Chapter 1: Linear Regression with One Predictor* (STA4210)

Rohit Patra

September 5, 2018

1 Deterministic vs statistical relationship

We have two variables X and Y, measured on each of n subjects, and we want to see how Y depends on X. We think the value of X "decides" or "predicts" the value of Y.

- X is called the **predictor** or **independent** variable.
- Y is the **dependent** or the **response** variable.

1.1 Deterministic Relationship

Table 1: Profits from Selling Toys

Data	Units Sold (X)	Profit in USD (Y)
1	75	75 $\times$50
2	25	25 $\times$50
3	130	130 $\times$50
4	3	3 $\times$50
5	40	40 $\times$50
6	1	1 $\times$50

Table 2: Stopping distance of a car and initial speed

Data	Initial Speed (X)	Stopping Distance in ft(Y)
1	4	2
2	4	10
3	7	4
4	7	22
⋮	⋮	⋮
50	25	85

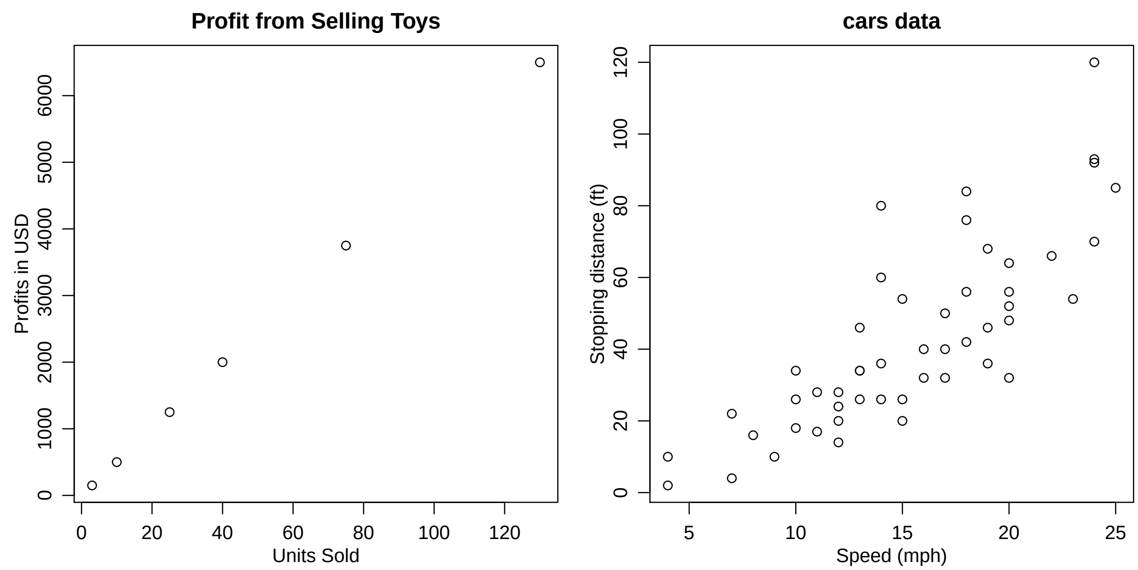

Figure 1: Plot of two datasets. Left panel: A deterministic model. Right panel: A statistical relationship

1.2 Statistical Relationship

2 Linear Regression Model

To model the relationship for the data set in 2 we will use a the following regression model. We will assume the following relationship between X and Y

$$Y = \beta_0 + \beta_1 X + \epsilon \tag{1}$$

where ϵ is random variable.

- X is the “predictor” or “covariate” or “independent variable”.

[*]Notes adapted and borrowed from classes taught by Deb Burr and Larry Winner at University of Florida

- Y is the "dependent variable" or the "response".
- ϵ is the unobserved noise.
- We assume that ϵ is independent of X.
- We assume that ϵ has normal distribution with mean 0 and variance σ^2, i.e., $\epsilon \sim N(0, \sigma^2)$.
- Note that the variance of ϵ does not depend on X. This assumptions is known as the "assumption of constant variance."
- The dependence of Y on X is linear.
- Is this relationship deterministic?
- The data is n observations of X (call them $X_1, \ldots, X_n$) and n observations of Y (call them $Y_1, \ldots, Y_n$)
- Each Y_i depends on β_0, β_1, X_i, and ϵ_i.
- Our data $(X_1, Y_1), \ldots, (X_n, Y_n)$.
- We do not know $\epsilon_1, \ldots, \epsilon_n$.

Goal: Estimate β_0 and β_1 using the data we have.

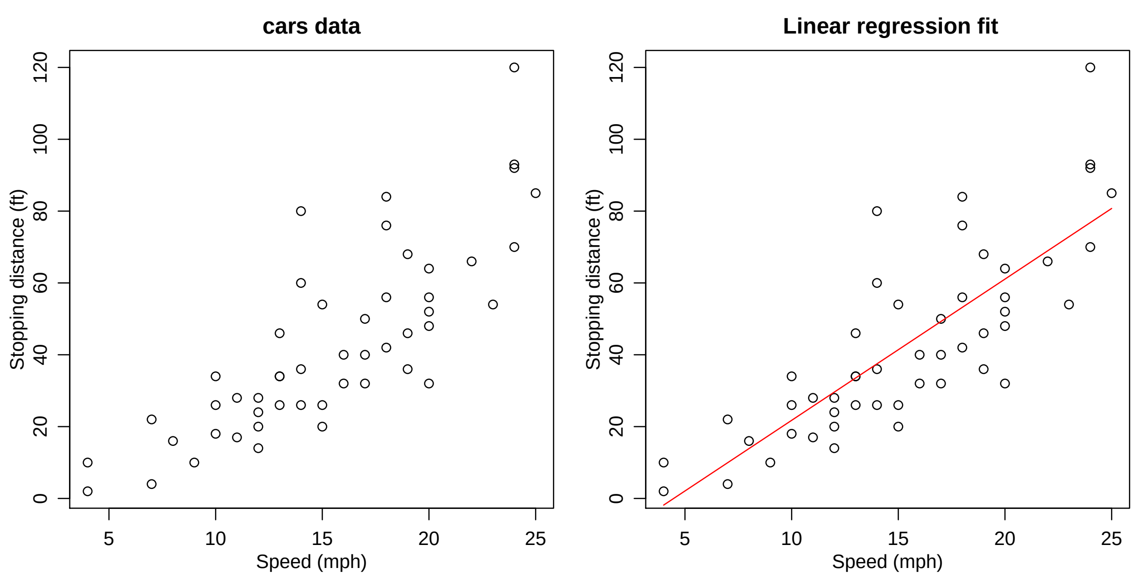

Figure 2: Linear fit for the cars data in Section 1.2

Why linear? The regression is assumed linear. Why is this a good assumption?

- Any curve is, over short portions, nearly a straight line.

- It is often true that if the regression line is not linear, it is possible to transform y, x or both, into new variables in such a way that the relationship in terms of the new variables is linear.

- The methods used to analyze linear regression can be generalized to handle the linear regression of y on more than one variable, and these, in turn can be used to handle many forms of nonlinear regressions.

2.1 Method of Least squares

For the correct parameter values β_0 and β_1, the *deviation* of the observed values to its expected value, i.e.,

$$Y_i - \beta_0 - \beta_1 X_i,$$

should be *small*.

We try to *minimize* the sum of the n squared deviations, i.e., we can try to minimize

$$Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

as a function of β_0 and β_1. In other words, we want to minimize the sum of the squares of the vertical deviations of all the points from the line.

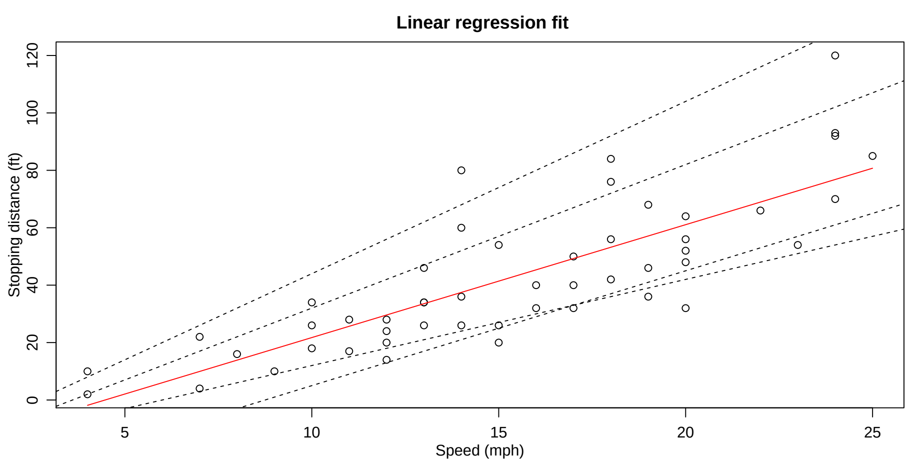

Figure 3: Looking at the fit of different regression lines.

The value of β_0 and β_1 that minimize the score (Q) is:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}, \qquad (2)$$

$$\hat{\beta}_0 = \overline{Y} - \beta_1 \overline{X}, \qquad (3)$$

where $\overline{X} = \sum_{i=1}^{n} X_i/n$ and $\overline{Y} = \sum_{i=1}^{n} Y_i/n$.

For the car dataset, $\hat{\beta}_1 = 3.932$, $\hat{\beta}_0 = -17.759$ and $Q(\hat{\beta}_0, \hat{\beta}_1) = 11,353.52$.

$Q(-18, 5) = 25,747$