

MBA
USP
ESALQ

*Supervised Machine
Learning:
Análise de Regressão
Simples e Múltipla*

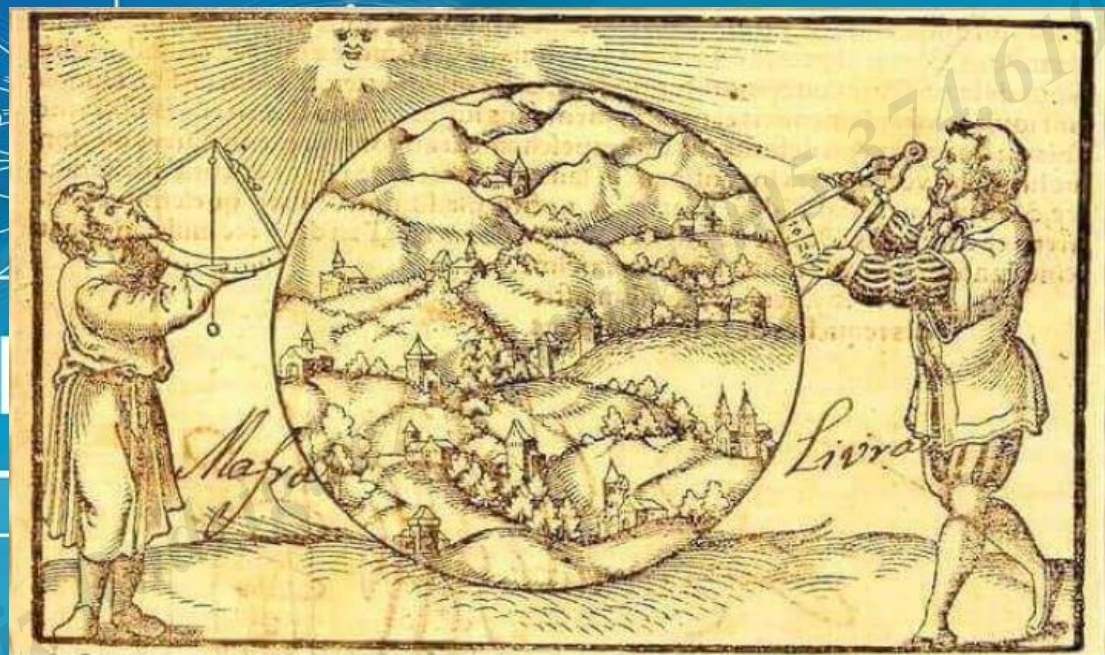
Prof. Dr. Luiz Paulo Fávero



$$z = \frac{Z_L}{Z_0}$$

$$\Gamma = \frac{V_{\text{reflected}}}{V_{\text{incident}}}$$

$$\frac{a}{b+c} = a \div (b+c) \neq \frac{a}{b} + \frac{a}{c}$$



$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

Reflexão

Modelos Supervisionados



“Diferentes pesquisadores, a partir de uma mesma base de dados, podem estimar diferentes modelos e, conseqüentemente, obter diferentes valores previstos do fenômeno em estudo. O objetivo é estimar modelos que, embora simplificações da realidade, apresentem a melhor aderência possível entre os valores reais e os valores previstos”.

Silberzahn, R.; Uhlmann, E. L. Many hands make tight work. **Nature**, v. 526, p. 189-191, Out 2015.

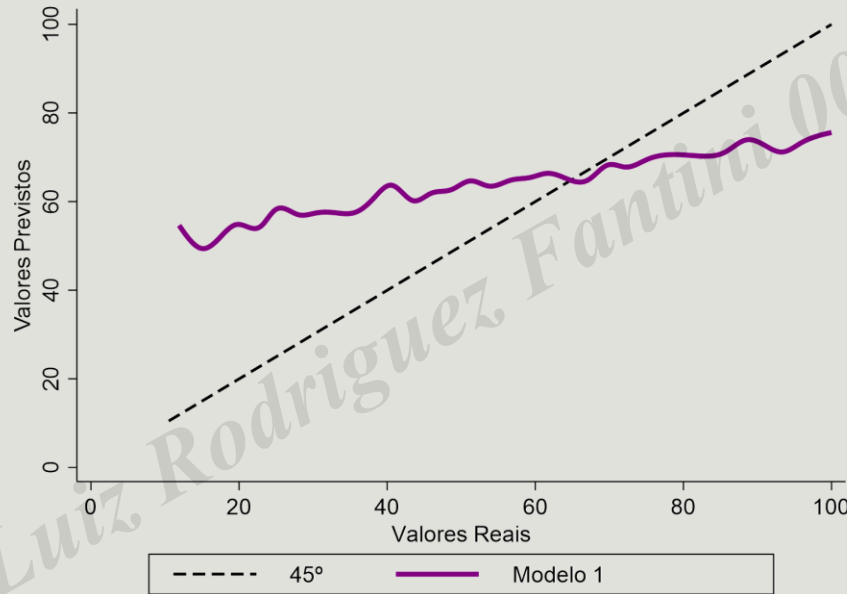
Reflexão

Modelos Supervisionados



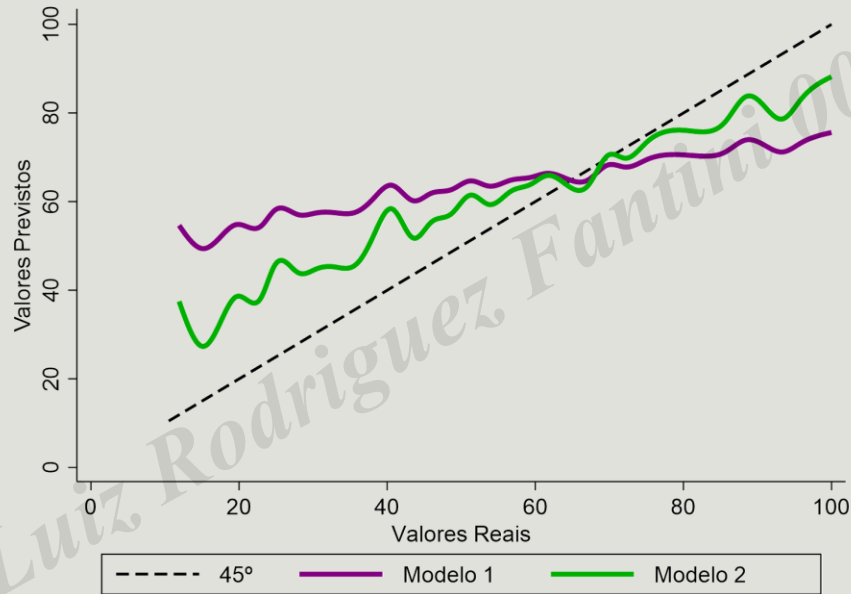
Reflexão

Modelos Supervisionados



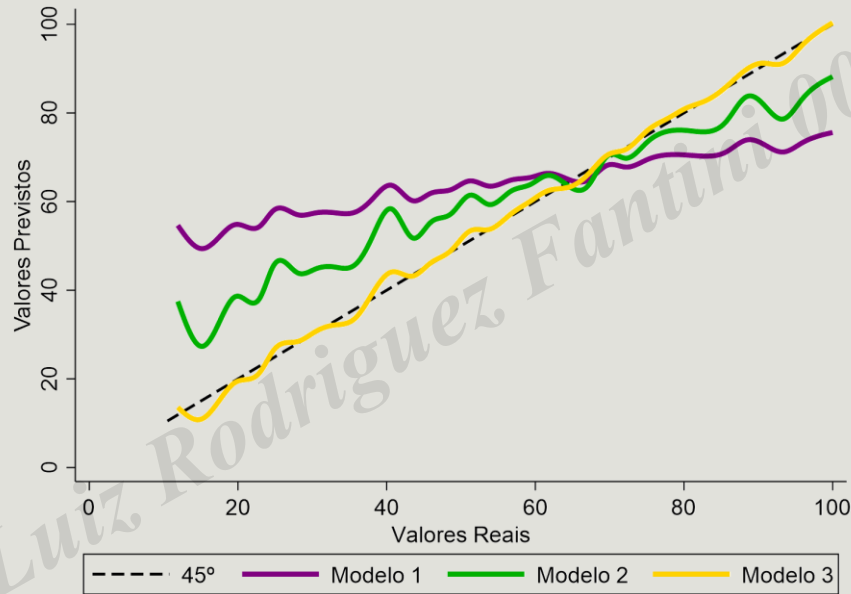
Reflexão

Modelos Supervisionados

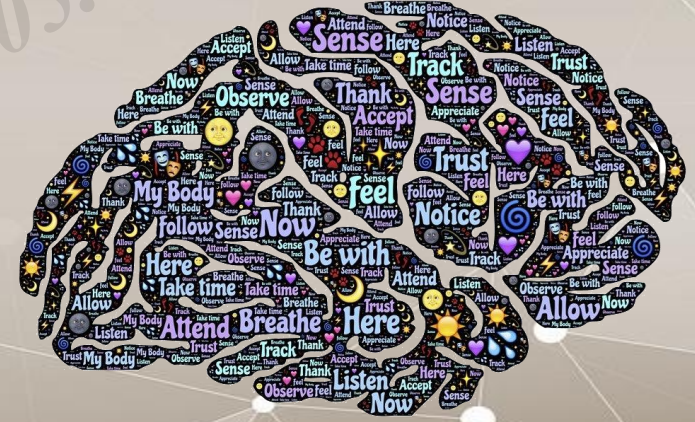
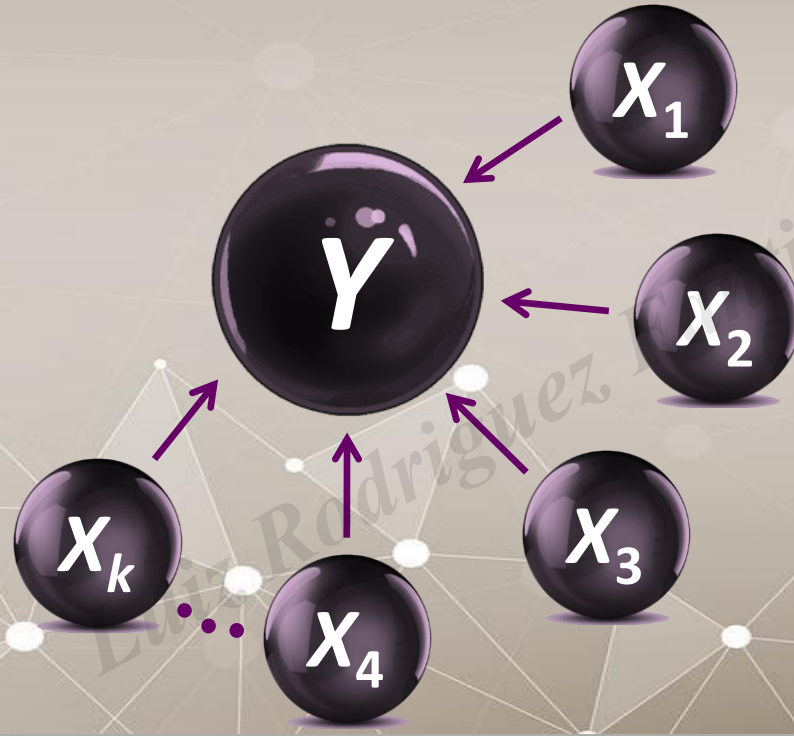


Reflexão

Modelos Supervisionados



Modelos Supervisionados de Machine Learning: Modelos Lineares Generalizados (GLM)



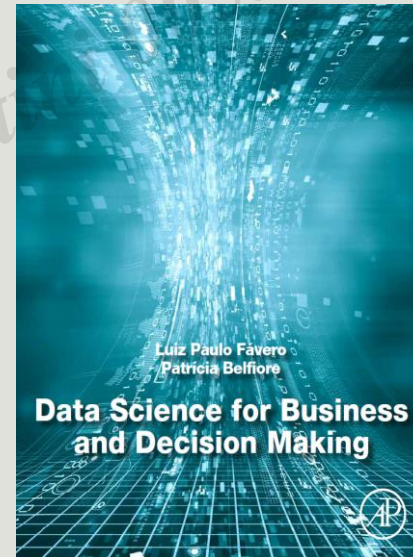
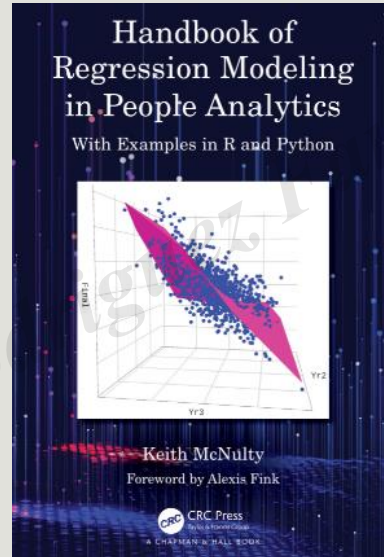
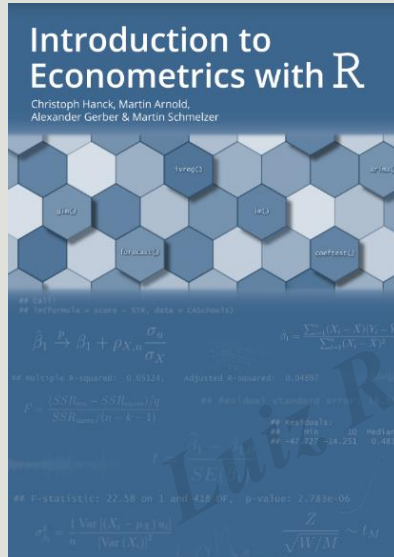
Modelos Lineares Generalizados (GLM)

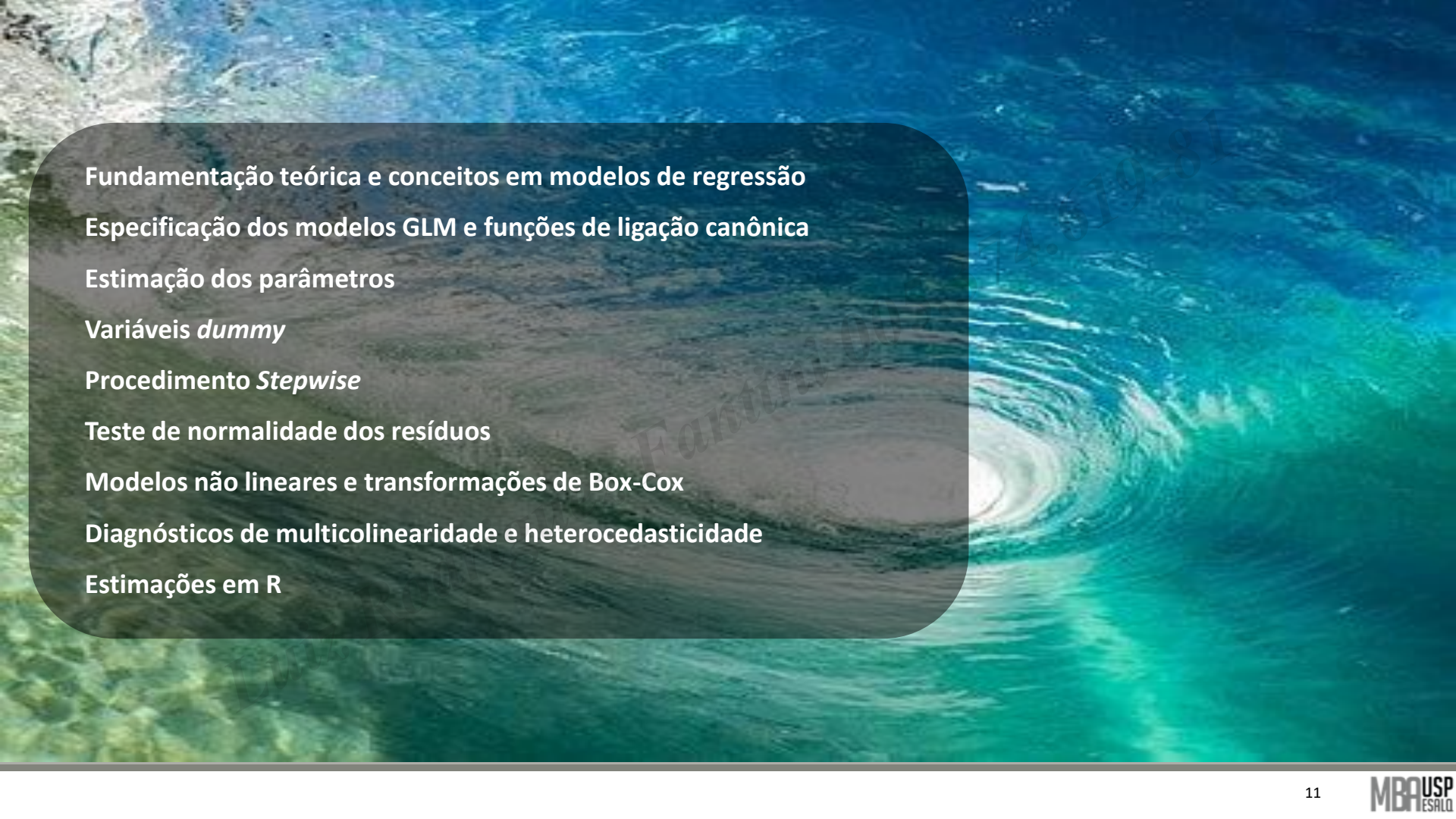
$$\eta_i = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki}$$

Modelos lineares generalizados, características da variável dependente e funções de ligação canônica.

Modelo de Regressão	Característica da Variável Dependente	Distribuição	Função de Ligação Canônica (η)
Linear	Quantitativa	Normal	\hat{Y}
Com Transformação de Box-Cox	Quantitativa	Normal Após a Transformação	$\frac{\hat{Y}^\lambda - 1}{\lambda}$
Logística Binária	Qualitativa com 2 Categorias (<i>Dummy</i>)	Bernoulli	$\ln\left(\frac{p}{1-p}\right)$
Logística Multinomial	Qualitativa M ($M > 2$) Categorias	Binomial	$\ln\left(\frac{p_m}{1-p_m}\right)$
Poisson	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson	$\ln(\lambda_{poisson})$
Binomial Negativo	Quantitativa com Valores Inteiros e Não Negativos (Dados de Contagem)	Poisson-Gama	$\ln(\lambda_{neg})$

Modelos Supervisionados: Modelos Lineares Generalizados (GLM)





Fundamentação teórica e conceitos em modelos de regressão

Especificação dos modelos GLM e funções de ligação canônica

Estimação dos parâmetros

Variáveis *dummy*

Procedimento *Stepwise*

Teste de normalidade dos resíduos

Modelos não lineares e transformações de Box-Cox

Diagnósticos de multicolinearidade e heterocedasticidade

Estimações em R

Regressão Linear Simples

Objetivo:

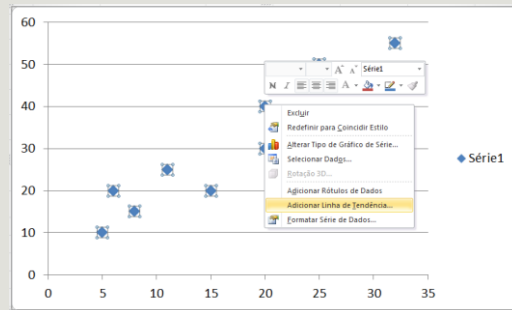
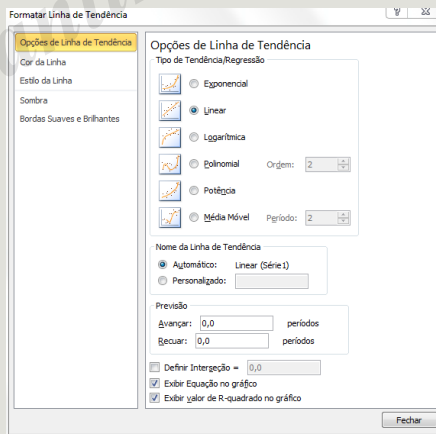
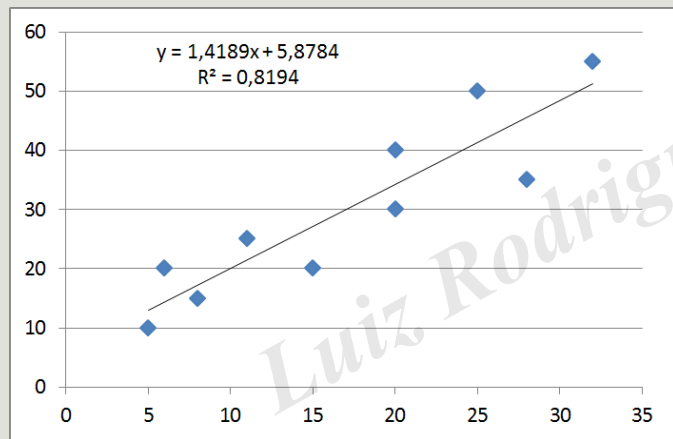
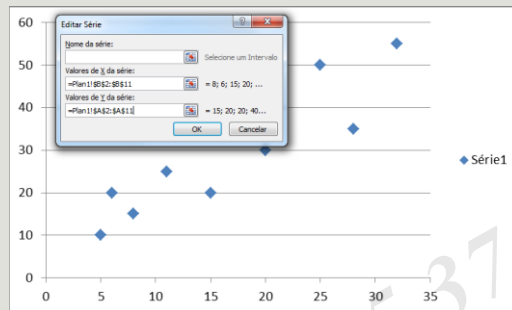
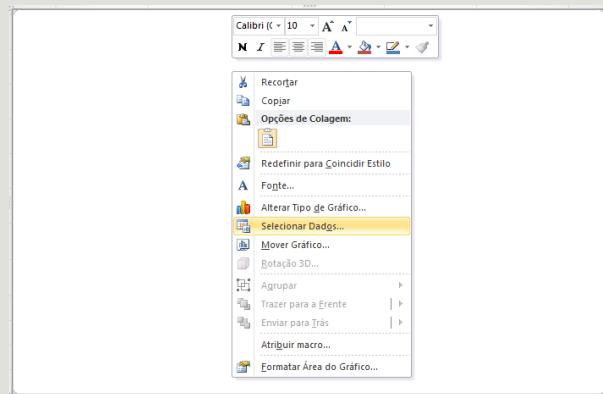
Desenvolver uma equação linear que apresente a relação entre uma variável dependente e uma variável explicativa.

Equação linear de uma reta num plano cartesiano:

$$Y_i = \alpha + \beta \cdot X_i + u_i$$



em que temos um intercepto (α), um coeficiente de inclinação da reta (β), uma variável explicativa X e um termo de erro u .

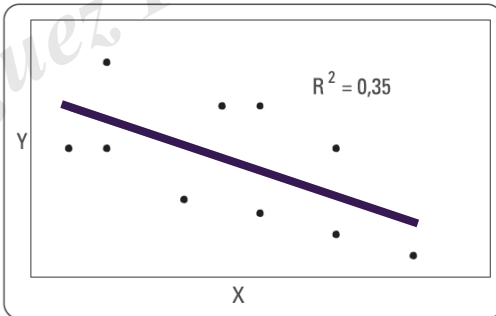
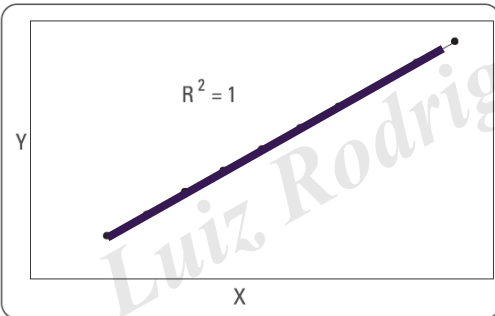
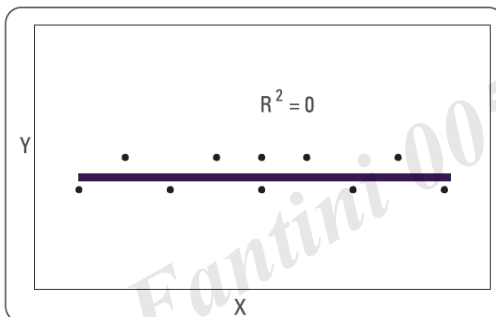
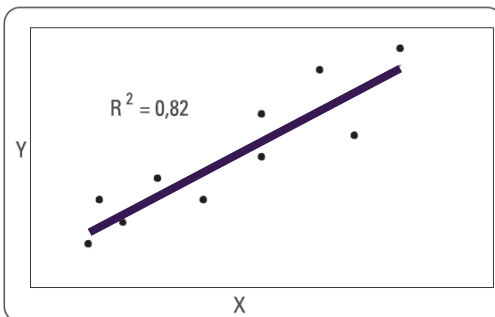


Análise de Regressão: Coeficiente de Ajuste do Modelo (R^2)

Indica o percentual de variância da variável Y que é devido ao comportamento de variação conjunta da(s) variável(is) explicativa(s) X . Varia de 0 a 1 e, quanto maior o coeficiente, maior o poder preditivo do modelo de regressão, ou seja, maior o poder de explicação do comportamento da variável dependente frente ao comportamento da(s) variável(is) explicativa(s).



Análise de Regressão: Coeficiente de Ajuste do Modelo (R^2)



Comportamento do R^2 para diferentes regressões lineares simples.

Análise de Regressão: Estimação dos Parâmetros

Critérios:

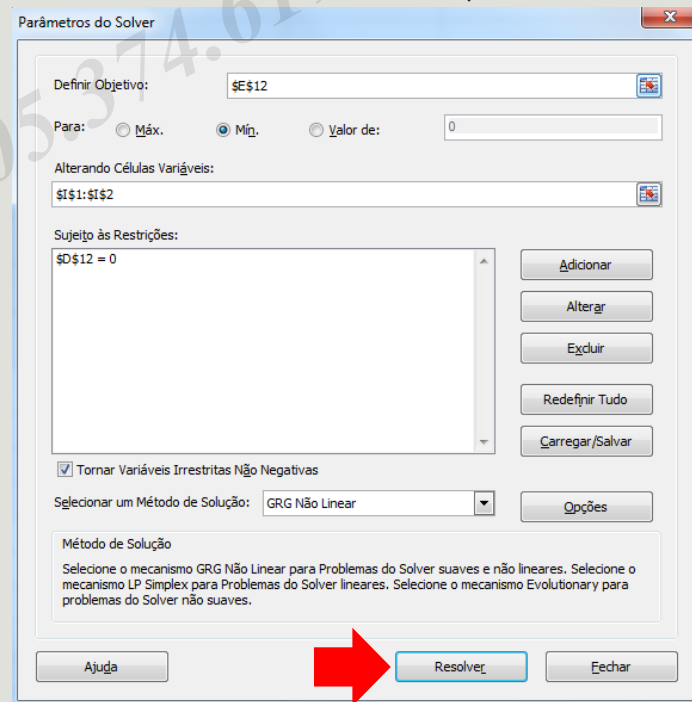
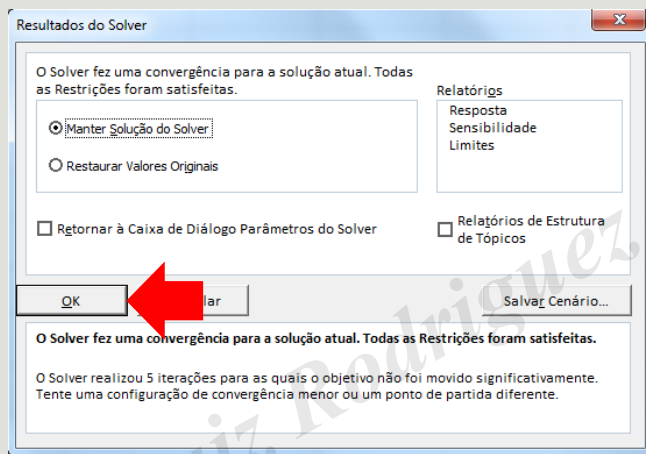
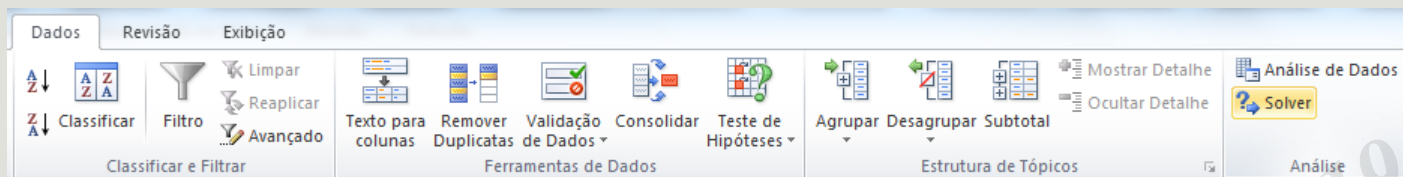
1 – Soma dos erros igual a zero:

$$\sum_{i=1}^n u_i = 0$$

2 – Soma dos erros ao quadrado sendo a mínima possível:

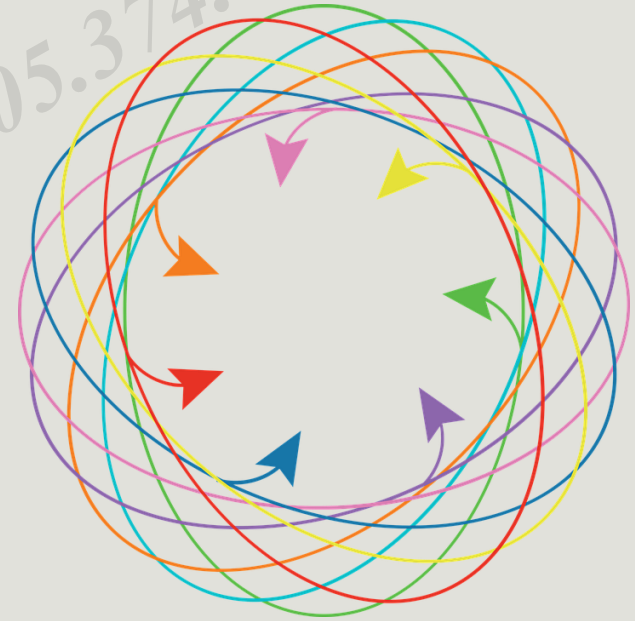
$$\sum_{i=1}^n u_i^2 = \text{mín}$$

Parâmetros α e β podem ser estimados por meio do método dos mínimos quadrados ordinários (MQO), em que a somatória dos quadrados dos termos de erro é minimizada.

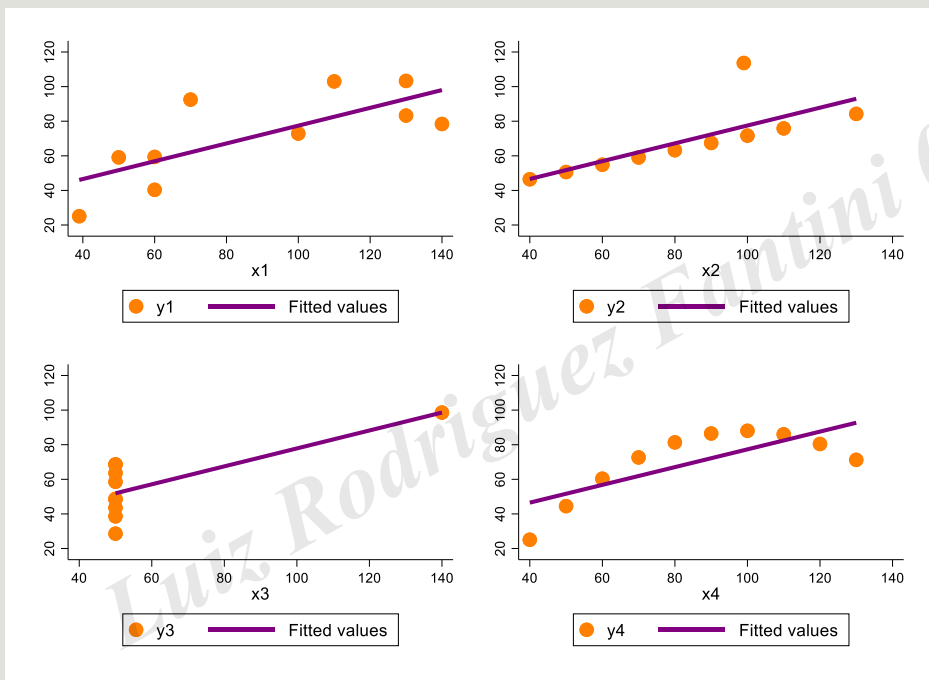


Análise de Regressão: Cálculo do R^2

$$R^2 = \frac{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2}{\sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (u_i)^2}$$



Apenas Parâmetros e R^2 ?





07. Fantini 005.374.619-81

Modelos de Regressão no

Significância Estatística do Modelo

- **Teste F :** Permite analisar se pelo menos um dos β s é estatisticamente significativo para a explicação do comportamento de Y .
- **Hipóteses:** $H_0: \beta_1 = \beta_2 = \beta_3 = \dots = \beta_k = 0$ H_1 : pelo menos um $\beta \neq 0$

Na rejeição da hipótese nula, pelo menos um dos β 's será estatisticamente diferente de zero para explicar o comportamento de Y -> p-valor abaixo do nível crítico (0,05, usualmente).

Significância Estatística dos Parâmetros do Modelo

- **Teste t:** Permite analisar se cada um dos parâmetros, individualmente, é estatisticamente diferente de zero (no caso de regressão simples, apresenta a mesma significância da estatística F).

- **Hipóteses:** $H_0: \beta = 0$ $H_1: \beta \neq 0$



Avalia-se a significância estatística de cada parâmetro do modelo, para determinado nível de significância (0,05, usualmente).

Comparação entre Modelos

Quando houver o intuito de se compararem os resultados das estimações de dois modelos com quantidades distintas de parâmetros e/ou obtidos a partir de amostras com tamanhos diferentes, faz-se necessário o uso do R^2 ajustado.

$$R^2_{ajust.} = 1 - \frac{n-1}{n-k} \cdot (1 - R^2)$$



Regressão Múltipla

Qual a diferença entre um modelo de regressão simples para um modelo de regressão múltipla?

A inclusão de novas variáveis explicativas no modelo!



A forma funcional passa a ser a seguinte:

$$Y_i = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \beta_3 \cdot X_{3i} \dots + \beta_k \cdot X_{ki} + u_i$$

Variáveis Explicativas (X) Qualitativas

- Modelando com variáveis explicativas (X) qualitativas.
- É muito comum observar que diversas variáveis explicativas podem se apresentar de maneira qualitativa (exemplo: rating de crédito, setor de atuação, etc.).
- Dado que tais características não possuem média e nem variância, como incorporá-las ao modelo de regressão?

A large, dense, and slightly curved wall of binary code (0s and 1s) in a 3D perspective, creating a sense of depth and digital space. The code is arranged in horizontal lines that recede into the distance, giving the impression of a vast, digital landscape. The lighting is soft, highlighting the texture of the binary digits.

Variáveis Explicativas (X) Qualitativas

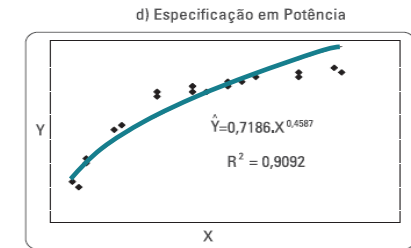
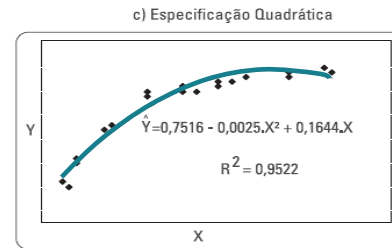
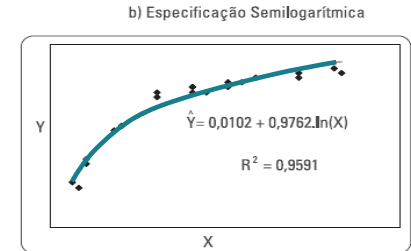
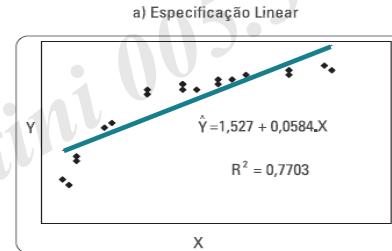
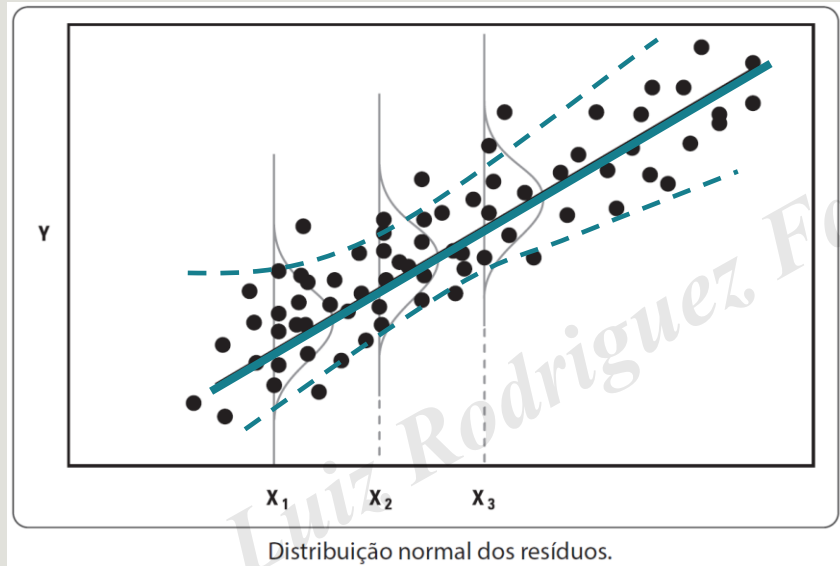
Variáveis *dummy*

São variáveis categóricas que representam um atributo por meio de combinação binária (0 para a ausência ou 1 para presença).

E quando tivermos uma variável categórica com mais de uma categoria?

Neste caso, devemos incluir $n - 1$ *dummies*, em que n é a quantidade de categorias existentes na variável original.

Modelos Não Lineares



Resultados da aplicação de quatro diferentes formas funcionais em regressão.

Modelos Não Lineares e Transformações de Box-Cox

An Analysis of Transformations

G. E. P. Box and D. R. Cox

Journal of the Royal
Statistical Society.

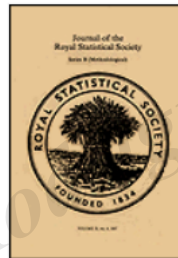
Series B

(Methodological)

Vol. 26, No. 2 (1964),

pp. 211-252 (42 pages)

Published By: Wiley



<https://www.jstor.org/stable/2984418>

$$Y_{Box-Cox}^* = \frac{Y^\lambda - 1}{\lambda}$$

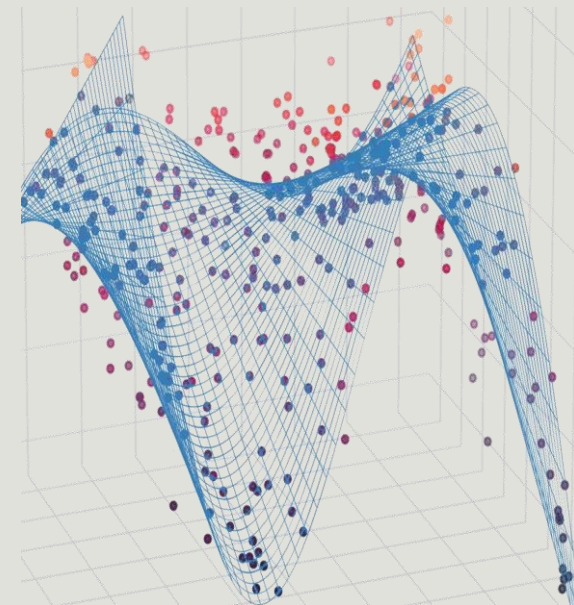
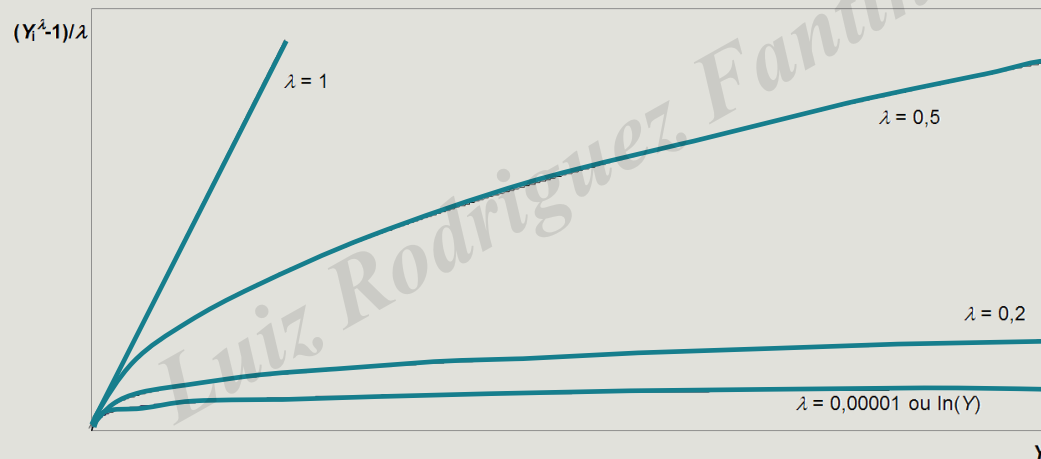
Qual o valor de λ (λ varia entre $-\infty$ e $+\infty$) que maximiza a aderência da distribuição da nova variável Y^* à normalidade?

Modelos Não Lineares e Transformações de Box-Cox

$Y_i = \alpha + \beta_1.X_1 + \beta_2.X_2 + \dots + \beta_k.X_k$	Especificação Linear ($\lambda = 1$)
$Y_i^2 = \alpha + \beta_1.X_1 + \beta_2.X_2 + \dots + \beta_k.X_k$	Especificação Quadrática ($\lambda = 2$)
$Y_i^3 = \alpha + \beta_1.X_1 + \beta_2.X_2 + \dots + \beta_k.X_k$	Especificação Cúbica ($\lambda = 3$)
$\sqrt{Y_i} = \alpha + \beta_1.X_1 + \beta_2.X_2 + \dots + \beta_k.X_k$	Especificação de Raiz ($\lambda = 0,5$)
$\frac{1}{Y_i} = \alpha + \beta_1.X_1 + \beta_2.X_2 + \dots + \beta_k.X_k$	Especificação Inversa ($\lambda = -1$)
$\ln(Y_i) = \alpha + \beta_1.X_1 + \beta_2.X_2 + \dots + \beta_k.X_k$	Especificação Semilogarítmica ($\lambda = 0$) Expansão de Taylor

Modelos Não Lineares e Transformações de Box-Cox

$$\frac{Y_i^\lambda - 1}{\lambda} = \alpha + \beta_1.X_1 + \beta_2.X_2 + \dots + \beta_k.X_k + u_i$$



Diagnóstico de Multicolinearidade

- **Multicolinearidade:** consequência da existência de alta correlação entre duas ou mais variáveis explicativas (preditoras).
- Possibilidade de interpretações erradas pela eventual distorção dos sinais dos parâmetros.
- Erros nas previsões.
- **Como detectar a multicolinearidade?**
 - Sinais inesperados dos coeficientes.
 - Testes t não significantes e teste F significativo.



Diagnóstico de Multicolinearidade

$$Y_i = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i} + \dots + \beta_k \cdot X_{ki} + u_i$$

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{U}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \dots \\ Y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1k} \\ 1 & X_{21} & X_{22} & \dots & X_{2k} \\ 1 & X_{31} & X_{32} & \dots & X_{3k} \\ \dots & \dots & \dots & \dots & \dots \\ 1 & X_{n1} & X_{n2} & \dots & X_{nk} \end{bmatrix}_{n \times k+1} \cdot \begin{bmatrix} a \\ b_1 \\ b_2 \\ \dots \\ b_k \end{bmatrix}_{k+1 \times 1} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ \dots \\ u_n \end{bmatrix}_{n \times 1}$$

$$\beta = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$$



Fontes Geradoras da Multicolinearidade

- 1 - Existência de variáveis que apresentam a mesma tendência durante alguns períodos, em decorrência da seleção de uma amostra que inclua apenas observações referentes a estes períodos.**
- 2 - Utilização de amostras com reduzido número de observações.**
- 3 - Utilização de valores defasados em algumas das variáveis explicativas como “novas” explicativas.**

Consequências da Multicolinearidade

$$Y_i = \alpha + \beta_1 \cdot X_{1i} + \beta_2 \cdot X_{2i}$$

(a) Correlação Perfeita:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 2 & 8 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 20 \\ 20 & 80 \end{bmatrix}$$

e, portanto, $\det(\mathbf{X}'\mathbf{X}) = 0$, ou seja, $(\mathbf{X}'\mathbf{X})^{-1}$ não pode ser definida.

Consequências da Multicolinearidade

(b) Correlação Muito Alta, porém Não Perfeita:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 2 & 7,9 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 19,8 \\ 19,8 & 78,41 \end{bmatrix}$$

de onde vem que $\det(\mathbf{X}'\mathbf{X}) = 0,01$ e, portanto:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 7.841 & -1.980 \\ -1.980 & 500 \end{bmatrix}$$

Consequências da Multicolinearidade

(c) Correlação Baixa:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 2 & 3 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 5 & 10 \\ 10 & 25 \end{bmatrix}$$

de onde vem que $\det(\mathbf{X}'\mathbf{X}) = 25$ e, portanto:

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1 & -0,4 \\ -0,4 & 0,2 \end{bmatrix}$$

Consequências da Multicolinearidade

- 1 – As significâncias estatísticas dos parâmetros $\beta = (X'X)^{-1}X'Y$ são sensíveis às correlações entre as variáveis explicativas.
- 2 – Os elementos da diagonal principal da matriz $(X'X)^{-1}$ aparecem no denominador da estatística t . Como a presença da multicolinearidade gera valores muito altos na diagonal da referida matriz, como vimos, ocorre a redução no valor da estatística t , sem alteração no cálculo da estatística F .

Identificação da Multicolinearidade

Regressões auxiliares entre cada uma das explicativas e as demais explicativas:

$$X_2 = \beta_1 + \beta_2.X_3 + \dots + \beta_{k-1}.X_k$$

$$X_3 = \beta_1 + \beta_2.X_2 + \dots + \beta_{k-1}.X_k$$

...

$$X_k = \beta_1 + \beta_2.X_2 + \dots + \beta_{k-1}.X_{k-1}$$

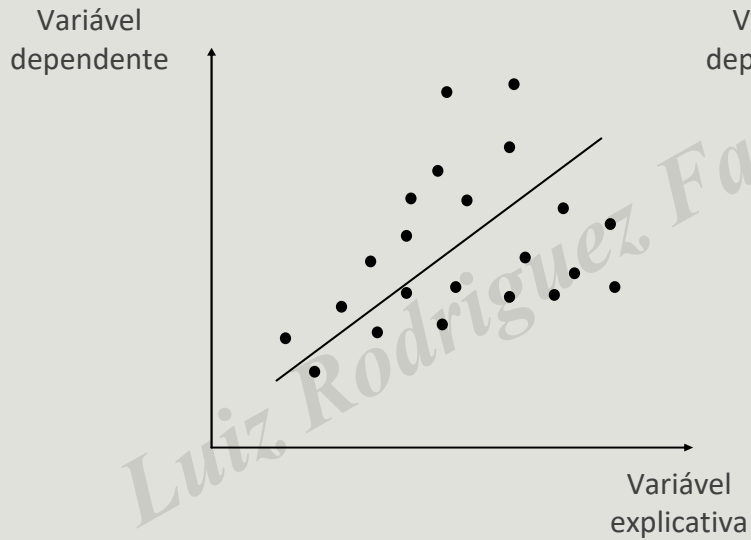
Estatísticas VIF (*Variance Inflation Factor*) e *Tolerance*:

$$Tolerance = 1 - R_p^2$$

$$VIF = 1 / Tolerance$$

em que o R_p^2 é o coeficiente de ajuste da regressão da variável explicativa X_p ($p = 2, 3, \dots, k$) com as demais variáveis explicativas.

Diagnóstico de Heterocedasticidade





Obrigado

Prof. Dr. Luiz Paulo Fávero
LinkedIn