Linear Regression

Outline

- Introduction
- Simple Linear Regression
- Estimating the Coefficients
- Assessing the Accuracy
- Applying mathematical analysis to practice
- Multiple Linear regression

Introduction

- This lesson is about linear regression, a very simple approach for supervised learning. In particular, linear regression is a useful tool for predicting a quantitative response.
- It serves as a good jumping-off point for newer approaches.
- The importance of having a good understanding of linear regression before studying more complex learning methods cannot be overstated.
- This lesson reviews some of the key ideas underlying the linear regression model, as well as the least squares approach that is most commonly used to fit this model.

Simple Linear Regression

- It is a very straightforward simple linear approach for predicting a quantitative response Y on the basis of a single predictor variable X.
- It assumes that there is approximately a linear relationship between X and Y.

$$Y \approx \beta_0 + \beta_1 X.$$

- β0 and β1 are two unknown constants that represent the intercept and slope terms in the linear model. Together, β0 and β1 are intercept known as the model coefficients or parameters.

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

Estimating the Coefficients

□ Let $\hat{y}_i = \widehat{\beta_0} + \widehat{\beta_1} x_i$. Then e$_i$=$y_i - \hat{y}_i$ represents the ith residual

□ We define the **residual sum of squares**(RSS)as

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

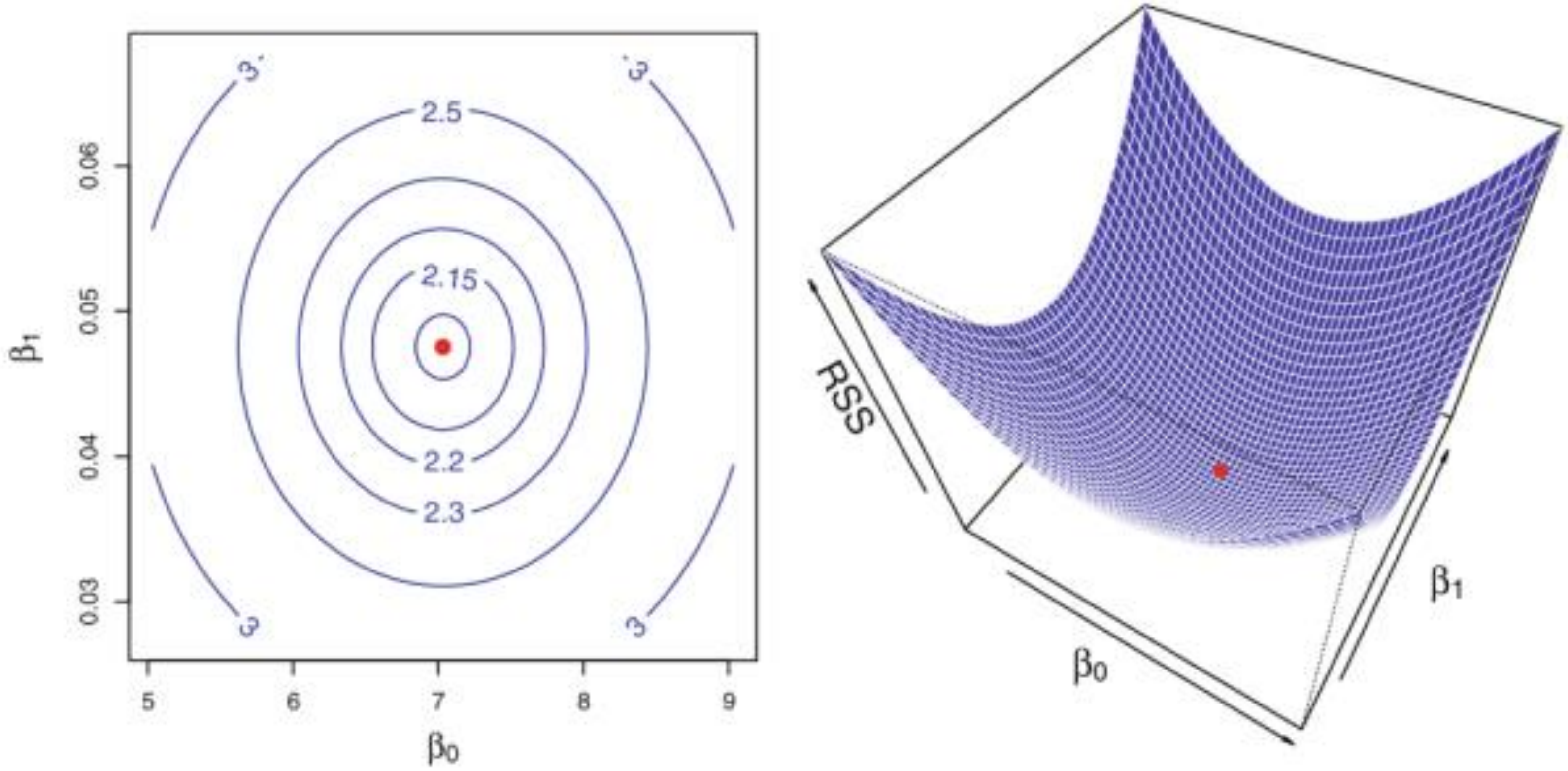
□ The least squares approach chooses □□ 0 and □□ 1 to minimize the RSS. Using some calculus, one can show that the minimizers are

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

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

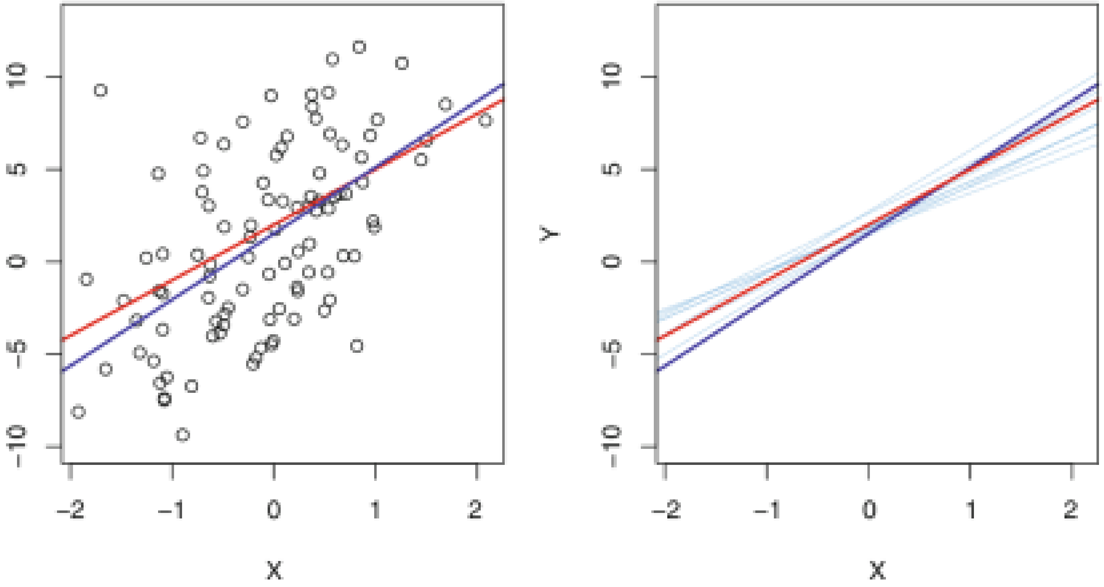
where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means.

Example β_0=7.03 and β_1=0.0475



Assessing the Accuracy of the Coefficient Estimates

- The true relationship between X and Y takes the form Y = f(X)+ □□ for some unknown function f, where □□ is a mean-zero random error term.



- The red line represents the true relationship, which is known as the population regression line

Assessing the Accuracy of the Coefficient Estimates (Cont.)

- The population mean μ of some random variable Y
- A reasonable estimate is $\hat{\mu} = \bar{y}$,
- $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean
- the standard error of μ:

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

- In a similar vein, we can wonder how close ⍰⍰ 0 and ⍰⍰ 1 are to the true values β0 and β1

$$\sigma^2 = \mathrm{Var}(\epsilon)$$

$$\mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right], \quad \mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

- For linear regression, the 95% confidence interval for β1 approximately takes the form

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

$$\hat{\beta}_0 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_0)$$

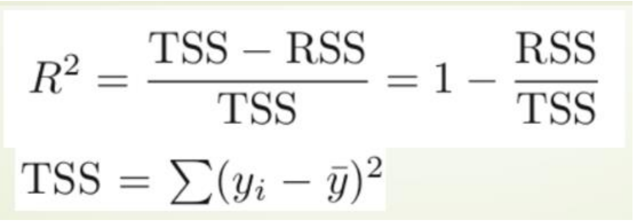
Assessing the Accuracy of the Model

□ The quality of a linear regression fit is typically assessed using two related quantities: the residual standard error (RSE) and the R^2statistic.

□ The RSE is an estimate of the standard deviation of □□. if □□ □□ ≈ □□□□ $y_i$for i =1,...,n then RSE will be small, and we can conclude that the model fits the data very well

□ R^2 measures the proportion of variability in Y that can be explained using X. An R^2statistic that is close to 1 indicates that a large proportion of the variability in the response has been explained by the regression. A number near 0 indicates that the regression did not explain much of the variability in the response.

Total sum of squares:



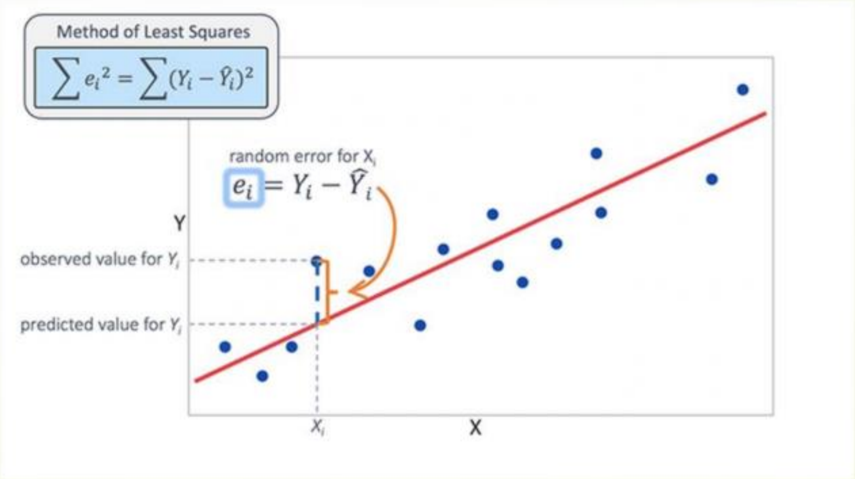
$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

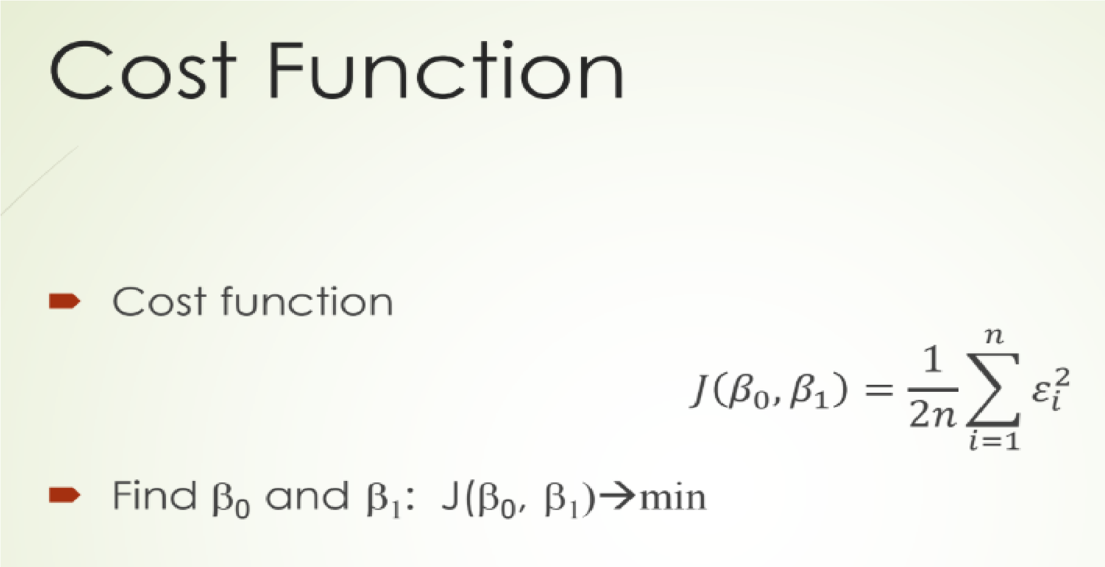
$$\text{TSS} = \sum(y_i - \bar{y})^2$$

Applying mathematical analysis to practice:

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$





Cost Function

- Cost function

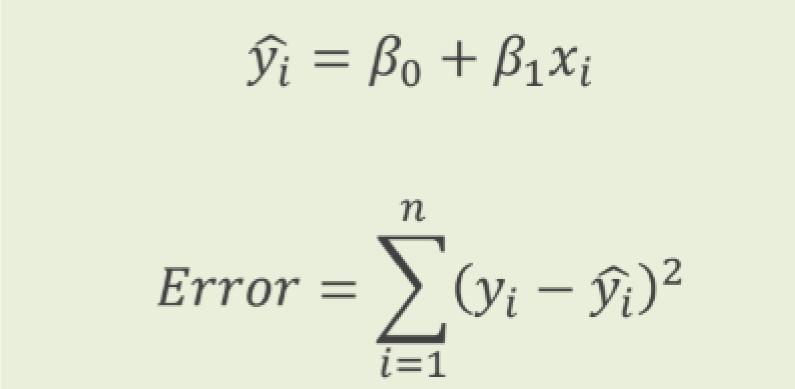
$$J(\beta_0, \beta_1) = \frac{1}{2n}\sum_{i=1}^{n} \varepsilon_i^2$$

- Find β_0 and β_1: $J(\beta_0, \beta_1) \rightarrow \min$

Problem

□ The nature of the problem: examine the cost function and determine the minimum and extract the regression coefficients.

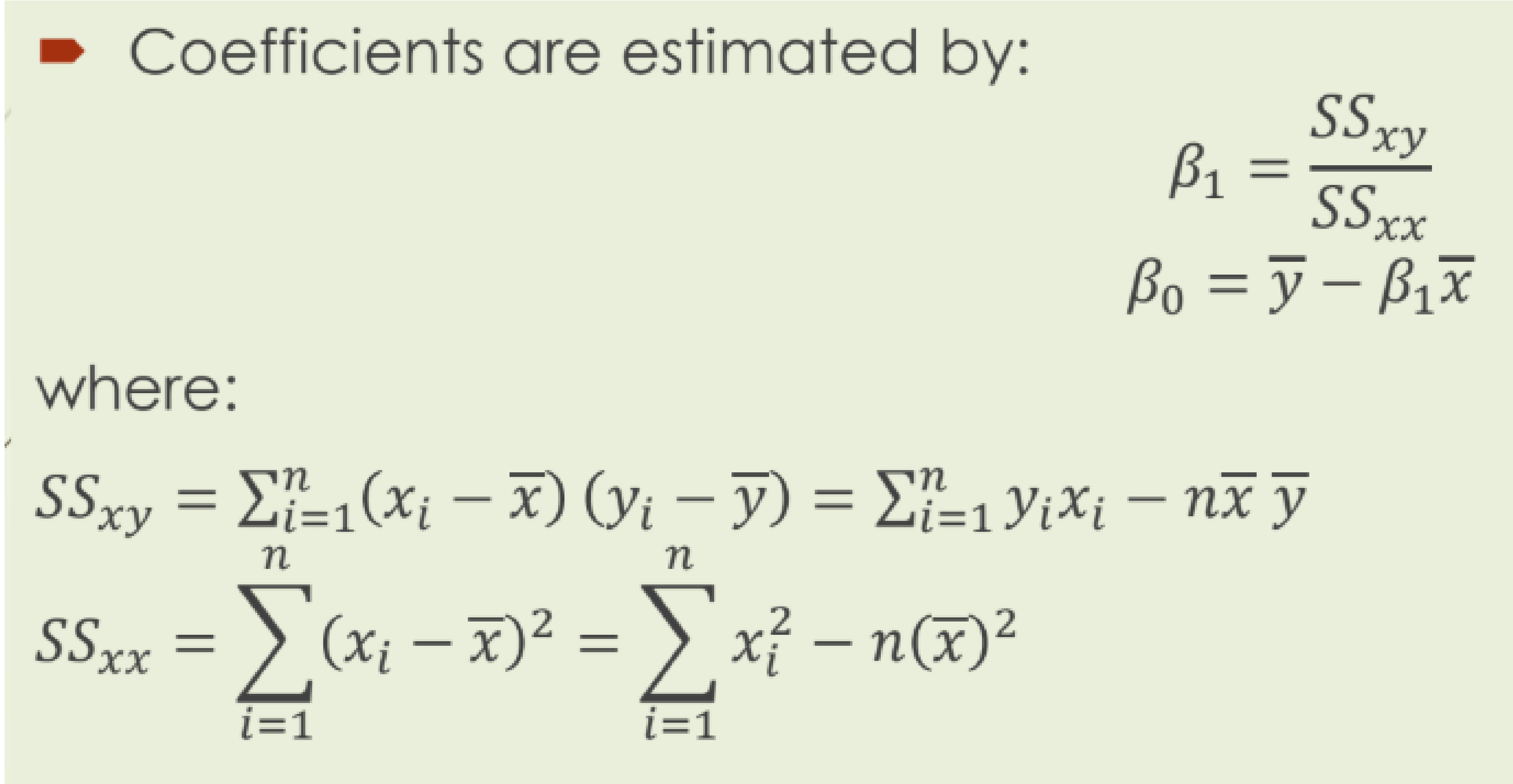
□ Details:



$$\hat{y}_i = \beta_0 + \beta_1 x_i$$

$$Error = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Solution



- Coefficients are estimated by:

$$\beta_1 = \frac{SS_{xy}}{SS_{xx}}$$

$$\beta_0 = \overline{y} - \beta_1\overline{x}$$

where:

$$SS_{xy} = \sum_{i=1}^{n}(x_i - \overline{x})\,(y_i - \overline{y}) = \sum_{i=1}^{n} y_i x_i - n\overline{x}\,\overline{y}$$

$$SS_{xx} = \sum_{i=1}^{n}(x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - n(\overline{x})^2$$

Implementation (python code)

```python
import numpy as np


def estimate_coef(x, y):
    # number of observations/points
    n = np.size(x)


    # mean of x and y vector
    m_x = np.mean(x)
    m_y = np.mean(y)


    # calculating cross-deviation and deviation about x
    SS_xy = np.sum(y*x) - n*m_y*m_x
    SS_xx = np.sum(x*x) - n*m_x*m_x


    # calculating regression coefficients
    b_1 = SS_xy / SS_xx
    b_0 = m_y - b_1*m_x


    return (b_0, b_1)
```

Multiple Linear Regression

- Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable. However, in practice we often have more than one predictor
- The multiple linear regression model takes the form

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

- One option is to run separate simple linear regressions, each of which uses a different type of X as a predictor. However, the approachof fitting a separate simple linear regression model for each predictor is not entirely satisfactory
- Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend the simple linear regression model so that it can directly accommodate multiple predictors.
- Estimating the Regression Coefficients

Estimating the Regression Coefficients

□ As was the case in the simple linear regression setting, the regression coefficients $\beta_0,\beta_1,...,\beta_p$ are unknown, and must be estimated. Given estimates □□ 0, □□ 1, ... , □□ □□, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p.$$

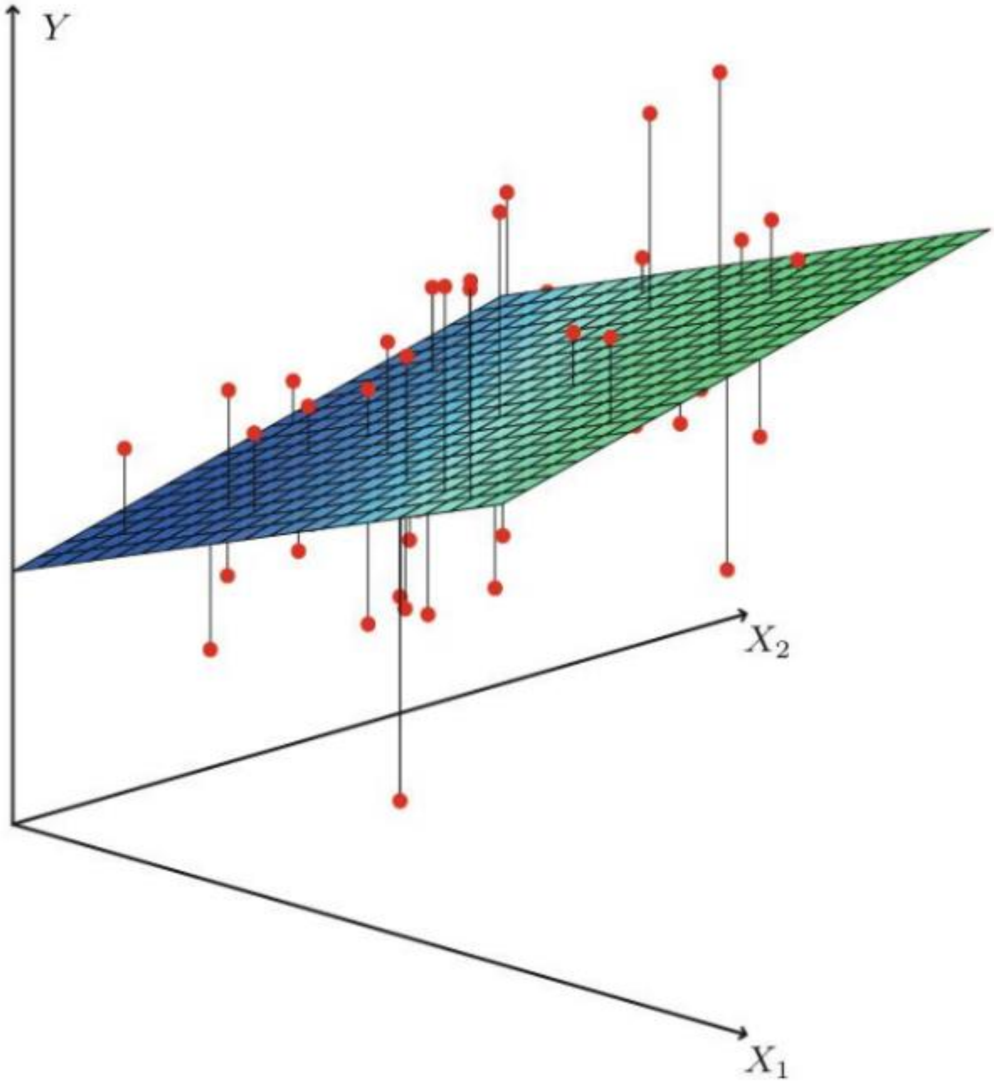
□ The parameters are estimated using the same least squares approach that we saw in the context of simple linear regression. We choose $\beta_0,\beta_1,...,\beta_p$to minimize the sum of squared residuals

$$\begin{aligned} \mathrm{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned}$$

An example of the least squares fit to a toy dataset with p=2 predictors

□ In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane.

The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red)and the plane.



Multiple Linear Regression: Applying mathematical analysis to practice

□ Multiple linear regression tries to model the relationship between two or more independent variables (features) and a response (dependent variable) by fitting a linear expression to observed data.

□ Considering a dataset with p attributes and a response.

□ Datasets have n rows/observations.

Definitions

- X(feature matrix) = matrix size nxp where x_{ij} represents the value of feature j^{th} in the observation i^{th}

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \vdots & x_{np} \end{pmatrix}$$

- y (response vector) = A vector of size n where $y_i$represents the response value of the ith observation.

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

Regression Line Equation

- The regression line for p features is represented as:

$$h(x_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip}$$

- Where h(xi) is the predicted response value for the i^{th} observation and β_0, β_1, …, β_p are model coefficients. Alternatively, one can write:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i$$

or

$$y_i = h(x_i) + \epsilon_i \rightarrow \epsilon_i = y_i - h(x_i)$$

Multiple Linear Regression Model

- The multiple linear regression model can be generalized by representing the feature matrix X as:

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \vdots & x_{np} \end{pmatrix}$$

- The multiple linear regression model can be represented in matrix form as follows:

$$y = X\beta + \epsilon$$

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \vdots \\ \beta_p \end{bmatrix} \qquad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Solution

- The task of determining β, i.e. finding ⍰⍰ using the Least Squares method. As explained, the Least Squares method tends to determine ⍰⍰ so that the total error is minimized.
- The multiple linear regression model can be estimated as:

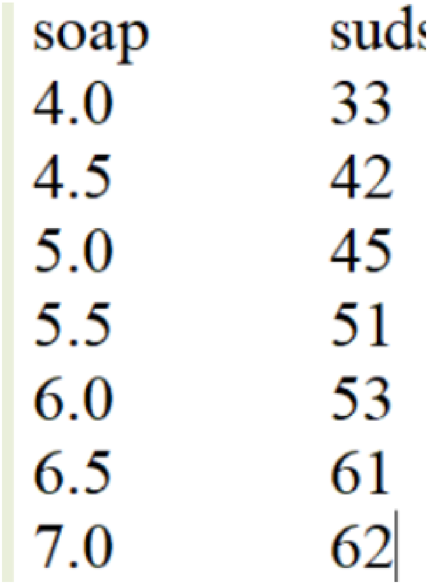
$$\hat{\beta} = (X'X)^{-1}X'y$$

- where ⍰⍰ is the estimated response vector.

$$\hat{y} = X\hat{\beta}$$

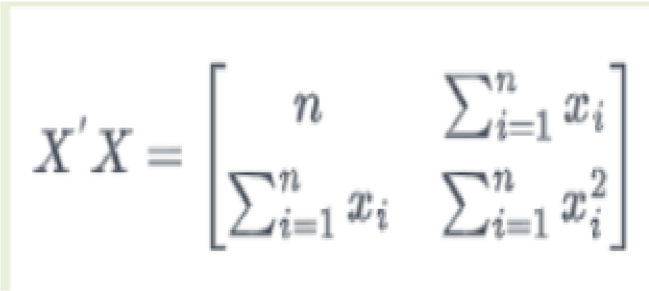
For Example

- Let's consider the data in the **Soap Suds dataset** (Draper and Smith, 1998), in which the height of suds (*y* = *suds*) in a standard dishpan was recorded for various amounts of soap (x = *soap*, in grams)



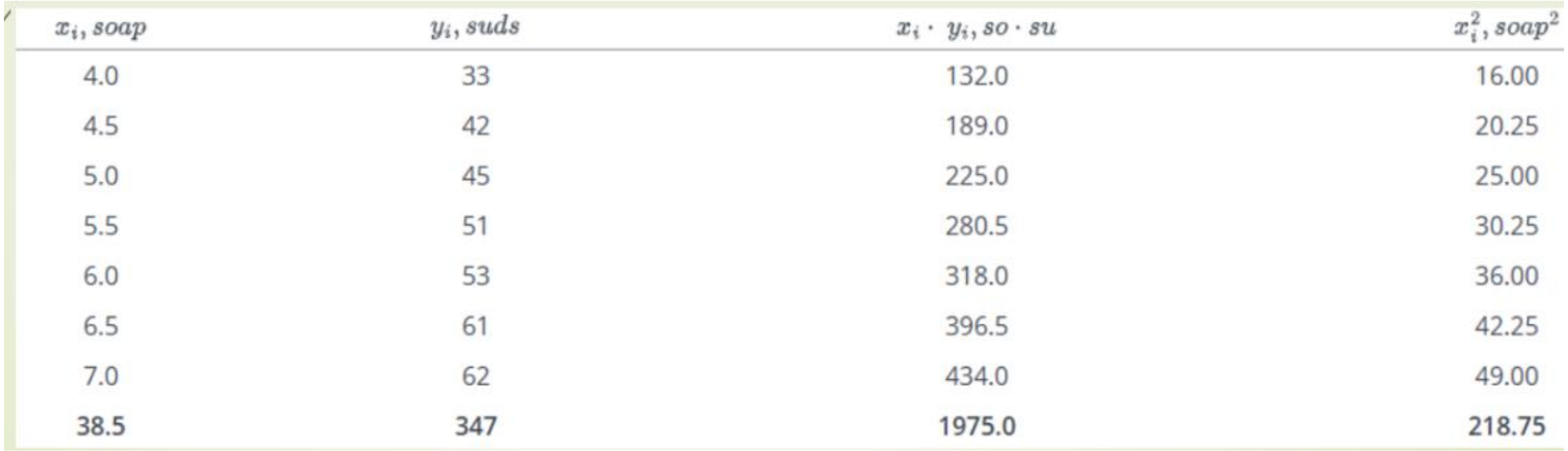
soap	suds
4.0	33
4.5	42
5.0	45
5.5	51
6.0	53
6.5	61
7.0	62

For Example (cont.)



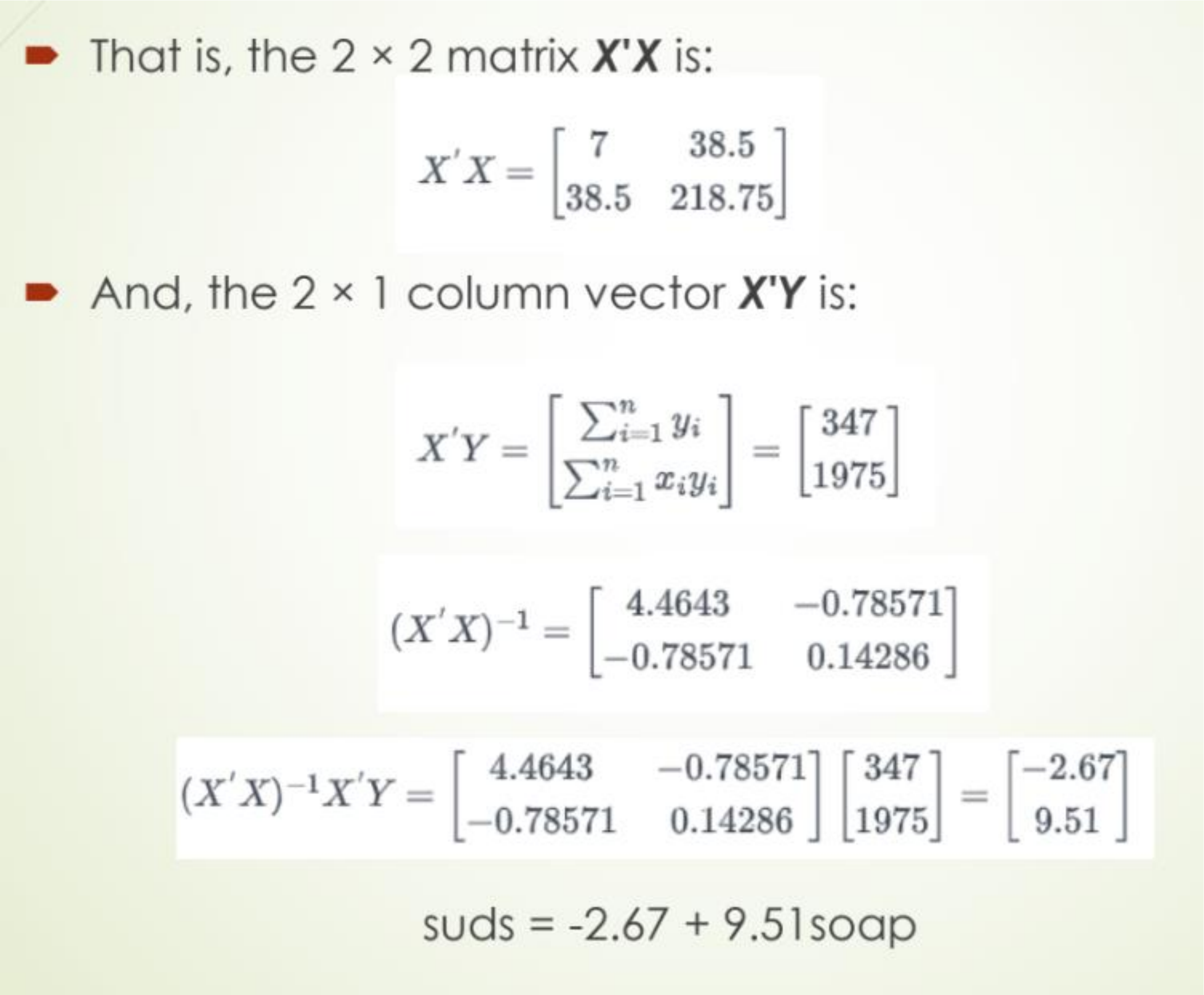
$$X'X = \begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix}$$

- we can easily calculate some parts of this formula:



$x_i, soap$	$y_i, suds$	$x_i \cdot y_i, so \cdot su$	$x_i^2, soap^2$
4.0	33	132.0	16.00
4.5	42	189.0	20.25
5.0	45	225.0	25.00
5.5	51	280.5	30.25
6.0	53	318.0	36.00
6.5	61	396.5	42.25
7.0	62	434.0	49.00
38.5	**347**	**1975.0**	**218.75**

For Example (cont.)



- That is, the 2 × 2 matrix ***X'X*** is:

$$X'X = \begin{bmatrix} 7 & 38.5 \\ 38.5 & 218.75 \end{bmatrix}$$

- And, the 2 × 1 column vector ***X'Y*** is:

$$X'Y = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix} = \begin{bmatrix} 347 \\ 1975 \end{bmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix}$$

$$(X'X)^{-1}X'Y = \begin{bmatrix} 4.4643 & -0.78571 \\ -0.78571 & 0.14286 \end{bmatrix} \begin{bmatrix} 347 \\ 1975 \end{bmatrix} = \begin{bmatrix} -2.67 \\ 9.51 \end{bmatrix}$$

suds = -2.67 + 9.51soap