STAT 571A — Advanced Statistical Regression Analysis

Chapter 8 NOTES

Quantitative and Qualitative Predictors for MLR

§8.1: Polynomial Regression

Mentioned in passing in §6.1, we now study polynomial regression in more detail.

This is technically a special form of MLR, since it has more than one β_k parameter.

Simplest case: 2nd-order/single predictor model:

$$Y_i = \beta_0 + \beta_1(X_i - \overline{X}) + \beta_2(X_i - \overline{X})^2 + \varepsilon_i$$

(i = 1,...,n) with $\varepsilon_i \sim \text{i.i.d.}N(0,\sigma^2)$.

Polynomial Regression (cont'd)

For simplicity, write $x_i = (X_i - \overline{X})$:

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

(Why? Centering usually reduces multi-collinearity with 2nd-order, and higher, predictors. Just do it.)

This quadratic regression can be a useful approximation to data that deviate from strict linearity. See Fig. 8.1 ⟶

Quadratic Regression

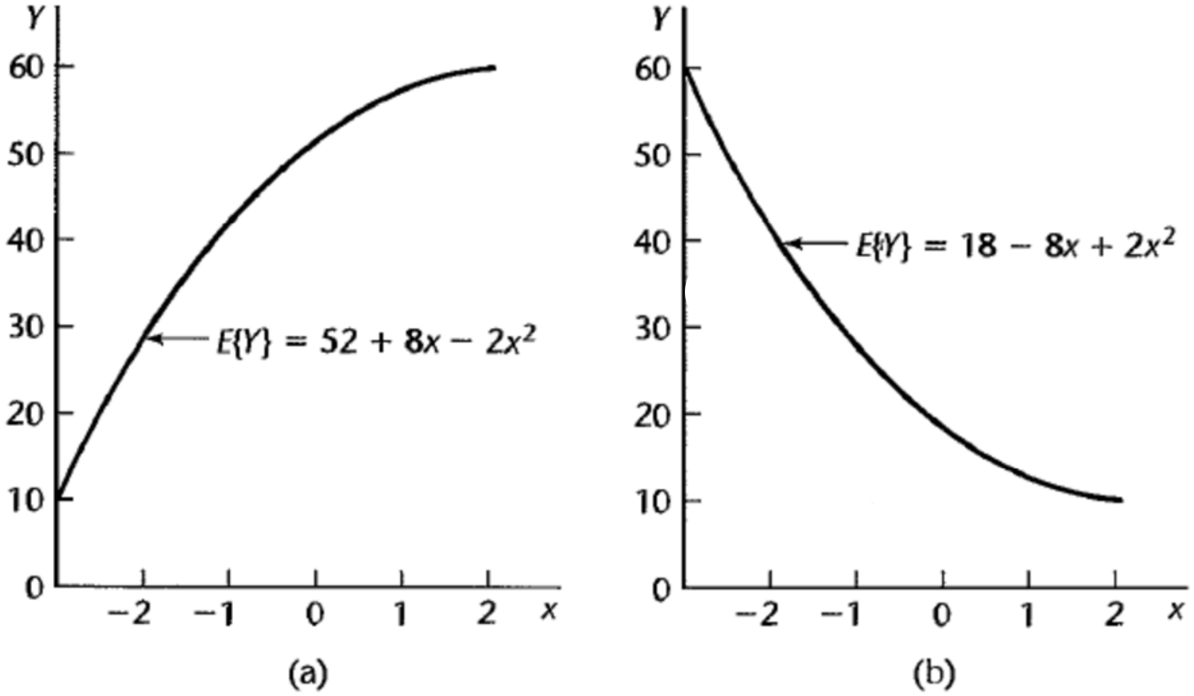
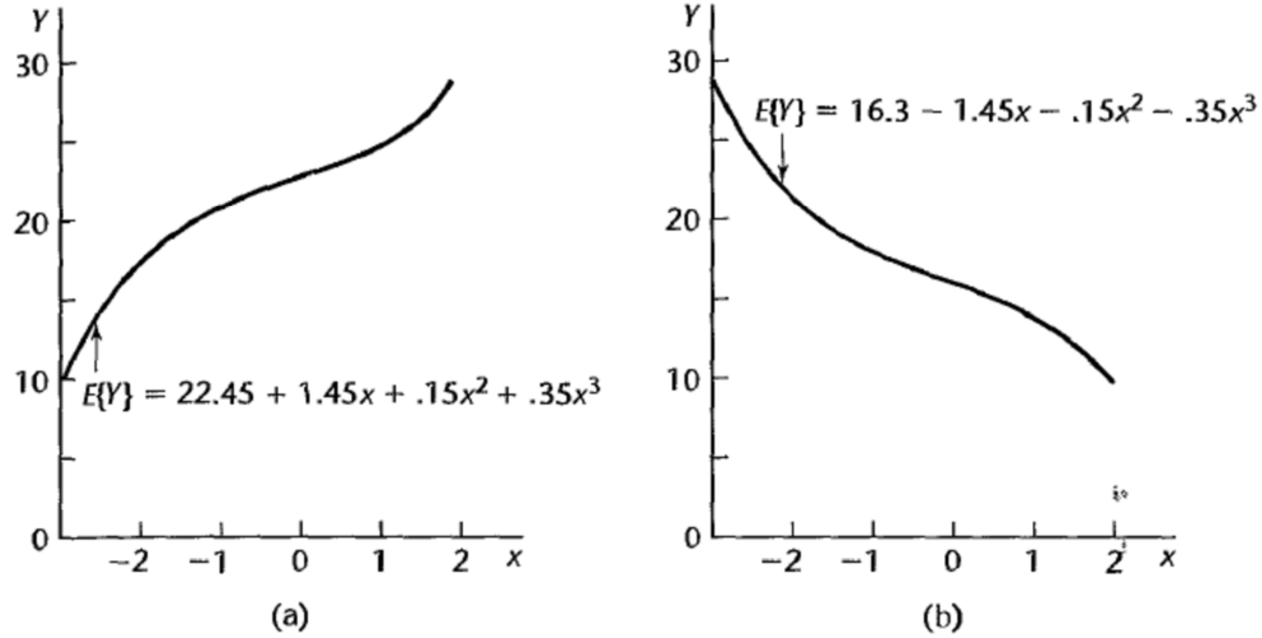


FIGURE 8.1 Examples of Second-Order Polynomial Response Functions.

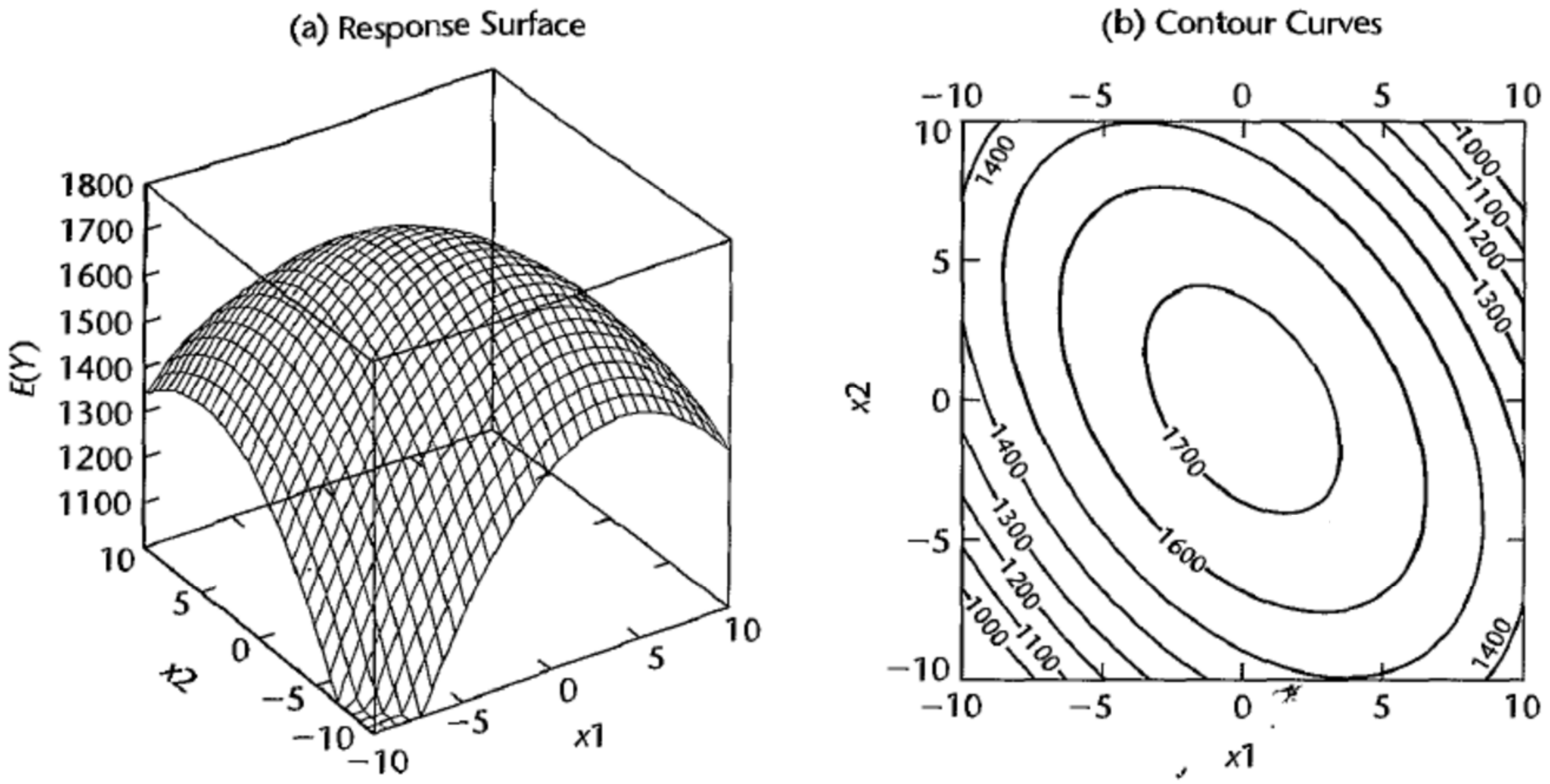
Cubic Regression

(Fig. 8.2. Examples of 3rd-order, cubic regression polynomials.)



2nd-Order Response Surface

(Fig. 8.3. Examples of 2nd-order response surface, as in §6.1.)



Testing Polynomial Models

- **Sequential testing: for testing purposes, we start with the highest-order term and work down the order ('up the ladder').**
- **Suppose $E\{Y_i\} = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3$**
  - **First test $H_o{:}\beta_{111}=0$ via partial F-test and $SSR(x_3|x_1,x_2)$. If signif., STOP and conclude cubic polynomial is significant.**
  - **If $H_o{:}\beta_{111}=0$ is NOT signif., drop β_{111} and go 'up ladder' to test $H_{o:}\beta_{11}=0$ via $SSR(x_2|x_1)$.**

Polynomial Regression (cont'd)

- $E\{Y_i\} = \beta_0 + \beta_1 x_i + \beta_{11} x_i^2 + \beta_{111} x_i^3$ (cont'd)
  - If H_o:β_{11}=0 is signif., STOP and conclude quadratic polynomial is significant.
  - If H_o:β_{11}=0 is NOT signif., drop β_{11} and go 'up ladder' to test $H_{o:}\beta_1$=0 via SSR(x_1).
  - If H_o:β_1=0 is signif., STOP and conclude simple linear model is significant. Etc.
- Once sequential testing is complete, we usually go back and fit the final model in terms of the orig. X_k's to get cleaner b_k's and std. errors.

Example: Power Cell Data (CH08TA01)

Power Cell Data example: Y = {# cycles} and we have 2 predictors (X_1 = charge rate & X_2 = temp.); see Table 8.1. Consider a 2nd-order "response surface" MLR:

```
> Y = c(150, 86, 49, ..., 279, 235, 224)
> X1 = c(0.6, 1.0, 1.4, ..., 0.6, 1.0, 1.4)
> X2 = c( rep(10,3), rep(20,5), rep(30,3) )
> x1 = (X1 - mean(X1))/0.4
> x2 = (X2 - mean(X2))/min(X2)
>
> x1sq = x1*x1
> x2sq = x2*x2
> x1x2 = x1*x2
```

Power Cell Data (CH08TA01) (cont'd)

Selection of X_1 and X_2 was controlled.

⇒ note the zero/near-zero correlations among the (transformed) *x*-variables:

```
> cor( cbind(x1, x2, x1sq, x2sq, x1x2) )
           x1        x2      x1sq      x2sq      x1x2
x1   1.00e+00  0.00e+00 -4.04e-16 -1.99e-17  0.00e+00
x2   0.00e+00  1.00e+00  0.00e+00  0.00e+00 -9.06e-17
x1sq -4.04e-16  0.00e+00  1.00e+00  2.67e-01  0.00e+00
x2sq -1.99e-17  0.00e+00  2.67e-01  1.00e+00  0.00e+00
x1x2  0.00e+00 -9.06e-17  0.00e+00  0.00e+00  1.00e+00
```

Power Cell Data (CH08TA01) (cont'd)

Compare this to (non-trivial) correlations among orig. X_ks, etc :

```
> cor( cbind(X1, X2, X1sq, X2sq, X1X2) )
              X1      X2    X1sq     X2sq   X1X2
 X1     1.00e+00  0.000  0.9910 -4.2e-18  0.605
 X2     0.00e+00  1.000  0.0000  0.09861  0.757
 X1sq   9.91e-01  0.000  1.0000  0.00592  0.600
 X2sq  -4.16e-18  0.986  0.0059  1.0e+00  0.746
 X1X2   6.05e-01  0.757  0.5999  7.5e-01  1.000
```

Power Cell Data (CH08TA01) (cont’d)

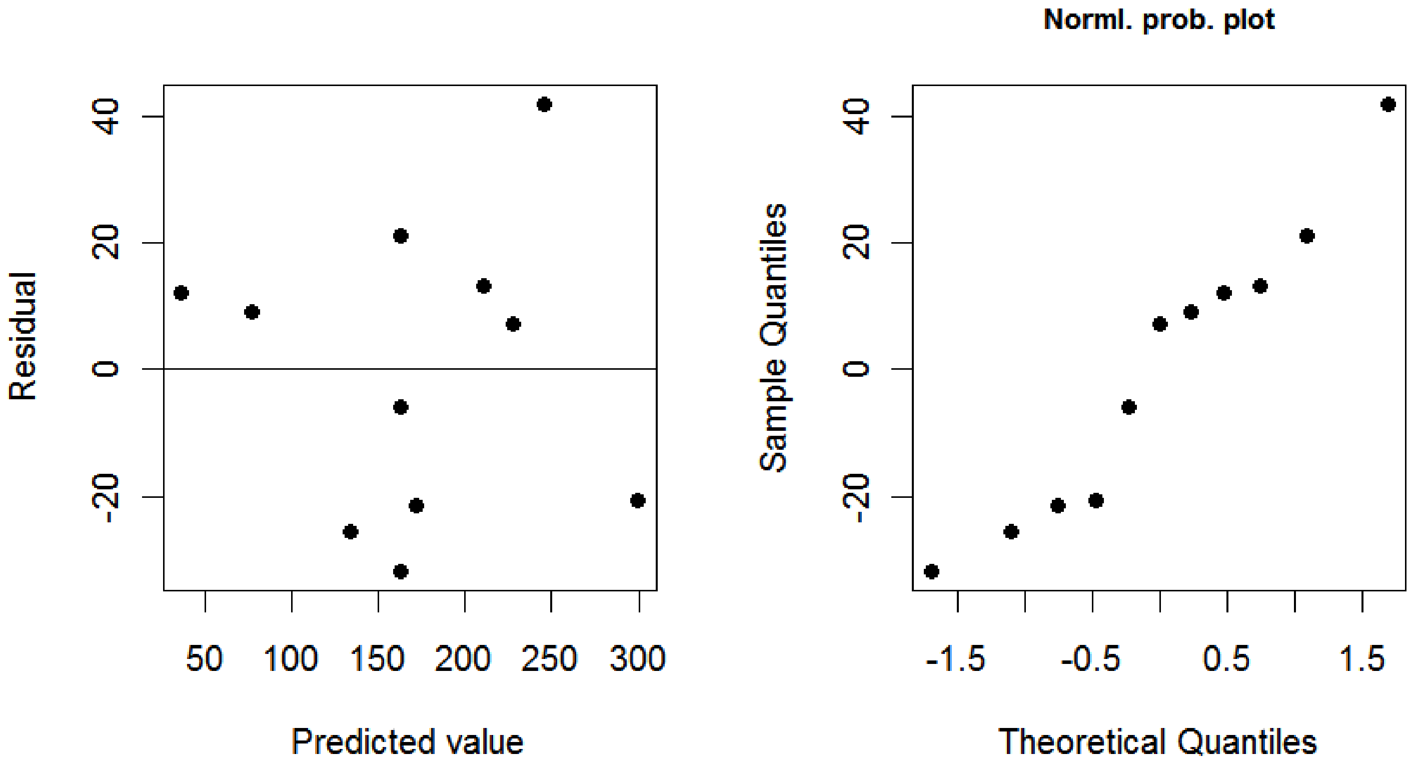
Full 2nd-order model fit with transfrm’d *x*’s:

```
> summary( lm(Y ~ x1 + x2 + x1sq + x2sq + x1x2) )
Call:
lm(formula = Y ~ x1 + x2 + x1sq + x2sq + x1x2)
Coefficients:
            Estimate  Std.Error t value  Pr(>|t|)
(Intercept)   162.84       16.6     9.81  0.00019
x1            -55.83       13.2    -4.22  0.00829
x2             75.50       13.2     5.71  0.00230
x1sq           27.39       20.3     1.35  0.23586
x2sq          -10.61       20.3    -0.52  0.62435
x1x2           11.50       16.2     0.71  0.50918
```

Power Cell Data (CH08TA01) (cont'd)

Residual analysis shows no serious issues:

```
> plot( resid(CH08TA01.lm) ~ fitted(CH08TA01.lm) )
> abline( h=0 )
> qqnorm( resid(CH08TA01.lm) )
```



Power Cell Data (CH08TA01) (cont'd)

Lack of Fit test. Only joint replication is at x_1=x_2=0, so need to set up the `factor` term <u>carefully</u> in R:

```
> LOFfactor = factor( c(seq(-4,-1), rep(0,3),
                                     seq(1,4)) )
> anova( CH08TA01.lm, lm(Y ~ LOFfactor) )
```

```
Analysis of Variance Table
Model 1: Y ~ x1 + x2 + x1sq + x2sq + x1x2
Model 2: Y ~ LOFfactor
  Res.Df     RSS  Df  Sum of Sq     F  Pr(>F)
1      5  5240.4
2      2  1404.7   3     3835.8  1.82  0.374
```

LOF stat. is F^* = 1.82 (P = 0.374). No signif. lack of fit.

Power Cell Data (CH08TA01) (cont'd)

Partial F-test of 2nd-order terms (H_o:$\beta_{11} = \beta_{22} = \beta_{12} = 0$):

```
> anova( lm(Y ~ x1+x2), CH08TA01.lm )

Analysis of Variance Table
Model 1: Y ~ x1 + x2
Model 2: Y ~ x1 + x2 + x1sq + x2sq + x1x2
  Res.Df     RSS Df Sum of Sq     F  Pr(>F)
1      8 7700.33
2      5 5240.44  3   2459.89 0.782  0.553
```

Partial 3 df F-statistic is *F = 0.78 (*P* = 0.553).**

No signif. deviation from 0 seen in 2nd-order terms.

Power Cell Data (CH08TA01) (cont’d)

Fit reduced 1st-order model:

```
> summary( lm(Y ~ x1+x2) )
Call:
lm(formula = Y ~ x1 + x2)
Coefficients:
          Estimate Std.Error t value  Pr(>|t|)
(Intercept) 172.00     9.3543 18.3872 7.88e-08
x1          -55.83    12.6658 -4.4082 0.002262
x2           75.50    12.6658  5.9609 0.000338

Multiple R-squared: 0.87294
Adjusted R-squared: 0.84118
F-stat.: 27.482 on 2 and 8 DF, p-val.: 0.00026
```

Power Cell Data (CH08TA01) (cont’d)

Bonferroni-adjusted simultaneous conf. intervals on 1st-order β-parameters (using original X-variables):

```
> g = length( coef(lm(Y ~ X1+X2)) ) - 1
> confint( lm(Y ~ X1+X2),
                      level = 1-(.10/g) )
```

```
X1          -212.6020565  -66.564610
X2             4.6292511   10.470749
```

(cf. Textbook p. 305)

Interaction Terms

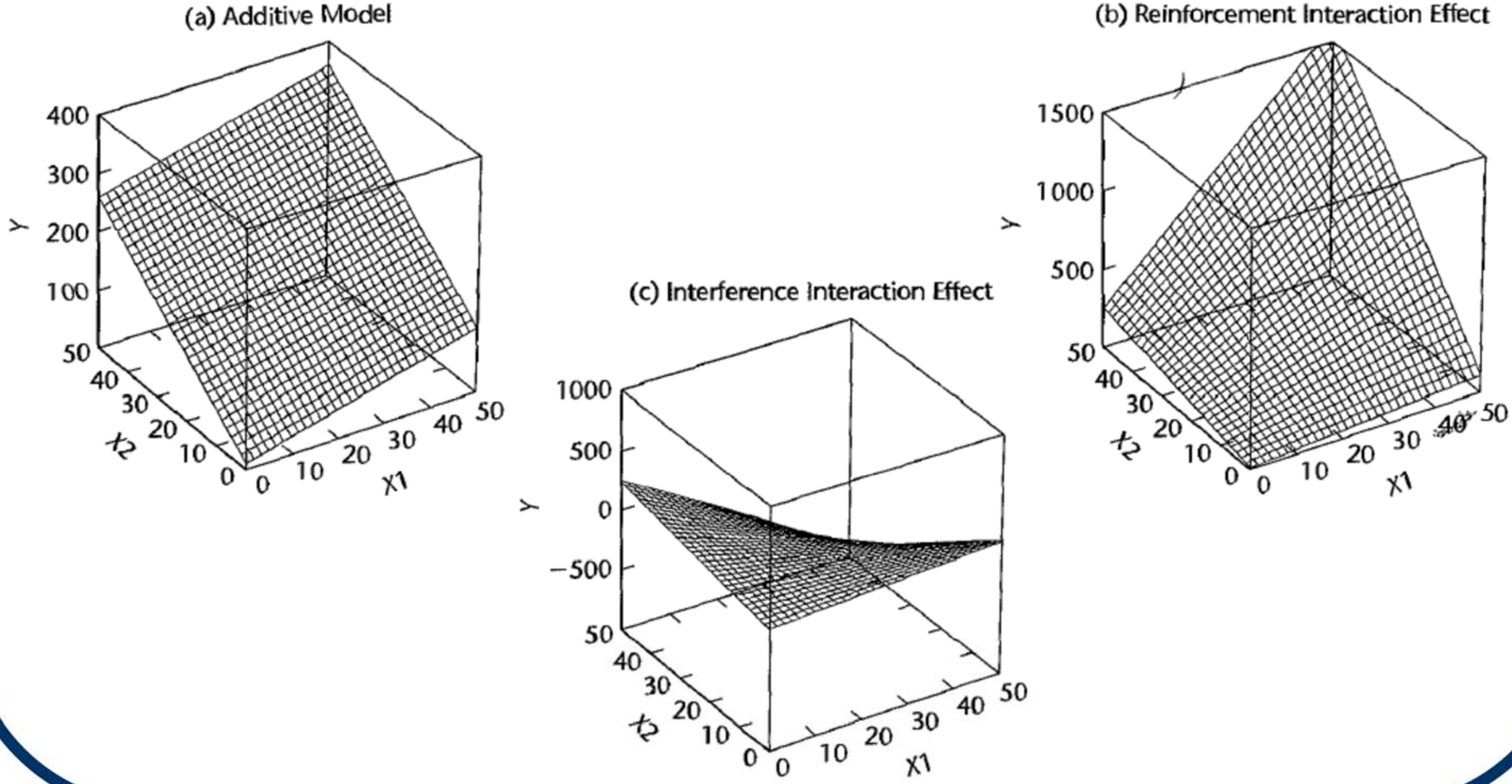
- Interaction cross-product terms can be included in any MLR to allow for **interactions** between the X_k-variables.
- E.g., $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2$
- The cross-product creates a <u>departure from additivity</u> in the mean response. If $\beta_3 = 0$, the mean response is strictly additive in X_1 and X_2.

Interaction Terms (cont'd)

- **Notice that the usual interpretation for the β_k parameters is muddied here.**
  - **What does it mean to increase X_1 by +1 unit while holding X_1X_2 fixed?!?**
- **Alt. interpretation: cross-product terms allow for '*synergistic*' or '*antagonistic*' interactions between the X_k-variables.**
- **It's a special kind of departure from additivity: synergy occurs for $\beta_3 > 0$, antagonism for $\beta_3 < 0$.**

Figure 8.8

Graphics for (a) additive, (b) synergistic, or (c) antagonistic response surfaces.



(a) Additive Model

(b) Reinforcement Interaction Effect

(c) Interference Interaction Effect

Interaction Caveats

Need to be careful with interactions.

- **If they exist and they are ignored, very poor inferences on E{Y} will result.**
- **On the other hand, adding a 'kitchen sink' of all possible interactions can overwhelm the MLR.**
  - → **With 3 X_k-variables there are 3 possible pairwise interactions (not incl. the tri-way!)**
  - → **With 8 X_k-variables there are 28 possible pairwise interactions (not incl. multi-ways!)**
  - → **Things get unwieldy fast...**

Body Fat Data (CH07TA01) (cont'd)

- To the 3 original X_k-variables now include all pairwise interactions.
- Center each X-variable (about its mean) first to assuage problems with multicollinearity: $x_{ik} = X_{ik} - \bar{X}_k$ (k = 1,2,3)
- MLR now has six predictor terms and p=7 β-parameters: $E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \beta_6 x_2 x_3$

Body Fat Data (CH07TA01) (cont'd)

- R can fit interaction terms using a special `*` operator:
  e.g., `x1*x2` fits `x1` <u>and</u> `x2` <u>and</u> `x1:x2` all with just 1 term.
- For the Body Fat data, construct centered *x*-variables as `x1 = X1 – mean(X1)`, etc. Then call

```
> anova( lm(Y ~ x1 + x2 + x3),
    lm(Y ~ x1*x2 + x1*x3 + x2*x3) )
```

Output follows ⟶

Body Fat Data (CH07TA01) (cont'd)

Output from partial F-test of all pairwise interactions:

```
Analysis of Variance Table
Model 1: Y ~ x1 + x2 + x3
Model 2: Y ~ x1 * x2 + x1 * x3 + x2 * x3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     16 98.405
2     13 87.690  3    10.715 0.5295 0.6699
```

Partial 3 df F-statistic is $F^* = 0.53$ ($P = 0.670$).

No signif. pairwise interactions are seen.

Body Fat Data (CH07TA01) (cont'd)

Can also include tri-way interactions:

```
> anova( lm(Y ~ x1+x2+x3), lm(Y ~ x1*x2*x3) )
Analysis of Variance Table
Model 1: Y ~ x1 + x2 + x3
Model 2: Y ~ x1 * x2 * x3
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     16 98.405
2     12 85.571  4    12.834 0.4499 0.7707
```

4 d.f. partial F-statistic is *F = 0.45 (*P* = 0.771).**

⇒ no signif. pairwise or tri-way interactions.

Qualitative Predictors

- We've seen cases where the X-variable was either 0 or 1 (called a binary indicator). If this indicated a qualitative state (say, 1 = ♀ or 0 = ♂) then the numbering is arbitrary. The predictor is actually qualitative, not quantitative.
- (Still 0 vs. 1 is usually as good a pseudo-quantification as any.)
- Question: What happens when binary indicators are combined with true quantitative predictors?

Example: Insur. Innov'n Data

- Suppose we study
  Y = Insurance method adoption time (mos.)
  in insurance companies, with
  X_1 = size of firm (quantitative)
  X_2 = type of firm: 1 = stock, 0 otherwise
  X_3 = type of firm: 1 = mutual, 0 otherwise

- Design matrix is (n = 4):

$$\mathrm{X} = \begin{array}{cccc} \text{“}X_0\text{”} & X_1 & X_2 & X_3 \\ \hline \end{array} \left[\begin{array}{cccc} 1 & X_{11} & 1 & 0 \\ 1 & X_{21} & 1 & 0 \\ 1 & X_{31} & 0 & 1 \\ 1 & X_{41} & 0 & 1 \end{array}\right]$$

Design Matrix Problem

- But wait, there's a **problem with this design matrix** $\mathbf{X}$. Notice that $X_0 = X_2 + X_3$ so the predictors are not linearly independent: rank($\mathbf{X}'\mathbf{X}$) = 3 < 4 = p. (See p. 314.)
  - The **MLR will fail**!

- Solution is (usually) to eliminate X_3 and model $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$.

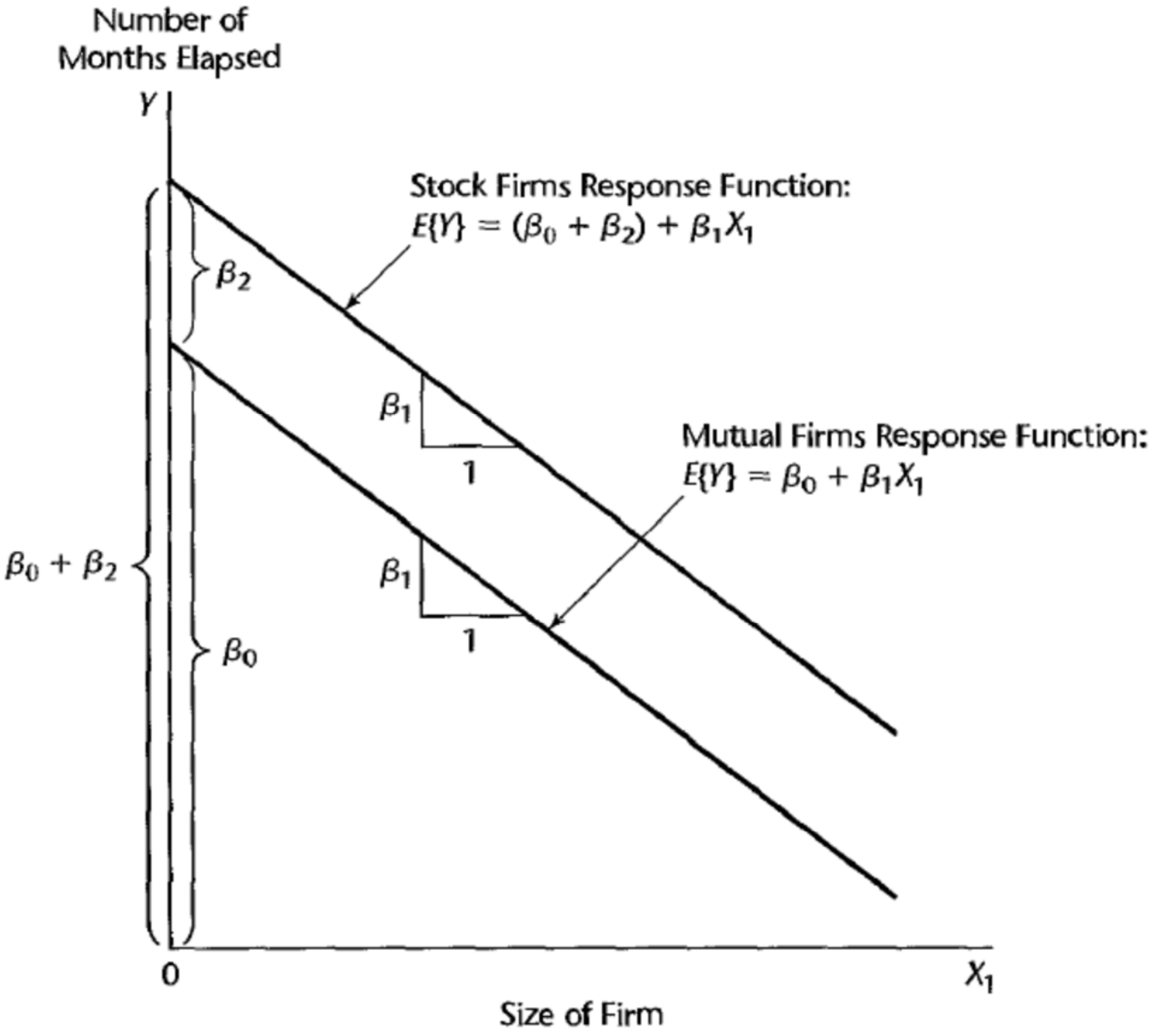
- Model interpretation here is actually sorta' intriguing →

Two Straight Lines

- For $E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 X_2$, with X_1 quantitative and X_2 a 0-1 indicator, consider:
  - When $X_2 = 0$ (mutual firm), $\boxed{E\{Y\} = \beta_0 + \beta_1 X_1}$, an SLR on X_1 with slope β_1 and Y-intercept β_0.
  - When $X_2 = 1$ (stock firm), $\boxed{E\{Y\} = (\beta_0 + \beta_2) + \beta_1 X_1}$, an SLR on X_1 with same slope β_1 but new Y-intercept $(\beta_0 + \beta_2)$.
- So we have two parallel straight lines—each with same σ^2—one for stock firms ($X_2 = 1$) and one for mutual firms ($X_2 = 0$).

ANCOVA Graphic

FIGURE 8.11 **Illustration of Meaning of Regression Coefficients for Regression Model (8.33) with Indicator Variable X_2—Insurance Innovation Example.**



Tests in Equal-Slopes ANCOVA

- For this equal-slopes ANCOVA model, some obvious hypotheses are
  - (first) H_o: $\beta_2 = 0$
    (i.e., no diff. between type ⇒ lines are same)
  - (next) H_o: $\beta_1 = 0$
    (i.e., no effect of size ⇒ lines are flat)
- Data are in Table 8.2; n = 20.

R code/analysis follows →

Insur. Innov’n Data (CH08TA02)

Y = Insurance method adoption time
X_1 = size of firm
X_2 = type of firm (mutual vs. stock)

```
> Y = c(17, 26, ..., 30, 14)
> X1 = c(151, 92 , ..., 124, 246)
> X2 = c( rep(0,10), rep(1,10) )
```

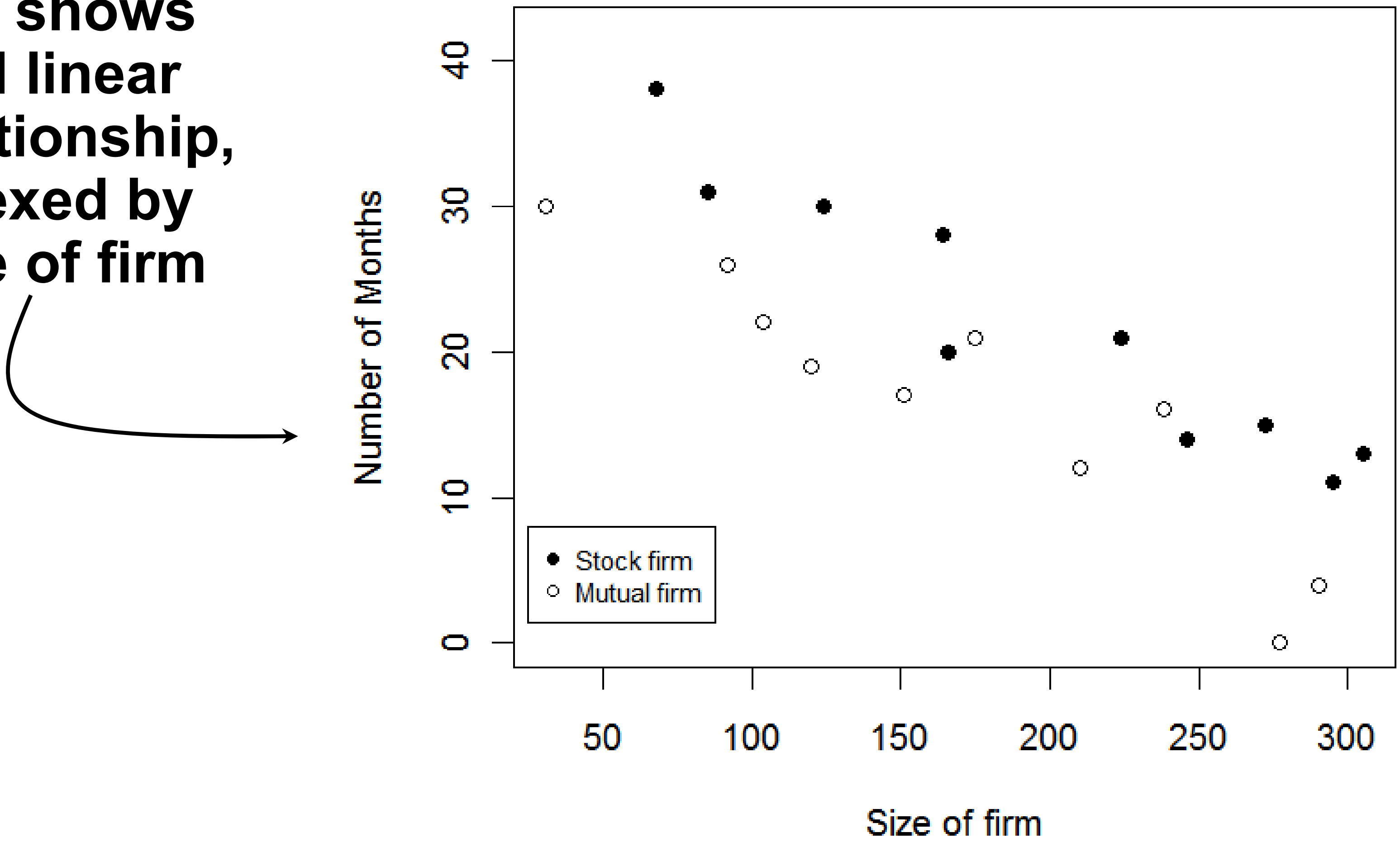
Scatterplot using

```
> plot( Y ~ X1, pch=1+(18*X2) )
```

(next slide →) shows two separate scatterlines, one for each type of firm.

Insur. Innov’n Data (CH08TA02) Scatterplot

Plot shows dual linear relationship, indexed by type of firm



Insur. Innov'n (CH08TA02) (cont'd)

Equal-slopes ANCOVA in R:

```
> CH08TA02.lm = lm( Y ~ X1 + X2 )
> summary( CH08TA02.lm )
```

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

Coefficients:
             Estimate Std. Error t value  Pr(>|t|)
(Intercept) 33.874069   1.813858  18.675  9.15e-13
X1          -0.101742   0.008891 -11.443  2.07e-09
X2           8.055469   1.459106   5.521  3.74e-05

Multiple R-squared: 0.8951, Adjusted R-squ.: 0.8827
F-statistic: 72.5 on 2 and 17 DF, p-value: 4.77e-09
```

Insur. Innov’n (CH08TA02) (cont’d)

ANOVA table (with sequential SSRs):

```
> CH08TA02.lm = lm( Y ~ X1 + X2 )
> anova( CH08TA02.lm )

 Analysis of Variance Table
 Response: Y
        Df  Sum Sq  Mean Sq F value    Pr(>F)
 X1      1 1188.17  1188.17  114.51  5.68e-09
 X2      1  316.25   316.25   30.48  3.74e-05
 Resid. 17  176.39    10.38
```

Partial F^* = 30.48 for X_2 ($P = 3.7\times10^{-5}$), so the two ‘types’ are signif. different (cf. Table 8.3).

Insur. Innov’n (CH08TA02) (cont’d)

__Pointwise__ conf. intervals from ANCOVA:

```
> CH08TA02.lm = lm( Y ~ X1 + X2 )
> confint( CH08TA02.lm )
                 2.5 %      97.5 %
(Intercept) 30.0471625 37.70097553
X1          -0.1205009 -0.08298329
X2           4.9770253 11.13391314
```

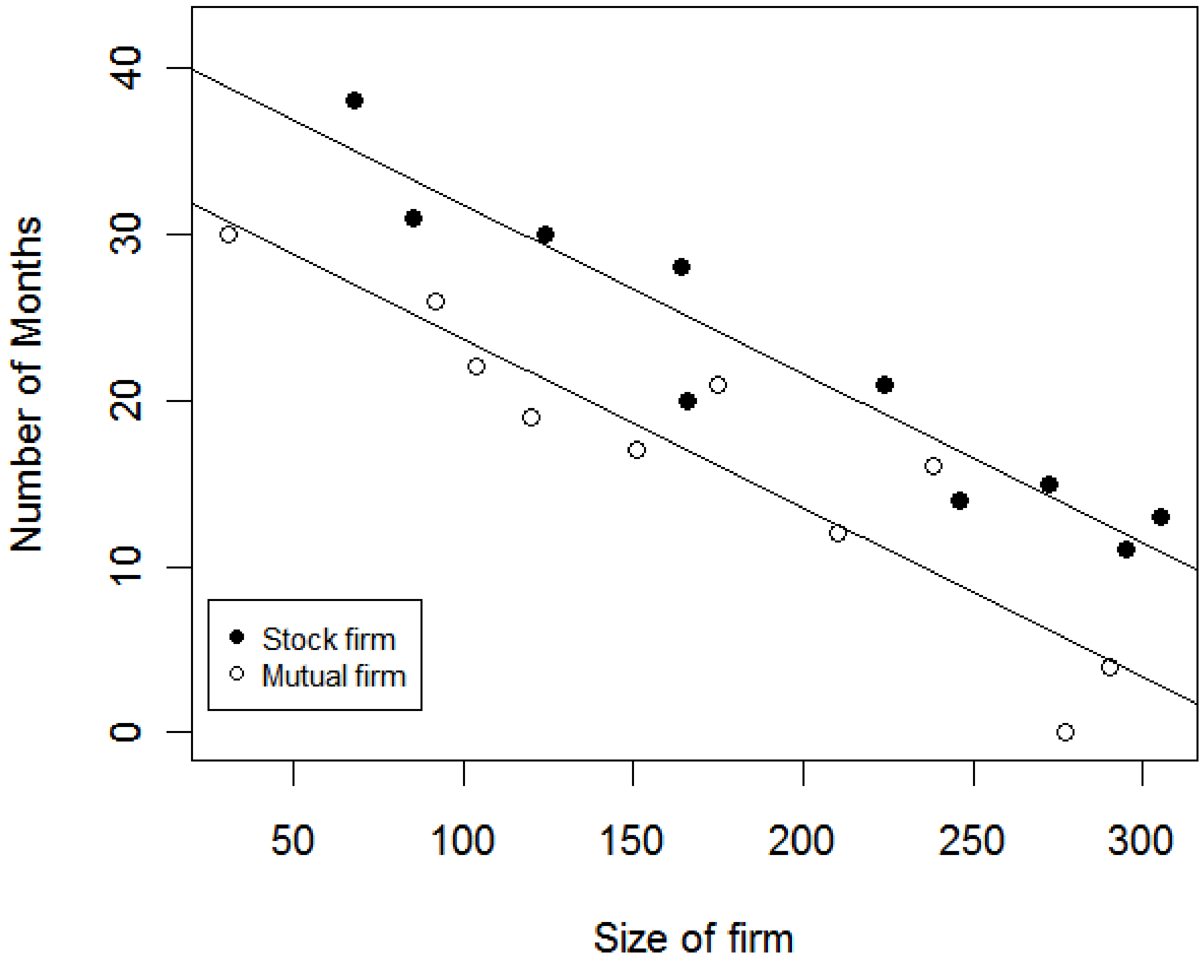
So, e.g., if interest is in effect of type of firm (X_2), we see stock firms take between

$$4.98 \leq \beta_2 \leq 11.13$$

months longer to adopt the innovation.

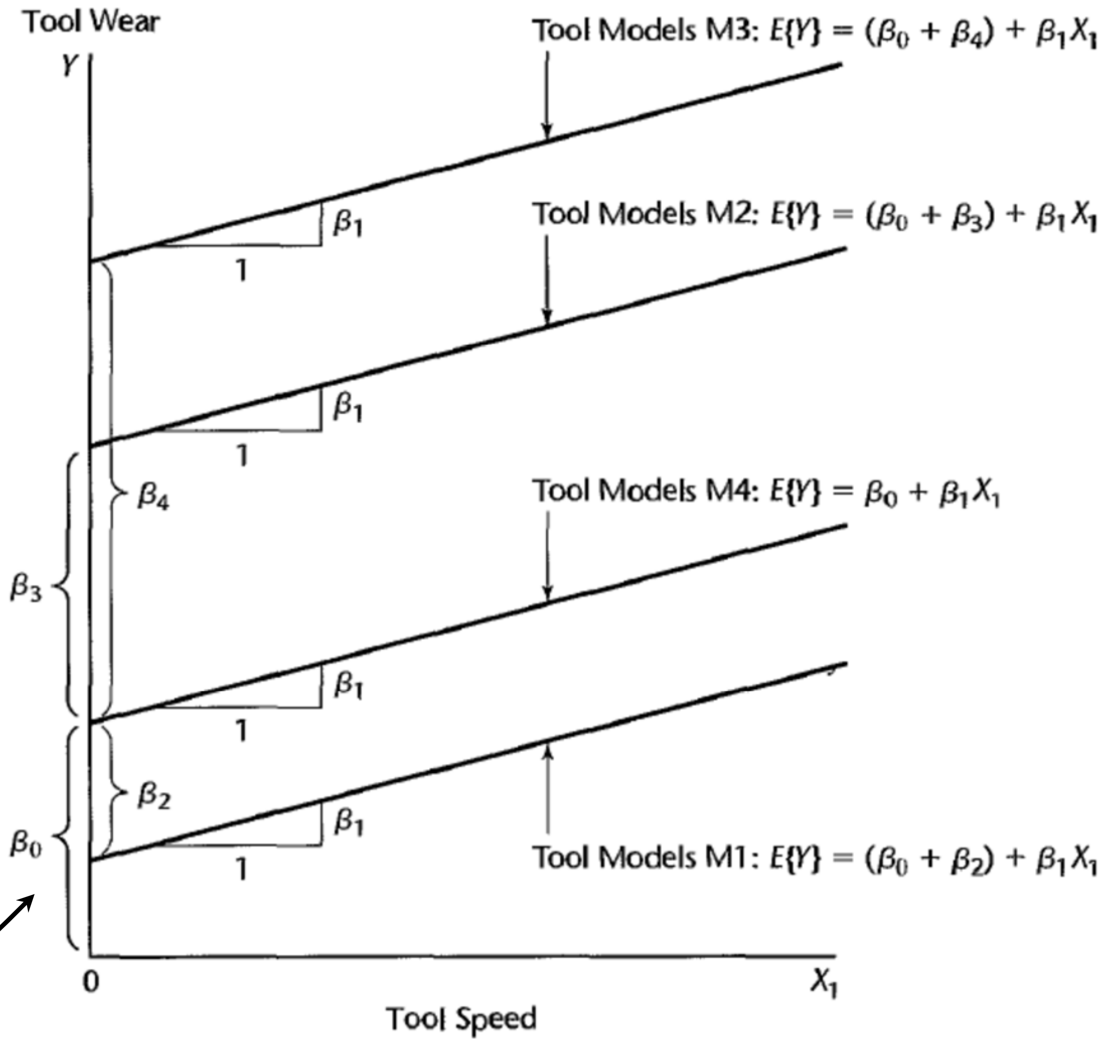
Insur. Innov'n Data (CH08TA02) (cont'd)

Scatterplot with separate, equal-slope lines overlaid (cf. Fig. 8.12)



Multiple-Level ANCOVA

If more than 2 levels are represented by the qualitative factor, just include more (parallel) lines: one line for each level of the factor. See, e.g., Fig. 8.13



Unequal-Slopes ANCOVA

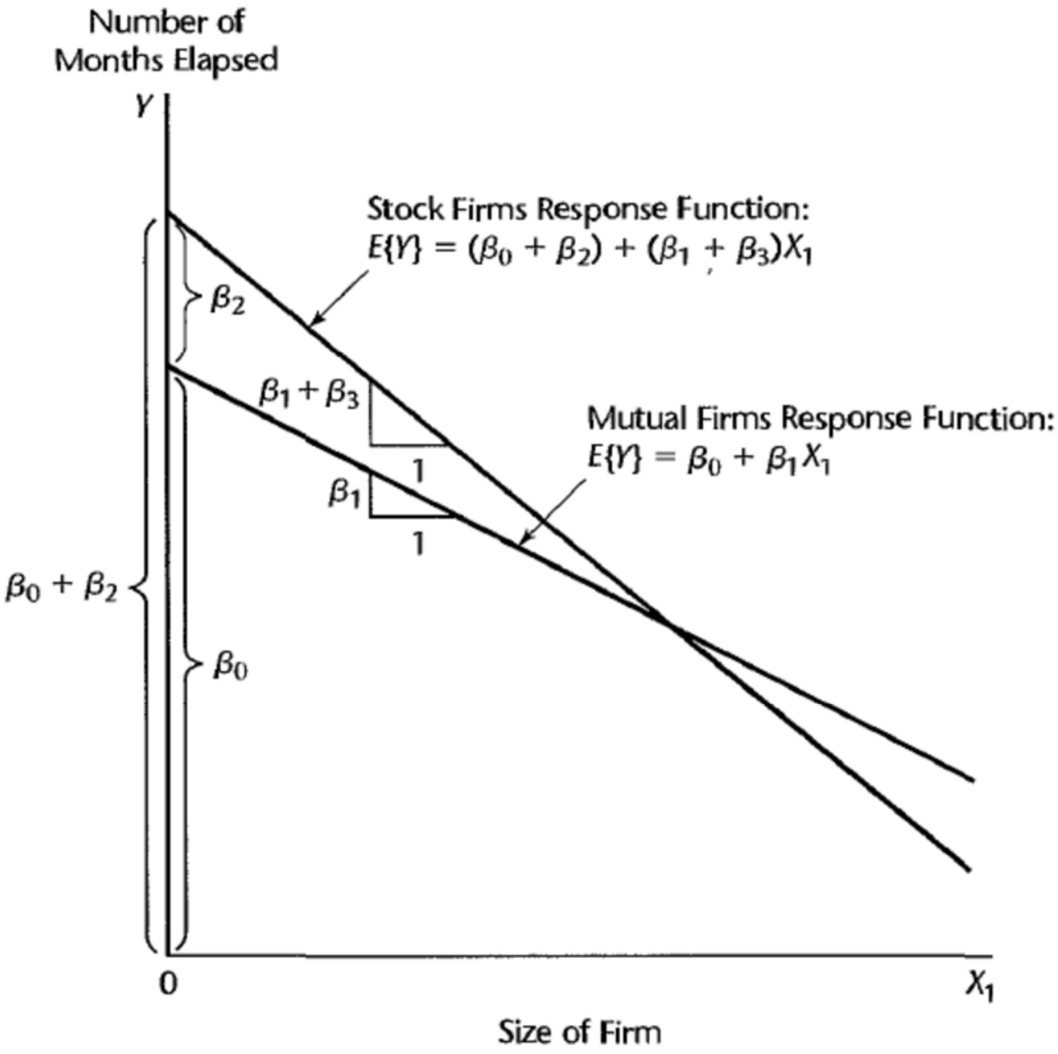
- How to incorporate differential slopes in a (two-factor/two-level) ANCOVA?
- Easy! Just add an X_1X_2 interaction term:

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

  - When $X_2 = 0$, $\boxed{E\{Y\} = \beta_0 + \beta_1 X_1}$, an SLR on X_1 with slope β_1 and Y-intercept β_0.
  - When $X_2 = 1$, $\boxed{E\{Y\} = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)X_1}$, an SLR on X_1 with new slope $(\beta_1 + \beta_3)$ <u>and</u> new Y-intercept $(\beta_0 + \beta_2)$.

Unequal-Slopes ANCOVA

For instance, with the Insur. Innov'n Data (CH08TA02), Fig. 8.14 conceptualizes the unequal-slopes model →



Insur. Innov’n (CH08TA02) (cont’d)

Unequal-slopes ANCOVA in R, via interaction term and * operator:

```
> summary( lm(Y ~ X1*X2) )
Call:
lm(formula = Y ~ X1 * X2)
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
Multiple R-squared: 0.8951, Adjusted R-squ.: 0.8754
F-statistic: 45.49 on 3 and 16 DF, p-val.: 4.68e-08
```

Insur. Innov’n (CH08TA02) (cont’d)

ANOVA table (with sequential SSRs) for unequal-slopes ANCOVA model:

```
> anova( lm(Y ~ X1*X2) )
```

```
Analysis of Variance Table
Response: Y
       Df  Sum Sq Mean Sq  F value   Pr(>F)
X1      1 1188.17 1188.17 107.7819 1.63e-08
X2      1  316.25  316.25  28.6875 6.43e-05
X1:X2   1    0.01    0.01   0.0005   0.9821
Resid. 16  176.38   11.02
```

X1*X2 interaction P-value > .05, so no signif. departure from equal slopes is indicated (cf. Table 8.4)

§8.6: Multi-Factor ANCOVA

- **The ANCOVA model can be extended to more than one quantitative X-variable.**
- **The concepts are essentially unchanged, just in a higher-dimensional space: the hyperplanes are all parallel and the qualitative predictor changes locations of the hyper-intercepts.**
- **Sounds trickier, but not really that different <u>and</u> not much harder to program.**

Only Qualitative Predictors

- What if all the predictor variables are qualitative (0-1) indicators?
- In effect, the model structure is more circumspect, since we are now just comparing the mean responses across the levels of each qualitative factor.
- This is known as ANOVA modeling, and is studied in STAT 571B.

§8.7: Comparing Multiple Regression Curves

- Let’s do a fully coordinated example of how to compare two regression functions.
- Example: Production Line Data (CH08TA05) with

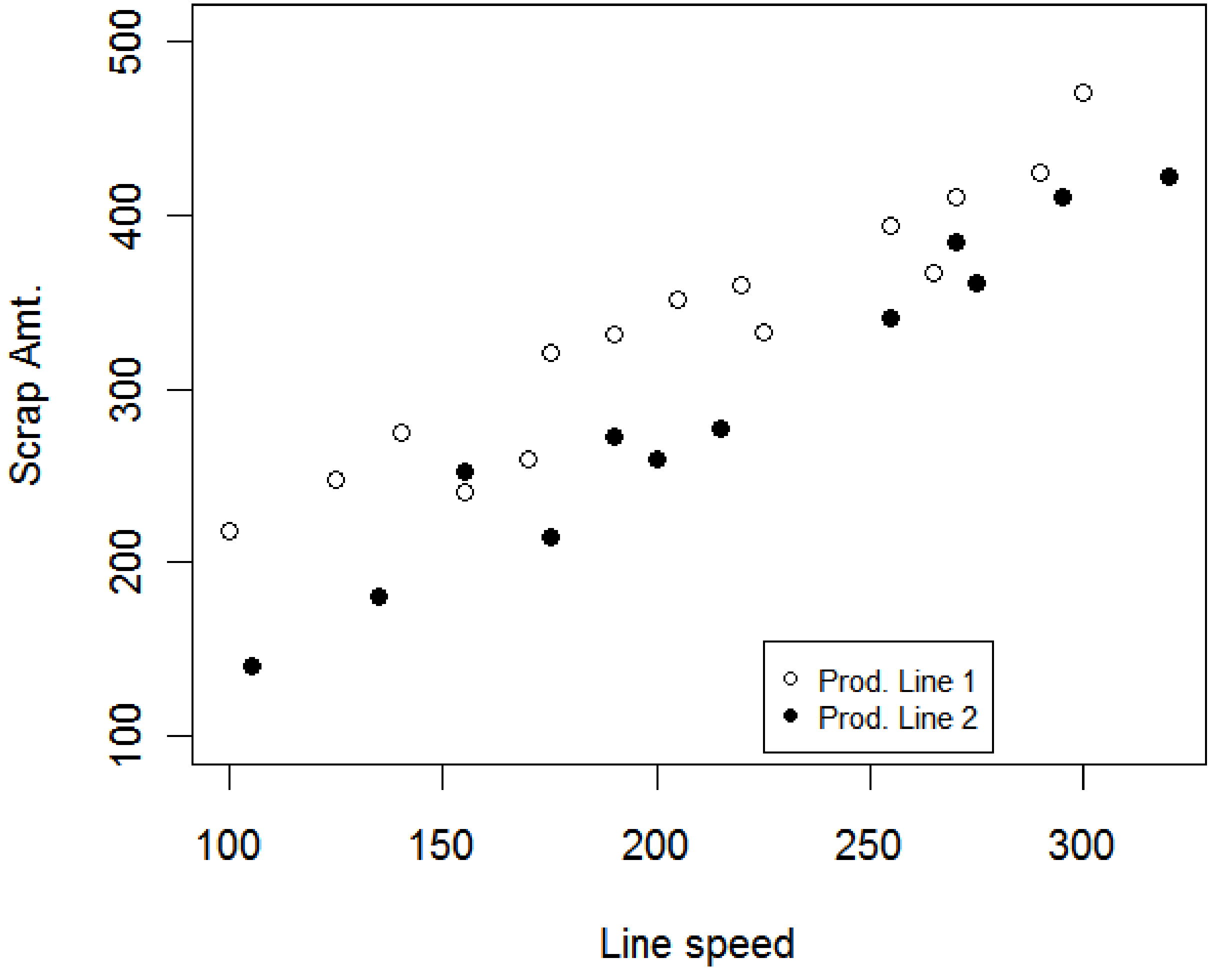
  Y = Soap Production ‘Scrap’

  X_1 = Product’n Line Speed

  X_2 = Line Indicator (Line 1 vs. Line 2)
- Start with a scatterplot →

Sec. 8.7: Product’n Line Data (CH08TA05) Scatterplot

Plot shows dual linear relationship, indexed by prod’n line



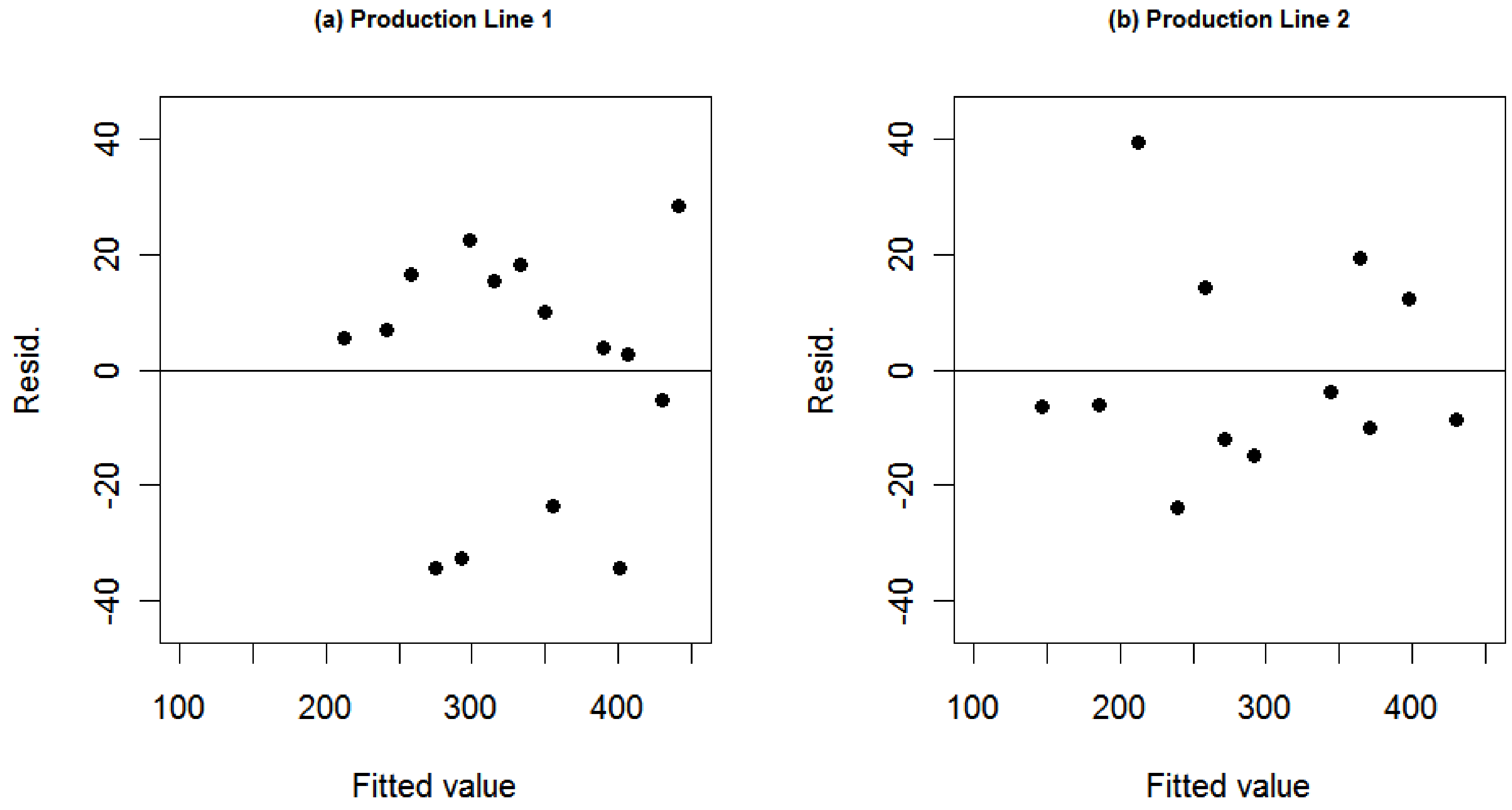
Product’n Line Data (CH08TA05) (cont’d)

Unequal-slopes ANCOVA in R, via interaction term and * operator:

```
> summary( lm(Y ~ X1*X2) )

Call:
lm(formula = Y ~ X1 * X2)

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)
(Intercept)  7.57446   20.86970   0.363   0.71996
X1           1.32205    0.09262  14.273  6.45e-13
X2          90.39086   28.34573   3.189   0.00409
X1:X2       -0.17666    0.12884  -1.371   0.18355

Multiple R-squ.: 0.9447,  Adjusted R-squ.: 0.9375
F-statistic: 130.9 on 3 and 23 DF, p-val.: 1.34e-14
```

Product'n Line Data (CH08TA05) (cont'd)

Per-line residual plots (cf. Fig. 8.17):



Product’n Line Data (CH08TA05) (cont’d)

Brown-Forsythe test for equal σ^2 between the two product’n lines shows insignif. *P* = 0.53:

```
> library( lawstat )
> BF.htest = levene.test( resid( CH08TA05.lm ),
              group=X2, location=“median” )
modified robust Brown-Forsythe Levene-type test
based on the absolute deviations from the median
data:  ei
Test Statistic = 0.4047, p-value = 0.5304
> sqrt( BF.htest$statistic )
Test Statistic
     0.6361795       #BF t*-stat. (cf. p.333)
```

Product’n Line Data (CH08TA05) (cont’d)

ANOVA table (with sequential SSRs) for unequal-slopes model:

```
> anova( lm(Y ~ X1*X2) )
```

```
Analysis of Variance Table
Response: Y
       Df  Sum Sq Mean Sq F value    Pr(>F)
X1         1 149661  149661 347.5548 2.224e-15
X2         1  18694   18694  43.4129 1.009e-06
X1:X2      1    810     810   1.8802    0.1835
Residuals 23   9904     431
```

(cf. Table 8.6)

Product’n Line Data (CH08TA05) (cont’d)

ANOVA partial F-test for identity of lines ($H_o:\beta_2=\beta_3=0$):

```
> anova( lm(Y ~ X1), lm(Y ~ X1*X2) )

Analysis of Variance Table
Model 1: Y ~ X1
Model 2: Y ~ X1 * X2
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     25 29407.8
2     23  9904.1  2     19504 22.646 3.669e-06
```

2 d.f. partial F^* = 22.65 ($P = 3.7\times10^{-6}$), so two lines are significantly different (somehow).

Product'n Line Data (CH08TA05) (cont'd)

ANOVA partial F-test for identity of slopes (H_o:$\beta_3 = 0$):

```
> anova( lm(Y ~ X1+X2), lm(Y ~ X1*X2) )
```

```
Analysis of Variance Table
Model 1: Y ~ X1 + X2
Model 2: Y ~ X1 * X2
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     24 10713.7
2     23  9904.1  1    809.62 1.8802 0.1835
```

1 d.f. partial *F = 1.88 (*P* = 0.1835), so two lines have insignificantly different slopes.**

(NB: Should adjust the 2 inferences on β_2 and β_3 for multiplicity.)