Estatística Aplicada
EAD Ao Vivo

Tema da aula

Regressão Logística

LABDATA | FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO BUSINESS SCHOOL

04 -11/11/2020

APLICAÇÕES DE ESTATÍSTICA PARA TOMADA DE DECISÃO

Professora:
Drª Karin Ayumi Tamura

Coordenadores:
Profª Drª Alessandra de Álvila Montini
Profª Dr. Adolpho Walter Pimazoni Canton

Currículo - Prof.ª Drª Karin Ayumi Tamura

FORMAÇÃO ACADÊMICA | EXPERIÊNCIA PROFISSIONAL

Prof.ª Dra.
Karin Ayumi Tamura

Contato: karin.tamura@fia.com.br

- **FORMAÇÃO ACADÊMICA:** Pós-doutora (2015), Doutora (2012), mestre (2007) e bacharel (2003) em Estatística pelo Instituto de Matemática e Estatística da USP, tendo como área de pesquisa modelos de regressão, análise multivariada de dados e algoritmos de *machine learning*.
- **ATUAÇÃO PROFISSIONAL:** Foi *Head* de *Analytics* por 14 anos, e atualmente é Conselheira Executiva e *Head* de Inovação na *Marketdata Solutions*, uma empresa do grupo WPP, e Professora Doutora no LABDATA FIA.
- **HISTÓRICO:** Atuação no mercado por 17 anos, com experiência profissional no segmento bancário (Bradesco) e consultoria *(Marketdata Solutions)*. Atuou como docente em cursos de pós-graduação (2010-16) no LABDATA FIA e ABEMD. Especialista em Estatística e *Advanced Analytics* trabalhando em projetos de diversos segmentos do mercado. Participante de congressos nacionais e internacionais voltados a área de Estatística, Dados e Algoritmos de *Machine Learning*.

"Tenho duas paixões no meu trabalho: dados e pessoas. Voltar a lecionar no LABDATA FIA está sendo a realização de um sonho planejado desde a minha época de aluna de pós-graduação. Meu objetivo como professora é integrar a visão do mercado com as técnicas e tecnologias de análise de dados, por meio de uma atuação humanista no ensino aos alunos"

Projetos atendidos

FIA Business School

NOSSOS DIFERENCIAIS | QUEM SOMOS

BUSINESS SCHOOL
Graduação, pós-graduação, MBA, Pós- MBA, Mestrado Profissional, Curso In *Company* e EAD

CONSULTING
Consultoria personalizada que oferece soluções baseada em seu problema de negócio

RESEARCH
Atualização dos conhecimentos e do material didático oferecidos nas atividades de ensino

FUNDAÇÃO INSTITUTO DE ADMINISTRAÇÃO

Líder em Educação Executiva, referência de ensino nos cursos de graduação, pós-graduação e MBA, tendo excelência nos programas de educação. Uma das principais escolas de negócio do mundo, possuindo convênios internacionais com Universidades nos EUA, Europa e Ásia. +8.000 projetos de consultorias em organizações públicas e privadas.

- Único curso de graduação em administração a receber as notas máximas
- A primeira escola brasileira a ser finalista da maior competição de MBA do mundo
- Única *Business School* brasileira a figurar no *ranking* LATAM
- Signatária do Pacto Global da ONU
- Membro fundador da ANAMBA - Associação Nacional MBAs
- Credenciada pela AMBA - *Association of MBAs*
- Credenciada ao *Executive MBA Council*
- Filiada a AACSB - *Association to Advance Collegiate Schools of Business*
- Filiada a EFMD - *European Foundation for Management Development*
- Referência em cursos de MBA nas principais mídias de circulação

LABDATA FIA

O **Laboratório de Análise de Dados** – LABDATA é um Centro de Excelência que atua nas áreas de ensino, pesquisa e consultoria em análise de informação utilizando técnicas de ***Big Data***, ***Analytics*** e **Inteligência Artificial**.

Profª Drª Alessandra Montini

O LABDATA é um dos pioneiros no lançamento dos cursos de *Big Data* e *Analytics* no Brasil

Os diretores foram professores de grandes especialistas do mercado

+10 anos de atuação

+1000 alunos formados

Docentes

- Sólida formação acadêmica: doutores e mestres em sua maioria
- Larga experiência de mercado na resolução de *cases*
- Participação em Congressos Nacionais e Internacionais
- Professor assistente que acompanha o aluno durante todo o curso

Estrutura

- 100% das aulas realizadas em laboratórios
- Computadores para uso individual durante as aulas
- 5 laboratórios de alta qualidade (investimento +R$2MM)
- 2 Unidades próximas a estação de metrô (com estacionamento)

Conteúdo Programático do Curso

21 AULAS AO VIVO COM PROFA. KARIN | 27 PLANTÕES AO VIVO COM PROF. STEPHAN, 7 LISTAS DE EXERCÍCIOS E EAD VIDEO AULA EM PYTHON

Dia	Mês	Aula	EAD Ao Vivo	Plantão Prof. Stephan
5	Agosto	Introdução ao Curso e Análise Exploratória de Dados	Aula Prof. Karin	06/ago
12	Agosto	Análise Exploratória de Dados	Aula Prof. Karin	13/ago
19	Agosto	Análise Exploratória de Dados - Introdução ao R	Aula Prof. Karin	20/ago
26	Agosto	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	27/ago
2	Setembro	Regressão Linear Simples	Aula Prof. Karin	03/set
9	Setembro	Regressão Linear Simples e Múltipla	Aula Prof. Karin	10/set
16	Setembro	Regressão Linear Simples e Múltipla	Aula Prof. Karin	17/set
23	Setembro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	24/set
30	Setembro	Análise de Cluster	Aula Prof. Karin	01/out
7	Outubro	Análise de Cluster	Aula Prof. Karin	08/out
14	Outubro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	15/out
21	Outubro	Arvore de Decisão	Aula Prof. Karin	22/out
28	Outubro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	29/out
4	Novembro	Regressão Logística	Aula Prof. Karin	05/nov
11	Novembro	Regressão Logística	Aula Prof. Karin	11/nov
18	Novembro	Lista de Exercícios em Sala de Aula (19hs-23hs - com presença obrigatória)	-	19/nov
25	Novembro	estudo de caso	Aula Prof. Karin	26/nov
2	Novembro	estudo de caso	Aula Prof. Karin	30/dez
9	Dezembro	estudo de caso	Aula Prof. Karin	10/dez
16	Dezembro	Análise de Série Temporal - modelo auto regressivo	Aula Prof. Karin	17/dez
23	Dezembro	Lista de Exercícios em Sala de Aula (Frequência Liberada - véspera Natal)	-	-
Recesso Escolar		EAD - INTRODUÇÃO AO PYTHON EAD - INTRODUÇÃO AO PYTHON	EAD Video Aula (8 horas)	- -
6	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	07/jan
13	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	14/jan
20	Janeiro	Modelos estatísticos em Python	Aula Prof. Karin	20/jan
27	Janeiro	Introdução a Big Data - Aplicações de Machine Learning e Deep Learning	Aula Prof. Karin	28/jan
3	Fevereiro	Aplicações de Machine Learning	Aula Prof. Karin	04/fev
10	Fevereiro	Aplicações de Machine Learning	Aula Prof. Karin	11/fev
17	Fevereiro	Lista de Exercícios (Frequência Liberada - quarta de cinzas)	-	18/fev
24	Fevereiro	EXERCICIOS DE REVISÃO - EAD (19hs e 23hs - com presença obrigatória)	-	24/fev
3	Março	Prova (Plataforma On Line: 19hs e 23hs)	-	

Conteúdo da Aula

1. Introdução

2. Regressão Logística Simples
   i. Chance
   ii. Razão de Chances

3. Regressão Logística Múltipla
   i. Teste de hipótese sob os parâmetros
   ii. Processo de redução de variáveis
   iii. Multicolinearidade
   iv. Análise de desempenho (KS)

4. Exercícios para casa
   CASE: *Credit Score*

1. Introdução

Objetivo

Assim, como a Árvore de Decisão, o modelo de Regressão Logística tem o objetivo predizer um evento binário (1- evento de interesse e 0 – caso contrário), segundo as variáveis explicativas.

A diferença entre uma técnica e outra é que os algoritmos trabalham com regras ou lógicas diferentes.

Enquanto, a Árvore de Decisão trabalha com a partição da base de dados, segundo variáveis explicativas com maior relevância pelo Teste Qui-Quadrado (método CHAID), a Regressão Logística é uma equação matemática, assim como a Regressão Linear, com a diferença que a variável resposta agora é binária.

Aplicações

Apesar dos modelos de Regressão Logística e Árvore de Decisão trabalharem com o objetivo de predizer um evento binário, alguns segmentos de mercado utilizam a Regressão Logística com maior frequência devido à sua facilidade de interpretação, implementação e manutenção.

Exemplo de segmentos de negócios que utilizam tradicionalmente a Regressão Logística:

- Bancário: modelos de *credit score*, *behavior score* e *collection score*.
- Telecom: modelos de *churn*, *cross-sell*, migração e *up-sell*.
- Área Médica: modelos de propensão aos eventos de saúde (ataque cardíaco, desnutrição, óbito, câncer, etc.).

Case *Credit Score* em Banco

Exemplo
Identificar a probabilidade de uma pessoa que ainda não é cliente da instituição financeira se tornar inadimplente ao adquirir um crédito pessoal.

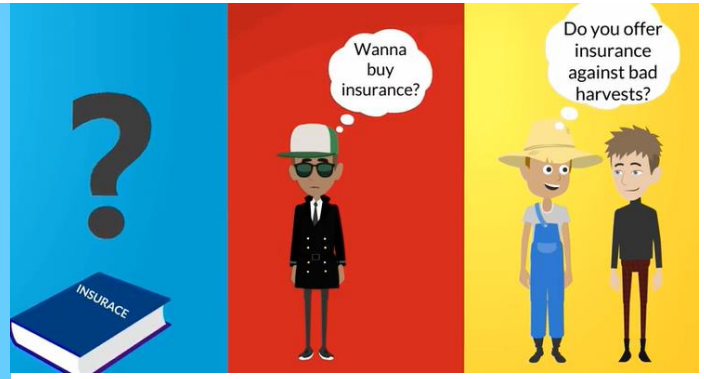
Aplicação
Segmento bancário.

https://www.youtube.com/watch?v=dwIGfhhgKOc&feature=youtu.be

Case Seguradora

Exemplo
Identificar a probabilidade de um indivíduo sofrer um sinistro com base em seu estilo de vida.

Aplicação
Segmento seguradoras.



https://www.youtube.com/watch?v=qYSdjUPCwwY&feature=youtu.be

Case Migração de Plano em Telecom

Exemplo
Identificar os clientes com maior propensão a migração de um plano controle para um plano pós-pago.

Aplicação
Segmento Telecom.

Case Doenças Cardíacas

Exemplo
Identificar a probabilidade de um paciente ter problemas coronários de acordo com seu hábito de vida: quantidade de horas de sono, quantidade de refeições diárias, frequência de consumo de frituras, frequência de consumo de doces, frequência de exercícios físicos, valor de colesterol total, valor de triglicérides, etc.

Aplicação
Área Médica.

Objetivo de Regressão Logística (Binária)

O modelo de Regressão Logística é amplamente utilizado no mercado e tem fácil aceitação pelas áreas de negócios devido a sua facilidade de entendimento.

Ele é muito usado quando se deseja obter um peso associado a cada variável explicativa para se obter uma probabilidade final do evento de interesse.

A função f(X) assume valores entre 0 e 1 e a variável X é a variável explicativa dada por:

$$f(X) = \frac{e^{X}}{1 + e^{X}},$$

→ Variável explicativa (ou covariável, ou variável preditora, ou variável independente).

em que **e** (ou exponencial) na função f(X) significa a base do logaritmo neperiano, ou seja, vale aproximadamente ~**2,7182**.

Exemplo: Regressão Logística (Binária)

Considere o case de *Credit Score*.

Um banco deseja identificar a probabilidade de um indivíduo, que ainda não é cliente da instituição financeira, pagar o valor de empréstimo de crédito pessoal no próximo ano. Seja Y a variável resposta:

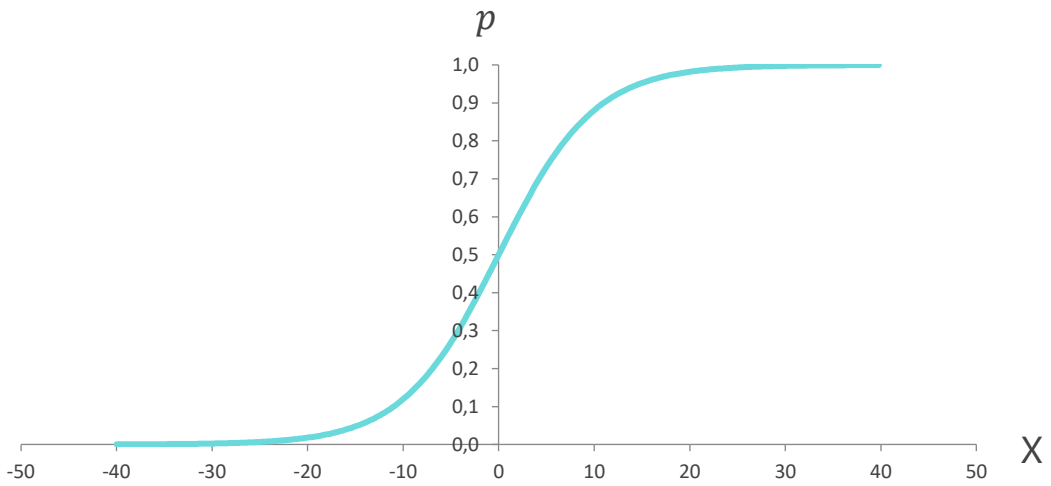
1 – cliente ficou inadimplente
0 – não ficou inadimplente

A **Função Logística** f(X) pode ser escrita como P(Y=1), sendo Y a variável resposta binária, dada por:

$$P(Y = 1) = f(X) = \frac{e^X}{1 + e^X} \quad .$$

Visualização gráfica do modelo

Função em formato S



$$p = P(Y = 1) = \frac{e^X}{1 + e^X}$$

Exercício: Regressão Logística (Binária)

Continuação do case *Credit Score*. Seja Y a variável resposta:

1 – cliente ficou inadimplente
0 – não ficou inadimplente

e a variável X assumindo dois valores: 1,7 para clientes que possuem empréstimo no mercado e 0 para aqueles que não possuem empréstimo no mercado.

A **Função Logística** pode ser escrita como P(Y=1), que é uma função que assume valores entre 0 e 1.

$$P(Y = 1) = \frac{e^X}{1+e^X} \ ,$$

em que **e** (ou exponencial) na função f(X) significa a base do logaritmo neperiano, ou seja, **e** vale aproximadamente ~**2,7182**.

Qual a probabilidade de inadimplência para os indivíduos que possuem empréstimo no mercado (em outras instituições)?

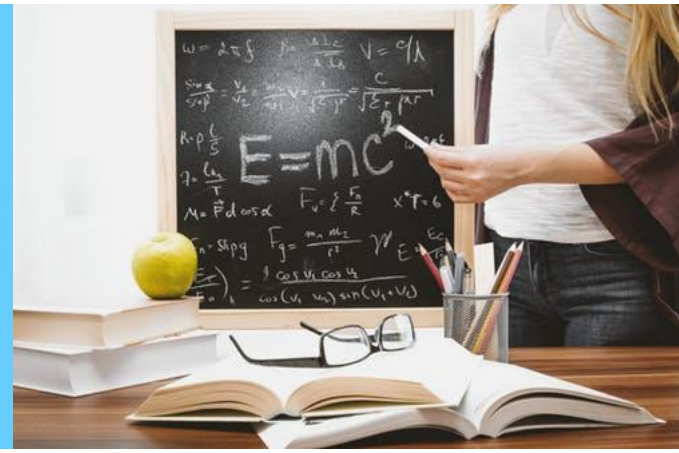
Qual a probabilidade de inadimplência para os indivíduos que não possuem empréstimo no mercado?

Regressão Logística (Binária)

Dado que entendemos o objetivo e como a Regressão Logística funciona, vamos aprender um pouco mais em detalhes sobre a função do modelo, seleção de covariáveis, interpretação de seus parâmetros e seu desempenho.

O tipo de variável resposta a ser estudada será a resposta binária (1 – evento de interesse e 0 – caso contrário) pela distribuição Binomial, apesar do método permitir o uso da resposta com múltiplas categorias (distribuição Mutinomial).

Este modelo pertence à classe dos modelos lineares generalizados (MLG), e é uma extensão do modelo de Regressão Linear.



2. Regressão Logística Simples

Exemplo

Considere o case de Fraude na transação de cartão de crédito.

A variável de interesse Y (fraude) é uma variável aleatória assumindo o valor Y=0 ou o valor Y=1.

Podemos considerar Y=0 para não fraude e Y=1 para fraude.

Y = 0

Y = 1

Definição da resposta

Considere o case de Fraude na transação de cartão de crédito.

A variável de interesse Y (fraude) é uma variável aleatória assumindo o valor Y=0 ou o valor Y=1.

Podemos considerar Y=0 para não fraude e Y=1 para fraude.

A regressão logística aloca uma nova observação em um dentre dois grupos por meio do cálculo de uma probabilidade p, que é dado por P(Y=1).

Uma probabilidade é um valor entre 0 e 1:

$$0 \leq p \leq 1$$

Variável Explicativa

Considere o case de Fraude na transação de cartão de crédito.

A variável de interesse Y (fraude) é uma variável aleatória assumindo o valor Y=0 ou o valor Y=1.

Podemos considerar Y=0 para não fraude e Y=1 para fraude.

O cálculo desta probabilidade é feito utilizando variáveis explicativas (X). No exemplo, podemos determinar a probabilidade de fraude dependendo do tempo de banco do funcionário.

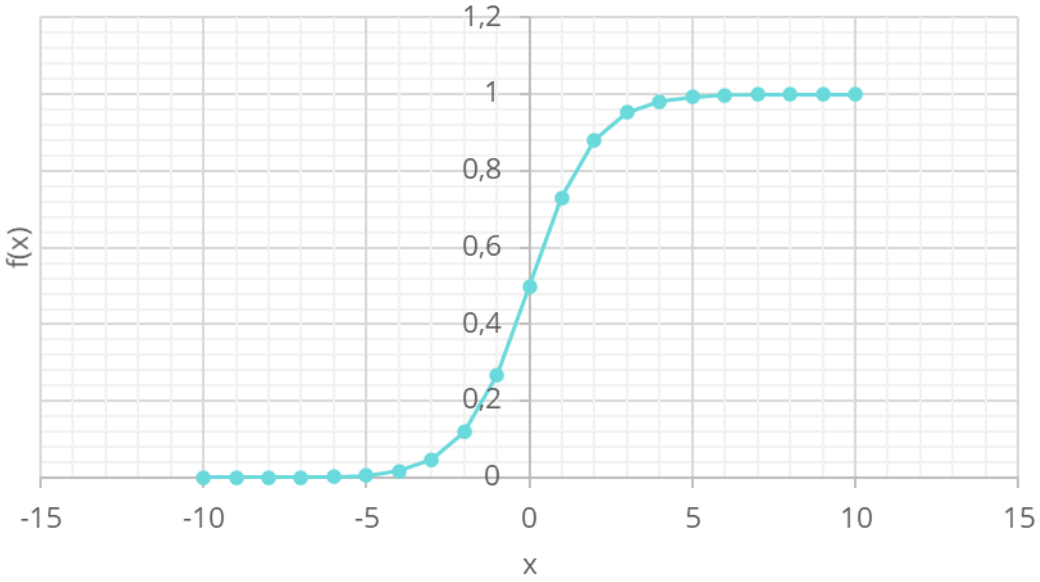
A variável X pode assumir qualquer valor. Pode-se dizer que X pode variar do menos infinito ao mais infinito.

$$-\infty \leq X \leq +\infty$$

Função Logística

A **Função Logística** - f(x) - é uma função que assume valores entre 0 e 1 e a variável X pode assumir qualquer valor.

O gráfico representa a **Função Logística**. Os valores de X estão variando entre -10 e 10 e os valores de f(X) variando entre 0 e 1.



A **Função Logística é dada por:**

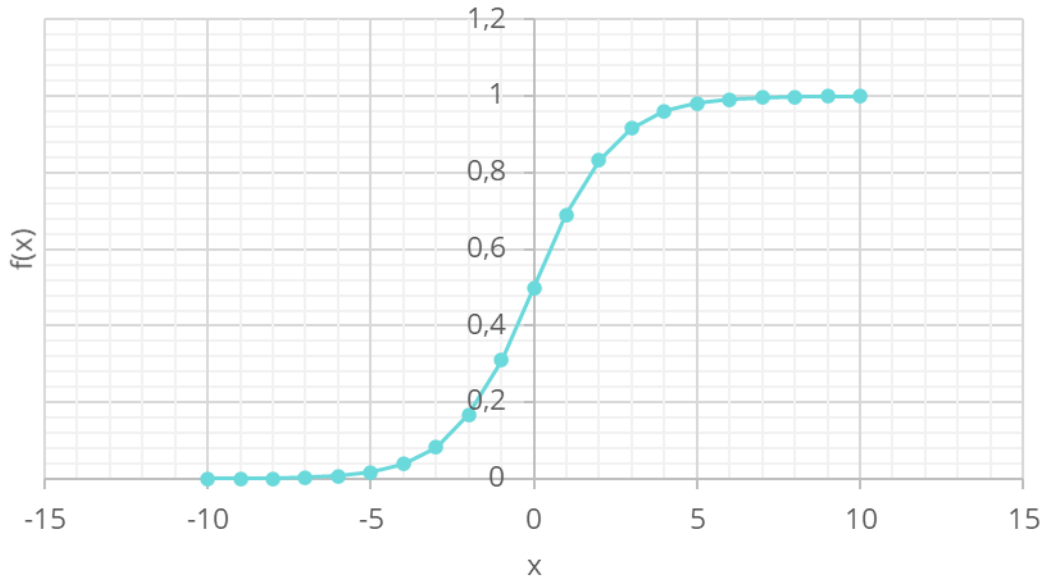
$$f(X) = \frac{e^X}{1 + e^X}$$

Sendo que o *e* na função significa a base do logaritmo neperiano, ou seja, aproximadamente 2,7182.

Função Logística

Quando o coeficiente da função logística β é positivo a probabilidade p cresce a medida que aumenta o valor de X.

A figura apresenta a função logística com β=0,8.

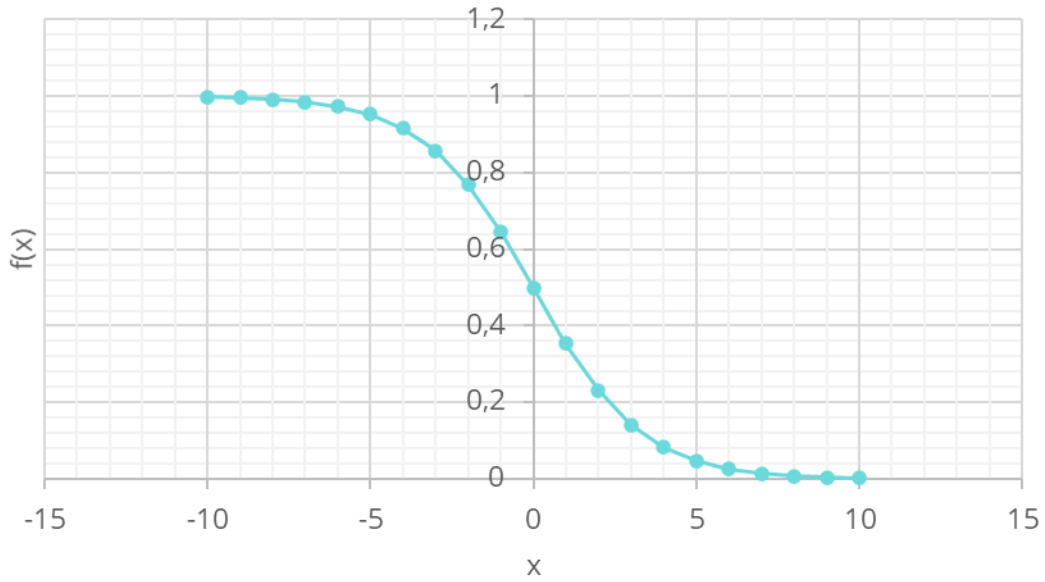


$$f(X) = \frac{e^{0,8*X}}{1 + e^{0,8*X}}$$

Função Logística

Quando o coeficiente da função logística β é negativo a probabilidade p decresce a medida que aumenta o valor de X.

A figura apresenta a função logística com β=-0,6.



$$f(X) = \frac{e^{-0,6*X}}{1 + e^{-0,6*X}}$$

Modelo de Regressão Logística

No exemplo da fraude, a probabilidade de fraude (p) pode ser obtida considerando a variável

X= tempo de relacionamento.

Considerando que a função logística pode ser utilizada para obter a probabilidade de fraude (p), tem-se que:

$$p = P(Y = 1) = f(X_1) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

Modelo de Regressão Logística

No exemplo da fraude, a probabilidade de fraude (p) pode ser obtida considerando a variável

X= tempo de relacionamento.

A probabilidade da operação ser uma fraude (p) também pode ser escrita como:

$$p = P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}$$

Regressão Logística Simples

Seja a variável resposta Y binária, assumindo os valores 1 – evento de interesse e 0 – caso contrário.
O modelo de Regressão Logística Simples (com uma covariável X) é dado por:

$$p = P(Y = 1) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

O mesmo modelo pode ser escrito de diversas formas

$$p = P(Y = 1) = \frac{1}{1+e^{-(\alpha+\beta X)}}$$

$$\frac{p}{1-p} = e^{\alpha+\beta X}$$

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$$

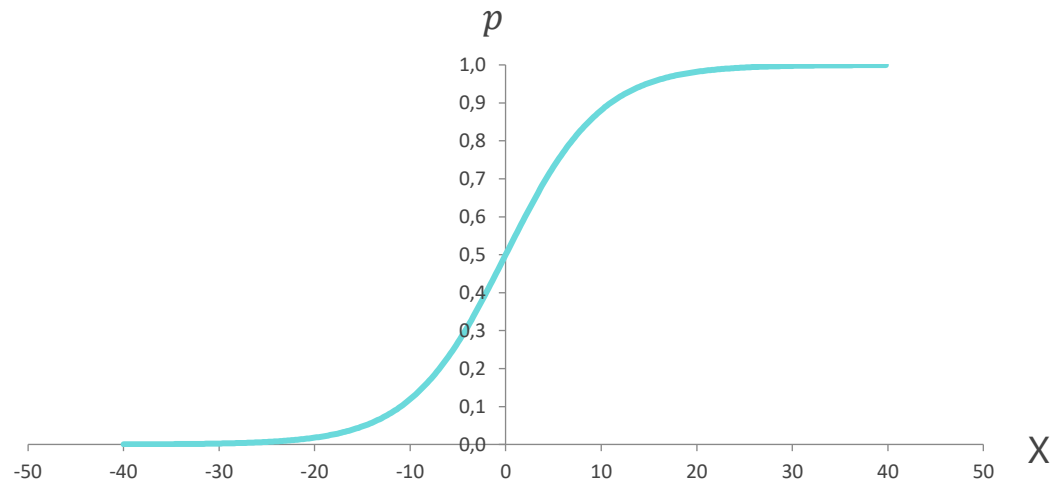
Em que :

- ✓ α e β são chamados **parâmetros do modelo**.
- ✓ **e** vale aproximadamente ~**2,7182**.

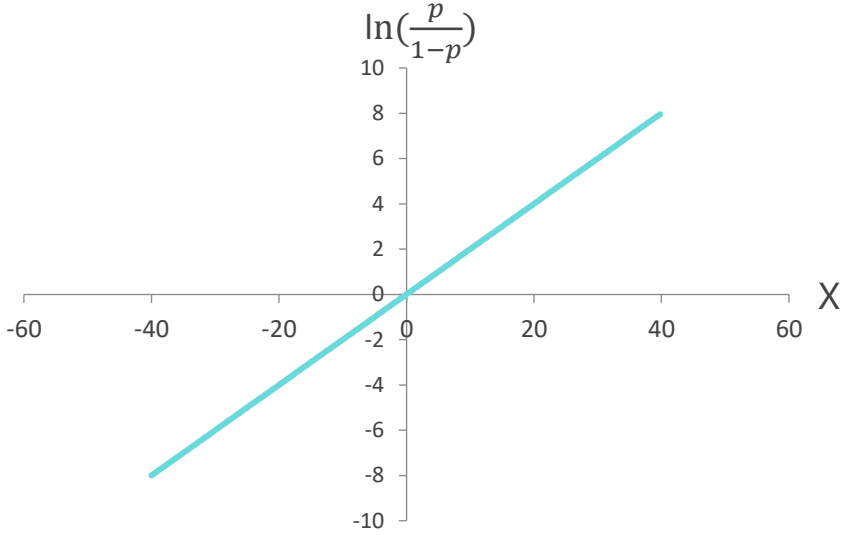
Visualização gráfica do modelo

Assumindo α =0, β =1 e X no intervalo do reais.

Função em formato S



Função Linear



$$p = P(Y = 1) = \frac{e^X}{1 + e^X}$$

Pode ser escrito →

$$Ln(\frac{p}{1-p}) = \textbf{logit (p)} = X$$

Função LOGIT

Chance (*Ratio*)

A chance é definida por: $\text{chance} = \frac{p}{1-p}$

Pode ser interpretada como a razão da probabilidade do evento de interesse pela probabilidade do evento complementar.

A relação entre a chance e o modelo de Regressão Logística é dada por:

$$\frac{p}{1-p} = e^{\alpha+\beta X}$$

Suponha que a variável X, sendo variável *dummy,* a presença do fator é dada como X =1 e a não presença do fator como X=0.

Presença do fator X=1: $\frac{p}{1-p} = e^{\alpha+\beta}$

Ausência do fator X=0: $\frac{p}{1-p} = e^{\alpha}$

Interpretação

Seja a variável original GÊNERO.
Criamos uma variável *dummy* (codificada):
X =1, se gênero FEMININO
X = 0, se gênero MASCULINO

Se gênero for FEMININO: $\frac{p}{1-p} = e^{\alpha+\beta}$

Se gênero for MASCULINO: $\frac{p}{1-p} = e^{\alpha}$

Assumindo α =0,5, β=1 e p a probabilidade de compra de um produto:

- A chance de uma mulher comprar o produto é 4,48 em relação a não comprar o produto (valor $e^{1,5}$=4,48) .
- A chance de um homem comprar o produto é 1,64 em relação a não comprar o produto (valor $e^{0,5}$=1,64).

Razão de Chances (*Odds Ratio*)

A razão de chances é definida por $\text{RC} = \frac{\text{chance}\ (\text{X} = 1)}{\text{chance}\ (\text{X} = 0)}$

Pode ser interpretada como a chance do evento resposta ocorrer no perfil X=1 em relação ao perfil X=0.

Na prática, podemos utilizar o valor absoluto da estimativa do parâmetro para interpretar a contribuição das categorias da variável; quanto maior o valor, maior a probabilidade do evento de interesse ocorrer.

No exemplo ao lado, a mulher tem maior probabilidade de comprar o produto em relação ao homem, pois o valor β associado ao seu perfil (X =1, equivale a β=1 com p=0,82) é maior do que o valor do coeficiente para os homens (X =0, equivale a β=0 com p=0,62).

Interpretação

Seja a variável original GÊNERO.
Criamos uma variável *dummy* (codificada):
X =1, se gênero FEMININO
X = 0, se gênero MASCULINO

Se gênero for FEMININO: $\frac{\text{p}}{1-\text{p}} = \text{e}^{\alpha+\beta}$

Se gênero for MASCULINO: $\frac{\text{p}}{1-\text{p}} = \text{e}^{\alpha}$

Assumindo α =0,5, β =1 e p a probabilidade de compra de um produto, a razão de chances pode ser escrita como:

$$\text{RC} = \text{e}^{\beta} = 2{,}72$$

A chance de uma mulher comprar o produto é 2,72 em relação ao homem de comprar o produto.

Tipos de covariáveis

O modelo de Regressão Logística estima os parâmetros tanto para covariáveis qualitativas como quantitativas.

Entretanto, a interpretação dos parâmetros do modelo se torna mais difícil por depender do conceito de chance e a razão de chances. Quando as covariáveis são quantitativas é usual utilizar a variável categorizada.

Desta forma, uma boa prática do mercado, é categorizar as variáveis quantitativas e entrar com todas as variáveis (quantitativas e qualitativas) como sendo *DUMMIES* para a estimação dos parâmetros.

$$\frac{p}{1-p} = e^{\alpha+\beta X}$$

3. Regressão Logística Múltipla

Modelo de Regressão Logística Múltipla

Seja a variável resposta Y binária assumindo os valores 1 – evento de interesse e 0 – caso contrário.
O modelo de Regressão Logística Múltipla (com a covariáveis X_1, X_2,..., X_p), é dado por:

$$p = P(Y = 1) = \frac{e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}{1 + e^{\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}}$$

Em que :

- ✓ β_0 , β_1 até β_p são chamados **parâmetros do modelo**.
- ✓ **e** vale aproximadamente ~**2,7182.**

Todas as equivalências vistas na Regressão Logística Simples (slide 29) podem ser estendidas de forma análoga para o caso multidimensional.

Case: Aumento do time de vendas

Com objetivo de aumentar seu time de vendas, uma indústria de cosméticos, por meio de vendas direta, quer identificar qual o perfil das pessoas que poderiam tornar-se consultores de seus produtos.

Seja Y o evento que tornou-se consultor (1 – SIM e 0 – NÃO) e as variáveis explicativas são:

Sexo: M (masculino) e F (feminino)

Idade: idade do cliente

Cidade: RJ ou SP

As primeiras linhas da base de dados são apresentada na tabela. Na coluna sinistro tem-se que **resposta = 1 para pessoas que se tornaram consultores** e **resposta = 0 para pessoas que não se tornaram consultores.**

Sexo	Idade	Cidade	Resposta
M	49	SP	0
M	32	SP	0
F	48	SP	0
F	32	RJ	0
F	64	SP	0
F	56	SP	0
F	69	SP	0

Arquivo Cosmeticos.xlsx

Case: Aumento do time de vendas

Com objetivo de aumentar seu time de vendas, uma indústria de cosméticos, por meio de vendas direta, quer identificar qual o perfil das pessoas que poderiam tornar-se consultores de seus produtos.

Seja Y o evento que tornou-se consultor (1 – SIM e 0 – NÃO) e as covariáveis:

Sexo: M (masculino) e F (feminino)

Idade: idade do cliente (<40 anos e >=40 anos)

Cidade: Capitais e Interior

A probabilidade da pessoa tornar-se consultor de cosméticos, p=P(Y=1) , pode ser escrita por meio do modelo:

$$p = \frac{e^{\beta_0 + \beta_1 sexo + \beta_2 idade + \beta_3 cidade}}{1 + e^{\beta_0 + \beta_1 sexo + \beta_2 idade + \beta_3 cidade}}$$

Teste de hipótese sob os parâmetros do modelo

Seja a hipótese nula (H_0) com os valores $\beta_i = 0$ e i=1, 2 e 3. Deseja-se rejeitar H_0, ao nível de significância 10%, com

H_1 : caso a variável sexo seja importante, o parâmetro β_1 será diferente de zero ($\beta_1 \neq 0$)

H_1 : caso a variável idade seja importante, o parâmetro β_2 será diferente de zero ($\beta_2 \neq 0$)

H_1 : caso a variável cidade seja importante, o parâmetro β_3 será diferente de zero ($\beta_3 \neq 0$)

A probabilidade da pessoa tornar-se consultor de cosméticos, p=P(Y=1) , pode ser escrita por meio do modelo:

$$p = \frac{e^{\beta_0+\beta_1 sexo+\beta_2 idade+\beta_3 cidade}}{1+e^{\beta_0+\beta_1 sexo+\beta_2 idade+\beta_3 cidade}}$$

Teste de hipótese sob os parâmetros do modelo

Para verificar se a **variável sexo** deve fazer parte do modelo deve-se testar a hipótese:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

```
Call:
glm(formula = Resposta ~ Sexo + Idade + Cidade, family = binomial(link = "logit"),
    data = cosmeticos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96839  -0.25510   0.02773   0.31641   2.39982

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
(Intercept)   8.0459     1.5376   5.233 0.00000016695 ***
SexoM         1.5793     0.5497   2.873       0.00407 **
Idade        -0.1958     0.0395  -4.956 0.00000072044 ***
CidadeSP     -3.2332     0.5477  -5.903 0.00000000357 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como o nível descritivo (**0,0047**) < **0,10** rejeita-se a hipótese H_0, evidenciando que $\beta_1 \neq 0$, ou seja, a variável sexo deve fazer parte do modelo.

Teste de hipótese sob os parâmetros do modelo

Para verificar se a **variável idade** deve fazer parte do modelo deve-se testar a hipótese:

$$H_0: \beta_2 = 0$$
$$H_1: \beta_2 \neq 0$$

```
Call:
glm(formula = Resposta ~ Sexo + Idade + Cidade, family = binomial(link = "logit"),
    data = cosmeticos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96839  -0.25510   0.02773   0.31641   2.39982

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
(Intercept)   8.0459     1.5376   5.233 0.00000016695 ***
SexoM         1.5793     0.5497   2.873       0.00407 **
Idade        -0.1958     0.0395  -4.956 0.00000072044 ***
CidadeSP     -3.2332     0.5477  -5.903 0.00000000357 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como o nível descritivo (**0,00000072**) < **0,10** rejeita-se a hipótese H_0, evidenciando que $\beta_2 \neq 0$, ou seja, a variável idade deve fazer parte do modelo.

Teste de hipótese sob os parâmetros do modelo

Para verificar se a **variável cidade** deve fazer parte do modelo deve-se testar a hipótese:

$$H_0: \beta_3 = 0$$
$$H_1: \beta_3 \neq 0$$

```
Call:
glm(formula = Resposta ~ Sexo + Idade + Cidade, family = binomial(link = "logit"),
    data = cosmeticos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96839  -0.25510   0.02773   0.31641   2.39982

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
(Intercept)   8.0459     1.5376   5.233 0.00000016695 ***
SexoM         1.5793     0.5497   2.873       0.00407 **
Idade        -0.1958     0.0395  -4.956 0.00000072044 ***
CidadeSP     -3.2332     0.5477  -5.903 0.00000000357 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Como o nível descritivo (**0,00000000357**) < **0,10** rejeita-se a hipótese H_0, evidenciando que $\beta_3 \neq 0$, ou seja, a variável cidade deve fazer parte do modelo.

Modelo com várias covariáveis

Output do modelo no R

```
Call:
glm(formula = Resposta ~ Sexo + Idade + Cidade, family = binomial(link = "logit"),
    data = cosmeticos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.96839  -0.25510   0.02773   0.31641   2.39982

Coefficients:
            Estimate Std. Error z value      Pr(>|z|)
(Intercept)   8.0459     1.5376   5.233 0.00000016695 ***
SexoM         1.5793     0.5497   2.873       0.00407 **
Idade        -0.1958     0.0395  -4.956 0.00000072044 ***
CidadeSP     -3.2332     0.5477  -5.903 0.00000000357 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Coeficientes estimados

Sexo = “M” – O peso positivo indica que se o profissional for do sexo masculino, a probabilidade de virar um consultor aumenta.

Idade – O peso negativo indica que, à medida que a idade aumenta, a probabilidade de virar um consultor diminui.

Cidade = “SP” – O peso negativo indica que se o profissional for da cidade de “São Paulo”, a probabilidade de virar um consultor diminui.

Teste de hipótese sob os parâmetros do modelo

Dado que todas as variáveis são importantes pode-se obter a equação para a probabilidade de a pessoa se tornar consultor.

$$p = \frac{e^{8{,}0459+1{,}5793*SexoM-0{,}1958*Idade-3{,}2332*CidadeSP}}{1+e^{8{,}0459+1{,}5793*SexoM-0{,}1958*Idade-3{,}2332*CidadeSP}}$$

A probabilidade da primeira pessoa da base de dados se tornar consultora é dada por:

$$p = \frac{e^{8,0459+1,5793*1-0,1958*49-3,2332*1}}{1+e^{8,0459+1,5793*1-0,1958*49-3,2332*1}} = 0,0391$$

Sexo	Idade	Cidade
M	49	SP

Exercício: Modelo com várias covariáveis

Com o *output* do modelo fornecido abaixo, calcule a probabilidade de um indivíduo tornar-se consultor da indústria de cosméticos, sendo:

(a) Sexo Feminino, com mais de 40 anos e do interior.

(b) Sexo Masculino, com menos de 40 anos e de capitais.

```
> #Modelo de Regressão Logística: family = binomial
> modelo_completo <- glm(Resposta ~ Sexo + Idade + Cidade, data = comesticos, family = "binomial")

> summary(modelo_completo)

Call:
glm(formula = Resposta ~ Sexo + Idade + Cidade, family = "binomial",
    data = comesticos)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-3.07948  -0.08476  -0.08476   0.13239   1.93312

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)     1.4233     0.6829   2.084   0.0371 *
SexoM          -3.1241     0.7250  -4.309 1.64e-05 ***
IdadeMenor_40  -3.9264     0.8014  -4.900 9.60e-07 ***
CidadeInterior  3.3096     0.7072   4.680 2.87e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Por meio de uma equação matemática, é possível selecionar as variáveis explicativas, fornecendo como resultado a probabilidade do indivíduo apresentar o evento de interesse.

Case: *Credit Score*

BANCO DE DADOS EM .TXT | EXERCÍCIO EM SALA

Uma *fintech* está preocupada com os clientes novos que entram em sua carteira e que apresentam *'default'* (não pagamento do empréstimo) após um certo período. Esta análise trata da aprovação de Crédito Pessoal para novos clientes. O objetivo é fazer a aprovação de crédito de maneira automática por meio de um algoritmo, deixando apenas alguns casos duvidosos para a análise de crédito manual. Para a aprovação ou não do cliente na instituição é utilizado o modelo de *Credit Score* que fornece a probabilidade do cliente apresentar *'default'* por meio das variáveis cadastrais e informações de restritivos de mercado que a *fintech* consulta no momento da análise de crédito.

Fonte: base simulada, inspirada em problemas reais de consultoria da Prof.ª Karin Ayumi Tamura

(a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados na visão do negócio.
(b) Faça uma análise do % de *default*.
(c) Faça a análise bivariada das variáveis explicativas (covariáveis) vs. a variável resposta. Quais variáveis discriminam o evento resposta? Como você poderia tratar as categorias com *missings value* na análise bivariada?
(d) Rode o modelo de Regressão Logística. Selecione um modelo final no qual a interpretação dos parâmetros esteja de acordo com a análise bivariada.
(e) Faça a análise de multicolinearidade entre as covariáveis. Reajuste o modelo caso seja necessário, garantindo que as estimativas dos parâmetros fiquem condizentes com a análise exploratória bivariada.
(f) Qual o perfil, a probabilidade e a representatividade do cliente mais propenso a apresentar o *default*? E do menos propenso?
(g) Analise a sensibilidade, especificidade e acurácia pela tabela de classificação.
(h) Como você classifica o desempenho do modelo?
(i) Pelo valor do KS, você indicaria para área de negócios utilizar o modelo?
(j) Qual o percentual da base de dados que seria rejeitado?

Case: *Credit Score*

Análise das covariáveis

```
          FX_IDADE           CEP_GRUPO_RISCO                FX_RENDA
(1) <=24       : 4941   (1) baissiximo: 4595   (1) nao informado : 4401
(2) 25 a 34    :14279   (2) baixo     :13387   (2) <1500          : 2606
(3) 35 a 39    : 4450   (3) medio     :14330   (3) De 1500 a 2500:17359
(4) 40 a 45    : 4162   (4) alto      : 7544   (4) De 2501 a 5000:13913
(5) >=46       :16575   (5) altissimo : 9962   (5) 5001 a 7000   : 8615
sem informacao: 5411                           (6) >7000         : 2924
INDICADOR_RESTRITIVO QTDE_CONSULTAS_CREDITO    RESPOSTA
Min.   :0.00000      0         :28648         Min.   :0.0000
1st Qu.:0.00000      1         :12991         1st Qu.:0.0000
Median :0.00000      2         : 4790         Median :0.0000
Mean   :0.02082      3 ou mais: 3389          Mean   :0.1207
3rd Qu.:0.00000                               3rd Qu.:0.0000
Max.   :1.00000                               Max.   :1.0000
```

Como tratamos os *missing values*?

Análise da variável resposta

Da base de **49.818** clientes, **12,07%** apresentaram *default*.

Case: *Credit Score*

Faixa de Idade (anos)	Grupo CEP (agrupado por risco)	Faixa de Renda (Reais)	Flag de Restritivo	Quantidade de consultas de Crédito

	0	1
(1) <=24	0.81	0.19
(2) 25 a 34	0.88	0.12
(3) 35 a 39	0.90	0.10
(4) 40 a 45	0.90	0.10
(5) >=46	0.92	0.08
sem informacao	0.77	0.23

	0	1
(1) baissiximo	0.95	0.05
(2) baixo	0.91	0.09
(3) medio	0.88	0.12
(4) alto	0.87	0.13
(5) altissimo	0.80	0.20

	0	1
(1) nao informado	0.77	0.23
(2) <1500	0.85	0.15
(3) De 1500 a 2500	0.86	0.14
(4) De 2501 a 5000	0.89	0.11
(5) 5001 a 7000	0.93	0.07
(6) >7000	0.95	0.05

	0	1
0	0.88	0.12
1	0.85	0.15

	0	1
0	0.89	0.11
1	0.89	0.11
2	0.85	0.15
3 ou mais	0.78	0.22

- Pela **análise bivariada** das covariáveis *versus* a resposta, descritivamente, todas as variáveis parecem ter relação com o evento resposta.

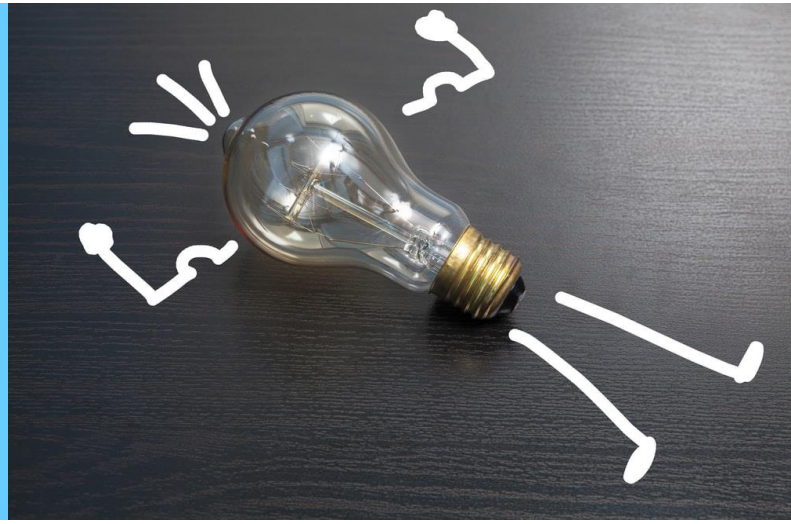
Case: *Credit Score*


```
Call:
glm(formula = RESPOSTA ~ FX_IDADE + CEP_GRUPO_RISCO + FX_RENDA +
    INDICADOR_RESTRITIVO + QTDE_CONSULTAS_CREDITO, family = binomial(link = "logit"),
    data = credit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4533  -0.5399  -0.4330  -0.3305   2.7307

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.49603    0.11498 -21.709  < 2e-16 ***
FX_IDADE(2) 25 a 34           -0.54122    0.04539 -11.925  < 2e-16 ***
FX_IDADE(3) 35 a 39           -0.68764    0.06226 -11.046  < 2e-16 ***
FX_IDADE(4) 40 a 45           -0.72516    0.06463 -11.219  < 2e-16 ***
FX_IDADE(5) >=46              -0.89433    0.04746 -18.843  < 2e-16 ***
FX_IDADEsem informacao         0.24865    0.07050   3.527  0.00042 ***
CEP_GRUPO_RISCO(2) baixo       0.33977    0.08008   4.243 2.21e-05 ***
CEP_GRUPO_RISCO(3) medio       0.57981    0.08009   7.239 4.51e-13 ***
CEP_GRUPO_RISCO(4) alto        0.69835    0.08353   8.361  < 2e-16 ***
CEP_GRUPO_RISCO(5) altissimo   0.97472    0.08504  11.461  < 2e-16 ***
FX_RENDA(2) <1500              0.39437    0.09308   4.237 2.27e-05 ***
FX_RENDA(3) De 1500 a 2500     0.37754    0.07735   4.881 1.06e-06 ***
FX_RENDA(4) De 2501 a 5000     0.17037    0.08074   2.110  0.03485 *
FX_RENDA(5) 5001 a 7000       -0.14448    0.09085  -1.590  0.11175
FX_RENDA(6) >7000             -0.31380    0.12022  -2.610  0.00905 **
INDICADOR_RESTRITIVO           0.50085    0.09158   5.469 4.53e-08 ***
QTDE_CONSULTAS_CREDITO1        0.20270    0.03447   5.881 4.08e-09 ***
QTDE_CONSULTAS_CREDITO2        0.51581    0.04645  11.104  < 2e-16 ***
QTDE_CONSULTAS_CREDITO3 ou mais  1.02277    0.04728  21.632  < 2e-16 ***
```

Processo de redução de variáveis

De forma análoga ao modelo de Regressão Linear, podemos realizar a redução das covariáveis do modelo passo-a-passo, eliminando a cada passo, as covariáveis com maiores p-valor.

Até que todos os parâmetros das variáveis sejam significantes.



Investigação da Multicolinearidade e Estatística V de *Cramer*

De forma análoga ao modelo de Regressão Linear, podemos fazer a análise da correlação entre as covariáveis do modelo. Como agora a associação será entre duas variáveis qualitativas, utilizaremos a **Estatística V de *Cramer***.

A **Estatística V de *Cramer*** mede a associação entre duas variáveis qualitativas em uma tabela de contingência (tabela de 2 variáveis qualitativas), e é dada por

$$V = V(X,Y) = \sqrt{\frac{\chi^2}{n \min(R-1, C-1)}}$$,em que $$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$ com

Sendo:

- X: variável X (linha)
- Y: variável Y (coluna)
- R: quantidade de categorias da variável X
- C: quantidade de categorias da variável Y
- n: quantidade total de observações
- O_{ij}: valor observado dentro da casela *ij*
- E_{ij}: valor esperado dentro da casela *ij*
- *i*: o índice que percorre as *R* linhas
- *j*: o índice que percorre as *C* colunas
- *n*: a quantidade total de observações

A **Estatística V de *Cramer*** varia entre 0<=V<=1, com valores próximos de zero indicando fraca associação, e valores próximos de 1 indicando forte associação.

Case: *Credit Score*

Uma empresa está preocupada com os clientes novos que entram em sua carteira e apresentam um *'default'* (não pagamento da dívida) após um certo período. Esta análise trata da aprovação de um empréstimo de dinheiro para novos clientes em uma instituição financeira. O objetivo é fazer a aprovação de crédito de maneira automática, deixando apenas alguns casos para a análise de crédito manual. Para a aprovação ou não do cliente na instituição é utilizado o modelo de *Credit Score* que fornece a probabilidade do cliente apresentar um *'default'* por meio das informações cadastrais fornecidas pelo cliente e informações restritivas de mercado que a instituição consulta no momento da análise de crédito.

Fonte: base simulada, inspirada em problemas reais de consultoria da Prof.ª Karin Ayumi Tamura

(a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados na visão do negócio.
(b) Faça uma análise do % de *default*.
(c) Faça a análise bivariada das variáveis explicativas (covariáveis) vs. a variável resposta. Quais variáveis discriminam o evento resposta? Como você poderia tratar as categorias com *missings value* na análise bivariada?
(d) Rode o modelo de Regressão Logística. Selecione um modelo final no qual a interpretação dos parâmetros esteja de acordo com a análise bivariada.
(e) Faça a análise de multicolinearidade entre as covariáveis. Reajuste o modelo caso seja necessário, garantindo que as estimativas dos parâmetros fiquem condizentes com a análise exploratória bivariada.
(f) Qual o perfil, a probabilidade e a representatividade do cliente mais propenso a apresentar o *default*? E do menos propenso?
(g) Analise a sensibilidade, especificidade e acurácia pela tabela de classificação.
(h) Como você classifica o desempenho do modelo?
(i) Pelo valor do KS, você indicaria para área de negócios utilizar o modelo?
(j) Qual o percentual da base de dados que seria rejeitado?

Case: *Credit Score*


```
> library(lsr)#biblioteca para o cálculo da estatística de Cramers'V
> cramersV(table(credit$FX_IDADE,credit$CEP_GRUPO_RISCO))
[1] 0.2656272
> cramersV(table(credit$FX_IDADE,credit$FX_RENDA))
[1] 0.3737505
> cramersV(table(credit$FX_IDADE,credit$INDICADOR_RESTRITIVO))
[1] 0.05818821
> cramersV(table(credit$FX_IDADE,credit$QTDE_CONSULTAS_CREDITO))
[1] 0.07021716
> cramersV(table(credit$CEP_GRUPO_RISCO,credit$FX_RENDA))
[1] 0.394642
> cramersV(table(credit$CEP_GRUPO_RISCO,credit$INDICADOR_RESTRITIVO))
[1] 0.04852818
> cramersV(table(credit$CEP_GRUPO_RISCO,credit$QTDE_CONSULTAS_CREDITO))
[1] 0.04619015
> cramersV(table(credit$FX_RENDA,credit$INDICADOR_RESTRITIVO))
[1] 0.0660219
> cramersV(table(credit$FX_RENDA,credit$QTDE_CONSULTAS_CREDITO))
[1] 0.07320021
> cramersV(table(credit$INDICADOR_RESTRITIVO,credit$QTDE_CONSULTAS_CREDITO))
[1] 0.03992982
```

Case: *Credit Score*


```
glm(formula = RESPOSTA ~ FX_IDADE + CEP_GRUPO_RISCO + INDICADOR_RESTRITIVO +
    QTDE_CONSULTAS_CREDITO, family = binomial(link = "logit"),
    data = credit)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2894  -0.5392  -0.4336  -0.3588   2.6494

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -2.53823    0.08050 -31.532  < 2e-16 ***
FX_IDADE(2) 25 a 34           -0.55341    0.04523 -12.235  < 2e-16 ***
FX_IDADE(3) 35 a 39           -0.70554    0.06206 -11.369  < 2e-16 ***
FX_IDADE(4) 40 a 45           -0.74688    0.06445 -11.589  < 2e-16 ***
FX_IDADE(5) >=46              -0.94092    0.04721 -19.929  < 2e-16 ***
FX_IDADEsem informacao         0.05073    0.05424   0.935     0.35
CEP_GRUPO_RISCO(2) baixo       0.57438    0.07655   7.504 6.20e-14 ***
CEP_GRUPO_RISCO(3) medio       0.88713    0.07483  11.856  < 2e-16 ***
CEP_GRUPO_RISCO(4) alto        1.03685    0.07796  13.299  < 2e-16 ***
CEP_GRUPO_RISCO(5) altissimo   1.27567    0.07746  16.470  < 2e-16 ***
INDICADOR_RESTRITIVO           0.46471    0.09103   5.105 3.31e-07 ***
QTDE_CONSULTAS_CREDITO1        0.20002    0.03440   5.814 6.09e-09 ***
QTDE_CONSULTAS_CREDITO2        0.50299    0.04631  10.861  < 2e-16 ***
QTDE_CONSULTAS_CREDITO3 ou mais 1.00651   0.04704  21.395  < 2e-16 ***
---
```

Como avaliar o desempenho da Regressão Logística?

Uma vez o modelo interpretado e fazendo sentido na visão de negócios (avaliado o p-valor e o sinal de cada estimativa do parâmetro), o próximo passo é avaliar o desempenho de classificação das regras do modelo, que significa avaliar o resultado predito em comparação com a resposta observada.

Assim, pode-se ter uma ideia se o modelo, dado as variáveis explicativas presentes, é capaz de explicar o evento resposta de forma satisfatória para que a equação possa ser utilizada como regra para predizer o evento resposta.

Análise de Desempenho

A Tabela de Classificação apresenta o cruzamento da variável resposta observada em comparação com a variável resposta predita pelo modelo. Ela também é conhecida como Matriz de Confusão.

Um bom ajuste de modelo apresenta grande concentração de casos na diagonal principal.

Tabela de Classificação avaliada no ponto de corte:

		Variável Resposta Predita		Total
		0	1	
Variável Resposta Observada	0	VN	FP	VN+FP
	1	FN	VP	FN+VP
	Total	VN+FN	FP+VP	VN+FN+ FP+VP[1]

[1] *VP*: verdadeiro-positivo; *VN*: verdadeiro-negativo; *FP*: falso-positivo e *FN*: falso-negativo

Análise de Desempenho

Tabela de Classificação avaliada no ponto de corte:

		Variável Resposta Predita		Total
		0	1	
Variável Resposta Observada	0	VN	FP	VN+FP
	1	FN	VP	FN+VP
	Total	VN+FN	FP+VP	VN+FN+ FP+VP[1]

[1] *VP*: verdadeiro-positivo; *VN*: verdadeiro-negativo; *FP*: falso-positivo e *FN*: falso-negativo

Os índices de **acurácia**, **sensibilidade** e **especificidade** variam de 0 a 1 (ou de 0% a 100%).

Esperamos que os índices sejam superiores a 50% (acima do acerto aleatório), sendo os valores mais próximos de 100% com maior poder preditivo.

Na prática, valores acima 70%-75% com ótimo desempenho.

- Acurácia $$Acur = \frac{VP+VN}{VP+VN+FP+FN}$$
- Sensibilidade $$Sens = \frac{VP}{VP + FN}$$
- Especificidade $$Espec = \frac{VN}{FP + VN}$$

Case: *Credit Score*

Uma *fintech* está preocupada com os clientes novos que entram em sua carteira e que apresentam *'default'* (não pagamento do empréstimo) após um certo período. Esta análise trata da aprovação de Crédito Pessoal para novos clientes. O objetivo é fazer a aprovação de crédito de maneira automática por meio de um algoritmo, deixando apenas alguns casos duvidosos para a análise de crédito manual. Para a aprovação ou não do cliente na instituição é utilizado o modelo de *Credit Score* que fornece a probabilidade do cliente apresentar *'default'* por meio das variáveis cadastrais e informações de restritivos de mercado que a *fintech* consulta no momento da análise de crédito.

Fonte: base simulada, inspirada em problemas reais de consultoria da Prof.ª Karin Ayumi Tamura

(a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados na visão do negócio.
(b) Faça uma análise do % de *default*.
(c) Faça a análise bivariada das variáveis explicativas (covariáveis) vs. a variável resposta. Quais variáveis discriminam o evento resposta? Como você poderia tratar as categorias com *missings value* na análise bivariada?
(d) Rode o modelo de Regressão Logística. Selecione um modelo final no qual a interpretação dos parâmetros esteja de acordo com a análise bivariada.
(e) Faça a análise de multicolinearidade entre as covariáveis. Reajuste o modelo caso seja necessário, garantindo que as estimativas dos parâmetros fiquem condizentes com a análise exploratória bivariada.
(f) Qual o perfil, a probabilidade e a representatividade do cliente mais propenso a apresentar o *default*? E do menos propenso?
(g) Analise a sensibilidade, especificidade e acurácia pela tabela de classificação.
(h) Como você classifica o desempenho do modelo?
(i) Qual o percentual da base de dados que seria rejeitado?

Análise de Desempenho

Tabela de Classificação avaliada no ponto de corte (12,07%):

		Predito		
Tabela de Classificação		0	1	**Total**
Observado	0	29.570	14.235	**43.805**
	1	2.556	3.457	**6.013**
	Total	**32.126**	**17.692**	**49.818**

- Acurácia: $\frac{29.570+3.457}{49.818} = 66,3\%$
- Sensibilidade $\frac{3.457}{6.013} = 57,5\%$
- Especificidade $\frac{29.570}{43.805} = 67,5\%$

Os índices de **acurácia**, **sensibilidade** e **especificidade,** apresentaram desempenho regular, sendo possível predizer, de maneira geral, que quase 62% dos eventos de default e não default são preditos pelo modelo corretamente.

Deve-se salientar que é mais difícil predizer um comportamento e pagamento antes do cliente entrar na instituição e, geralmente, é esperado um acerto baixo, caso estejam disponíveis poucas informações do cliente na entrada.

4. Exercícios para casa

4. Exercícios para casa

DATA DE ENTREGA 29/11/2020 | 1 EXERCÍCIO-CASE

CASE: *Credit Score*

PARTE 1 (4,0 pontos)

PARTE 2 (6,0 pontos)

Instruções importantes:

- A lista vale nota (0-10) e deve ser entregue até 29/11/2020. Lista entregue até 06/12/2020 valerá 80% da nota. Posteriormente, não será mais aceita a lista para correção. Não serão aceitas listas parciais.
- O exercício será considerado como "realizado", quando tiver, além das análises, a interpretação do resultados.
- Soluções técnicas "elegantes e mais completas" serão considerados como ponto extra para o aluno (+0,5 na lista geral).
- Caso o aluno tire nota >10, considerando os pontos extras, os pontos extras poderão ser acumulados para listas seguintes, sendo a média geral de todas as listas realizadas no curso, com valor máximo igual a 10.

BOM ESTUDO ☺

Case: *Credit Score*

Uma *fintech* está preocupada com os clientes novos que entram em sua carteira e que apresentam *'default'* (não pagamento do empréstimo) após um certo período. Esta análise trata da aprovação de Crédito Pessoal para novos clientes. O objetivo é fazer a aprovação de crédito de maneira automática por meio de um algoritmo, deixando apenas alguns casos duvidosos para a análise de crédito manual. Para a aprovação ou não do cliente na instituição é utilizado o modelo de *Credit Score* que fornece a probabilidade do cliente apresentar *'default'* por meio das variáveis cadastrais e informações de restritivos de mercado que a *fintech* consulta no momento da análise de crédito.

Fonte: base simulada, inspirada em problemas reais de consultoria da Prof.ª Karin Ayumi Tamura

Os *outputs* em R já foram gerados em aula. Em conjunto com os resultados, interpretar e concluir na visão de negócios.

(a) Faça a análise exploratória univariada e interprete todas as variáveis do banco de dados na visão do negócio.
(b) Faça uma análise do % de *default*.
(c) Faça a análise bivariada das variáveis explicativas (covariáveis) vs. a variável resposta. Quais variáveis discriminam o evento resposta? Como você poderia tratar as categorias com *missings value* na análise bivariada?
(d) Rode o modelo de Regressão Logística. Selecione um modelo final no qual a interpretação dos parâmetros esteja de acordo com a análise bivariada.
(e) Faça a análise de multicolinearidade entre as covariáveis. Reajuste o modelo caso seja necessário, garantindo que as estimativas dos parâmetros fiquem condizentes com a análise exploratória bivariada.
(f) Qual o perfil, a probabilidade e a representatividade do cliente mais propenso a apresentar o *default*? E do menos propenso?
(g) Analise a sensibilidade, especificidade e acurácia pela tabela de classificação.
(h) Como você classifica o desempenho do modelo?
(i) Qual o percentual da base de dados que seria rejeitado?

Case: *Credit Score*

Uma *fintech* está preocupada com os clientes novos que entram em sua carteira e que apresentam *'default'* (não pagamento do empréstimo) após um certo período. Esta análise trata da aprovação de Crédito Pessoal para novos clientes. O objetivo é fazer a aprovação de crédito de maneira automática por meio de um algoritmo, deixando apenas alguns casos duvidosos para a análise de crédito manual. Para a aprovação ou não do cliente na instituição é utilizado o modelo de *Credit Score* que fornece a probabilidade do cliente apresentar *'default'* por meio das variáveis cadastrais e informações de restritivos de mercado que a *fintech* consulta no momento da análise de crédito.

Fonte: base simulada, inspirada em problemas reais de consultoria da Prof.ª Karin Ayumi Tamura

Continuando o mesmo problema solucionado em sala de aula, teste o modelo com a variável faixa de renda (retirando a faixa de idade).

(k) Refaça os itens (f) a (i) do slide anterior.

(l) Sugira para área de negócios qual modelo você recomendaria para ser utilizado pela instituição.

Referências

- Agresti, A. (2002). *Categorical data analysis* (Vol. 359). Wiley-interscience.
- Conover, W. J. (1999). *Practical nonparametric statistics*. New York: Wiley.
- Cramér, H. (1945). *Mathematical methods of statistics* (Vol. 9). Princeton university press.
- Hosmer, D. W. e Lemeshow, S. (2000). *Applied Logistic Regression*, 2nd ed. New York: Wiley.