# STP 530
# Lecture 4: Multiple Regression I with Matrix Approach

Yi Zheng, Ph.D.
yi.isabel.zheng@asu.edu

ARIZONA STATE UNIVERSITY
SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES

September 19, 2023

# Outline

## Why use multiple regression model instead of simple regression model?

- If your primary purpose is **description**:

  A single predictor variable in the model would have provided an inadequate description since a number of key variables affect the response variable in important and distinctive ways.

- If your primary purpose is **prediction**:

  Predictions of the response variable based on a model containing only a single predictor variable are too imprecise to be useful.

(Textbook p.214)

**First-order additive multiple regression model:**

The hypothesized model:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

With the assumption $E\{\varepsilon_i\} = 0$, we have:

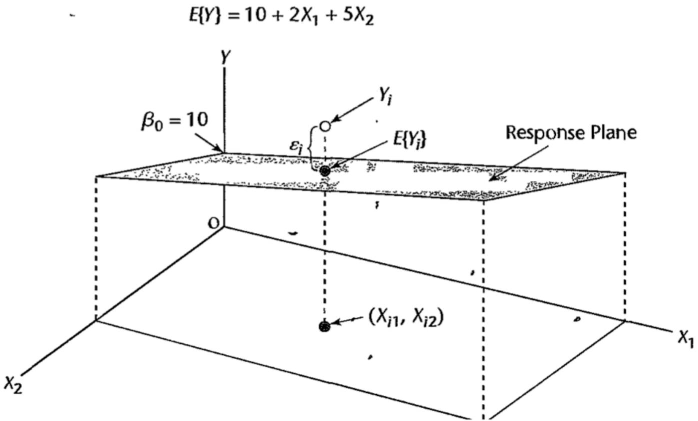
$$E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1}$$

The fitted regression function/equation:

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i2} + \cdots + b_{p-1} X_{i,p-1}$$

# First-order additive multiple regression model

**FIGURE 6.1 Response Function is a Plane—Sales Promotion Example.**



# First-order additive multiple regression model

$$E\{Y_i\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{p-1} X_{i,p-1} = \sum_{k=0}^{p-1} \beta_k X_{ik}, \text{ where } X_{i0} \equiv 1$$

- $p-1$: total number of predictors or independent variables.
- $X_{i1}, X_{i2}, \cdots, X_{i,p-1}$: $p-1$ predictors of the $i$th observation.
- $Y_i$: the dependent/response variable of the $i$th observation.
- $\varepsilon_i$: the random error term. In normal error regression models, we assume $\varepsilon_i \ iid \sim N(0, \sigma^2)$
- $\beta_0$: the $Y$-intercept — **The expected $Y$ value when all predictors are 0.**
- $\beta_k$: the slope of $X_k$ holding all other predictors constant. **The amount of increase in the expected $Y$ value for every 1-unit increase in $X_k$, holding all other predictors constant.**

# Proof of the interpretation of slopes in multiple regression

Assume we have two data points with 1-unit difference in $X_1$ holding all other predictors constant:

Data point 1: $x_1, \quad x_2, \quad x_3, \quad y$

Data point 2: $x_1^* = x_1 + 1, \quad x_2^* = x_2, \quad x_3^* = x_3, \quad y^*$

Based on the hypothesized linear model, we have:

Data point 1: $E\{y\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

Data point 2:
$$\begin{aligned} E\{y^*\} &= \beta_0 + \beta_1 x_1^* + \beta_2 x_2^* + \beta_3 x_3^* \\ &= \beta_0 + \beta_1 (x_1 + 1) + \beta_2 x_2 + \beta_3 x_3 \\ &= \beta_0 + \beta_1 x_1 + \beta_1 + \beta_2 x_2 + \beta_3 x_3 \end{aligned}$$

$$E\{y^*\} - E\{y\} = \beta_1$$

To estimate the parameters ($\beta$'s) in the model with the calculus method, we need to solve the equation system

$$\begin{cases} \dfrac{\partial Q}{\partial \beta_0} = 0 \\ \dfrac{\partial Q}{\partial \beta_1} = 0 \\ \vdots \\ \dfrac{\partial Q}{\partial \beta_{p-1}} = 0 \end{cases}$$

where

$$Q = \text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{i1} - \cdots - \beta_{p-1} X_{i,p-1})^2$$

That quickly becomes too complicated to solve analytically, which is why we now introduce the matrix approach to linear regression models.

## Multiple regression in matrix notation

In matrix terms, a multiple regression model is expressed by:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

## Multiple regression in matrix notation

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon$$

$$\mathbf{Y}_{n\times 1} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \qquad \mathbf{X}_{n\times p} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1,p-1} \\ 1 & X_{21} & X_{22} & \cdots & X_{2,p-1} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{n,p-1} \end{bmatrix}$$

$$\boldsymbol{\beta}_{p\times 1} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} \qquad \boldsymbol{\varepsilon}_{n\times 1} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

# Multiple regression in matrix notation

Hypothesized model:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Hypothesized model (expected outcome):

$$E\{\mathbf{Y}\} = \mathbf{X}\boldsymbol{\beta}$$

Fitted regression function:

$$\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$$

## Outline

## Matrix approach to least square estimation of multiple regression

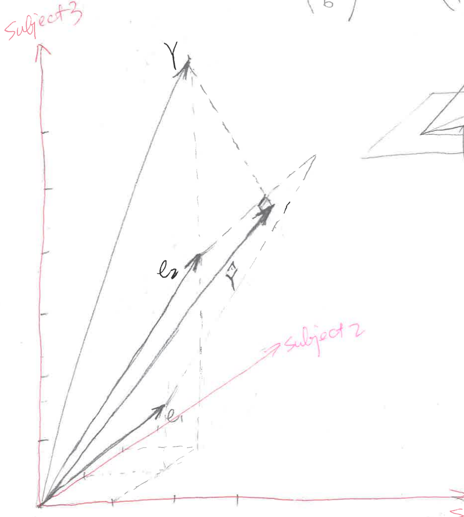
**Least square estimation:** Solve for model parameters $\beta_0, \beta_1, \cdots, \beta_{p-1}$ that minimize $Q = \text{SSE} = \sum_{i=1}^{n} \varepsilon_i^2$.

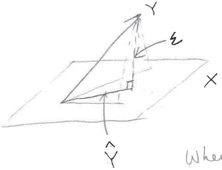
Note that within the matrix context, the **norm** of the $\boldsymbol{\varepsilon}$ vector, which is also the length of the vector, is defined as $\|\boldsymbol{\varepsilon}\| := \sqrt{\sum_{i=1}^{n} \varepsilon_i^2} = \sqrt{\text{SSE}}$.
So minimizing SSE is equivalent with **minimizing the length of the vector** $\boldsymbol{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$.

Also note that the fitted regression function $\hat{\mathbf{Y}} = \mathbf{Xb}$ implies that $\hat{\mathbf{Y}}$ lies in the space spanned by the columns of $\mathbf{X}$. Thus the length of $\boldsymbol{\varepsilon} = \mathbf{Y} - \hat{\mathbf{Y}}$ is minimized when $\hat{\mathbf{Y}}$ is the **projection** of $\mathbf{Y}$ onto the space spanned by the columns of $\mathbf{X}$, or equivalently, $\boldsymbol{\varepsilon}$ is orthogonal to the space spanned by the columns of $\mathbf{X}$:

$$\mathbf{X}'\boldsymbol{\varepsilon} = \mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}$$

$Y = \begin{pmatrix}1\\2\\6\end{pmatrix}$ $X = \begin{pmatrix}1&1\\1&2\\1&3\end{pmatrix}$ $\hat{Y} = X\beta = \begin{pmatrix}1&1\\1&2\\1&3\end{pmatrix}\begin{pmatrix}\beta_0\\\beta_1\end{pmatrix}$




$= \begin{pmatrix}\beta_0+\beta_1\\\beta_0+2\beta_1\\\beta_0+3\beta_1\end{pmatrix}$

$= \beta_0 e_1 + \beta_1 e_2$

Where $e_1 = \begin{pmatrix}1\\1\\1\end{pmatrix}$ $e_2 = \begin{pmatrix}1\\2\\3\end{pmatrix}$

are the two basis vectors of the X plane

Minimizing $SSE = \varepsilon_1^2+\varepsilon_2^2+\varepsilon_3^2$ is equivalent to minimizing $\|\varepsilon\| = \sqrt{\varepsilon_1^2+\varepsilon_2^2+\varepsilon_3^2}$ which is length of

$\varepsilon = Y - \hat{Y}$

so $\varepsilon$ must be orthogonal to the X plane

From the orthogonal relationship we have

$$\mathbf{X}'(\mathbf{Y} - \hat{\mathbf{Y}}) = \mathbf{0}$$

Thus,

$$\mathbf{X}'\mathbf{Y} - \mathbf{X}'\hat{\mathbf{Y}} = \mathbf{0}$$
$$\mathbf{X}'\hat{\mathbf{Y}} = \mathbf{X}'\mathbf{Y}$$
$$\mathbf{X}'\mathbf{X}\mathbf{b} = \mathbf{X}'\mathbf{Y}$$

Solve for **b**:

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

where

$$\mathbf{b} := \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

# Outline

# Expectation of b

Because in linear regression models, $X$'s are treated as known constants,

$$\begin{aligned} E\{\mathbf{b}\} &= E\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'E\{\mathbf{Y}\} \\ &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta}) \\ &= (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\boldsymbol{\beta} \\ &= \boldsymbol{\beta} \end{aligned}$$

This means $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$ is an **unbiased** estimator of $\boldsymbol{\beta}$.

# Variance-covariance matrix of b

$$\boldsymbol{\sigma}^2\{\mathbf{b}\} = \begin{bmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} & \cdots & \sigma\{b_0, b_{p-1}\} \\ \sigma\{b_1, b_0\} & \sigma^2\{b_1\} & \cdots & \sigma\{b_1, b_{p-1}\} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma\{b_{p-1}, b_0\} & \sigma\{b_{p-1}, b_1\} & \cdots & \sigma^2\{b_{p-1}\} \end{bmatrix}_{p\times p}$$

Note: The standard errors of each parameter estimate are the square roots of the respective diagonal elements of the above matrix.

## Derivation of the variance-covariance matrix of b

First note that

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \boldsymbol{\sigma}^2\{\mathbf{X}\beta + \varepsilon\} = \boldsymbol{\sigma}^2\{\varepsilon\} = \begin{bmatrix} \sigma^2 & & \\ & \ddots & \\ & & \sigma^2 \end{bmatrix} = \sigma^2\mathbf{I}_{n\times n}$$

Thus

$$\begin{aligned}
\boldsymbol{\sigma}^2\{\mathbf{b}\} =& \boldsymbol{\sigma}^2\{(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}\} \\
=& (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\boldsymbol{\sigma}^2\{\mathbf{Y}\}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\
=& (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}((\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')' \\
=& (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(((\mathbf{X}'\mathbf{X})^{-1})' \\
=& (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(((\mathbf{X}'\mathbf{X})')^{-1} \\
=& (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\sigma^2\mathbf{I}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
=& \sigma^2\mathbf{I}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\
=& \sigma^2(\mathbf{X}'\mathbf{X})^{-1}
\end{aligned}$$

The previous slide shows that the variance-covariance matrix of the vector **b** is:

$$\boldsymbol{\sigma}^2\{\mathbf{b}\} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$$

This matrix is **estimated** by replacing $\sigma^2$ with $s^2$:

$$\begin{aligned} \boldsymbol{s}^2\{\mathbf{b}\} &= s^2(\mathbf{X}'\mathbf{X})^{-1} \\ &= \frac{\text{SSE}}{n-p}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \frac{\mathbf{e}'\mathbf{e}}{n-p}(\mathbf{X}'\mathbf{X})^{-1} \end{aligned}$$

where $\underset{n\times 1}{\mathbf{e}} = [e_1, e_2, \cdots, e_n]' = \mathbf{Y} - \hat{\mathbf{Y}}$

## Confidence intervals for regression coefficients

The $1-\alpha$ confidence limits for $\beta_k$ for $k = 0, 1, \cdots, p-1$, are:

$$b_k \pm t(1-\alpha/2; df = n-p) \cdot s\{b_k\}$$

- $s\{b_k\}$ can be obtained by taking the **square root of the $k$th element of the main diagonal of the variance-covariance matrix $s^2\{\mathbf{b}\}$**.
- $t(1-\alpha/2; df = n-p)$ can be obtained by R code: `qt(p=(1-alpha/2),df=(n-p))`.

**Interpretation:**

- We are $(1-\alpha)100\%$ sure that the true population parameter $\beta_k$ falls within this interval.
- (For the slope parameters only) With a $(1-\alpha)100\%$ confidence, we estimate that expected $Y$ value changes by somewhere between LL and UL for each 1-unit increase in $X_k$ while holding all other predictors constant.

## Confidence interval of $E\{Y_h\}$ given $\mathbf{X}_h$

Assume we would like to estimate the mean response value $E\{Y_h\}$ given the predictor vector $\mathbf{X}_h' = [1, X_{h1}, X_{h2}, \cdots, X_{h,p-1}]$.

The **point estimator** of $E\{Y_h\}$ is: $\hat{Y}_h = \mathbf{X}_h'\mathbf{b}$

The **standard error** of $\hat{Y}_h$ as an estimator of $E\{Y_h\}$ is:

$$\sigma\{\hat{Y}_h\} = \sigma\sqrt{\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h}$$

The above is estimated by replacing $\sigma$ with $s$:

$$s\{\hat{Y}_h\} = s\sqrt{\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h}$$

where

$$s = \sqrt{\frac{\text{SSE}}{n-p}} = \sqrt{\frac{\mathbf{e}'\mathbf{e}}{n-p}}$$

## Confidence interval of $E\{Y_h\}$ given $\mathbf{X}_h$

The $1-\alpha$ **confidence interval** for $E\{Y_h\}$ is given by the following confidence limits:

$$\hat{Y}_h \pm t(1-\alpha/2; df = n-p) \cdot s\{\hat{Y}_h\}$$

**Interpretation:** We are $(1-\alpha)100\%$ sure that the mean value of $Y$ for all observations with $\mathbf{X}_h$ falls within this interval.

# Derivation of $\sigma^2\{\hat{Y}_h\}$:

$$
\begin{aligned}
\sigma^2\{\hat{Y}_h\} &= \sigma^2\{\mathbf{X}_h'\mathbf{b}\} \\
&= \mathbf{X}_h'\boldsymbol{\sigma}^2\{\mathbf{b}\}\mathbf{X}_h \\
&= \mathbf{X}_h'\sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h \\
&= \sigma^2\mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h
\end{aligned}
$$

## Prediction variance of $\hat{Y}_h$ given $\mathbf{X}_h$

Note that using $\hat{Y}_h$ to predict a single case given $\mathbf{X}_h$ involves two steps:

$$\hat{Y}_h \longrightarrow E\{Y_h\} \longrightarrow Y_h$$

So the **prediction variance** of $\hat{Y}_h$ comes from two **independent** sources:

1. The **sampling variance** when using $\hat{Y}_h$ to predict $E\{Y_h\}$ (i.e., the variation of the fitted line), that is, $\sigma^2\{\hat{Y}_h\}$.
2. The **residual variance** when using $E\{Y_h\}$ to predict $Y_h$ (i.e., the variation from the line to the point), that is, $\sigma^2$.

Thus the **prediction variance** of $\hat{Y}_h$ (i.e., the variance of using $\hat{Y}_h$ to predict a single case) is:

$$\sigma^2_{\text{pred}}\{\hat{Y}_h\} = \sigma^2\{\hat{Y}_h\} + \sigma^2 = \sigma^2(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

## Prediction interval of $\hat{Y}_h$

The variance of using $\hat{Y}_h$ to predict a single case is:

$$\sigma^2_{\text{pred}}\{\hat{Y}_h\} = \sigma^2(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

which is estimated by replacing $\sigma^2$ with $s^2$:

$$s^2_{\text{pred}}\{\hat{Y}_h\} = s^2(1 + \mathbf{X}_h'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_h)$$

The $1-\alpha$ **prediction interval** for $Y_h$ is given by the following prediction limits:

$$\hat{Y}_h \pm t(1-\alpha/2; df = n-p) \cdot s_{\text{pred}}\{\hat{Y}_h\}$$

**Interpretation:** With a $(1-\alpha)100\%$ confidence, we predict that the value of $Y$ for the next observation with a $\mathbf{X}_h$ falls within this interval.

# The Hat-Matrix H

We define the Hat-Matrix **H** by

$$\hat{\mathbf{Y}} = \mathbf{H}\mathbf{Y}$$

Because $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$ and $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$, we have

$$\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$$

So that Hat-Matrix

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

Note the Hat-Matrix is symmetric and idempotent:

$$\mathbf{H}\mathbf{H} = \mathbf{H}$$

# The variance-covariance matrix of the residual vector

Recall $\underset{n\times 1}{\mathbf{e}} = [e_1, e_2, \cdots, e_n]' = \mathbf{Y} - \hat{\mathbf{Y}}$

$$\begin{aligned}
\boldsymbol{\sigma}^2\{\mathbf{e}\} =& \boldsymbol{\sigma}^2\{\mathbf{Y} - \hat{\mathbf{Y}}\} \\
=& \boldsymbol{\sigma}^2\{(\mathbf{I} - \mathbf{H})\mathbf{Y}\} \\
=& (\mathbf{I} - \mathbf{H})\boldsymbol{\sigma}^2\{\mathbf{Y}\}(\mathbf{I} - \mathbf{H})' \\
& \qquad \text{(Because } \mathbf{I} - \mathbf{H} \text{ is a constant matrix)} \\
=& (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' \\
=& \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I} - \mathbf{H})' \\
=& \sigma^2(\mathbf{I} - \mathbf{H}) \quad \text{(Because } \mathbf{I} - \mathbf{H} \text{ is symmetric and idempotent)}
\end{aligned}$$

The variance-covariance matrix of the residual vector $\boldsymbol{\sigma}^2\{\mathbf{e}\} = \sigma^2(\mathbf{I} - \mathbf{H})$ is estimated by replacing $\sigma^2$ with $s^2$:

$$s^2\{\mathbf{e}\} = s^2(\mathbf{I} - \mathbf{H}) = \frac{\text{SSE}}{n-p}(\mathbf{I} - \mathbf{H}) = \frac{\mathbf{e}'\mathbf{e}}{n-p}(\mathbf{I} - \mathbf{H})$$

The main diagonal elements of this matrix are used to compute the "real" studentized residuals:

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{s\sqrt{1 - h_{ii}}}$$

Note: Compare the above with the "semistudentized" residual:

$$e_i^* = \frac{e_i}{s}$$

# Outline

# Matrix transpose

$$\underset{3\times 2}{\mathbf{A}} = \begin{bmatrix} 2 & 5 \\ 7 & 10 \\ 3 & 4 \end{bmatrix}$$

$$\underset{2\times 3}{\mathbf{A}'} = \begin{bmatrix} 2 & 7 & 3 \\ 5 & 10 & 4 \end{bmatrix}$$

R code:

```
A = matrix(c(2, 7, 3, 5, 10, 4), nrow=3, ncol=2)
t(A)
```

# Matrix addition and subtraction

$$\underset{3\times 2}{\mathbf{A}} = \begin{bmatrix} 1 & 4 \\ 2 & 5 \\ 3 & 6 \end{bmatrix} \qquad \underset{3\times 2}{\mathbf{B}} = \begin{bmatrix} 1 & 2 \\ 2 & 3 \\ 3 & 4 \end{bmatrix}$$

$$\underset{3\times 2}{\mathbf{A}+\mathbf{B}} = \begin{bmatrix} 1+1 & 4+2 \\ 2+2 & 5+3 \\ 3+3 & 6+4 \end{bmatrix} = \begin{bmatrix} 2 & 6 \\ 4 & 8 \\ 6 & 10 \end{bmatrix}$$

## Scalar multiplication

$$k\mathbf{A} = k \begin{bmatrix} 2 & 7 \\ 9 & 3 \end{bmatrix} = \begin{bmatrix} 2k & 7k \\ 9k & 3k \end{bmatrix}$$

## Matrix multiplication

$$\mathbf{AB} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix} \begin{bmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{bmatrix}$$
$$= \begin{bmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{bmatrix}$$

Note: $\mathbf{AB} \neq \mathbf{BA}$

# Regression model in matrix notation

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

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$= \begin{bmatrix} \beta_0 + \beta_1 X_1 \\ \beta_0 + \beta_1 X_2 \\ \vdots \\ \beta_0 + \beta_1 X_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 + \varepsilon_1 \\ \beta_0 + \beta_1 X_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 X_n + \varepsilon_n \end{bmatrix}$$

# Diagonal matrix

$$\begin{bmatrix} a_1 & 0 & 0 \\ 0 & a_2 & 0 \\ 0 & 0 & a_3 \end{bmatrix}$$

# Identity matrix

$$\mathbf{I}_1 = \begin{bmatrix} 1 \end{bmatrix}, \quad \mathbf{I}_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \cdots, \mathbf{I}_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

$$\mathbf{AI} = \mathbf{IA} = \mathbf{A}$$

## Scalar matrix

$$\begin{bmatrix} k & 0 & 0 \\ 0 & k & 0 \\ 0 & 0 & k \end{bmatrix} = k \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = k\mathbf{I}$$

## The zero vector, the one vector, and the matrix of ones

$$\underset{r\times 1}{\mathbf{0}} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad \underset{r\times 1}{\mathbf{1}} = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \qquad \underset{r\times r}{\mathbf{J}} = \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix} = \mathbf{1}\mathbf{1}'$$

## Inverse of a matrix

$$\mathbf{A}^{-1}\mathbf{A} = \mathbf{A}\mathbf{A}^{-1} = \mathbf{I}$$

## Inverse of a diagonal matrix

$$\underset{3\times 3}{\mathbf{A}} = \begin{bmatrix} 3 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 2 \end{bmatrix}$$

$$\underset{3\times 3}{\mathbf{A}^{-1}} = \begin{bmatrix} 1/3 & 0 & 0 \\ 0 & 1/4 & 0 \\ 0 & 0 & 1/2 \end{bmatrix}$$

How the inverse of the other forms of matrices is calculated manually is nuanced. We will leave it to the computer.

## Using matrix inverse to solve equation systems

To solve for $\mathbf{Y}$ in the following equation system

$$\mathbf{AY} = \mathbf{C}$$

we pre-multiply the inverse of $\mathbf{A}$ to both sides:

$$\mathbf{A}^{-1}\mathbf{AY} = \mathbf{A}^{-1}\mathbf{C}$$

Because $\mathbf{A}^{-1}\mathbf{A} = \mathbf{AA}^{-1} = \mathbf{I}$, and $\mathbf{IY} = \mathbf{YI} = \mathbf{Y}$, we have obtained the solution:

$$\mathbf{Y} = \mathbf{A}^{-1}\mathbf{C}$$

## Regression model in matrix notation

$$\mathbf{A} + \mathbf{B} = \mathbf{B} + \mathbf{A} \tag{5.25}$$

$$(\mathbf{A} + \mathbf{B}) + \mathbf{C} = \mathbf{A} + (\mathbf{B} + \mathbf{C}) \tag{5.26}$$

$$(\mathbf{AB})\mathbf{C} = \mathbf{A}(\mathbf{BC}) \tag{5.27}$$

$$\mathbf{C}(\mathbf{A} + \mathbf{B}) = \mathbf{CA} + \mathbf{CB} \tag{5.28}$$

$$k(\mathbf{A} + \mathbf{B}) = k\mathbf{A} + k\mathbf{B} \tag{5.29}$$

$$(\mathbf{A}')' = \mathbf{A} \tag{5.30}$$

$$(\mathbf{A} + \mathbf{B})' = \mathbf{A}' + \mathbf{B}' \tag{5.31}$$

$$(\mathbf{AB})' = \mathbf{B}'\mathbf{A}' \tag{5.32}$$

$$(\mathbf{ABC})' = \mathbf{C}'\mathbf{B}'\mathbf{A}' \tag{5.33}$$

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1} \tag{5.34}$$

$$(\mathbf{ABC})^{-1} = \mathbf{C}^{-1}\mathbf{B}^{-1}\mathbf{A}^{-1} \tag{5.35}$$

$$(\mathbf{A}^{-1})^{-1} = \mathbf{A} \tag{5.36}$$

$$(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})' \tag{5.37}$$

# Expectation of a random vector

$$\mathbf{E}\{\mathbf{Y}\} = \begin{bmatrix} E\{Y_1\} \\ E\{Y_2\} \\ \vdots \\ E\{Y_n\} \end{bmatrix}_{n\times 1}$$

$$\mathbf{E}\{\boldsymbol{\varepsilon}\} = \underset{n\times 1}{\mathbf{0}}$$

which means:

$$\begin{bmatrix} E\{\varepsilon_1\} \\ E\{\varepsilon_2\} \\ \vdots \\ E\{\varepsilon_n\} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}_{n\times 1}$$

## Variance-covariance matrix of a random vector

$$\boldsymbol{\sigma}^2\{\mathbf{Y}\} = \begin{bmatrix} \sigma^2\{Y_1\} & \sigma\{Y_1, Y_2\} & \cdots & \sigma\{Y_1, Y_n\} \\ \sigma\{Y_2, Y_1\} & \sigma^2\{Y_2\} & \cdots & \sigma\{Y_2, Y_n\} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma\{Y_n, Y_1\} & \sigma\{Y_n, Y_2\} & \cdots & \sigma^2\{Y_n\} \end{bmatrix}_{n\times n}$$

# Variance-covariance matrix of the error vector

$$\boldsymbol{\sigma}^2\{\varepsilon\} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix}_{n\times n}$$

OR

$$\boldsymbol{\sigma}^2\{\varepsilon\} = \sigma^2\mathbf{I}$$

Some properties of **expectation matrices** and **variance-covariance matrices**:

Assume **A** is a constant matrix, **Y** is a random matrix,

$$\mathbf{E}\{\mathbf{A}\} = \mathbf{A}$$

$$\mathbf{E}\{\mathbf{AY}\} = \mathbf{A}\,\mathbf{E}\{\mathbf{Y}\}$$

$$\boldsymbol{\sigma}^2\{\mathbf{AY}\} = \mathbf{A}\,\boldsymbol{\sigma}^2\{\mathbf{Y}\}\,\mathbf{A}'$$

## Orthogonality

Two vectors $\boldsymbol{v}$ and $\boldsymbol{w}$ are orthogonal when their dot/inner product $\boldsymbol{v} \cdot \boldsymbol{w} = 0$, or

$$\boldsymbol{v}'\boldsymbol{w} = 0$$

A vector $\boldsymbol{w}$ is orthogonal to the space spanned by the columns of a matrix $\mathbf{A}$ when

$$\mathbf{A}'\boldsymbol{w} = \mathbf{0}$$