Multiple Linear Regression

Multiple linear regression allows you to determine the linear relationship between a dependent variable (Y) and a series of independent variables (X_1, X_2, X_3, . . . X_n).

The linear model is: $Y=\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots + \beta X_{in} + \varepsilon_i$

Where: Yi is the observed response of the ith individual,
Xi1, Xi2, Xi3, . . . Xin are the levels of the different independent variables,
β_0 is the Y-intercept,
$\beta_1, \beta_2, \beta_3, \ldots \beta X_n$ are the regression coefficients for the respective independent variables, and
ε_i is the random error (i.e., residual).

We will use a least squares method to,

1. Estimate $\beta_1, \beta_2, \beta_3, \ldots \beta X_n$,

2. Test to see if the independent variables significantly contribute to explaining the the variation in the dependent variable Y.

This will be accomplished using the PROC GLM command of SAS.

Interpretation of the PROC GLM Output

1. ANOVA at the top of the page
   - *F*=Model MS/Error MS

   - Hypothesis tested is: H_o: The model is not significant; all regression coefficients =0.(i.e., H_o: $\beta_1, \beta_2, \beta_3, \ldots \beta X_n = 0$)

     H_A: The model is significant; at least one of the regression coefficients does not equal zero.

2. ANOVA with the Type I Sum of Squares.
   - Type I SS are the sequential sum of squares.

   - Given three independent variables X_1, X_2, and X_3

     SS X_1= Sum of square of X_1

     SS ($X_2|X_1$) = SS when X_2 is added to the model that already includes X_1

     SS ($X_3|X_1, X_2$) = SS when X_3 is added to the model that already includes X_1 and X_2.

   - The sum of all the Type I SS = the Model SS.

3. The Type III SS are the represent the sum of squares that would be calculated if the listed independent variable was added to the model with all the other independent variables.
   - $SS(X_1|X_2, X_3)$ Improvement in SS if X_1 is added to the model already including X_2 and X_3.
   - $SS(X_2|X_1, X_3)$ Improvement in SS if X_2 is added to the model already including X_1 and X_3.
   - $SS(X_3|X_1, X_2)$ Improvement in SS if X_3 is added to the model already including X_1 and X_2.
   - *F*-test for Type III Analysis
     - ❖ $F_{X1}=MS(X_1|X_2, X_3)/\text{Residual MS}$
     - ❖ H_o: The addition of X_1 to the model involving X_2 and X_3 does not significantly improve the model.

       H_A: The addition of X_1 to the model involving X_2 and X_3 does not significantly improve the model.

Selecting the Best Multivariate Model

- The best model can be selected using the Stepwise procedure of SAS.
- For a model that has three possible independent variables, the process would work as follows:

1. Step 1. Of all the one-variable models, the one that yields the largest *R*-square is selected (e.g., X_2). If the *F*-test for this variable is significant, the X_2 is kept in the model. This will be the best one-variable model.
2. Step 2. X_2 is kept in the model and the Partial *F*-tests for $F(X_1|X_2)$ and $F(X_3|X_2)$ are determined. The variable that gives the largest partial *F* is then considered for entry into the model (e.g., X_3). If $F(X_3|X_2)$ is significant, then X_3 is entered into the model and this is the best two-variable model. If $F(X_3|X_2)$ is not significant, the stepwise procedure stops. You should then report the one-variable model as the best model and use it for all interpretation.
3. Step 3. X_2 and X_3 are kept in the model and the Partial *F*-test for $F(X_1|X_2, X_3)$ is calculated. If it is significant, then X_1 is added to the model. This becomes the best three-variable model, and it should be used for all interpretation. If $F(X_1|X_2, X_3)$ is not significant, you should use the best two-variable model.

SAS Commands for Stepwise Regression

- This problem will be done assuming there are nine independent variables (i.e., x1, x2, x3, . . ., xn) and one dependent variable Y.

```
proc glm;
model Y=x1 x2 x3 x4 x5 x6 x7 x8 x9;
title 'GLM Analysis';
run;
proc stepwise;
model Y=x1 x2 x3 x4 x5 x6 x7 x8 x9;
title 'Stepwise Analysis';
run;
```

Stepwise Regression Analysis

The GLM Procedure

GLM Analysis 09:59 Friday, December 3, 2004 1

The GLM Procedure

Number of observations 154

Stepwise Regression Analysis

The GLM Procedure

GLM Analysis 09:59 Friday, December 3, 2004 2

The GLM Procedure

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	9	52.24342807	5.80482534	19.59	<.0001
Error	144	42.67975549	0.29638719		
Corrected Total	153	94.92318356			

R-Square	Coeff Var	Root MSE	Y Mean
0.550376	0.698928	0.544415	77.89275

Source	DF	Type I SS	Mean Square	F Value	Pr > F
x1	1	32.80358368	32.80358368	110.68	<.0001
x2	1	5.82063268	5.82063268	19.64	<.0001
x3	1	0.79936064	0.79936064	2.70	0.1027
x4	1	3.80270359	3.80270359	12.83	0.0005
x5	1	0.00017295	0.00017295	0.00	0.9808
x6	1	4.37834137	4.37834137	14.77	0.0002
x7	1	4.30245443	4.30245443	14.52	0.0002
x8	1	0.00141191	0.00141191	0.00	0.9451
x9	1	0.33476682	0.33476682	1.13	0.2897

Source	DF	Type III SS	Mean Square	F Value	Pr > F
x1	1	4.96988578	4.96988578	16.77	<.0001
x2	1	0.41163775	0.41163775	1.39	0.2405
x3	1	0.37033218	0.37033218	1.25	0.2655
x4	1	6.78264058	6.78264058	22.88	<.0001
x5	1	0.00143995	0.00143995	0.00	0.9445
x6	1	4.64492653	4.64492653	15.67	0.0001
x7	1	1.06042557	1.06042557	3.58	0.0606
x8	1	0.25123345	0.25123345	0.85	0.3588
x9	1	0.33476682	0.33476682	1.13	0.2897

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	85.14069242	2.40772087	35.36	<.0001
x1	-0.03457906	0.00844442	-4.09	<.0001
x2	-0.01390324	0.01179745	-1.18	0.2405
x3	-0.72822097	0.65147413	-1.12	0.2655
x4	-0.01047793	0.00219031	-4.78	<.0001

The GLM Procedure

GLM Analysis 09:59 Friday, December 3, 2004 3

The GLM Procedure

Dependent Variable: Y

Parameter	Estimate	Standard Error	t Value	Pr > |t|
x5	0.00311248	0.04465420	0.07	0.9445
x6	0.07524964	0.01900836	3.96	0.0001
x7	-0.31506598	0.16656783	-1.89	0.0606
x8	-0.55923000	0.60740931	-0.92	0.3588
x9	-0.03089684	0.02907185	-1.06	0.2897

The GLM Procedure

Stepwise Analysis 09:59 Friday, December 3, 2004 4

The STEPWISE Procedure
Model: MODEL1
Dependent Variable: Y

Stepwise Selection: Step 1

Variable x2 Entered: R-Square = 0.3799 and C(p) = 48.6131

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	36.05680	36.05680	93.10	<.0001
Error	152	58.86638	0.38728		
Corrected Total	153	94.92318			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	81.68095	0.39579	16494	42590.3	<.0001
x2	-0.05220	0.00541	36.05680	93.10	<.0001

Bounds on condition number: 1, 1

Stepwise Selection: Step 2

Variable x7 Entered: R-Square = 0.4274 and C(p) = 35.3731

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	40.57374	20.28687	56.36	<.0001
Error	151	54.34945	0.35993		
Corrected Total	153	94.92318			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	83.41949	0.62164	6481.50273	18007.7	<.0001
x2	-0.05154	0.00522	35.10497	97.53	<.0001
x7	-0.43920	0.12398	4.51693	12.55	0.0005

The GLM Procedure

Stepwise Analysis 09:59 Friday, December 3, 2004 5

The STEPWISE Procedure
Model: MODEL1
Dependent Variable: Y

Stepwise Selection: Step 2

Bounds on condition number: 1.0013, 4.0051

Stepwise Selection: Step 3

Variable x1 Entered: R-Square = 0.4651 and C(p) = 25.3215

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	44.14569	14.71523	43.47	<.0001
Error	150	50.77749	0.33852		
Corrected Total	153	94.92318			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	83.23484	0.60554	6395.97351	18894.1	<.0001
x1	-0.02562	0.00789	3.57196	10.55	0.0014
x2	-0.03009	0.00832	4.43022	13.09	0.0004
x7	-0.48962	0.12123	5.52148	16.31	<.0001

Bounds on condition number: 2.7084, 19.294

Stepwise Selection: Step 4

Variable x4 Entered: R-Square = 0.4864 and C(p) = 20.4859

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	46.17167	11.54292	35.28	<.0001
Error	149	48.75151	0.32719		
Corrected Total	153	94.92318			

Stepwise Analysis 09:59 Friday, December 3, 2004 6

The STEPWISE Procedure
Model: MODEL1
Dependent Variable: Y

Stepwise Selection: Step 4

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	83.72286	0.62680	5837.64650	17841.7	<.0001
x1	-0.02604	0.00776	3.68807	11.27	0.0010
x2	-0.02502	0.00843	2.88376	8.81	0.0035
x4	-0.00435	0.00175	2.02598	6.19	0.0139
x7	-0.47699	0.11930	5.23084	15.99	0.0001

Bounds on condition number: 2.873, 31.008

Stepwise Selection: Step 5

Variable x6 Entered: R-Square = 0.5427 and C(p) = 4.4650

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	51.51283	10.30257	35.12	<.0001
Error	148	43.41035	0.29331		
Corrected Total	153	94.92318			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	81.97570	0.72099	3791.76868	12927.4	<.0001
x1	-0.03429	0.00759	5.98080	20.39	<.0001
x2	-0.01943	0.00809	1.69249	5.77	0.0175
x4	-0.01014	0.00214	6.58520	22.45	<.0001
x6	0.07730	0.01812	5.34116	18.21	<.0001
x7	-0.46347	0.11299	4.93466	16.82	<.0001

Bounds on condition number: 2.9506, 53.336

All variables left in the model are significant at the 0.1500 level.

No other variable met the 0.1500 significance level for entry into the model.

Stepwise Regression Analysis

The GLM Procedure

Stepwise Analysis 09:59 Friday, December 3, 2004 7

The STEPWISE Procedure
Model: MODEL1
Dependent Variable: Y

Summary of Stepwise Selection

Step	Variable Entered	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	x2		1	0.3799	0.3799	48.6131	93.10	<.0001
2	x7		2	0.0476	0.4274	35.3731	12.55	0.0005
3	x1		3	0.0376	0.4651	25.3215	10.55	0.0014
4	x4		4	0.0213	0.4864	20.4859	6.19	0.0139
5	x6		5	0.0563	0.5427	4.4650	18.21	<.0001

Model would be:

$$Y=81.98 – 0.03X1 – 0.02X2 -0.01X4 + 0.08X6 – 0.46X7$$