Recall that the overall hypotheses we want to test are

$$H_0 : \beta_1 = \beta_2 = 0 \;\; vs \;\; H_A : \text{ at least one is non-zero}$$

This is the test done in the ANOVA table given in the output from a MLR model. This is called the **global F-test** as it tests whether at least one of the terms in the model is important for predicting the response.

The ANOVA table for MLR follows the same ideas as in SLR. We are taking the total amount of variation in the response (SS(Tot)) and partitioning it into a part due to the model (SS(R)) and a part due to experimental error (SS(E)). In fact, the formulas for the sums of squares remain the same, only the degrees of freedom and the F-distribution used for finding the p-value change.

The full ANOVA table for MLR is given below:

Source	Sum of squares	df	Mean Square	F-Ratio
Regression	$SS(R)$	p	$MS(R)$	$MS(R)/MS(E)$
Error	$SS(E)$	$n-p-1$	$MS(E)$	
Total	$SS(Tot)$	$n-1$		

How to do MLR in SAS?
The following code will produce output appropriate for analysis:

```
proc reg data=adexp ;
model adsorp=aluminum iron/clb;
run;
```

Output From Proc Reg for Adsorption Example 1

The REG Procedure
Model: MODEL1
Dependent Variable: adsorp

Number of Observations Read	14
Number of Observations Used	13
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	3529.90308	1764.95154	92.03	<.0001
Error	10	191.78922	19.17892		
Corrected Total	12	3721.69231			

Root MSE	4.37937	R-Square	0.9485
Dependent Mean	29.84615	Adj R-Sq	0.9382
Coeff Var	14.67316		

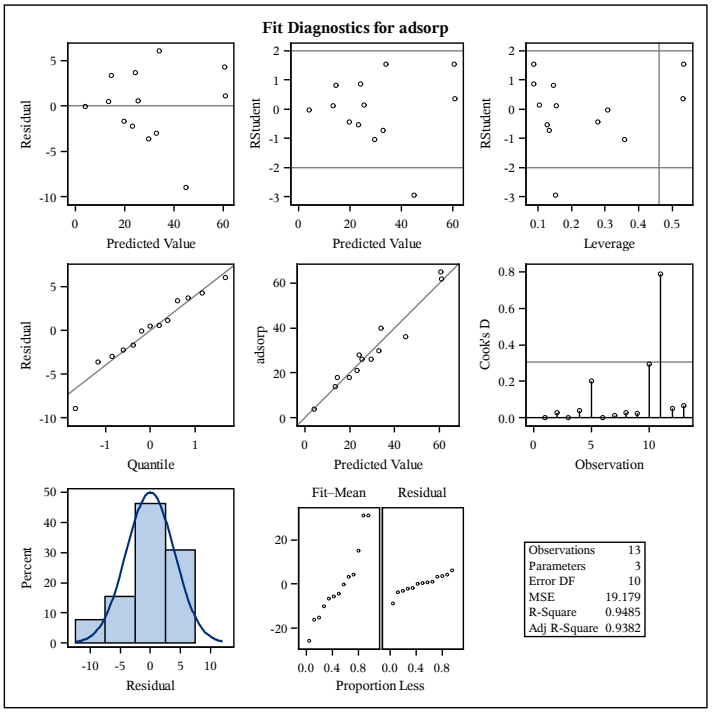
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-7.35066	3.48467	-2.11	0.0611	-15.11498	0.41366
aluminum	1	0.34900	0.07131	4.89	0.0006	0.19012	0.50788
iron	1	0.11273	0.02969	3.80	0.0035	0.04658	0.17889

Note! The tests in the parameter estimate table are tests for that β coefficient being 0 ***after accounting for the other predictors in the model.***

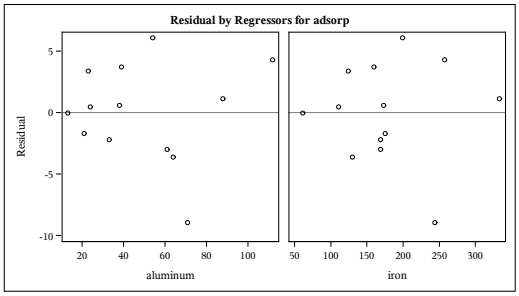
Output From Proc Reg for Adsorption Example 2

The REG Procedure
Model: MODEL1
Dependent Variable: adsorp



Output From Proc Reg for Adsorption Example 3

The REG Procedure
Model: MODEL1
Dependent Variable: adsorp



A non-additive model example:
A random sample of students taking the same exam:

IQ	Study TIME	GRADE
105	10	75
110	12	79
120	6	68
116	13	85
122	16	91
130	8	79
114	20	98
102	15	76

Consider regressing GRADE on IQ (X_1), TIME(X_2), and TI ($X_1 * X_2$), where TI = TIME*IQ. That is, we fit the model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + E$$

```
proc reg;
model Grade = IQ Time TI;
run;
```

```
                                The SAS System                                1
                              The REG Procedure

                           Analysis of Variance

                                    Sum of           Mean
Source                      DF      Squares         Square    F Value    Pr > F

Model                        3    610.81033      203.60344      26.22    0.0043
Error                        4     31.06467        7.76617
Corrected Total              7    641.87500

                           Parameter Estimates

                    Parameter     Standard
Variable    DF       Estimate        Error  t Value  Pr > |t|
Intercept    1       72.20608     54.07278     1.34    0.2527
IQ           1       -0.13117      0.45530    -0.29    0.7876
Time         1       -4.11107      4.52430    -0.91    0.4149
TI           1        0.05307      0.03858     1.38    0.2410
```

Discussion of the interaction model:
We call the product TI = Time*IQ an "interaction" term. That is, our explanatory variables do not have an independent effect on the response.

$$\widehat{MeanGrade} = 72.21 - 0.13 * IQ - 4.11 * Time + 0.0531 * TI$$

Now if IQ = 100 we get

$$\widehat{MeanGrade} = (72.21 - 13.1) + (-4.11 + 5.31) * Time$$

and if IQ 120 we get

$$\widehat{MeanGrade} = (72.21 - 15.7) + (-4.11 + 6.37) * Time.$$

Thus we expect an extra hour of study to increase the grade by 1.20 points for someone with IQ = 100 and by 2.26 points for someone with IQ = 120 if we use this interaction model.

Generally, we can interpret the (true) β parameters in the model as:

- β_0 - Average value of Grade when IQ and Study Time are 0
- β_1 - Average change in Grade for a unit increase in IQ when Study Time is 0
- β_2 - Average change in Grade for a unit increase in Study Time when IQ is 0
- β_3 - Average change in the slope for IQ (or Study Time) for a given value of Study Time (or IQ).

The interpretation of the interaction 'slope' can be seen by looking at the following:

$$\mu(x_1+1,x_2)-\mu(x_1,x_2)=\beta_0+\beta_1(x_1+1)+\beta_2x_2+\beta_3(x_1+1)(x_2)-\beta_0-\beta_1x_1-\beta_2x_2-\beta_3x_1(x_2)$$

$$=\beta_1+\beta_3x_2$$

So β_3 is the amount the slope for x_1 changes per unit change in x_1 while x_2 is held constant.

Note: The global p-value is significant, but none of our individual terms are. This gives evidence that our model is over-fit. we may want to go back to the simpler "main effects" model.

The next idea to tackle is what model to use if we are unsure of the predictors we want in our model. This idea is called ***model selection****.*

Model Selection:

x_1, x_2, x_3 denote p independent variables. Consider several models:

1. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1$
2. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2$
3. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_3 x_3$
4. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
5. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_3 x_3$
6. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
7. $\mu(x_1, x_2, x_3) = E(Y|x_1, x_2, x_3) = \beta_0 + \beta_2 x_2 + \beta_3 x_3$

A is nested in B means model A can be obtained by restricting (e.g. setting to 0 or setting equal β's) parameter values in model B.

True or false:

- Model 1 nested in Model 4 True Model 1 nested in Model 5 True
- Model 2 nested in Model 4 True Model 4 nested in Model 1 False
- Model 3 nested in Model 4 True Model 5 nested in Model 4 True
- Model 3 nested in Model 7 True Model 1 nested in Model 7 False

A nested in $B \longrightarrow A$ called reduced model, B called full model.

p - number of regression parameters in full model
q - number of regression parameters in reduced model
$p - q$ - number of regression parameters being tested.

In comparing two models, suppose

$$\beta_1, \ldots, \beta_q \text{ in reduced model } (A)$$
$$\beta_1, \ldots, \beta_q, \beta_{q+1}, \ldots, \beta_p \text{ in full model } (B).$$

Comparison of models A and B amounts to testing

$$H_0 : \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0 \text{ (model } A \text{ ok)}$$

$$H_1 : \beta_{q+1}, \beta_{q+2}, \ldots, \beta_p \text{ not all 0 (model } B \text{ adds something)}$$

To test this hypothesis we can use the F distribution with $p - q$ numerator df and $n - p - 1$ denominator df

$$F = \frac{(SS(E)_r - SS(E)_f)/(p-q)}{MS(E)_f} = \frac{(SS(R)_f - SS(R)_r)/(p-q)}{MS(E)_f}$$

(r and f abbreviate reduced and full, respectively.)

Difference in the numerator called an **extra regression sum of squares**:

$$R(\beta_{q+1}, \beta_{q+2}, \ldots, \beta_p | \beta_0, \beta_1, \beta_2, \ldots, \beta_q) = SS(R)_f - SS(R)_r.$$

(ok to supress β_0 in these extra SS terms.)

Consider why this test stat makes sense. $SS(R)_f - SS(R)_r$ can be thought of as the amount of variation in Y (or part of SS(Tot)) that can be attributed to the variables in the alternative hypothesis. If the variables in the alternative are really meaningful, this should a relatively large quantity compared to MS(E).

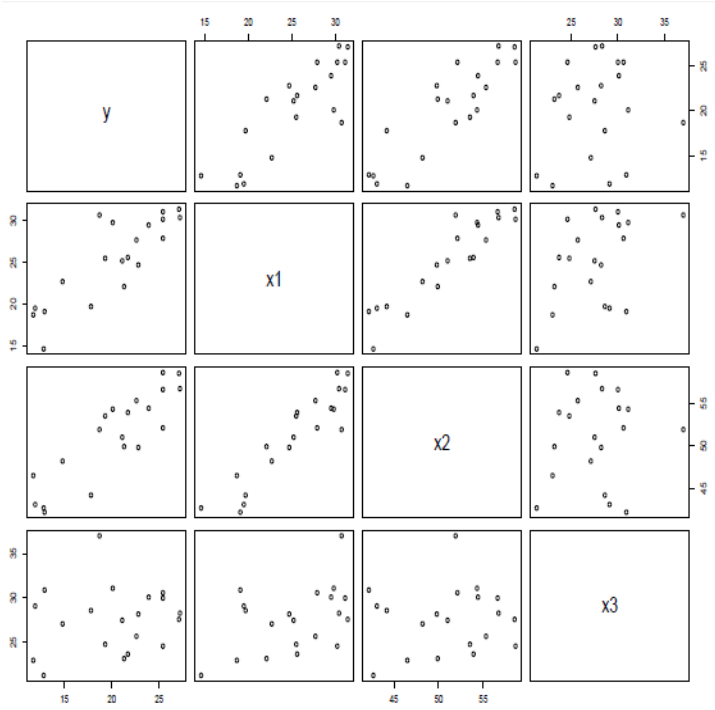
Let's get a handle on this notation. Give the extra regression SS terms for comparing some of the nested models on preceding page:

- Model 1 in model 4: $R(\beta_2, \beta_3|\beta_1)$
- Model 2 in model 4: $R(\beta_1, \beta_3|\beta_2)$
- Model 3 in model 4: $R(\beta_1, \beta_2|\beta_3)$
- Model 1 in model 5: $R(\beta_3|\beta_1)$
- Model 5 in model 4: $R(\beta_2|\beta_1, \beta_3)$

An example: How to measure body fat? For each of $n = 20$ healthy individuals, the following measurements were made: bodyfat percentage y_i, triceps skinfold thickness, x_1, thigh circumference x_2, midarm circumference x_3.

x1	x2	x3	y
19.5	43.1	29.1	11.9
24.7	49.8	28.2	22.8
30.7	51.9	37.0	18.7
29.8	54.3	31.1	20.1
19.1	42.2	30.9	12.9
25.6	53.9	23.7	21.7
31.4	58.5	27.6	27.1
27.9	52.1	30.6	25.4
22.1	49.9	23.2	21.3
25.5	53.5	24.8	19.3
31.1	56.6	30.0	25.4
30.4	56.7	28.3	27.2
18.7	46.5	23.0	11.7
19.7	44.2	28.6	17.8
14.6	42.7	21.3	12.8
29.5	54.4	30.1	23.9
27.7	55.3	25.7	22.6
30.2	58.6	24.6	25.4
22.7	48.2	27.1	14.8
25.2	51.0	27.5	21.1

```
ods graphics on;
proc corr plots=matrix;
var y x1 x2 x3;
run;
```




```
Pearson Correlation Coefficients, N = 20
        Prob > |r| under H0: Rho=0
```

	y	x1	x2	x3
y	1.00000	0.84327	0.87809	0.14244
		<.0001	<.0001	0.5491
x1	0.84327	1.00000	0.92384	0.45778
	<.0001		<.0001	0.0424
x2	0.87809	0.92384	1.00000	0.08467
	<.0001	<.0001		0.7227
x3	0.14244	0.45778	0.08467	1.00000
	0.5491	0.0424	0.7227	

Looking at the scatter plots and the correlation output, marginal associations between y and x_1 and between y and x_2 are highly significant, providing evidence of a strong $r \approx 0.85$ linear association between average bodyfat and triceps skinfold and between average bodyfat and thigh circumference.

Notice the scatter plot between x_1 and x_2, there is a strong linear relationship. This means that triceps skinfold and thigh circumference are giving some of the same information. This can lead to issues when fitting a model.

Multicollinearity: linear associations among the independent variables; causes problems such as inflated sampling variances for $\hat{\boldsymbol{\beta}}$.

```
proc reg data=bodyfat;
```

```
   model y=x1/covb;
   model y=x2/covb;
   model y=x3/covb;
   model y=x1 x2/covb;
   model y=x1 x2 x3/covb;
run;
```

Yields the following output:

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	352.26980	352.26980	44.30	<.0001
Error	18	143.11970	7.95109		
Corrected Total	19	495.38950			

Root MSE	2.81977	R-Square	0.7111
Dependent Mean	20.19500	Adj R-Sq	0.6950
Coeff Var	13.96271		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.49610	3.31923	-0.45	0.6576
x1	1	0.85719	0.12878	6.66	<.0001

Covariance of Estimates

Variable	Intercept	x1
Intercept	11.01731839	-0.419670565
x1	-0.419670565	0.0165844918

The REG Procedure
Model: MODEL2
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

Root MSE	2.51024	R-Square	0.7710
Dependent Mean	20.19500	Adj R-Sq	0.7583
Coeff Var	12.43002		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-23.63449	5.65741	-4.18	0.0006
x2	1	0.85655	0.11002	7.79	<.0001

Covariance of Estimates

Variable	Intercept	x2
Intercept	32.006329324	-0.619332881
x2	-0.619332881	0.0121034372

The REG Procedure
Model: MODEL3
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	10.05160	10.05160	0.37	0.5491
Error	18	485.33790	26.96322		
Corrected Total	19	495.38950			

Root MSE	5.19261	R-Square	0.0203
Dependent Mean	20.19500	Adj R-Sq	-0.0341
Coeff Var	25.71236		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	14.68678	9.09593	1.61	0.1238
x3	1	0.19943	0.32663	0.61	0.5491

Covariance of Estimates		
Variable	Intercept	x3
Intercept	82.735867956	-2.946694682
x3	-2.946694682	0.1066869907

The REG Procedure
Model: MODEL4
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	385.43871	192.71935	29.80	<.0001
Error	17	109.95079	6.46769		
Corrected Total	19	495.38950			

Root MSE	2.54317	R-Square	0.7781
Dependent Mean	20.19500	Adj R-Sq	0.7519
Coeff Var	12.59305		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-19.17425	8.36064	-2.29	0.0348
x1	1	0.22235	0.30344	0.73	0.4737
x2	1	0.65942	0.29119	2.26	0.0369

Covariance of Estimates			
Variable	Intercept	x1	x2
Intercept	69.900312587	1.8469661215	-2.273097628
x1	1.8469661215	0.0920751757	-0.081628463
x2	-2.273097628	-0.081628463	0.0847900309

The REG Procedure
Model: MODEL5
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	R-Square	0.8014
Dependent Mean	20.19500	Adj R-Sq	0.7641
Coeff Var	12.28017		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	117.08469	99.78240	1.17	0.2578
x1	1	4.33409	3.01551	1.44	0.1699
x2	1	-2.85685	2.58202	-1.11	0.2849
x3	1	-2.18606	1.59550	-1.37	0.1896

Covariance of Estimates

Variable	Intercept	x1	x2	x3
Intercept	9956.5279384	300.1979628	-257.3823153	-158.6704127
x1	300.1979628	9.0933087788	-7.779145105	-4.7880263
x2	-257.3823153	-7.779145105	6.6668028532	4.0946155019
x3	-158.6704127	-4.7880263	4.0946155019	2.545617053

Question: Why is the global p-value in the last model significant, i.e. at least one predictor is useful, but the individual tests are all nonsignificant?

Each individual test is a test for that variable given the other factors are retained in the model.

Output and significance results discussed. Note the covariance of the parameter estimates are inflated in the last model. This is exactly the type of situation where we might want to employ a model selection strategy.

In the bodyfat data, consider comparing the simple model that Y depends only on x_1 (triceps) versus the full model that it depends on all three.

$$\begin{aligned} \text{Model } A: \mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 \\ \text{Model } B: \mu(x_1, x_2, x_3) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \end{aligned}$$

or the null hypothesis

$$H_0 : \beta_2 = \beta_3 = 0 \quad \text{vs} \quad H_1 : \beta_2, \beta_3 \text{ not both } 0$$

after accounting for x_1. Our F statistic can be used

$$F = \frac{(396.9 - 352.3)/2}{6.15} = \frac{22.3}{6.15} = 3.64$$

How many *df* for numerator and denominator?

The 95^{th} percentile is $F(0.05, 2, 16) = 3.63$.

Our conclusion about the hypotheses?
Reject H_0 in favor of H_A. The full model gives us something more than just the SLR model with x1.
That is, after accounting for the linear dependence between triceps and bodyfat, there is still some linear association between mean bodyfat and at least one of x_2, x_3 (thigh,midarm).

To get the nested model F-ratio in SAS:

```
proc reg data=bodyfat;
     model y=x1 x2 x3;
     test x2=0,x3=0;
run;
```

Full mode vs only Triceps 4

The REG Procedure
Model: MODEL1

Test 1 Results for Dependent Variable y

Source	DF	Mean Square	F Value	Pr > F
Numerator	2	22.35741	3.64	0.0500
Denominator	16	6.15031		

However, we saw in the previous output that a model with all three variables is no good. This is due to the multicollinearity. We will now very briefly look at a few automated model selection techniques.

Using proc reg to perform variable selection:
We'll discuss three hypothesis testing methods for selecting variables (there are many other ways to accomplish this we won't discuss).

1. **Forward Selection** - Start with nothing and work forward.
   (a) Begin with a model with only β_0
   (b) Calculate $R(\beta_i|\beta_0)$ for all possible predictors and find p-values for each
   (c) Take most significant p-value less than a cutoff (say 0.3), add predictor into model.
   (d) Say β_j was added in the last step, repeat above process with added predictor. That is, calculate $R(\beta_i|\beta_0, \beta_j)$ for all other predictors, etc.
   (e) Stop when no predictors are below the cutoff or if the full model is selected.
2. **Backward Selection** - Start with everything and work backward.
   (a) Start with full model.
   (b) Locate variable with largest p-value greater than a cutoff (say 0.1), remove that variable.
   (c) Repeat until all p-values are less than the cut off or the null model (intercept only model) is chosen.
3. **Subset Selection** - Compute all possible models, pick best.
   (a) Compare each of the models using a criterion.
   (b) Choose model that minimizes that criterion. Possible criteria include:
      - *Adjusted* $R^2 = 1 - \frac{n-1}{n-p-1}(1 - R^2)$ (takes into account the addition of more predictors)
      - Mallow's C_P, AIC, AICc, or BIC (all take into account the model complexity, not just how well the model fits the data)

How to do these model selection methods in SAS?

```
proc reg data=bodyfat plots=none;
    model y=x1 x2 x3/selection=cp ;
    model y=x1 x2 x3/selection=forward SLentry=0.3;
    model y=x1 x2 x3/selection=backward SLstay=0.1;
   model y=x1 x2 x3/selection=adjrsq;
run;
```

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL1
Dependent Variable: y

C(p) Selection Method

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Number in Model	C(p)	R-Square	Variables in Model
1	2.4420	0.7710	x2
2	3.2242	0.7862	x1 x3
2	3.8773	0.7781	x1 x2
3	4.0000	0.8014	x1 x2 x3
2	4.0657	0.7757	x2 x3
1	7.2703	0.7111	x1
1	62.9128	0.0203	x3

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL2
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Forward Selection: Step 1

Variable x2 Entered: R-Square = 0.7710 and C(p) = 2.4420

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	381.96582	381.96582	60.62	<.0001
Error	18	113.42368	6.30132		
Corrected Total	19	495.38950			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	-23.63449	5.65741	109.97344	17.45	0.0006
x2	0.85655	0.11002	381.96582	60.62	<.0001

Bounds on condition number: 1, 1

No other variable met the 0.3000 significance level for entry into the model.

Summary of Forward Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x2	1	0.7710	0.7710	2.4420	60.62	<.0001

The REG Procedure
Model: MODEL3
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.8014 and C(p) = 4.0000

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	117.08469	99.78240	8.46816	1.38	0.2578
x1	4.33409	3.01551	12.70489	2.07	0.1699
x2	-2.85685	2.58202	7.52928	1.22	0.2849
x3	-2.18606	1.59550	11.54590	1.88	0.1896

Bounds on condition number: 708.84, 4133.4

Backward Elimination: Step 1

Variable x2 Removed: R-Square = 0.7862 and C(p) = 3.2242

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	389.45533	194.72767	31.25	<.0001
Error	17	105.93417	6.23142		
Corrected Total	19	495.38950			

Variable Selection Methods on Bodyfat Example

The REG Procedure
Model: MODEL3
Dependent Variable: y

Backward Elimination: Step 1

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	6.79163	4.48829	14.26834	2.29	0.1486
x1	1.00058	0.12823	379.40373	60.89	<.0001
x3	-0.43144	0.17662	37.18554	5.97	0.0258

Bounds on condition number: 1.2651, 5.0605

All variables left in the model are significant at the 0.1000 level.

Summary of Backward Elimination							
Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x2	2	0.0152	0.7862	3.2242	1.22	0.2849

The REG Procedure
Model: MODEL4
Dependent Variable: y

Adjusted R-Square Selection Method

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Number in Model	Adjusted R-Square	R-Square	Variables in Model
3	0.7641	0.8014	x1 x2 x3
2	0.7610	0.7862	x1 x3
1	0.7583	0.7710	x2
2	0.7519	0.7781	x1 x2
2	0.7493	0.7757	x2 x3
1	0.6950	0.7111	x1
1	-.0341	0.0203	x3

Models selected discussed. Notice that they are not the same. You should bring some subject matter knowledge into play here.

If you notice, we now really have multiple tests for a given slope term. A different test for each set of variables already being accounted for. Let's discuss this idea in more detail.

Types of Sums of Squares

Given that we have 4 predictors, $X_1 - X_4$ we really can have a number of tests based on nested models for $\beta_4 = 0$ (or for any other β for that matter). Let's write them down in terms of extra regression sums of squares:

$R(\beta_4|\beta_0)$ (SLR test)
$R(\beta_4|\beta_0, \beta_1)$ (test after accounting for X_1)
$R(\beta_4|\beta_0, \beta_2)$ (test after accounting for X_2)
$R(\beta_4|\beta_0, \beta_3)$ (test after accounting for X_3)
$R(\beta_4|\beta_0, \beta_1, \beta_2)$ (test after accounting for X_1 and X_2)
$R(\beta_4|\beta_0, \beta_1, \beta_3)$ (test after accounting for X_1 and X_3)
$R(\beta_4|\beta_0, \beta_2, \beta_3)$ (test after accounting for X_2 and X_3)
$R(\beta_4|\beta_0, \beta_1, \beta_2, \beta_3)$ (test after accounting for X_1, X_2, and X_3)

Some of these tests can be easily found using different types of sums of squares.

- Type I sums of squares - sequential, test for adding the variable after all *previous* variables are accounted for (order of variables in model determine the tests).
- Type II sums of squares - partial, test for adding the variable once all other terms not containing a function of that variable are accounted for (i.e. interactions/quadratics/etc).
- Type III sums of squares - partial, test for adding the variable after all other terms in the model are accounted for.

The tests given for the parameter estimates are all type III tests and this is the test usually done to determine if a slope term has significance. However, type I tests are very useful for model building. For example, if we wanted to look at building a model for the bodyfat example and we thought the order of importance for the variables was X_1 (triceps), X_3 (midarm), and X_2 (thigh), we could get sequential tests for these models using type I sums of squares.

In SAS proc reg use the following code:

```
proc reg data=bodyfat;
   model y=x1 x3 x2/ss1; *Note the order of variables is important for Type I;
run;
```

Sequential tests for bodyfat example **1**

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	21
Number of Observations Used	20
Number of Observations with Missing Values	1

Analysis of Variance					
Source	**DF**	**Sum of Squares**	**Mean Square**	**F Value**	**Pr > F**
Model	3	396.98461	132.32820	21.52	<.0001
Error	16	98.40489	6.15031		
Corrected Total	19	495.38950			

Root MSE	2.47998	**R-Square**	0.8014
Dependent Mean	20.19500	**Adj R-Sq**	0.7641
Coeff Var	12.28017		

Parameter Estimates						
Variable	**DF**	**Parameter Estimate**	**Standard Error**	**t Value**	**Pr > |t|**	**Type I SS**
Intercept	1	117.08469	99.78240	1.17	0.2578	8156.76050
x1	1	4.33409	3.01551	1.44	0.1699	352.26980
x3	1	-2.18606	1.59550	-1.37	0.1896	37.18554
x2	1	-2.85685	2.58202	-1.11	0.2849	7.52928

Let's label the Type I SS in terms of extra regression sums of squares (R notation).

Note: we will soon use proc glm for our model analysis and this gives even better output for type I sums of squares. (The tests given for type I sums of squares use the *full model* MS(E) rather than the full model MS(E) up to that point. This test still works because MS(E) from each model is an unbiased estimate of σ^2. The tests using the different MS(E) terms could give different results, but will usually agree.

```
proc glm data=bodyfat;
   model y=x1 x3 x2;
run;
```

Sequential tests for bodyfat example using GLM 2

The GLM Procedure

Dependent Variable: y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	396.9846118	132.3282039	21.52	<.0001
Error	16	98.4048882	6.1503055		
Corrected Total	19	495.3895000			

R-Square	Coeff Var	Root MSE	y Mean
0.801359	12.28017	2.479981	20.19500

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x1	1	352.2697968	352.2697968	57.28	<.0001
x3	1	37.1855371	37.1855371	6.05	0.0257
x2	1	7.5292779	7.5292779	1.22	0.2849

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x1	1	12.70489278	12.70489278	2.07	0.1699
x3	1	11.54590217	11.54590217	1.88	0.1896
x2	1	7.52927788	7.52927788	1.22	0.2849

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	117.0846948	99.78240295	1.17	0.2578
x1	4.3340920	3.01551136	1.44	0.1699
x3	-2.1860603	1.59549900	-1.37	0.1896
x2	-2.8568479	2.58201527	-1.11	0.2849