```
* MSBA 635 - Data Analytics II;

* print data;
*logit/probit model
*Logit and probit models are appropriate when attempting to model a dichotomous dependent
variable, e.g. yes/no, agree/disagree, like/dislike etc.
*issues with the traditional approach when counting # of beers consumed: 1. The
regression line may lead to predictions outside the range of zero and one 2. The
functional form assumes that beer has the same marginal effect on each beer consumed 3. A
residuals plot would quickly reveal heteroscedasticity. Logit and probit models solve
each of these problems by fitting a nonlinear function to the data. It 1; respects the
boundaries of the dependent variable 2. Allows for different rates of change at the high
and low ends 3. Does away with heteroscedasticity
*The logit model uses the cumulative distribution function of the logistic distribution
*the probit model uses the cumulative distribution function of the standard normal
distribution. Logit and probit take any number and rescale it to fall between 0 and 1
*this is consumer packaged goods
*in store advertising can be a feature or display. Disp_coke=coke display
*what they are saying is 1 if pepsi on display in week of product purchased
*p-ratio is 1 if they are in parity. If greater than 1 price of coke is greater than
price of pepsi, if under 1 then price of coke is less than pepsi
*p-ratio=1 means same price

proc print data=tmp1.coke (obs=10);
run;
```


Obs	coke	pr_pepsi	pr_coke	disp_ pepsi	disp_ coke	pratio
1	1	1.79	1.79	0	0	1.00000
2	1	1.79	0.89	0	1	0.49721
3	1	1.41	0.89	0	0	0.63121
4	1	1.79	1.33	0	0	0.74302
5	1	1.79	1.79	0	0	1.00000
6	1	0.99	1.79	1	0	1.80808
7	1	0.77	1.79	1	0	2.32468
8	1	1.33	1.79	1	0	1.34587
9	1	1.79	0.99	0	0	0.55307
10	1	1.79	1.29	0	0	0.72067

```
* display data attributes;

proc contents data=tmp1.coke;
run;
```


The CONTENTS Procedure

Data Set Name	TMP1.COKE	Observations	1140
Member Type	DATA	Variables	6
Engine	V9	Indexes	0
Created	06/20/2010 11:33:33	Observation Length	48
Last Modified	06/20/2010 11:33:33	Deleted Observations	0
Protection		Compressed	NO
Data Set Type		Sorted	NO
Label			
Data Representation	WINDOWS_32		
Encoding	wlatin1 Western (Windows)		

Engine/Host Dependent Information

Data Set Page Size	4096
Number of Data Set Pages	15
First Data Page	1
Max Obs per Page	84
Obs in First Data Page	46
Number of Data Set Repairs	0
Filename	C:\Users\nxnguy01\Desktop\coke.sas7bdat
Release Created	9.0202M0
Host Created	XP_PRO

Alphabetic List of Variables and Attributes

#	Variable	Type	Len	Label
1	coke	Num	8	=1 if coke chosen, =0 if pepsi chosen
5	disp_coke	Num	8	= 1 if coke is displayed at time of purchase, otherwise = 0
4	disp_pepsi	Num	8	= 1 if pepsi is displayed at time of purchase, otherwise = 0
3	pr_coke	Num	8	price of 2 liter bottle of coke
2	pr_pepsi	Num	8	price of 2 liter bottle of pepsi
6	pratio	Num	8	price coke relative to price pepsi

```
* produce frequencies;

proc freq data=tmp1.coke;
tables coke;
run;
```


The FREQ Procedure

=1 if coke chosen, =0 if pepsi chosen

coke	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	630	55.26	630	55.26
1	510	44.74	1140	100.00

```
* estimate linear probability model using proc reg;
*this is where you show why you don't want to run a proc reg based on 0s and 1s.
*you'll see that here there are negative probabilities, and greater than 1 which is not
allowed
*if dependent variable is an indicator variable then can't use proc reg
*lpm=Linear Probability Model
options nolabel;
proc reg data=tmp1.coke;
model coke = pratio disp_coke disp_pepsi;
output out=lpmout p=phat_lpm;
run;
quit;
```


The REG Procedure
Model: MODEL1
Dependent Variable: coke

Number of Observations Read	1140
Number of Observations Used	1140

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	33.83781	11.27927	51.67	<.0001
Error	1136	248.00430	0.21831		
Corrected Total	1139	281.84211			

Root MSE	0.46724	R-Square	0.1201
Dependent Mean	0.44737	Adj R-Sq	0.1177
Coeff Var	104.44198		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.89022	0.06548	13.59	<.0001
pratio	1	-0.40086	0.06135	-6.53	<.0001
disp_coke	1	0.07717	0.03439	2.24	0.0250
disp_pepsi	1	-0.16566	0.03560	-4.65	<.0001

```
* print data;

proc print data=work.lpmout (obs=10);
run;
```


Obs	coke	pr_pepsi	pr_coke	disp_ pepsi	disp_ coke	pratio	phat_lpm
1	1	1.79	1.79	0	0	1.00000	0.48935
2	1	1.79	0.89	0	1	0.49721	0.76808
3	1	1.41	0.89	0	0	0.63121	0.63719
4	1	1.79	1.33	0	0	0.74302	0.59237
5	1	1.79	1.79	0	0	1.00000	0.48935
6	1	0.99	1.79	1	0	1.80808	**-0.00024**
7	1	0.77	1.79	1	0	2.32468	**-0.20732**
8	1	1.33	1.79	1	0	1.34587	0.18505
9	1	1.79	0.99	0	0	0.55307	0.66851
10	1	1.79	1.29	0	0	0.72067	0.60133

```
* estimate probit with proc qlim;
*the model is non-linear so you can't do an F-test here
*this is an equivalent of the f-test but for a non-linear model
*its saying the display variables that the first test, that the coefficient on cokes
display=the negative of the coefficient on pepsi's display so we will test this
hypothesis
*if pepsi isn't displaying is it effective at taking away sales from coke?
*coefficients are equal in magnitude but opposite in sign
*is a rival merchandising event-does it take as much business away as when coke does a
merchandising event, is rival merchandising as impacting as my own merchandising?
*if we end up at the same market position then we have wasted a lot of money
*take the coefficient on the coke display and add it to the pepsi display, are those two
marketing effects offset or equal?
*next test says each display of coke and of pepsi equal zero. They have no influence over
people choosing coke. Should you be spending marketing dollars on displays?
*

options nolabel;
proc qlim data=tmp1.coke;
model coke = pratio disp_coke disp_pepsi / discrete;
output out=probitout marginal;
test 'H0: 1' disp_coke+disp_pepsi=0 / wald;
test 'H0: 2' disp_coke=0, disp_pepsi=0 / wald;
run;
quit;
```


The QLIM Procedure

Discrete Response Profile of coke

Index	Value	Total Frequency
1	0	630
2	1	510

Model Fit Summary

Number of Endogenous Variables	1
Endogenous Variable	coke
Number of Observations	1140
Log Likelihood	-710.94858
Maximum Absolute Gradient	4.33302E-6
Number of Iterations	9
Optimization Method	Quasi-Newton
AIC	1430
Schwarz Criterion	1450

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	145.82	2 * (LogL - LogL0)
Upper Bound of R (U)	1567.7	- 2 * LogL0
Aldrich-Nelson	0.1134	R / (R+N)
Cragg-Uhler 1	0.1201	1 - exp(-R/N)
Cragg-Uhler 2	0.1607	(1-exp(-R/N)) / (1-exp(-U/N))

Estrella	0.1256	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.1189	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.093	R / U
Veall-Zimmermann	0.1959	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.1895	

N = # of observations, K = # of regressors

*the higher the price of coke relative to pepsi the less likely you will be to choose coke because of the pratio (-1.145963). the probability of choosing coke is going down
*if coke is on display in the week of purchase, you have a higher probability of choosing coke (0.217187), cokes demand output will shift to the right
*if display for pepsi is out, then the demand curve for coke shifts to the left (-0.447297) lowers the probability that you will choose coke
*the magnitudes are 0.22 and 0.45. they are opposite and equal in magnitude
*look at the **wald test H0:1** here, reject the null hypothesis because they are not of the same magnitude. Your rival merchandising event is more effective at taking business away from you then you are at getting your own business
*__Wald test H0:2__ reject the null. Both the disp_coke and disp_pepsi have an effect on whether or not a customer purchases the coke
*magnitudes = estimates
*merchandising at all doesn't matter at the market place is rejected. Why would they be insignificant jointly? That wouldn't make sense

Algorithm converged.

The QLIM Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	1.108060	0.189959	5.83	<.0001
pratio	1	**-1.145963**	0.180883	-6.34	<.0001
disp_coke	1	**0.217187**	0.096608	2.25	0.0246
disp_pepsi	1	**-0.447297**	0.101403	-4.41	<.0001

Test Results

Test	Type	Statistic	Pr > ChiSq	Label
'H0: 1'	Wald	5.40	0.0201	disp_coke + disp_pepsi = 0
'H0: 2'	Wald	19.46	<.0001	disp_coke = 0 , disp_pepsi = 0

```
* print probitout data;

proc print data=work.probitout (obs=10);
run;
```


Obs	coke	pr_pepsi	pr_coke	disp_ pepsi	disp_ coke	Meff_P1_ pratio	Meff_P2_ pratio	Meff_P1_ disp_coke	Meff_P2_ disp_ coke	Meff_P1_ disp_ pepsi	Meff_P2_ disp_ pepsi
1	1	1.79	1.79	0	0	1.00000	0.45684	-0.45684	-0.086583	0.086583	0.17832	-0.17832
2	1	1.79	0.89	0	1	0.49721	0.34368	-0.34368	-0.065135	0.065135	0.13414	-0.13414
3	1	1.41	0.89	0	0	0.63121	0.42456	-0.42456	-0.080464	0.080464	0.16572	-0.16572
4	1	1.79	1.33	0	0	0.74302	0.44237	-0.44237	-0.083839	0.083839	0.17267	-0.17267
5	1	1.79	1.79	0	0	1.00000	0.45684	-0.45684	-0.086583	0.086583	0.17832	-0.17832
6	1	0.99	1.79	1	0	1.80808	0.16889	-0.16889	-0.032010	0.032010	0.06592	-0.06592
7	1	0.77	1.79	1	0	2.32468	0.06147	-0.06147	-0.011651	0.011651	0.02399	-0.02399
8	1	1.33	1.79	1	0	1.34587	0.30998	-0.30998	-0.058748	0.058748	0.12099	-0.12099
9	1	1.79	0.99	0	0	0.55307	0.40854	-0.40854	-0.077429	0.077429	0.15946	-0.15946
10	1	1.79	1.29	0	0	0.72067	0.43933	-0.43933	-0.083263	0.083263	0.17148	-0.17148

```
* produce means;
*if you have a 1 unit increase, if coke was $1 more for a 2-liter then a pepsi 2-liter
then the probability of you choosing coke will fall by 40.1%
*if coke were $1 cheaper than a 2-liter then there would be a 40.1% increase
*if you put a merchandising event into the market place then the probability of choosing
coke increases by roughly 8%
*marginally- take every observation in the data set and calculate the phat value
*if you put a pepsi merchandising event in the week of purchase you will decrease the
probability of buying coke by almost 16%

options nolabel;
proc means data=work.probitout;
var meff_p1_pratio meff_p2_pratio meff_p1_disp_coke meff_p2_disp_coke meff_p1_disp_pepsi
meff_p2_disp_pepsi;
run;
```


The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Meff_P1_pratio	1140	**0.4096951**	0.0667242	0.0614731	0.4569719
Meff_P2_pratio	1140	**-0.4096951**	0.0667242	-0.4569719	-0.0614731
Meff_P1_disp_coke	1140	-0.0776469	0.0126458	-0.0866070	-0.0116506
Meff_P2_disp_coke	1140	**0.0776469**	0.0126458	0.0116506	0.0866070
Meff_P1_disp_pepsi	1140	0.1599139	0.0260441	0.0239945	0.1783672
Meff_P2_disp_pepsi	1140	**-0.1599139**	0.0260441	-0.1783672	-0.0239945

```
* estimate logit using proc qlim;


options nolabel;
proc qlim data=tmp1.coke;
model coke = pratio disp_coke disp_pepsi / discrete (d=logit);
output out=logitout marginal;
test 'H0: 1' disp_coke+disp_pepsi=0 / wald;
test 'H0: 2' disp_coke=0, disp_pepsi=0 / wald;
run;
quit;
```


The QLIM Procedure

Discrete Response Profile of coke

Index	Value	Total Frequency
1	0	630
2	1	510

Model Fit Summary

Number of Endogenous Variables	1
Endogenous Variable	coke
Number of Observations	1140
Log Likelihood	-709.44614
Maximum Absolute Gradient	8.16552E-7
Number of Iterations	9
Optimization Method	Quasi-Newton
AIC	1427
Schwarz Criterion	1447

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	148.83	2 * (LogL - LogL0)
Upper Bound of R (U)	1567.7	- 2 * LogL0
Aldrich-Nelson	0.1155	R / (R+N)
Cragg-Uhler 1	0.1224	1 - exp(-R/N)
Cragg-Uhler 2	0.1638	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.1282	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.1214	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.0949	R / U
Veall-Zimmermann	0.1994	(R * (U+N)) / (U * (R+N))
McKelvey-Zavoina	0.4036	

N = # of observations, K = # of regressors

Algorithm converged.

The QLIM Procedure

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
Intercept	1	1.922972	0.325832	5.90	<.0001
pratio	1	-1.995742	0.314586	-6.34	<.0001
disp_coke	1	0.351599	0.158540	2.22	0.0266
disp_pepsi	1	-0.730986	0.167837	-4.36	<.0001

Test Results

Test	Type	Statistic	Pr > ChiSq	Label
'H0: 1'	Wald	5.61	0.0179	disp_coke + disp_pepsi = 0
'H0: 2'	Wald	18.97	<.0001	disp_coke = 0 , disp_pepsi = 0

```
* print logitout data;

proc print data=work.logitout (obs=10);
run;
```


Obs	coke	pr_pepsi	pr_coke	disp_ pepsi	disp_ coke	pratio	Meff_P1_ pratio	Meff_P2_ pratio	Meff_P1_ disp_coke	Meff_P2_ disp_ coke	Meff_P1_ disp_ pepsi	Meff_P2_ disp_ pepsi
1	1	1.79	1.79	0	0	1.00000	0.49828	-0.49828	-0.087784	0.087784	0.18250	-0.18250
2	1	1.79	0.89	0	1	0.49721	0.33928	-0.33928	-0.059773	0.059773	0.12427	-0.12427
3	1	1.41	0.89	0	0	0.63121	0.44785	-0.44785	-0.078900	0.078900	0.16404	-0.16404
4	1	1.79	1.33	0	0	0.74302	0.47553	-0.47553	-0.083777	0.083777	0.17418	-0.17418
5	1	1.79	1.79	0	0	1.00000	0.49828	-0.49828	-0.087784	0.087784	0.18250	-0.18250
6	1	0.99	1.79	1	0	1.80808	0.15011	-0.15011	-0.026445	0.026445	0.05498	-0.05498
7	1	0.77	1.79	1	0	2.32468	0.05966	-0.05966	-0.010510	0.010510	0.02185	-0.02185
8	1	1.33	1.79	1	0	1.34587	0.29879	-0.29879	-0.052639	0.052639	0.10944	-0.10944
9	1	1.79	0.99	0	0	0.55307	0.42378	-0.42378	-0.074659	0.074659	0.15522	-0.15522
10	1	1.79	1.29	0	0	0.72067	0.47074	-0.47074	-0.082933	0.082933	0.17242	-0.17242

```
* produce means;

options nolabel;
proc means data=work.logitout;
var meff_p1_pratio meff_p2_pratio meff_p1_disp_coke meff_p2_disp_coke meff_p1_disp_pepsi
meff_p2_disp_pepsi;
run;
```


The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Meff_P1_pratio	1140	0.4332631	0.0854584	0.0596590	0.4988652
Meff_P2_pratio	1140	-0.4332631	0.0854584	-0.4988652	-0.0596590
Meff_P1_disp_coke	1140	-0.0763300	0.0150556	-0.0878875	-0.0105104
Meff_P2_disp_coke	1140	0.0763300	0.0150556	0.0105104	0.0878875
Meff_P1_disp_pepsi	1140	0.1586925	0.0313011	0.0218515	0.1827207
Meff_P2_disp_pepsi	1140	-0.1586925	0.0313011	-0.1827207	-0.0218515