MBA Business Analytics e Big Data

Análise Preditiva

Prof. Dr. João Rafael Dias

1º semestre - 2020

Agenda
Na aula de hoje...

Aprendizagem supervisionada
Regressão e classificação
Formas de treino e validação
Bias-variance trade-off
Avaliação e comparação de modelos
Prática no RStudio

Estrutura de uma árvore de decisão
Intuição
Particionamento dos nós na regressão e classificação
Poda da árvores vs *overfitting*

Introdução e motivações
Feature engineering
Tratamento de variáveis
Transformação de variáveis
Arquivos de trabalho
Prática no RStudio

Regressão linear múltipla
Coeficiente de determinação
Regressão logística
Odds e log odds
Comparação entre as regressões
Multicolinearidade
Seleção de variáveis *step-wise*
Prática no RStudio

Modelos de *ensemble*
Bootstrap
Random forest
Adaptive boosting
Prática no RStudio

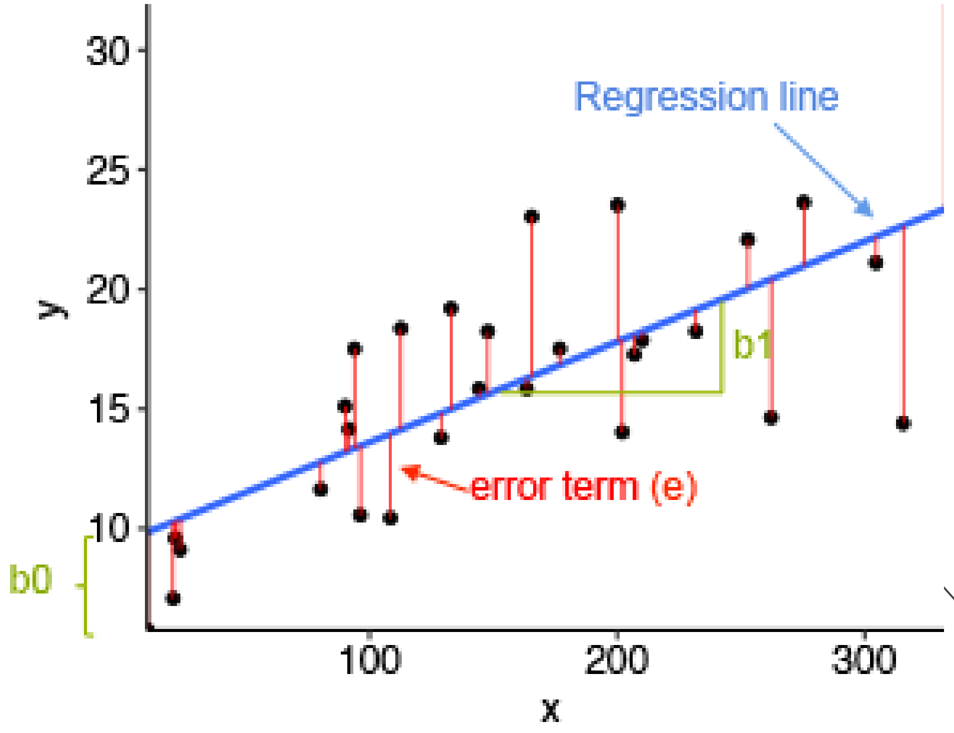
Regressão linear múltipla

Regressão linear múltipla

Overview

- O objetivo da análise de regressão linear é **estimar o valor** de variável *target* y dado que os valores das variáveis explicativas **x** sejam **conhecidos**
- Em outras palavras, a regressão é usada para prever uma variável *y* através de outras x_1, x_2, ..., x_n sejam elas quantitativas ou qualitativas

...lembrando da regressão simples



$$\hat{y} = b_0 + b_1 x_1$$

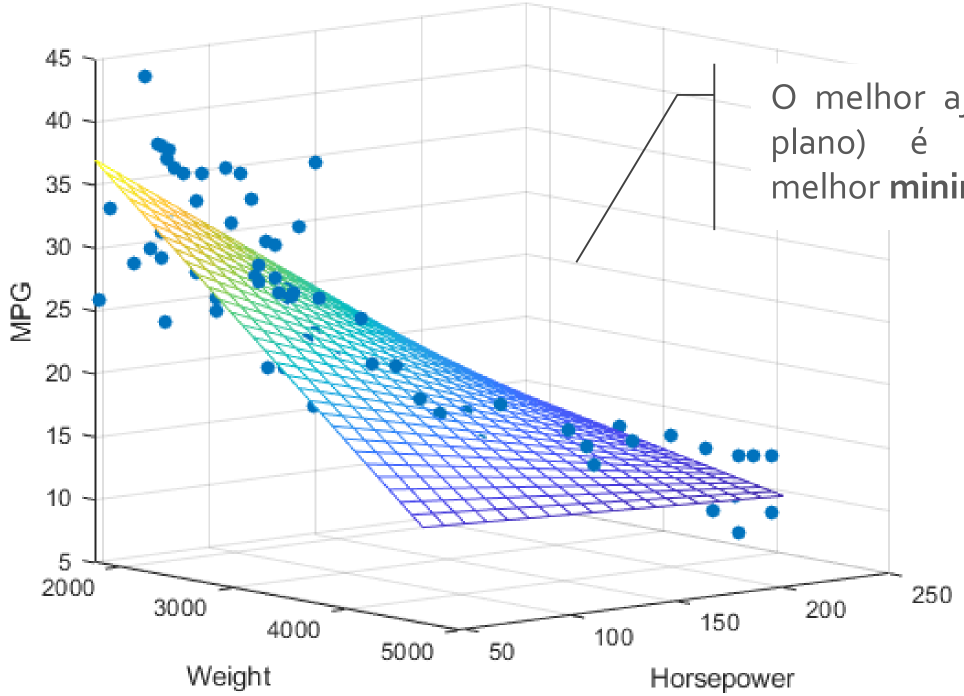
A previsão do target y contínuo, sendo representada por um ajuste linear, onde:
b_0 é o intercepto
b_1 é o coeficiente da variável x_1
x_1 variável independente (qualitativa ou quantitativa)

O melhor ajuste (melhor reta) é aquela que melhor **minimiza** o MSE!

Regressão linear múltipla
Relação funcional

- O objetivo da análise de regressão linear é **estimar o valor** de variável *target* y dado que os valores das variáveis explicativas **x** sejam **conhecidos**
- Em outras palavras, a regressão é usada para prever uma variável *y* através de outras x_1, x_2, ..., x_n sejam elas quantitativas ou qualitativas

Na regressão múltipla temos n variáveis



O melhor ajuste (melhor plano) é aquela que melhor **minimiza** o MSE!

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_n x_n$$

A previsão do target y contínuo, sendo representada por um ajuste linear, onde:

b_o é o intercepto

b_1 , b_2, ..., b_n são os coeficientes das variáveis

x_1 , x_2, ..., x_n variáveis independentes (qualitativas* ou quantitativas)

*na forma de *dummy variable*

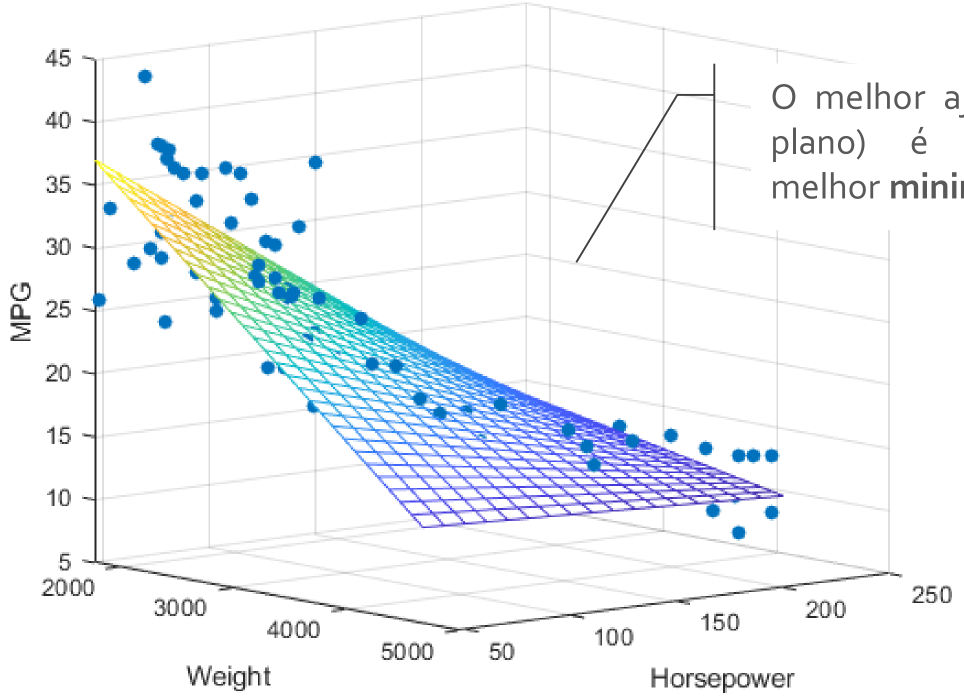
https://medium.com/analytics-vidhya/new-aspects-to-consider-while-moving-from-simple-linear-regression-to-multiple-linear-regression-dad06b3449ff

Regressão linear múltipla

Relação funcional

- O objetivo da análise de regressão linear é **estimar o valor** de variável *target* y dado que os valores das variáveis explicativas **x** sejam **conhecidos**
- Em outras palavras, a regressão é usada para prever uma variável *y* através de outras x_1, x_2, ..., x_n sejam elas quantitativas ou qualitativas

Na regressão múltipla temos n variáveis



O melhor ajuste (melhor plano) é aquela que melhor **minimiza** o MSE!

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \cdots + b_nx_n$$

b_0 é o intercepto ou **coeficiente linear**, É o valor previsto de y quando todas as variáveis forem nulas

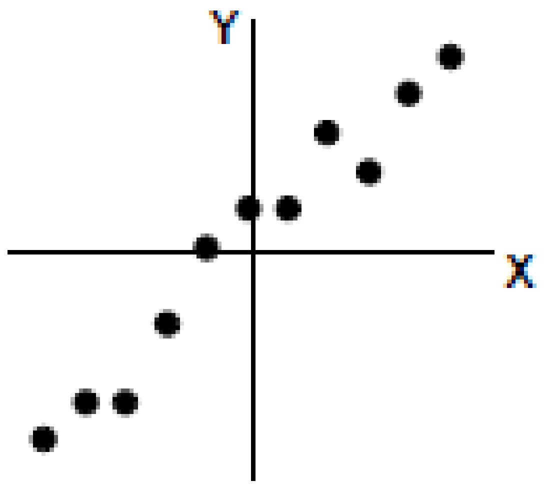
b_1 , b_2, ..., b_n são os **coeficientes angulas** ou pesos das variáveis. É a variação em y quando x tem um incremento de uma unidade

https://medium.com/analytics-vidhya/new-aspects-to-consider-while-moving-from-simple-linear-regression-to-multiple-linear-regression-dad06b3449ff

Regressão linear múltipla

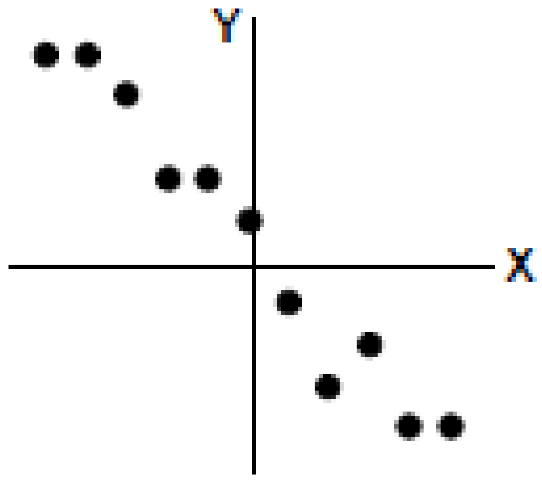
Coeficiente de determinação

- Na ausência de uma relação linear perfeita entre duas variáveis sempre existe uma incerteza remanescente
- Em outras palavras, existe sempre uma proporção de variância (incerteza) na variável *y* que permanece não esclarecida após o ajuste linear ter sido realizado
- Dessa forma o coeficiente de determinação R^2 indica a proporção da variância em *y* que estatisticamente explicada pela regressão



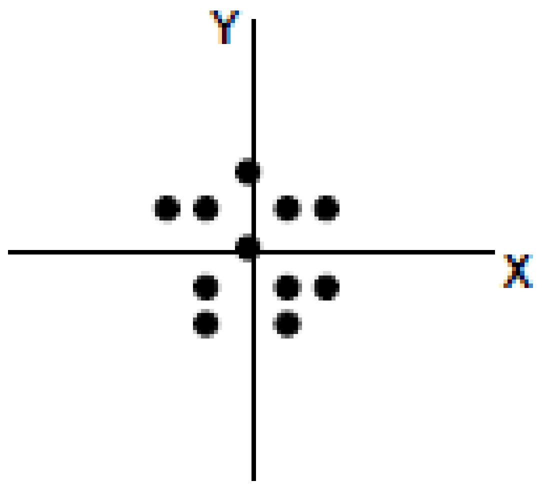
$$R^2 \cong 1$$

$$r \cong 1$$



$$R^2 \cong 1$$

$$r \cong -1$$

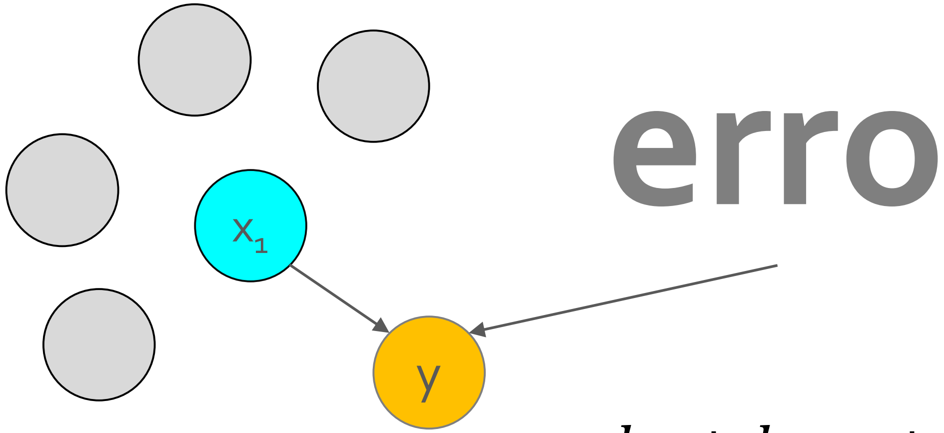


$$R^2 \cong 0$$

$$r \cong 0$$

Regressão linear múltipla

A lógica das *n* variáveis



$$y = \ b_0 + b_1 x_1 + erro$$

$$y = \ b_0 + b_1 x_1 + b_2 x_2 + erro$$

$$y = \ b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + erro$$

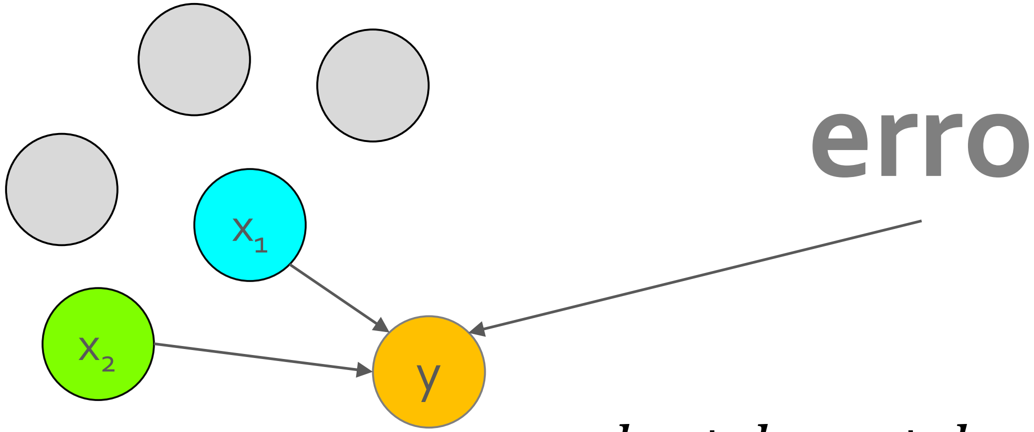
Aqui o ajuste de apenas uma variável pode levar à um R^2 baixo e um erro elevado para as previsões do modelo

Regressão linear múltipla

A lógica das *n* variáveis

$$y = \ b_0 + b_1x_1 + erro$$

$$y = \ b_0 + b_1x_1 + b_2x_2 + b_3x_3 + erro$$



$$y = \ b_0 + b_1x_1 + b_2x_2 + erro$$

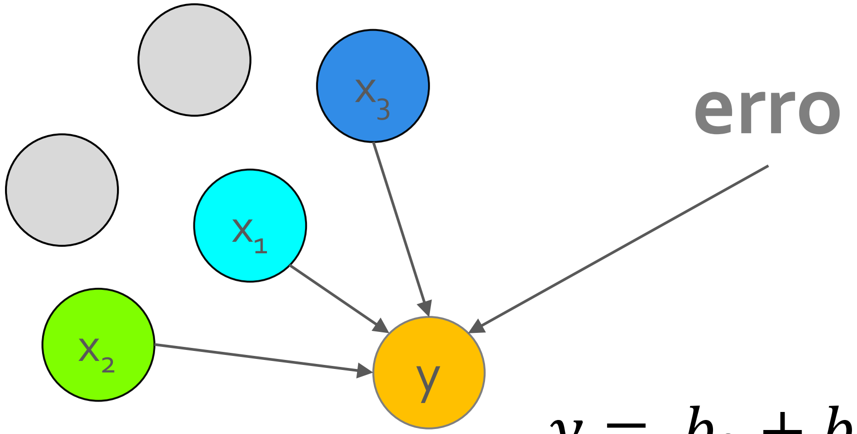
Aqui R^2 aumenta com a entrada de nova variável, o erro diminui. Quanto mais a variável "contribuir" para o modelo, maior o aumento do R^2 e a diminuição do erro do modelo

Regressão linear múltipla

A lógica das *n* variáveis

$$y = b_0 + b_1 x_1 + erro$$

$$y = b_0 + b_1 x_1 + b_2 x_2 + erro$$



$$y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + erro$$

É necessário encontrar variáveis relacionadas com y para construir modelo com R^2 alto e erro baixo!

Regressão linear múltipla
Prós e contras do algoritmo

Prós	Contras
É um método extremamente simples e intuitivo	É um método muito simples, pois assume que a relação é linear
Mesmo que não ajuste-se aos dados exatamente, ela consegue medir natureza da relação entre x e y	Assume que os resíduos são independentes (distribuição aleatória) e possuem uma distribuição normal e variância constante
Pode-se adicionar interação entre os termos ou termos com potência	Apresenta problemas de ajuste quando existem variáveis correlacionadas entre si
Permite uma interpretação clara dos pesos e as relações das variáveis	Descarta registros com *missing values* para a estimação dos parâmetros
	Para utilizar varáveis categóricas necessário usar o conceito de *dummy variables*

Regressão logística

Regressão logística

Overview

- O algoritmo da regressão logística fornece uma relação funcional f(x) e um vetor de parâmetros **b** para expressar a probabilidade de y dado o conjunto de variáveis explicativas **x**
- Suponha que uma observação/indivíduo possa pertencer a um dos dois *labels* pré-determinados A e B. Com isso a variável reposta é

***label* A**

Evento
Categoria +

$y = 1$

$y = 0$

***label* B**

Não Evento
Categoria -

A regressão logística permite estimar as probabilidades que uma observação pertença a cada um dos grupos a partir das características da observação (i.e. variáveis explicativas ou *features*)

$y = ?$

$P(y = 1 \mid \mathbf{x})$

Regressão logística

Regressão logística
Probabilidades e chance

- O conceito matemático central que permeia esse tipo de regressão é o *logit* – o logaritmo natural da chance
- Considere uma observação com variável *target* y e explicativas $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n)$

  - probabilidade de que x pertença ao *label* A: $P(y = 1 \mid \mathbf{x})$

  - probabilidade de que x pertença ao *label* B: $P(y = 0 \mid \mathbf{x})$

  - podemos considerar: $P(y = 0 \mid \mathbf{x}) = 1 - P(y = 1 \mid \mathbf{x})$

Chance (*odds*)

$$odds = \frac{P(y = 1|\boldsymbol{x})}{P(y = 0|\boldsymbol{x})}$$

$$odds = \frac{P(y = 1|\boldsymbol{x})}{1 - P(y = 1|\boldsymbol{x})}$$

Exemplo

Se $P(y = 1) = 0{,}25$ então $P(y = 0) = 0{,}75$, com isso a razão de chances será de 1 : 3. Podemos interpretar que, de cada 4 observações com as mesmas características (i.e. variáveis explicativas), 1 pertence ao *label A* e 3 ao *label B*.

Regressão logística
Relação funcional

- A regressão logística assume que o *logit* é linearmente relacionado com as variáveis independentes do modelo. Dessa forma temos

$$ln\left[\frac{P(y=1|\,\boldsymbol{x})}{1\ -P(y=1|\,\boldsymbol{x})}\right]=relação\ linear\ x_1, x_2, \ldots, xn$$

$$ln\left[\frac{P(y=1|\,\boldsymbol{x})}{1\ -P(y=1|\,\boldsymbol{x})}\right]=b_0+b_1x_1+\cdots+b_nx_n$$

A probabilidade de ocorrência do evento de interesse (ou pertença a categoria +) será dada pela seguinte relação funciona (simplificando a notação)

$$P=\frac{1}{1+e^{-Z}} \qquad P=\frac{e^{Z}}{1+e^{Z}}$$

onde, $Z=b_0+b_1x_1+\cdots+b_nx_n$

Regressão logística

Interpretação dos parâmetros vs chance [extra]

- A chance estabelece a relação entre a probabilidade de pertencer a categoria positiva vs a probabilidade de não pertencer a categoria positiva

$$odds = \frac{P}{1 - P}$$

Probability	Odds	Log Odds
0.100	0.111	-2.197
0.200	0.250	-1.386
0.300	0.428	-0.847
0.400	0.667	-0.405
0.500	1.000	0.000
0.600	1.500	0.405
0.700	2.333	0.847
0.800	4.000	1.386
0.900	9.000	2.197

$$ln[odds] = b_0 + b_1 x_1 + \cdots + b_n x_n$$

$$odds = e^{(b_0 + b_1 x_1 + \cdots + b_n x_n)}$$

$$odds = e^{b_0} e^{b_1 x_1} e^{b_2 x_2} \ldots e^{b_n x_n}$$

Dessa forma, o sinal dos coeficientes afeta a chance de forma um pouco diferente de como é relacionado y com x na regressão linear múltipla

$$b_n > 0 \Rightarrow e^{b_n x_n} > 1$$

a chance **aumenta** conforme x **aumenta**

$$b_n < 0 \Rightarrow e^{b_n x_n} < 1$$

a chance **diminui** conforme x **aumenta**

Regressão logística

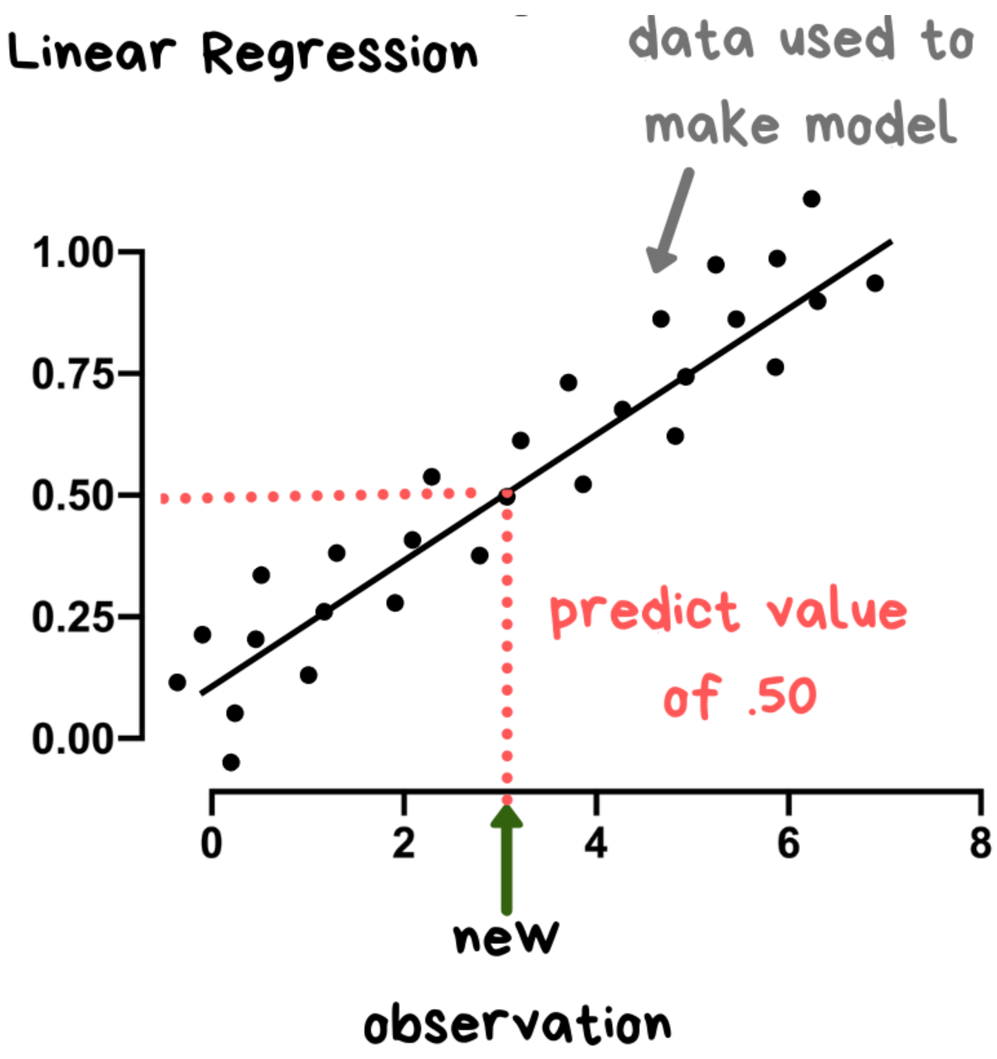
Prós e contras do algoritmo

Prós	Contras
É uma técnica robusta: as variáveis não precisam ser normalmente distribuídas ou ter uma variância igual em cada grupo	Ela não é útil quando não são identificadas todas as variáveis independentes
Não é assumida uma relação linear entre a variável dependente com as variáveis independentes	Na maior parte das aplicações trabalha-se com uma variável resposta categórica binária
Pode-se adicionar interação entre os termos ou termos com potência	Apresenta problemas de ajuste quando existem variáveis correlacionadas entre si
Prevê probabilidade diretamente	Descarta registros com *missing values* para a estimação dos parâmetros
Permite uma interpretação clara dos pesos e as relações das variáveis	Para utilizar varáveis categóricas necessário usar o conceito de *dummy variables*

Considerações para as regressões

Consideração para as regressões

Comparações entre os algoritmos



https://thenextweb.com/neural/2020/04/25/machine-learning-models-explained-to-a-five-year-old-syndication/

- Usada para problemas de regressão, onde a variável *target* é contínua
- Estima a variável dependente y quando há variação na(s) variável(eis) independentes x
- *Output* contínuo
- Assume relação linear entre as variáveis

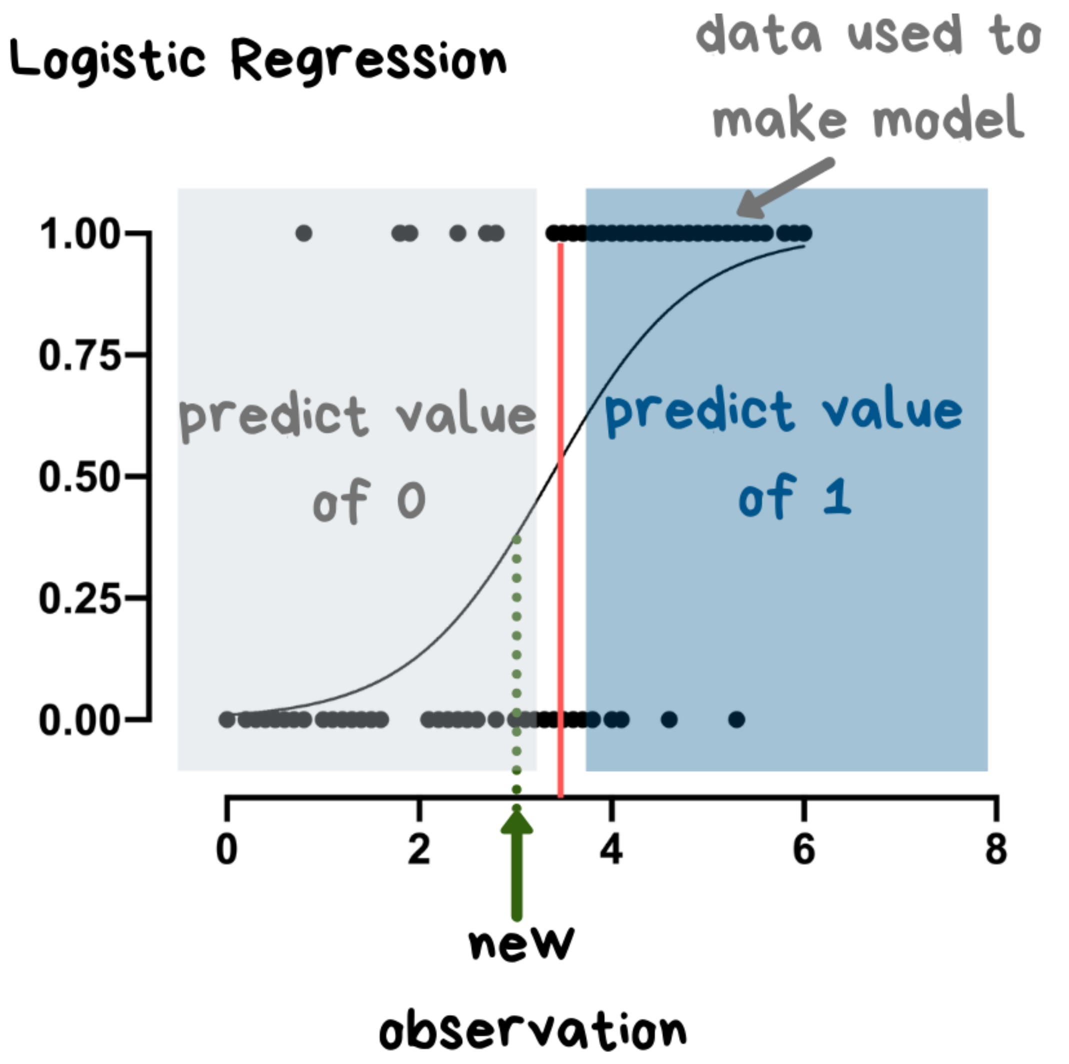
MSE – Mean Square Error

O MSE mede a média das diferenças quadráticas entre o valor real do *target* e o valor previsto pelo ajuste dado pelos coeficientes ($b_0, b_1, \ldots, b_n$)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Consideração para as regressões

Comparações entre os algoritmos



https://thenextweb.com/neural/2020/04/25/machine-learning-models-explained-to-a-five-year-old-syndication/

- Usada para problemas de classificação, onde o *target* é uma variável categórica
- Estima a probabilidade de pertencer a um *label*
- *Output* discreto
- Assume relação pode não ser linear entre as variáveis

MLE – Maximum Likelihood Estimate

A regressão logística não usa o critério de MSE, pois esse critério não tem as mesmas boas propriedades quando a variável resposta deixa de ser continua e passa a ser binária. Utiliza-se o critério de máxima verossimilhança para estimar os coeficientes.

$$MLE = \sum_{i=1}^{n} y_i ln p_i + (1 - y_i) ln(1 - p_i)$$

Considerações para as regressões

Significância estatística dos coeficientes

- A relação encontrada pelas regressões é estatisticamente significante?
- Em outras palavras, com base na amostra de desenvolvido conseguimos extrapolar para a população?
- Para decidir o *p*-valor e R^2 são métricas muito importantes (para logística não é aplicado)
- Lembrando: quando menor o *p*-valor maior é a evidência de relação

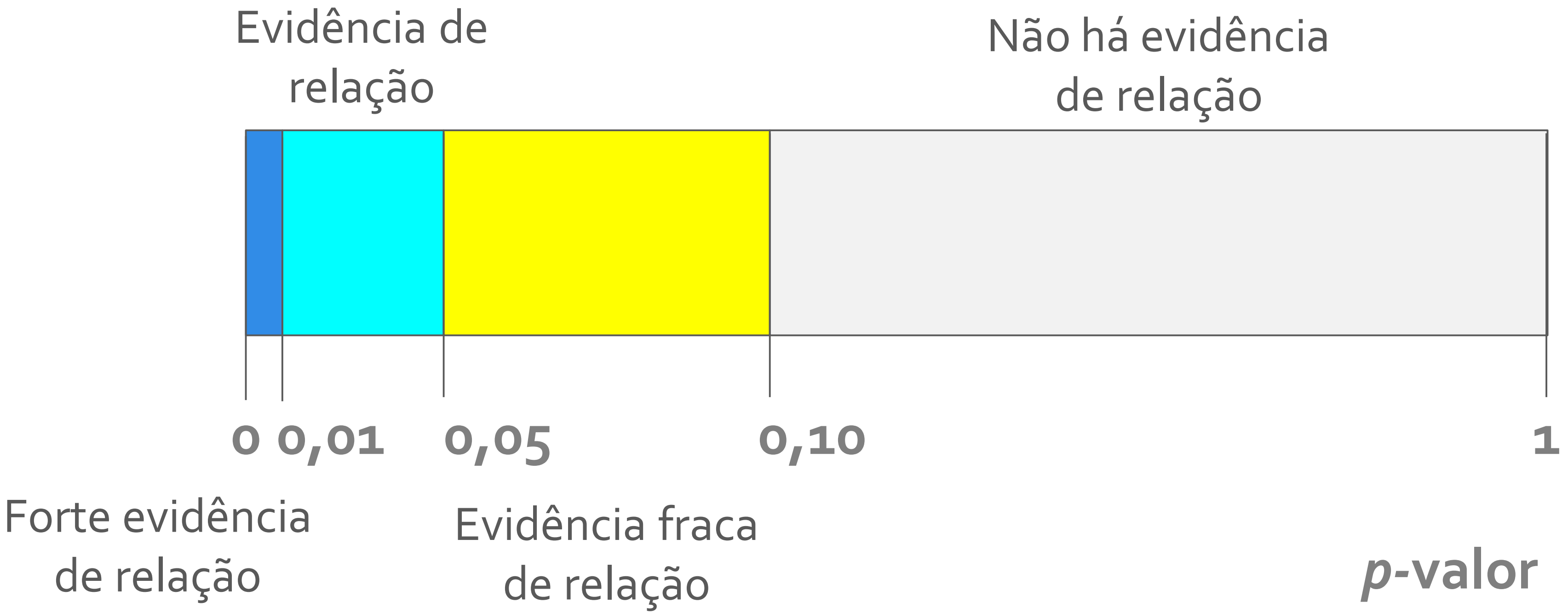
Teste de hipótese

H_0: não há relação entre as variáveis

H_a: existe relação entre as variáveis

$$H_0: b_n = 0$$

$$H_a: b_n \neq 0$$

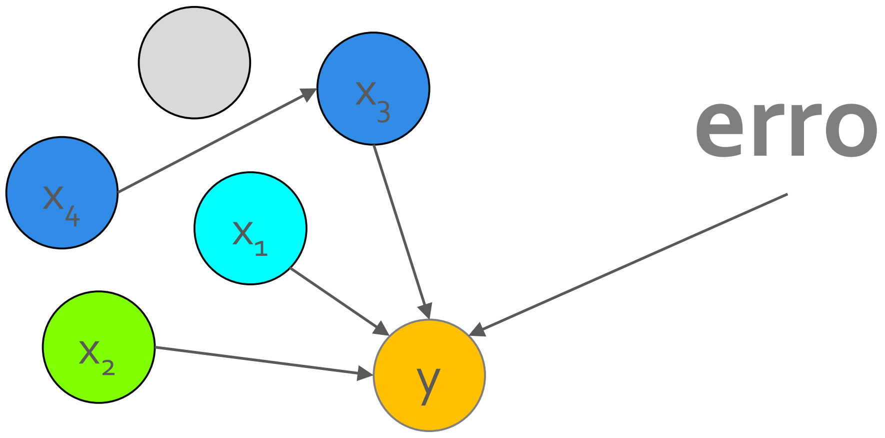


Prof. Dr. João Rafael Dias Pinto

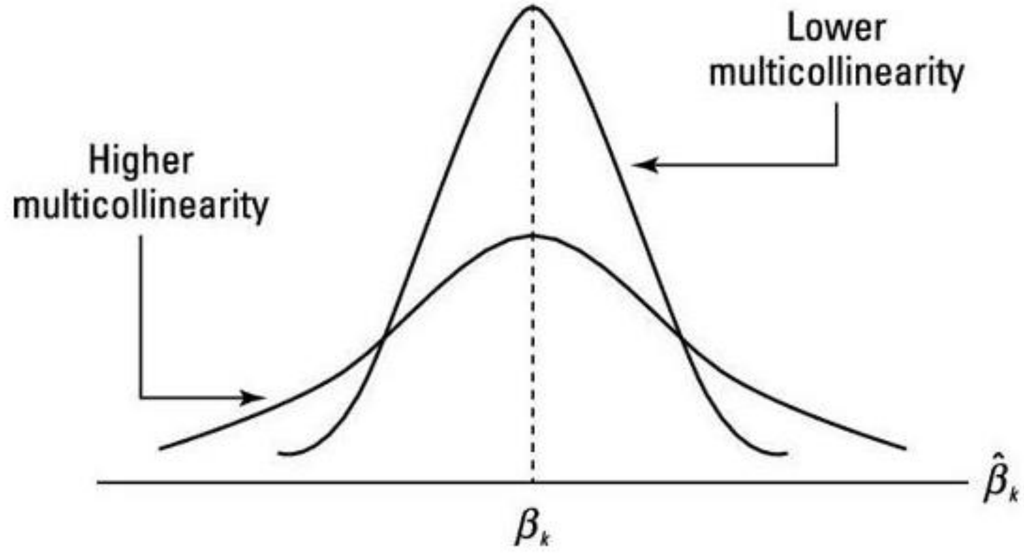
Consideração para as regressões

Multicolinearidade

- Variáveis independentes correlacionadas tendem a trazer problema de estimação dos parâmetros no modelo de regressão linear múltipla e logístico
- Quando existe correlação entre variáveis temos redundância na informação que elas carregam (uma consegue explicar outra(s))
- Para identificar multicolinearidade no modelo logístico podemos usar o VIF (*variance inflation factor*). Esse score mede quanto da variância de um coeficiente da regressão é inflada devido à multicolinearidade no modelo



$$y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4 + erro$$



https://medium.com/high-data-stories/the-concept-of-multicollinearity-359fccc9ae14

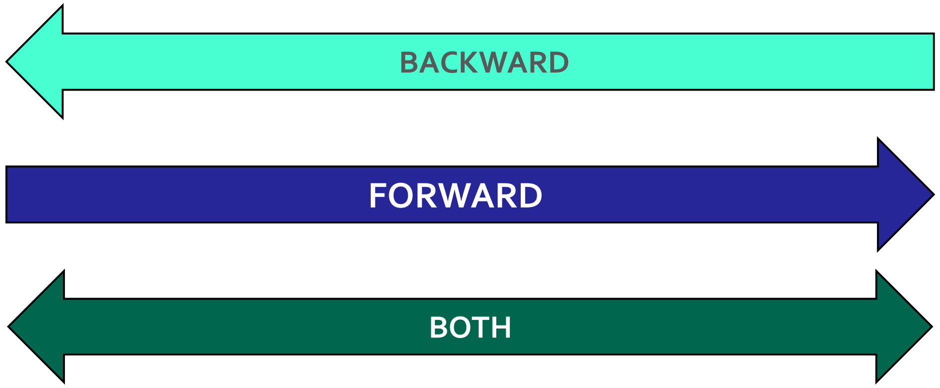
Consideração para as regressões

Seleção de variáveis

- Os coeficientes (b) de todas as variáveis mudam quando uma outra variável é adicionada ou removida do momento da aprendizagem do algoritmo
- Uma variável sozinha pode ser significante para explicar y, mas quando uma outra variável muito correlacionada com x entra no modelo, x pode tornar-se insignificante e instável
- Dessa forma deve-se remover uma variável por vez, para verificar se as outras variáveis mantêm-se relevantes ou não
- Usaremos aqui o processo denominado ***stepwise regression***

Descrição

Nesse processo, ao invés de se utilizar o *p*-valor, usaremos uma métrica denominada AIC (*Akaike Information Criterion*). Essa métrica é utilizada para comparar modelos e mede a perda relativa de informação por um determinado ajuste. Aqui quanto menor melhor.



A cada passo mantém apenas a variável que mais minimizou o **AIC**

$$Z = b_0$$

apenas o intercepto

$$Z = b_0 + b_1 x_1 + \cdots + b_p x_p$$

modelo *full*

Prática no RStudio

...foco de hoje

- **Treinando os algoritmos de regressão linear e logística sobre as bases de estudo**

  Criando as amostras de treino e teste. Ajuste dos algoritmos, aprimoramento dos resultados. Avaliações dos *outputs* dos modelos