EKT 720
LPM/LOGIT/PROBIT

Modeling binary response variables:

Grouped data:

Linear Probability modeling:

1) Calculate $P_i = \frac{n}{N}$ per group/category
2) Model $P_i = \beta_0 + \beta_1 X_i + u_i$ by the usual OLS estimate $\hat{P}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
3) Use the valid $\hat{P}$ values (between 0 and 1) to calculate $w_i = \sqrt{N_i \hat{P}_i (1 - \hat{P}_i)}$
   Why?
4) Transform model as in 1. by dividing by w_i
5) New model is
   $P_i / w_i = \beta_0 / w_i + \beta_1 X_i / w_i + u_i / w_i$
   $P^*_i = \beta_0 Z^*_1 + \beta_1 Z^*_2 + u^*_i, \quad where$
   $P^*_i = P_i / w_i$
   $Z^*_1 = 1 / w_i$
   $Z^*_2 = X_i / w_i$
   $u^*_i = u_i / w_i$
6) Estimate the relevant parameters and transform back
7) Evaluate significance and goodness of fit

In the textbook $w_i = N_i \hat{P}_i (1 - \hat{P}_i)$ and therefore they divide with $\sqrt{w_i}$

Example (LPM grouped) - owning ??

```
                        The REG Procedure
                          Model: MODEL1
                     Dependent Variable: p1

           Number of Observations Read           8
           Number of Observations Used           8

                      Analysis of Variance
                              Sum of         Mean
Source                  DF     Squares       Square   F Value   Pr > F

Model                    1     0.69645      0.69645    137.08   <.0001
Error                    6     0.03048      0.00508
Corrected Total          7     0.72693

        Root MSE              0.07128    R-Square     0.9581
        Dependent Mean        0.41255    Adj R-Sq     0.9511
        Coeff Var            17.27774
```

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.36008	0.07064	-5.10	0.0022
x	1	0.00429	0.00036662	11.71	<.0001

Obs	x	n1	n2	p1	pred
1	75	2	30	0.06667	-0.03816
2	105	4	45	0.08889	0.09062
3	135	10	60	0.16667	0.21939
4	165	20	70	0.28571	0.34816
5	195	30	75	0.40000	0.47693
6	225	35	60	0.58333	0.60570
7	255	40	50	0.80000	0.73447
8	285	50	55	0.90909	0.86325

Obs	x	n1	n2	p1	pred	w	pstar	z1star	z2star
1	105	4	45	0.08889	0.09062	1.92567	0.04616	0.51930	54.526
2	135	10	60	0.16667	0.21939	3.20553	0.05199	0.31196	42.115
3	165	20	70	0.28571	0.34816	3.98574	0.07168	0.25089	41.398
4	195	30	75	0.40000	0.47693	4.32552	0.09247	0.23119	45.081
5	225	35	60	0.58333	0.60570	3.78545	0.15410	0.26417	59.438
6	255	40	50	0.80000	0.73447	3.12267	0.25619	0.32024	81.661
7	285	50	55	0.90909	0.86325	2.54812	0.35677	0.39245	111.847

The REG Procedure
Model: MODEL1
Dependent Variable: pstar

Number of Observations Read	7
Number of Observations Used	7

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.23406	0.11703	518.58	<.0001
Error	5	0.00113	0.00022567		
Uncorrected Total	7	0.23519			

Root MSE	0.01502	R-Square	0.9952
Dependent Mean	0.14705	Adj R-Sq	0.9933
Coeff Var	10.21566		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
z1star	1	-0.42536	0.04578	-9.29	0.0002
z2star	1	0.00464	0.00023319	19.90	<.0001

Final estimated model: $\hat{P} = -0.42536 + 0.00464X$

SAS program:

```
options ls=72 nodate pageno=1 ;

data a ;
input       x      n1    n2     ;
p1=n1/n2 ;
cards;
     75      2    30
    105      4    45
    135     10    60
    165     20    70
    195     30    75
    225     35    60
    255     40    50
    285     50    55
;


proc reg data=a ;
model p1 = x ;
output out=b p=pred ;
run ;

proc print data=b;
run ;



data b ;
set b ;
if pred >= 1 then delete ;
if pred <=0 then delete ;
w = sqrt(n2*pred*(1-pred)) ;
pstar = p1/w ;
z1star=1/w ;
z2star=x/w ;
run ;

proc print data=b;
run ;


proc reg data=b ;
model pstar = z1star z2star / noint ;
run ;
```

LOGIT modeling:

General:

- Estimation of the following non linear regression model
- Model not linear in parameters
- Linear transformation leads to a linear model in terms of the ln(odds) and X.

Linear Transformation:

$$P_i = \frac{1}{1+e^{-Z_i}}, \quad with \quad Z_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...+ \beta_k X_{ki}$$

$$\frac{1}{P_i} = 1 + e^{-Z_i}$$

$$\frac{1}{P_i} - 1 = e^{-Z_i}$$

$$\frac{1-P_i}{P_i} = e^{-Z_i}$$

$$\ln(\frac{1-P_i}{P_i}) = -Z_i$$

$$-\ln(\frac{1-P_i}{P_i}) = Z_i$$

$$\ln(\frac{P_i}{1-P_i}) = Z_i$$

$$\ln(odds) = Z_i$$

$$\ln(odds) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...+ \beta_k X_{ki}$$

$$l = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ...+ \beta_k X_{ki} \quad with \quad l = \ln(odds)$$

Process:

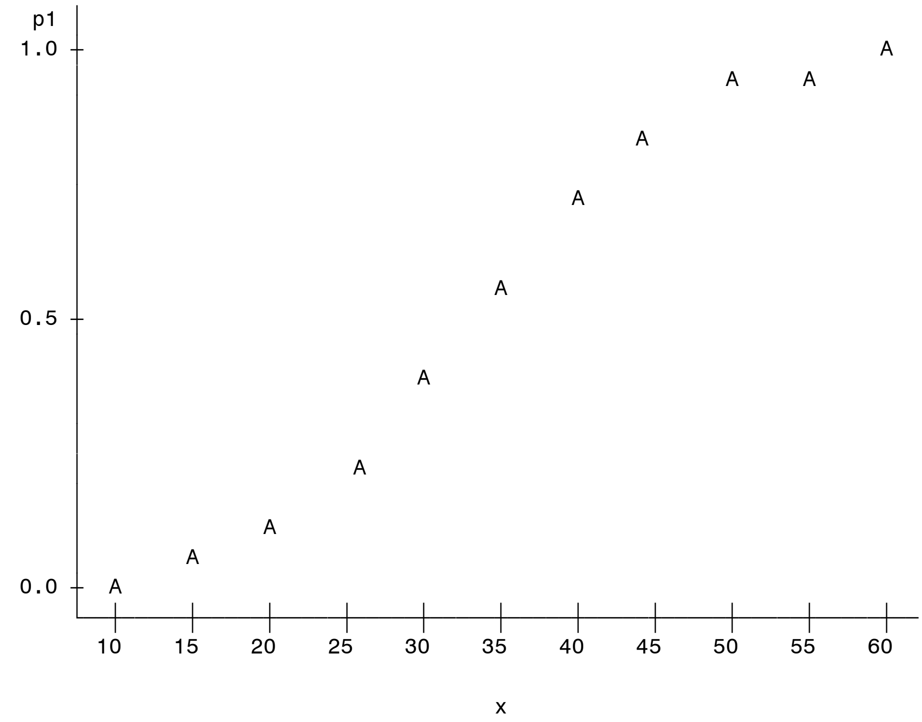
1) Estimate the model in transformed form $l = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}$
2) Calculate $\hat{l}$
3) Estimate $\hat{P}$ from $\hat{l}$
   a. $odds = e^{\hat{l}}$
   b. $\hat{P} = \dfrac{odds}{1 + odds}$
4) Calculate $w_i = \sqrt{\dfrac{1}{N_i \hat{P}_i (1 - \hat{P}_i)}}$
5) Transform model as in 1. by dividing by w_i
6) New model is
   $$l / w_i = \frac{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_k X_{ki}}{w_i}$$
7) Estimate the heteroscedasticity transformed model and transform back.
8) Transform back into non linear form.
9) Evaluate significance and goodness of fit

In the textbook $w_i = N_i \hat{P}_i (1 - \hat{P}_i)$ and therefore they divide with $\sqrt{w_i}$

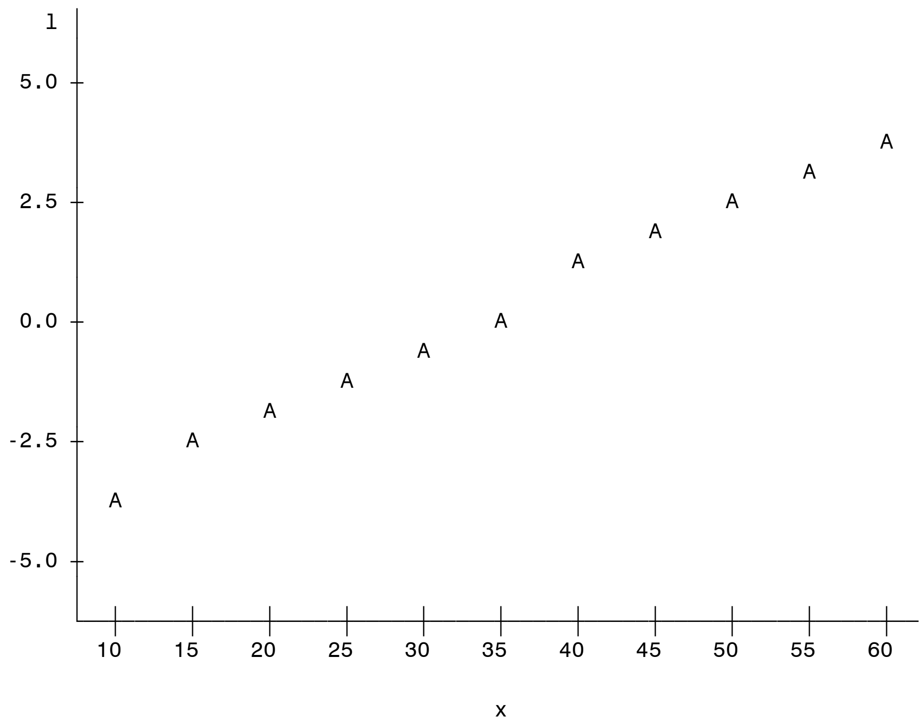
Example 1 (LOGIT grouped) - buying a product

Obs	x	n	r	p1	l
1	10	50	1	0.02	-3.89182
2	15	50	3	0.06	-2.75154
3	20	50	6	0.12	-1.99243
4	25	50	11	0.22	-1.26567
5	30	50	19	0.38	-0.48955
6	35	50	28	0.56	0.24116
7	40	50	37	0.74	1.04597
8	45	50	43	0.86	1.81529
9	50	50	46	0.92	2.44235
10	55	50	48	0.96	3.17805
11	60	50	49	0.98	3.89182

Plot of p1*x. Legend: A = 1 obs, B = 2 obs, etc.



Plot of l*x. Legend: A = 1 obs, B = 2 obs, etc.



The REG Procedure
Model: MODEL1
Dependent Variable: l

Number of Observations Read	11
Number of Observations Used	11

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	63.59426	63.59426	3984.94	<.0001
Error	9	0.14363	0.01596		
Corrected Total	10	63.73789			

Root MSE	0.12633	R-Square	0.9977
Dependent Mean	0.20215	Adj R-Sq	0.9975
Coeff Var	62.49226		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-5.12029	0.09252	-55.34	<.0001
x	1	0.15207	0.00241	63.13	<.0001

Obs	x	n	r	p1	l	pred	odds
1	10	50	1	0.02	-3.89182	-3.59959	0.0273
2	15	50	3	0.06	-2.75154	-2.83924	0.0585
3	20	50	6	0.12	-1.99243	-2.07890	0.1251
4	25	50	11	0.22	-1.26567	-1.31855	0.2675
5	30	50	19	0.38	-0.48955	-0.55820	0.5722
6	35	50	28	0.56	0.24116	0.20215	1.2240
7	40	50	37	0.74	1.04597	0.96250	2.6182
8	45	50	43	0.86	1.81529	1.72285	5.6004
9	50	50	46	0.92	2.44235	2.48319	11.9795
10	55	50	48	0.96	3.17805	3.24354	25.6243
11	60	50	49	0.98	3.89182	4.00389	54.8110

Obs	lh	w	lstar	z1star	z2star
1	0.02661	0.87876	-4.42878	1.13797	11.380
2	0.05524	0.61905	-4.44475	1.61537	24.231
3	0.11117	0.44990	-4.42856	2.22269	44.454
4	0.21106	0.34657	-3.65199	2.88543	72.136
5	0.36396	0.29393	-1.66552	3.40216	102.065
6	0.55037	0.28429	0.84830	3.51755	123.114
7	0.72362	0.31623	3.30759	3.16223	126.489
8	0.84850	0.39444	4.60224	2.53526	114.087
9	0.92296	0.53034	4.60526	1.88559	94.279
10	0.96244	0.74382	4.27261	1.34441	73.943
11	0.98208	1.06611	3.65050	0.93799	56.280

The REG Procedure
Model: MODEL1
Dependent Variable: lstar

Number of Observations Read	11
Number of Observations Used	11

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	160.45807	80.22903	2723.56	<.0001
Error	9	0.26512	0.02946		
Uncorrected Total	11	160.72319			

Root MSE	0.17163	R-Square	0.9984
Dependent Mean	0.24244	Adj R-Sq	0.9980
Coeff Var	70.79209		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
z1star	1	-5.05654	0.07283	-69.43	<.0001
z2star	1	0.15160	0.00206	73.73	<.0001

Final estimated model: $\hat{P} = \dfrac{1}{1 + e^{(-5.05654 + 0.15160X)}}$

Obs	x	n	r	p1	l	lh	odds	ph	w	lstar
1	10	50	1	0.02	-3.89182	-3.59959	0.0273	0.02661	0.87876	-4.42878
2	15	50	3	0.06	-2.75154	-2.83924	0.0585	0.05524	0.61905	-4.44475
3	20	50	6	0.12	-1.99243	-2.07890	0.1251	0.11117	0.44990	-4.42856
4	25	50	11	0.22	-1.26567	-1.31855	0.2675	0.21106	0.34657	-3.65199
5	30	50	19	0.38	-0.48955	-0.55820	0.5722	0.36396	0.29393	-1.66552
6	35	50	28	0.56	0.24116	0.20215	1.2240	0.55037	0.28429	0.84830
7	40	50	37	0.74	1.04597	0.96250	2.6182	0.72362	0.31623	3.30759
8	45	50	43	0.86	1.81529	1.72285	5.6004	0.84850	0.39444	4.60224
9	50	50	46	0.92	2.44235	2.48319	11.9795	0.92296	0.53034	4.60526
10	55	50	48	0.96	3.17805	3.24354	25.6243	0.96244	0.74382	4.27261
11	60	50	49	0.98	3.89182	4.00389	54.8110	0.98208	1.06611	3.65050
		===								
		550								

```
Obs  z1star  z2star  lstarh   lh_het  odds_het  ph_het buy correct

  1 1.13797  11.380 -4.02899 -3.54050   0.0290 0.02818  0      49
  2 1.61537  24.231 -4.49474 -2.78248   0.0619 0.05828  0      47
  3 2.22269  44.454 -4.49975 -2.02446   0.1321 0.11666  0      44
  4 2.88543  72.136 -3.65420 -1.26643   0.2818 0.21987  0      39
  5 3.40216 102.065 -1.72970 -0.50841   0.6014 0.37557  0      31
  6 3.51755 123.114  0.87802  0.24961   1.2835 0.56208  1      28
  7 3.16223 126.489  3.18636  1.00763   2.7391 0.73256  1      37
  8 2.53526 114.087  4.47640  1.76565   5.8454 0.85392  1      43
  9 1.88559  94.279  4.75861  2.52368  12.4744 0.92579  1      46
 10 1.34441  73.943  4.41196  3.28170  26.6209 0.96380  1      48
 11 0.93799  56.280  3.78922  4.03972  56.8104 0.98270  1      49
                                                            =======
                                                              461
```

Goodness of Fit: $\mathit{Count}\ R^2 = \dfrac{461}{550} = 0.8382$

SAS program:

```
options ls=72 ps=60 nodate pageno=1 ;

data a;
input x  n   r  ;
p1=r/n ;
l=log(p1/(1-p1)) ;
cards;
10   50   1
15   50   3
20   50   6
25   50  11
30   50  19
35   50  28
40   50  37
45   50  43
50   50  46
55   50  48
60   50  49
;

proc print data=a ;
run ;

proc plot data=a ;
plot (p1 l)*x ;
run ;

proc reg data=a ;
model l=x ;
output out=b p=lh ;
run ;


data b ;
 set b ;

odds = exp(lh) ;
ph = odds/(1+odds) ;
```

```
if ph <= 0 then delete ;
if ph >= 1 then delete ;

w = sqrt(1/(n*ph*(1-ph))) ;

lstar = l/w ;
z1star = 1/w ;
z2star = x/w ;


run ;


proc print data=b ;
run ;


proc reg data=b ;
model lstar = z1star z2star / noint ;
output out=c p=lstarh ;
run ;

data c ;
 set c ;
 lh_het = lstarh*w ;
 odds_het = exp(lh_het) ;
 ph_het = odds_het/(1+odds_het) ;

 buy=0 ;
 if ph_het  >= 0.5 then buy = 1 ;

 correct= 0 ;
 if ph_het < 0.5 then correct = n-r ;
 if ph_het >= 0.5 then correct = r ;


run ;

proc print data=c;
sum n correct ;
run ;
```

Example 2 (LOGIT grouped) - buying a product

Logit model (2) without adjusting for heteroscedasticity
Two explanatory variables

Obs	x1	x2	r	n	p	l
1	100	20	10	100	0.10	-2.19722
2	120	20	15	100	0.15	-1.73460
3	140	20	45	100	0.45	-0.20067
4	160	20	50	100	0.50	0.00000
5	100	25	15	100	0.15	-1.73460
6	120	25	20	100	0.20	-1.38629
7	140	25	45	100	0.45	-0.20067
8	160	25	55	100	0.55	0.20067
9	100	30	20	100	0.20	-1.38629
10	120	30	25	100	0.25	-1.09861
11	140	30	55	100	0.55	0.20067

12	160	30	60	100	0.60	0.40547
13	100	40	50	100	0.50	0.00000
14	120	40	55	100	0.55	0.20067
15	140	40	75	100	0.75	1.09861
16	160	40	80	100	0.80	1.38629

Logit model (2) without adjusting for heteroscedasticity
Two explanatory variables

The REG Procedure
Model: MODEL1
Dependent Variable: p

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.66629	0.33314	73.96	<.0001
Error	13	0.05856	0.00450		
Corrected Total	15	0.72484			

Root MSE	0.06712	R-Square	0.9192
Dependent Mean	0.42188	Adj R-Sq	0.9068
Coeff Var	15.90881		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.99545	0.11854	-8.40	<.0001
x1	1	0.00694	0.00075037	9.25	<.0001
x2	1	0.01793	0.00227	7.90	<.0001

Logit model (2) without adjusting for heteroscedasticity
Two explanatory variables

The REG Procedure
Model: MODEL2
Dependent Variable: l

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	15.56428	7.78214	82.33	<.0001
Error	13	1.22888	0.09453		
Corrected Total	15	16.79315			

Root MSE	0.30746	R-Square	0.9268
Dependent Mean	-0.40291	Adj R-Sq	0.9156
Coeff Var	-76.30843		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-7.25395	0.54304	-13.36	<.0001

x1	1	0.03356	0.00344	9.76	<.0001
x2	1	0.08655	0.01039	8.33	<.0001

Probit modeling:

Modeling a binary response variables using the CDF of the normal distribution.

Process

1. Estimate the initial probabilities / ratios
2. Convert / transform to associated standard normal *z-values*.
3. Fit a linear model between the *z-values* and the X variables
4. Use to model to estimate the *z-value* for applicable X variables
5. Use appropriate software to transform the estimated *z-values* to probabilities

Example 1 (LOGIT grouped)

Obs	X	T	n
1	6	40	8
2	8	50	12
3	10	60	18
4	13	80	28
5	15	100	45
6	20	70	36
7	25	65	39
8	30	50	33
9	35	40	30
10	40	25	20

Logit model

RESULTS

bh1	-1.65867
bh2	0.0791661
si2h	0.0216904
seb1	0.0957771
seb2	0.0041431
t1	-17.31802
t2	19.107789
prt1	1.2591E-7
prt2	5.8296E-8
r2	0.9785585
ar2	0.9758783
f	365.1076
prf	5.8296E-8

X		Y	YH	P	PH
1	6	-1.386294	-1.183674	0.2	0.2343922
1	8	-1.15268	-1.025342	0.24	0.2639882
1	10	-0.847298	-0.86701	0.3	0.2958769
1	13	-0.619039	-0.629511	0.35	0.3476213
1	15	-0.200671	-0.471179	0.45	0.3843372
1	20	0.0571584	-0.075349	0.5142857	0.4811717
1	25	0.4054651	0.3204814	0.6	0.5794416
1	30	0.6632942	0.7163118	0.66	0.6717943
1	35	1.0986123	1.1121422	0.75	0.7525283
1	40	1.3862944	1.5079726	0.8	0.8187605

```
                      Probit model

                        RESULTS

                   bh1   -1.015578
                   bh2   0.0484664
                   si2h  0.0079694
                   seb1  0.0580551
                   seb2  0.0025113
                   t1    -17.49335
                   t2    19.298953
                   prt1  1.1638E-7
                   prt2  5.3916E-8
                   r2    0.9789722
                   ar2   0.9763438
                   f     372.44957
                   prf   5.3916E-8


         X                   Y         YH          P        PH

         1         6 -0.841621   -0.72478        0.2 0.2342936
         1         8 -0.706303 -0.627847        0.24 0.2650521
         1        10 -0.524401 -0.530914        0.3 0.2977392
         1        13  -0.38532 -0.385515        0.35  0.349928
         1        15 -0.125661 -0.288582        0.45 0.3864506
         1        20 0.0358166  -0.04625 0.5142857 0.4815555
         1        25 0.2533471 0.1960819        0.6  0.577727
         1        30 0.4124631 0.4384138        0.66 0.6694568
         1        35 0.6744898 0.6807458        0.75 0.7519838
         1        40 0.8416212 0.9230778        0.8 0.8220167
```

SAS program:

```
options ls=72 nodate pageno=1  ;

data a ;
 input X T n ;
 cards;
  6  40  8
  8  50 12
 10  60 18
 13  80 28
 15 100 45
 20  70 36
 25  65 39
 30  50 33
 35  40 30
 40  25 20
  ;
  run;
proc print data = a ;
 run;

 proc iml ;
  use a ;
  read all into Xtn ;
```

```
  *print xtn ;

  n=nrow(xtn) ;

  p=xtn[,3]#recip(xtn[,2]) ;


*  model=2 ;

  do model=1 to 2 ;

  if model=1 then do ;
    y=log(p/(1-p));
  end;

  if model=2 then do ;
    y=probit(p) ;
  end;

  x = J(n,1,1) || xtn[,1] ;

  k=ncol(x) ;

  bh = inv(x`*x)*x`*y ;
  yh = x*bh ;
  uh = y - x*bh ;
  si2h = (uh`*uh)/(n-k) ;
  covb = si2h*inv(x`*x) ;

  sebv = sqrt(vecdiag(covb)) ;
  t = 1/sebv#bh ;

  prt = 2*(1 - probt(abs(t),n-k)) ;


  td = n*(sum(y)/n)**2 ;

  r2 = (bh`*x`*y-td) / (y`*y - td) ;
  ar2 = 1-(1-r2)*(n-1)/(n-k) ;

  F = (r2/(k-1))/((1-r2)/(n-k)) ;
  PrF = 1 - probf(f,k-1,n-k);


  *print bh si2h  sebv t prt r2 ar2 f prf ;

  nm = { "bh1" "bh2" "si2h" "seb1"  "seb2" "t1" "t2" "prt1" "prt2"
"r2" "ar2" "f" "prf"} ;

  results = (bh` || si2h || sebv` || t` ||  prt` || r2 || ar2 || f ||
prf)` ;



  if model = 1 then do ;
   Print "Logit model" ;
   print results[rowname=nm] ;
   ph= exp(yh)/(1+exp(yh)) ;
   print x y yh p ph ;
  end;
```

```
if model = 2 then do ;
 Print "Probit model" ;
 print results[rowname=nm] ;
 results = (bh` || si2h || sebv` || t` ||  prt` || r2 || ar2 ||
            f || prf)` ;
 ph = probnorm(yh) ;
 print x y yh p ph;
end;


end ;
```

Individual data:

Linear Probability modeling:

1) Model $Y_i = \beta_0 + \beta_1 X_i + u_i$ by the usual OLS estimate $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ with Y=0 or 1.

2) Use the valid $\hat{Y}$ values (between 0 and 1) to calculate $w_i = \sqrt{\hat{Y}_i(1-\hat{Y}_i)}$
3) Transform model as in 1. by dividing by w_i
4) New model is

$$Y_i / w_i = \beta_0 / w_i + \beta_1 X_i / w_i + u_i / w_i$$
$$Y_i^* = \beta_0 Z_1^* + \beta_1 Z_2^* + u_i^*, \quad where$$
$$Y_i^* = Y_i / w_i$$
$$Z_1^* = 1 / w_i$$
$$Z_2^* = X_i / w_i$$
$$u_i^* = u_i / w_i$$

5) Estimate the relevant parameters and transform back

In the textbook $w_i = \hat{Y}_i(1-\hat{Y}_i)$ and therefore they divide with $\sqrt{w_i}$

Example (LPM individual) - owning a home

```
                             The SAS System                            1

                           The REG Procedure
                             Model: MODEL1
                        Dependent Variable: y

              Number of Observations Read          37
              Number of Observations Used          37


                         Analysis of Variance

                                      Sum of          Mean
  Source                   DF        Squares        Square   F Value   Pr > F

  Model                     1        7.41340       7.41340    141.80   <.0001
  Error                    35        1.82985       0.05228
  Corrected Total          36        9.24324


          Root MSE              0.22865    R-Square     0.8020
          Dependent Mean        0.51351    Adj R-Sq     0.7964
          Coeff Var            44.52680


                         Parameter Estimates

                      Parameter       Standard
```

Variable	DF	Estimate	Error	t Value	Pr > |t|
Intercept	1	-0.97698	0.13069	-7.48	<.0001
X	1	0.10327	0.00867	11.91	<.0001

Obs	Family	y	X	yh	w	ystar	z1star	z2star
1	1	0	8	-0.15079	.	.	.	.
2	21	1	22	1.29505	.	.	.	.
3	2	1	16	0.67540	0.46822	2.13573	2.13573	34.172
4	22	1	16	0.67540	0.46822	2.13573	2.13573	34.172
5	3	1	18	0.88195	0.32267	3.09917	3.09917	55.785
6	23	0	12	0.26231	0.43989	0.00000	2.27330	27.280
7	4	0	11	0.15903	0.36571	0.00000	2.73443	30.079
8	24	0	11	0.15903	0.36571	0.00000	2.73443	30.079
9	5	0	12	0.26231	0.43989	0.00000	2.27330	27.280
10	25	1	16	0.67540	0.46822	2.13573	2.13573	34.172
11	6	1	19	0.98522	0.12065	8.28818	8.28818	157.475
12	26	0	11	0.15903	0.36571	0.00000	2.73443	30.079
13	7	1	20	1.08850	.	.	.	.
14	27	1	20	1.08850	.	.	.	.
15	8	0	13	0.36558	0.48159	0.00000	2.07644	26.994
16	28	1	18	0.88195	0.32267	3.09917	3.09917	55.785
17	9	0	9	-0.04752	.	.	.	.
18	29	0	11	0.15903	0.36571	0.00000	2.73443	30.079
19	10	0	10	0.05576	0.22945	0.00000	4.35815	43.582
20	30	0	10	0.05576	0.22945	0.00000	4.35815	43.582
21	11	1	17	0.77868	0.41514	2.40884	2.40884	40.950
22	31	1	17	0.77868	0.41514	2.40884	2.40884	40.950
23	12	1	18	0.88195	0.32267	3.09917	3.09917	55.785
24	32	0	13	0.36558	0.48159	0.00000	2.07644	26.994
25	13	0	14	0.46885	0.49903	0.00000	2.00389	28.054
26	33	1	21	1.19177	.	.	.	.
27	14	1	20	1.08850	.	.	.	.
28	34	1	20	1.08850	.	.	.	.
29	15	0	6	-0.35734	.	.	.	.
30	35	0	11	0.15903	0.36571	0.00000	2.73443	30.079
31	16	1	19	0.98522	0.12065	8.28818	8.28818	157.475
32	36	0	8	-0.15079	.	.	.	.
33	17	1	16	0.67540	0.46822	2.13573	2.13573	34.172
34	37	1	17	0.77868	0.41514	2.40884	2.40884	40.950
35	18	0	10	0.05576	0.22945	0.00000	4.35815	43.582
36	38	1	16	0.67540	0.46822	2.13573	2.13573	34.172
37	19	0	8	-0.15079	.	.	.	.

The REG Procedure
Model: MODEL1
Dependent Variable: ystar

Number of Observations Read	37
Number of Observations Used	26
Number of Observations with Missing Values	11

NOTE: No intercept in model. R-Square is redefined.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	200.61403	100.30701	414.87	<.0001
Error	24	5.80265	0.24178		
Uncorrected Total	26	206.41667			

Root MSE	0.49171	R-Square	0.9719
Dependent Mean	1.68381	Adj R-Sq	0.9695
Coeff Var	29.20213		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
z1star	1	-1.28382	0.11696	-10.98	<.0001
z2star	1	0.12286	0.00727	16.89	<.0001

Final estimated model: $\hat{Y} = -1.28382 + 0.12286X$

LOGIT modeling – MLE / not including heteroscedasticity.

Econometrics 720

1. Logistic Regression

The logistic regression model arises from the desire to model posterior probabilities of K classes via a linear functions in x, while at the same time ensuring that they sum to one and remain in the inerval $[0, 1]$.

A model that complies to the above is:

$$\begin{aligned}
log\frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} &= \beta_{10}+\beta_1^T x \\
log(odds_1) &= \beta_{10}+\beta_1^T x \quad \text{where} \quad odds_1 = \frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} \\
odds_1 &= e^{\beta_{10}+\beta_1^T x}
\end{aligned}$$

$$\begin{aligned}
log\frac{Pr(G=2|X=x)}{Pr(G=K|X=x)} &= \beta_{20}+\beta_2^T x \\
log(odds_2) &= \beta_{20}+\beta_2^T x \quad \text{where} \quad odds_2 = \frac{Pr(G=1|X=x)}{Pr(G=K|X=x)} \\
odds_2 &= e^{\beta_{20}+\beta_2^T x}
\end{aligned}$$

...

...

$$\begin{aligned}
log\frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} &= \beta_{(K-1)0}+\beta_{K-1}^T x \\
log(odds_{K-1}) &= \beta_{(K-1)0}+\beta_{K-1}^T x \quad \text{where} \quad odds_{K-1} = \frac{Pr(G=K-1|X=x)}{Pr(G=K|X=x)} \\
odds_{K-1} &= e^{\beta_{(K-1)0}+\beta_{K-1}^T x}
\end{aligned}$$

Consider the following:

$$\begin{aligned}
\Sigma_{l=1}^{K-1} log(odds_l) &= \Sigma_{l=1}^{K-1}\beta_{l0}+\beta_l^T x \\
\frac{1}{Pr(G=K|X=x)}\Sigma_{l=1}^{K-1}Pr(G=l|X=x) &= \Sigma_{l=1}^{K-1}e^{\beta_{l0}+\beta_l^T x} \\
\frac{1}{Pr(G=K|X=x)}(1-Pr(G=K|X=x)) &= \Sigma_{l=1}^{K-1}e^{\beta_{l0}+\beta_l^T x} \\
\frac{1}{Pr(G=K|X=x)} &= 1+\Sigma_{l=1}^{K-1}e^{\beta_{l0}+\beta_l^T x} \\
Pr(G=K|X=x) &= \frac{1}{1+\Sigma_{l=1}^{K-1}e^{\beta_{l0}+\beta_l^T x}}
\end{aligned}$$

$$\begin{aligned}
Pr(G=k|X=x) &= e^{\beta_{k0}+\beta_k^T x} pr(G=K|X=x) \\
Pr(G=k|X=x) &= e^{\beta_{k0}+\beta_k^T x}\frac{1}{1+\Sigma_{l=1}^{K-1}e^{\beta_{l0}+\beta_l^T x}}
\end{aligned}$$

Clearly $\Sigma_{l=1}^{K} Pr(G=l|X=x)$ is equal to 1, and all probabilities depend on the full parameter set $\theta = \{\beta_{10}, \beta_1, \beta_{20}, \beta_2, \ldots, \beta_{(K-1)0}, \beta_{K-1}\}$. These probabilities are denoted by $Pr(G=k|x=x) = p_k(x;\theta)$

Estimation of Logistic Regression models using Newton-Raphs

The Log-likelihood function

Consider only cases where $K = 2$.The log-likelihood function for N observations is:

$$l(\theta) = \Sigma_{i=1}^{N} log p_{g_i}(x_i;\theta)$$

where $p_{g_i}(x_i;\theta)$ is the probability of being in the group g that is associated with the i'th observation.

The log-likelihood function can be simplified by using the following code:

$$\begin{aligned}
y_i &= 1 \quad \text{when} \quad g_i = 1 \\
y_i &= 0 \quad \text{when} \quad g_i = 2 \\
p_1(x;\theta) &= p(x;\theta) \\
p_2(x;\theta) &= 1 - p(x;\theta) \quad \text{for two groups}
\end{aligned}$$

The log-likelihood function then is:

$$\begin{aligned}
l(\beta) &= \Sigma_{i=1}^{N} [y_i \, logp(x_i,\beta) + (1-y_i) log(1-p(x_i;\beta)] \\
&= \Sigma_{i=1}^{N} [y_i \, logp(x_i,\beta) + log(1-p(x_i;\beta)) - y_i \, log(1-p(x_i;\beta)] \\
&= \Sigma_{i=1}^{N} \left[y_i \, log \frac{p(x_i,\beta)}{(1-p(x_i;\beta)} + log(1-p(x_i;\beta) \right] \\
&= \Sigma_{i=1}^{N} \left[y_i \, \beta^T x_i + log(1 - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}}) \right] \\
& \quad \text{with} \quad \beta = \{\beta_{10}, \beta_1\} \quad \text{and} \quad x \quad \text{coded to include the intercept} \\
&= \Sigma_{i=1}^{N} \left[y_i \, \beta^T x_i + log(\frac{1}{1+e^{\beta^T x_i}}) \right] \\
&= \Sigma_{i=1}^{N} \left[y_i \, \beta^T x_i - log(1+e^{\beta^T x_i}) \right]
\end{aligned}$$

Maximum Likelihood Estimation

In order to maximise the log-likelihood function we set its derivatives equal to zero.

$$\frac{\partial l(\beta)}{\partial \beta} = \Sigma_{i=1}^{N} x_i(y_i - p(x_i;\beta)) = 0$$

which are $p+1$ score equations *non-linear* in β.

In order to solve the score equations , we use the Newton-Raphson algorithm, which requires the second derivative or Hessian matrix

$$\frac{\partial^2 l(\beta)}{\partial \beta \partial \beta^t} = -\Sigma_{i=1}^{N} x_i x_i^T p(x_i;\beta)(1-p(x_i;\beta))$$

The Newton-Raphson algorithm relates to the following update formula:

$$\beta^{new} = \beta^{old} - \left(\frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^t}\right)^{-1} \frac{\partial l(\beta)}{\partial\beta}$$

where the derivatives are evaluated at β^{old}.

In matrix notation we get:

$$\begin{aligned} \frac{\partial l(\beta)}{\partial\beta} &= X^T(y-p) \\ \frac{\partial^2 l(\beta)}{\partial\beta\partial\beta^t} &= -X^T W X \end{aligned}$$

with

y	the vector of y_i values
X	the $N \times (p+1)$ matrix of x_i values
p	the vector of fitted probabilities with i'th element $p(x_i;\beta^{old})$
W	a $N \times N$ diagonal matrix of weights with i'th diagonal element $p(x_i;\beta^{old})(1-p(x_i;\beta^{old}))$

The Newton-Raphson update formula is:

$$\begin{aligned} \beta^{new} &= \beta^{old} + (X^T W X)^{-1} X^T (y-p) \\ &= (X^T W X)^{-1}(X^T W X)\beta^{old} + (X^T W X)^{-1} X^T (y-p) \\ &= (X^T W X)^{-1} X^T W (X\beta^{old} + W^{-1}(y-p)) \\ &= (X^T W X)^{-1} X^T W z \end{aligned}$$

---

where

$$z = (X\beta^{old} + W^{-1}(y-p))$$

The vector z is sometimes called the adjusted response, and the equations are solved repeatedly.

Example (LOGIT individual)

GR8 Marketing has requested you to support them in developing a propensity to buy model for a new product that they are launching. They have specifically asked you to develop a statistical model that can be used

- as a scoring model of prospective clients
- to classify clients as likely buyers or not.

The following information is available for each of 3000 individuals in the sample:

Buy Indication	Age	Gender	Region
Yes	a: up to 19	Male	Northern
No	b: 20 to 29	Female	Southern
	c: 30 to 39		Western
	d: 40 +		Eastern

The following results are available:

```
                Table of buy by region
buy       region

Frequency|
Percent  |
Row Pct  |
Col Pct  |Eastern |Northern|Southern|Western |  Total
---------+--------+--------+--------+--------+
No       |    114 |     93 |    503 |    393 |   1103
         |   3.80 |   3.10 |  16.77 |  13.10 |  36.78
 ---------+--------+--------+--------+--------+
Yes      |    191 |    227 |    726 |    753 |   1897
         |   6.37 |   7.57 |  24.21 |  25.08 |  63.22
---------+--------+--------+--------+--------+
Total         305      320     1229     1146     3000
            10.17    10.67    40.98    38.18   100.00

        Statistics for Table of buy by region
Statistic                     DF       Value      Prob
------------------------------------------------------
Chi-Square                     3     20.3084    0.0001
Likelihood Ratio Chi-Square    3     20.4987    0.0001
Mantel-Haenszel Chi-Square     1      0.0806    0.7765
```

Table of buy by age

buy / age Frequency Percent Row Pct Col Pct	a: up to 19	b: 20 - 29	c: 30 - 39	d: 40 +	Total
No	283	263	291	266	1103
	9.43	8.77	9.70	8.87	36.77
Yes	453	460	425	559	1897
	15.10	15.33	14.17	18.63	63.23
Total	736	723	716	825	3000
	24.53	24.10	23.87	27.50	100.00

Statistics for Table of buy by age

Statistic	DF	Value	Prob
Chi-Square	3	12.8354	0.0050
Likelihood Ratio Chi-Square	3	12.9075	0.0048
Mantel-Haenszel Chi-Square	1	3.9103	0.0480

Table of buy by gender

buy / gender Frequency Percent Row Pct Col Pct	Female	Male	Total
No	745	358	1103
	24.83	11.93	36.77
Yes	1257	640	1897
	41.90	21.33	63.23
Total	2002	998	3000
	66.73	33.27	100.00

Statistics for Table of buy by gender

Statistic	DF	Value	Prob
Chi-Square	1	0.5152	0.4729
Likelihood Ratio Chi-Square	1	0.5161	0.4725
Continuity Adj. Chi-Square	1	0.4591	0.4980
Mantel-Haenszel Chi-Square	1	0.5150	0.4730

The CATMOD Procedure

Data Summary

Response	buy	Response Levels	2
Weight Variable	None	Populations	30
Data Set	A	Total Frequency	2999
Frequency Missing	1	Observations	2999

One-Way Frequencies

Variable	Value	Frequency
buy	No	1103
	Yes	1896
region	Eastern	305
	Northern	320
	Southern	1229
	Western	1145
age	a: up to 19	736
	b: 20 - 29	723
	c: 30 - 39	716
	d: 40 +	824
gender	Female	2002
	Male	997

Maximum Likelihood Analysis

Maximum likelihood computations converged.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	1	134.25	<.0001
region	3	14.66	0.0021
age	3	5.15	0.1611
gender	1	2.80	0.0945
age*gender	3	11.34	0.0100
Likelihood Ratio	19	18.37	0.4978

The CATMOD Procedure

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		-0.6073	0.0524	134.25	<.0001
region	Eastern	0.00367	0.1192	0.00	0.9754
	Northern	-0.1605	0.1168	1.89	0.1693
	Southern	0.2314	0.0702	10.87	0.0010
age	a: up to 19	0.0574	0.0844	0.46	0.4962
	b: 20 - 29	-0.0964	0.0739	1.70	0.1923
	c: 30 - 39	0.1291	0.0731	3.12	0.0775
gender	Female	0.0704	0.0421	2.80	0.0945
age*gender	a: up to 19 Female	0.1621	0.0694	5.45	0.0195
	b: 20 - 29 Female	0.0981	0.0717	1.87	0.1716
	c: 30 - 39 Female	-0.0799	0.0703	1.29	0.2550

The results above can be summarized as follows:

Parameter	Level	Estimate	Index
Intercept		-0.6073	0.54482
region	Eastern	0.00367	1.003677
	Northern	-0.1605	0.851718
	Southern	0.2314	1.260363
	Western	-0.07457	0.928143
age	a: up to 19	0.0574	1.059079
	b: 20 - 29	-0.0964	0.908101
	c: 30 - 39	0.1291	1.137804
	d: 40+	-0.0901	0.91384
gender	Female	0.0704	1.072937
	Male	-0.0704	0.932021
age*gender	a: up to 19 Female	0.1621	1.175978
	b: 20 - 29 Female	0.0981	1.103073
	c: 30 - 39 Female	-0.0799	0.923209
	d: 40+ Female	-0.1803	0.83502
	a: up to 19 Male	-0.1621	0.850356
	b: 20 - 29 Male	-0.0981	0.906558
	c: 30 - 39 Male	0.0799	1.083179
	d: 40+ Male	0.1803	1.197577

A few examples:

1:
Odds(Not Buy|Male) = 0.545*0.932 = 0.508
P(Not Buy|Male)=0.337
Classification: Buy

2:
Odds(Not Buy| age=22, Female) = 0.545*0.908*1.073*1.103=0.586
P(Not Buy| age=22, Female) = 0.369
Classification: Buy

Measuring goodness of fit:

Example:

EKT 720

The FREQ Procedure

Table of inc_c by lbuy

inc_c (Frequency / Percent / Row Pct / Col Pct)	lbuy: Buy	None	Total
a:	4889 2.44 15.12 9.88	27443 13.72 84.88 18.23	32332 16.17
b:	6998 3.50 23.83 14.14	22374 11.19 76.17 14.86	29372 14.69
c:	13197 6.60 33.49 26.67	26213 13.11 66.51 17.41	39410 19.71
d:	8436 4.22 35.19 17.05	15537 7.77 64.81 10.32	23973 11.99
e:	7495 3.75 25.11 15.15	22354 11.18 74.89 14.85	29849 14.92
f:	5058 2.53 19.30 10.22	21151 10.58 80.70 14.05	26209 13.10
g:	3405 1.70 18.06 6.88	15450 7.73 81.94 10.26	18855 9.43
Total	49478 24.74	150522 75.26	200000 100.00

EKT 720

The FREQ Procedure

Statistics for Table of inc_c by lbuy

Statistic	DF	Value	Prob
Chi-Square	6	5516.0401	<.0001
Likelihood Ratio Chi-Square	6	5540.2908	<.0001
Mantel-Haenszel Chi-Square	1	0.0751	0.7840
Phi Coefficient		0.1661	
Contingency Coefficient		0.1638	
Cramer's V		0.1661	

Sample Size = 200000

EKT 720

The FREQ Procedure

Table of age_c by lbuy

age_c / lbuy (Frequency, Percent, Row Pct, Col Pct)	Buy	None	Total
a: <= 25	1856 0.93 7.76 3.75	22071 11.04 92.24 14.66	23927 11.96
b: 26 - 35	15794 7.90 26.94 31.92	42825 21.41 73.06 28.45	58619 29.31
c: 36 - 45	17090 8.55 31.08 34.54	37902 18.95 68.92 25.18	54992 27.50
d: 46 - 59	11696 5.85 27.45 23.64	30916 15.46 72.55 20.54	42612 21.31
e: 60 +	3042 1.52 15.32 6.15	16808 8.40 84.68 11.17	19850 9.93
Total	49478 24.74	150522 75.26	200000 100.00

EKT 720

The FREQ Procedure

Statistics for Table of age_c by lbuy

Statistic	DF	Value	Prob
Chi-Square	4	6158.4205	<.0001
Likelihood Ratio Chi-Square	4	7166.0509	<.0001
Mantel-Haenszel Chi-Square	1	355.0063	<.0001
Phi Coefficient		0.1755	
Contingency Coefficient		0.1728	
Cramer's V		0.1755	

Sample Size = 200000

Table of Gender by lbuy

Gender / lbuy Frequency Percent Row Pct Col Pct	Buy	None	Total
Female	24707 12.35 27.26 49.94	65931 32.97 72.74 43.80	90638 45.32
Male	24771 12.39 22.65 50.06	84591 42.30 77.35 56.20	109362 54.68
Total	49478 24.74	150522 75.26	200000 100.00

EKT 720

The FREQ Procedure

Statistics for Table of Gender by lbuy

Statistic	DF	Value	Prob
Chi-Square	1	565.3509	<.0001
Likelihood Ratio Chi-Square	1	563.7879	<.0001
Continuity Adj. Chi-Square	1	565.1034	<.0001
Mantel-Haenszel Chi-Square	1	565.3481	<.0001
Phi Coefficient		0.0532	
Contingency Coefficient		0.0531	
Cramer's V		0.0532	

Fisher's Exact Test

Cell (1,1) Frequency (F)	24707
Left-sided Pr <= F	1.0000

```
 Right-sided Pr >= F     7.123E-125

 Table Probability (P)   1.995E-125
 Two-sided Pr <= P       1.302E-124

        Sample Size = 200000

              EKT 720

         The FREQ Procedure

      Table of sgroup by lbuy
```

sgroup   lbuy

Frequency / Percent / Row Pct / Col Pct	Buy	None	Total
1	4357 2.18 11.42 8.81	33799 16.90 88.58 22.45	38156 19.08
2	8583 4.29 26.11 17.35	24287 12.14 73.89 16.14	32870 16.44
3	18579 9.29 32.22 37.55	39076 19.54 67.78 25.96	57655 28.83
4	15185 7.59 26.16 30.69	42864 21.43 73.84 28.48	58049 29.02
5	2774 1.39 20.90 5.61	10496 5.25 79.10 6.97	13270 6.64
Total	49478 24.74	150522 75.26	200000 100.00

EKT 720

The FREQ Procedure

Statistics for Table of sgroup by lbuy

Statistic	DF	Value	Prob
Chi-Square	4	5572.0322	<.0001
Likelihood Ratio Chi-Square	4	6117.6639	<.0001
Mantel-Haenszel Chi-Square	1	1657.9477	<.0001
Phi Coefficient		0.1669	
Contingency Coefficient		0.1646	
Cramer's V		0.1669	

Sample Size = 200000

Table of pcheque by lbuy

pcheque Frequency Percent Row Pct Col Pct	lbuy Buy	 None	 Total
0	33705 16.85 23.73 68.12	108302 54.15 76.27 71.95	142007 71.00
1	15773 7.89 27.20 31.88	42220 21.11 72.80 28.05	57993 29.00
Total	49478 24.74	150522 75.26	200000 100.00

EKT 720

The FREQ Procedure

Statistics for Table of pcheque by lbuy

Statistic	DF	Value	Prob
Chi-Square	1	265.2770	<.0001
Likelihood Ratio Chi-Square	1	262.0319	<.0001
Continuity Adj. Chi-Square	1	265.0910	<.0001
Mantel-Haenszel Chi-Square	1	265.2757	<.0001
Phi Coefficient		-0.0364	
Contingency Coefficient		0.0364	
Cramer's V		-0.0364	

Fisher's Exact Test

Cell (1,1) Frequency (F)	33705
Left-sided Pr <= F	3.371E-59

Right-sided Pr >= F	1.0000
Table Probability (P)	5.647E-60
Two-sided Pr <= P	6.584E-59

Sample Size = 200000

EKT 720

The FREQ Procedure

Table of SIC_CDE by lbuy

SIC_CDE / lbuy Frequency Percent Row Pct Col Pct	Buy	None	Total
9110	41896 20.95 27.34 84.68	111371 55.69 72.66 73.99	153267 76.63
9120	2374 1.19 17.35 4.80	11311 5.66 82.65 7.51	13685 6.84
9130	5208 2.60 15.76 10.53	27840 13.92 84.24 18.50	33048 16.52
Total	49478 24.74	150522 75.26	200000 100.00

Statistics for Table of SIC_CDE by lbuy

Statistic	DF	Value	Prob
Chi-Square	2	2387.8447	<.0001
Likelihood Ratio Chi-Square	2	2555.6211	<.0001
Mantel-Haenszel Chi-Square	1	2272.6902	<.0001
Phi Coefficient		0.1093	
Contingency Coefficient		0.1086	
Cramer's V		0.1093	

Sample Size = 200000

EKT 720

The FREQ Procedure

Table of MRTL_STAT_CDE by lbuy

MRTL_STAT_CDE / lbuy Frequency Percent Row Pct Col Pct	Buy	None	Total
D	1632 0.82 23.31 3.30	5369 2.68 76.69 3.57	7001 3.50
M	25534 12.77 26.32 51.61	71489 35.74 73.68 47.49	97023 48.51
S	338 0.17 26.72 0.68	927 0.46 73.28 0.62	1265 0.63
U	20269 10.13 23.03 40.97	67750 33.88 76.97 45.01	88019 44.01
W	1705 0.85 25.48 3.45	4987 2.49 74.52 3.31	6692 3.35
Total	49478 24.74	150522 75.26	200000 100.00

EKT 720

The FREQ Procedure

Statistics for Table of MRTL_STAT_CDE by lbuy

Statistic	DF	Value	Prob
Chi-Square	4	280.5319	<.0001
Likelihood Ratio Chi-Square	4	281.0198	<.0001
Mantel-Haenszel Chi-Square	1	168.2952	<.0001
Phi Coefficient		0.0375	
Contingency Coefficient		0.0374	
Cramer's V		0.0375	

Sample Size = 200000

EKT 720

The FREQ Procedure

Table of avg_hh_sizec by lbuy

avg_hh_sizec / lbuy Frequency Percent Row Pct Col Pct	Buy	None	Total
A: =0 - 3	7915 3.96 21.64 16.00	28668 14.33 78.36 19.05	36583 18.29
B: 3 - 3.6	10193 5.10 24.75 20.60	30989 15.49 75.25 20.59	41182 20.59
C: 3.6 - 4.3	12512 6.26 29.51 25.29	29894 14.95 70.49 19.86	42406 21.20
D: 4.3 -	5540 2.77 29.93 11.20	12967 6.48 70.07 8.61	18507 9.25
Z: Not known	13318 6.66 21.72 26.92	48004 24.00 78.28 31.89	61322 30.66
Total	49478 24.74	150522 75.26	200000 100.00

EKT 720

The FREQ Procedure

Statistics for Table of avg_hh_sizec by lbuy

Statistic	DF	Value	Prob
Chi-Square	4	1275.5060	<.0001
Likelihood Ratio Chi-Square	4	1258.7840	<.0001
Mantel-Haenszel Chi-Square	1	2.7386	0.0980
Phi Coefficient		0.0799	
Contingency Coefficient		0.0796	
Cramer's V		0.0799	

Sample Size = 200000

EKT 720

The FREQ Procedure

Table of avg_hh_sizec by lbuy

avg_hh_sizec / lbuy (Frequency, Percent, Row Pct, Col Pct)	Buy	None	Total
A: =0 - 3	7915 3.96 21.64 16.00	28668 14.33 78.36 19.05	36583 18.29
B: 3 - 3.6	10193 5.10 24.75 20.60	30989 15.49 75.25 20.59	41182 20.59
C: 3.6 - 4.3	12512 6.26 29.51 25.29	29894 14.95 70.49 19.86	42406 21.20
D: 4.3 -	5540 2.77 29.93 11.20	12967 6.48 70.07 8.61	18507 9.25
Z: Not known	13318 6.66 21.72 26.92	48004 24.00 78.28 31.89	61322 30.66
Total	49478 24.74	150522 75.26	200000 100.00

The FREQ Procedure

Statistics for Table of avg_hh_sizec by lbuy

Statistic	DF	Value	Prob
Chi-Square	4	1275.5060	<.0001
Likelihood Ratio Chi-Square	4	1258.7840	<.0001
Mantel-Haenszel Chi-Square	1	2.7386	0.0980
Phi Coefficient		0.0799	
Contingency Coefficient		0.0796	
Cramer's V		0.0799	

Sample Size = 200000

EKT 720

The CATMOD Procedure

Data Summary

Response	lbuy	Response Levels	2
Weight Variable	None	Populations	8526
Data Set	MODEL	Total Frequency	200000
Frequency Missing	0	Observations	200000

Maximum Likelihood Analysis

Maximum likelihood computations converged.

Maximum Likelihood Analysis of Variance

Source	DF	Chi-Square	Pr > ChiSq
Intercept	1	4150.74	<.0001
inc_c	6	1967.79	<.0001
age_c	4	1070.27	<.0001
Gender	1	574.62	<.0001
sgroup	4	889.19	<.0001
pcheque	1	91.54	<.0001
SIC_CDE	2	370.77	<.0001
MRTL_STAT_CDE	4	59.15	<.0001
avg_hh_sizec	4	575.96	<.0001
inc_c*age_c	24	684.84	<.0001
sgroup*pcheque	4	22.05	0.0002
Likelihood Ratio	8E3	12353.22	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept		-1.3597	0.0211	4150.74	<.0001
inc_c	a:	0.0592	0.0263	5.05	0.0246
	b:	0.3815	0.0234	266.27	<.0001
	c:	0.7313	0.0190	1475.04	<.0001
	d:	0.4084	0.0217	353.30	<.0001
	e:	-0.2411	0.0236	104.24	<.0001
	f:	-0.6009	0.0286	441.29	<.0001
age_c	a: <= 25	-0.6516	0.0310	441.18	<.0001
	b: 26 - 35	0.0269	0.0149	3.26	0.0710
	c: 36 - 45	0.3570	0.0136	694.14	<.0001
	d: 46 - 59	0.3446	0.0139	614.13	<.0001
Gender	Female	0.1350	0.00563	574.62	<.0001
sgroup	1	-0.5658	0.0319	314.64	<.0001

EKT 720

The CATMOD Procedure

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
sgroup	2	-0.2799	0.0418	44.88	<.0001
	3	-0.2120	0.0189	125.37	<.0001
	4	0.4001	0.0199	404.91	<.0001
pcheque	0	-0.1152	0.0120	91.54	<.0001
SIC_CDE	9110	0.2017	0.0121	278.67	<.0001
	9120	-0.0773	0.0189	16.81	<.0001
MRTL_STAT_CDE	D	-0.1798	0.0277	41.96	<.0001
	M	0.0144	0.0170	0.72	0.3949
	S	0.0337	0.0538	0.39	0.5301
	U	0.000890	0.0178	0.00	0.9601
avg_hh_sizec	A: =0 - 3	-0.1490	0.0118	158.96	<.0001
	B: 3 - 3.6	-0.0178	0.0108	2.72	0.0990
	C: 3.6 - 4.3	0.1606	0.0103	241.48	<.0001
	D: 4.3 -	0.1419	0.0142	100.20	<.0001
inc_c*age_c	a: a: <= 25	-0.7370	0.0508	210.85	<.0001
	a: b: 26 - 35	0.1130	0.0296	14.60	0.0001
	a: c: 36 - 45	0.2827	0.0291	94.44	<.0001
	a: d: 46 - 59	0.3699	0.0295	157.23	<.0001
	b: a: <= 25	-0.3645	0.0538	45.93	<.0001
	b: b: 26 - 35	0.1051	0.0295	12.73	0.0004
	b: c: 36 - 45	0.2068	0.0273	57.54	<.0001
	b: d: 46 - 59	0.2625	0.0284	85.56	<.0001
	c: a: <= 25	0.1190	0.0447	7.09	0.0077
	c: b: 26 - 35	0.2101	0.0237	78.82	<.0001
	c: c: 36 - 45	0.0594	0.0233	6.51	0.0107
	c: d: 46 - 59	-0.1567	0.0234	44.83	<.0001
	d: a: <= 25	-0.0116	0.0555	0.04	0.8345
	d: b: 26 - 35	0.0501	0.0272	3.39	0.0656
	d: c: 36 - 45	0.1032	0.0267	14.91	0.0001
	d: d: 46 - 59	-0.0607	0.0299	4.10	0.0428
	e: a: <= 25	0.0977	0.0627	2.42	0.1195
	e: b: 26 - 35	-0.0626	0.0277	5.12	0.0237
	e: c: 36 - 45	-0.1873	0.0272	47.31	<.0001
	e: d: 46 - 59	-0.0310	0.0295	1.10	0.2940
	f: a: <= 25	0.3585	0.0829	18.69	<.0001
	f: b: 26 - 35	-0.2674	0.0348	58.91	<.0001
	f: c: 36 - 45	-0.2614	0.0323	65.58	<.0001
	f: d: 46 - 59	-0.1625	0.0341	22.75	<.0001
sgroup*pcheque	1 0	-0.0382	0.0221	3.00	0.0831
	2 0	0.0478	0.0385	1.54	0.2142
	3 0	-0.0426	0.0144	8.75	0.0031
	4 0	0.0197	0.0142	1.93	0.1650

EKT 720

The LOGISTIC Procedure

Model Information

Data Set	EKT.MODEL
Response Variable	lbuy
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	200000
Number of Observations Used	200000

Response Profile

Ordered Value	lbuy	Total Frequency
1	Buy	49478
2	None	150522

Probability modeled is lbuy='Buy'.

Class Level Information

Class	Value	Design Variables					
inc_c	a:	1	0	0	0	0	0
	b:	0	1	0	0	0	0
	c:	0	0	1	0	0	0
	d:	0	0	0	1	0	0
	e:	0	0	0	0	1	0
	f:	0	0	0	0	0	1
	g:	-1	-1	-1	-1	-1	-1
age_c	a: <= 25	1	0	0	0		
	b: 26 - 35	0	1	0	0		
	c: 36 - 45	0	0	1	0		
	d: 46 - 59	0	0	0	1		
	e: 60 +	-1	-1	-1	-1		
Gender	Female	1					
	Male	-1					
sgroup	1	1	0	0	0		
	2	0	1	0	0		
	3	0	0	1	0		
	4	0	0	0	1		
	5	-1	-1	-1	-1		

The LOGISTIC Procedure

Class Level Information

Class	Value	Design Variables			
pcheque	0	1			
	1	-1			
SIC_CDE	9110	1	0		
	9120	0	1		
	9130	-1	-1		
MRTL_STAT_CDE	D	1	0	0	0
	M	0	1	0	0
	S	0	0	1	0
	U	0	0	0	1
	W	-1	-1	-1	-1
avg_hh_sizec	A: =0 - 3	1	0	0	0
	B: 3 - 3.6	0	1	0	0
	C: 3.6 - 4.3	0	0	1	0
	D: 4.3 -	0	0	0	1
	Z: Not known	-1	-1	-1	-1

Model Convergence Status

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

Model Fit Statistics

Criterion	Intercept Only	Intercept and Covariates
AIC	223781.82	208083.25
SC	223792.03	208644.58
-2 Log L	223779.82	207973.25

Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	15806.5779	54	<.0001
Score	14417.2632	54	<.0001
Wald	12239.1820	54	<.0001

The LOGISTIC Procedure

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
inc_c	6	1967.7847	<.0001
age_c	4	1070.2675	<.0001
Gender	1	574.6176	<.0001
sgroup	4	889.1894	<.0001
pcheque	1	91.5400	<.0001
SIC_CDE	2	370.7700	<.0001
MRTL_STAT_CDE	4	59.1507	<.0001
avg_hh_sizec	4	575.9614	<.0001
inc_c*age_c	24	684.8263	<.0001
sgroup*pcheque	4	22.0530	0.0002

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square
Intercept		1	-1.3597	0.0211	4150.7387
inc_c	a:	1	0.0592	0.0263	5.0546
inc_c	b:	1	0.3815	0.0234	266.2663
inc_c	c:	1	0.7313	0.0190	1475.0379
inc_c	d:	1	0.4084	0.0217	353.3004
inc_c	e:	1	-0.2411	0.0236	104.2373
inc_c	f:	1	-0.6009	0.0286	441.2950
age_c	a: <= 25	1	-0.6516	0.0310	441.1816
age_c	b: 26 - 35	1	0.0269	0.0149	3.2599
age_c	c: 36 - 45	1	0.3570	0.0136	694.1355

Analysis of Maximum Likelihood Estimates

Parameter		Pr > ChiSq
Intercept		<.0001
inc_c	a:	0.0246
inc_c	b:	<.0001
inc_c	c:	<.0001
inc_c	d:	<.0001
inc_c	e:	<.0001
inc_c	f:	<.0001
age_c	a: <= 25	<.0001
age_c	b: 26 - 35	0.0710
age_c	c: 36 - 45	<.0001

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter			DF	Estimate	Standard Error	Wald Chi-Square
age_c	d: 46 - 59		1	0.3446	0.0139	614.1300
Gender	Female		1	0.1350	0.00563	574.6176
sgroup	1		1	-0.5658	0.0319	314.6386
sgroup	2		1	-0.2799	0.0418	44.8794
sgroup	3		1	-0.2120	0.0189	125.3668
sgroup	4		1	0.4001	0.0199	404.9060
pcheque	0		1	-0.1152	0.0120	91.5400
SIC_CDE	9110		1	0.2017	0.0121	278.6742
SIC_CDE	9120		1	-0.0773	0.0189	16.8075
MRTL_STAT_CDE	D		1	-0.1798	0.0277	41.9603
MRTL_STAT_CDE	M		1	0.0144	0.0170	0.7239
MRTL_STAT_CDE	S		1	0.0337	0.0538	0.3942
MRTL_STAT_CDE	U		1	0.000890	0.0178	0.0025
avg_hh_sizec	A: =0 - 3		1	-0.1490	0.0118	158.9632
avg_hh_sizec	B: 3 - 3.6		1	-0.0178	0.0108	2.7223
avg_hh_sizec	C: 3.6 - 4.3		1	0.1606	0.0103	241.4793
avg_hh_sizec	D: 4.3 -		1	0.1419	0.0142	100.1974
inc_c*age_c	a:	a: <= 25	1	-0.7370	0.0508	210.8399
inc_c*age_c	a:	b: 26 - 35	1	0.1130	0.0296	14.5985
inc_c*age_c	a:	c: 36 - 45	1	0.2827	0.0291	94.4417
inc_c*age_c	a:	d: 46 - 59	1	0.3699	0.0295	157.2290
inc_c*age_c	b:	a: <= 25	1	-0.3645	0.0538	45.9252
inc_c*age_c	b:	b: 26 - 35	1	0.1051	0.0295	12.7327
inc_c*age_c	b:	c: 36 - 45	1	0.2068	0.0273	57.5358
inc_c*age_c	b:	d: 46 - 59	1	0.2625	0.0284	85.5618
inc_c*age_c	c:	a: <= 25	1	0.1190	0.0447	7.0943
inc_c*age_c	c:	b: 26 - 35	1	0.2101	0.0237	78.8166
inc_c*age_c	c:	c: 36 - 45	1	0.0594	0.0233	6.5129
inc_c*age_c	c:	d: 46 - 59	1	-0.1567	0.0234	44.8310
inc_c*age_c	d:	a: <= 25	1	-0.0116	0.0555	0.0436
inc_c*age_c	d:	b: 26 - 35	1	0.0501	0.0272	3.3900
inc_c*age_c	d:	c: 36 - 45	1	0.1032	0.0267	14.9096
inc_c*age_c	d:	d: 46 - 59	1	-0.0607	0.0299	4.1037
inc_c*age_c	e:	a: <= 25	1	0.0977	0.0627	2.4234
inc_c*age_c	e:	b: 26 - 35	1	-0.0626	0.0277	5.1154
inc_c*age_c	e:	c: 36 - 45	1	-0.1873	0.0272	47.3130
inc_c*age_c	e:	d: 46 - 59	1	-0.0310	0.0295	1.1014
inc_c*age_c	f:	a: <= 25	1	0.3585	0.0829	18.6856
inc_c*age_c	f:	b: 26 - 35	1	-0.2674	0.0348	58.9099
inc_c*age_c	f:	c: 36 - 45	1	-0.2614	0.0323	65.5758
inc_c*age_c	f:	d: 46 - 59	1	-0.1625	0.0341	22.7515
sgroup*pcheque	1	0	1	-0.0382	0.0221	3.0032
sgroup*pcheque	2	0	1	0.0478	0.0385	1.5426
sgroup*pcheque	3	0	1	-0.0426	0.0144	8.7529
sgroup*pcheque	4	0	1	0.0197	0.0142	1.9274

EKT 720

The LOGISTIC Procedure

Odds Ratio Estimates

Effect		Point Estimate
Gender	Female vs Male	1.310
SIC_CDE	9110 vs 9130	1.386
SIC_CDE	9120 vs 9130	1.048
MRTL_STAT_CDE	D vs W	0.733
MRTL_STAT_CDE	M vs W	0.890
MRTL_STAT_CDE	S vs W	0.908
MRTL_STAT_CDE	U vs W	0.878
avg_hh_sizec	A: =0 - 3 vs Z: Not known	0.987
avg_hh_sizec	B: 3 - 3.6 vs Z: Not known	1.125
avg_hh_sizec	C: 3.6 - 4.3 vs Z: Not known	1.345
avg_hh_sizec	D: 4.3 - vs Z: Not known	1.320

Odds Ratio Estimates

95% Wald Confidence Limits	
1.281	1.339
1.337	1.436
0.987	1.114
0.675	0.796
0.837	0.947
0.786	1.048
0.824	0.936
0.955	1.020
1.091	1.160
1.305	1.385
1.270	1.372

Association of Predicted Probabilities and Observed Responses

Percent Concordant	67.7	Somers' D	0.359
Percent Discordant	31.8	Gamma	0.360
Percent Tied	0.5	Tau-a	0.134
Pairs	7447527516	c	0.679

EKT 720

The LOGISTIC Procedure

Partition for the Hosmer and Lemeshow Test

		lbuy = Buy		lbuy = None	
Group	Total	Observed	Expected	Observed	Expected
1	19992	1000	971.58	18992	19020.42
2	20022	2605	2602.17	17417	17419.83
3	20001	3305	3319.07	16696	16681.93
4	19998	3829	3880.73	16169	16117.27
5	20043	4403	4475.95	15640	15567.05
6	20023	5051	5139.67	14972	14883.33
7	20020	5923	5862.38	14097	14157.62
8	20051	6777	6641.38	13274	13409.62
9	19996	7638	7497.30	12358	12498.70
10	19854	8947	9087.78	10907	10766.22

Hosmer and Lemeshow Goodness-of-Fit Test

Chi-Square	DF	Pr > ChiSq
18.6666	8	0.0167

EKT 720

The FREQ Procedure

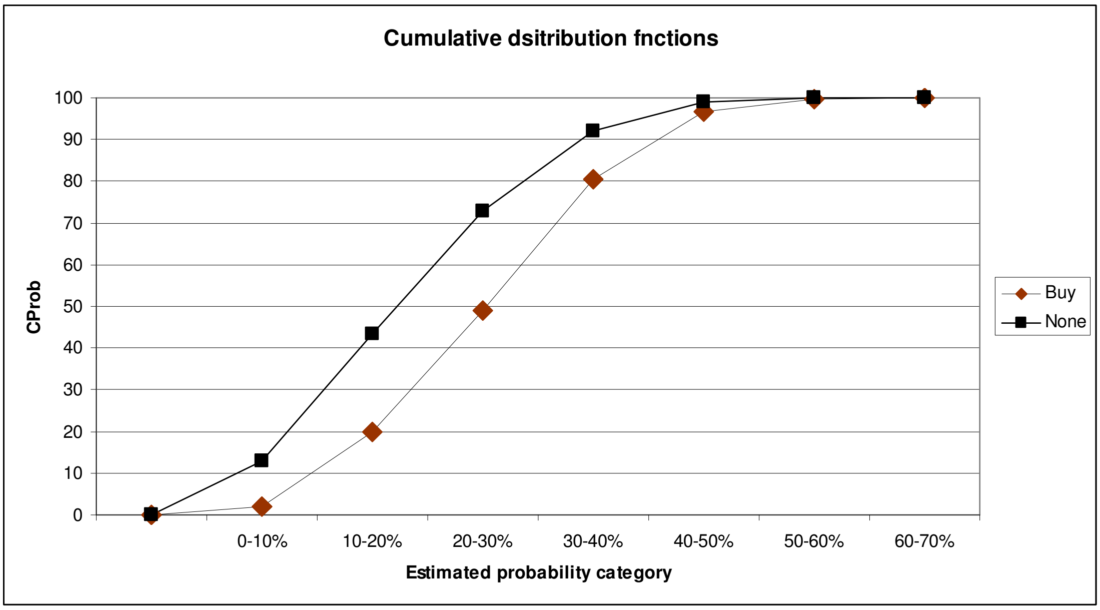
pr	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0-10%	20472	10.24	20472	10.24
10-20%	54674	27.34	75146	37.57
20-30%	58898	29.45	134044	67.02
30-40%	44378	22.19	178422	89.21
40-50%	18206	9.10	196628	98.31
50-60%	3140	1.57	199768	99.88
60-70%	232	0.12	200000	100.00

EKT 720
The FREQ Procedure

Table of pr by lbuy

pr / Frequency / Percent / Row Pct / Col Pct	lbuy: Buy	None	Total
0-10%	1052 0.53 5.14 2.13	19420 9.71 94.86 12.90	20472 10.24
10-20%	8707 4.35 15.93 17.60	45967 22.98 84.07 30.54	54674 27.34
20-30%	14561 7.28 24.72 29.43	44337 22.17 75.28 29.46	58898 29.45
30-40%	15443 7.72 34.80 31.21	28935 14.47 65.20 19.22	44378 22.19
40-50%	8009 4.00 43.99 16.19	10197 5.10 56.01 6.77	18206 9.10
50-60%	1592 0.80 50.70 3.22	1548 0.77 49.30 1.03	3140 1.57
60-70%	114 0.06 49.14 0.23	118 0.06 50.86 0.08	232 0.12
Total	49478 24.74	150522 75.26	200000 100.00

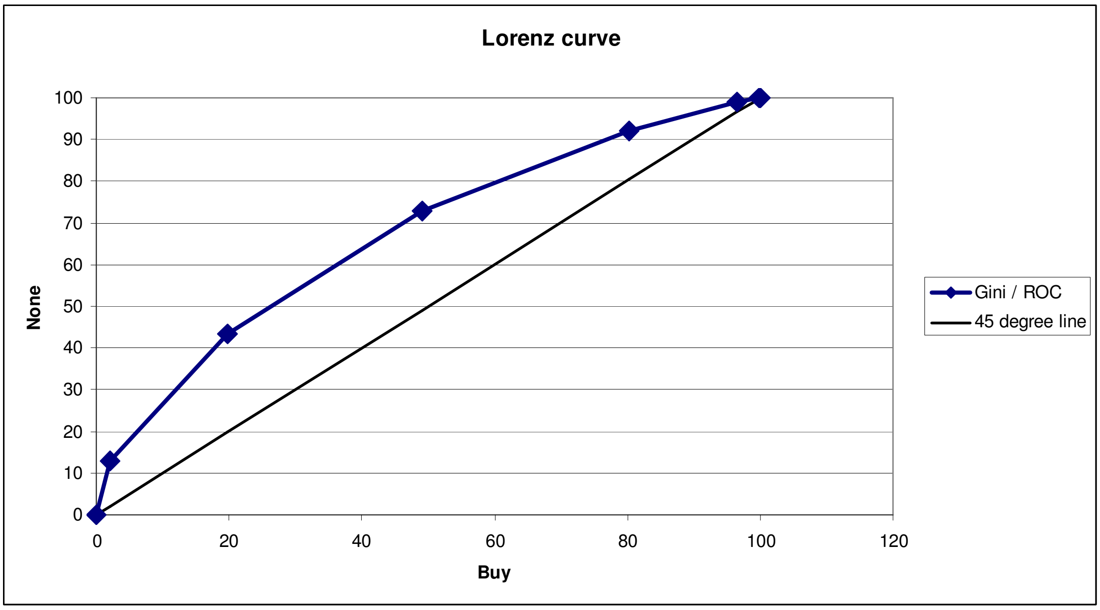
K-S values



K-S value is the maximum difference between the two CDF functions.

Critical values?

Gini Index values



Gini index is the area between the 45 degree line and the Lorenz curve (one definition).

Hosmer and Lemshow:
Source SAS Help

Sufficient replication within subpopulations is required to make the Pearson and deviance goodness-of-fit tests valid. When there are one or more continuous predictors in the model, the data are often too sparse to use these statistics. Hosmer and Lemeshow (2000) proposed a statistic that they show, through simulation, is distributed as chi-square when there is no replication in any of the subpopulations. This test is only available for binary response models.

First, the observations are sorted in increasing order of their estimated event probability. The event is the response level specified in the response variable option EVENT=, or the response level which is not specified in the REF= option, or, if neither of these options were specified, then the event is the response level identified in the "Response Profiles" table as "Ordered Value 1". The observations are then divided into approximately ten groups according to the following scheme. Let N be the total number of subjects. Let M be the target number of subjects for each group given by

$$M = [0.1 \times N + 0.5]$$

where $[x]$ represents the integral value of x. If the *single-trial* syntax is used, blocks of subjects are formed of observations with identical values of the explanatory variables. Blocks of subjects are not divided when being placed into groups.

Suppose there are n_1 subjects in the first block and n_2 subjects in the second block. The first block of subjects is placed in the first group. Subjects in the second block are added to the first group if

$$n_1 < M \quad \text{and} \quad n_1 + [0.5 \times n_2] \le M$$

Otherwise, they are placed in the second group. In general, suppose subjects of the (j-1)th block have been placed in the kth group. Let c be the total number of subjects currently in the kth group. Subjects for the jth block (containing n_j subjects) are also placed in the kth group if

$$c < M \quad \text{and} \quad c + [0.5 \times n_j] \le M$$

Otherwise, the n_j subjects are put into the next group. In addition, if the number of subjects in the last group does not exceed $[0.05 \times N]$ (half the target group size), the last two groups are collapsed to form only one group.

Note that the number of groups, g, may be smaller than 10 if there are fewer than 10 patterns of explanatory variables. There must be at least three groups in order for the Hosmer-Lemeshow statistic to be computed.

The Hosmer-Lemeshow goodness-of-fit statistic is obtained by calculating the Pearson chi-square statistic from the $2 \times g$ table of observed and expected frequencies, where g is the number of groups. The statistic is written

$$\chi^2_{HL} = \sum_{i=1}^{g} \frac{(O_i - N_i\bar{\pi}_i)^2}{N_i\bar{\pi}_i(1-\bar{\pi}_i)}$$

where N_i is the total frequency of subjects in the ith group, O_i is the total frequency of event outcomes in the ith group, and $\bar{\pi}_i$ is the average estimated predicted probability of an event outcome for the ith group. The Hosmer-Lemeshow statistic is then compared to a chi-square distribution with $(g-n)$ degrees of freedom, where the value of n can be specified in the LACKFIT option in the MODEL statement. The default is n=2. Large values of χ^2_{HL} (and small p-values) indicate a lack of fit of the model.

SAS program / GOF example

```
options ls=72 nodate pageno=1 ;


libname  ekt  "c:\departement\ekt720\logitlpm" ;



title "EKT 720" ;

proc freq data=ekt.model ;
 tables (inc_c age_c gender sgroup pcheque sic_cde mrtl_stat_cde
avg_hh_sizec avg_hh_sizec)*lbuy  / chisq;
run;


proc catmod data=ekt.model ;
 model lbuy = inc_c age_c gender sgroup pcheque sic_cde mrtl_stat_cde
avg_hh_sizec   inc_c*age_c  sgroup*pcheque / noprofile;
 response out=ekt.resp ;
run ;

quit ;


proc logistic data=ekt.model ;
 class  inc_c age_c gender sgroup pcheque sic_cde mrtl_stat_cde
avg_hh_sizec  ;
 model lbuy = inc_c age_c gender sgroup pcheque sic_cde mrtl_stat_cde
avg_hh_sizec   inc_c*age_c  sgroup*pcheque / lackfit;
run ;




data ekt.resp1(drop=lbuy) ; ;
 set ekt.resp ;

 *product=substr(product,1,9) ;

 if lbuy="Buy" ;

 if     _pred_ <= 0.1 then  pr="0-10%    " ;
 if 0.1 <_pred_ <= 0.2 then pr="10-20%   " ;
 if 0.2 <_pred_ <= 0.3 then pr="20-30%   " ;
 if 0.3 <_pred_ <= 0.4 then pr="30-40%   " ;
 if 0.4 <_pred_ <= 0.5 then pr="40-50%   " ;
 if 0.5 <_pred_ <= 0.6 then pr="50-60%   " ;
 if 0.6 <_pred_ <= 0.7 then pr="60-70%   " ;
 if 0.7 <_pred_ <= 0.8 then pr="70-80%   " ;
 if 0.8 <_pred_ <= 0.9 then pr="80-90%   " ;
 if _pred_ > 0.9 then pr="90+  %   " ;
run ;


proc sort data=ekt.model ;
 by inc_c age_c gender sgroup pcheque sic_cde mrtl_stat_cde
avg_hh_sizec;
run ;
```

```
proc sort data=ekt.resp1 ;
 by inc_c age_c gender sgroup pcheque sic_cde mrtl_stat_cde
avg_hh_sizec;
run ;


data ekt.saam ;
 merge ekt.model ekt.resp1 ;
 by inc_c age_c gender sgroup pcheque sic_cde mrtl_stat_cde
avg_hh_sizec;
run ;



proc freq data=ekt.saam ;
 tables pr pr*lbuy ;
run ;
```