Predicted Values from Regression Output[1]

It is possible to use the output from OLS regression, and means of variables, to calculate the predicted values of your dependent variable for different subgroups. 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. (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 REG Procedure
Model: MODEL1
Dependent Variable: SEI

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	91438221	22859555	59072.1	<.0001
Error	537891	208151596	386.97728		
Corrected Total	537895	299589817			

Root MSE	19.67174	R-Square	0.3052
Dependent Mean	42.07443	Adj R-Sq	0.3052
Coeff Var	46.75462		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-11.78439	0.18634	-63.24	<.0001
AGE	1	0.13094	0.00298	43.88	<.0001
female	1	1.61912	0.05410	29.93	<.0001
black	1	-7.59608	0.08833	-86.00	<.0001
higrader	1	3.89376	0.00842	462.41	<.0001

The MEANS Procedure

Variable	N	Mean
SEI	537896	42.0744289
AGE	537896	43.4276533
female	537896	0.4419553
black	537896	0.1049422
higrader	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. The bottom row (Predicted Values) is the sum of the values in the columns. For example, the formula in D13, (the predicted value for white men), is =SUM(D7:D11)

	A	B	C	D	E	F	G	H	I	J
1										
2										
3	**Continous Variable, SEI**									
4										
5				White	Black	White	Black		White Men	42.156
6		Coeff	Means	Men	Men	Women	Women		Black Men	34.560
7	Intercept	-11.78	1	-11.784	-11.784	-11.784	-11.784		White Women	43.775
8	AGE	0.1309	43.428	5.686	5.686	5.686	5.686		Black Women	36.179
9	female	1.6191	0	0.000	0.000	1.619	1.619			
10	black	-7.596	0	0.000	-7.596	0.000	-7.596			
11	higrader	3.8938	12.393	48.254	48.254	48.254	48.254			
12										
13	Predicted Values			42.156	34.560	43.775	36.179			

For this example model, the predicted values of SEI are 42.156 for white men, 34.560 for black men, 43.775 for white women, and 36.179 for black women. These are different from the actual mean values of SEI for each group. The predicted values assume that the slopes and means for age and education are the same for these four groups. The actual mean values for SEI are 43.25 for white men, 28.84 for black men, 43.57 for white women, and 32.82 for black women.

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/predval.xls