MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

Linear Regression Models

Segment 1 – Simple Linear Regression Model

Topic 1 – Data Generation Process: Sample and Population

Sudarsan N.S. Acharya (sudarsan.acharya@manipal.edu)

Topics

MANIPAL
ACADEMY *of* HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

1. Questions from Data
2. Output & Input Variables in Linear Regression
3. Population & Sample
4. Population Model
5. A Linear Population Model

Questions from Data

Questions from Data

- Output and Input variables.

Questions from Data

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Output and Input variables.
- Is there a relationship between output and Input variables?

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship?

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship? *Accurate prediction.*

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship? *Accurate prediction.*
- Can we quantify the effect of the relationship?

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship? *Accurate prediction.*
- Can we quantify the effect of the relationship?
- Is the relationship approximately linear?

Questions from Data

- Output and Input variables.
- Is there a relationship between output and Input variables?
- How strong is the relationship? *Accurate prediction.*
- Can we quantify the effect of the relationship?
- Is the relationship approximately linear? *Linear Regression.*

Output & Input Variables in Linear Regression

Output & Input Variables in Linear Regression

- Output variables have other names:

Output & Input Variables in Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Output variables have other names: dependent variables,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.

Output & Input Variables in Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples:

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg),

Output & Input Variables in Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.

Output & Input Variables in Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names:

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.

Output & Input Variables in Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.

Output & Input Variables in Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.
- Examples:

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.
- Examples: advertisement budget in Dollars,

Output & Input Variables in Linear Regression

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.
- Examples: advertisement budget in Dollars, horse power of vehicle,

Output & Input Variables in Linear Regression

- Output variables have other names: dependent variables, outcomes, response variables, target variables.
- Output variables in linear regression should be *continuous*.
- Examples: sales in number of units, fuel economy in Miles per gallon (mpg), Body mass index (BMI) etc.
- Input variables also have other names: independent variables, features, predictors, covariates.
- Input variables in linear regression can be a mix of *continuous* and *categorical* variables.
- Examples: advertisement budget in Dollars, horse power of vehicle, individual's height, weight, education level, gender etc.

Population & Sample

Population & Sample

- In data science, it is important to distinguish between population and sample.

Population & Sample

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- In data science, it is important to distinguish between population and sample.
- Example of a population parameter:

Population & Sample

- In data science, it is important to distinguish between population and sample.
- Example of a population parameter: the average height of all biological females in a city.

Population & Sample

- In data science, it is important to distinguish between population and sample.
- Example of a population parameter: the average height of all biological females in a city.
- Example of a sample statistic:

Population & Sample

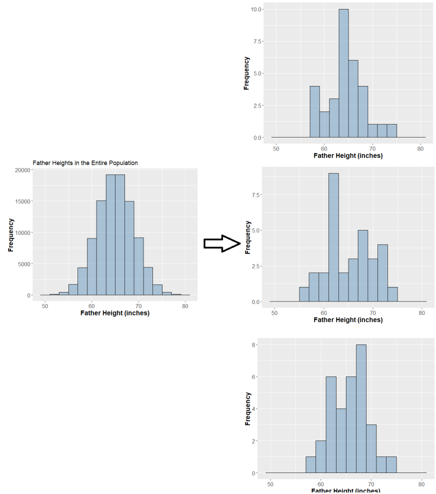
- In data science, it is important to distinguish between population and sample.
- Example of a population parameter: the average height of all biological females in a city.
- Example of a sample statistic: the average height of n randomly chosen biological females in a city.

Population & Sample

- In data science, it is important to distinguish between population and sample.
- Example of a population parameter: the average height of all biological females in a city.
- Example of a sample statistic: the average height of n randomly chosen biological females in a city.
- Note that sample statistic (or just statistic) is a *random variable*.

Population & Sample - Example with Sample Size = 32

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)



Population Model

Population Model

- We can use a *probabilistic model* for understanding the population.

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example:

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.

Population Model

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear:

Population Model

MANIPAL ACADEMY of HIGHER EDUCATION (Institution of Eminence Deemed to be University)

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear: $Y = f(X) + \epsilon$ for an unknown nonlinear function f,

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear: $Y = f(X) + \epsilon$ for an unknown nonlinear function f, where ϵ is a *random error term*.

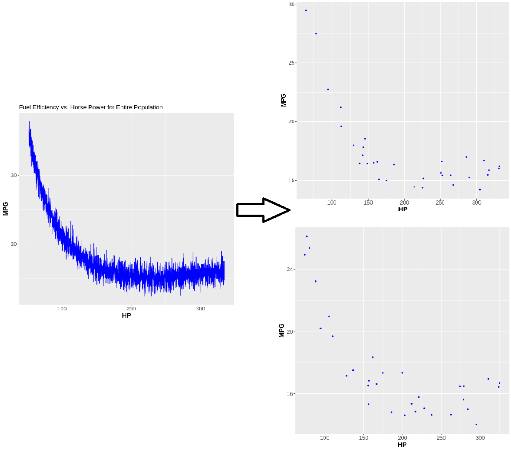
Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear: $Y = f(X) + \epsilon$ for an unknown nonlinear function f, where ϵ is a *random error term*.
- Example of a population model for mpg and hp:

Population Model

- We can use a *probabilistic model* for understanding the population.
- Example: Let fuel efficiency (mpg) be the response variable and horse power (hp) be the only input variable.
- For a random car from the population, let Y represent the mpg and X represent the hp.
- There is a probability distribution of X in the population.
- Y has a conditional probability distribution given X.
- Population models are typically nonlinear: $Y = f(X) + \epsilon$ for an unknown nonlinear function f, where ϵ is a *random error term*.
- Example of a population model for mpg and hp: $Y = \frac{1.8}{X} - 0.03X + \epsilon$.

Population & Sample - Another Example with Sample Size = 32



A Linear Population Model

A Linear Population Model

MANIPAL ACADEMY of HIGHER EDUCATION
(Institution of Eminence Deemed to be University)

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches:

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches:

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches: $\boxed{Y \mid (X = x)} \sim N(\mu = 42 + 0.4x, \sigma^2 = 9)$.

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches: $\boxed{Y \mid (X = x)} \sim N(\mu = 42 + 0.4x, \sigma^2 = 9)$.
- The population model for Y as a function of X is a linear one:

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches: $\boxed{Y \mid (X = x)} \sim N(\mu = 42 + 0.4x, \sigma^2 = 9)$.
- The population model for Y as a function of X is a linear one: $Y = 42 + 0.4X + \epsilon,$

A Linear Population Model

- For a random father-son pair in a population, let Y represent the son's height and X represent the father's height.
- Suppose that in the population, father's heights are normally distributed with mean 65 inches and standard deviation 4 inches: $X \sim N(\mu = 65, \sigma^2 = 16)$.
- Given the father's height $X = x$, suppose the son's height Y is also normally distributed with mean $42 + 0.4 \times x$ and standard deviation 3 inches: $\boxed{Y \mid (X = x)} \sim N(\mu = 42 + 0.4x, \sigma^2 = 9)$.
- The population model for Y as a function of X is a linear one: $Y = 42 + 0.4X + \epsilon$, where $\epsilon \sim N(\mu = 0, \sigma^2 = 9)$.