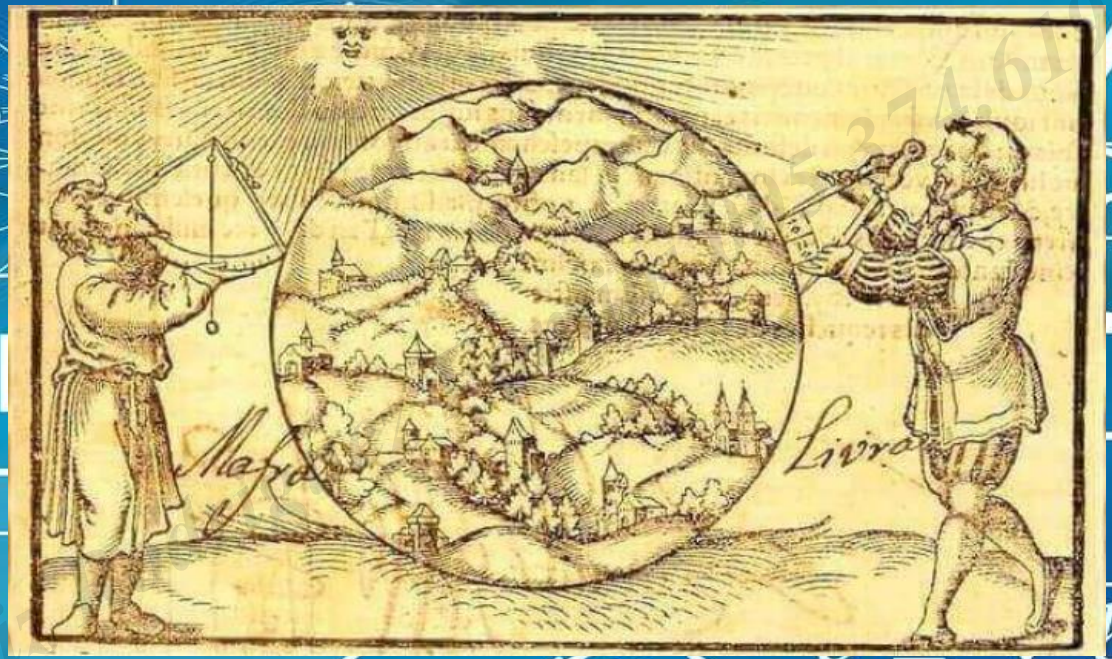
MBA USP ESALQ

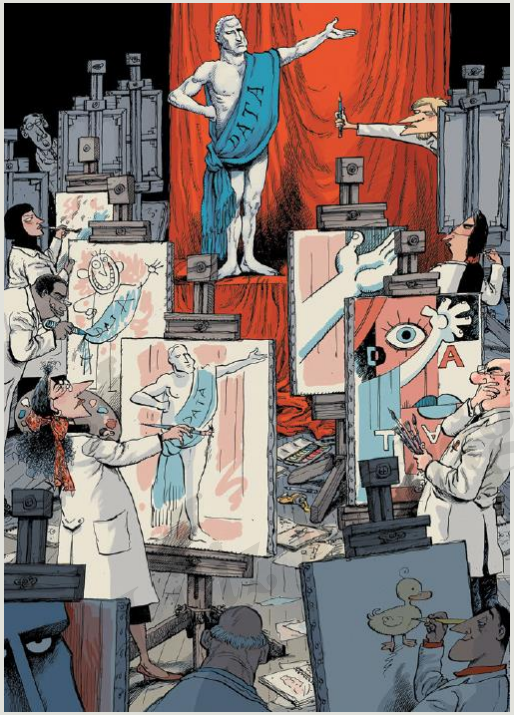
Supervised Machine Learning: Análise de Regressão Simples e Múltipla

Prof. Dr. Luiz Paulo Fávero

Luiz Rodriguez Fantini 205.374.619-81


Mafra
Livro

Reflexão
Modelos Supervisionados

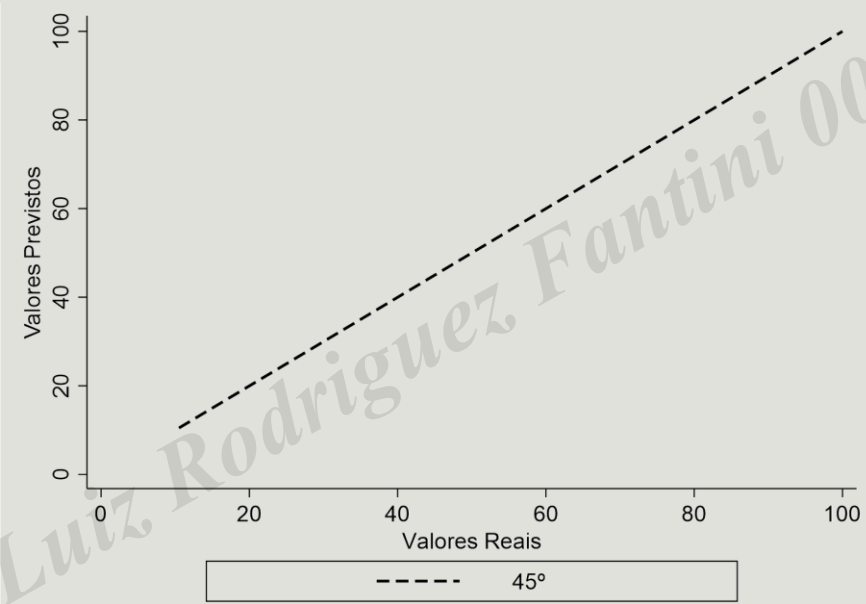


“Diferentes pesquisadores, a partir de uma mesma base de dados, podem estimar diferentes modelos e, consequentemente, obter diferentes valores previstos do fenômeno em estudo. O objetivo é estimar modelos que, embora simplificações da realidade, apresentem a melhor aderência possível entre os valores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

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

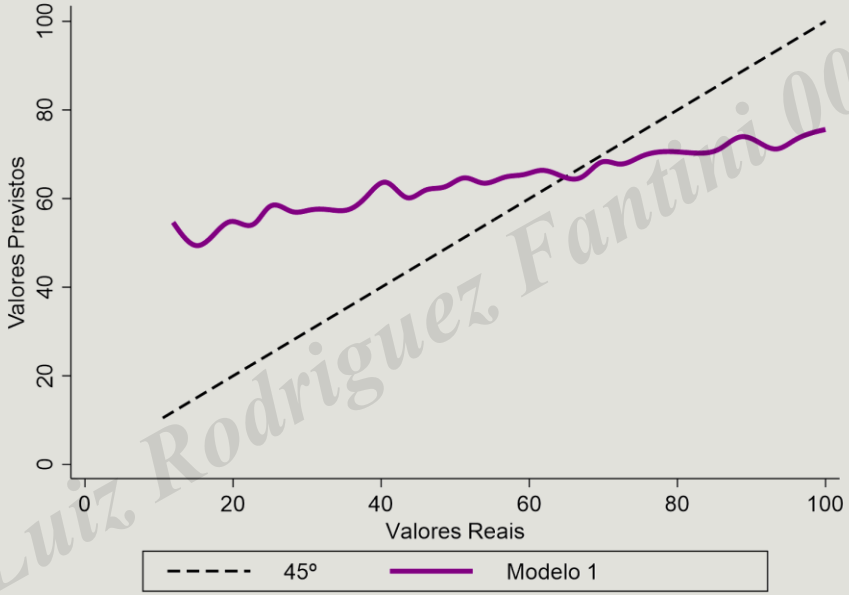
Rodrigues Fantini 005.374.619-81

Reflexão
Modelos Supervisionados



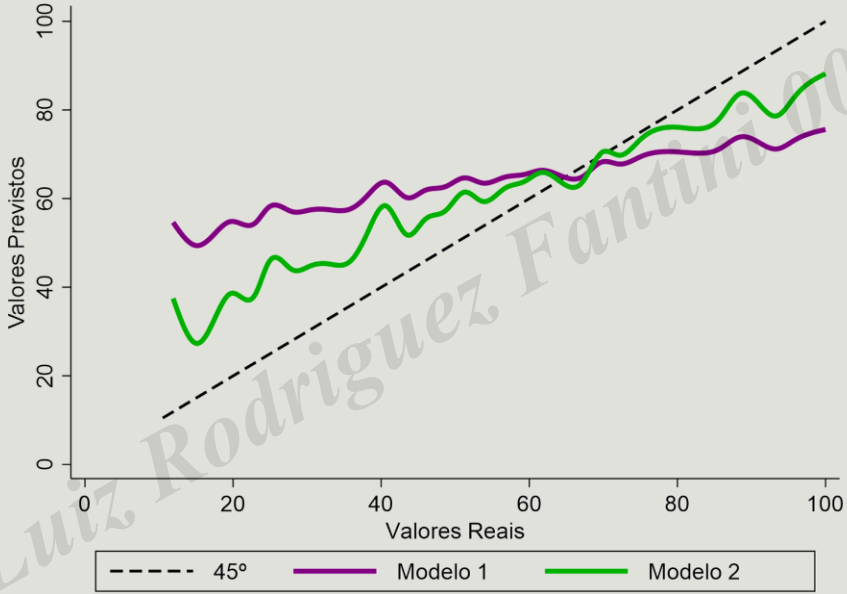
Luiz Rodriguez Fantini 005.374.619-81

Reflexão
Modelos Supervisionados



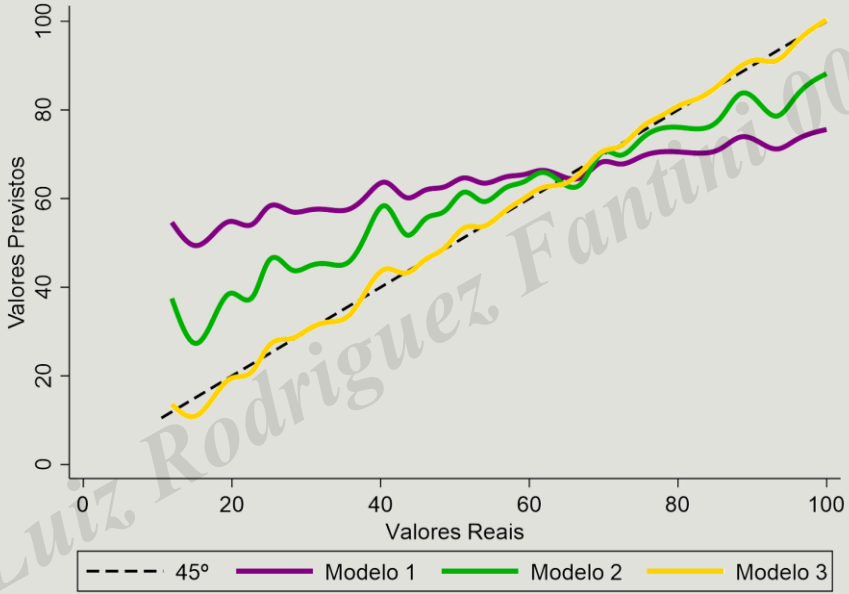
Luiz Rodriguez Fantini 005.374.619-81

Reflexão
Modelos Supervisionados



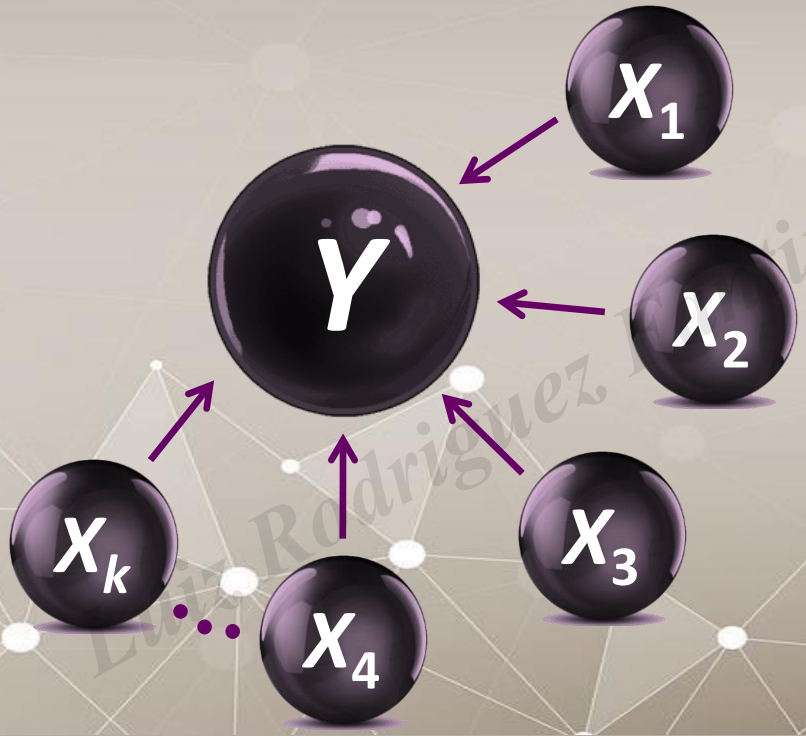
Luiz Rodriguez Fantini 005.374.619-81

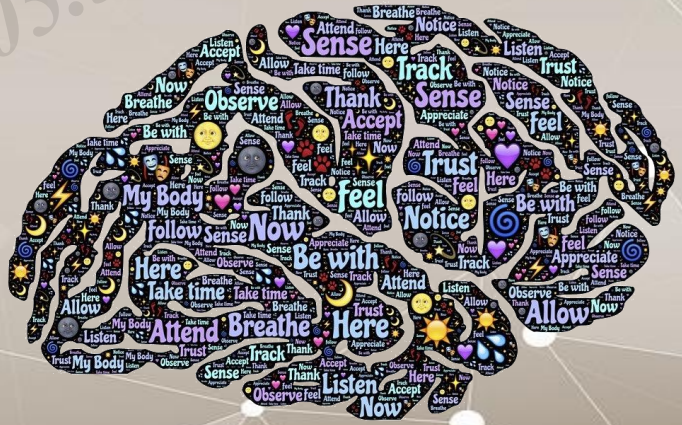
Reflexão
Modelos Supervisionados



Luiz Rodriguez Fantini 005.374.619-81

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





Luiz Rodriguez Marcini 005.374.619-81

Modelos Lineares Generalizados (GLM)

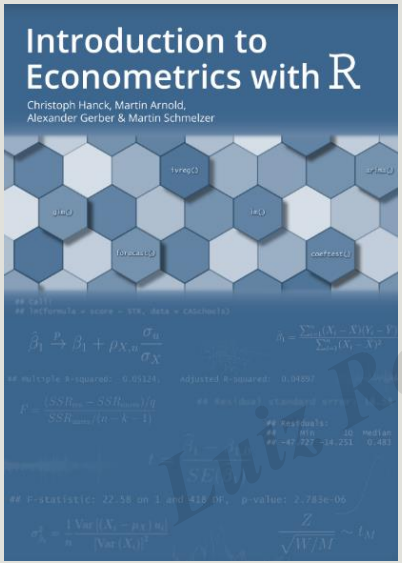
$$\eta_i = \alpha + \beta_1 . X_{1i} + \beta_2 . X_{2i} + \ldots \beta_k . X_{ki}$$

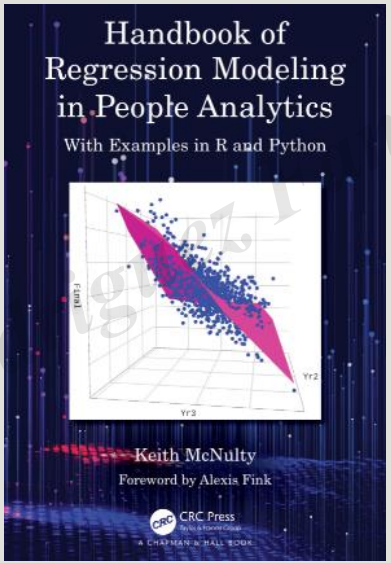
Modelos lineares generalizados, características da variável dependente e funções de ligação canônica.

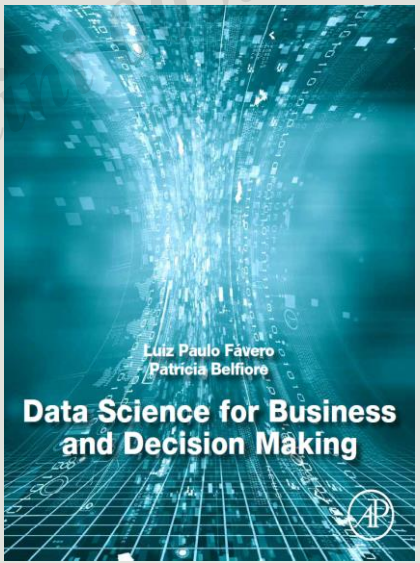
Modelo de Regressão	Característica da Variável Dependente	Distribuição	Função de Ligação Canônica (η)
Linear	Quantitativa	Normal	$\hat{Y}$
Com Transformação de Box-Cox	Quantitativa	Normal Após a Transformação	$\frac{\hat{Y}^{\lambda} - 1}{\lambda}$
Logística Binária	Qualitativa com 2 Categorias (*Dummy*)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

Luiz Rodriguez Fantini 005.374.619-81

Modelos Supervisionados:
Modelos Lineares Generalizados (GLM)







Luiz Rodriguez, 105.374.619-81

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

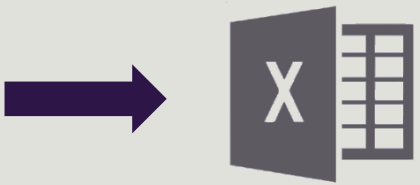
Regressão Linear Simples

Objetivo:

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

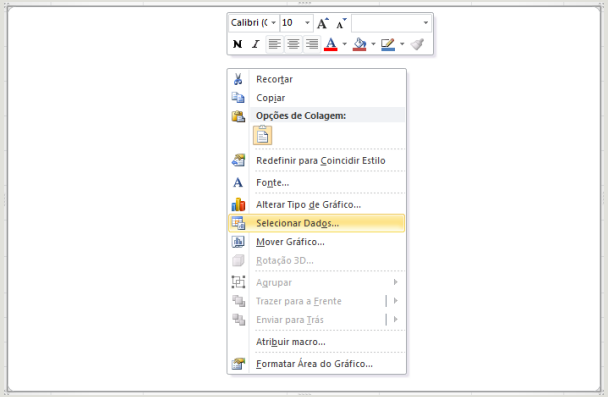
Equação linear de uma reta num plano cartesiano:

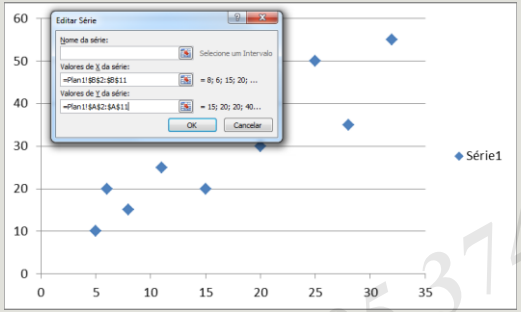
$$Y_i = \alpha + \beta . X_i + u_i$$

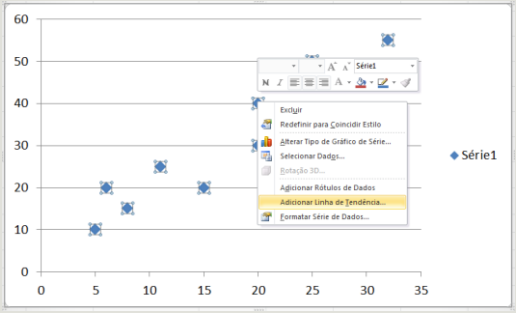


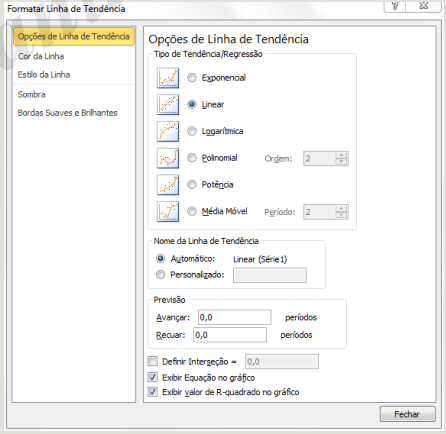
em que temos um intercepto (α), um coeficiente de inclinação da reta (β), uma variável explicativa X e um termo de erro u.

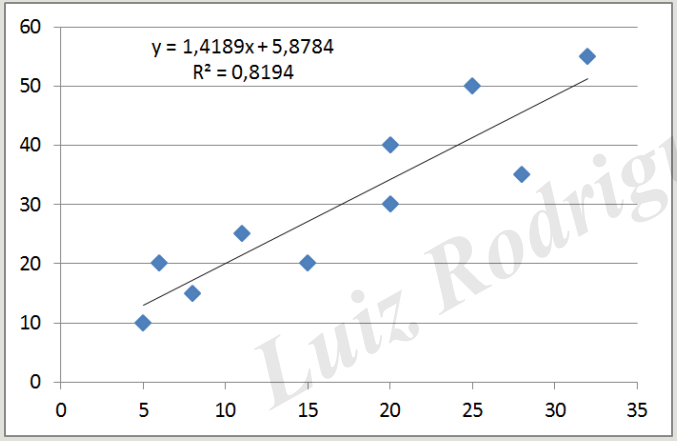
Luiz Rodrigues Fantini 005.374.619-81











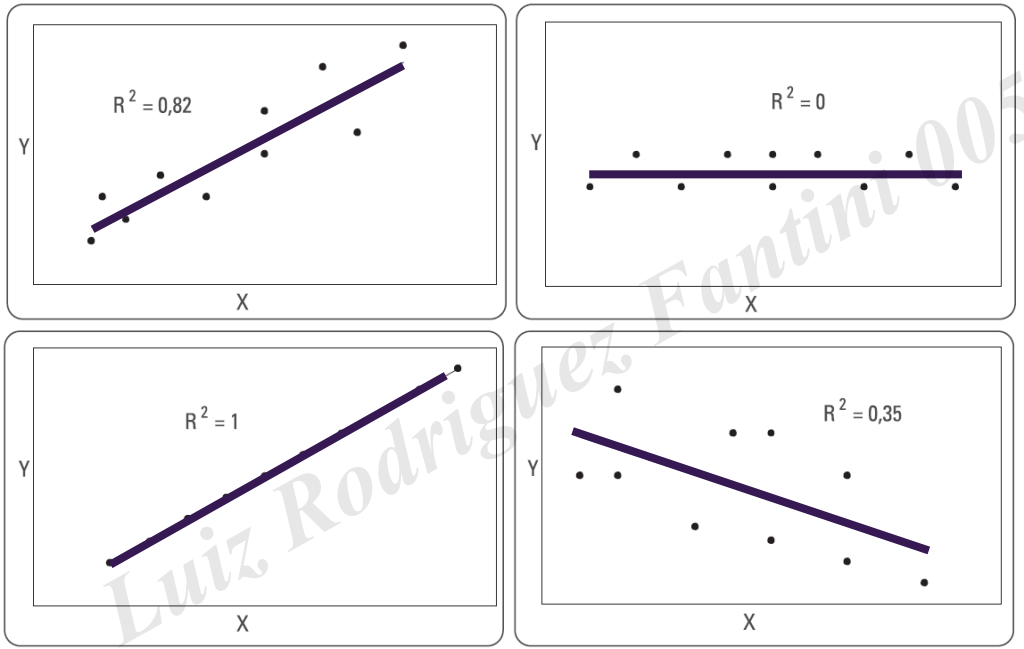
Luiz Rodriguez Fontini 005.374.619-81

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).

Luiz Rodriguez Fantini 005.374.619-81

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



Comportamento do R^2 para diferentes regressões lineares simples.

Luiz Rodriguez Fantini 005.374.619-81

Análise de Regressão:
Estimação dos Parâmetros

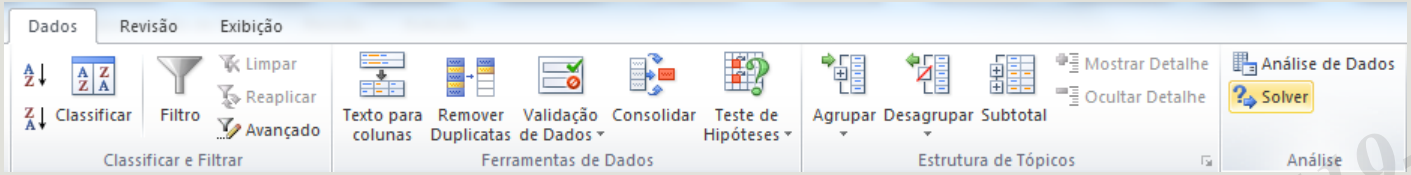
Critérios:

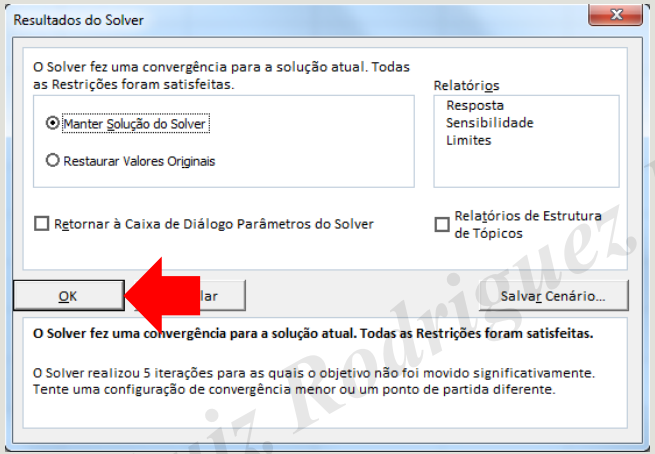
1 – Soma dos erros igual a zero: $\sum_{i=1}^{n} u_i = 0$

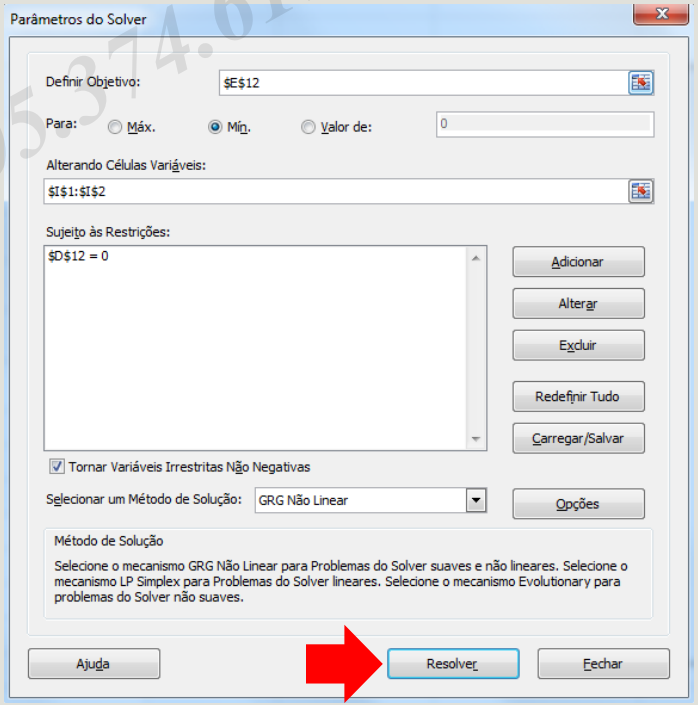
2 – Soma dos erros ao quadrado sendo a mínima possível: $\sum_{i=1}^{n} u_i^2 = \text{mín}$

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.

Luiz Rodriguez Fontini 005.374.619-81



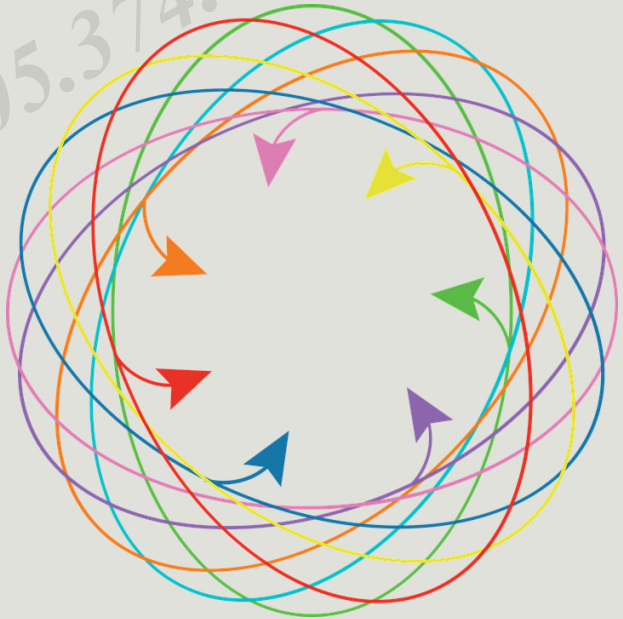




Luiz Rodriguez Fantini 005.374.619-81

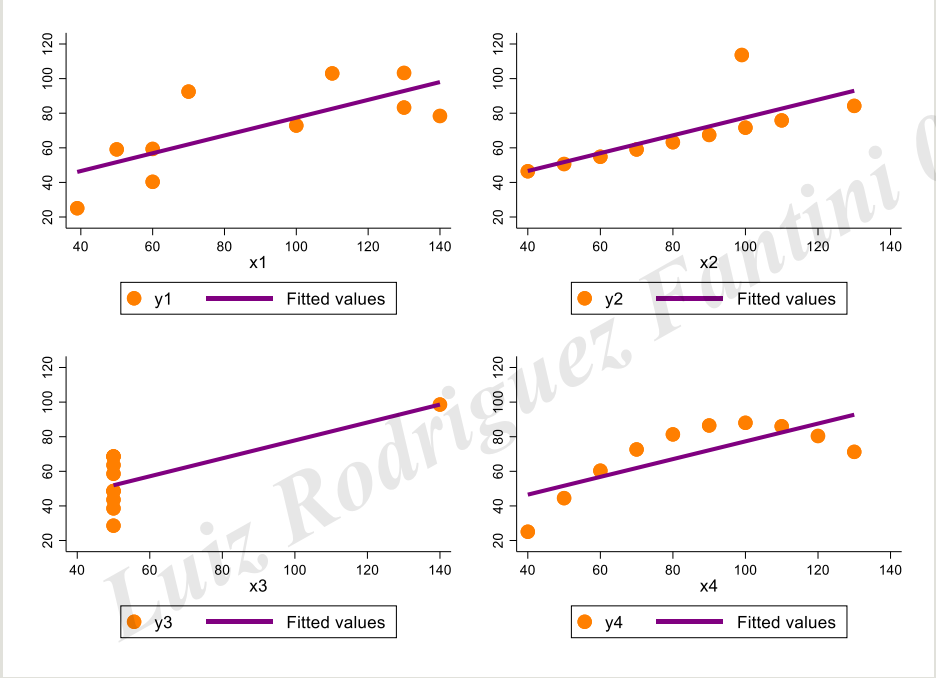
Análise de Regressão:
Cálculo do R^2

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



Luiz Rodrigues Fantini 005.374.619-81

Apenas Parâmetros e R^2 ?



Luiz Rodriguez Fontini 005.374.619-81

Modelos de Regressão no R

ez, Fantini 005.374.619-81

Significância Estatística do Modelo

- **Teste *F*:** Permite analisar se pelo menos um dos β's é estatisticamente significante para a explicação do comportamento de *Y*.

  - **Hipóteses:** H_0: $\beta_1 = \beta_2 = \beta_3 = \cdots = \beta_k = 0$ 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).

Luis Rodriguez Fantini 005.374.619-81

Significância Estatística dos Parâmetros do Modelo

Luis Rodriguez Funtini 005.374.619-81

- **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.

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

Luiz Rodriguez Fantini 065.574.619-81

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:

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

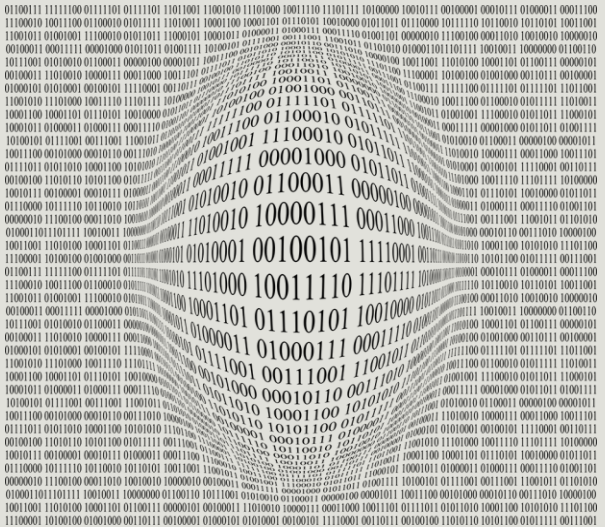
Luiz Rodriguez Fantini 005.374.619-81

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?



Luiz Rodrigues Fantini 005.374.619-81

Variáveis Explicativas (*X*) Qualitativas

Variáveis *dummy*

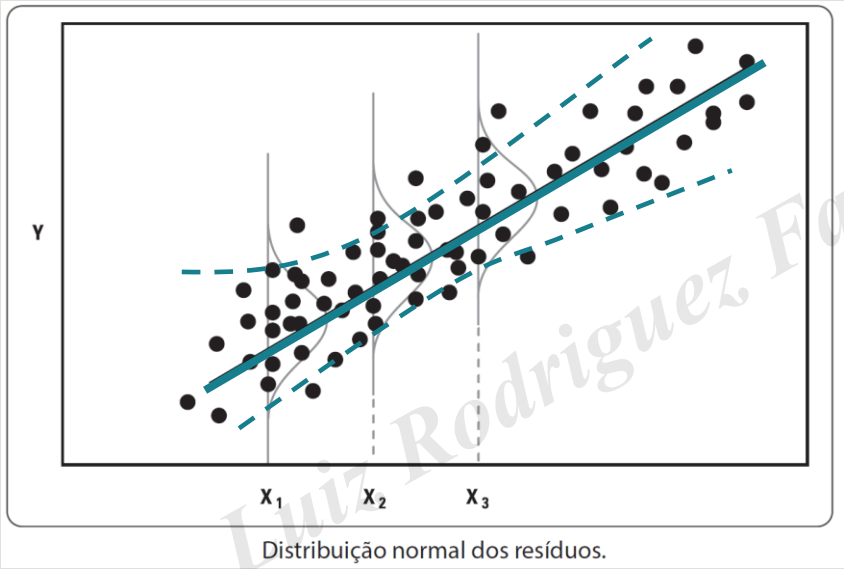
São variáveis categóricas que representam um atributo por meio de combinação binária (0 para a ausência ou 1 para presença).

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

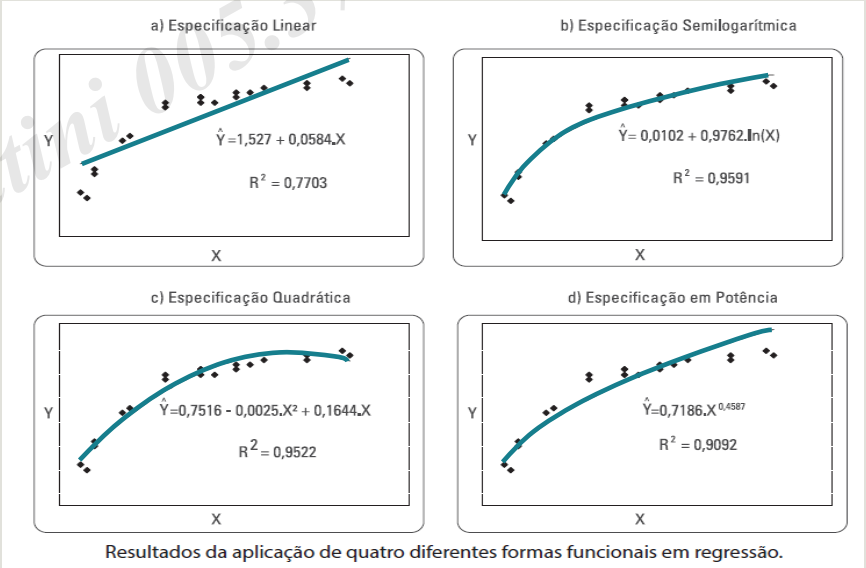
Neste caso, devemos incluir *n* – 1 *dummies*, em que *n* é a quantidade de categorias existentes na variável original.

Luiz Rodrigues Fantini 005.374.619-81

Modelos Não Lineares



Distribuição normal dos resíduos.



Resultados da aplicação de quatro diferentes formas funcionais em regressão.

Luiz Rodriguez Fantini 005.374.619-81

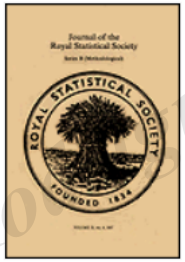
Modelos Não Lineares e Transformações de Box-Cox

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?

Luiz Rodriguez Fantini 005.374.619-81

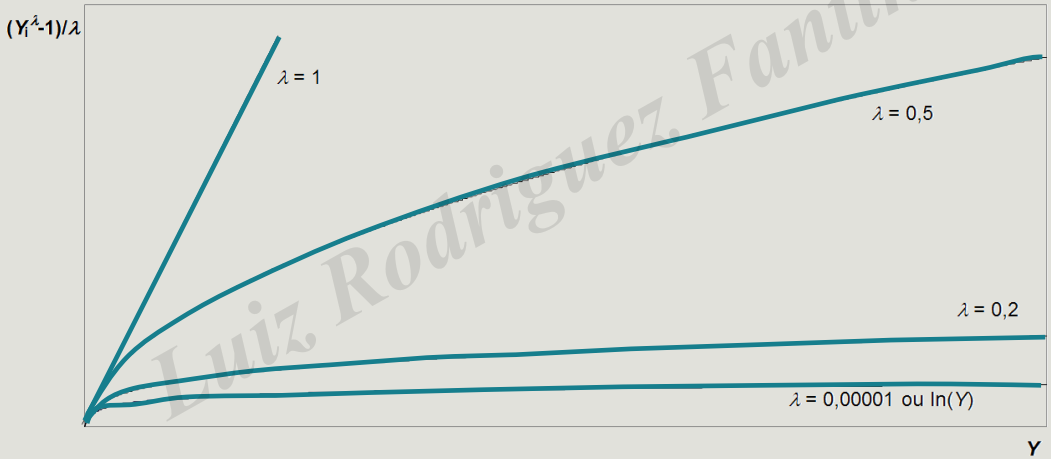
Modelos Não Lineares e Transformações de Box-Cox

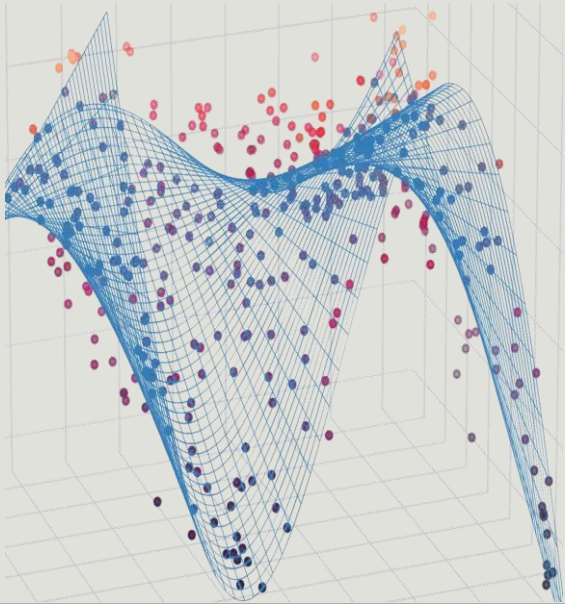
$Y_i = \alpha + \beta_1.X_1 + \beta_2.X_2 + ... + \beta_k.X_k$	Especificação Linear ($\lambda = 1$)
$Y_i^2 = \alpha + \beta_1.X_1 + \beta_2.X_2 + ... + \beta_k.X_k$	Especificação Quadrática ($\lambda = 2$)
$Y_i^3 = \alpha + \beta_1.X_1 + \beta_2.X_2 + ... + \beta_k.X_k$	Especificação Cúbica ($\lambda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

Luiz Rodrigues Fantini 003.374.619-81

Modelos Não Lineares e Transformações de Box-Cox

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





Luiz Rodriguez Fantini 005.374.619-81

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 predições.**

- **Como detectar a multicolinearidade?**
  - Sinais inesperados dos coeficientes.
  - Testes *t* não significantes e teste *F* significante.

Luis Rodriguez Fantini 005 374.619-81

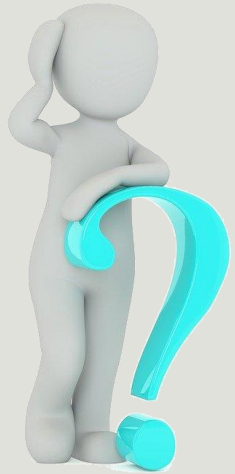
Diagnóstico de Multicolinearidade

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

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

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ ... \\ Y_n \end{bmatrix}_{nx1} = \begin{bmatrix} 1 & X_{11} & X_{12} & ... & X_{1k} \\ 1 & X_{21} & X_{22} & ... & X_{2k} \\ 1 & X_{31} & X_{32} & ... & X_{3k} \\ ... & ... & ... & ... & ... \\ 1 & X_{n1} & X_{n2} & ... & X_{nk} \end{bmatrix}_{nxk+1} . \begin{bmatrix} a \\ b_1 \\ b_2 \\ ... \\ b_k \end{bmatrix}_{k+1x1} + \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ ... \\ u_n \end{bmatrix}_{nx1}$$

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



Luiz Rodrigues Fantini 005.374.619-81

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.

Luiz Rodriguez Farina: 005.374.619-81

Consequências da Multicolinearidade

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

(a) Correlação Perfeita:

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

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

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

Luiz Rodriguez Furtini 005.374.619-81

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(X'X)** = 0,01 e, portanto:

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

Luiz Rodriguez Fantini 005.374.619-81

Consequências da Multicolinearidade

(c) Correlação Baixa:

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

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

de onde vem que **det(X'X)** = 25 e, portanto:

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

Luiz Rodriguez Fantini 005.374.619-81

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.*

Luiz Roberto Fantini 005.374.619-81

Identificação da Multicolinearidade

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

$$X_2 = \beta_1 + \beta_2 X_3 + \ldots + \beta_{k-1} X_k$$
$$X_3 = \beta_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_k$$
$$\ldots$$
$$X_k = \beta_1 + \beta_2 X_2 + \ldots + \beta_{k-1} X_{k-1}$$

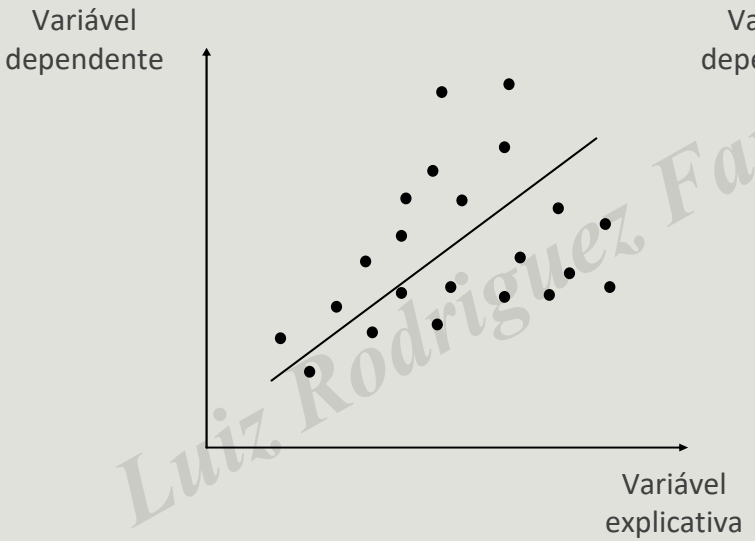
Estatísticas VIF (*Variance Inflation Factor*) e *Tolerance*:

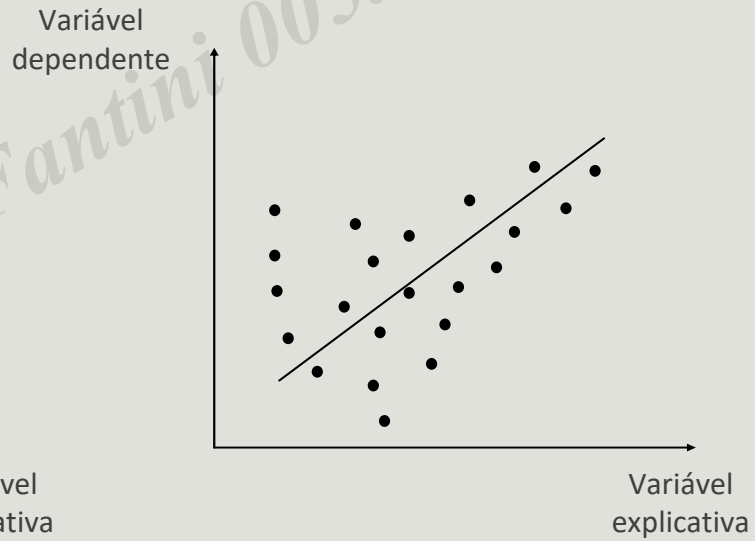
$$\textit{Tolerance} = 1 - R_p^2$$
$$\text{VIF} = 1 / \textit{Tolerance}$$

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

Luiz Rodrigues Fantini 005.374.619-81

Diagnóstico de Heterocedasticidade





Luiz Rodriguez Fantini 005.374.619-81

Obrigado
Prof. Dr. Luiz Paulo Fávero
LinkedIn