Building Multiple Linear Regression (MLR) Models - Food for Thought

Vancouver SAS Users Group Meeting – May 2009

Colleen McGahan
Biostatistician
BC Cancer Agency
cmcgahan@bccancer.bc.ca

Building Multiple Linear Regression (MLR) Models - Food for Thought

Does not cover:

- Model assumptions
- Assessing the adequacy of the model
- Considerations when the model does not fit well
- Statistical theory

Assumes knowledge of MLR

Step 1: Get to know your data!

- Compute descriptive statistics
- Do cross tabulations
- Identify potential interactions
- Use graphical displays of the data
- Identify outliers, extreme values, missing obsv^{n}s
- Look at the distribution of the variables

Step 1: Get to know your data!

- Identify the number of missing values.
- Understand the relationship/correlation between:
  (a) the response variable and the independent variables
  (b) the independent variables
- Consider the number of variables you have in relation to the number of observations.
- Know what the primary objective of the modeling is.

Step 1: Get to know your data!

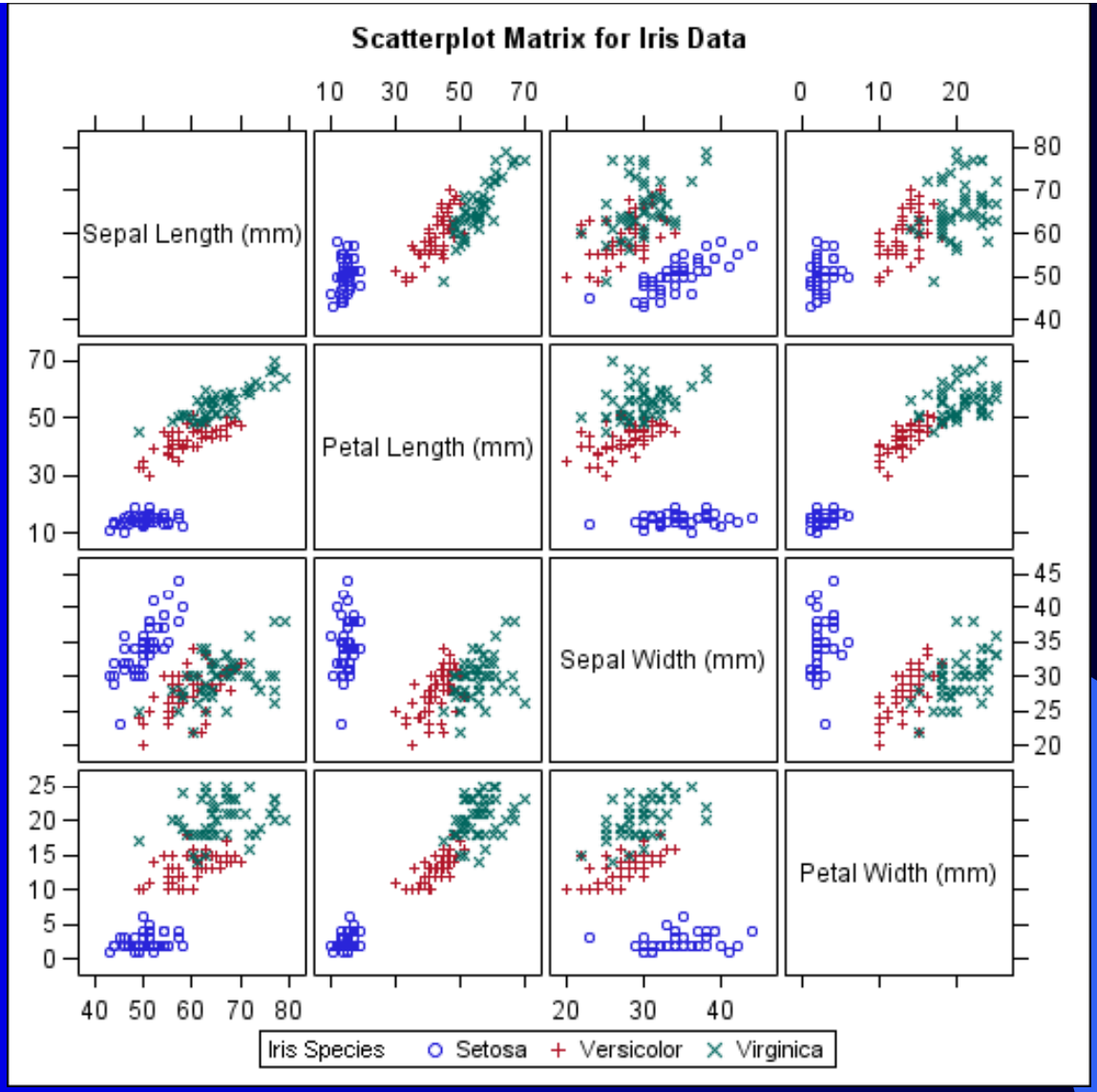
- Talk to the experts of the topic
- Use prior knowledge to help determine variables to choose

Step 1: Get to know your data!

Some SAS procedures to help:

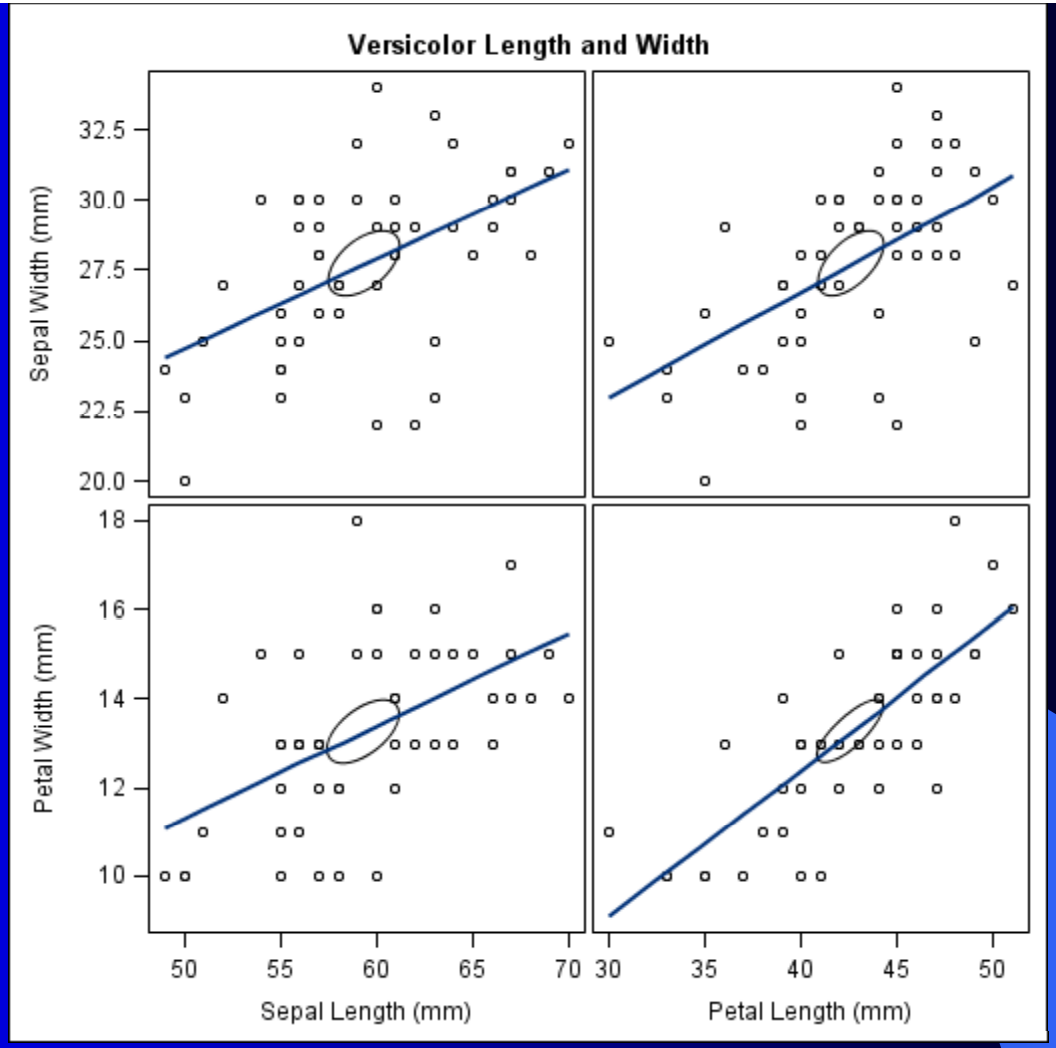
- UNIVARIATE: for continuous variables, has the CLASS statement
- FREQ: for categorical variables
- CORR: produces correlations
- SGSCATTER: paneled scatter plots, histograms
- SGPLOT: statistical graphics eg. Histograms, box plots

PROC SGSCATTER



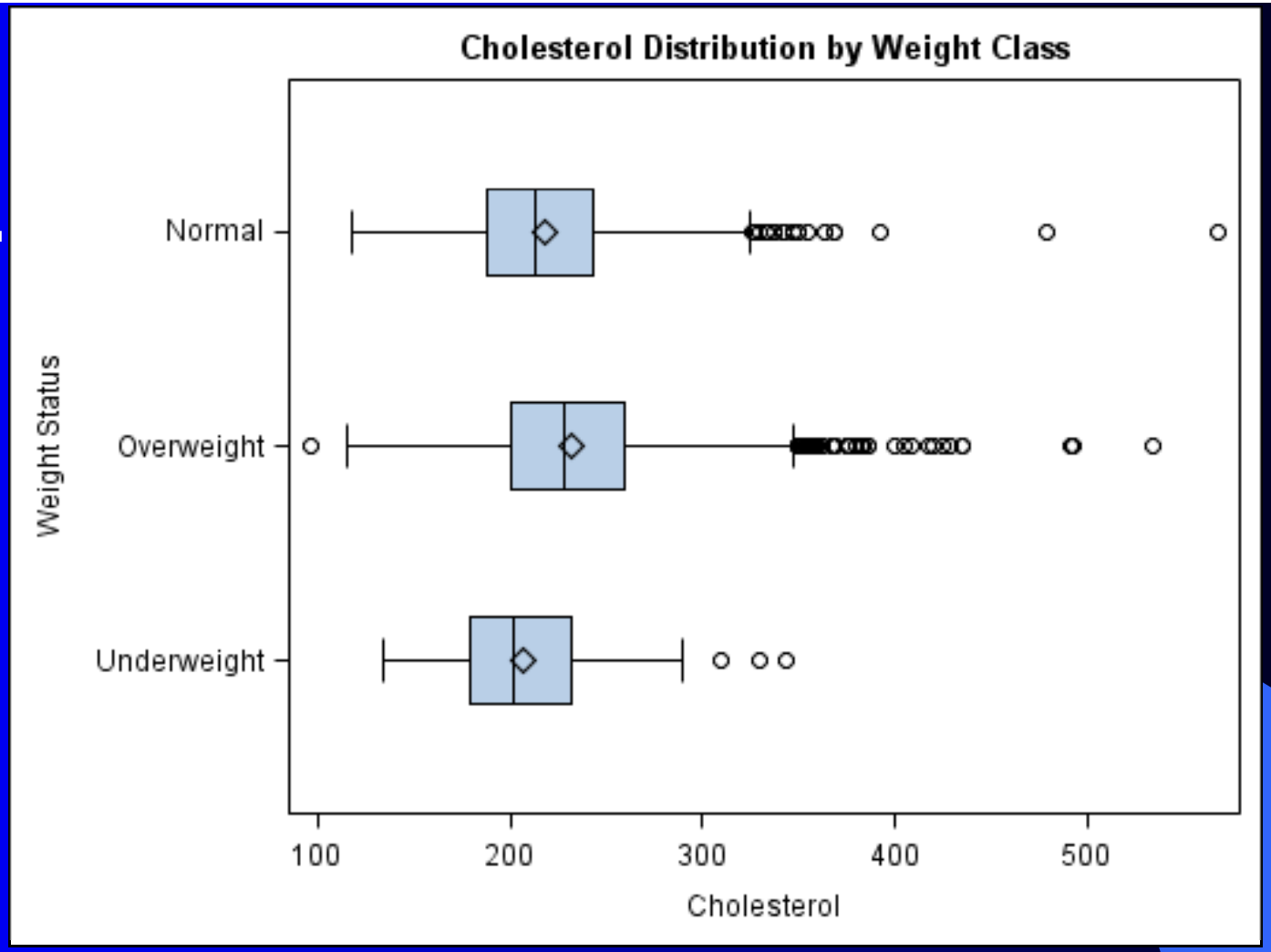

```
proc sgscatter data=dset;
        matrix <vble list> / group=category-vble;
run;
```

PROC SGSCATTER




```
proc sgscatter data=dset;
    compare x=(<vble list>)  y=(<vble list>)
                / reg ellipse=mean
run;
```

PROC SGPLOT




```
proc sgplot data=dset;
    hbox <vble> / category=<category-vble>;
run;
```

Why use multiple linear regression

- You want to investigate a collection of factors for their potential association with the outcome of interest

- You want to investigate a collection of known relevant factors for their ability to predict the outcome of interest.

The GOAL

To obtain a parsimonious set of variables that efficiently predicts the response variable of interest.

Step 2: Model Selection

PROC REG supports a variety of model selection methods but does not support a CLASS statement.

PROC GLM supports the CLASS statement but does not include the model selection methods

PROC GLMSELECT supports the CLASS statement and includes model selection methods but does not include regression diagnostics or hypothesis testing, LS-means etc.

Step 2: Model Selection

PROC GLMSELECT:

- Only available in SAS 9.2
- Can download from SAS website for 9.1

http://support.sas.com/rnd/app/da/glmselect.html

Model Selection Methods

	PROC REG	PROC GLMSELECT
Forward Selection <FORWARD>	✓	✓
Backward Elimination <BACKWARD>	✓	✓
Stepwise Selection <STEPWISE>	✓	✓
Maximum R^2 Improvement <MAXR>	✓	
Minimum R^2 Improvement <MINR>	✓	
R^2 Selection <RSQUARE>	✓	
Adjusted R^2 Selection <ADJRSQ>	✓	
Mallow's C_p Selection <CP>	✓	
Least Angle Regression Selection <LARS>		✓
Lasso Selection <LASSO>		✓

Model Selection Methods:
Least Angle Regression Selection (LARS)

- First presented in 2004
- Useful when number of parameters is large compared to number of observations
- Developed by:
  - Bradley Efron, Trevor Hastie, Iain Johnstone and Robert Tibshirani
- Reference:
  - The Annals of Statistics 2004, Vol 32(2); 407-499

Step 2: Model Selection SAS Code

- PROC REG:

model y = <*regressors*> / **selection=***method* **;**

- PROC GLMSELECT:

model *y* = <*regressors*>
/ **selection=***method*(<*method options*>)

SAS Code: Example 1

- PROC REG:

Example:

model y = *x1 x2 x3*
/ **selection=***forward;* **SLE=***0.05***;**

- PROC GLMSELECT:

Example:

model y = x1 x2 x3
/ **selection=***forward***(select=***SL***)SLE=***0.05***);**

Step 2: Model Selection using Significance Level

	PROC REG	PROC GLMSELECT
Significance Level <SL>	Do not need to specify	
Used with selection methods FORWARD, BACKWARD & STEPWISE		✓
Criterion for entry into the model <SLE>		
Used with selection methods:	Defaults	Defaults
FORWARD	0.50	0.15
STEPWISE	0.15	0.15
Criterion for staying in the model <SLS>		
Used with selection methods:	Defaults	Defaults
BACKWARD	0.10	0.15
STEPWISE	0.15	0.15

Step 2: Model Selection Criterion in PROC GLMSELECT

- Adjusted R2 Statistic <ADJRSQ>
- Akaike Information Criteria <AIC>
- Corrected AIC <AICC>
- Mallow's C_p Statistic <CP>
- Schwarz Bayesian Information Criteria <SBC>
- Significance Level <SL>
- Predicted Residual SS statistic <PRESS>
- Sawa Bayesian Information Criteria <BIC>

SAS Code: Example 2

- PROC GLMSELECT:

Example:

```
model y = x1 x2 x3
        / selection=forward(select=SL
                            SLE=0.05
                            stop=PRESS);
```

Effects are added to the model based on significance level of 0.05

Selection terminates if adding any effect increases the predicted residual SS (PRESS)

Example 3: INCLUDE statement

Includes the first n variables in the model

- PROC REG:

Example:

model y = *x1 x2 x3 x4 x5 x6*
/ **selection=***forward* **include**=*2;*

- PROC GLMSELECT:

Example:

model y = *x1 x2 x3 x4 x5 x6*
/ **selection=***forward*(**select=***SL* **include=***2*);

Example 4: Categorical Variables

Response Variable = lntm

Independent Variables;

- 'Skill' is categorical with 3 levels
- 'Fieldgp' is categorical with 4 levels

Example 4: Categorical Variables

In PROC REG:

You need to produce 7 dummy variables;

skill1 skill2 skill3 fieldgp1 fieldgp2 fieldgp3 fieldgp4

	Skill1	**skill2**	**skill3**
category 1	1	0	0
category 2	0	1	0
category 3	0	0	1

	fileldgp1	**fieldgp2**	**fieldgp3**	**fieldgp4**
category 1	1	0	0	0
category 2	0	1	0	0
category 3	0	0	1	0
category 4	0	0	0	1

Example 4: Categorical Variables

- PROC REG:

```
proc reg data=dset;
   model y = {skill1 skill2 skill3} {fieldgp1 fieldgp2 fieldgp3 fieldgp4}
                / selection=forward SLE=0.05
                  groupnames='Skill' 'Fieldgp';
run;
```

- PROC GLMSELECT:

```
proc glmselect data=dset;
   class skill fieldgp;
   model y = x1 x2 / selection=forward(select=SL SLE=0.05);
run;
```

Window Help

Output - (Untitled)

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: lntm

Forward Selection: Step 2

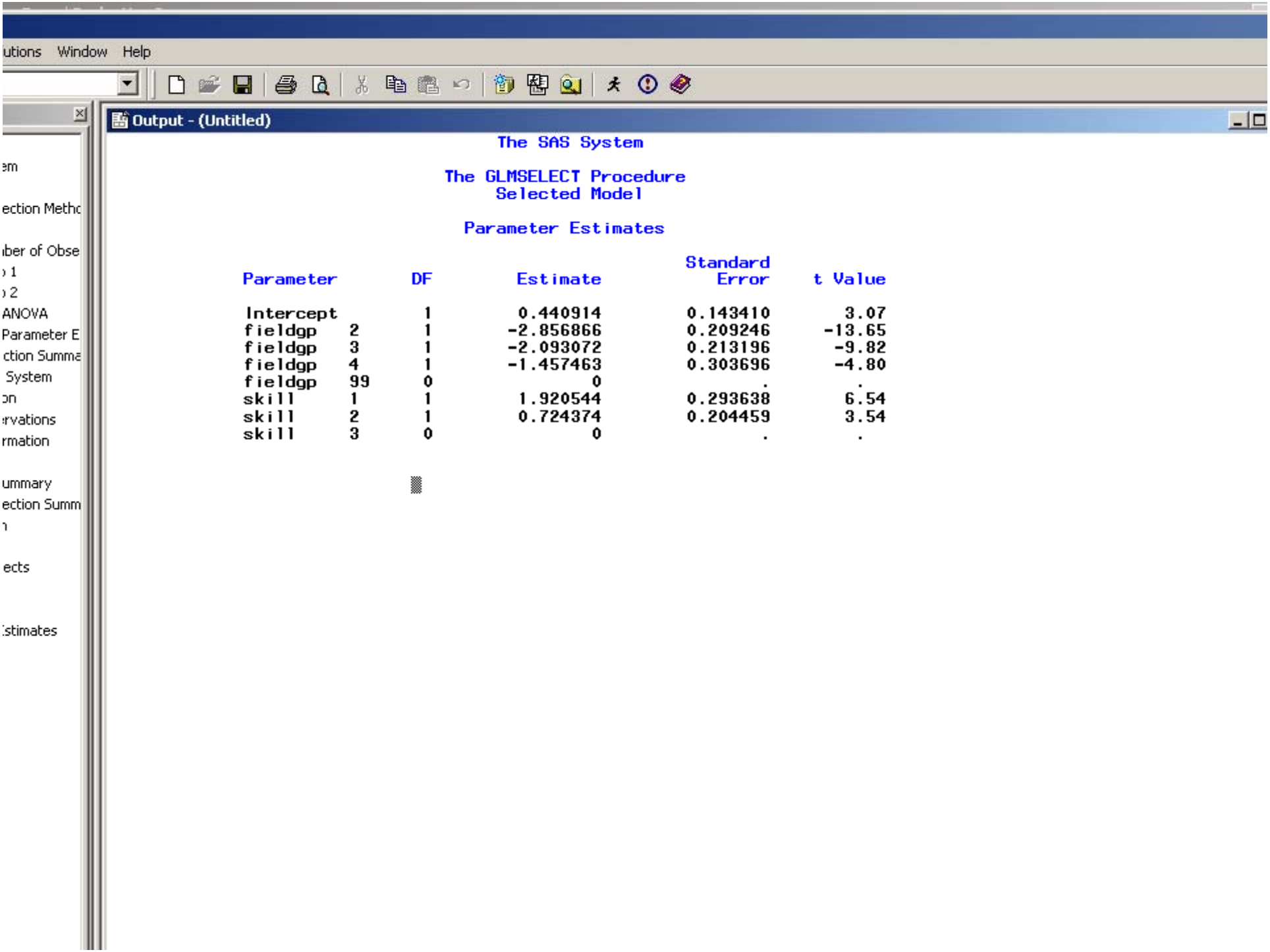
Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	0.44091	0.14341	19.00885	9.45	0.0023
--- Group skill ---			94.40786	23.47	<.0001
skill1	1.92054	0.29364	86.02636	42.78	<.0001
skill2	0.72437	0.20446	25.24178	12.55	0.0005
--- Group fieldgp ---			396.15822	65.67	<.0001
fieldgp1	-2.85687	0.20925	374.86296	186.41	<.0001
fieldgp2	-2.09307	0.21320	193.82826	96.39	<.0001
fieldgp3	-1.45746	0.30370	46.31518	23.03	<.0001

Bounds on condition number: 1.7554, 33.953

No other group of variables met the 0.0500 significance level for entry into the model.

Summary of Forward Selection

Step	Group Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	fieldgp	3	0.3039	0.3039	48.9462	49.48	<.0001
2	skill	5	0.0849	0.3888	6.0000	23.47	<.0001



The SAS System

The GLMSELECT Procedure
Selected Model

Parameter Estimates

Parameter		DF	Estimate	Standard Error	t Value
Intercept		1	0.440914	0.143410	3.07
fieldgp	2	1	-2.856866	0.209246	-13.65
fieldgp	3	1	-2.093072	0.213196	-9.82
fieldgp	4	1	-1.457463	0.303696	-4.80
fieldgp	99	0	0	.	.
skill	1	1	1.920544	0.293638	6.54
skill	2	1	0.724374	0.204459	3.54
skill	3	0	0	.	.

Summary

- PROC REG
  - Can carry out the full modeling process within the same procedure
  - Need to create dummy variables
  - Less control over model selection technique
- PROC GLMSELECT
  - Utilizes the CLASS statement
  - Modeling techniques are very flexible, more control
  - Need to export information into PROC REG or PROC GLM to investigate model further

Remember…

You know more than the computer!

And you need to incorporate this knowledge into the analysis.