Data modeling: CSCI E-106

Applied Linear Statistical Models

Chapter 8 – Regression Models for Quantitative and Qualitative Predictors

Uses Polynomial Models

- Polynomial regression models for quantitative predictor variables
  - the most frequently used curvilinear response models
  - handled easily
  - special case of the general LRM

Uses Polynomial Models, cont'd

- Tw obasic types
  - The true curvilinear response function is a polynomial function
  - The true curvilinear response function is unknown; a polynomial function is a good approximation to the true function

- Polynomial regression models may provide good fits for the data at hand, but may turn in unexpected directions when extrapolated beyond the range of the data

Polynomial Regression Models

- Polynomial regression models may contain one, two, or more than two predictor variables.
- Each predictor variable may be present in various powers

A second order model with one predictor variable

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i$$

where x_i=$(X_i - \bar{X})$.

  - Predictor variable is centered ($x_i = (X_i - \bar{X})$): X and X^2 will be highly correlated.
  - Centering the predictor variable reduces the multicollinearity

Second order with predictor variable

- The regression coefficients in polynomial regression are frequently written in a slightly different fashion, to reflect the pattern of the exponents:

$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \varepsilon_i$, where x_i=$(X_i - \bar{X})$.

- The quadratic response function: parabola → $E\{Y\} = \beta_o + \beta_1 x + \beta_{11} x^2$

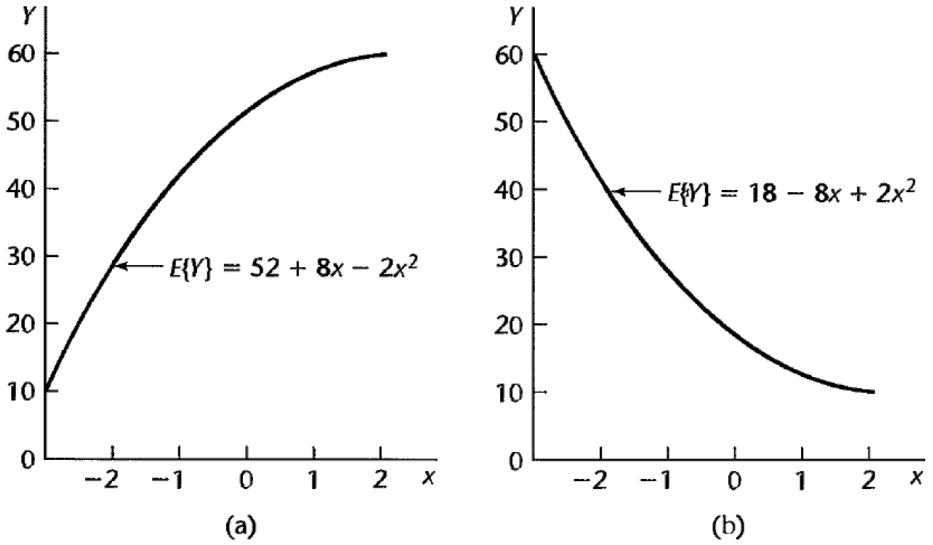


Figure : Examples of Second-Order Polynomial Response Functions.

Second order with predictor variable, cont'd

$$E\{Y\} = \beta_o + \beta_1 x + \beta_{11} x^2$$

Parameters:

- β_o: the mean response of Y when $x = 0$ (i.e $X = \overline{X}$)
- β_1: the linear effect coefficient
- β_{11}: the quadratice effect coefficent

Third order with predictor variable

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \varepsilon_i$$

where x_i=$(X_i - \bar{X})$.

The response function for regression model:

$$E\{Y\} = \beta_o + \beta_1 x + \beta_{11} x^2 + \beta_{111} x^3$$

Higher orders with one predictor variable

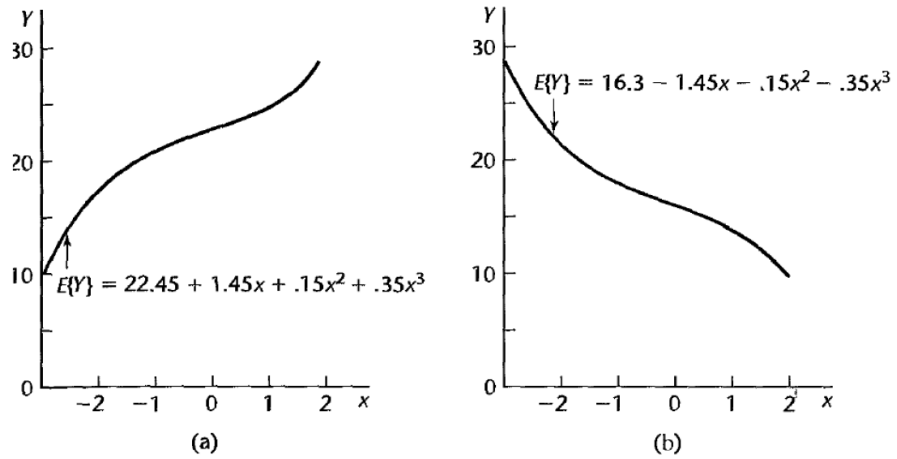


Figure : Examples of Third-Order Polynomial Response Functions.

- Employed with special caution
- The interpretation of the coefficients becomes difficult
- The models may be highly erratic for interpolations and small extrapolations
- A polynomial model of sufficiently high order can always be found to fit data containing no repeat observations perfectly.
  - Illustration: the fitted polynomial regression function for one predictor variable of order n − 1 will pass through all n observed Y values.

Second order with two predictor variables

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i \,,$$

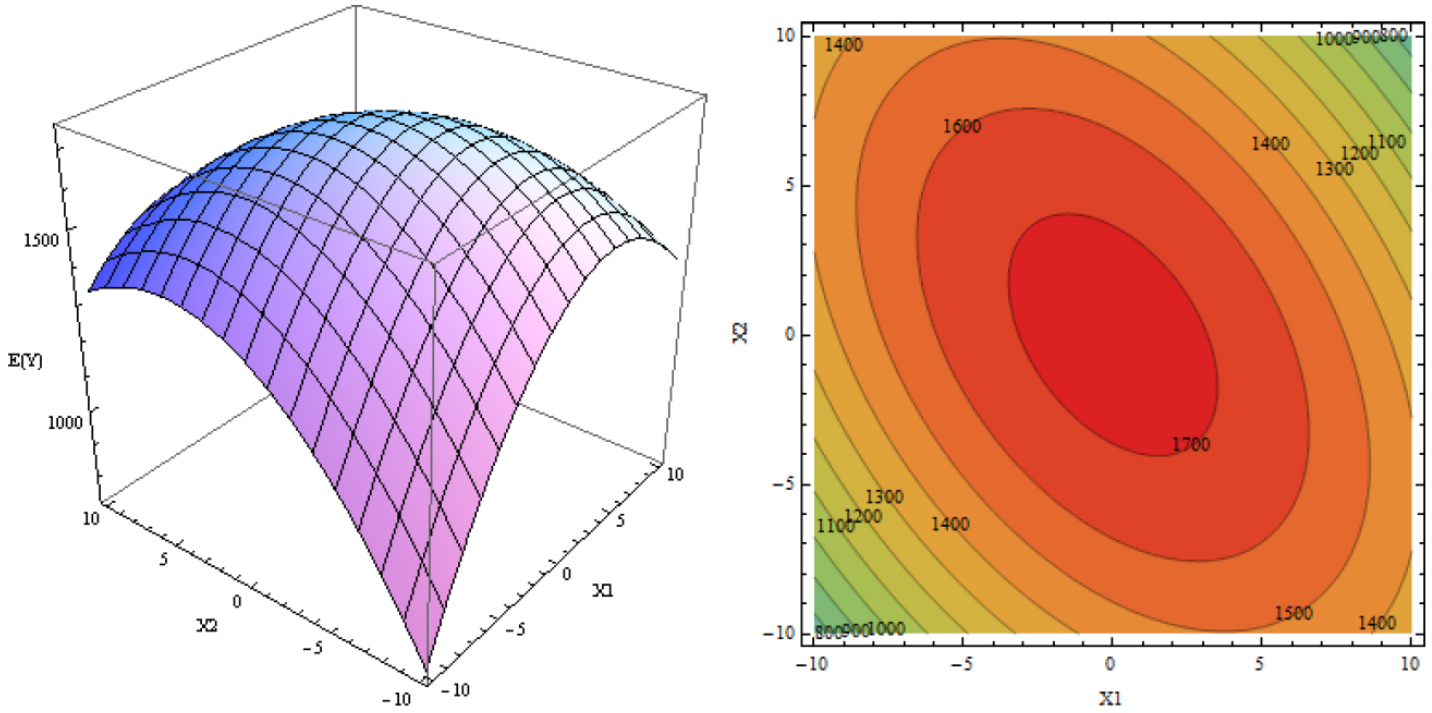
where $x_{i1}=(X_{i1} - \bar{X}_1)$, and $x_{i2}=(X_{i2} - \bar{X}_2)$

- The response function: conic section

$$E\{Y\} = \underbrace{\beta_0 + \beta_1 x_1 + \beta_2 x_2}_{\text{Linear component}} + \underbrace{\beta_{11} x_1^2 + \beta_{22} x_2^2}_{\text{Quadratic component}} + \underbrace{\beta_{12} x_1 x_2}_{\text{Cross product}}$$

- β_{12}: the interaction effect coefficient
- The second-order model with three predictor variables is similar

Second order with two predictor variables, cont'd



(a) Response Surface (b) Contour Curves

Figure : Examples -$EY = 1,740 - 4x_1^2 - 3x_2^2 - 3x_1x_2$. (By mathematica)

maximum: $x_1 = 0;\ x_2 = 0$

Second order with two predictor variables, cont’d

```
## Figure of 8.3
require(grDevices)
x1<-seq(-10,10,0.1)
x2<-seq(-10,10,0.1)
y <- function(x1, x2) {1740-4*x1^2-3*x2^2-3*x1*x2}
z <- outer(x1, x2, y)
persp(x1, x2, z, theta = -30, phi = 20, expand = 1,col = "red",ticktype = "detailed",,xlab="x1",ylab = "x2", zlab = "E(y)")
image(x1, x2, z, col =terrain.colors(100))
contour(x1, x2, z, col = "blue", add = TRUE, method = "edge",vfont = c("sans serif", "plain"))
```

Hierarchical Approach to Fitting

- Special cases of the general linear regression model
- Fitting of polynomial models: no new problems
- All earlier results on fitting apply on making inferences.

Implementation of Polynomial Regression Models

Ideas to find an approximation to the true regression function:

1. often fit a second-order or third-order model

$$Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \varepsilon_i$$

2. explore whether a lower-order model is adequate
   - Test $\beta_{111} = 0$
   - Or test not both $\beta_{11} = 0; \beta_{111} = 0$
3. The decomposition of SSR into extra sums of squares:

$$SSR(x); \quad SSR(x^2|x); \quad SSR(x^3|x, x^2)$$

4. Test whether $\beta_{111} = 0$: $SSR(x^3|x, x^2)$
   Test whether $\beta_{11} = \beta_{111} = 0$:
   $SSR(x^2, x^3|x) = SSR(x^2|x) + SSR(x^3|x, x^2)$

Implementation of Polynomial Regression Models, cont'd

- If a polynomial term of a given order is retained, then all related terms of lower order are also retained in the model.
  - providing more basic information about the shape of the response function

- Wish to express the model in terms of the original variables:

$$\hat{Y} = b_0 + b_1 x + b_{11} x^2$$

$$\hat{Y} = b_0 + b_1 (X - \bar{X}) + b_{11} (X - \bar{X})^2$$

$$\hat{Y} = (b_0 - b_1 \bar{X} + b_{11} \bar{X}^2) + (b_1 - 2 b_{11} \bar{X}) X + b_{11} X^2$$

$$\hat{Y} = b_0' + b_1' X + b_{11}' X^2$$

  - The fitted values and residuals for the regression function in terms of x or X are the same.

- Centered observations: to reduce potential calculation difficulties; multicollinearity;

Case Example: the life of power cell

- Studied the effects of the charge rate and temperature on the
- life of a new type of power cell
- X_1: the charge rate at three level (0.6, 1.0, 1.4 amperes)
- X_2: the ambient temperature at three levels (10, 20,300C)
- Y : the life of the power cell
- don't know the nature of the response function in the range of the factors studied
- fit the second-order polynomial regression model

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{11} x_{i1}^2 + \beta_{22} x_{i2}^2 + \beta_{12} x_{i1} x_{i2} + \varepsilon_i \,,$$

$$\Rightarrow$$

$$E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

Case Example: the life of power cell, cont'd

Table : Data-Power Cells Example.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Number of Cycles	Charge Rate		Coded	Values			
Cell	(thousands)	(thousands)	Temperature					
i	Y_i	X_{i1}	X_{i2}	x_{i1}	x_{i2}	x_{i1}^2	x_{i2}^2	$x_{i1}x_{i2}$
1	150	0.6	10	-1	-1	1	1	1
2	86	1.0	10	0	-1	0	1	0
3	49	1.4	10	1	-1	1	1	-1
4	288	0.6	20	-1	0	1	0	0
5	157	1.0	20	0	0	0	0	0
6	131	1.0	20	0	0	0	0	0
7	184	1.0	20	0	0	0	0	0
8	109	1.4	20	1	0	1	0	0
9	279	0.6	30	-1	1	1	1	-1
10	235	1.0	30	0	1	0	1	0
11	224	1.4	30	1	1	1	1	1
		$\bar{X}_1 = 1.0$	$\bar{X}_2 = 20$					

$$x_{i1} = \frac{X_{i1} - \bar{X}_1}{0.4}; \qquad x_{i2} = \frac{X_{i2} - \bar{X}_2}{10}$$

Case Example: the life of power cell, cont'd

```
ex<-Dataset_08TA01
attach(ex)
X1sqr<-X1^2
X2sqr<-X2^2
X1X2<-X1*X2
x1<-(ex$X1-mean(ex$X1))/0.4
x2<-(ex$X2-mean(ex$X2))/10
x1sqr<-x1^2
x2sqr<-x2^2
x1x2<-x1*x2
```

Case Example: the life of power cell, cont'd

Correlation between	
X_1 and X_1^2:	.991
x_1 and x_1^2:	0.0

Correlation between	
X_2 and X_2^2:	.986
x_2 and x_2^2:	0.0

```
> round(cor(cbind(X1,X2,X1sqr,X2sqr,X1X2)),4)
          X1     X2  X1sqr  X2sqr   X1X2
X1    1.0000 0.0000 0.9910 0.0000 0.6053
X2    0.0000 1.0000 0.0000 0.9861 0.7566
X1sqr 0.9910 0.0000 1.0000 0.0059 0.5999
X2sqr 0.0000 0.9861 0.0059 1.0000 0.7461
X1X2  0.6053 0.7566 0.5999 0.7461 1.0000
```

```
> round(cor(cbind(x1,x2,x1sqr,x2sqr,x1x2)),4)
      x1 x2  x1sqr  x2sqr x1x2
x1     1  0 0.0000 0.0000    0
x2     0  1 0.0000 0.0000    0
x1sqr  0  0 1.0000 0.2667    0
x2sqr  0  0 0.2667 1.0000    0
x1x2   0  0 0.0000 0.0000    1
```

- Interested in whether interaction effects and curvature effects are required in the model
- fitted model:

$$\hat{Y} = 162.84 - 55.83x_1 + 75.50x_2 + 27.39x_1^2 - 10.61x_2^2 + 11.50x_1x_2$$

Case Example: the life of power cell, cont'd

```
> summary(lm(Y~x1+x2+x1sqr+x2sqr+x1x2))
Call:
lm(formula = Y ~ x1 + x2 + x1sqr + x2sqr + x1x2)

Residuals:
      1       2       3       4       5       6       7       8       9
-21.465   9.263  12.202  41.930  -5.842 -31.842  21.158 -25.404 -20.465
     10      11
  7.263  13.202

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   162.84      16.61   9.805 0.000188 ***
x1            -55.83      13.22  -4.224 0.008292 **
x2             75.50      13.22   5.712 0.002297 **
x1sqr          27.39      20.34   1.347 0.235856
x2sqr         -10.61      20.34  -0.521 0.624352
x1x2           11.50      16.19   0.710 0.509184
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 32.37 on 5 degrees of freedom
Multiple R-squared:  0.9135,    Adjusted R-squared:  0.8271
F-statistic: 10.57 on 5 and 5 DF,  p-value: 0.01086
```

Case Example: the life of power cell, cont'd

Model: MODEL1
Dependent Variable: Y

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob>F
Model	5	55365.56140	11073.11228	10.565	0.0109
Error	5	5240.43860	1048.08772		
C Total	10	60606.00000			

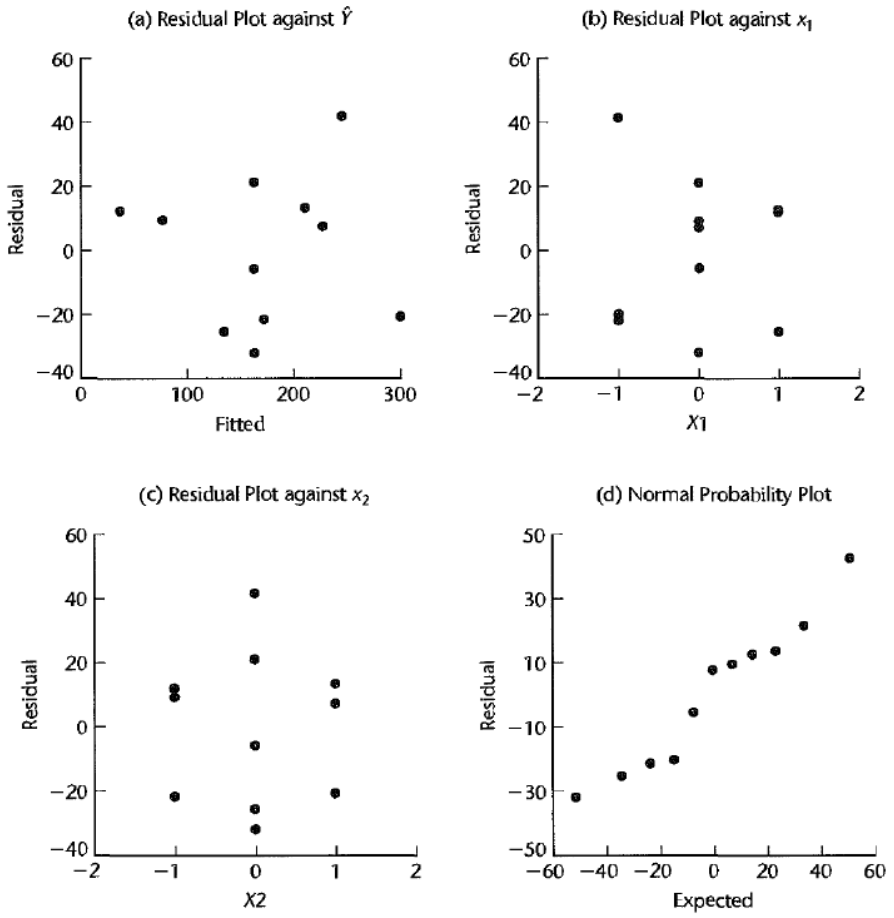
Root MSE	32.37418	R-square	0.9135
Dep Mean	172.00000	Adj R-sq	0.8271
C.V.	18.82220		

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	T for H0: Parameter=0	Prob > |T|
INTERCEP	1	162.842105	16.60760542	9.805	0.0002
X1	1	-55.833333	13.21670483	-4.224	0.0083
X2	1	75.500000	13.21670483	5.712	0.0023
X1SQ	1	27.394737	20.34007956	1.347	0.2359
X2SQ	1	-10.605263	20.34007956	-0.521	0.6244
X1X2	1	11.500000	16.18709146	0.710	0.5092

Variable	DF	Type I SS
INTERCEP	1	325424
X1	1	18704
X2	1	34202
X1SQ	1	1645.966667
X2SQ	1	284.928070
X1X2	1	529.000000

Case Example: the life of power cell, cont'd



- None of the plots suggest any gross inadequacies of the model
- The coefficient of correlation between the ordered residuals and their expected values: 0.974

Figure : Diagnostic Residual Plots-Power Cells Example.

Case Example: the life of power cell, cont'd

- Test of fit: replications at $x_1 = 0$, $x_2 = 0$

$$SSPE = 1,\ 404.67 \quad (df = n - c = 2)$$
$$SSLF = 3,\ 835.77 \quad (df = c - p = 3)$$
$$F^* = 1.82 \leq F(0.95;\ 3,\ 2) = 19.2$$

conclude that the second-order polynomial regression function is a good fit

- Coefficient of Multiple Determination: R^2 = %91 and R_a^2=%83

The variation in the lives of the power cells is reduced by about 91% when the first-order and second-order relations to the change rate and ambient temperature are utilized.

Case Example: the life of power cell, cont'd

```
## Multiple regression: lack of fit
### Method 1
fit<-lm(Y~x1+x2+x1sqr+x2sqr+x1x2)
exfactor = factor( c(seq(-4,-1), rep(0,3),seq(1,4)) )
#fit full model
anova( fit, lm(Y ~ exfactor))
### Method 2
library(alr3) # for lack of fit
pureErrorAnova(fit)
```

Case Example: the life of power cell, cont'd

```
Analysis of Variance Table

Response: Y
            Df Sum Sq Mean Sq F value  Pr(>F)
x1           1  18704   18704 26.6315 0.03556 *
x2           1  34202   34202 48.6970 0.01992 *
x1sqr        1   1646    1646  2.3436 0.26546
x2sqr        1    285     285  0.4057 0.58935
x1x2         1    529     529  0.7532 0.47696
Residuals    5   5240    1048
 Lack of fit 3   3836    1279  1.8205 0.37378
 Pure Error  2   1405     702
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Case Example: the life of power cell, cont'd

Partial *F* Test: whether a first-order model would be sufficient?

$$H_0: \beta_{11} = \beta_{22} = \beta_{12}$$
$$H_a: \text{not all } \beta\text{s in } H_0 \text{ equal to zero}$$

$$\Rightarrow F^* = \frac{SSR(x_1^2, x_2^2, x_1x_2 | x_1, x_2)}{3} \div MSE = \frac{2{,}459.9}{3} \div 1048.1$$

$$= 0.78 \leq F(0.95; 3,5) = 5.41$$

⇒ conclude H_0: no curvature and interaction effects are needed

```
> anova( lm(Y~x1+x2), fit)
Analysis of Variance Table
Model 1: Y ~ x1 + x2
Model 2: Y ~ x1 + x2 + x1sqr + x2sqr + x1x2
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      8 7700.3
2      5 5240.4  3    2459.9 0.7823 0.5527
```

Case Example: the life of power cell, cont'd

First-order Model:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$$

$$\Rightarrow \hat{Y} = 172.00 - \underset{(12.67)}{55.83} x_1 + \underset{(12.67)}{75.50} x_2$$

$$\Rightarrow \hat{Y} = 160.58 - 139.58 X_1 + 7.55 X_2 \quad \text{the original variables}$$

Case Example: the life of power cell, cont'd

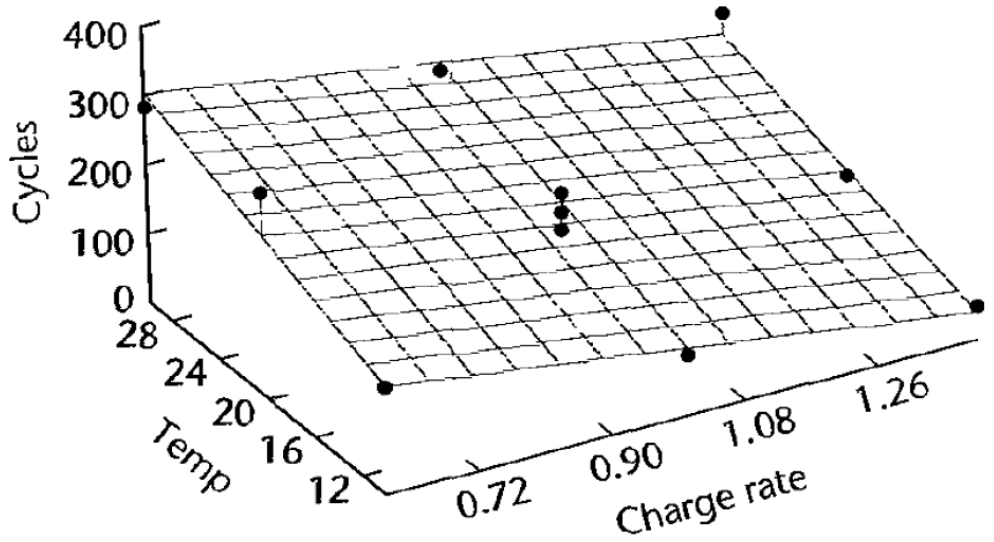


Figure : S-Plus Plot of Fitted Response Plane (8.19)-Power Cells Example.

Case Example: the life of power cell, cont'd

- Estimation of Regression Coefficients:
  to estimate the linear effect of the two predictor variables with a 90% family confidence coefficients by the Bonferroni method.
  ($g = 2$)

$$B = t\left(1 - 0.1/2(2)\right) = 2.306$$

$$s\{b_1'\} = \left(\frac{1}{0.4}\right)s\{b_1\} = 31.68$$

$$s\{b_2'\} = \left(\frac{1}{10}\right)s\{b_2\} = 1.267$$

$$\Rightarrow -212.6 \leq \beta_1 \leq -66.5 \quad 4.6 \leq \beta_2 \leq 10.5$$

Case Example: the life of power cell, cont'd

$$-212.6 \le \beta_1 \le -66.5 \qquad 4.6 \le \beta_2 \le 10.5$$

- With confidence 0.90, we conclude that the mean number of charge/discharge cycles before failure decreases by 66.5 to 212.6 cycles with a unit increase in the charge rate for given ambient temperature, and increase by 4.6 to 10.5 cycles with a unit increase of ambient temperature for given charge rate.

```
> g = length(coef(lm(Y ~ X1+X2)) ) - 1
> alpha=0.1
> confint( lm(Y ~ X1+X2), level = 1-(alpha/g) )
                 2.5 %    97.5 %
(Intercept)   64.617930 256.54874
X1          -212.602056 -66.56461
X2             4.629251  10.47075
```

Some further Comments on polynomial regression

Drawbacks: polynomial models

- Such models can be more expensive in degrees of freedom than alternative nonlinear models or linear models with transformed variables.
- serious multicollinearity may be present even when the predictor variables are centered.
- An alternative to using centered variables is to use orthogonal polynomials. Orthogonal polynomials are uncorrelated. (Some computer packages used orthogonal polynomials).
- Sometimes a quadratic response function is fitted for the purpose of establishing the linearity of the response function when repeat observations are not available.

Interaction Regression Models

Interaction effects:

- A regression model with p − 1 predictor variables contains additive effects:

$$E\{Y\} = f_1(X_1) + f_2(X_2) + \ldots + f_{p-1}(X_{p-1})$$

- f_i , i = 1, . . . , p − 1: any functions, not necessarily simple one
- Illustration:

$$E\{Y\} = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_1^2}_{f_1(X_1)} + \underbrace{\beta_3 X_2}_{f_2(X_2)}$$

$\Rightarrow$ the effects of X_1, X_2 on Y are additive

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 \quad (\times)$$

Interaction Regression Models, cont'd

- a cross-product term: modelling the interaction effect of two predictor variables on the response variable; ex: $\beta_3 X_1 X_2$
- The cross-product term: called an interaction term; a linear-by-linear or a bilinear interaction term

Interaction Regression Models, cont'd

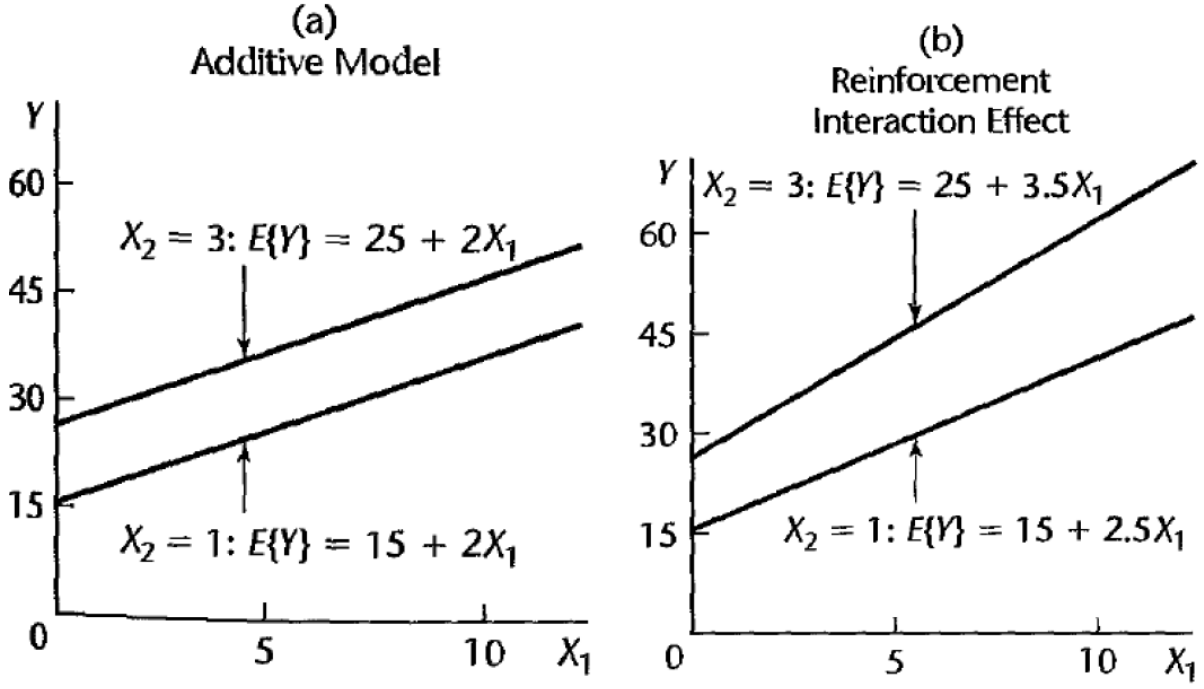
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

- The meaning of β_1, β_2 is not the same as that given earlier ($\because \beta_3 X_{i1} X_{i2}$)
- β_1, β_2 no longer indicate the change in the mean response with a unit increase of the predictor variable with the other predictor variable held constant.
- the change in the mean response with a unit increase in X_1 when X_2 is held constant:

$$\frac{\partial E\{Y\}}{\partial X_1} = \beta_1 + \beta_3 X_2 \quad \text{(depend on } X_2\text{)}$$

Interaction Regression Models, cont'd

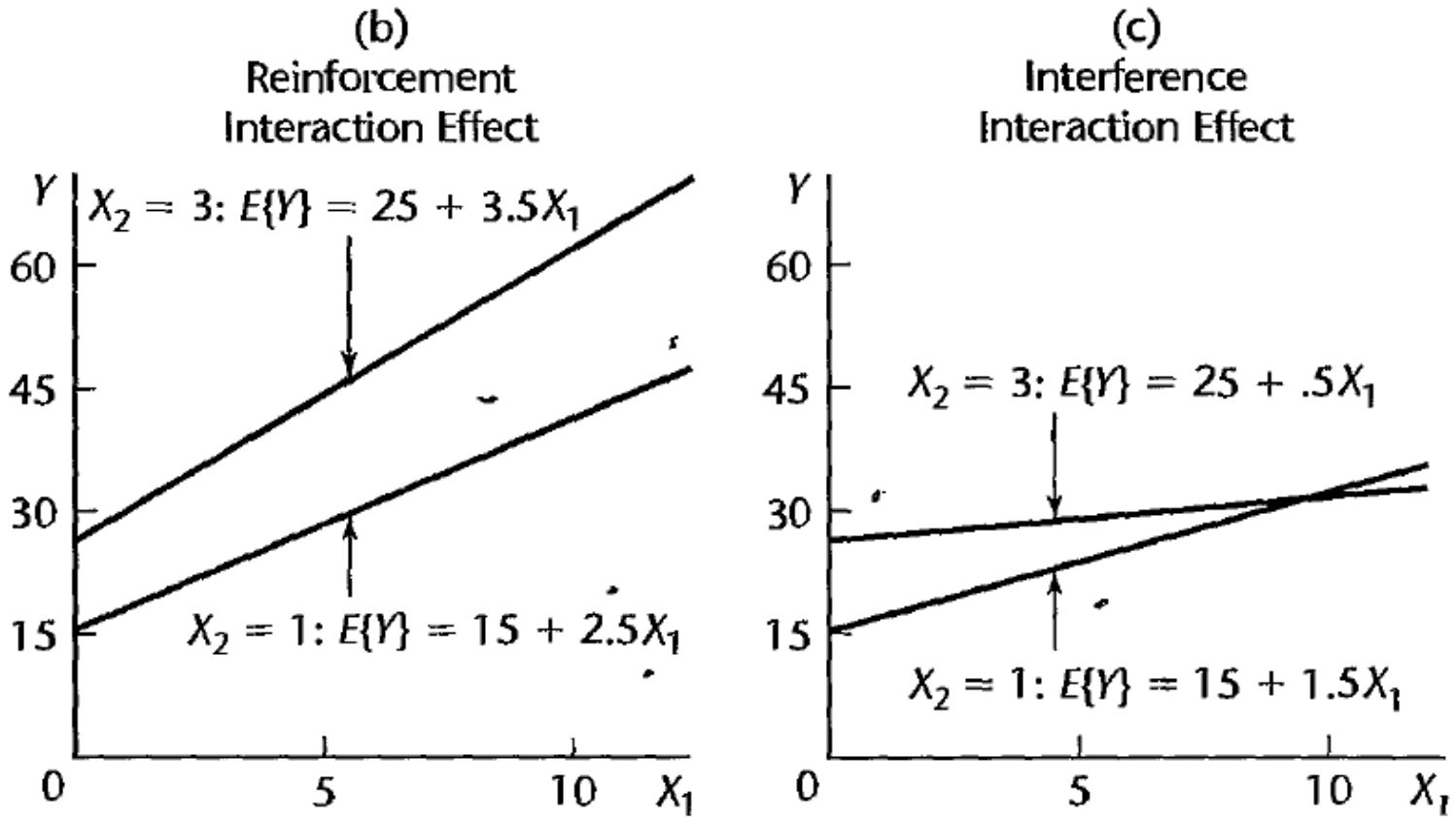
- (a): parallel; condition effect plot
- (b): the slope (2.5, 3.5) of the response function vs. X_1 differ for $X_2 = 1, 3$



$$E\{Y\} = 10 + 2X_1 + 5X_2$$
$$E\{Y\} = 10 + 2X_1 + 5X_2 + .5X_1X_2$$

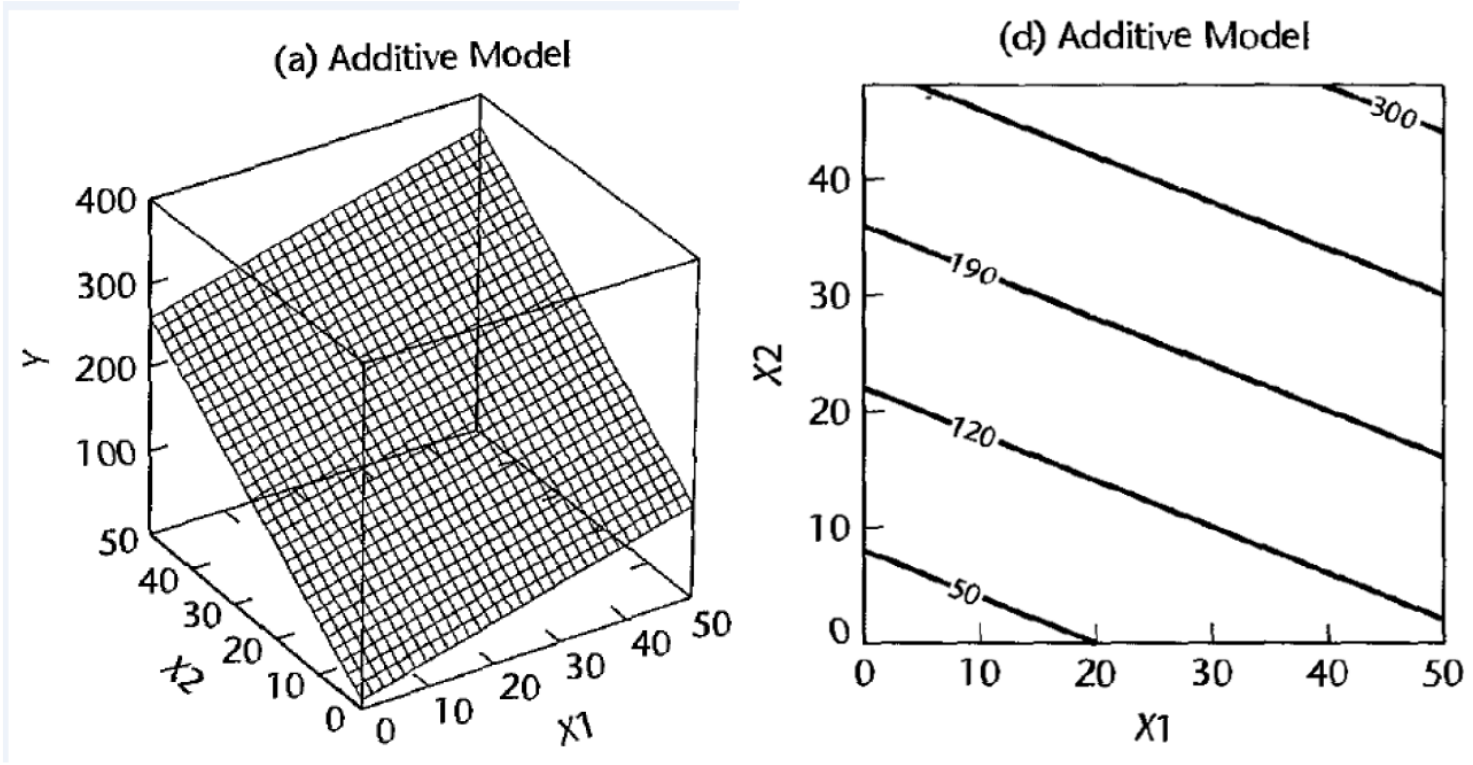
Interaction Regression Models, cont'd

- β_1, β_2 are positive:
  - the interaction effect is called a reinforcement or synergistic type
  - the interaction effect is called an interference or antagonistic type



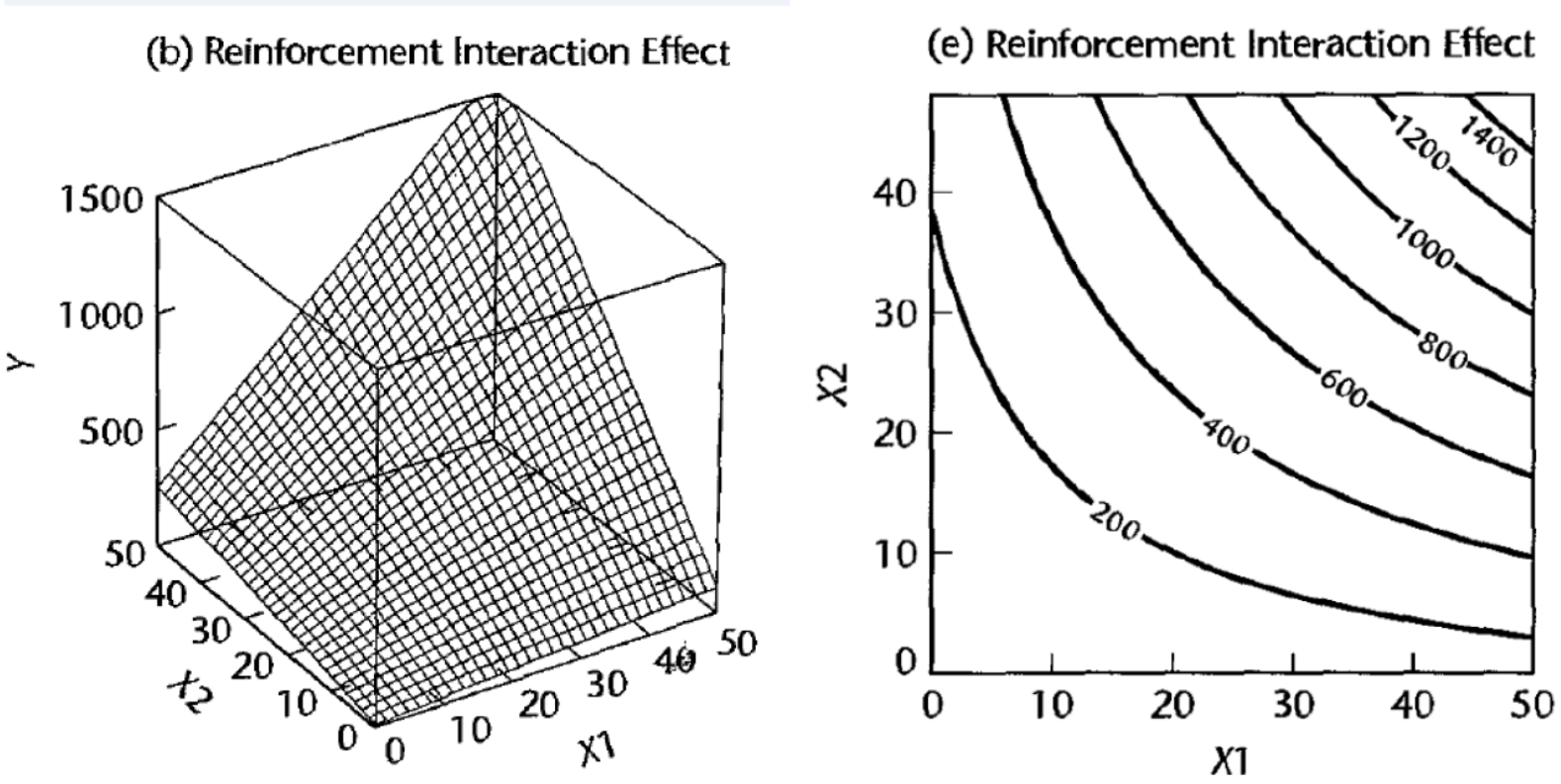
Shape of response function

$$E\{Y\} = 10 + 2X_1 + 5X_2$$



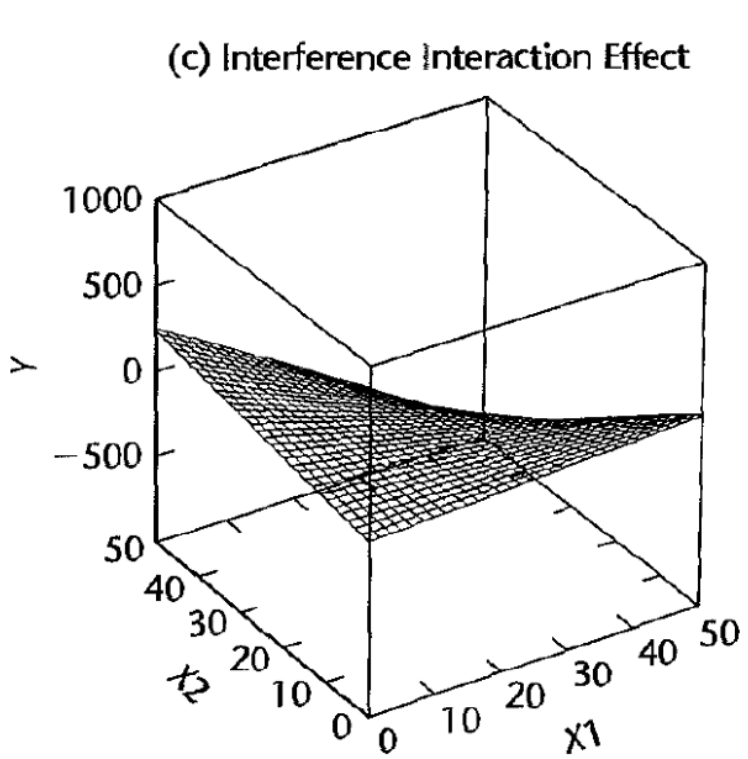
Shape of response function, cont'd

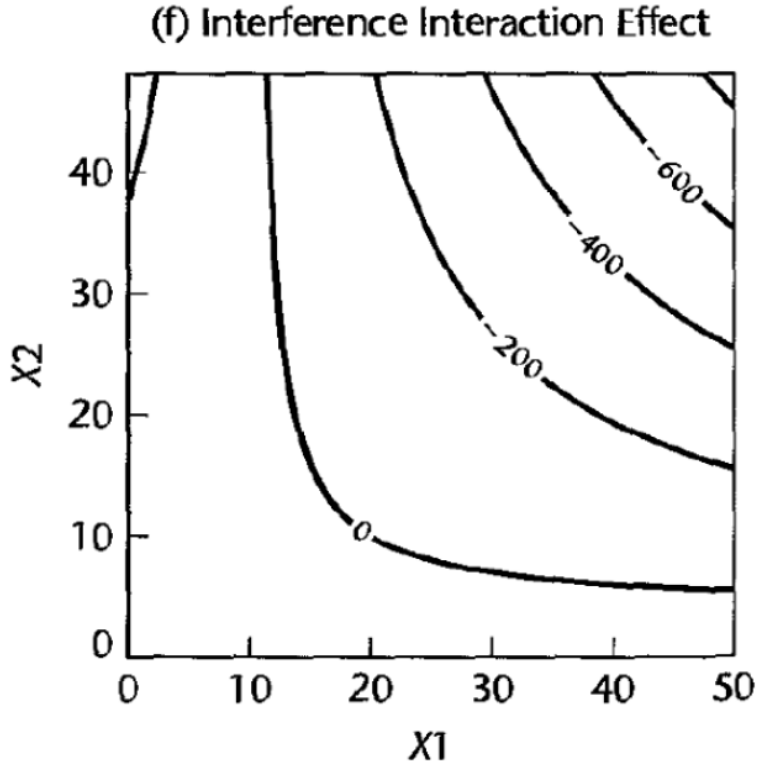
$$E\{Y\} = 10 + 2X_1 + 5X_2 + .5X_1X_2$$



Shape of response function, cont'd

$$E\{Y\} = 10 + 2X_1 + 5X_2 - .5X_1X_2$$





Shape of response function, cont'd

$$E\{Y\} = 65 + 3X_1 + 4X_2 - 10X_1^2 - 15X_2^2 + 35X_1X_2$$

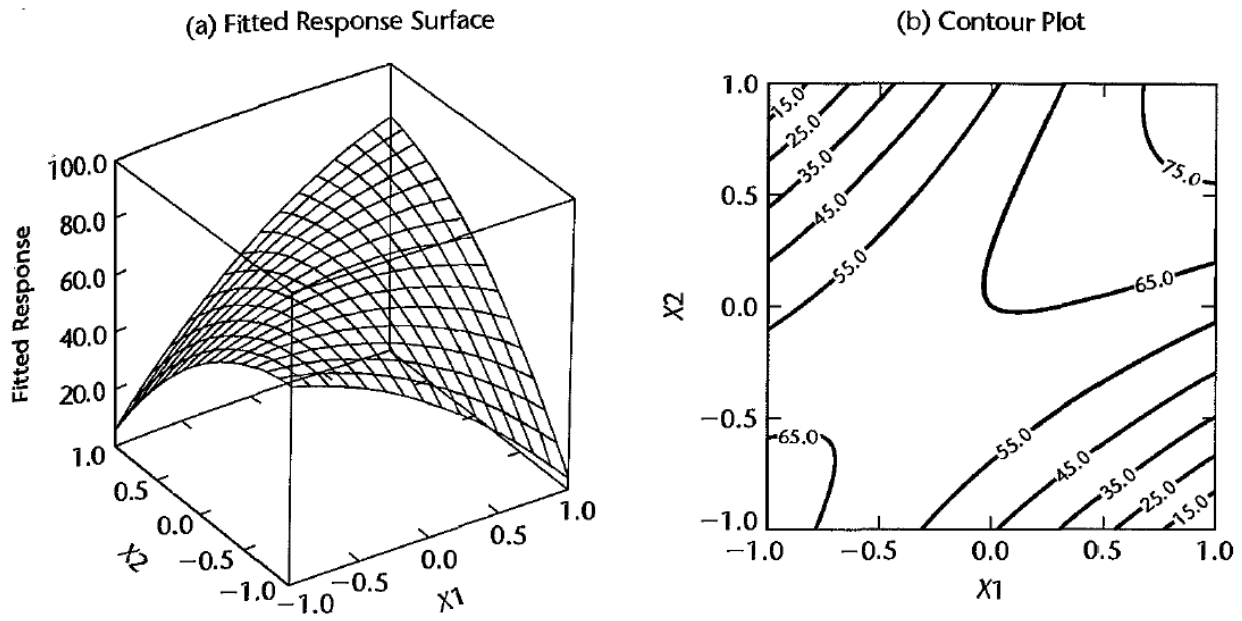


Figure : Response Surfaces and Contour Curves for Curvilinear Regression Model with Interaction Effect-Quick Bread Volume Example.

Curvilinear Effects

$$E\{Y\} = 65 + 3X_1 + 4X_2 - 10X_1^2 - 15X_2^2 + 35X_1X_2$$

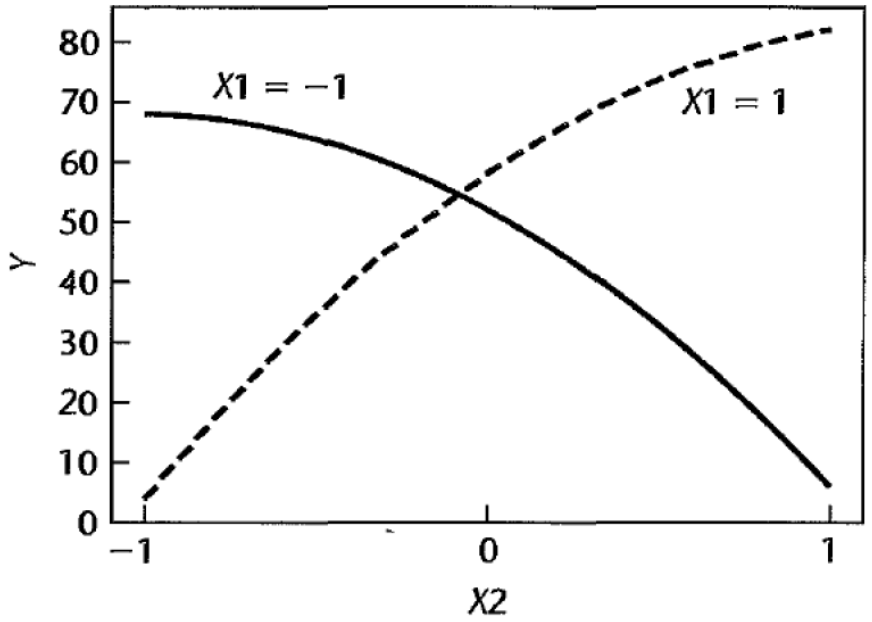


Figure : Conditional Effect Plot for Curvilinear Regression Model with Interaction Effect-Quick Bread Volume Example.

Implementation of interaction regression models

Considerations:

- high multicollinearities: center the predictor variables

$$x_{ik} = X_{ik} - \bar{X}_{ik}$$

- The number of predictor variables is large ⇒ the potential number of interaction terms becomes very large

- utilizing a priori knowledge

- plot the residuals for the additive regression model vs. the different interaction terms to determine which ones appear to be influential

Implementation of interaction regression models, cont'd

Body fat example:

- Model: Three predictor variables

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + \varepsilon_i$$

- Some of the predictor variables are highly correlated with some of the interaction terms

- Centered variables:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i1} x_{i2} + \beta_5 x_{i1} x_{i3} + \beta_6 x_{i2} x_{i3} + \varepsilon_i$$

$$\Rightarrow \quad \hat{Y} = 20.53 + 3.438 x_1 - 2.095 x_2 - 1.616 x_3 + .00888 x_1 x_2 - .08479 x_1 x_3 + .09042 x_2 x_3$$

$$MSE = 6.745$$

Implementation of interaction regression models, cont'd

Variable	Extra Sum of Squares
x_1	$SSR(x_1) = 352.270$
x_2	$SSR(x_2\|x_1) = 33.169$
x_3	$SSR(x_3\|x_1, x_2) = 11.546$
x_1x_2	$SSR(x_1x_2\|x_1, x_2, x_3) = 1.496$
x_1x_3	$SSR(x_1x_3\|x_1, x_2, x_3, x_1x_2) = 2.704$
x_2x_3	$SSR(x_2x_3\|x_1, x_2, x_3, x_1x_2, x_1x_3) = 6.515$

$$H_0: \beta_4 = \beta_5 = \beta_6$$

H_a: not all βs in H_0 equal to zero

$$\Rightarrow F^* = \frac{SSR(x_1x_2, x_1x_3, x_2x_3|x_1, x_2, x_3)}{3} \div MSE$$

= 0.53 ≤ F(0.95; 3,13) = 3.41

⇒ conclude H_0:the interaction terms are not needed in the regression model. (*P*–value=0.67)

Qualitative Predictors

- Field: (Qualitative variables) business, economics, the social, biological sciences

- Examples:
  - gender (M,F);
  - purchase status: purchase, no purchase
  - Disability status: not; partial; fully

- A study of innovation in the insurance industry
  - related the speed with which a particular insurance innovation is adopted (Y)
  - the size of the insurance firm (X_1)
  - the type of the firm (X_2)

Qualitative Predictors, cont'd

- Qualitative predictor with two classes:
  - indicator variable: take on 0 and 1
  - Ex: two indicator variables X_2, X_3

$$X_2 = \begin{cases} 1 \text{ if stock company} \\ \\ 0 \text{ otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 \text{ if mutual company} \\ \\ 0 \text{ otherwise} \end{cases}$$

$$\Rightarrow \quad Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

Qualitative Predictors, cont'd

- intuitive approach of setting up an indicator variable: leads to computation difficulties
- Design matrix $\boldsymbol{X}$ with $n = 4$:

$$\boldsymbol{X} = \begin{array}{c} \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccc} X_1 & & X_2 & X_3 \end{array} \\ \begin{bmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{bmatrix} \end{array}$$

$$\boldsymbol{X}'\boldsymbol{X} = \begin{bmatrix} 4 & \sum X_{i=1}^{4} X_{i1} & 2 & 2 \\ \sum_{i=1}^{4} X_{i1} & \sum_{i=1}^{4} X_{i1}^2 & \sum_{i=1}^{2} X_{i1} & \sum_{i=3}^{4} X_{i1} \\ 2 & \sum_{i=1}^{2} X_{i1} & 2 & 0 \\ 2 & \sum_{i=3}^{4} X_{i1} & 0 & 2 \end{bmatrix}$$

  - $\boldsymbol{X}$: the first column = the sum of the last two columns$\Rightarrow$ linearly dependent

- $\boldsymbol{X}'\boldsymbol{X}$: no inverse

Qualitative Predictors, cont'd

Simple way out of the difficulty: drop one indicator variable

Principle:

A qualitative variable with *c* classes will be represented by $c - 1$ indicator variables, each taking on the values 0 and 1.

- Indicator variables: called dummy variables or binary variables

Drop X_3 $\Rightarrow$ $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

$\Rightarrow$ $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

Qualitative Predictors, cont'd

Meaning of the regression coefficients:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

- X_2=0

$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(0) = \beta_0 + \beta_1 X_1$ Mutual Firms

- X_2=1

$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 X_1$ Stock Firms

Qualitative Predictors, cont'd

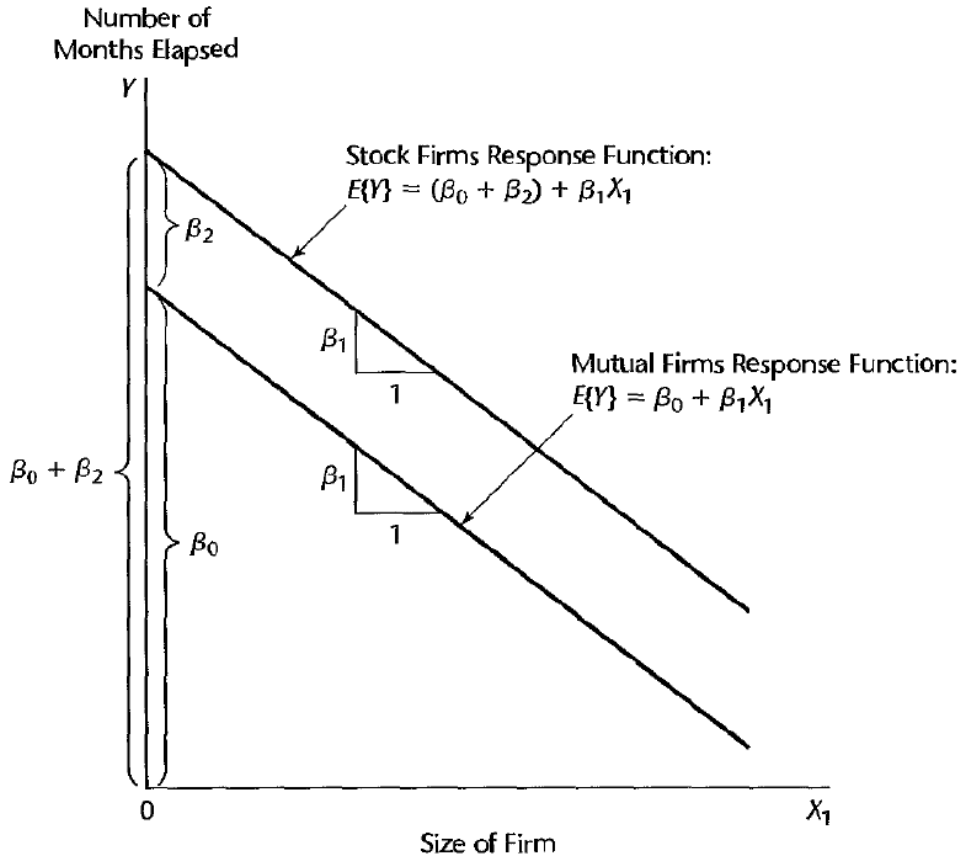


Figure : Illustration of Meaning of Regression Coefficient for Regression Model (8.33) with Indicator Variable X_2-Insurance Innovation Example.

Qualitative Predictors, cont'd

Firm i	(1) Number of Months Elapsed Y_i	(2) Size of Firm (million dollars) X_{i1}	(3) Type of Firm	(4) Indicator Code X_{i2}	(5) $X_{i1}X_{i2}$
1	17	151	Mutual	0	0
2	26	92	Mutual	0	0
3	21	175	Mutual	0	0
4	30	31	Mutual	0	0
5	22	104	Mutual	0	0
6	0	277	Mutual	0	0
7	12	210	Mutual	0	0
8	19	120	Mutual	0	0
9	4	290	Mutual	0	0
10	16	238	Mutual	0	0
11	28	164	Stock	1	164
12	15	272	Stock	1	272
13	11	295	Stock	1	295
14	38	68	Stock	1	68
15	31	85	Stock	1	85
16	21	224	Stock	1	224
17	20	166	Stock	1	166
18	13	305	Stock	1	305
19	30	124	Stock	1	124
20	14	246	Stock	1	246

Qualitative Predictors, cont'd

- studied 10 mutual firms and 10 stock firms
- The fitted of regression model:

$$\hat{Y} = 33.87407 - 0.10174X_1 + 8.05547X_2$$

- Interested in the effect of type of firm (X_2)
- a 95% confidence interval for β_2:

$$4.98 \leq \beta_2 \leq 11.13 \quad (t(0.975; 17) = 2.110)$$

- Test:

$$H_0 : \beta_2 = 0 \quad \text{vs.} H_a : \beta_2 \neq 0$$

$\Rightarrow$ lead to H_a: type of firm has an effect ($\alpha = 0.05$)

Qualitative Predictors, cont'd

Figure : Regression Results for Fit of Regression Model (8.33)-Insurance Innovation Example.

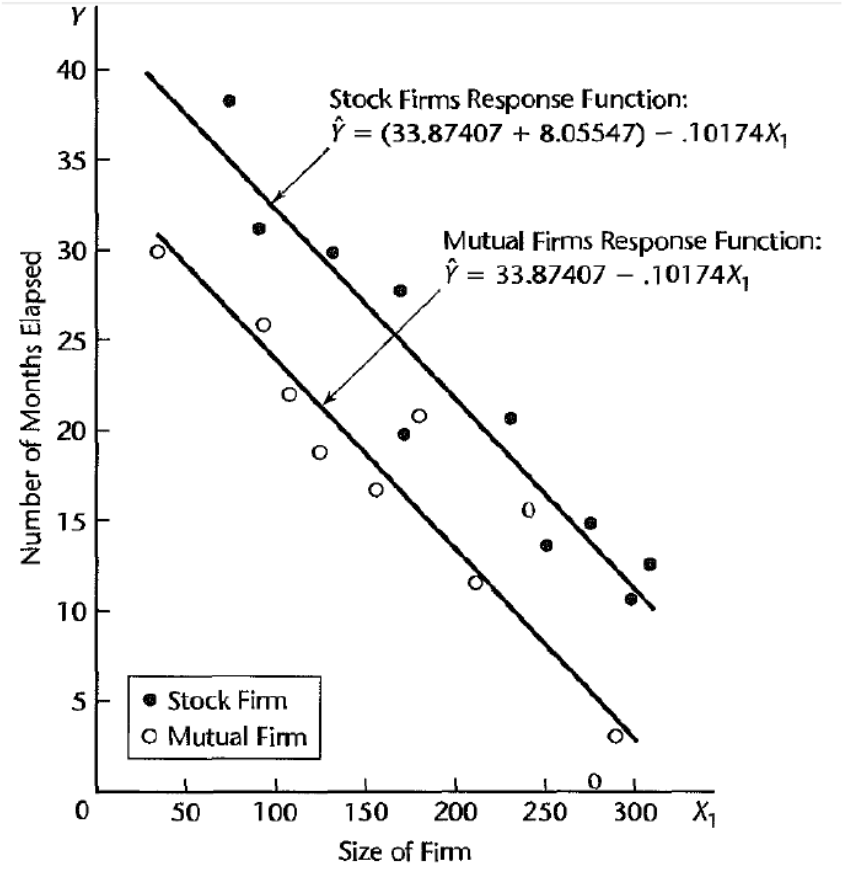
(a) Regression Coefficients

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
β_0	33.87407	1.81386	18.68
β_1	−.10174	.00889	−11.44
β_2	8.05547	1.45911	5.52

(b) Analysis of Variance

Source of Variation	*SS*	*df*	*MS*
Regression	1,504.41	2	752.20
Error	176.39	17	10.38
Total	1,680.80	19	

Qualitative Predictors, cont'd



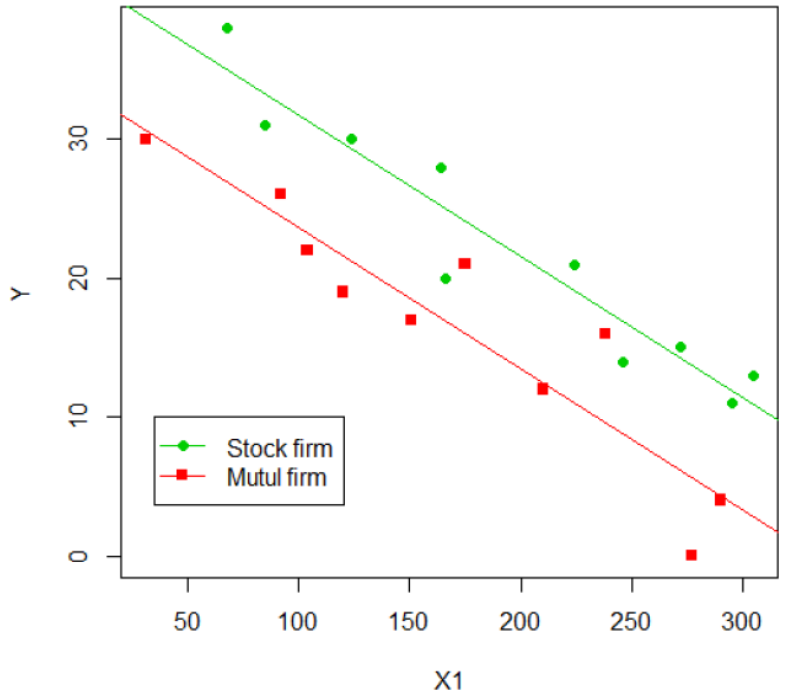


Figure : Fitted Regression Function for Regression Model (8.33) Indicator Innovation Example.

Qualitative Predictors, cont’d

```
ex<-Dataset_08TA02
attach(ex)
cX2<-as.factor(X2)
fit<-lm(Y~X1 +cX2)
coef <- coefficients(fit)
plot( Y ~ X1, pch=(15+X2),col = as.character(2+X2))
abline(coef["(Intercept)"] + coef["cX21"], coef["X1"], col = 3)
abline(coef["(Intercept)"] , coef["X1"], col = 2)
legend(35,10,c("Stock firm","Mutual firm"),col=c(3,2), lwd=1,lty=c(1,1), pch=c(16,15))
```

Qualitative Predictors, cont'd

Why did we not simply fit separate regressions for stock firms and mutual firms

1. The model assumes equal slopes and the same constant error term variance for each type of firm, the common slope β_1 can best be estimated by pooling the two types of firms
2. Other inferences, β_0, β_2, can be made more precisely by working with one regression model containing an indicator variable since more *df* associated with *MSE*.

More than two classes

- The regression of tool wear (Y) on tool speed X_1 and tool Model (qualitative: M1, M2, M3, M4)
- Require three indicator variables:

$$X_2 = \begin{cases} 1 & \text{if tool model M1} \\ 0 & \text{otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 & \text{if tool model M2} \\ 0 & \text{otherwise} \end{cases}$$

$$X_4 = \begin{cases} 1 & \text{if tool model M3} \\ 0 & \text{otherwise} \end{cases}$$

More than two classes, cont'd

First-order model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \varepsilon_i$$

$$\Rightarrow E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

$\Leftrightarrow$ Tool model M4: $E\{Y\} = \beta_0 + \beta_1 X_1$

Tool model M1: $E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1$

Tool model M2: $E\{Y\} = (\beta_0 + \beta_3) + \beta_1 X_1$

Tool model M3: $E\{Y\} = (\beta_0 + \beta_4) + \beta_1 X_1$

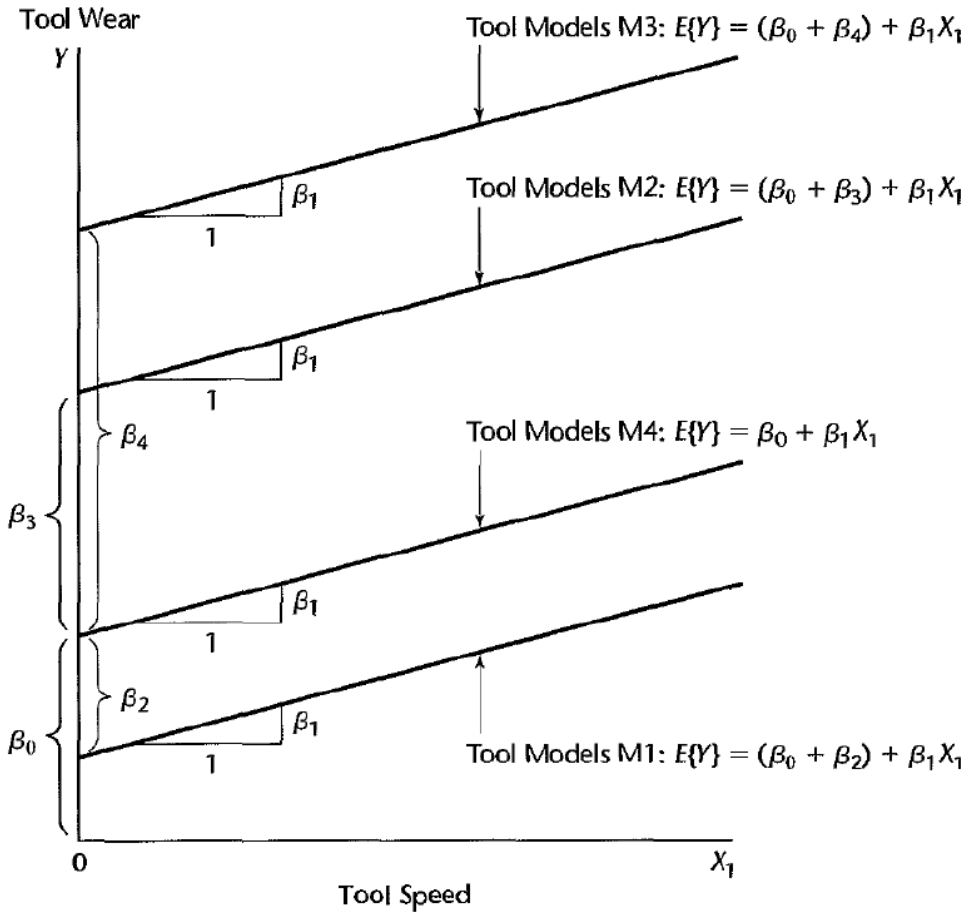
Tool Model	X_1	X_2	X_3	X_4
M1	X_{i1}	1	0	0
M2	X_{i1}	0	1	0
M3	X_{i1}	0	0	1
M4	X_{i1}	0	0	0

More than two classes, cont'd

- the regression of tool wear on tool speed is linear, with the same slope for all four tool models
- β_2, β_3, β_4: how much higher (lower) the response functions for toll models (M1,M2,M3) are than the one for M4
- always compared with the class for which $X_2 = X_3 = X_4 = 0$

More than two classes, cont'd

Figure : Illustration of Regression Model (8.36) - Tool Wear Example.



More than two classes, cont'd

- wish to estimate differential effects other than against tool models M4:

$\Rightarrow$ can be done by estimating differences between regression coefficients (e.g. $\beta_4 - \beta_3$)

- The point estimator is $b_4 - b_3$
- the estimated variance of this estimator:

$$s^2\{b_4 - b_3\} = s^2\{b_4\} + s^2\{b_3\} - 2s\{b_4, b_3\}$$

Time Series Applications

- Economists; business analysis
- Using indicator variables:

  - year: peacetime; wartime

$$Y_t = \beta_0 + \beta_1 X_{t1} + \beta_2 X_{t2} + \varepsilon_t, \quad t = 1, \ldots, n$$

where $X_{t1} = \textit{income}$

$$X_{t2} = \begin{cases} 1 & \text{if period } t \text{ peace time} \\ 0 & \textit{otherwise} \end{cases}$$

  - monthly or quarterly: seasonal effect
- Time series data: susceptible to correlated error terms (Chap. 12)

Some Considerations in Using Indicator Variables

Allocated codes: arbitrary; other numbers;

- define a metric for the classes of the qualitative variable

Class	X_1	$E\{Y\}$
Frequent user	3	$E\{Y\} = \beta_0 + \beta_1(3) = \beta_0 + 3\beta_1$
Occasional user	2	$E\{Y\} = \beta_0 + \beta_1(2) = \beta_0 + 2\beta_1$
Nonuser	1	$E\{Y\} = \beta_0 + \beta_1(1) = \beta_0 + \beta_1$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$$

The key implication:

$$E\{Y|frequent\ user\} - E\{Y|ocasional\ user\}$$
$$= E\{Y|ocasional\ user\} - E\{Y|nonuser\} = \beta_1$$

Some Considerations in Using Indicator Variables, cont'd

Indicator variables: make no assumptions about the spacing of the classes;

Class	X_1	X_2
Frequent user	1	0
Occasional user	0	1
Nonuser	0	0

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

β_1: $E\{Y|\textit{frequent user}\} - E\{Y|\textit{nonuser}\}$

β_2: $E\{Y|\textit{ocasional user}\} - E\{Y|\textit{nonuser}\}$

- $\beta_1 - \beta_2$: the differential effect between frequent user and occasional user
- if $\beta_1 = 2\beta_2 \Rightarrow$ equal spacing between the three classes

Some Considerations in Using Indicator Variables, cont'd

- Indicator variables can be used even if the predictor variable is quantitative.
  - transformed by grouping into classes
  - age: 21, 21-34, 35-49

- information about the original variable may be thrown away

- additional parameters into the model: reducing the df associated with *MSE*

Some Considerations in Using Indicator Variables, cont'd

Other codings:

- First coding:

$$X_{i2} = \begin{cases} 1 & \text{if stock company} \\ -1 & \text{if mutual company} \end{cases}$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$$

$$E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1 \quad \text{Stock firms}$$

$$E\{Y\} = (\beta_0 - \beta_2) + \beta_1 X_1 \quad \text{Mutual firms}$$

$\Rightarrow \beta_0$: "average" intercept of the regression line

- two regression lines are the same:

$$H_0 : \beta_2 = 0 \text{ vs. } H_a : \beta_2 \neq 0$$

Some Considerations in Using Indicator Variables, cont'd

- Second coding:
  - using indicator variables for each of the c classes of the qualitative variable
  - drop the intercept term

$$Y_i = \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$$

where: $X_{i1} = \textit{size of firm}$; $X_{i2} = \begin{cases} 1 & \textit{if stock company} \\ 0 & \textit{otherwise} \end{cases}$

$$X_{i3} = \begin{cases} 1 & \textit{if mutual company} \\ 0 & \textit{otherwise} \end{cases}$$

$$\Rightarrow E\{Y\} = \beta_2 + \beta_1 X_1 \quad \textit{Stock firms}$$

$$E\{Y\} = \beta_3 + \beta_1 X_1 \quad \textit{Mutual firms}$$

- The same regression line: $H_0 : \beta_2 = \beta_3$ vs. $H_a : \beta_2 \neq \beta_3$

Modeling Interactions between Quantitative and Qualitative Predictors

- the possibility of interaction effects: the size of firm and type of firm

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

where $X_{i1} = \textit{size of firm}$

$$X_{i2} = \begin{cases} 1 & \textit{if stock company} \\ 0 & \textit{otherwise} \end{cases}$$

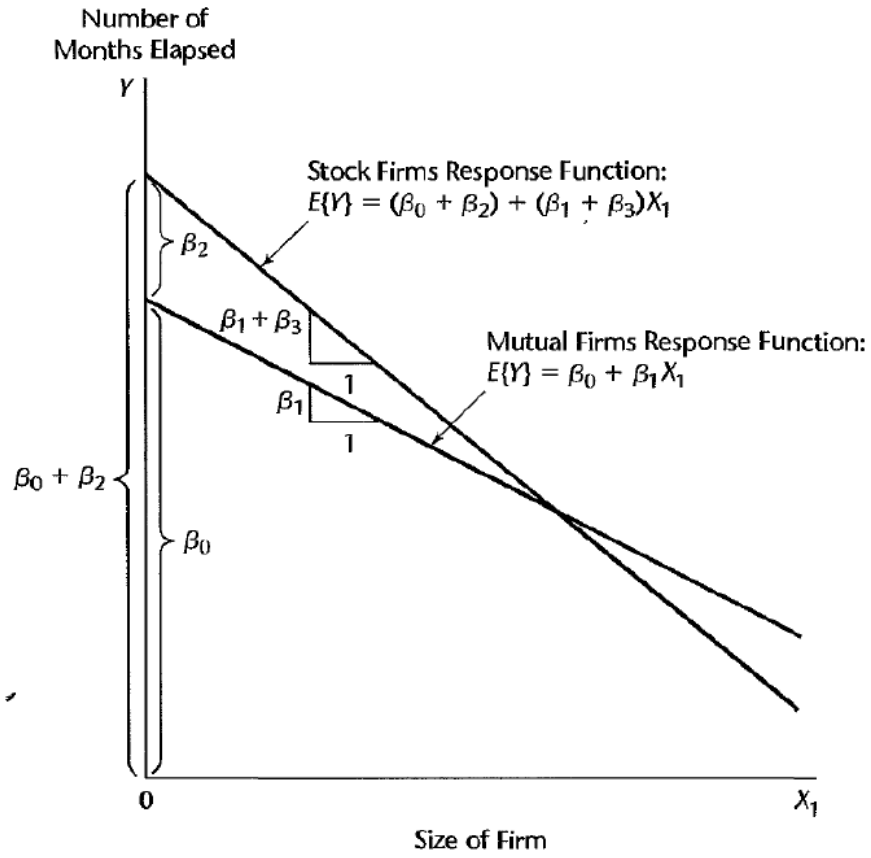
$$\Rightarrow E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$$

$$\Leftrightarrow E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 X_1 \ \textit{Mutual firms}$$

$$E\{Y\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) X_1 \ \textit{Stock firms}$$

Modeling Interactions between Quantitative and Qualitative Predictors, cont'd

disordinal interaction



- Both the intercept and the slope differ for the two classed in the response function
- The effect of the qualitative predictor variable can be studied only by comparing the regression functions within the scope of the model for the different classed of the qualitative variable

Modeling Interactions between Quantitative and Qualitative Predictors, cont'd

ordinal interaction

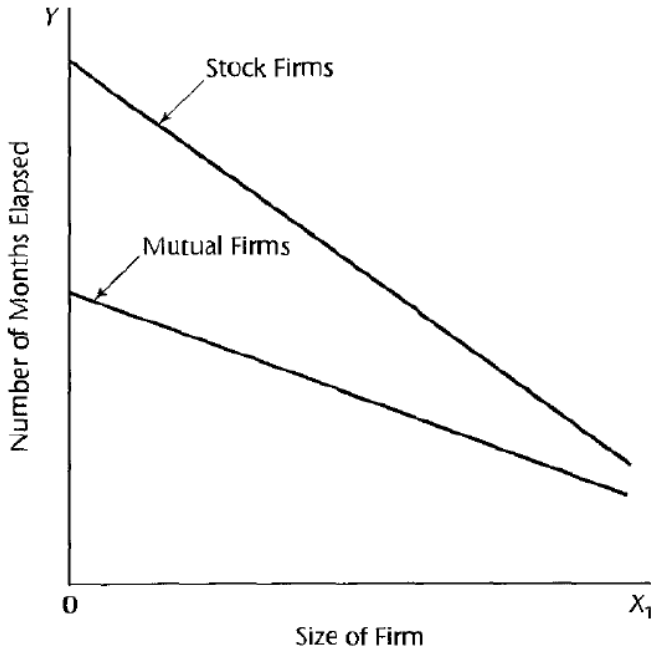


Figure : Another Illustration of Regression Model (8.49) with Indicator Variable X_2 and Interaction Term- Insurance Innovation Example.

Modeling Interactions between Quantitative and Qualitative Predictors, cont'd

Example: Insurance Innovation

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i1} X_{i2} + \varepsilon_i$$

Figure : Regression Result for Fit of Regression Model (8.49) with Interaction Term - Insurance Innovation Example.

(a) Regression Coefficients

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation	t^*
β_0	33.83837	2.44065	13.86
β_1	−.10153	.01305	−7.78
β_2	8.13125	3.65405	2.23
β_3	−.0004171	.01833	−.02

(b) Analysis of Variance

Source of Variation	*SS*	*df*	*MS*
Regression	1,504.42	3	501.47
Error	176.38	16	11.02
Total	1,680.80	19	

Modeling Interactions between Quantitative and Qualitative Predictors, cont'd

Example: Insurance Innovation

$$H_0 : \beta_3 = 0 \text{ vs. } H_a : \beta_3 \neq 0 \Rightarrow |t^*| = 0.02 < t(0.975; 16)$$

$\Rightarrow$ no interaction effects

```
> summary(lm(Y~X1+X2+X1*X2))
Call:
lm(formula = Y ~ X1 + X2 + X1 * X2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7144 -1.7064 -0.4557  1.9311  6.3259

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 33.8383695  2.4406498  13.864 2.47e-10 ***
X1          -0.1015306  0.0130525  -7.779 7.97e-07 ***
X2           8.1312501  3.6540517   2.225   0.0408 *
X1:X2       -0.0004171  0.0183312  -0.023   0.9821
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.32 on 16 degrees of freedom
Multiple R-squared:  0.8951,    Adjusted R-squared:  0.8754
F-statistic: 45.49 on 3 and 16 DF,  p-value: 4.675e-08
```

More Complex Models

- two or more of the predictor are qualitative

$$X_2 = \begin{cases} 1 & \textit{if firm incorporated} \\ 0 & \textit{otherwise} \end{cases}$$

$$X_3 = \begin{cases} 1 & \textit{if quality of sales management high} \\ 0 & \textit{otherwise} \end{cases}$$

first-order: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$

interaction: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i1} X_{i2} + \beta_5 X_{i1} X_{i3} + \beta_6 X_{i2} X_{i3} + \varepsilon_i$

Type of Firm	Quality of Sales Management	Response Function
Incorporated	High	$E\{Y\} = (\beta_0 + \beta_2 + \beta_3 + \beta_6) + (\beta_1 + \beta_4 + \beta_5)X_1$
Not incorporated	High	$E\{Y\} = (\beta_0 + \beta_3) + (\beta_1 + \beta_5)X_1$
Incorporated	Low	$E\{Y\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_4)X_1$
Not incorporated	Low	$E\{Y\} = \beta_0 + \beta_1 X_1$

More Complex Models, cont'd

- Qualitative predictor variables only

$$Y_i = \beta_0 + \beta_2 X_{i2} + \beta_3 X_{i3}$$

- all explanatory variables are qualitative: *Analysis Of Variance (ANOVA) models*

- some qualitative & some qualitative explanatory variables: *Analysis Of Covariance (ANCOVA) models*

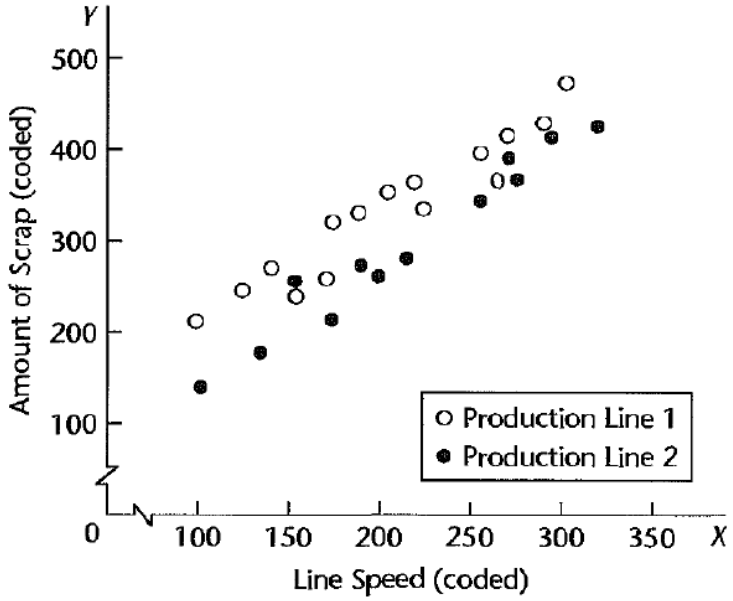
Comparison of Two or More Regression Functions

Three examples:

- two production lines for making soap bars
- family income study from urban and rural areas
- instrument calibration study

Production Lines for Making Soap Bar

- Two production lines(X_2): X_1 =production line speed & Y = the amount of the scrap
- linear but not the same for the two production lines: the same slopes & the different height
- A formal test is desired to determine whether or not the two regression lines are identical.



Family Income Study

- modeled by linear regression

- wish to compare whether, at given income level, urban and rural families tend to save the same amount — i.e. whether the slopes of the two regression lines are the same

Instrument Calibration Study

- Two instruments were constructed
- relation between gauge readings & actual pressures
- If the two regression lines are the same, a single calibration schedule can be developed for the two instruments; otherwise, two different calibration schedules will be required.