

## Multiple Linear Model

designed to help answer questions like "is  $y$  related to  $x$ , after accounting for  $x_2, x_3, \dots, x_p$ ?"

response:  $y_1, y_2, \dots, y_n$

covariates:  $x_{11}, x_{21}, \dots, x_{n1}$

$x_{12}, x_{22}, \dots, x_{n2}$

$x_{13}, x_{23}, \dots, x_{n3}$

Model for  $y_i$ :

$$\varepsilon_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$$

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + \varepsilon_i \quad b_0, b_1, b_2, b_3, \sigma^2 \text{ unknown numbers}$$

why do we say "after accounting for  $x_2, x_3, \dots$ ?"

Suppose covariates for subjects 1 and 2 satisfy

$$x_{21} = x_{11} + u, \quad x_{22} = x_{12} + v, \quad x_{23} = x_{13} + w$$

$$\begin{aligned} E(Y_2 - Y_1) &= E(Y_2) - E(Y_1) & E(a+bX) &= a+bE(X) \\ &= E(b_0 + b_1 x_{21} + b_2 x_{22} + b_3 x_{23} + \varepsilon_2) \\ &\quad - E(b_0 + b_1 x_{11} + b_2 x_{12} + b_3 x_{13} + \varepsilon_1) \nearrow 0 \\ &= b_0 + b_1 x_{21} + b_2 x_{22} + b_3 x_{23} + E(\varepsilon_2) \nearrow 0 \\ &\quad - (b_0 + b_1 x_{11} + b_2 x_{12} + b_3 x_{13} + E(\varepsilon_1)) \\ &= b_1(x_{21} - x_{11}) + b_2(x_{22} - x_{12}) + b_3(x_{23} - x_{13}) \\ &= b_1 u + b_2 v + b_3 w \end{aligned}$$

The total effect is  $b_1 u + b_2 v + b_3 w$ , but the effect of changing  $x_1$  after accounting for the changes due to  $x_2$  and  $x_3$  is  $b_1 u$

The model for  $y_1, \dots, y_n$  can be written together in matrix form:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & x_{12} & x_{13} \\ x_{20} & x_{21} & x_{22} & x_{23} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & x_{n3} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$x_{10} = 1$

$$\underline{Y} = \underline{X} \underline{b} + \underline{\epsilon}$$

$n \times 1 \quad n \times 4 \quad 4 \times 1 \quad n \times 1$

In general, we have  $p$  covariates and model  $y_i$  as

$$Y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \dots + b_p x_{ip} + \epsilon_i, \quad \epsilon_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$$

write together in matrix form as

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & \dots & x_{1p} \\ \vdots & \vdots & & \vdots \\ x_{n0} & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$n \times 1 \quad n \times (p+1) \quad (p+1) \times 1 \quad n \times 1$

given the data  $\underline{y}$  and  $\underline{X}$ , we estimate  $\underline{b}$  by minimizing squared residuals

$$RSS(\underline{b}^*) = \sum_{i=1}^n \left( y_i - \sum_{j=0}^p b_j^* x_{ij} \right)^2$$

$$\frac{\partial RSS(\underline{b}^*)}{\partial b_k^*} = -2 \sum_{i=1}^n \left( y_i - \sum_{j=0}^p b_j^* x_{ij} \right) x_{ik}$$

set each derivative equal to 0

$$\begin{aligned} \sum_{i=1}^n y_i x_{ik} &= \sum_{i=1}^n \sum_{j=0}^p \hat{b}_j x_{ij} x_{ik} \\ &= \sum_{j=0}^p \hat{b}_j \sum_{i=1}^n x_{ij} x_{ik} \end{aligned}$$

this is one equation.  
We have  $p+1$  of them total.

Normal Equations  $\rightarrow$

$$\sum_{i=1}^n y_i x_{i0} = \sum_{j=0}^p \hat{b}_j \sum_{i=1}^n x_{ij} x_{i0}$$

$$\vdots$$

$$\sum_{i=1}^n y_i x_{ip} = \sum_{j=0}^p \hat{b}_j \sum_{i=1}^n x_{ij} x_{ip}$$

exercise: verify that these can be written in matrix form as

$$X^T \underline{y} = X^T X \hat{\underline{b}}$$

if  $\text{rank}(X) = p+1$  (full column rank), then  $(X^T X)^{-1}$  exists, and

$$(X^T X)^{-1} X^T \underline{y} = (X^T X)^{-1} (X^T X) \hat{\underline{b}} \rightarrow \hat{\underline{b}} = (X^T X)^{-1} X^T \underline{y}$$

$$\hat{\underline{b}} = (X^T X)^{-1} X^T \underline{y} \quad \text{The most important equation in all of statistics}$$

fitted values:  $\hat{\underline{y}} = X \hat{\underline{b}} = X (X^T X)^{-1} X^T \underline{y} = P \underline{y}$ .

residuals:  $\hat{\underline{e}} = \underline{y} - X \hat{\underline{b}} = (I - X(X^T X)^{-1} X^T) \underline{y} = (I - P) \underline{y}$

P and I - P are very special matrices called projection matrices

## Projections :

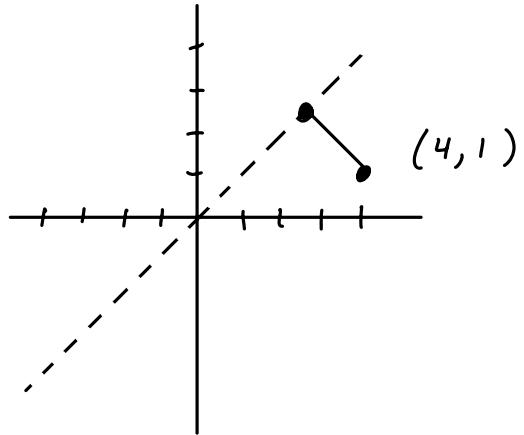
A projection is a special linear transformation. (i.e  $y \rightarrow Ay$ )

### Example :

$$P = \begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{bmatrix} \quad y = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

$$Py = \begin{bmatrix} 2 + 1/2 \\ 2 + 1/2 \end{bmatrix} = \begin{bmatrix} 5/2 \\ 5/2 \end{bmatrix}$$

What does this projection "do"?



Moves the point  $(4, 1)$  onto the line described by  $a(1, 1)$

- $(5/2, 5/2) = 5/2(1, 1)$ .
- We say that  $5/2$  is the coordinates of  $(5/2, 5/2)$  in the linear space spanned by  $(1, 1)$

In fact, the projection finds the closest point on  $a(1, 1)$  to  $(4, 1)$ .

This works for any input vector  $y$  (try it!)

In general, an  $n \times n$  projection matrix  $P$  maps vectors to the closest point in the linear subspace spanned by the columns of  $P$ .

Projection matrices must satisfy 2 properties

1. symmetric:  $P = P^T$

2. idempotent:  $PP = P$

Example :

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}, \quad P^T = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \quad PP = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} + \frac{1}{4} & \frac{1}{4} + \frac{1}{4} \\ \frac{1}{4} + \frac{1}{4} & \frac{1}{4} + \frac{1}{4} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix}$$

projections are the secret sauce of linear regression

Example :  $P = X(X^T X)^{-1} X^T$  ✓

symmetric:  $P^T = X(X^T X)^{-1} X^T = P$

↑  
 $X^T X$  is symmetric, therefore so is inverse

idempotent:  $PP = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X(X^T X)^{-1} X^T = P$  ✓

Some properties of projections

① if  $P$  is a projection, so is  $I - P$

symmetric:  $I - P = I - P^T = I - P$  ✓

idempotent:  $(I - P)(I - P) = I - P - P + PP = I - P - P + P = I - P$  ✓

Consequence: if  $P = X(X^T X)^{-1} X^T y$

$$(I - P)y = y - Py = y - X\hat{b} = e$$

\* The residuals are a projection onto  $I - P$

$$e^T \hat{y} = \hat{y}^T (I - P)Py = \hat{y}^T (P - PP)y = \hat{y}^T (P - P)y = 0$$

\* residuals are orthogonal to fitted values

② if you have a matrix  $M$  with  $k$  linearly independent columns

( $k < n$ ,  $M = [\underline{m}_1 \ \underline{m}_2 \ \dots \ \underline{m}_k]$ ) then  $M(M^T M)^{-1} M^T$  is a projection

onto the space spanned by  $\underline{m}_1, \dots, \underline{m}_k$  (onto the columns of  $M$ )

$$\underbrace{M(M^T M)^{-1} M^T}_{n \times k \quad k \times 1} y = M \underline{d} = [\underline{m}_1 \ \dots \ \underline{m}_k] \begin{bmatrix} d_1 \\ \vdots \\ d_k \end{bmatrix} = d_1 \underline{m}_1 + d_2 \underline{m}_2 + \dots + d_k \underline{m}_k$$

Consequence:  $n \times 1$

if  $P = X(X^T X)^{-1} X^T$  then  $\hat{y} = Py = X \hat{b} = \hat{b}_0 \underline{x}_0 + \hat{b}_1 \underline{x}_1 + \cdots + \hat{b}_p \underline{x}_p$

\* regression coefficients are the coordinates of the fitted values  
in the space spanned by the columns of  $X$ .

③ An  $n \times n$  projection matrix  $P$  with rank  $k$  can be decomposed as

$$P = \underset{n \times n}{U} \underset{n \times k}{U^T} \quad \text{where the columns of } U (\underline{u}_1, \dots, \underline{u}_k)$$

are an orthonormal basis for the columns of  $P$

- $\underline{u}_i^T \underline{u}_i = 1$  and  $\underline{u}_i^T \underline{u}_j = 0$  for  $i \neq j$

## Properties of linear transformations of random vectors

We want to know sampling dist of  $\hat{\underline{B}} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y} = \underline{M} \underline{Y}$

Need to know about distribution of random vectors ( $\hat{\underline{B}}$  and  $\underline{Y}$ )  
and distributions of linear functions of random vectors ( $\underline{M} \underline{Y}$ )

### Expectation

$$\underline{z} = (z_1, \dots, z_n)^T \quad \text{column vector}$$

$$E(\underline{z}) := (E(z_1), \dots, E(z_n))^T \quad \text{definition of expectation of random vector}$$

$$E(M\underline{z}) = E \left[ \begin{array}{c} \sum_{j=1}^n m_{1j} z_j \\ \vdots \\ \sum_{j=1}^n m_{pj} z_j \end{array} \right] = \left[ \begin{array}{c} E\left(\sum_{j=1}^n m_{1j} z_j\right) \\ \vdots \\ E\left(\sum_{j=1}^n m_{pj} z_j\right) \end{array} \right]$$

$$= \left[ \begin{array}{c} \sum_{j=1}^n m_{1j} E(z_j) \\ \vdots \\ \sum_{j=1}^n m_{pj} E(z_j) \end{array} \right] = \underbrace{M}_{p \times n} \underbrace{E(\underline{z})}_{n \times 1} \quad \text{follows from the linearity property of expectation}$$

## Covariance

The covariance matrix of a vector  $\underline{z}$  is defined as

$$\underset{n \times n}{\text{Cov}}(\underline{z}) = \begin{bmatrix} \text{Cov}(z_1, z_1) & \text{Cov}(z_1, z_2) & \cdots & \text{Cov}(z_1, z_n) \\ \text{Cov}(z_2, z_1) & \text{Cov}(z_2, z_2) & \cdots & \text{Cov}(z_2, z_n) \\ \vdots & & & \\ \text{Cov}(z_n, z_1) & \text{Cov}(z_n, z_2) & \cdots & \text{Cov}(z_n, z_n) \end{bmatrix}$$

Covariance matrices are symmetric and can also be written as

$$\underset{n \times n}{\text{Cov}}(\underline{z}) = E \left[ (\underline{z} - E(\underline{z})) (\underline{z} - E(\underline{z}))^T \right]$$

$E[\text{matrix}]$  means take  $E$  of each entry (as in  $E(\text{vector})$ )

$$\begin{aligned} \underset{p \times n}{M} \underset{n \times 1}{\text{Cov}}(\underline{z}) &= E \left( (M\underline{z} - M E(\underline{z})) (M\underline{z} - M E(\underline{z}))^T \right) \\ &= E \left\{ [M(\underline{z} - E(\underline{z}))] [M(\underline{z} - E(\underline{z}))]^T \right\} \\ &= E \left\{ M(\underline{z} - E(\underline{z})) (\underline{z} - E(\underline{z}))^T M^T \right\} \\ &= M E \left[ (\underline{z} - E(\underline{z})) (\underline{z} - E(\underline{z}))^T \right] M^T \\ &= M \text{Cov}(\underline{z}) M^T \end{aligned}$$

## Multivariate normal distribution

If  $\underline{z}$  has multivariate normal distribution, we use notation

$$\underline{z} \sim N(\underline{\mu}, \Sigma)$$

$n \times 1$

where  $\underline{\mu} = E(\underline{z})$  and  $\Sigma = \text{Cov}(\underline{z})$

$\underline{z}$  has probability density

$$p(\underline{z}) = \frac{1}{(2\pi)^n/2} \det(\Sigma)^{-1/2} \exp\left(-\frac{1}{2}(\underline{z}-\underline{\mu})^T \Sigma^{-1} (\underline{z}-\underline{\mu})\right)$$

$$\det(\Sigma) = \text{determinant of } \Sigma \quad \exp(a) = e^a$$

Example:  $E(\underline{z}) = \underline{\mu} = \begin{bmatrix} \mu \\ \vdots \\ \mu \end{bmatrix}$   $\text{Cov}(\underline{z}) = \begin{bmatrix} \sigma^2 & & & \\ & \ddots & & 0 \\ & & \ddots & \\ 0 & & & \sigma^2 \end{bmatrix} = \sigma^2 I$

$$\begin{aligned} p(\underline{z}) &= \frac{1}{(2\pi)^n/2} \det(\sigma^2 I)^{-1/2} \exp\left(-\frac{1}{2}(\underline{z}-\underline{\mu})^T (\sigma^2 I)^{-1} (\underline{z}-\underline{\mu})\right) \\ &= \frac{1}{(2\pi)^n/2} \left((\sigma^2)^n\right)^{-1/2} \exp\left(-\frac{1}{2}(\underline{z}-\underline{\mu})^T \left(\frac{1}{\sigma^2} I\right) (\underline{z}-\underline{\mu})\right) \\ &= \frac{1}{(2\pi)^{n/2}} \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2}(\underline{z}-\underline{\mu})^T (\underline{z}-\underline{\mu})\right) \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2} \sum_{i=1}^n (z_i - \mu_i)^2\right) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(z_i - \mu_i)^2}{2\sigma^2}\right) = \prod_{i=1}^n p(z_i) \end{aligned}$$

Therefore  $z_1, \dots, z_n \stackrel{\text{ind}}{\sim} N(\mu, \sigma^2)$

## Properties of multivariate normal

- completely determined by mean vector and covariance matrix
- if  $\underline{z} \sim N(\underline{\mu}, \Sigma)$ , then  $\underset{p \times 1}{\underline{a}} + M \underset{p \times n}{\underline{z}}$  is multivariate normal  
$$\underset{p \times 1}{\underline{a}} + M \underset{p \times n}{\underline{z}} \sim N\left(\underset{p \times 1}{\underline{a}} + M \underline{\mu}, M \Sigma M^T\right)$$

exercise: verify this by using transformation  
of variables

Example: distribution of average:

$$\underline{Y} \sim N(b_0 \mathbf{1}, \sigma^2 I) \longrightarrow Y_1, \dots, Y_n \stackrel{\text{ind}}{\sim} N(b_0, \sigma^2)$$

$$\bar{Y} = \left(\frac{1}{n} \mathbf{1}^T\right) \underline{Y}$$

$$E(\bar{Y}) = E\left(\frac{1}{n} \mathbf{1}^T \underline{Y}\right) = \frac{1}{n} \mathbf{1}^T E(\underline{Y}) = \frac{b_0}{n} \mathbf{1}^T \mathbf{1} = \frac{b_0}{n} n = b_0$$

$$\begin{aligned} \text{Cov}(\bar{Y}) &= \left(\frac{1}{n} \mathbf{1}^T\right) (\sigma^2 I) \left(\frac{1}{n} \mathbf{1}\right) = \frac{\sigma^2}{n^2} (\mathbf{1}^T I \mathbf{1}) \\ &= \frac{\sigma^2}{n^2} (\mathbf{1}^T \mathbf{1}) = \frac{\sigma^2}{n^2} \cdot n = \frac{\sigma^2}{n} \end{aligned}$$

$$\Rightarrow \bar{Y} \sim N\left(b_0, \frac{\sigma^2}{n}\right)$$

## Properties of multiple linear regression estimators

$$\underline{Y} = \underline{X} \underline{b} + \underline{\varepsilon}, \quad \underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T \quad \varepsilon_i \stackrel{\text{ind}}{\sim} N(0, \sigma^2)$$

$$\underline{\varepsilon} \sim N(0, \sigma^2 I) \quad \text{multivariate normal}$$

$$\underline{Y} = \underline{X} \underline{b} + \underline{\varepsilon} \quad \text{is multivariate normal}$$

why?

$$E(\underline{Y}) = \underline{X} \underline{b}, \quad \text{Cov}(\underline{Y}) = \sigma^2 I$$

$$\underline{Y} \sim N(\underline{X} \underline{b}, \sigma^2 I)$$

$$\hat{\underline{B}} = (X^T X)^{-1} X^T \underline{Y} = [(X^T X)^{-1} X^T] \underline{Y}$$

$$\Rightarrow \hat{\underline{B}} \text{ is MVN}$$

$$E(\hat{\underline{B}}) = [(X^T X)^{-1} X^T] \underline{X} \underline{b} = (X^T X)^{-1} (X^T X) \underline{b} = I \underline{b} = \underline{b}$$

$$\begin{aligned} \text{Cov}(\hat{\underline{B}}) &= [(X^T X)^{-1} X^T] (\sigma^2 I) [X (X^T X)^{-1}] \\ &= \sigma^2 (X^T X)^{-1} (X^T X) (X^T X)^{-1} = \sigma^2 (X^T X)^{-1} \end{aligned}$$

$$\hat{\underline{B}} \sim N(\underline{b}, \sigma^2 (X^T X)^{-1})$$

$$\hat{\underline{Y}} = \underline{X}(\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{Y} = P \underline{Y}$$

$$\hat{\underline{e}} = \underline{Y} - \hat{\underline{Y}} = (\mathbb{I} - P) \underline{Y}$$

$$\hat{\underline{Y}} \sim N\left(\underline{X}\underline{b}, \sigma^2 P\right)$$

$$\hat{\underline{e}} \sim N\left(\underline{0}, \sigma^2 (\mathbb{I} - P)\right)$$

what is  $\text{Cov}(\hat{\underline{Y}}, \hat{\underline{e}}) = E((\hat{\underline{Y}} - \underline{X}\underline{b})\hat{\underline{e}}^T)$  ?

$$= E(\hat{\underline{Y}}\hat{\underline{e}}^T - (\underline{X}\underline{b})\hat{\underline{e}}^T)$$

$$= E(P(\underline{Y} - \underline{X}\underline{b})(\underline{Y} - \underline{X}\underline{b})^T(\mathbb{I} - P)) \quad P = \underline{X}(\underline{X}^T \underline{X})^{-1} \underline{X}^T$$

$$= P E[(\underline{Y} - \underline{X}\underline{b})(\underline{Y} - \underline{X}\underline{b})^T](\mathbb{I} - P)$$

$$= P(\sigma^2 \mathbb{I})(\mathbb{I} - P) = \sigma^2 (P\mathbb{I} - PP) = \sigma^2 (P - P) = \underline{0}$$

residuals are uncorrelated with the fitted values