Big Data Analysis
(MA60306)

Bibhas Adhikari

Spring 2022-23, IIT Kharagpur

Lecture 4
January 11, 2023

Statistical learning of data

Plots for two quantitative variables

Example of scatterplot of height and weight data of the *nutri* data

Statistical learning of data

Plots for two quantitative variables

Example of scatterplot of height and weight data of the *nutri* data

Example of scatterplot of the *Advertising* data

Statistical learning of data

Statistical Learning Mathematical analysis of the data - interpret the model and quantify the uncertainty in the data

Statistical learning of data

Statistical Learning Mathematical analysis of the data - interpret the model and quantify the uncertainty in the data

Machine Learning - the emphasis is on making predictions using large-scale data

Statistical learning of data

Statistical Learning Mathematical analysis of the data - interpret the model and quantify the uncertainty in the data

Machine Learning - the emphasis is on making predictions using large-scale data

Modeling data

- ▷ accurately predict some future quantity of interest, given some observed data - prediction - exp. direct-marketing campaign

Statistical learning of data

Statistical Learning Mathematical analysis of the data - interpret the model and quantify the uncertainty in the data

Machine Learning - the emphasis is on making predictions using large-scale data

Modeling data

- ▷ accurately predict some future quantity of interest, given some observed data - prediction - exp. direct-marketing campaign
- ▷ discover patterns in the data - inference - exp. advertising

Statistical learning of data

Statistical Learning Mathematical analysis of the data - interpret the model and quantify the uncertainty in the data

Machine Learning - the emphasis is on making predictions using large-scale data

Modeling data

- ▷ accurately predict some future quantity of interest, given some observed data - prediction - exp. direct-marketing campaign
- ▷ discover patterns in the data - inference - exp. advertising
- ▷ both prediction and inference - exp. real estate data

Statistical learning of data

Statistical Learning Mathematical analysis of the data - interpret the model and quantify the uncertainty in the data

Machine Learning - the emphasis is on making predictions using large-scale data

Modeling data

- ▷ accurately predict some future quantity of interest, given some observed data - prediction - exp. direct-marketing campaign
- ▷ discover patterns in the data - inference - exp. advertising
- ▷ both prediction and inference - exp. real estate data

Tools:

- ▷ Functional approximation - how one data variable depends on another data variable

Statistical learning of data

Statistical Learning Mathematical analysis of the data - interpret the model and quantify the uncertainty in the data

Machine Learning - the emphasis is on making predictions using large-scale data

Modeling data

- ▷ accurately predict some future quantity of interest, given some observed data - prediction - exp. direct-marketing campaign
- ▷ discover patterns in the data - inference - exp. advertising
- ▷ both prediction and inference - exp. real estate data

Tools:

- ▷ Functional approximation - how one data variable depends on another data variable
- ▷ Optimization - best possible model in a class of models

Statistical learning

Prediction function g: Let $\boldsymbol{x}$ be a feature/input vector. Then one of the fundamental problems in machine learning is to predict an output response variable y. The prediction for y is based on a function $g(\boldsymbol{x})$ such that g encompasses all the information about the relationship between the variables $\boldsymbol{x}$ and y. - exp. digitized signature

Statistical learning

Prediction function g: Let $\boldsymbol{x}$ be a feature/input vector. Then one of the fundamental problems in machine learning is to predict an output response variable y. The prediction for y is based on a function $g(\boldsymbol{x})$ such that g encompasses all the information about the relationship between the variables $\boldsymbol{x}$ and y. - exp. digitized signature

Regression In regression problems, the response variable y can take any real value

Statistical learning

Prediction function g: Let $\boldsymbol{x}$ be a feature/input vector. Then one of the fundamental problems in machine learning is to predict an output response variable y. The prediction for y is based on a function $g(\boldsymbol{x})$ such that g encompasses all the information about the relationship between the variables $\boldsymbol{x}$ and y. - exp. digitized signature

Regression In regression problems, the response variable y can take any real value

Classification If y lies in a finite set then predicting $y \in \{0, \ldots, k-1\}$ is same as classifying the input into one of the k categories

Statistical learning

Prediction function g: Let $\boldsymbol{x}$ be a feature/input vector. Then one of the fundamental problems in machine learning is to predict an output response variable y. The prediction for y is based on a function $g(\boldsymbol{x})$ such that g encompasses all the information about the relationship between the variables $\boldsymbol{x}$ and y. - exp. digitized signature

Regression In regression problems, the response variable y can take any real value

Classification If y lies in a finite set then predicting $y \in \{0, \ldots, k-1\}$ is same as classifying the input into one of the k categories

Loss function Let $\widehat{y}$ denote a prediction corresponding to a given response. Then we do we measure the accuracy of the prediction?
We write $Loss(y, \widehat{y})$ to denote the measure

- regression: $Loss(y, \widehat{y}) = (y - \widehat{y})^2$
- classification: $Loss(y, \widehat{y}) = \mathbb{1}\{y \neq \widehat{y}\}$
- in general, $Loss(y, \widehat{y}) = \|y - \widehat{y}\|^2$

Supervised learning technique

The word 'regression' - In 1889, Francis Galton observed that the heights of the adult offsprings have more average heights/intelligence than their parents - used the term 'regression' to indicate "return to mediocrity"

Supervised learning technique

The word 'regression' - In 1889, Francis Galton observed that the heights of the adult offsprings have more average heights/intelligence than their parents - used the term 'regression' to indicate "return to mediocrity"

Let $\mathbf{x} = [x_1, \ldots, x_p]^T$ be an input with p features and we need to predict a quantitative response/output y via a function $g(\mathbf{x})$, x_i could be discrete or continuous

Supervised learning technique

The word 'regression' - In 1889, Francis Galton observed that the heights of the adult offsprings have more average heights/intelligence than their parents - used the term 'regression' to indicate "return to mediocrity"

Let $\mathbf{x} = [x_1, \ldots, x_p]^T$ be an input with p features and we need to predict a quantitative response/output y via a function $g(\mathbf{x})$, x_i could be discrete or continuous

Example

- predict birth weight of a baby from the weight of the mother, her socio-economic status, her smoking habits (sometimes known as explanatory variables)
- Consider the advertising data for advertising budgets for the product in each of 200 markets for three different media: TV, radio, and newspaper. Can we develop a model to predict sales based on the three media budgets?

Regression

Defining it as a statistical problem

- It is unlikely that the prediction function g will make accurate prediction

Regression

Defining it as a statistical problem

- It is unlikely that the prediction function g will make accurate prediction
- in reality, even for the same input $\mathbf{x}$, the output y may be different

Regression

Defining it as a statistical problem

- ▷ It is unlikely that the prediction function g will make accurate prediction
- ▷ in reality, even for the same input $\mathbf{x}$, the output y may be different
- ▷ we adopt a probabilistic approach: we assume that each pair $(\mathbf{x}, y)$ is an instance of a random variable $(\mathbf{X}, Y)$ that has the joint pdf $f(\mathbf{x}, y)$

Regression

Defining it as a statistical problem

- It is unlikely that the prediction function g will make accurate prediction
- in reality, even for the same input $\mathbf{x}$, the output y may be different
- we adopt a probabilistic approach: we assume that each pair $(\mathbf{x}, y)$ is an instance of a random variable $(\mathbf{X}, Y)$ that has the joint pdf $f(\mathbf{x}, y)$
- Then the performance of the prediction can be measured via the expected loss, known as *risk*:

$$l(g) = \mathbb{E}[\text{Loss}(Y, g(\mathbf{X}))]$$

Regression

Defining it as a statistical problem

- It is unlikely that the prediction function g will make accurate prediction
- in reality, even for the same input $\mathbf{x}$, the output y may be different
- we adopt a probabilistic approach: we assume that each pair $(\mathbf{x}, y)$ is an instance of a random variable $(\mathbf{X}, Y)$ that has the joint pdf $f(\mathbf{x}, y)$
- Then the performance of the prediction can be measured via the expected loss, known as *risk*:

$$l(g) = \mathbb{E}[\text{Loss}(Y, g(\mathbf{X}))]$$

- In the classification case, the risk equals the probability of incorrect classification:

$$l(g) = \mathbb{P}[Y \neq g(\mathbf{X})]$$

Regression

Defining it as a statistical problem

- It is unlikely that the prediction function g will make accurate prediction
- in reality, even for the same input $\mathbf{x}$, the output y may be different
- we adopt a probabilistic approach: we assume that each pair $(\mathbf{x}, y)$ is an instance of a random variable $(\mathbf{X}, Y)$ that has the joint pdf $f(\mathbf{x}, y)$
- Then the performance of the prediction can be measured via the expected loss, known as *risk*:

$$l(g) = \mathbb{E}[\text{Loss}(Y, g(\mathbf{X}))]$$

- In the classification case, the risk equals the probability of incorrect classification:

$$l(g) = \mathbb{P}[Y \neq g(\mathbf{X})]$$

In this case, g is called the classifier

Statistical learning

Given a rv $(\mathbf{X}, Y)$ and a loss function, in principle, we determine

$$g^* = \arg\min_g \mathbb{E}[\text{Loss}(Y, g(\mathbf{X}))]$$

which yields the smallest risk $l^* = l(g^*)$

Statistical learning

Given a rv $(\mathbf{X}, Y)$ and a loss function, in principle, we determine

$$g^* = \arg\min_g \mathbb{E}[\text{Loss}(Y, g(\mathbf{X}))]$$

which yields the smallest risk $l^* = l(g^*)$

Theorem If $\text{Loss}(y, \widehat{y}) = (y - \widehat{y})^2$ then the optimal prediction function g^* is equal to the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$:

$$g^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$$

Statistical learning

Given a rv $(\mathbf{X}, Y)$ and a loss function, in principle, we determine

$$g^* = \arg\min_g \mathbb{E}[\text{Loss}(Y, g(\mathbf{X}))]$$

which yields the smallest risk $l^* = l(g^*)$

Theorem If $\text{Loss}(y, \widehat{y}) = (y - \widehat{y})^2$ then the optimal prediction function g^* is equal to the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$:

$$g^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$$

Proof

$$\mathbb{E}[(Y - g(\mathbf{X}))^2] = \mathbb{E}[(Y - g^*(\mathbf{X}) + g^*(\mathbf{X}) - g(\mathbf{X}))^2]$$

Statistical learning

Given a rv $(\mathbf{X}, Y)$ and a loss function, in principle, we determine

$$g^* = \arg\min_g \mathbb{E}[\text{Loss}(Y, g(\mathbf{X}))]$$

which yields the smallest risk $l^* = l(g^*)$

Theorem If $\text{Loss}(y, \widehat{y}) = (y - \widehat{y})^2$ then the optimal prediction function g^* is equal to the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$:

$$g^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$$

Proof

$$\begin{aligned} & \mathbb{E}[(Y - g(\mathbf{X}))^2] = \mathbb{E}[(Y - g^*(\mathbf{X}) + g^*(\mathbf{X}) - g(\mathbf{X}))^2] \\ = \quad & \mathbb{E}[(Y - g^*(\mathbf{X}))^2] + 2\mathbb{E}[(Y - g^*(\mathbf{X}))(g^*(\mathbf{X}) - g(\mathbf{X}))] \\ & + \mathbb{E}[(g^*(\mathbf{X}) - g(\mathbf{X}))^2] \end{aligned}$$

Statistical learning

Given a rv $(\mathbf{X}, Y)$ and a loss function, in principle, we determine

$$g^* = \arg\min_g \mathbb{E}[\text{Loss}(Y, g(\mathbf{X}))]$$

which yields the smallest risk $l^* = l(g^*)$

Theorem If $\text{Loss}(y, \widehat{y}) = (y - \widehat{y})^2$ then the optimal prediction function g^* is equal to the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$:

$$g^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$$

Proof

$$\begin{aligned}
& \mathbb{E}[(Y - g(\mathbf{X}))^2] = \mathbb{E}[(Y - g^*(\mathbf{X}) + g^*(\mathbf{X}) - g(\mathbf{X}))^2] \\
= \; & \mathbb{E}[(Y - g^*(\mathbf{X}))^2] + 2\mathbb{E}[(Y - g^*(\mathbf{X}))(g^*(\mathbf{X}) - g(\mathbf{X}))] \\
& + \mathbb{E}[(g^*(\mathbf{X}) - g(\mathbf{X}))^2] \\
\geq \; & \mathbb{E}[(Y - g^*(\mathbf{X}))^2] + 2\mathbb{E}[(Y - g^*(\mathbf{X}))(g^*(\mathbf{X}) - g(\mathbf{X}))]
\end{aligned}$$

Statistical learning

Given a rv $(\mathbf{X}, Y)$ and a loss function, in principle, we determine

$$g^* = \arg\min_g \mathbb{E}[\text{Loss}(Y, g(\mathbf{X}))]$$

which yields the smallest risk $l^* = l(g^*)$

Theorem If $\text{Loss}(y, \widehat{y}) = (y - \widehat{y})^2$ then the optimal prediction function g^* is equal to the conditional expectation of Y given $\mathbf{X} = \mathbf{x}$:

$$g^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$$

Proof

$$\begin{aligned}
& \mathbb{E}[(Y - g(\mathbf{X}))^2] = \mathbb{E}[(Y - g^*(\mathbf{X}) + g^*(\mathbf{X}) - g(\mathbf{X}))^2] \\
= \; & \mathbb{E}[(Y - g^*(\mathbf{X}))^2] + 2\mathbb{E}[(Y - g^*(\mathbf{X}))(g^*(\mathbf{X}) - g(\mathbf{X}))] \\
& + \mathbb{E}[(g^*(\mathbf{X}) - g(\mathbf{X}))^2] \\
\geq \; & \mathbb{E}[(Y - g^*(\mathbf{X}))^2] + 2\mathbb{E}[(Y - g^*(\mathbf{X}))(g^*(\mathbf{X}) - g(\mathbf{X}))] \\
= \; & \mathbb{E}[(Y - g^*(\mathbf{X}))^2] + 2\mathbb{E}\{(g^*(\mathbf{X}) - g(\mathbf{X}))\mathbb{E}[Y - g^*(\mathbf{X})|\mathbf{X}]\}
\end{aligned}$$

Statistical learning

Now

$$\mathbb{E}[Y - g^*(\mathbf{X})|\mathbf{X}] = 0$$

and hence

$$\mathbb{E}[(Y - g(\mathbf{X}))^2] \geq \mathbb{E}[(Y - g^*(\mathbf{X}))^2]$$

which implies g^* yields the smallest squared-error.

Statistical learning

Now

$$\mathbb{E}[Y - g^*(\mathbf{X})|\mathbf{X}] = 0$$

and hence

$$\mathbb{E}[(Y - g(\mathbf{X}))^2] \geq \mathbb{E}[(Y - g^*(\mathbf{X}))^2]$$

which implies g^* yields the smallest squared-error.

Consequence

- The conditional $\mathbf{X} = \mathbf{x}$, the random response Y can be written as

$$Y = g^*(\mathbf{x}) + \epsilon(\mathbf{x})$$

where $\epsilon(\mathbf{x})$ can be thought of as a random deviation of the response from its conditional mean at $\mathbf{x}$.

Statistical learning

Now

$$\mathbb{E}[Y - g^*(\mathbf{X})|\mathbf{X}] = 0$$

and hence

$$\mathbb{E}[(Y - g(\mathbf{X}))^2] \geq \mathbb{E}[(Y - g^*(\mathbf{X}))^2]$$

which implies g^* yields the smallest squared-error.

Consequence

- The conditional $\mathbf{X} = \mathbf{x}$, the random response Y can be written as

$$Y = g^*(\mathbf{x}) + \epsilon(\mathbf{x})$$

  where $\epsilon(\mathbf{x})$ can be thought of as a random deviation of the response from its conditional mean at $\mathbf{x}$.
- $\mathbb{E}[\epsilon(\mathbf{x})] = 0$

Statistical learning

Now

$$\mathbb{E}[Y - g^*(\mathbf{X})|\mathbf{X}] = 0$$

and hence

$$\mathbb{E}[(Y - g(\mathbf{X}))^2] \geq \mathbb{E}[(Y - g^*(\mathbf{X}))^2]$$

which implies g^* yields the smallest squared-error.

Consequence

- The conditional $\mathbf{X} = \mathbf{x}$, the random response Y can be written as

$$Y = g^*(\mathbf{x}) + \epsilon(\mathbf{x})$$

  where $\epsilon(\mathbf{x})$ can be thought of as a random deviation of the response from its conditional mean at $\mathbf{x}$.
- $\mathbb{E}[\epsilon(\mathbf{x})] = 0$
- $\mathbb{V}\text{ar}[\epsilon(\mathbf{x})] = \nu^2(\mathbf{x})$ for some function $\nu(\mathbf{x})$

Statistical learning

Learning

- The optimal prediction function g^* depends on the joint distribution of $(\mathbf{X}, Y)$, which is not available in practice

Statistical learning

Learning

- The optimal prediction function g^* depends on the joint distribution of $(\mathbf{X}, Y)$, which is not available in practice
- available : finite number of (usually) independent realizations from the joint density $f(\mathbf{x}, y)$

Statistical learning

Learning

- The optimal prediction function g^* depends on the joint distribution of $(\mathbf{X}, Y)$, which is not available in practice
- available : finite number of (usually) independent realizations from the joint density $f(\mathbf{x}, y)$
- Let $\mathcal{T} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ be a sample, and call it the *training set* with n examples

Statistical learning

Learning

- The optimal prediction function g^* depends on the joint distribution of $(\mathbf{X}, Y)$, which is not available in practice
- available : finite number of (usually) independent realizations from the joint density $f(\mathbf{x}, y)$
- Let $\mathcal{T} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ be a sample, and call it the *training set* with n examples
- a given sample point: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$

Statistical learning

Learning

- The optimal prediction function g^* depends on the joint distribution of $(\mathbf{X}, Y)$, which is not available in practice
- available : finite number of (usually) independent realizations from the joint density $f(\mathbf{x}, y)$
- Let $\mathcal{T} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ be a sample, and call it the *training set* with n examples
- a given sample point: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- Goal: 'learn' the unknown function g^* using the n examples from the training set $\mathcal{T}$

Statistical learning

Learning

- The optimal prediction function g^* depends on the joint distribution of $(\mathbf{X}, Y)$, which is not available in practice
- available : finite number of (usually) independent realizations from the joint density $f(\mathbf{x}, y)$
- Let $\mathcal{T} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ be a sample, and call it the *training set* with n examples
- a given sample point: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- Goal: 'learn' the unknown function g^* using the n examples from the training set $\mathcal{T}$
- Denote $g_{\mathcal{T}}$ as an approximation for g^* that can be constructed from $\mathcal{T}$

Statistical learning

Learning

- The optimal prediction function g^* depends on the joint distribution of $(\mathbf{X}, Y)$, which is not available in practice
- available : finite number of (usually) independent realizations from the joint density $f(\mathbf{x}, y)$
- Let $\mathcal{T} = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ be a sample, and call it the *training set* with n examples
- a given sample point: $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$
- Goal: 'learn' the unknown function g^* using the n examples from the training set $\mathcal{T}$
- Denote $g_{\mathcal{T}}$ as an approximation for g^* that can be constructed from $\mathcal{T}$Obviously, $g_{\mathcal{T}}$ is a random function and a particular outcome is g_{τ}

Regression models

Linear regression model The response Y depends on a d-dimensional explanatory vector $\mathbf{x} = [x_1, \ldots, x_d]^T$ via the linear relationship

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d + \epsilon = \boldsymbol{\beta}^T \mathbf{x} + \epsilon = \mathbf{x}^T \boldsymbol{\beta} + \epsilon$$

Regression models

Linear regression model The response Y depends on a d-dimensional explanatory vector $\mathbf{x} = [x_1, \ldots, x_d]^T$ via the linear relationship

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d + \epsilon = \boldsymbol{\beta}^T \mathbf{x} + \epsilon = \mathbf{x}^T \boldsymbol{\beta} + \epsilon := g(\mathbf{x}) + \epsilon,$$

where $\mathbb{E}[\epsilon] = 0$ and $\mathbb{V}\text{ar}[\epsilon] = \sigma^2$.

Regression models

Linear regression model The response Y depends on a d-dimensional explanatory vector $\mathbf{x} = [x_1, \ldots, x_d]^T$ via the linear relationship

$$Y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d + \epsilon = \boldsymbol{\beta}^T \mathbf{x} + \epsilon = \mathbf{x}^T \boldsymbol{\beta} + \epsilon := g(\mathbf{x}) + \epsilon,$$

where $\mathbb{E}[\epsilon] = 0$ and $\mathbb{V}\text{ar}[\epsilon] = \sigma^2$.

The model for a training set $\mathcal{T} = \{(\mathbf{x}_1, Y_1), \ldots, (\mathbf{x}_n, Y_n)\}$ is given by

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

where $\boldsymbol{\epsilon} = [\epsilon_1, \ldots, \epsilon_n]^T$ is a vector of iid copies of ϵ, and $\mathbf{X}$ is the *model matrix* or *regression matrix* given by

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1d} \\ 1 & x_{21} & x_{22} & \ldots & x_{2d} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{nd} \end{bmatrix} = \begin{bmatrix} 1 & \mathbf{x}_1^T \\ 1 & \mathbf{x}_2^T \\ \vdots & \vdots \\ 1 & \mathbf{x}_n^T \end{bmatrix}$$

Regression models

- In most cases x_{ij}s are generally chosen so that the columns of $\mathbf{X}$ are linearly independent i.e rank of $\mathbf{X}$ is $d + 1$

Regression models

- In most cases x_{ij}s are generally chosen so that the columns of $\mathbf{X}$ are linearly independent i.e rank of $\mathbf{X}$ is $d+1$
- However, in some experimental design situations $x_{ij} \in \{0, 1\}$ and columns of $\mathbf{X}$ are linearly dependent, and $\mathbf{X}$ is called the *design matrix*

Regression models

- In most cases x_{ij}s are generally chosen so that the columns of $\mathbf{X}$ are linearly independent i.e rank of $\mathbf{X}$ is $d+1$
- However, in some experimental design situations $x_{ij} \in \{0, 1\}$ and columns of $\mathbf{X}$ are linearly dependent, and $\mathbf{X}$ is called the *design matrix*
- Note that $\mathbf{x}_j$ are known and called the explanatory variable and Y as the response variable

Regression models

- In most cases x_{ij}s are generally chosen so that the columns of $\mathbf{X}$ are linearly independent i.e rank of $\mathbf{X}$ is $d+1$
- However, in some experimental design situations $x_{ij} \in \{0,1\}$ and columns of $\mathbf{X}$ are linearly dependent, and $\mathbf{X}$ is called the *design matrix*
- Note that $\mathbf{x}_j$ are known and called the explanatory variable and Y as the response variable

Question How do we estimate β?

Regression models

- In most cases x_{ij}s are generally chosen so that the columns of $\mathbf{X}$ are linearly independent i.e rank of $\mathbf{X}$ is $d+1$
- However, in some experimental design situations $x_{ij} \in \{0,1\}$ and columns of $\mathbf{X}$ are linearly dependent, and $\mathbf{X}$ is called the *design matrix*
- Note that $\mathbf{x}_j$ are known and called the explanatory variable and Y as the response variable

Question How do we estimate β?

Note that the first column of $\mathbf{X}$ need not be 1.

Regression models

- In most cases x_{ij}s are generally chosen so that the columns of $\mathbf{X}$ are linearly independent i.e rank of $\mathbf{X}$ is $d+1$
- However, in some experimental design situations $x_{ij} \in \{0,1\}$ and columns of $\mathbf{X}$ are linearly dependent, and $\mathbf{X}$ is called the *design matrix*
- Note that $\mathbf{x}_j$ are known and called the explanatory variable and Y as the response variable

Question How do we estimate β?

Note that the first column of $\mathbf{X}$ need not be 1.

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} x_{10} & x_{11} & x_{12} & \cdots & x_{1,d-1} \\ x_{20} & x_{21} & x_{22} & \cdots & x_{2,d-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{n0} & x_{n1} & x_{n2} & \cdots & x_{n,d-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{d-1} \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$