

**Modelling Qualitative Variables**

Most of models presented in this course require that dependent variable(s) be normally distributed. One obvious question arises, what analysis should one pursue, if variables are not normally distributed. To provide some structure to this topic, following summary might be helpful starting point. Note also that in all instances in this table, only one dependent variable is considered.

**Alternative Qualitative Variables**

Com- pared on	Measurement Properties				
	Binary	Ordinal	Nominal	Continu- ous and not normal	Mixed
Model	Logit or logistic analysis	Ordered logit	multinomial logit, conditional logit, Nested logit	Survival or Event history analysis	Tobit, Switching regression, Binary and normal
Examples	Ownership of a durable good, e.g. Cellphone	Preference ranks for brands	Type of residence, Brand choice, Choice of profession	Failure or Purchase incidence	Spending on automobile or eating out expenses

Word logit may be replaced with probit. With exception of multinomial situation or model for nominal data, estimated model parameters are similar for logit and probit models. Although probit is preferred model when dependent variable is multinomial, computational problems associated it, have prevented wide spread usage of probit model.

**Logistic Analysis or Binary Logit Model**

• **Objectives**

1. Identify a set of variables that “best” predict group membership into two or more groups.
2. Classify existing or future observations.
3. Use predictive logistic function to understand the impact of independent variable(s) on group membership.

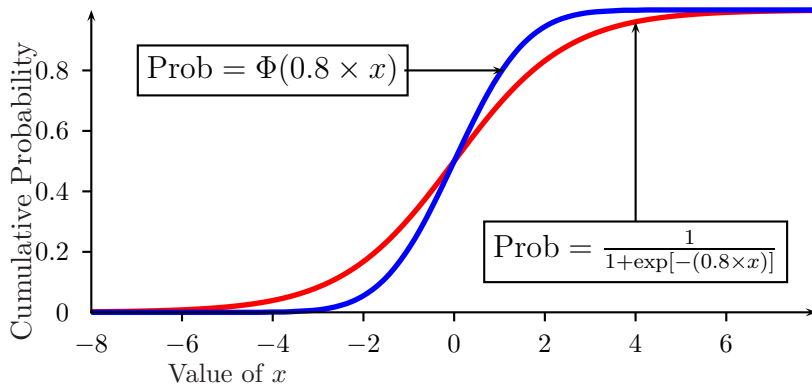
• **Applications**

1. Credit risk assessment, firms as well as individuals.
2. Multiple brand purchasers or single brand purchasers.
3. Heavy, moderate or light users of product category.
4. Predicting choice among competing brands.
5. Cluster analysis followed by logistic analysis.

• **Assumptions**

The logistic or logit analysis does not require many of assumptions used in discriminant analysis, especially distribution of multivariate normality of independent variables.

• **Illustrative Example**



Consider the example about “Factors influencing Bankruptcy”. In logistic analysis, our interest lies with predicting the likelihood that a firm may declare bankruptcy (BANKRUP = 0) and we may use various financial ratios to predict that likelihood. Suppose we are looking at ratio of retained earnings to total assets (RE). Then we may write

$$\text{Prob}(\text{BANKRUP}_i = 0) = \frac{1}{1 + \exp[-(b_0 + b_1 \times \text{RE}_i)]}$$

where  $i$  subscript denotes  $i$ th sample firm. This curve is called logistic and it is ‘S’-shaped. In comparison to normal distribution, logistic curve has “thicker” tails. Note also that when inside parenthesis expression  $\exp[ ]$  takes a large positive value, then the likelihood approaches to zero. This will occur when  $(b_0 + b_1 \times \text{RE}_i)$  is a large negative number. On the other hand, when  $\exp[ ]$  takes a large negative value, then the likelihood approaches one. This is illustrated graphically above.

We also know that the likelihood of bankruptcy can also be used to predict the likelihood that firm has not declared bankruptcy. That is,

$$\text{Prob}(\text{BANKRUP}_i = 1) = 1 - \text{Prob}(\text{BANKRUP}_i = 0)$$

$$\begin{aligned}
&= 1 - \frac{1}{1 + \exp(-(b_0 + b_1 \times RE_i))} \\
&= \frac{1 + \exp(-(b_0 + b_1 \times RE_i)) - 1}{1 + \exp(-(b_0 + b_1 \times RE_i))} \\
&= \frac{\exp(-(b_0 + b_1 \times RE_i))}{1 + \exp(-(b_0 + b_1 \times RE_i))}
\end{aligned}$$

Suppose that we computed

$$\frac{\text{Prob}(\text{BANKRUP}_i = 0)}{1 - \text{Prob}(\text{BANKRUP}_i = 0)} = \exp(b_0 + b_1 \times RE_i),$$

and this is called odds ratio. Suppose we took logarithms of odds ratio. Then we may write

$$\log \left( \frac{\text{Prob}(\text{BANKRUP}_i = 0)}{1 - \text{Prob}(\text{BANKRUP}_i = 0)} \right) = b_0 + b_1 \times RE_i.$$

Note that the relationship between the likelihood that a firm may go bankrupt and the ratio of retained earning to total assets is *non-linear*, whereas the relationship between the log of the odds and the ratio retained earning to total assets is *linear*. As a result of this observation, interpretation of linear parameters ( $b_0$  and  $b_1$ ) should be with respect to the log odds and not on the likelihood.

Let us look at procedure for obtaining parameters. One such procedure is based on the concept of maximum likelihood estimation. We will illustrate this method using a simple coin tossing example. I will provide numerical example as well as derivation based on calculus and then we will conclude with an example using SAS.

#### • Coin Toss

Consider the case where a coin is tossed and probability of obtaining a head,  $H$ , is  $p$  and the probability of obtaining a tail,  $T$ , is  $(1 - p)$ . Consider a situation in which the coin is tossed five times with following outcomes:  $H, T, H, H$  and  $T$ . If the outcome at each toss is independent of the previous outcomes and probability  $p$  does not change after each outcome, then the joint probability of obtaining three heads and two tails ( $\mathcal{L}$ ) is given by

$$\begin{aligned}
\mathcal{L} = \text{Prob}(H, T, H, H, T) &= p \times (1 - p) \times p \times p \times (1 - p) \\
&= p^3(1 - p)^2
\end{aligned}$$

In this equation,  $p$  is the parameter that we would like to estimate. The mathematical narrow question is: What is the value of the parameter  $p$  that maximizes the joint probability given in equation for  $\mathcal{L}$ ?

The maximum likelihood estimate of parameter  $p$  is defined as that estimate of the parameter that results in the maximum likelihood or probability of observing the given the sample data; that is it is the value of  $p$  for which the sample data will occur the most often.

Value of the Likelihood Function for various value of $p$		
$p$	$\mathcal{L}$	$\mathcal{L}\mathcal{L}$
0.001	0.0000	-20.7253
0.1	0.0008	-7.1185
0.2	0.0051	-5.2746
0.3	0.0132	-4.3253
0.4	0.0230	-3.7705
0.5	0.0313	-3.4657
0.6	0.0346	-3.3651
0.7	0.0309	-3.4780
0.8	0.0205	-3.8883
0.9	0.0073	-4.9213
0.999	0.0000	-13.8185

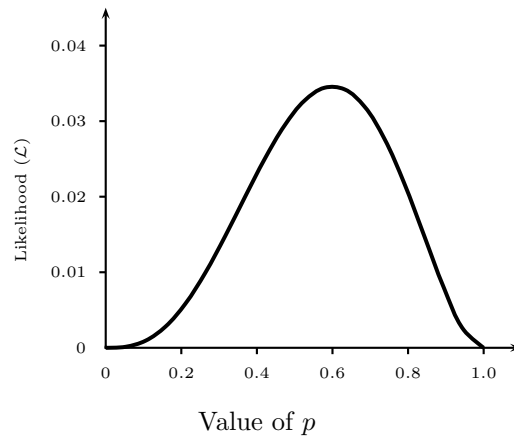


Figure 1: Illustration of Likelihood Function

The above equation of  $\mathcal{L}$  is known as the likelihood function. The value of  $p$  can be obtained by trial-and-error, that is we could choose different values  $p$  and check to see whether that is maximum and repeat our search until we find  $\mathcal{L}$  to be maximum.

As you might expect that the maximum does seem to occur when  $p$  is 0.6 and the likelihood function value at this point is 0.0346. Since the likelihood function is reasonable (continuous and differentiable), the estimate of  $p$  can also be obtained by differentiating the likelihood function with respect to the parameter  $p$  and equating it to zero. That is,

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial p} &= 3p^2(1-p)^2 + p^3[2(1-p)(-1)] = 0 \\ &= (1-p)p^2[3(1-p) - 2p] = 0\end{aligned}$$

Note that above expression is satisfied at  $\hat{p} = 0, 0.6$  and 1. Without obtaining second derivative, we intuitively should realize that  $p = 0$  and  $p = 1$  refer to minimum and only remaining value is the maximum.

Instead of maximizing  $\mathcal{L}$ , the natural logarithm ( $\log$ ) of the likelihood function is maximized, often denoted by  $\mathcal{L}\mathcal{L}$ . Maximizing  $\mathcal{L}\mathcal{L}$  instead of  $\mathcal{L}$  does not affect the estimates and may in fact improve computational precision.

Let us make our problem of finding parameters slightly more interesting. Suppose we had two individuals tossing coins and we observed five events for each individual. Consider that the first individual obtained  $H, H, T, H$  and  $H$  while the second individual realized  $T, T, H, T$  and  $T$ . In this situation, we are interested in two parameters, the probability that head comes up for the first individual ( $p_1$ ) and the same event for the second individual ( $p_2$ ). Intuitively, we would expect that  $\hat{p}_1$  to be equal to 0.8 and  $\hat{p}_2$  to be 0.2. Note that if we ignored the individual differences then we would get overall likelihood that head comes which in this case is likely to

be  $\hat{p} = .5$  Let us see whether we get these results using idea of maximum likelihood estimation. Let us denote  $\mathcal{L}_0$  when we estimate overall likelihood that heads will come up and  $\mathcal{L}_1$  when we estimate individual specific parameters. Note that

$$\begin{aligned}\mathcal{L}_0 &= p \times p \times (1-p) \times p \times p(1-p) \times (1-p) \times p \times (1-p) \times (1-p) \\ &= p^5 \times (1-p)^5 \text{ or in logarithms} \\ \mathcal{L}\mathcal{L}_0 &= 5 \times \log(p) + 5 \times \log(1-p)\end{aligned}$$

The maximum of such function does exist at point  $\hat{p} = 0.5$  and the maximum of  $\mathcal{L}\mathcal{L}_0$  is equal to  $-6.93$ . Let us now look at what would happen when we estimate separate parameters for each individual.

$$\begin{aligned}\mathcal{L}_1 &= p_1 \times p_1 \times (1-p_1) \times p_1 \times p_1(1-p_2) \times (1-p_2) \times p_2 \times (1-p_2) \times (1-p_2) \\ &= p_1^4 \times (1-p_1) \times p_2 \times (1-p_2)^4 \text{ or in logarithms} \\ \mathcal{L}\mathcal{L}_1 &= 4 \times \log(p_1) + \log(1-p_1) + \log(p_2) + 4 \times \log(1-p_2)\end{aligned}$$

The maximum of such function does exist at point  $\hat{p}_1 = 0.8$  and  $\hat{p}_2 = 0.2$ . Moreover, the maximum for the logarithm of likelihood function at such point is  $-5.004$ . There is statistical test to determine whether adding separate parameter for individual improves fit of the model. This statistical test is given by

$$\chi^2 = -2 \times (\mathcal{L}\mathcal{L}_0 - \mathcal{L}\mathcal{L}_k)$$

where  $k$  different parameters estimates. This statistic is distributed with  $\chi^2$  with  $k$  degrees of freedom. For our example,  $\chi^2$  is 3.855 and we would reject the null hypothesis at probability level 0.05 that parameters are equal for two individuals (Critical value of  $\chi^2$  is 3.84 at prob. 0.05 and 1 degree of freedom).

Let us see how we can do this analysis using SAS.

### SAS Input

```
options ls=75 ps = 65 nocenter nodate;
data logist;
input headtail person;
datalines;
1 0
1 0
1 0
0 0
1 0
0 1
0 1
1 1
0 1
0 1
;;;
proc logistic descending;
model headtail = person ;
run;
```

Note that keyword `descending` is used estimate the likelihood that head (denoted by 1) shows. If you do not use this keyword, SAS will estimate the likelihood that tail shows.

### SAS Output

The LOGISTIC Procedure

Data Set: WORK.LOGIST  
 Response Variable: HEADTAIL  
 Response Levels: 2  
 Number of Observations: 10  
 Link Function: Logit

#### Response Profile

Ordered Value	HEADTAIL	Count
1	1	5
2	0	5

#### Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	15.863	14.008	.
SC	16.166	14.613	.
-2 LOG L	13.863	10.008	3.855 with 1 DF (p=0.0496)
Score	.	.	3.600 with 1 DF (p=0.0578)

#### Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	1.3863	1.1180	1.5375	0.2150	.	.
PERSON	1	-2.7726	1.5811	3.0749	0.0795	-0.805647	0.063

#### Association of Predicted Probabilities and Observed Responses

Concordant = 64.0%	Somers' D = 0.600
Discordant = 4.0%	Gamma = 0.882
Tied = 32.0%	Tau-a = 0.333
(25 pairs)	c = 0.800

Note that

- Instead of printing  $\mathcal{LL}$ , SAS prints  $-2\mathcal{LL}$ .
- Note also that `Intercept only` refers to condition when there are no individual differences.
- Further, `Intercept and Covariate` refers to condition when independent variables are included.

- Note that the log of likelihood function increases with inclusion of more variables in the model. There are number of alternatives that penalize fit function for estimating more parameters. For example, if  $k$  is number of categories in the dependent variable and  $s$  is number of independent variables, then Akaike information criteria (AIC), is  $-2\mathcal{LL} + 2(k + s - 1)$ . Similarly, Schwartz criteria (SC) is  $-2\mathcal{LL} + (k + s - 1) \log(N)$  where  $N$  is sample size. Note that both AIC and SC may not increase with more parameters.
- SAS does not directly provide us the estimate of  $\hat{p}_1$  or  $\hat{p}_2$ . Instead SAS provides estimates of log-odds ratio or  $\log\left(\frac{\hat{p}_1}{1 - \hat{p}_1}\right)$ . That is, for person 1,

$$\log\left(\frac{\hat{p}_1}{1 - \hat{p}_1}\right) = 1.386$$

$$\hat{p}_1 = \frac{1}{1 + \exp(-1.386)} = 0.8$$

Similarly for person 2,

$$\log\left(\frac{\hat{p}_2}{1 - \hat{p}_2}\right) = 1.386 - 2.773$$

$$\hat{p}_2 = \frac{1}{1 + \exp(1.387)} = 0.2$$

- SAS also prints various measures of association which may provide descriptive indication of association between independent and dependent variable. If  $N$  is sample size,  $t$  is pairs of different responses for the dependent variable, NC is number of concordant observations and ND is number of discordant<sup>1</sup> observations, then tied pair of observations is  $t - NC - ND$ . Moreover,

$$\text{Somers's D} = \frac{NC - ND}{t},$$

$$\text{Gamma} = \frac{NC - ND}{NC + ND},$$

$$\text{Tau - a} = \frac{NC - ND}{0.5 \times N(N - 1)},$$

$$c = \frac{NC + 0.5 \times (t - NC - ND)}{t}.$$

---

<sup>1</sup>A pair of observations with different responses is said to be concordant if the larger response has a lower predicted probability that the smaller response.

**Latent Segmentation: An Example**

Idea of latent segments or group of consumers with similar responsiveness to marketing decisions has received attention in the marketing and statistics literature. Moreover, segmentation idea provides a method of grouping individuals into several homogenous groups by using each individual’s responses. Suppose there are eight individuals in a sample and each individual made 20 choice among two brands of detergent, Tide (T) and Sunlight (S). The resulting outcomes and associated log-likelihood statistics are summarized in Table below.

**Effect of One and Two Segment Solution on Log-Likelihood Statistics**

Person	T S		One Segment	Two Segment		
			$\mathcal{LL}$	$\mathcal{LL}$	$\mathcal{LL}$	$\mathcal{LL}$
			with $p = 0.5$	with $p = 0.25$	with $p = 0.75$	Average
1	2	18	-13.8629	-7.9509	-25.5287	-8.6440
2	4	16	-13.8629	-10.1481	-23.3314	-10.8412
3	6	14	-13.8629	-12.3453	-21.1342	-13.0383
4	8	12	-13.8629	-14.5425	-18.9370	-15.2234
5	12	8	-13.8629	-18.9370	-14.5425	-15.2234
6	14	6	-13.8629	-21.1342	-12.3453	-13.0383
7	16	4	-13.8629	-23.3314	-10.1481	-10.8412
8	18	2	-13.8629	-25.5287	-7.9509	-8.6440

In this example, there are two sets of parameters that we need to estimate. First, the probability that Tide will be chosen ( $p$ ) and this probability will vary by group or individual. Second, for each individual, determine group membership and in the process determine number of groups. To begin this process, one could argue that

there is one segment. Since there are total 160 choices and 80 times Tide is chosen, the probability estimate is 0.5. Consequently, log of likelihood function for individual  $i$  is  $\mathcal{LL}_i = n_i \times \log(p) + (20 - n_i) \times \log(1 - p)$  where  $n_i$  is number times Tide is chosen. Because number of choices and the probability of Tide is chosen ( $p$ ) is 0.5, log of likelihood function is same for all individuals. This results in the sample log-likelihood which is sum of individual log-likelihoods. Thus, the sample log-likelihood and Bayesian Information Criteria<sup>2</sup> (BIC) equal to -110.90 and 113.44 respectively. As is indicated below, these indicators may be used to evaluate number of segments.

To consider a possibility of two segments or groups, we need to divide individuals in two groups. One possibility is to group first four observations in one segment. For the first group, the probability that Tide is chosen is 0.25. The last four observations have probability that Tide is chosen is 0.75. It is possible to determine the optimal groups (one that maximizes log of likelihood function) with any combination of group membership. Because individual might belong to one of two groups, we need to compute two values of log of likelihood function, one with probability of 0.25 and another with probability of 0.75. These estimates as well as ‘average’ of likelihoods is summarized below. Note that maximized value of log-likelihood is equal to -95.49 and BIC is equal to 98.03. Note that in creating these two groups, we have estimated three parameters, probability that Tide will be chosen by the first group ( $\hat{p}_1 = 0.25$ ) and by the second group ( $\hat{p}_2 = 0.75$ ) and proportion of observations in the first group.

---

<sup>2</sup>This is also called Schwartz criteria and BIC is equal to  $-\mathcal{LL} + \frac{1}{2}k \log(N)$  where  $k$  is number of parameters estimated and  $N$  is total number of observations.

To form three segments, there are number of possibilities. We could divide person 1 and 2 in group one, 3, 4, 5 and 6 in group two and observations 7 and 8 in group three. The optimal groupings (one that maximized log of likelihood function) involves observation 1, 2, 3 and 4 in group one with  $p_1 = 0.25$ , 5 and 6 in group two with  $p_2 = 0.65$  and 7 and 8 in group three with  $p_3 = 0.85$ . At such point, log-likelihood value is equal to  $-95.3462$  and BIC is 108.0341. Although three groups in this example provide interesting insight about segmentation, statistically three groups with six parameters is not simple description of observed data. This is because first BIC is increasing and second the log-likelihood has not changed much with addition of three more parameters in comparison to two groups solution. Consequently, we conclude that two groups is the optimal sample description of observed data.

### Back to Bankruptcy Example.

We could use PROC LOGISTIC to determine the effect of various factors (working capital, retained earning etc.) on the likelihood of bankruptcy.

### SAS Input

```
options nocenter nodate ps = 65 ls=75;
data fin;
infile "c:\26-606\bankrup.dat";
input bankrupt id wc re ebit mktval sales ;
proc logistic simple descending;
  model bankrupt = wc re ebit mktval sales /details selection=f ;
run;
```

Note that option on model statement `selection=f` requests stepwise inclusion of independent variables. Also note that following message is printed SAS's log window.

NOTE: PROC LOGISTIC is modeling the probability that BANKRUP=1.

### SAS Output

The LOGISTIC Procedure

Data Set: WORK.FIN  
 Response Variable: BANKRUP  
 Response Levels: 2  
 Number of Observations: 66  
 Link Function: Logit

#### Response Profile

Ordered Value	BANKRUP	Count
1	1	33
2	0	33

Forward Selection Procedure

Simple Statistics for Explanatory Variables

Variable	BANKRUP	Mean	Standard Deviation	Minimum	Maximum
WC	1	41.384848	14.218586	14.000000	69.000000
	0	-6.047273	45.553467	-185.100000	72.400000
	Total	17.668788	41.136727	-185.100000	72.400000
RE	1	35.251515	16.507747	-3.300000	68.600000
	0	-62.512121	71.312529	-308.900000	20.800000
	Total	-13.630303	71.161567	-308.900000	68.600000
EBIT	1	15.318182	10.867769	-14.400000	34.100000
	0	-31.781818	51.353502	-280.000000	6.800000
	Total	-8.231818	43.813076	-280.000000	34.100000
MKTVAL	1	254.369697	206.077846	53.400000	771.700000
	0	40.045455	54.938318	0.700000	267.900000
	Total	147.207576	184.536302	0.700000	771.700000
SALES	1	1.939394	0.930033	0.900000	5.500000
	0	1.503030	1.162294	0.100000	6.700000
	Total	1.721212	1.067350	0.100000	6.700000

Step 0. Intercept entered:

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	0	0.2462	0.0000	1.0000	.	.

Residual Chi-Square = 41.7699 with 5 DF (p=0.0001)

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > Chi-Square
WC	22.2741	0.0001
RE	31.6212	0.0001
EBIT	19.3619	0.0001
MKTVAL	22.5992	0.0001
SALES	2.8003	0.0942

Step 1. Variable RE entered:

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Intercept

Criterion	Intercept Only	and Covariates	Chi-Square for Covariates
AIC	93.495	19.803	.
SC	95.685	24.182	.
-2 LOG L	91.495	15.803	75.692 with 1 DF (p=0.0001)
Score	.	.	31.621 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	-1.1666	0.8164	2.0419	0.1530	.	.
RE	1	0.1767	0.0571	9.5776	0.0020	6.933297	1.193

Association of Predicted Probabilities and Observed Responses

Concordant = 99.1%      Somers' D = 0.983  
 Discordant = 0.8%      Gamma = 0.983  
 Tied = 0.1%      Tau-a = 0.499  
 (1089 pairs)      c = 0.991

Residual Chi-Square = 7.1822 with 4 DF (p=0.1266)

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > Chi-Square
WC	1.1158	0.2908
EBIT	4.3956	0.0360
MKTVAL	1.9331	0.1644
SALES	2.5613	0.1095

Step 2. Variable EBIT entered:

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	93.495	15.472	.
SC	95.685	22.041	.
-2 LOG L	91.495	9.472	82.024 with 2 DF (p=0.0001)
Score	.	.	32.698 with 2 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	-0.5503	0.9510	0.3349	0.5628	.	.
RE	1	0.1574	0.0749	4.4120	0.0357	6.173924	1.170
EBIT	1	0.1947	0.1224	2.5302	0.1117	4.704092	1.215

Association of Predicted Probabilities and Observed Responses

Concordant = 99.7%                      Somers' D = 0.994  
 Discordant = 0.3%                      Gamma = 0.994  
 Tied = 0.0%                              Tau-a = 0.505  
 (1089 pairs)                              c = 0.997

Residual Chi-Square = 6.2949 with 3 DF (p=0.0981)

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > Chi-Square
WC	0.6648	0.4149
MKTVAL	4.4847	0.0342
SALES	2.7277	0.0986

Step 3. Variable MKTVAL entered:

WARNING: There is a complete separation in the sample points. The maximum likelihood estimate does not exist.  
 WARNING: The LOGISTIC procedure continues in spite of the above warning. Results shown are based on the last maximum likelihood iteration.

The LOGISTIC Procedure

Validity of the model fit is questionable.

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	93.495	11.370	.
SC	95.685	20.128	.
-2 LOG L	91.495	3.370	88.126 with 3 DF (p=0.0001)
Score	.	.	37.895 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
INTERCPT	1	-5.4710	5.6467	0.9387	0.3326	.	.
RE	1	0.2549	0.2296	1.2324	0.2669	9.999710	1.290
EBIT	1	0.3985	0.3141	1.6095	0.2046	9.626001	1.490
MKTVAL	1	0.0664	0.0602	1.2139	0.2706	6.753153	1.069

Association of Predicted Probabilities and Observed Responses

Concordant =100.0%                      Somers' D = 1.000  
 Discordant = 0.0%                      Gamma = 1.000  
 Tied = 0.0%                              Tau-a = 0.508  
 (1089 pairs)                              c = 1.000

Residual Chi-Square = 0.8527 with 2 DF (p=0.6529)

Analysis of Variables Not in the Model

Variable	Score Chi-Square	Pr > Chi-Square
WC	0.8370	0.3602
SALES	0.7968	0.3721

NOTE: No (additional) variables met the 0.05 significance level for entry into the model.

WARNING: The validity of the model fit is questionable.

Summary of Forward Selection Procedure

Step	Variable Entered	Number In	Score Chi-Square	Pr > Chi-Square
1	RE	1	31.6212	0.0001
2	EBIT	2	4.3956	0.0360
3	MKTVAL	3	4.4847	0.0342

Our analysis indicate following observations.

- Retained earnings to total assets (RE), Earning before interest and taxes to total assets (EBIT) and market value of equity to book value of liabilities (MKTVAL) are important predictors of whether firm declare bankruptcy. RE is the most important in contributing to explaining in differences between bankrupt and non-bankrupt firms.
- Parameter for RE indicate that increase in RE results in decrease in log odd ratio of that firm will become bankrupt (see parameters on page 9, bottom section). More specifically, we find that (Note that changes in signs)

$$\text{Prob}(\text{BANKRUP}_i = 1) = \frac{1}{1 + \exp(1.167 - 0.177 \times \text{RE}_i)}$$

Consider the average firm that went bankrupt. That firm's ratio of retained earnings to total assets was  $-62.51$  while firms that did not go bankrupt had the average ratio of  $35.25$ . Based on our equation, we can write

$$\begin{aligned} \text{Prob}(\text{BANKRUP}_i = 1) &= \frac{1}{1 + \exp(1.167 - 0.177 \times (-62.51))}, \\ &= \frac{1}{1 + \exp(1.167 + 11.0643)} \approx 0 \end{aligned}$$

We could also perform similar calculations to determine the likelihood that firm may not go bankrupt based on the group's average. Suppose we performed similar calculations for all 66 observations. Then we can derive likelihood that firm will go bankrupt for each

firm and then make comparison to actual outcome. Following table provides details of these calculations.

Comparison of Prediction and Actual

Obs.	Bank. Status	RE Ratio	Pred. Prob.	Obs.	Bank. Status	RE Ratio	Pred. Prob.	Obs.	Bank. Status	RE Ratio	Pred. Prob.
1	0	-62.80	0.000	2	0	3.30	0.358	3	0	-120.80	0.000
4	0	-18.10	0.013	5	0	-3.80	0.137	6	0	-61.20	0.000
7	0	-20.30	0.009	8	0	-194.50	0.000	9	0	20.80	0.925
10	0	-106.10	0.000	11	0	-39.40	0.000	12	0	-164.10	0.000
13	0	-308.90	0.000	14	0	7.20	0.526	15	0	-118.30	0.000
16	0	-185.90	0.000	17	0	-34.60	0.001	18	0	-27.90	0.002
19	0	-48.20	0.000	20	0	-49.20	0.000	21	0	-19.20	0.010
22	0	-18.10	0.013	23	0	-98.00	0.000	24	0	-129.00	0.000
25	0	-4.00	0.133	26	0	-8.70	0.063	27	0	-59.20	0.000
28	0	-13.10	0.030	29	0	-38.00	0.000	30	0	-57.90	0.000
31	0	-8.80	0.062	32	0	-64.70	0.000	33	0	-11.40	0.040
34	1	43.00	0.998	35	1	47.00	0.999	36	1	-3.30	0.148
37	1	35.00	0.993	38	1	46.70	0.999	39	1	20.80	0.925
40	1	33.00	0.991	41	1	26.10	0.969	42	1	68.60	1.000
43	1	37.30	0.996	44	1	59.00	1.000	45	1	49.60	0.999
46	1	12.50	0.739	47	1	37.30	0.996	48	1	35.30	0.994
49	1	49.50	0.999	50	1	18.10	0.884	51	1	31.40	0.988
52	1	21.50	0.933	53	1	8.50	0.583	54	1	40.60	0.998
55	1	34.60	0.993	56	1	19.90	0.913	57	1	17.40	0.871
58	1	54.70	1.000	59	1	53.50	1.000	60	1	35.90	0.994
61	1	39.40	0.997	62	1	53.10	1.000	63	1	39.80	0.997
64	1	59.50	1.000	65	1	16.30	0.847	66	1	21.70	0.935

Bank. Status 0 indicate that firm declared bankruptcy while 1 indicate otherwise.

- We conclude that bankruptcy prediction based on RE is very good and we predict 63 out of 66 cases correctly. For observation numbers 9, 14 and 36, our prediction is incorrect.
- Ratio of earning before interest and taxes to total assets (EBIT) is second important variable in predicting the likelihood of bankruptcy (see page 10). Including this variable appears to improve prediction for observation number 14.
- Final ratio included is market value of equity to book value of liabilities. With inclusion of this ratio, we have predictive model that perfectly predicts the likelihood of bankruptcy. The perfect prediction creates difficult estimation problem, since likelihood at such point is not very well defined. SAS indicates that in the output and gives warning that finding may be misleading.

• **Who owns Personal Computer at Home?**

There is an indication that personal computers are for those with higher income and better education. Some have argued that personal computers and internet<sup>3</sup> is likely to create “digital divide” or those with “haves” and “have-nots.” To better understand notion of digital divide following analysis examines influence of **demographic factors** on the likelihood of owning personal computer. To estimate parameters, I will use the survey data collected by the Energy Information Administration (EIA) of the U.S. Department of Energy for 1993. The Residential Energy Consumption Survey (RECS) contains basic data on housing unit characteristics,

<sup>3</sup>Work by Hoffman, Donna and Tom Novak (1998) “Bridging the Racial Divide on the Internet, *Science*, volume 280, April 17, pp. 390-391, provide further details about such comparisons.

annualized 1993 fuel consumption and expenditures and estimates for these energy end uses: space heating, air conditioning, water heating, appliances, refrigerators, freezers, lighting, electric clothes dryers, dishwashers, and electric ranges/ovens. The data file used below contained 7,041 records representing households in the 48 States and the District of Columbia. The households are weighted to represent 96.1 million households, as of July 1993, the midpoint of data collection activity. Households in Alaska and Hawaii have been removed by EIA from public use files for confidentiality reasons.

Following table provides summary of households that own personal computers as well as those that do not own them. Note that looking at one variable at a time, we note that on every socio-demographic variables, the owners are different from non-owners. Such analysis, however, ignores correlations among various independent variables. Subsequent logistic analysis indicates that only seven variables are statistically significant (ignoring couple of squared variables) at probability of less than or equal to 0.05. One final note on selection of variables. One would expect that age of residence, respondent's age and household's income to have non-linear influence on the likelihood of computer ownership. To test nonlinear effect of these factors, I have included squared terms of these variables.

**The Sample Description**

Descriptor	Owners	Non-owners	Sample Average
Urban	39.57%	46.04%	44.51%
Town	13.47%	17.52%	16.56%
Suburban	28.86%	17.76%	20.38%
Home Ownership	76.07%	61.08%	64.62%
High School Education	24.17%	62.48%	53.43%
College Education	69.21%	30.10%	39.34%
Married	76.37%	56.58%	61.26%
Full Time Employed	69.99%	49.31%	54.20%
Part-time Employed	9.80%	8.72%	8.98%
Region: North east	21.95%	21.68%	21.74%
Region: Mid West	22.67%	21.64%	21.89%
Region: South	30.55%	36.93%	35.42%
Spanish speaking	5.65%	9.41%	8.52%
Race: White	91.10%	80.10%	82.70%
Race: Black	4.39%	14.58%	12.17%
Age of Residence	22.65	28.76	27.32
Respondent's Age	42.55	48.90	47.40
Household size	3.05	2.59	2.70
Drivers in household	2.04	1.53	1.65
Household Income(\$'000)	52.93	30.44	35.75

## • SAS Input

```

options ls=80 ps=70 nocenter nodate;
data demo;
infile "d:\ener\1993\demograp.asc" lrecl = 800 missover;
input HHID REGIONC DIVISION CNTLCELL URBRUR SHARELQT LIVEATSP
      KOWNRENT KOWNCOND HUPROJ RENTHELP YEARMAD E OCCUPY OCCUPYY
      OCCUPYM ANSWERHH HOUSEHLD HHSEX HHAGE EMPLOYHH ANSWER02
      KIN02 SEX02 YEARS02 EMPLOY02 ANSWER03 KIN03 SEX03 YEARS03
      EMPLOY03 ANSWER04 KIN04 SEX04 YEARS04 EMPLOY04 ANSWER05 KIN05
      SEX05 YEARS05 EMPLOY05 ANSWER06 KIN06 SEX06 YEARS06 EMPLOY06
      ANSWER07 KIN07 SEX07 YEARS07 EMPLOY07 ANSWER08 KIN08 SEX08
      YEARS08 EMPLOY08 ANSWER09 KIN09 SEX09 YEARS09 EMPLOY09 ANSWER10
      KIN10 SEX10 YEARS10 EMPLOY10 ANSWER11 KIN11 SEX11 YEARS11
      EMPLOY11 ANSWER12 KIN12 SEX12 YEARS12 EMPLOY12 NHSLDMEM
      DRIVEMON GRADHH MARRIED SDESCENT ORIGIN WAGES SELFHIRE
      SSECURTY PENSIONS FSTAMPS AFDC UNEMPLOY SSI OTHERAID
      MONEYPY INC45PLU DRIVECAR VEHICLES FAMSIZE POOR100
      POOR125 POOR150 HOWPAID NWEIGHT;
proc sort data=demo; by hhid;
data applia;
infile "d:\ener\1993\applianc.asc" lrecl=256 missover;
input HHID REGIONC DIVISION CNTLCELL LGT12 FLRLGT12 LGT4 FLRLGT4
      LGT1 FLRLGT1 OUTLGTNO OUTLGTVEV OUTLGTNT OUTLGTTI GASLIGHT
      OUTLGTTHI OUTLGTLW OUTLGTDK ELSTOVE NGSTOVE LPSTOVE OTHSTOVE
      OVEN ELOVEN NGOVEN LPOVEN OTHOVEN TOASTER GRILL BARBECUE
      LPGBROIL MICRO AMTMICRO NUMFRIG SEPFREEZ NUMFREEZ AGERFRI1
      SIZRFRI1 TYPERFRI1 REFRIGT1 AGERFRI2 SIZRFRI2 TYPERFR2
      REFRIGT2 AGEFRZR SIZFREEZ FREEZER ICE MONRFRI2 UPRTFRZR
      CWASHER ELDRYER NGDRYER LPDRYER DRYER DISHWASH TVCOLOR
      TVBLACK WATERBED NOWTBDHT AQUARIUM NOTMOIST MOISTURE WELLPUMP
      EXHFAN CLEANER SWAMPOL COMPUTER PRINTER FAX COPIER SWIMPOOL
      POOL RECBATH WINDFAN PORTFAN ATTICEXH ATTICFAN CFAN NUMCFAN
      NWEIGHT ;
proc sort; by hhid; run;
data combined;
  merge applia demo;
by hhid;

if computer = 9 then computer = .;
urban = 0;
rural = 0;
suburb = 0;
if urbrur = 1 then urban = 1;
if urbrur = 2 then rural = 1;
if urbrur = 3 then suburb = 1;
gradhh1 = 0;
gradhh2 = 0;
if gradhh ge 4 and gradhh le 12 then gradhh1 = 1;
if gradhh ge 14 and gradhh le 18 then gradhh2 = 1;
if married = 1 or married = 5 then married = 1;
if married = 2 or married = 3 or married = 4 then married = 0;
hhagesq = hhage*hhage;
/* Whites, Blacks and Others */
white = 0;
black = 0;
if origin = 1 then white = 1;
if origin = 2 then black = 1;
/* Employment status */
EmpFul = 0;
EmpPart = 0;

```

```

if EMPLOYHH = 1 then Empful = 1;
if employhh = 2 then EmpPart = 1;

/*      Income from categories to mid-points      */

if moneypy = 1 then income = 1.5;
if moneypy = 2 then income = 4;
if moneypy = 3 then income = 4.5;
if moneypy = 4 then income = 5.5;
if moneypy = 5 then income = 6.75;
if moneypy = 6 then income = 8.25;
if moneypy = 7 then income = 9.5;
if moneypy = 8 then income = 10.5;
if moneypy = 9 then income = 11.75;
if moneypy = 10 then income = 13.25;
if moneypy = 11 then income = 14.5;
if moneypy = 12 then income = 16.25;
if moneypy = 13 then income = 18.75;
if moneypy = 14 then income = 21.25;
if moneypy = 15 then income = 23.75;
if moneypy = 16 then income = 26.25;
if moneypy = 17 then income = 28.75;
if moneypy = 18 then income = 31.25;
if moneypy = 19 then income = 33.75;
if moneypy = 20 then income = 37.5;
if moneypy = 21 then income = 42.5;
if moneypy = 22 then income = 47.5;
if moneypy = 23 then income = 62.5;
if moneypy = 24 then income = 87.5;
if moneypy = 25 then income = 120;
income2 = income*income;
/*      Year house built from category to mid-points      */
if yearmade = 1 then yearhous = 1940;
if yearmade = 2 then yearhous = 1945;
if yearmade = 3 then yearhous = 1955;
if yearmade = 4 then yearhous = 1965;
if yearmade = 5 then yearhous = 1975;
if yearmade = 6 then yearhous = 1982;
if yearmade = 7 then yearhous = 1985.5;
if yearmade = 8 then yearhous = 1987;
if yearmade = 9 then yearhous = 1988;
if yearmade = 10 then yearhous = 1989;
if yearmade = 11 then yearhous = 1990;
if yearmade = 12 then yearhous = 1991;
if yearmade = 13 then yearhous = 1992;
if yearmade = 14 then yearhous = 1993;
if yearmade = 15 then yearhous = 1994;
yearhous = 1994 - yearhous;
yearh2 = yearhous*yearhous;
/*      Regional differences      */
norest = 0;
mdwest = 0;
south = 0;
if regionc = 1 then norest = 1;
if regionc = 2 then mdwest = 1;
if regionc = 3 then south = 1;
run;

proc logistic descending simple covout outest=estim data=combined;
model computer = urban rural suburb kownrent yearhous yearh2 hhage hhagesq empful empPart
              nhsldmem drivemon gradhh1
              gradhh2 married income income2  norest mdwest south sdescent white black;

```

run;

• SAS Output

The LOGISTIC Procedure

Data Set: WORK.COMBINED  
 Response Variable: COMPUTER  
 Response Levels: 2  
 Number of Observations: 7041  
 Link Function: Logit

Response Profile

Ordered Value	COMPUTER	Count
1	1	1663
2	0	5378

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept Only	Intercept and Covariates	Chi-Square for Covariates
AIC	7699.869	6137.180	.
SC	7706.728	6301.808	.
-2 LOG L Score	7697.869	6089.180	1608.689 with 23 DF (p=0.0001) 1436.959 with 23 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

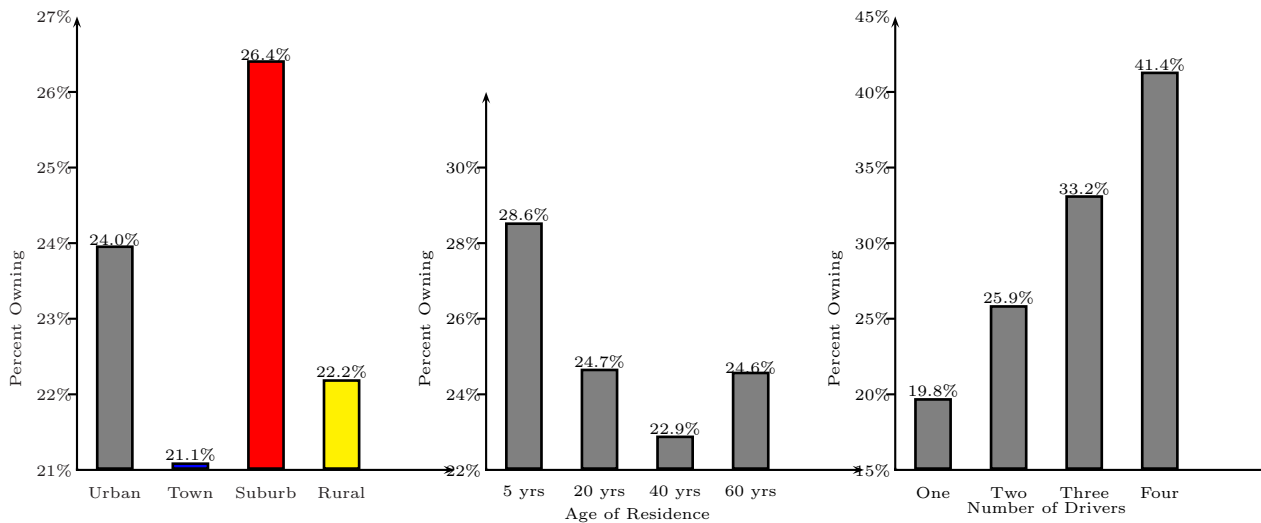
Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	Standardized Estimate	Odds Ratio
INTERCPT	1	-3.2642	0.4383	55.4572	0.0001	.	.
URBAN	1	0.0991	0.0949	1.0917	0.2961	0.027157	1.104
RURAL	1	-0.0659	0.1126	0.3428	0.5582	-0.013509	0.936
SUBURB	1	0.2291	0.0990	5.3518	0.0207	0.050885	1.257
KOWNRENT	1	-0.1033	0.0805	1.6481	0.1992	-0.028759	0.902
YEARHOUS	1	-0.0192	0.00748	6.5837	0.0103	-0.194714	0.981
YEARH2	1	0.000239	0.000127	3.5048	0.0612	0.145695	1.000
HHAGE	1	0.0617	0.0147	17.6130	0.0001	0.589385	1.064
HHAGESQ	1	-0.00083	0.000154	29.1046	0.0001	-0.833146	0.999
EMPFUL	1	0.0471	0.0902	0.2730	0.6013	0.012947	1.048
EMPPART	1	0.2362	0.1247	3.5864	0.0583	0.037219	1.266
NHSLDMEM	1	0.0128	0.0290	0.1930	0.6604	0.010339	1.013
DRIVEMON	1	0.3507	0.0509	47.4652	0.0001	0.174691	1.420
GRADHH1	1	-0.5681	0.1279	19.7373	0.0001	-0.156246	0.567
GRADHH2	1	0.7514	0.1227	37.4834	0.0001	0.202381	2.120
MARRIED	1	0.0559	0.0869	0.4128	0.5205	0.015004	1.057
INCOME	1	0.0201	0.00427	22.0846	0.0001	0.307552	1.020
INCOME2	1	-0.00006	0.000033	2.8637	0.0906	-0.094753	1.000
NOREST	1	-0.1915	0.0998	3.6849	0.0549	-0.043564	0.826
MDWEST	1	-0.1777	0.0971	3.3474	0.0673	-0.040501	0.837
SOUTH	1	-0.4118	0.0906	20.6523	0.0001	-0.108589	0.662
SDESCENT	1	-0.2635	0.1326	3.9450	0.0470	-0.040559	0.768
WHITE	1	0.2177	0.1555	1.9601	0.1615	0.045401	1.243
BLACK	1	-0.6160	0.2020	9.3045	0.0023	-0.111054	0.540

Association of Predicted Probabilities and Observed Responses

Concordant = 80.5%	Somers' D = 0.612
Discordant = 19.3%	Gamma = 0.614
Tied = 0.2%	Tau-a = 0.221
(8943614 pairs)	c = 0.806

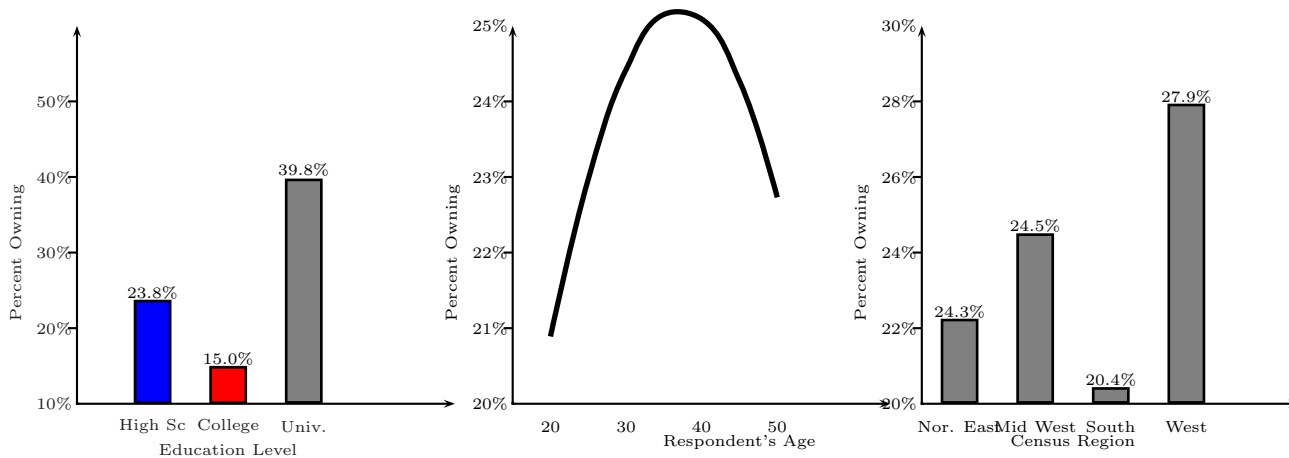
• Interpreting Parameters

One challenge with logistic regression is that estimated parameters do not have immediate interpretation. One approach that is useful to interpret is to consider an average observation. In my case, household that lives in urban environment, own's his or her residence and lives in residence that is built about five years prior to the survey. Moreover, such a respondent is about 48 years old, has full time employment with high school level education, and lives with two other members, one of whom is also driver with annual household income of \$36K. Based on our estimated logistic model, we would conclude that such a household would have 23.62% of chance of owning computer in 1993.



Suppose such household lived in smaller sized city (Town), then the likelihood of owning computer for such household will go down by 1.8%. On the other hand, if such household lived in suburban area, that household's likelihood of owning would go up by 2.5%. All these conditions, along with fourth condition (Rural) is given in above graph.

We could apply such marginal analysis to all the variables that are statistically significant at probability of 0.05. First consider age of residence. Our marginal analysis indicates that as age of residence increases, upto 40 years, the likelihood of owning computer decreases and then the likelihood increases. On the other hand, the effect associated with respondent's age is concave such that about age of 35, the likelihood of computer ownership is at the maximum level.



The effect associated with number of drivers in the household is dramatic. One would expect that household with more high school and college going children, result in greater likelihood of computer ownership. Our model supports this notion very well.

The effect of education is somewhat surprising. One would expect that the higher the education, need for computer at home increases. Our model parameters indicate that those respondent's who did not complete high school have higher likelihood of owning computer at home than those who have not completed college education. Those respondents with university education appear to have the highest likelihood of owning computer.

As one might expect, if household is located in the western region, household would have about 28% chance of owning home computer. On the other hand, household located in Southern states, their likelihood of owning computer is about 20%. The rust belt states (North east) and middle American (Mid-west) states have ownership rates close to the average.

The effect of household income is in the expected direction. For every \$10 thousand increase in the household income, the likelihood of owning computer increases by about three percent around the sample average. Those households with family income of \$180 thousand have the maximum likelihood of owning computer (51.3%). Note also that various race based groups differ significantly in terms of likelihood of owning computer. Those with Spanish descended had likelihood of owning home computer of about 20% whereas Black household had chance of 15.3%, a strong support for "Digital Divide".

In conclusion, from above example we find that respondent's age, age of residence and income to be non-linearly related to the likelihood of computer ownership. Respondent's education level as well as number of individuals with driving age have stronger influence on computer ownership. There are also regional as well as various race-based group differences in the likelihood of owning computer. Whether respondent lives in own home or rented home, employment status as well as marital status had no significant impact on computer ownership.

- **Example from Major League Baseball**

The World Series of baseball games in the North America is a best-of-seven series. Therefore, the first team to win four games is the victor. To try to make the series fair, the first two games are played at one team's home park, the next three games are played at the other team's park (just two may be played if one team wins four in a row), and the final two games (if needed) are scheduled at the first team's park. It is therefore possible that one team will play four games in its home park while the other team may only play three games at home.

It is generally assumed that there is some advantage to playing at home. Each baseball park is different, so there may be some advantage to playing on a familiar field. The hometown fans provide emotional support for the team. Is there an advantage to playing World Series games on your home field? In addition, there is also debate about game rules. When game is played in American League (AL) park, rules are based on AL (designated hitter and pitcher may not bat) and vice versa for National League (NL). Consequently, there may be differential advantage of playing at home field for AL or NL team. Using outcomes from the past World Series games, it is possible to test (1) whether there is a home field advantage and (2) whether there is a differential home field advantage for AL or NL team.

Using variety of sources, the dataset containing information on the year, the two teams competing in the world series, whether the AL or NL team won the game, location of game played, and the AL and NL teams' winning percentages at home or away was put together. Note that there were 461 games played between 1922 to 2001. Furthermore, the home team won 55.97% of games there by indicating a small home field advantage. To understand formally above questions the logistic regression may be used. That is, the probability that the AL team would win a game is a function that game is played in the AL park.

$$\text{Prob(AL = W)} = \frac{1}{1 + \exp(-(a + b \times \text{AL Park}))}$$

The logistic regression provided following estimates,  $\hat{a} = 0.4643$  and  $\hat{b} = 0.4501$ . This equation suggest that AL team playing in on its own field is likely to win 61.4% of games and AL team playing on NL field is likely to win 49.36% of games, about 12% drop in game outcome. One concludes from above example that home field advantage is slightly favoured to AL teams than NL teams.

- **Demographic Reasons to Own Multiple TV Sets.**

The basic logistic model can be easily modified to accomodate ordered categories of dependent variable. We will use previous dataset to illustrate the effect of demographic factors on household's decision to own multiple TV sets. Suppose there are  $k$  ordered categories (number of

sets owned). Then, the probability that the household  $j$  owns  $i$  TV sets is given by

$$\text{Prob}(Y_j = i) = \begin{cases} F(\alpha_1 + \sum_{m=1}^M \beta_m X_{jm}) & \text{for } i = 0 \\ F(\alpha_i + \sum_{m=1}^M \beta_m X_{jm}) - F(\alpha_{i-1} + \sum_{m=1}^M \beta_m X_{jm}) & \text{for } 1 < i \leq k - 1 \\ 1 - F(\alpha_{k-1} + \sum_{m=1}^M \beta_m X_{jm}) & \text{for } i = k \end{cases}$$

Note that  $Y_j$  is the categorical response variable and corresponding demographic variables,  $M$  of them are in matrix  $X_{jm}$ . Note that parameter vector  $\beta$  does not vary by alternative. Moreover, we assume that parameter vector is fixed across different categories. Let us look at descriptive information on some of demographic variables. Finally  $F(\cdot)$  can be logistic or cumulative normal distribution function.

Descriptive Summary of Demographic Variables

	Number of Colour TV Sets Owned						Average
	0	1	2	3	4	5 or more	
Urban	54.94%	47.37%	42.45%	42.13%	40.57%	38.60%	44.51%
Town	14.81%	18.39%	16.74%	13.80%	11.32%	10.53%	16.56%
Suburb	8.64%	15.23%	21.52%	27.69%	31.45%	38.60%	20.38%
NorthEast	17.90%	21.20%	23.13%	20.83%	19.81%	23.68%	21.74%
Mid-west	21.60%	21.34%	21.21%	22.96%	27.99%	23.68%	21.89%
South	40.12%	34.84%	35.87%	35.56%	33.33%	37.72%	35.42%
Renters	66.67%	50.28%	28.62%	18.43%	10.38%	4.39%	35.38%
Full Time	38.89%	46.95%	56.96%	62.31%	70.75%	70.18%	54.20%
Part Time	10.49%	9.38%	8.90%	8.33%	7.55%	8.77%	8.98%
School Grad	66.05%	58.88%	51.59%	46.67%	43.40%	34.21%	53.43%
College	29.63%	33.98%	41.12%	45.74%	48.74%	58.77%	39.34%
Married	31.48%	46.59%	68.33%	78.24%	83.33%	85.09%	61.26%
Year Home Built	1960	1965	1968	1969	1972	1972	1967
Age	46.9	48.0	47.0	47.0	47.2	46.2	47.4
Household Size	2.09	2.28	2.81	3.25	3.52	3.96	2.70
N of Drivers	0.93	1.32	1.76	2.08	2.30	2.46	1.65
Family Income	19.45	26.07	38.40	47.99	57.68	61.50	35.75

• SAS Output

The LOGISTIC Procedure

Data Set: WORK.COMBINED  
 Response Variable: TVCOLOR  
 Response Levels: 6  
 Number of Observations: 7041  
 Link Function: Logit

Response Profile

Ordered  
 Value TVCOLOR Count

1	5	114
2	4	318
3	3	1080
4	2	2551
5	1	2816
6	0	162

Score Test for the Proportional Odds Assumption

Chi-Square = 95.5476 with 76 DF (p=0.0643)

Model Fitting Information and Testing Global Null Hypothesis BETA=0

Criterion	Intercept and Covariates		Chi-Square for Covariates
	Intercept Only	Intercept and Covariates	
AIC	18532.968	16480.622	.
SC	18567.266	16645.250	.
-2 LOG L	18522.968	16432.622	2090.346 with 19 DF (p=0.0001)
Score	.	.	1808.791 with 19 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	-8.5478	0.2949	840.0091	0.0001	.	.
INTERCP2	1	-7.0881	0.2824	630.0472	0.0001	.	.
INTERCP3	1	-5.4571	0.2767	389.0440	0.0001	.	.
INTERCP4	1	-3.4552	0.2719	161.5199	0.0001	.	.
INTERCP5	1	0.4495	0.2766	2.6418	0.1041	.	.
URBAN	1	0.6700	0.0677	97.8173	0.0001	0.183602	1.954
RURAL	1	0.3405	0.0791	18.5432	0.0001	0.069794	1.406
SUBURB	1	0.7194	0.0744	93.4684	0.0001	0.159788	2.053
KOWNRENT	1	-0.5737	0.0554	107.3481	0.0001	-0.159749	0.563
YEARHOUS	1	-0.00596	0.00137	18.8244	0.0001	-0.060436	0.994
HHAGE	1	0.0889	0.00873	103.5990	0.0001	0.848647	1.093
HHAGESQ	1	-0.00077	0.000085	81.6435	0.0001	-0.771596	0.999
EMPFUL	1	-0.00700	0.0631	0.0123	0.9117	-0.001922	0.993
EMPPART	1	0.0136	0.0898	0.0230	0.8794	0.002147	1.014
NHSLDMEM	1	0.2576	0.0198	168.8674	0.0001	0.208741	1.294
DRIVEMON	1	0.3896	0.0351	123.5110	0.0001	0.194033	1.476
GRADHH1	1	-0.1173	0.0910	1.6610	0.1975	-0.032258	0.889
GRADHH2	1	-0.1402	0.0928	2.2822	0.1309	-0.037768	0.869
MARRIED	1	0.0315	0.0602	0.2748	0.6001	0.008473	1.032
INCOME	1	0.0249	0.00308	65.2430	0.0001	0.381098	1.025
INCOME2	1	-0.0001	0.000024	16.6524	0.0001	-0.171300	1.000
NOREST	1	0.2009	0.0724	7.7006	0.0055	0.045699	1.223
MDWEST	1	0.1937	0.0709	7.4569	0.0063	0.044162	1.214
SOUTH	1	0.1080	0.0646	2.7948	0.0946	0.028472	1.114

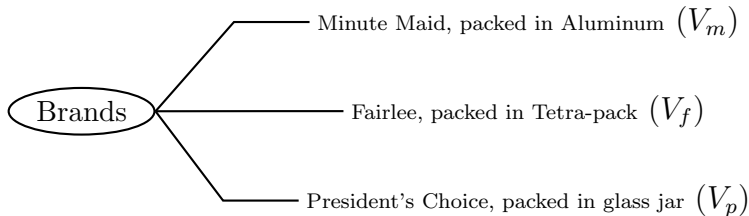
The LOGISTIC Procedure

Association of Predicted Probabilities and Observed Responses

Concordant = 70.8%	Somers' D = 0.477
Discordant = 23.1%	Gamma = 0.508
Tied = 6.1%	Tau-a = 0.326
(16915730 pairs)	c = 0.739

### Multinomial Logit

The basic logit model can be extended to situation involving multiple categorical variables, and a dependent variable does not have ordinal structure. Consider following situation one may face when purchasing single serving Orange juice.



Suppose prices of all three brands are identical in this store but juice is prepared with three different sources; freshly prepared juice, juice prepared from concentrate and juice prepared after pasterization. In such situation, the likelihood that a consumer might choose Minute Maid is given by

$$\text{Prob}(\text{Minute Maid}) = \frac{\exp(V_m)}{\exp(V_m) + \exp(V_f) + \exp(V_p)}.$$

Suppose we are able to create an experimental design such that all attributes, price (\$0.60, \$0.80 or \$1.00), packaging material (Glass, Aluminium can, tetra-pak) and preparation (fresh frozen, from frozen concentrate, pasteurized frozen concentrate) from are varied and we ask respondent to choose from such alternatives. Then, we may be able to parameterize utility function. In following example, a group of students were provided with 10 situations and were asked to choose one alternative out of three. Resulting observations were analyzed using SAS software. Collected data is shown below. Note that the first column is used to identify respondent. Next 10 columns indicate particular brand chosen when faced with a particular choice occasion. The last column identify each respondent who participated in the study.

```

1 1 2 2 3 1 3 1 1 3 2 Melanie
2 1 1 2 1 1 3 1 1 3 2 Wenli
3 1 2 2 2 1 3 3 1 1 3 Brian
4 1 1 1 3 1 1 2 3 3 2 Francis
5 1 1 2 1 1 2 3 1 1 3 Diane
6 1 2 2 2 1 3 1 1 3 2 John
7 1 1 1 1 2 1 2 3 1 2 Vicki
8 1 2 1 2 1 3 2 3 3 3 Alina
9 2 2 2 1 1 3 3 1 1 3 Quang
  
```

### SAS Input for Analyzing Choices

```

options nocenter nodate ps=65 ls=75;
data ojprof;                                /* Read Choice Profiles */
infile "c:\pricing\ojch1.prn";
input situat brand prep pacmat price order;
datalines;
  1 1 1 1 0.60 1
  1 2 1 2 0.60 2
  1 3 1 1 0.60 3

  2 1 2 2 1.00 3
  2 2 3 3 0.60 2
  2 3 2 2 1.00 1

  3 1 3 3 0.80 3
  3 2 2 1 0.60 1
  3 3 3 3 0.80 2

  4 1 1 2 0.80 1
  4 2 3 1 0.80 3
  4 3 3 1 0.80 2

  5 1 2 3 0.60 2
  5 2 2 3 0.80 1
  5 3 2 2 1.00 3

  6 1 1 3 1.00 3
  6 2 1 2 0.80 2
  6 3 1 3 0.60 1

  7 1 1 3 1.00 2
  7 2 3 1 1.00 1
  7 3 2 1 0.60 3

  8 1 2 1 0.80 1
  8 2 1 3 1.00 3
  8 3 3 3 1.00 2

  9 1 3 2 0.80 3
  9 2 1 2 1.00 2
  9 3 1 1 0.60 1

 10 1 2 2 1.00 2
 10 2 1 1 0.80 1
 10 3 3 2 0.60 3
run;
data ojch ;                                /* Read Choices */
infile "c:\pricing\oj_clas.dat";
input id ch1-ch10 Name $;
run;
/* Transpose Choices, */
/* There will 10 Records per person */
proc transpose data = ojch
  out = res1(rename=(col1=ch) drop =_name_) ;
  by id;
  var ch1-ch10;
run;
/* Convert Choice observations 1 or 2 */
data res2(drop=ch situat);
i = mod(_n_ - 1,10) + 1;
situa = i;

```

```

set res1;
  do j = 1 to 3;
    choice = 2 - (j eq ch);
    ij = (i - 1)*3 + j;
    set ojprof point=ij;      /* Read in choice profile information */
    output ;
  end ;
run;

/* Compute Various dummy variables */

data res3;
  set res2;
  minute = 0;
  fairlee = 0;
  if brand = 1 then minute = 1;
  if brand = 2 then fairlee = 1;
  prep1 = 0;
  prep2 = 0;
  if prep = 1 then prep1 = 1; /* prep1 dummy variable for Fresh Frozen Preparation */
  if prep = 2 then prep2 = 1; /* prep2 dummy variable for from frozen concentrate */
  pacmat1 = 0;
  pacmat2 = 0;
  if pacmat = 1 then pacmat1 = 1; /* pacmat1 dummy variable for Glass packaging */
  if pacmat = 2 then pacmat2 = 1; /* pacmat2 dummy variable for aluminum can */
proc sort ; by id;
proc phreg data=res3 outest = betas nosummary;
model
  choice*choice(2) = minute fairlee
                    prep1 prep2 pacmat1 pacmat2 price / ties = breslow;
strata id situa;
run;

```

## SAS Output

The PHREG Procedure

Data Set: WORK.RES3  
 Dependent Variable: CHOICE  
 Censoring Variable: CHOICE  
 Censoring Value(s): 2  
 Ties Handling: BRESLOW

Testing Global Null Hypothesis: BETA=0

Criterion	Without Covariates	With Covariates	Model Chi-Square
-2 LOG L	197.750	150.512	47.238 with 7 DF (p=0.0001)
Score	.	.	43.386 with 7 DF (p=0.0001)
Wald	.	.	31.158 with 7 DF (p=0.0001)

Analysis of Maximum Likelihood Estimates

Variable	DF	Parameter Estimate	Standard Error	Wald Chi-Square	Pr > Chi-Square	
MINUTE	1	1.471557	0.36040	16.67207	0.0001	Minute Maid
FAIRLEE	1	0.194885	0.37066	0.27645	0.5990	Fairlee
PREP1	1	-0.011471	0.35775	0.00103	0.9744	Fresh Frozen
PREP2	1	-0.955293	0.60693	2.47738	0.1155	From Concentrate
PACMAT1	1	1.175472	0.63943	3.37939	0.0660	Glass jar

PACMAT2	1	0.213025	0.57521	0.13715	0.7111	Aluminum can
PRICE	1	-4.482758	1.05526	18.04548	0.0001	Price of serving

### • Interpretation of Results

Because multinomial logit model is non-linear, interpretation of estimated parameters requires more efforts than a linear models. We will use estimated model to interpret brand constants, own and cross-price elasticities, brand and attribute premiums as well as attribute importance. Suppose that utility or attraction for Minute Maid can be written as  $V_m = \exp(a_m + b_1 \text{Price}_m)$  where  $a_m$  is brand constant for Minute Maid and  $b_1$  is price coefficient. Similarly attraction of Fairlee and President's Choice are  $V_f = \exp(a_f + b_1 \text{Price}_f)$  and  $V_p = \exp(b_1 \text{Price}_p)$  respectively. Note that in this particular situation President's Choice is considered to be the base brand and its brand constant is assumed to zero. Then according to multinomial logit model, we may write

$$\text{Prob}_m = \frac{V_m}{V_m + V_f + V_p}.$$

You may write this as

$$\text{Prob}_m = \frac{\exp(a_m + b_1 \text{Price}_m)}{\exp(a_m + b_1 \text{Price}_m) + \exp(a_f + b_1 \text{Price}_f) + \exp(b_1 \text{Price}_p)}$$

If prices of three brands are identical, then above expression can be simplified to

$$\text{Prob}_m = \frac{\exp(a_m)}{\exp(a_m) + \exp(a_f) + \exp(0)}$$

or

$$\text{Prob}_m = \frac{\exp(a_m)}{1 + \exp(a_m) + \exp(a_m)}$$

In our example, Minute Maid brand has constant parameter of  $a_m = 1.472$  and Fairlee brand has constant parameter of  $a_f = 0.195$ . This would imply that Minute Maid would have choice proportion of

$$\begin{aligned} \text{Prob}_m &= \frac{\exp(1.472)}{1 + \exp(1.472) + \exp(0.195)} \\ &= \frac{4.358}{1 + 4.358 + 1.215} \\ &= 0.6629 \end{aligned}$$

Similar calculation for Fairlee would result in choice proportion of 0.1849 and for President's

Choice to be 0.1522.

- **Deriving Own and Cross Elasticities**

To obtain own price elasticity<sup>4</sup> for the first brand, we need to obtain the partial derivative of this expression with respect to the first price. That is,

$$\frac{\partial \text{Prob}_m}{\partial \text{Price}_1} = \frac{b_1 V_m}{V_m + V_f + V_p} - \frac{(V_m)^2 b_1}{(V_m + V_f + V_p)^2}$$

Note that above expression can be reduced to

$$\frac{\partial \text{Prob}_m}{\partial \text{Price}_m} = b_1 \text{Prob}_m (1 - \text{Prob}_m).$$

To get own price elasticity ( $\eta_{mm}$ ), then we multiply both sides of above equation by  $\text{Price}_m / \text{Prob}_m$ , or

$$\eta_{mm} = (1 - \text{Prob}_m) b_1 \text{Price}_m.$$

A similar approach also can be used to derive cross price elasticity or the percent change in probability of choice for Minute Maid, as result of 1% change in price of Fairlee. If you do all above steps, we would get

$$\eta_{mf} = -b_1 \text{Prob}_f \text{Price}_f.$$

Consider a situation where all three brands are priced at \$0.85, then own and cross price elasticities would be

$$\boldsymbol{\eta} = \begin{pmatrix} -1.284 & 0.705 & 0.580 \\ 2.526 & -3.106 & 0.580 \\ 2.526 & 0.705 & -3.230 \end{pmatrix}$$

From this example, we would conclude that Minute maid has low own price elasticity and other two brands have relatively high own price elasticity. When price of Minute Maid is lowered, note that there is big shift in changes in probability that an individual would choose Fairlee and / or President's Choice. This is because Fairlee and President's Choice have high cross elasticities with respect to changes in Minute Maid prices. Intuitively, if Minute Maid lowers price, more of Fairlee and President's Choice buyers will switch away from them. On the other hand, if Fairlee lowers price, impact on Minute Maid is relatively small. This is because cross price elasticity is small.

### Deriving Brand Price Premiums

One managerial question that is often asked with choice model is price premium associated with branded products. That is, holding everything else (product formulation, packaging and

---

<sup>4</sup>Own price elasticity is percent change in choice share as a result of one percent change in own price. On the other hand, cross price elasticity is percent change in choice share of brand 1 as result of one percent change in price of brand 2 or brand 3. We would expect that own price elasticity to be negative and the cross price elasticities to be positive for substitute products.

distribution) how much discount should Fairlee brand offer in order to get same share as Minute Maid brand? To answer this question, we would need to equate brand utility of two brands and solve resulting equation for prices. That is,

$$\begin{aligned} V_m &= V_f \quad \text{which implies that} \\ \exp(a_m + b_1 \text{Price}_m) &= \exp(a_f + b_1 \text{Price}_f) \quad \text{or} \\ \text{Price}_m - \text{Price}_f &= \frac{a_f - a_m}{b_1} \end{aligned}$$

Note that we would generally expect that  $b_1$  to be less than zero. This would imply that if  $a_f$  is less than  $a_m$  (these are parameters associated with brands), the difference between two prices will be positive. In this case we would conclude that the second brand, Fairlee should lower its price to get same choice shares as the first brand. Note that the right hand side in the above expression is equal to \$0.285 which means that Fairlee should reduce its price to \$0.56 to get about 66% choice share. On the other hand, if  $a_f$  is greater than  $a_m$ , we would expect that the second brand could increase its price to keep same choice shares.

If two brands have differing prices, then one may want to compute the relative price premium by dividing both sides of above expression by  $\text{Price}_m$ . This would result in expression,

$$\frac{\text{Price}_m - \text{Price}_f}{\text{Price}_m} = \frac{a_f - a_m}{b_1 \text{Price}_m}$$

and can be interpreted as a percent premium (discount) over another brand. At price of \$0.85 Minute Maid could get 33.5% premium over Fairlee brand. Note that if Minute Maid price would have been lower than \$0.85 (say \$0.80), then price premium would be higher (at price of \$0.80 premium would be 35.61%).

### Deriving Price Premiums for Product Attributes

Some product features are preferred by consumers while other features may be less valued. Moreover, to quantify and assign price premium associated with product features, expression for relative price premium may be used. This price premium amount often is termed respondents' are willingness to pay for brand or product feature. For example, in Orange juice example, coefficient for *packaging material* glass has estimate of 1.175 and that for aluminum can is 0.213. In this situation, third packaging material, tetra-pak is assumed to be zero. Suppose President's Choice is considering glass as packaging material instead of tetra-pak, and tetra-pak is priced at the base price of \$0.80. To determine, discount amount tetra-pak should offer in order to get same share as glass packaging, we need to equate utilities and solve for prices as it is done above. This would result in relative price premium for glass to be

$$\begin{aligned} \text{Glass Premium} &= - \frac{\text{Estimate for Glass}}{\text{Price} \times \text{Price Coefficient}} \\ &= - \frac{1.175}{0.8 \times -4.483} \\ &= 0.3276 \end{aligned}$$

**Attribute Levels and Their Price Premiums**

Attribute Descriptor	Utility	Price Premium
<b>Brand</b>		
Minute Maid	1.472	41.03%
Fairlee	0.195	5.43%
President's Choice	0.000	0.00%
<b>Packaging Material</b>		
Glass	1.175	32.78%
Aluminum can	0.213	5.94%
Tetra-pak	0.000	0.00%
<b>Preparation</b>		
Fresh Frozen	-0.011	-0.32%
From Concentrate	-0.955	-26.64%
Pastuerized	0.000	0.00%

**Relative Importance of Attributes**

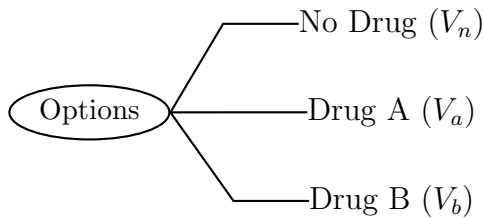
Attribute	Minimum Utility	Maximum Utility	Range	Relative Importance
Brand	0.000	1.472	1.472	27.27%
Packaging	0.000	1.175	1.175	21.79%
Preparation	-0.955	0.000	0.955	17.71%
Price†	-4.483	-2.690	1.793	33.23%
Total			5.395	

† Because the maximum price in this situation is \$1.00 and the minimum price is \$0.60, to obtain minimum and maximum utility, multiply price coefficient by these prices

In words, glass packaging material thus may get 32.76% price premium over tetra-pak. Similar calculations for aluminum can would have price premium of 5.4%. Note that in this example tetra-pak gets zero percent or no price premium. Tables provide summary for all design attributes and willingness to pay for them. Note that negative price premiums imply that price discount may be needed to keep utility at the same level. In addition, price premiums are always relative to some other product feature and changing the base feature (one that has zero premium) does not alter relative price premiums. Finally, to compute overall price premium for a brand with variety of features, one may add all price premiums to determine overall valuation of new or revised brand. For example, if Minute Maid (premium of 41.03%) is available in glass (premium of 32.78%), and prepared from fresh frozen (discount of 0.32%), then these respondents will be willing to pay 73.49% price premium over President's Choice juice sold in tetra-pak and prepared using pasteurizing.

### Comparing Multinomial to Nested Logit Model

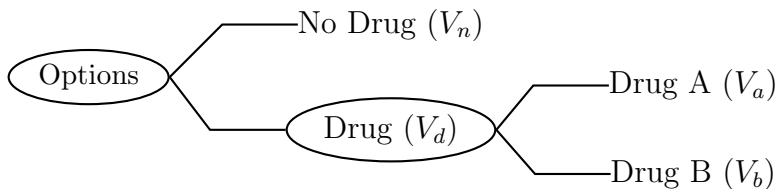
Consider following situation in which a respondent is facing three alternatives.



Let us denote respondent’s utility associated with alternative “No Drug” to be  $V_n$  and  $V_a$  and  $V_b$  to be utility associated with drug A and B respectively. Suppose that we are interested in determining the likelihood that a respondent would choose No Drug option. For above described model, we would write

$$\text{Prob(No Drug)} = \frac{\exp(V_n)}{\exp(V_n) + \exp(V_a) + \exp(V_b)}$$

An alternative to above model is to consider two level tree structure. In the first level, we may consider that respondent makes a choice between “No Drug” and Drug. In the second level, respondent makes a choice among drugs, conditional on the fact that at the first level Drug is chosen alternative.



### Nested Logit Model

This model formulation requires us to examine two interrelated equations. First we must examine the likelihood of choosing between No Drug (utility associated with this alternative denoted by  $V_n$ ) and Drug (utility associated with it denoted by  $V_d$ ). Note that utility associated with Drug option depends upon attributes associated with both drugs. We will examine below procedure for determining this utility. Thus, we may write

$$\text{Prob(No Drug)} = \frac{\exp(V_n)}{\exp(V_n) + \exp(V_d)}$$

Then, we must indicate the likelihood of choosing Drug A, given that individual has chosen Drug as an alternative. Thus, we may write

$$\text{Prob(Drug A | Drug)} = \frac{\exp(V_a)}{\exp(V_a) + \exp(V_b)}$$

This model with two equations is called nested logit model. Let us examine comparison among two models: nested and non-nested logit. To illustrate model comparison, consider the likelihood of choosing Drug A given by both models. For non-nested model, it is

$$\text{Prob}(\text{Drug A}) = \frac{\exp(V_a)}{\exp(V_n) + \exp(V_a) + \exp(V_b)}, \quad (1)$$

and for nested model it is

$$\begin{aligned} \text{Prob}(\text{Drug A}) &= \text{Prob}(\text{Drug}) \times \text{Prob}(\text{Drug A} \mid \text{Drug}) \\ &= \frac{\exp(V_d)}{\exp(V_n) + \exp(V_d)} \times \frac{\exp(V_a)}{\exp(V_a) + \exp(V_b)} \end{aligned} \quad (2)$$

Suppose we assigned utility associated with Drug option as a function of two alternatives. More specifically,

$$\begin{aligned} V_d &= \theta \log[\exp(V_a) + \exp(V_b)] \\ &= [\exp(V_a) + \exp(V_b)]^\theta. \end{aligned}$$

Then it can be shown that when  $\theta = 1$  results in non-nested logit model while  $\theta \neq 1$  results in nested logit model. As a result of this formulation, we may use above hypothesis test to determine whether the choice model is nested or non-nested logit. In addition, note also that if there are two alternatives and  $\theta$  is about  $1/2$ , then we would think that both drug options are equally weighted to arrive at overall utility of drug option. One of the important advantage associated with the nested logit model is that addition of new alternative, only affects branch within which the alternative is added. Let us illustrate this particular idea. Consider we have new drug (call it Drug C) and it is similar to Drug A and B. As result of this change in the choice set, choice probabilities of Drug A and B will be strongly influenced but probability that no drug option is chosen would remain unaffected.

### Duration Data and Hazard Function.

#### Preliminaries and Definitions

Suppose the random variable  $T$  denotes the purchase time for a household. Consider a purchase occasion at time  $t$  where  $t \geq 0$  and it is the elapsed time since the last purchase. Let us denote cumulative distribution function (CDF) by  $F(t) = \text{Prob}(T \leq t)$ . That is the likelihood that a purchase will occur at time  $t$ . We can also look at  $f(t)$ , the likelihood of purchase at time  $t$ . This is often called probability density or distribution function (PDF). We can also define couple of more functions. Suppose we want to know a ‘survival’ function. That is the likelihood that household will not purchase up to time  $t$ . This is the complementary function to the cumulative distribution function, namely,  $S(t) = 1 - F(t)$ . Note that

$$f(t) = \frac{\partial F(t)}{\partial t} = -\frac{\partial S(t)}{\partial t}.$$

We can also define the odds ratio. That is, the likelihood of purchase at time  $t$  to the likelihood that purchase has not occurred up to time  $t$ . This can be written in number of equivalent mathematical forms.

$$\begin{aligned} h(t) &= \frac{f(t)}{S(t)} \\ &= \frac{f(t)}{1 - F(t)} \\ &= -\frac{\partial S(t)}{\partial t} \frac{1}{S(t)} \\ &= -\frac{\partial \log S(t)}{\partial t}. \end{aligned}$$

Note that  $h(t)$  is the rate at which random event is expected to re-occur. Some like to call it as the hazard rate or hazard function. The hazard function provides a convenient definition of duration dependence. *Positive duration dependence* exists at the point  $t^*$  if  $\partial h(t)/\partial t > 0$  at  $t = t^*$ . Positive duration dependence means that probability of a purchase increase with increase in the length of non purchase. Purchase behaviour in a *product category* is believed to be with a positive duration dependence. *Negative duration dependence* exists at the point  $t^*$  if  $\partial h(t)/\partial t < 0$  at  $t = t^*$ . For minor brands, the likelihood of purchasing that brand decreases, as the elapsed time increases.

### Truncated or Censored Distributions

Most of purchase (event) history data are often collected over a restricted period of time. In other situation, an event may not occur until some minimum time is elapsed. That is, instead of observing the value of random variable  $T$  over entire range of values, we observe only over a range  $a < t \leq b$ . If  $F(t)$  is the original (untruncated) distribution of  $T$ , then the truncated distribution is

$$F(t|a < t \leq b) = \begin{cases} 0 & \text{for } t \leq a, \\ \frac{F(t) - F(a)}{F(b) - F(a)} & \text{for } a < t \leq b, \\ 1 & \text{for } t > b. \end{cases} \quad (3)$$

In terms of survival distribution function, we have similar results. That is,

$$S(t|a < t \leq b) = \begin{cases} 1 & \text{for } a \leq t, \\ \frac{S(t) - S(b)}{S(a) - S(b)} & \text{for } a < t \leq b, \\ 0 & \text{for } t > b. \end{cases} \quad (4)$$

Truncation (censoring) by exclusion of values less than  $a$  is called *truncation from below*, or *left-hand truncation*. In some instances, left-hand truncation may occur because product may not fail before certain time. On the other hand, truncation by exclusion values greater than  $b$  is called *truncation from above* or *right-hand truncation*. In equation (3) and (4) we have

a double truncation. Finally, if the original distribution function (PDF) is  $f(t)$ , then the truncated distribution has PDF

$$f(t|a < t \leq b) = \begin{cases} \frac{f(t)}{F(b)-F(a)} = \frac{f(t)}{S(a)-S(b)} & \text{for } a < t \leq b, \\ 0 & \text{elsewhere.} \end{cases} \quad (5)$$

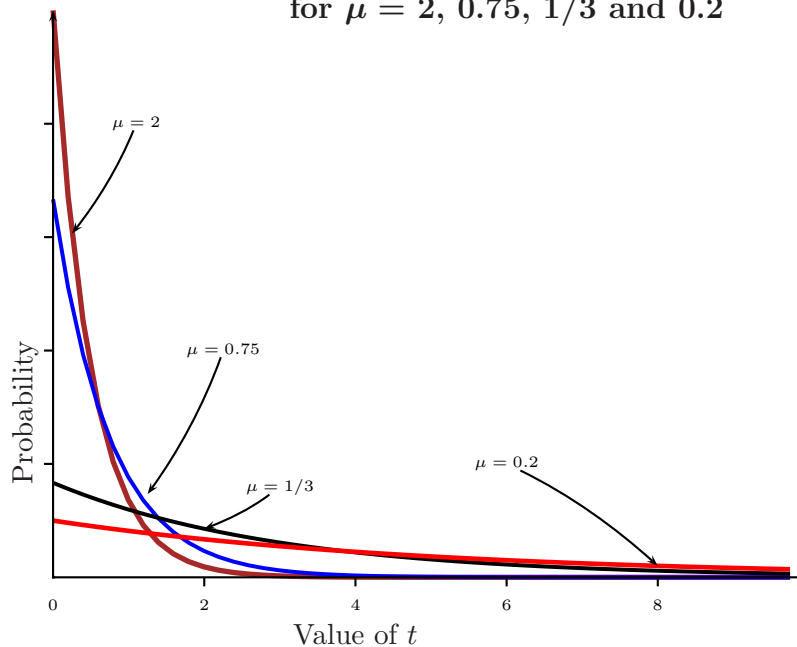
We can now define the hazard function in such instance which is

$$h(t|a < t < b) = \frac{f(t|a < t < b)}{S(t|a < t < b)} = \frac{f(t)}{S(t) - S(b)} = h(t) \frac{S(t)}{S(t) - S(b)} \quad (6)$$

Note that the hazard function for the truncated and untruncated distribution depends only on  $b$ , the upper truncation point, and not on  $a$ , the lower truncation point. Note also that when  $t = b$ , value of the hazard function is not meaningful (because we have chosen to stop collecting data at this point).

### Exponential Probability Distribution Function

for  $\mu = 2, 0.75, 1/3$  and  $0.2$



### Exponential Distribution

The PDF of an exponential distribution is

$$f(t) = \mu \exp(-\mu t), \quad \mu > 0, \text{ and } t > 0.$$

Since CDF of this distribution is  $F(t) = 1 - \exp(-\mu t)$ , the survival distribution is  $S(t) = \exp(-\mu t)$  and the hazard rate is equal to  $\mu$ , constant for all  $t > 0$ . It can be shown that (see below application of the maximum likelihood estimation)  $1/\mu$  is the average purchase rate or

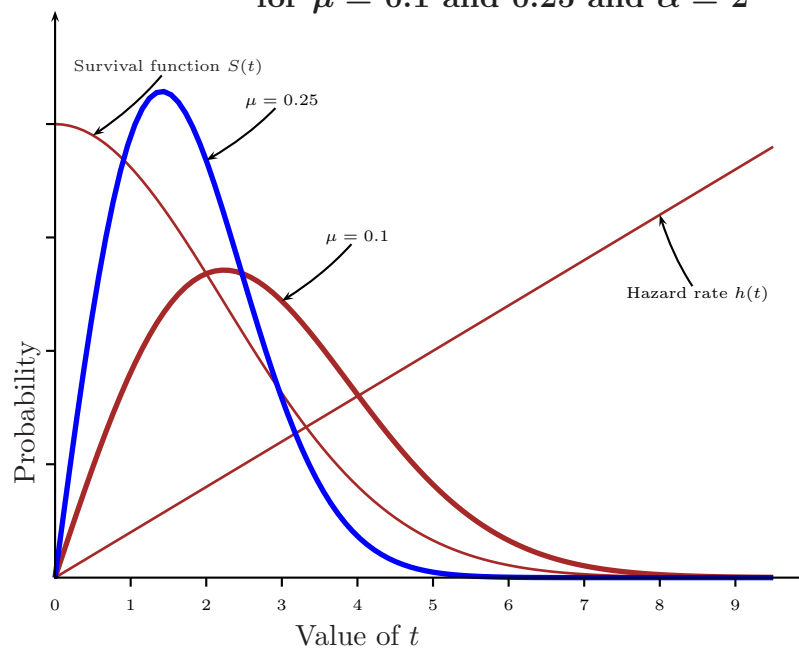
failure rate. An illustration of such distribution is given above when  $\mu = 2, 0.75, 1/3$  and  $0.2$ . Note the effect of censoring changes exponential distribution to a logistic distribution with the mean shift. Specifically, substituting values of various terms in equation (6) gives

$$h(t|a < t < b) = \mu \frac{\exp(-\mu t)}{\exp(-\mu t) - \exp(-\mu b)}$$

which can be further simplified to

$$h(t|a < t < b) = \frac{\mu}{1 - \exp[-\mu(b - t)]}$$

**Weibull Probability Distribution Function  
for  $\mu = 0.1$  and  $0.25$  and  $\alpha = 2$**



**Weibull Distribution**

In this distribution<sup>5</sup> we have added one extra parameter to an exponential distribution. PDF for Weibull distribution is given by

$$f(t) = \mu \alpha t^{\alpha-1} \exp(-\mu t^\alpha), \quad \mu, \alpha > 0, \text{ and } t > 0.$$

Note that when  $\alpha = 1$ , we obtain an exponential distribution. Since CDF of this distribution is  $F(t) = 1 - \exp(-\mu t^\alpha)$ , the survival distribution is  $S(t) = \exp(-\mu t^\alpha)$  and the hazard rate is

<sup>5</sup>A general Weibull distribution has three parameters, location, scale and shape. We are assuming that location parameter is zero. By replacing  $t$  with  $t - \theta$ , where  $\theta$  is a location parameter, we obtain generalized Weibull distribution.

equal to  $h(t) = \mu\alpha t^{\alpha-1}$ . Note that  $\mu$  in this case is scale parameter and  $\alpha$  is shape parameter. An illustration of such distribution with  $\mu = 0.75$  and  $\alpha = 2.5$  is given in exhibit 2.

There are number of such distribution one could employ to construct the survival function. Gamma, logistic, lognormal, log-logistic and Erlang-2 are some of these distributions commonly used in the literature. In choosing a particular distribution one must make trade-off between flexibility and ease of estimation. A more complicated distribution may require computer intensive estimation but may offer variety of shapes and forms by changing parameter values.

### Maximum Likelihood Estimation

Suppose that the family of duration distribution under consideration has been specified, so that the data distribution is known up to a vector of parameters  $\theta$ . For an exponential distribution  $\theta$  is equal to  $\mu$  whereas a Weibull distribution  $\theta$  is equal to  $\{\mu, \alpha\}$ . If a sample of  $n$  completed spells were available and each individual spell independent of the others<sup>6</sup>, the likelihood function is  $\mathcal{L}(\theta) = \prod_{i=1}^n f(t_i, \theta)$ . The likelihood function is the joint probability distribution of the sample as a function of parameters  $\theta$ . When a spell is censored, at duration  $t_u$  for example, the only information available is that the duration was at least  $t_u$ . Consequently the contribution to likelihood from that observation is the value of the survivor function,  $S(t_u, \theta)$ , the probability that the duration is longer than  $t_u$ . Let  $d_i = 1$  if the  $i$ th spell is uncensored,  $d_i = 0$  if censored. Then the logarithm of the likelihood function ( $\mathcal{LL}$ ) is

$$\mathcal{LL} = \sum_{i=1}^n d_i \log f(t_i, \theta) + \sum_{i=1}^n (1 - d_i) \log S(t_i, \theta),$$

which has completed spells contributing a density term  $f(t_i, \theta)$  and censored spells contributing a probability  $S(t_i, \theta)$ . But we also know that  $h(t_i, \theta) = f(t_i, \theta)/S(t_i, \theta)$ . Thus, we can write the log-likelihood function as

$$\mathcal{LL} = \sum_{i=1}^n d_i \log h(t_i, \theta) + \sum_{i=1}^n \log S(t_i, \theta). \quad (7)$$

This is very useful form of the likelihood function. Note that if know the survival and hazard function, we can specify the log-likelihood function. Hence, model specification, estimation and interpretation all depend upon two simple concepts, survival and hazard rate. For example, consider an exponential distribution, we know that  $h(t_i, \theta) = \mu$  and  $S(t_i, \theta) = \exp(-\mu t_i)$ . Then, the maximum of the likelihood function will occur at

$$\hat{\mu} = \frac{\sum_{i=1}^n d_i}{\sum_{i=1}^n t_i}.$$

Note that  $\sum_{i=1}^n d_i$  is equal to the number of completed spells. If we treat censored spells as complete, the maximum likelihood estimate would be

$$\hat{\mu} = \frac{n}{\sum_{i=1}^n t_i}.$$

---

<sup>6</sup>We are also assuming that each event comes from the same distribution.

Since  $n \geq \sum_{i=1}^n d_i$ , ignoring censoring leads to upward asymptotic bias in the estimated hazard function, or overstatement of the conditional probability of ending a spell. Let us look at how this might help us estimate parameters for a real example. Table below contains data for the US Industrial Strikes during 1996 (Feb) to 1997 (March). Note that there are 33 total events and two strikes have not ended at the end of data collection. In order for us to obtain unbiased estimate of  $\mu$ . We need to divide  $\sum d_i = (33 - 2)$  by  $\sum t_i$  which is easy to add. Note that in this instance, up to the point of truncation the sum is 1782. This gives us estimate of  $\hat{\mu} = 0.0174$ . That is, implied expected duration of a strike is  $1/0.0174$  or about 57 days. If we ignored 5 strikes that lasted more than 100 days, then we get  $\hat{\mu} = 26/1158 = 0.0384$  or implied expected duration of a strike of about 26 days.

**Work Stoppages in the U.S. involving  
1,000 or more workers, 1996-97**

Begin Date	End Date	Duration	N of Employees	Strike Against
13/7/95	19/2/97	588	2,500	Detroit Free Press
4/1/96	4/2/96	32	30,000	Commercial Building Realty
21/1/96	3/7/96	165	1,100	Trailmobile
1/2/96	8/2/96	8	5,000	San Diego Public Schools
7/2/96	6/3/96	29	2,600	Yale University, Clerical
14/2/96	15/2/96	2	7,600	Litton Industries
15/2/96	20/3/96	35	3,500	Oakland Public Schools
17/2/96	19/2/96	3	1,800	Chrysler Corp
4/3/96	22/3/96	19	1,000	Nursing Home Industries
8/3/96	22/3/96	15	136,000	General Motors
27/3/96	23/4/96	28	1,100	Yale University, Hotel & Rest
15/4/96	16/4/96	2	1,500	General Motors
19/4/96	8/6/96	51	1,000	Crown Cork and Seal Co
22/4/96	29/4/96	8	1,000	United Technologies
14/5/96	25/6/96	43	14,500	Grocery Industry
30/5/96	9/6/96	11	1,000	Southeast Michigan Roofing
1/6/96	14/6/96	14	1,300	NBC Merchants
1/6/96	12/6/96	12	5,000	Northern Illinois Mason
3/6/96	2/7/96	30	4,900	Bay Area Cleaning Company
24/6/96	28/8/96	66	6,000	League of Voluntary Hospitals
5/6/96	16/9/96	104	6,700	McDonnell Douglas Aerospace
7/17/96	No Info.		2,500	National Steel and Shipbuilding
22/7/96	21/3/97	243	1,100	Pemco Aeroplex
12/9/96	12/10/96	31	3,800	Boise Cascade
5/9/96	29/9/96	25	1,300	Aluminum Company of America
1/10/96	Strike Cont.		3,800	Wheeling Pittsburgh Steel Corp
7/10/96	24/10/96	18	1,200	Holt Cargo Systems
30/10/96	6/11/96	8	5,100	General Motors
30/10/96	3/11/96	5	2,700	General Motors
1/11/96	2/3/97	122	1,500	Elevator Industries
12/11/96	29/11/96	18	1,800	IBP Inc
18/12/96	19/1/97	33	2,100	Giant Food Inc
14/3/97	27/3/97	14	2,700	General Motors

### Adding Explanatory Variables

Explanatory variables can affect the distribution of duration in many different ways. The proportional hazard specification is popular and simple to interpret. The effect of regressors

is to multiply the hazard function by a scale factor. These models are called the proportional hazard functions. In these models, a vector of explanatory variables  $x$  with unknown coefficients  $\beta$  and  $h_0$  is factored as  $h(t, x, \beta, h_0) = \phi(x, \beta)h_0(t)$  where  $h_0$  is a “baseline” hazard corresponding to  $\phi(\cdot) = 1$ . If we set the regressors at the mean value, then  $h(\cdot)$  function has an interpretation as the hazard function for the mean observation in the sample. Note that the coefficients designated by  $\theta$  previously have been separated into  $\beta$  and  $h_0$ . In this specification the effect of explanatory variables is to multiply the hazard  $h_0$  by a factor  $\phi$  which does not depend on duration of  $t$ . A specification of  $\phi$  in general use is  $\phi(x, \beta) = \exp(x'\beta)$ . This specification is convenient because non negativity of  $\phi$  does not require any restrictions on  $\beta$ . However, various steps in estimation and inference stay same as discussed above. We will now consider incorporation of explanatory variables to an exponential distribution.

We already noted that  $h(t, x, \beta\mu) = \mu = \exp(x'\beta)$ . Suppose that we assume that  $h_0(t) = 1$ . Then, log-likelihood function (see equation (7) above) can be re-written as

$$\mathcal{LL} == \sum_{i=1}^n d_i \log \exp(x'_i \beta) + \sum_{i=1}^n \log \exp(-\mu t_i).$$

If we substitute  $\mu = \exp(x'_i \beta)$  and simplify to obtain

$$\mathcal{LL} == \sum_{i=1}^n d_i x'_i \beta - \sum_{i=1}^n t_i \exp(x'_i \beta).$$

Although the function looks more complicated, the conceptual idea of obtaining parameters in this case is no different from our earlier instance<sup>7</sup>.

### Software for Estimating Duration time models

Many alternative software packages are available to estimate hazard functions. As one would expect, considerable effort goes into collecting and understanding data of this sort. Researching historical archives is often fun but tedious and time consuming. Modelling then becomes add-on but insightful. In SAS, there is PROC LIFEREG and LIMDEP also contains procedure called SURVIVAL. If you know how things work in terms of the likelihood function and number of observations are small, you could do estimation using LOTUS or EXCEL (that is what I did for the US strike data).

### Dependent Variables with Mixed Character

In most product categories consumers are either buyers or non-buyers. Much of Marketing is based on 80:20 rule which states that 20% of consumers purchase 80% in product category. Consider an example of understanding consumption of Tobacco and related products. We would suspect that a large proportion of respondents do not buy any tobacco and related products

<sup>7</sup>The maximum for this function occur at  $\sum_{i=1}^n d_i x'_i - \sum_{i=1}^n t_i \exp(x'_i \beta) x' = 0$ . A solution to these non-linear equations provide parameter estimates.

and some may not have purchased over their lifetime. Since our interest is on consumption, non-buyers by definition consume zero units and spend zero, while those consuming tobacco products would have non-zero expenditure and distribution of remaining observation may even be normal. Modelling such variable entails, examining two inter-related dependent variables. First, explaining factors that influence whether one makes any purchase(1) or not (0) and then explaining variations among those who consume any positive quantity. Formally, these two related models can be stated as

$$\mathbf{y} = \begin{cases} \mathbf{x}'\boldsymbol{\beta} + \mathbf{u} & \text{if } y > 0 \\ 0 & \text{if } y = 0 \end{cases}$$

where  $\mathbf{y}$  is a vector of size  $n \times 1$  or  $n$  observations,  $\mathbf{x}$  is a matrix of size  $p \times n$  observations and  $\boldsymbol{\beta}$  and  $\mathbf{u}$  are unobserved parameter vector ( $p \times 1$ ) and random noise respectively. We may assume that  $\mathbf{u}$ , residual vector is independently and normally distributed with mean zero and a common variance of  $\sigma^2$ . Our task is to estimate  $\boldsymbol{\beta}$  and  $\sigma^2$  using  $n$  observations on  $\mathbf{y}$  and  $\mathbf{x}$ . To broadly appreciate issues of estimation and parameter interpretation, suppose there are  $n_0$  observations for which dependent variable is zero and  $n_1$  are remaining observations for which dependent variable values are positive. For the observations that have zero dependent variable value, we want to know,  $\text{Prob}(y_i = 0)$  which is equal to  $\text{Prob}(u_i < -\boldsymbol{\beta}\mathbf{x}_i)$ . For the observations that have positive values associated with dependent variable, we need to have to account for two probabilities, probability that dependent variable is greater than zero and conditional probability that dependent variable takes specific value. Formally, two probabilities jointly determine outcome, that is,  $\text{Prob}(y_i > 0) \times \text{Prob}(y_i | y_i > 0)$ . Note that  $\text{Prob}(y_i > 0)$  can be estimated using logit or probit model that we discussed earlier and note that for a sample,  $\text{Prob}(y_i > 0) = \frac{n_1}{n}$ . Moreover,  $\text{Prob}(y_i | y_i > 0)$  reflects regression equation model that include  $n_1$  observations. Consequently, it is possible to estimate such a model, if we just have  $n_1$  observations and we know total sample size. However, if factors influencing the likelihood that dependent variable is zero are different than factors influencing positive values of dependent variable, then it would necessary to estimate two separate equations. These ideas are illustrated below using data from Statistics Canada for Family Expenditure Survey. We will use expenditure on Tobacco and related products by 1051 households for year 1996.

```
--> RESET
--> read;file=f:\multiv\tob963.dat;
    nvar=7;nobs=1052;
    names=gender,city,educ,agegr,marst,prov,tobspend$
--> create;age = 27 + (agegr-1)*5; agesq = age*age;
    if(marst=1)marst1=1;      | Married
    if(marst=2)marst1=1;      | Common Law
    if(educ=1)educ1=1;        | Less than 9 years
    if(educ=2)educ2=1;        | Completed Secondary school
    if(educ=3)educ3=1;        | Some Post Secondary
    if(educ=4)educ4=1;        | Complete Post Sec,
    if(prov=10)prov1=1;       | NF
```

```

if(prov=11)prov2=1;      | PEI
if(prov=12)prov3=1;      | NS
if(prov=13)prov4=1;      | NB
if(prov=24)prov5=1;      | Quebec
if(prov=35)prov6=1;      | ONT
if(prov=46)prov7=1;      | Manitoba
if(prov=47) prov8=1;      | Sask
if(prov=48) prov9=1;      | Alberta
if(prov=59)prov10=1$     | BC
--> create;if(tobspend>0)tob0 = 1$
--> create;if(tobspend>0)lgtob=log(tobspend);if(tobspend=0)lgtob=0$
--> logit;lhs=tob0;rhs=one,marst1,educ1,educ2,educ3,educ4,age,agesq,
    prov1,prov2,prov3,prov4,prov5,prov6,prov7,prov8,prov9,prov10;hold$

```

```

+-----+
| Multinomial Logit Model |
| Maximum Likelihood Estimates |
| Dependent variable      TOB0 |
| Weighting variable      ONE  |
| Number of observations   1051 |
| Iterations completed     5   |
| Log likelihood function  -660.3911 |
| Restricted log likelihood -723.0415 |
| Chi-squared             125.3009 |
| Degrees of freedom      18   |
| Significance level      .0000000 |
+-----+

```

```

+-----+-----+-----+-----+-----+
|Variable | Coefficient | Standard Error |b/St.Er.|P[|Z|>z] | Mean of X|
+-----+-----+-----+-----+-----+
Characteristics in numerator of Prob[Y = 1]
Constant - .8227860436      .95076623      -.865      .3868
MARST1   - .5383592658      .18145368      -2.967     .0030      .64700285
MARST2   - .5314443307      .25064040      -2.120     .0340      .13510942
EDUC1     .9185246663      .28525849      3.220     .0013     .10941960
EDUC2     .9922411291      .19857083      4.997     .0000     .40437678
EDUC3     .8196340354      .28921916      2.834     .0046     .76117983E-01
EDUC4     .6447808937      .21018353      3.068     .0022     .24072312
AGE       .7797101071E-01   .31586504E-01   2.468     .0136     53.346337
AGESQ    - .1084275973E-02 .28420721E-03   -3.815     .0001     3103.5576
PROV1    - .2188769188      .45103574      -.485     .6275     .55185538E-01
PROV2    - .3042012517      .46847183      -.649     .5161     .48525214E-01
PROV3    - .2727770470      .43451753      -.628     .5302     .66603235E-01
PROV4    - .4372843975      .44089499      -.992     .3213     .63748811E-01
PROV5    - .5685975393      .39244338      -1.449     .1474     .15509039
PROV6    - .6402346542      .38141419      -1.679     .0932     .22835395
PROV7    - .7175243353      .44666859      -1.606     .1082     .59942912E-01
PROV8    - .3998978613      .42898936      -.932     .3512     .73263559E-01
PROV9    - .6764257375      .42379070      -1.596     .1105     .80875357E-01
PROV10   - .6839296835      .40001119      -1.710     .0873     .13415794

```

Frequencies of actual & predicted outcomes  
 Predicted outcome has maximum probability.

Actual	Predicted		Total
	0	1	
0	391	188	579
1	176	296	472
Total	567	484	1051

```
--> SELECTION;Lhs=LGT0B;Rhs=ONE,CITY,PROV1,PROV2,PROV3,PROV4,PROV5,PROV6,PROV7,PROV8,PROV9,PROV10$
```

```
+-----+
| Sample Selection Model |
| Logit selection equation based on TOB0 |
| Selection rule is: Observations with TOB0 = 1 |
| Results of selection: |
| Data points Sum of weights |
| Data set 1051 1051.0 |
| Selected sample 472 472.0 |
+-----+
```

```
+-----+
| Sample Selection Model |
| Two stage least squares regression Weighting variable = none |
| Dep. var. = LGTOB Mean= 6.492784151 , S.D.= 1.670924972 |
| Model size: Observations = 472, Parameters = 13, Deg.Fr.= 459 |
| Residuals: Sum of squares= 1206.201524 , Std.Dev.= 1.62108 |
| Fit: R-squared= .056777, Adjusted R-squared = .03212 |
| (Note: Not using OLS. R-squared is not bounded in [0,1] |
| Model test: F[ 12, 459] = 2.30, Prob value = .00743 |
| Diagnostic: Log-L = -891.1666, Restricted(b=0) Log-L = -911.5526 |
| LogAmemiyaPrCrt.= .993, Akaike Info. Crt.= 3.831 |
| Standard error corrected for selection..... 1.6618 |
| Correlation of disturbance in regression |
| and Selection Criterion (Rho)..... -.35563 |
+-----+
```

Variable	Coefficient	Standard Error	b/St.Er.	P[ Z >z]	Mean of X
Constant	8.155086422	.47526130	17.159	.0000	
CITY	-.5204952732	.21433768	-2.428	.0152	1.1567797
PROV1	-.3966208745	.47352914	-.838	.4023	.63559322E-01
PROV2	.2555606614	.50177365	.509	.6105	.48728814E-01
PROV3	-.8388960431	.45770589	-1.833	.0668	.76271186E-01
PROV4	-.4694947038	.47004006	-.999	.3179	.67796610E-01
PROV5	-.5111182225	.41223648	-1.240	.2150	.16101695
PROV6	-.8732614075	.40513160	-2.156	.0311	.21610169
PROV7	-.7913110912	.49537539	-1.597	.1102	.52966102E-01
PROV8	-.5460820728	.45469689	-1.201	.2298	.78389831E-01
PROV9	-.3020612700	.47228242	-.640	.5224	.72033898E-01
PROV10	-.8570875187	.43087539	-1.989	.0467	.11864407
LAMBDA	-.5909911063	.31048021	-1.903	.0570	.79942909

```
--> tobit;lhs=lgtoB;rhs=one,city,marst1,marst2,educ1,educ2,educ3,educ4,age,ag...
prov2,prov3,prov4,prov5,prov6,prov7,prov8,prov9,prov10$
```

```
+-----+
| Limited Dependent Variable Model - CENSORED Regression |
| Ordinary least squares regression Weighting variable = none |
| Dep. var. = LGTOB Mean= 2.915884033 , S.D.= 3.419380963 |
| Model size: Observations = 1051, Parameters = 20, Deg.Fr.= 1031 |
| Residuals: Sum of squares= 10865.20585 , Std.Dev.= 3.24631 |
| Fit: R-squared= .114979, Adjusted R-squared = .09867 |
| Model test: F[ 19, 1031] = 7.05, Prob value = .00000 |
| Diagnostic: Log-L = -2718.7796, Restricted(b=0) Log-L = -2782.9661 |
| LogAmemiyaPrCrt.= 2.374, Akaike Info. Crt.= 5.212 |
+-----+
```

Variable	Coefficient	Standard Error	b/St.Er.	P[ Z >z]	Mean of X
----------	-------------	----------------	----------	----------	-----------

Constant	2.921327394	1.4594867	2.002	.0453	
CITY	-.2785756645	.29943303	-.930	.3522	1.1512845
MARST1	-.7764907289	.26686061	-2.910	.0036	.64700285
MARST2	-.7810881624	.37846490	-2.064	.0390	.13510942
EDUC1	1.437113871	.42528457	3.379	.0007	.10941960
EDUC2	1.712149258	.29992327	5.709	.0000	.40437678
EDUC3	1.546232414	.44507878	3.474	.0005	.76117983E-01
EDUC4	1.174224750	.32060695	3.663	.0002	.24072312
AGE	.9890944631E-01	.45556133E-01	2.171	.0299	53.346337
AGESQ	-.1419634700E-02	.39661182E-03	-3.579	.0003	3103.5576
PROV1	-.6296806698	.69236394	-.909	.3631	.55185538E-01
PROV2	-.5019064284	.71239370	-.705	.4811	.48525214E-01
PROV3	-.9391489192	.67218137	-1.397	.1624	.66603235E-01
PROV4	-.9924552744	.67761237	-1.465	.1430	.63748811E-01
PROV5	-1.247905853	.60115280	-2.076	.0379	.15509039
PROV6	-1.493019585	.58318342	-2.560	.0105	.22835395
PROV7	-1.577269522	.68076416	-2.317	.0205	.59942912E-01
PROV8	-.9783236208	.65738542	-1.488	.1367	.73263559E-01
PROV9	-1.321013217	.64770278	-2.040	.0414	.80875357E-01
PROV10	-1.526874284	.60985629	-2.504	.0123	.13415794

Normal exit from iterations. Exit status=0.

```

+-----+
| Limited Dependent Variable Model - CENSORED |
| Maximum Likelihood Estimates                |
| Dependent variable          LGTOB          |
| Weighting variable          ONE            |
| Number of observations      1051          |
| Iterations completed        6            |
| Log likelihood function     -1897.735     |
| Threshold values for the model:           |
| Lower= .0000      Upper=++infinity      |
+-----+

```

Variable	Coefficient	Standard Error	b/St.Er.	P[ Z >z]	Mean of X
Primary Index Equation for Model					
Constant	-2.814062997	3.1888361	-.882	.3775	
CITY	-.2761657729	.64702043	-.427	.6695	1.1512845
MARST1	-1.760714404	.58324983	-3.019	.0025	.64700285
MARST2	-1.669279753	.81362135	-2.052	.0402	.13510942
EDUC1	3.312788578	.97169469	3.409	.0007	.10941960
EDUC2	3.663004937	.67716402	5.409	.0000	.40437678
EDUC3	3.215005566	.96662513	3.326	.0009	.76117983E-01
EDUC4	2.570960692	.71804922	3.580	.0003	.24072312
AGE	.3081573261	.10249954	3.006	.0026	53.346337
AGESQ	-.4083656727E-02	.91952842E-03	-4.441	.0000	3103.5576
PROV1	-.8458466088	1.4366519	-.589	.5560	.55185538E-01
PROV2	-.8607615033	1.4993160	-.574	.5659	.48525214E-01
PROV3	-1.321600759	1.3948665	-.947	.3434	.66603235E-01
PROV4	-1.717494738	1.4174103	-1.212	.2256	.63748811E-01
PROV5	-2.208540254	1.2489457	-1.768	.0770	.15509039
PROV6	-2.581038653	1.2125142	-2.129	.0333	.22835395
PROV7	-2.901702811	1.4482502	-2.004	.0451	.59942912E-01
PROV8	-1.636791323	1.3729061	-1.192	.2332	.73263559E-01
PROV9	-2.325326772	1.3614506	-1.708	.0876	.80875357E-01
PROV10	-2.867735834	1.2809649	-2.239	.0252	.13415794
Disturbance standard deviation					
Sigma	6.212619904	.23434311	26.511	.0000	