Research Design - - Topic 16
Multiple Regression: Applications

General Overview

Curve Fitting

Mediation analysis

Moderation Analysis

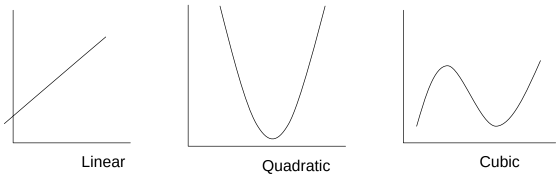
General Overview

As we have seen multiple regression is a very powerful and useful data analytic procedure. It is the basis of MRC analysis as applied to standard analysis of variance designs as well as those that include continuous factors. Earlier, it was mentioned that is was the basis of path analysis. This section describes three further applications:

1. Curve Fitting (polynomial regression)
2. Mediation analysis
3. Moderation analysis

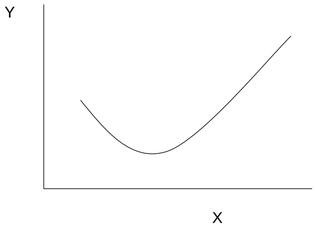
Curve Fitting

The investigation of non-linear functions using multiple regression was introduced by Cohen (1978). It permits researchers to determine the nature of the functional relationship between two variables, Y and X. One procedure on SPSS that performs this function is the Regression – Curve Estimation program. As indicated in the window in slide 5, there are a number of options. We will focus on the one using orthogonal power functions. Below, are examples of three types of power functions.

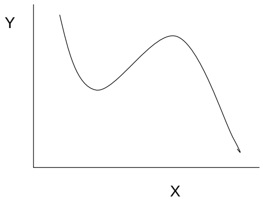


Functions with combinations of components are indicated by significant components for various models. Two examples are shown on slide 4.

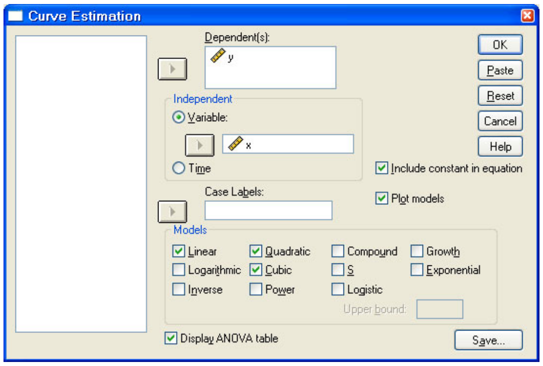
A combination of a positive linear and a positive quadratic function indicated by a significant positive linear trend in model 1 and a significant positive X **2 trend in model 2.



A combination of a negative linear and a negative cubic component indicated by a significant negative linear coefficient in model 1 and a significant negative coefficient for X **3 in model 3.



The Window for the Curve Estimation program is below. To test for linear, quadratic, and cubic components, check the three models indicated, and also check “display ANOVA table”. You should not select any of the other options with these three; the other options deal with other types of functions.



Curve Estimation

Dependent(s): y

Independent

● Variable: x

○ Time

☑ Include constant in equation

Case Labels:

☑ Plot models

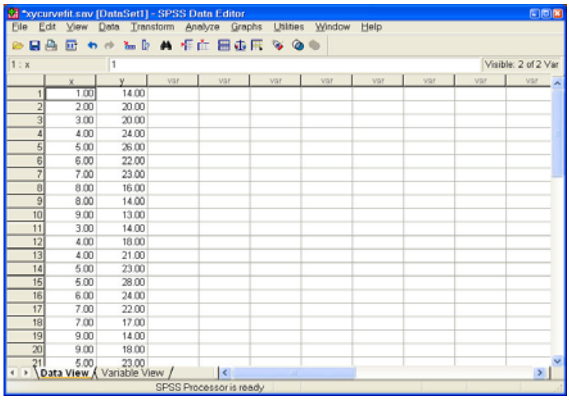
Models

☑ Linear	☑ Quadratic	□ Compound	□ Growth
□ Logarithmic	☑ Cubic	□ S	□ Exponential
□ Inverse	□ Power	□ Logistic	

Upper bound:

☑ Display ANOVA table

Data Set for the Curve fitting examples



*xycurvefit.sav [DataSet1] - SPSS Data Editor

File Edit View Data Transform Analyze Graphs Utilities Window Help

1 : x 1 Visible: 2 of 2 Var

	x	y	var	var	var	var	var	var	var	var
1	1.00	14.00								
2	2.00	20.00								
3	3.00	20.00								
4	4.00	24.00								
5	5.00	26.00								
6	6.00	22.00								
7	7.00	23.00								
8	8.00	16.00								
9	8.00	14.00								
10	9.00	13.00								
11	3.00	14.00								
12	4.00	18.00								
13	4.00	21.00								
14	5.00	23.00								
15	5.00	28.00								
16	6.00	24.00								
17	7.00	22.00								
18	7.00	17.00								
19	9.00	14.00								
20	9.00	18.00								
21	5.00	23.00								

SPSS Processor is ready

Selecting the three models and the ANOVA table option yields information about the F-ratios (not shown here) for the three models and the tests of the components.

Linear component. The coefficient for X is not significant, indicating that there is no linear component.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
x	-.399	.425	-.211	-.939	.360
(Constant)	21.938	2.560		8.571	.000

Quadratic component. X **2 is significant, indicating a significant quadratic component.

Coefficients

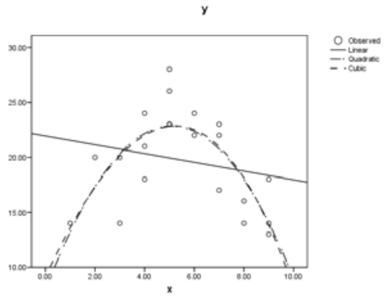
	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
x	5.871	1.352	3.097	4.343	.000
x ** 2	-.577	.121	-3.387	-4.749	.000
(Constant)	7.893	3.437		2.296	.034

Cubic component. X **3 is not significant indicating that there is no cubic component.

Coefficients

	Unstandardized Coefficients		Standardized Coefficients		
	B	Std. Error	Beta	t	Sig.
x	4.843	4.207	2.554	1.151	.266
x ** 2	-.345	.901	-2.029	-.383	.706
x ** 3	-.015	.058	-.837	-.259	.799
(Constant)	9.099	5.843		1.557	.138

The Curve Estimation program plots the three models, even though of course, there is no evidence of a linear or cubic trend here. Note that the quadratic component in these data describes an inverted U shape, which corresponds to the negative sign associated with the regression coefficient for the quadratic component in the previous table. Note too that the tests of significance for the other components (i.e., X in the quadratic table, and X and X **2 in the Cubic table) are not tests of linear or quadratic components. They are also tests of quadratic and cubic components respectively.



Curve Fitting: How it Works

The Curve Estimation program is a simple application of Multiple Regression. To see how it works, we compute two more variables, Xsq and Xcu (X-squared and X-cubed), and perform a hierarchical multiple regression analysis as described in the Syntax file below. Unlike the Curve Estimation program, however, we could add higher order functions if desired (cf., Cohen, 1978).

```
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT y
  /METHOD=ENTER x  /METHOD=ENTER xsq  /METHOD=ENTER xcu  .
```

This would yield a full set of results. I present only the tests of the regression coefficients.

Note these results are identical to those from the Curve Estimation program.
Model 1 shows that there is no evidence of a linear relationship (b = -.399, ns).
Model 2 shows a significant negative quadratic component (b = -.577, p<.0004).
Model 3 shows that there is no evidence of a cubic component (b = -.015, ns).

Coefficients[a]

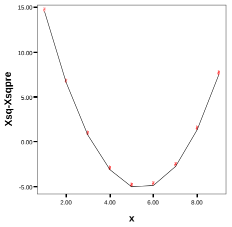
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	21.938	2.560		8.571	.000
	x	-.399	.425	-.211	-.939	.360
2	(Constant)	7.893	3.437		2.296	.034
	x	5.871	1.352	3.097	4.343	.000
	xsq	-.577	.121	-3.387	-4.749	.000
3	(Constant)	9.099	5.843		1.557	.138
	x	4.843	4.207	2.554	1.151	.266
	xsq	-.345	.901	-2.029	-.383	.706
	xcu	-.015	.058	-.837	-.259	.799

a. Dependent Variable: y

Rationale: The residual of Xsq from the value of Xsc predicted from X is determined by calculating the regression of Xsq on X. For these data, that equation is:

Xsqpre = -24.362 + 10.877*X.

The residuals are Xsq – Xsqpre. The plot is as shown.



Thus, if the regression coefficient for Y against xsq in model 2 is positive, it would indicate this type of function. If the regression coefficient is negative, it indicates an inverted U type of function because of the negative semipartial correlation of Y with Xsq.

Similarly, we could investigate the function for Xcu. The regression for Xcupre = $b_0 + b_1$*X + b_2*Xsq, and a plot of the residuals would produce a curve like that in Slide 3.

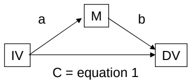
Mediation Analysis

Mediation Analysis is concerned with testing the hypothesis that the relationship between two variables, IV and DV, is due to the influence of a third variable, M. There are more complex forms of mediation and you are referred to a webpage at http://www.davidakenny.net/cm/mediate.htm for further information on the topic.

Baron & Kenny (1986) define a mediator variable as one that accounts for (much of) the relation between an independent variable and a dependent variable. It assumes a causal model where:

1. The independent variable (IV) causes the dependent variable (DV)
2. The independent variable (IV) causes the mediator (M)
3. The mediator (M) causes the dependent variable (DV)

Consider the diagram



C = equation 1

C' = equation 3

Baron & Kenny propose a three step procedure.

1. Regress the DV on the IV $\quad IV \xrightarrow{c} DV$

$$DV = b_{02} + b_{12}IV \qquad where\ b_{12} = c$$

2. Regress the Mediator on the IV $\quad IV \xrightarrow{a} M$

$$M = b_{01} + b_{11}IV \qquad where\ b_{11} = a$$

3. Regress the DV on both the Mediator and the IV

$$DV = b_{03} + b_{13}IV + b_{23}M \qquad where\ b_{13} = c' \quad b_{23} = b$$

$$\begin{tikzcd}
M \arrow[rd, "b"] & \\
& DV \\
IV \arrow[uu, bend left] \arrow[ru, "c'"'] &
\end{tikzcd}$$

For Mediation:

1. b_{11} = a must be significant
2. b_{12} = c must be significant
3. b_{23} = b must be significant

and b_{13} = c' must not be significant or at least significantly smaller than c.

Tests of Significance for Mediation

If c’ is significant, one can still test the difference c – c’ to see if it is significantly smaller than c. It can be shown that c-c’ = a*b from the previous notation. Therefore, a test of c-c’ is a test of the null hypothesis that a*b = 0. There are three forms of a Z test that have been suggested, and the test can be performed on the website shown on slide 12. If the Z test is significant, one concludes that c’ is significantly smaller, and that mediation has been demonstrated. There are minor differences in the tests, and generally the Aroian test is the one most recommended.

The Sobel Test
$$Z = \frac{a*b}{\sqrt{b^2S_a^2 + a^2S_b^2}}$$

The Aroian Test (Goodman I)
$$Z = \frac{a*b}{\sqrt{b^2S_a^2 + a^2S_b^2 + S_a^2S_b^2}}$$

The Goodman II Test
$$Z = \frac{a*b}{\sqrt{b^2S_a^2 + a^2S_b^2 - S_a^2S_b^2}}$$

A Numerical Example of Mediation

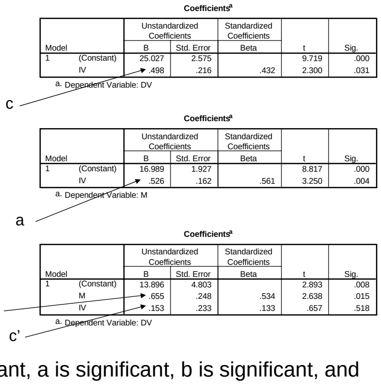
Given the data for variables, IV, DV and M, we could test the mediation model. The correlation matrix for the three variables is:

Correlations

		DV	M	IV
Pearson Correlation	DV	1.000	.609	.432
	M	.609	1.000	.561
	IV	.432	.561	1.000
Sig. (1-tailed)	DV	.	.001	.015
	M	.001	.	.002
	IV	.015	.002	.
N	DV	25	25	25
	M	25	25	25
	IV	25	25	25

IV	DV	M
5	26	21
5	21	15
6	24	16
7	25	18
7	30	18
8	34	27
8	32	27
8	34	24
9	27	15
10	29	23
10	22	24
10	36	25
10	38	27
11	32	25
12	35	22
12	31	23
13	29	24
13	33	22
14	30	25
15	36	21
16	29	25
17	38	28
17	28	24
18	36	25
19	30	28

Performing the three multiple regression runs produces the regression information on slide 16.



Step 1. Regression of DV on IV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	25.027	2.575		9.719	.000
	IV	.498	.216	.432	2.300	.031

a. Dependent Variable: DV

Step 2. Regression of M on IV

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	16.989	1.927		8.817	.000
	IV	.526	.162	.561	3.250	.004

a. Dependent Variable: M

Step 3. Regression of DV on M and IV

Coefficients[a]

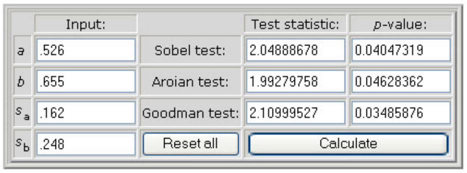
Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	13.896	4.803		2.893	.008
	M	.655	.248	.534	2.638	.015
	IV	.153	.233	.133	.657	.518

a. Dependent Variable: DV

Note c is significant, a is significant, b is significant, and c’ is not significant. Also note that c-c ‘ = .498-.153 = .345 and that a*b = (.526)(.655) = .345, as stated earlier.

Applying the Tests from the Webpage shown on Slide 12

Inputting the values of a and b, and their standard errors from the computer output yields the following values. In each case, the Z value is significant, so we can conclude that M is acting as a mediator. That is, it is meaningful to conclude that the relation between the IV and the DV is due to the mediating effects of M.



	Input:		Test statistic:	*p*-value:
a	.526	Sobel test:	2.04888678	0.04047319
b	.655	Aroian test:	1.99279758	0.04628362
s_a	.162	Goodman test:	2.10999527	0.03485876
s_b	.248	Reset all	Calculate	

Moderation Analysis

Moderation is synonymous with interaction. A **moderator variable** is a variable that affects the direction and/or strength of the relation between a predictor variable and a criterion. Thus, a moderator is a third variable that changes the relation between the criterion and a predictor. Using analysis of variance terminology a variable is a moderator if it interacts with the other variable to effect values on the criterion (cf., Cohen, 1978).

Given three variables X1 = the criterion
X2 = predictor (1)
X3 = predictor (2)

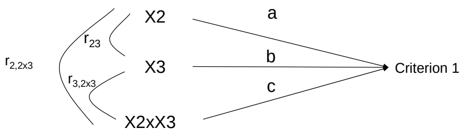
We can compute a fourth variable, X2X3 as the product of X2 and X3. We can then compute two multiple regressions:

$$R^2_{1.2,3,2x3} \ and \ R^2_{1.2,3}$$

Then if $R^2_{1.2,3,2x3} - R^2_{1.2,3}$

is significant, variable X3 moderates the relation between variables X1 and X2, and variable X2 moderates the relation between variables X1 and X3.

This can be shown as a path diagram



If c is significant, variable 3 is said to be a moderator. It is not necessary for any of the other coefficients to be significant. In fact, Cohen notes that in the presence of a product term, the regression coefficients for the non-product terms (a and b) are arbitrary nonsense (i.e., they have no interpretative value).

It is desirable that the moderator be uncorrelated with the predictor, though this simply makes the interpretation of the interaction clearer. As r_{23} increases, the interaction gets less clear.

A Numerical Example of Moderation

Input Data

Z = Criterion
X = Predictor 1
Y = predictor 2
XY = product term

Z	X	Y	XY
22.00	10.00	1.00	10.00
20.00	11.00	2.00	22.00
15.00	12.00	3.00	36.00
14.00	12.00	3.00	36.00
15.00	13.00	4.00	52.00
16.00	13.00	4.00	52.00
14.00	13.00	4.00	52.00
16.00	14.00	5.00	70.00
17.00	14.00	5.00	70.00
15.00	14.00	5.00	70.00
22.00	20.00	6.00	120.00
21.00	19.00	6.00	114.00
20.00	18.00	6.00	108.00
15.00	12.00	6.00	72.00
17.00	11.00	6.00	66.00
14.00	16.00	5.00	80.00
18.00	16.00	5.00	80.00
17.00	16.00	5.00	80.00
15.00	17.00	4.00	68.00
14.00	17.00	4.00	68.00
18.00	17.00	4.00	68.00
20.00	18.00	3.00	54.00
17.00	18.00	3.00	54.00
13.00	19.00	2.00	38.00
13.00	20.00	1.00	20.00

To test for moderation, compute the multiple regression of Z on X,Y, and XY

Following is the table of regression coefficients for this run. R = .729

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	32.572	4.391		7.417	.000
	X	-1.109	.279	-1.221	-3.971	.001
	Y	-4.739	1.094	-2.617	-4.332	.000
	XY	.326	.069	3.232	4.744	.000

a. Dependent Variable: Z

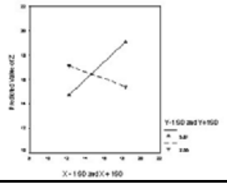
These coefficients indicate that there is a significant interaction between X and Y in the prediction of Z, because the regression coefficient for XY = .326 is significant at the .0004 level. The regression coefficients for X and Y have no particular meaning. **In fact, Cohen (1978, p. 861) states they are arbitrary nonsense.**

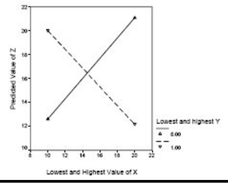
Viewing the interaction

It is common practice to view the interaction by solving the regression equation for high and low values of X and Y, and then plotting the results. Two plots are shown. Given mean x = 15.2, s.d. = 3.04, and mean y = 4.08, s.d. = 1.53, low and high values could be the means ± 1 sd, or alternatively, the highest and lowest values:

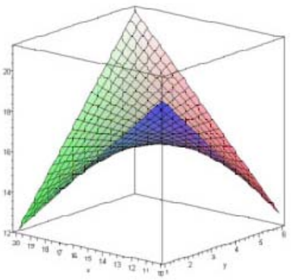
	Low	High	Low	High
X	11.16	18.24	10	20
Y	2.55	5.61	1	6

$$Z = 32.572 - 1.109 * X - 4.739 * Y + .326 * XY$$





Obviously, the interpretation can be slightly different, depending on what is chosen as high or low, and, of course, many such figures could be plotted. In fact, the complete plot considering X and Y as continuous values is:



Centering

It is often recommended that the analysis be done after X and Y have been centered by subtracting the mean from each value, and forming the product term from the centered values. This changes the values of the regression coefficients, but the test of significance of the interaction is identical to what it was with the uncentered data, as demonstrated in the following table of regression coefficients. As before, the multiple correlation is .729.

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	16.595	.405		40.976	.000
	cx	.221	.141	.243	1.567	.132
	cy	.216	.272	.119	.795	.435
	cxcy	.326	.069	.734	4.744	.000

a. Dependent Variable: z

Thus, as far as the interaction is concerned, it doesn't matter if X and Y are centred. Generally, however, they are.

Issues

As indicated earlier, the regression coefficients for the continuous variables have no interpretative value because they are residualized from the product term as well as each other. If one is interested in these values, one must use a hierarchical procedure, and the nature of the hierarchy will influence the interpretation. There are three approaches:

1. Enter X followed by Y
2. Enter Y followed by X
3. Enter X and Y in the first step.

The interpretation differs with the approach used.

When, one variable is entered first, a significant regression coefficient for that variable indicates that there is a direct relationship between that variable and the criterion.

A significant regression coefficient for the second variable on step 2 indicates that the second variable adds significantly to the prediction accounted for by the first variable on that step.

If both variables are added on the first step, the regression coefficients reflect the extent to which each variable contributes relative to the other.

Adding X first

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	15.803	2.926		5.400	.000
	x	.060	.189	.066	.320	.752
2	(Constant)	14.816	3.230		4.587	.000
	x	.048	.191	.053	.250	.805
	y	.288	.382	.159	.755	.458
3	(Constant)	32.572	4.391		7.417	.000
	x	-1.109	.279	-1.221	-3.971	.001
	y	-4.739	1.094	-2.617	-4.332	.000
	xy	.326	.069	3.232	4.744	.000

a. Dependent Variable: z

Results obtained if both X and Y were added in one step

Adding Y first

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	15.510	1.618		9.585	.000
	y	.297	.372	.164	.796	.434
2	(Constant)	14.816	3.230		4.587	.000
	y	.288	.382	.159	.755	.458
	x	.048	.191	.053	.250	.805
3	(Constant)	32.572	4.391		7.417	.000
	y	-4.739	1.094	-2.617	-4.332	.000
	x	-1.109	.279	-1.221	-3.971	.001
	xy	.326	.069	3.232	4.744	.000

a. Dependent Variable: z

Except for the Constants, the same values would be obtained in steps 1 and 2 if centred values were used.

Calculating Simple Slopes

A simple slope is the slope of the criterion against one of the predictors for given values of the other, for example the slope of Z against X for different values of Y. The general equation is:

$$Z = b_0 + b_1 X + b_2 Y + b_3 X * Y$$

which can be reordered to express Z as a function of X as follows:

$$Z = \underbrace{(b_0 + b_2 Y)}_{\text{Intercept}} + \underbrace{(b_1 + b_3 Y)}_{\text{Slope}} * X$$

Thus:

$$Z = (32.572 - 4.739 * Y) + (-1.109 + .326 * Y) X$$

For Low Y = 2.55 intercept = 20.488 slope = -.278

For High Y = 5.61 intercept = 5.986 slope = .720

Tests of Significance for the Simple Slopes

These tests make use of estimates of the standard error of the slope by calculating the standard deviation of the aggregate (b1+b3Y in this example) using covariances of the regression coefficients that can be obtained from the multiple regression analysis.

$$\text{Given} \quad t = \frac{\text{slope}}{\text{S.E.}}$$

$$S.E. = \sqrt{SE_{b1}^2 + SE_{b3}^2 Y^2 + 2\,\mathrm{cov}_{b1,b3}\,Y}$$

Coefficient Correlations[a]

Model			xy	x	y
1	Correlations	xy	1.000	-.873	-.969
		x	-.873	1.000	.835
		y	-.969	.835	1.000
	Covariances	xy	.005	-.017	-.073
		x	-.017	.078	.255
		y	-.073	.255	1.197

a. Dependent Variable: z

Computing Standard Errors for Low and High Values of Y

For Low Y = 2.55

$$S.E. = \sqrt{.078 + (.005)(2.55^2) + 2(-.017)(2.55)} = \sqrt{.024} = .155$$

For high Y = 5.61

$$S.E. = \sqrt{.078 + (.005)(5.61^2) + 2(-.017)(5.61)} = \sqrt{.044} = .210$$

Note: This example makes use of the non-centered data. Similar computations could be performed using the centered data. The slopes and the test of significance would be the same.

t-tests of the significance of each of the slopes
df = N-p-1=25-3-1=21

For Low Y $t = \frac{-.278}{.155} = -1.79$

For High Y $t = \frac{.720}{.210} = 3.43 p < .01$

Thus, for a low value of Y (i.e., 1 standard deviation below the mean), there is a slight negative (but not significant) slope of Z on X, while for a high value of Y there is a significant positive slope. Of course, such tests could be performed on any values of Y (or X). It should be noted that although Cohen, Cohen, West, & Aiken (2003) believe such tests are meaningful, they state "**There exists no test of significance of difference between simple slopes computed at single values (points) along a continuum…" (**p.280).

A Cautionary Note

McClelland and Judd (1993) demonstrated that tests of interactions in field studies often have less than 20% of the efficiency of controlled experiments, and they discuss problems associated with tests of moderator effects as well as curve fitting. They conclude (pp. 387-388):

"Our analysis of the relative superiority of experimental designs for detecting interactions implies that unless researchers can select, oversample, or control the levels of the predictor variables, detection of reliable interactions or quadratic effects explaining an appreciable proportion of the variation of the dependent variable will be difficult. This does not mean that researchers should not seek interactions in such conditions; however, they should be aware that the odds are against them."

References

Baron, R.M. & Kenny, D.A. (1986). The moderator-mediator variable distinction in social psychological research: conceptual, strategic and statistical considerations. *Journal of Personality and Social Psychology, 51,* 1173-1182.

Cohen, J. (1978). Partialed products *are* interactions, partialed powers *are* curve components. *Psychological Bulletin, 85*, 858-866.

Cohen, J., Cohen, P., West, S. G. & Aiken, L.S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences (third edition).* Mahwah, NJ: Lawrence Erlbaum.

McClelland, G. H. & Judd, C. M. (1993). Statistical difficulties of detecting interactions and moderator effects. *Psychological Bulletin, 114*, 376-390.