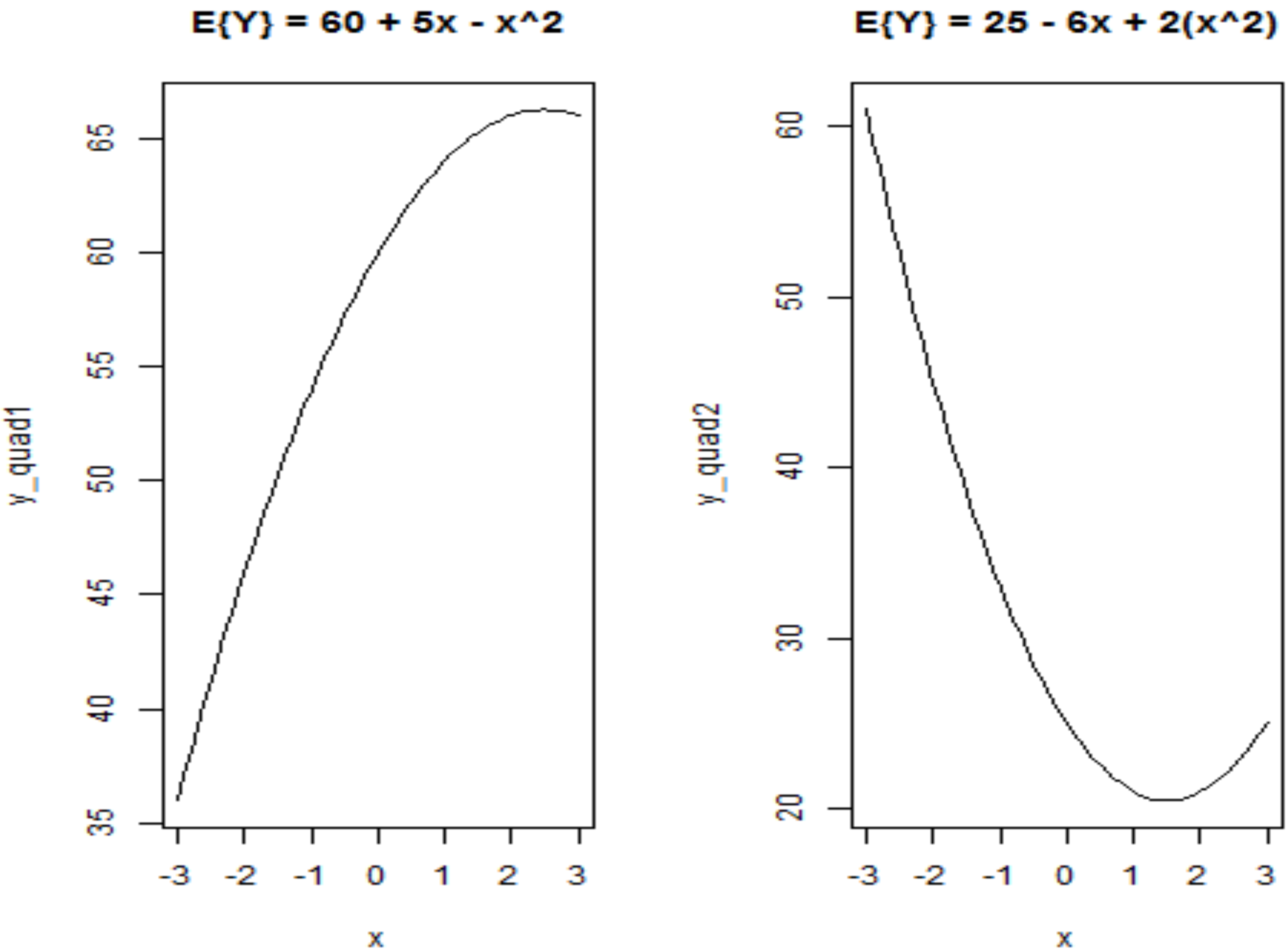
Regression Models for Quantitative/Numerical and Qualitative/Categorical Predictors

KNNL – Chapter 8

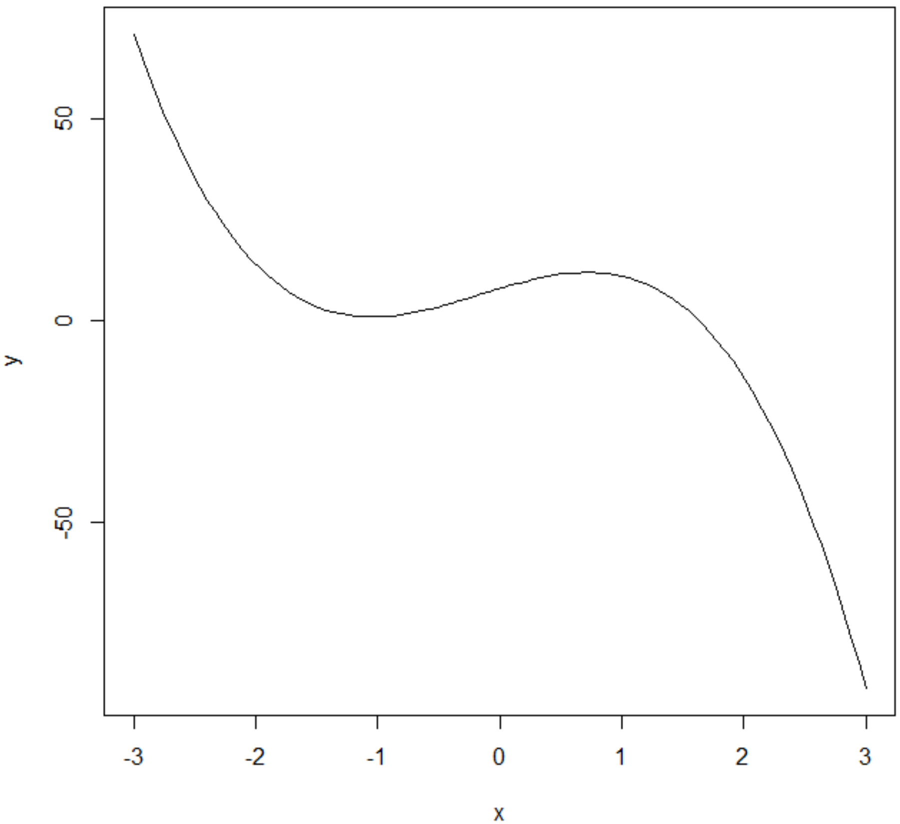
Polynomial Regression Models

- Useful in two 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 in model, creates *p*-1 "bends"
  - 2nd order Model: $E\{Y\} = \beta_0 + \beta_1 x + \beta_2 x^2$ has one bend
  - 3rd order Model: $E\{Y\} = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3$ has two bends
- Response Surfaces with 2 (or more) predictors
  - 2nd order model with 2 Predictors:

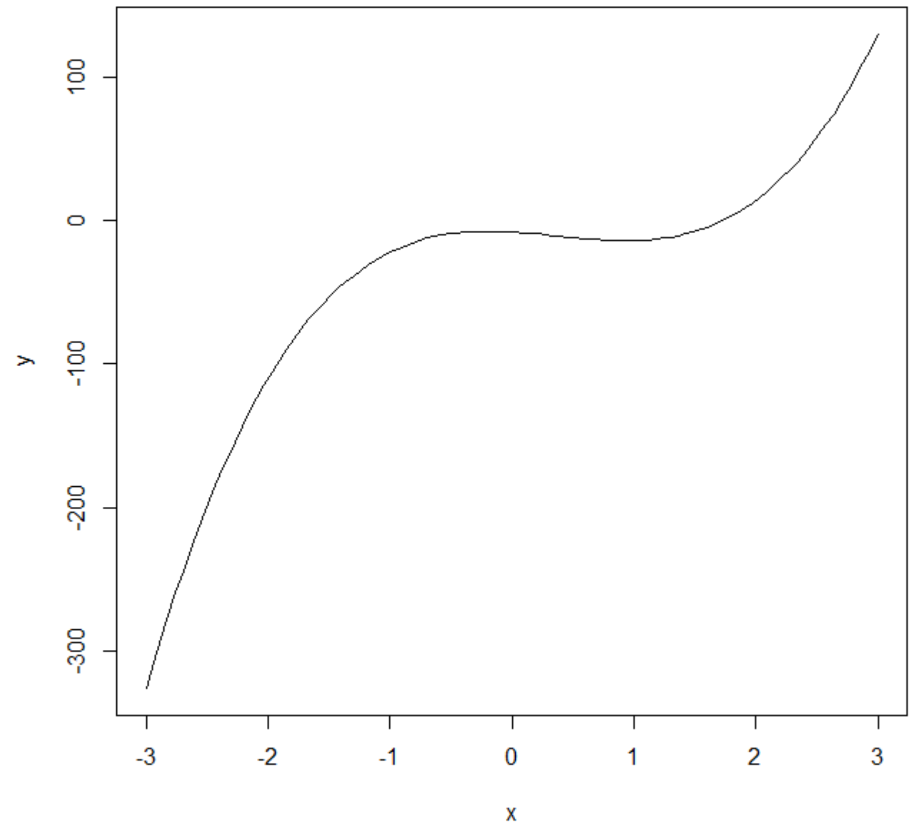
$$E\{Y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{11} x_1^2 + \beta_{22} x_2^2 + \beta_{12} x_1 x_2 \qquad x_1 = X_1 - \overline{X}_1 \quad x_2 = X_2 - \overline{X}_2$$

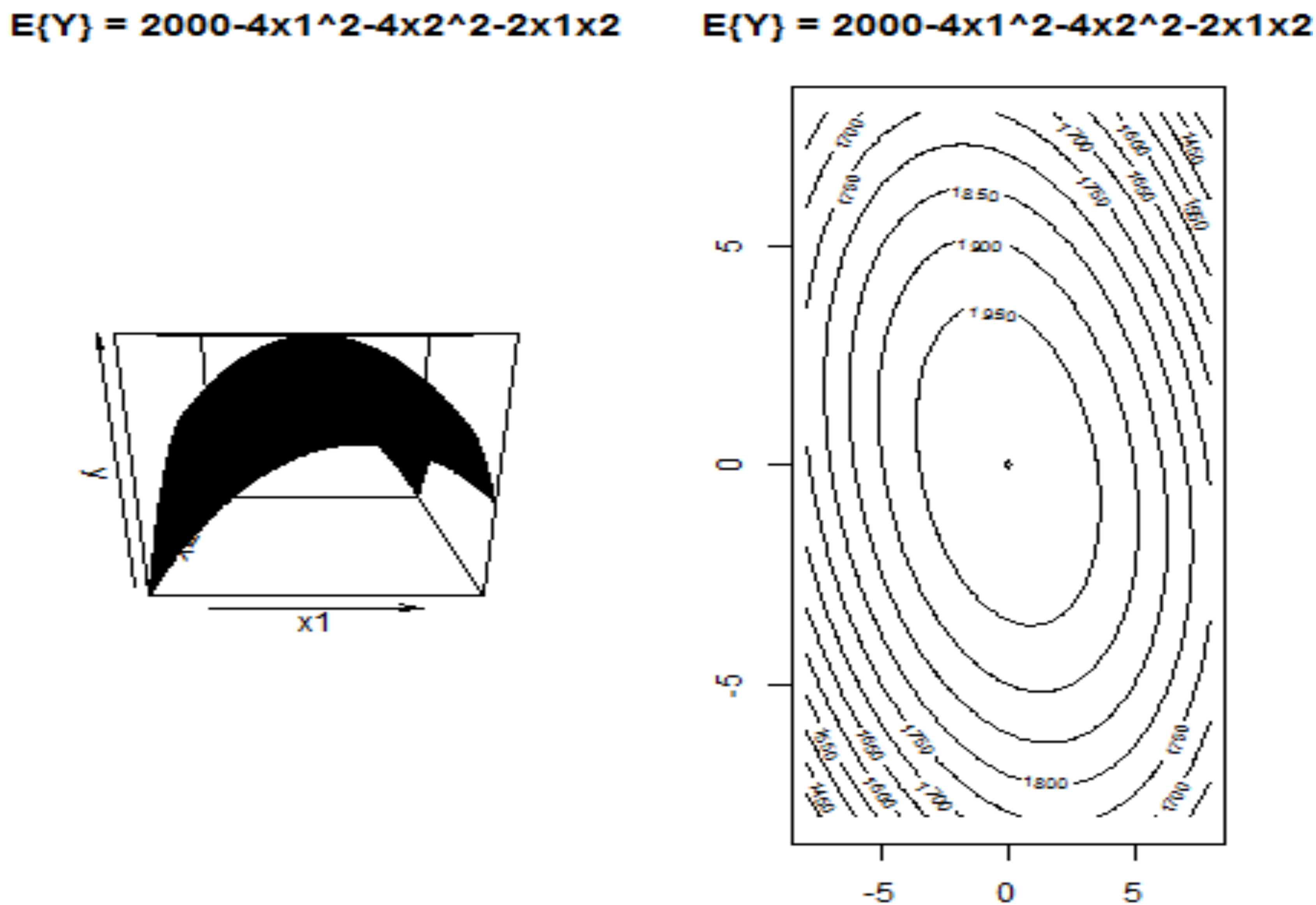

E{Y} = 60 + 5x - x^2
y_quad1
35
40
45
50
55
60
65
-3
-2
-1
0
1
2
3
x
E{Y} = 25 - 6x + 2(x^2)
y_quad2
20
30
40
50
60
-3
-2
-1
0
1
2
3
x

$-4x^3 - 2x^2 + 9x + 8$



$9x^3 - 10x^2 - 5x - 8$




E{Y} = 2000-4x1^2-4x2^2-2x1x2
y
x1
E{Y} = 2000-4x1^2-4x2^2-2x1x2
1700
1750
1850
1900
1950
1800
1650
1600
1450
5
0
-5
-5
0
5

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.
- Fit models in original units, or back-transform for plotting on original scale* (see below for quadratic)

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

$$= b_0 + b_1 X - b_1 \overline{X} + b_{11} X^2 - 2b_{11} X \overline{X} + b_{11} \overline{X}^2 = \left(b_0 - b_1 \overline{X} + b_{11} \overline{X}^2\right) + \left(b_1 - 2b_{11} \overline{X}\right) X + b_{11} \overline{X}^2 = b_o^{'} + b_1^{'} X + b_2^{'} X^2$$

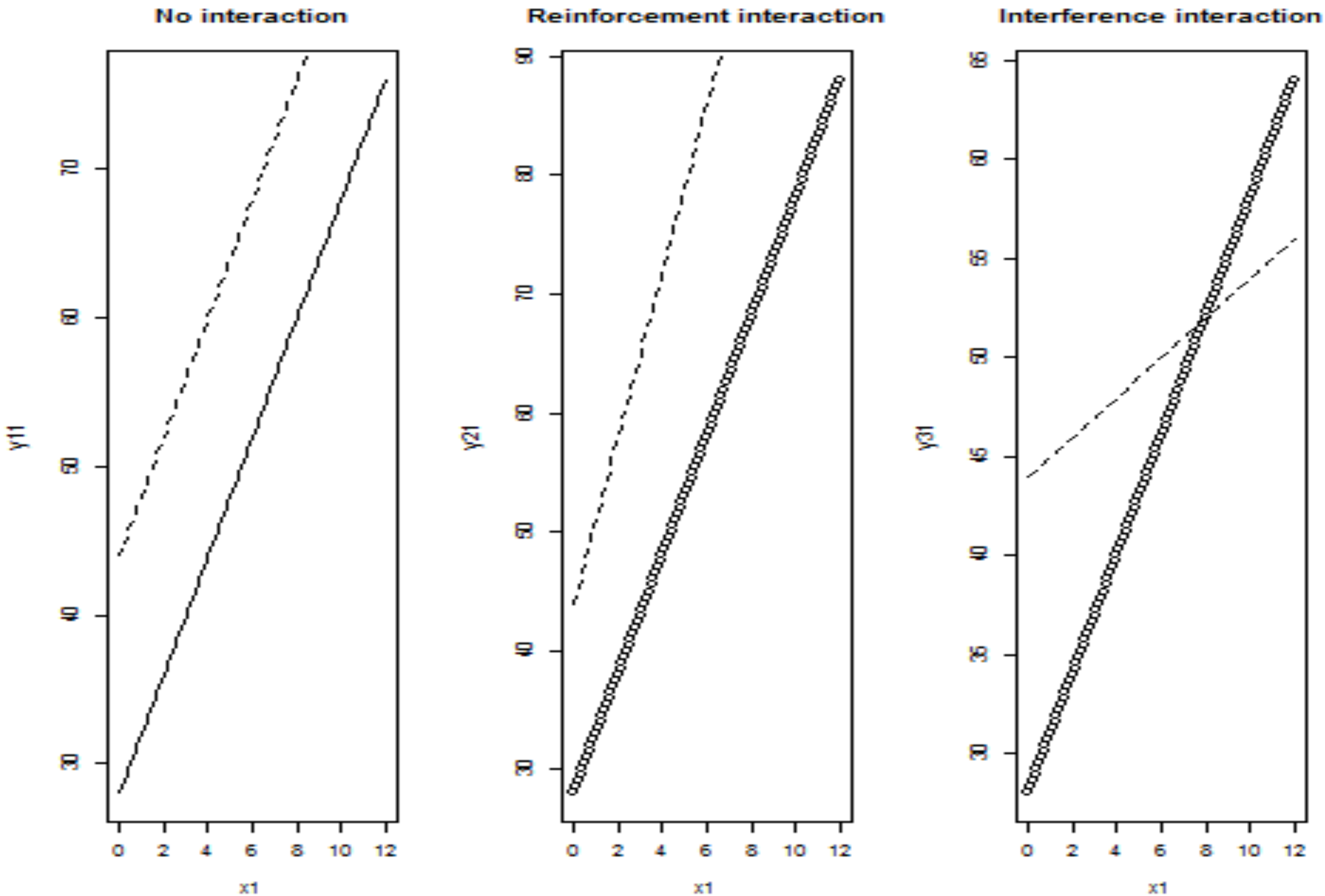
Regression Models with Interaction Term(s)

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

$$X_2 = 0 \Rightarrow E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 (0) + \beta_3 (X_1(0)) = \beta_0 + \beta_1 X_1$$

$$X_2 = 10 \Rightarrow E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2 (10) + \beta_3 (X_1(10)) = \beta_0 + \beta_1 X_1 + 10\beta_2 + 10\beta_3 X_1 =$$
$$= (\beta_0 + 10\beta_2) + (\beta_1 + 10\beta_3) X_1$$

Testing Hypothesis of no interaction: $H_0 : \beta_3 = 0 \quad H_A : \beta_3 \neq 0$



No interaction
Reinforcement interaction
Interference interaction
y11
y21
y31
x1
x1
x1

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
- 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

Example – Salary vs Experience by Region

Suppose Y = salary, X_1 = experience, and we also want to include 3 regions as predictors. Define the following:

$$X_2 = \begin{cases} 1 & \text{if Region 1} \\ 0 & \text{otherwise} \end{cases}$$

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

$$X_4 = \begin{cases} 1 & \text{if Region 3} \\ 0 & \text{otherwise} \end{cases}$$

If we include all three of the dummy variables X_2, X_3, and X_4, the problem is that the design matrix $\boldsymbol{X}$ will not be of full rank, and $\boldsymbol{X'X}$ will not be invertible:

$$\boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & 1 & 0 & 0 \\ 1 & X_{12} & 1 & 0 & 0 \\ 1 & X_{13} & 0 & 1 & 0 \\ 1 & X_{14} & 0 & 1 & 0 \\ 1 & X_{15} & 0 & 0 & 1 \\ 1 & X_{16} & 0 & 0 & 1 \end{bmatrix}$$

Example – Salary vs Experience by Region

Define the following:

$$X_2 = \begin{cases} 1 & \text{if Region 1} \\ 0 & \text{otherwise} \end{cases}$$

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

$$X_4 = \begin{cases} 1 & \text{if Region 3} \\ 0 & \text{otherwise} \end{cases}$$

Solution:
Choose a "reference" region, say Region 3. Then use just the Region 1 dummy (X_2) and the Region 2 dummy (X_3)
(Note: it is arbitrary which region is the reference)
Then the first-order regression model will be:

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

Example – Salary vs Experience by Region

3 regions, Y=salary, X_1 = 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

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
- Can conduct General Linear Test to determine whether slopes differ (or t-test when qualitative predictor has *c*=2 levels)
- These models generalize to any number of quantitative and qualitative predictors

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: $(X_2 = 1, X_3 = 0): E\{Y\} = \beta_0 + \beta_1 X_1 + \beta_2(1) + \beta_3(0) + \beta_4 X_1(1) + \beta_5 X_1(0) = (\beta_0 + \beta_2) + (\beta_1 + \beta_4) X_1$

Region 2: $(X_2 = 0, X_3 = 1): E\{Y\} = (\beta_0 + \beta_3) + (\beta_1 + \beta_5) X_1$ Region 3: $(X_2 = 0, X_3 = 0): E\{Y\} = \beta_0 + \beta_1 X_1$