Chapter 8

Quantitative and Qualitative Predictors

Instructor: Li C.X.

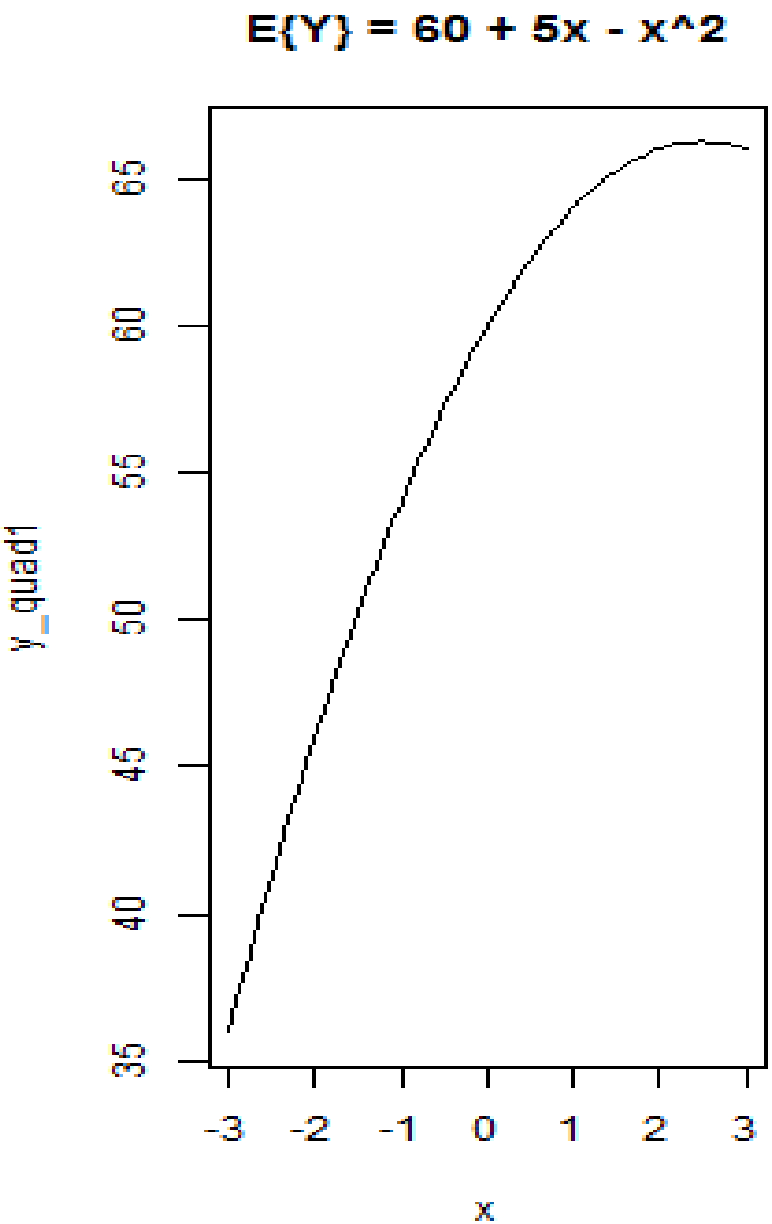
Outline

- Two types of predictors
  - Quantitative
  - Qualitative

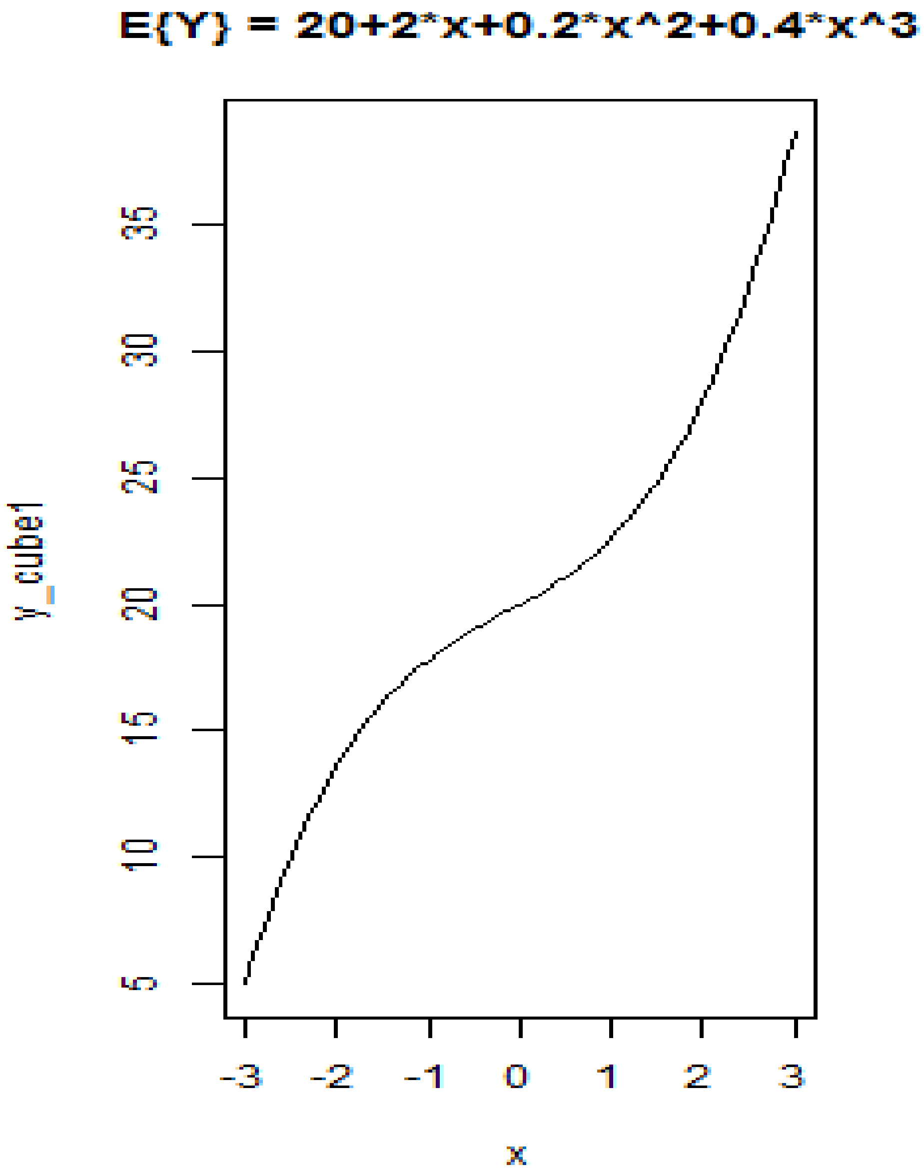
- Models
  - Polynomial Regression Models
  - Interaction Regression Models

8.1 Polynomial Regression Models

- Useful in 2 Settings:
  - True relation between response and predictor is polynomial
  - True relation is complex nonlinear function that can be approximated by polynomial in specific range of X-levels
- Models with 1 Predictor: Including p polynomial terms
- 2nd order Model: $E\{Y\} = \beta_0 + \beta_1 x + \beta_2 x^2$, where $x = X - \overline{X}$
  - X is centered due to the possible high correlation between X and X^2.
  - β_0 is the mean response when x = 0.
  - β_1 is called the linear effect.
  - β_2 is called the quadratic effect.
- 3rd order Model: $E\{Y\} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$



E{Y} = 60 + 5x - x^2
y_quad1
35
40
45
50
55
60
65
-3
-2
-1
0
1
2
3
x



E{Y} = 20+2*x+0.2*x^2+0.4*x^3
y_cube1
5
10
15
20
25
30
35
-3
-2
-1
0
1
2
3
x

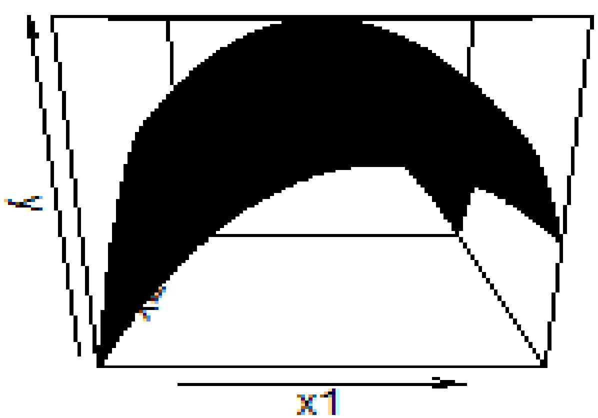
Polynomial Regression Models

- Response Surfaces with 2 (or more) predictors
  - 2^{nd} order model with 2 Predictors:

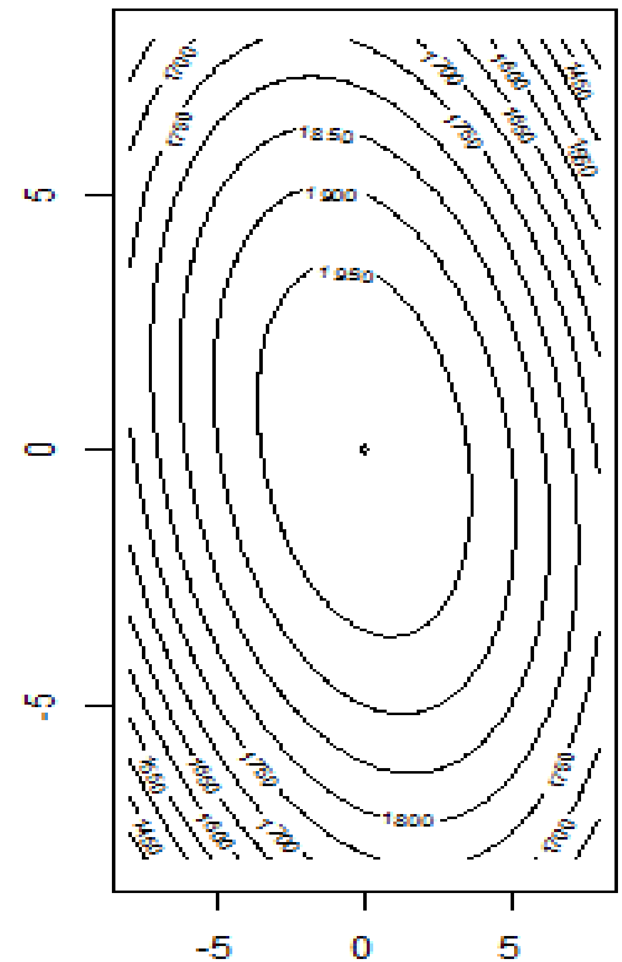
$$E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2$$

  - The coecient β_{12} is called the interaction effect coefficient.

E{Y} = 2000-4x1^2-4x2^2-2x1x2



E{Y} = 2000-4x1^2-4x2^2-2x1x2



Implementation of Polynomial Regression Models

- Fitting----Very easy, just use the least squares for multiple linear regressions since they can all be seen as a multiple regression.
- Determine the order----Very important step!
- e.g. $Y_i = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3 + \epsilon_i$
- Naturally, we want to test whether or not $\beta_{111} = 0$, or whether or not both $\beta_{11} = 0$ and $\beta_{111} = 0$.
- How to do the test?
  - Extra Sums of Squares and General linear test
  - To test $\mathrm{H}_0 : \beta_{111} = 0$

$$t^* = \frac{b_2}{s\{b_2\}} \quad \text{or} \quad F^* = \frac{SSR(x_3 \mid x_1, x_2)/1}{SSE(x_1, x_2, x_3)/(n-4)} = \frac{MSR(x_3 \mid x_1, x_2)}{MSE(x_1, x_2, x_3)}$$

  - To test $\mathrm{H}_0 : \beta_{11} = \beta_{111} = 0$

$$F^* = \frac{SSR(x_2, x_3 \mid x_1)/2}{SSE(x_1, x_2, x_3)/(n-4)} = \frac{MSR(x_2, x_3 \mid x_1)}{MSE(x_1, x_2, x_3)}$$

Modeling Strategies

- Use Extra Sums of Squares and General Linear Tests to compare models of increasing complexity (higher order)
- Use coding in fitting models (centered/scaled) predictors to reduce multicollinearity when conducting testing.
- Back-transform for plotting on original scale* (see below for quadratic)

Centered variables: $\hat{Y} = b_0 + b_1 x + b_{11} x^2 = b_0 + b_1\left(X - \overline{X}\right) + b_{11}\left(X - \overline{X}\right)^2$

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

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

$$\hat{Y} = b_o' + b_1' X + b_2' X^2$$

Example: Power Cell (p.300)

- Response variable is the life (in cycles) of a power cell
- Explanatory variables are
  - Charge rate (3 levels)
  - Temperature (3 levels)
- This is a designed experiment
- Standardizing the explanatory variables

$$x_{i1} = \frac{X_{i1} - \bar{X}_1}{.4} = \frac{X_{i1} - 1.0}{.4}$$

$$x_{i2} = \frac{X_{i2} - \bar{X}_2}{10} = \frac{X_{i2} - 20}{10}$$

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

Cell i	(1) Number of Cycles Y_i	(2) Charge Rate X_{i1}	(3) Temperature X_{i2}	(4) Coded Values x_{i1}	(5) x_{i2}	(6) x_{i1}^2	(7) x_{i2}^2	(8) $x_{i1}x_{i2}$
1	150	.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	.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	.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
		$\bar{X}_1 = 1.0$	$\bar{X}_2 = 20$					

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

- Using original data

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	337.72149	149.96163	2.25	0.0741
chrate	1	-539.51754	268.86033	-2.01	0.1011
temp	1	8.91711	9.18249	0.97	0.3761
chrate2	1	171.21711	127.12550	1.35	0.2359
temp2	1	-0.10605	0.20340	-0.52	0.6244
ct	1	2.87500	4.04677	0.71	0.5092

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	55366	11073	10.57	0.0109
Error	5	5240.43860	1048.08772		
Corrected Total	10	60606			

- Using standardized data

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	162.84211	16.60761	9.81	0.0002
schrate	1	-43.24831	10.23762	-4.22	0.0083
stemp	1	58.48205	10.23762	5.71	0.0023
schrate2	1	16.43684	12.20405	1.35	0.2359
stemp2	1	-6.36316	12.20405	-0.52	0.6244
sct	1	6.90000	9.71225	0.71	0.5092

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	55366	11073	10.57	0.0109
Error	5	5240.43860	1048.08772		
Corrected Total	10	60606			

8.2 Interaction Regression Models

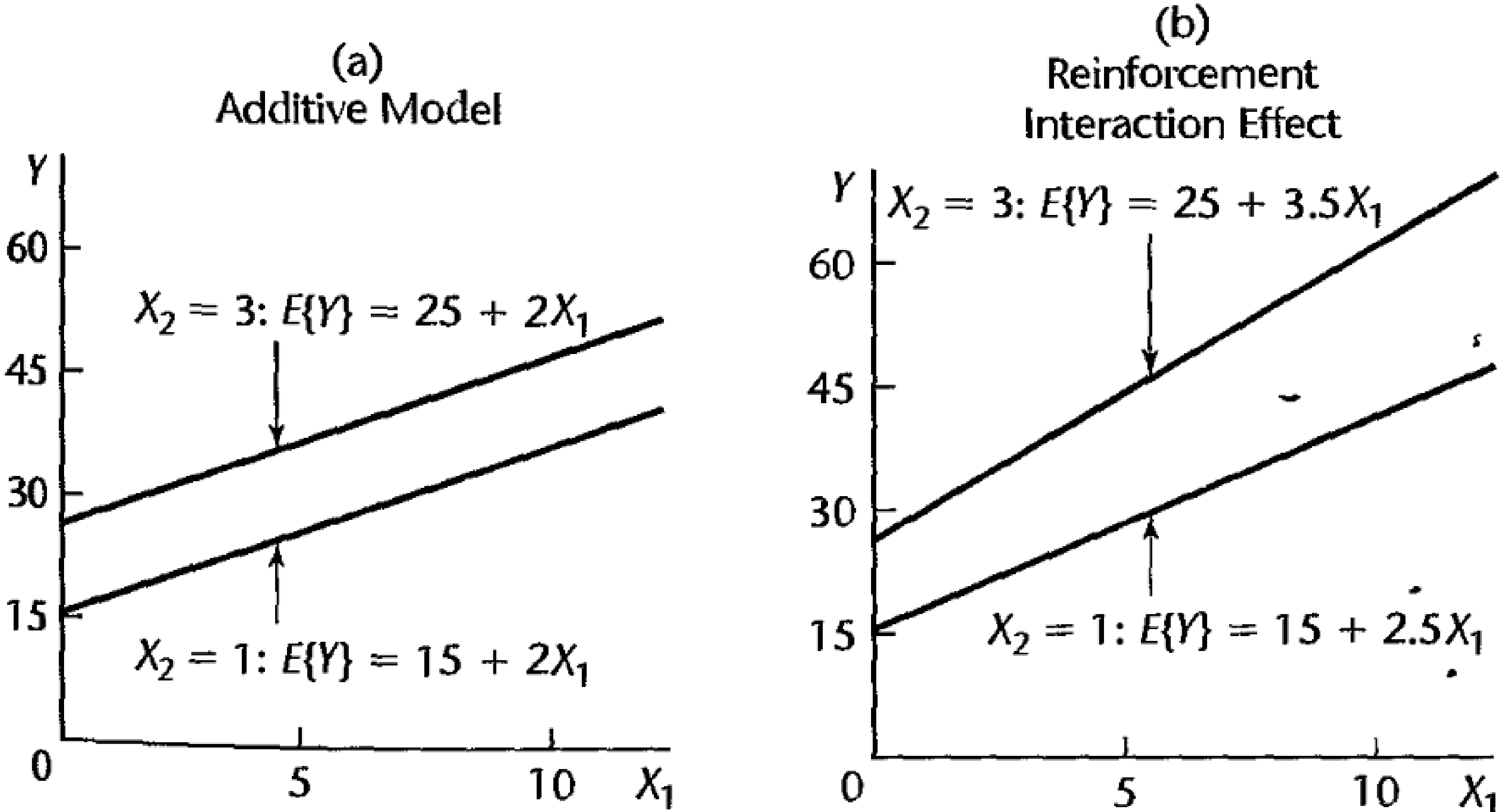
- Interaction $\Rightarrow$ Effect (Slope) of one predictor variable depends on the level other predictor variable(s)
- Formulated by including cross-product term(s) among predictor variables

- 2 Variable Models: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- The change in mean response with a unit increase in X_1 when X_2 is held constant is $\beta_1 + \beta_3 X_2$
- Similarly, a unit increase in X_2 when X_1 is constant is: $\beta_2 + \beta_3 X_1$

Type of interaction

- Reinforcement (synergistic) type:
- Conditional Effects Plot:

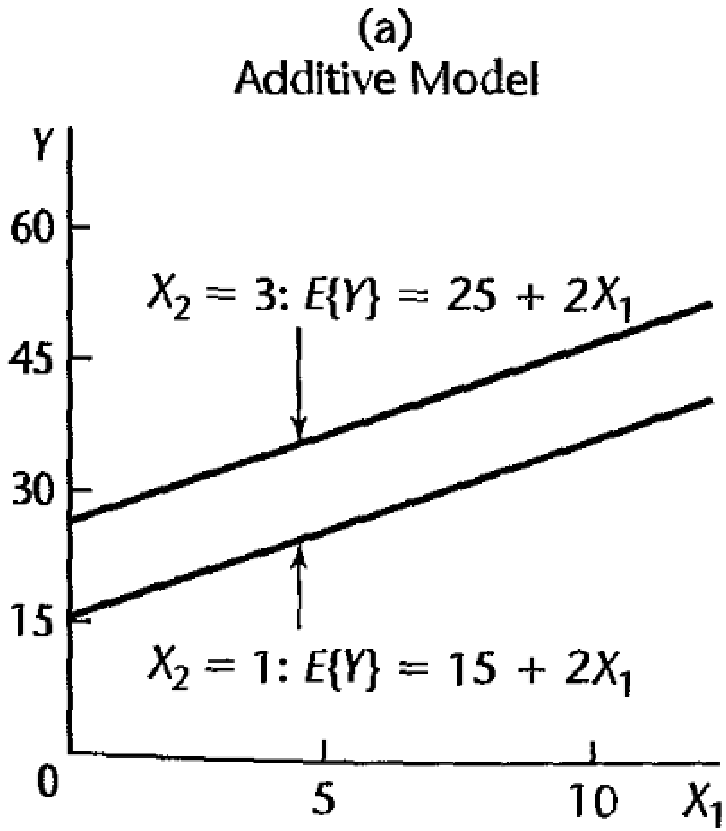
$$E\{Y\} = 10 + 2X_1 + 5X_2 + .5X_1X_2$$

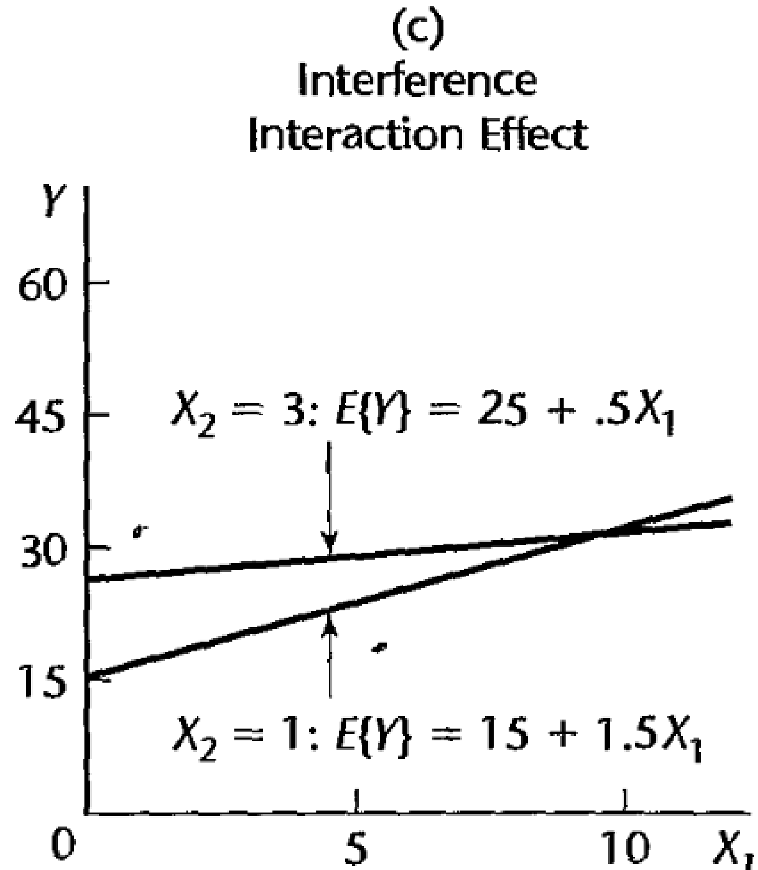


Type of interaction

- Interference (antagonistic) type:

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





8.3 Qualitative Predictors

- Often, we wish to include categorical variables as predictors (e.g. gender, region of country, ...)
- Trick: Create dummy (indicator) variable(s) to represent effects of levels of the categorical variables on response

- **A study of innovation in insurance industry**: related the speed with which a particular insurance innovation is adopted (Y) to the size of the insurance firm (X_1) and the type of the firm.
- Response Y : quantitative, continuous
- Predictor X_1: quantitative,
- Second predictor :type of firm(stock companies and mutual companies).

A study of innovation in insurance industry

- Predictors:

X_1=the size of the insurance firm

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

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

- Then, we have the model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \epsilon_i$$

Qualitative Predictor with Two Classes

- Suppose, we have n = 4 observations, the rst two being stock firms, the second two be mutual firms. Then

$$\mathbf{X} = \begin{pmatrix} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{pmatrix}$$

- Observation: first column is equal to the sum of the X2 and X3 columns, linear dependent…
- Problem: If variable has *c* categories, and we create *c* dummy variables, the model is not full rank when we include intercept
- Solution: Create $c - 1$ dummy variables, leaving one level as the control/baseline/reference category

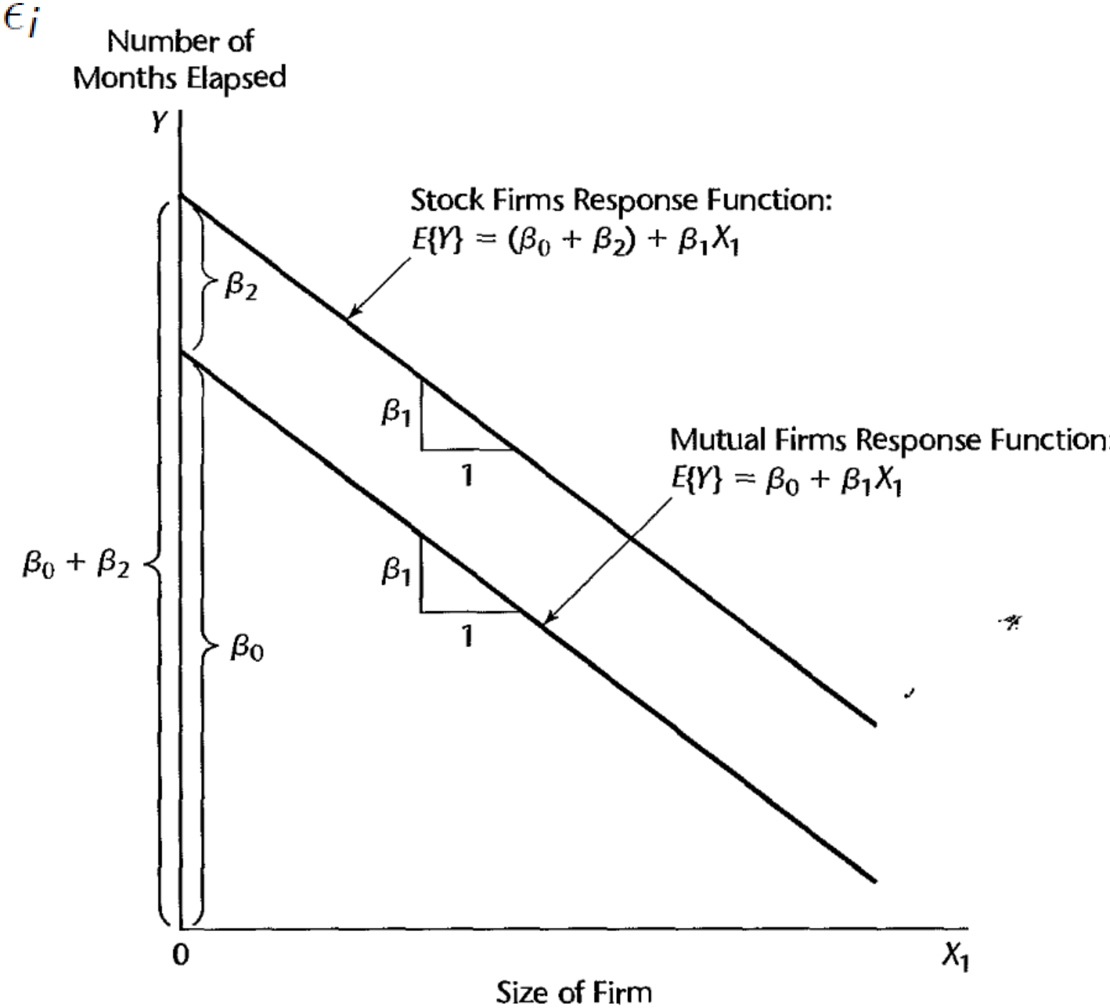
Qualitative Predictor with Two Classes

Now, we drop the X3 from the regression model:

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

X_1: the size of the firm

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



More than Two Classes

- Example – Salary vs Experience by Region
- Y: salary; Predictors: experience(X_1), Region (1,2,3)
- Solution, just use the Region 1 dummy (X_2) and the region 2 dummy (X_3), making Region 3 the "reference" region (Note: it is arbitrary which region is the reference)

$$Y = \text{salary},\ X_1 = \text{experience}$$

$$X_2 = \begin{cases} 1 \text{ if Region 1} \\ 0 \text{ otherwise} \end{cases} \qquad X_3 = \begin{cases} 1 \text{ if Region 2} \\ 0 \text{ otherwise} \end{cases}$$

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

Example – Salary vs Experience by Region

$X_1 = \text{experience} \quad X_2 = \begin{cases} 1 \text{ if Region 1} \\ 0 \text{ otherwise} \end{cases} \quad X_3 = \begin{cases} 1 \text{ if Region 2} \\ 0 \text{ otherwise} \end{cases}$

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

Region 1: $X_2 = 1, X_3 = 0 \quad \Rightarrow \quad E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3(0) = (\beta_0 + \beta_2) + \beta_1 X_1$

Region 2: $X_2 = 0, X_3 = 1 \quad \Rightarrow \quad E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(1) = (\beta_0 + \beta_3) + \beta_1 X_1$

Region 3: $X_2 = 0, X_3 = 0 \quad \Rightarrow \quad E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(0) + \beta_3(0) = \beta_0 + \beta_1 X_1$

$\beta_2 \equiv$ Difference between Regions 1 and 3, controlling for experience

$\beta_3 \equiv$ Difference between Regions 2 and 3, controlling for experience

$\beta_2 - \beta_3 \equiv$ Difference between Regions 1 and 2, controlling for experience

$\beta_2 = \beta_3 = 0 \quad \Rightarrow \quad$ No differences among Regions 1,2,3 wrt Salary, Controlling for Experience

- To test $H_0: \beta_2 = 0$ (no difference between regions 1 and 3)
  - t-test or partial F test (General linear test)

$$t^* = \frac{b_2}{s\{b_2\}} \qquad F^* = \frac{SSR(X_2 \mid X_1, X_3)/1}{SSE(X_1, X_2, X_3)/(n-4)} = \frac{MSR(X_2 \mid X_1, X_3)}{MSE(X_1, X_2, X_3)}$$

- To test $H_0: \beta_3 = 0$ (no difference between regions 2 and 3)
  - t-test or partial F test (General linear test)
- To test $H_0: \beta_2 = \beta_3$ (no difference between regions 1 and 2)
  - Full model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$
  - Reduced model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X'_{i2} + \varepsilon_i$, with $X'_{i2} = X_{i2} + X_{i3}$
  - General linear test: $$F^* = \frac{[SSE(R) - SSE(F)]/1}{SSE(F)/(n-4)}$$
- To test $H_0: \beta_2 = \beta_3 = 0$ (no difference between 3 regions)

$$F^* = \frac{SSR(X_2, X_3 \mid X_1)/2}{SSE(X_1, X_2, X_3)/(n-4)} = \frac{MSR(X_2, X_3 \mid X_1)}{MSE(X_1, X_2, X_3)}$$

8.4 Some Considerations in Using Indicator Variables

- An alternative: allocated codes.
- For example, the predictor variable "frequency of product use" has three classes: frequent user, occasional user, nonuser. We can use a single X_1 variable to denote it as follows:

$$X_1 = \begin{cases} 3, & \text{Frequent User;} \\ 2, & \text{Occasional User;} \\ 1, & \text{Nonuser.} \end{cases}$$

Class	$E\{Y\}$
Frequent User	$\beta_0 + 3\beta_1$
Occasional User	$\beta_0 + 2\beta_1$
Nonuser	$\beta_0 + \beta_1$

- Then, we have the regression model: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$
- The mean response with the regression function will be:
- Using indicator variables doesn't have this restriction since it has one more variable to denote them.

Other Codings for Indicator Variables

- For the stock company and mutual company data:

$$X_2 = \begin{cases} 1, & \text{if stock company;} \\ -1, & \text{if mutual company.} \end{cases}$$

- Another alternative: use indicator variable for each of the c classes and drop the intercept term:

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

X_1=the size of the insurance firm

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

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

ANOVA and linear regression models

- If there are only qualitative predictors, the linear regression model is equivalent to one-way or multi-way ANOVA analysis
- Eg1. Y: salary; A qualitative predictor: Region (1,2,3)

$$X_1 = \begin{cases} 1 \text{ if Region 1} \\ 0 \text{ otherwise} \end{cases} \quad X_2 = \begin{cases} 1 \text{ if Region 2} \\ 0 \text{ otherwise} \end{cases}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \varepsilon$$

  - Can be analyzed by one-way ANOVA
- Eg2. Two qualitative predictors: Region (1,2,3), education level(1,2),

$$X_1 = \begin{cases} 1 \text{ if Region 1} \\ 0 \text{ otherwise} \end{cases} \quad X_2 = \begin{cases} 1 \text{ if Region 2} \\ 0 \text{ otherwise} \end{cases} \quad X_3 = \begin{cases} 1 \text{ if education level 2} \\ 0 \text{ otherwise} \end{cases}$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$$

  - Can be analyzed by two-way ANOVA

8.5 Modeling Interactions Between Qualitative and Quantitative Predictors

- We can allow the slope wrt to a Quantitative Predictor to differ across levels of the Categorical Predictor
- Trick: Create cross-product terms between Quantitative Predictor and each of the c-1 dummy variables

Salary (Y), Expediture (X_1), and regions (X_2, X_3):

Additive Model: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

Interaction Model: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$

Region 1: $\left(X_2 = 1, X_3 = 0\right)$:

$$E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3(0) + \beta_4 X_1(1) + \beta_5 X_1(0) = \left(\beta_0 + \beta_2\right) + \left(\beta_1 + \beta_4\right) X_1$$

Region 2: $\left(X_2 = 0, X_3 = 1\right): E\{Y\} = \left(\beta_0 + \beta_3\right) + \left(\beta_1 + \beta_5\right) X_1$

Region 3: $\left(X_2 = 0, X_3 = 0\right): E\{Y\} = \beta_0 + \beta_1 X_1$

Interactions between Quantitative and Qualitative Variables

- Can conduct General Linear Test to determine whether slopes differ (or t-test when qualitative predictor has c=2 levels)
- **Insurance industry example**

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon$$

X_1: the size of the firm

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

  - To test $\mathrm{H}_0 : \beta_3 = 0$

- **Salary example**

Salary (Y), Expediture (X_1), and regions (X_2, X_3):

Interaction Model: $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_1 X_3$

  - To test $\mathrm{H}_0 : \beta_4 = \beta_5 = 0$

- These models generalize to any number of quantitative and qualitative predictors

8.7 Comparison of Two or More Regression Functions

- Soap Production Lines Example
- A company operates two productions lines for making soap bars. For each line, the relationship between the speed of the line and the amount of scrap for the day was studied.
- Y : scrap, X1: line speed. X2: code for production line.
- Interaction model:

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

$$X_{i1} = \text{line speed}$$

$$X_{i2} = \begin{cases} 1, & \text{if production line 1;} \\ 0, & \text{if production line 2.} \end{cases}$$

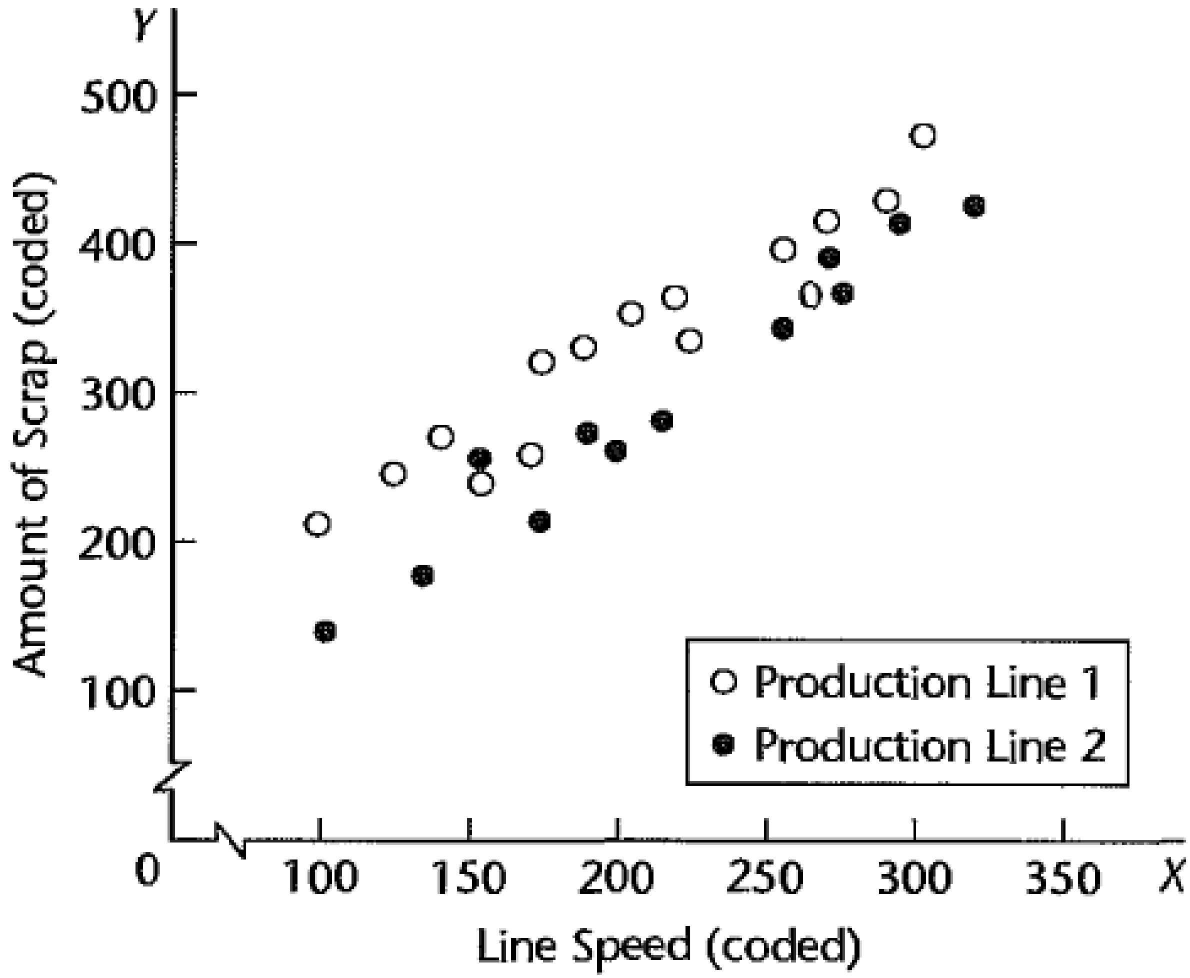
$$i = 1, 2, \cdots, 27$$

Production Line 1

Case i	Amount of Scrap Y_i	Line Speed X_{i1}	X_{i2}
1	218	100	1
2	248	125	1
3	360	220	1
4	351	205	1
5	470	300	1
6	394	255	1
7	332	225	1
8	321	175	1
9	410	270	1
10	260	170	1
11	241	155	1
12	331	190	1
13	275	140	1
14	425	290	1
15	367	265	1

Production Line 2

Case i	Amount of Scrap Y_i	Line Speed X_{i1}	X_{i2}
16	140	105	0
17	277	215	0
18	384	270	0
19	341	255	0
20	215	175	0
21	180	135	0
22	260	200	0
23	361	275	0
24	252	155	0
25	422	320	0
26	273	190	0
27	410	295	0



$$\hat{Y} = 7.57 + 1.322X_1 + 90.39X_2 - .1767X_1X_2$$

- Inference for identity of regression functions for the two production lines

H_0: $\beta_2 = \beta_3 = 0$

H_a: not both $\beta_2 = 0$ and $\beta_3 = 0$

$$F^* = \frac{SSR(X_2, X_1X_2|X_1)}{2} \div \frac{SSE(X_1, X_2, X_1X_2)}{n-4}$$

(b) Analysis of Variance

Source of Variation	SS	df
Regression	169,165	3
X_1	149,661	1
$X_2 \mid X_1$	18,694	1
$X_1X_2 \mid X_1, X_2$	810	1
Error	9,904	23
Total	179,069	26

$$SSR(X_2, X_1X_2|X_1) = SSR(X_2|X_1) + SSR(X_1X_2|X_1, X_2)$$
$$= 18{,}694 + 810 = 19{,}504$$

$$F^* = \frac{19{,}504}{2} \div \frac{9{,}904}{23} = 22.65$$ >F(0.99; 2, 23)=5.67

- Conclusion: the regression functions for the two production lines are not identical

(a) Regression Coefficients

Regression Coefficient	Estimated Regression Coefficient	Estimated Standard Deviation
β_0	7.57	20.87
β_1	1.322	.09262
β_2	90.39	28.35
β_3	−.1767	.1288

H_0: $\beta_3 = 0$
H_a: $\beta_3 \neq 0$

$$F^* = \frac{SSR(X_1X_2|X_1, X_2)}{1} \div \frac{SSE(X_1, X_2, X_1X_2)}{n-4}$$

$$= \frac{810}{1} \div \frac{9{,}904}{23} = 1.88 \quad < F(0.99; 1, 23) = 7.88$$

Or : t test: t*= –0.1767/0.1288 = – 1.37; |t*|<t(0.99; 23)=2.8

- 95% CI for β2:

$90.39 \pm 2.069(28.35)$ =(31.7,149.0)

R Code

```
#########First Example######
dat = read.table('cell.txt')
X1 = dat[,2]; X2 = dat[,3];Y = dat[,1]
x1 = (X1-mean(X1))/0.4 ;  x2 = (X2-mean(X2))/10
cor(X1, X1^2);  cor(x1, x1^2)
cor(X2, X2^2); cor(x2, x2^2)
x1sq = x1^2;  x2sq = x2^2;  x1x2 = x1*x2
fit = lm(Y ~ x1 + x2 + x1sq + x2sq + x1x2)
summary(fit)
resi = residuals(fit);  yhat = fitted(fit)
par(mfrow=c(2,2))
plot( yhat, resi);  plot(x1, resi);  plot(x2, resi);   qqnorm(resi)
```

R code

```
##Partial F-Test to test whether a first-order model would be sufficient
fit1 = lm(Y~x1+x2)
###one way of testing
anotab = anova(fit);  anotab[3:5,2]
Fstar = sum(anotab[3:5,2])/3/1048
qf(0.95, 3,5)
###an easier way to do it
anova(fit1,fit)
#####transfer back since first-order model is sufficient
fito = lm(Y~X1+X2)
summary(fito)
```

```
######### Example soap data
 dat = read.table('soap.txt')
X1 = dat[,2]; X2 = dat[,3];Y = dat[,1]
plot(X1[X2==1],Y[X2==1],xlim=c(100,350),ylim=c(100,500),
    xlab='Line Speed', ylab='Amount of Scrap')
points(X1[X2==0],Y[X2==0],pch=19)
legend("bottomright",c('Production Line 1','Production Line
2'),pch=c(1,19))
X12 = X1*X2
fit = lm(Y ~ X1+X2+X12)
```

R code

```
###Inference about two regression lines
fit0 <- lm(Y~X1)
anova(fit0,fit)

###Inference about two slopes or interaction
fit12 <- lm(Y~X1+X2)
anova(fit12, fit)
```

Homework

- P340

  8.31 8.34(a)(b) 8.35 8.38