Predicted Probability from Logistic Regression Output[1]

It is possible to use the output from Logistic regression, and means of variables, to calculate the predicted probability of different subgroups in your analysis falling into a category. The following example will use a subset of 1980 IPUMS data to demonstrate how to do this. http://www.ipums.umn.edu/usa/index.html

I used SAS create the following output (which I have abbreviated). (Please note, I am using the subset of cases and variables that I am using only because it was convenient to do so for this document. I am not suggesting that the model is properly specified).

The LOGISTIC Procedure

Model Information

Data Set	WORK.F506	
Response Variable	highsei	SEI in top 50%
Number of Response Levels	2	
Number of Observations	537896	

Ordered Value	highsei	Total Frequency
1	1	257852
2	0	280044

Probability modeled is highsei=1.

The LOGISTIC Procedure

Analysis of Maximum Likelihood Estimates

Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-5.6842	0.0257	49010.9112	<.0001
AGE	1	0.0141	0.000347	1661.4577	<.0001
female	1	0.2873	0.00626	2107.5317	<.0001
black	1	-0.8269	0.0111	5513.2191	<.0001
higrader	1	0.3959	0.00139	81129.4148	<.0001

The MEANS Procedure

Variable	Label	N	Mean
highsei	SEI in top 50%	537896	0.4793715
AGE		537896	43.4276533
female		537896	0.4419553
black		537896	0.1049422
higrader	higrade recoded (higradeg - 3)	537896	12.3926410

[1]Prepared by Patty Glynn , University of Washington, 03/05/03.

I put this information into a spreadsheet. By manipulating the values in the spreadsheet, I was able to calculate the predicted values for white men, black men, white women, and black women. Column A has the variable names. Column B has the coefficients from the regression equation. Column C has the means for the variables, EXCEPT, 1 is put in the the Intercept, and zeros are put in for the variables that are to be manipulated. Columns D through G are the product of the values in B and C. For example, the values in d8, e8, f8 and g8 are =$B8*$C8

The values for D9 through G9 and D10 through G10 are manipulated depending on the group for whom predicted values are bing calculated. For white men, the cells D9 and D10 are left at 0. For black men, the coefficient for black is entered into the cell E10, but the cell for e9 is left at 0. For black women, the coefficients for both female, and black are entered into the appropriate cells. Next the predicted probabilities must be calculated. Row 14 is the sum of the values in the columns. In row 15, the formula =EXP(-D14) is entered for column D, and =EXP(-E14) for column E, and so on. Finally, in row 16, the formula =1/(1+D15) is entered for column E, and =1/(1+E15) for column D, and so on. This row is the predicted probability of the group being in the high SEI.

	A	B	C	D	E	F	G	H	I	J
1	Predval.xls									
2										
3	Dichotomous Variable, High SEI									
4										
5				White	Black	White	Black			
6		Coeff	Means	Men	Men	Women	Women			
7	Intercept	-5.684	1	-5.684	-5.684	-5.684	-5.684			
8	AGE	0.0141	43.428	0.612	0.612	0.612	0.612		White Men	0.459
9	female	0.2873	0	0.000	0.000	0.287	0.287		Black Men	0.270
10	black	-0.827	0	0.000	-0.827	0.000	-0.827		White Women	0.530
11	higrader	0.3959	12.393	4.906	4.906	4.906	4.906		Black Women	0.331
12										
13										
14	sum of column			-0.166	-0.993	0.122	-0.705			
15	=EXP(-D14)			1.180	2.698	0.885	2.024			
16	=1/(1+D15), Predicted Probability			0.459	0.270	0.530	0.331			
17										

It is then possible to graph the predicted values for each group.

You can find the sample spreadsheet used for this document at:

http://staff.washington.edu/glynn/predprob.xls