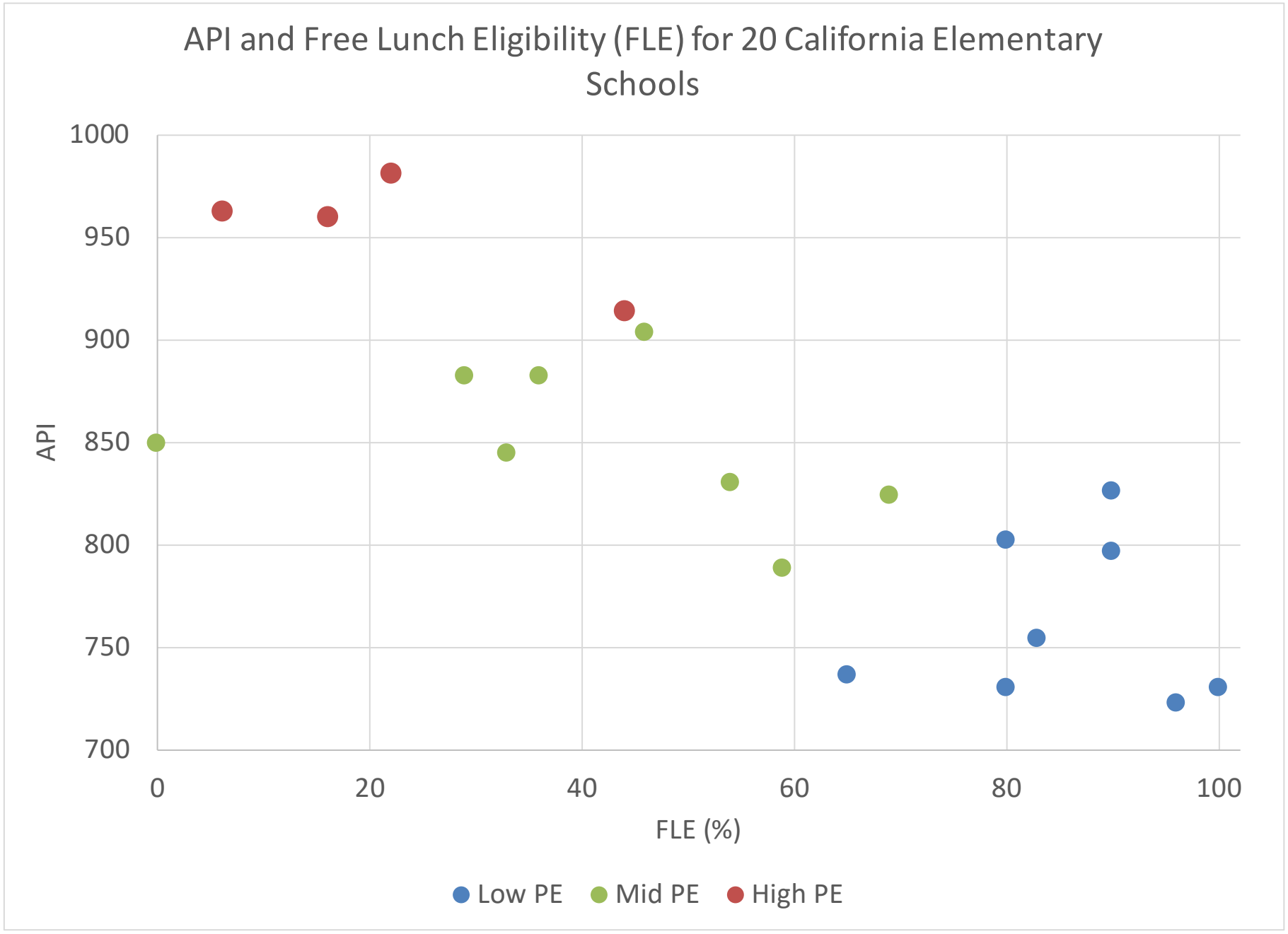
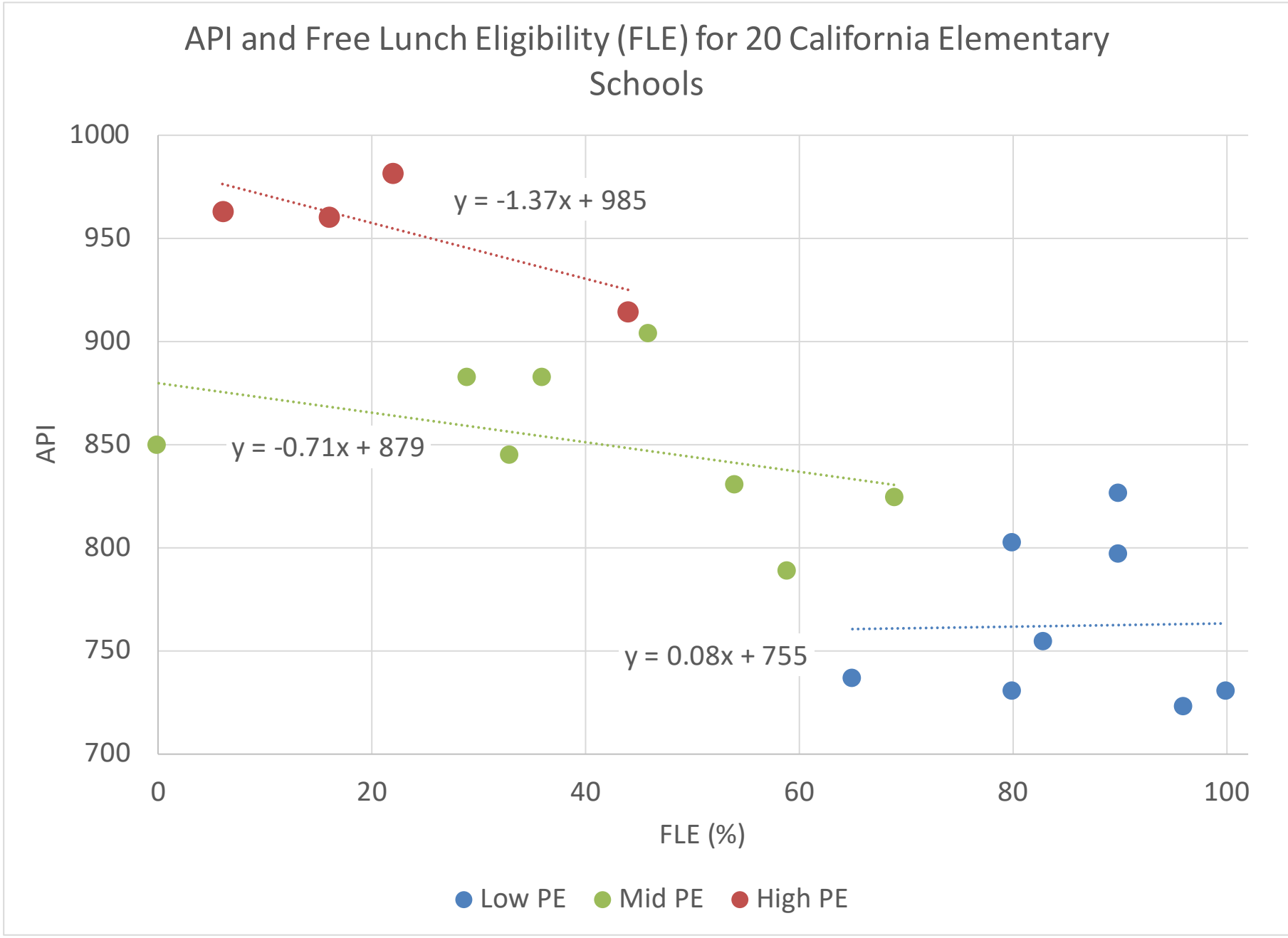
Chapter 3

Multiple Regression

Learning Objectives

- Derive the ordinary least-squares estimators of the regression coefficients when there are two or more right-hand variables in our model
- Fit a multiple regression model using the least-squares criterion
- Identify the conditions under which a multiple regression estimate is the same as the simple regression estimate
- Interpret a multiple regression coefficient


API and Free Lunch Eligibility (FLE) for 20 California Elementary Schools
1000
950
900
850
800
750
700
API
0
20
40
60
80
100
FLE (%)
Low PE
Mid PE
High PE


API and Free Lunch Eligibility (FLE) for 20 California Elementary Schools
1000
950
900
850
800
750
700
API
0
20
40
60
80
100
FLE (%)
y = -1.37x + 985
y = -0.71x + 879
y = 0.08x + 755
Low PE
Mid PE
High PE

Multiple Regression Model with K+1 Parameters

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_K X_{Ki} + e_i$$

If we have two right-hand variables, our model looks like this:

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

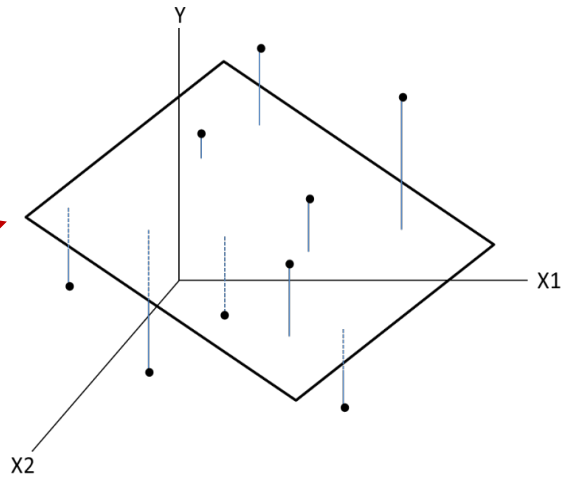


Figure 3.1. Illustration of a regression plane for a model with two explanatory variables, *X1* and *X2*

Deriving Least-Squares Estimator in Multiple Regression

$$\min_{b_k, k=0,1,\ldots,K} SSE = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_{1i} - \ldots - b_K X_{Ki})^2$$

$$b_0 : \sum_{i=1}^{N} 2(Y_i - b_0 - b_1 X_{1i} - \ldots - b_K X_{Ki})(-1) = 0$$

$$b_1 : \sum_{i=1}^{N} 2(Y_i - b_0 - b_1 X_{1i} - \ldots - b_K X_{Ki})(-X_{1i}) = 0$$

$$b_2 : \sum_{i=1}^{N} 2(Y_i - b_0 - b_1 X_{1i} - \ldots - b_K X_{Ki})(-X_{2i}) = 0$$

$$\vdots$$

$$b_K : \sum_{i=1}^{N} 2(Y_i - b_0 - b_1 X_{1i} - \ldots - b_K X_{Ki})(-X_{Ki}) = 0$$

Two RHS Variables

3 Equations, 3 Unknowns (b_0, b_1, b_2, after dividing by 2)

$$b_0 : \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})(-1) = 0$$

$$b_1 : \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})(-X_{1i}) = 0$$

$$b_2 : \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_{1i} - b_2 X_{2i})(-X_{2i}) = 0$$

OLS Formula, 2 RHS Variables

$$b_0 = \bar{Y} - b_1 \overline{X_1} - b_2 \overline{X_2}$$

$$b_1 = \frac{\sum_{i=1}^{N} x_{1i} y_i \sum_{i=1}^{N} x_{2i}^2 - \sum_{i=1}^{N} x_{1i} x_{2i} \sum_{i=1}^{N} x_{2i} y_i}{\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{2i}^2 - \left(\sum_{i=1}^{N} x_{1i} x_{2i} \right)^2}$$

$$b_2 = \frac{\sum_{i=1}^{N} x_{2i} y_i \sum_{i=1}^{N} x_{1i}^2 - \sum_{i=1}^{N} x_{1i} x_{2i} \sum_{i=1}^{N} x_{1i} y_i}{\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{2i}^2 - \left(\sum_{i=1}^{N} x_{1i} x_{2i} \right)^2}$$

When do these collapse to the simple OLS formulas?

$$b_1^s = \frac{\sum_{i=1}^{N} x_{1i} y_i}{\sum_{i=1}^{N} x_{1i}^2}$$

Perfect Multicollinearity:

Least squares fails if X_1 and X_2 are perfectly correlated (i.e., $X_1 = c*X_2$)

$$b_1 = \frac{\sum_{i=1}^{N} x_{1i} y_i \sum_{i=1}^{N} (cx_{1i})^2 - \sum_{i=1}^{N} x_{1i} cx_{1i} \sum_{i=1}^{N} cx_{1i} y_i}{\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} (cx_{1i})^2 - (\sum_{i=1}^{N} x_{1i} cx_{1i})^2}$$

$$= \frac{c^2 \left(\sum_{i=1}^{N} x_{1i} y_i \sum_{i=1}^{N} x_{1i}^2 - \sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{1i} y_i \right)}{c^2 \left(\sum_{i=1}^{N} x_{1i}^2 \sum_{i=1}^{N} x_{1i}^2 - (\sum_{i=1}^{N} x_{1i}^2)^2 \right)}$$

$$= \frac{0}{0}$$

20 Elementary Schools

- We have two regression models

$$API_i = 951.87 - 2.11FLE_i + e_i$$

$$API_i = 777.17 - 0.51FLE_i + 2.34PE_i + e_i$$

- Why is the coefficient on *FLE* smaller?
- Which model has the larger R^2? Why?

Interpreting MR Coefficients

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$$

- Suppose two individuals have the same X_2 value and their X_1 values differ by one unit.
- The predicted difference in their *Y* values is $\boldsymbol{b_1}$

Alternate Way to Do Least Squares in Multiple Regression

- Model: $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + e_i$

- You want to compute b_1. You can do it in two steps and you will get the same number as if you use the MR formula on slide 8.

- Step 1: $X_{1i} = c_0 + c_1 X_{2i} + v_{1i}$

- Step 2: $Y_i = b_0 + b_1 v_{1i} + e_i$

$$b_1 = \frac{\sum_{i=1}^{N} v_{1i} y_i}{\sum_{i=1}^{N} v_{1i}^2}$$

Simplifying Notation with Matrix Algebra

- One equation for each observation

$$Y_1 = b_0 + b_1 X_{11} + b_2 X_{21} + \ldots + b_K X_{K1} + e_1$$
$$Y_2 = b_0 + b_1 X_{12} + b_2 X_{22} + \ldots + b_K X_{K2} + e_2$$
$$\vdots$$
$$Y_N = b_0 + b_1 X_{1N} + b_2 X_{2N} + \ldots + b_K X_{KN} + e_N$$

- Combine into matrices

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & \cdots & X_{K1} \\ 1 & X_{12} & \cdots & X_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1N} & \cdots & X_{KN} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_K \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_N \end{bmatrix}$$

$$Y = X b + e$$

What We Learned

- How to solve the least-squares problem to fit a MR model.
- MR estimates differ from simple regression estimates if the right-hand-side variables are correlated with each other.
- How to apply the MR least-squares formula using a spreadsheet.
- How to interpret a MR coefficient
- R^2 increases when you add an X variable to a model