Multicolinearidade

Gilberto A. Paula

Departamento de Estatística
IME-USP, Brasil
giapaula@ime.usp.br

$1^{\underline{o}}$ Semestre 2023

Sumário

Objetivos

Neste material serão apresentados os seguintes tópicos relacionados à multicolinearidade em Regressão Linear Múltipla:

- Ortogonalidade
- Fontes de Multicolinearidade
- Efeitos da Multicolinearidade
- Como detectar Multicolineridade
- Tratamentos da Multicolinearidade
- Referências

Sumário

Regressão Linear Múltipla

Supor o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ são valores observados da variável resposta, $x_{i1}, x_{i2}, \ldots, x_{ip}$ são valores observados de variáveis explicativas e $\epsilon_i \stackrel{\text{iid}}{\sim} \text{N}(0, \sigma^2)$.

Regressão Linear Múltipla

Tem-se ortogonalidade entre as colunas da matriz modelo **X** se

$$\sum_{i=1}^{n} x_{ij}x_{i\ell} = 0,\ \forall j \neq \ell = 1, \ldots, p.$$

Quando a matriz modelo **X** tem posto coluna completo tem-se sob ortogonalidade que

$$\mathbf{X}^\top\mathbf{X} = \mathrm{diag}\{\sum_{i=1}^{n} x_{i1}^2, \ldots, \sum_{i=1}^{n} x_{ip}^2\} \ \text{e} \ \mathbf{X}^\top\mathbf{y} = (\sum_{i=1}^{n} x_{i1}y_i, \ldots, \sum_{i=1}^{n} x_{ip}y_i)^\top,$$

em que $\mathbf{y} = (y_1, \ldots, y_n)^\top$.

Regressão Linear Múltipla

Portanto, sob ortogonalidade, segue que

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_1 \\ \vdots \\ \widehat{\beta}_p \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} = \begin{bmatrix} \frac{\sum_{i=1}^n x_{i1} y_i}{\sum_{i=1}^n x_{i1}^2} \\ \vdots \\ \frac{\sum_{i=1}^n x_{ip} y_i}{\sum_{i=1}^n x_{ip}^2} \end{bmatrix}.$$

Portanto, $\widehat{\beta}_j = \frac{\sum_{i=1}^n x_{ij} y_i}{\sum_{i=1}^n x_{ij}^2}$ depende apenas dos valores $y_1, \ldots, y_n$ e de $x_{1j}, \ldots, x_{nj}$, para $j = 1, \ldots, p$.

Regressão Linear Múltipla

Além disso, a matriz de variância-covariância para $\widehat{\beta}$ fica dada por

$$\mathrm{Var}(\widehat{\beta}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{bmatrix} \frac{\sigma^2}{\sum_{i=1}^n x_{i1}^2} & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & \frac{\sigma^2}{\sum_{i=1}^n x_{ip}^2} \end{bmatrix}.$$

Portanto, $\mathrm{Var}(\widehat{\beta}_j) = \frac{\sigma^2}{\sum_{i=1}^n x_{ij}^2}$ e $\mathrm{Cov}(\widehat{\beta}_j, \widehat{\beta}_\ell) = 0$, para $j \neq \ell$ e $j, \ell = 1, \ldots, p$. Tem-se independência mútua entre os estimadores dos coeficientes.

Regressão Linear Múltipla

Em particular, quando tem-se um intercepto na regressão linear múltipla segue que $\widehat{\beta}_1 = \overline{y}$ e $\text{Var}(\widehat{\beta}_1) = \frac{\sigma^2}{n}$.

Multicolinearidade

Multicolinearidade é o oposto da ortogonalidade. Ocorre quando há uma alta correlação linear entre variáveis explicativas e consequentemente entre os estimadores dos coeficientes da regressão linear múltipla. Uma consequência prática é que $\det(\mathbf{X}^\top\mathbf{X}) \cong 0$.

Sumário

Regressão Linear Múltipla

Agumas fontes de multicolinearidade:

- Método empregado na coleta de dados
  Os dados são coletados de um estrato da população onde há uma alta correlação linear entre duas variáveis explicativas. Por exemplo, num estudo de regressão em que tem-se como variáveis explicativas o consumo de um produto alimentício e o preço do produto alimentício. É razoável esperar nos estratos de renda mais baixa uma correlação mais alta entre as duas variáveis explicativas.
- Restrições no modelo ou na população
  Duas variáveis explicativas que têm uma correlação linear alta são incluídas no modelo. Por exemplo, consumo de energia elétrica e renda percapita. Notas referentes às avaliações sobre qualidade e clareza das aulas de um instrutor.

Regressão Linear Múltipla

- Especificação do modelo
  No modelo são incluídos vários termos que estão em função de uma mesma variável explicativa. Por exemplo, numa regressão polinomial em que são incluídos termos $x + x^2 + x^3 + \cdots$.
- Modelo superdimensionado
  Estudos com amostras pequenas e uma grande quantidade de variáveis explicativas. Por exemplo, na área médica em geral tem-se amostras pequenas com uma grande quantidade de informações por paciente.

Sumário

Avaliação de Docentes

Como motivação serão considerados os dados apresentados no livro de Weisberg (2014) e também na biblioteca `alr4` do R, referentes a notas médias recebidas por 364 instrutores de uma universidade norte americana durante um período de 10 anos. O objetivo do estudo é relacionar o interesse do avaliador (`RaterInterest`) (escore de 1 a 5) com as seguintes avaliações feitas pelo avaliador:

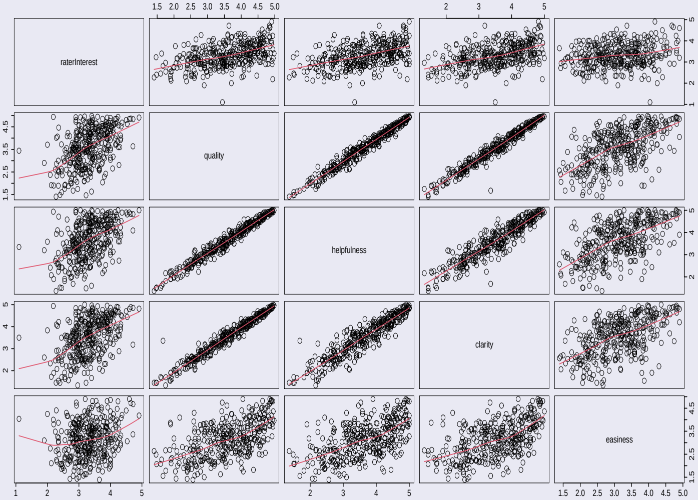
- `Quality`: qualidade das aulas do instrutor (escore de 1 a 5)
- `Helpfulness`: prestatividade do instrutor (escore de 1 a 5)
- `Clarity`: clareza das aulas do instrutor (escore de 1 a 5)
- `Easiness`: facilidade que o instrutor tem com a matéria (escore de 1 a 5).

Matriz de Correlações Lineares

	RInterest	Quality	Helpfulness	Clarity	Easiness
RInterest	1,0000	0,4707	0,4630	0,4611	0,2052
Quality		1,0000	0,9810	0,9760	0,5651
Helpfulness			1,0000	0,9208	0,5635
Clarity				1,0000	0,5359
Easiness					1,0000

Notas-se altas correlações lineares entre Quality e Helpfulness e entre Quality e Clarity.

Diagramas de Dispersão



Sumário

Exemplo

Para ilustrar considere o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ são valores observados da variável resposta com comprimento unitário, x_{i1} e x_{i2} são valores observados de variáveis explicativas com comprimento unitário[a] e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$, para $i = 1, \ldots, n$.

[a] $\sum_{i=1}^{n} x_{ij} = 0$ e $\sum_{i=1}^{n} x_{ij}^2 = 1$ para $j = 1, 2$

Exemplo

Para esse exemplo tem-se que

$$\mathbf{X}^\top\mathbf{X} = \begin{bmatrix} \sum_{i=1}^n x_{i1}^2 & \sum_{i=1}^n x_{i1}x_{i2} \\ \sum_{i=1}^n x_{i2}x_{i1} & \sum_{i=1}^n x_{i2}^2 \end{bmatrix} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix},$$

em que r_{12} denota a correlação linear amostral entre X_1 e X_2. Além disso

$$\mathbf{X}^\top\mathbf{y} = \begin{bmatrix} \sum_{i=1}^n x_{i1}y_i \\ \sum_{i=1}^n x_{i2}y_i \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix},$$

em que r_{1y} e r_{2y} denotam, respectivamente, as correlações lineares amostrais entre X_1 e Y e X_2 e Y.

Exemplo

Portanto, as estimativas de mínimos quadrados ficam dadas por

$$\widehat{\boldsymbol{\beta}} = \begin{bmatrix} \widehat{\beta}_1 \\ \widehat{\beta}_2 \end{bmatrix} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y} = \begin{bmatrix} \frac{r_{1y} - r_{12} r_{2y}}{(1 - r_{12}^2)} \\ \frac{r_{2y} - r_{12} r_{1y}}{(1 - r_{12}^2)} \end{bmatrix}.$$

Portanto, as estimativas de mínimos quadrados $\widehat{\beta}_1$ e $\widehat{\beta}_2$ dependem das correlações lineares r_{12}, r_{1y} e r_{2y}.

Exemplos

Além disso, a matriz de variância-covariância para $\widehat{\beta}$ assume a forma

$$\text{Var}(\widehat{\beta}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1} = \begin{bmatrix} \frac{\sigma^2}{(1-r_{12}^2)} & -\frac{\sigma^2 r_{12}}{(1-r_{12}^2)} \\ -\frac{\sigma^2 r_{12}}{(1-r_{12}^2)} & \frac{\sigma^2}{(1-r_{12}^2)} \end{bmatrix}.$$

Ou seja, $\text{Var}(\widehat{\beta}_1) = \text{Var}(\widehat{\beta}_2) = \frac{\sigma^2}{(1-r_{12}^2)}$ e $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) = -\frac{\sigma^2 r_{12}}{(1-r_{12}^2)}$.

Consequências

- Se $|r_{12}| \to 1$ então $\text{Var}(\widehat{\beta}_1)$ e $\text{Var}(\widehat{\beta}_2)$ ficam grandes.
- Se $r_{12} \to 1$ então $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) \to -\infty$.
- Se $r_{12} \to -1$ então $\text{Cov}(\widehat{\beta}_1, \widehat{\beta}_2) \to \infty$.

Sumário

Fator de Inflação da Variância

Supor o seguinte modelo de regressão linear múltipla:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

em que $y_1, \ldots, y_n$ são valores observados da variável resposta, $x_{i1}, x_{i2}, \ldots, x_{ip}$ são valores observados de variáveis explicativas com comprimento unitário[a] e $\epsilon_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$.

[a] $(\mathbf{X}^\top \mathbf{X})$ é a matriz de correlações.

Fator de Inflação da Variância

Seja $\mathbf{C} = (\mathbf{X}^\top\mathbf{X})^{-1} = \{C_{j\ell}\}$, em que $C_{j\ell}$ denota o (j, ℓ)-ésimo elemento da matriz $\mathbf{C}$, para $j, \ell = 1, \ldots, p$. É possível mostrar que

$$\mathrm{Var}(\widehat{\beta}_j) = \sigma^2 C_{jj} = \sigma^2(1 - \mathrm{R}_j^2)^{-1},$$

em que R_j^2 denota o coeficiente de determinação da regressão linear da variável explicativa X_j contra as demais variáveis explicativas X_ℓ, em que $j \neq \ell$, para $j, \ell = 1, \ldots, p$.

Fator de Inflação da Variância

O fator de inflação de variância da j-ésima variável explicativa é definido por

$$\mathrm{VIF}_j = (1 - \mathrm{R}_j^2)^{-1}.$$

Assim, se $\mathrm{R}_j^2 \to 1$ então $\mathrm{VIF}_j \to \infty$, para $j = 1, \ldots, p$.

Fator de Inflação da Variância

Supor três variáveis explicativas X_1, X_2 e X_3 cujos valores amostrais têm comprimento unitário. Os VIFs saem das seguintes regressões:

- VIF_1: da regressão $x_{i1} = \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$
- VIF_2: da regressão $x_{i2} = \beta_1 x_{i1} + \beta_3 x_{i3} + \epsilon_i$
- VIF_3: da regressão $x_{i3} = \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$, para $i = 1, \ldots, n$.

Critério: se $\text{VIF}_j \geq 10$ indica que $\widehat{\beta}_j$ está com variância inflacionada.

Análise de Autovalores de $(\mathbf{X}^\top\mathbf{X})$

Sejam $\lambda_1, \lambda_2, \ldots, \lambda_p$ os autovalores da matrix $(\mathbf{X}^\top\mathbf{X})$. Autovalores próximos de zero é indício de multicolinearidade.

Número da Condição

Uma medida resumo de multicolinearidade entre as colunas da matriz $\mathbf{X}$ é o número da condição definido por

$$k = \frac{\lambda_{\max}}{\lambda_{\min}}.$$

Critério: (i) se $k \leq 100$ não há indícios de multicolinearidade, (ii) se $100 < k \leq 1000$ há indícios moderados de multicolinearidade e (iii) se $k > 1000$ há indícios fortes de multicolinearidade.

Índice da Condição

Quando há indícios de multicolinearidade através do número da condição, pode-se avaliar a contribuição de cada variável explicativa através do índice da condição definido por

$$k_j = \frac{\lambda_{\text{max}}}{\lambda_j},$$

para $j = 1, \ldots, p$. Os mesmos critérios usados para o número da condição são usados para o índice da condição.

Determinante da Matrix $(\mathbf{X}^\top\mathbf{X})$

Se as variáveis explicativas têm comprimento unitário então

$$0 \leq \det(\mathbf{X}^\top\mathbf{X}) \leq 1.$$

Tem-se que $\det(\mathbf{X}^\top\mathbf{X}) = 1$ indica ortogonalidade entre as colunas da matriz $\mathbf{X}$, enquanto $\det(\mathbf{X}^\top\mathbf{X}) = 0$ indica dependência linear entre as colunas da matrix $\mathbf{X}$. Então, valores próximos de zero são indícios de multicolinearidade.

Sumário

Regressão Linear Múltipla

Alguns tratamentos para a multicolinearidade

- Coletar mais dados.
- Eliminação de variáveis explicativas.
- Transformação de variáveis explicativas.
- Regressão Ridge.
- Regressão através de componentes principais.

Regressão Ridge

O objetivo da regressão ridge é utilizar um estimador tendencioso que produza variâncias mais estáveis para os estimadores dos coeficientes da regressão.

Erro Quadrático Médio

Seja $\widehat{\boldsymbol{\beta}}^*$ um estimador tendencioso de $\boldsymbol{\beta}$. O erro quadrático médio de $\widehat{\boldsymbol{\beta}}^*$ pode ser expresso na forma

$$\text{EQM}(\widehat{\boldsymbol{\beta}}^*) = \text{Var}(\widehat{\boldsymbol{\beta}}^*) + [\text{Viés}][\text{Viés}]^\top,$$

em que $\text{Viés} = \text{E}(\widehat{\boldsymbol{\beta}}^*) - \boldsymbol{\beta}$.

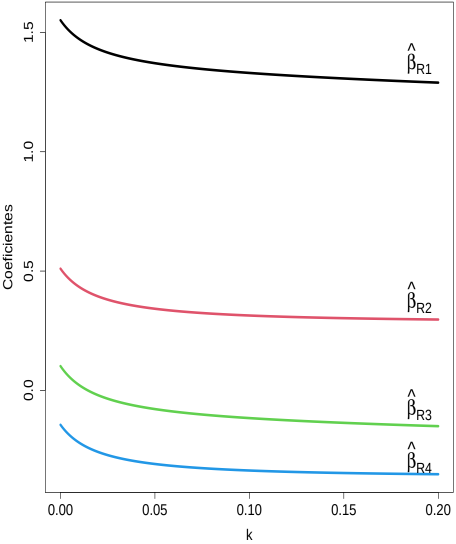
Estimador Ridge

A fim de estabilizar as estimativas dos coeficientes da regressão linear múltipla bem com as respectivas variâncias é proposto o seguinte estimador:

$$\widehat{\beta}_R = (\mathbf{X}^\top\mathbf{X} + k\mathbf{I}_p)^{-1}\mathbf{X}^\top\mathbf{y},$$

em que $k > 0$ é uma constante desconhecida que é estimada separadamente. Em particular quando $k = 0$ recupera-se o estimador de mínimos quadrados. Estima-se k até estabilizar as estimativas dos coeficientes.

Ilustração Regressão Ridge



Propriedades

Denote $\widehat{\boldsymbol{\beta}}_R = \mathbf{Z}_k\widehat{\boldsymbol{\beta}}$, em que $\mathbf{Z}_k = (\mathbf{X}^\top\mathbf{X} + k\mathbf{I}_p)^{-1}(\mathbf{X}^\top\mathbf{X})$. Tem-se as seguintes propriedades:

- $\mathrm{E}(\widehat{\boldsymbol{\beta}}_R) = \mathrm{E}(\mathbf{Z}_k\widehat{\boldsymbol{\beta}}) = \mathbf{Z}_k\mathrm{E}(\widehat{\boldsymbol{\beta}}) = \mathbf{Z}_k\boldsymbol{\beta}$.
- $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_R) = \mathrm{Var}(\mathbf{Z}_k\widehat{\boldsymbol{\beta}}) = \mathbf{Z}_k\mathrm{Var}(\widehat{\boldsymbol{\beta}})\mathbf{Z}_k^\top = \sigma^2(\mathbf{X}^\top\mathbf{X} + k\mathbf{I}_p)^{-1}(\mathbf{X}^\top\mathbf{X})(\mathbf{X}^\top\mathbf{X} + k\mathbf{I}_p)^{-1}$.

Em particular, se $(\mathbf{X}^\top\mathbf{X}) = \mathbf{I}_p$ tem-se que $\mathbf{Z}_k = (1+k)^{-1}\mathbf{I}_p$. Portanto, $\mathrm{E}(\widehat{\boldsymbol{\beta}}_R) = (1+k)^{-1}\boldsymbol{\beta}$ e $\mathrm{Var}(\widehat{\boldsymbol{\beta}}_R) = \sigma^2(1+k)^{-2}\mathbf{I}_p$. Ou seja, à medida que k cresce o estimador ridge fica mais tendencioso havendo um encolhimento com relação ao estimador de mínimos quadrados. A variância diminui com o aumento de k.

Propriedades

- Tem-se ainda que $\widehat{\boldsymbol{\beta}}_R \sim \mathrm{N}_p(\mathrm{E}(\widehat{\boldsymbol{\beta}}_R), \mathrm{Var}(\widehat{\boldsymbol{\beta}}_R))$.
  Daí segue que $\widehat{\beta}_{R_j}$ são normais de média $\mathrm{E}(\widehat{\beta}_{R_j})$ e variância $\mathrm{Var}(\widehat{\beta}_{R_j})$, para $j = 1, \ldots, p$.
- É possível mostrar que $\mathrm{SQRes}(k) = (\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_R)^\top(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}_R) = \mathrm{SQRes} + (\widehat{\boldsymbol{\beta}}_R - \widehat{\boldsymbol{\beta}})^\top(\mathbf{X}^\top\mathbf{X})(\widehat{\boldsymbol{\beta}}_R - \widehat{\boldsymbol{\beta}})$.

Portanto, com a regressão ridge há um aumento na soma de quadrados de resíduos, logo uma redução no valor de R^2.

Estimação de k

A constante k pode ser estimada através do processo iterativo

$$k^{(m+1)} = \frac{p\widehat{\sigma}^2}{\widehat{\beta}_R^\top(k^{(m)})\widehat{\beta}_R(k^{(m)})},$$

para $m = 0, 1, \ldots$, em que $\widehat{\sigma}^2$ é obtido através do estimador de mínimos quadrados $\widehat{\beta}$. Para valor inicial utiliza-se o estimador de HKB (Hoerl, Kennard e Baldwin, 1975) dado por $k^{(0)} = p\widehat{\sigma}^2/\widehat{\beta}^\top\widehat{\beta}$.

Procedimentos de Diagnóstico

Métodos de diagnóstico para a regressão ridge podem ser obtidos dos métodos de diagnóstico para regressão não paramétrica.

Forma Canônica

A forma canônica da regressão linear múltipla $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ é definida por

$$\mathbf{y} = \mathbf{Z}\boldsymbol{\alpha} + \boldsymbol{\epsilon},$$

em que $\boldsymbol{\epsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2\mathbf{I}_n)$, $\mathbf{Z} = \mathbf{X}\mathbf{T}$, $\boldsymbol{\alpha} = \mathbf{T}^\top\boldsymbol{\beta}$ e $\mathbf{Z}^\top\mathbf{Z} = \mathbf{T}^\top\mathbf{X}^\top\mathbf{X}\mathbf{T} = \boldsymbol{\Lambda}$, com $\boldsymbol{\Lambda} = \mathrm{diag}\{\lambda_1, \ldots, \lambda_p\}$ sendo a matriz diagonal $p \times p$ com os autovalores da matriz $\mathbf{X}^\top\mathbf{X}$ e $\mathbf{T}$ a matriz $p \times p$ cujas colunas são os autovetores ortogonais correspondentes aos autovalores $\lambda_1, \ldots, \lambda_p$. Como $\mathbf{X}^\top\mathbf{X}$ é positiva definida todos os seus autovalores são positivos. Sugere-se que $\mathbf{y}$ e a matriz $\mathbf{X}$ sejam centralizadas, assim não precisa de intercepto.

Efeito na Variância de $\widehat{\alpha}_j$

Portanto, tem-se que

$$
\begin{aligned}
\widehat{\boldsymbol{\alpha}} &= (\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{y} \\
&= \boldsymbol{\Lambda}^{-1}\mathbf{Z}^\top\mathbf{y} \ \text{ e} \\
\text{Var}(\widehat{\boldsymbol{\alpha}}) &= \sigma^2(\mathbf{Z}^\top\mathbf{Z})^{-1} \\
&= \sigma^2\boldsymbol{\Lambda}^{-1}.
\end{aligned}
$$

Pode-se mostrar que $\text{Var}(\widehat{\alpha}_j) = \sigma^2\lambda_j^{-1}$. Assim, λ_j próximo de zero inflaciona a variância de $\widehat{\alpha}_j$.

Efeito na Variância de $\widehat{\beta}_j$

Similarmente, segue que

$$\begin{aligned} \text{Var}(\widehat{\boldsymbol{\beta}}) &= \text{Var}(\mathbf{T}\widehat{\boldsymbol{\alpha}}) \\ &= \mathbf{T}\text{Var}(\widehat{\boldsymbol{\alpha}})\mathbf{T}^\top \\ &= \sigma^2\mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}^\top. \end{aligned}$$

E daí pode-se mostrar que $\text{Var}(\widehat{\beta}_j) = \sigma^2 \sum_{\ell=1}^{p} t_{j\ell}^2/\lambda_\ell$, em que $t_{j\ell}$ denota o (j,ℓ)-ésimo elemento da matriz $\mathbf{T}$. Verifica-se aqui também o efeito de autovalores próximos de zero na variância de $\widehat{\beta}_j$.

Seleção dos Coeficientes

A proposta é considerar os coeficientes estimados

$$\widehat{\alpha}_1^{CP}, \ldots, \widehat{\alpha}_{p-s}^{CP}$$

com os $p - s$ maiores autovalores e os demais s coeficientes como sendo iguais a zero. Assim, da relação $\widehat{\boldsymbol{\beta}} = \mathbf{T}\widehat{\boldsymbol{\alpha}}$ obtém-se os coeficientes estimados

$$\widehat{\beta}_1^{CP}, \ldots, \widehat{\beta}_p^{CP}$$

que serão considerados na regressão linear múltipla. Ou seja, os coeficientes estimados segundo o critério de componentes principais serão dados por $\widehat{\boldsymbol{\beta}}^{CP} = \mathbf{T}\widehat{\boldsymbol{\alpha}}^{CP}$. Porém, os erros padrão desses coeficientes estimados devem ser extraidos da relação $\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2\mathbf{T}\boldsymbol{\Lambda}^{-1}\mathbf{T}^\top$.

Interpretação dos Componentes Principais

Da relação $\mathbf{Z} = \mathbf{XT}$ segue que

$$\mathbf{Z}_j = \sum_{\ell=1}^{p} \mathbf{X}_\ell t_{\ell j},$$

em que $\mathbf{Z}_1, \ldots, \mathbf{Z}_p$ e $\mathbf{X}_1, \ldots, \mathbf{X}_p$ denotam, respectivamente, as colunas de $\mathbf{Z}$ e $\mathbf{X}$, enquanto $t_{1j}, \ldots, t_{pj}$ denotam os componentes do autovetor correspondente ao autovalor λ_j.Assim, se λ_j for próximo de zero os componentes de $\mathbf{Z}_j$ devem ser aproximadamente constantes. Deve-se portanto escolher os $p - s$ componentes principais $\mathbf{Z}_1, \ldots, \mathbf{Z}_{(p-s)}$ que correspondem aos $p - s$ maiores autovalores.

Sumário

Referências

- Montgomery, D. C.; Peck, E. A. e Vining, G. G. (2021). *Introduction to Linear Regression Analysis, 6th Edition*.
- Weisberg, S. (2014) *Applied Linear Regression, 5th Edition*. Wiley. Hoboken: Wiley.