Serviço Geológico do Brasil – CPRM

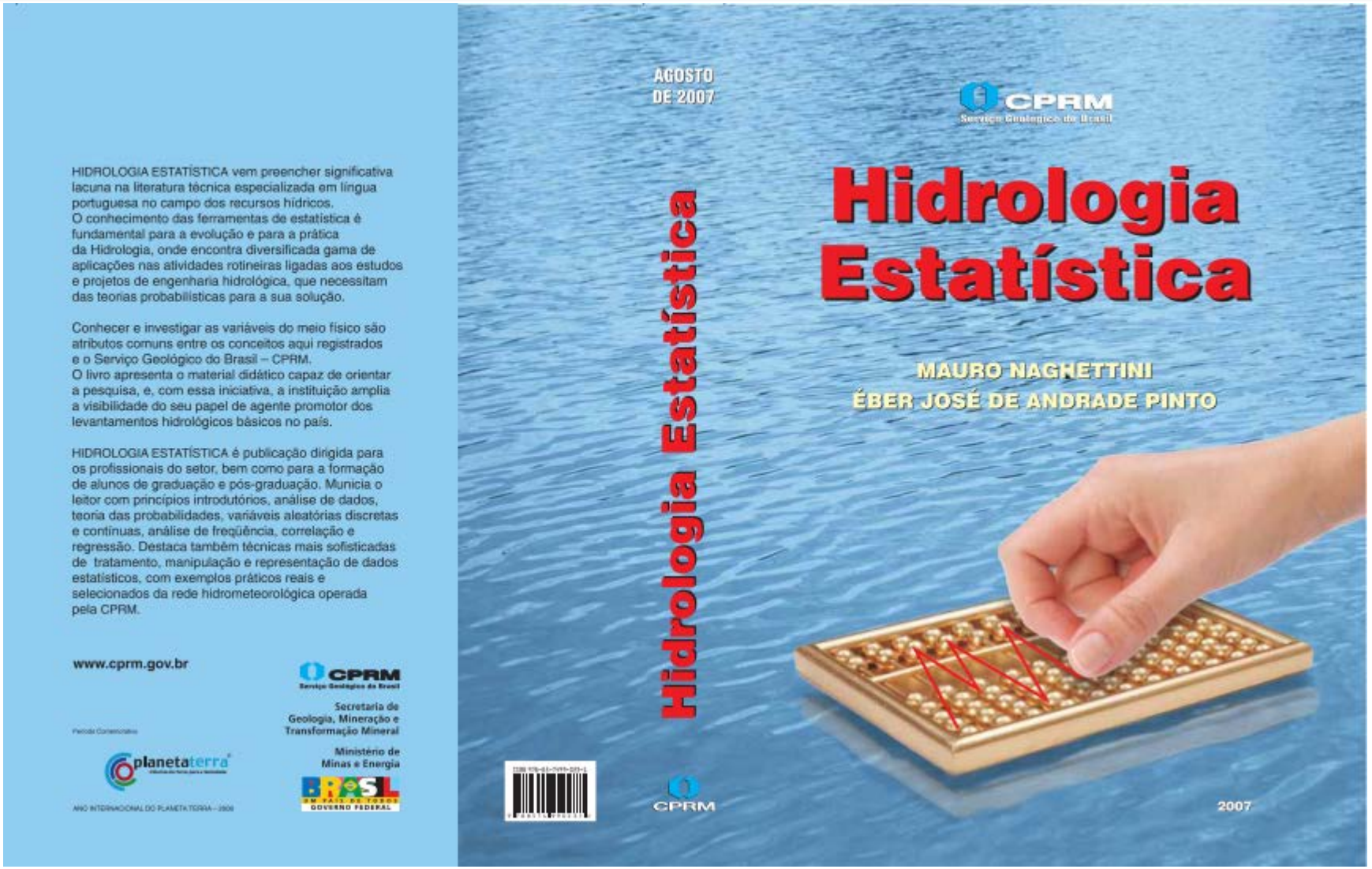
Correlação e Regressão Linear

Aula 05 : RLM

Eber José de Andrade Pinto
Coordenador Executivo do DEHID
eber.andrade@cprm.gov.br
www.cprm.gov.br

Belo Horizonte, 22 de outubro de 2020

Livro Texto



http://www.cprm.gov.br/publique/Hidrologia/Mapas-e-Publicacoes/Livro-%22Hidrologia-Estatistica%22-981.html

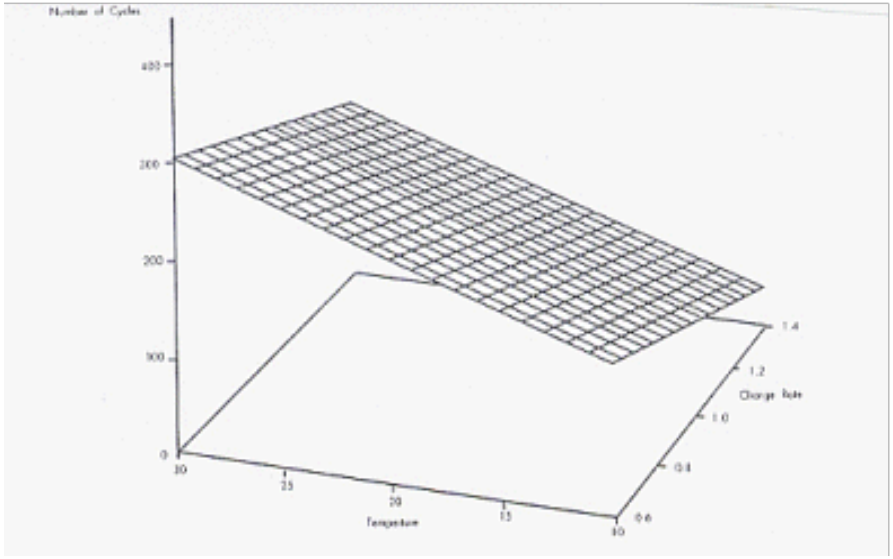
Regressão Múltipla

RM → comportamento da VA dependente Y em função de duas ou mais VA's independentes X_i.

RLM → $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_P X_P$ (modelo geral)

Exemplo: Plano de Regressão

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$



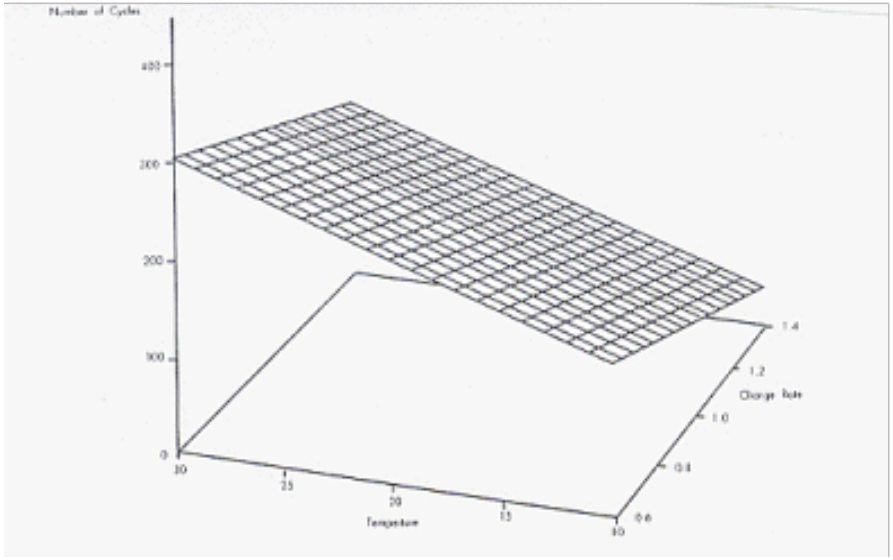
Regressão Múltipla

Plano de Regressão

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

em notação matricial

$$[Y] = [X].[\beta]$$



Equações Normais do Plano de Regressão
(exemplo de MMQ para RLM):

$$\sum Y = N\beta_0 + \beta_1 \sum X_1 + \beta_2 \sum X_2$$

$$\sum X_1 Y = \beta_0 \sum X_1 + \beta_1 \sum X_1^2 + \beta_2 \sum X_1 X_2$$

$$\sum X_2 Y = \beta_0 \sum X_2 + \beta_1 \sum X_1 X_2 + \beta_2 \sum X_2^2$$

Matriz dos Coeficientes de Correlação Simples ou de Ordem Zero

$$\begin{array}{c|ccc} & Y & X_1 & X_2 \\ \hline Y & 1 & r_{YX_1} & r_{YX_2} \\ X_1 & r_{X_1Y} & 1 & r_{X_1X_2} \\ X_2 & r_{X_2Y} & r_{X_2X_1} & 1 \end{array}$$

$$r_{YX_1} = \frac{n\sum YX_1 - \sum Y \sum X}{\sqrt{\left[n\sum X_1^2 - \left(\sum X_1\right)^2\right]\left[n\sum Y^2 - \left(\sum Y\right)^2\right]}}$$

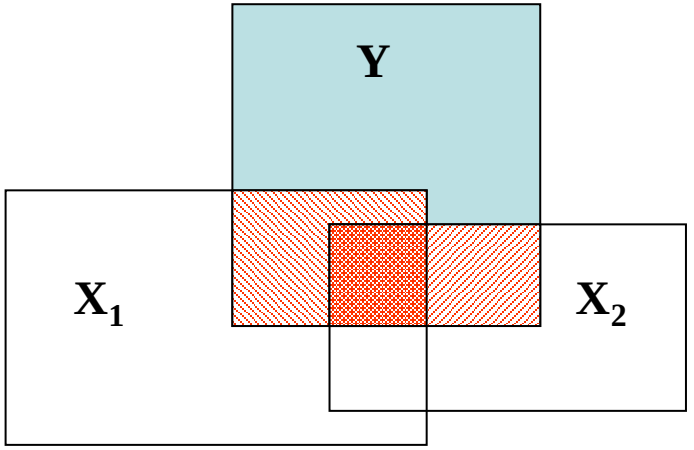
- A variável explicativa com maior correlação linear com Y entra no modelo
- Verificar se a correlação linear entre X_1 e X_2 é maior do que 0,85;
  - se positivo, haverá grande chance de multicolineariade, o que pode impor tendências no cálculo dos coeficientes de regressão.
  - Correção possível: retirar a variável explicativa com menor correlação parcial com *Y.*

Coeficiente de Determinação Múltipla R^2

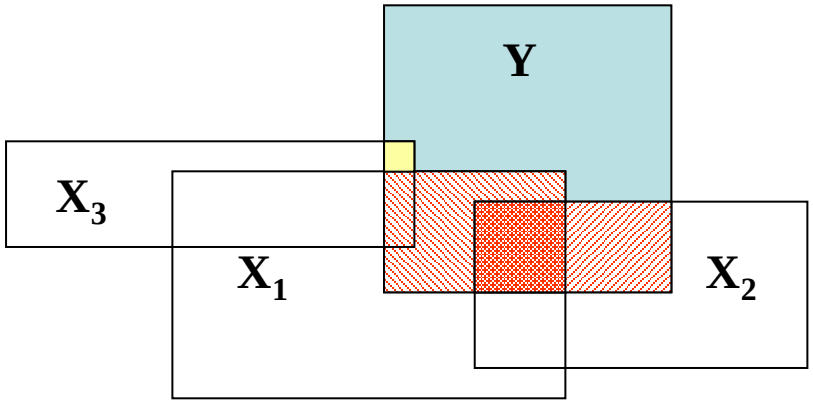
R^2 = parte da variância de Y explicada por X_1 e X_2

$$R^2 = \frac{VEM}{VT} = \frac{\sum\left(\hat{Y} - \bar{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)} \text{ ou}$$

$$R^2 = 1 - \frac{VRES}{VT} = 1 - \frac{\sum\left(Y - \hat{Y}\right)^2}{\sum\left(Y - \bar{Y}\right)^2}$$



Coeficiente de Determinação Múltipla Ajustado R^2_{ajust}



$$R^2_{ajust} = 1 - \frac{(n-1)}{(n-p)}\left(1 - R^2\right)$$

$$Exemplo: \quad R^2_2 = 0{,}92 \quad R^2_3 = 0{,}94 \quad n = 11$$

$$R^2_{2,ajust} = 0{,}9 \quad R^2_{3,ajust} = 0{,}89$$

Coeficiente de Determinação Parcial

É a proporção da variância de *Y* que é explicada por uma variável independente X_k, enquanto se mantém constante as outras variáveis explicativas.

$$R^2_{Yk(P-k)} = \frac{SQ\,Reg(X_k)}{SQT - SQ\,Reg + SQ\,Reg(X_k)}$$

$SQ\,Reg(X_k) = SQ\,Reg$ (todas as variáveis com X_k) - $SQ\,Reg$(todas as variáveis sem X_k)

TABELA ANOVA

Como visto na regressão simples,

$$\sum \left(Y_i - \bar{Y}\right)^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 + \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

(Variância total) = **(Variância Residual)** + **(Variância da Regressão)**

Fonte	Graus de liberdade	Somatório dos quadrados	Quadrado médio
Regressão	P	$SQ\,Reg = \left[\hat{\beta}\right]^T \left[X\right]^T \left[Y\right] - n\bar{Y}^2$	$QM\,Reg = \frac{SQ\,Reg}{P}$
Resíduos	n – P -1	$SQ\,Res = \left[Y\right]^T \left[Y\right] - \left[\hat{\beta}\right]^T \left[X\right]^T \left[Y\right]$	$QM\,Res = \frac{SQ\,Res}{n-P-1}$
Total	**n - 1**	$SQT = \left[Y\right]^T \left[Y\right] - n\bar{Y}^2$	

TESTE DO MODELO

Fonte	Graus de liberdade	Somatório dos quadrados	Quadrado médio
Regressão	P	$SQ\,Reg = \left[\hat{\beta}\right]^T \left[X\right]^T \left[Y\right] - n\overline{Y}^2$	$QM\,Reg = \frac{SQ\,Reg}{P}$
Resíduos	n – P -1	$SQ\,Res = \left[Y\right]^T \left[Y\right] - \left[\hat{\beta}\right]^T \left[X\right]^T \left[Y\right]$	$QM\,Res = \frac{SQ\,Res}{n-P-1}$
Total	**n - 1**	$SQT = \left[Y\right]^T \left[Y\right] - n\overline{Y}^2$	

Teste da Hipótese:

H_0: o modelo de regressão, como um todo, NÃO é significativo

H1 : o modelo de regressão, como um todo, é significativo

$$ET = F_{total} = \frac{QM\,\mathrm{Re}\,g}{QM\,\mathrm{Re}\,s} \sim F_{\nu_1=P,\nu_2=n-P-1,\alpha}$$

quanto maior o F_{total}, mais significativo o modelo

Decisão: se $F_{total} > F_{\nu_1=P,\nu_2=n-P-1,\alpha}$ **Rejeita-se a hipótese nula.**

TESTE DE INCLUSÃO DE NOVAS VARIÁVEIS

Teste da Hipótese:

H0 : a variável X_k NÃO melhora significativamente o modelo

H1 : a variável X_k melhora significativamente o modelo

$$ET = F_{parcial} = \frac{\frac{R_P^2 - R_{P-1}^2}{1 - R_{P-1}^2}}{N - P - 2} \sim F_{\nu_1=1,\nu_2=n-P-2,\alpha}$$

qto. maior o $F_{parcial}$, mais significativa a inclusão

Decisão: se $F_P > F(\alpha, 1, n - p - 1)$ Rejeita-se a hipótese nula

SELEÇÃO DE VARIÁVEIS EXPLICATIVAS

- $R^2_{ajustado}$
- **'*Step backward regression*' → inicia com todas as variáveis explicativas. A menos importante (*r* ou $r_{parcial}$) é retirada e calcula-se o $F_{parcial}$. Se a variável não é significativa, repete-se o processo para a próxima. Quando se encontrar uma variável significativa, a eq. de regressão anterior à retirada é a opção.**
- **'*Step forward regression*' → inicia com apenas a variável + importante. A seguinte + importante (*r* ou $r_{parcial}$) é incluída e calcula-se o $F_{parcial}$. Se a variável é significativa, repete-se o processo para a próxima. Quando se encontrar uma variável não significativa, a eq. de regressão anterior à inclusão é a opção.**
- **'*Stepwise regression*' → combina as etapas *back* e *forward*. Vai um passo *forward*, seguido de um *backward*. Quando todas as possibilidades tiverem sido testadas, a mais significativa é a opção.**

MODELOS NÃO LINEARES

$$Y = \beta_0 . X_1^{\beta_1} . X_2^{\beta_2} . \varepsilon$$

$$ln\,Y = ln\,\beta_0 + \beta_1\, ln\,X_1 + \beta_2\, ln\,X_2 + ln\,\varepsilon$$

$$Y = e^{(\beta_0 + \beta_1 X_1 + \beta_2 X_2)} . \varepsilon$$

$$ln\,Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ln\,\varepsilon$$

O procedimento para análise da RLM:

Etapa 1 Selecione as variáveis preditoras (Xi) que estão relacionadas à variável a ser prevista (Y) por alguma relação física.

Etapa 2 Plote as variáveis preditoras (Xi) em relação à variável a ser prevista (Y)

Etapa 3 Determine a forma da equação desejada; isto é, linear ou curvilíneo.

Etapa 4 Calcule os coeficientes de correlação entre as variáveis. Matriz de correlação.

Etapa 5 Calcule os coeficientes de regressão.

No EXCEL: Função PROJ.LIN(), PROJ.LOG() e a ferramenta Análise de Dados/Regressão

Etapa 6 Calcule o erro padrão da estimativa, Se; desvio padrão da variável a ser prevista, Sy; e o coeficiente de determinação, r^2; o r^2 parcial.

O procedimento para análise da RLM:

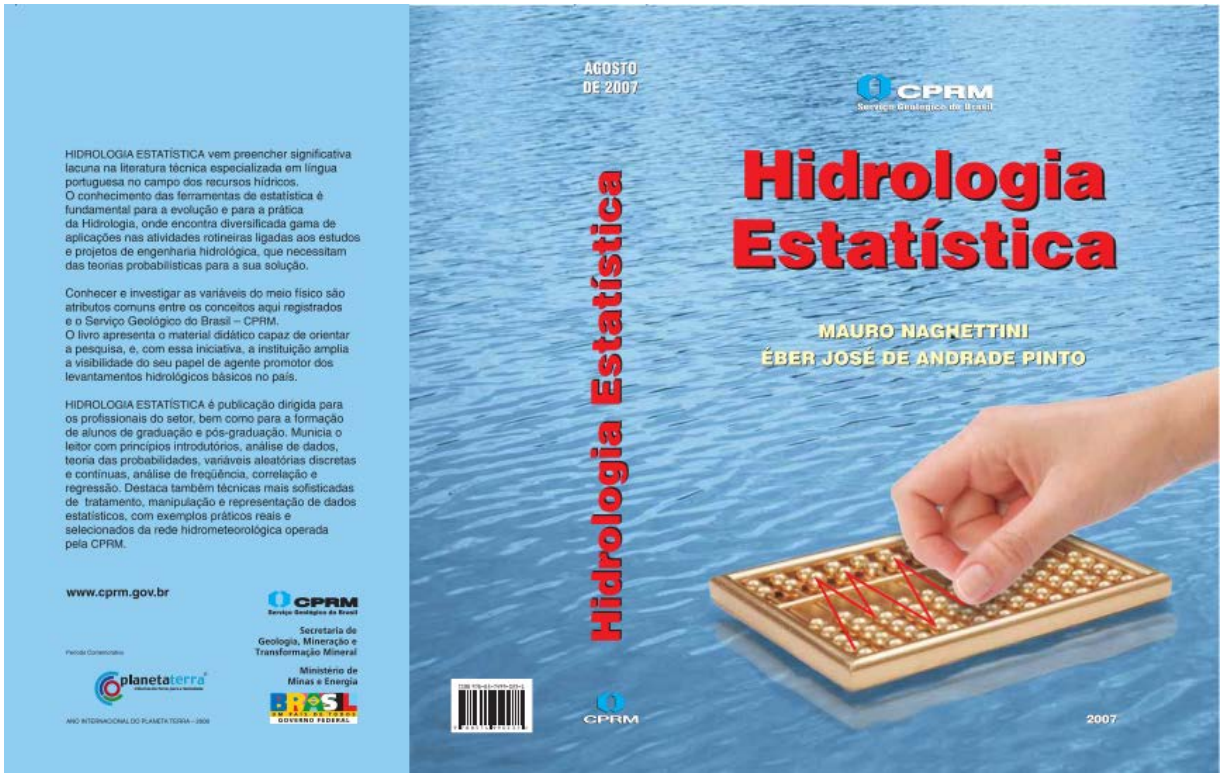
Etapa 7 Avalie a equação de regressão pelos seguintes métodos:

- O erro padrão da estimativa tem os limites $0 \leq Se \leq Sy$; se Se -> 0 maior parte da variância é explicada pela regressão.
- Coeficiente de determinação tem limites $0 \leq r^2 \leq 1$; quando r^2 -> 1, melhor será o “ajuste” da linha de regressão aos dados.
- Os testes F parciais e totais são usados para avaliar cada preditor e a significância total da equação.
- O sinal de cada coeficiente de regressão deve ser comparado com o coeficiente de correlação para o critério de predição apropriado. Os sinais devem ser os mesmos.
- Examine os resíduos para identificar deficiências na equação de regressão e verifique as suposições do modelo.

Etapa 8 Se a precisão da equação de regressão não for aceitável, reformule a equação de regressão ou transforme as variáveis. Uma solução satisfatória nem sempre é possível a partir dos dados disponíveis.

Recomendações

Para consolidar conhecimentos estudar no livro texto o iten 9.8



http://www.cprm.gov.br/publique/Hidrologia/Mapas-e-Publicacoes/Livro-%22Hidrologia-Estatistica%22-981.html

Serviço Geológico do Brasil – CPRM

Departamento de Hidrologia da CPRM

Eber José de Andrade Pinto
Coordenador Executivo do DEHID
eber.andrade@cprm.gov.br
www.cprm.gov.br

Belo Horizonte, 22 de outubro de 2020