Incomplete Data Analysis:
Entrega do Trabalho Prático

Manuel Curral

February 2025

1 Classificação do Mecanismo de Dados em Falta

Analisemos cada situação para determinar o mecanismo de dados em falta:

(a) Inquérito sobre rendimentos

- A probabilidade de um dado estar em falta depende diretamente do valor real do rendimento (rendimentos mais elevados têm menor probabilidade de ser reportados); - A ausência de valores depende diretamente da variável que teria sido observada; - Trata-se de um caso claro de **MNAR** (Missing Not At Random - Não Ausente Aleatoriamente), pois a probabilidade de ausência depende do próprio valor não observado.

(b) Estudo sobre Grelina

- A contaminação da amostra no laboratório é um evento aleatório; - A ausência de valores é completamente independente de variáveis observadas ou não observadas; - Assim, trata-se de um caso de **MCAR** (Missing Completely At Random - Ausente Completamente ao Acaso), pois a falta de dados ocorre devido a problemas técnicos aleatórios.

(c) Inquérito sobre emprego

- Os dados demográficos (idade, género, raça e educação) são completamente observados; - A maior taxa de não resposta entre hispânicos sugere que a ausência de dados depende de uma variável observada (raça); - Assim, trata-se de um caso de **MAR** (Missing At Random - Ausente Aleatoriamente), pois a ausência de dados depende de variáveis observadas, mas não do estatuto de emprego em si (variável não observada).

2 Análise de Caso Completo

Consideremos um conjunto de dados com 500 sujeitos e 20 variáveis, em que cada variável tem 5% de valores em falta.

Dados fornecidos: - 500 sujeitos (linhas); - 20 variáveis (colunas); - Cada variável tem 5% de valores em falta.

Para resolver este problema, analisemos os dados em falta e o impacto da análise de casos completos.

- Cada variável tem 95% de valores completos (5% de dados em falta); - Para que um caso seja completo, todas as 20 variáveis devem estar presentes;

Entendimento da Análise de Casos Completos: Uma análise de casos completos (ou eliminação por linha) inclui apenas os sujeitos que não têm valores em falta em nenhuma das 20 variáveis.

- Probabilidade de um caso ser completo: $(0.95)^{20}$; - Em termos "médios"podíamos dizer que: da amostra com casos completos: $500 \times (0.95)^{20} \approx 179*$ sujeitos; (se os valores ausentes/NAs forem distribuídos aleatoriamente)

Se assumirmos que os valores em falta são independentes, a probabilidade de que um determinado sujeito tenha dados completos para uma variável específica é 95% (dado que 5% dos valores estão em falta).

Como há 20 variáveis independentes, a probabilidade de um sujeito não ter valores em falta em nenhuma delas é:

$$(0.95)^{20}$$

Calculando este valor:

$$(0.95)^{20} \approx 0.3585$$

Assim, a proporção esperada de sujeitos com dados completos é de aproximadamente 35.85%.

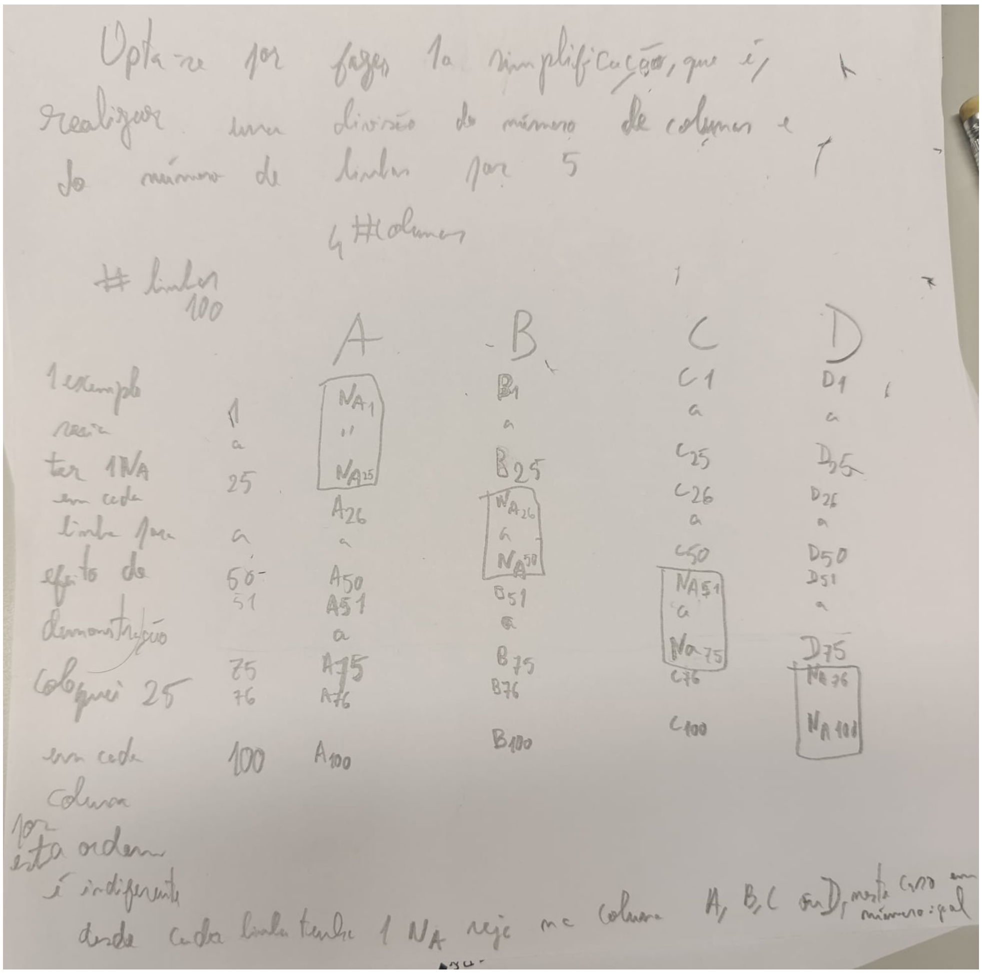
O tamanho máximo da subamostra possível é:

$$500 \times 0.3585 \approx 179.25$$

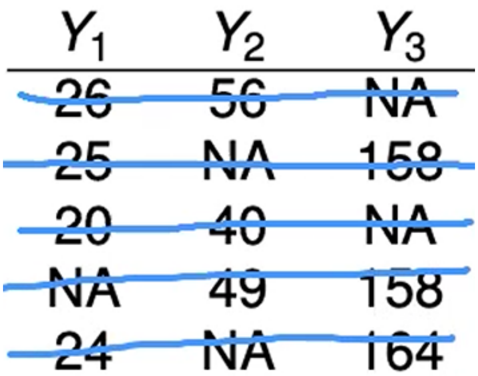
Como o tamanho da amostra deve ser um número inteiro, aproximamos para 179 sujeitos.

Menor subamostra possível: No pior cenário, cada sujeito tem pelo menos um valor em falta, o que significa que nenhum sujeito tem dados completos.

Neste caso, caso todas as linhas tiverem exatamente 1 valor NA em todas linha, de forma que: - O de NA em todas as colunas for a mesmo e igual a 25, perfaz o caso mínimo e extremo, exemplificado na figura seguinte:

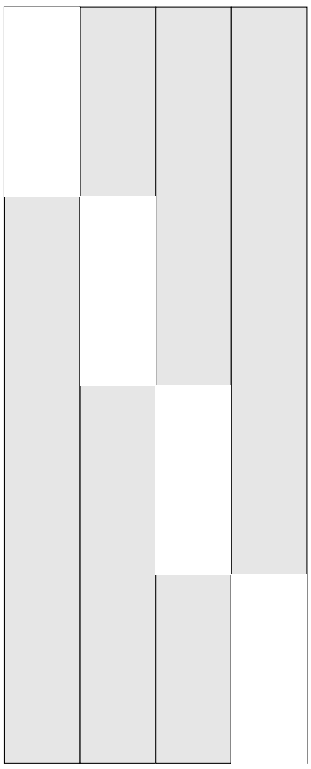


Neste caso 500 linhas excluído 25linhas * 20 por todas as colunas resultaria 500 - 500 = 0, a análise de casos completos excluiria todos os sujeitos. Assim, o menor tamanho possível da subamostra é 0 sujeitos. Teríamos todas as línhas com NA's



Y_1	Y_2	Y_3
26	56	NA
25	NA	158
20	40	NA
NA	49	158
24	NA	164

A B C D



Presente

NA

Maior subamostra possível:

Fazendo a mesma analogia que anteriormente, partimos de um caso mais simples, para descobrir a fórmula visualmente.

Assim sendo temos, a divisão por 5 do número de linhas e também das colinas, retando:

4 colunas, de 20 / 5 100 linhas , de 500 / 5

O caso mais extremos seria em todas as colunas os NA(s) coincidirem exatamente nas mesmas linhas.

Sendo X a representação de NA graficamente e O representação do valor

$$\begin{matrix} X & X & X & X \\ X & X & X & X \\ X & X & X & X \\ X & X & X & X \\ X & X & X & X \end{matrix}$$

com 95 colunas de $O \quad O \quad O \quad O$

A ordem das linhas será indiferente, as 5 linhas de 4 X seguidos na linha, pode aparecer na linha 7, 34, 39, 57 e 99 ou outro conjunto de 5 valores desde que estejam representados no intervalo [1,100] Caso consideremos que desta forma não acrescenta qq informação, teríamos então o caso mais largo como sendo:

$$
\begin{array}{cccc}
O & X & X & X \\
X & O & X & X \\
X & X & O & X \\
X & X & X & O \\
X & X & O & X \\
X & O & X & X \\
O & X & X & O
\end{array}
$$

Na imagem fotográfica abaixo, coloquei o X na última linha representada mas está errado, pois para ser exatamente 5% de missing values, só poderemos ter 5 NA em cada coluna.

O número necessário para conter pelo menos um valor (not NA) em cada linha, teríamos de ter 2(n-1) + 1 linhas, no caso mais largo possível. n é o número de colunas. Na representação temos a ideia de como a ideia surgiu, 19+19+1=39 listwise deletion

500-39 = 461

a fórmula pode ser ser apresentada também no formato 2n-1, em que n é o número de colunas que é percorrida, e na volta a coluna no extremo pode aparecer simultaneamente com o aparecimento do valor em outra linha. De forma similar estas 39 linhas podem aparecer de forma aleatória ao longo das 500 linhas. A representação desta forma "agregada", serve meramente para ser facilitada a visualização do raciocínio para chegar à minha proposta de resolução para chegar ao valor pedido como resposta.

Conclusão: - Maior subamostra possível: 461 sujeitos (todas as linhas completas e as listwise deletions rows tem 1 único valor excepto uma linha que acaba por ter 2); - Menor subamostra possível: 0 sujeitos (se todos os sujeitos tiverem pelo menos e exatamente um valor ausente).

Justificação: A probabilidade de um sujeito ter dados completos é $(0.95)^{20} \approx 0.3585$, e no pior caso, todos os sujeitos têm valores em falta.

Distribuição de Dados Ausentes (NA)

isolado os Na em 5% das linhas

Caso extremo

X X X X
X X X X
X X X X
X X X X
X X X X

Conteúdo se concluirmos que isto não são dados, pois não aglomeram nenhuma informação, podemos partir para o caso em que tem algum registo na linha.

Começado na diagonal

O X X X (n-1)x2+1
X O X X
X X O X
X X X O
X X O X
X O X X
O X X X ← ou O

(3x2+1=7) é 30 então

→ 19

+1

19 ↑ ou seja 39

500-39=461

← 19

3 Questão Bónus

3.1 (a) Mecanismo de Dados em Falta

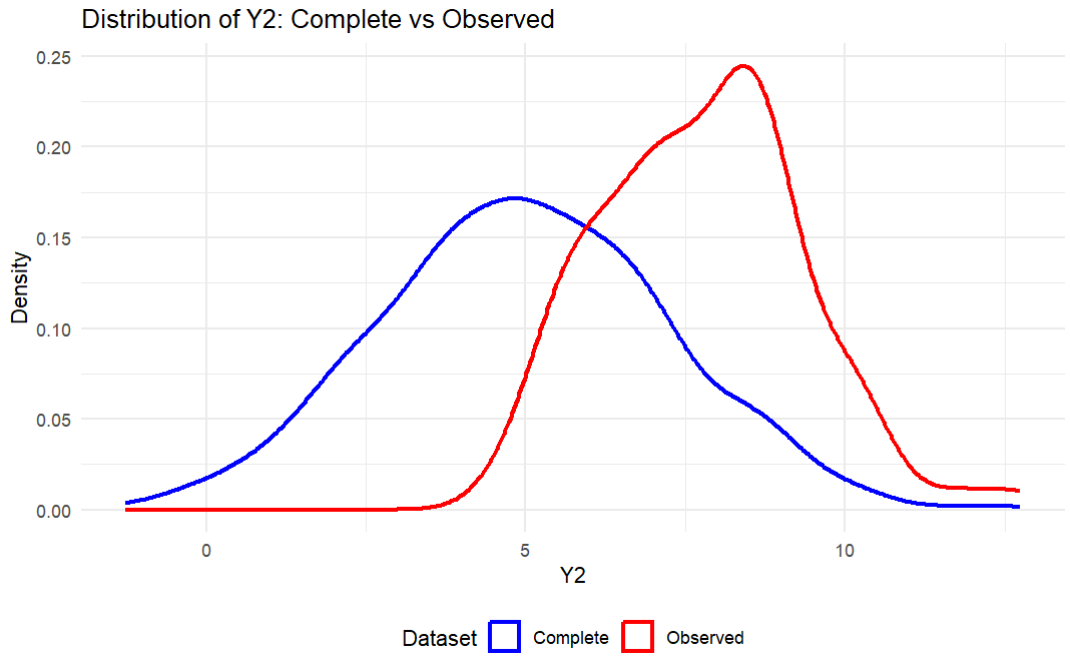
- Neste problema, após simular um conjunto de dados completo de tamanho 500 para (Y1, Y2) e considerando a = 4 e b = 0, o mecanismo de dados em falta é **MAR** (Missing At Random - Em Falta de Forma Aleatória) porque:
- Y_2 está em falta quando:

$$a \times (Y_1 - 2) + b \times (Y_2 - \theta) + Z_3 < 0 \tag{1}$$

- Y_1 está completamente observado.
- A ausência de Y_2 depende de Y_1 (observado) e Z_3 (variável aleatória).
- Com $a = 4$ e $b = 0$, a condição simplifica-se para:

$$4 \times (Y_1 - 2) + Z_3 < 0$$

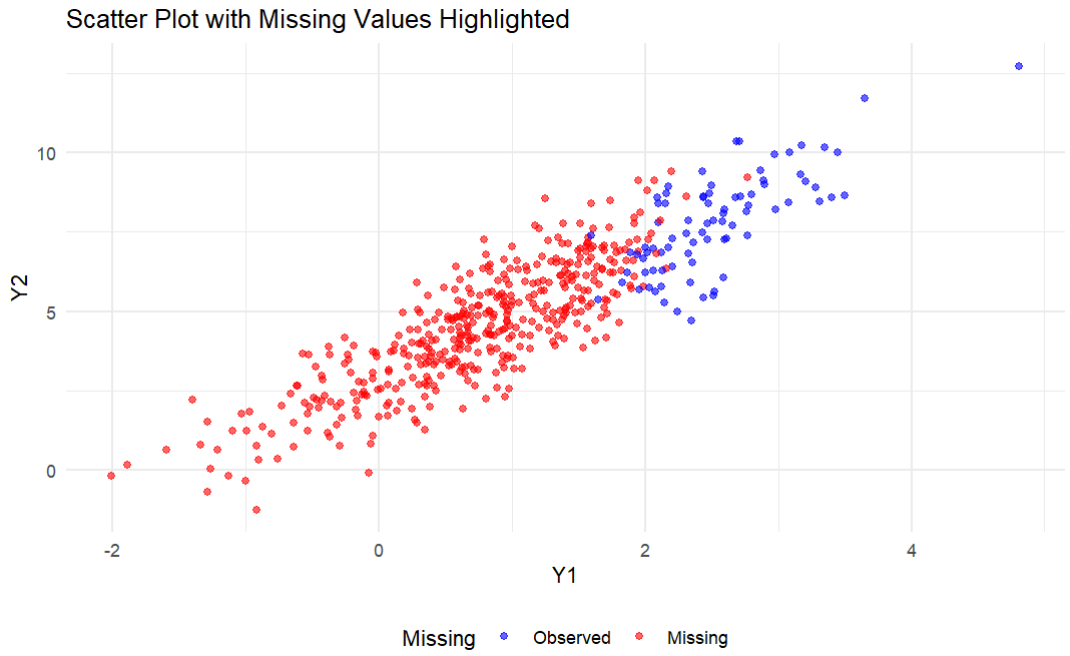
  a ausência de dados depende apenas de Y_1 e Z_3, tornando-o um caso MAR.
- Isto significa que a probabilidade de Y_2 estar em falta depende apenas de Y_1 (que é totalmente observado) e de Z_3 (um termo aleatório).
- Como a falta não depende do próprio valor de Y_2 (pois $b = 0$), mas apenas de variáveis observadas (Y_1) e de ruído aleatório (Z_3), o mecanismo é MAR.



Analisando a distribuição de Y_2 para os dados completos e observados:

- A distribuição de Y_2 para os dados observados (após impor a falta) é diferente da distribuição original completa.

- Especificamente, existe um enviesamento na amostra observada, pois os valores em falta são determinados pelo valor de Y_1.
- Quando Y_1 é menor (abaixo de 2, considerando o termo $Y_1 - 2$), a probabilidade de Y_2 estar em falta aumenta.
- Como Y_1 e Y_2 estão correlacionados (devido ao termo $2 \times Z_1$ em Y_2), isto causa uma diferença sistemática entre as distribuições.



Este comportamento é característico do mecanismo MAR - os dados em falta não são completamente aleatórios (como seria em MCAR), mas a probabilidade de ausência pode ser explicada inteiramente por variáveis observadas (neste caso, Y1).

3.2 (b) Imputação por Regressão Estocástica

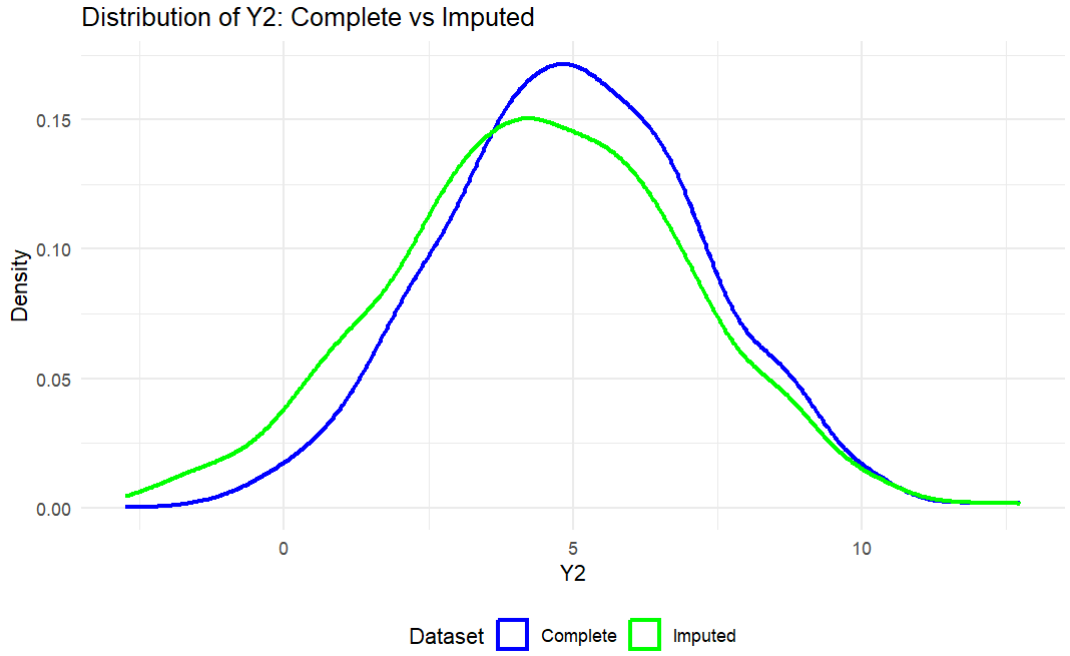
Para o conjunto de dados observados simulado na alínea (a), a imputação dos valores em falta utilizando a imputação por regressão estocástica segue os seguintes passos:

1. Primeiro, ajustamos um modelo de regressão linear utilizando os casos completos, onde Y_2 é a variável dependente e Y_1 é a variável explicativa.
   - Utilizaríamos a relação entre Y_1 e Y_2 para imputar os valores em falta.
   - A distribuição marginal de Y_2 deveria preservar a relação:

$$Y_2 = 5 + 2 \times Z_1 + Z_2 \tag{2}$$

2. Em seguida, para cada valor em falta de Y_2, fazemos a previsão utilizando o modelo ajustado e adicionamos um termo de erro aleatório para preservar a variabilidade.
   - Os valores imputados teriam maior variância do que os valores originais devido à incerteza adicional no processo de imputação.

A distribuição marginal de Y_2 para os dados completos (originalmente simulados) e completados (após imputação) revela aspetos importantes:

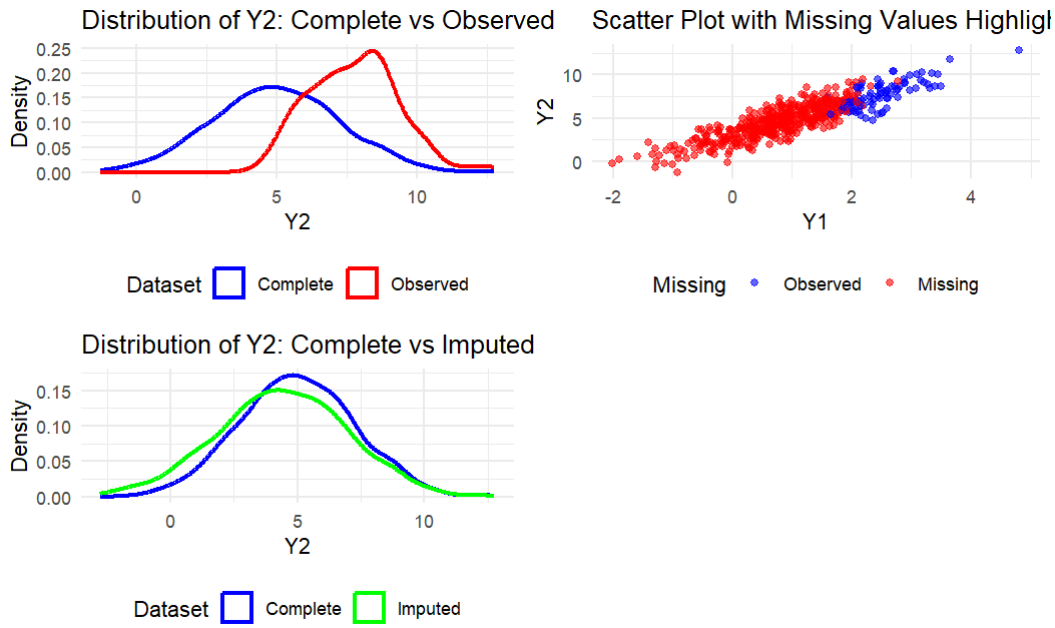


- **Regression model:** $Y_2 = \beta_0 + \beta_1 \times Y_1 + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$.
- **Estimated** β_0: 2.174
- **Estimated** β_1: 2.219
- **Estimated** σ: 1.082
- **Mean of complete** Y_2: 4.999
- **Mean of imputed** Y_2: 4.41
- **SD of complete** Y_2: 2.244
- **SD of imputed** Y_2: 2.56
- A distribuição dos dados imputados aproxima-se bastante da distribuição original, o que indica que a imputação por regressão estocástica foi eficaz.
- A média dos valores imputados é próxima da média dos valores originais, demonstrando que a imputação consegue preservar a tendência central.
- A variabilidade (desvio padrão) da distribuição imputada também é semelhante à original, o que é uma consequência direta da adição de erros aleatórios durante a imputação.

Isto acontece porque:

- O modelo de regressão captura a relação entre Y_1 e Y_2 (lembrando que $Y_2 = 5 + 2 \times Z_1 + Z_2$, onde Z_1 influencia tanto Y_1 como Y_2).

- O componente estocástico (adição de ruído aleatório baseado no desvio padrão residual do modelo) ajuda a preservar a variabilidade natural dos dados.



Uma vez que o mecanismo de dados em falta é MAR (dependente apenas de Y_1 e Z_3), a imputação baseada em Y_1 é particularmente adequada, pois consegue "explicar"o padrão de ausência.

Isto ilustra uma vantagem importante da imputação por regressão estocástica sob o mecanismo MAR: quando a ausência depende de variáveis observadas, podemos utilizar essas mesmas variáveis para fazer imputações bastante precisas.

Se compararmos com a imputação determinística (sem componente aleatório), notaríamos uma subestimação da variabilidade, enquanto a imputação estocástica preserva tanto a média como a dispersão da distribuição original.

Código R para Simulação e Imputação por Regressão Estocástica

(a) Simulação e Identificação do Mecanismo de Falta

Listing 1: Simulação do conjunto de dados completo

```r
# Definir a semente para reprodutibilidade
set.seed(1)

# Gerar o conjunto de dados completo
n <- 500  # tamanho da amostra
Z1 <- rnorm(n)
Z2 <- rnorm(n)
Z3 <- rnorm(n)

# Criar Y1 e Y2 de acordo com as equations especificadas
Y1 <- 1 + Z1
Y2 <- 5 + 2 * Z1 + Z2

```

```r
# Criar o conjunto de dados completo
dados_completos <- data.frame(Y1, Y2)
```

Listing 2: Implementação do mecanismo de dados em falta

```r
# Implementar o mecanismo de falta com a=4 e b=0
a <- 4
b <- 0
indicador_falta <- (a * (Y1 - 2) + b * (Y2 - 6) + Z3) < 0

# Criar o conjunto de dados observado com valores em falta
Y2_obs <- Y2
Y2_obs[indicador_falta] <- NA
dados_observados <- data.frame(Y1, Y2_obs)

# Calcular statistics dos dados completos e observados
num_faltantes <- sum(indicador_falta)
percentagem_faltantes <- mean(indicador_falta) * 100
media_completa_Y2 <- mean(Y2)
media_observada_Y2 <- mean(Y2_obs, na.rm = TRUE)
dp_completo_Y2 <- sd(Y2)
dp_observado_Y2 <- sd(Y2_obs, na.rm = TRUE)
```

Listing 3: Visualização das distribuições de Y2

```r
# Criar graphic de densidade para comparar as distribuitions
p1 <- ggplot() +
  geom_density(aes(x = Y2, color = "Completo"), linewidth = 1) +
  geom_density(aes(x = Y2_obs[!is.na(Y2_obs)], color = "Observado"),
      linewidth = 1) +
  theme_minimal() +
  labs(title = "Distribuition de Y2: Completo vs Observado",
       x = "Y2",
       y = "Densidade",
       color = "Conjunto de Dados") +
  scale_color_manual(values = c("Completo" = "blue", "Observado" = "red
      ")) +
  theme(legend.position = "bottom")
```

(b) Imputation por Regressão Estocástica

Listing 4: Ajuste do modelo de regressão linear

```r
# Ajustar um modelo de regression linear usando os dados observados
modelo <- lm(Y2_obs ~ Y1, data = dados_observados)
beta0 <- coef(modelo)[1]
beta1 <- coef(modelo)[2]
sigma <- summary(modelo)$sigma
```

Listing 5: Imputação por regressão estocástica

```r
# Realizar a imputation por regress o estoc stica
Y2_imp <- Y2_obs
indices_faltantes <- which(is.na(Y2_obs))
Y2_imp[indices_faltantes] <- beta0 + beta1 * Y1[indices_faltantes] +
                             rnorm(length(indices_faltantes), mean = 0,
                                 sd = sigma)
```

```r

# Criar o conjunto de dados completado
dados_completados <- data.frame(Y1, Y2_imp)

# Calcular estat sticas dos dados imputados
media_imputada_Y2 <- mean(Y2_imp)
dp_imputado_Y2 <- sd(Y2_imp)
```

Listing 6: Visualização das distribuições de Y2 completo e imputado

```r
# Criar gr fico de densidade para comparar os dados completos e
    imputados
p3 <- ggplot() +
  geom_density(aes(x = Y2, color = "Completo"), size = 1) +
  geom_density(aes(x = Y2_imp, color = "Imputado"), size = 1) +
  theme_minimal() +
  labs(title = "Distribui  o de Y2: Completo vs Imputado",
       x = "Y2",
       y = "Densidade",
       color = "Conjunto de Dados") +
  scale_color_manual(values = c("Completo" = "blue", "Imputado" = "
     green")) +
  theme(legend.position = "bottom")
```

Listing 7: Sumário dos resultados e conclusões

```r
# Apresentar as estat sticas e conclus es
cat("\n======= PARTE (a) =======\n")
cat("An lise do Mecanismo de Dados em Falta:\n")
cat("N mero de valores em falta:", num_faltantes, "\n")
cat("Percentagem de valores em falta:", round(percentagem_faltantes, 2)
    , "%\n")
cat("M dia de Y2 completo:", round(media_completa_Y2, 3), "\n")
cat("M dia de Y2 observado:", round(media_observada_Y2, 3), "\n")
cat("DP de Y2 completo:", round(dp_completo_Y2, 3), "\n")
cat("DP de Y2 observado:", round(dp_observado_Y2, 3), "\n\n")

cat("Mecanismo de Falta: MAR (Missing At Random)\n")
cat("Justifica  o: Com a=4 e b=0, a aus ncia depende apenas de Y1 (
    totalmente observado) e de Z3 (ru do aleat rio),\n")
cat("              mas n o dos valores de Y2 propriamente ditos.\n\n")

cat("\n======= PARTE (b) =======\n")
cat("Resultados da Imputa  o por Regress o Estoc stica:\n")
cat("Modelo de regress o: Y2 =  0  +  1     Y1 +   , onde    ~ N(0,
       )\n")
cat("Estimativa de  0 :", round(beta0, 3), "\n")
cat("Estimativa de  1 :", round(beta1, 3), "\n")
cat("Estimativa de   :", round(sigma, 3), "\n")
cat("M dia de Y2 completo:", round(media_completa_Y2, 3), "\n")
cat("M dia de Y2 imputado:", round(media_imputada_Y2, 3), "\n")
cat("DP de Y2 completo:", round(dp_completo_Y2, 3), "\n")
cat("DP de Y2 imputado:", round(dp_imputado_Y2, 3), "\n")
```