

# Lecture 7

## Multivariate Linear Regression

Readings: DSA 8020 Lectures 1-4; Zelterman, 2015, Chapter 9

*DSA 8070 Multivariate Analysis*

September 27 - October 1, 2021

Whitney Huang  
Clemson University

## 1 Model and Assumptions

## 2 Parameter Estimation

## 3 Inference and Prediction

## Example: Motor Trend Car Road Tests

```
> head(mtcars)
```

	mpg	cyl	disp	hp	drat	wt	qsec	vs	am	gear	carb
Mazda RX4	21.0	6	160	110	3.90	2.620	16.46	0	1	4	4
Mazda RX4 Wag	21.0	6	160	110	3.90	2.875	17.02	0	1	4	4
Datsun 710	22.8	4	108	93	3.85	2.320	18.61	1	1	4	1
Hornet 4 Drive	21.4	6	258	110	3.08	3.215	19.44	1	0	3	1
Hornet Sportabout	18.7	8	360	175	3.15	3.440	17.02	0	0	3	2
Valiant	18.1	6	225	105	2.76	3.460	20.22	1	0	3	1

Suppose we would like to study the (linear) relationship between mpg, disp, hp, wt (responses) and cyl, am, carb (predictors)

The multiple linear regression model has the form:

$$y_i = \beta_0 + \sum_{j=1}^p \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \dots, n,$$

where

- $y_i$  is the **response** for the  $i$ -th observation
- $x_{ij}$  is the  $j$ -th **predictor** for the  $i$ -th observation
- $\beta_0$  and  $\beta_j$ 's are the **regression intercept** and **slopes** for the response, respectively
- $\varepsilon_i$  is the **error** term for the response of the  $i$ -th observation

# The Multivariate Linear Regression Model: Scalar Form

The multivariate (multiple) linear regression model has the form:

$$y_{ik} = \beta_{0k} + \sum_{j=1}^p \beta_{jk} x_{ij} + \varepsilon_{ik}, \quad i = 1, \dots, n, \quad k = 1, \dots, d,$$

where

- $y_{ik}$  is the  $k$ -th **response** for the  $i$ -th observation
- $x_{ij}$  is the  $j$ -th **predictor** for the  $i$ -th observation
- $\beta_{0k}$  and  $\beta_{jk}$ 's are the **regression intercept** and **slopes** for  $k$ -th response, respectively
- $\varepsilon_{ik}$  is the **error** term for the  $k$ -th response of the  $i$ -th observation

# The Multivariate Linear Regression Model: Assumptions

The assumptions of the model are:

- Relationship between  $\{x_{jk}\}_{j=1}^p$  and  $Y_k$  is **linear** for each  $k \in \{1, \dots, d\}$
- $(\varepsilon_{i1}, \dots, \varepsilon_{id})^T \stackrel{i.i.d.}{\sim} N(\mathbf{0}, \Sigma)$  is an **unobserved random vector**
- $[Y_{ik} | x_{i1}, \dots, x_{ip}] \sim N(\beta_{0k} + \sum_{j=1}^p \beta_{jk} x_{ij}, \sigma_{kk})$  for each  $k \in \{1, \dots, d\}$

# The Multivariate Linear Regression Model: Matrix Form

The multivariate multiple linear regression model has the form

$$Y = XB + E,$$

where

- $Y = [y_1, \dots, y_d]$  is the  $n \times d$  **response matrix**, where  $y_k = (y_{1k}, \dots, y_{nk})^T$  is the  $k$ -th response vector
- $X = [1, x_1, \dots, x_p]$  is the  $n \times (p + 1)$  **design matrix**
- $B = [\beta_1, \dots, \beta_d]$  is the  $(p + 1) \times d$  **matrix of regression coefficients**
- $E = [\epsilon_1, \dots, \epsilon_d]$  is the  $n \times d$  **error matrix**

Matrix form writes the multivariate linear regression model for all  $n \times d$  points simultaneously as

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E}$$

$$\begin{bmatrix} y_{11} & \cdots & y_{1d} \\ y_{21} & \cdots & y_{2d} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nd} \end{bmatrix} = \begin{bmatrix} 1 & \cdots & x_{1p} \\ 1 & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ 1 & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_{01} & \cdots & \beta_{0d} \\ \beta_{11} & \cdots & \beta_{1d} \\ \vdots & \ddots & \vdots \\ \beta_{p1} & \cdots & \beta_{pd} \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} & \cdots & \varepsilon_{1d} \\ \varepsilon_{21} & \cdots & \varepsilon_{2d} \\ \vdots & \ddots & \vdots \\ \varepsilon_{n1} & \cdots & \varepsilon_{nd} \end{bmatrix}$$

Assuming that  $n$  subjects are **independent**, we have

$$\bullet \varepsilon_k \sim N(\mathbf{0}, \sigma_{kk}\mathbf{I}), \quad k \in \{1, \dots, d\}$$

$$\bullet \varepsilon_i \stackrel{i.i.d.}{\sim} N(\mathbf{0}, \Sigma), \quad i = 1, \dots, n$$



The **ordinary least squares** OLS estimate is

$$\operatorname{argmin}_{\mathbf{B} \in \mathbb{R}^{(p+1) \times d}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 = \operatorname{argmin}_{\mathbf{B} \in \mathbb{R}^{(p+1) \times d}} \sum_{i=1}^n \sum_{k=1}^d \left( y_{ik} - \beta_{0k} - \sum_{j=1}^p \beta_{jk} x_{ij} \right)^2,$$

where  $\|\cdot\|$  denotes the Frobenius norm.

- $\text{OLS}(\mathbf{B}) = \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|^2 = \text{tr}(\mathbf{Y}^T \mathbf{Y}) - 2\text{tr}(\mathbf{Y}^T \mathbf{X}\mathbf{B}) + \text{tr}(\mathbf{B}^T \mathbf{X}^T \mathbf{X} \mathbf{B})$
- $\frac{\partial \text{OLS}(\mathbf{B})}{\partial \mathbf{B}} = -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \mathbf{B}$

The OLS estimate has the form

$$\hat{\mathbf{B}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \Rightarrow \hat{\beta}_k = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}_k, \quad k \in \{1, \dots, d\}$$

The expected value of the estimated coefficients is given by

$$\begin{aligned}\mathbb{E}(\hat{B}) &= \mathbb{E}[(X^T X)^{-1} X^T Y] \\ &= (X^T X)^{-1} X^T \mathbb{E}(Y) \\ &= (X^T X)^{-1} X^T X B \\ &= B\end{aligned}$$

$\Rightarrow \hat{B}$  is an unbiased estimator of  $B$

- Fitted values are given by

$$\hat{Y} = X\hat{B},$$

$$\text{i.e., } \hat{y}_{ik} = \hat{\beta}_{0k} + \sum_{j=1}^p \hat{\beta}_{jk} x_{ij}, \quad i = 1, \dots, n, \quad k = 1, \dots, d$$

- Residuals are given by

$$\hat{E} = Y - \hat{Y},$$

$$\text{i.e., } \hat{e}_{ik} = y_{ik} - \hat{y}_{ik}, \quad i = 1, \dots, n, \quad k = 1, \dots, d$$

Just like in univariate linear regression we can write the fitted values as

$$\begin{aligned}\hat{Y} &= X\hat{B} \\ &= X(X^T X)^{-1} X^T Y \\ &= HY,\end{aligned}$$

where  $H = X(X^T X)^{-1} X^T$  is the **hat matrix**

$\Rightarrow H$  projects  $y_k$  onto the column space of  $X$  for  $k \in \{1, \dots, d\}$

## Partitioning the Total Variation

We can partition the total covariation in  $\{\mathbf{y}_i\}_{i=1}^n$  ( $\text{SSCP}_{\text{Tot}}$ ) as

$$\begin{aligned}\text{SSCP}_{\text{tot}} &= \sum_{i=1}^n (\mathbf{y}_i - \bar{\mathbf{y}})^T (\mathbf{y}_i - \bar{\mathbf{y}}) \\&= \sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i + \hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \hat{\mathbf{y}}_i + \hat{\mathbf{y}}_i - \bar{\mathbf{y}})^T \\&= \underbrace{\sum_{i=1}^n (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\hat{\mathbf{y}}_i - \bar{\mathbf{y}})^T}_{\text{SSCP}_{\text{Reg}}} + \underbrace{\sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)(\mathbf{y}_i - \hat{\mathbf{y}}_i)^T}_{\text{SSCP}_{\text{Err}}} \\&\quad + \underbrace{2 \sum_{i=1}^n (\hat{\mathbf{y}}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \hat{\mathbf{y}}_i)^T}_{=0} \\&= \text{SSCP}_{\text{Reg}} + \text{SSCP}_{\text{Err}}\end{aligned}$$

The corresponding **degrees of freedom** are  $d(n-1)$  for  $\text{SSCP}_{\text{Tot}}$ ;  $dp$  for  $\text{SSCP}_{\text{Reg}}$ ; and  $d(n-p-1)$  for  $\text{SSCP}_{\text{Err}}$

The estimated error variance is

$$\begin{aligned}\hat{\Sigma} &= \frac{\sum_{i=1}^n (\mathbf{y}_i - \hat{\mathbf{y}}_i)(\mathbf{y}_i - \hat{\mathbf{y}}_i)^T}{n - p - 1} \\ &= \frac{\text{SSCP}_{Err}}{n - p - 1}\end{aligned}$$

- $\hat{\Sigma}$  is an unbiased estimate of  $\Sigma$
- The estimate  $\hat{\Sigma}$  is the mean  $\text{SSCP}_{Err}$

## Sampling Distributions of $\hat{B}$ , $\hat{Y}$ , and $\hat{E}$

We would need to figure out the **sampling distributions** of estimator and predictor in order to drawn inference

Given the model assumptions, we have

$$\text{vec}(\hat{B}) \sim N(\text{vec}(B), \Sigma \otimes (X^T X)^{-1})$$

$$\text{vec}(\hat{Y}) \sim N(\text{vec}(XB), \Sigma \otimes H)$$

$$\text{vec}(\hat{E}) \sim N(0, \Sigma \otimes (I - H)),$$

where  $\text{vec}(\cdot)$  is the vectorization operator and  $\otimes$  is the Kronecker product

## Inference about Multiple $\hat{\beta}_{jk}$

Assume that  $q < p$  and want to test if a reduced model is sufficient:

$$H_0 : B_2 = \mathbf{0}_{p-q} \times d, \quad \text{versus} \quad H_a : B_2 \neq \mathbf{0}_{p-q} \times d,$$

where

$$B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix}$$

is the partitioned of the coefficient vector

We can compare the  $\text{SSCP}_{Err}$  for the **full model**:

$$y_{ik} = \beta_{0k} + \sum_{j=1}^p \beta_{jk} x_{ij} + \varepsilon_{ik}, \quad k = 1, \dots, d$$

and the **reduced model**:

$$y_{ik} = \beta_{0k} + \sum_{j=1}^q \beta_{jk} x_{ij} + \varepsilon_{ik}, \quad k = 1, \dots, d$$



## Some Test Statistics

Let  $\tilde{E} = n\tilde{\Sigma}$  denote the  $\text{SSCP}_{Err}$  matrix from the **full model**,  
and let  $\tilde{H} = n(\tilde{\Sigma}_1 - \tilde{\Sigma})$  denote the hypothesis  $\text{SSCP}_{Err}$  matrix  
Some test statistics for

$$H_0 : B_2 = \mathbf{0}_{p-q} \times d, \quad \text{versus} \quad H_a : B_2 \neq \mathbf{0}_{p-q} \times d :$$

- Wilks Lambda

$$\Lambda^* = \frac{|\tilde{E}|}{|\tilde{H} + \tilde{E}|}$$

Reject  $H_0$  if  $\Lambda^*$  is “small”

- Hotelling-Lawley Trace

$$T_0^2 = \text{tr}(\tilde{H}\tilde{E}^{-1})$$

Reject  $H_0$  if  $T_0^2$  is “large”

- Pillai Trace

$$V = \text{tr}(\tilde{H}(\tilde{H} + \tilde{E})^{-1})$$

Reject  $H_0$  if  $V$  is “large”

We would like to estimate the **expected value of the response** for a given predictor  $\mathbf{x}_h = (1, x_{h1}, \dots, x_{hp})$ .

Note that we have

$$\hat{\mathbf{y}}_h \sim N(\mathbf{B}^T \mathbf{x}_h, \mathbf{x}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_h \Sigma)$$

We can exploit the duality between interval estimation and hypothesis testing. That is, we can test

$$H_0 : \mathbb{E}(\mathbf{y}_h) = \mathbf{y}_h^* \text{ versus } H_a : \mathbb{E}(\mathbf{y}_h) \neq \mathbf{y}_h^*$$

The  $100(1 - \alpha)\%$  confidence interval is the collection of  $\mathbf{y}_h^*$  values that fail to reject  $H_0$  at  $\alpha$  level

### Test statistics:

$$T^2 = \left( \frac{\hat{B}^T \mathbf{x}_h - B^T \mathbf{x}_h}{\sqrt{\mathbf{x}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_h}} \right)^T \hat{\Sigma}^{-1} \left( \frac{\hat{B}^T \mathbf{x}_h - B^T \mathbf{x}_h}{\sqrt{\mathbf{x}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_h}} \right)$$
$$\stackrel{H_0}{\sim} \frac{d(n-p-1)}{n-p-d} F_{d,n-p-d}$$

Therefore, the  $100(1 - \alpha)\%$  simultaneous **confidence interval** for  $y_{hk}$  is

$$\hat{y}_{hk} \pm \sqrt{\frac{d(n-p-1)}{n-p-d} F_{d,n-p-d}} \sqrt{\mathbf{x}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_h \hat{\sigma}_{kk}},$$

$$k \in \{1, \dots, d\}$$

## Predicting New Observations

Here we want to predict the **observed value of response** for a given predictor

- **Note:** interested in actual  $\hat{y}_h$  instead of  $\mathbb{E}(\hat{y}_h)$
- Given  $\mathbf{x}_h = (1, x_{h1}, \dots, x_{hp})$ , the fitted value is still  $\hat{y}_h = \hat{\mathbf{B}}^T \mathbf{x}_h$

We can exploit the duality between interval estimation and hypothesis testing. That is, we can test

$$H_0 : \mathbf{y}_h = \mathbf{y}_h^* \text{ versus } H_a : \mathbf{y}_h \neq \mathbf{y}_h^*$$

The  $100(1 - \alpha)\%$  prediction interval is the collection of  $\mathbf{y}_h^*$  values that fail to reject  $H_0$  at  $\alpha$  level

### Test statistics:

$$T^2 = \left( \frac{\hat{B}^T \mathbf{x}_h - B^T \mathbf{x}_h}{\sqrt{1 + \mathbf{x}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_h}} \right)^T \hat{\Sigma}^{-1} \left( \frac{\hat{B}^T \mathbf{x}_h - B^T \mathbf{x}_h}{\sqrt{1 + \mathbf{x}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_h}} \right)$$
$$\stackrel{H_0}{\sim} \frac{d(n-p-1)}{n-p-d} F_{d,n-p-d}$$

Therefore, the  $100(1 - \alpha)\%$  simultaneous **prediction interval** for  $y_{hk}$  is

$$\hat{y}_{hk} \pm \sqrt{\frac{d(n-p-1)}{n-p-d} F_{d,n-p-d}} \sqrt{(1 + \mathbf{x}_h^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_h)} \hat{\sigma}_{kk},$$

$$k \in \{1, \dots, d\}$$