You Had Me at Linear Regression

Greg Damico
8/23/23

// FLATIRON SCHOOL

Why Linear Regression ?

- LR is a fundamental tool in the data scientist’s kit.
- LR can be done in an inferential mode in an attempt to discover the relationship between one variable and another.
- LR illustrates a technique that is relevant to many machine learning algorithms.

Linear Regression

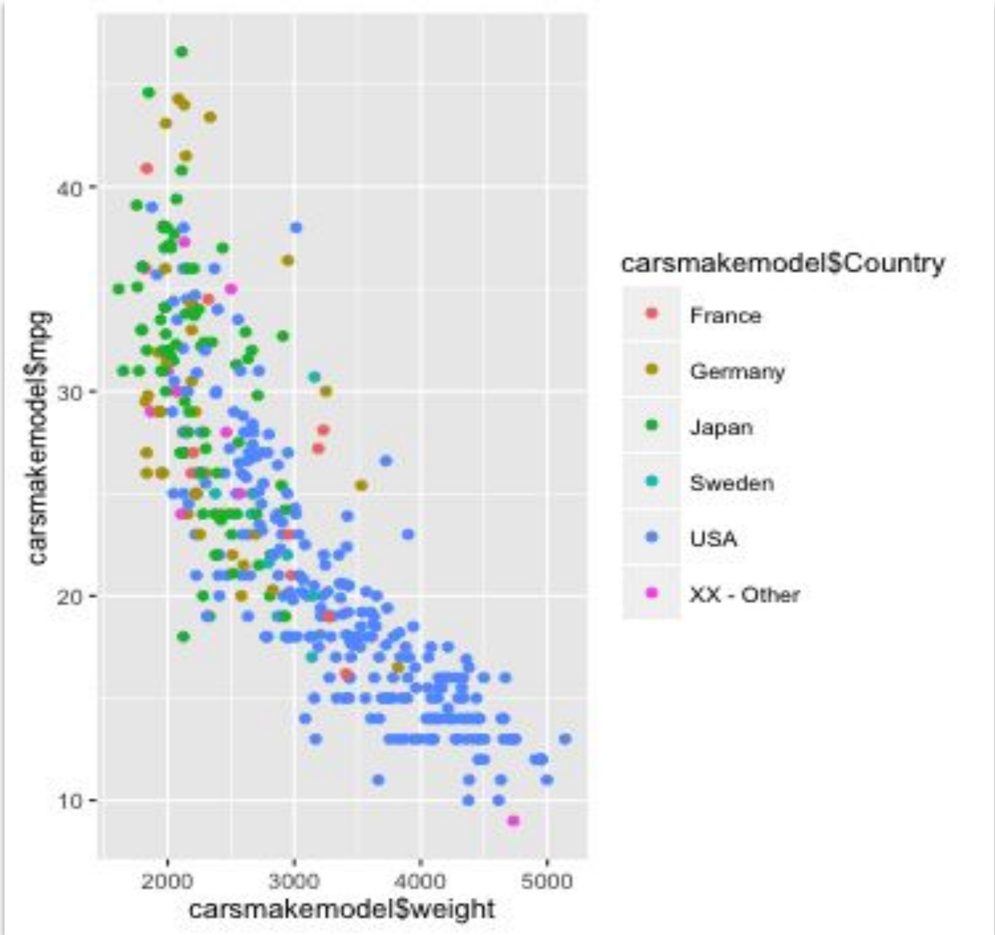
- Linear: models are lines
- Regression: dependent variable is continuously-valued
- Simple (the line is a function of a single variable) OR Multiple (the line is a function of multiple variables)

Inference and Prediction

- As population density increases, so do housing prices.
- As the number of trees decreases, the concentration of

CO_2 goes up.

Car Weight and MPG



Best-Fit Line

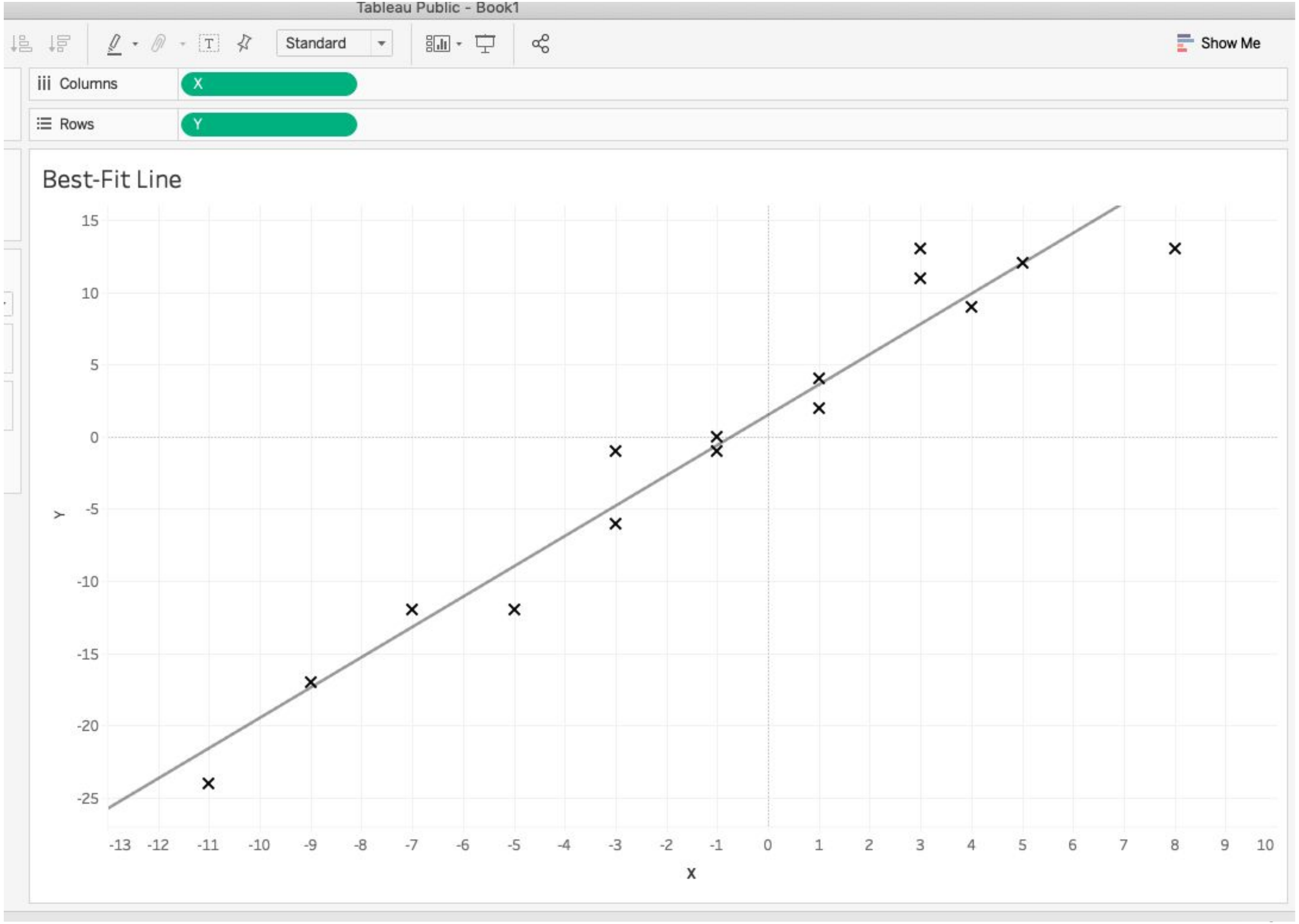


Tableau Public - Book1

Columns X

Rows Y

Models as Idealizations

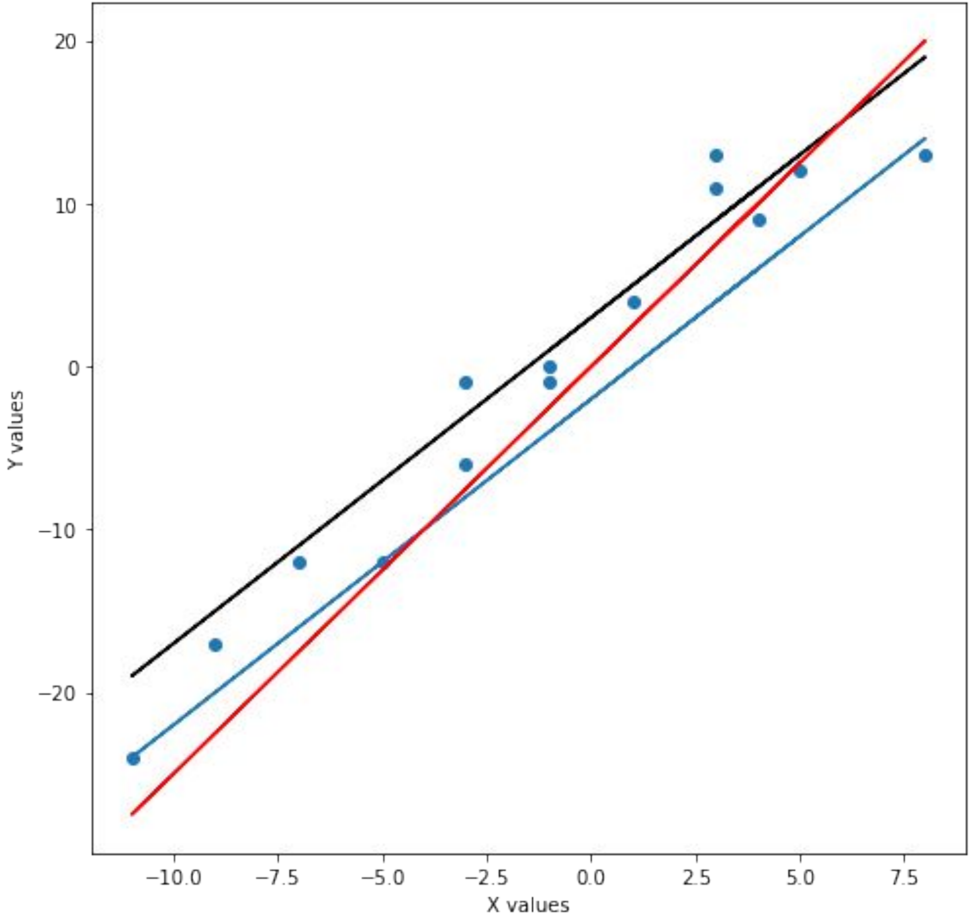
- Richard Levins in the 1960s: scientific models have ***tradeoffs*** between generality and precision.
  - The more precise I make my model, trying to get ever more accurate predictions, the less general it will tend to be.
  - An extremely simple model won't be very precise, but it will have more applications. An extremely complex model will be more precise, but it will have much less application.
- Linear regression is a prime example of a simple model.

A Line as a Model

- Predictions for *all* values of the X variable
  - Model shape: $\hat{y} = \beta_1 x + \beta_0$
- Error as the distance between real and predicted values $E = y - \hat{y}$
  $E^2 = (y - \hat{y})^2$

Goal: Minimize Error

- Which of these lines fits the data best?



Demo!

Let’s run some code to illustrate this idea of a linear regression.