(Multiple) Regression in SAS: REG vs. GLM

- There are two major procedures in SAS that fit regression models
  - **proc reg** (reg is short for 'regression')
  - **proc glm** (glm stands for 'general linear model')
- Some of the pros and cons of each are:
  - REG will allow you to fit several different models in a single **proc** statement; GLM will not
  - GLM will allow you to directly include special categorical predictor variables (using a class statement) and interactions in the model statement; to include such terms using REG, you first need to create new variables in the **data** step.
- Note: the **`clb cli clm`** options on the **model** statement still work for multiple regression, in both REG and GLM (see the notes on Simple Linear Regression in SAS).

Example SAS code using the BAC data:

```
DATA bac;
input BAC weight sex $ beers;
datalines;
0.12  192   male  8
0.1   132   female      5
0.03  128   female      2
0.19  110   female      9
0.04  172   male  3
0.095 250   female      7
0.07  125   female      3
0.06  175   male  5
0.02  175   female      3
0.05  275   male  5
0.07  130   female      4
0.1   168   male  6
0.085 128   female      5
0.09  246   male  7
0.01  164   male  1
0.05  175   male  4
;
run;

DATA bac;
     set bac; * this tells SAS to copy the bac dataset we just created;

/* recode sex from 'female/male' to '0/1' */
if sex='male' then sex_r=1;
      else if sex='female'  then sex_r=0;

/* quadratic terms for beers and weight */
beers2 = beers*beers;
weight2 = weight*weight;

/* create interaction terms */
int_sb = beers*sex_r;
int_wb = beers*weight;

run;
```

```
/*****************************************************************/
/* Use PROC REG to create models                                 */
/* Note: Can use multiple MODEL statements                       */
/* Note: Need to use recoded variables from 2nd DATA step above */
/*****************************************************************/


/* Basic BAC Models from the notes */
PROC reg data=bac;
      model bac = beers;
      model bac = beers weight;
      model bac = beers sex_r; * note inclusion of recoded variable;
      model bac = beers sex_r weight;
      model bac = beers weight beers2 weight2 int_wb;
run;



/**************************************************************/
/* Use PROC GLM to create models                              */
/* Note: Only one MODEL statement per proc                    */
/* Note: Do not need to use recoded variables from DATA step */
/**************************************************************/

PROC glm data=bac;
  model bac = beers;
run;

PROC glm data=bac;
  class sex;   * this tells SAS that sex is a categorical variable;
  model bac = beers sex / solution;
/* the 'solution' option requests that the parameter estimates be displayed, these are
hidden by default when you include a categorical predictor */
run;

PROC glm data=bac;
      model bac = beers weight beers*beers weight*weight weight*beers; * the last 3 terms
all represent interactions;
run;
quit;  * GLM will tend to run endlessly if you do not include a ‘quit’ statement;



/**********************/
/* Too much output??? */
/**********************/

ODS HTML CLOSE; *this stops SAS from writing any more output to the Results Viewer
window;
ODS HTML; *this opens a new Results Viewer window where new output will be displayed;
```

Output Produced by this Code (Graphics Not Included):

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: BAC

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.02338	0.02338	55.94	<.0001
Error	14	0.00585	0.00041783		
Corrected Total	15	0.02922			

Root MSE	0.02044	R-Square	0.7998
Dependent Mean	0.07375	Adj R-Sq	0.7855
Coeff Var	27.71654		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.01270	0.01264	-1.00	0.3320
beers	1	0.01796	0.00240	7.48	<.0001

The REG Procedure
Model: MODEL2
Dependent Variable: BAC

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.02782	0.01391	128.33	<.0001
Error	13	0.00141	0.00010838		
Corrected Total	15	0.02922			

Root MSE	0.01041	R-Square	0.9518
Dependent Mean	0.07375	Adj R-Sq	0.9444
Coeff Var	14.11574		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.03986	0.01043	3.82	0.0021
beers	1	0.01998	0.00126	15.82	<.0001
weight	1	-0.00036282	0.00005668	-6.40	<.0001

The REG Procedure
Model: MODEL3
Dependent Variable: BAC

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.02494	0.01247	37.79	<.0001
Error	13	0.00429	0.00032991		
Corrected Total	15	0.02922			

Root MSE	0.01816	R-Square	0.8532
Dependent Mean	0.07375	Adj R-Sq	0.8307
Coeff Var	24.62821		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-0.00348	0.01200	-0.29	0.7767
beers	1	0.01810	0.00214	8.48	<.0001
sex_r	1	-0.01976	0.00909	-2.18	0.0487

The REG Procedure
Model: MODEL4
Dependent Variable: BAC

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	0.02785	0.00928	80.81	<.0001
Error	12	0.00138	0.00011486		
Corrected Total	15	0.02922			

Root MSE	0.01072	R-Square	0.9528
Dependent Mean	0.07375	Adj R-Sq	0.9410
Coeff Var	14.53212		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.03871	0.01097	3.53	0.0042
beers	1	0.01990	0.00131	15.20	<.0001
sex_r	1	-0.00324	0.00629	-0.52	0.6156
weight	1	-0.00034440	0.00006842	-5.03	0.0003

The REG Procedure
Model: MODEL5
Dependent Variable: BAC

Number of Observations Read	16
Number of Observations Used	16

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.02843	0.00569	71.53	<.0001
Error	10	0.00079491	0.00007949		
Corrected Total	15	0.02922			

Root MSE	0.00892	R-Square	0.9728
Dependent Mean	0.07375	Adj R-Sq	0.9592
Coeff Var	12.08920		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	0.13855	0.04909	2.82	0.0181
beers	1	0.01364	0.00724	1.88	0.0892
weight	1	-0.00132	0.00044525	-2.98	0.0139
beers2	1	0.00051929	0.00045388	1.14	0.2792
weight2	1	0.00000255	0.00000103	2.48	0.0324
int_wb	1	0.00000203	0.00003050	0.07	0.9483

The SAS System

The GLM Procedure

Number of Observations Read 16

Number of Observations Used 16

The GLM Procedure

Dependent Variable: BAC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	0.02337535	0.02337535	55.94	<.0001
Error	14	0.00584965	0.00041783		
Corrected Total	15	0.02922500			

R-Square	Coeff Var	Root MSE	BAC Mean
0.799841	27.71654	0.020441	0.073750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
beers	1	0.02337535	0.02337535	55.94	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
beers	1	0.02337535	0.02337535	55.94	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	-.0127006040	0.01263750	-1.00	0.3320
beers	0.0179637619	0.00240170	7.48	<.0001

Note: 'Type I' vs. 'Type III' Sums of Squares

- **Type I SS = sequential (order dependent)**
  - **Accounts for new term in model after all <u>previous</u> terms have been accounted for**
- **Type III SS = overall (not order dependent)**
  - **Accounts for new term in model after all other terms have been accounted for**
- **Both printed by default with PROC GLM**

The GLM Procedure

Class	Levels	Values
sex	2	female male

Number of Observations Read 16

Number of Observations Used 16

The GLM Procedure

Dependent Variable: BAC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	0.02493623	0.01246811	37.79	<.0001
Error	13	0.00428877	0.00032991		
Corrected Total	15	0.02922500			

R-Square	Coeff Var	Root MSE	BAC Mean
0.853250	24.62821	0.018163	0.073750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
beers	1	0.02337535	0.02337535	70.85	<.0001
sex	1	0.00156088	0.00156088	4.73	0.0487

Source	DF	Type III SS	Mean Square	F Value	Pr > F
beers	1	0.02371123	0.02371123	71.87	<.0001
sex	1	0.00156088	0.00156088	4.73	0.0487

Recall from class: for a categorical predictor, you need 1 fewer variables than there are categories. That is basically what GLM is telling you here.

Also note: 'male' is used as the baseline category.

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-.0232383420	B	0.01222983	-1.90	0.0798
beers	0.0181001727		0.00213501	8.48	<.0001
sex female	0.0197625216	B	0.00908557	2.18	0.0487
sex male	0.0000000000	B	.	.	.

This table would not be printed without the solutions option, since sex is categorical.

Note: The X'X matrix has been found to be singular, and a generalized inverse was used to solve the normal equations. Terms whose estimates are followed by the letter 'B' are not uniquely estimable.

The GLM Procedure

Dependent Variable: BAC

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	0.02843009	0.00568602	71.53	<.0001
Error	10	0.00079491	0.00007949		
Corrected Total	15	0.02922500			

R-Square	Coeff Var	Root MSE	BAC Mean
0.972800	12.08920	0.008916	0.073750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
beers	1	0.02337535	0.02337535	294.06	<.0001
weight	1	0.00444077	0.00444077	55.86	<.0001
beers*beers	1	0.00012247	0.00012247	1.54	0.2428
weight*weight	1	0.00049115	0.00049115	6.18	0.0322
beers*weight	1	0.00000035	0.00000035	0.00	0.9483

Source	DF	Type III SS	Mean Square	F Value	Pr > F
beers	1	0.00028170	0.00028170	3.54	0.0892
weight	1	0.00070396	0.00070396	8.86	0.0139
beers*beers	1	0.00010405	0.00010405	1.31	0.2792
weight*weight	1	0.00048975	0.00048975	6.16	0.0324
beers*weight	1	0.00000035	0.00000035	0.00	0.9483

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.1385530174	0.04909102	2.82	0.0181
beers	0.0136361629	0.00724370	1.88	0.0892
weight	-.0013249957	0.00044525	-2.98	0.0139
beers*beers	0.0005192901	0.00045388	1.14	0.2792
weight*weight	0.0000025500	0.00000103	2.48	0.0324
beers*weight	0.0000020291	0.00003050	0.07	0.9483