Sample Means

The model provides estimates

$$b_0 = \overline{x}_C = \mathbf{96.5}$$

$$b_1 = \overline{x}_A - \overline{x}_C = \mathbf{9.3}$$

$$b_2 = \overline{x}_B - \overline{x}_C = \mathbf{0.7}$$

So the drug means are:

Drug A = 96.5 + 9.3 = 105.8

Drug B = 96.5 + 0.7 = 97.2

Drug C = 96.5

Module 32: Multiple Regression

This module reviews simple linear regression and then discuses multiple regression. The next module contains several examples.

Module Content

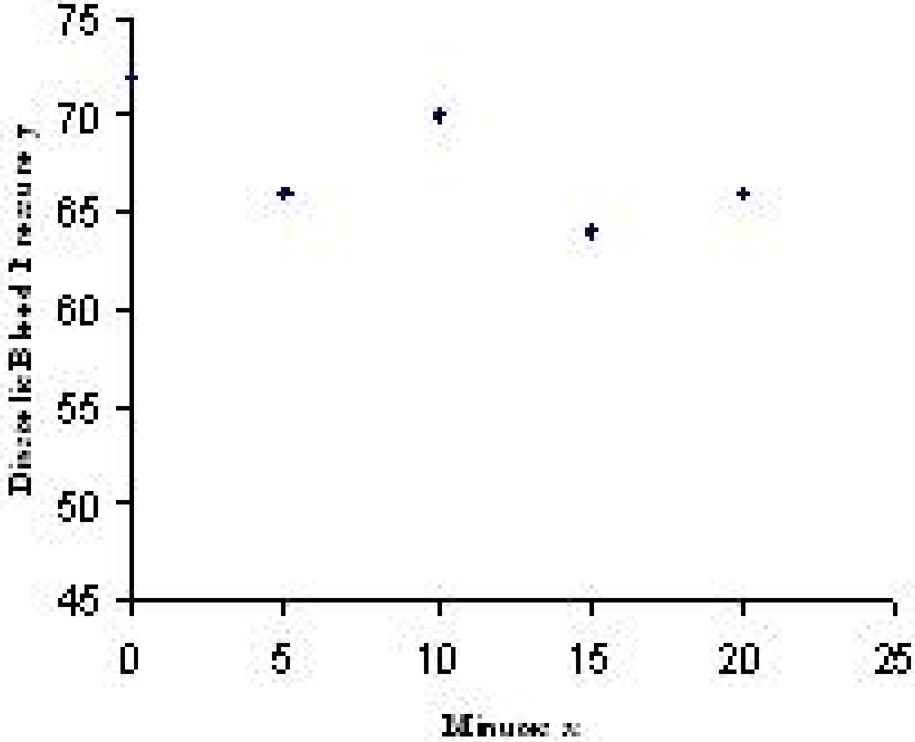
A. Review of Simple Linear Regression

B. Multiple Regression

C. Relationship to ANOVA and Analysis of Covariance

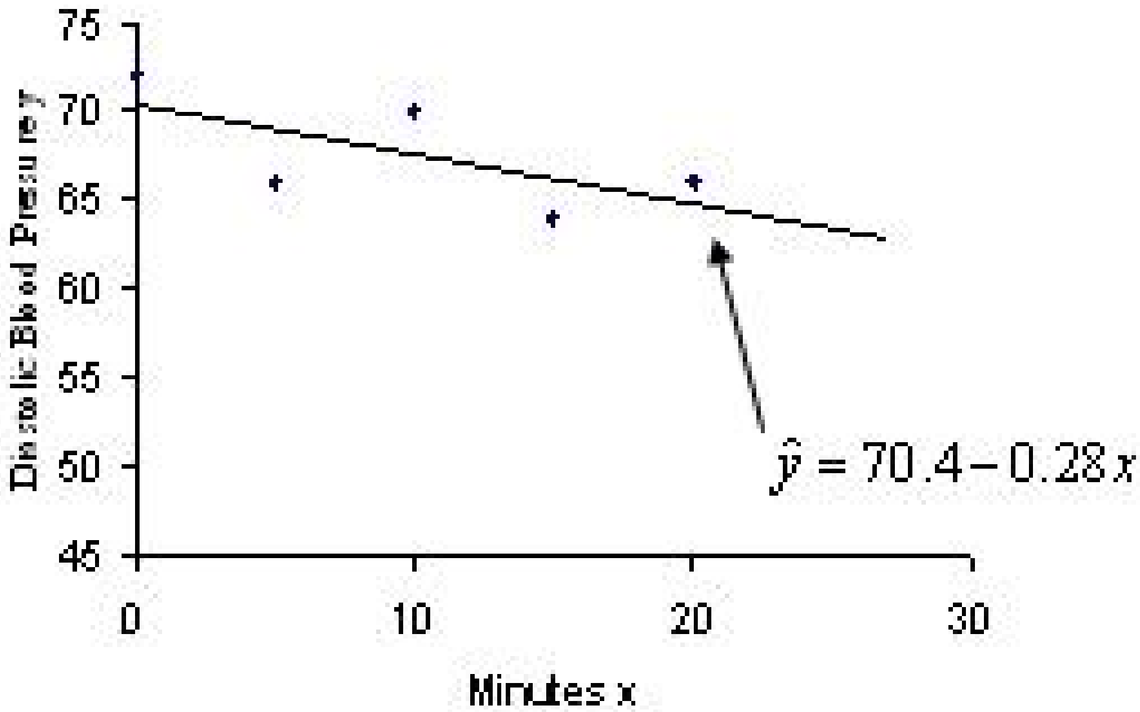
A. Review of Simple Linear Regression

Patient	Time x	DBP y
1	0	72
2	5	66
3	10	70
4	15	64
5	20	66



75
70
65
60
55
50
45
0
5
10
15
20
25

Patient	Time x	DBP y	x^2	y^2	xy
1	0	72	0	5,184	0
2	5	66	25	4,356	330
3	10	70	100	4,900	700
4	15	64	225	4,096	960
5	20	66	400	4,356	1,320
Sum	50	338	750	22,892.00	3,310
Mean	10	67.6			
SD	7.91	3.29			
SS	250	43.20	SS(xy)	-70	
b	-0.28				
a	70.4				


75
70
65
60
55
50
45
Diastolic Blood Pressure y
0
10
20
30
Minutes x
$\hat{y} = 70.4 - 0.28x$

ANOVA

Source	df	SS	MS	F
Regression	1	19.6	19.6	2.49
Residual	3	23.6	7.89	
Total	4	43.2		

SS(Total) = SS(y) = 43.2

SS(Regression) = b SS(xy) = (-0.28)(-70) = 19.6

SS(Residual) = SS(Total) – SS(Regression) = 43.2 – 19.6 = 23.6

F = MS(Regression) / MS(Residual) = 2.49 $F_{0.95}(1,3) = 10.13$

Accept H_0: $\beta = 0$ since $F = 2.49 < F_{0.95}(1,3) = 10.13$

$$R^2 = \frac{SS(Regression)}{SS(Total)} = \frac{19.6}{43.2} = 0.4537$$

B. Multiple Regression

For simple linear regression, we used the formula for a straight line, that is, we used the model:

$$Y = \alpha + \beta x$$

For multiple regression, we include more than one independent variable and for each new independent variable, we need to add a new term to the model, such as:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_k x_k$$

Population Equation

The population equation is:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

the β's are ***coefficients for the independent variables*** in the true or population equation and the x's are the values of the independent variables for the member of the population.

Sample Equation

The sample equation is:

$$\hat{y}_j = b_0 + b_1 x_1 + b_2 x_2 + \ldots + b_k x_k ,$$

where $\hat{y}_j$ represents the regression estimate of the dependent variable for the jth member of the sample and the b's are estimates of the β's.

The Multiple Regression Process

The process involves using data from a sample to obtain an overall expression of the relationship between the dependent variable y and the independent variables, the x's.

This is done in such a manner that the impact of the relationship of the x's collectively and individually on the value of y can be estimated.

The Multiple Regression Concept

Conceptually, multiple regression is a straight forward extension of the simple linear regression procedures.

Simple linear regression is a bivariate situation, that is, it involves two dimensions, one for the dependent variable Y and one for the independent variable x.

Multiple regression is a multivariable situation, with one dependent variable and multiple independent variables.

CARDIA Example

The data in the table on the following slide are:

Dependent Variable

y = BMI

Independent Variables

x_1 = Age in years

x_2 = FFNUM, a measure of fast food usage,

x_3 = Exercise, an exercise intensity score

x_4 = Beers per day

OBS	AGE	BMI	FFNUM	EXERCISE	BEER
1	26	23.2	0	621	3
2	30	30.2	9	201	6
3	32	28.1	17	240	10
4	27	22.7	1	669	5
5	33	28.9	7	1,140	12
6	29	22.4	3	445	9
7	32	23.2	1	710	15
8	33	20.3	0	783	11
9	31	25.6	1	454	0
10	33	21.2	3	432	2
11	26	22.3	5	1,562	13
12	34	23.0	2	697	1
13	33	26.3	4	280	2
14	31	22.2	1	449	5
15	31	19.0	0	689	4
16	27	20.8	2	785	3
17	36	20.9	2	350	7
18	35	36.4	14	48	11
19	31	28.6	11	285	12
20	36	27.5	8	85	5
Total	626	492.8	91	10,925	136
Mean	31.3	24.6	4.6	546.3	6.8

The REG Procedure

Model: MODEL1
Dependent Variable: bmi

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7932 and C(p) = 5.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	273.74877	68.43719	14.38	<.0001
Error	15	71.37923	4.75862		
Corrected Total	19	345.12800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	18.47774	6.45406	39.00436	8.20	0.0119
age	0.08424	0.18931	0.94239	0.20	0.6627
ffood	0.42202	0.13671	45.53955	9.57	0.0074
exercise	-0.00107	0.00170	1.87604	0.39	0.5395
beer	0.32601	0.11518	38.12111	8.01	0.0127

One df for each independent variable in the model

The REG Procedure

Model: MODEL1
Dependent Variable: bmi

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7932 and C(p) = 5.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	273.74877	68.43719	14.38	<.0001
Error	[illegible]	[illegible]	[illegible]		
Corrected Total	19	346.12500			

	Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
b_0	Intercept	18.47774	6.45406	39.00436	8.20	0.0119
b_1	age	0.08424	0.18931	0.94259	0.20	0.6627
b_2	ffnum	0.42202	0.13671	45.53065	9.57	0.0074
b_3	exercise	-0.00107	0.00170	1.87604	0.39	0.5396
b_4	beer	0.32601	0.11518	38.12111	8.01	0.0127

The REG Procedure

Model: MODEL1
Dependent Variable: bmi

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7932 and C(p) = 5.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	273.74877	68.43719	14.38	<.0001
Error	15	71.37923	4.75862		
Corrected Total	19	345.12800			

	Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
b_0	Intercept	18.47774	6.45406	39.00436	8.20	0.0119
b_1	age	0.08424	0.18931	0.94259	0.20	0.6621
b_2	ffnum	0.42202	0.13671	45.53065	9.57	0.0074
b_3	exercise	-0.00107	0.00170	1.87604	0.39	0.5395
b_4	beer	0.32601	0.11518	38.12111	8.01	0.0127

The Multiple Regression Equation

Age

We have,

$b_0 = 10.478, \quad b_1 = 0.084, \quad b_2 = 0.422,$

$b_3 = -0.001, \quad b_4 = 0.326$

So,

$$\hat{y} = 10.478 + 0.084x_1 + 0.422x_2 - 0.001x_3 + 0.326x_4$$

The Multiple Regression Coefficient

The interpretation of the multiple regression coefficient is similar to that for the simple linear regression coefficient, except that the phrase "adjusted for the other terms in the model" should be added.

For example, the coefficient for Age in the model is $b_1 = 0.084$, for which the interpretation is that for every unit increase in age, that is for every one year increase in age, the BMI goes up 0.084 units, *adjusted for the other three terms in the model.*

Global Hypothesis

The first step is to test the global hypothesis:

$$H_0: \ \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$$

$$\text{vs} \quad H_1: \ \beta_1 \neq \beta_2 \neq \beta_3 \neq \beta_4 \neq 0$$

The ANOVA highlighted in the green box at the top of the next slide tests this hypothesis:

$$F = 14.33 > F_{0.95}(4,15) = 3.06,$$

so the hypothesis is rejected. Thus, we have evidence that at least on of the $\beta_i \neq 0$.

The REG Procedure

Model: MODEL1
Dependent Variable: bmi

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7932 and C(p) = 5.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	273.74877	68.43719	14.38	<.0001
Error	15	71.37923	4.75862		
Corrected Total	19	345.12800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	18.47774	6.45406	39.00436	8.20	0.0119
age	0.08424	0.18931	0.94239	0.20	0.6627
ffnum	0.42292	0.13671	45.53958	9.57	0.0074
exercise	-0.00107	0.00170	1.87604	0.39	0.5395
beer	0.32601	0.11518	38.12111	8.01	0.0127

Variation in BMI Explained by Model

The amount of variation in the dependent variable, BMI, explained by its regression relationship with the four independent variables is

R^2 = SS(Model)/SS(Total) = 273.75/345.13

= 0.79 or 79%

Tests for Individual Parameters

If the global hypothesis is rejected, it is then appropriate to examine hypotheses for the individual parameters, such as

H_0: $\beta_1 = 0$ vs H_1: $\beta_1 \neq 0$.

$P = 0.6627$ for this test is greater than $\alpha = 0.05$,

so we accept H_0: $\beta_1 = 0$

Outcome of Individual Parameter Tests

From the ANOVA, we have

$$b_1 = 0.084, \quad P = 0.66$$

$$b_2 = 0.422, \quad P = 0.01$$

$$b_3 = -0.001, \quad P = 0.54$$

$$b_4 = 0.326, \quad P = 0.01$$

So $b_2 = 0.422$ and $b_4 = 0.326$ appear to represent terms that should be explored further.

Stepwise Multiple Regression

Backward elimination

Start with all independent variables, test the global hypothesis and if rejected, eliminate, step by step, those independent variables for which $\beta = 0$.

Forward

Start with a " core " subset of essential variables and add others step by step.

Backward Elimination

The next few slides show the process and steps for the backward elimination procedure.

Global hypothesis

The REG Procedure

Model: MODEL1
Dependent Variable: bmi

Backward Elimination: Step 0

All Variables Entered: R-Square = 0.7932 and C(p) = 5.0000

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	273.74877	68.43719	14.38	<.0001
Error	15	71.37923	4.75862		
Corrected Total	19	345.12800			

	Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
b_0	Intercept	18.47774	6.45406	39.00436	8.20	0.0119
b_1	age	0.08424	0.18931	0.94259	0.20	0.6627
b_2	ffnum	0.42202	0.13671	45.53965	9.57	0.0074
b_3	exercise	-0.00107	0.00170	1.87604	0.39	0.5395
b_4	beer	0.32601	0.11518	38.12111	8.01	0.0127

Backward Elimination: Step 1

Variable age Removed: R-Square = 0.7904 and C(p) = 3.1080

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	272.50638	90.83546	20.12	<.0001
Error	16	[illegible]	[illegible]		
Corrected Total	19	345.12500			

The REG Procedure
Model: MODEL1
Dependent Variable: bmi

Backward Elimination: Step 1

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	21.28788	1.30004	1211.08839	268.13	<.0001
kltrm	0.42963	0.13243	47.57610	10.53	0.0051
exercise	-0.00140	0.00149	4.00750	0.89	0.3604
beer	0.32275	0.11203	37.51501	8.30	0.0109

Bounds on condition number: 1.7880, 14.023

Backward Elimination: Step 2

Variable exercise Removed: R-Square = 0.7788 and C(p) = 2.0402

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	268.73688	134.36844	29.93	<.0001
Error	17	76.39312	4.49395		
Corrected Total	19	345.13000			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	26.29386	0.75579	3237.63255	726.27	<.0001
ffnum	0.46382	0.12493	53.84876	13.35	0.0020
tear	0.33375	0.11105	40.55414	9.03	0.0080

Bounds on condition number: 1.054, 4.4161

All variables left in the model are significant at the 0.0500 level.

The SAS System

Model: MODEL1
Dependent Variable: bmi

Summary of Backward Elimination

Step	Variable Removed	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	age	3	0.0027	0.7904	3.1000	0.20	0.6627
2	exercise	2	0.0116	0.7788	2.0402	0.88	0.3604

Forward Stepwise Regression

The next two slides show the process and steps for Forward Stepwise Regression.

In this procedure, the first independent variable entered into the model is the one with the highest correlation with the dependent variable.

The REG Procedure

Model: MODEL1
Dependent Variable: bmi

Stepwise Selection: Step 1

Variable ffnum Entered: R-Square = 0.6613 and C(p) = 8.5625

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	228.24473	228.24473	35.15	<.0001
Error	18	116.88327	6.49351		
Corrected Total	19	345.12800			

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	21.43827	0.78506	4842.33805	745.72	<.0001
ffnum	0.70368	0.11868	228.24473	35.15	<.0001

Stepwise Selection: Step 2

Variable [illegible] Entered: R-Square = 0.7788 and C(p) = 2.0402

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	268.79888	134.39944	29.93	<.0001
Error	17	76.32912	4.48995		
Corrected Total	19	345.12800			

Model: MODEL1
Dependent Variable: bmi

Stepwise Selection: Step 2

Variable	Parameter Estimate	Standard Error	Type II SS	F Value	Pr > F
Intercept	20.20360	0.75579	3237.09859	720.97	<.0001
ffnum	0.46380	0.12658	59.94878	13.35	0.0020
bmar	0.33375	0.11106	40.55414	9.03	0.0080

Bounds on condition number: 1.654, 6.6161

All variables left in the model are significant at the 0.0500 level.

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

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	ffnum		1	0.6613	0.6613	8.5625	35.15	<.0001
2	bmar		2	0.1175	0.7788	2.0402	9.03	0.0080

C. Relationship to ANOVA and Analysis of Covariance

Multiple regression procedures can be used to analyze data from one-way ANOVA, randomized block, or factorial designs simply by setting up the independent variables properly for the regression analyses. To demonstrate this process, we will work with the one-way ANOVA problem for diastolic blood pressure on the following slide.

Blood pressure measurements for n = 30 children randomly assigned to receive one of three drugs

	Drug		
	A	B	C
	100	104	105
	102	88	112
	96	100	90
	106	98	104
	110	102	96
	110	92	110
	120	96	98
	112	100	86
	112	96	80
	90	96	84
Mean	105.8	97.2	96.5

The ANOVA Approach

$H_0: \mu_A = \mu_B = \mu_C$ vs $H_1: \mu_A \neq \mu_B \neq \mu_C$

ANOVA

Source	df	SS	MS	F
Among	2	536.47	268.23	3.54
Within	27	2043.70	75.69	
Total	29	2580.17		

Reject $H_0: \mu_A = \mu_B = \mu_C$

since $F = 3.54$, is greater than $F_{0.95}(2,27) = 3.35$

Multiple Regression Approach

The multiple regression approach requires a data table such as the following, which means we need to code the drug groups in such a manner that they can be handled as independent variables in the regression model. That is, we need to prepare a data table such as the one below.

Person	y	x_1	x_2
1	100	?	?
2	102	?	?
...	...	...	...
n	84	?	?

Coding the Independent Variables

We can use a coding scheme for the xs to indicate the drug group for each participant. For three drugs we need two xs, with

$x_1 = 1$ if the person received drug A

$= 0$ otherwise

$x_2 = 1$ if the person received drug B

$= 0$ otherwise

Implications of Coding Scheme

The values for x_1 and x_2 for the three drug groups are:

Drug	x_1	x_2
A	1	0
B	0	1
C	0	0

It takes only two xs to code the three drugs.

Use of Coding Scheme

Person 1 has (y = BP) = 100 and receives Drug A

Person 2 has (y = BP) = 102 and receives Drug B

Person 3 has (y = BP) = 105 and receives Drug C

Person	y	x_1	x_2
1	100	1	0
2	102	0	1
3	105	0	0

Indicator Variables

These "indicator" variables provide a mechanism for including categories into analyses using multiple regression techniques. If they are used properly, they can be made to represent complex study designs.

Adding such variables to a multiple regression analysis is readily accomplished. For proper interpretation, one needs to keep in mind how the different variables are defined; otherwise, the process is straight forward multiple regression.

Complete Data Table

Person	y	x_1	x_2
1	100	1	0
2	102	1	0
3	96	1	0
4	106	1	0
5	110	1	0
6	110	1	0
7	120	1	0
8	112	1	0
9	112	1	0
10	90	1	0
11	104	0	1
12	88	0	1
13	100	0	1
14	98	0	1
15	102	0	1
16	92	0	1
17	96	0	1
18	100	0	1
19	96	0	1
20	96	0	1
21	105	0	0
22	112	0	0
23	90	0	0
24	104	0	0
25	96	0	0
26	110	0	0
27	98	0	0
28	86	0	0
29	80	0	0
30	84	0	0

Drug		
A	B	C
100	104	105
102	88	112
96	100	90
106	98	104
110	102	96
110	92	110
120	96	98
112	100	86
112	96	80
90	96	84

Coding Scheme and Means

$x_1 = 1$ if the person received drug A

$= 0$ otherwise

$x_2 = 1$ if the person received drug B

$= 0$ otherwise

$$\beta_0 = \mu_C \qquad b_0 = \bar{x}_C$$

$$\beta_1 = \mu_A - \mu_C \qquad b_1 = \bar{x}_A - \bar{x}_C$$

$$\beta_2 = \mu_B - \mu_C \qquad b_2 = \bar{x}_B - \bar{x}_C$$

$\beta_1 = \beta_2 = 0$ implies $\mu_A = \mu_B = \mu_C$

The SAS System

General Linear Models Procedure

Same as ANOVA

Dependent Variable: Y

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	536.46667	268.23333	3.54	0.0430
Error	27	2043.70000	75.69259		
Corrected Total	29	2580.16667			

R-Square	C.V.	Root MSE	Y Mean
0.207919	8.714673	8.7001	99.833

Source	DF	Type I SS	Mean Square	F Value	Pr > F
X1	1	534.01667	534.01667	7.06	0.0131
X2	1	2.45000	2.45000	0.03	0.8586

Source	DF	Type III SS	Mean Square	F Value	Pr > F
X1	1	432.45000	432.45000	5.71	0.0241
X2	1	2.45000	2.45000	0.03	0.8586

Parameter	Estimate	T for H0: Parameter=0	Pr > |T|	Std Error of Estimate
INTERCEPT	96.50000000	35.08	0.0001	2.75122868
X1	9.30000000	2.39	0.0241	3.89082491
X2	0.70000000	0.18	0.8586	3.89082491