Optimización combinatoria para machine learning
Universidad de los Andes

Andrés Gómez

June 20, 2023

Linear regression

Consider the dataset given in the book from Thomas, G.S. (1990). *The rating guide to life in America's small cities*, summarized in the figure below.

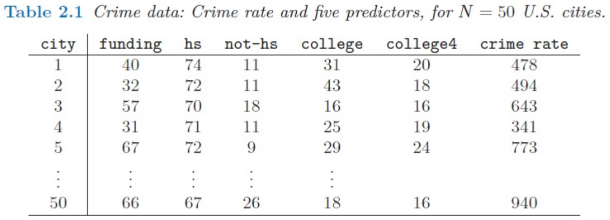

Table 2.1 *Crime data: Crime rate and five predictors, for $N = 50$ U.S. cities.*

city	funding	hs	not-hs	college	college4	crime rate
1	40	74	11	31	20	478
2	32	72	11	43	18	494
3	57	70	18	16	16	643
4	31	71	11	25	19	341
5	67	72	9	29	24	773
⋮	⋮	⋮	⋮	⋮		
50	66	67	26	18	16	940

The full dataset is in supporting file. The dataset includes information about 50 cities in the US, and in particular contains the following information about each city.

funding Annual police funding

hs Percentage of 25 years or older with four years of high school

not-hs Percentage of 16- to 19-old not in high school

college Percentage of 18- to 24-years old in college

college4 Percentage of 25+ years or older with a least four years of college

Finally, it also contains the crime rate at each city. The goal is to understand how the crime rate is affected by each predictor, and perhaps predict the crime rate of new cities.

Which model is best?

Consider the following models.

1. Model 1 postulates that

   $$\text{crime\_rate} \approx 163.6 + 10.0 \times \text{funding} - 1.2 \times \text{hs} + 11.0 \times \text{not\_hs} + 2.0 \times \text{college} + 1.0 \times \text{college4}$$

   The coefficient of determination (or R^2) of this model is 0.904

2. Model 2 postulates that

   $$\text{crime\_rate} \approx 10.4 \times \text{funding} + 14.3 \times \text{not\_hs} + 3.3 \times \text{college}$$

   The coefficient of determination (or R^2) of this model is 0.903.

3. Model 3 postulates that

   $$\text{crime\_rate} \approx 299.7 + 11.0 \times \text{funding}$$

   The coefficient of determination (or R^2) of this model is 0.899

Questions Answer the following questions:

1. Which model do you prefer? Why?

2. According to your preferred model: does funding cause an increase in the crime rate?

3. According to your preferred model: does a large population with college degrees (or at least four year of college) cause an increase in crime?