MANIPAL
ACADEMY of HIGHER EDUCATION
(Deemed to be University under Section 3 of the UGC Act, 1956)

Handling Missing Values in Regression

Kodipaka Chathurya (202402004)
Aarya K (202402006)
Greeshma John (202402013)

Semester II, M.Sc. Biostatistics

Under the Guidance of

Mr. Himanshu Pokhriyal
Assistant Professor
Department of Data Science
PSPH, MAHE, Manipal

Outline

INTRODUCTION

Introduction

Why missing values arise

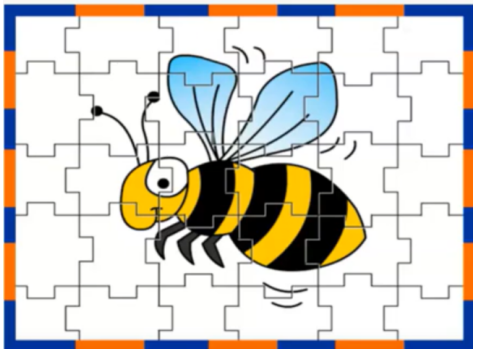
- Global or unit non response
- Item non response
- Errors in data collection or processing
- Missing data may be built into the design of a study
- Missing data should be distinguished from data that are conditionally undefined

TYPES OF MISSING DATA

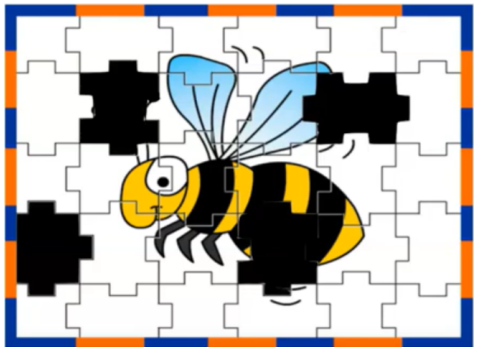
Types of missing data

1. **Missing Completely At Random (MCAR):** Missing data are MCAR if the complete cases are a random sample of the originally identified set of cases. The probability that a data value is missing is unrelated to the data value itself or to any other value, missing or observed in the data set.
2. **Missing At Random (MAR):** MAR describes the data that are missing for reasons related to completely observed variables in the data set.
3. **Missing Not At Random (MNAR):** If the probability that a data value is missing is related to the missing values themselves then missing data are said to be MNAR.

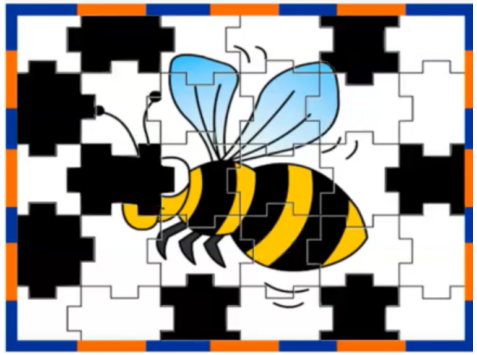
Examples



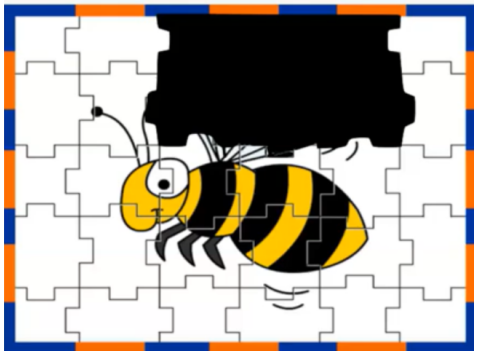
(a) Complete Data



(b) MCAR



(c) MAR



(d) MNAR

Figure 1: Example for different types of missing values

An Illustration

250 observations are sampled from a bivariate-normal distribution with parameters,

$$\boldsymbol{\mu} = \begin{bmatrix} 10 \\ 20 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 9 & 8 \\ 8 & 16 \end{bmatrix}.$$

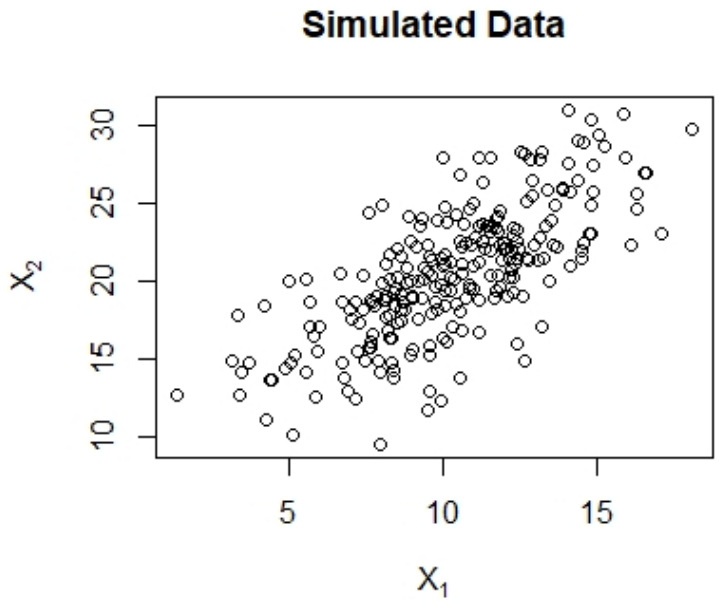


Figure 2: Simulated data: X_1 *vs.* X_2

The population correlation between X_1 and X_2 is

$$\rho_{12} = \frac{8}{\sqrt{9 \times 16}} = \frac{2}{3}$$

.

The slope for the regression of X_1 on X_2 is

$$\beta_{12} = \frac{8}{16} = \frac{1}{2}$$

.

The slope for the regression of X_2 on X_1 is

$$\beta_{21} = \frac{8}{9}$$

.

The variable X_1 is completely observed, but missing data on X_2 will be generated in different ways.

100 of the observations on X_2 are selected at random and set to missing.

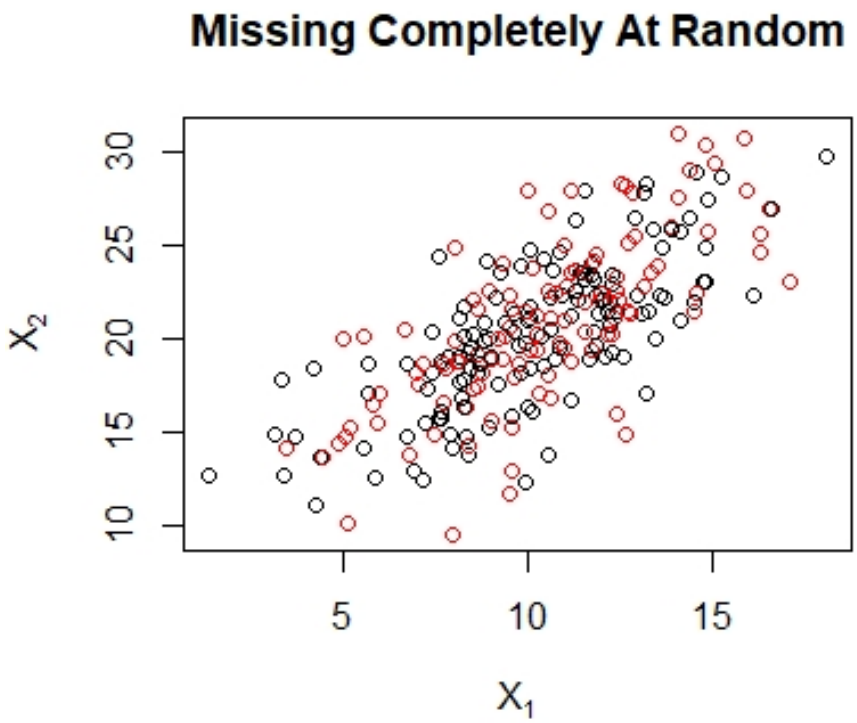


Figure 3: MCAR

An observation's missingness on X_2 is related to its (observed) value of X_1:

$$Pr(X_{i2} \ is \ missing) = \left[1 + exp\left[\frac{1}{2} + \frac{2}{3}(X_{i1} - 10)\right]\right]^{-1} \quad (1)$$

Figure 4: MAR

An observation's missingness on X_2 is related to the (potentially unobserved) value of X_2 itself:

$$Pr(X_{i2} \ is \ missing) = \left[1 + exp\left[\frac{1}{2} + \frac{2}{3}(X_{i2} - 20)\right]\right]^{-1} \quad (2)$$

Figure 5: MNAR

Missing Data Handling Techniques

1. Traditional Approach
   - **List-wise deletion**
   - **Pair-wise deletion**
   - Imputation
     - **Mean imputation**
     - **Regression imputation**
     - **Hot deck imputation**
     - K-nearest neighbour imputation
     - K-mean clustering method
     - Fuzzy K-mean clustering imputation
2. Modern Approach
   - **Maximum Likelihood**
   - **Multiple imputation**
   - EM algorithm

TRADITIONAL APPROACHES FOR HANDLING MISSING DATA

Traditional Approaches

- Complete-Case Analysis
- Available-Case Analysis
- Mean Imputation
- Regression Imputation
- Hot Deck Imputation

Using following data we will demonstrate different Traditional Techniques:

x_1	x_2
5	4
6	NA
13	7
9	NA
11	2
3	8

Complete-Case Analysis

- Also called list-wise or case-wise deletion of missing data.
- Ignores observation with any missing data on the variables included in the analysis.
- The below table is an example for Complete-case Analysis:

x_1	x_2
5	4
6	NA
13	7
9	NA
11	2
3	8

$\rightarrow$

x_1	x_2
5	4
13	7
11	2
3	8

Complete-Case Analysis

- Advantages
  - Simple to implement
  - Provides consistent estimates and valid inferences when missing data are MCAR
  - Provides consistent estimates of regression coefficients and valid inferences when missingness does not depend on the response variable (even if data are not MCAR)
- Disadvantages
  - Generally not efficient because it discards some valid data
  - When data are MAR or MNAR, complete case analysis usually provides biased results and invalid inferences.
  - if many variables are included in a model, there may be very few complete cases, so that most of the data would be discarded for the sake of a sample analysis.

Available-Case Analysis

- Also called pair-wise deletion; Uses all non- missing observation to compute each statistic of interest
- In Available Case Analysis, an analyst would use all the observations for a category where the observation is recorded for that category, even if that response has missing values for other categories.

x_1	x_2
5	4
6	NA
13	7
9	NA
11	2
3	8

$\rightarrow$

x_1	x_2
5	4
6	
13	7
9	
11	2
3	8

Available-Case Analysis

- Advantages
  - Simple
  - All available data is used
  - Smaller loss of cases than in list-wise deletion
- Disadvantages
  - Appears to use more information than complete-case analysis, but these estimators can be less efficient than those based on complete-case analysis
  - The populations of each analysis would be different and possibly non-comparable.
  - Generally provides biased estimate and invalid inferences when data are MAR or MNAR

Unconditional Mean Imputation

- Replacing missing values with plausible imputed values.
- The resulting completed data set is then analyzed using standard methods.
- **Unconditional mean imputation** (or mean substitution) replaces each missing value with the mean of the observed data for the variable.
- Mean imputation preserves the means of variables, but it makes their distributions less variable and tends to weaken relationships between variables.

Example

Mean Substitution

x_1	x_2
5	4
6	NA
13	7
9	NA
11	2
3	8

$\rightarrow$

x_1	x_2
5	4
6	5.25
13	7
9	5.25
11	2
3	8

Unconditional Mean Imputation

- Advantages
  - Simple
  - Full sample size is preserved
- Disadvantages
  - Theoretically illogical
  - Sometimes overestimates and sometimes underestimates coefficients
  - Mean is often a bad estimate
  - Attenuates variance

Conditional mean imputation

- Replaces missing data with predicted values obtained
- **Regression imputation**-uses regression to predict what a missing value would be, using the observed to inform the regression equation.
- From a regression equation using available data,we regress each variable with missing data on other variables in the data set.The resulting regression equation is used to produce predicted values that replace the missing data.

Example

The regression equation corresponding to the given data is,

$$x_2 = 6.5441 - 0.1618x_1$$

x_1	x_2
5	4
6	NA
13	7
9	NA
11	2
3	8

$\rightarrow$

x_1	x_2
5	4
6	5.57
13	7
9	5.09
11	2
3	8

Conditional mean imputation

Advantages

- Easy to implement
- Contrary to mean imputation, regression imputation can also be used when more than 10% of the data is missing and when the data contains highly correlated variables

Disadvantages

- Imputed observations tend to be less variable than real data because they lack residual variation.
- failed to account for uncertainty in the estimation of the regression coefficients used to obtain the imputed values

Hot Deck Imputation

- Replaces each missing value with an observed response from a similar unit
- The idea is to use some idea of similarity to cluster the data before executing imputation.
- The more typical application of hot-deck imputation replaces each missing value with a random draw from a sub-sample of respondents that scored similarly on a set of matching variables.

Hot Deck Imputation

Advantages

- Full sample size is preserved
- Preserves the univariate distribution of the data
- Does not reduce the variability of the filled in data to the same extent as other imputation methods

Disadvantages

- We are not imputing any distinct values.
- Artificially increases statistical power by assuming that similar units are actually identical

Example

Hot Deck Imputation

x_1	x_2
5	4
6	NA
13	7
9	NA
11	2
6	8
9	15

$\rightarrow$

x_1	x_2
5	4
6	8
13	7
9	15
11	2
6	8
9	15

MCAR Dataset

	μ_1	μ_2	σ_1^2	σ_2^2	σ_{12}	ρ_{12}	β_{12}	β_{21}
Population	10.0000	20.0000	9.0000	16.0000	8.0000	0.6670	0.5000	0.8890
Sample	10.2603	20.5311	8.7212	17.9604	9.0206	0.7208	0.5022	1.0343
C-MCAR	10.1279	20.2249	8.9553	16.1249	8.8211	0.7341	0.5471	0.9850
A-MCAR	10.2603	20.2249	8.7212	16.1249	8.8211	0.7438	0.5471	0.9850
I-MCAR	10.2603	20.2249	8.7212	8.7424	4.7825	0.5477	0.5471	0.5484
R-MCAR	10.2603	20.2249	8.7212	8.7424	4.7825	0.5477	0.5471	0.5484

MAR Dataset

	μ_1	μ_2	σ_1^2	σ_2^2	σ_{12}	ρ_{12}	β_{12}	β_{21}
Population	10.0000	20.0000	9.0000	16.0000	8.0000	0.6670	0.5000	0.8890
Sample	10.2603	20.5311	8.7212	17.9604	9.0206	0.7208	0.5022	1.0343
C-MAR	12.0017	22.3503	3.6708	13.6906	3.9609	0.5587	0.2893	1.0790
A-MAR	10.26034	22.3503	8.7212	13.6901	3.9609	0.3625	0.2893	1.0790
I-MAR	10.2603	20.5311	8.7212	17.9604	9.0206	0.7208	0.5022	1.0343
R-MAR	10.2603	20.5311	8.7212	17.9604	9.0206	0.7208	0.5022	1.0343

MNAR Dataset

	μ_1	μ_2	σ_1^2	σ_2^2	σ_{12}	ρ_{12}	β_{12}	β_{21}
Population	10.0000	20.0000	9.0000	16.0000	8.0000	0.6670	0.5000	0.8890
Sample	10.2603	20.5311	8.7212	17.9604	9.0206	0.7208	0.5022	1.0343
C-MNAR	11.5797	22.9692	5.5848	8.2798	3.8869	0.5716	0.4694	0.6960
A-MNAR	10.2603	22.9692	8.7212	8.2798	3.8869	0.4574	0.4694	0.6960
I-MNAR	10.2603	20.5311	8.7212	17.9604	9.0206	0.7208	0.5022	1.0343
R-MNAR	10.2603	20.5311	8.7212	17.9604	9.0206	0.7208	0.5022	1.0343

MAXIMUM LIKELIHOOD ESTIMATION

Maximum Likelihood Estimation

Let $p(\mathbf{X};\boldsymbol{\theta}) = p(\mathbf{X}_{obs}, \mathbf{X}_{mis};\boldsymbol{\theta})$ represent the joint probability density for the complete data $\mathbf{X}$, where the observed and missing components are denoted by $\mathbf{X}_{obs}$ and $\mathbf{X}_{mis}$, respectively.

$\boldsymbol{\theta}$ contains the unknown parameters on which the complete data distribution depends.

The ML estimate of $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be obtained from the marginal distribution of the observed data, if missing data is missing at random.

Consider bivariate normally distributed variables X_1 and X_2. Here, X_1 is completely observed in a sample of n observation and X_2 has m $<$ n observations missing at random.

For notational convenience, the first m observations are taken as missing.

Then, from the univariate-normal distribution,

$$p_1(x_{i1}; \mu_1, \sigma_1^2) = \frac{1}{\sigma_1\sqrt{2\pi}} exp\left[-\frac{(x_{i1} - \mu_1)^2}{2\sigma_1^2}\right] \tag{3}$$

is the marginal probability density for observation i on variable X_1, and from the bivariate normal distribution,

$$p_{12}(x_{i1}, x_{i2}; \Theta) = \frac{1}{2\pi\sqrt{det\Sigma}} exp\left[-\frac{1}{2}[(\boldsymbol{x_i} - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x_i} - \boldsymbol{\mu})]\right] \tag{4}$$

where $\Theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \sigma_{12})$ is the joint probability density for observation i on variables X_1 and X_2 .

In Eq. (4), $x_i = (x_{i1}, x_{i2})$ is a vector giving a pair of values for x_{i1} and x_{i2}.

$\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ is a vector of means for the two variables and $\boldsymbol{\Sigma} = \begin{bmatrix} {\sigma_1}^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$ is the covariance matrix.

Therefore the log-likelihood for the observed data is,

$$log_e\ L(\Theta) = \Sigma_{i=1}^{m} log_e\ p_1(x_{i1}; \mu_1, \sigma_1^2) + \Sigma_{m+1}^{n} log_e\ p_{12}(x_{i1}, x_{i2}, \Theta) \qquad (5)$$

Θ which maximizes the log likelihood, $log_e\ L(\Theta)$ will be the ML estimate of Θ. We have derived the ML estimates by differentiating the log likelihood w.r.t the parameters and equating it to zero.

The statistics,

$$\bar{x_1}^* = \frac{\sum_{m+1}^{n} x_{i1}}{n-m}, \quad s_1^2* = \frac{\sum_{m+1}^{n} (x_{i1} - \bar{x_1}^*)^2}{n-m} \tag{6}$$

$$\bar{x_2}^* = \frac{\sum_{m+1}^{n} x_{i2}}{n-m}, \quad s_2^2* = \frac{\sum_{m+1}^{n} (x_{i2} - \bar{x_2}^*)^2}{n-m} \tag{7}$$

$$s_{12}^* = \frac{\sum_{m+1}^{n} (x_{i1} - \bar{x_1}^*)(x_{12} - \bar{x_2}^*)}{n-m} \tag{8}$$

are the means,variances and covariance for the two variables computed from the n-m complete cases.

$$\overline{x_1} = \frac{\sum_{1}^{n} x_{i1}}{n}, \quad s_1^2 = \frac{\sum_{1}^{n} (x_{i1} - \bar{x_1})^2}{n} \tag{9}$$

are the mean and variance of x_1 computed from all the n available cases.

The ML estimators of the parameter of the bivariate-normal model are,

$$\hat{\mu_1} = \bar{x_1} \tag{10}$$

$$\hat{\mu_2} = \bar{x_2}^* + \frac{s_{12}^*}{s_1^{2*}}(\bar{x_1} - \bar{x_2}^*) \tag{11}$$

$$\hat{\sigma_1}^2 = s_1^2 \tag{12}$$

$$\hat{\sigma_2}^2 = s_2^2 + (\frac{s_{12}^*}{s_1^{2*}})^2(s_1^2 - s_1^{2*}) \tag{13}$$

$$\hat{\sigma_{12}} = \sigma_{12}^*(\frac{s_1^2}{s_1^{2*}}) \tag{14}$$

Thus,the ML estimates combine information from the complete-case and available-case statistics.

Multiple Imputation

Multiple imputation for missing data is an attractive method for handling missing data in multivariate analysis. The idea of multiple imputation for missing data was first proposed by Rubin (1977) [6].

- MI has three basic phases:
- Imputation or Fill-in Phase: The missing data are filled in with estimated values and a complete data set is created. This process of fill-in is repeated m times.
- Analysis Phase: Each of the m complete data sets is then analyzed using a statistical method of interest (e.g. linear regression).
- Pooling Phase: The parameter estimates (e.g. coefficients and standard errors) obtained from each analyzed data set are then combined for inference.
  The imputation method you choose depends on the pattern of missing information as well as the type of variable(s) with missing information.

Multiple Imputation

In R, we can use the mice (Multiple Imputation with Chained Equations) to perform multiple imputation and the subsequent analysis. The three steps above are performed via the mice, with and pool functions respectively.

Advantages

- Accounts for uncertainty due to missing data.
- No biases (if imputation model is correct).
- Easy to use.
- increases the power of statistical tests by increasing the number of observations used
- The researcher can perform multiple imputation for missing data with any kind of data in any kind of analysis with well equipped software.

Simulation Study

In this section, we perform a simulation study to compare the performance of different imputation techniques studied so far for MAR missing values.

The following steps design the full simulation experiment:

Step 1: Generate 250 observations from $N_2\ (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\mu} = \begin{bmatrix} 10 \\ 20 \end{bmatrix}, \quad \boldsymbol{\Sigma} = \begin{bmatrix} 9 & 8 \\ 8 & 16 \end{bmatrix}.$$

Step 2: We set some observations on X_2 missing, using the following:

$$\pi = Pr(X_{i2}\ is\ missing) = \left[1 + exp\left[\frac{1}{2} + \frac{2}{3}(X_{i1} - 10)\right]\right]^{-1} \tag{15}$$

If $\pi > 0.5$, X_{i2} is set to be missing.

Step 3: Compute the estimates of $\Theta = (\mu_1, \mu_2, \beta_{12}, \beta_{21})$ using all the imputation methods discussed earlier. Let us denote the estimate as $\hat{\Theta}$. For i^{th} sample, the squared error is given by:

$$SQ_i = \left(\hat{\Theta} - \Theta\right)^2 \tag{16}$$

Step 4: Repeat the steps 1 to 3, 1000 times.

Step 5: From the sampling distribution of the estimates, calculate the mean parameter estimates, root mean squared error, confidence interval, average width of the interval and the coverage probability.

Parameter	Complete Cases	Mean Imp.	Regression Imp.	Multiple Imp.
$\mu_1 = 10$	11.476	10.001	10.001	10.001
	(1.489)	(0.189)	(0.189)	(0.189)
$\mu_2 = 20$	21.222	21.322	20.008	20.008
	(1.355)	(1.355)	(0.326)	(0.344)
$\beta_{12} = 0.5$	0.391	0.391	0.645	0.498
	(0.117)	(0.117)	(0.151)	(0.041)
$\beta_{21} = 0.888$	0.891	0.353	0.891	0.890
	(0.100)	(0.538)	(0.100)	(0.106)

Table 1: Mean Parameter Estimates and Confidence Interval Coverage for a Simulation Experiment With Data Missing at Random (MAR)

Parameter	Complete Cases	Mean Imp.	Regression Imp.	Multiple Imp.
$\mu_1 = 10$	0 (0.792)	.951 (0.750)	.951 (0.750)	.951 (0.746)
$\mu_2 = 20$	.005 (1.194)	0 (0.711)	.823 (0.881)	.947 (1.451)
$\beta_{12} = 0.5$	.304 (0.174)	.629 (0.246)	.037 (0.140)	.955 (0.175)
$\beta_{21} = 0.888$	.953 (0.396)	0 (0.220)	.661 (0.191)	.939 (0.463)

Table 2: Confidence Interval Coverage for a Simulation Experiment With Data Missing at Random (MAR)

Real Data Analysis: Admissions Data

Variable Name	Type	Description
Chance.of.Admit	Continuous	Chance of admission in the postgraduate course
CGPA	Continuous	CGPA secured in the undergraduate course
GRE Score	Continuous	Marks obtained in the GRE Exam
TOEFL Score	Continuous	Marks obtained in the TOEFL Exam
University.Rating	Categorical (Ordinal)	Undergraduate university rating
SOP	Continuous	Marks obtained in the SOP
LOR	Continuous	Marks obtained in the LOR
Research	Categorical (Binary)	Involvement in any research during under-graduation

Chance of Admit	CGPA	GRE Score	TOEFL Score	University Rating	SOP	LOR	Research
0.98	9.06	316	105	3	3	3.5	0
0.87	8.86	329	119	4	4.5	4.5	1
0.84	7.45	314	103	2	2	3	0
0.84	8.1	322	110	5	5	4	1
0.84	8.7	324	111	3	2.5	2	1

Variable Name	Chance.of.Admit (Y)	CGPA (X)
Min	0.4100	5.060
1st Qu.	0.5900	6.570
Median	0.6400	7.080
Mean	0.6432	7.062
3rd Qu.	0.7000	7.522
Max.	0.9800	9.060
S.D.	0.0841	0.7425
Shapiro Wilk P-Value	0.1274	-

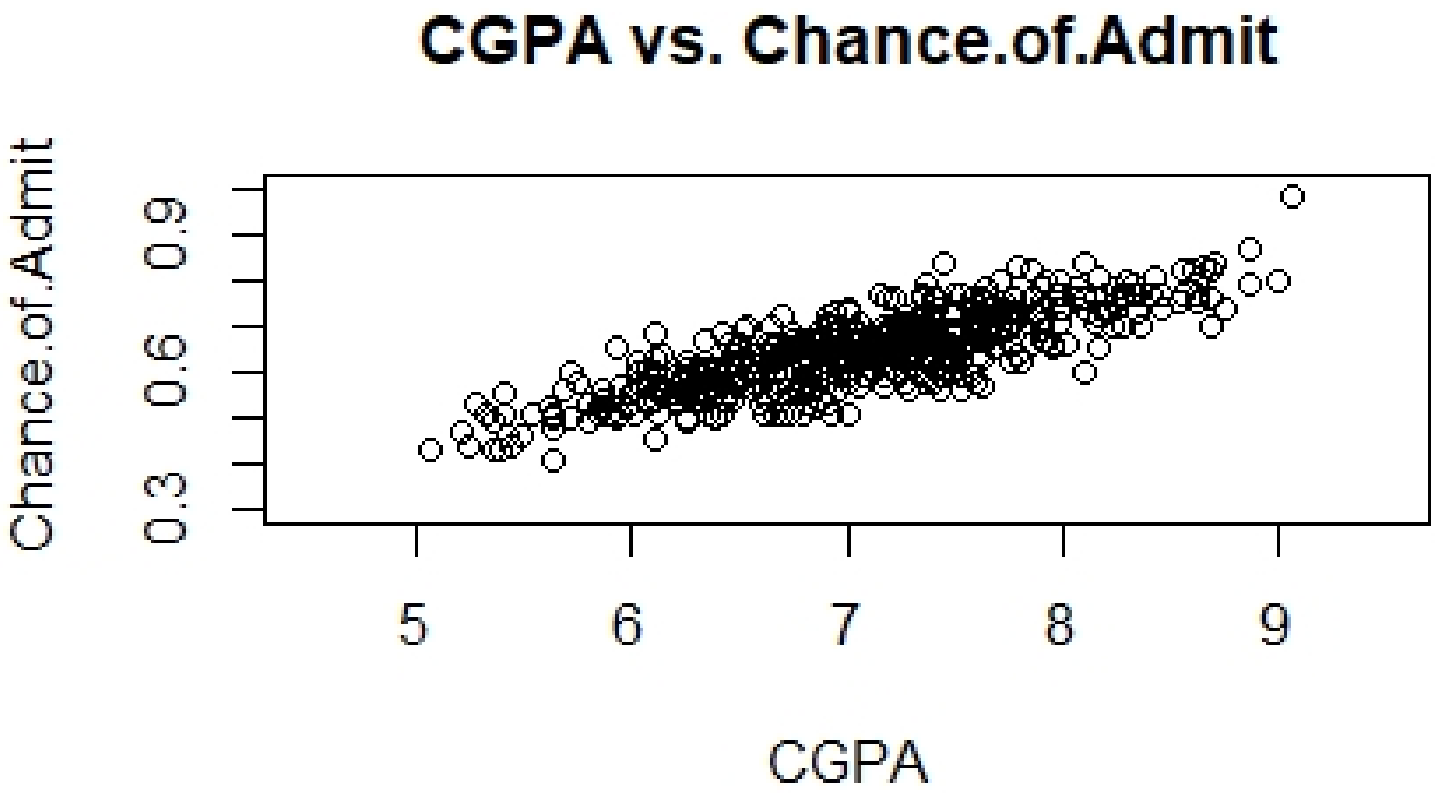


Figure 6: CGPA vs. Chance.of.Admit

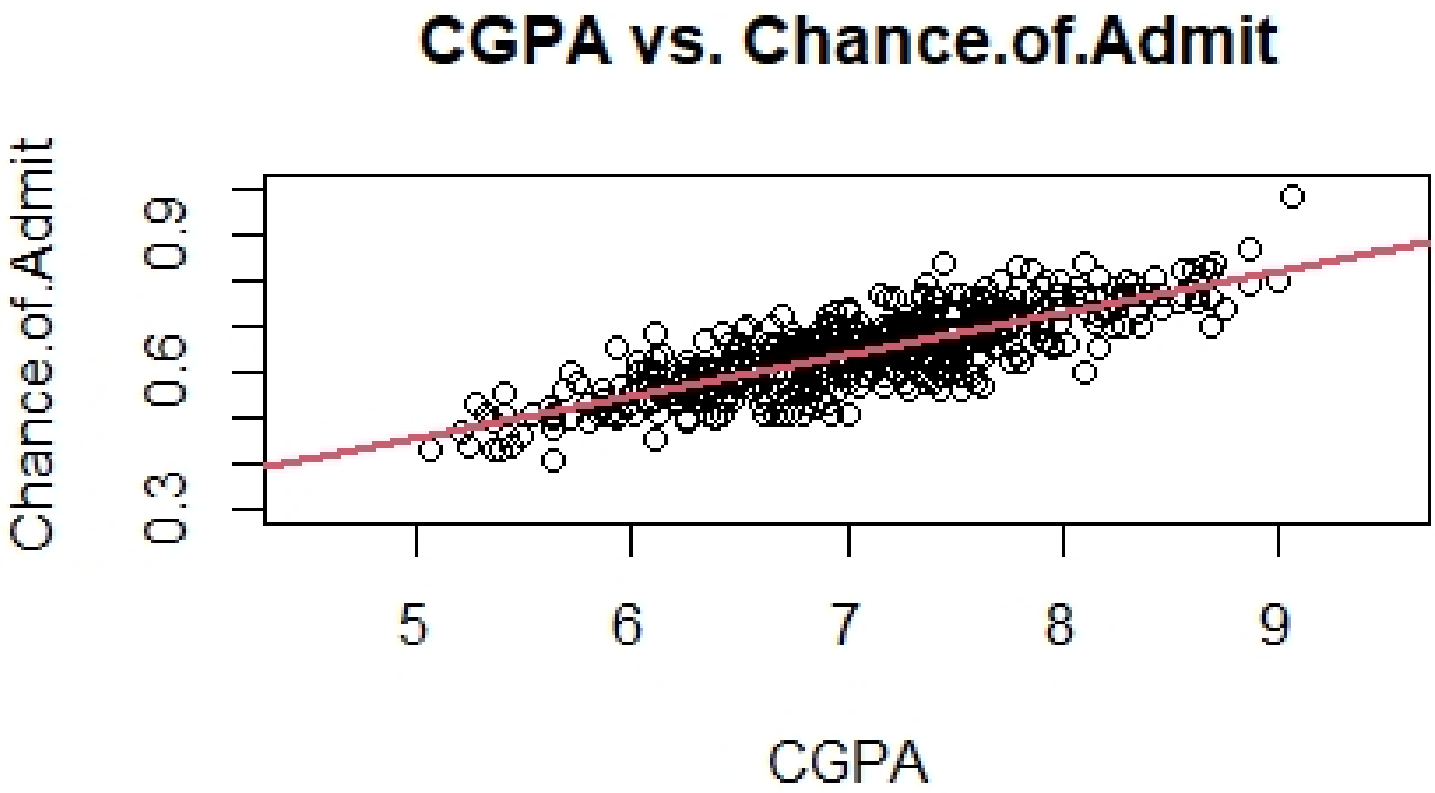


Figure 7: Regression Fit: Y = 0.002 + 0.091 X

MCAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}	β_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7998	0.0907
10%	0.6443	7.0741	0.0069	0.5374	0.7928	0.0901
25%	0.6476	7.0963	0.0075	0.5595	0.8006	0.0925
50%	0.6448	7.0499	0.0069	0.5909	0.8226	0.0889
75%	0.6467	7.1740	0.0063	0.5059	0.7694	0.0861

Table 3: Complete Case Analysis Results

MCAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}	β_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7998	0.0907
10%	0.6432	7.0741	0.0071	0.5374	0.7846	0.0901
25%	0.6432	7.0963	0.0071	0.5595	0.8219	0.0925
50%	0.6432	7.0499	0.0071	0.5909	0.8120	0.0889
75%	0.6432	7.1740	0.0071	0.5059	0.7275	0.0861

Table 4: Available Case Analysis Results

MCAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}	β_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7999	0.0907
10%	0.6432	7.0741	0.0071	0.4771	0.7393	0.0901
25%	0.6432	7.0963	0.0071	0.4361	0.7257	0.0925
50%	0.6432	7.0499	0.0071	0.2889	0.5678	0.0889
75%	0.6432	7.1740	0.0071	0.1328	0.3728	0.0861

Table 5: Mean Imputation Results

MCAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}	β_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7998	0.0907
10%	0.6432	7.0741	0.0071	0.4771	0.7393	0.0901
25%	0.6432	7.0963	0.0071	0.4361	0.7257	0.0925
50%	0.6432	7.0499	0.0071	0.2889	0.5678	0.0889
75%	0.6432	7.1738	0.0071	0.1328	0.3728	0.0861

Table 6: Regression Imputation Results

MAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}	β_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7998	0.0907
55%	0.6629	7.2053	0.0049	0.4290	0.7293	0.0776
60%	0.6831	7.3350	0.0037	0.3754	0.6973	0.0690
65%	0.7118	7.5320	0.0028	0.3397	0.6493	0.0589

Table 7: Complete Case Analysis Results

MAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}	β_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7998	0.0907
55%	0.6432	7.2053	0.0071	0.4290	0.6037	0.0776
60%	0.6432	7.3349	0.0071	0.3754	0.5027	0.0690
65%	0.6432	7.5320	0.0071	0.3398	0.4080	0.0589

Table 8: Available Case Analysis Results

MAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}	β_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7998	0.0907
55%	0.6432	7.2053	0.0071	0.3748	0.5643	0.0776
60%	0.6432	7.3350	0.0071	0.2678	0.4246	0.0690
65%	0.6432	7.5320	0.0071	0.1634	0.2829	0.0589

Table 9: Mean Imputation Results

MAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}	β_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7998	0.0907
55%	0.6432	7.2053	0.0071	0.3748	0.5643	0.0776
60%	0.6432	7.3350	0.0071	0.2678	0.4246	0.0690
65%	0.6432	7.5309	0.0071	0.1641	0.2828	0.0588

Table 10: Regression Imputation Results

MAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7998
55%	0.6432	7.0699	0.0071	0.5339	0.0486
60%	0.6432	7.0542	0.0071	0.5441	0.0499
65%	0.6432	7.0408	0.0071	0.5593	0.0507

Table 11: ML Estimation Results

MAR Missing	μ_X	σ_X^2	μ_Y	σ_Y^2	ρ_{XY}
0%	0.6432	7.0618	0.0071	0.5513	0.7998
55%	0.6432	7.0699	0.0071	0.5085	0.0486
60%	0.6432	7.0542	0.0071	0.4888	0.0499
65%	0.6432	7.0408	0.0071	0.4573	0.0507

Table 12: Multiple Imputation Results

Bibliography

- Norman R Draper and Harry Smith. *Applied regression analysis*. Vol. 326. John Wiley & Sons, 1998.
- John Fox. *Applied regression analysis and generalized linear models*. Sage Publications, 2015.
- Paul Lodder. "To impute or not impute: That's the question". In: *Advising on research methods: Selected topics* (2013), pp. 1–7.
- Douglas C Montgomery, Elizabeth A Peck, and G Geoffrey Vining. *Introduction to linear regression analysis*. John Wiley & Sons, 2021.
- Therese D Pigott. "A review of methods for missing data". In: *Educational research and evaluation* 7.4 (2001), pp. 353–383.
- Donald B Rubin. *Multiple imputation for nonresponse in surveys*. Vol. 81. John Wiley & Sons, 2004.


THANK
YOU