

# Linear Regression (part 1)

Predictive Modeling & Statistical Learning

Gaston Sanchez

CC BY-SA 4.0

# Linear Regression

# Advertising Data

*# file in folder data/ of github repo*

```
Advertising <- read.csv("data/Advertising.csv", row.names = 1)
```

	TV	Radio	Newspaper	Sales
1	230.1	37.8	69.2	22.1
2	44.5	39.3	45.1	10.4
3	17.2	45.9	69.3	9.3
4	151.5	41.3	58.5	18.5
5	180.8	10.8	58.4	12.9
6	8.7	48.9	75.0	7.2
7	57.5	32.8	23.5	11.8
8	120.2	19.6	11.6	13.2

(first 8 rows)

# Advertising Data

**Advertising** consists of:

- ▶ the Sales of a product in 200 different markets
- ▶ the advertising budgets for three different media:
  - TV
  - Radio
  - Newspaper
- ▶ It is not possible to directly increase the sales of the product
- ▶ On the other hand, it is possible to control the advertising expenditure in each of the 3 media

# Introduction

- ▶ Suppose we observe a quantitative response  $Y$  and  $p$  different predictors,  $X_1, X_2, \dots, X_p$
- ▶ We assume there is some relationship between  $Y$  and  $[X_1, \dots, X_p]$ . that can be written in a general form as

$$Y = f(X_1, X_2, \dots, X_p) + \epsilon$$

- ▶  $f$  represents the systematic information that the predictors provide about  $Y$
- ▶  $\epsilon$  represents an *error* term that is a catch-all for what we miss with the model

# Data set Advertising

Response:

- ▶  $Y$ : Sales

Predictors:

- ▶  $X_1$ : TV
- ▶  $X_2$ : Radio
- ▶  $X_3$ : Newspaper

Relationship:

$$\text{Sales} = f(\text{TV}, \text{Radio}, \text{Newspaper}) + \epsilon$$

# Linear relationship

- ▶ One possible form for  $f()$  is a linear relationship:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

- ▶ It assumes a linear dependence of  $Y$  on the predictors
- ▶  $\beta_0, \beta_1, \dots, \beta_p$  are unknown constants also known as the model *coefficients* or *parameters*
- ▶ The linearity is in the parameters (i.e. coefficients)

# Linear relationship

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon$$

$$\text{Sales} = \beta_0 + \beta_1 \text{TV} + \beta_2 \text{Radio} + \beta_3 \text{Newspaper} + \epsilon$$



# Introduction

The challenge involves finding parameter estimates denoted by

$$\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$$

that provide the “best” approximation for  $Y$ :

$$Y \approx \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$$

or more commonly

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \dots + \hat{\beta}_p X_p$$

# Introduction

- ▶ Linearity is a BIG assumption.
- ▶ True regression functions are rarely linear.
- ▶ Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

# Simple Linear Regression

# Simple Linear Regression with one predictor

- ▶ Simple Linear Regression = Univariate regression
- ▶ One predictor variable  $X$  and one response variable  $Y$
- ▶ One predictor variable  $x$  and one response variable  $y$

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

# Simple Linear Regression with one predictor

We assume a linear model

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

where:

- ▶  $\beta_0$  and  $\beta_1$  are two unknown constants also known as *coefficients* or *parameters*
- ▶  $\beta_0$  represents the *intercept*
- ▶  $\beta_1$  represents the *slope*
- ▶  $\varepsilon$  is a vector of error terms

# Simple Linear Regression with one predictor

In vector notation:

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \boldsymbol{\varepsilon}$$

where:

- ▶  $\mathbf{y}$  is the vector representing the response variable
- ▶  $\mathbf{x}$  is the vector representing the predictor variable
- ▶  $\boldsymbol{\varepsilon}$  is the vector representing the error term

# Some vector-matrix notation

In matrix notation:

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times 2}{\mathbf{X}} \times \underset{2 \times 1}{\boldsymbol{\beta}} + \underset{n \times 1}{\boldsymbol{\varepsilon}}$$

which can be represented by:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} \\ 1 & x_{21} \\ 1 & x_{31} \\ \vdots & \vdots \\ 1 & x_{n1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

# Simple Linear Regression with one predictor

Note that if the data is center (mean = 0)

$$\mathbf{y} = \beta_1 \mathbf{x} + \varepsilon$$

then there is no intercept term  $\beta_0$



## Some vector-matrix notation

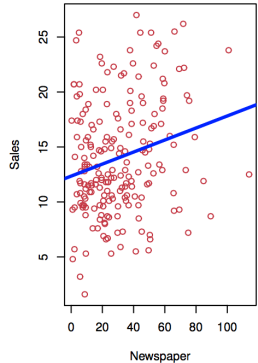
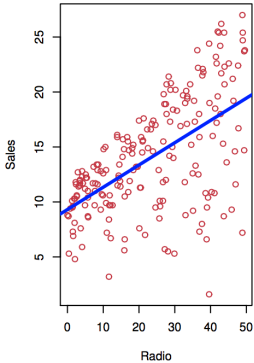
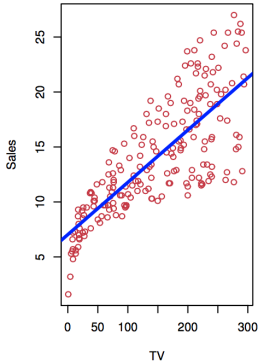
With centered data we have:

$$\underset{n \times 1}{\mathbf{y}} = \underset{n \times 1}{\mathbf{x}} \times \underset{n \times 1}{\beta} + \underset{n \times 1}{\boldsymbol{\varepsilon}}$$

which can be represented by:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{11} \\ x_{21} \\ x_{31} \\ \vdots \\ x_{n1} \end{bmatrix} [\beta_1] + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \varepsilon_3 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

# Various simple regressions



# Simple Linear Regression with one predictor

Assuming the model

$$\mathbf{y} = \beta_0 + \beta_1 \mathbf{x} + \epsilon$$

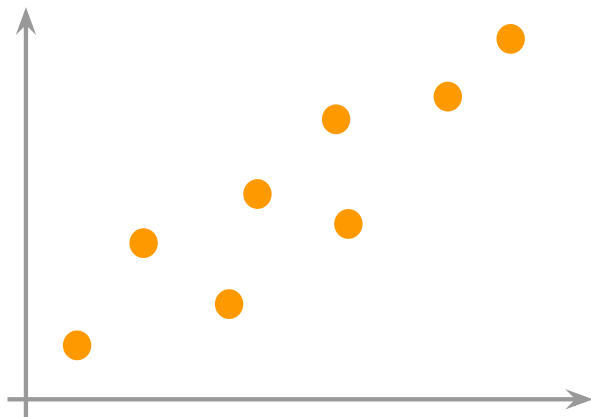
and given some estimates  $\hat{\beta}_0$  and  $\hat{\beta}_1$  for the model coefficients, we predict future sales using

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \hat{\beta}_1 \mathbf{x}$$

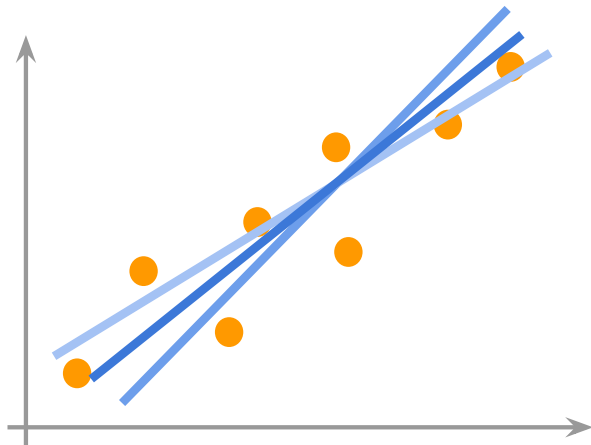
where  $\hat{\mathbf{y}}$  indicates a prediction of  $\mathbf{y}$

# Fitting a Line

## Fitting a Line

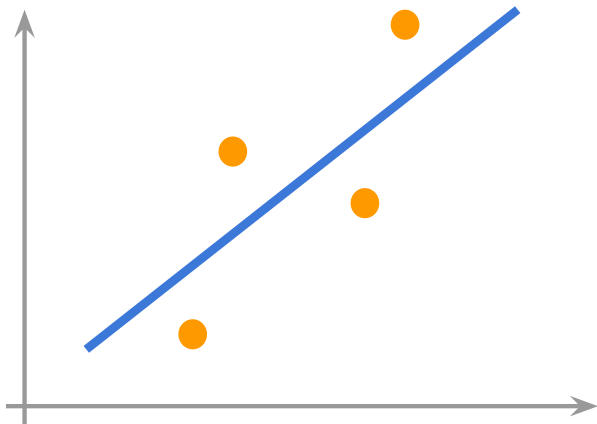


# Fitting a Line



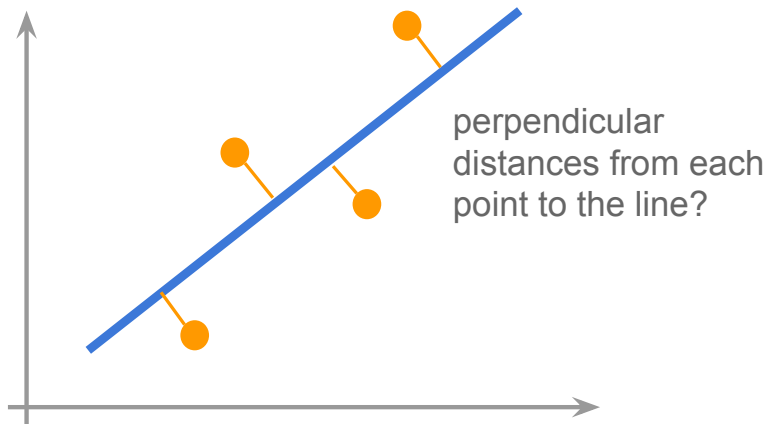
How to find the “best”  
fitting line?

# Fitting a Line

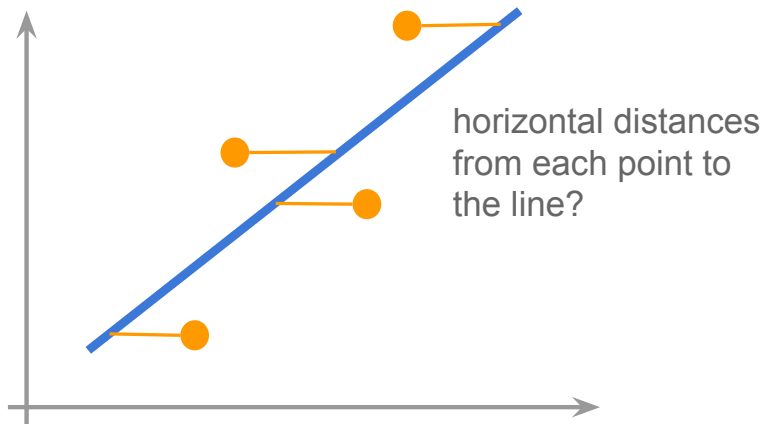




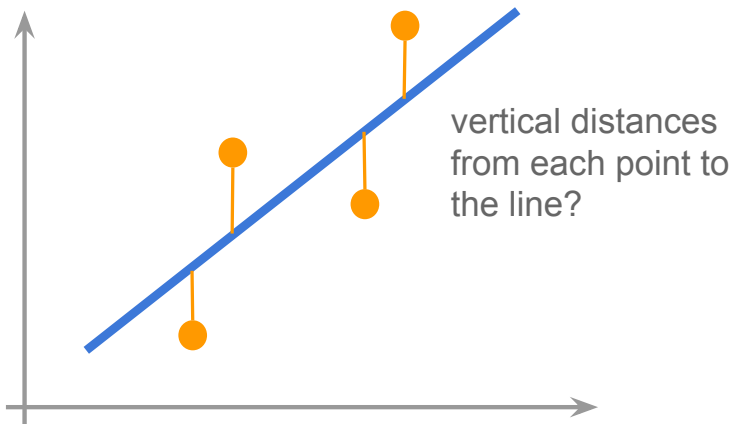
# Fitting a Line



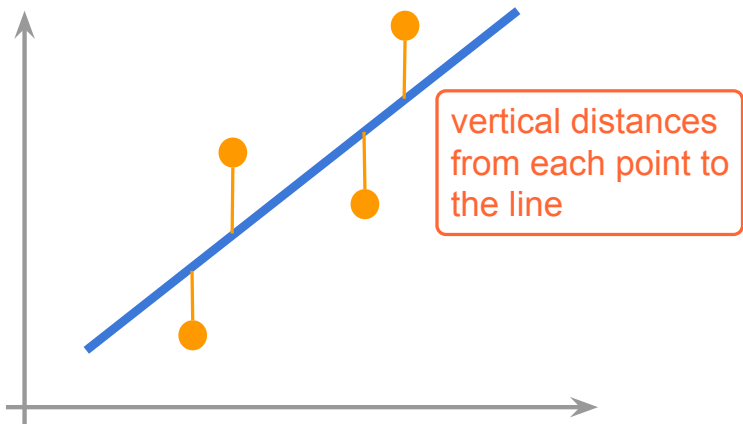
# Fitting a Line



# Fitting a Line



# Fitting a Line



# Estimation of Parameters

# Estimation of the parameters

- ▶ Let  $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  be the prediction for  $y$  based on the  $i$ th value of  $x$

# Estimation of the parameters

- ▶ Let  $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  be the prediction for  $y$  based on the  $i$ th value of  $x$
- ▶ Then  $e_i = y_i - \hat{y}_i$  represents the  $i$ th residual

# Estimation of the parameters

- ▶ Let  $y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$  be the prediction for  $y$  based on the  $i$ th value of  $x$
- ▶ Then  $e_i = y_i - \hat{y}_i$  represents the  $i$ th **residual**
- ▶ We define the **Residual Sum of Squares** (RSS) as

$$RSS = e_1^2 + e_2^2 + \cdots + e_n^2$$

- ▶ The **Least Squares** approach chooses  $\hat{\beta}_0$  and  $\hat{\beta}_1$  to minimize the RSS



# Estimation of the parameters

The starting point is to write the model as:

$$\mathbf{e} = \mathbf{y} - (\beta_0 + \beta_1 \mathbf{x})$$

For convenience we define a quadratic loss function

$$L = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

To minimize  $L$  we take partial derivatives with respect to each of the two parameters

# Estimation of the parameters

Thus,

$$\frac{\partial L}{\partial \beta_0} = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)(-1) = 0$$

and

$$\frac{\partial L}{\partial \beta_1} = 2 \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)(-x_i) = 0$$

# Estimation of the parameters

The solutions for  $\beta_0$  and  $\beta_1$  would be obtained by solving the so-called *normal equations*

$$\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i) = 0$$

and

$$\sum_{i=1}^n (x_i y_i - x_i \beta_0 - \beta_1 x_i^2) = 0$$

# Estimation of the parameters by OLS

The **Least Squares** coefficients are:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \quad \text{and} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

# Estimation of the parameters by OLS

Notice that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

is equivalent to:

$$\hat{\beta}_1 = \frac{\text{cov}(\mathbf{x}, \mathbf{y})}{\text{var}(\mathbf{x})}$$

# Example: Advertising Data

```
# number of observations  
n <- nrow(Advertising)  
  
# model matrix  
x <- Advertising$TV  
  
# response variable  
y <- Advertising$Sales
```

# Example: Advertising Data

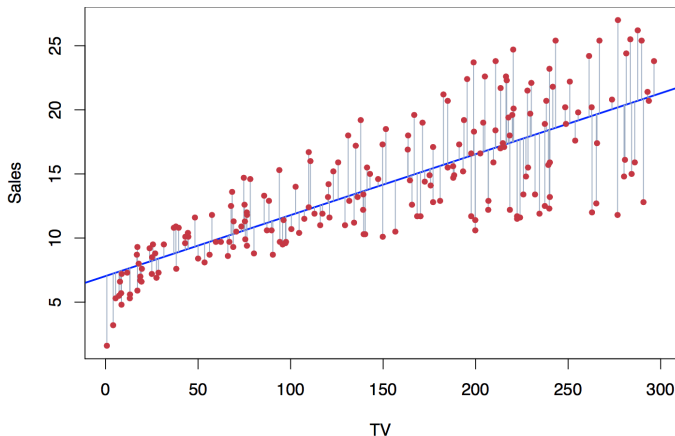
```
# slope
b1 <- cov(x, y) / var(x)
b1

## [1] 0.04753664

# intercept
b0 <- mean(y) - b1 * mean(x)
b0

## [1] 7.032594
```

# Example: Advertising Data



The least squares fit for the regression of Sales on TV.

In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.



# Another perspective

# Projection

Notice that:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

Can be expressed in vector notation as:

$$\hat{\beta}_1 = \frac{\mathbf{y}^\top \mathbf{x}}{\mathbf{x}^\top \mathbf{x}}$$

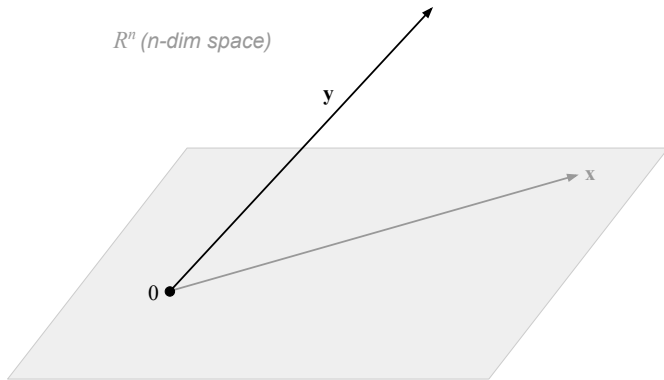
with  $\mathbf{x}$  and  $\mathbf{y}$  mean-centered.

# Projection

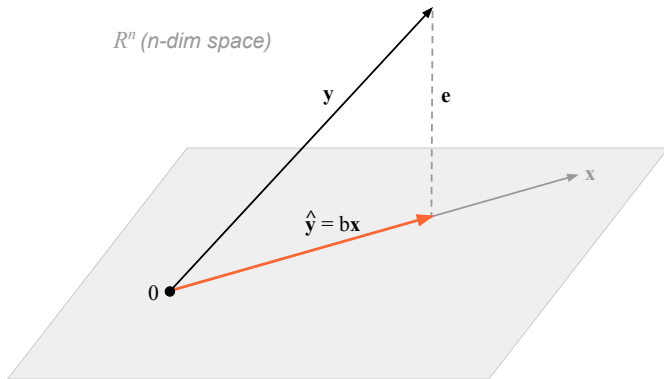
Thus, with centered variables  $\mathbf{x}$  and  $\mathbf{y}$ , the fitted values  $\hat{\mathbf{y}}$  are given by:

$$\begin{aligned}\hat{\mathbf{y}} &= \hat{\beta}_1 \mathbf{x} \\ &= \left( \frac{\mathbf{y}^\top \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \right) \\ &= \mathbf{x} \left( \frac{\mathbf{y}^\top \mathbf{x}}{\mathbf{x}^\top \mathbf{x}} \right) \\ &= \mathbf{x} (\mathbf{x}^\top \mathbf{x})^{-1} \mathbf{x}^\top \mathbf{y}\end{aligned}$$

# From variables perspective



# From variables perspective



# Example: Advertising Data

```
# number of observations  
n <- nrow(Advertising)  
  
# model matrix  
x <- Advertising$TV - mean(Advertising$TV)  
  
# reponse variable  
y <- Advertising$Sales - mean(Advertising$Sales)
```

# Example: Advertising Data

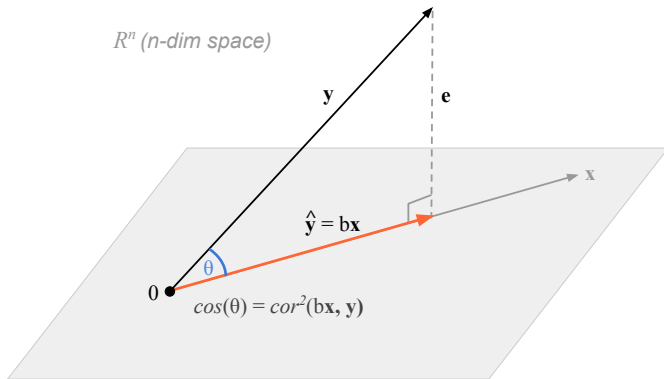
```
# slope
b1 <- sum(x * y) / sum(x * x)
b1

## [1] 0.04753664

# intercept
b0 <- mean(Advertising$Sales) - b1 * mean(Advertising$TV)
b0

## [1] 7.032594
```

# From variables perspective





# Some Remarks

- ▶ There is nothing in the Least Squares method that requires statistical inference: formal tests of null hypotheses or confidence intervals.
- ▶ In its simplest form, regression analysis can be performed without statistical inference.
- ▶ The inferential part can sometimes be very useful but goes beyond the definition of a regression analysis.

# Some Comments

- ▶ Linear Regression is a “simple” approach to supervised learning.
- ▶ Don’t get fooled by the word “simple”.
- ▶ “simple”  $\neq$  easy / boring / uninteresting.
- ▶ I will use the terms *Regression Analysis* and *Regression Model* interchangeably.