Tomi Seppälä and Robert Plante
Purdue University

Multiple regression analysis

Supplementary course material

MULTIPLE REGRESSION

* SOME KNOWLEDGE OF MATRIX ALGEBRA IS REQUIRED.

A. SIMPLE LINEAR REGRESSION MODEL IN MATRIX TERMS

WE WILL USE MATRIX ALGEBRA IN OUR DISCUSSION OF MULTIPLE REGRESSION. TO MOTIVATE THE USE OF MATRIX ALGEBRA LET'S CONSIDER HOW WE COULD REPRESENT OUR SIMPLE LINEAR REGRESSION MODEL USING VECTORS AND MATRICES.

REMEMBER THAT OUR MODEL FOR EACH OBSERVATION IS,

$$Y_i = \alpha + \beta X_i + \varepsilon_i$$

WITHOUT CHANGING THE NATURE OF THIS EQUATION WE CAN REWRITE IT AS,

$$Y_i = \alpha(1) + \beta X_i + \varepsilon_i$$

NOW LET'S DEFINE A SET OF VECTORS AND MATRICES THAT CAN BE USED TO REPRESENT ALL OF OUR OBSERVATIONS. GIVEN A RANDOM SAMPLE OF n OBSERVATIONS, LET

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}$$

THIS IS THE COLUMN VECTOR OF OUR OBSERVATIONS ON THE DEPENDENT OR RESPONSE VARIABLE.

LET,

$$X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

THIS IS THE n x 2 MATRIX OF MY INDEPENDENT VARIABLE AND THE CONSTANT TERM.

LET,

$$\beta = \begin{bmatrix} \alpha \\ \beta \end{bmatrix}$$

THIS IS THE 2 x 1 COLUMN VECTOR OF MY LINEAR REGRESSION COEFFICIENTS.

FINALLY LET,

$$\varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

THIS IS THE COLUMN VECTOR OF THE ERROR TERMS IN OUR MODEL FOR Y_i.

NOW WE CAN WRITE (INCLUDE) ALL OF OUR OBSERVATIONS IN ONE EQUATION USING MATRIX TERMS AS,

$$Y = X\beta + \varepsilon$$

WHICH REPRESENTS,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

OR AFTER MULTIPLICATION AND ADDITION OF MATRICES,

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} \alpha + \beta X_1 + \varepsilon_1 \\ \alpha + \beta X_2 + \varepsilon_2 \\ \vdots \\ \alpha + \beta X_n + \varepsilon_n \end{bmatrix}$$

WHICH YIELDS OUR ORIGINAL SET OF EQUATIONS,

$$Y_1 = \alpha + \beta X_1 + \varepsilon_1$$

$$Y_2 = \alpha + \beta X_2 + \varepsilon_2$$

$$\vdots$$

$$Y_n = \alpha + \beta X_n + \varepsilon_n$$

CONSEQUENTLY WE NOW HAVE REPRESENTED OUR MODEL VERY COMPACTLY AS

$$Y = X\beta + \varepsilon$$

TO OBTAIN OUR LEAST SQUARES ESTIMATES OF α & β WE CAN USE MATRIX ALGEBRA AND CALCULUS TO MINIMIZE THE SUM OF THE SQUARED ERROR TERMS.

REMEMBER THAT THE SUM OF THE SQUARED ERRORS IS JUST:

$$\sum_{i=1}^{n} \varepsilon_i^2$$

WE CAN OBTAIN THIS BY MULTIPLYING

$$\varepsilon'\varepsilon$$

WHERE ε' IS THE TRANSPOSE OF ε. HOWEVER,

$$\varepsilon = Y - X\beta$$

THUS,

$$\varepsilon'\varepsilon = (Y - X\beta)'(Y - X\beta)$$

AFTER MULTIPLICATION THE EXPRESSION ON THE RIGHT IS JUST,

$$\sum_{i=1}^{n} [Y_i - \alpha - \beta X_i]^2$$

WHICH IS EXACTLY WHAT WE HAVE MINIMIZED FOR α & β PREVIOUSLY TO OBTAIN a AND b.

LET,

$$b = \begin{bmatrix} a \\ b \end{bmatrix}$$

THIS IS OUR ESTIMATES OF α AND β. BY MINIMIZING $\varepsilon'\varepsilon$ WE CAN SHOW THAT,

$$b = (X'X)^{-1}(X'Y)$$

WHERE $(\ )^{-1}$ REPRESENTS THE INVERSE OF A MATRIX.

AFTER SOME MULTIPLICATION WE WOULD OBTAIN,

$$b = \begin{bmatrix} \bar{Y} - b\bar{X} \\ \dfrac{\sum_{i=1}^{n} (X_i-\bar{X})(Y_i-\bar{Y})}{\sum_{i=1}^{n} (X_i-\bar{X})^2} \end{bmatrix} = \begin{bmatrix} a \\ b \end{bmatrix}$$

WHICH ARE EXACTLY THE LS ESTIMATES WE OBTAINED FOR α AND β PREVIOUSLY.

IN ADDITION, IT CAN BE SHOWN THAT,

$$SST = Y'Y - n\bar{Y}^2$$

$$SSR = b'X'Y - n\bar{Y}^2$$

AND,

$$SSE = Y'Y - b'X'Y$$

AND REMEMBER, (AS IS EASILY SEEN)

$$SST = SSR + SSE.$$

IT CAN ALSO BE SHOWN THAT THE VARIANCE-COVARIANCE MATRIX OF b,

$$\sigma^2(b) = \begin{bmatrix} \sigma_a^2 & \sigma_{ab} \\ \sigma_{ab} & \sigma_b^2 \end{bmatrix}$$

IS OBTAINED BY

$$\sigma^2(b) = \sigma^2(X'X)^{-1}$$

WHERE σ^2 IS THE VARIANCE OF OUR ERROR TERM ε_i. Thus,

$$s^2(b) = MSE\ (X'X)^{-1}$$

$$SS_X = \Sigma(X_i - \bar{X})^2$$

$$SSE = \Sigma(Y_i - \hat{Y}_i)^2$$

$$MSE = \frac{SSE}{n-2}$$

AND AFTER MULTIPLICATION WE OBTAIN,

$$s^2(b) = \begin{bmatrix} MSE\ [\frac{1}{n} + \frac{\bar{X}^2}{SS_X}] & \frac{-\bar{X}\ MSE}{SS_X} \\ \frac{-\bar{X}\ MSE}{SS_X} & \frac{MSE}{SS_X} \end{bmatrix}$$

AFTER ALL OF THIS, ONE MAY FEEL THAT THERE IS A GREAT DEAL OF COMPUTATIONAL EFFORT INVOLVED WITH MATRIX ALGEBRA (INVERSION, ETC.). YOU WOULD NOT BE ALONE IN THIS FEELING. HOWEVER, COMPUTER PACKAGES HAVE BEEN DEVELOPED TO ALLEVIATE THIS TEDIOUS ENDEAVOR. NOT ALL PACKAGES GIVE RESULTS IN EXACTLY THE SAME MANNER. THUS, YOU WILL STILL NEED A GOOD METHODOLOGICAL BASIS TO INTERPRET THE RESULTS. THIS IS NOT THE ONLY REASON, BUT IS MOTIVATION ENOUGH FOR LEARNING THE ESSENTIALS OF THE MATRIX APPROACH TO REGRESSION.

B. USING AN ANOVA TABLE TO PERFORM TESTS IN REGRESSION ANALYSIS

WHEN WE PERFORM TESTS IN MULTIPLE REGRESSION, THE t DISTRIBUTION WILL NOT BE AS USEFUL AS IN SIMPLE REGRESSION. CONSEQUENTLY WE WILL CONSIDER ANOTHER APPROACH USING,

$$SST = SSR + SSE.$$

NOTICE THAT THIS BREAKS DOWN OUR INITIAL TOTAL VARIATION INTO EXPLAINED (SSR) AND UNEXPLAINED (SSE) PORTIONS. WE CAN AND WILL CONSTRUCT A TABLE USING THE ABOVE AS FOLLOWS,

** (THIS IS CALLED AN ANALYSIS OF VARIANCE (ANOVA) TABLE.)

SOURCE OF VARIATION	SUM OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE
REGRESSION	SSR	1	SSR/1
ERROR	SSE	n-2	SSE/(n-2
TOTAL	SST	n-1	

WHERE,

MSR = SSR/1 (MEAN SQUARE REGRESSION)

AND,

MSE = SSE/(n-2) (MEAN SQUARE ERROR)

IF WE TEST,

$H_0: \beta = 0$

$H_1: \beta \neq 0$

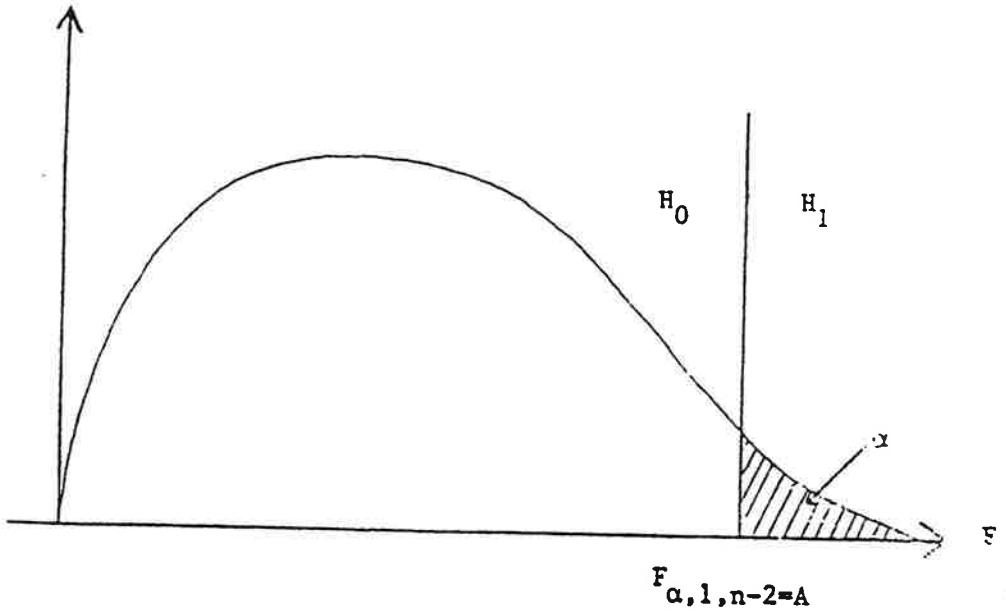
THAT IS, IF WE TEST TO SEE IF THE REGRESSION LINE IS SIGNIFICANT (SHOULD BE RETAINED FOR FORECASTING, ETC.), THEN WHEN H_0 IS TRUE, THE RATIO OF,

$$\frac{MSR}{MSE}$$

IS AN F RANDOM VARIABLE WITH PARAMETERS (DEGREES OF FREEDOM FOR SSR, DEGREES OF FREEDOM FOR SSE), SUCH THAT

$$F = \frac{MSR}{MSE} \sim F_{1,n-2}$$

GRAPHICALLY, FOR AN α LEVEL TEST,



CAN YOU EXPLAIN WHY OUR DECISION RULE IS,

IF $F \leq A$ REJECT H_1
OTHERWISE ACCEPT H_1 ?

<u>EXAMPLE</u>,

SOURCE	SS	DIF.	MS
R	124.25	1	124.25
E	79.75	5	15.95
T	204	6	

TO TEST,

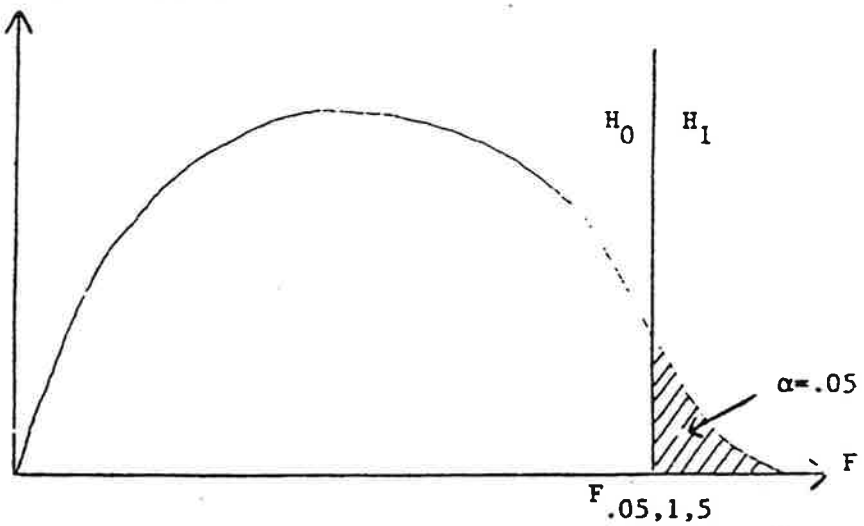
$H_0: \beta = 0$

$H_1: \beta \neq 0$

WITH A LEVEL OF SIGNIFICANCE OF .05,

THEN

$$F = \frac{MSR}{MSE} \sim F_{1,n-2}$$



$$F_{.05,1,5} = 6.61$$

DECISION RULE.

IF $F \leq 6.61$, REJECT H_1

OTHERWISE ACCEPT H_1.

Give p-value!

$$F = \frac{MSR}{MSE} = \frac{124.25}{15.95} \cong 7.79$$

DECISION,

ACCEPT H_1 (RETAIN ESTIMATED REGRESSION LINE)

SINCE $F > 6.61$.

<u>NOTE</u>:

$$F_{1,r} = (t_r)^2$$

C. MULTIPLE REGRESSION ANALYSIS

MOST "USEFUL" REGRESSION MODELS NEED MORE THAN ONE INDEPENDENT VARIABLE. SUPPOSE WE USE p INDEPENDENT VARIABLES. OUR MULTIPLE REGRESSION MODEL FOR EACH Y_i OBSERVATION IS THEN,

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \varepsilon_i$$

CONSIDER, FOR EXAMPLE, THE PRODUCTION FORECASTING CASE IN M&W, THE DATA FOR THAT CASE IS,

TABLE 1
DATA FROM 20 DAYS OF PRODUCTION UNDER CONTROLLED CONDITIONS

DAY	PRODUCTION (Y)	WIDTH (X_1)	DENSITY (X_2)	STRENGTH (X_3)
1	763	19.8	128	85
2	650	20.9	110	72
3	554	15.1	95	62
4	742	19.8	123	82
5	470	21.4	77	52
6	651	19.5	107	72
7	756	25.2	123	84
8	563	26.2	95	63
9	681	26.8	116	76
10	579	28.8	100	64
11	716	22.9	119	80
12	650	24.2	107	72
13	761	24.9	125	85
14	549	25.6	89	61
15	641	24.7	108	71
16	606	26.2	103	67
17	696	21.0	119	77
18	795	29.4	133	88
19	582	21.6	96	65
20	559	20.0	91	62

FOR THIS EXAMPLE WE HAVE p=3 INDEPENDENT VARIABLES AND OUR MODEL IS,

$$Y_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i.$$

THESE CAN BE REPRESENTED IN MATRIX NOTATION AS

$$Y = X\beta + \epsilon$$

WHERE,

$$X = \begin{bmatrix} 1 & X_{11} & X_{12} & X_{13} \\ 1 & X_{21} & X_{22} & X_{23} \\ & & \vdots & \\ 1 & X_{n1} & X_{n2} & X_{n3} \end{bmatrix}$$

AND,

$$\beta = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}$$

THE LEAST SQUARES ESTIMATES OF α, β_1, β_2, β_3 ARE,

$$b = \begin{bmatrix} a \\ b_1 \\ b_2 \\ b_3 \end{bmatrix}$$

WHERE,

$$b = (X'X)^{-1}(X'Y).$$

USING A COMPUTER PACKAGE WE OBTAIN

$$b = \begin{bmatrix} -1.051 \\ 0.072 \\ 0.376 \\ 8.429 \end{bmatrix}$$

THUS $\hat{Y}$, OUR ESTIMATED MODEL IS,

$$\hat{Y} = -1.051 + .072X_1 + .376X_2 + 8.429X_3$$

THE ANOVA TABLE FOR A MULTIPLE REGRESSION PROBLEM IS SIMILAR TO THAT FOR A SIMPLE REGRESSION PROBLEM. IT DIFFERS ONLY IN THE DEGREES OF FREEDOM COLUMN.

ANOVA

SOURCE	SS	DF	MS
R	SSR	p	SSR/p
E	SSE	n-p-1	SSE/(n-p-1)
T	SST	n-1	

AND,

$$MSR = SSR/p$$

$$MSE = SSE/(n-p-1)$$

THUS, IF WE WISH TO TEST FOR RETENTION OF OUR REGRESSION LINE WE HAVE,

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

$$H_1: \text{NOT ALL } \beta_j = 0$$

WHEN H_0 IS TRUE,

$$F = \frac{MSR}{MSE} \sim F_{p,n-p-1} \cdot$$

EXAMPLE,

THE ANOVA TABLE FOR OUR PROBLEM IS,

COMPUTE,

$$SST = Y'Y - n\bar{Y}$$

$$SSR = b'X'Y - n\bar{Y}^2$$

$$SSE = Y'Y - b'X'Y$$

AGAIN, USING A COMPUTER PACKAGE,

SST = 152,535.58

SSR = 152,386.00

SSE = 149.58

ANOVA

Source	SS	DF	MS
R	152,386.00	3	50,795.33
E	149.58	16	9.349
T	152,535.58	19	

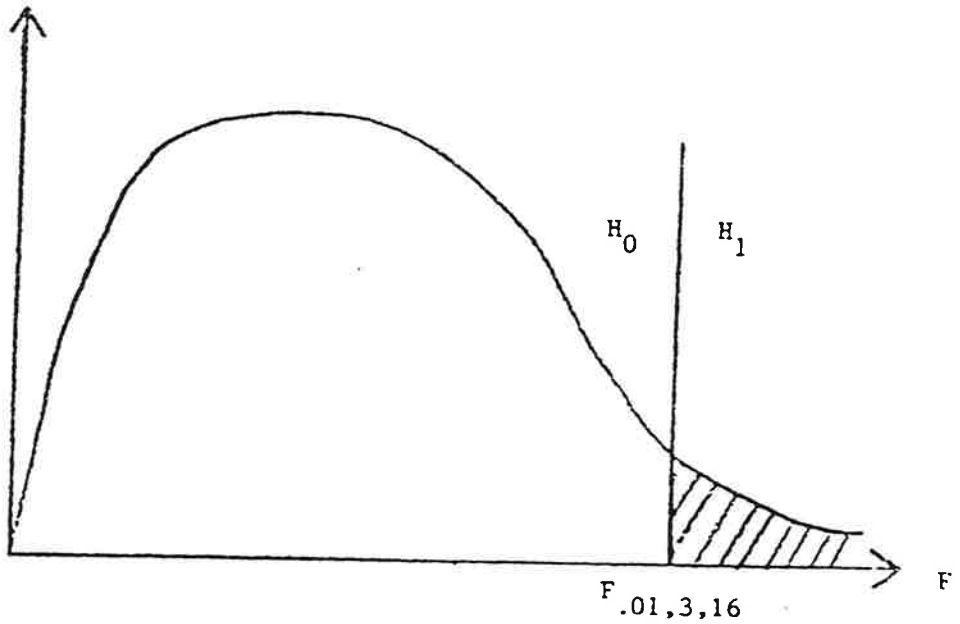
THEN,

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

H_1: NOT ALL $\beta_j = 0$.

$$F = \frac{MSR}{MSE} \sim F_{3,16}$$

TEST WITH $\alpha = .01$.



$F_{.01,3,16} = 5.29$

DECISION RULE:

IF $F \leq 5.29$, REJECT H_1 OTHERWISE, ACCEPT H_1.

DECISION:

SINCE

$$F = \frac{MSR}{MSE} = \frac{50,795.33}{9.349} = 5,433.2, \text{ ACCEPT } H_1.$$

OBVIOUSLY THIS IS MUCH GREATER THAN 5.29 AND THUS WE CONCLUDE THAT THERE IS A SIGNIFICANT RELATIONSHIP BETWEEN PRODUCTION OUTPUT AND OUR SET OF INDEPENDENT VARIABLES.

1. <u>COEFFICIENT OF MULTIPLE DETERMINATION</u>

SIMILAR TO THE R^2 VALUE FOR SIMPLE REGRESSION, THE R^2 VALUE FOR MULTIPLE REGRESSION INDICATES THE PORTION OF TOTAL VARIANCE EXPLAINED BY OUR REGRESSION MODEL.

REMEMBER,

$$R^2 = \frac{SSR}{SST}$$

OR,

$$R^2 = 1 - \frac{SSE}{SST}$$

FOR OUR EXAMPLE,

$$R^2 = \frac{152,386}{152,535} \cong 0.9990$$

2. THE ADJUSTED R^2

IT SO HAPPENS, THAT, AS WE INCREASE THE NUMBER OF INDEPENDENT VARIABLES IN OUR REGRESSION MODEL R^2 ALSO INCREASES, EVEN THOUGH THE ADDED INDEPENDENT VARIABLES OR VARIABLE ARE/IS NOT RELATED TO OUR DEPENDENT VARIABLE. TO ACCOUNT FOR THIS WE CAN USE THE ADJUSTED R^2 WHICH IS,

$$R_a^2 = 1 - \frac{SSE}{SST}\left[\frac{n-1}{n-p}\right].$$

FOR OUR EXAMPLE,

$$R_a^2 = 1 - \frac{149.58}{152,535.58}\left[\frac{19}{17}\right] \cong 0.9988$$

3. TEST FOR AN INDIVIDUAL COEFFICIENT β_j

EVEN THOUGH THE SET OF INDEPENDENT VARIABLES IS CORRELATED WITH OUR DEPENDENT VARIABLE, IT MAY BE TRUE THAT ONE OR MORE OF THE VARIABLES ARE NOT.

FOR EXAMPLE CONSIDER THE FOLLOWING CORRELATION MATRIX (OBTAINED VIA COMPUTER PACKAGE OUTPUT),

	Y	X_1	X_2	X_3
Y	1.000	0.171	0.991	0.999
X_1	0.171	1.000	0.181	0.167
X_2	0.991	0.181	1.000	0.990
X_3	0.999	0.167	0.990	1.000

NOTICE THAT X_2 AND X_3 ARE HIGHLY CORRELATED WITH Y, NAMELY 0.991 AND 0.999. HOWEVER, THE CORRELATION BETWEEN Y AND X_1 IS ONLY 0.171. THIS IMPLIES THAT THE WIDTH OF A STAINLESS STEEL SHEET DOES NOT INFLUENCE PRODUCTION LEVEL

THERE ARE SOME OTHER INTERESTING OBSERVATIONS (i.e., X_2 AND X_3 ARE STRONGLY CORRELATED AT 0.990) IN THIS CORRELATION MATRIX WHICH WE WILL DISCUSS LATER.

IF WE WISH TO TEST WHETHER OR NOT A SINGLE β_j COEFFICIENT IS SIGNIFICANT WE CAN USE THE FOLLOWING,

$$H_0: \beta_1 = 0$$

$$H_1: \beta_1 \neq 0$$

THE RELEVANT STATISTIC FOR THIS TEST IS,

$$T = \frac{b_1}{S(b_1)}$$

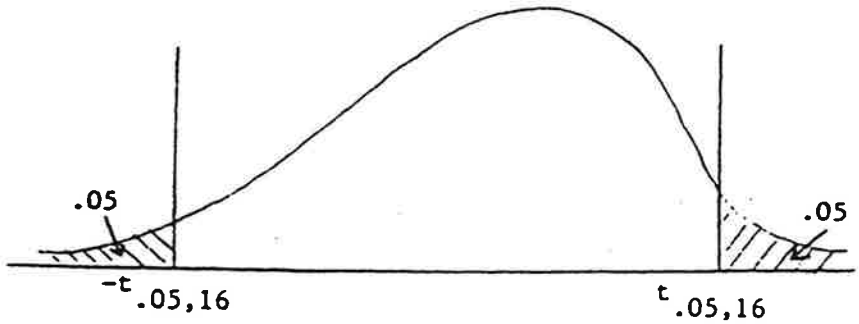
WHICH, WHEN H_0 IS TRUE, HAS THE FOLLOWING DISTRIBUTION,

$$T = \frac{b_1}{S(b_1)} \sim t_{n-p-1}$$

WHERE $S(b_1)$ CAN BE OBTAINED FROM THE VARIANCE-COVARIANCE MATRIX. HOWEVER, BY COMPUTER OUTPUT WE HAVE,

$S(b_1) \cong .1984$

FOR A LEVEL OF SIGNIFICANCE $\alpha = .10$.



$t_{.05,16} = 1.746$

DECISION RULE:

IF $-1.746 \leq T \leq 1.746$, REJECT H_1

OTHERWISE ACCEPT H_1.

DECISION:

$$T = \frac{.072}{.1984} \cong .365$$

THUS REJECT H_1 OR ACCEPT $\beta_1 = 0$.

THUS, WE CAN DROP X_1 FROM OUR MODEL. ONCE WE DO THIS HOWEVER, WE MUST OBTAIN THE ESTIMATES, FOR,

$$Y_i = \alpha + \beta_2 X_2 + \beta_3 X_3 + \varepsilon_i$$

WHICH, FROM COMPUTER OUTPUT IS,

$$\hat{Y} = 0.3085 + .390X_2 + 8.413X_3.$$

COMPARE THIS TO OUR PREVIOUS MODEL WHICH WAS

$$\hat{Y} = -1.051 + .072X_1 + .376X_2 + 8.429X_3.$$

NOTICE THAT THE ESTIMATED COEFFICIENTS b_2 AND b_3 (IN PARTICULAR) HAVE CHANGED. WHY IS THIS SO?

ESSENTIALLY b_j AND $S(b_j)$ ARE SENSITIVE TO THE INCLUSION OR EXCLUSION OF OTHER INDEPENDENT VARIABLES. THIS IS PRECISELY THE REASON WHY WE CAN ONLY REJECT ONE OF OUR INDEPENDENT VARIABLES WITH A t-TEST. SINCE, ONCE WE ELIMINATE IT, ANOTHER INDEPENDENT VARIABLE MAY BECOME IMPORTANT. HOWEVER, WITHOUT RE-FITTING THE MODEL, WE MAY NOT BE AWARE OF THIS.

FROM COMPUTER OUTPUT WE OBTAIN,

$$R^2 = .999$$

FOR OUR NEW MODEL. LET'S TEST TO SEE IF β_2 IS SIGNIFICANTLY DIFFERENT FROM ZERO. REMEMBER THAT X_2 IS HIGHLY CORRELATED WITH Y AT .991 - WOULD YOU EXPECT β_2 TO BE SIGNIFICANTLY DIFFERENT FROM ZERO?

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

$$T = \frac{b_2}{S(b_2)} \sim t_{n-p-1} \sim t_{17}$$

FOR $\alpha = .10$

$$t_{.05,17} = 1.74$$

$$-t_{.05,17} = -1.74$$

DECISION RULE:

IF $-1.74 \leq T \leq 1.74$, REJECT H_1

OTHERWISE ACCEPT H_1.

COMPUTATION:

FROM OUR COMPUTER OUTPUT, $S(b_2) = .3316$

THEN,

$$T = \frac{.390}{.3316} \doteq 1.175$$

$\therefore$ REJECT H_1 AND CONCLUDE THAT $\beta_2 = 0$!!!!!

HOW CAN THIS BE WHEN X_2 IS SO HIGHLY CORRELATED, AND THUS RELATED, TO Y.

THE ANSWER LIES IN OUR CORRELATION MATRIX. NOTICE THAT X_2 AND X_3 ARE CORRELATED AT .99. THUS, THE TOTAL VARIANCE EXPLAINED BY X_2 IS LARGELY EXPLAINED BY X_3. THUS INCLUSION OF X_2 DOES NOT EFFECTIVELY REDUCE THE UNEXPLAINED PORTION OF THE TOTAL VARIANCE GIVEN X_3 IS IN THE MODEL.

FORMALLY,

LET,

$SSE(X_3)$ = THE UNEXPLAINED PORTION OF THE VARIANCE WHEN X_3 IS IN THE MODEL. CALL THIS THE REDUCED MODEL (i.e., X_2 IS NOT INCLUDED).

$SSE(X_3,X_2)$ = THE UNEXPLAINED PORTION OF THE TOTAL VARIANCE WHEN X_3 AND X_2 ARE IN THE MODEL. CALL THIS THE FULL MODEL.

THEN,

$SSE(X_3)-SSE(X_3,X_2)$ REPRESENTS THE REDUCTION IN THE UNEXPLAINED PORTION OF THE VARIANCE WHEN X_2 IS INCLUDED.

THEN,

$$r^2_{X_2/X_3} = \frac{SSE(X_3)-SSE(X_3,X_2)}{SSE(X_3)}$$

REPRESENTS THE PROPORTION OF THE UNEXPLAINED VARIANCE WHEN X_3 IS IN THE MODEL [$SSE(X_3)$] EXPLAINED BY INTRODUCING X_2. THIS IS CALLED THE <u>COEFFICIENT OF PARTIAL DETERMINATION</u>.

NOTE: $r_{X_2/X_3} = \sqrt{r^2_{X_2/X_3}}$ $\equiv$ PARTIAL CORRELATION COEFFICIENT.

IF THE COEFFICIENT OF PARTIAL DETERMINATION IS SMALL, THEN THE INCLUSION OF X_2 IN THE MODEL IS NOT SIGNIFICANT IN REDUCING TOTAL VARIANCE.

4. MULTICOLLINEARITY

WHEN TWO INDEPENDENT VARIABLES ARE CORRELATED THEN MULTICOLLINEARITY IS PRESENT. MULTICOLLINEARITY CAUSES A NUMBER OF PROBLEMS. IN PARTICULAR THE STANDARD ERROR $[S(b_2), S(b_3)]$ OF OUR ESTIMATORS ARE NOT PRECISE. FURTHER, IF USED FOR FORECASTING, ONE MUST ASSUME THAT THE CORRELATION BETWEEN THE TWO VARIABLES X_2 AND X_3 WILL CONTINUE.

FOR EXAMPLE,

$$S(b_3) = .499$$

WHEN X_2 IS IN THE MODEL.

IF WE FIT THE MODEL WITH JUST X_3 IN IT WE OBTAIN,

$$\hat{Y} = .697 + 8.993X_3$$

WITH

$$S(b_3) = 0.0694$$

WHICH IS MUCH LOWER THAN .499.

FOR INSTANCE IF WE OBTAIN A 90% CONFIDENCE INTERVAL FOR β_3 WE OBTAIN

$$(b_3 - S(b_3)t_{.05,n-p-1} \leq \beta_3 \leq b_3 + S(b_3)t_{.05,n-p-1},$$

FOR X_2 AND X_3 IN THE MODEL,

$$(8.413 - (.499)t_{.05,17} \leq \beta_3 \leq 8.413 + (.499)t_{.05,17}$$

OR,

$$(7.54 \leq \beta_3 \leq 9.28).$$

FOR X_3 IN THE MODEL,

$$(8.993-(0.0694)t_{.05,18} \leq \beta_3 \leq 8.993+(0.0694)t_{.05,18}$$

OR,

$$(8.873 \leq \beta_3 \leq 9.11)$$

WHICH IS A MUCH TIGHTER 90% CONFIDENCE LIMIT (i.e., MORE PRECISE).

THE POINT HERE IS, IF MULTICOLLINEARITY IS PRESENT, THEN THE β ESTIMATES ARE NOT VERY PRECISE AND ARE NOT GOOD INDICATIONS OF THE RELATIONSHIP BETWEEN Y AND THE INDEPENDENT VARIABLE.

5. EXTRA SUM OF SQUARES

USING A CONCEPT SIMILAR TO THE PARTIAL COEFFICIENT OF DETERMINATION WE MAY TEST WHETHER A GROUP OF VARIABLES MAY BE INCLUDED (RETAINED) IN A MODEL, OR NOT.

FOR EXAMPLE, SUPPOSE WE WISH TO TEST SIMULTANEOUSLY, WHETHER TO INCLUDE X_1 AND X_2 IN OUR MODEL.

FIRST, WE FIT A REDUCED MODEL WITH JUST X_3 IN IT AND OBTAIN $SSE(X_3)$.

THEN WE FIT A FULL MODEL WITH X_1, X_2 AND X_3 IN IT AND OBTAIN $SSE(X_1, X_2, X_3)$.

THEN,

$$\frac{SSE(X_3)-SSE(X_1,X_2,X_3)}{SSE(X_1,X_2,X_3)}$$

REPRESENTS THE REDUCTION IN THE UNEXPLAINED PORTION OF THE VARIANCE WHEN X_1 AND X_2 ARE INCLUDED.

FROM OUR COMPUTER OUTPUT WE HAVE,

$$SSE(X_3) = 172.17$$

$$SSE(X_1, X_2, X_3) = 149.58$$

THEN,

$$F = \frac{[172.17 - 149.58]/2}{9.349} = \frac{[SSE(X_3) - SSE(X_1, X_2, X_3)]/q}{SSE(X_1, X_2, X_3)/(n-4)}$$

$$= 1.208.$$

SINCE $F < 4.69$ WE CONCLUDE H_0, THAT $\beta_1 = \beta_2 = 0$. $F_{0.05,2,16} = 4.69$

THUS, THE REGRESSION MODEL WE WILL RETAIN FOR THE PRODUCTION OF STAINLESS STEEL SHEETS IS,

$$\hat{Y} = .697 + 8.993X_3$$

WHERE

X_3 IS THE STRENGTH OF THE STEEL.

TRY THE FOLLOWING PROBLEM,

A COMPANY SELLS A SPECIAL SKIN CREAM THROUGH DRUGSTORES EXCLUSIVELY. IT OPERATES IN 15 MARKETING DISTRICTS AND IS INTERESTED IN PREDICTING DISTRICT SALES. THE COMPANY HAS DATA ON SALES BY DISTRICT, AS WELL AS DISTRICT DATA ON TARGET POPULATION AND PER CAPITA INCOME. EXPLORE THE FEASIBILITY OF PREDICTING DISTRICT SALES FROM TARGET POPULATION AND PER CAPITA INCOME

THE TEST WE WILL CONDUCT IS,

H_0: $\beta_1 = \beta_2 = 0$

H_1: NOT ALL $\beta_j = 0$

LET $\alpha = .05$

THE RELEVANT STATISTIC IS,

$$F = \frac{[SSE(X_3)-SSE(X_1,X_2,X_3)]/q}{SSE(X_1,X_2,X_3)/(n-p-1)}$$

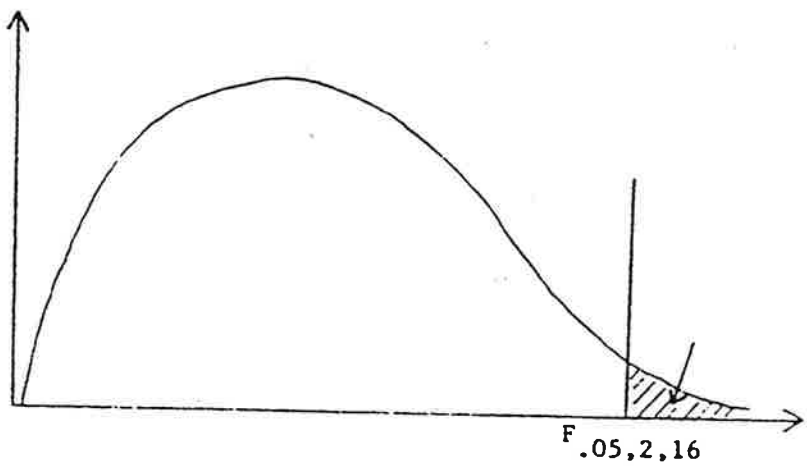
WHERE q = THE NUMBER OF VARIABLES BEING TESTED (q=2 HERE).

THUS,

$$F = \frac{[SSE(X_3)-SSE(X_1,X_2,X_3)]/q}{MSE}$$

WHEN H_0 HOLDS,

$$F \sim F_{q,n-p-1} = F_{2,16}$$



$F_{.05,2,16} = 4.69.$

DISTRICT	SALES	TARGET POP.	PER CAPITA INCOME
1	162	274	2,450
2	120	180	3,254
3	223	375	3,802
4	131	205	2,828
5	67	86	2,347
6	169	265	3,782
7	81	98	3,008
8	192	330	2,450
9	116	195	2,137
10	55	53	2,560
11	252	430	4,020
12	232	372	4,427
13	144	236	2,660
14	103	157	2,088
15	212	370	2,605

TRY IT!!

6. DUMMY VARIABLES

SO FAR WE HAVE DISCUSSED MULTIPLE REGRESSION ANALYSIS USING "SO-CALLED" QUANTITATIVE VARIABLES WHICH TAKE ON VALUES ON A WELL-DEFINED SCALE.

MANY VARIABLES OF INTEREST IN BUSINESS AND ECONOMICS, HOWEVER, ARE NOT QUANTITATIVE BUT RATHER ARE QUALITATIVE. EXAMPLES ARE SEX(MALE, FEMALE), PURCHASE STATUS (PURCHASE, NO PURCHASE), AND DISABILITY STATUS (NOT DISABLED, PARTLY DISABLED, FULLY DISABLED).

QUANTITATIVE VARIABLES CAN BE USED TO REPRESENT QUALITATIVE VARIABLES. THESE VARIABLES ARE CALLED DUMMY VARIABLES (ALSO INDICATOR OR BINARY VARIABLES).

A. EXAMPLES

(i) SUPPOSE WE WISH TO REPRESENT SEX WITH A QUANTITATIVE VARIABLE. THE QUANTITATIVE VARIABLE WE WILL USE IS A BINARY VARIABLE X, SUCH THAT

$X_1 = 1$ IF MALE;

$X_1 = 0$ IF FEMALE.

NOTICE THAT ONE BINARY VARIABLE REPRESENTS BOTH CLASSES OF SEX.

(ii) SUPPOSE WE WISH TO REPRESENT THE QUALITATIVE VARIABLE OF DISABILITY STATUS WITH A QUANTITATIVE VARIABLE. TO DO THIS WE MUST USE TWO BINARY VARIABLES X_1 AND X_2, SUCH THAT,

$X_1 = 1$ IF NOT DISABLED;

$X_2 = 1$ IF PARTLY DISABLED.

ALL OF THE DISABILITY CLASSES ARE THEN REPRESENTED BY THE FOLLOWING JOINT VALUES OF X_1 AND X_2,

X_1	X_2	
1	0	NOT DISABLED
0	1	PARTLY DISABLED
0	0	FULLY DISABLED

NOTE: IN GENERAL, IF THERE ARE C CLASSES FOR A QUALITATIVE VARIABLE, THEN (C-1) BINARY OR DUMMY VARIABLES WILL BE REQUIRED TO REPRESENT THE C CLASSES.

B. EXAMPLE

AN ECONOMIST WISHED TO RELATE THE SPEED WITH WHICH A PARTICULAR INSURANCE INNOVATION IS ADOPTED (Y) TO THE SIZE OF THE FIRM (X_1) AND THE TYPE OF FIRM.

$Y \equiv$ # OF MONTHS ELAPSED BETWEEN THE TIME THE FIRST FIRM ADOPTED THE INNOVATION AND THE TIME THE GIVEN FIRM ADOPTED THE INNOVATION.

$X_1 \equiv$ TOTAL ASSETS OF THE FIRM.

TYPE OF FIRM $\equiv$ STOCK COMPANIES OR MUTUAL COMPANIES.

LET $X_2 = 1$ IF STOCK

$X_2 = 0$ IF MUTUAL

THE FOLLOWING DATA WAS OBTAINED,

FIRM	Y_i	X_{1i}	TYPE	X_{2i}
1	17	151	MUTUAL	0
2	26	92	MUTUAL	0
3	28	164	STOCK	1
4	15	272	STOCK	1
5	11	295	STOCK	1
6	21	175	MUTUAL	0
7	30	31	MUTUAL	0
8	22	104	MUTUAL	0
9	38	68	STOCK	1
10	0	277	MUTUAL	0
11	12	210	MUTUAL	0
12	31	85	STOCK	1
13	21	224	STOCK	1
14	20	166	STOCK	1
15	13	305	STOCK	1
16	19	120	MUTUAL	0
17	30	124	STOCK	1
18	4	290	MUTUAL	0
19	16	238	MUTUAL	0
20	14	246	STOCK	1

SUPPOSE WE FIT THE FOLLOWING REGRESSION MODEL,

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

THEN THE RESPONSE FUNCTION OF OUR MODEL IS,

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 X_2 \qquad \text{WHY?}$$

WHEN $X_2 = 1 \Rightarrow$ <u>STOCK COMPANY</u>, THE MODEL IS,

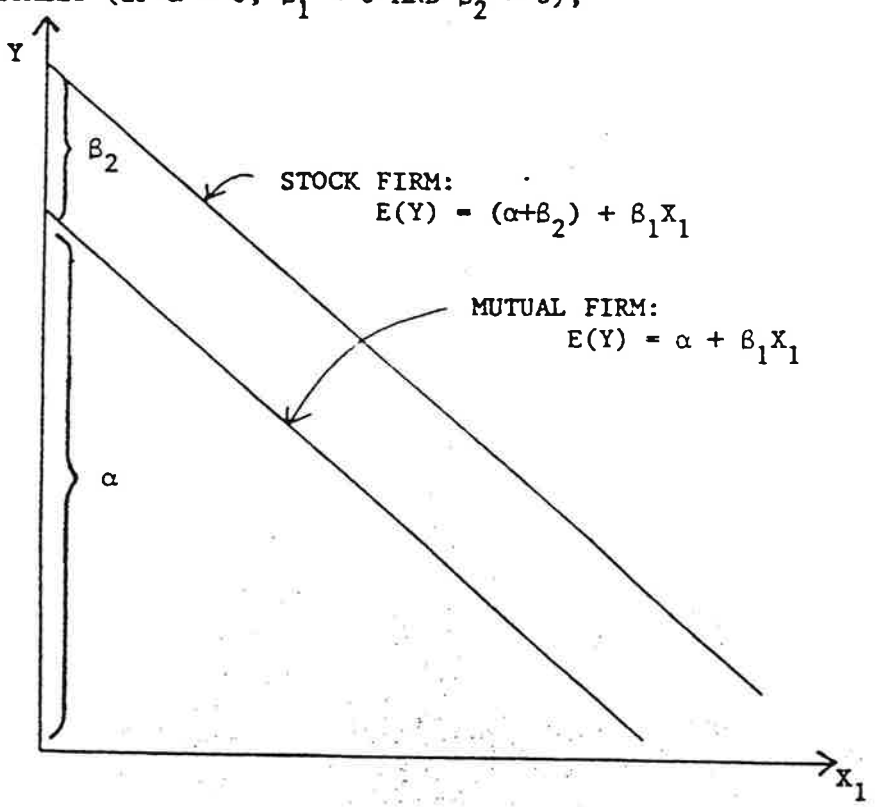
$$E(Y) = \alpha + \beta_1 X_1 + \beta_2$$
$$= (\alpha + \beta_2) + \beta_1 X_1$$

WHEN $X_2 = 0 \Rightarrow$ <u>MUTUAL COMPANY</u>, THE MODEL IS,

$$E(Y) = \alpha + \beta_1 X_1$$

NOTICE THAT THE SLOPE (β_1) IS THE SAME FOR BOTH COMPANIES. THE MODELS DIFFER ONLY IN THE INTERCEPT POINT.

GRAPHICALLY (IF $\alpha > 0$, $\beta_1 < 0$ AND $\beta_2 > 0$),



CONSEQUENTLY WE ASSUME THAT THE RESPONSE FUNCTIONS FOR BOTH FIRMS ARE PARALLEL AND DIFFER ONLY WITH RESPECT TO THE INTERCEPT TERMS (α) AND ($\alpha+\beta_2$).

FITTING THE DATA, WE OBTAIN

	REGRESSION COEFFICIENTS		
COEFFICIENT	ESTIMATE	STD. DEV.	t
α	33.874	1.814	18.68
β_1	-.102	.009	-11.44
β_2	8.055	1.459	5.52

ANOVA

SOURCE	SS	DF	MS
REGRESSION	1,504.41	2	752.20
ERROR	176.39	17	10.38
TOTAL	1,680.80	19	

SINCE β_2 REPRESENTS THE DIFFERENCE BETWEEN MUTUAL AND STOCK COMPANIES FOR ADOPTION OF INNOVATION TIME, WE WILL CONSTRUCT A 95% CONFIDENCE INTERVAL FOR β_2 AS FOLLOWS,

$$t_{(\cancel{.975,17})\ 0.025,17} = 2.110$$

THEN,

$$\text{UPPER LIMIT} = 8.055 + (2.110)(1.459)$$
$$= 11.134$$

$$\text{LOWER LIMIT} = 8.055 - (2.110)(1.459)$$
$$= 4.977$$

THEN,

$$(4.977 \leq \beta_2 \leq 11.134).$$

THUS, WITH 95% CONFIDENCE, WE WOULD CONCLUDE THAT MUTUAL COMPANIES TENDED TO ADOPT THE INNOVATION SOMEWHERE BETWEEN 5 AND 11 MONTHS EARLIER, ON THE AVERAGE, THAN STOCK COMPANIES.

SUPPOSE WE TESTED,

$$H_0: \beta_2 = 0$$

$$H_1: \beta_2 \neq 0$$

WITH $\alpha = 0.05$.

WHAT WOULD BE THE RESULT?

C. <u>OTHER USES OF DUMMY VARIABLES</u>

(i) SUPPOSE, IN OUR PREVIOUS EXAMPLE WE COULD NOT ASSUME THAT THE TWO RESPONSE FUNCTIONS HAD THE SAME SLOPE β_1.

WE COULD EXAMINE THE IMPACT OF AN <u>INTERRACTION</u> TERM AS FOLLOWS,

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \varepsilon_i$$

WHERE,

$X_{1i}X_{2i} \equiv$ AN INTERRACTION TERM,

WHEN $X_2 \overset{=}{\neq} 1 \Rightarrow$ <u>STOCK</u>, THEN

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 + \beta_3 X_1$$

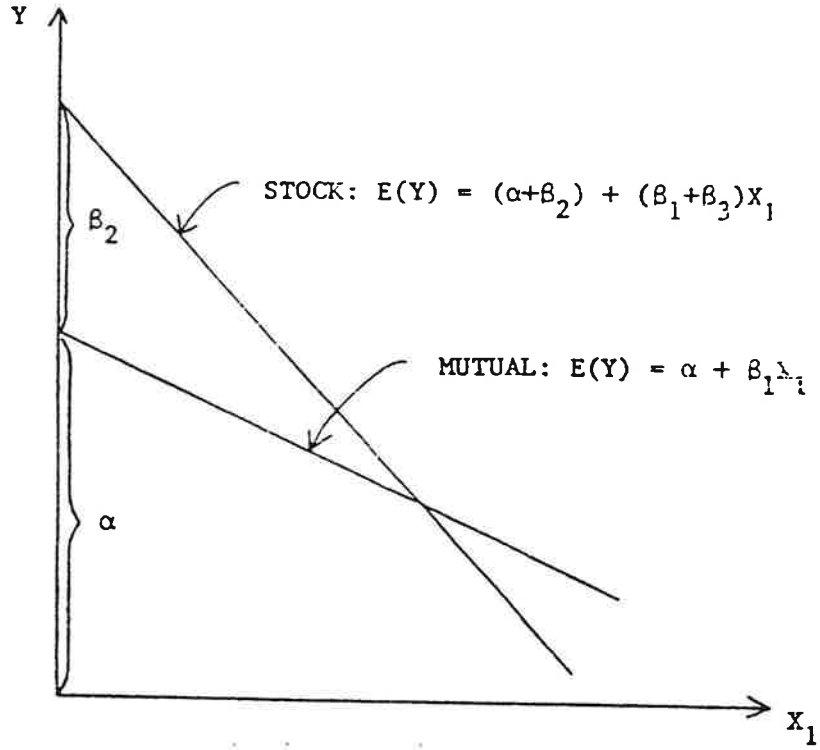
$$= (\alpha+\beta_2) + (\beta_1+\beta_3)X_1.$$

WHEN $X_2 = 0 \Rightarrow$ <u>MUTUAL</u>, THEN

$$E(Y) = \alpha + \beta_1 X_1.$$

NOTICE THAT THE TWO RESPONSE FUNCTIONS DIFFER <u>NOW</u> BOTH WITH RESPECT TO INTERCEPT AND SLOPE.

GRAPHICALLY (SUPPOSE $\beta_3 < 0$)



HOW WOULD YOU TEST TO SEE WHETHER OR NOT THE TWO REGRESSION FUNCTIONS DIFFERED WITH RESPECT TO THE SLOPE?

TRY IT!!

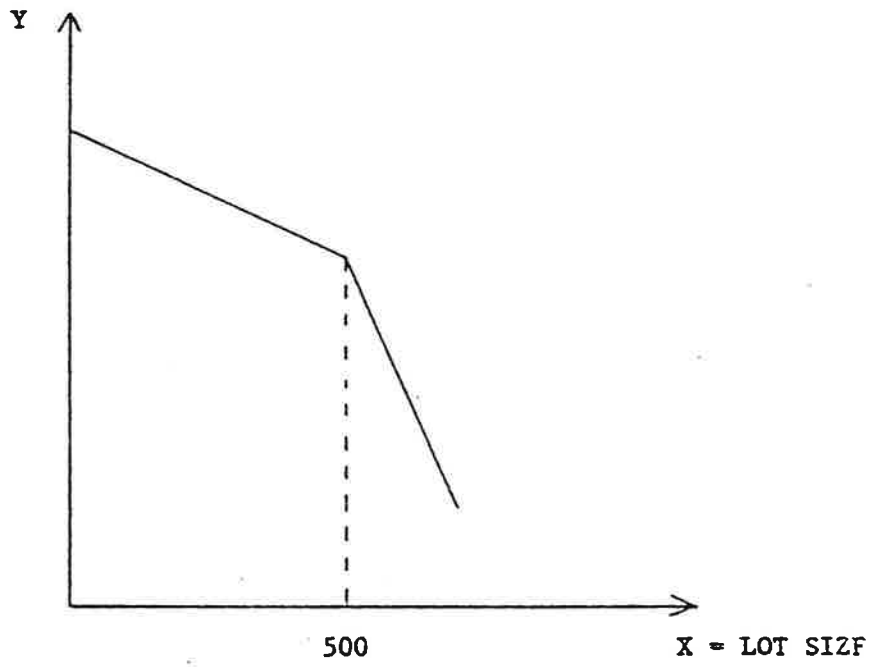
(ii) USING DUMMY VARIABLES IN PIECEWISE LINEAR REGRESSION

SOMETIMES THE REGRESSION OF Y ON X FOLLOWS A PARTICULAR LINEAR RELATION IN SOME RANGE OF X, BUT FOLLOWS A DIFFERENT LINEAR RELATION ELSEWHERE.

EXAMPLES

1. SUPPOSE IN A PRODUCTION PROCESS PRODUCING VARIOUS LOT SIZES OF GOODS, THE UNIT PRODUCTION COST (Y) FOLLOWS A CERTAIN LINEAR REGRESSION UP TO LOTS OF SIZE 500, AT WHICH POINT THE SLOPE CHANGES BECAUSE OF SOME OPERATING EFFICIENCIES ONLY POSSIBLE WITH LOT SIZES OF MORE THAN 500.

GRAPHICALLY



THE MODEL FOR THIS PROBLEM CAN BE EXPRESSED AS,

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - 500) X_{2i} + \epsilon_i$$

WHERE,

$$X_{2i} = 1 \text{ IF } X_{1i} > 500$$
$$= 0 \text{ IF } X_{1i} \leq 500.$$

WHEN $X_2 = 1 \Rightarrow \underline{X_1 > 500}$, THEN

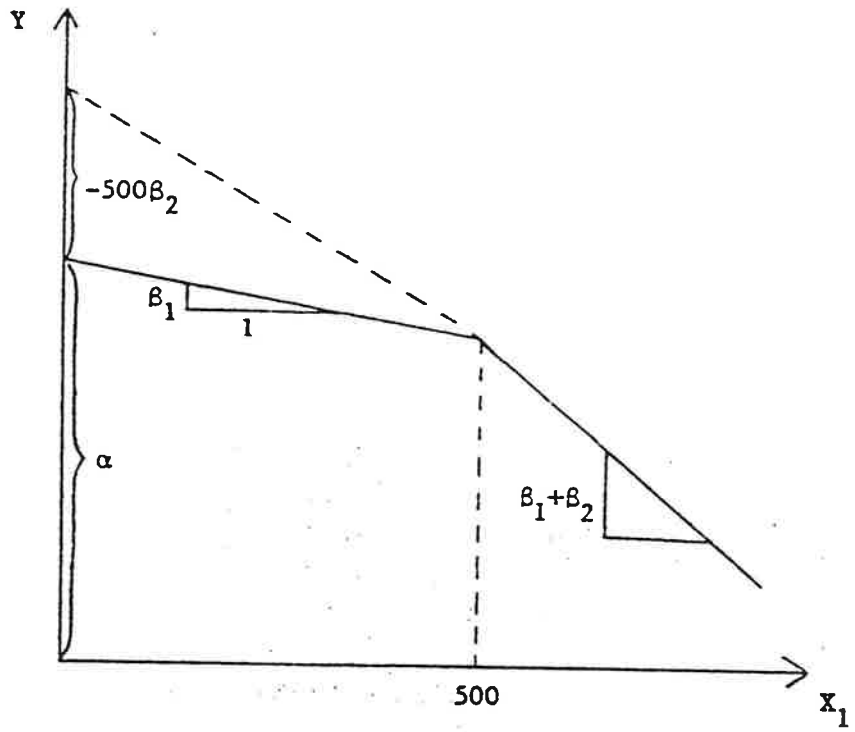
$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 (X_1 - 500)$$
$$= (\alpha - 500\beta_2) + (\beta_1 + \beta_2) X_1.$$

WHEN $X_2 = 0 \Rightarrow \underline{X_1 \leq 500}$, THEN

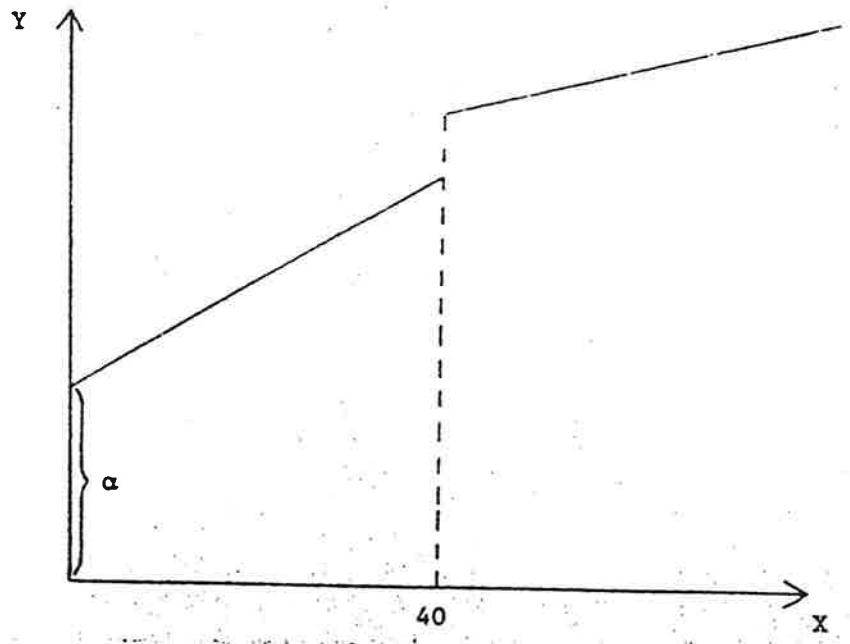
$$E(Y) = \alpha + \beta_1 X_1.$$

NOTICE THAT THE TWO RESPONSE FUNCTIONS DIFFER BOTH WITH RESPECT TO INTERCEPT AND SLOPE.

GRAPHICALLY (SUPPOSE $\alpha > 0$, $\beta_1 < 0$, $\beta_2 < 0$)



2. SUPPOSE WE HAD A DISCONTINUOUS LINEAR REGRESSION WHICH LOOKED LIKE,



THE MODEL WOULD BE,

$$Y_i = \alpha + \beta_1 X_{1i} + \beta_2 (X_{1i} - 40) X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

WHERE,

$$X_{2i} = 1 \text{ IF } X_{1i} > 40$$
$$= 0 \text{ IF } X_{1i} \leq 40$$

$$X_{3i} = 1 \text{ IF } X_{1i} > 40$$
$$= 0 \text{ IF } X_{1i} \leq 40$$

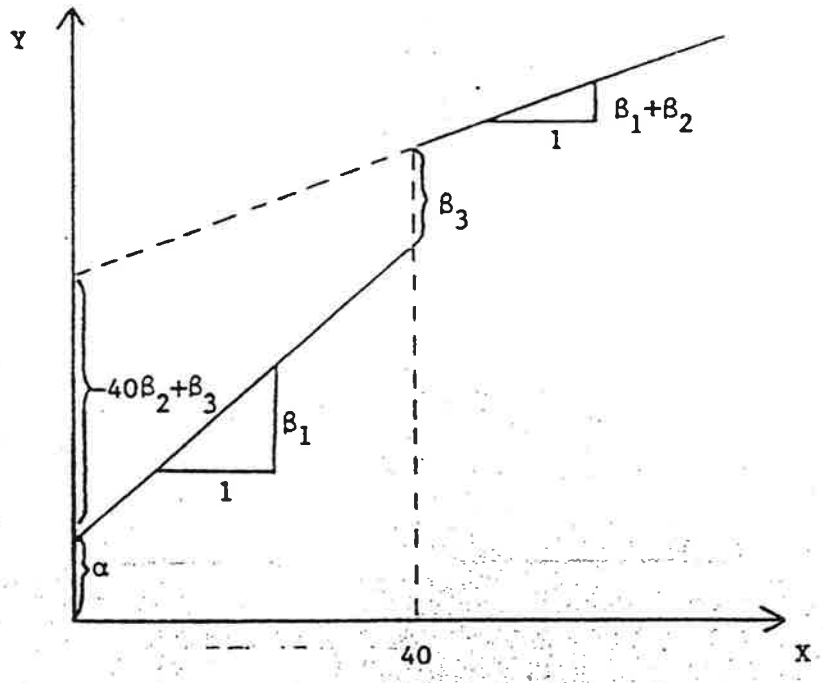
WHEN $X_2 = 0$ AND $X_3 = 0 \Rightarrow X_1 \leq 40$, THEN

$$E(Y) = \alpha + \beta_1 X_1.$$

WHEN $X_2 = 1$ AND $X_3 = 1 \Rightarrow X_1 > 40$, THEN

$$E(Y) = \alpha + \beta_1 X_1 + \beta_2 (X_1 - 40) + \beta_3$$
$$= (\alpha - 40\beta_2 + \beta_3) + (\beta_1 + \beta_2) X_1.$$

GRAPHICALLY,



3. SAMPLE DATA FOR THE PRODUCTION PROBLEM,

LOT	UNIT COST	LOT SIZE	X_2
1	2.57	650	1
2	4.40	340	0
3	4.52	400	0
4	1.39	800	1
5	4.75	300	0
6	3.55	570	1
7	2.49	720	1
8	3.77	480	0

THE FITTED REGRESSION EQUATION IS,

$$\hat{Y} = 5.895 - .004X_1 - .004(X_1-500)X_2$$

A. HOW WOULD WE TEST THE CONJECTURE THAT THE SLOPE INDEED DOES CHANGE FOR LOT SIZES GREATER THAN 500?

B. HOW WOULD WE TEST, INCLUDING THE APPROPRIATE REGRESSION MODEL TO FIT, THE CONJECTURE THAT THERE IS A DISCONTINUITY AT 500?