phi(L_i)]$$

where  $L_i = (L - \mu_i)/\sigma_i$  and  $R_i = (R - \mu_i)/\sigma_i$ . The truncated Box-Cox regression model adds more complication in estimating parameters, though the truncated likelihood function does not provide that much advantage over non-truncated Box-Cox regression. Therefore, the Box-Cox regression model is estimated using the uncorrected log-likelihood function ( $\ell$ ).

When the dependent variable is discrete, censored, or truncated, the Box-Cox transformation is only applied to explanatory variables.

---

## Binary Discrete Choice Modeling

### **Probit and Logit Model**

The binary choice model is written

$$y_i^* = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$$

where the sign of the dependent variable is only observed as follows:

$$\begin{aligned} y_i &= 1 && \text{if } y_i^* > 0 \\ &= 0 && \text{otherwise} \end{aligned}$$

The disturbance,  $\epsilon_i$ , of the probit model has standard normal distribution with the distribution function (CDF)

$$\Phi(x) = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} \exp(-t^2/2) dt$$

The disturbance of the logit model has standard logistic distribution with the CDF

$$\Lambda(x) = \frac{\exp(x)}{1 + \exp(x)} = \frac{1}{1 + \exp(-x)}$$

The binary discrete choice model has the following probability that the event  $\{y = 1\}$  occurs:

$$P(y_i = 1) = \begin{cases} \Phi(\mathbf{x}_i' \boldsymbol{\beta}) & \text{(probit)} \\ \Lambda(\mathbf{x}_i' \boldsymbol{\beta}) & \text{(logit)} \end{cases}$$

The log-likelihood function is written

$$\ell = \sum_{i=1}^N \{y_i \log[F(\mathbf{x}_i' \boldsymbol{\beta})] + (1 - y_i) \log[1 - F(\mathbf{x}_i' \boldsymbol{\beta})]\}$$

where the CDF  $F(x)$  is defined as  $\Phi(x)$  for the probit model while  $F(x) = \Lambda(x)$  for logit. The first and second derivatives of the logit model are

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N (y_i - \Lambda(\mathbf{x}_i' \boldsymbol{\beta})) \mathbf{x}_i$$

$$\frac{\partial^2 \ell}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}'} = - \sum_{i=1}^N \Lambda(\mathbf{x}_i' \boldsymbol{\beta}) (1 - \Lambda(\mathbf{x}_i' \boldsymbol{\beta})) \mathbf{x}_i \mathbf{x}_i'$$

The probit model has more complicated derivatives

$$\frac{\partial \ell}{\partial \beta} = \sum_{i=1}^N \left[ \frac{(2y_i - 1)\phi((2y_i - 1)\mathbf{x}'_i\beta)}{\Phi(\mathbf{x}'_i\beta)} \right] \mathbf{x}_i = \sum_{i=1}^N r_i \mathbf{x}_i$$

$$\frac{\partial^2 \ell}{\partial \beta \partial \beta'} = - \sum_{i=1}^N r_i (r_i + \mathbf{x}'_i\beta) \mathbf{x}_i \mathbf{x}'_i$$

where

$$r_i = \frac{(2y_i - 1)\phi((2y_i - 1)\mathbf{x}'_i\beta)}{\Phi(\mathbf{x}'_i\beta)}$$

Note that logit maximum likelihood estimates are greater than probit maximum likelihood estimates by approximately  $\frac{\pi}{\sqrt{3}}$ , since the probit parameter estimates ( $\beta$ ) are standardized and the error term with logistic distribution has a variance of  $\frac{\pi^2}{3}$ .

---

## Multinomial Discrete Choice Modeling

When the dependent variable takes multiple discrete values, multinomial discrete choice modeling can be used to analyze the data. Ordinal choice models are explained in the following “Ordinal Probit/Logit” section. Unordered multinomial data is analyzed using the probit or logit link function. However, the multinomial probit model requires burdensome computation since multi-dimensional integration is involved when the likelihood function is computed. In addition, the multinomial probit model fits more parameters compared to multinomial logit models. Therefore, multinomial logit models are used frequently, though multinomial logit models are derived from the random utility function whose random component is more restrictively defined than the multinomial probit model. Let the random utility function be defined as

$$U_{ij} = V_{ij} + \epsilon_{ij}$$

where  $V_{ij}$  is a non-stochastic utility function and  $\epsilon_{ij}$  is a random component. If you assume that  $V_{ij}$  has a linear utility function, then  $V_{ij} = \mathbf{x}'_{ij}\beta$ . With most restrictive assumptions of the random component of the utility, the conditional logit model is derived. For conditional logit models, the error disturbances are assumed to have type I extreme value distribution with the distribution function,  $\exp(-\exp(-\epsilon_{ij}))$ . The event of selecting an alternative,  $\{y_i = j\}$ , can be expressed in terms of a random utility function as follows:

$$U_{ij} > \max_{k \in C_i, k \neq j} U_{ik}$$

Using properties of the type I extreme value distribution, the probability of choosing an alternative  $j$  among  $n_i$  choices of individual  $i$  can be written

$$\begin{aligned} P_i(j) &= P[\mathbf{x}'_{ij}\beta + \epsilon_{ij} \geq \max_{k \in C_i} (\mathbf{x}'_{ik}\beta + \epsilon_{ik})] \\ &= \frac{\exp(\mathbf{x}'_{ij}\beta)}{\sum_{k \in C_i} \exp(\mathbf{x}'_{ik}\beta)} \end{aligned}$$

**Ordinal Probit/Logit**

When the dependent variable is observed in sequence with  $M$  categories, binary discrete choice modeling is not appropriate for data analysis. McKelvey and Zavoina (1975) proposed the ordinal (or ordered) probit model.

Consider the following regression equation:

$$y_i^* = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$$

where error disturbances,  $\epsilon_i$ , have the distribution function  $F$ . The unobserved continuous random variable,  $y_i^*$ , is identified as  $M$  categories. Suppose there are  $M + 1$  real numbers,  $\mu_0, \dots, \mu_M$ , where  $\mu_0 = -\infty$ ,  $\mu_1 = 0$ ,  $\mu_M = \infty$ , and  $\mu_0 \leq \mu_1 \leq \dots \leq \mu_M$ . Define that

$$R_{i,j} = \mu_j - \mathbf{x}_i' \boldsymbol{\beta}$$

The probability that the unobserved dependent variable is contained in the  $j$ th category can be written

$$P[\mu_{j-1} < y_i^* \leq \mu_j] = F(R_{i,j}) - F(R_{i,j-1})$$

The log-likelihood function is

$$\ell = \sum_{i=1}^N \sum_{j=1}^M d_{ij} \log [F(R_{i,j}) - F(R_{i,j-1})]$$

where

$$d_{ij} = \begin{cases} 1 & \text{if } \mu_{j-1} < y_i \leq \mu_j \\ 0 & \text{otherwise} \end{cases}$$

The first derivatives are written

$$\frac{\partial \ell}{\partial \boldsymbol{\beta}} = \sum_{i=1}^N \sum_{j=1}^M d_{ij} \left[ \frac{f(R_{i,j-1}) - f(R_{i,j})}{F(R_{i,j}) - F(R_{i,j-1})} \mathbf{x}_i \right]$$

$$\frac{\partial \ell}{\partial \mu_k} = \sum_{i=1}^N \sum_{j=1}^M d_{ij} \left[ \frac{\delta_{j,k} f(R_{i,j}) - \delta_{j-1,k} f(R_{i,j-1})}{F(R_{i,j}) - F(R_{i,j-1})} \right]$$

where  $f(x) = \frac{dF(x)}{dx}$  and  $\delta_{j,k} = 1$  if  $j = k$ . When the ordinal probit is estimated, it is assumed that  $F(R_{i,j}) = \Phi(R_{i,j})$ . The ordinal logit model is estimated if  $F(R_{i,j}) = \Lambda(R_{i,j})$ . The first threshold parameter,  $\mu_1$ , is estimated when the LIMIT1=VARYING option is specified. By default (LIMIT1=ZERO),  $M - 2$  threshold parameters ( $\mu_2, \dots, \mu_{M-1}$ ) are estimated.

The ordered probit models are analyzed by Aitchison and Silvey (1957), and Cox (1970) discussed ordered response data using the logit model. They defined the probability that  $y_i^*$  belongs to  $j$ th category as

$$P[\mu_{j-1} < y_i \leq \mu_j] = F(\mu_j + \mathbf{x}_i' \boldsymbol{\theta}) - F(\mu_{j-1} + \mathbf{x}_i' \boldsymbol{\theta})$$

where  $\mu_0 = -\infty$  and  $\mu_M = \infty$ . Therefore, the ordered response model analyzed by Aitchison and Silvey can be estimated if the LIMIT1=VARYING option is specified. Note that  $\boldsymbol{\theta} = -\boldsymbol{\beta}$ .

### **Multinomial and Conditional Logit**

When explanatory variables contain only individual characteristics, the simple multinomial logit model is defined as

$$P[y_i = j] = P_{ij} = \frac{\exp(\mathbf{x}_i' \boldsymbol{\beta}_j)}{\sum_{k=0}^M \exp(\mathbf{x}_i' \boldsymbol{\beta}_k)} \quad \text{for } j = 0, \dots, M$$

For model identification, we assume that  $\boldsymbol{\beta}_0 = 0$ . The simple multinomial logit model is reduced to the binary logit model if  $M = 1$ . The log-odds ratio of alternative  $j$  and  $k$  is

$$\ln \left[ \frac{P_{ij}}{P_{ik}} \right] = \mathbf{x}_i' (\boldsymbol{\beta}_j - \boldsymbol{\beta}_k)$$

This type of simple multinomial choice modeling has a couple of weaknesses: it has too many parameters and it is difficult to interpret. The log-likelihood function of the simple multinomial logit model is written

$$\ell = \sum_{i=1}^N \sum_{j=0}^M d_{ij} \ln P[y_i = j]$$

where

$$d_{ij} = \begin{cases} 1 & \text{if individual } i \text{ chooses an alternative } j \\ 0 & \text{otherwise} \end{cases}$$

The conditional logit model is similarly defined when the outcome-varying data,  $\mathbf{x}_{ik}$ , is available.

$$P[y_i = j] = \frac{\exp(\mathbf{x}_{ij}' \boldsymbol{\beta})}{\sum_{k \in C_i} \exp(\mathbf{x}_{ik}' \boldsymbol{\beta})}$$

where there are  $n_i$  choices in each individual's choice set,  $C_i$ . The log-likelihood function is written

$$\ell = \sum_{i=1}^N \sum_{j \in C_i} d_{ij} \ln P(y_i = j)$$

Using properties of type I extreme value distribution, the probability of choosing an alternative  $j$  from  $n_i$  choices of individual  $i$  can be defined as follows:

$$P_i(j) = P[\mathbf{x}'_{ij}\boldsymbol{\beta} + \epsilon_{ij} > \max_{k \in C_i, k \neq j}(\mathbf{x}'_{ik}\boldsymbol{\beta} + \epsilon_{ik})] = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta})}{\sum_{k \in C_i} \exp(\mathbf{x}'_{ik}\boldsymbol{\beta})}$$

The problematic aspect of the conditional logit model lies in the independence from irrelevant alternatives (IIA) property. The IIA problem can be explained using the probability ratio of any two choices.

$$\frac{P_i(j)}{P_i(l)} = \frac{\exp(\mathbf{x}'_{ij}\boldsymbol{\beta}) / \sum_{k \in C_i} \exp(\mathbf{x}'_{ik}\boldsymbol{\beta})}{\exp(\mathbf{x}'_{il}\boldsymbol{\beta}) / \sum_{k \in C_i} \exp(\mathbf{x}'_{ik}\boldsymbol{\beta})} = \exp[(\mathbf{x}_{ij} - \mathbf{x}_{il})'\boldsymbol{\beta}]$$

It is evident that the probability ratio is only affected by choices  $j$  and  $l$ . Note that this IIA property is caused by an assumption of an independent and identical distribution of the random utility function.

---

## Goodness-of-Fit Measures

McFadden (1974) suggested a likelihood ratio index that is analogous to the  $R^2$  in the linear regression model.

$$R_M^2 = 1 - \frac{\ln L}{\ln L_0}$$

where  $L$  is the value of the maximum likelihood function at the maximum and  $L_0$  is a likelihood function when regression coefficients except an intercept term are zero. McFadden's likelihood ratio index is bounded by 0 and 1.

Estrella (1998) proposes the following requirements for a goodness-of-fit measure to be desirable in discrete choice modeling:

- The measure must take values in  $[0, 1]$ , where 0 represents no fit and 1 corresponds to perfect fit.
- The measure should be directly related to the valid test statistic for significance of all slope coefficients.
- The derivative of the measure with respect to the test statistic should comply with corresponding derivatives in a linear regression.

Estrella's measure is written

$$R_{E1}^2 = 1 - \left( \frac{\ln L}{\ln L_0} \right)^{-\frac{2}{N} \ln L_0}$$

Estrella suggests an alternative measure

$$R_{E2}^2 = 1 - [(\ln L - K) / \ln L_0]^{-\frac{2}{N} \ln L_0}$$

where  $\ln L_0$  is computed with null slope parameter values,  $N$  is the number observations used, and  $K$  represents the number of estimated parameters.

Other goodness-of-fit measures are summarized as follows:

$$R_{CU1}^2 = 1 - \left( \frac{L_0}{L} \right)^{\frac{2}{N}} \quad (\text{Cragg-Uhler 1})$$

$$R_{CU2}^2 = \frac{1 - (L_0/L)^{\frac{2}{N}}}{1 - L_0^{\frac{2}{N}}} \quad (\text{Cragg-Uhler 2})$$

$$R_A^2 = \frac{2(\ln L - \ln L_0)}{2(\ln L - \ln L_0) + N} \quad (\text{Aldrich-Nelson})$$

$$R_{VZ}^2 = R_A^2 \frac{2 \ln L_0 - N}{2 \ln L_0} \quad (\text{Veall-Zimmermann})$$

$$R_{MZ}^2 = \frac{\sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2}{N + \sum_{i=1}^N (\hat{y}_i - \bar{\hat{y}})^2} \quad (\text{McKelvey-Zavoina})$$

where  $\hat{y}_i = \mathbf{x}_i' \hat{\boldsymbol{\beta}}$  and  $\bar{\hat{y}} = \sum_{i=1}^N \hat{y}_i / N$ .

---

## Limited Dependent Variable Models

### **Censored and Truncated Regression Models**

When the range of dependent variables is limited, tobit models are used to analyze the data. The standard tobit model can be defined as

$$y_i^* = \mathbf{x}_i' \boldsymbol{\beta} + \epsilon_i$$

$$y_i = \begin{cases} y_i^* & \text{if } y_i^* > 0 \\ 0 & \text{if } y_i^* \leq 0 \end{cases}$$

where  $\epsilon_i \sim iidN(0, \sigma^2)$ . The dependent variable of a standard tobit or censored regression model is observed when  $y_i^* > 0$  while exogenous variables are observed for  $i = 1, \dots, N$ . The log-likelihood function of the standard censored regression model is written

$$\ell = \sum_{i \in \{y_i=0\}} \ln[1 - \Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma)] + \sum_{i \in \{y_i>0\}} \ln \left[ \frac{\phi(y_i - \mathbf{x}_i' \boldsymbol{\beta})}{\sigma} \right]$$

When neither a dependent variable nor exogenous variables are observed for  $y_i^* \leq 0$ , the truncated regression model can be specified. The log-likelihood function of the truncated regression model is written

$$\ell = \sum_{i \in \{y_i>0\}} \left\{ -\ln \Phi(\mathbf{x}_i' \boldsymbol{\beta} / \sigma) + \ln \left[ \frac{\phi(y_i - \mathbf{x}_i' \boldsymbol{\beta})}{\sigma} \right] \right\}$$

The tobit model can be generalized to handle observation-by-observation censoring and truncation. The censored model on both of the lower and upper limits can be defined as follows:

$$y_i = \begin{cases} R_i & \text{if } y_i^* \geq R_i \\ y_i^* & \text{if } L_i < y_i^* < R_i \\ L_i & \text{if } y_i^* \leq L_i \end{cases}$$

The log-likelihood function can be written

$$\begin{aligned} \ell = & \sum_{i \in \{L_i < y_i < R_i\}} \ln \phi\left(\frac{y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) / \sigma + \sum_{i \in \{y_i = R_i\}} \ln \Phi\left(-\frac{R_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) + \\ & \sum_{i \in \{y_i = L_i\}} \ln \Phi\left(\frac{L_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) \end{aligned}$$

Log-likelihood functions of the lower- or upper-limit censored model are easily derived from the two-limit censored model. The log-likelihood function of the lower-limit censored model is

$$\ell = \sum_{i \in \{y_i > L_i\}} \ln \phi\left(\frac{y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) / \sigma + \sum_{i \in \{y_i = L_i\}} \ln \Phi\left(\frac{L_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right)$$

The log-likelihood function of the upper-limit censored model is

$$\ell = \sum_{i \in \{y_i < R_i\}} \ln \phi\left(\frac{y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) / \sigma + \sum_{i \in \{y_i = R_i\}} \ln \left[ 1 - \Phi\left(\frac{R_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) \right]$$

The two-limit truncation model is defined as

$$y_i = y_i^* \quad \text{if } L_i < y_i^* < R_i$$

The log-likelihood function of the two-limit truncated regression model can be written

$$\ell = \sum_{i=1}^N \left\{ \ln \phi\left(\frac{y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) / \sigma - \ln \left[ \Phi\left(\frac{R_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) - \Phi\left(\frac{L_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) \right] \right\}$$

The log-likelihood functions of the lower- and upper-limit truncation model are

$$\begin{aligned} \ell &= \sum_{i=1}^N \left\{ \ln \left[ \phi\left(\frac{y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) / \sigma \right] - \ln \left[ 1 - \Phi\left(\frac{L_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) \right] \right\} \quad (\text{lower}) \\ \ell &= \sum_{i=1}^N \left\{ \ln \left[ \phi\left(\frac{y_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) / \sigma \right] - \ln \left[ \Phi\left(\frac{R_i - \mathbf{x}'_i \boldsymbol{\beta}}{\sigma}\right) \right] \right\} \quad (\text{upper}) \end{aligned}$$

Amemiya (1984) classified tobit models into five types based on the characteristics of the likelihood function. For notational convenience, let  $P$  denote a distribution or density function, assuming that  $y_{ji}^*$  is normally distributed with a mean of  $\mathbf{x}'_{ji}\beta_j$  and a variance of  $\sigma_j^2$ .

### Type 1 Tobit

The Type 1 tobit model, discussed in the preceding “Censored and Truncated Regression Models” section, is defined as

$$\begin{aligned} y_{1i}^* &= \mathbf{x}'_{1i}\beta_1 + u_{1i} \\ y_{1i} &= y_{1i}^* \text{ if } y_{1i}^* > 0 \\ &= 0 \text{ if } y_{1i}^* \leq 0 \end{aligned}$$

The likelihood function is characterized as  $P(y_1 < 0)P(y_1)$ .

### Type 2 Tobit

The Type 2 tobit model is defined as

$$\begin{aligned} y_{1i}^* &= \mathbf{x}'_{1i}\beta_1 + u_{1i} \\ y_{2i}^* &= \mathbf{x}'_{2i}\beta_2 + u_{2i} \\ y_{1i} &= 1 \text{ if } y_{1i}^* > 0 \\ &= 0 \text{ if } y_{1i}^* \leq 0 \\ y_{2i} &= y_{2i}^* \text{ if } y_{1i}^* > 0 \\ &= 0 \text{ if } y_{1i}^* \leq 0 \end{aligned}$$

where  $(u_{1i}, u_{2i}) \sim N(0, \Sigma)$ . The likelihood function is described as  $P(y_1 < 0)P(y_1 > 0, y_2)$ .

### Type 3 Tobit

The Type 3 tobit model is different from the Type 2 tobit in that  $y_{1i}^*$  of the Type 3 tobit is observed when  $y_{1i}^* > 0$ .

$$\begin{aligned} y_{1i}^* &= \mathbf{x}'_{1i}\beta_1 + u_{1i} \\ y_{2i}^* &= \mathbf{x}'_{2i}\beta_2 + u_{2i} \\ y_{1i} &= y_{1i}^* \text{ if } y_{1i}^* > 0 \\ &= 0 \text{ if } y_{1i}^* \leq 0 \\ y_{2i} &= y_{2i}^* \text{ if } y_{1i}^* > 0 \\ &= 0 \text{ if } y_{1i}^* \leq 0 \end{aligned}$$

where  $(u_{1i}, u_{2i})' \sim iidN(0, \Sigma)$ .

The likelihood function is characterized as  $P(y_1 < 0)P(y_1, y_2)$ .

**Type 4 Tobit**

The Type 4 tobit model consists of three equations.

$$\begin{aligned}
 y_{1i}^* &= \mathbf{x}'_{1i}\beta_1 + u_{1i} \\
 y_{2i}^* &= \mathbf{x}'_{2i}\beta_2 + u_{2i} \\
 y_{3i}^* &= \mathbf{x}'_{3i}\beta_3 + u_{3i} \\
 y_{1i} &= y_{1i}^* \text{ if } y_{1i}^* > 0 \\
 &= 0 \text{ if } y_{1i}^* \leq 0 \\
 y_{2i} &= y_{2i}^* \text{ if } y_{1i}^* > 0 \\
 &= 0 \text{ if } y_{1i}^* \leq 0 \\
 y_{3i} &= y_{3i}^* \text{ if } y_{1i}^* \leq 0 \\
 &= 0 \text{ if } y_{1i}^* > 0
 \end{aligned}$$

where  $(u_{1i}, u_{2i}, u_{3i})' \sim iidN(0, \Sigma)$ . The likelihood function of the Type 4 model is characterized as  $P(y_1 < 0, y_3)P(y_1, y_2)$ .

**Type 5 Tobit**

The Type 5 tobit model is defined as

$$\begin{aligned}
 y_{1i}^* &= \mathbf{x}'_{1i}\beta_1 + u_{1i} \\
 y_{2i}^* &= \mathbf{x}'_{2i}\beta_2 + u_{2i} \\
 y_{3i}^* &= \mathbf{x}'_{3i}\beta_3 + u_{3i} \\
 y_{1i} &= 1 \text{ if } y_{1i}^* > 0 \\
 &= 0 \text{ if } y_{1i}^* \leq 0 \\
 y_{2i} &= y_{2i}^* \text{ if } y_{1i}^* > 0 \\
 &= 0 \text{ if } y_{1i}^* \leq 0 \\
 y_{3i} &= y_{3i}^* \text{ if } y_{1i}^* \leq 0 \\
 &= 0 \text{ if } y_{1i}^* > 0
 \end{aligned}$$

where  $(u_{1i}, u_{2i}, u_{3i})'$  are from *iid* trivariate normal distribution. The likelihood function of the Type 5 model is characterized as  $P(y_1 < 0, y_3)P(y_1 > 0, y_2)$

## Examples

### Example 1. Ordered Data Modeling

Cameron and Trivedi (1986) studied an Australian Health Survey data. Variable definitions are given in Cameron and Trivedi (1998, p. 68). The dependent variable, `dvisits`, has nine ordered values. The following SAS statements estimate the ordinal probit model:

```
proc qlim data=docvisit;
  model dvisits = sex age agesq income levyplus
    freepoor freerepa illness actdays hscore
    chcond1 chcond2 / type=oprobit;
run;
```

The model fit summary is shown in Output 1.1. The Newton-Raphson technique converges in 18 iterations. The maximum log-likelihood value is  $-3138$ .

**Output 1.1.** Fit Summary Table of Ordinal Probit

The QLIM Procedure	
Ordinal Probit Estimates	
Model Fit Summary	
Dependent Variable	DVISITS
Number of Observations	5190
Log Likelihood	-3138
Maximum Absolute Gradient	9.70889E-6
Number of Iterations	18
Optimization Method	Newton-Raphson
AIC	6316
Schwarz Criterion	6447

The “Discrete Response Profile” of `dvisits` is shown in Output 1.2. The highest frequency case is no visit (79.79%), while the lowest frequency case is more than or equal to eight visits (0.12%).

**Output 1.2.** Ordinal Choice Frequency

The QLIM Procedure				
Ordinal Probit Estimates				
Discrete Response Profile				
Index	DVISITS	Frequency	Percent	
0	0	4141	79.79	
1	1	782	15.07	
2	2	174	3.35	
3	3	30	0.58	
4	4	24	0.46	
5	5	9	0.17	
6	6	12	0.23	
7	7	12	0.23	
8	8	6	0.12	

The pseudo- $R^2$  measures are shown in Output 1.3. The restricted log-likelihood function value (LogL0) is computed assuming that there are no slope coefficients. Therefore, even with a 0.5% significance level the likelihood ratio statistic fails to accept the null hypothesis that all slope parameters are zero ( $78.73 > \chi_{005,12}^2=28.3$ ).

**Output 1.3.** Pseudo- $R^2$  Measures

The QLIM Procedure		
Ordinal Probit Estimates		
Goodness-of-Fit Measures for Discrete Choice Models		
Measure	Value	Formula
Likelihood Ratio (R)	789.73	$2 * (\text{LogL} - \text{LogL0})$
Upper Bound of R (U)	7065.9	$- 2 * \text{LogL0}$
Aldrich-Nelson	0.1321	$R / (R+N)$
Cragg-Uhler 1	0.1412	$1 - \exp(-R/N)$
Cragg-Uhler 2	0.1898	$(1 - \exp(-R/N)) / (1 - \exp(-U/N))$
Estrella	0.1490	$1 - (1 - R/U)^{(U/N)}$
Adjusted Estrella	0.1442	$1 - ((\text{LogL}-K)/\text{LogL0})^{(-2/N*\text{LogL0})}$
McFadden's LRI	0.1118	$R / U$
Veall-Zimmermann	0.2291	$(R * (U+N)) / (U * (R+N))$
McKelvey-Zavoina	0.2036	

N = # of observations, K = # of regressors

Finally, the parameter estimates are shown in Output 1.4. Cameron and Trivedi (1998) also reported rescaled ordinal probit estimates (p. 92), but they do not show threshold parameter estimates.

Output 1.4. Ordinal Probit Parameter Estimates

The QLIM Procedure						
Ordinal Probit Estimates						
Parameter Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t	Gradient
Intercept	1	-1.3787	0.1474	-9.35	<.0001	-8.34E-7
SEX	1	0.1319	0.0438	3.01	0.0026	-4.08E-7
AGE	1	-0.5342	0.8158	-0.65	0.5126	-2.74E-7
AGESQ	1	0.8573	0.8983	0.95	0.3399	-2.35E-7
INCOME	1	-0.0622	0.0680	-0.91	0.3604	-5.86E-7
LEVYPLUS	1	0.1370	0.0533	2.57	0.0101	-5.43E-7
FREEPOOR	1	-0.3460	0.1296	-2.67	0.0076	-3.87E-8
FREEREPA	1	0.1784	0.0743	2.40	0.0164	-1.7E-7
ILLNESS	1	0.1505	0.0157	9.56	<.0001	-1.48E-6
ACTDAYS	1	0.1006	0.005850	17.19	<.0001	-7.62E-7
HSCORE	1	0.0319	0.009201	3.46	0.0005	-1.19E-6
CHCOND1	1	0.0616	0.0490	1.26	0.2089	-3.96E-7
CHCOND2	1	0.1353	0.0677	2.00	0.0457	-2.06E-7
LIMIT2	1	0.9389	0.0312	30.07	<.0001	-8.42E-6
LIMIT3	1	1.5143	0.0493	30.70	<.0001	9.709E-6
LIMIT4	1	1.7117	0.0581	29.44	<.0001	2.285E-7
LIMIT5	1	1.9529	0.0720	27.12	<.0001	5.956E-8
LIMIT6	1	2.0874	0.0816	25.57	<.0001	1.076E-7
LIMIT7	1	2.3338	0.1017	22.94	<.0001	5.124E-8
LIMIT8	1	2.7898	0.1562	17.86	<.0001	2.122E-8

The same data is analyzed using the ordinal logit model. Estimated parameters are shown in Output 1.5.

Output 1.5. Ordinal Logit Parameter Estimates

The QLIM Procedure						
Ordinal Logit Estimates						
Parameter Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t	Gradient
Intercept	1	-2.3837	0.2715	-8.78	<.0001	2.481E-6
SEX	1	0.2415	0.0809	2.99	0.0028	1.49E-6
AGE	1	-1.0663	1.4849	-0.72	0.4727	9.169E-7
AGESQ	1	1.6822	1.6307	1.03	0.3023	1.008E-6
INCOME	1	-0.0790	0.1255	-0.63	0.5292	1.186E-6
LEVYPLUS	1	0.2794	0.0992	2.82	0.0048	1.083E-6
FREEPOOR	1	-0.7112	0.2612	-2.72	0.0065	-1.57E-7
FREEREPA	1	0.3517	0.1359	2.59	0.0097	6.591E-7
ILLNESS	1	0.2668	0.0281	9.51	<.0001	5.119E-6
ACTDAYS	1	0.1719	0.0107	16.02	<.0001	6.609E-6
HSCORE	1	0.0620	0.0165	3.75	0.0002	4.961E-6
CHCOND1	1	0.1206	0.0903	1.34	0.1815	1.244E-6
CHCOND2	1	0.2521	0.1231	2.05	0.0406	4.621E-7
LIMIT2	1	1.7698	0.0634	27.94	<.0001	-1.37E-6
LIMIT3	1	2.9593	0.1083	27.32	<.0001	0.000014
LIMIT4	1	3.3790	0.1306	25.88	<.0001	0.000019
LIMIT5	1	3.8930	0.1646	23.65	<.0001	-8.9E-6
LIMIT6	1	4.1755	0.1871	22.32	<.0001	-8.85E-6
LIMIT7	1	4.7166	0.2399	19.66	<.0001	-6.16E-6
LIMIT8	1	5.8422	0.4111	14.21	<.0001	-5.24E-6

By default, ordinal probit/logit models are estimated assuming that the first threshold or limit parameter ( $\mu_1$ ) is 0. However, this parameter can also be estimated when the LIMIT1=VARYING option is specified. The probability that  $y_i^*$  belongs to the  $j$ th category is defined as

$$P[\mu_{j-1} < y_i^* < \mu_j] = F(\mu_j - \mathbf{x}_i' \boldsymbol{\beta}) - F(\mu_{j-1} - \mathbf{x}_i' \boldsymbol{\beta})$$

where  $F(\cdot)$  is the logistic or standard normal CDF,  $\mu_0 = -\infty$  and  $\mu_9 = \infty$ . Output 1.6 lists ordinal or cumulative logit estimates. Note that the intercept term is suppressed for model identification when  $\mu_1$  is estimated.

Output 1.6. Ordinal Logit Parameter Estimates with LIMIT1=VARYING

The QLIM Procedure						
Ordinal Logit Estimates						
Parameter Estimates						
Parameter	DF	Estimate	Standard Error	t Value	Approx Pr >  t	Gradient
SEX	1	0.2415	0.0809	2.99	0.0028	3.005E-8
AGE	1	-1.0663	1.4851	-0.72	0.4728	2.172E-8
AGESQ	1	1.6822	1.6309	1.03	0.3023	1.23E-8
INCOME	1	-0.0790	0.1255	-0.63	0.5292	2.385E-8
LEVYPLUS	1	0.2794	0.0992	2.82	0.0048	2.232E-8
FREERPOOR	1	-0.7112	0.2612	-2.72	0.0065	6.77E-10
FREEREPA	1	0.3517	0.1359	2.59	0.0097	1.466E-8
ILLNESS	1	0.2668	0.0281	9.51	<.0001	1.086E-7
ACTDAYS	1	0.1719	0.0107	16.02	<.0001	8.942E-8
HSCORE	1	0.0620	0.0165	3.75	0.0002	9.305E-8
CHCOND1	1	0.1206	0.0903	1.34	0.1815	2.498E-8
CHCOND2	1	0.2521	0.1231	2.05	0.0406	8.958E-9
LIMIT1	1	2.3837	0.2715	8.78	<.0001	2.739E-7
LIMIT2	1	4.1535	0.2792	14.88	<.0001	-1.11E-7
LIMIT3	1	5.3431	0.2935	18.21	<.0001	-1.99E-7
LIMIT4	1	5.7628	0.3025	19.05	<.0001	-4.36E-9
LIMIT5	1	6.2767	0.3188	19.69	<.0001	-802E-12
LIMIT6	1	6.5593	0.3310	19.81	<.0001	-2.51E-9
LIMIT7	1	7.1003	0.3636	19.53	<.0001	1.179E-8
LIMIT8	1	8.2259	0.4937	16.66	<.0001	-1.38E-8

---

## References

- Abramowitz, M. and Stegun, A. (1970), *Handbook of Mathematical Functions*, New York: Dover Press.
- Aitchison, J. and Silvey, S. (1957), "The Generalization of Probit Analysis to the Case of Multiple Responses," *Biometrika*, 44, 131–140.
- Amemiya, T. (1978), "The Estimation of a Simultaneous Equation Generalized Probit Model," *Econometrica*, 46, 1193–1205.
- Amemiya, T. (1978), "On a Two-Step Estimate of a Multivariate Logit Model," *Journal of Econometrics*, 8, 13–21.
- Amemiya, T. (1981), "Qualitative Response Models: A Survey," *Journal of Economic Literature*, 19, 483–536.
- Amemiya, T. (1984), "Tobit Models: A Survey," *Journal of Econometrics*, 24, 3–61.
- Amemiya, T. (1985), *Advanced Econometrics*, Cambridge: Harvard University Press.
- Ben-Akiva, M. and Lerman, S.R. (1987), *Discrete Choice Analysis*, Cambridge: MIT Press.
- Bera, A.K., Jarque, C.M., and Lee, L.-F. (1984), "Testing the Normality Assumption in Limited Dependent Variable Models," *International Economic Review*, 25, 563–578.
- Bloom, D.E. and Killingsworth, M.R. (1985), "Correcting for Truncation Bias Caused by a Latent Truncation Variable," *Journal of Econometrics*, 27, 131–135.
- Box, G.E.P. and Cox, D.R. (1964), "An Analysis of Transformations," *Journal of the Royal Statistical Society, Series B.*, 26, 211–252.
- Cameron, A.C. and Trivedi, P.K. (1986), "Econometric Models Based on Count Data: Comparisons and Applications of Some Estimators," *Journal of Applied Econometrics*, 1, 29–53.
- Cameron, A.C. and Trivedi, P.K. (1998), *Regression Analysis of Count Data*, Cambridge: Cambridge University Press.
- Copley, P.A., Doucet, M.S., and Gaver, K.M. (1994), "A Simultaneous Equations Analysis of Quality Control Review Outcomes and Engagement Fees for Audits of Recipients of Federal Financial Assistance," *The Accounting Review*, 69, 244–256.
- Cox, D.R. (1970), *Analysis of Binary Data*, London: Methuen.
- Cox, D.R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society, Series B*, 20, 187–220.
- Cox, D.R. (1975), "Partial Likelihood," *Biometrika*, 62, 269–276.
- Deis, D.R. and Hill, R.C. (1998), "An Application of the Bootstrap Method to the Simultaneous Equations Model of the Demand and Supply of Audit Services," *Contemporary Accounting Research*, 15, 83–99.

- Estrella, A. (1998), "A New Measure of Fit for Equations with Dichotomous Dependent Variables," *Journal of Business and Economic Statistics*, 16, 198–205.
- Godfrey, L.G. (1988), *Misspecification Tests in Econometrics*, Cambridge: Cambridge University Press.
- Gourieroux, C., Monfort, A., Renault, E., and Trognon, A. (1987), "Generalized Residuals," *Journal of Econometrics*, 34, 5–32.
- Green, W.H. (1997), *Econometric Analysis*, Upper Saddle River, N.J.: Prentice Hall.
- Hajivassiliou, V.A. (1993), "Simulation Estimation Methods for Limited Dependent Variable Models," in *Handbook of Statistics*, vol. 11, ed. G.S. Maddala, C.R. Rao, and H.D. Vinod, New York: Elsevier Science Publishing.
- Heckman, J.J. (1978), "Dummy Endogenous Variables in a Simultaneous Equation System," *Econometrica*, 46, 931–959.
- Hinkley, D.V. (1975), "On Power Transformations to Symmetry," *Biometrika*, 62, 101–111.
- Kim, M. and Hill, R.C. (1993), "The Box-Cox Transformation-of-Variables in Regression," *Empirical Economics*, 18, 307–319.
- King, G. (1989b), *Unifying Political Methodology: The Likelihood Theory and Statistical Inference*, Cambridge: Cambridge University Press.
- Lee, L.-F. (1981), "Simultaneous Equations Models with Discrete and Censored Dependent Variables," in *Structural Analysis of Discrete Data with Econometric Applications*, eds. C.F. Manski and D. McFadden, Cambridge: MIT Press
- Long, J.S. (1997), *Regression Models for Categorical and Limited Dependent Variables*, Thousand Oaks, CA : SAGE Publications, Inc.
- McFadden, D. (1974), "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. P. Zarembka, New York: Academic Press.
- McFadden, D. (1981), "Econometric Models of Probabilistic Choice," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. C.F. Manski and D. McFadden, Cambridge: MIT Press.
- McKelvey, R.D. and Zavoina, W. (1975), "A Statistical Model for the Analysis of Ordinal Level Dependent Variables," *Journal of Mathematical Sociology*, 4, 103–120.
- Mroz, T.A. (1987), "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765–799.
- Mroz, T.A. (1999), "Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome," *Journal of Econometrics*, 92, 233–274.
- Nawata, K. (1994), "Estimation of Sample Selection Bias Models by the Maximum Likelihood Estimator and Heckman's Two-Step Estimator," *Economics Letters*, 45, 33–40.

Parks, R.W. (1967), "Efficient Estimation of a System of Regression Equations When Disturbances Are Both Serially and Contemporaneously Correlated," *Journal of the American Statistical Association*, 62, 500–509.

Powers, D.A. and Xie, Y. (2000), *Statistical Methods for Categorical Data Analysis*, San Diego: Academic Press.