Least Squares Estimation[1]

STA 302 Fall 2020

[1] See last slide for copyright information.

Overview

1. The Model
2. Scalar Least Squares
3. Matrix Version
4. More Notation
5. R^2
6. Estimating σ^2
7. Curve Fitting

Reading in In Rencher and Schaalje's *Linear Models In Statistics*

- Glance at Ch. 6 first.
- Sections 7.1, 7.2, 7.3.1 (pp. 137-145).
- Section 7.3.3 (pp. 149-151) on estimation of σ^2.
- Section 7.7 on R^2, but they use material in a section we will cover later.

Multiple regression in scalar form

For $i = 1, \ldots, n$, let $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$, where

- x_{ij} are observed, known constants.
- $\epsilon_1, \ldots, \epsilon_n$ are independent random variables with $E(\epsilon_i) = 0$ and $Var(\epsilon_i) = \sigma^2$.
- $\beta_0, \ldots \beta_k$ and σ^2 are unknown constants, with $\sigma^2 > 0$.

For example

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$

For customer $i = 1, \ldots, n$,

- y_i is purchases in $.
- x_{i1} is income.
- x_{i2} is age.
- x_{i3} is advertising recall.

Multiple regression in matrix form

Compare $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$

$$\begin{array}{ccccc} \mathbf{y} & = & \mathbf{X} & \boldsymbol{\beta} & + & \boldsymbol{\epsilon} \end{array}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 14.2 & \cdots & 1 \\ 1 & 11.9 & \cdots & 0 \\ 1 & 3.7 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 6.2 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

where

$\mathbf{X}$ is an $n \times (k+1)$ matrix of observed constants

$\boldsymbol{\beta}$ is a $(k+1) \times 1$ matrix of unknown constants

$E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $cov(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}_n$, where σ^2 is an unknown constant.

Least Squares Estimation: The idea

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$

- Statistical model says the distribution of y_i is determined by parameter θ (could be a vector).
- Calculate $E(y_i)$.
- Expected value depends on θ, so write $E_\theta(y_i)$.
- How should we estimate θ?
- Choose the value of θ that gets the observed y_i as close as possible to their expected values ,
- By minimizing

$$Q(\theta) = \sum_{i=1}^{n} (y_i - E_\theta(y_i))^2$$

over θ.

Least Squares

$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \qquad E_{\boldsymbol{\beta}}(y_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik}$

Estimate β_j by minimizing

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - E_{\boldsymbol{\beta}}(y_i))^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2$$

$$\begin{aligned} \frac{\partial Q}{\partial \beta_0} &\stackrel{\text{set}}{=} 0 \\ \frac{\partial Q}{\partial \beta_1} &\stackrel{\text{set}}{=} 0 \\ &\vdots \\ \frac{\partial Q}{\partial \beta_k} &\stackrel{\text{set}}{=} 0 \end{aligned}$$

Solve $k+1$ equations in $k+1$ unknowns.

Differentiate with respect to β_0

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_0} &= \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_0} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \\
&= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})(-1) \\
&= -2 \left(\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_{i1} - \cdots - \beta_k \sum_{i=1}^{n} x_{ik} \right) \stackrel{\text{set}}{=} 0
\end{aligned}
$$

Differentiate with respect to β_1

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_1} &= \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \\
&= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})(-x_{i1}) \\
&= -2 \sum_{i=1}^{n} (x_{i1} y_i - \beta_0 x_{i1} - \beta_1 x_{i1}^2 - \cdots - \beta_k x_{i1} x_{ik}) \\
&= -2 \left(\sum_{i=1}^{n} x_{i1} y_i - \beta_0 \sum_{i=1}^{n} x_{i1} - \beta_1 \sum_{i=1}^{n} x_{i1}^2 - \cdots - \beta_k \sum_{i=1}^{n} x_{i1} x_{ik} \right) \\
&\stackrel{\text{set}}{=} 0
\end{aligned}
$$

Differentiate with respect to β_j

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_j} &= \frac{\partial}{\partial \beta_j} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_j x_{ij} - \cdots - \beta_k x_{ik})^2 \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_j} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_j x_{ij} - \cdots - \beta_k x_{ik})^2 \\
&= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_j x_{ij} - \cdots - \beta_k x_{ik})(-x_{ij}) \\
&= -2 \sum_{i=1}^{n} (x_{ij} y_i - \beta_0 x_{ij} - \beta_1 x_{i1} x_{ij} - \cdots - \beta_j x_{ij}^2 - \cdots - \beta_k x_{ij} x_{ik}) \\
&= -2 \left(\sum_{i=1}^{n} x_{ij} y_i - \beta_0 \sum_{i=1}^{n} x_{ij} - \beta_1 \sum_{i=1}^{n} x_{i1} x_{ij} - \cdots - \sum_{i=1}^{n} \beta_j x_{ij}^2 - \cdots - \beta_k \sum_{i=1}^{n} x_{ij} x_{ik} \right) \\
&\overset{\text{set}}{=} 0
\end{aligned}
$$

Differentiate with respect to β_k

$$
\begin{aligned}
\frac{\partial Q}{\partial \beta_k} &= \frac{\partial}{\partial \beta_k} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \\
&= \sum_{i=1}^{n} \frac{\partial}{\partial \beta_1} (y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \\
&= \sum_{i=1}^{n} 2(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})(-x_{ik}) \\
&= -2 \sum_{i=1}^{n} (x_{ik} y_i - \beta_0 x_{ik} - \beta_1 x_{i1} x_{ik} - \cdots - \beta_k x_{ik}^2) \\
&= -2 \left(\sum_{i=1}^{n} x_{ik} y_i - \beta_0 \sum_{i=1}^{n} x_{ik} - \beta_1 \sum_{i=1}^{n} x_{i1} x_{ik} - \cdots - \beta_k \sum_{i=1}^{n} x_{ik}^2 \right) \\
&\overset{\text{set}}{=} 0
\end{aligned}
$$

Have $k+1$ equations in $k+1$ unknowns

Solve for $\beta_0, \ldots, \beta_k$

$$
\begin{aligned}
-2\left(\sum_{i=1}^{n} y_i - n\beta_0 - \beta_1 \sum_{i=1}^{n} x_{i1} - \cdots - \beta_k \sum_{i=1}^{n} x_{ik}\right) &= 0 \\
-2\left(\sum_{i=1}^{n} x_{i1}y_i - \beta_0 \sum_{i=1}^{n} x_{i1} - \beta_1 \sum_{i=1}^{n} x_{i1}^2 - \cdots - \beta_k \sum_{i=1}^{n} x_{i1}x_{ik}\right) &= 0 \\
&\vdots \\
-2\left(\sum_{i=1}^{n} x_{ik}y_i - \beta_0 \sum_{i=1}^{n} x_{ik} - \beta_1 \sum_{i=1}^{n} x_{i1}x_{ik} - \cdots - \beta_k \sum_{i=1}^{n} x_{ik}^2\right) &= 0
\end{aligned}
$$

Divide by -2 and re-arrange, obtaining

$$
\begin{array}{lcllllcl}
\beta_0 n & + & \beta_1 \sum_{i=1}^n x_{i1} & + & \cdots & + & \beta_k \sum_{i=1}^n x_{ik} & = & \sum_{i=1}^n y_i \\
\beta_0 \sum_{i=1}^n x_{i1} & + & \beta_1 \sum_{i=1}^n x_{i1}^2 & + & \cdots & + & \beta_k \sum_{i=1}^n x_{i1}x_{ik} & = & \sum_{i=1}^n x_{i1}y_i \\
\beta_0 \sum_{i=1}^n x_{i2} & + & \beta_1 \sum_{i=1}^n x_{i1}x_{i2} & + & \cdots & + & \beta_k \sum_{i=1}^n x_{i2}x_{ik} & = & \sum_{i=1}^n x_{i2}y_i \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\beta_0 \sum_{i=1}^n x_{ik} & + & \beta_1 \sum_{i=1}^n x_{i1}x_{ik} & + & \cdots & + & \beta_k \sum_{i=1}^n x_{ik}^2 & = & \sum_{i=1}^n x_{ik}y_i
\end{array}
$$

- These are called the *normal equations.*
- It has nothing to do with the normal distribution.
- Wikipedia says" "In geometry, a normal is an object such as a line, ray, or vector that is perpendicular to a given object."
- The normal equations are a system of $k+1$ *linear* equations in $k+1$ unknowns. All the x_{ij} and y_i are constants.

Solve the Normal Equations

The normal equations are

$$\begin{array}{ccccccccc}
\beta_0 n & + & \beta_1 \sum_{i=1}^n x_{i1} & + & \cdots & + & \beta_k \sum_{i=1}^n x_{ik} & = & \sum_{i=1}^n y_i \\
\beta_0 \sum_{i=1}^n x_{i1} & + & \beta_1 \sum_{i=1}^n x_{i1}^2 & + & \cdots & + & \beta_k \sum_{i=1}^n x_{i1}x_{ik} & = & \sum_{i=1}^n x_{i1}y_i \\
\beta_0 \sum_{i=1}^n x_{i2} & + & \beta_1 \sum_{i=1}^n x_{i1}x_{i2} & + & \cdots & + & \beta_k \sum_{i=1}^n x_{i2}x_{ik} & = & \sum_{i=1}^n x_{i2}y_i \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
\beta_0 \sum_{i=1}^n x_{ik} & + & \beta_1 \sum_{i=1}^n x_{i1}x_{ik} & + & \cdots & + & \beta_k \sum_{i=1}^n x_{ik}^2 & = & \sum_{i=1}^n x_{ik}y_i
\end{array}$$

In matrix form,

$$\begin{pmatrix}
n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\
\sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\
\sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
\sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2
\end{pmatrix}
\begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix}
=
\begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{pmatrix}$$

$$\mathbf{X}'\mathbf{X} \qquad \boldsymbol{\beta} \quad = \quad \mathbf{X}'\mathbf{y}$$

Multiple regression in matrix form

Compare $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$

$$
\begin{array}{ccccccc}
\mathbf{Y} & = & \mathbf{X} & \boldsymbol{\beta} & + & \boldsymbol{\epsilon} \\
\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} & = & \begin{pmatrix} 1 & 14.2 & \cdots & 1 \\ 1 & 11.9 & \cdots & 0 \\ 1 & 3.7 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 6.2 & \cdots & 1 \end{pmatrix} & \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} & + & \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}
\end{array}
$$

The $\mathbf{X}$ Matrix

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$\mathbf{X}'\mathbf{X} =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$\mathbf{X}'\mathbf{X} =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$$\mathbf{X}'\mathbf{X} =$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$$\mathbf{X}'\mathbf{X} =$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$$\mathbf{X}'\mathbf{X} =$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$\mathbf{X}'\mathbf{X} =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$\mathbf{X}'\mathbf{X} =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$\mathbf{X}'\mathbf{X} =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$\mathbf{X}'\mathbf{X} =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$$\mathbf{X}'\mathbf{X} =$$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$\mathbf{X}'\mathbf{X} =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$\mathbf{X}'\mathbf{X} =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{X}$ Matrix

$\mathbf{X}'\mathbf{X} =$

$$\begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1k} \\ 1 & x_{21} & x_{22} & \cdots & x_{2k} \\ 1 & x_{31} & x_{32} & \cdots & x_{3k} \\ 1 & x_{41} & x_{42} & \cdots & x_{4k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{nk} \end{pmatrix}$$

$$= \begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix}$$

The $\mathbf{X}'\mathbf{y}$ Matrix

$$\mathbf{X}'\mathbf{y} = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1} y_i \\ \sum_{i=1}^n x_{i2} y_i \\ \vdots \\ \sum_{i=1}^n x_{ik} y_i \end{pmatrix}$$

The Normal Equations in Matrix Form

$$\begin{pmatrix} n & \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i2} & \cdots & \sum_{i=1}^n x_{ik} \\ \sum_{i=1}^n x_{i1} & \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} & \cdots & \sum_{i=1}^n x_{i1}x_{ik} \\ \sum_{i=1}^n x_{i2} & \sum_{i=1}^n x_{i1}x_{i2} & \sum_{i=1}^n x_{i2}^2 & \cdots & \sum_{i=1}^n x_{i2}x_{ik} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \sum_{i=1}^n x_{ik} & \sum_{i=1}^n x_{i1}x_{ik} & \sum_{i=1}^n x_{i2}x_{ik} & \cdots & \sum_{i=1}^n x_{ik}^2 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^n y_i \\ \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \\ \vdots \\ \sum_{i=1}^n x_{ik}y_i \end{pmatrix}$$

$$\mathbf{X}'\mathbf{X} \qquad \boldsymbol{\beta} \quad = \quad \mathbf{X}'\mathbf{y}$$

Solve the Normal Equations

$$
\begin{aligned}
\mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= \mathbf{X}'\mathbf{y} \\
\Rightarrow\ (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
\Rightarrow\ \boldsymbol{\beta} &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}
\end{aligned}
$$

Provided $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

What is $\boldsymbol{\beta}$??

- We have arrived at $\boldsymbol{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$, provided $(\mathbf{X}'\mathbf{X})^{-1}$ exists.
- But $\boldsymbol{\beta}$ is an unknown parameter, while $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ is a statistic that can be calculated exactly from the sample data. What's going on?
- Almost always, $\boldsymbol{\beta}$ is a vector of unknown parameters in the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.
- But just temporarily, for least squares estimation, $\boldsymbol{\beta}$ is a vector of variables over which we are minimizing the sum of squares Q.
- The solution is an *estimate*, so we write $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$.
- Provided $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

When Does $(\mathbf{X}'\mathbf{X})^{-1}$ Exist?

Theorem The following 3 conditions are equivalent:

(a) The columns of $\mathbf{X}$ are linearly independent.

(b) $\mathbf{X}'\mathbf{X}$ is positive definite.

(c) $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

Proof of equivalence

(a) The columns of $\mathbf{X}$ are linearly independent.

(b) $\mathbf{X}'\mathbf{X}$ is positive definite.

(c) $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

- Assume the columns of $\mathbf{X}$ are linearly independent.
- Columns of $\mathbf{X}$ linearly independent means $\mathbf{X}\mathbf{v} = \mathbf{0}$ implies $\mathbf{v} = \mathbf{0}$.
- Seek to show $\mathbf{X}'\mathbf{X}$ positive definite, meaning $\mathbf{v}'(\mathbf{X}'\mathbf{X})\mathbf{v} > 0$ for $\mathbf{v} \neq \mathbf{0}$.
- First, $\mathbf{X}'\mathbf{X}$ is non-negative definite, because $\mathbf{v}'(\mathbf{X}'\mathbf{X})\mathbf{v} = (\mathbf{X}\mathbf{v})'(\mathbf{X}\mathbf{v}) = \mathbf{z}'\mathbf{z} = \sum_{i=1}^{n} z_i^2 \geq 0$.
- And if $\mathbf{v}'(\mathbf{X}'\mathbf{X})\mathbf{v} = 0$, then $\mathbf{X}\mathbf{v} = \mathbf{z} = \mathbf{0} \Rightarrow \mathbf{v} = \mathbf{0}$.
- Thus $\mathbf{v}'(\mathbf{X}'\mathbf{X})\mathbf{v} > 0$ for $\mathbf{v} \neq \mathbf{0}$.
- Proving $\mathbf{X}'\mathbf{X}$ positive definite.
- This establishes (a) $\Rightarrow$ (b).

Showing (b) $\Rightarrow$ (c)

(a) The columns of $\mathbf{X}$ are linearly independent.

(b) $\mathbf{X}'\mathbf{X}$ is positive definite.

(c) $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

- $\mathbf{X}'\mathbf{X}$ is symmetric, for $(\mathbf{X}'\mathbf{X})' = \mathbf{X}'\mathbf{X}$.
- Thus we have the spectral decomposition $\mathbf{X}'\mathbf{X} = \mathbf{CDC}'$.
- And because $\mathbf{X}'\mathbf{X}$ is positive definite, its eigenvalues are all positive, $\mathbf{D}^{-1}$ is defined, and $(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{CD}^{-1}\mathbf{C}'$.
- So $(\mathbf{X}'\mathbf{X})^{-1}$ exists.
- This establishes (b) $\Rightarrow$ (c).

Showing (c) $\Rightarrow$ (a)

(a) The columns of $\mathbf{X}$ are linearly independent.

(b) $\mathbf{X}'\mathbf{X}$ is positive definite.

(c) $(\mathbf{X}'\mathbf{X})^{-1}$ exists.

- Let $\mathbf{X}\mathbf{v} = \mathbf{0}$. Seek to show $\mathbf{v} = \mathbf{0}$.
- $\mathbf{X}\mathbf{v} = \mathbf{0} \Rightarrow \mathbf{X}'\mathbf{X}\mathbf{v} = \mathbf{X}'\mathbf{0} = \mathbf{0}$
- $\Rightarrow (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\mathbf{v} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{0} = \mathbf{0}$.
- $\Rightarrow \mathbf{v} = \mathbf{0}$.
- And the columns of $\mathbf{X}$ are linearly independent.
- This establishes (c) $\Rightarrow$ (a).

■

The Message

- The least squares estimate $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ exists if and only if the columns of $\mathbf{X}$ are linearly independent.
- This just means the explanatory variables are not redundant.
- Example: Predicting final exam score.
- We will always assume that the columns of $\mathbf{X}$ are linearly independent. If not, fix it up.

"Predicted" $\mathbf{y}$

$$\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$$

- More like an estimated $E(\mathbf{y}) = E(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \mathbf{X}\boldsymbol{\beta}$.
- Could be denoted $\widehat{E(\mathbf{y})}$, but it's not.
- It would be *predicted* $\mathbf{y}$ only for a new sample with the same set of $\mathbf{X}$ values.
- $\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_k x_{ik}$ is a point on the best-fitting hyper-plane.

Residuals

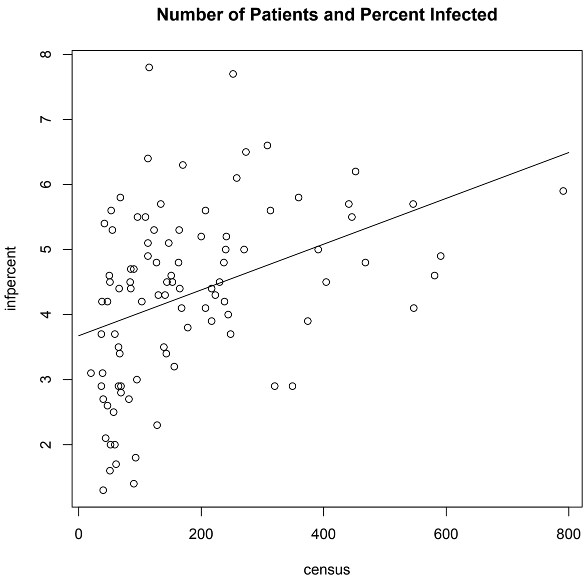
Vertical distances of the points from the hyperplane

$$\widehat{\boldsymbol{\epsilon}} = \mathbf{y} - \widehat{\mathbf{y}}$$

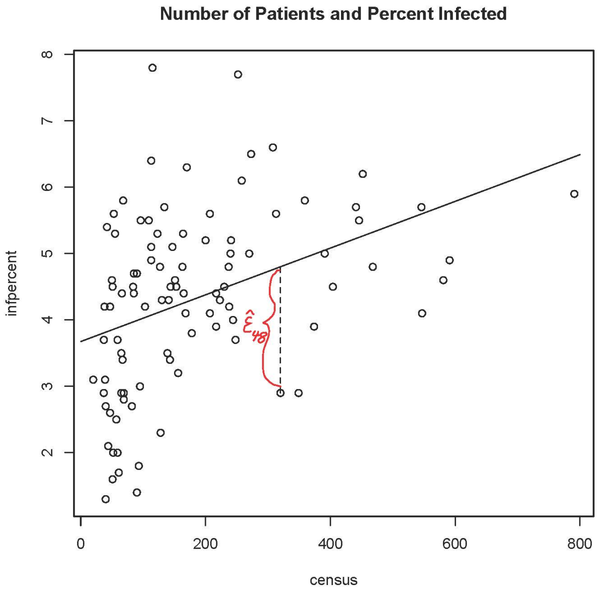
Why the funny notation?

$$\begin{aligned} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \;\Leftrightarrow\; \boldsymbol{\epsilon} &= \mathbf{y} - \mathbf{X}\boldsymbol{\beta} \\ \widehat{\boldsymbol{\epsilon}} &= \mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}} \end{aligned}$$

Hospital-Acquired Infection



Residual for Hospital 48



The Hat Matrix: $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

The hat matrix puts a hat on $\mathbf{y}$

$$
\begin{aligned}
\widehat{\mathbf{y}} &= \mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \mathbf{H}\mathbf{y}
\end{aligned}
$$

The Hat Matrix is Symmetric

Recall $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$

$$\begin{aligned}
\mathbf{H}' &= \left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right)' \\
&= \mathbf{X}''(\mathbf{X}'\mathbf{X})^{-1'}\mathbf{X}' \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{'-1}\mathbf{X}' \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{H}
\end{aligned}$$

The Hat Matrix is Idempotent

Meaning $\mathbf{HH} = \mathbf{H}$

$$
\begin{aligned}
\mathbf{HH} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \;\; \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{H}
\end{aligned}
$$

$\widehat{\boldsymbol{\epsilon}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$

$$\begin{aligned} \widehat{\boldsymbol{\epsilon}} &= \mathbf{y} - \widehat{\mathbf{y}} \\ &= \mathbf{y} - \mathbf{H}\mathbf{y} \\ &= \mathbf{I}\mathbf{y} - \mathbf{H}\mathbf{y} \\ &= (\mathbf{I} - \mathbf{H})\mathbf{y} \end{aligned}$$

$(\mathbf{I} - \mathbf{H})$ is also symmetric and idempotent.

$\mathbf{X}'\widehat{\boldsymbol{\epsilon}} = \mathbf{0}$

An important result

$$\begin{aligned}
\mathbf{X}'\widehat{\boldsymbol{\epsilon}} &= \mathbf{X}'(\mathbf{y} - \widehat{\mathbf{y}}) \\
&= \mathbf{X}'(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}) \\
&= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} \\
&= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \\
&= \mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{y} \\
&= \mathbf{0}
\end{aligned}$$

$\mathbf{X}'\widehat{\boldsymbol{\epsilon}} = \mathbf{0}$

The vector $\mathbf{0}$ is $(k+1) \times 1$

- The inner product of each row of $\mathbf{X}'$ and the vector of residuals is *zero.*
- The vector of residuals $\widehat{\boldsymbol{\epsilon}}$ is at right angles (orthogonal) to each column of $\mathbf{X}$, as vectors in $\mathbb{R}^n$.
- Also, this little formula makes certain calculations much easier.

Is it really a minimum?

- We have found that all the derivatives of

$$\begin{aligned} Q(\boldsymbol{\beta}) &= \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \end{aligned}$$

  are zero at $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$.
- Is the function really a minimum there, and not a maximum or saddle point?
- Multivariable second derivative test is to check whether all the eigenvalues of the Hessian matrix $\left(\frac{\partial^2 Q}{\partial\beta_i \partial\beta_j}\right)$ are positive.
- No thank you!

Minimize $Q(\boldsymbol{\beta})$ without calculus

Using $\mathbf{X}'\widehat{\boldsymbol{\epsilon}} = \mathbf{0}$

$$
\begin{aligned}
Q(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\
&= (\mathbf{y} - \widehat{\mathbf{y}} + \widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{y} - \widehat{\mathbf{y}} + \widehat{\mathbf{y}} - \mathbf{X}\boldsymbol{\beta}) \\
&= (\widehat{\boldsymbol{\epsilon}} + \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta})'(\widehat{\boldsymbol{\epsilon}} + \mathbf{X}\widehat{\boldsymbol{\beta}} - \mathbf{X}\boldsymbol{\beta}) \\
&= \left(\widehat{\boldsymbol{\epsilon}} + \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)' \left(\widehat{\boldsymbol{\epsilon}} + \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) \\
&= \left(\widehat{\boldsymbol{\epsilon}}' + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\right) \left(\widehat{\boldsymbol{\epsilon}} + \mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right) \\
&= \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} + \widehat{\boldsymbol{\epsilon}}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\widehat{\boldsymbol{\epsilon}} + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} + \quad 0 \quad + \quad 0 \quad + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \\
&= \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})
\end{aligned}
$$

$Q(\boldsymbol{\beta}) = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$

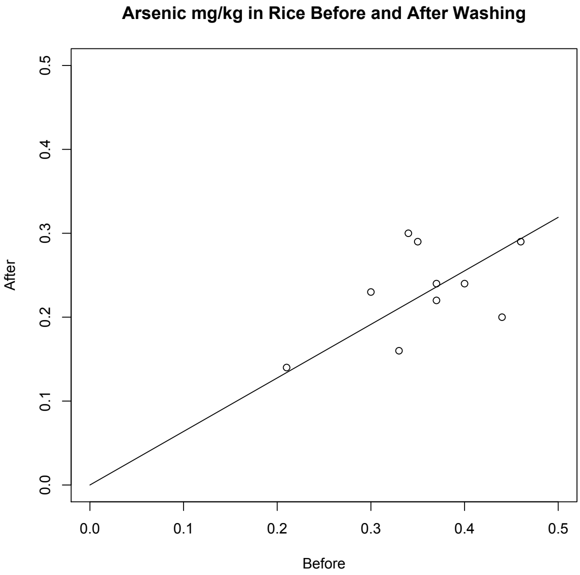
- The first term, $\sum_{i=1}^n \hat{\epsilon}_i^2 = \sum_{i=1}^n (y_i - \widehat{y}_i)^2$, is called the residual sum of squares, or sum of squares error.
- It does not depend (functionally) on $\boldsymbol{\beta}$.
- In the second term,
    - The columns of $\mathbf{X}$ are linearly independent, so $\mathbf{X}'\mathbf{X}$ is positive definite.
    - This means the second term is strictly positive except when $\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} = \mathbf{0}$.
  - $\Leftrightarrow$ $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$. Then, the second term equals zero.
    - So, $Q(\boldsymbol{\beta})$ has a unique minimum over $\boldsymbol{\beta}$ when $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}$.
- $\widehat{\boldsymbol{\beta}}$ really is the least squares estimate.

Regression through the origin

$$Q(\boldsymbol{\beta}) = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}} + (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\mathbf{X}'\mathbf{X}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

- This requires only that the columns of **X** be linearly independent.
- The first column does not have to be all ones.
- We can have regression models without an intercept.

Regression through the origin: $\widehat{y}_i = \widehat{\beta} x_i$



If there *is* an Intercept

First column of **X** is all ones and

$$\mathbf{X}'\widehat{\boldsymbol{\epsilon}} = \begin{pmatrix} 1 & 1 & 1 & 1 & \cdots & 1 \\ x_{11} & x_{21} & x_{31} & x_{41} & \cdots & x_{n1} \\ x_{12} & x_{22} & x_{32} & x_{42} & \cdots & x_{n2} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{1k} & x_{2k} & x_{3k} & x_{4k} & \cdots & x_{nk} \end{pmatrix} \begin{pmatrix} \widehat{\epsilon}_1 \\ \widehat{\epsilon}_2 \\ \widehat{\epsilon}_3 \\ \vdots \\ \widehat{\epsilon}_n \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

$$\sum_{i=1}^{n} \widehat{\epsilon}_i = 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} (y_i - \widehat{y}_i) = 0$$

$$\Leftrightarrow \quad \sum_{i=1}^{n} y_i = \sum_{i=1}^{n} \widehat{y}_i$$

Analysis of Variance

- Variance just means variation.
- Variation in a phenomenon means there is something to explain.
- Some businesses make more money than others. Why?
- Some students get higher grades. Why?
- Some covid-19 patients get a lot sicker. Why?
- We will measure variation to explain by variation around the sample mean:

$$SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

Theorem

If the regression model has an intercept,

$$\begin{array}{ccccc} SST & = & SSR & + & SSE \\ \sum_{i=1}^{n}(y_i - \overline{y})^2 & = & \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 & + & \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 \end{array}$$

Interpretation:

- With no predictor variables, the best guess at y_i is $\overline{y}$.
- SST is variation to be explained.
- With predictor variables, best guess is $\widehat{y}_i$
- SSE is variation still unexplained.
- So SSR must be variation that was explained.

Proof of $SST = SSR + SSE$

$$
\begin{aligned}
SST &= \sum_{i=1}^{n}(y_i - \overline{y})^2 = \sum_{i=1}^{n}(y_i - \widehat{y}_i + \widehat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^{n}(\widehat{\epsilon}_i + \widehat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^{n}\left(\hat{\epsilon}_i^2 + 2\,\widehat{\epsilon}_i(\widehat{y}_i - \overline{y}) + (\widehat{y}_i - \overline{y})^2\right) \\
&= \sum_{i=1}^{n}\hat{\epsilon}_i^2 + 2\sum_{i=1}^{n}\widehat{\epsilon}_i(\widehat{y}_i - \overline{y}) + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \\
&= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 + 0 + \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \\
&= SSE + SSR
\end{aligned}
$$

because ...

Showing middle term equals zero

$$\begin{aligned}
\sum_{i=1}^{n} \widehat{\epsilon}_i(\widehat{y}_i - \overline{y}) &= \sum_{i=1}^{n} \widehat{\epsilon}_i \widehat{y}_i - \sum_{i=1}^{n} \widehat{\epsilon}_i \overline{y} \\
&= \sum_{i=1}^{n} \widehat{y}_i \widehat{\epsilon}_i - \overline{y} \sum_{i=1}^{n} \widehat{\epsilon}_i \\
&= \widehat{\mathbf{y}}' \widehat{\boldsymbol{\epsilon}} + 0 \\
&= \left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right)' \widehat{\boldsymbol{\epsilon}} \\
&= \widehat{\boldsymbol{\beta}}' \mathbf{X}' \widehat{\boldsymbol{\epsilon}} \\
&= \widehat{\boldsymbol{\beta}}' \mathbf{0} \\
&= 0
\end{aligned}$$

Proportion of Variation Explained by Predictor Variables

Using $SST = SSR + SSE$

$$R^2 = \frac{SSR}{SST}$$

In simple regression $R^2 = r^2$

$$r = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}} \text{ and } \widehat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2}, \text{ so } \widehat{\beta}_1 = r\frac{s_y}{s_x}$$

$$\begin{aligned} r\frac{s_y}{s_x} &= \left(\frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2}}\right)\left(\frac{\sqrt{\sum_{i=1}^n (y_i - \overline{y})^2/(n-1)}}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2/(n-1)}}\right) \\ &= \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2}} \\ &= \widehat{\beta}_1 \end{aligned}$$

Still showing $R^2 = r^2$, using $\widehat{\beta}_1 = r\frac{s_y}{s_x}$ and $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$

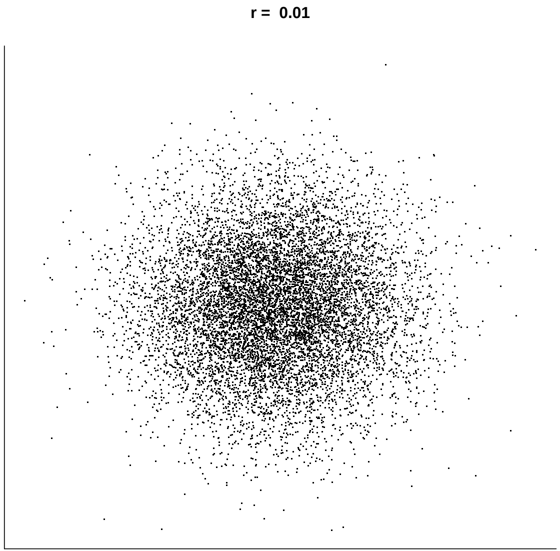
$$
\begin{aligned}
R^2 &= \frac{SSR}{SST} \\
&= \frac{1}{SST}\sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2 \\
&= \frac{1}{SST}\sum_{i=1}^{n}(\widehat{\beta}_0 + \widehat{\beta}_1 x_i - \overline{y})^2 \\
&= \frac{1}{SST}\sum_{i=1}^{n}(\overline{y} - \widehat{\beta}_1\overline{x} + \widehat{\beta}_1 x_i - \overline{y})^2 \\
&= \frac{1}{SST}\sum_{i=1}^{n}\left(\widehat{\beta}_1(x_i - \overline{x})\right)^2 \\
&= \frac{\widehat{\beta}_1^2}{SST}\sum_{i=1}^{n}(x_i - \overline{x})^2
\end{aligned}
$$

Continued . . .

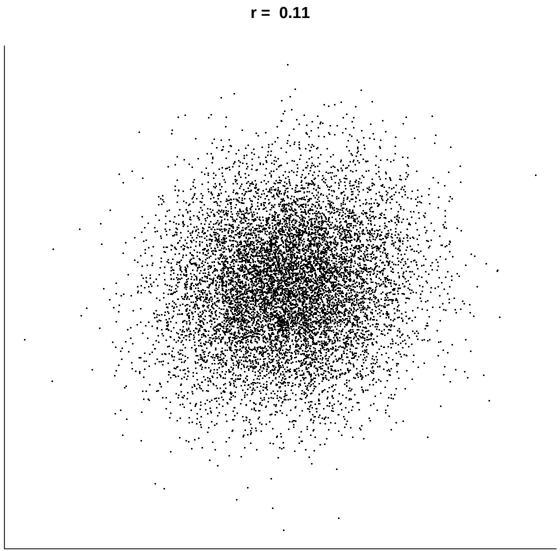
Using $\widehat{\beta}_1 = r\frac{s_y}{s_x}$

$$
\begin{aligned}
&= \frac{\widehat{\beta}_1^2}{SST}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \left(r\frac{s_y}{s_x}\right)^2 \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= r^2 \frac{s_y^2}{s_x^2} \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2} \\
&= r^2
\end{aligned}
$$

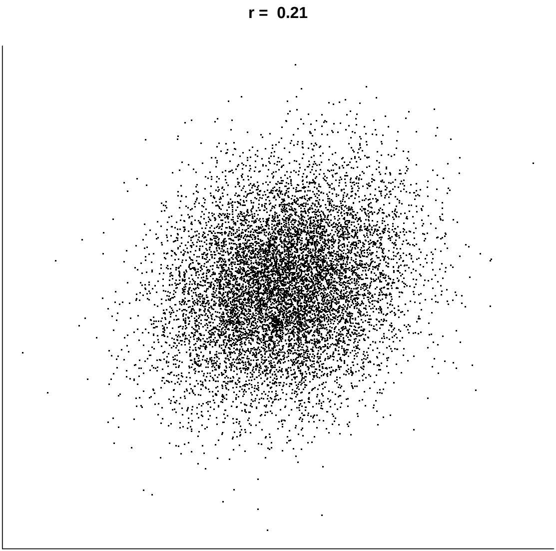
$R^2 = r^2$ helps with interpretation of R^2



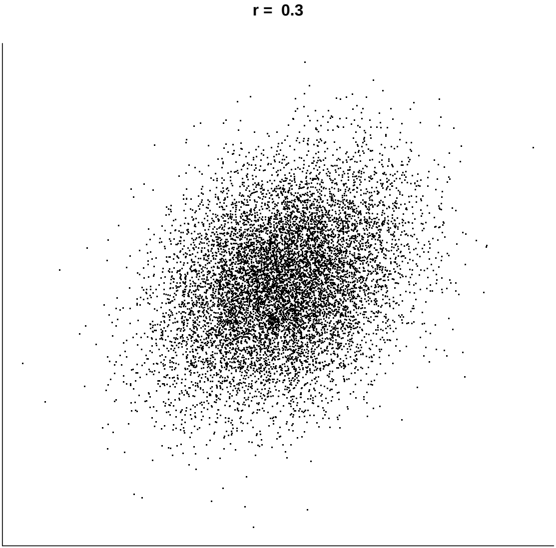
$R^2 = r^2$ helps with interpretation of R^2



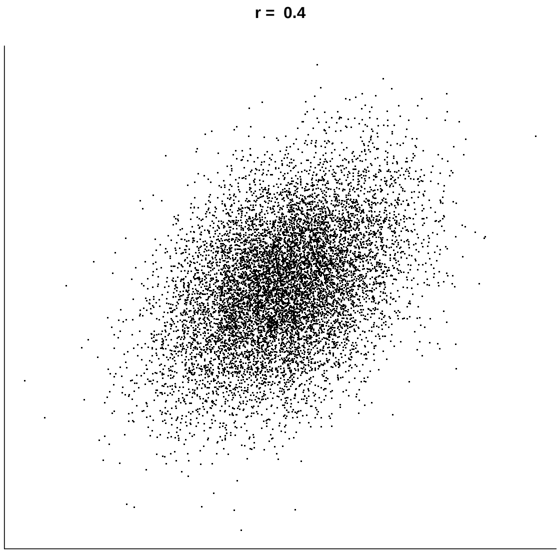
$R^2 = r^2$ helps with interpretation of R^2



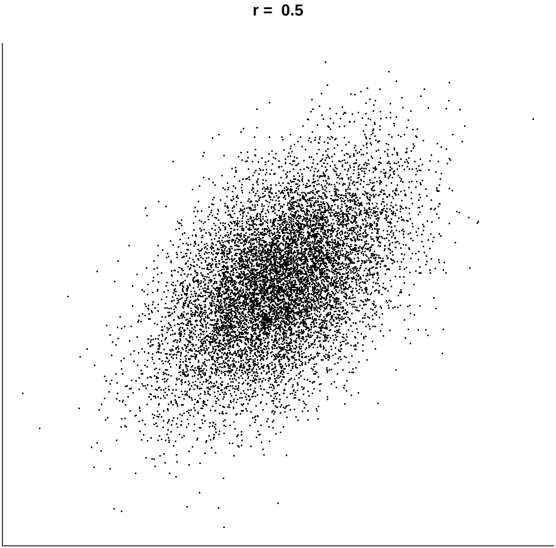
$R^2 = r^2$ helps with interpretation of R^2



$R^2 = r^2$ helps with interpretation of R^2



$R^2 = r^2$ helps with interpretation of R^2



Lesson

- Since I start to see a relationship at around $r = 0.3$, I start to get interested in a multiple regression when $R^2 > 0.09$.
- Also, the squared sample correlation between y_i and $\widehat{y}_i$ is R^2.

Estimating σ^2

Why estimate σ^2?

- The model says that $y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i$.
- The response is an expected value plus a piece of random noise, ϵ_i.
- $Var(\epsilon_i) = \sigma^2$, so σ^2 is how loud the noise is.
- The more noisy the data, the less precise the estimated β_j.
- Need to estimate how precise our estimates are.
- Estimated σ^2 appears in all the tests and confidence intervals.

Base estimate of σ^2 on SSE

- Can't estimate σ^2 by least squares, because $E(y_i)$ is not a function of σ^2.
- But think of $s^2 = \frac{\sum_{i=1}^{n}(y_i-\overline{y}_n)^2}{n-1}$
- Seek an estimator based on $SSE = \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$.
- But first, some preliminary results:
  - $tr(\mathbf{H}) = k + 1$
  - $(\mathbf{I} - \mathbf{H})\mathbf{y} = (\mathbf{I} - \mathbf{H})\boldsymbol{\epsilon}$

Preliminaries

$$
\begin{aligned}
tr(\mathbf{H}) &= tr\left(\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\right) \\
&= tr\left(\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\right) \\
&= tr\left(\mathbf{I}_{k+1}\right) \\
&= k+1
\end{aligned}
$$

Preliminaries

$$\begin{aligned}
(\mathbf{I}-\mathbf{H})\mathbf{y} &= (\mathbf{I}-\mathbf{H})(\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}) \\
&= (\mathbf{I}-\mathbf{H})\mathbf{X}\boldsymbol{\beta}+(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon} \\
&= \mathbf{X}\boldsymbol{\beta}-\mathbf{H}\mathbf{X}\boldsymbol{\beta}+(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon} \\
&= \mathbf{X}\boldsymbol{\beta}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}+(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon} \\
&= \mathbf{X}\boldsymbol{\beta}-\mathbf{X}\boldsymbol{\beta}+(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon} \\
&= (\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon}
\end{aligned}$$

Seek an unbiased estimator of σ^2

$$
\begin{aligned}
E\left\{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\right\} &= E\left\{\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}\right\} \\
&= E\left\{tr\left(\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}\right)\right\} \\
&= E\left\{tr\left([(\mathbf{I}-\mathbf{H})\mathbf{y}]'(\mathbf{I}-\mathbf{H})\mathbf{y}\right)\right\} \\
&= E\left\{tr\left([(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon}]'(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon}\right)\right\} \\
&= E\left\{tr\left(\boldsymbol{\epsilon}'(\mathbf{I}-\mathbf{H})'(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon}\right)\right\} \\
&= E\left\{tr\left(\boldsymbol{\epsilon}'(\mathbf{I}-\mathbf{H})\boldsymbol{\epsilon}\right)\right\} \\
&= E\left\{tr\left(\boldsymbol{\epsilon}\boldsymbol{\epsilon}'(\mathbf{I}-\mathbf{H})\right)\right\} \\
&= tr\left(E\left\{\boldsymbol{\epsilon}\boldsymbol{\epsilon}'\right\}(\mathbf{I}-\mathbf{H})\right) \\
&= tr\left(E\left\{(\boldsymbol{\epsilon}-\mathbf{0})(\boldsymbol{\epsilon}-\mathbf{0})'\right\}(\mathbf{I}-\mathbf{H})\right) \\
&= tr\left(cov(\boldsymbol{\epsilon})(\mathbf{I}-\mathbf{H})\right) \\
&= tr\left(\sigma^2\mathbf{I}_n(\mathbf{I}-\mathbf{H})\right) \\
&= \sigma^2 tr(\mathbf{I}-\mathbf{H}) = \sigma^2\left(tr(\mathbf{I}) - tr(\mathbf{H})\right) \\
&= \sigma^2\left(n-(k+1)\right)
\end{aligned}
$$

$E\left(\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2\right) = \sigma^2(n-k-1)$

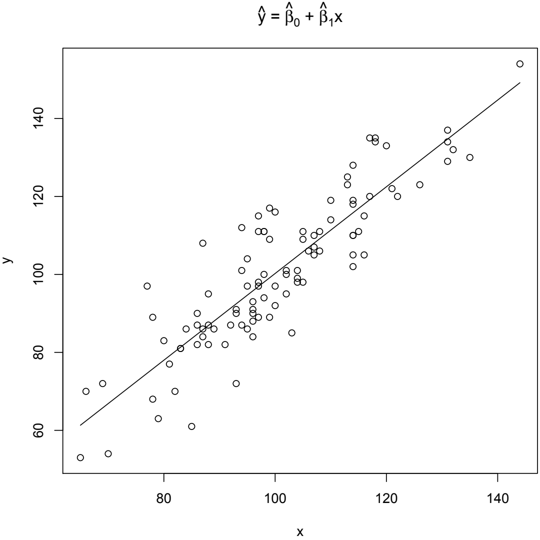
- $E(SSE) = \sigma^2(n-k-1)$, so $E\left(\frac{SSE}{n-k-1}\right) = \sigma^2$.
- $E(MSE) = \sigma^2 \quad (MSE = \text{Mean Squared Error})$
- s^2 is an unbiased estimator of σ^2, where

$$s^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2}{n-k-1} = \frac{\widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}}{n-k-1} = MSE$$

- We are estimating σ^2 with average squared vertical distance from the points to the plane.
- To avoid confusion, we will usually call it *MSE*.

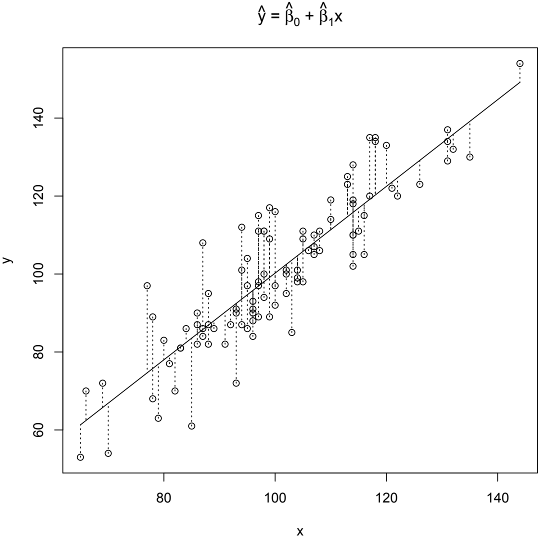
Least Squares Estimation is Curve Fitting

Minimizing $\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$

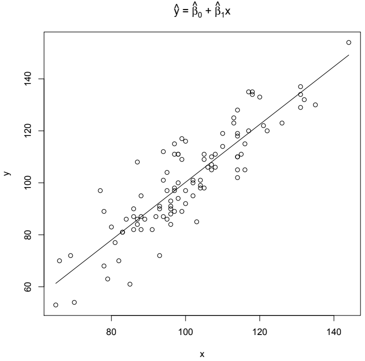


Least Squares Estimation is Curve Fitting

Minimizing $\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$



Best Fitting Line or Plane

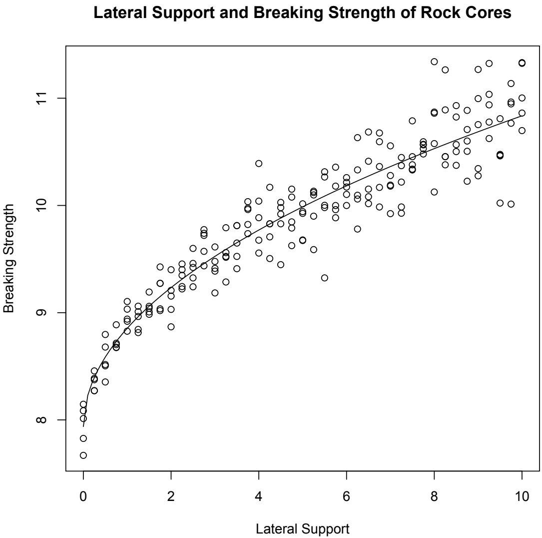


- $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$ is the equation of a line.
- $\widehat{y}_i$ is the point on the line corresponding to x_i.
- $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2$ is the equation of a plane.
- $\widehat{y}_i$ is the point on the plane corresponding to $(x_{i,1}, x_{i,2})$.

$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k$ is the equation of a hyper-plane.

Fitting a curve: $\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 \sqrt{x}$

Transform the explanatory variable



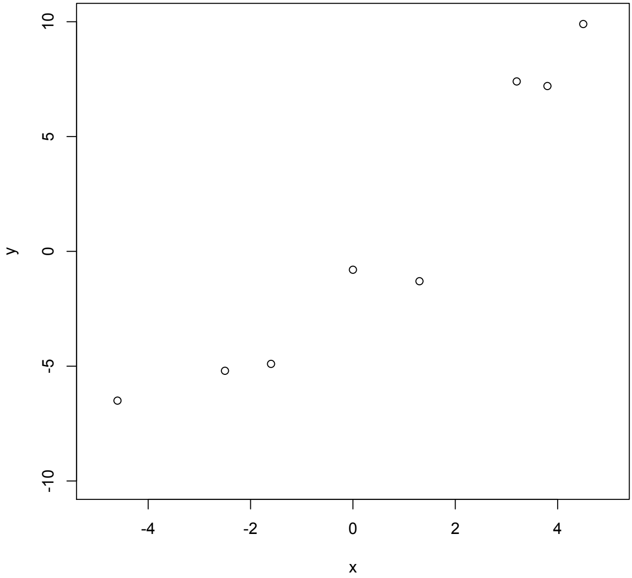
R Code for the Record

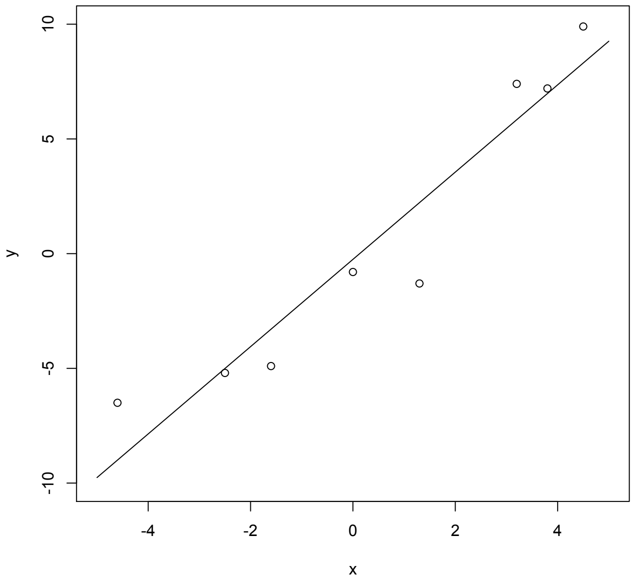
```
rm(list=ls())
rocks =
read.table('http://www.utstat.toronto.edu/~brunner/data/legal/rock1.data.txt')
head(rocks); attach(rocks)
plot(support,bforce, xlab = 'Lateral Support', ylab='Breaking Strength',
main = 'Lateral Support and Breaking Strength of Rock Cores')
sqrtsup = sqrt(support)
# Fit the model bforce_i = beta_0 + beta_1 sqrtsup_i + epsilon_i
fit = lm(bforce ~ sqrtsup)
betahat = coefficients(fit); betahat
xx = seq(from=0,to=10,by=0.1); yy = betahat[1] + betahat[2]*sqrt(xx)
lines(xx,yy)
```

Our text emphasizes curve fitting

In the presentation of least squares

- They minimize over $\widehat{\beta}_j$ rather than β_j right from the beginning. They minimize $Q(\widehat{\boldsymbol{\beta}}) = \widehat{\boldsymbol{\epsilon}}'\widehat{\boldsymbol{\epsilon}}$.
- We minimize over β_j and put hats on the answer.
- Their point is that the curve fitting can be useful (maybe for prediction) even if you don't believe the model at all.


10
5
0
-5
-10
y
-4
-2
0
2
4
x


10
5
0
-5
-10
y
-4
-2
0
2
4
x

Machine Learning

- Machine learning algorithms are often based on statistical models, but the models are often not mentioned.
- Prediction is emphasized over tests and confidence intervals.
- "Learning" means parameter estimation.
- The algorithm "learns" by minimizing a "loss function."
- In our case, the loss function is $\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \cdots - \beta_k x_{ik})^2$.
- In machine learning, the loss function is usually minimized numerically, but here we can do it explicitly.
- Sometimes, disregarding the model can lead to important new methods and insights.
- But even hard core machine learning hackers should know the details of one good model-based method.

Copyright Information