Chapter 2

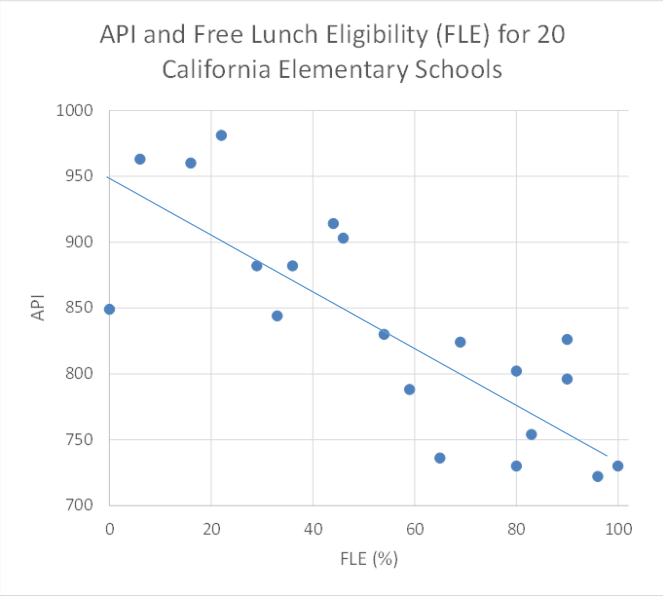
Simple Regression

Learning Objectives

- Explain what a simple regression model is
- Fit a simple regression model using the least-squares criterion
- Compute R^2 to measure how well the model fits the data
- Interpret the results from a simple regression model

How We Estimate This Model (Find the Best Line)

Figure 1.5. It sure looks like API decreases with FLE.



Simple regression model (note: no Greek letters!)

Prediction part

$$Y_i = b_0 + b_1 X_i + e_i$$

Outcome variable

RHS variable

Error, how much we miss by if we use X_i to predict Y_i.

The Least-squares Criterion: Minimize the Sum of Squared Errors (Why?)

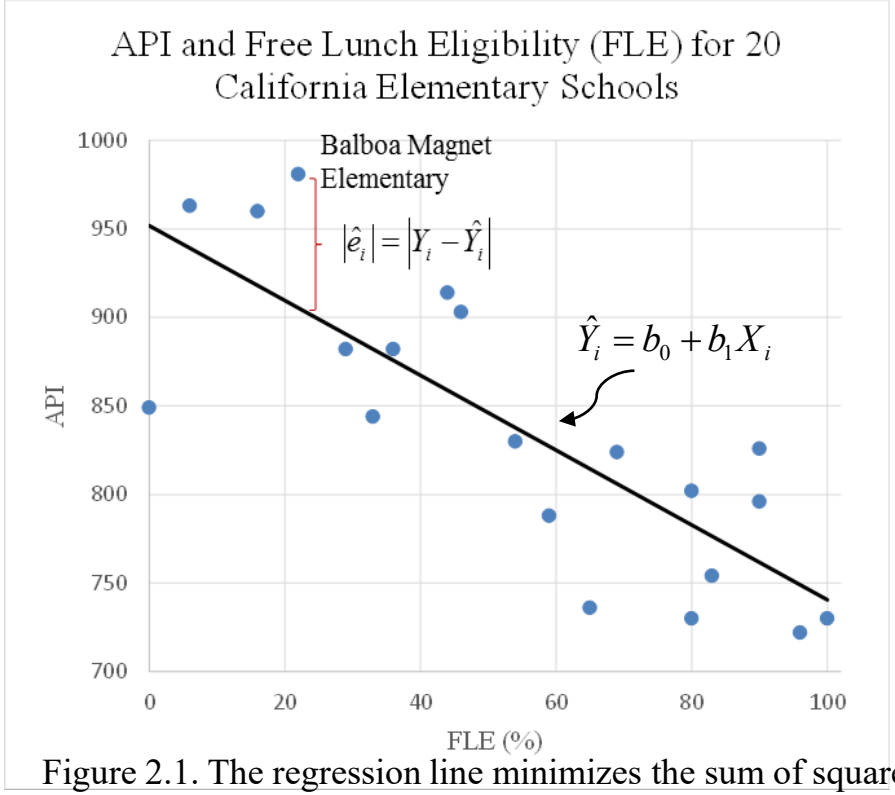


Figure 2.1. The regression line minimizes the sum of squared errors

Sum of Squared Errors (SSE):

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Goal: Find the b_0 and b_1 that minimize SSE:

$$\min_{b_0, b_1} SSE = \sum_{i=1}^{N} (Y_i - b_0 - b_1 X_i)^2$$

Derive the Normal Equations

$$\min_{b_0, b_1} SSE = \sum_{i=1}^{N} \left(Y_i - b_0 - b_1 X_i \right)^2$$

$$b_0 : \sum_{i=1}^{N} 2\left(Y_i - b_0 - b_1 X_i \right)(-1) = 0$$

$$b_1 : \sum_{i=1}^{N} 2\left(Y_i - b_0 - b_1 X_i \right)(-X_i) = 0$$

Solve the Normal Equations for b_0

$$\sum_{i=1}^{N}(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{N}(Y_i) - \sum_{i=1}^{N}(b_0) - \sum_{i=1}^{N}(b_1 X_i) = 0$$

$$\sum_{i=1}^{N}(Y_i) - Nb_0 - b_1 \sum_{i=1}^{N} X_i = 0$$

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Solve the Normal Equations for b_1 *(Substituting for b_0)*

$$\sum_{i=1}^{N}(Y_i - (\overbrace{\overline{Y} - b_1\overline{X}}^{b_0}) - b_1 X_i)(-X_i) = 0$$

$$\sum_{i=1}^{N}\left[(Y_i - \overline{Y}) - b_1(X_i - \overline{X})\right](X_i) = 0$$

$$\sum_{i=1}^{N}\left[\overbrace{X_i(Y_i - \overline{Y})}^{y_i} - b_1 X_i \overbrace{(X_i - \overline{X})}^{x_i}\right] = 0$$

$$b_1 = \frac{\sum_{i=1}^{N} X_i y_i}{\sum_{i=1}^{N} X_i x_i}$$

Where small-x and small-y are the deviations of X_i and Y_i from their means

Two equivalent ways to write the OLS formula for b_1

$$b_1 = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right) y_i}{\sum_{i=1}^{N} \left(X_i - \bar{X}\right) x_i} = \frac{\sum_{i=1}^{N} X_i y_i - \sum_{i=1}^{N} \bar{X} y_i}{\sum_{i=1}^{N} X_i x_i - \sum_{i=1}^{N} \bar{X} x_i} = \frac{\sum_{i=1}^{N} X_i y_i - \bar{X} \sum_{i=1}^{N} y_i}{\sum_{i=1}^{N} X_i x_i - \bar{X} \sum_{i=1}^{N} x_i} = \frac{\sum_{i=1}^{N} X_i y_i - \bar{X} * 0}{\sum_{i=1}^{N} X_i x_i - \bar{X} * 0} = \frac{\sum_{i=1}^{N} X_i y_i}{\sum_{i=1}^{N} X_i x_i}$$

OLS Estimator of $Y_i = b_0 + b_1 X_i + e_i$

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

$$b_0 = \overline{Y} - b_1 \overline{X}$$

Difference between Sample and Population Regression

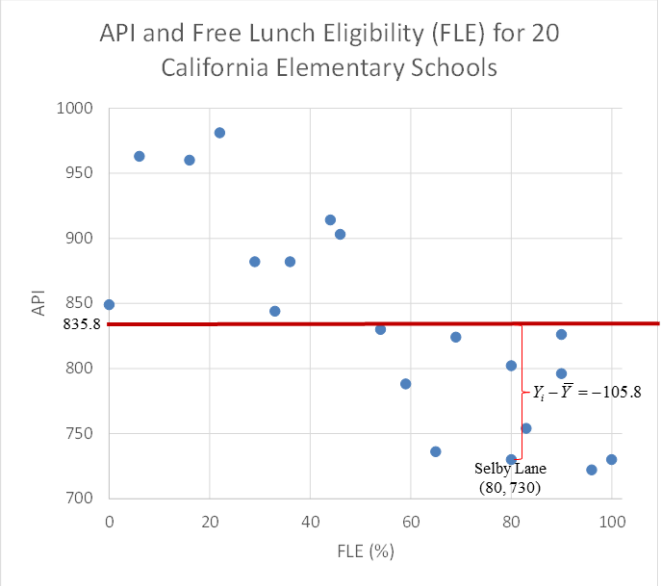
$$Y_i = b_0 + b_1 X_i + e_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Sometimes econometricians use hats "^" to indicate things they've actually estimated

$$Y_i = \hat{b}_0 + \hat{b}_1 X_i + \hat{e}_i$$

How Good Is Your Fit? Predicting with the Mean



Take Selby Lane Elementary (X,Y) = (80,730).

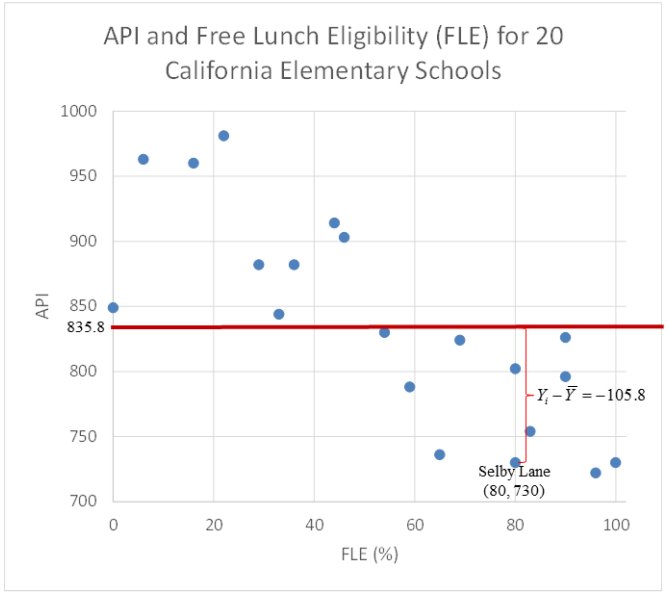
The deviation from the mean is 730 – 835.8 = –105.8.

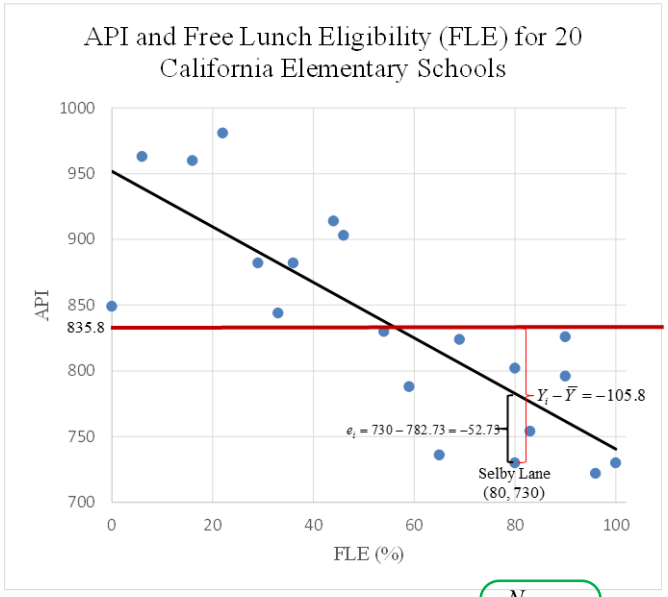
Without a regression model, we would only have the mean to work with, and it would not be a very good predictor of *Y*.

$$TSS = \sum_{i=1}^{N} y_i^2$$

Misses if predict with sample mean (Numerator in sample variance)

How Good Is Your Fit? The R-squared





$$TSS = \sum_{i=1}^{N} y_i^2$$

Variation in Y not explained by mean (Numerator in sample variance)

$$R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} y_i^2}$$

Variation in Y not explained by regression model

Interpreting Regression Coefficients

$$Y_i = b_0 + b_1 X_i + e_i$$

- Suppose two individuals have *X* values that differ by one unit.
- The predicted difference in their *Y* values is $\boldsymbol{b_1}$

What We Learned

- How to solve the least-squares problem to fit a simple regression model.
- How to apply the least-squares formula using a spreadsheet.
- R^2 is a useful characteristic of your model, but maximizing R^2 is often not the objective of your analysis.