ESTATÍSTICA APLICADA AO DATA SCIENCE

PREDIÇÕES COM REGRESSÃO LOGÍSTICA

Autor: Dr. Antonio Gomes de Mattos Neto

Revisor: Rafael Maltempe

INICIAR

introdução

Introdução

Na primeira unidade, vimos como aplicar modelos da estatística à ciência dos dados. Especificamente, estudamos um caso de aplicação de modelos de regressão linear simples e múltipla na predição do valor esperado de venda de imóveis. Nesse caso, a variável resposta, o valor do imóvel, é quantitativa. Denominamos modelos de regressão a classe de modelos que produzem, como saída, uma variável resposta quantitativa.

Nesta unidade, veremos outra classe de modelos preditivos, aqueles que produzem, como saída, o resultado de uma variável qualitativa. Modelos desse tipo são chamados modelos de classificação. Daremos início ao estudo de algoritmos de classificação examinando um dos mais famosos: o de regressão logística. Mas há uma curiosidade, afinal, regressão logística é um algoritmo de regressão (que faz predição de valores para variáveis quantitativas) ou um algoritmo de classificação (que faz predição de valores de variáveis qualitativas). Regressão logística é, de fato, empregada como um algoritmo de classificação, mas parte de sua construção funciona como um modelo de regressão. Esses pontos ficarão mais notórios ao longo da unidade.

O fato é que regressão logística é muito popular e aplicada à ciência dos dados com enorme frequência.

Tipos de Aprendizagem de Máquina

Nesta seção, veremos aprendizagem supervisionada e não supervisionada, jargões típicos da área de *machine learning*, que acabou sendo empregada também na ciência dos dados. Em seguida, como nosso foco são modelos preditivos, veremos quais são os dois tipos principais de aprendizagem supervisionada, regressão e classificação.

Aprendizagem não Supervisionada e Supervisionada

Em primeiro lugar, antes de descrevermos quais são os dois principais tipos de aprendizagem supervisionada, precisamos entender o que são aprendizagem supervisionada e não supervisionada. Vamos começar com esta última, utilizando dados estruturados, organizados em uma tabela, na qual as variáveis são dispostas nas colunas e as observações, nas linhas.

Tabela 2.1 - Organização dos dados em aprendizagem não supervisionada
Fonte: Elaborada pelo autor.

LINHAS	Observações, repetições, realizações, instâncias, exemplos
COLUNAS	Variáveis qualitativas ou quantitativas
CÉLULA	x_{ij} = Resultado da i-ésima observação da variável X_j

Na aprendizagem não supervisionada, não apontamos uma das variáveis como uma variável resposta, sobre a qual gostaríamos de predizer o resultado para diferentes valores das variáveis de entrada. O foco é nas observações, e o objetivo do aprendizado é o de procurar padrões comuns entre as observações da amostra.

Chamamos uma linha da tabela de observação. Uma observação i é um vetor de registros

$$(x_{i1}, x_{i2}, x_{i3}, \ldots, x_{im}) \qquad i = 1, 2, \ldots, n$$

dos valores das variáveis $X_1, X_2, X_3, \ldots, X_m$ da observação i, e n é o tamanho da amostra, igual ao número de linhas na tabela. Cada um desses vetores representa uma observação individual. Uma das mais frequentes abordagens da aprendizagem não supervisionada é a de tentar identificar similaridades entre essas observações X_i (isto é, similaridades entre os vetores de observações X_i) e, ao encontrar similaridades, agrupá-las. Aos indivíduos de um mesmo grupo, podemos dar um nome. A partir desse ponto, qualquer novo indivíduo observado será classificado como pertencente a um dos grupos previamente identificados. Esse ponto será elucidado na Unidade 4, quando estudaremos algoritmos de agrupamento, que fazem parte dos métodos de aprendizagem não supervisionada. Entretanto, se você quiser já ler algo sobre esse tema, recomendamos o livro *Introdução à Mineração de Dados com o R*, de Leandro Augusto da Silva *et al*. (2016) ou *Estatística Prática para Cientistas de Dados*, de Peter Bruce e Andrew Bruce (2019) .

Tabela 2.2 - Organização dos dados em aprendizagem supervisionada
Fonte: Elaborada pelo autor.

LINHAS	Observações, repetições, realizações, instâncias, exemplos
COLUNAS	Variáveis qualitativas ou quantitativas

CÉLULA x_{ij} = Resultado da i-ésima observação da variável X_j

y_i = Resultado da i-ésima observação da variável Y

Na aprendizagem supervisionada, as variáveis $X_1, X_2, ..., X_p$ são as variáveis de entrada, enquanto a variável Y é a variável de saída. Ao coletarmos os dados, consideramos uma das variáveis, que chamamos de Y, como uma resposta aos valores (dados de entrada) assumidos pelas outras variáveis, denominadas $X_1, X_2, ..., X_p$ e procuramos descobrir uma função que, alimentada com os dados de entrada, produza a resposta (Y). Os dados da variável resposta agem como supervisores ou exemplos das tentativas de acertar qual função leva aos valores de saída. Comparamos nossas estimativas $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_n$ obtidas por meio do modelo escolhido, com os valores observados $y_1, y_2, \ldots, y_n$ Essa comparação funciona com um supervisor, que nos diz quão boas são nossas estimativas. Quando fazemos essas tentativas, podemos aprimorar nosso modelo, usando algum critério de medida de *performance*, até ficarmos satisfeitos. Cada modelo preditivo possui um ou mais critérios de medida da sua *performance*.

Na aprendizagem supervisionada, outros nomes dados para as variáveis de entrada são: variáveis regressoras, explanatórias, preditoras ou independentes; e, para a variável resposta: variável de saída, dependente ou *target variable* (este último nome é um jargão da ciência da computação). Na Unidade 1, usamos esse arranjo de variáveis para fazer o ajuste dos modelos de regressão linear simples e múltipla aos dados observados. Especificamente, ajustamos os modelos aos dados usando o Método dos Mínimos Quadrados, já automaticamente embutido no *software* estatístico R, para determinar os coeficientes do modelo.

Dois Principais Tipos de Aprendizagem Supervisionada

Na aprendizagem supervisionada, usamos valores conhecidos das variáveis de entrada, obtidos por meio de uma amostra de tamanho n, e tentamos predizer o valor da variável resposta Y. Se anotarmos as variáveis de entrada como um vetor

$$X = (X_1, X_2, \ldots, X_p)$$

podemos escrever esse processo como aquele de procurar uma função preditiva f que faça esse trabalho, qual seja,

$$Y = f(X) + \epsilon$$

em que ϵ é um termo de erro aleatório. Representa a aleatoriedade do fenômeno estudado, ruídos ambientais, erros de medições, efeitos de variáveis que não sabemos existir, mas que influenciam o fenômeno estudado.

Quando Y é uma variável aleatória quantitativa, assume valores quantitativos, que são coisas que a gente consegue medir:

massa, comprimento, temperatura, preço, área, densidade, inflação...

Quando Y é uma variável qualitativa, assume como valores suas classes ou níveis, que são coisas que se consegue contar, por exemplo: quantas pessoas moram no Centro, Zona Leste, Zona Sul, Zona Norte ou Zona Oeste, na sua turma da faculdade. A sua turma é a sua amostra, e você conta quantos dos alunos são de uma ou de outra classe (zona onde moram). Outros exemplos são:

classes sociais (A, B, C, D e E), escolaridade (fundamental, médio, superior), sexo (feminino ou masculino), cor (vermelho, azul ou verde), gravidade de uma doença (leve, moderada, grave)...

Quando em um problema de aprendizagem supervisionada a variável resposta que queremos predizer é quantitativa, denominamos regressão. Quando em um problema de aprendizagem supervisionada a variável resposta é qualitativa, é denominado classificação.

Note que, para ambos os problemas, as variáveis de entrada podem ser quantitativas ou qualitativas, conforme já estudamos, na Unidade 1, para modelos de regressão linear simples e múltipla, na predição de valores de venda esperados para imóveis.

Finalmente, observamos que classificação é tanto ou mesmo mais frequente que regressão. Alguns exemplos nos ajudarão a perceber isso (JAMES *et al*., 2013):

i. Os sintomas apresentados por uma pessoa (X = batimentos cardíacos, pressão arterial, ritmo respiratório, movimentação ocular, ...) levam à suspeita de que pode estar tendo um dentre três tipos de ataques: Y = overdose, ataque cardíaco, ataque epiléptico.

ii. Um serviço de banco *on-line* pode suspeitar que a operação que está sendo realizada (X = IP do cliente, localização, valor, padrão de digitação, ...) é fraudulenta (Y = sim ou não fraudulenta).

iii. Um teste de sequência de DNA (X = sequência) pode indicar se o paciente tem alguma doença genética (Y = sim ou não doença genética).

praticar

Vamos Praticar

A análise preditiva é uma tarefa de mineração de dados aplicável em um grande número de domínios. Alguns exemplos de áreas nas quais a análise preditiva está presente são: análise do comportamento e expressão das emoções em redes sociais, realizada com base no vocabulário usado nas manifestações de opiniões sobre produtos; na biometria, com reconhecimento de íris, impressão digital, face ou assinatura; na predição de subida ou queda de ações no mercado financeiro; na Biologia, mediante a classificação de novas espécies de organismos vivos; na Medicina, com aplicação de modelos de predição categórica para auxiliar no diagnóstico de um tumor como maligno ou benigno.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados**: com aplicações em R. Rio de Janeiro: Elsevier Editora, 2016.

Está correto o que se afirma em:

- ○ **a)** Análise de comportamento é um problema de regressão.
- ○ **b)** Reconhecimento de íris, impressão digital, face ou assinatura são um problema de regressão.
- ○ **c)** A predição se um tumor é maligno ou benigno é um problema de regressão.
- ○ **d)** Todas as aplicações descritas são problemas de classificação.
- ○ **e)** Apenas a predição de subida e queda de ações são um problema de classificação.

Estudo de Caso - Predição de Inadimplência

Nesta seção, veremos como fazer predição de classes de uma variável qualitativa com modelos de regressão logística. Regressão logística é um método de classificação da estatística, de emprego muito comum na ciência dos dados. Explicaremos a aparente contradição do nome regressão, usado em problemas de classificação.

Regressão Logística e Outros Classificadores

São muitos os algoritmos de classificação disponíveis para o desenvolvimento de modelos preditivos: regressão logística, análise discriminante linear (LDA = Linear Discriminant Analysis), árvores de decisão para classificação, máquinas de vetores de suporte (SVM = *support vector machines*), k-vizinhos mais próximos (KNN = k-*nearest neighbors*). A lista não acaba aqui. Nesses listados estão alguns métodos de classificação clássicos e outros mais recentes, originados da fusão de métodos da estatística com métodos de aprendizado de máquina (ML = Machine Learning) da ciência da computação.

Este último termo, *machine learning*, você já deve ter ouvido ou lido. Classificação por métodos de *machine learning* estão, hoje, muito presentes no nosso dia a dia. São algoritmos como os que a Netflix usa para recomendar o próximo filme a ser assistido, ou que o Facebook usa para sugerir uma nova amiga ou um novo amigo para nossa rede de relacionamento social, ou bancos usam para detectar operações potencialmente fraudulentas com cartões de débito ou crédito, ou que concessionárias de distribuição de energia elétrica usam para identificar casos potenciais de roubo de energia da rede, os famosos “gatos”.

Entraremos nesse mundo via regressão logística. Nada melhor para demonstrar a aplicação da estatística à ciência dos dados. Porém, antes de apresentar o modelo de classificação por regressão logística, devemos entender duas formas diferentes de fazer classificação: determinística ou probabilística. Tome, por exemplo, a variável resposta qualitativa Y com dois

níveis (classes); o indivíduo está infectado pelo vírus HIV ($Y = 1$) ou não está infectado ($Y = 0$), dado um conjunto de sintomas $x = (x_1, x_2, \ldots, x_p)$ que apresenta.

No jargão da estatística, escrever $Y = y$ significa que a variável aleatória Y resultou no valor y, em que y é um dos possíveis valores que a variável aleatória Y pode assumir (ou seja, uma de suas classes, no caso das variáveis qualitativas). Também, nesse mesmo jargão, escrever $P(Y = y \mid X = x)$ significa a probabilidade de Y ser igual a um dos seus possíveis valores y quando a variável de entrada X é igual a x (dado que $X = x$).

Um classificador determinístico confirmará se o indivíduo está ou não está infectado, dados os sintomas que apresenta. Um classificador probabilístico determinará a probabilidade de o indivíduo estar ou não infectado, dados os sintomas que apresenta. Veja que, no primeiro caso, a variável resposta é claramente uma qualidade, estar ou não infectado, e o classificador classificará o indivíduo em uma das duas classes da variável resposta: sim ou não infectado. No segundo caso, o classificador produz, como saída, a probabilidade de cada um dos níveis (classes) se manifestar, ou seja, a probabilidade de o indivíduo estar ou não infectado.

Ambos os tipos de classificadores precisam ser treinados com base em dados que lhes são passados. No exemplo aqui discutido, são dados relativos a pessoas com sintomas indicativos de possível infecção por HIV, $x = (x_1, x_2, \ldots, x_p)$, e o resultado exato de um teste diagnóstico padrão ouro, que confirmou se essas pessoas estavam ou não com o vírus ($Y = 1$ ou 0)

O modelo de classificação por regressão logística é um classificador probabilístico. Indica a probabilidade de uma determinada classe e, em sintonia com o exemplo que acabamos de ver, é mais usado para o caso de variáveis respostas qualitativas com duas classes, apenas, ditas dicotômicas. Pode ser usado para variáveis respostas qualitativas com mais de duas classes, dita politômicas, mas isso é menos frequente no caso de classificação por regressão logística.

Finalmente, todos os modelos ou algoritmos da estatística ou de *machine learning* aplicados à ciência de dados erram. Em outras palavras, apresentam uma *performance* com maior ou menor nível de acertos e erros. Isso depende dos dados com os quais foram treinados e testados, e também do próprio jeito de funcionar do algoritmo. Cada um tem seu jeito próprio de funcionar, que pode ser melhor ou pior do que outro algoritmo, para cada situação específica.

Adiante, apresentaremos a técnica de classificação da regressão logística com a ajuda de um estudo de caso simulado. Mesmo sendo simulado, reflete bem situações vividas no mundo real para a aprovação de créditos bancários, tais como cartões de créditos. A diferença é que no mundo real a classificação é feita com um grande número de variáveis de entrada, enquanto neste estudo de caso simulado por conta de seu propósito didático, trabalharemos com um conjunto pequeno de variáveis preditivas. Esse estudo abrirá um grande leque de possíveis aplicações da estatística e das ciências dos dados no mundo no qual vivemos hoje. De fato, é exatamente isso que já está acontecendo em, praticamente, todas as áreas da atividade humana, seja no mundo acadêmico ou no mundo dos negócios.

Predição de Inadimplência com Cartões de Crédito

Um gerente de pessoas físicas de um banco de varejo vive em um ambiente onde questões relativas à análise de aprovação de cartão de crédito para seus clientes e à inadimplência no pagamento das faturas mensais desses cartões são frequentes.

A atividade de venda de produtos financeiros por bancos de varejo, tais como cartões de crédito, requer que se faça uma avaliação do cliente. É foco dessa avaliação aprovar ou não um cartão de crédito para o cliente e, se aprovado, definir o limite do cartão, ou seja, o valor do crédito a conceder.

O primeiro problema é um problema de classificação: aprovar (sim ou não) o cartão de crédito, uma variável qualitativa dicotômica, com dois níveis (classes). O segundo problema é um problema de regressão, predizer o valor do limite (do crédito) do cartão. Como já explicado, aqui, daremos atenção ao primeiro problema.

Como se fazia isso nos bancos e ainda se faz, ao menos em parte? Por meio da definição de regras que devem ser atendidas por cada cliente, tais como idade, emprego estável, renda fixa, dívidas pequenas, nome “limpo”, casa própria etc.

Como se faz isso com algoritmos ou modelos preditivos? Uma alternativa frequentemente usada são algoritmos de aprendizagem supervisionada. Para isso, precisamos de dados. Ensinamos ao algoritmo, com base nos dados que lhes são passados, a predizer clientes que são maus pagadores potenciais das faturas do cartão. Dessa forma, se o algoritmo, ao ser alimentado com os dados referentes a um novo cliente, classificar esse cliente como um mau pagador potencial, o banco não aprovará o cartão.

Para equipes de análise de crédito, poder contar com a ajuda de um *software* com a capacidade de recomendar a aprovação ou não da concessão do cartão é de grande valor. A recomendação feita pelo *software* poderá ser tratada ao lado de outras regras de crédito, para uma decisão final sobre a concessão de cartão para o cliente.

Dados de Inadimplência com Cartões de Crédito

Usaremos um conjunto de dados de um banco fictício de nome “Banco Mais com Menos”. O gerente desse banco decidiu investigar a possibilidade de trabalhar com um algoritmo de predição de potencial de inadimplência referente ao pagamento das faturas de cartão de crédito. Para isso, contratou uma jovem cientista de dados, que solicitou uma amostra, colhida randomicamente da base cadastral do banco, de pessoas físicas, dos últimos dois anos. A cientista de dados pediu uma amostra pequena, de tamanho 200, com o propósito de realizar alguns testes iniciais. Se tivesse sucesso, solicitaria mais dados ao banco para melhor treinar e testar o seu algoritmo, para uma futura validação pelo seu cliente: o banco.

A amostra lhe foi passada na forma de uma tabela, com 200 observações de 4 variáveis. Alguns de seus valores encontram-se exibidos no Quadro 2.3. As variáveis observadas foram a renda mensal da pessoa (R$), seu gasto médio com cartão de crédito (R$), se a pessoa tinha um emprego estável (Sim ou Não), e se a pessoa havia, ao longo do período pesquisado, ficado inadimplente com o pagamento de faturas do cartão ao menos uma vez (Sim ou Não).

Tabela 2.3 - Dados de inadimplência com cartões de crédito
Fonte: Elaborada pelo autor.

A jovem cientista de dados usou a seguinte nomenclatura, com o objetivo de usar uma notação mais compacta para as variáveis a serem analisadas:

$$n = \text{Tamanho da amostra} = 200$$

$$X_1 = \text{Renda mensal da pessoa (R\$)}$$

$$X_2 = \text{Gasto médio com cartão (R\$)}$$

$$X_3 = \text{Se a pessoa tem emprego estável (1 = Sim ou 0 = Não)}$$

$$Y = \text{Se a ficou inadimplente nos últimos 2 anos (1 = Sim ou 0 = Não)}$$

Decidiu realizar, em primeiro lugar, uma análise descritiva dos dados amostrados. Como a variável resposta de interesse é uma variável qualitativa dicotômica, também decidiu que adotaria, como modelo preditivo, o de classificação por regressão logística. Veremos adiante como ela prosseguiu com seu trabalho.

praticar

Vamos Praticar

"Tipicamente, uma base de dados usada em sistemas informatizados convencionais é organizada de forma que se tenham dados armazenados em estruturas tabulares, em que as linhas armazenam uma ocorrência de um evento caracterizado por um conjunto de colunas que representam características que descrevem um exemplar (instância) daquele evento."

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados**: com aplicações em R. Rio de Janeiro: Elsevier Editora, 2016, p. 7.

- ○ **a)** O trecho refere-se a dados não estruturados, tais como textos, imagens, vídeos e sons. Outros tipos de dados são chamados dados estruturados.
- ○ **b)** O trecho refere-se a dados qualitativos, que são os únicos que podem ser organizados em forma tabular. Dados quantitativos não podem ser organizados em forma tabular.
- ○ **c)** O trecho refere-se a dados quantitativos, que são os únicos que podem ser organizados em forma tabular. Dados qualitativos não podem ser organizados em forma tabular.
- ○ **d)** O trecho refere-se a dados dicotômicos ou binários, os únicos que podem ser organizados em forma tabular. Dados politômicos não podem ser organizados em forma tabular.
- ○ **e)** O trecho refere-se a dados estruturados, que são aqueles que podem ser organizados em forma tabular. Podem conter tanto dados quantitativos como qualitativos.

Análise Descritiva dos Dados

Nesta seção, faremos uma análise descritiva dos dados. Iniciaremos com a análise descritiva de cada variável da amostra, isoladamente. Depois, examinaremos a relação entre algumas dessas variáveis.

Análise Descritiva de Cada Variável da Amostra

São quatro as variáveis observadas neste estudo. A primeira delas é a renda mensal das pessoas. Usando as funções min(), mean() e max() do R, a cientista de dados obteve:

min(x1) = 1.137,02 mean(x1) = 3.405,56 max(x1) = 9.086,15

Para visualizarmos a distribuição de frequências desses dados, como se trata de uma variável quantitativa, recorreu à função gráfica hist() do R. Obteve o histograma exibido na Figura 2.4.

A segunda delas é o gasto médio das pessoas com cartão de crédito. Usando as funções min(), mean() e max() do R, obteve:

min(x2) = 379,79 mean(x2) = 1.180,87 max(x2) = 3.118,27

Para visualizar a distribuição de frequências desses dados, como também se trata de uma variável quantitativa, novamente recorreu à função gráfica hist() do R e obteve o histograma exibido na Figura 2.5.

A terceira é uma variável qualitativa dicotômica, que indica se a pessoa tem ou não um emprego estável. Para contar a frequência desses valores no conjunto de dados da amostra, a jovem cientista de dados empregou a função table() do R:

```
table(x3)      N       S

               94     106
```

Ou seja, das 200 pessoas da amostra, 94 não tinham emprego estável, enquanto 106 tinham um emprego estável.

A quarta e última refere-se à variável resposta (também qualitativa dicotômica), que indica se a pessoa ficou ou não inadimplente ao longo do período estudado. A jovem cientista de dados também usou, nesse caso, a função table() do *software* estatístico R, para contar a frequência de aparecimento desses valores na amostra coletada e obteve:

```
table(y)       N       S
```

```
122  78
```

Ou seja, das 200 pessoas da amostra, 122 não tinham ficado inadimplentes nos dois anos do período selecionado e 78, sim, e falharam em pagar a fatura do cartão ao menos uma vez ao longo desse período.

Análise Descritiva da Relação entre Variáveis

Até agora, a jovem cientista de dados havia examinado as variáveis uma a uma. Decidiu analisar a relação entre algumas delas. Começou com a tentativa de visualizar a relação entre três variáveis: a renda mensal das pessoas, gasto médio mensal com o cartão de crédito dessas mesmas pessoas, e se haviam ou não ficado inadimplentes. Para isso, recorreu à função plot() do R base e obteve o gráfico exibido na Figura 2.6. Esse gráfico exibe as duas variáveis quantitativas nos eixos horizontal e vertical, respectivamente, e a variável qualitativa, que indica se a pessoa ficou ou não inadimplente, foi exibida com as cores azul-claro para os adimplentes e laranja para os inadimplentes.

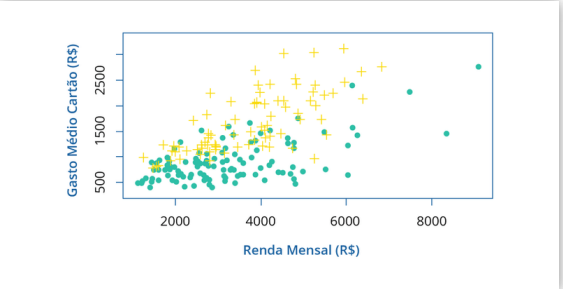


Figura 2.6 - Renda mensal x gasto médio cartão x inadimplência
Fonte: Elaborada pelo autor.

Para visualizar a relação da renda mensal (variável quantitativa) com o status de adimplência das pessoas (variável qualitativa), decidiu usar a função gráfica boxplot() do R. Fez o mesmo para visualizar a relação entre o gasto médio mensal das pessoas com cartão de crédito com seu status de adimplência. As Figuras 2.7 e 2.8 exibem esses gráficos:

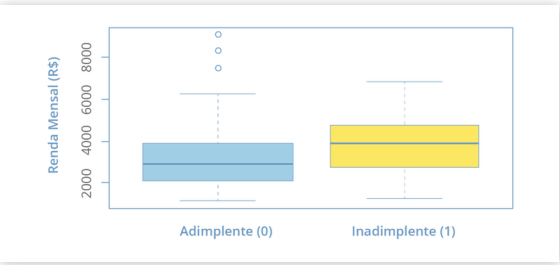


Figura 2.7 - Boxplot renda mensal x inadimplência
Fonte: Elaborada pelo autor.

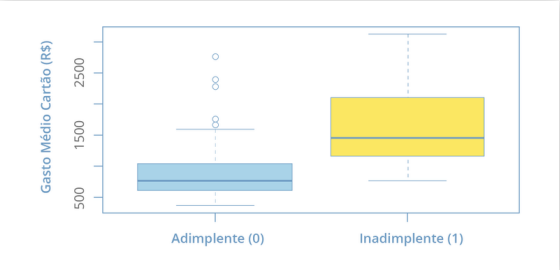


Figura 2.8 - Boxplot gasto médio mensal com cartão x inadimplência
Fonte: Elaborada pelo autor.

É fácil perceber que há um efeito de aumento da inadimplência, tanto com um aumento da renda média da pessoa quanto com o aumento de seus gastos com cartão de crédito. Esse efeito, porém, é mais pronunciado pelo aumento dos gastos com cartão do que com a renda mensal das pessoas. Por meio do *software* R é possível verificar que as pessoas adimplentes têm uma renda média de R$ 3.188,31 e as inadimplentes, de R$ 3.745,33, ou seja, 17.5% a mais. Igualmente, é possível verificar que as pessoas adimplentes gastaram, em média, R$ 879,01 por mês com o cartão de crédito e que as inadimplentes gastaram, em média, R$ 1.653,00, ou seja, 88.1% a mais, quase o dobro.

Para finalizar a análise descritiva, a jovem cientista resolveu investigar a relação entre as duas variáveis qualitativas emprego estável (S ou N) e inadimplência (S ou N). Para isso, empregou, novamente, a função table() do R e obteve:

	y	
x3	N	S
N	41	53
S	81	25

Esse resultado indica que, das 53 + 25 = 78 pessoas com emprego estável, apenas 25 ficaram inadimplentes (32,1%). Por outro lado, do total das 41 + 81 = 122 pessoas sem emprego estável,

81 ficaram inadimplentes (66,4%). O efeito da instabilidade de empregos no nível de inadimplência das pessoas é muito forte, ao menos para essa pequena amostra de 200 pessoas. A jovem cientista de dados resolveu visualizar esse resultado recorrendo à função gráfica mosaicplot() do R, como mostra a Figura 2.9.

Com isso, a análise descritiva inicial foi finalizada. A nossa jovem cientista de dados, já tendo decidido anteriormente desenvolver seu modelo preditivo para esse caso, adotando a regressão logística, respirou fundo e deu partida a esse desafiante passo, com veremos adiante.

praticar

Vamos Praticar

"Estatística é a ciência dos dados. Um aspecto importante de lidar com dados é organizar e resumir os dados em maneiras que facilitem sua interpretação e análise subsequente [...] Veremos que há métodos numéricos para resumir dados e um número de técnicas gráficas poderosas. As técnicas gráficas são particularmente importantes. Qualquer boa análise estatística deve sempre começar plotando os dados."

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. 5. ed. Rio de Janeiro: LTC, 2013, p. 128.

Quanto a esse assunto, analise as afirmativas a seguir:

I. A organização de dados em tabelas e o cálculo de resumos estatísticos são um aspecto importante para a interpretação e análise de dados.

II. Resumos estatísticos são calculados com base em amostras de dados e também são chamados sumários estatísticos.

III. Técnicas gráficas são poderosas para a interpretação e análise de dados, e qualquer análise estatística deve sempre começar plotando gráficos.

IV. O termo estatística descritiva refere-se a um conjunto de técnicas de organização de dados, cálculo de resumos e exposição gráfica dos dados.

Está correto o que se afirma em:

- ○ **a)** II, III e IV, apenas.
- ○ **b)** II e III, apenas.

- ○ **c)** III e IV, apenas.
- ○ **d)** I, II, III, apenas.
- ○ **e)** I, II, III e IV.

Predições com Modelos de Regressão Logística

Nesta seção, ajustaremos modelos de regressão logística simples e múltipla aos dados. Com os modelos prontos, realizaremos predições de classes, usando o caso para o qual foi contratada, pelo gerente do banco, a nossa jovem cientista de dados.

Modelo de Regressão Logística Simples

Como já dito, a regressão logística é um classificador muito usado em situações nas quais a variável qualitativa é dicotômica. Não somente isso, mas também quando as classes se misturam um pouco e não há uma fronteira muito clara de divisão entre elas. Aqui, entre adimplentes e inadimplentes. Para enxergar esse ponto, volte à Figura 2.6 e veja como os pontos azuis dos adimplentes se misturam um pouco com os pontos laranja dos inadimplentes. Nessa situação, a regressão logística também se mostra uma opção interessante.

Outro ponto a observar é que nossa jovem cientista de dados verificou, na sua análise descritiva dos dados, que dentre as duas variáveis quantitativas (renda) e (gastos com cartão), a segunda é mais influente. Dessa forma, decidiu começar com um modelo de regressão logística simples, com uma só variável de entrada, justamente a mais influente, a variável (gastos médios mensais com o cartão). O modelo preditivo que vai tentar desenvolver será um de predição da probabilidade de a pessoa ficar inadimplente em função dos seus gastos médios com o cartão.

Na discussão que se segue, anotaremos $p(x_2)$ para a probabilidade esperada da variável resposta Y do *status* de adimplência das pessoas ser igual a sua classe $y = 1$ (inadimplente), quando o valor da variável de entrada X_2 for igual a um determinado valor x_2 de gasto médio mensal dessa pessoa com o seu cartão de crédito, isto é, $P(Y = 1 \mid X = x_2)$. Essa notação, típica dos campos da probabilidade e da estatística, lê-se: probabilidade de $Y = 1$ dado $X = x_2$.

A palavra regressão, em regressão logística, tem relação com a regressão linear, que já vimos na Unidade 1. Por outro lado, nesta unidade, já vimos que a regressão logística é um classificador probabilístico. Porém, não é boa ideia a nossa cientista de dados tentar desenvolver um modelo preditivo de probabilidade usando uma equação como

$$p(x_2) = b_0 + b_2 x_2$$

pois essa equação representa uma reta e $p(x_2)$, sendo uma reta, poderia assumir valores menores do que 0 (probabilidades negativas) ou maiores do que 1 (maiores do que 100%), o que não é possível para probabilidades.

Para resolver esse problema, os estatísticos recorreram a outro modelo, substituindo a probabilidade $p(x_2)$ na equação acima pelo logaritmo de sua chance, escrita como:

$$\log [\, p(x_2) \,/\, (1 - p(x_2)) \,] = b_0 + b_2 x_2$$

Dessa equação, podemos isolar $p(x_2)$ com alguns poucos passos de álgebra, para obter

$$p(x_2) = [\, \exp(b_0 + b_2 x_2) \,] \,/\, [\, 1 + \exp(b_0 + b_2 x_2) \,]$$

Essa última equação garante que a probabilidade $p(x_2)$ ficará contida entre os limites 0 e 1, para qualquer valor da variável de entrada x_2.

A partir desse ponto, o procedimento seguido pela jovem cientista de dados foi o de estimar os coeficientes b_0 e b_2 usando o *software* estatístico R. O método que o R usa, aqui, é o de Minimização da Função de Verossimilhança, algo parecido ao que faz o Método dos Mínimos Quadrados para o caso dos modelos de regressão linear. A cientista obteve os seguintes valores para os coeficientes do modelo:

$b0 = -4{,}16$ e $b2 = 0{,}00314$

Podemos, agora, predizer a probabilidade de uma pessoa ser inadimplente em função dos seus gastos médios com o cartão de crédito. Analisemos duas pessoas: uma com gastos mensais médios com cartão de crédito de R$ 500,00, e outra de R$ 1000,00, ou seja, o dobro da primeira. Usando os valores calculados para seus coeficientes, a equação fica

$$p(x_2) = [\exp(-4{,}16 + 0{,}0314\, x_2)\ ] / [\ 1 + \exp(-4{,}16 + 0{,}00314 x_2)\ ]$$

e obtemos

$p(500) = 0{,}07$ e $p(1000) = 0{,}27$

Em palavras, a probabilidade de a primeira pessoa ficar inadimplente gastando R$ 500,00 por mês com cartão de crédito é de 7%, enquanto para a segunda pessoa com gastos de R$ 1000,00, essa mesma probabilidade é de 27%. Ou seja, a segunda pessoa tem uma probabilidade 3,8 vezes maior de ficar inadimplente do que a primeira pessoa.

Para visualizar esse resultado, a jovem cientista de dados construiu um gráfico no qual plotou, simultaneamente:

1. na cor azul-claro não inadimplentes $y = 0$ *versus* gastos com cartão x_2
2. na cor laranja inadimplentes $y = 1$ *versus* gastos com cartão x_2
3. na cor salmão probabilidade $p(x_2)$ de inadimplência ($y = 1$) *versus* gastos com cartão x_2

Obteve o gráfico exibido na Figura 2.10:

Esse exemplo deixa claro o que queremos dizer quando denominamos, a regressão logística, classificador probabilístico, que estima a probabilidade de acontecer uma das classes da variável resposta, em função do valor da variável de entrada. No caso, o da pessoa ficar inadimplente. A predição para a probabilidade de a outra classe acontecer é, simplesmente,

$$1 - p(x2)$$

que é a probabilidade de a pessoa não ficar inadimplente. O banco pode, então, decidir usar um valor limite superior, tal como $p(x2) = 0.05$, isto é, 5%, como seu critério de aprovação do cartão. Se a probabilidade de a pessoa ficar inadimplente for menor ou igual a esse valor, o banco aprova o cartão e desaprova se essa probabilidade for superior a esse valor limite.

Ao chegar a esse ponto, a jovem cientista de dados já estava bastante satisfeita com esses resultados parciais. Decidiu dar mais um passo adiante.

Modelo de Regressão Logística Múltipla

Regressão múltipla é aquela cujo modelo de regressão considera mais de uma variável de entrada. Na amostra cedida pelo gerente do banco à jovem cientista de dados, há 200 observações de 4 variáveis. Dessa forma, ela escreveu o modelo de regressão logística múltipla da seguinte forma:

$$\log [\, p(x) \,/\, (1 - p(x)) \,] = b0 + b1x1 + b2x2 + b3x3$$

em que

$$x = (x1, x2, x3)$$

$$x1 = \text{Renda mensal da pessoa (R\$)}$$

x_2 = Gasto médio mensal com cartão de crédito (R$)

x_3 = Se a pessoa tem um emprego estável (1 = Sim ou 0 = Não)

$p(x)$ = Probabilidade esperada da pessoa ficar inadimplente $(0 \leq p(x) \leq 1)$

Assim como no caso do modelo de regressão logística simples, a jovem cientista de dados fez o ajuste desse modelo aos dados da amostra com a ajuda do *software* R, e obteve, para os coeficientes:

b0 = - 2,85

b1= 0,000920

b2 = 0,00519

b3 = -1,52

Ela sabia que a relação entre $\log[p(x) / (1 - p(x))]$ e $p(x)$ é tal que, quando uma cresce, a outra também cresce. Com isso, pôde deduzir que $p(x_2)$ a probabilidade de a pessoa ser inadimplente:

Isso nós já sabíamos, da análise descritiva feita pela nossa jovem cientista de dados, mas, agora, ela foi muito além. Quantificou essas relações por meio desse modelo de regressão logística múltipla. E se apressou em mostrar esse resultado ao gerente do banco, aquele que a havia contratado. Este logo pediu, à cientista, que demonstrasse o poder de predição desse modelo e, para isso, ela apresentou duas situações:

Primeira situação

Duas pessoas, com renda de R$ 1000,00 ao mês e gastos médios mensais com cartão de crédito de R$ 400,00; porém, uma com emprego estável e a outra sem emprego estável, como segue:

Pessoa A: $x_1 = 1200,\ x_2 = 400$ e $x_3 = 1 \quad \Rightarrow \quad p(1200, 400, 1) = 0,032$

Pessoa B: $x_1 = 1200,\ x_2 = 400$ e $x_3 = 0 \quad \Rightarrow \quad p(1200, 400, 0) = 0,133$

A pessoa A tem uma probabilidade de 3,2% de ficar inadimplente com o cartão. Para a pessoa B, essa probabilidade é de 13,3%. Desse modo, se o banco usar o critério do limite máximo de 5%, aprovaria o cartão para A, e não para B.

Segunda situação

Duas pessoas, com renda de R$ 8.000,00 ao mês e ambas com empregos estáveis; porém, uma com gastos médios mensais com cartão de crédito de R$ 1.500,00 e a outra, R$ 3.000,00:

Pessoa C: $x_1 = 8000,\ x_2 = 1500$ e $x_3 = 1 \quad \Rightarrow \quad p(8000, 1500, 1) = 0,019$

Pessoa D: $x_1 = 8000,\ x_2 = 3000$ e $x_3 = 1 \quad \Rightarrow \quad p(8000, 3000, 1) = 0,979$

A pessoa C tem uma probabilidade de 1,9% de ficar inadimplente com o cartão. Para a pessoa B, essa probabilidade é de 97,9%. Obviamente, para essa pessoa, o banco não aprovaria o cartão de crédito.

O gerente do banco ficou tão contente que convidou a jovem cientista de dados a ingressar definitivamente para sua equipe de inteligência de negócios e também pediu-lhe para liderar um time no uso desse tipo de ferramentas de estatística e *machine learning*. Concluiu que já estava na hora de mudar alguns dos processos do banco com esse tipo de tecnologia.

praticar

Vamos Praticar

Considere os seguintes exemplos de aplicação de regressão logística, assim como as afirmativas, a seguir: 1) previsão de risco na área tributária – calcular a probabilidade de o contribuinte ser inadimplente ou adimplente após o parcelamento de tributos; 2) utilizada para classificar se a empresa encontra-se no grupo de empresas solvente ou insolvente; 3) determinar quais características levam as empresas a adotarem o *balanced scorecard*.

UNIVERSIDADE DE SÃO PAULO. Sistemas de Apoio às Disciplinas. **Regressão Logística**. [2019]. Disponível em: https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf. Acesso em: 26 dez. 2019.

I. É uma técnica recomendada para situações em que a variável dependente é de natureza dicotômica ou binária. Quanto às independentes, podem ser categóricas ou não.

II. A regressão logística é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explanatórias.

III. Busca estimar a probabilidade de a variável dependente assumir um determinado valor em função dos conhecidos de outras variáveis.

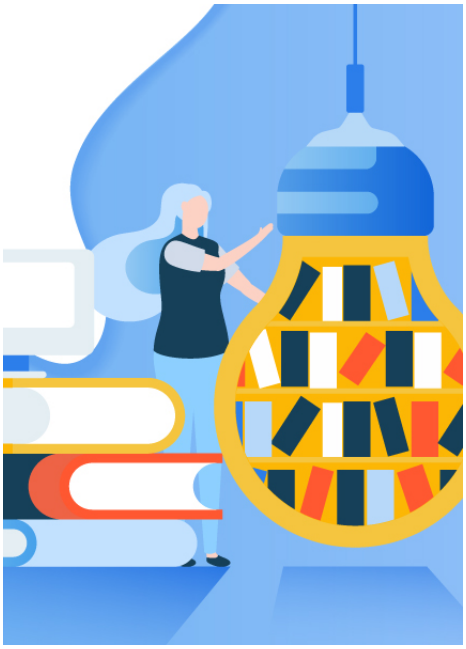
IV. Os resultados da análise ficam contidos no intervalo entre zero a um.

Está correto o que se afirma em:

- ○ **a)** II, III e IV, apenas.
- ○ **b)** II e III, apenas.
- ○ **c)** I, II, III e IV.
- ○ **d)** I, II, III, apenas.
- ○ **e)** III e IV, apenas.

indicações

Material Complementar



LIVRO

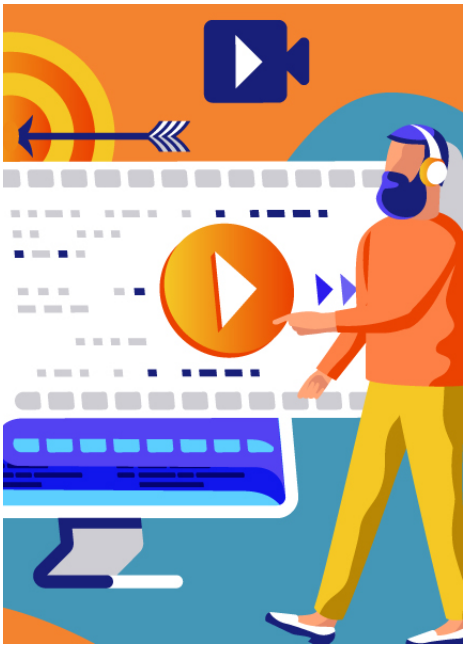
Análise Estatística com R para Leigos

Joseph Schmuller

Editora: Alta Books

ISBN: 978-85-508-0485-9

Comentário: nesse livro, o autor procura apresentar a estatística de maneira fácil, com o uso do *software* estatístico R. Faz um balanço entre conceitos estatísticos e programação em R, de forma a tornar o mais fácil possível o aprendizado.



FILME

O homem que mudou o jogo

Ano: 2011

Comentário: esse filme discorre sobre um treinador de um time de beisebol que decidiu usar a análise de dados e estatística no processo de tomada de decisão para melhorar a *performance* do seu time.

Para saber mais sobre o filme, acesse o *trailer*.

TRAILER

conclusão

Conclusão

Nesta unidade, vimos um caso simulado. Casos reais são similares ao caso abordado. Porém, é comum encontrarmos, nos casos reais, muito mais dados, tanto em número de observações, que chegam à casa de milhares ou milhões, como também em número de variáveis, que facilmente chegam a algumas dezenas ou mesmo centenas. No caso aqui estudado, a amostra possuía apenas 200 observações e somente quatro variáveis. Para problemas em dimensões maiores – ou muito maiores, como os problemas chamados de *big data* –, mais importante ainda é o uso intensivo de técnicas computacionais na aplicação da estatística à ciência dos dados.

A ideia, nesta unidade, como também foi a da unidade anterior, foi a de mostrar poder dessas técnicas, modelos e algoritmos, quando usados em favor da sociedade humana, dos seus negócios, das suas pesquisas. Não seria produtivo tentar, nessa introdução, cobrir em mais profundidade detalhes importantes para a construção desses modelos, como treiná-los, como testá-los e como validá-los, por exemplo. Sendo assim, há muitas coisas que não vimos e que deixamos para você, ao longo da sua trajetória como estudante e, no futuro, como profissional, explorar e aprender, se assim desejar.

Concluindo, essa área, hoje, é muito promissora e valorizada pelo mercado de trabalho, e que qualquer profissional pode se apoderar dessas ferramentas e aplicá-las à sua área de especialização.

referências

Referências Bibliográficas

BRUCE, P.; BRUCE, A. **Estatística prática para cientistas de dados**: 50 conceitos essenciais. Rio de Janeiro: Alta Books, 2019.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning**: with applications in R. New york: Springer, 2013.

MENDES, C. A. B.; VEGA, F. A. C. Técnicas de regressão logística aplicada à análise ambiental. **Revista Geografia (Londrina)**, v. 20, n. 1, 2011. Disponível em: http://www.uel.br/revistas/uel/index.php/geografia/article/view/6878. Acesso em: 26 dez. 2019.

MONTGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**. 5. ed. Rio de Janeiro: LTC, 2013.

SILVA, L. A. da; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados**: com aplicações em R. Rio de Janeiro: Elsevier, 2016.

UNIVERSIDADE DE SÃO PAULO. Sistemas de Apoio às Disciplinas. **Regressão Logística**. [2019]. Disponível em: https://edisciplinas.usp.br/pluginfile.php/3769787/mod_resource/content/1/09_RegressaoLogistica.pdf Acesso em: 5 dez. 2019.