# Detecção de Multicolinearidade em Modelos de Regressão por Autovetores e Autovalores

Felipe Ferreira  Fernanda Rahal  Fernando Fukui  Marcelo Baraldo
Matheus Siniscarchio  Vitor Jensen

Instituto de Matemática e Estatística (IME)
Universidade de São Paulo (USP)

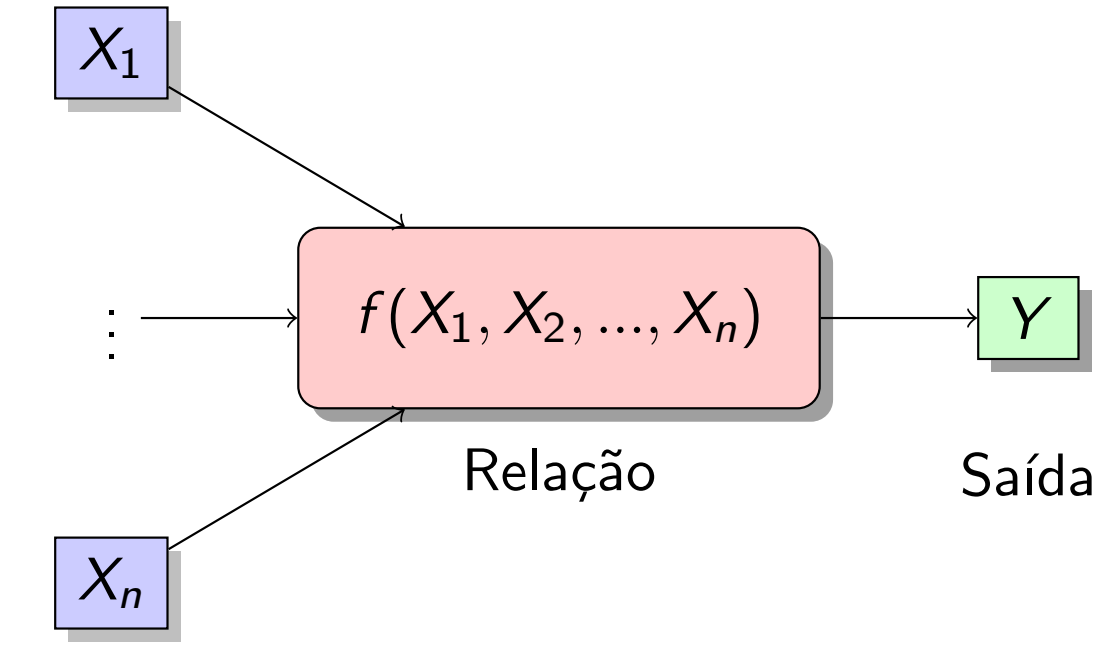
11 de julho de 2022

# Sumário

# Introdução

- Estamos interessados em estudar a relação de uma variável chamada de **regressora**, **independente**, **explicativa** ou **X** *(inputs)* e outra chamada de **dependente**, **explicada**, **resposta** ou **Y** *(output)*.
- Essa relação é representada por um modelo matemático que associa essas duas variáveis. O modelo assume que a função de regressão $E(Y|X)$ é linear.

# Introdução



# Regressão Linear

Nosso cenário é descrito por um vetor com valores de observações das variáveis *(inputs)* $X = (X_1, X_2, ..., X_n)$ e queremos prever o verdadeiro valor da variável explicada $Y$ *(outputs)*. O modelo de regressão linear é da forma:

$$Y = f(X) = \beta_0 + \sum_{i=1}^{n} X_i \beta_i + \varepsilon. \tag{1}$$

onde:

- Os $\beta_j$'s são parâmetros desconhecidos.
- as variáveis $X_i$ podem vir de diferentes fontes como: observações quantitativas, transformações em observações quantitativas como log, raiz quadrada ou o quadrado.
- O modelo assume que a função de regressão $E(Y|X)$ é linear.

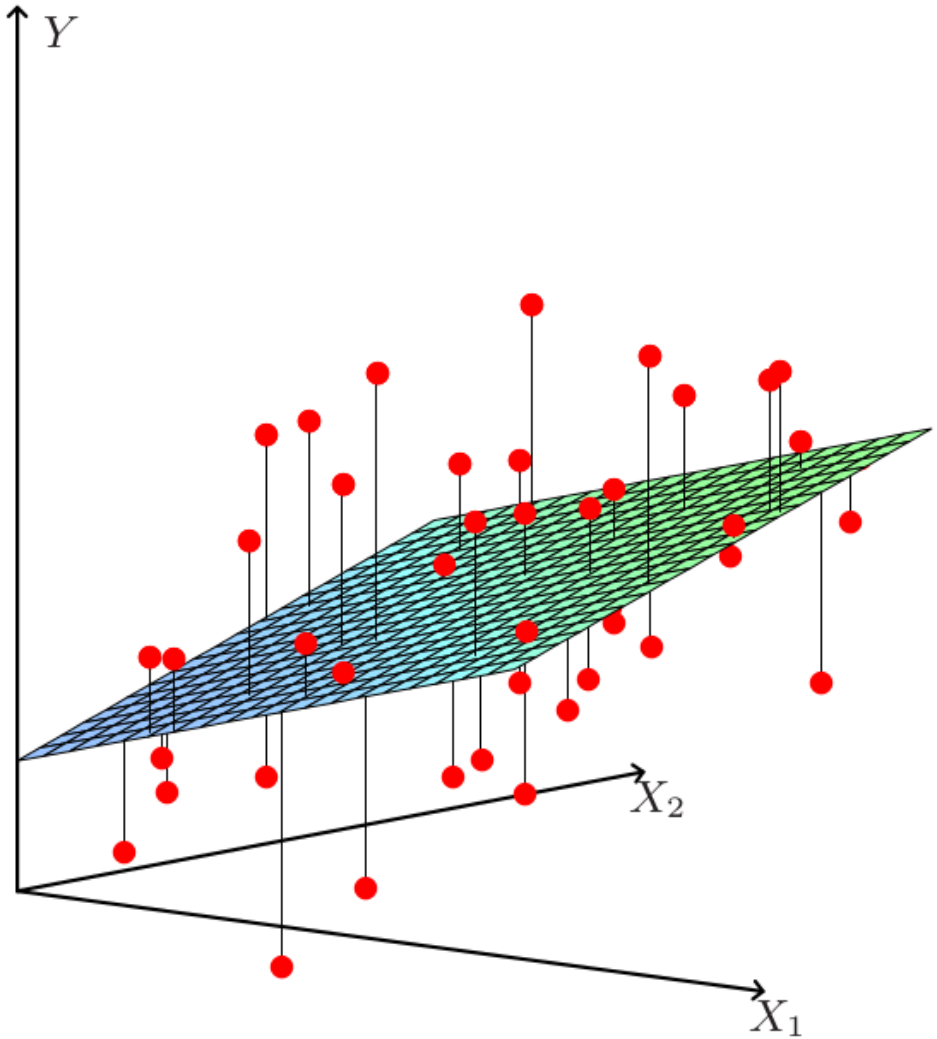
# Método dos Mínimos Quadrados

- Nós escolhemos os coeficientes $\beta = (\beta_o, \beta_1, ...\beta_n)^T$ de modo que minimize o resíduo do quadrado da soma.

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{P} (y_i - f(x_i))^2 \\ &= \sum_{i=1}^{P} \Big( y_i - \beta_0 - \sum_{j=1}^{n} (x_{ij}\beta_j) \Big)^2. \end{aligned} \tag{2}$$

# Método dos Mínimos Quadrados

Figura: Mínimos quadrados encaixando plano em $X \in \mathbb{R}^2$.



# Método dos Mínimos Quadrados

Relembrando, nós procuramos uma função de $X$ que minimize o resíduo da soma dos quadrados de $Y$:

$$\text{RSS}(\beta) = \sum_{i=1}^{P} \left( y_i - \beta_0 - \sum_{j=1}^{n} (x_{ij}\beta_j) \right)^2 \tag{3}$$

Reescrevendo (2) na forma matricial:

$$\text{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T (\mathbf{y} - \mathbf{X}\beta) \tag{4}$$

# Método dos Mínimos Quadrados

Assumindo que $\mathbf{X}$ tem posto coluna completo, então $\mathbf{X}^T\mathbf{X}$ será positiva definida, igualando a segunda derivada à zero:

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0 \tag{5}$$

e obtemos a única solução

$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}. \tag{6}$$

# Método dos Mínimos Quadrados

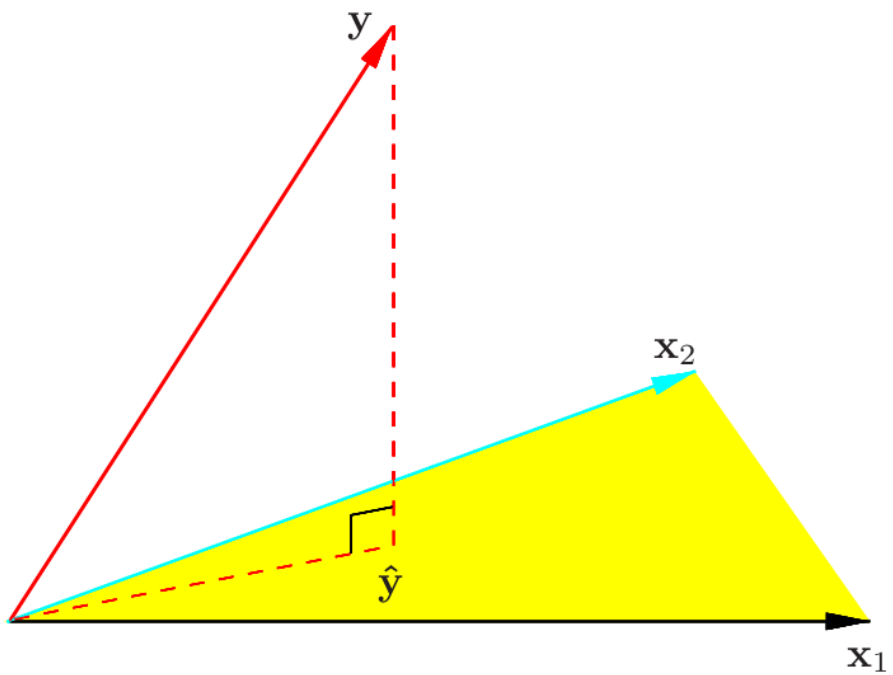
Os valores previstos em um vetor de entrada $x_0$ são dados por $\hat{f} = (1 : x_0)^T \hat{\beta}$; os valores encaixados pelas entradas são

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} \tag{7}$$

onde $\hat{y}_i = \hat{f}(x_i)$. A matriz $\mathrm{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ em (7) é chamada de matriz de projeção.

# Método dos Mínimos Quadrados

A geometria P-dimensional da regressão de mínimos quadrados com dois preditores. O vetor resultante $\mathbf{y}$ é ortogonalmente projetado sobre o hiperplano gerado pelos vetores de entrada $\mathbf{x}_1$ e $\mathbf{x}_2$. A projeção $\hat{\mathbf{y}}$ representa o vetor previsto pelos mínimos quadrados.



**Recapitulando**

- Denotamos os vetores colunas de $\mathbf{X}$ por $\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_n$, onde $\mathbf{x}_0 \equiv 1$
- Esses vetores geram um subespaço de $\mathbb{R}^P$, conhecido como espaço coluna de $\mathbf{X}$.
- Nós minimizamos $\text{RSS}(\beta) = \|\mathbf{y} - \mathbf{X}\beta\|^2$ escolhendo um $\hat{\beta}$ tal que o vetor resíduo $\mathbf{y} - \hat{\mathbf{y}}$ seja ortogonal a esse subespaço.
- $\mathbf{H}$ é a matriz de projeção.

# Multicolinearidade

Modelo de regressão linear multiplo:

$$y = X\beta + \varepsilon \tag{8}$$

Quando existem dependências lineares ou **proximas** de serem lineares entre os regressores, diz-se que existe o problema da multicolinearidade.
Seja a $j$ coluna da matrix X, tal que $X = [X_1, X_2, ..., X_p]$

$$\sum_{j=1}^{p} t_j X_j = 0 \tag{9}$$

Os vetores $X_1, X_2, ..., X_p$ são linearmente dependentes se temos um conjunto de constantes $t_1, t_2, ..., t_p$, com $t_i$ não todos nulos.
Se a equação (2) ocorrer em algum subconjunto das colunas de X, então o rank de $X'X$ é menor que p e $(X'X)^{-1}$ não existe. Portanto, temos problema para estimar $\beta$

# Multicolinearidade

Suponha que a equação (2) é aproximadamente verdade para algum subconjunto das colunas de X. Então teremos aproximadamente uma dependência linear em $X'X$ e o problema da multicolinearidade existe.

# Tipos de multicolinearidade

- Multicolinearidade estrutural - feature engineering
- Multicolinearidade baseada nos dados
  - Experimentos mal projetados
  - Dependência de dados puramente observacionais
  - Incapacidade de manipular o sistema no qual os dados são coletados

O primeiro é mais facíl de corrigir, nfelizmente, o mais comum na pratica é o segundo.

1. Relembrando Regressão Linear Simples

2. Multicolinearidade
   - Tipos de multicolinearidade
   - Efeitos da Multicolinearidade

3. Métodos para detectar multicolinearidade

# Efeitos da Multicolinearidade

A presença de multicolinearidade tem potencial de causar sérios efeitos nos coeficientes da regressão dos mínimos quadrados.

# Efeitos da Multicolinearidade

Suponha que existam duas variáveis regressoras, $x_1$ e $x_2$. O modelo, assumindo que $x_1$, $x_2$ e $y$ estão normalizados, é:

$$y = \beta_1 x_1 + \beta_2 x_2 + \epsilon \tag{10}$$

E as equações normais de mínimos quadrados são:

$$X'X\hat{\beta} = X'y$$

$$\begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} r_{1y} \\ r_{2y} \end{bmatrix}$$

- $r_{12}$ é a correlação entre $x_1$ e $x_2$.
- $x_{jy}$ é a correlação entre $x_j$ e $y$, onde $j = 1, 2$.

# Efeitos da Multicolinearidade

$$C = X'X^{-1} = \begin{bmatrix} \frac{1}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix} \tag{11}$$

E as estimativas dos coeficientes de regressão são:

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{1-r_{12}^2}, \quad \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{1-r_{12}^2}$$

# Efeitos da Multicolinearidade

Podemos fazer algumas conclusões nesse exemplo:

$$\hat{\beta}_1 = \frac{r_{1y} - r_{12}r_{2y}}{1 - r_{12}^2}, \quad \hat{\beta}_2 = \frac{r_{2y} - r_{12}r_{1y}}{1 - r_{12}^2}$$

- Se há uma forte multicolinearidade entre $x_1$ e $x_2$, então o coeficiente de correlação $r_{12}$ vai ser grande;
- Na matriz inversa $C$, note que se $|r_{12}| \to 1$, então $Var(\hat{\beta}_j) = C_{jj}\sigma^2 \to \infty$ e $Cov(\hat{\beta}_1, \hat{\beta}_2) = C_{12}\sigma^2 \to \pm\infty$

# Efeitos da Multicolinearidade

- Forte multicolinearidade entre $x_1$ e $x_2$ resulta em altas variâncias e covariâncias para os estimadores de mínimos quadrados dos coeficientes de regressão.
- Isso implica que diferente amostras tomadas nos mesmos níveis de $x$ podem levar a estimativas muito diferentes dos parâmetros do modelo.

Quando há mais de duas variáveis regressoras, a multicolinearidade produz efeitos similares.

# Efeitos da Multicolinearidade

Podemos mostrar que os elementos da diagonal de $C = (X'X)^{-1}$ são:

$$C_{jj} = \frac{1}{1 - R_j^2}, \quad j = 1, 2, \cdots, p \tag{12}$$

onde $R_j^2$ é o coeficiente de determinação múltipla de regressão de $x_j$ nas restantes $p - 1$ variáveis regressoras.

- Se há uma forte multicolinearidade entre $x_j$ nas restantes $p - 1$ variáveis regressoras, então o valor de $R_j^2$ será próxima de 1.
- Como a variância de $\hat{\beta}_j$ é $Var(\hat{\beta}_j) = C_{jj}\sigma^2 = (1 - R_j^2)^{-1}\sigma$.
- Uma forte multicolinearidade implica que a variância da estimativa de mínimos quadrados do coeficiente de regressão $B_j$ é muito grande.

# Efeitos da Multicolinearidade

- Geralmente, a covariância de $\hat{\beta}_i$ e $\hat{\beta}_j$ também será se os regressores $x_i$ e $x_j$ são envolvidos em um relacionamento multicolinear.
- A multicolinearidade também tende a produzir estimativas de mínimo quadrados $\hat{\beta}_j$ que são muito grandes em valor absoluto.

# Efeitos da Multicolinearidade

Considere que o quadrado da distância de $\hat{\beta}$ ao verdadeiro valor do parâmetro $\beta$:

$$L_1^2 = (\hat{\beta} - \beta)'(\hat{\beta} - \beta) \tag{13}$$

A distância ao quadrado esperada, $E(L_1^2)$, é

$$\begin{aligned} E(L_1^2) = E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) &= \sum_{j=1}^{p} E(\hat{\beta}_j - \beta_j)^2 \\ &= \sum_{j=1}^{p} Var(\hat{\beta}_j) \\ &= \sigma^2 Tr[(X'X)^{-1}] \end{aligned} \tag{14}$$

# Efeitos da Multicolinearidade

Quando há multicolinearidade, alguns autovalores de $X'X$ serão pequenos. Como o traço de uma matriz é a soma de seus autovalores, temos:

$$E(L_1^2) = \sigma^2 \sum_{j=1}^{p} \frac{1}{\lambda_j}, \text{ onde } \lambda_j > 0,\ j = 1, 2, \cdots, p \text{ são autovalores de } X'X \tag{15}$$

Portanto, se a matriz $X'X$ é mal condicionada devido à multicolinearidade, pelo menos um dos seus autovalores $(\lambda_j)$ será pequeno.

- Isso implica que a distância do estimador de mínimos quadrados $\hat{\beta}$ do real parâmetro $\beta$ será grande.

# Efeitos da Multicolinearidade

Nos podemos equivalentemente mostrar que

$$E(L_1^2) = E(\hat{\beta} - \beta)'(\hat{\beta} - \beta) = E(\hat{\beta}'\hat{\beta} - 2\hat{\beta}'\beta + \beta'\beta) \tag{16}$$

ou

$$E(\hat{\beta}'\hat{\beta}) = \beta'\beta + \sigma^2 Tr[(X'X)^{-1}] \tag{17}$$

Isto é, o vetor $\hat{\beta}$ é geralmente maior que o vetor $\beta$. Isso implica que o método dos mínimos quadrados produz coeficientes de regressão que são muito grandes em valor absoluto.

# Efeitos da Multicolinearidade

- Embora o método dos mínimos quadrados geralmente produza estimativas ruins dos parâmetros individuais do modelo individual quando há multicolinearidade, isso não implica que necessariamente o modelo seja um preditor ruim.
- Se as previsões estiverem confinadas a região do espaço $x$ onde a multicolinearidade se mantém aproximadamente, o modelo ajustado geralmente produz previsões satisfatórias.
- Isso pode acontecer porque a combinação linear $\sum_{j=1}^{p} \beta_j x_{ij}$ pode ter estimado muito bem, mesmo o parâmetro individual $\beta_j$ seja mal estimado.
- Isto é, se os dados originais estiverem aproximadamente ao longo do hiperplano definido por (10), então futuras observações que também estão próximas desse hiperplano podem ser previstas com precisão, apesar das estimativas individuais inadequadas dos parâmetros do modelo.

1. Relembrando Regressão Linear Simples

2. Multicolinearidade

3. Métodos para detectar multicolinearidade
   - Análise da matriz de Correlação
   - Fator de inflação da variância (VIF)
   - Análise dos autovalores e autovetores

# Métodos para detectar multicolinearidade

- Várias técnicas já foram propostas para detectar multicolinearidade.
- Características desejáveis de um procedimento diagnóstico para identificação de multicolinearidade é que ele reflita diretamente o grau de intensidade do problema e que ele forneça informações úteis para determinar quais variáveis regressoras estão envolvidas no problema de multicolinearidade.

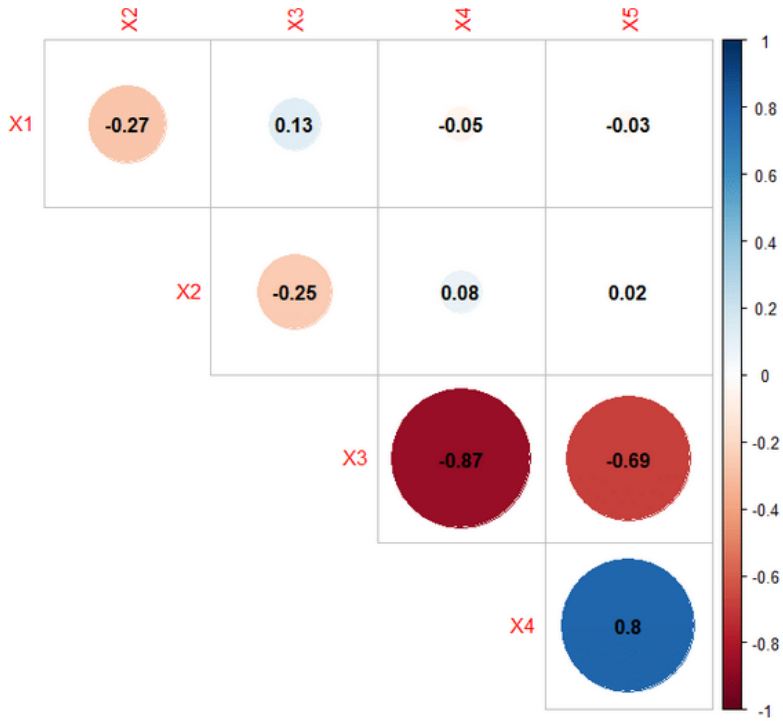
# Métodos para detectar multicolinearidade

Vamos discutir e ilustrar três dessas medidas de diagnóstico:

- Análise da matriz de Correlação
- Fator de inflação da variância (variance inflation factor - VIF)
- Análise dos autovalores de $X'X$

# Análise da matriz de Correlação

Uma forma simples de identificar uma forma de multicolinearidade é a inspeção dos coeficientes de correlação $r_{ij}$ da matriz $X'X$.




Valores altos de $|r_{ij}|$ ($>$0,7) indicam a existência de relação quase linear entre as variáveis $x_i$ e $x_j$.

# Análise da matriz de Correlação

De forma geral, a inspeção dos elementos $r_{ij}$ só é suficiente para detectar colinearidade entre pares de variáveis, no entanto, colinearidade pode ocorrer entre 3 variáveis ou mais (nesse caso chamamos de multicolinearidade) mesmo quando não há alta correlação entre nenhum par de variáveis, logo, a matriz de correlação não pode ser usada para detectar todos os casos de multicolinearidade.

# Fator de inflação da variância (VIF)

O VIF, para cada termo do modelo, mede o efeito combinado das dependências entre as variáveis regressoras sobre a variância daquele termo. Os VIF's são dados pelos elementos da diagonal principal da matriz $C = (X'X)^{-1}$. Já mostramos que o VIF associado à variável $x_j$ do modelo pode ser calculado por:

$$VIF_j = C_{jj} = \frac{1}{(1 - R_j^2)}$$

Onde $R_j$ é o coeficiente de determinação múltipla obtido através da regressão de $x_j$ nas demais variáveis regressoras.

# Fator de inflação da variância (VIF)

- Um VIF de 1 para uma determinada variável regressora indica a total ausência de multicolinearidade entre essa variável e outras variáveis regressoras no modelo.
- Claramente, se $x_j$ for quase linearmente dependente de alguns dos outros regressores (ou seja, $x_j$ pode ser quase perfeitamente previsto por outras variáveis do modelo), então $R_j$ será próximo de 1 e consequentemente $VIF_j$ será "grande".
- Experimentos práticos mostram que valores de VIF maiores do que 10, para qualquer uma das variáveis regressoras, normalmente são sinais da existência de multicolinearidade (esse valor é relativo e alguns autores consideram limites mais conservadores como valores de VIF maiores do que 5 ou até mesmo 2,5).

# Fator de inflação da variância (VIF)

**Exemplo: Multicolinearidade sem alta correlação entre nenhum par de variáveis**

```
library(corrplot)
library(regclass)

# Primeiro definimos os regressores de forma que x5 esteja
    "ligeiramente" relacionado a todos os outros
set.seed(1)
x1 = rnorm(100)
x2 = rnorm(100)
x3 = rnorm(100)
x4 = rnorm(100)
x5 = 0.1*x1 + 0.1*x2 + 0.1*x3 + 0.1*x4 + rnorm(100)*0.03

# y sera nossa variavel dependente
y = rnorm(100)
```

# Fator de inflação da variância (VIF)

## Código Parte 2

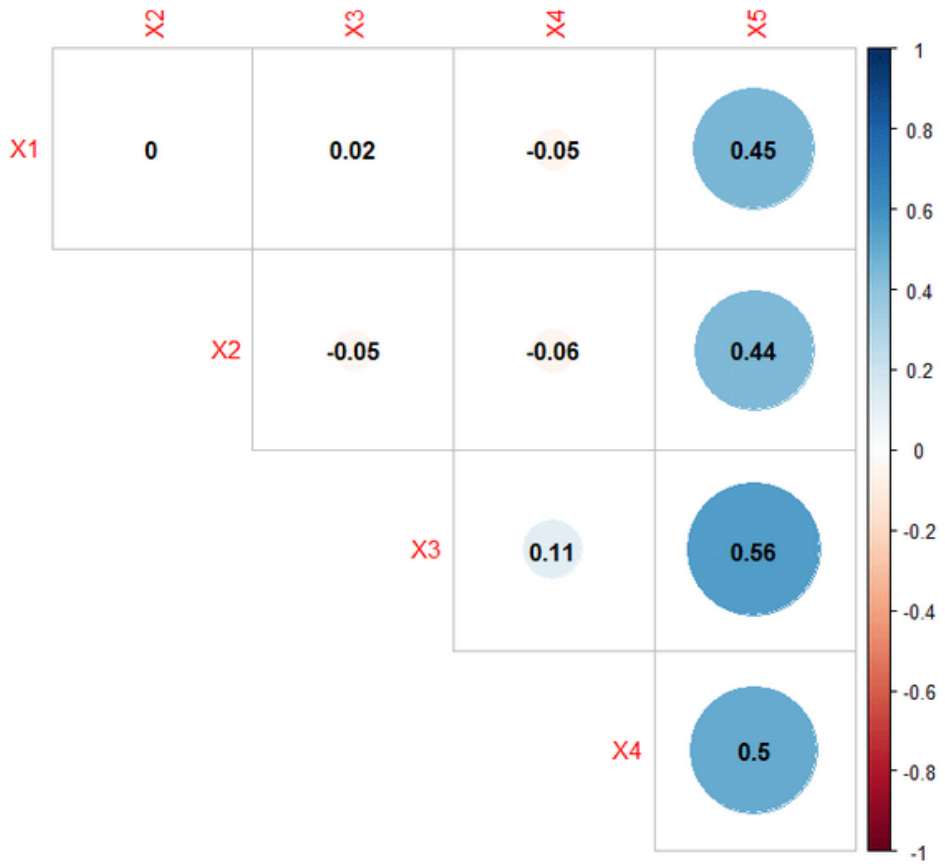
```
# junta todas as variaveis em um data frame
df = data.frame(X1=x1, X2=x2, X3=x3, X4=x4, X5=x5, Y=y)

# plota a matriz de correlacao
corrplot(cor(df[,c("X1", "X2", "X3", "X4", "X5")]), \
    diag = FALSE, type="upper",addCoef.col = "black")

# mostra o VIF associado a cada variavel regressora
VIF(lm(Y ~ X1 + X2 + X3 + X4 + X5, data=df))
```

# Fator de inflação da variância (VIF)

**Exemplo: Multicolinearidade sem alta correlação entre nenhum par de variáveis** A matriz de correlação não indica sinais de colinearidade entre nenhum par de variáveis já que todos os coeficientes de correlação estão abaixo de 0,7.



# Fator de inflação da variância (VIF)

**Exemplo: Multicolinearidade sem alta correlação entre nenhum par de variáveis**

No entanto, olhando para o VIF de cada variável:

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ |
|---|---|---|---|---|
| 8.662448 | 9.599640 | 10.272285 | 9.574024 | 34.629869 |

Podemos ver que 2 delas apresentam VIF$>$10, sinalizando a existência de multicolinearidade no modelo.

1. Relembrando Regressão Linear Simples

2. Multicolinearidade

3. Métodos para detectar multicolinearidade
   - Análise da matriz de Correlação
   - Fator de inflação da variância (VIF)
   - Análise dos autovalores e autovetores

# Autovalores e Autovetores

Podemos utilizar raízes característica ou autovalores para verificar a existencia de multicolinearidade na matriz.
Caso haja um ou mais autovalores próximos de zero, eles indicam a quase dependencia linear.

# Autovalores e Autovetores

Exemplos:

$$A = \begin{bmatrix} 1 & 2 & 8 & 3 \\ 2 & 4 & 5 & 6 \\ 8 & 2 & 7 & 10 \\ 3 & 3 & 7 & 6 \end{bmatrix}$$

$$\lambda_1 \approx 20.448, \quad \lambda_2 \approx -5.153, \quad \lambda_3 \approx 2.704, \quad \lambda_4 = 0$$

$$A = \begin{bmatrix} 1 & 2 & 8 & 2.98 \\ 2 & 4 & 5 & 5.96 \\ 8 & 2 & 7 & 10.02 \\ 3 & 3 & 7 & 6.03 \end{bmatrix}$$

$$\lambda_1 \approx 20.455, \quad \lambda_2 \approx -5.153, \quad \lambda_3 \approx 2.700, \quad \lambda_4 \approx 0.027$$

# Autovalores e Autovetores

Seja os autovalores de uma matriz $X'X$, dados por $\lambda_1$, $\lambda_2$, $\cdots$, $\lambda_p$, podemos verificar se existe multicolinearidade nos dados pela presença de um ou mais autovalores próximos de zero, que podem implicar a quase dependência linear das colunas X.

A análise direta dos autovalores pode causar problemas como o quão próximo de zero $\lambda$ precisa ser, e se todos os autovalores forem pequenos.

# Autovalores e Autovetores

Para quantificar se $\lambda$ é próximo o suficiente de zero usamos o **número de condicionamento** (*condition number*) de $X'X$ dado por:

$$k = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} \tag{18}$$

que, por sua vez é uma medida de proporção dos autovalores.

- Geralmente se o $k$ for menor que 100, não há um problema sério de multicolinearidade;
- Se $k$ for entre 100 e 1000, indica moderada ou forte multicolinearidade;
- Se $k$ for maior que 1000, implica severa multicolinearidade.

Definimos também os **índices de condicionamento** (*condition indices*) de $X'X$ como:

$$k_j = \frac{\lambda_{\max}}{\lambda_j}, \quad j = 1, 2, \cdots, p \tag{19}$$

É notável que o maior índice de condicionamento será o igual ao número de condicionamento.

# Autovalores e Autovetores

Considerando o exemplo tirado do livro Webster,Gunst, and Mason [1974]:

| Observation, $i$ | $y_i$ | $x_{i1}$ | $x_{i2}$ | $x_{i3}$ | $x_{i4}$ | $x_{i5}$ | $x_{i6}$ |
|---|---|---|---|---|---|---|---|
| 1 | 10.006 | 8.000 | 1.000 | 1.000 | 1.000 | 0.541 | –0.099 |
| 2 | 9.737 | 8.000 | 1.000 | 1.000 | 0.000 | 0.130 | 0.070 |
| 3 | 15.087 | 8.000 | 1.000 | 1.000 | 0.000 | 2.116 | 0.115 |
| 4 | 8.422 | 0.000 | 0.000 | 9.000 | 1.000 | –2.397 | 0.252 |
| 5 | 8.625 | 0.000 | 0.000 | 9.000 | 1.000 | –0.046 | 0.017 |
| 6 | 16.289 | 0.000 | 0.000 | 9.000 | 1.000 | 0.365 | 1.504 |
| 7 | 5.958 | 2.000 | 7.000 | 0.000 | 1.000 | 1.996 | –0.865 |
| 8 | 9.313 | 2.000 | 7.000 | 0.000 | 1.000 | 0.228 | –0.055 |
| 9 | 12.960 | 2.000 | 7.000 | 0.000 | 1.000 | 1.380 | 0.502 |
| 10 | 5.541 | 0.000 | 0.000 | 0.000 | 10.000 | –0.798 | –0.399 |
| 11 | 8.756 | 0.000 | 0.000 | 0.000 | 10.000 | 0.257 | 0.101 |
| 12 | 10.937 | 0.000 | 0.000 | 0.000 | 10.000 | 0.440 | 0.432 |

Figura: Variáveis de regressão e resposta não padronizadas de Webster, Gunst e Mason [1974]

Teremos a matrix:

$$X'X = \begin{bmatrix} 1 & 0.052 & -0.343 & -0.498 & 0.417 & -0.192 \\ 0.052 & 1 & -0.432 & -0.371 & 0.485 & -0.317 \\ -0.343 & -0.432 & 1 & -0.355 & -0.505 & 0.494 \\ -0.498 & -0.371 & -0.355 & 1 & -0.215 & -0.087 \\ 0.417 & 0.485 & -0.505 & -0.215 & 1 & -0.123 \\ -0.192 & -0.317 & 0.493 & -0.087 & -0.123 & 1 \end{bmatrix}$$

# Autovalores e Autovetores

Os autovalores da matriz $X'X$, serão:

$$\lambda_1 = 2.42879, \qquad \lambda_2 = 1.54615, \qquad \lambda_3 = 0.92208$$
$$\lambda_4 = 0.79398, \qquad \lambda_5 = 0.30789, \qquad \lambda_6 = 0.00111$$

E seus índices de condicionamento serão:

$$k_1 = \frac{2.42879}{2.42879} = 1 \qquad k_2 = \frac{2.42879}{1.54615} = 1.57086$$
$$k_3 = \frac{2.42879}{0.92208} = 2.63403 \qquad k_4 = \frac{2.42879}{0.79398} = 3.05900$$
$$k_5 = \frac{2.42879}{0.30789} = 7.88849 \qquad k_6 = \frac{2.42879}{0.00111} = 2188.099$$

Então, temos que:

$$k = \frac{\lambda_{\text{max}}}{\lambda_{\text{min}}} = \frac{2.42879}{0.00111} = 2188.099$$

Como o número de condicionamento ($k$) é maior que 1000, isso indica uma forte multicolinearidade, e como os outros valores são pequenos (menores que 100), podemos afirmar que existe apenas uma quase-dependência linear nos dados.

# Análise de Autovalores e Autovetores

A análise dos autovalores e autovetores pode ser usada para identificar a natureza das quase-dependências nos dados. A matriz $X'X$ pode ser diagonalizada como:

$$X'X = T\Lambda T' \tag{20}$$

em que:

- $\Lambda$: matriz diagonal $p \times p$ cujo os elementos diagonais são os autovalores $\lambda_j (j = 1, 2, \cdots, p)$ de $X'X$;
- $T$: matriz ortonormal $p \times p$ cujas colunas $\mathbf{t_1}, \mathbf{t_2}, \cdots, \mathbf{t_p}$ são os autovetores de $X'X$.

# Análise de Autovalores e Autovetores

Por exemplo, usando os dados de Webster,Gunst, and Mason [1974], encontramos a seguinte $T$.

Tabela: Autovetores de Webster, Gunst and Mason

| $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ |
|---|---|---|---|---|---|
| -.39072 | -.33968 | .67980 | .07990 | -.25104 | -.44768 |
| -.45560 | -.05392 | -.70013 | .05769 | -.34447 | -.42114 |
| .48264 | -.45333 | -.16078 | .19103 | .45364 | -.54169 |
| .18766 | .73547 | .13587 | -.27645 | .01521 | -.57337 |
| -.49773 | -.09714 | -.03185 | -.56356 | .65128 | -.00605 |
| .35195 | -.35476 | -.04864 | -.74818 | -.43375 | -.00217 |

Se o autovalor $\lambda_j$ for próximo de zero, então os elementos do vetor $\mathbf{t_j}$ são os coeficientes $t_1, t_2, \cdots, t_p$ na seguinte equação:

$$\sum_{j=1}^{p} t_j X_j = 0 \tag{21}$$

que descreve a natureza da dependência linear.

# Análise de Autovalores e Autovetores

No nosso exemplo temos que $\lambda_{\text{min}} : \lambda_6 = 0.0011$. Portanto os elementos do autovetor $\mathbf{t_6}$ serão os coeficientes dos regressores.

| $\mathbf{t_6}$ |
|---|
| -.44768 |
| -.42114 |
| -.54169 |
| -.57337 |
| -.00605 |
| -.00217 |

# Análise de Autovalores e Autovetores

Retomando a equação: $\sum_{j=1}^{p} t_j X_j = 0$, teremos:

$$-0.44768x_1 - 0.42114x_2 - 0.54169x_3 - 0.57337x_4 - 0.00605x_5 - 0.00217x_6 = 0$$

Assumindo 0.00605 e 0.0027 como coeficientes nulos, podemos rearranjar os termos da seguinte maneira:

$$x_1 \simeq -0.941x_2 - 1.120x_3 - 1.281x_4$$

Ou seja, os quatro primeiras variáveis de regressão somadas resultam aproximadamente em uma constante. Assim, os elementos de $\mathbf{t_6}$ refletem diretamente a relação usada para gerar $x_1$, $x_2$, $x_3$ e $x_4$.

# Proposta de Besley, Kuh, and Welsch

Outra maneira semelhante de diagnosticar multicolinearidade foi proposta por Besley, Kuh, and Welsch [1980]. Nesta abordagem, a matriz $X$, $n \times p$, deve ser fatorada na forma SVD (*Singular Value Decomposition*):

$$X = UDT' \tag{22}$$

- $U$: matriz ortonormal $n \times p$ cujas colunas são os autovetores associados aos $p$ autovalores não-nulos de $XX'$
- $D$: matriz diagonal $p \times p$ cujos elementos diagonais $\mu_j$, $j = 1, 2, \cdots, p$ são chamados valores singulares de $X$
- $T$: matriz $p \times p$ de autovetores de $X'X$

A vantagem da fatoração SVD é permitir o uso de algoritmos mais eficientes e numericamente estáveis; além de ser feita diretamente sobre a matriz de dados e não sobre a matriz de correlação.

Note que partir de

$$X = UDT' \tag{23}$$

podemos obter

$$X'X = (UDT')'UDT' = TD'(U'U)DT' = TD^2T' = T\Lambda T' \tag{24}$$

e portanto

$$\mu_j^2 = \lambda_j \tag{25}$$

De maneira análoga ao uso de autovalores, para cada valor pequeno de $\mu$ teremos uma quase dependência linear. Definiremos o **índice de condicionamento** da matriz X como:

$$\eta_j = \frac{\mu_{\max}}{\mu_j}, \quad j = 1, 2, \cdots, p \tag{26}$$

O maior valor de $\eta_j$ será o número de condicionamento de X.

# *Singular Value Decomposition*

A matriz de covariância de $\hat{\beta}$ é:

$$Var(\hat{\beta}) = \sigma^2 (X'X)^{-1}$$

Lembrando que:

$$X'X = T \Lambda T'$$

Portanto,

$$Var(\hat{\beta}) = \sigma^2 T \Lambda T'$$

E a variância do j-ésimo coeficiente regressor é o j-ésimo elemento diagonal dessa matriz.

$$Var(\hat{\beta}) = \sigma^2 \sum_{i=1}^{p} \left( \frac{t_{ji}^2}{\mu_i^2} \right) = \sigma^2 \sum_{i=1}^{p} \left( \frac{t_{ji}^2}{\lambda_i} \right)$$

# *Singular Value Decomposition*

- Autovalores pequenos podem inflar a variância de $\hat{\beta}_j$
- Belsley, Kuh e Welsch sugerem o uso de proporções de decomposição de variâncias, definidas por:

$$\pi_{ij} = \frac{t_{ji}^2/\mu_i^2}{VIF_j}, j = 1, 2, \cdots, p$$

- Medidas de multicolinearidade
- $\pi$ como uma matriz $p \times p \rightarrow \pi_{ij} =$ proporção da variância de $\hat{\beta}_j$ (ou VIF), contribuído pelo i-ésimo autovalor
- Multicolinearidade indicada quando uma alta proporção da variância para dois ou mais coeficientes de regressão é associado com um pequeno valor singular.
- Recomendado: índices de condicionamento $> 30$; proporções de decomposição de variância $> 0,5$

# Grafico

Figura: Variance Decomposition Proportions for the Webster, Gunst, and Mason [1974] Data

| Number | Eigenvalue | Condition Indices | Variance Decomposition Proportions | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | $X_1$ | $X_2$ | $X_3$ | $X_4$ | $X_5$ | $X_6$ |
| *A. Regressors Centered* | | | | | | | | |
| 1 | 2.42879 | 1.00000 | 0.0003 | 0.0005 | 0.0004 | 0.0000 | 0.0531 | 0.0350 |
| 2 | 1.54615 | 1.25334 | 0.0004 | 0.0000 | 0.0005 | 0.0012 | 0.0032 | 0.0559 |
| 3 | 0.92208 | 1.62297 | 0.0028 | 0.0033 | 0.0001 | 0.0001 | 0.0006 | 0.0018 |
| 4 | 0.79398 | 1.74900 | 0.0000 | 0.0000 | 0.0002 | 0.0003 | 0.2083 | *004845* |
| 5 | 0.30789 | 2.80864 | 0.0011 | 0.0024 | 0.0025 | 0.0000 | 0.7175 | 004199 |
| 6 | 0.00111 | 46.86052 | 0.9953 | 0.9937 | 0.9964 | 0.9984 | 0.0172 | 0.0029 |

# *Singular Value Decomposition*

A tabela mostra os índices de condicionamento de X e as proporções de decomposição de variância (usando os dados de Webster, Gunst e Mason);

- Painel A: regressores centralizados para que essas variáveis sejam $(x_{ij} - \bar{x}_j)$, $j = 1, \cdots, 6$
- VIFs em um modelo polinomial são afetados por centralizar os termos lineares no modelo, antes de gerar os termos polinomiais de ordem maior.
- Centralizar afeta as proporções de decomposição de variância (e os autovalores e autovetores)
- Remove mal condicionamento resultante da interceptação

# *Singular Value Decomposition*

$\eta_6 = 46,86 > 30 \rightarrow$ uma dependência nas colunas de $X$
$\pi_{61}, \pi_{62}, \pi_{63}, \pi_{64} > 0,5 \rightarrow$ os primeiros quatro regressores estão envolvidos numa relação multicolinear.
Belsley, Kuh e Welsch sugerem:

- Escalonar os regressores para uma unidade de medida
- Não centralizados quando computando as proporções de decomposição de variância (p/ o papel da interceptação ser diagnosticado em casos dependências quase-lineares)

# Detecção de Multicolinearidade em Modelos de Regressão por Autovetores e Autovalores

Felipe Ferreira  Fernanda Rahal  Fernando Fukui  Marcelo Baraldo
Matheus Siniscarchio  Vitor Jensen

Instituto de Matemática e Estatística (IME)
Universidade de São Paulo (USP)

11 de julho de 2022

Obrigado!