

GUIA MANGÁ ANÁLISE DE REGRESSÃO

SHIN TAKAHASHI,
IROHA INOUE E
TREND-PRO CO., LTD.



Novatec

The Manga Guide to Regression Analysis is a translation of the Japanese original, *Manga de wakaru tōkei-gaku kaiki bunseki-hen*, published by Ohmsha, Ltd. of Tokyo, Japan, © 2005 by Shin Takahashi and TREND-PRO Co., Ltd. The English edition is co-published by No Starch Press, Inc. and Ohmsha, Ltd. Portuguese-language rights arranged with Ohmsha, Ltd. and No Starch Press, Inc. for *Guia Mangá Análise de Regressão* ISBN 978-85-7522-698-8, published by Novatec Editora Ltda.

Edição original em japonês *Manga de wakaru tōkei-gaku kaiki bunseki-hen*, publicado pela Ohmsha, Ltd. de Tóquio, Japão © 2005 por Shin Takahashi e TREND-PRO Co., Ltd. Edição em inglês *The Manga Guide to Regression Analysis*, co-publicação da No Starch Press, Inc. e Ohmsha, Ltd. Direitos para a edição em português acordados com a Ohmsha, Ltd. e No Starch Press, Inc. para *Guia Mangá Análise de Regressão* ISBN 978-85-7522-698-8, publicada pela Novatec Editora.

Copyright © 2018 da Novatec Editora Ltda.

Todos os direitos reservados e protegidos pela Lei 9.610, de 19/02/1998.

É proibida a reprodução desta obra, mesmo parcial, por qualquer processo, sem prévia autorização, por escrito, do autor e da editora.

Editor: Rubens Prates

Tradução: BrodTec

Revisão gramatical: Tássia de Carvalho

Editoração eletrônica: Carolina Kuwabata

ISBN: 978-85-7522-698-8

Histórico de impressões:

Setembro/2018 Primeira edição

NOVATEC EDITORA LTDA.

Rua Luís Antônio dos Santos 110

02460-000 – São Paulo, SP – Brasil

Tel.: +55 11 2959-6529

E-mail: novatec@novatec.com.br

Site: www.novatec.com.br

Twitter: twitter.com/novateceditora

Facebook: facebook.com/novatec

LinkedIn: linkedin.com/in/novatec

SUMÁRIO

PREFÁCIOxi
PRÓLOGO	
ACEITA MAIS CHÁ?	1
1	
UM COPO REFRESCANTE DE MATEMÁTICA.	11
Construindo uma base.	12
Funções inversas.	14
Expoentes e Logaritmos.	19
Regras para exponenciais e logaritmos.	21
Cálculo diferencial.	24
Matrizes.	37
Somando matrizes.	39
Multiplicando matrizes.	40
Regras para a multiplicação de matrizes.	43
Matriz Identidade e Matriz Inversa.	44
Tipos de Dados Estatísticos.	46
Teste de Hipóteses.	48
Medindo a variação.	49
Somatório dos desvios ao quadrado.	50
Variância.	50
Desvio Padrão.	51
Função Densidade de Probabilidade.	52
Distribuição Normal.	53
Distribuição Qui-quadrado.	54
Tabelas de Distribuição Densidade de Probabilidade.	55
Distribuição F.	57
2	
ANÁLISE DE REGRESSÃO SIMPLES	61
Primeiros passos.	62
Plotando os dados.	64
A equação de regressão.	66
Procedimento geral para análise de regressão.	68
Passo 1: Desenhe um gráfico de dispersão da variável independente versus a variável dependente. Se os pontos estiverem alinhados, as variáveis podem estar correlacionadas.	69
Passo 2: Calcule a equação de regressão.	71
Passo 3: Calcule o coeficiente de correlação (R) e avalie nossa população e premissas.	78
Amostras e populações.	82
Hipóteses de normalidade.	85
Passo 4: Conduza a análise da variância.	87

Passo 5: Calcule os intervalos de confiança.....	91
Passo 6: Faça uma previsão!	95
Quais são os passos necessários?	100
Resíduo Padronizado	100
Interpolação e extrapolação	102
Autocorrelação.....	102
Regressão não linear	103
Transformando equações não lineares em equações lineares	104

3

ANÁLISE DE REGRESSÃO MÚLTIPLA..... 107

Fazendo previsões com diversas variáveis	108
A equação de regressão múltipla	112
Procedimento para a análise de regressão múltipla	112
Passo 1: Desenhe um gráfico de dispersão de cada variável preditora e da variável resultado para analisar se elas estão relacionadas.	113
Passo 2: Calcule a equação de regressão múltipla.....	115
Passo 3: Examine a acurácia da equação de regressão múltipla.....	119
O problema do R^2	122
R^2 ajustado	124
Teste de Hipóteses na regressão múltipla.....	127
Passo 4: Conduza a análise do teste de variância (ANOVA)	128
Encontrando S_{11} e S_{22}	132
Passo 5: Calcule o intervalo de confiança da população.	133
Passo 6: Faça uma previsão!	136
Escolhendo a melhor combinação de variáveis preditoras.	138
Analisando a População com a Análise de Regressão Múltipla	142
Resíduos padronizados	143
Distância de Mahalanobis	144
Passo 1	144
Passo 2	145
Passo 3	146
Utilizando dados categóricos na análise de regressão múltipla	147
Multicolinearidade.....	149
Determinando a influência relativa das variáveis preditoras na variável resultado	149

4

ANÁLISE DE REGRESSÃO LOGÍSTICA 153 |

A última aula.....	154
O método da máxima verossimilhança	160
Encontrando a máxima verossimilhança utilizando a função de verossimilhança.	163
Escolhendo variáveis preditoras	164
Análise de regressão logística em ação!	168
Procedimento para a análise de regressão logística	168
Passo 1: Desenhe um gráfico de dispersão para as variáveis preditoras e variáveis resultado para ver se existe relação entre elas.	169
Passo 2: Calcule a equação de regressão logística.....	170

Passo 3: Obtenha a acurácia da equação.	173
Passo 4: Faça um teste de hipóteses.	178
Passo 5: Prever se a Norns venderá ou não o prato especial.	182
Análise de Regressão Logística no Mundo Real.	190
Função Logit, Razão de Possibilidades e Risco Relativo	190
Função Logit	190
Razão de Possibilidades.	191
Razão de Possibilidades Ajustada	192
Teste de Hipóteses com Possibilidades.	194
Intervalo de Confiança para Razão de Possibilidades	194
Risco Relativo	195
APÊNDICE	
CÁLCULOS DE REGRESSÃO COM O EXCEL	197
Número de Euler	198
Potências	200
Logaritmos Naturais	200
Multiplicação de matrizes	201
Matrizes inversas.	202
Calculando a estatística qui-quadrado a partir de um p-valor	204
Calculando um p-valor a partir de uma estatística qui-quadrado	205
Calculando uma estatística F a partir de um p-valor	206
Calculando um p-valor a partir de uma estatística F	208
Coeficiente de regressão parcial de uma análise de regressão múltipla	209
Coeficiente de regressão de uma análise de regressão logística.	210
ÍNDICE REMISSIVO.	213

PREFÁCIO

Este livro é uma introdução à análise de regressão, cobrindo a análise de regressão simples, múltipla e logística.

As análises de regressão simples e múltiplas são métodos estatísticos para prever valores: por exemplo, você pode utilizar a regressão simples para prever quantos chás gelados serão pedidos, baseando-se na temperatura máxima do dia ou utilizar a análise de regressão múltipla para prever os salários mensais de uma loja, baseando-se no seu tamanho e na distância até a estação de trem mais próxima.

A análise de regressão logística é um método para prever a probabilidade de se vender um bolo específico baseando-se no dia da semana, por exemplo. O público-alvo deste livro são estudantes de matemática e estatística ou qualquer pessoa que esteja começando a aprender sobre previsões estatísticas e probabilidades. Você precisará de alguns conceitos básicos sobre estatística antes de começar. O *Guia Mangá de Estatística* (Novatec Editora, 2010) é um excelente guia para prepará-lo para este livro. Esta obra consiste em quatro capítulos:

- Capítulo 1: Um Copo Refrescante de Matemática
- Capítulo 2: Análise de Regressão Simples
- Capítulo 3: Análise de Regressão Múltipla
- Capítulo 4: Análise de Regressão Logística

Cada capítulo tem uma seção em mangá e uma parte mais técnica em texto. Você pode obter uma visão geral a partir do mangá e alguns detalhes mais úteis bem como definições específicas a partir das seções de texto.

Gostaria de mencionar alguns pontos sobre o Capítulo 1. Embora muitos leitores tenham conhecimento prévio sobre os tópicos deste capítulo, como diferenciação e operações com matrizes, o capítulo revisa estes tópicos dentro do conceito de análise de regressão, o que será útil nas seções seguintes. Se o Capítulo 1 for uma simples revisão para você, ótimo. Se nunca estudou esses tópicos ou faz muito tempo que estudou sobre eles, vale a pena se empenhar um pouco e ter a certeza de que entendeu o Capítulo 1 antes de continuar.

Neste livro, a matemática para os cálculos é tratada em detalhes. Se você é bom em matemática, será capaz de acompanhar e entender os cálculos. Se não for tão bom em matemática, pode obter uma visão geral dos procedimentos e utilizar as instruções passo a passo para encontrar as respostas. Você não precisa obrigar-se a entender a parte matemática agora. Mantenha-se relaxado, mas dê uma olhada no desenvolvimento dos cálculos.

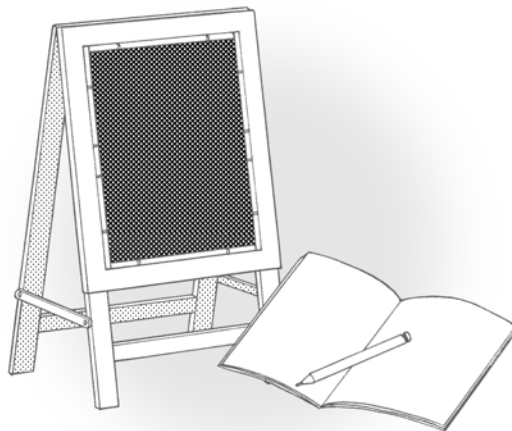
Nós arredondamos alguns números neste livro para facilitar a leitura, o que significa que alguns valores podem ser inconsistentes com os valores que você encontrará fazendo os cálculos por conta, mas eles ficarão próximos. Pedimos sua compreensão.

Gostaria de agradecer ao meu editor, Ohmsha, por me dar a oportunidade de escrever este livro. Também gostaria de agradecer a TREND-PRO, Co., Ltd. por transformar o meu manuscrito neste mangá, ao escritor de cenário re_akino e ao ilustrador Iroha Inoue. Por último, mas não menos importante, gostaria de agradecer ao Dr. Sakaori Fumitake da Faculdade de Relações Sociais Rikkyo University. Ele me deu conselhos imensuráveis, mais ainda do que já havia me dado quando eu estava preparando o meu livro anterior. Gostaria de expressar a minha profunda admiração.

Shin Takahashi
Setembro, 2005

1

**UM COPO REFRESCANTE DE
MATEMÁTICA**



CONSTRUINDO UMA BASE

O CHEFE FINALMENTE FOI EMBORA, E NÓS TEMOS QUE IR TAMBÉM!

UFA!

UHM, RISA, PODEMOS COMEÇAR AS AULAS HOJE À NOITE?

SÉRIO?

VOCÊ QUER COMEÇAR AGORA?

ACENO

EU NUNCA VI VOCÊ TÃO EMPOLGADA EM APRENDER! GERALMENTE VOCÊ DORME NAS AULAS!

É, BEM... É QUE...

DESCULPE...

EU NÃO QUERIA ENVERGONHÁ-LA...

REGRAS PARA EXPONENCIAIS E LOGARITMOS

1. REGRA DA POTÊNCIA

$(e^a)^b$ E $e^{a \times b}$ SÃO IGUAIS.



Vamos fazer uma tentativa. Nós vamos confirmar que $(e^a)^b$ e $e^{a \times b}$ são iguais quando $a = 2$ e $b = 3$.

$$(e^2)^3 = \underbrace{e^2 \times e^2 \times e^2}_3 = \underbrace{(e \times e) \times (e \times e) \times (e \times e)}_3 = \underbrace{e \times e \times e \times e \times e \times e}_6 = e^{2 \times 3}$$

Isso também significa que $(e^a)^b = e^{a \times b} = (e^b)^a$.



2. REGRA DO QUOCIENTE

$\frac{e^a}{e^b}$ E e^{a-b} SÃO IGUAIS.



Agora, vamos tentar o seguinte: Nós vamos confirmar que $\frac{e^a}{e^b}$ e e^{a-b} são iguais quando $a = 3$ e $b = 5$.

$$\frac{e^3}{e^5} = \frac{e \times e \times e}{e \times e \times e \times e \times e} = \frac{\cancel{e} \times \cancel{e} \times \cancel{e}}{e \times e \times \cancel{e} \times \cancel{e} \times \cancel{e}} = \frac{1}{e^2} = e^{-2} = e^{3-5}$$

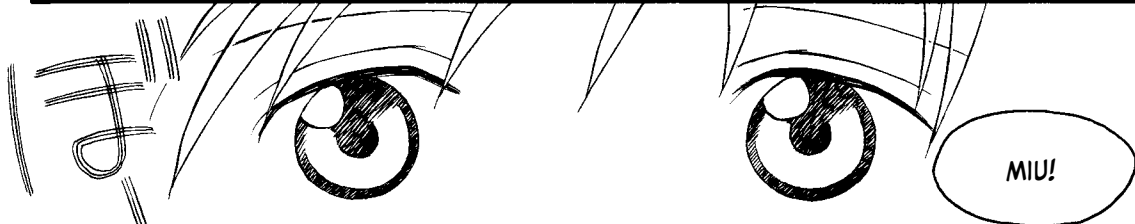
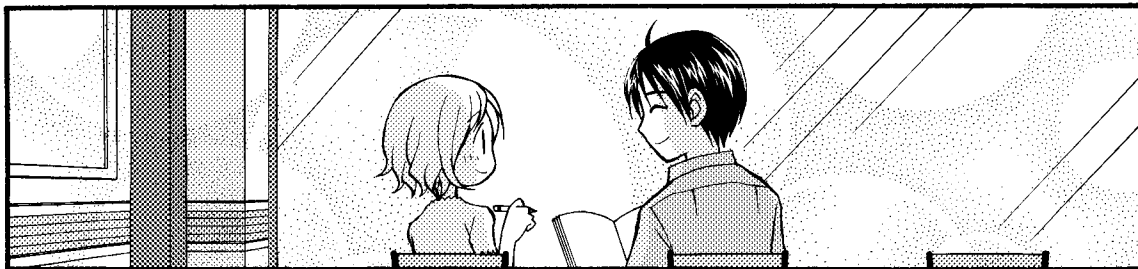
PRIMEIROS PASSOS

ISSO SIGNIFICA QUE...

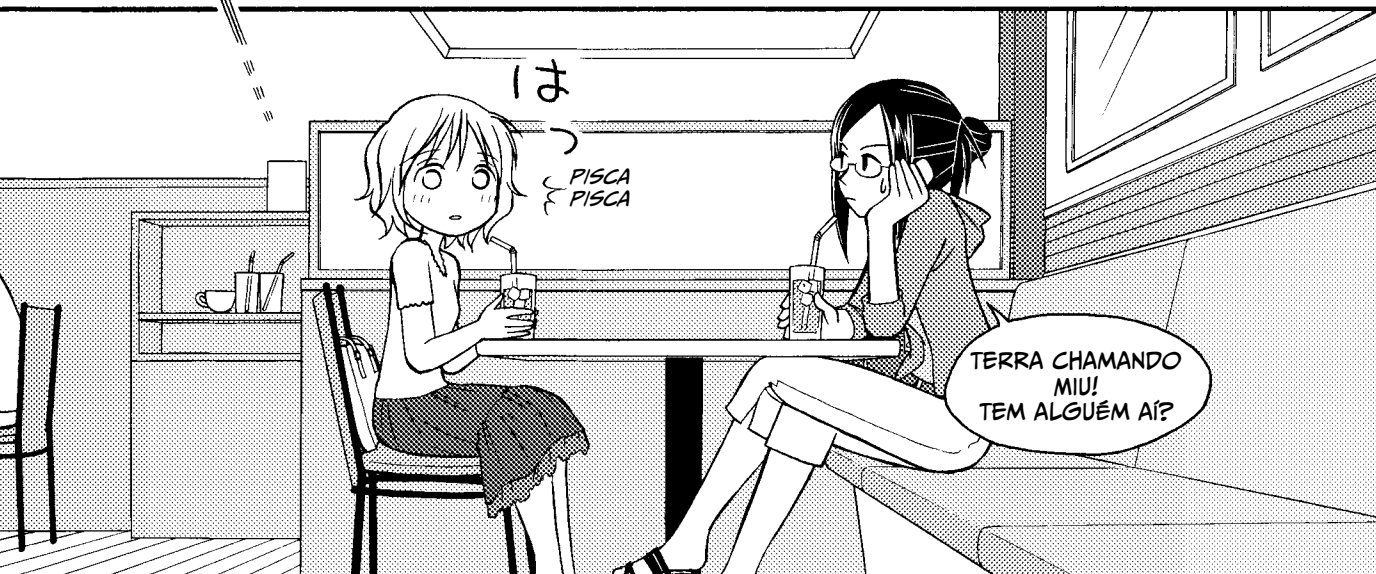
EXISTEM UMA LIGAÇÃO ENTRE AMBOS, CERTO?

ISSO MESMO!

ONDE VOCÊ APRENDEU TANTO SOBRE ANÁLISE DE REGRESSÃO, MIU?



MIU!



は

PISCA PISCA

TERRA CHAMANDO MIU! TEM ALGUÉM AÍ?

Passo 3

Derive S_e em relação a a e b , e iguale a derivada a zero. A derivada de $y = (ax + b)^{n-1}$ em relação a x é $\frac{dy}{dx} = n(ax + b)^{n-1} \times a$.

- Derive em relação a a .

$$\frac{dS_e}{da} = 2[77 - (29a + b)] \times (-29) + \dots + 2[84 - (30a + b)] \times (-30) = 0 \quad \text{①}$$

- Derive em relação a b .

$$\frac{dS_e}{db} = 2[77 - (29a + b)] \times (-1) + \dots + 2[84 - (30a + b)] \times (-1) = 0 \quad \text{②}$$

Passo 4

Rearranje as equações ① e ②.

Rearranje ①.

$$2[77 - (29a + b)] \times (-29) + \dots + 2[84 - (30a + b)] \times (-30) = 0$$

$$[77 - (29a + b)] \times (-29) + \dots + [84 - (30a + b)] \times (-30) = 0$$

DIVIDA AMBOS OS LADOS POR 2

$$29[(29a + b) - 77] + \dots + 30[(30a + b) - 84] = 0$$

$$(29 \times 29a + 29 \times b - 29 \times 77) + \dots + (30 \times 30a + 30 \times b - 30 \times 84) = 0$$

MULTIPLIQUE POR -1

$$(29^2 + \dots + 30^2)a + (29 + \dots + 30)b - (29 \times 77 + \dots + 30 \times 84) = 0$$

MULTIPLIQUE

③

ISOLE a E b

Rearranje ②.

$$2[77 - (29a + b)] \times (-1) + \dots + 2[84 - (30a + b)] \times (-1) = 0$$

$$[77 - (29a + b)] \times (-1) + \dots + [84 - (30a + b)] \times (-1) = 0$$

DIVIDA AMBOS OS LADOS POR 2

$$[(29a + b) - 77] + \dots + [(30a + b) - 84] = 0$$

MULTIPLIQUE POR -1

$$(29 + \dots + 30)a + \underbrace{b + \dots + b}_{14} - (77 + \dots + 84) = 0$$

ISOLE a E b

$$(29 + \dots + 30)a + 14b - (77 + \dots + 84) = 0$$

$$14b = (77 + \dots + 84) - (29 + \dots + 30)a$$

SUBTRAIA $14b$ DE AMBOS OS LADOS E MULTIPLIQUE POR -1

$$b = \frac{77 + \dots + 84}{14} - \frac{29 + \dots + 30}{14}a$$

$$b = \bar{y} - \bar{x}a$$

ISOLE b NO LADO ESQUERDO DA EQUAÇÃO

④ $b = \bar{y} - \bar{x}a$

⑤

OS COMPONENTES EM ④ SÃO AS MÉDIAS DE y E x

FAZENDO PREVISÕES COM DIVERSAS VARIÁVEIS

OBRIGADA
POR TRAZER
OS DADOS.

NÃO HÁ DE QUÊ.
VOCÊ QUE ESTÁ
SENDO LEGAL EM
AJUDAR A SUA AMIGA.

BEM...

... EU TENHO
MEUS MOTIVOS.

RISA...

UFA

RESPIRA

AH, ELA
CHEGOU!

AQUI!

DESCULPE
PELO ATRASO!

A MINHA AULA
TERMINOU MAIS
TARDE, TIVE QUE
CORRER ATÉ
AQUI.

TUDO BEM.
NÓS ACABAMOS
DE CHEGAR.

ANALISANDO A POPULAÇÃO COM A ANÁLISE DE REGRESSÃO MÚLTIPLA

Vamos revisar o procedimento utilizado na análise de regressão múltipla, mostrado na página 112.

1. Desenhe um gráfico de dispersão de cada variável preditora e da variável resultado para analisar se elas estão relacionadas.
2. Calcule a equação de regressão múltipla.
3. Examine a acurácia da equação de regressão múltipla.
4. Faça a análise do teste de variância (ANOVA).
5. Calcule o intervalo de confiança da população.
6. Faça uma previsão!

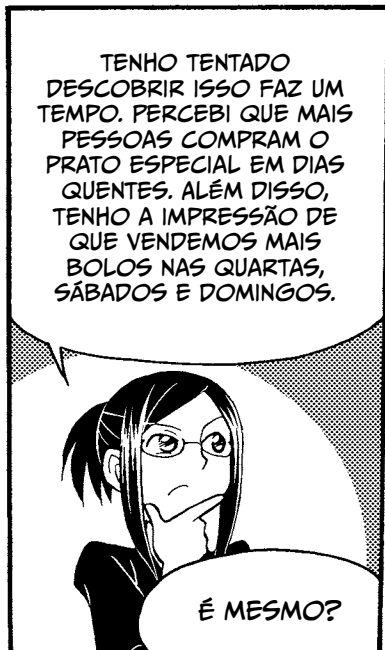
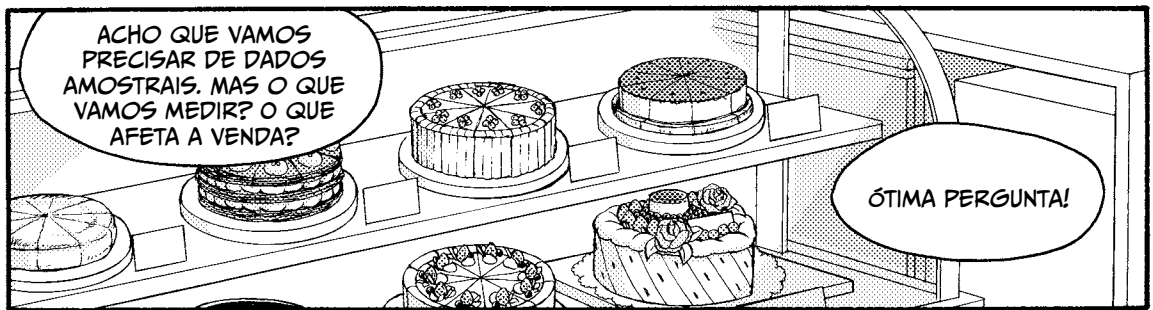
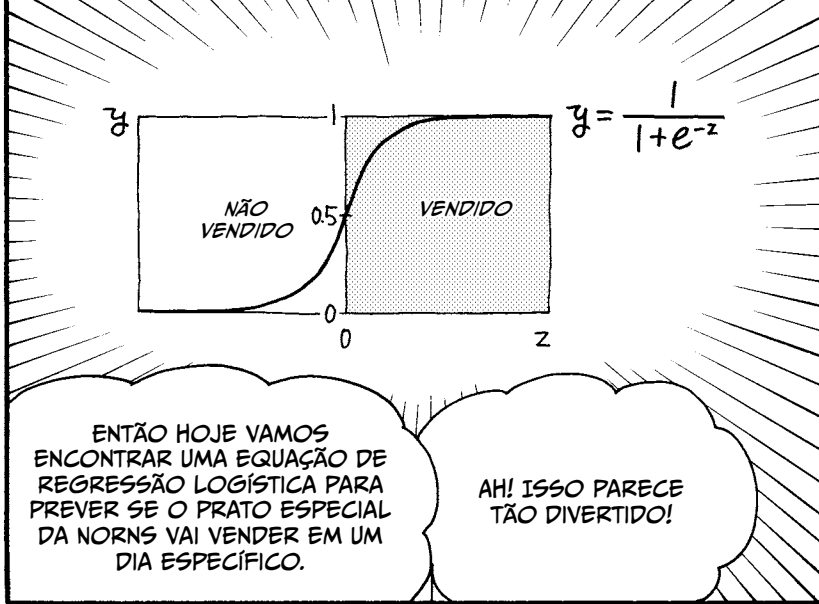
Assim como no Capítulo 2, falamos sobre todos os passos como se eles fossem obrigatórios. Porém, na realidade os passos 4 e 5 podem ser ignorados na análise de determinados conjuntos de dados.

As Padarias Kazami são, atualmente, compostas de 10 filiais, e entre essas filiais apenas uma (a padaria Yumenooka) possui área de 10 tsubo¹ e está a 80 metros da estação mais próxima. Porém, Risa calculou o intervalo de confiança para uma população de padarias que possuem área de 10 tsubo e que estão a 80 metros da estação mais próxima. Por que ela fez isso?

Bem, é possível que as Padarias Kazami inaugurem outra padaria com 10 tsubo de área e que também estejam a 80 metros da estação de trem mais próxima. Se as Padarias Kazami continuarem crescendo, poderá haver dezenas de padarias que se encaixarão nessa descrição. Quando Risa fez essa análise, ela presumiu que mais padarias de 10 tsubo a 80 metros de uma estação de trem seriam inauguradas no futuro.

A utilidade dessa hipótese é discutida. A padaria Yumenooka vende mais do que qualquer outra padaria, o que pode fazer com que a família Kazami abra mais padarias iguais a ela. Porém, a próxima padaria que será inaugurada, a Padaria Isebashi, terá 10 tsubo de área e estará a 110 metros da estação mais próxima. Na verdade, provavelmente não seria necessário analisar uma população tão específica de padarias, pois Risa poderia ter ignorado o cálculo do R^2 ajustado para realizar a previsão. Porém, ela o calculou para demonstrar a Miu todos os passos.

1. Lembre-se de que 1 tsubo é equivalente a aproximadamente 3,34 metros quadrados.



FAREMOS O TESTE DA RAZÃO DE VEROSSIMILHANÇAS. ESSE TESTE NOS PERMITE EXAMINAR TODOS OS COEFICIENTES DE UMA VEZ SÓ E OBTER A RELAÇÃO ENTRE ELES.



OS PASSOS DO TESTE DA RAZÃO DE VEROSSIMILHANÇAS

Passo 1	Defina a população.	Todos os dias em que o prato especial da Norns é vendido, comparando quartas, sábados e domingos com os dias restantes, para cada temperatura.
Passo 2	Defina uma hipótese nula e uma hipótese alternativa.	Hipótese nula é $A_1 = 0$ e $A_2 = 0$. Hipótese alternativa $A_1 \neq 0$ ou $A_2 \neq 0$.
Passo 3	Escolha qual teste de hipóteses conduzir.	Vamos calcular o teste da razão de verossimilhanças.
Passo 4	Escolha o nível de significância	Vamos usar um nível de significância de 0,05.
Passo 5	Calcule a estatística de teste a partir dos dados amostrais.	A estatística de teste é: $2[L - n_1 \log_e(n_1) - n_0 \log_e(n_0) + (n_1 + n_0) \log_e(n_1 + n_0)]$ Ao colocarmos nossos dados, obtemos: $2[-8.9010 - 8 \log_e 8 - 13 \log_e 13 + (8 + 13) \log_e (8 + 13)] = 10.1$ A estatística de teste segue uma distribuição qui-quadrado com 2 graus de liberdade (o número de variáveis preditoras), caso a hipótese nula seja verdadeira.
Passo 6	Determine se o p -valor para a estatística de teste obtida no Passo 5 é menor do que o nível de significância.	O nível de significância é 0,05. O valor da estatística de teste é 10,1, então o p -valor é 0,006. Finalmente, $0,006 < 0,05$.*
Passo 7	Decida se você pode rejeitar a hipótese nula.	Uma vez que o p -valor é menor do que o nível de significância, rejeitamos a hipótese nula.

* Os passos para obter o p -valor em uma distribuição qui-quadrado estão explicados na página 205.