# Data-analysis and Retrieval
# Linear Regression

Ad Feelders

Universiteit Utrecht

# Literature

Required Literature: Chapter 3 of ISLR by James et al. (videos by authors available through a link on the course webpage).

The slides of this lecture *complement* the book, they do not *cover* the book! We give a bit more detail on the derivation of the least squares estimates for the linear regression model.

Please ask questions about the material of part 2 in the channel *Hoorcolleges deel 2* (or during the lecture of course).

# Regression

In regression problems we want to predict a *numeric* target variable from one or more predictor variables (features).

Examples:

- Predict sale price of a house from lot size, location, has garage?, etc.
- Predict a person's income from education level, gender, age, etc.
- Predict the number of bugs in a computer program from code-complexity measures.
- Assignment: predict relevance score of product for a query from match between query text and product description.

# Linear Regression Model

The central assumption of linear regression is

$$E(Y \mid X) = f(X) = \beta_0 + \beta_1 X$$

Or, alternatively

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

with $E(\varepsilon \mid X) = 0$.

Usually, it is also assumed that $\mathrm{Var}(Y \mid X) = \sigma^2$, that is, $Y$ has the same variance for each value of $X$.

# Minimizing empirical loss

Given training data

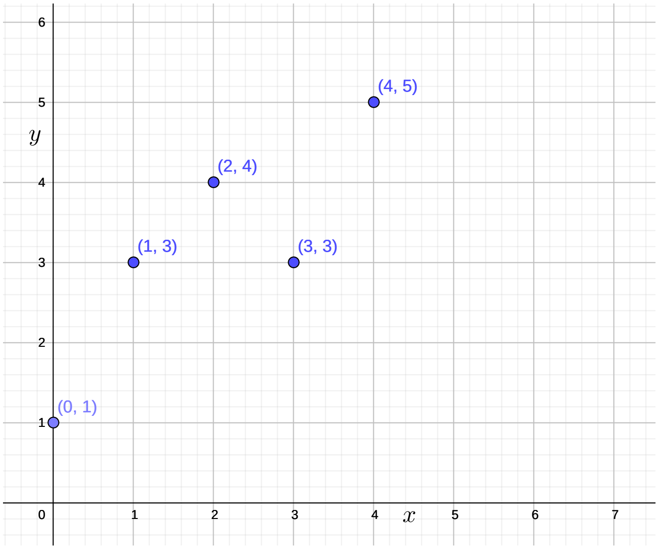
$$D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\},$$

find the values of $b_0$ and $b_1$ such that the residual sum of squares

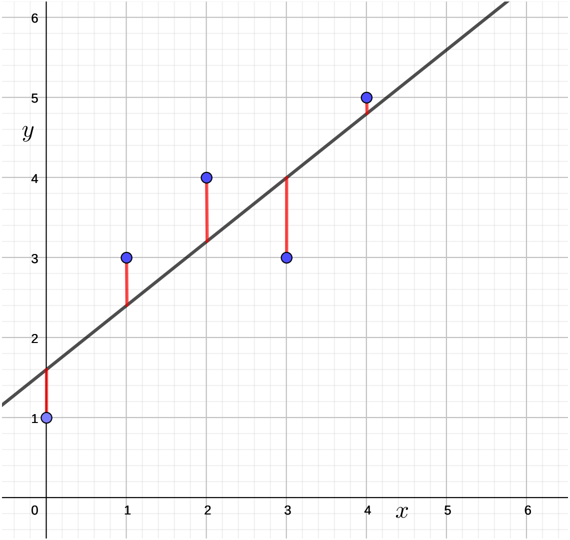
$$\mathrm{RSS}(b_0, b_1) = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - (b_0 + b_1 x_i))^2$$

is minimized, where $\hat{y}_i = b_0 + b_1 x_i$ is the predicted value for $y_i$.

# Scatterplot of Training Data



# Error of Line



# Example

| $i$ | $x$ | $y$ | $\hat{y} = b_0 + b_1 x$ | $e = y - \hat{y}$ | $e^2 = (y - \hat{y})^2$ |
|---|---|---|---|---|---|
| 1 | 0 | 1 | $b_0$ | $1 - b_0$ | $(1 - b_0)^2$ |
| 2 | 1 | 3 | $b_0 + b_1$ | $3 - b_0 - b_1$ | $(3 - b_0 - b_1)^2$ |
| 3 | 2 | 4 | $b_0 + 2b_1$ | $4 - b_0 - 2b_1$ | $(4 - b_0 - 2b_1)^2$ |
| 4 | 3 | 3 | $b_0 + 3b_1$ | $3 - b_0 - 3b_1$ | $(3 - b_0 - 3b_1)^2$ |
| 5 | 4 | 5 | $b_0 + 4b_1$ | $5 - b_0 - 4b_1$ | $(5 - b_0 - 4b_1)^2$ |

$$\begin{aligned} \text{RSS}(b_0, b_1) &= (1 - b_0)^2 + (3 - b_0 - b_1)^2 \\ &\quad + (4 - b_0 - 2b_1)^2 + (3 - b_0 - 3b_1)^2 \\ &\quad + (5 - b_0 - 4b_1)^2 \end{aligned}$$

# Minimizing RSS (single coefficient)

Suppose RSS only depends on a single coefficient $b$. From calculus we know that a necessary condition for a minimum is:

$$\frac{d\ RSS}{d\ b} = 0 \tag{1}$$

This condition is not sufficient, since maxima and points of inflection also satisfy equation (1). Together with the second-order condition:

$$\frac{d^2\ RSS}{d\ b^2} > 0, \tag{2}$$

we have a sufficient condition for a local minimum.

# Minimizing RSS (multiple coefficients)

Usually RSS depends on multiple coefficients $b_1, \ldots, b_p$.
Analogous to the single-parameter case a necessary condition for a minimum is:

$$\frac{\partial RSS}{\partial b_j} = 0, \text{ for all } j = 1, \ldots, p \tag{3}$$

Again this condition is not sufficient, since maxima and saddle points also satisfy (3).

Together with the second-order condition that the Hessian matrix (the matrix of second order partial derivatives) is positive definite, we have a sufficient condition for a local minimum.

# Example

$$\frac{\partial \text{RSS}}{\partial b_0} = [2(1-b_0)(-1)] + [2(3-b_0-b_1)(-1)]$$
$$+ [2(4-b_0-2b_1)(-1)] + [2(3-b_0-3b_1)(-1)]$$
$$+ [2(5-b_0-4b_1)(-1)]$$
$$= -32 + 10b_0 + 20b_1$$

$$\frac{\partial \text{RSS}}{\partial b_1} = 0 + [2(3-b_0-b_1)(-1)]$$
$$+ [2(4-b_0-2b_1)(-2)] + [2(3-b_0-3b_1)(-3)]$$
$$+ [2(5-b_0-4b_1)(-4)]$$
$$= -80 + 20b_0 + 60b_1$$

# Example

Setting partial derivatives to zero gives

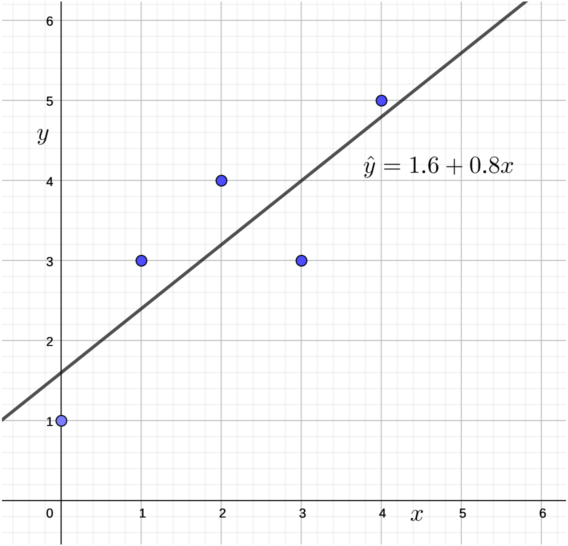
$$
\begin{aligned}
10b_0 + 20b_1 &= 32 \\
20b_0 + 60b_1 &= 80
\end{aligned}
$$

which gives $b_0 = 1.6$ and $b_1 = 0.8$.

So the least squares fitted line is

$$\hat{y} = 1.6 + 0.8x$$

# Fitted Line



# General Solution

We want to minimize

$$\text{RSS}(b_0, b_1) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_i)^2$$

Consider an arbitrary term from this sum:

$$\text{RSS}_i = (y_i - b_0 - b_1 x_i)^2 = e_i^2,$$

where $e_i = y_i - b_0 - b_1 x_i$. Using the chain rule, we have

$$\begin{aligned} \frac{\partial \text{RSS}_i}{\partial b_0} &= \frac{\partial e_i^2}{\partial e_i} \frac{\partial e_i}{\partial b_0} \\ &= (2e_i)(-1) = -2(y_i - b_0 - b_1 x_i) \end{aligned}$$

# General Solution

Partial derivative with respect to intercept:

$$\begin{aligned}\frac{\partial \text{RSS}}{\partial b_0} &= \frac{\partial}{\partial b_0}\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)^2 \\ &= \sum_{i=1}^{n}\frac{\partial}{\partial b_0}(y_i - b_0 - b_1 x_i)^2 \\ &= -2\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i)\end{aligned}$$

Equate to zero

$$\sum_{i=1}^{n}(y_i - b_0 - b_1 x_i) = \sum_{i=1}^{n} e_i = 0$$

In the optimal solution, the sum of the errors is zero.

# General Solution

Partial derivative with respect to slope:

$$\frac{\partial \text{RSS}}{\partial b_1} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-x_i)$$

$$= -2\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i)$$

Equate to zero

$$\sum_{i=1}^{n} x_i(y_i - b_0 - b_1 x_i) = \sum_{i=1}^{n} x_i e_i = 0$$

# Normal Equations

Expand and collect terms:

$$\sum_{i=1}^{n} y_i = nb_0 + b_1 \sum_{i=1}^{n} x_i \qquad (4)$$

$$\sum_{i=1}^{n} x_i y_i = b_0 \sum_{i=1}^{n} x_i + b_1 \sum_{i=1}^{n} x_i^2 \qquad (5)$$

To solve for $b_0$ divide (4) by $n$:

$$b_0 = \bar{y} - b_1 \bar{x},$$

where $\bar{y} = \frac{1}{n} \sum y_i$. Note that $\bar{y} = b_0 + b_1 \bar{x}$, so the line goes through the "point of means" $(\bar{x}, \bar{y})$.

# Normal Equations

To solve for $b_1$, multiply (4) by $\sum x_i$ and (5) by $n$

$$\sum x_i \sum y_i = nb_0 \sum x_i + b_1 \left(\sum x_i\right)^2 \tag{6}$$

$$n \sum x_i y_i = nb_0 \sum x_i + nb_1 \sum x_i^2 \tag{7}$$

Subtract (6) from (7) and solve for $b_1$:

$$b_1 = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - \left(\sum x_i\right)^2}$$

# Linear regression through the origin

Suppose we know that the population regression line goes through the origin, i.e.

$$E(Y \mid X) = \beta X$$

Find the value of $b$ such that the sum of squared errors

$$\text{RSS}(b) = \sum_{i=1}^{n} (y_i - b x_i)^2$$

is minimized.

# Linear regression through the origin: calculus solution

Take the derivative

$$\frac{d\text{RSS}}{db} = -2\sum(y_i - bx_i)x_i$$

and equate to zero

$$\sum x_i y_i - b\sum x_i^2 = 0$$

so we get

$$b = \frac{\sum x_i y_i}{\sum x_i^2}$$

# Regression through the origin: Geometrical Solution

Regression through the origin: $\hat{y}_i = bx_i$

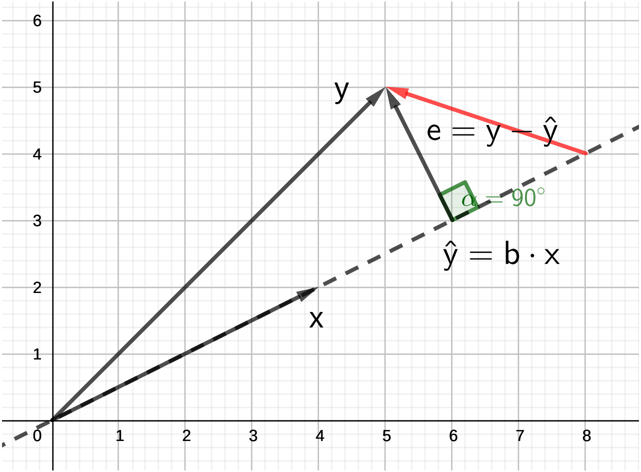
$D = \{(x_1, y_1), (x_2, y_2)\} = \{(4, 5), (2, 5)\}$ contains only two observations.

$$X = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 4 \\ 2 \end{bmatrix} \text{ and } Y = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$$

$$e = \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \text{ and } \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \end{bmatrix}$$

$$\hat{Y} = bX \text{ and } e = Y - \hat{Y} = Y - bX$$

# Least Squares Solution ($n$ dimensional space!)



# Least Squares Solution

The length of $e = \sqrt{e \cdot e} = \sqrt{e_1^2 + e_2^2} = \sqrt{\text{RSS}}$.

So to minimize RSS, $e$ must be perpendicular to $X$, i.e. $X \cdot e = 0$.

$$X \cdot e = X \cdot (Y - bX) = X \cdot Y - bX \cdot X = 0$$

Therefore

$$b = \frac{X \cdot Y}{X \cdot X} = \frac{\sum x_i y_i}{\sum x_i^2}$$

Of course we obtained the same solution as with calculus.

Matrix notation

$$b = \frac{X^T Y}{X^T X} \qquad \text{or} \qquad b = (X^T X)^{-1} X^T Y$$

# Solution

Solution of the numerical example

$$X^T Y = [4 \;\; 2] \begin{bmatrix} 5 \\ 5 \end{bmatrix} = 30$$

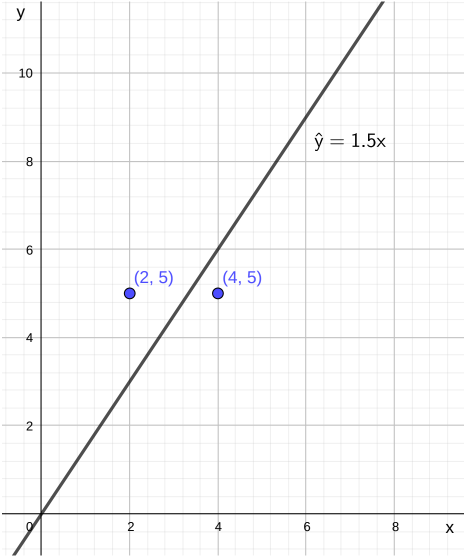
and

$$X^T X = [4 \;\; 2] \begin{bmatrix} 4 \\ 2 \end{bmatrix} = 20$$

which yields

$$b = \frac{X^T Y}{X^T X} = \frac{30}{20} = 1.5$$

# Fitted line



# Simple linear regression in matrix terms

We can write the observed $y$ values as

$$y_i = b_0 + b_1 x_i + e_i, \qquad\qquad i = 1, \ldots, n$$

which is short for

$$\begin{aligned} y_1 &= b_0 + b_1 x_1 + e_1 \\ y_2 &= b_0 + b_1 x_2 + e_2 \\ &\vdots \\ y_n &= b_0 + b_1 x_n + e_n \end{aligned}$$

# Matrix Notation

We can write this more compactly using matrix notation.
Define:

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} b = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

Then we can simply write

$$Y = Xb + e$$

# Check

$$Y = Xb + e$$

$$\begin{aligned}
\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \\ 1 & x_n \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \\
&= \begin{bmatrix} b_0 + b_1 x_1 \\ b_0 + b_1 x_2 \\ \vdots \\ b_0 + b_1 x_n \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} b_0 + b_1 x_1 + e_1 \\ b_0 + b_1 x_2 + e_2 \\ \vdots \\ b_0 + b_1 x_n + e_n \end{bmatrix}
\end{aligned}$$

# Least Squares Solution

$\hat{Y}$ is a linear combination of the columns of $X$:

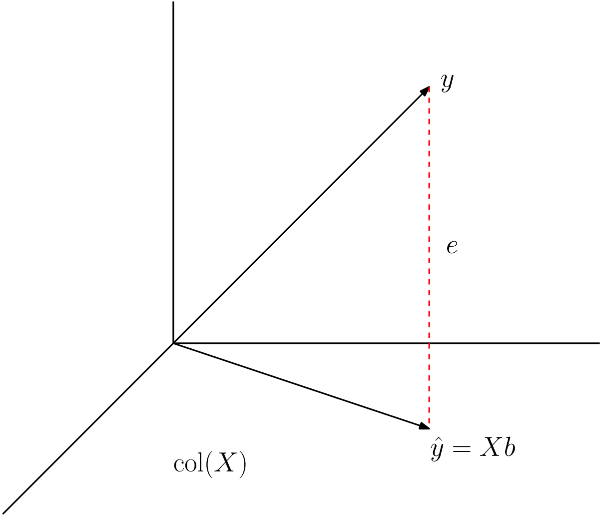
$$\hat{Y} = Xb$$

Typically, $Y$ is not in the column space of $X$. Find the value of $\hat{Y}$ that is closest to $Y$. For this to be the case, the error vector

$$e = Y - Xb$$

must be orthogonal to *all columns* of $X$.

# Least Squares Solution (Cartoon!)



# Least Squares Solution

In other words,

$$X^T e = X^T(Y - Xb) = X^T Y - X^T Xb = \begin{bmatrix} 0 \\ 0 \end{bmatrix},$$

from which it follows that

$$X^T Xb = X^T Y$$

Premultiply both sides by the inverse of $X^T X$:

$$(X^T X)^{-1} X^T Xb = (X^T X)^{-1} X^T Y$$

We then find, since $(X^T X)^{-1} X^T X = I$ and $Ib = b$:

$$\boxed{b = (X^T X)^{-1} X^T Y}$$

# Numeric example

$$D = \{(0,1),(1,1),(2,2),(3,2)\}$$

$$X = \begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \end{bmatrix} \quad Y = \begin{bmatrix} 1 \\ 1 \\ 2 \\ 2 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 4 & 6 \\ 6 & 14 \end{bmatrix} \quad X^T Y = \begin{bmatrix} 6 \\ 11 \end{bmatrix}$$

# Numeric Example

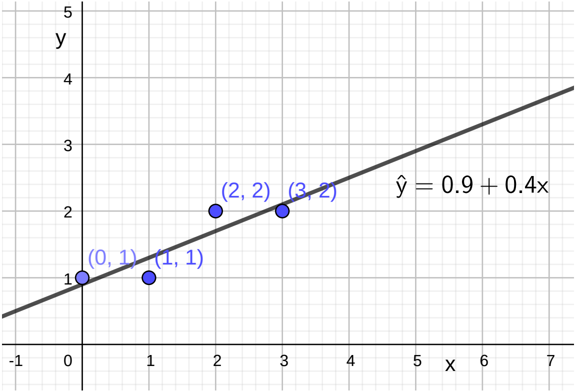
Now, since

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

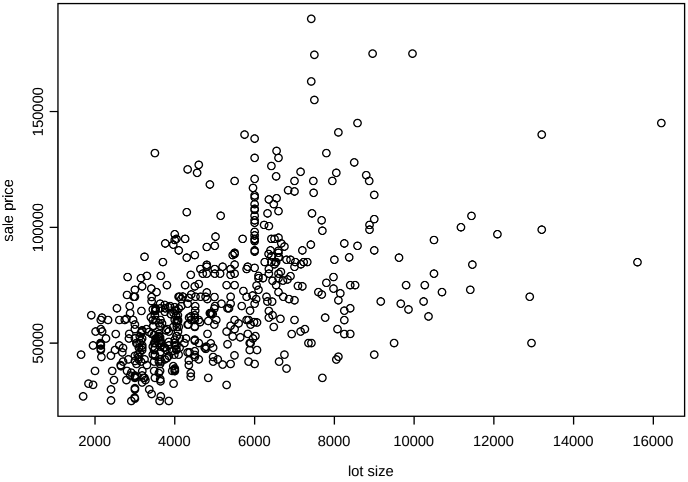
we get

$$\begin{aligned} b = (X^T X)^{-1} X^T Y &= \frac{1}{20} \begin{bmatrix} 14 & -6 \\ -6 & 4 \end{bmatrix} \begin{bmatrix} 6 \\ 11 \end{bmatrix} \\ &= \frac{1}{20} \begin{bmatrix} 18 \\ 8 \end{bmatrix} = \begin{bmatrix} 9/10 \\ 4/10 \end{bmatrix} \end{aligned}$$

# Fitted Line



# Scatterplot of lot size and sale price



# Least Squares fitted line

Using `R` we find:

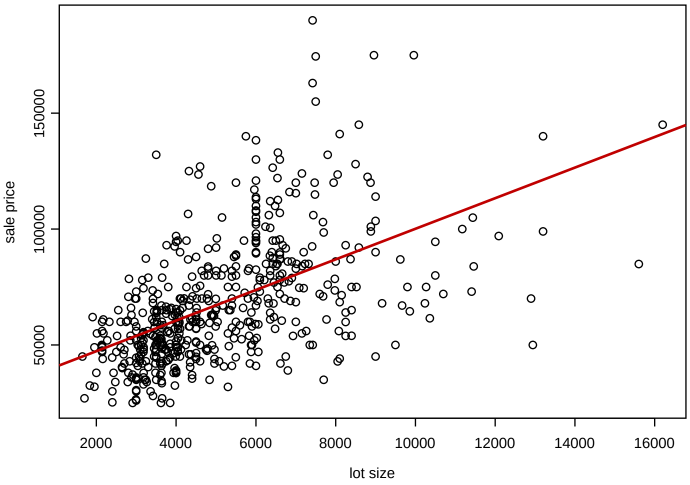
$$\text{sale price} = 34,136 + 6.6 \times \text{lot size}$$

$R^2 = 0.2871$

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2}$$

There is still room for improvement!

# Least Squares fitted line



# Multiple Linear Regression

Usually, you want to use more that one input variable to predict $Y$.

The basic assumption is

$$E(Y|\mathbf{X}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_{p-1} X_{p-1}$$

# Multiple Linear Regression

We can write the observed $y$ values as

$$y_i = b_0 + b_1 x_{i,1} + b_2 x_{i,2} + \ldots + b_{p-1} x_{i,p-1} + e_i$$

which is short for

$$\begin{aligned}
y_1 &= b_0 + b_1 x_{1,1} + b_2 x_{1,2} + \ldots + b_{p-1} x_{1,p-1} + e_1 \\
y_2 &= b_0 + b_1 x_{2,1} + b_2 x_{2,2} + \ldots + b_{p-1} x_{2,p-1} + e_2 \\
&\vdots \\
y_n &= b_0 + b_1 x_{n,1} + b_2 x_{n,2} + \ldots + b_{p-1} x_{n,p-1} + e_n
\end{aligned}$$

# Notation and Least Squares Solution

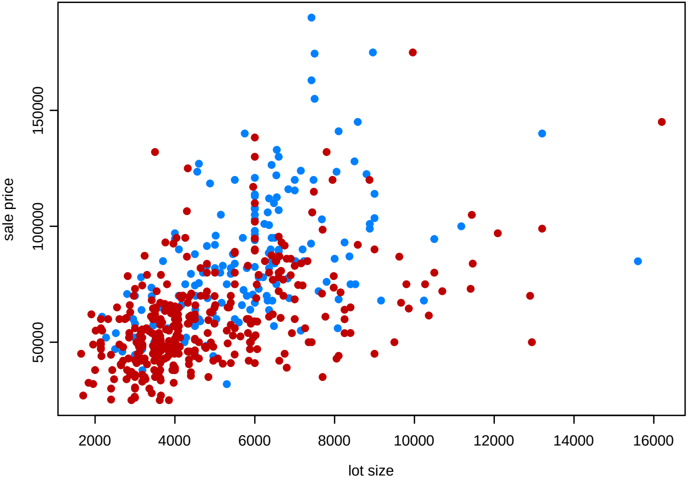
$$X = \begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \dots & x_{1,p-1} \\ 1 & x_{2,1} & x_{2,2} & \dots & x_{2,p-1} \\ \vdots & & & & \\ 1 & x_{n,1} & x_{n,2} & \dots & x_{n,p-1} \end{bmatrix}$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

Then we can write

$$Y = Xb + e, \quad b = (X^T X)^{-1} X^T Y$$

# Scatterplot of lot size, airco, and sale price



# Fitted Equation

The fitted regression line is:

$$\text{sale.price} = 32,693 + 5.6 \times \text{lot.size} + 20,175 \times \text{air.cond}$$

If air.cond=0:

$$\text{sale.price} = 32,693 + 5.6 \times \text{lot.size}$$

If air.cond=1:

$$\text{sale.price} = (32,693 + 20,175) + 5.6 \times \text{lot.size}$$

$R^2 = 0.4048$

The premium for air conditioning is 20,175 Canadian dollars.

# Fitted Equation