Modelo de regressão linear
Teoria e aplicações

Fernando Lucambio Pérez
Fevereiro de 2024

Conteúdo

Prefácio **v**

1 Introdução **1**
1.1 Modelo de regressão linear . . . 2
1.2 Modelos lineares . . . 7
1.2.1 Propriedades dos modelos lineares . . . 8
1.3 Distribuição normal . . . 10
1.3.1 Distribuição normal multivariada . . . 13
1.4 Exemplos . . . 18
1.4.1 Quanto custa ser deputado? . . . 18
1.4.2 Peso dos mexilhões . . . 20
1.4.3 Biomassa do capim-marinho . . . 21
1.5 Exercícios . . . 23

2 Estimação e testes de hipóteses **27**
2.1 Estimação . . . 28
2.1.1 Estimador de mínimos quadrados . . . 28
2.1.2 Estimador de máxima verossimilhança . . . 30
2.1.3 Estimadores do modelo de regressão . . . 31
2.1.4 Propriedades dos estimadores . . . 38
2.2 Estimação por intervalos de confiança . . . 41
2.2.1 Formas quadráticas . . . 42
2.2.2 Predição em modelos de regressão . . . 44
2.3 Teste de hipóteses . . . 52
2.3.1 Teste de hipóteses paramétrico . . . 52
2.3.2 Testes acerca dos parâmetros da regressão . . . 54
2.3.3 Análise de Variância da regressão . . . 57
2.3.4 Verossimilhança perfilada . . . 61
2.4 Exercícios . . . 68

3 Medidas de diagnóstico **73**
3.1 Resíduos ordinários . . . 73
3.1.1 Matriz de predição . . . 74
3.2 Resíduos independentes . . . 80
3.2.1 Testes de normalidade . . . 89
3.3 Métodos gráficos . . . 95
3.3.1 Gráfico segundo a ordem de observação . . . 96
3.3.2 Gráfico de probabilidade normal . . . 97

3.3.3 Valores ajustados . . . 97
3.3.4 Pontos influentes . . . 97
3.3.5 Distância de Cook . . . 98
3.4 Exemplo: Biomassa do capim-marinho . . . 100
3.5 Exercícios . . . 105

4 Qualidade do ajuste **109**
4.1 Coeficiente de correlação ρ . . . 109
4.1.1 Coeficiente de correlação amostral . . . 112
4.1.2 Momentos do coeficiente de correlação amostral . . . 113
4.1.3 Coeficiente de correlação parcial . . . 114
4.2 Coeficiente de determinação R^2 . . . 116
4.2.1 Propriedades do coeficiente de determinação . . . 120
4.2.2 Coeficientes de determinação ajustados . . . 128
4.3 Importância relativa das covariáveis . . . 130
4.3.1 R^2 parcial . . . 131
4.3.2 Diferentes métricas de importância relativa . . . 135
4.4 Transformação de variáveis . . . 137
4.4.1 Teste de linearidade . . . 138
4.4.2 Transformações das variáveis explicativas . . . 139
4.4.3 Transformações na variável resposta . . . 145
4.4.4 Modelos de retas paralelas . . . 146
4.4.5 Regressão segmentada . . . 146
4.5 Exercícios . . . 149

5 Seleção de modelos **153**
5.1 Critérios de seleção . . . 154
5.2 Critério de Informação de Akaike . . . 155
5.2.1 Informação de Kullback-Liebler . . . 155
5.2.2 O Critério de Informação de Akaike . . . 156
5.3 Outros critérios de seleção de modelos . . . 158
5.3.1 C_p de Mallow . . . 158
5.3.2 Estatística PRESS . . . 159
5.4 Métodos de seleção de modelos . . . 160
5.4.1 forward . . . 160
5.4.2 backguard . . . 161
5.4.3 Exemplo: Quanto custa ser um deputado? . . . 161
5.5 Critério de Informação Bootstrap . . . 161
5.5.1 Método Bootstrap . . . 162
5.5.2 O Critério de Informação Bootstrap (EIC) . . . 162
5.5.3 Exemplo: peso dos mexilhões . . . 163
5.6 Exercícios . . . 164

Referências Bibliográficas **165**

Prefácio

Pensei, desde o começo deste projeto, criar um material de consulta para estudantes de graduação em estatística e ciências de dados em particular e, de maneira geral, para qualquer pessoa que tenha conhecimentos mais do que a nível básico de estatística. Queremos dizer que este material foi pensado para ser apropriado a compreender do assunto, como também para aprofundar os conhecimentos em modelos de regressão, em especial os modelos de regressão lineares com resposta gaussiana ou normal. Consideramos então adequado este material para até níveis de especialização e mestrado em outras áreas do conhecimento, desde que o interessado entenda elementos de álgebra, cálculo, probabilidade e inferência estatística.

O livro pode ser dividido em duas grandes partes. Na primeira dela, formada pelos primeiros dois capítulos, abordamos a teoria básica de modelos de regressão lineares. Quem domine este conteúdo pode considerar que conhece os fundamentos desta teoria e está em condições de aplicar estes modelos em diversas áreas.

A segunda parte do livro, os restantes capítulos, dedica-se a aprofundar detalhes e extensões dos modelos lineares que tentam dilucidar problemas que podem aparecer quando aplicada esta modelagem a dados reais.

O auxilio computacional é baseado na linguagem de programação **R**. Exemplos desenvolvidos no texto são disponibilizados no endereço `http://leg.ufpr.br/~lucambio/Linear/`, neste endereço podem ser encontradas resoluções dos exemplos mostrados no texto, todos resolvidos utilizando esta linguagem. Para melhor aproveitamento destes exemplos o leitor deve ter conhecimentos básicos desta linguagem. Este texto foi redigido utilizando LaTeX.

Este livro ha sido resultado de um trabalho intenso durante anos. Comecei a escrever durante a estadia de pós-doutorado na Universidade Federal de Pernambuco, Recife para o qual contei com ajuda do programa de apoio REUNI em 2012, graças ao qual pude iniciar e estruturar o trabalho. Percebe-se que somente terminei de escrever e corrigir o texto vários anos depois.

Agradeço em particular à minha esposa e filho, pelo apoio e paciência durante todo o trabalho.

Curitiba
Fevereiro de 2024.

Capítulo 1

Introdução

O termo "regressão" foi utilizado pela primeira vez por Francis Galton[1] no século XIX para descrever um fenômeno biológico. Ele foi o primeiro a aplicar métodos estatísticos para o estudo das diferenças de heranças humanas e introduziu a utilização de questionários em pesquisas para coletar dados sobre as comunidades humanas.

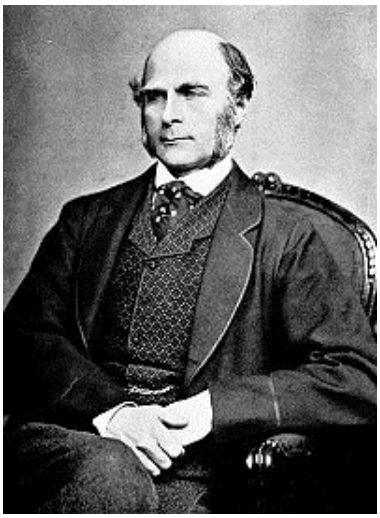

Figura 1.1: Francis Galton

O fenômeno em questão por ele estudado foi que as alturas dos descendentes de ancestrais altos tendem a regredir em direção a média normal, um fenômeno também conhecido como regressão para a média. Para Galton, a regressão somente tinha esse significado biológico, mas seu trabalho foi posteriormente prorrogado por Udny Yule[2] e Karl Pearson[3] para um contexto mais geral de estatística. Nos trabalhos de Yule e Pearson, a distribuição conjunta da resposta e das variáveis explicativas é assumido como sendo gaussiana. Esta hipótese foi enfraquecida por Ronald Fisher[4], em seus trabalhos de 1922 e 1925. Fisher assumiu que a distribuição condicional da variável resposta é gaussiana, mas a distribuição conjunta não precisa ser. Nesse sentido a suposição de Fisher é mais próxima à formulação original de Gauss[5] em 1821.

Nas últimas décadas, novos métodos foram desenvolvidos para modelos de regressão, como modelos lineares generalizados, modelos de regressão simétricos a regressão envolvendo respostas correlacionadas e outros.

Neste livro abordamos de maneira moderna a teoria clássica de modelos de regressão lineares. Nosso objetivo é que o leitor possa encontrar neste texto os mais modernos desenvolvimentos destes modelos e assim entender melhor todos os esforços de estende-los a situações mais complexas. Os

[1]Francis Galton (1822-1911) foi um antropólogo, meteorologista, matemático e estatístico inglês.

[2]George Udny Yule (1871-1951) foi um estatístico britânico.

[3]Karl Pearson (1857 - 1936) foi um grande contribuidor para o desenvolvimento da estatística como uma disciplina científica séria e independente. Foi o fundador do Departamento de Estatística Aplicada na University College London em 1911; foi o primeiro departamento universitário dedicado à estatística em todo o mundo.

[4]Ronald Aylmer Fisher (1890 - 1962) foi um estatístico, biólogo evolutivo e geneticista inglês. Foi descrito como "um gênio que criou praticamente sozinho as fundações para a moderna ciência estatística".

[5]Johann Carl Friedrich Gauss (1777 - 1855) foi um matemático, astrônomo e físico alemão que contribuiu muito em diversas áreas da ciência, dentre elas a teoria dos números, estatística, análise matemática, geometria diferencial, geodésia, geofísica, eletroestática, astronomia e óptica.

primeiros dois capítulos do livro são dedicados a explanação dos princípios teóricos; os três últimos capítulos são dedicados a identificar e solucionar problemas que eventualmente podem acontecer e invalidar o modelo.

1.1 Modelo de regressão linear

A análise de regressão é uma das ferramentas mais comumente utilizadas em diferentes situações, sejam problemas na biologia, engenharia, econometria, ciências humanas, sociais e outras. Começamos então nossa discussão no chamado modelo de regressão linear simple.

Agora, uma pergunta: o que é análise de regressão? a descrição e quantificação da relação entre uma variável, chamada de regressora ou dependente, e uma ou mais outras variáveis, chamadas de independentes ou explicativas, é uma análise de regressão.

Exemplo 1.1. *Considere um pesquisador observando a velocidade de uma partícula que se move ao longo de uma linha. Ele toma observações a dados momentos de tempo $t_1, t_2, \cdots, t_n$. Seja β_0 a velocidade inicial da partícula e β_1 a aceleração; então a velocidade da partícula V no tempo t é dada por $V = \beta_0 + \beta_1 t + \epsilon$, onde ϵ é uma variável aleatória não observável, como um erro na medição. Na prática o observador desconhece os valores de β_0 e β_1 e tem que utilizar as observações $v_1, v_2, \cdots, v_n$ da velocidade da partícula nos instantes $t_1, t_2, \cdots, t_n$, respectivamente, para obter alguma informação sobre os parâmetros desconhecidos β_0 e β_1.*

Neste exemplo podemos perceber detalhes do modelo probabilístico de regressão, descrevemos a variável aleatória V como uma função determinística do tempo mais um termo aleatório. Ainda esclarecemos que o termo determinístico é nosso interesse fundamental, indicando o procedimento de coleta de informações ou estimação dos parâmetros que o definem.

A literatura deste tema é extensa, livros clássicos deste tema são Searle (1971), Rao (1973), Graybill (1976), Seber (1977), Wetherill (1986), Sen & Srivastava (1990), Cook & Weisberg (1994), Draper & Smith (1998), Scheffé (1959), Montgomery, Peck & Vining (2001), Weisberg (2005) e outros. Devemos mencionar também que existe muita literatura abordando temas específicos e aplicações, mais recentemente surgiu uma literatura que ensina como utilizar em situações práticas. Nosso objetivo é geral, queremos mostrar a teoria dos modelos de regressão e resolver problemas práticos utilizando a linguagem de programação **R**[6] (Team, 2024).

Escritos em português mencionamos os livros de Cordeiro & Paula (1989) e Charnet, Bonvino, Freire & Charnet (2008) de professores do IMPA e UNICAMP, respectivamente. Uma referência mais abrangente é o livro "Modelos de Regressâo com apoio computacional" do Prof. Gilberto A. Paula do IME-USP disponível na página pessoal do professor no Instituto de Matemática e Estatística, USP.

Definição 1.1 (Modelo de Regressão Linear Simples)**.** *O modelo clássico de regressão linear simples é escrito como*

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

onde $\beta_0 + \beta_1 x_i$ $(i = 1, \ldots, n)$ é a parte determinística e $\epsilon_1, \cdots, \epsilon_n$ o erro aleatório. Neste modelo a variável aleatória Y representa a resposta, a variável determinística x representa a variável preditora, explicativa ou covariável e β_0 e β_1 os parâmetros da regressão desconhecidos.

[6] R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Um modelo de regressão, novamente, descreve a variável de interesse, que chamaremos variável resposta, como a soma de uma parte considerada determinística e uma parte aleatória, sendo a parte determinística uma função de uma ou várias variáveis explicativas, também chamadas de regressoras, e a parte aleatória o erro. Constituem uma das metodologias mais utilizadas na análise de dados devido a sua estrutura simples, aplicabilidade e interpretabilidade.

Exemplo 1.2. *Se estivéssemos fazendo um estudo de tráfego em uma determinada cidade, por exemplo, poderíamos querer saber quais seriam os limites de velocidade segura para permitir em diferentes ruas. Nesse contexto, poderíamos precisar de saber em que distância um automóvel poderia ser parado quando viajávamos em velocidades diferentes, de modo que comparando essa distância com a largura das diferentes ruas e o comprimento da visão nas interseções poderíamos julgar quão rápido os carros poderiam ser capazes de viajar sem risco de colisões nos cruzamentos da rua.*

Uma maneira de determinar qual é a relação entre velocidade e distância de parada seria fazer um número de testes em diferentes partes da cidade, usando diferentes tipos de máquinas e diferentes motoristas. Vamos supor que, como resultado de tal série de testes, obtivemos a série de observações mostradas na Tabela 1.1.

Velocidade quando	4	7	17	14	12	11	20	15	17	13	15	19	10
o sinal é dado	18	22	18	8	4	12	20	23	18	12	16	18	19
(milhas por hora)	24	14	12	9	10	15	24	25	20	19	13	10	7
	16	14	20	24	24	17	13	11	13	14	20		
Distância percorrida após	2	4	50	36	20	28	48	54	40	34	26	68	26
o sinal antes de parar	56	66	84	16	10	14	56	54	76	24	32	42	46
(pés)	93	26	28	10	34	20	70	85	64	36	26	18	22
	40	60	52	120	92	32	34	17	46	80	32		

Tabela 1.1: Relação entre a velocidade do automóvel e a distância para parar após o sinal, como mostrado pelas 50 observações individuais apresentadas aqui. Estas observações da distância percorrida foram feitas antes que os freios de 4 rodas fossem comuns.

Com os comandos **R** a seguir construímos o gráfico na Figura 1.2, nele mostramos a relação entre a velocidade e a distância percorrida. Pode-se observar que as variáveis velocidade e distância apresentadas na Tabela 1.1 estão em unidades alheias a nós, por esse motivo fizemos a transformação adequada: 1 milha = 1,60934 quilômetros e 1 pé = 0,3048 metros.

```
> veloc = c(4,7,17,14,12,11,20,15,17,13,15,19,10,18,22,18,8,4,12,20,23,18,12,16,18,
            19,24,14,12,9,10,15,24,25,20,19,13,10,7,16,14,20,24,24,17,13,11,13,14,20)*1.60934
> distan = c(2,4,50,36,20,28,48,54,40,34,26,68,26,56,66,84,16,10,14,56,54,76,24,32,42,
            46,93,26,28,10,34,20,70,85,64,36,26,18,22,40,60,52,120,92,32,34,17,46,80,32)*0.3048
> par(mar=c(5,5,1,1), pch=19, cex.axis=0.6)
> plot(veloc, distan, xlab="Velocidade em quilômetros por hora \n  quando o sinal é dado",
       ylab=" Distância percorrida após o sinal \n antes de parar, em metros")
```

Esta antiquíssimo exemplo aparece no livro "Methods of Correlation Analysis" de Mordecai Ezekiel escrito 1929 e publicado pela John Wiley & Sons em 1930. As unidades de medição não são utilizadas por nós, por esse motivo na Figura 1.2, transformamos a velocidade quando o sinal é dado em quilômetros por hora e a distância percorrida em metros.

É evidente, a partir da tabela, que existem grandes variações na distância que carros diferentes ou diferentes condutores precisam para parar, mesmo quando viajam na mesma velocidade. Isso é

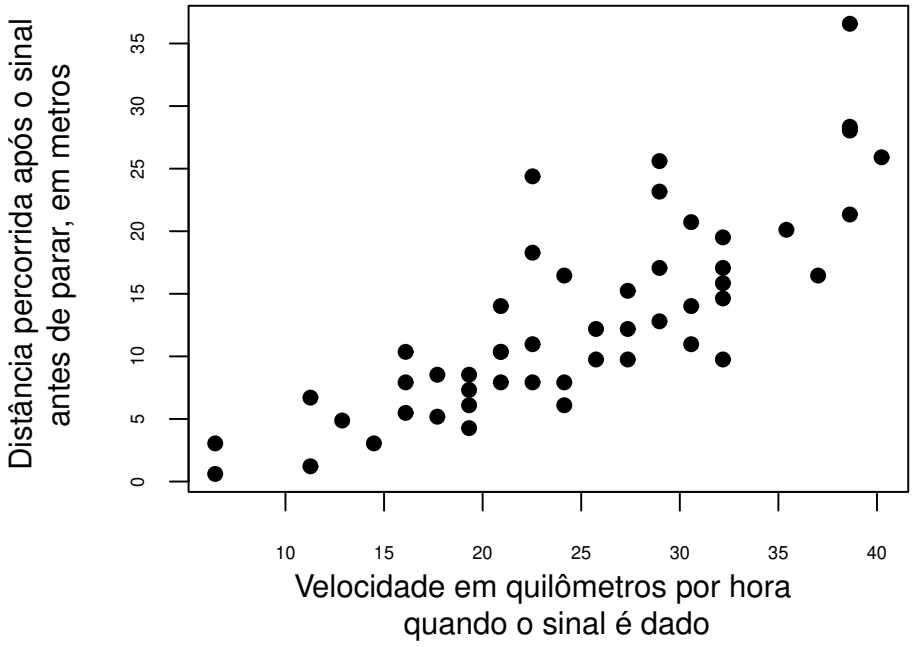

Figura 1.2: Gráfico da relação entre a velocidade do automóvel com a distância que leva para parar, como mostrado por observações individuais.

mostrado ainda mais claramente quando fazemos um gráfico de pontos dos dados exatamente da mesma maneira mostrada na Figura 1.2.

Não há dificuldade particular em entender por que a relação não é mais definida. Os dados representam uma grande variedade de elementos diferentes: carros com freios em duas rodas e carros com freios nas quatro rodas; carros com freios em ajuste e carros com freios bem desgastados; carros quase vazios e carros fortemente carregados; carros com pneus quase murchos e carros com pneus com alta pressão.

Além disso, os motoristas diferem: alguns são condutores experientes, alguns inexperientes; alguns fortes e alguns incapazes de pressionar os freios completamente para baixo. alguns com quase instantânea reação ao nosso sinal para parar, alguns com hesitação ou atraso na resposta; alguns brilhantes e bem acordados, outros cansados e desatentos; alguns calmos e firmes, outros nervosos e erráticos. Finalmente, as condições dos testes podem ser diferentes: algumas no pavimento de concreto, outras no asfalto; alguns em aclive, alguns em declive.

Existem diversas maneiras diferentes pelas quais podemos decidir exatamente o que essas observações variadas mostraram. Uma maneira seria dividir os dados de modo que o efeito de alguns dos diferentes fatores mencionados fossem removidos dos resultados. Assim, se separássemos as observações em diferentes grupos de acordo com a marca do carro e, em seguida, reportássemos cada um desses grupos de acordo com o modelo ou o ano de fabricação, a relação entre velocidade e distância para qualquer grupo individual não seria mais afetada pelas diferenças de equipamento de frenagem.

A maioria dos fatores restantes, no entanto, ainda estaria presente para afetar os resultados, de modo que, mesmo dentro de cada subdivisão, os registros ainda mostrariam grande diversidade na relação. Somente se continuássemos o processo de subdivisão de nossa amostra até chegarmos a observações sucessivas de um único carro operado por um único motorista no mesmo local. Dife-

renças na prontidão com que o motorista respondeu ao sinal, na precisão com que a velocidade no momento de dar o sinal foi observada e, possivelmente, na força com a qual o motorista aplicava seus freios, tudo isso poderia influenciar o resultado que, mesmo assim, os resultados poderiam ser menos consistentes, a curva menos definida definitivamente, do que em uma série de experimentos de laboratório, onde todas as variáveis externas importantes poderiam ser definitivamente controladas e, assim, impedidas de afetar o resultado obtido.

Úteis como eles são, o modelo de regressão linear simples é limitado. Geralmente, pretende-se estudar a influência de várias outras variáveis sobre Y. Em situações nas quais exista mais do que uma variável para explicar a resposta temos o chamado de modelo de regressão múltipla.

Definição 1.2 (Modelo de Regressão Linear Múltipla)**.** *O modelo clássico de regressão linear múltipla é escrito como*

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

onde $\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$ $(i = 1, \cdots, n)$ é a parte determinística e $\epsilon_1, \cdots, \epsilon_n$ o erro aleatório. Neste modelo a variável aleatória Y representa a resposta, as variáveis determinísticas $x_1, \cdots, x_p$ representam as variáveis preditoras e $\beta_0, \beta_1, \cdots, \beta_p$ os parâmetros da regressão desconhecidos. Diz-se também que Y é a variável dependente e $x_1, \cdots, x_p$ as regressoras do modelo.

Em muitos tipos de problemas as diferenças em uma variável podem ser devidas às influências de várias outras variáveis, todas agindo ao mesmo tempo. Assim, as diferenças nos rendimentos do milho de ano para ano são o resultado combinado das diferenças de precipitação, temperatura, ventos e insolação, mês a mês ou mesmo semana a semana durante a estação de crescimento. Os prêmios ou descontos aos quais diferentes lotes de trigo são vendidos no mesmo dia variam de acordo com o teor de proteína, o peso por alqueire, a quantidade de docagem ou matéria estranha e o teor de umidade. A velocidade com que um motorista reage a uma situação perigosa pode variar com sua agudeza de visão, sua velocidade de reação nervosa, sua inteligência e sua familiaridade com tais situações. O preço pelo qual o açúcar é vendido no atacado pode depender da produção daquela safra, do excesso da temporada anterior, do nível geral de preços e da prosperidade dos consumidores. O peso de uma criança varia de acordo com sua idade, altura e sexo. O volume de um determinado peso de gás varia com a temperatura e a pressão barométrica. Dentre outras muitas situações. Vejamos o seguinte exemplo de uma possível relação de regressão linear múltipla, na Seção 1.4 consideramos situações mais complexas as quais serão tratadas no decorrer deste livro.

Exemplo 1.3. *O problema das relações múltiplas é ilustrado pelos dados da Tabela 1.2 os quais foram coletados em 20 fazendas em uma área, com áreas cultivadas variadas, vacas leiteiras e como resposta a renda obtida por hectare. Para determinar a partir desses registros qual renda pode ser esperada, em média, com um determinado tamanho de fazenda e com um dado número de vacas é necessário estimar os efeitos das diferenças no número de hectares sobre a renda e também os efeitos das diferenças no número de vacas na renda.*

As linhas de comandos **R** nos permitem obter os correlogramas ou gráficos de dispersão para duas variáveis.

```
> Tamanho = c(24,89,73,32,48,40,69,44,65,93,28,48,97,65,36,44,89,44,65,32)
> Vacas = c(18,0,14,6,1,9,6,12,7,2,17,15,7,0,12,16,2,6,12,15)
> Renda = c(960,830,1260,610,590,900,820,880,860,760,1020,1080,960,700,800,1130,760,740,980,800)
> library(ggplot2)
> par(mar=c(5,4,1,1),pch=19,cex.axis=0.6)
> qplot(Tamanho, Renda, ylab='Renda por hectare', xlab="Tamanho da fazendo em hectare")
> qplot(Vacas, Renda, ylab='Renda por hectare', xlab="Número de vacas")
```

Tamanho da fazenda	24	89	73	32	48	40	69	44	65	93	28	48
em hectares	97	65	36	44	89	44	65	32				
Números de vacas	18	0	14	6	1	9	6	12	7	2	17	15
	7	0	12	16	2	6	12	15				
Renda por hectare	960	830	1260	610	590	900	820	880	860	760	1020	1080
	960	700	800	1130	760	740	980	800				

Tabela 1.2: Hectares, número de vacas e renda por hectare para 20 fazendas.

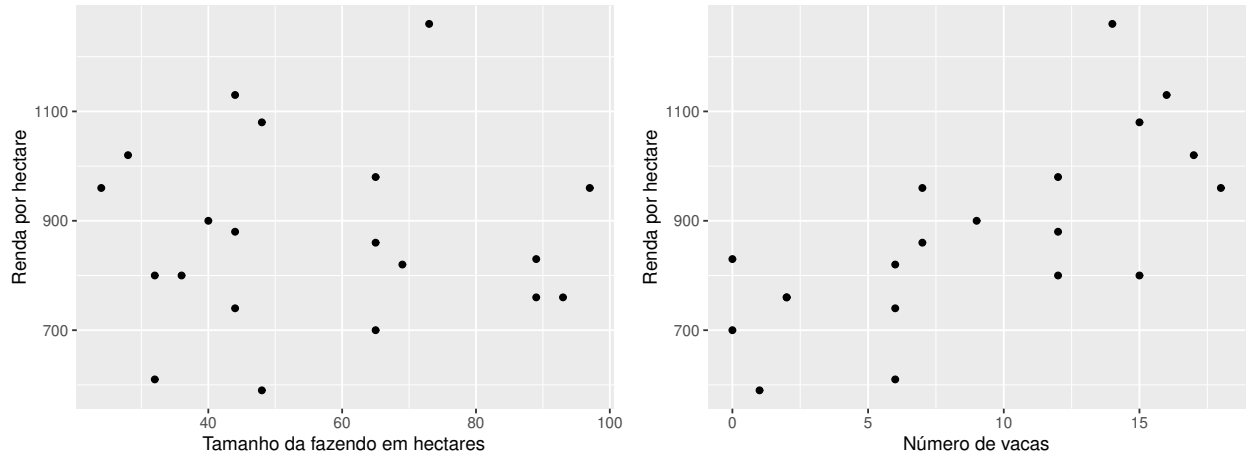

Figura 1.3: Gráfico descritivo da possível relação entre o tamanho em hectares da fazenda e o número de vacas com a renda obtida. Este gráfico é chamado de gráfico de dispersão.

A partir desses dados parece que, tanto o tamanho da fazenda quanto o tamanho do rebanho leiteiro influenciam a renda agrícola, a julgar pelos pontos que mostram a relação da renda com os hectares e a renda com o número de vacas. Parece, a partir desses gráficos, que pode haver uma leve tendência de as fazendas com maior área plantada terem rendimentos maiores e uma tendência bastante marcada para as fazendas, com o maior número de vacas a ter maiores rendimentos.

A fim de interpretar o resultado de um estudo que envolve assumir um modelo linear para o resultado de uma experiência, é importante compreender os princípios subjacentes a sua utilização. Para esclarecer estas hipóteses vamos derivar o modelo linear por uma cadeia de simplificações de um início geral. Estamos interessados em estudar o comportamento da média μ de alguns y quantitativo variável, que é, em parte, determinadas aleatoriamente e em parte em função de um conjunto de outras variáveis $x_1, \cdots, x_n$. Nós controlamos o nonrandom variáveis $x_1, \cdots, x_n$ e está interessado em média, em função destas variáveis.

A comparação simples, por si só, no entanto, não é suficiente para dizer exatamente como os rendimentos mudam com os hectares e com o número de vacas. Isso é porque aqui há uma relação marcada entre o tamanho das fazendas e o número de vacas, como ilustrado no gráfico a direita.

Há uma tendência definida para as fazendas maiores terem rebanhos leiteiros menores. Como resultado, a diferença de receitas observadas que parece ser devido diretamente a diferenças em áreas, pode ser devido em parte às diferenças nos tamanhos dos rebanhos leiteiros nas fazendas com áreas diferentes em colheitas.

Se, por outro lado, deveríamos tentar determinar o quanto a renda varia com as diferenças no número de vacas, classificando os registros em relação ao número de vacas e calculando a média dos rendimentos, poderíamos assegurar os resultados mostrados na tabela.

Mesmo que a renda seja maior nas fazendas com muitas vacas, os gráficos não indicam quanto disso pode ser creditado às vacas e quanto a outros fatores. É a partir da mesa que, à medida que o número de vacas aumenta, o número de acres diminui; e as diferenças de renda associadas a mudanças no número de vacas, em um ou dois hectares, ou em parte com as duas?

1.2 Modelos lineares

Tratamos aqui das propriedades probabilísticas dos chamados modelos lineares para as médias e certas estruturas específicas para a matriz de variâncias e covariâncias das observações.

Referências importantes são os livros de Searle (1971) e Graybill (1976). Apresentamos a seguir a definição probabilística de modelo linear.

Definição 1.3 (Modelo Linear)**.** *Seja $Y = (Y_1, Y_2, \cdots, Y_n)^\top$ um vetor aleatório de componentes independentes e seja $X = (x_{ij})$ uma matriz $n \times p + 1$ $(p < n)$, de constantes conhecidas x_{ij}, $i = 1, \cdots, n$; $j = 0, \cdots, p$. Dizemos que a distribuição de Y satisfaz um modelo linear se*

$$\mathrm{E}(Y) = X\beta, \tag{1.1}$$

onde $\beta = (\beta_0, \beta_1, \cdots, \beta_p)^\top$ é um vetor de parâmetros desconhecidos de dimensão $p + 1$.

Podemos perceber que um modelo linear é uma representação para a média da variável resposta. Convém escrever o modelo linear de maneira matricial em termos das variáveis aleatórias envolvidas como

$$Y = X\beta + \epsilon, \tag{1.2}$$

onde $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^\top$ é um vetor não observável conhecido como vetor de erros aleatórios, tal que $\mathrm{E}(\epsilon_i) = 0$, e $\mathrm{Var}(\epsilon_i) = \sigma^2$ para $i = 1, \cdots, n$. Podemos então dizer que, se as variáveis aleatórias $Y_1, Y_2, \cdots, Y_n$ satisfazem um modelo linear, então elas podem ser escritas como

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \epsilon_i, \tag{1.3}$$

onde $\mathrm{E}(\epsilon_i) = 0$ e $i = 1, \cdots, n$.

As situações apresentadas nos exemplos 1.1 e 1.3 correspondem os casos particulares do modelo linear em (1.3) onde $p = 1$ e $p = 2$, respectivamente. Quando o vetor de parâmetros da regressão seja composto unicamente por β_0 e β_1 o modelo em (1.3) é chamado de modelo linear de regressão simples, caso contrário é chamado de modelo de regressão linear múltipla. O coeficiente β_0 costuma chamar-se de intercepto e recomenda-se sua inclusão nos modelos de regressão a menos que o problema em questão exija o contrário, observe que no modelo de regressão simples se $x = 0$ então $\mathrm{E}(Y) = \beta_0$. A existência deste coeficiente implica que a primeira coluna da matriz X seja de 1, já que este coeficiente não está relacionado a nenhuma das variáveis explicativas.

Devemos esclarecer que o fato de qualquer variável aparecer de maneira linear no modelo (1.3) não significa que seja, necessariamente linear, ou seja, pode ser que $x_{i1} = \log(z_{i1})$ sendo z_{i1} uma outra variável positiva e que $x_{i2} = z_{i2}^3$. Portanto

$$Y_i = \beta_0 + \beta_1 \log(z_{i1}) + \beta_2 z_{i2}^3 + \cdots + \beta_p x_{ip} + \epsilon_i,$$

onde $\mathrm{E}(\epsilon_i) = 0$ e $i = 1, \cdots, n$, continuando sendo um modelo linear.

1.2.1 Propriedades dos modelos lineares

Existe uma situação mais geral da considerada na definição do modelo linear. Aitken (1935) definiu modelos lineares como aqueles nos quais

$$\mathrm{E}(Y) = X\beta \qquad \text{e} \qquad \mathrm{Var}(Y) = \sigma^2 G,$$

para $|G| \neq 0$, uma matriz quadrada conhecida, introduzindo assim correlações entre as observações.

Este modelo pode ser reduzido para o modelo em (1.1) considerando $Z = G^{-1/2}Y$, obtendo-se

$$\mathrm{E}(Z) = G^{-1/2}X\beta = U\beta \qquad \text{e} \qquad \mathrm{Var}(Y) = \sigma^2 \mathrm{I},$$

onde I é a matriz identidade. Isso significa que neste livro basta considerar somente modelos lineares segundo definidos em (1.1), casos onde a estrutura de correlação entre as variáveis resposta seja conhecida podem ser reduzidos a modelos lineares independentes.

Desde os anos 1980 o termo "modelo linear" se tornou padrão em grande parte da literatura. Seu uso difundiu-se muito rapidamente. Muitos trabalhos contribuíram para encontrar propriedades probabilísticas, por exemplo, os livros Searle (1971) e Rao (1973). Algumas das propriedades destes modelos as apresentamos nos resultados a seguir.

Teorema 1.1. *Consideremos um modelo linear* $Y = X\beta + \epsilon$, *satisfazendo* $\mathrm{E}(Y) = X\beta$ *e seja* L *um vetor coluna. Então*

$$\mathrm{E}(L^\top Y) = L^\top X\beta\cdot$$

Demonstração. Exercício. □

Teorema 1.2. *Consideremos um modelo linear* $Y = X\beta+\epsilon$, *satisfazendo* $\mathrm{E}(Y) = X\beta$ *e* $\mathrm{Var}(Y) = \sigma^2 \mathrm{I}$ *e seja* L *um vetor coluna. Então*

$$\mathrm{Var}(L^\top Y) = \sigma^2 L^\top L\cdot$$

Demonstração.

$$\mathrm{Var}(L^\top Y) = L^\top \mathrm{Var}(Y)L = \sigma^2 L^\top L\cdot$$

□

Estes dois teoremas apresentam propriedades dos momentos de primeira e segunda ordens de combinações lineares da variável resposta. Vejamos agora resultados de relações entre combinações lineares e formas quadráticas.

Teorema 1.3. *Consideremos um modelo linear* $Y = X\beta+\epsilon$, *satisfazendo* $\mathrm{E}(Y) = X\beta$ *e* $\mathrm{Var}(Y) = \sigma^2 \mathrm{I}$ *e sejam* L *e* M *vetores coluna. Então*

$$\mathrm{Cov}(L^\top Y, M^\top Y) = \sigma^2 L^\top M\cdot$$

Demonstração.

$$\mathrm{Cov}(L^\top Y, M^\top Y) = L^\top \mathrm{Cov}(Y, Y^\top)M = L^\top \mathrm{Var}(Y)M = \sigma^2 L^\top M\cdot$$

□

Teorema 1.4. *Consideremos um modelo linear* $Y = X\beta+\epsilon$, *satisfazendo* $\mathrm{E}(Y) = X\beta$ *e* $\mathrm{Var}(Y) = \sigma^2 \mathrm{I}$. *Seja* G *uma matriz quadrada. Então*

$$\mathrm{E}(Y^\top GY) = \sigma^2 \,\mathrm{tr}(G) + \beta^\top X^\top GX\beta,$$

onde tr *denota a função traço de uma matriz.*

Observemos que utilizamos neste teorema uma nova função, chamada de traço[7].

Demonstração. Para simplificar a notação, seja $\mathrm{E}(Y) = \mu$. Podemos escrever então

$$\begin{aligned} Y^\top GY &= (Y-\mu)^\top GY + \mu^\top GY \\ &= (Y-\mu)^\top G(Y-\mu) + \mu^\top G\mu + (Y-\mu)^\top G\mu\cdot \end{aligned}$$

Tomando esperança o último termo desaparece e

$$\mathrm{E}(Y^\top GY) = \mathrm{E}[(Y-\mu)^\top G(Y-\mu)] + \mu^\top G\mu\cdot,$$

bastando provar que $\mathrm{E}\big((Y-\mu)^\top G(Y-\mu)\big) = \sigma^2\,\mathrm{tr}(G)$, mas este é um cálculo direto (exercício). □

Suposições dos modelos de regressão lineares

Como qualquer modelo estatístico devemos fazer certas suposições para poder obtermos estimadores, realizar testes de hipóteses de interesse e verificarmos a adequação deste modelo aos dados, a final, a utilidade de qualquer modelo é sua qualidade em representar os dados.

1. **Suposição de Linearidade:** esta suposição é implícita na definição do modelo (1.1) e significa que cada valor observado y_i da variável resposta Y_i pode ser escrito como uma função linear da i-ésima linha da matriz X, x_i, isto é

$$y_i = x_i\beta + \epsilon_i, \qquad i = 1, \cdots, n\cdot$$

2. **Suposição Computacional:** para encontrar estimador único do vetor de parâmetros β é necessário que $(X^\top X)^{-1}$ exista, ou equivalentemente que

$$\mathrm{posto}(X) = p+1,$$

onde posto é a função posto[8] e $p+1$ é o número de parâmetros de regressão ou dimensão do vetor β.

3. **Suposição Distribucional:** a estimação dos parâmetros é por mínimos quadrados ou por máxima verossimilhança e para isso assume-se que:

   (a) X é mensurada sem erros,

   (b) ϵ_i não depende de x_i, $i = 1, \cdots, n$ e

   (c) $\epsilon \sim N_n(0, \sigma^2\mathrm{I})$, isto é, a distribuição de probabilidade do erro é normal multivariada de ordem n, com média zero e matriz de variâncias e covariâncias $\sigma^2\,\mathrm{I}$. Isto significa que $\mathrm{E}(\epsilon_i) = 0$ e

$$\mathrm{cov}(\epsilon_i, \epsilon_j) = \begin{cases} \sigma^2 & \text{se} \quad i \neq j \\ 0 & \text{se} \quad i = j \end{cases}, \qquad \text{para}\, i, j = 1, \cdots, n$$

4. **Suposição Implícita:** todas as observações são igualmente possíveis de serem observadas e tem a mesma importância na determinação dos resultados de mínimos quadrados ou de máxima verossimilhança. Significa que todas as observações têm a influência nas conclusões.

[7] O traço de uma matriz quadrada A, denotada como $\mathrm{tr}(A)$, é a função que associa a matriz $A = (a_{ij})$ à soma dos elementos da sua diagonal principal, ou seja, $\mathrm{tr}(A) = a_{11} + \cdots + a_{nn}$.

[8] Uma matriz pode ser considerada como um conjunto de vetores linha ou coluna escritos em uma ordem particular. O posto de uma matriz é definido como o número de linhas independentes ou de colunas independentes.

1.3 Distribuição normal

A distribuição normal ou gaussiana é a função de densidade mais amplamente usada em aplicações estatísticas numa grande variedade de áreas. No começo dos desenvolvimentos da teoria das probabilidades a distribuição binomial utilizou-se para resolver problemas do tipo "Se uma moeda é lançada 100 vezes, qual a probabilidade de obter 60 ou mais caras?". O cálculo exato deste número é realizado utilizando a expressão

$$P(X = x) = \frac{N!}{x!(N - x!)}\theta^x(1 - \theta)^{N-x},$$

onde x é o número de caras (60), N o número de lançamentos da moeda (100) e θ a probabilidade de obter-se cara ao lançar uma moeda balanceada (1/2).

Acontece que para resolver este problema devemos calcular a probabilidade de obter 60 caras, depois de obter 61 caras, 62 caras, e assim por diante e somar estes valores para obtermos a probabilidade procurada. Não é difícil imaginar o trabalhou-so deste simples cálculo sem a utilização das modernas calculadoras ou computadores.

Abraham de Moivre (1667-1754), no século 18, observou que quando o número de eventos aumenta (lançamentos da moeda), a forma da distribuição binomial aproxima-se de uma curva continua e suave, isto pode ser observado nos gráficos na Figura 1.4.

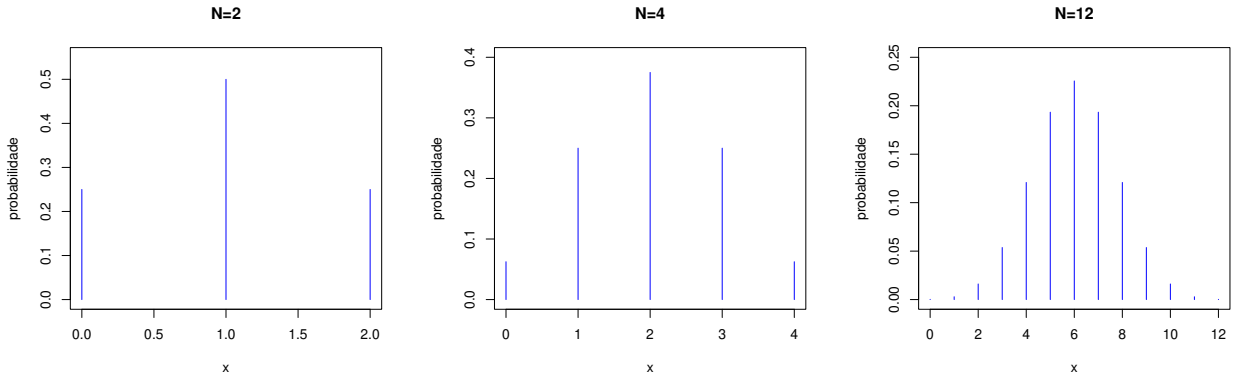

Figura 1.4: Exemplos da distribuição binomial. O tamanho das barras azuis representam os valores de probabilidade.

O pensamento então foi que se fosse possível encontrar a expressão matemática desta curva suave seria muito mais fácil o cálculo para encontrar a probabilidade de obter 60 ou mais caras em 100 lançamentos de uma moeda, e foi exatamente isso que de Moivre fez, descobrindo a densidade normal.

Definição 1.4 (Normal padrão)**.** *Considere a distribuição dos disparos ao alvo (idealmente pontos) e seja (X, Y) as coordenadas dos desvios (erros) do disparo com relação aos eixos ortogonais através do centro. Consideremos as seguintes hipóteses:*

(a) As funções de densidade marginal dos erros X e Y são contínuas,

(b) A função de densidade no ponto (x, y) depende somente da distância $z = (x^2 + y^2)^{1/2}$ desde a origem,

(c) Os erros nas direções em x e y são independentes.

A função de densidade do desvio Z em qualquer direção é a densidade normal

$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left(-z^2/2\right), \tag{1.4}$$

para $-\infty < z < \infty$.

A distribuição normal é importante pelo fato de ser a distribuição de probabilidade natural em muitos fenômenos ou pelo menos aproxima bem a distribuição de probabilidades natural do fenômeno. Uma das primeiras aplicações foi a análise de erros de medições. Galileo (Galileo Galilei, 1564-1642) no século 17 observou que os erros de medições, no caso medições astronômicas, aconteciam simetricamente e que pequenos erros ocorriam mais frequentemente do que erros grosseiros.

De maneira independente, os matemáticos Adrian (Robert Adrain, 1775-1843) em 1808 e Gauss (Johann Carl Friedrich Gauss, 1777-1855) em 1809 desenvolveram a expressão da densidade normal e mostraram que os erros de medição ajustam-se bem a ela. De diversas formas podemos encontrar a forma funcional desta densidade. Por exemplo, Maxwell[9] chegou à distribuição normal estudando a distribuição das velocidades de moléculas. Outras formas podem ser encontradas em Rao (1973). A Definição 1.4 deve-se a William Herschel (1738-1822). Resumidamente dizemos que $Z \sim N(0,1)$, com $\mathrm{E}(Z) = 0$ e $\mathrm{Var}(Z) = 1$, e chamamos de distribuição normal padrão.

Para obter uma variável aleatória normal Y com média arbitrária μ e variância σ^2, vamos utilizar a transformação $Z = (Y - \mu)/\sigma$ ou $Y = \sigma Z + \mu$, de modo que $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \sigma^2$. Agora encontramos a densidade $f(y)$ de $f(z)$ em (1.4). Para uma função crescente contínua, como $y = \sigma z + \mu$, ou para uma função decrescente contínua, a técnica de mudança de variável para uma integral definida fornece

$$f(y) = f(z) \left| \frac{\mathrm{d}z}{\mathrm{d}y} \right| \tag{1.5}$$

onde $|\,\mathrm{d}z/\,\mathrm{d}y|$ é o valor absoluto de $\mathrm{d}z/\,\mathrm{d}y$. Para usar (1.5) para encontrar a função de y, é claro que tanto z como $\mathrm{d}z/\,\mathrm{d}y$ no lado direito devem ser expressos em termos de y. Vamos aplicar (1.5) a $y = \sigma z + \mu$. A densidade $g(z)$ é dada em (1.4), e para $z = (y - \mu)/\sigma$, temos $|\,\mathrm{d}z/\,\mathrm{d}y| = 1;\sigma$. Portanto

$$f(y) = f(z) \left| \frac{\mathrm{d}z}{\mathrm{d}y} \right| = f\left(\frac{y-\mu}{\sigma}\right) \frac{1}{\sigma} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \tag{1.6}$$

que é a densidade normal com $\mathrm{E}(Y) = \mu$ e $\mathrm{Var}(Y) = \sigma^2$.

Quando Y tem densidade (1.6), dizemos que Y é distribuído como $N(\mu, \sigma^2)$ ou simplesmente que $Y \sim N(\mu, \sigma^2)$. A forma funcional desta densidade é mostrada na Figura 1.5, para diversos valores de variância e qualquer seja a média.

Observamos que a distribuição normal, qualquer seja o valor da esperança μ é simétrica em torno desse número e que quanto menor a variância mais concentrada a densidade ao redor de μ. O contrário acontece conforme a variância aumenta. Estas e outras propriedades da densidade normal podem ser encontradas em livros de probabilidade, aqui fazemos uma apresentação, tanto é que não faremos a demonstração de que a expressão em (1.4) ou em (1.6) é mesmo uma função de densidade.

Como encontrar a expressão (1.6) a partir daquela na Definição 1.4. Escrevendo $\mathrm{E}(X) = \mu$ e $\mathrm{Cov}(X) = \sigma^2$, obtemos uma transformação adequada que nos permite chegar à expressão em (1.4). Alguns detalhes importantes devem ser considerados:

[9] James Clerk Maxwell (1831-1879) foi um físico e matemático britânico.

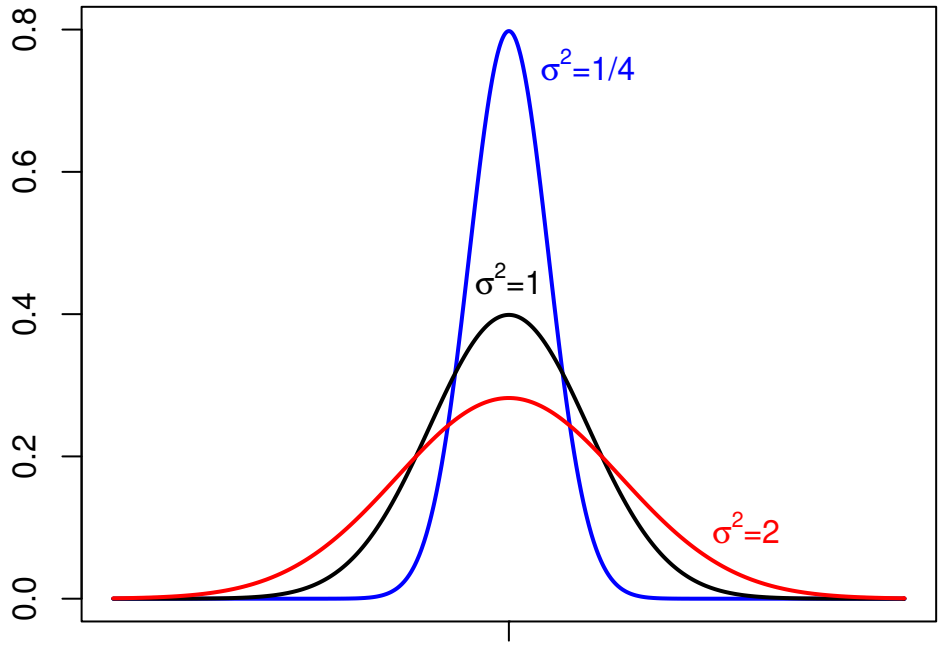

Figura 1.5: Densidade normal para diversos valores de variância e igual média.

- A forma da função de densidade é determinada pelos valores dos parâmetros μ e σ^2. Então, se conhecemos a média e variância populacionais e sabemos que a distribuição de probabilidades da variável aleatória é normal, podemos afirmar que conhecemos tudo acerca das probabilidades associadas já que conhecemos a função (1.6) completamente.
- A forma de (1.6) depende criticamente do termo

$$-\frac{(y-\mu)^2}{\sigma^2} = (y-\mu)(\sigma^2)^{-1}(y-\mu),$$

  e note que este termo depende do desvio da média ou diferença quadrática $(y-\mu)^2$.
- O desvio mencionado acima é padronizado pelo desvio padrão σ, o qual tem a mesma unidade do que Y, por isso a transforma em unidades padronizadas, ou seja, números reais.
- O desvio da média padronizada tem a interpretação de medida de distância, isto é, mede quão afastado Y está de μ e logo coloca o resultado em unidades padronizadas do distanciamento, de Y ao redor de μ, esperado.
- Desta forma a distribuição normal e o método de mínimos quadrados, o qual depende da minimização da soma dos desvios quadrados, estão fortemente relacionados. Alias, o matemático Robert Adrian chegou à expressão da densidade normal pela formulação do método de mínimos quadrados.

Na linguagem de programação R (Team, 2024) temos disponíveis diversas funções para calcular as probabilidades de variáveis normais. As funções básicas são: `dnorm`, `pnorm`, `qnorm` e `rnorm` as quais, respectivamente, nos fornecem a função de densidade, a função de distribuição acumulada, os quantis e nos permitem a geração de números aleatórios.

A forma de utilização destas funções é mostrada a seguir e pode ser consultada digitando `help(rnorm)`

```
dnorm(y, mean=0, sd=1, log = FALSE)
pnorm(q, mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)
qnorm(p, mean=0, sd=1, lower.tail = TRUE, log.p = FALSE)
rnorm(n, mean=0, sd=1)
```

onde `y` é um vetor de números reais, `q` um vetor de quantis, `p` um vetor de probabilidades e `n` o número de observações. Os parâmetros da distribuição são considerados, média 0 e variância 1 se não são especificados, caso queira-se especificá-los deve-se indicar o vetor de médias em `mean` e o vetor de desvio padrão em `sd`. Em algumas situações interessam os valores do logaritmo da função de densidade ou o logaritmo da probabilidade e, nesses casos, especifica-se `log=TRUE` ou `log.p=TRUE`.

A função de distribuição acumulada, por definição, nos fornece as probabilidades acumuladas até o ponto de interesse, ou seja, `pnorm(y)`$=P(Y \leq y)$. Se o interesse é encontrar $P(Y > y)$ temos duas formas de fazer este cálculo. Uma forma é digitar 1-`pnorm(y)`, outra é especificar `lower.tail=FALSE`, isto é, escrever 1-`pnorm(y,lower.tail=F)`.

Por exemplo, se $Y \sim N(2, 9)$ e queremos calcular $P(-2 < Y < 3)$ podemos obter o resultado escrevendo na linha de comandos do R

```
> pnorm(3,mean=2,sd=3)-pnorm(-2,mean=2,sd=3)
```

ou

```
> pnorm(-2,mean=2,sd=3,lower.tail=F)-pnorm(3,mean=2,sd=3,lower.tail=F),
```

obtendo-se em ambos os casos `0.5393474` como resposta. Se o objetivo é calcular $P(Y \geq 1)$ escrevemos a linha de códigos,

```
> 1-pnorm(1,mean=2,sd=3)
```

obtendo-se `0.6305587`; lembrando que $P(Y \geq 1) = P(Y > 1)$.

1.3.1 Distribuição normal multivariada

Faremos nesta seção uma introdução à distribuição de probabilidades normal para múltiplas variáveis. Primeiro estudaremos a forma de defini-la no caso de duas variáveis e depois para o caso mais geral e também mostraremos algumas das mais importantes propriedades aplicadas aos modelos de regressão.

Uma das estatísticas mais comuns na pesquisa científica é o coeficiente de correlação, coeficiente de correlação de Pearson e também chamado coeficiente de correlação de produtos de momentos de Pearson. Naturalmente, a correlação de Pearson é usada para examinar relações bivariadas simples mas a covariância, o equivalente não escalada, também é usada em várias técnicas multivariadas. Se a estimativa de correlação de Pearson for deflacionada ou inflada por não-normalidade, esse problema pode se generalizar para uma ampla variedade de técnicas estatísticas. O objetivo do presente relatório é (1) determinar quando a não-normalidade distorce a estimativa pontual do coeficiente de correlação de Pearson e (2) comparar sistematicamente as principais alternativas ao coeficiente de correlação de Pearson para determinar se elas podem atenuar esse problema.

A densidade da distribuição normal bivariada foi estudada por Robert Adrian em 1808, Pierre-Simon Laplace em 1811, Carl Friedrich Gauss em 1813, 1823 e Auguste Bravais em 1846 dentre

muitos outros. No entanto, nenhum desses autores descobriu o coeficiente de correlação como uma medida de associação e as características das distribuições condicionais, como linhas de regressão e homocedasticidade, como fez Francis Galton em 1889. Para um estudo detalhado do desenvolvimento das ideias de correlação, ver Helen Walker (1931).

Se a correlação de Pearson é realmente distorcida, a questão importante é: "Comparado com o quê?"(Efron, 1988). No lugar da correlação de Pearson, existem várias técnicas alternativas para lidar com a não-normalidade, mas não está claro qual delas, se alguma, se sairia melhor. As principais alternativas para a correlação de Pearson incluem bootstrapping, correlação de ordem de classificação de Spearman e outras transformações de dados não-lineares.

Antes de conhecermos a distribuição normal multivariada estudaremos o coeficiente de correlação, conceito importante nos modelos de regressão.

Definição 1.5. *Sejam X e Y variáveis aleatórias tais que $\mathrm{E}(X^2)$ e $\mathrm{E}(Y^2)$ existem. Definimos o coeficiente de correlação entre X e Y como*

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}, \tag{1.7}$$

onde $\mathrm{Cov}(X,Y) = \mathrm{E}(XY) - \mathrm{E}(X)\,\mathrm{E}(Y)$ é a covariância entre X e Y.

Observemos que o sinal de ρ coincide com o sinal da covariância entre X e Y. Um outro detalhe importante é que o cálculo da raiz quadrada das variâncias necessários na expressão (1.7) considera somente a raiz positiva.

Definição 1.6. *Dizemos que as variáveis X e Y são não correlacionadas se, e somente se, $\rho = 0$.*

Logicamente $\rho = 0$ se, e somente se, $\mathrm{Cov}(X,Y) = 0$. Se X e Y forem independentes[10] então $\mathrm{Cov}(X,Y) = 0$ e X e Y são não correlacionadas.

Devemos esclarecer que, se X e Y forem não correlacionadas e portanto $\mathrm{Cov}(X,Y) = 0$, não implica que X e Y sejam necessariamente independentes. Por exemplo, sejam $X = U + V$ e $Y = U - V$ onde U e V são duas variáveis com a mesma média e variância. Então

$$\mathrm{Cov}(X,Y) = \mathrm{E}(U^2 - V^2) - \mathrm{E}(U - V)\,\mathrm{E}(U + V) = 0,$$

logo, X e Y são não correlacionadas porém não necessariamente independentes.

Teorema 1.5. *Seja ρ o coeficiente de correlação entre X e Y. Então $|\rho| \leq 1$ e a igualdade $|\rho| = \pm 1$ acontece se, e somente se, existirem constantes β_0 e β_1 tais que*

$$P(Y = \beta_0 + \beta_1 X) = 1\cdot$$

Demonstração. Pela desigualdade de Cauchy-Schwarz (James, 2009)

$$\mathrm{Cov}(X,Y) \leq \mathrm{Var}(X)\,\mathrm{Var}(Y),$$

com igualdade se, e somente se, existirem constantes β_0 e β_1 tais que

$$P(Y = \beta_0 + \beta_1 X) = 1\cdot$$

□

[10]Dizemos que as variáveis aleatórias X e Y são independentes se, e somente se, $F(x,y) = F_1(x)F_2(y)$ para toda $(x,y) \in \mathbb{R}_2$, onde $F(x,y)$ é a distribuição conjunta do vetor (X,Y) e $F_1(x)$ e $F_2(y)$ são as distribuições marginais de X e Y, respectivamente.

Observemos que a situação em que $\rho = \pm 1$ é oposta à independência. Assim, quanto mais próximo de 1 ou -1 o valor do coeficiente de correlação mais forte é a relação linear entre as variáveis X e Y. Até aqui entendemos melhor o conceito de correlação para quaisquer variáveis aleatórias, vejamos agora a importância do coeficiente de correlação na definição da distribuição normal multivariada.

Definição 1.7. *Diz-se que o vetor de variáveis aleatórias (X, Y) tem função de densidade normal bivariada se a função de densidade conjunta é da forma*

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2}Q(x,y)\right], \tag{1.8}$$

onde $\sigma_1 > 0$, $\sigma_2 > 0$, $|\rho| < 1$ e $Q(x,y)$ é uma específica forma quadrática de expressão

$$Q(x,y) = \frac{1}{1-\rho^2}\left[\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - 2\rho\left(\frac{x-\mu_1}{\sigma_1}\right)\left(\frac{y-\mu_2}{\sigma_2}\right) + \left(\frac{y-\mu_2}{\sigma_2}\right)^2\right].$$

Interpretamos os parâmetros como $\mathrm{E}(X) = \mu_1$, $\mathrm{Var}(X) = \sigma_1^2$, $\mathrm{E}(Y) = \mu_2$, $\mathrm{Var}(Y) = \sigma_2^2$ e ρ o coeficiente de correlação entre X e Y. Vemos também que não é possível definir distribuição normal bivariada caso a relação entre estas variáveis seja perfeitamente linear. Interessante é notar que a covariância entre X e Y pode ser escrita como $\mathrm{Cov}(X,Y) = \rho\sigma_1\sigma_2$. Podemos então definir a matriz de variâncias e covariâncias entre X e Y como

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma^2 \end{pmatrix}, \tag{1.9}$$

de determinante $|\Sigma| = \sigma_1\sigma_2(1-\rho^2)$. A raiz quadrada positiva deste determinante aparece no denominador da expressão (1.8) que define a distribuição normal bivariada.

Teorema 1.6. *Seja (X, Y) um vetor aleatório com distribuição normal bivariada. As variáveis X e Y são independentes se, e somente se, $\rho = 0$.*

Demonstração. Exercício. □

Observando a expressão (1.8) concluímos que, no caso das variáveis X e Y serem independentes, a distribuição normal bivariada se reduz ao produto de duas distribuições normais univariadas

$$f(x,y) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_1)^2}{\sigma_1^2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right]\right\}.$$

Teorema 1.7. *Consideremos o vetor aleatório (X, Y) com distribuição normal bivariada. Então as densidades marginais de X e Y são, respectivamente, $N(\mu_1, \sigma_1^2)$ e $N(\mu_2, \sigma_2^2)$.*

Demonstração. Seja $f(x) = \displaystyle\int_{-\infty}^{\infty} f(x,y)\,\mathrm{d}y$, a função de densidade marginal de X. Observe que

$$\begin{aligned}(1-\rho^2)Q(x,y) &= \left(\frac{y-\mu_2}{\sigma_2} - \rho\frac{x-\mu_1}{\sigma_1}\right)^2 + (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2 \\ &= \left\{\frac{y-[\mu_2+\rho(\sigma_2/\sigma_1)(x-\mu_1)]}{\sigma_2}\right\}^2 + (1-\rho^2)\left(\frac{x-\mu_1}{\sigma_1}\right)^2.\end{aligned}$$

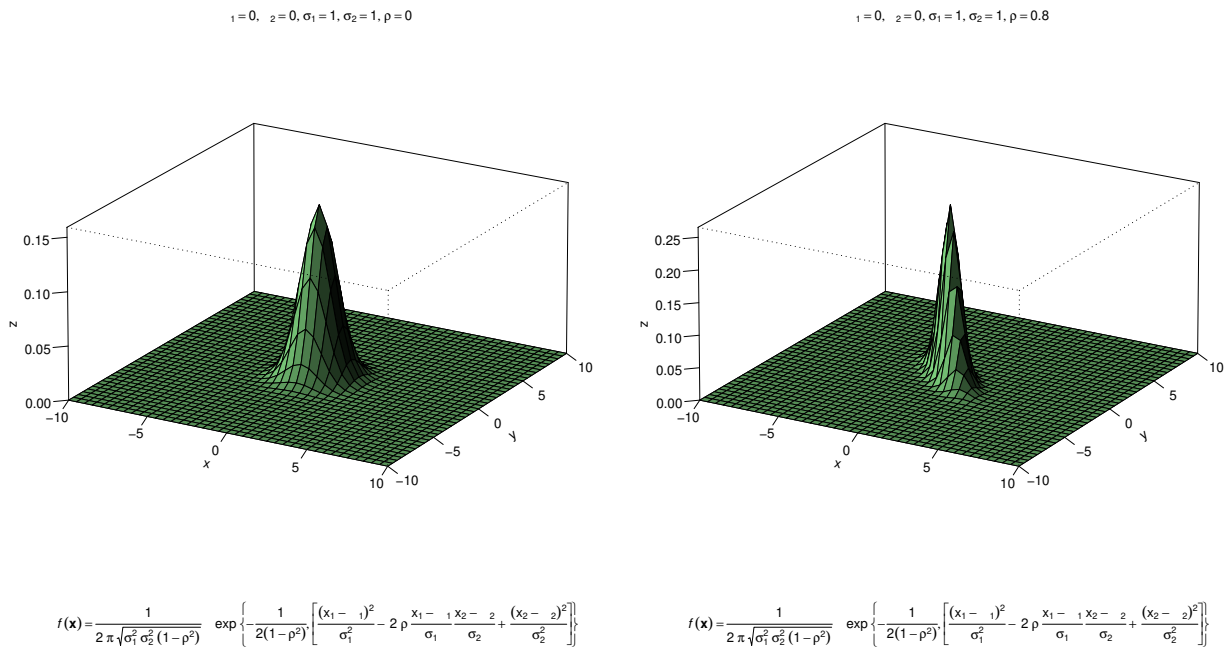

Figura 1.6: Densidade normal bi-variada. A esquerda marginais independentes e a direita marginais correlacionadas com coeficiente de correlação igual a 0,8.

Desta expressão segue que

$$f(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right] \int_{-\infty}^{\infty} -\frac{(y-\beta_x)^2/[2\sigma_2^2(1-\rho^2)]}{\sigma_2\sqrt{1-\rho^2}\sqrt{2\pi}}\,\mathrm{d}y,$$

onde

$$\beta_x = \mu_2 + \rho\left(\frac{\sigma_2}{\sigma_1}\right)(x-\mu_1)\cdot \tag{1.10}$$

A função na integral corresponde a distribuição $N[\beta_x, \sigma_2^2(1-\rho^2)]$, logo

$$f(x) = \frac{1}{\sigma_1\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2\right], \tag{1.11}$$

sendo $-\infty < x < \infty$. Similarmente para obter a função de densidade marginal de Y. □

Teorema 1.8. *Consideremos o vetor aleatório (X, Y) com distribuição normal bivariada. Então, $N[\beta_y, \sigma_1^2(1-\rho^2)]$ e $N[\beta_x, \sigma_2^2(1-\rho^2)]$ são as densidades condicionais de $X|Y = y$ e $Y|X = x$, respectivamente. A expressão de β_x foi obtida em (1.10) e $\beta_y = \mu_1 + \rho\sigma_1(y-\mu_2)/\sigma_2$.*

Demonstração. A função de densidade condicional de $Y|X = x$ é

$$f_{Y|X}(y|x) = \frac{f(x,y)}{f(x)} = \frac{1}{\sigma_2\sqrt{2\pi}\sqrt{1-\rho^2}} \exp\left[-\frac{(y-\beta_x)^2}{2\sigma_2^2(1-\rho^2)}\right],$$

onde $f(x,y)$ é a função de densidade conjunta definida em (1.8), $f(x)$ a densidade marginal de X, foi obtida em (1.11) e β_x foi definida em (1.10). Desta expressão fica claro que a densidade marginal de $Y|X = x$ é normal de média β_x e variância $\sigma_2^2(1-\rho^2)$. A média de $X|Y = y$ é $\beta_y = \mu_1 + \rho\sigma_1(y-\mu_2)/\sigma_2$. □

Francis Galton em 1889 explorou problemas bivariados em genética e introduziu as ideias de correlação e regressão no estudo de medidas pareadas. A função de densidade normal multivariada é uma extensão da expressão em (1.8) à situação em que Y é um vetor de dimensão n, no qual cada componente é normalmente distribuída possivelmente correlacionadas.

Definição 1.8. *Diz-se que o vetor de variáveis aleatórias $Y = (Y_1, \cdots, Y_n)^\top$ tem função de densidade normal multivariada se a função de densidade conjunta é da forma*

$$f(\underset{\sim}{y}) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(\underset{\sim}{y} - \mu)^\top \Sigma^{-1}(\underset{\sim}{y} - \mu)\right],$$

onde $\underset{\sim}{y} = (y_1, \cdots, y_n)^\top$ é o vetor de observações, $\mathrm{E}(Y) = \mu$ *e* $\mathrm{Var}(Y) = \Sigma$*, sendo Σ a matriz de variâncias e covariâncias de Y.*

Resumidamente escrevemos $Y \sim N_n(\mu, \Sigma)$, com o qual queremos dizer que o vetor Y tem como distribuição de probabilidades normal multivariada de dimensão n, com média μ e matriz de variâncias e covariâncias Σ. Logicamente, se a dimensão do vetor Y é dois, a expressão em (1.8) se reduz à situação bivariada definida em (1.8). As densidades marginais e condicionais são semelhantes ao caso bivariado. A matriz de variâncias e covariâncias Σ pode ser como em (1.9) no caso bivariado ou ainda $\Sigma = \sigma^2\,\mathrm{I}_n$, no caso das variáveis $Y_1, \cdots, Y_n$ serem independentes, nesta expressão I_n representa a matriz identidade de ordem n. Mais interessante no caso multivariado é o resultado a seguir, o qual nos permitirá deduzir diversas propriedades dos estimadores dos parâmetros dos modelos de regressão.

Teorema 1.9. *Seja $Y = (Y_1, \cdots, Y_n)^\top$ um vetor aleatório com distribuição $N_n(\mu, \Sigma)$. Então, qualquer combinação linear de CY satisfaz*

$$CY \sim N_p(C\mu, C\Sigma C^\top), \tag{1.12}$$

onde C é uma matriz $p \times n$ não singular.

Demonstração. Seja $Z = CY$. A distribuição de Z é obtida da transformação

$$\underset{\sim}{y} = C^{-1}\underset{\sim}{z}$$

utilizando o teorema do Jacobiano (Spivak, 1970). O determinante do jacobiano, denotado por $|J|$, desta transformação é

$$|J| = \sqrt{\frac{1}{|C|^2}} = \sqrt{\frac{|\Sigma|}{|C| \times |\Sigma| \times |C^\top|}} = \frac{|\Sigma|^{1/2}}{|C\Sigma C^\top|^{1/2}},$$

dado que C é uma matriz não singular, isto é, C é tal que $|C| \neq 0$. A forma quadrática $(\underset{\sim}{y} - \mu)^\top\Sigma^{-1}(\underset{\sim}{y} - \mu)$ assume a forma,

$$\begin{aligned}
(\underset{\sim}{y} - \mu)^\top\Sigma^{-1}(\underset{\sim}{y} - \mu) &= (C^{-1}\underset{\sim}{z} - \mu)^\top\Sigma^{-1}(C^{-1}\underset{\sim}{z} - \mu) \\
&= (C^{-1}\underset{\sim}{z} - C^{-1}C\mu)^\top\Sigma^{-1}(C^{-1}\underset{\sim}{z} - C^{-1}C\mu) \\
&= [C^{-1}(\underset{\sim}{z} - C\mu)]^\top\Sigma^{-1}[C^{-1}(\underset{\sim}{z} - C\mu)] \\
&= (\underset{\sim}{z} - C\mu)^\top(C^{-1})^\top\Sigma^{-1}C^{-1}(\underset{\sim}{z} - C\mu) \\
&= (\underset{\sim}{z} - C\mu)^\top(C\Sigma C^\top)^{-1}(\underset{\sim}{z} - C\mu),
\end{aligned}$$

desde que $(C^{-1})^\top = (C^\top)^{-1}$ dado que $CC^{-1} = \mathrm{I}$. Então a densidade de Z é

$$\begin{aligned} f(z) &= \frac{1}{(2\pi)^{n/2}}|\Sigma|^{-1/2}|J|\exp[-\tfrac{1}{2}(\underset{\sim}{y}-\mu)^\top\Sigma^{-1}(\underset{\sim}{y}-\mu)] \\ &= \frac{1}{(2\pi)^{n/2}}\frac{1}{|C\Sigma C^\top|^{1/2}}\exp\left[-\tfrac{1}{2}(\underset{\sim}{z}-C\mu)^\top(C\Sigma C^\top)^{-1}(\underset{\sim}{z}-C\mu)\right]. \end{aligned}$$

□

Por exemplo, aplicando este teorema, se escolhemos $C = (1, \underbrace{0, \cdots, 0}_{n-1})$ obtemos que a densidade de Y_1 é da forma apresentada em (1.6).

1.4 Exemplos

Durante todo o livro utilizaremos diversos exemplos com o intuito de melhor explanar os resultados teóricos apresentados. Todos os arquivos de dados mencionados assim como diversos ajustes de modelos podem ser consultados no endereço `http://leg.ufpr.br/~lucambio/Linear/`.

Alguns destes exemplos, por sua complexidade, necessitam serem apresentados com maiores detalhes do que outros e aqui os detalhamos.

1.4.1 Quanto custa ser deputado?

Para ser eleito deputado federal no Brasil há uma série de exigências formais e uma grande exigência informal. As formais estão descritas na Constituição: o candidato precisa ter mais de 21 anos, nacionalidade brasileira e filiação partidária. A exigência informal não está em nenhuma lei, mas é tão verdadeira quanto as demais: para ser eleito, é preciso ter muito dinheiro.

A revista Época cruzou os resultados das eleições de 2010 com os dados de financiamento de campanha de 3.767 candidatos a deputado em todo o país. A conclusão é que em todos os Estados há uma forte correlação entre arrecadação de dinheiro e sucesso eleitoral. As estatísticas provam que é até possível arrecadar muito dinheiro e mesmo assim perder a eleição. Mas parece dificílimo vencer sem arrecadar muito.

Como os repórteres Ricardo Mendonça, Alberto Cairo, Marco Vergotti e Gerardo Rodriguez da revista Época chegaram a estas conclusões? o resultado da pesquisa foi publicado em maio de 2012 e utilizaremos os dados coletados por eles para mostrar uma aplicação de modelos de regressão.

No arquivo de dados `EPOCA.csv` dispomos das seguintes informações para cada candidato a Deputado Federal em 2010:

- UF: Unidade da Federação;
- Nome: Nome do Deputado ou candidato a Deputado Federal;
- Partido: Partido de filiação;
- N.votos: Número de votos (variável resposta);
- Arrecadado: Dinheiro arrecadado;
- Eleito: Código de valores 1 ou 0, indicando se o candidato foi eleito Deputado Federal ou não, respectivamente.

Juntos, os 3.767 candidatos a deputado captaram R$ 887 milhões. Os 513 eleitos mais os 58 suplentes que assumiram alguma cadeira na Câmara após licença do titular foram responsáveis por 70% do montante.

O Estado que teve a eleição proporcionalmente mais cara foi Roraima, onde cada voto custou R$ 66,16. Entre os eleitos, o campeão em receitas foi Sandro Mabel (PR-GO), com R$ 4,9 milhões. Já o custo por voto mais alto foi de Edio Lopes (PMDB-RR), com uma média de R$ 152,14 para cada eleitor. Curiosamente, os dois candidatos que mais arrecadaram no país não foram eleitos.

Estes são dados declarados por cada candidato no Tribunal Superior Eleitoral. No endereço mencionado podem ser consultadas as linhas de comando R que foram utilizadas para gerar os gráficos descritivos assim como o ajuste do modelo de regressão linear.

Na Figura 1.7 mostramos o gráfico de dispersão entre a variável resposta, o número de votos obtidos na campanha para Deputado Federal em 2010, e duas variáveis explicativas o total em dinheiro arrecadado e o indicativo da conquista do cargo. Alguns estatísticas descritivas das variáveis número de votos e arrecadação são mostradas na tabela a seguir:

Variável	n	média	desvio padrão	mínimo	máximo
Número de votos	3775	30707.42	209900.4	1.00	7565377
Arrecadado	3775	257835.65	968806.6	8.25	38036984

Modificamos a escala das variáveis devido a que na figura a esquerda não é possível perceber nenhuma relação entre o número de votos e a arrecadação, quando aplicamos a função logaritmo a ambas variáveis podemos perceber que, de maneira geral, quanto mais arrecadado pelo candidato maior o número de votos obtidos. Um detalhe importante é que aqueles eleitos Deputados Federais tiveram sempre arrecadação muito grande.

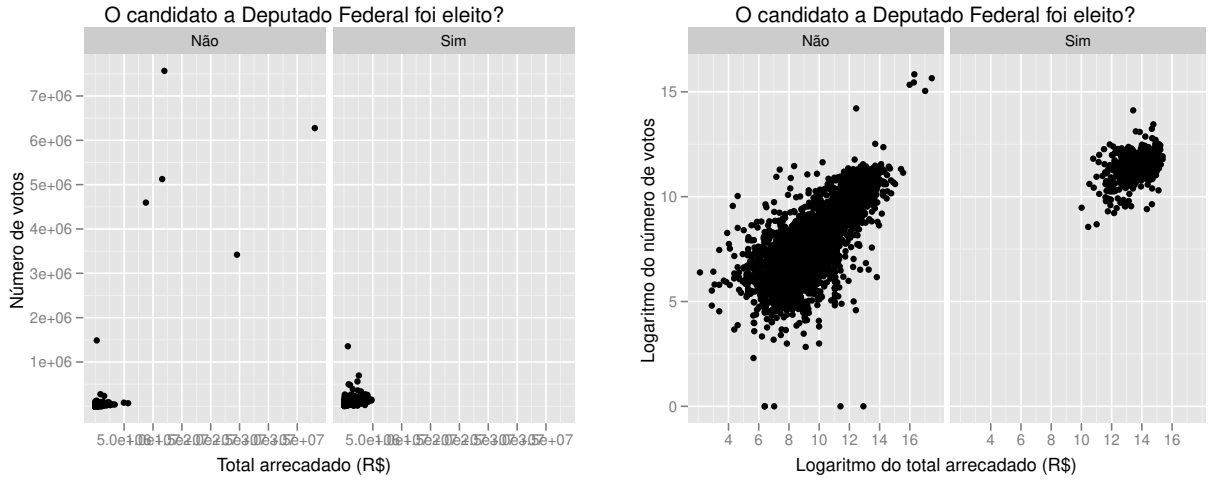

Figura 1.7: Gráficos de dispersão. A esquerda relação entre o valor arrecadado em reais (R$) para a campanha a Deputado Federal e o número de votos obtidos e a direita as mesmas variáveis em escala logarítmica. Os dados são apresentados separadamente segundo os valores da variável Eleito, isso é, mostramos separadamente os dados no caso dos candidatos eleitos e os que não foram eleitos Deputados Federais nas eleições de 2010.

Quando consideradas as outras informações disponíveis, como Estado da Federação e Distrito Federal e o partido de filiação do candidato em 2010, em ambos casos percebemos que existe relação positiva entre o valor arrecadado e o número de votos obtido, isto é, quanto maior arrecadação mais votos serão obtidos.

1.4.2 Peso dos mexilhões

Os dados deste exemplo vêm de um estudo de mexilhões cavalo amostrados nas Marlborough Sounds que são uma extensa rede de vias navegáveis e penínsulas localizados ao norte da Ilha Sul de Nova Zelândia.

A variável resposta é a masa muscular, a parte comestível do mexilhão, em gramas. Há quatro variáveis preditores todas relativas às características de conchas de mexilhões: largura, altura, comprimento da concha em milímetros (mm) e a masa ou peso da conha em gramas (g).

O objetivo é desenvolver um modelo que permita uma compreensão de como a distribuição de massa muscular depende das quatro variáveis preditoras. Espera-se que a função de regressão aumente com os valores dos preditores. Quantificar o quão esse aumento ocorre é também parte deste estudo.

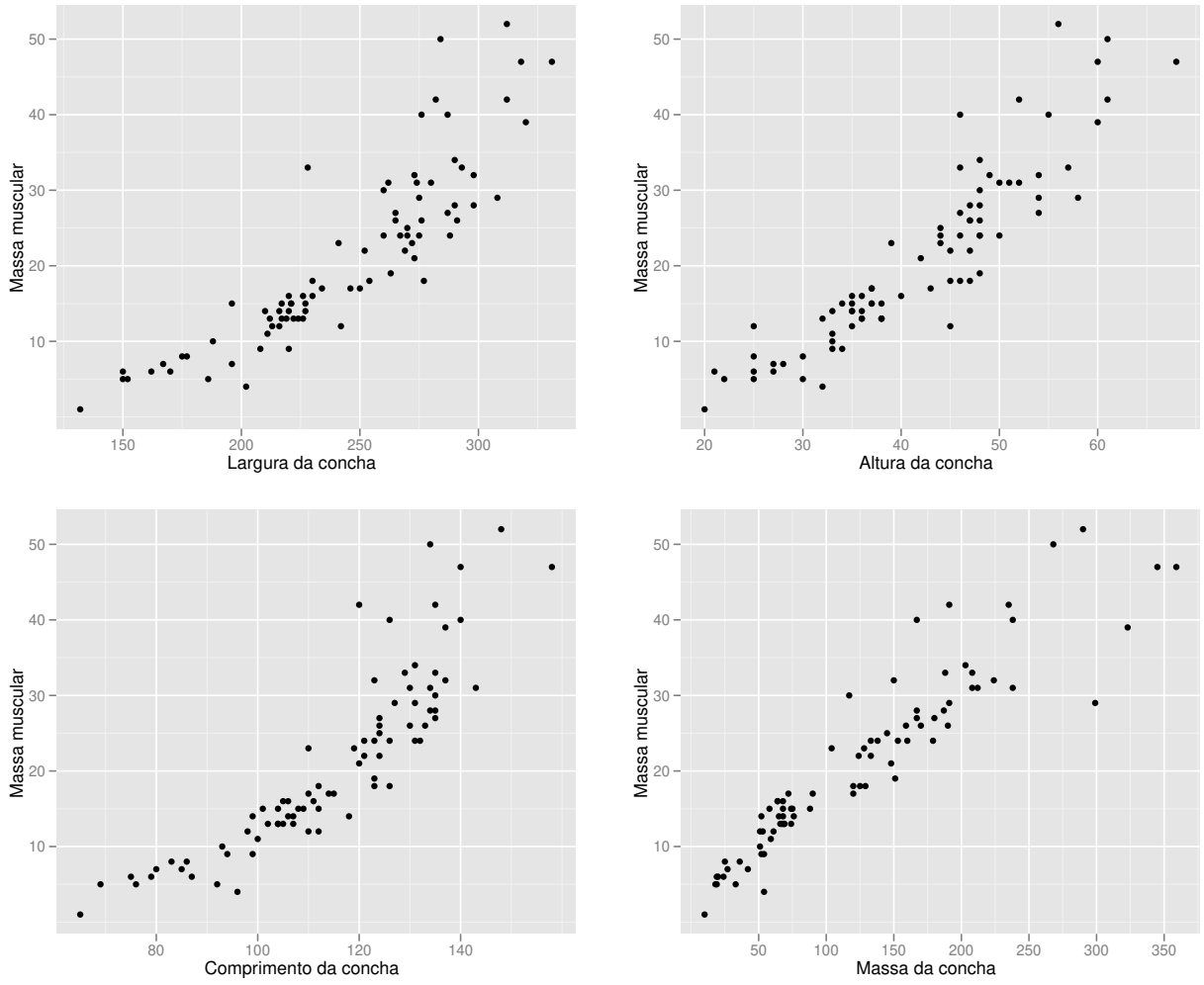

Figura 1.8: Gráficos de dispersão mostrando a relação entre a massa muscular do mexilhão (M), a variável resposta neste estudo, e as quatro variáveis preditoras ou explicativas: largura (W), altura (H), comprimento (L) e peso da concha (S).

Este estudo foi realizado 1986 por Mike Camden, Wellington Polytechnic, Wellington, Nova Zelândia e publicados no livro de Cook & Weisberg (1994). Os dados estão disponíveis no arquivo de dados `mexilhoes.csv` com as seguintes informações:

- W: largura da concha (mm);
- H: altura da concha (mm);
- L: comprimento da concha (mm);
- S: peso da concha (g);
- M: masa ou peso muscular (g) (variável resposta);

No site mencionado podem ser consultadas as linhas de comando R que foram utilizadas para gerar os gráficos descritivos na Figura 1.8, assim como o ajuste do modelo de regressão linear.

1.4.3 Biomassa do capim-marinho

Os dados considerados são parte de um estudo conduzido pelo Dr. Rick Linthurst em 1979 na Universidade Estadual da Carolina do Norte, nos Estados Unidos. O propósito de sua pesquisa foi identificar no solo importantes características que influenciam a produção de biomassa aérea do capim-marinho (Spartina alterniflora), que é uma planta herbácea presente em habitats lacustres, pântanos e zonas costeiras. É também conhecida como capim-da-praia, capim-da-roça e capim-paraturá.

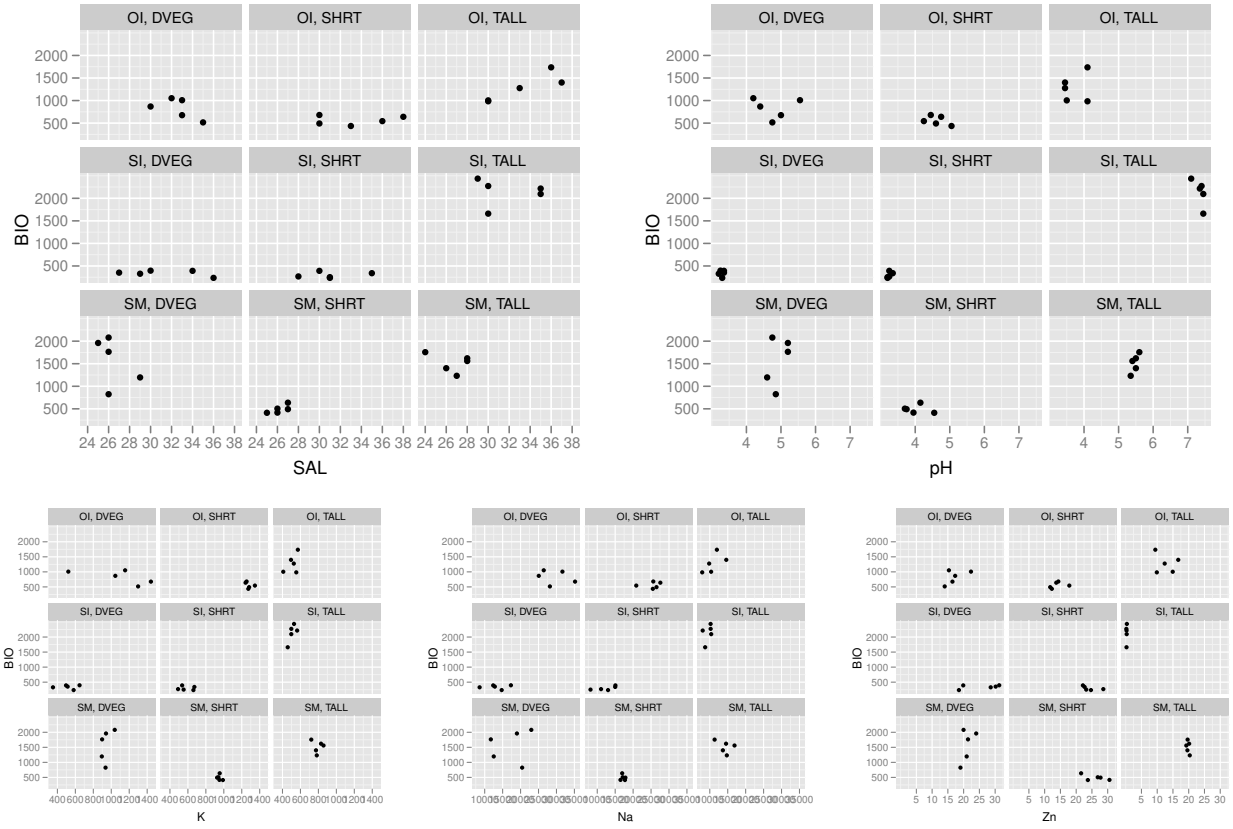

Figura 1.9: Gráficos de dispersão mostrando a relação entre a biomassa (BIO), variável resposta, e as distintas características observadas no solo: SAL, pH, K, Na e Zn. Considera-se a relação entre a resposta e as características do solo para cada combinação de local de amostragem e tipo de vegetação.

O estudo foi realizado no estuário do Cabo Fear[11] na Carolina do Norte. Uma fase da pesquisa consistiu em amostrar três tipos de vegetação, áreas regeneradas onde a vegetação foi morta denominada de DVEG, áreas com capim curto denominada de SHIRT e áreas com capim alto chamadas de TALL em cada um de três locais Oak Island (OI), Smith Island (SI) e Snows Marsh (SM).

As amostras de solo do substrato a partir de 5 sítios aleatórios dentro cada tipo de local de vegetação, que dá um total de 45 amostras, foram analisadas para as características físico-químicas do solo a cada mês durante vários meses, porém somente estão disponíveis os dados de setembro de 1979. Além disso, a biomassa acima do solo em cada local de cada amostra foi medida.

A variável dependente BIO é a biomassa aérea medida em grama por metro quadrado (gm^{-2}). O objetivo desta fase da pesquisa foi identificar quais substratos mostram as relações mais fortes para a produção de biomassa. Estas variáveis seriam então utilizadas em estudos controlados para investigar relações causais. O propósito deste estudo de caso é a utilização do modelo de regressão linear múltipla para relacionar a produção de biomassa com cinco variáveis de características de substrato e locais de amostragem.

Os dados estão disponíveis no arquivo de dados `capim-marinho.csv` com as seguintes informações:

- Loc.: localização de coleta Oak Island, Smith Island e Snows Marsh;
- Type: tipos de vegetação, podendo ser DVEG, SHIRT e TALL;
- BIO: biomassa aérea em gm^{-2} (variável resposta);
- SAL: percentual de salinidade;
- pH: acides da água, pH é o símbolo para o potencial hidrogeniônico que indica a acides, neutralidade ou alcalinidade de uma solução aquosa;
- K: potássio em partes por milhão (ppm);
- Na: sódio em ppm;
- Zn: zinco em ppm.

No endereço `http://leg.ufpr.br/~lucambio/Linear/` podem ser consultadas as linhas de comando R que foram utilizadas para gerar os gráficos descritivos na Figura 1.9, assim como o ajuste do modelo de regressão linear. Estes dados aparecem publicados com permissão do autor Dr. Rick A. Linthurst em Rawlings, Pantula & Dickey (1998).

[11] O Cabo Fear, em inglês Cape Fear que significa cabo do medo, é um promontório que penetra no Oceano Atlântico na costa da Carolina do Norte até ao sudeste dos Estados Unidos. Compõe-se principalmente de dunas num recife de coral e sedimentos do rio Cape Fear.

1.5 Exercícios

Exercícios da Seção 1.2

1. Sejam $\epsilon_1, \cdots, \epsilon_n$ variáveis aleatórias independentes com distribuição $N(0, \sigma^2)$. Quais dos seguintes modelos é linear? Explique. Sempre que possível, forneça uma transformação $Z_i = h(Y_i)$ que resulte num modelo linear para Z.

   (a) $Y_i = \beta_0 + \beta_1 x_i^2 + \epsilon_i$.

   (b) $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i^2$.

   (c) $Y_i = e^{\beta_0} e^{\beta_1 x_i} X_i^{\beta_2} e^{\epsilon_i}$.

   (d) $Y_i = \beta_0 + \beta_0 e^{\beta_1} x_i + \epsilon_i$.

   (e) $Y_i = \sqrt{\beta_0 + \beta_1 x_i + \epsilon_i}$.

   (f) $1/Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$.

2. Considere o modelo linear por partes

$$Y = \begin{cases} \alpha_0 + \alpha_1 x + \epsilon & \text{se} \quad x \leq x_0 \\ \beta_0 + \beta_1 x + \epsilon & \text{se} \quad x \geq x_0 \end{cases} \tag{1.13}$$

   Mostre que se x_0 é conhecido, este modelo pode ser escrito como um modelo de regressão linear com uma escolha apropriada da variável explicativa.

3. Observe que o modelo por partes em (1.13) pode ser descontínuo em x_0. Esta descontinuidade desaparece impondo a restrição $\beta_0 - \alpha_0 = (\alpha_1 - \beta_1)x_0$. Escreva este modelo de regressão por partes contínuo como um modelo de regressão linear com uma escolha adequada da variável explicativa.

4. Considere o modelo de regressão quadrático em x

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon,$$

   onde $\mathrm{Var}(\epsilon) = \sigma^2$ com observações independentes. Se $\beta_2 > 0$, encontre o valor de x que minimize a resposta esperada.

5. Seja Y um vetor aleatório de média μ e variância $\sigma^2\,\mathrm{I}$. Prove que

$$\mathrm{E}\big((Y-\mu)^\top G(Y-\mu)\big) = \sigma^2\,\mathrm{tr}(G),$$

   onde G é uma matriz quadrada.

6. Prove as principais propriedades da função traço:

   (a) O traço é comutativo, isto é, $\mathrm{tr}(AB) = \mathrm{tr}(BA)$,

   (b) O traço é linear, assim $\mathrm{tr}(A+B) = \mathrm{tr}(A) + \mathrm{tr}(B)$,

   (c) $\mathrm{tr}(\lambda A) = \lambda\,\mathrm{tr}(A)$, sendo λ um escalar,

   (d) $\mathrm{tr}(A) = \mathrm{tr}(A^\top)$ se A é uma matriz quadrada,

   (e) $Y^\top G Y = \mathrm{tr}(GYY^\top)$, onde Y é um vetor coluna.

7. Prove as principais propriedades da função posto:

   (a) $\mathrm{posto}(AA^\top) = \mathrm{posto}(A)$, para quaisquer matriz quadrada A,

   (b) Seja A uma matriz $m \times n$ de posto m e S uma matriz $r \times m$ de posto r. Então $\mathrm{posto}(SA) = r$,

   (c) $\mathrm{tr}(A) = \mathrm{posto}(A)$, se A é uma matriz idempotente, isto é, se $A = AA$,

8. Suponhamos que $\lambda_1, \lambda_2, \cdots, \lambda_n$ sejam os valores próprios ou autovalores de uma matriz A. Prove que

   (a) $\mathrm{tr}(A) = \lambda_1 + \lambda_2 + \cdots + \lambda_n$,

   (b) $|A| = \lambda_1 \lambda_2 \cdots \lambda_n$.

9. Seja A uma matriz quadrada de ordem p. Prove que $|A| = 0$ se, e somente se, $\mathrm{posto}(A) \neq p$.

10. Seja A uma matriz simétrica e idempotente. Prove que
    (a) Se $|A| \neq 0$, então $A = \mathrm{I}$, onde I é a matriz identidade,
    (b) Todo autovalor de A ou é 0 ou é 1.
11. Seja M uma matriz tal que $\text{posto}(M) = p$. Prove que
    (a) A matriz $M^\top M$ é simétrica,
    (b) $\text{posto}(M^\top M) = p$,
    (c) A matriz $M^\top M$ é definida semi-positiva.
12. Seja P uma matriz e P^{-1} sua inversa, ou seja, $P^{-1}P = PP^{-1} = \mathrm{I}$. Prove que
    (a) $|P^{-1}| = 1/|P|$,
    (b) $(P^{-1})^{-1} = P$,
    (c) $(P^\top)^{-1} = (P^{-1})^\top$,
    (d) Se a matriz P for simétrica, isto é, se $P^\top = P$ então $P^{-1} = (P^{-1})^\top$,
    (e) Se P é ortogonal, isto é, P é uma matriz tal que $P^\top P = PP^\top = \mathrm{I}$ então $P^\top = P^{-1}$,
13. Seja $f(\beta)$ uma função escalar de argumento o vetor coluna $\beta = (\beta_1, \cdots, \beta_n)^\top$ e denote por $\partial f(\beta)/\partial\beta$ o vetor coluna de derivadas $(\partial f(\beta)/\partial\beta_1, \cdots, \partial f(\beta)/\partial\beta_n)$. Similarmente, definimos $\partial^2 f(\beta)/\partial\beta\partial\beta^\top$ como a matriz de segundas derivadas $(\partial^2 f(\beta)/\partial\beta_i\partial\beta_j)$. Prove as seguintes propriedades de derivadas de vetores.
    (a) $\dfrac{\partial}{\partial\beta}(M^\top\beta) = M$ e $\dfrac{\partial^2}{\partial\beta\partial\beta^\top}(M^\top\beta) = 0$,
    (b) $\dfrac{\partial}{\partial\beta}(\beta^\top AZ) = AZ$ e $\dfrac{\partial^2}{\partial\beta\partial\beta^\top}(\beta^\top AZ) = 0$,
    (c) $\dfrac{\partial}{\partial\beta}\beta^\top\beta = 2\beta$ e $\dfrac{\partial^2}{\partial\beta\partial\beta^\top}\beta^\top\beta = 2\,\mathrm{I}$,
    (d) $\dfrac{\partial}{\partial\beta}(\beta^\top A\beta) = 2A\beta$ e $\dfrac{\partial^2}{\partial\beta\partial\beta^\top}(\beta^\top A\beta) = 2A$, sendo A uma matriz simétrica.
14. Seja $f(B)$ uma função escalar de argumento a matriz $B = (b_{ij})$ e denote por $\partial f(B)/\partial B$ a matriz de derivadas $(\partial f(B)/\partial b_{ij})$. Prove as seguintes propriedades de derivadas de matrizes.
    (a) $\dfrac{\partial}{\partial B}(Y^\top BZ) = YZ^\top$,
    (b) $\dfrac{\partial}{\partial B}\,\mathrm{tr}(B) = \mathrm{I}$,
    (c) $\dfrac{\partial}{\partial B}|B| = |B|[2B^{-1} - \,\mathrm{diag}(B^{-1})]$, se B é simétrica,
    (d) $\dfrac{\partial}{\partial B}\,\mathrm{tr}(BC) = C + C^\top - \,\mathrm{diag}(C)$, se B é simétrica,
    (e) $\dfrac{\partial}{\partial B}(Y^\top BY) = 2YY^\top - \,\mathrm{diag}(YY^\top)$, se B é simétrica.

Exercícios da Seção 1.3

1. Considere $Y_1, Y_2, \cdots, Y_n$ uma amostra aleatória de tamanho n da distribuição normal $N(\mu, \sigma^2)$. Encontre a média e a variância das seguintes funções:
    (a) $\overline{Y} = \dfrac{1}{n}\sum_{i=1}^n Y_i$.
    (b) Y_1.
    (c) $\dfrac{Y_1}{2} + \dfrac{1}{2(n-1)}\sum_{i=2}^n Y_i$.

2. Seja $Y \sim N(15, 16)$. Encontre os valores das probabilidades: $P(Y \leq 12)$, $P(10 \leq Y \leq 17)$, $P(|Y - 15| \geq 0.5)$ e $P(10 \leq Y \leq 19|Y \leq 17)$.

3. Se $Y \sim N(-1, 9)$, encontre os valores y tais que: (a) $P(Y \geq y) = 0.38$ e (b) $P(|Y + 1| < y) = 0.4$.

4. A resistência à compressão de amostras de cimento pode ser modelada por uma distribuição normal com média de 6.000 quilogramas por centímetro quadrado e um desvio padrão de 100 quilogramas por centímetro quadrado.

   (a) Calcule a probabilidade de que a resistência à compressão da amostra seja menor do que 6.250 Kg/cm^2?

   (b) Qual é a probabilidade da resistência da amostra estar entre 5.800 e 5.900 Kg/cm^2?

   (c) Que resistência é excedida por 95% das amostras?

5. Sejam $X_1, \cdots, X_n$ variáveis aleatórias independentes cada uma com distribuição $X_k \sim N(\mu_k, \sigma_k^2)$, $k = 1, 2, \cdots, n$. Prove que $S_n = \sum_{k=1}^n X_k$ é uma variável aleatória normal com distribuição $N(\sum_{k=1}^n \mu_k, \sum_{k=1}^n \sigma_k^2)$.

6. Se $X_1, X_2, \cdots, X_n$ são variáveis aleatórias independentes com distribuição normal padrão $N(0, 1)$, prove então que $n^{-1/2}S_n$ também tem distribuição normal padrão $N(0, 1)$.

7. Sejam X e Y variáveis aleatórias independentes. Prove que $X + Y$ é normalmente distribuída se, e somente se, X e Y são ambas normais.

8. Sejam X e Y variáveis aleatórias independentes normal padrão. Prove que $X+Y$ e $X-Y$ são independentes.

9. Considere as variáveis aleatórias X e Y independentes com a mesma distribuição de variância finita e também, considere $Z_1 = X + Y$ e $Z_2 = X - Y$ independentes. Demonstre que todas as variáveis aleatórias X, Y, Z_1 e Z_2 são normalmente distribuídas.

10. Prove que se as variáveis aleatórias normais $X_1, X_2, \cdots, X_n$ são tais que

$$\sum_{k=1}^n a_k b_k \text{Var}(X_k) = 0,$$

    então as variáveis aleatórias $L_1 = \sum_{k=1}^n a_k X_k$ e $L_2 = \sum_{k=1}^n b_k X_k$ são independentes. Os números $a_1, \cdots, a_n$ e $b_1, \cdots, b_n$ são reais fixos não negativos.

11. Sejam as variáveis aleatórias $X_1, X_2, \cdots, X_n$ independentes, os números $a_1, \cdots, a_n$ e $b_1, \cdots, b_n$ reais fixos, nenhum deles zero, e as formas lineares $L_1 = \sum_{k=1}^n a_k X_k$ e $L_2 = \sum_{k=1}^n b_k X_k$ independentes. Prove então que todas as variáveis aleatórias são normalmente distribuídas.

12. Sejam $X_1, X_2, \cdots, X_n$ variáveis aleatórias independentes e igualmente distribuídas com variância finita. Prove que a distribuição de probabilidades destas variáveis é normal se, e somente se,

$$S_n = \sum_{k=1}^n X_k \qquad \text{e} \qquad Y_n = \sum_{k=1}^n (X_k - n^{-1}S_n)^2,$$

    são independentes.

13. Sejam Y e Z variáveis aleatórias independentes, cada uma com distribuição normal. Prove que $Y + Z$ e $Y - Z$ são independentes se, e somente se, $\text{Var}(Y) = \text{Var}(Z)$.

14. Sejam Y e Z variáveis aleatórias independentes com distribuição normal padrão. Mostre que $U = (Y+Z)/\sqrt{2}$ e $V = (Y - Z)/\sqrt{2}$ também são independentes e com distribuição normal padrão.

15. As variáveis aleatórias Y e Z são independentes, identicamente distribuídas $N(0, \sigma^2)$.

    (a) Mostre que $Y^2 + Z^2$ e $\dfrac{Y}{\sqrt{Y^2 + Z^2}}$ são independentes.

    (b) Seja

$$\theta = \text{sen}^{-1}\left(\frac{Y}{\sqrt{Y^2 + Z^2}}\right).$$

    Mostre que θ tem distribuição uniforme $U(-\pi/2, \pi/2)$.

16. Seja $(Y, Z) \sim N(\mu, \Sigma)$, um vetor de variáveis aleatórias com densidade normal onde $\mu = (1, 2)$ e

$$\Sigma = \begin{pmatrix} 4 & 0.5 \\ 0.5 & 1 \end{pmatrix}.$$

Encontre:

(a) $P(Y + 2Z \leq 4)$, $P(Y \leq 2|Z = 1)$ e $P(5 \leq Z \leq 11|Y = 2)$.

(b) a função de distribuição conjunta de $Y + 2Z$ e $3Z - 2Y$.

(c) a distribuição de $Y - Z$.

17. Sejam $Y_i \sim N(\beta_0 + \beta_1 z_i, \sigma^2)$, variáveis aleatórias independentes, $i = 1, \cdots, n$; sendo $z_1, \cdots, z_n$ números conhecidos tais que $\sum_{i=1}^n z_i = 0$. Encontre:

(a) a função de distribuição conjunta do vetor $(Y_1, \cdots, Y_n)$.

(b) a função de distribuição conjunta de $\overline{Y}$ e $W = \sum_{i=1}^n z_i Y_i / \sum_{i=1}^n z_i^2$, quando $\sum_{i=1}^n z_i^2 > 0$.

18. Sejam X e Y variáveis aleatórias tais que $\mathrm{E}(X^2)$ e $\mathrm{E}(Y^2)$ existam e sejam $U = \beta_0 + \beta_1 X$ e $V = \alpha_0 + \alpha_1 Y$. Prove que

$$\rho_{X,Y} = \pm \rho_{U,V},$$

onde $\rho_{X,Y}$ e $\rho_{U,V}$ são, respectivamente, os coeficientes de correlação entre X e Y e U e V.

19. Considere ρ o coeficiente de correlação entre as variáveis aleatórias X e Y, ambas com média 0 e variância 1. Prove que

$$\mathrm{E}\big(\max(X^2, Y^2)\big) \leq 1 + \sqrt{1 - \rho^2}.$$

20. Os vetores $\begin{pmatrix} Y_i \\ Z_i \end{pmatrix}$ são independentes e cada um distribuído segundo a densidade

$$N\left(\begin{pmatrix} \mu \\ \nu \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{YZ} \\ \sigma_{YZ} & \sigma_Z^2 \end{pmatrix}\right), \qquad i = 1, 2, 3.$$

Encontre:

(a) a função de distribuição conjunta das seis variáveis aleatórias.

(b) a função de distribuição conjunta de $\begin{pmatrix} \overline{Y} \\ \overline{Z} \end{pmatrix}$.

21. Considere um vetor $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$ de variáveis aleatórias com distribuição

$$N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_{Y_1}^2 & \rho\sigma_{Y_1 Y_2} \\ \rho\sigma_{Y_1 Y_2} & \sigma_{Y_2}^2 \end{pmatrix}\right).$$

Definamos $Z_i = Y_i/\sigma_{Y_i}$ para $i = 1, 2$. Prove que $\mathrm{Var}(Z_1 - Z_2) = 2(1 - \rho)$. Como é que isto sugere que ρ é uma medida de associação entre Y_1 e Y_2?

22. Prove que, se X e Y são variáveis aleatórias com função de densidade normal bivariada, então X e Y são independentes se, e somente se, $\rho = 0$ (Teorema 1.6).

23. Seja $Z = X + Y$, onde $X \sim N(\mu_1, \sigma_1^2)$ e $Y \sim N(\mu_2, \sigma_2^2)$ e $\mathrm{Corr}(X, Y) = \rho$. Prove que

$$\mathrm{E}(Z) = \mathrm{E}(X) + \mathrm{E}(Y), \qquad \mathrm{Var}(Z) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X, Y),$$

ou seja,

$$\mathrm{E}(Z) = \mu_1 + \mu_2, \qquad \mathrm{Var}(Z) = \sigma_1^2 + \sigma_2^2 + 2\rho\sigma_1\sigma_2.$$

24. Prove que se $Z \sim N_n(\tau, \Gamma)$ e independente de $Y \sim N(\mu, \Sigma)$, então

$$Z + Y \sim N_n(\mu + \tau, \Sigma + \Gamma).$$

Capítulo 2

Estimação e testes de hipóteses

Para encontrar os estimadores dos parâmetros no modelo de regressão podem ser utilizados dois métodos estatísticos, um deles conhecido como de máxima verossimilhança o outro requer a minimização da soma de quadrados. Os estimadores obtidos por estes dois métodos na família exponencial de densidades coincidem (Bradley, 1973).

Figura 2.1: Carl Friedrich Gauss

O Método dos Mínimos Quadrados ou Mínimos Quadrados Ordinários é uma técnica de otimização matemática com propriedades estatísticas importantes que procura encontrar o melhor ajuste para um conjunto de dados tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados, tais diferenças são chamadas resíduos. Consiste em um estimador que minimiza a soma dos quadrados dos resíduos da regressão, de forma a maximizar o grau de ajuste do modelo aos dados observados.

Um requisito para o método dos mínimos quadrados é que o fator imprevisível ou erro seja distribuído aleatoriamente, essa distribuição seja normal e independente. Outro requisito é que o modelo é linear nos parâmetros, ou seja, as variáveis apresentam uma relação linear entre si. Caso contrário, deveria ser usado um modelo de regressão não-linear. Uma propriedade importante é que este estimador é não-enviesado de mínima variância linear na variável resposta.

Credita-se a Carl Friedrich Gauss como desenvolvedor das bases fundamentais do método dos mínimos quadrados, em 1795, quando Gauss tinha apenas dezoito anos. Entretanto, Adrien-Marie Legendre[1] foi o primeiro a publicar o método em 1805. Gauss publicou suas conclusões em 1809.

O procedimento de estimação por máxima verossimilhança foi vastamente popularizado por Ronald Aylmer Fisher nas primeiras décadas do século XX. Este procedimento fornece uma abordagem unificada para encontrar estimativas bem definidas no caso da distribuição normal e muitos outros problemas.

[1] Adrien-Marie Legendre (1752-1833) foi um matemático francês que fez importantes contribuições à estatística, teoria dos números, álgebra e análise matemática.

2.1 Estimação

A mais antiga forma de estimação de um modelo de regressão foi o método dos mínimos quadrados, publicado por Legendre em 1805 e por Gauss em 1809. Embora Legendre publicou primeiro artigo mostrando este método, Gauss publicou um maior desenvolvimento da teoria da mínimos quadrados em 1821.

Primeiro mostraremos a forma de estimação de mínimos quadrados, depois estudaremos o procedimento de estimação mais universalmente utilizado na estatística conhecido como máxima verossimilhança. Por último conheceremos testes de hipóteses que nos permitem validar afirmações acerca de valores dos parâmetros.

2.1.1 Estimador de mínimos quadrados

A teoria dos mínimos quadrados está preocupada em encontrar estimadores dos parâmetros em um modelo de regressão linear. Neste problema, para encontrar os estimadores, não é necessário fazer nenhuma exigência quanto a distribuição de probabilidades da variável resposta Y, somente vamos supor que a distribuição de Y satisfaça um modelo linear.

Definição 2.1. *Seja Y um vetor aleatório com distribuição de probabilidade satisfazendo um modelo linear da forma matricial (1.2), isto é,*

$$Y = X\beta + \epsilon,$$

onde $\mathrm{E}(\epsilon_i) = 0$ e $\mathrm{Var}(\epsilon_i) = \sigma^2$ para $i = 1, \cdots, n$. Dizemos que $\widehat{\beta}$ é o estimador de mínimos quadrados de β se minimiza a soma de quadrados

$$\sum_{i=1}^{n} \epsilon_i^2 = \epsilon^\top \epsilon = (Y - X\beta)^\top (Y - X\beta)\cdot \tag{2.1}$$

Observemos que

$$(Y - X\beta)^\top (Y - X\beta) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_{p-1} x_{ip-1})^2,$$

é a soma dos quadrados das diferenças entre as observações e as expectativas; da qual, por diferenciação, se obtém equação normal,

$$X^\top X\widehat{\beta} = X^\top Y\cdot \tag{2.2}$$

Teorema 2.1. *Seja $\widehat{\beta}$ uma solução da equação normal (2.2). Então*

$$(Y - X\beta)^\top (Y - X\beta) \geq (Y - X\widehat{\beta})^\top (Y - X\widehat{\beta})\cdot$$

Demonstração.

$$\begin{aligned}
(Y - X\beta)^\top (Y - X\beta) &= [Y - X\widehat{\beta} + X(\widehat{\beta} - \beta)]^\top [Y - X\widehat{\beta} + X(\widehat{\beta} - \beta)] \\
&= (Y - X\widehat{\beta})^\top (Y - X\widehat{\beta}) + (\widehat{\beta} - \beta)^\top X^\top X(\widehat{\beta} - \beta) \\
&\geq (Y - X\widehat{\beta})^\top (Y - X\widehat{\beta})\cdot
\end{aligned}$$

□

Este teorema mostra que o mínimo de $(Y - X\beta)^\top(Y - X\beta)$ é $(Y - X\widehat{\beta})^\top(Y - X\widehat{\beta})$, sendo alcançado em $\beta = \widehat{\beta}$ e que é único para todas as soluções $\widehat{\beta}$ de (2.2). Encontremos agora as equações normais e suas soluções.

Teorema 2.2. *Seja Y um vetor aleatório com distribuição de probabilidade satisfazendo um modelo linear da forma matricial. Se $\widehat{\beta}$ é o estimador de mínimos quadrados de β, então $\widehat{\beta}$ satisfaz as equações normais*

$$X^\top X\widehat{\beta} = X^\top Y\cdot$$

Demonstração. Seja $f(\beta) = (Y - X\beta)^\top(Y - X\beta)$ uma função escalar de argumento vetorial, então

$$\frac{\partial f(\beta)}{\partial \beta} = 2X^\top X\beta - 2X^\top Y\cdot$$

Veja o exercício 13 do Capítulo 1. A matriz de segundas derivadas, conhecida como matriz hessiana, é $\partial^2 f(\beta)/\partial\beta\partial\beta^\top = 2X^\top X$ e se $X^\top X$ for definida positiva (Luenberger, 1984), o ponto $\widehat{\beta}$ no qual

$$\left.\frac{\partial f(\beta)}{\partial \beta}\right|_{\beta=\widehat{\beta}} = 0,$$

é um mínimo estrito de $f(\beta)$ e satisfaz as equações normais. □

Exemplo 2.1. *No modelo de regressão linear simples, no qual*

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i,$$

para $i = 1, \cdots, n$ a soma de quadrados $(Y - X\beta)^\top(Y - X\beta)$ assume a forma

$$f(\beta) = (Y - X\beta)^\top(Y - X\beta) = \sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 x_i)^2\cdot$$

Derivando esta expressão em relação aos parâmetros de regressão temos

$$\frac{\partial f(\beta)}{\partial \beta_0} = n\overline{Y} - n\beta_0 - n\overline{x}\beta_1$$

e

$$\frac{\partial f(\beta)}{\partial \beta_1} = \sum_{i=1}^{n} x_i Y_i - n\overline{x}\beta_0 - \sum_{i=1}^{n} x_i^2\beta_1\cdot$$

Igualando estas derivadas a zero obtemos os estimadores de mínimos quadrados

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x} \tag{2.3}$$

e

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i Y_i - n\overline{x}\,\overline{Y}}{\sum_{i=1}^{n} x_i^2 - n\overline{x}^2}\cdot \tag{2.4}$$

2.1.2 Estimador de máxima verossimilhança

Vamos estudar um método frequentemente utilizado para encontrar estimadores o método de estimação de máxima verossimilhança. O princípio da máxima verossimilhança essencialmente pressupõe que a amostra é representativa da população e escolhe como a estimativa o valor do parâmetro que maximiza a função de probabilidade conjunta. Referências importantes desta metodologia de estimação são os livros Rao (1973) e Rohatgi (1976).

Definição 2.2. *Sejam $Y_1, \cdots, Y_n$ variáveis aleatórias com função de densidade ou de probabilidade conjunta $f(y_1, \cdots, y_n; \theta)$, sendo $\theta \in \Theta$. A função*

$$L(\theta; y_1, \cdots, y_n) = f(y_1, \cdots, y_n; \theta),$$

considerada como função de θ é chamada de função de verossimilhança.

Geralmente θ será um vetor de parâmetros. Se $Y_1, \cdots, Y_n$ forem variáveis aleatórias independentes igualmente distribuídas com função de densidade ou de probabilidade $f(y; \theta)$, a função de verossimilhança é

$$L(\theta; \underset{\sim}{y}) = \prod_{i=1}^{n} f(y_i; \theta)\cdot$$

Sendo $\theta \in \mathbb{R}_k$ e $\underset{\sim}{y} = (y_1, \cdots, y_n)$, o vetor de observações.

Definição 2.3. *O princípio da estimação por máxima verossimilhança consiste em escolher como estimador de θ a $\widehat{\theta}$ que maximiza $L(\theta; \underset{\sim}{y})$, isto é, encontrar uma função $\widehat{\theta}$ em $\mathbb{R}_n \to \mathbb{R}_k$ que satisfaz*

$$L(\widehat{\theta}; \underset{\sim}{y}) = \sup_{\theta \in \mathbb{R}_k} L(\theta; \underset{\sim}{y}) \tag{2.5}$$

Se existe um $\widehat{\theta}$ satisfazendo (2.5) chamamos-lhe um estimador de máxima verossimilhança para θ. O método de estimação por máxima verossimilhança tenta encontrar a moda da distribuição, isto é, o valor de θ que maximiza a função de probabilidade conjunta. O fato da moda ser geralmente uma estimativa mais pobre do que a média ou a mediana explica porque as propriedades em amostras pequenas dos estimadores de máxima verossimilhança são, em geral, pobres. Para grandes amostras, no entanto, a moda tende a se aproximar da média, se existir, e da mediana e o método tem muitas propriedades interessantes em grandes amostras (Wilks, 1962).

É conveniente trabalhar com o logaritmo da função de probabilidade. O princípio da estimação por máxima verossimilhança pode ser resumido pela maximização do logaritmo da função de verossimilhança. Dado que a função logaritmo é uma função monótona então

$$\log L(\widehat{\theta}; \underset{\sim}{y}) \;=\; \sup_{\theta \in \mathbb{R}_k} \ell(\theta; \underset{\sim}{y}), \tag{2.6}$$

onde $\ell(\theta; \underset{\sim}{y}) = \log L(\theta; \underset{\sim}{y})$.

Significa que, para encontrar o ponto que maximiza $L(\theta; \underset{\sim}{y})$, somente é necessário encontrar o ponto que maximiza $\ell(\theta; \underset{\sim}{y})$.

Exemplo 2.2. *Sejam $Y_1, \cdots, Y_n$ variáveis aleatórias independentes com distribuição $N(\mu, \sigma^2)$, onde tanto μ quanto σ^2 são desconhecidos. Nesta situação $\Theta = \{(\mu, \sigma^2), \mu \in \mathbb{R}, \sigma^2 > 0\}$. A função de verossimilhança é*

$$L(\mu, \sigma^2; \underset{\sim}{y}) = \frac{1}{\sigma^n (2\pi)^{n/2}} \exp\left[-\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} \right]$$

e o logaritmo da função de verossimilhança é

$$\ell(\mu, \sigma^2; \underset{\sim}{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2}\cdot$$

O procedimento para encontrar os estimadores de máxima verossimilhança requer que Θ seja um subconjunto aberto de $\mathbb{R}_n$ e que $f(y;\theta)$ seja uma função positiva e diferenciável de θ, isto é, que as derivadas parciais de primeira ordem existam nas componentes de θ.

Se existir um ponto $\widehat{\theta}$ tal que $\sup_{\theta\in\mathbb{R}_k}\ell(\theta;\underset{\sim}{y}) = \ell(\widehat{\theta};\underset{\sim}{y})$, então $\widehat{\theta}$ deve satisfazer as equações de verossimilhança

$$\left.\frac{\partial\ell(\theta;\underset{\sim}{y})}{\partial\theta_j}\right|_{\theta=\widehat{\theta}} = 0, \qquad j=1,\cdots,k, \qquad \theta=(\theta_1,\cdots,\theta_k)\cdot \tag{2.7}$$

Qualquer raiz não trivial das equações de verossimilhança (2.7) é chamada de estimador de máxima verossimilhança no sentido amplo. Um valor de parâmetro que fornece o máximo absoluto da função de verossimilhança é chamado de estimador de máxima verossimilhança no sentido estrito ou simplesmente o estimador de máxima verossimilhança.

Exemplo 2.3. *Continuando o exemplo 2.2. As equações de verossimilhança são*

$$\frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{n}(y_i-\widehat{\mu})^2 \;=\; 0 \qquad e \qquad -\frac{n}{2}\frac{1}{\widehat{\sigma}^2} \;+\; \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{n}(y_i-\widehat{\mu})^2 \;=\; 0\cdot$$

Resolvendo a primeira dessas equações para $\widehat{\mu}$ obtemos que $\widehat{\mu} = \overline{y}$ e substituindo na segunda $\widehat{\sigma}^2 = \sum_{i=1}^{n}(y_i-\overline{y})^2/n$. Vemos que $(\widehat{\mu},\widehat{\sigma}^2)\in\Theta$ e que este par maximiza a função de verossimilhança.

2.1.3 Estimadores do modelo de regressão

No caso do modelo de regressão linear simples $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$, onde $\epsilon_1,\cdots,\epsilon_n$ é uma amostra aleatória com distribuição $N(0,\sigma^2)$; podemos construir as equações de verossimilhança em (2.7). O logaritmo da função de verossimilhança nesta situação é

$$\ell(\beta_0,\beta_1,\sigma^2;\underset{\sim}{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i-\beta_0-\beta_1x_i)^2}{2\sigma^2}, \tag{2.8}$$

da qual, derivando em relação a β_0, β_1 e σ^2 obtemos as equações de verossimilhança.

Sendo

$$\frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{n}(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i) = 0,$$

quando derivamos em relação a β_0, obtendo-se que

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}\cdot \tag{2.9}$$

Em relação a β_1 obtemos a equação

$$\frac{1}{\widehat{\sigma}^2}\sum_{i=1}^{n}x_i(y_i-\widehat{\beta}_0-\widehat{\beta}_1x_i) = 0,$$

da qual temos

$$\widehat{\beta}_1 = \big(\sum_{i=1}^{n} x_i y_i - n\overline{xy}\big) \Big/ \big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\big), \tag{2.10}$$

depois de substituirmos a expressão de $\widehat{\beta}_0$ acima. As expressões dos estimadores dos parâmetros da regressão em (2.9) e (2.10), obtidos pelo método de máxima verossimilhança coincidem com aquelas em (2.3) e (2.4) de mínimos quadrados.

Continuando a aplicação do método de máxima verossimilhança temos que, quando derivamos (2.8) em relação a σ^2 obtemos

$$-\frac{n}{2\widehat{\sigma}^2} + \frac{1}{2\widehat{\sigma}^4}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2 = 0,$$

esta equação e substituindo as expressões de $\widehat{\beta}_0$ e $\widehat{\beta}_1$ chegamos a que, o estimador de máxima verossimilhança da variância no modelo de regressão linear simples é

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2.$$

Maiores detalhes acerca deste estimador serão estudados posteriormente.

Exemplo 2.4. *Este exemplo nos permitirá entender melhor os conceitos. Os dados foram coletados por um analista que esta investigando a relação entre o tamanho do lote produzido (variável explicativa) e o custo unitário (variável resposta) numa determinada indústria. Um estudo das operações mais recentes forneceu as seguintes observações:*

Produção	*100*	*120*	*140*	*160*	*180*	*200*	*220*	*240*	*260*	*280*	*300*
Custo	*9.73*	*9.61*	*8.15*	*6.98*	*5.87*	*4.98*	*5.09*	*4.79*	*4.02*	*4.46*	*3.82*

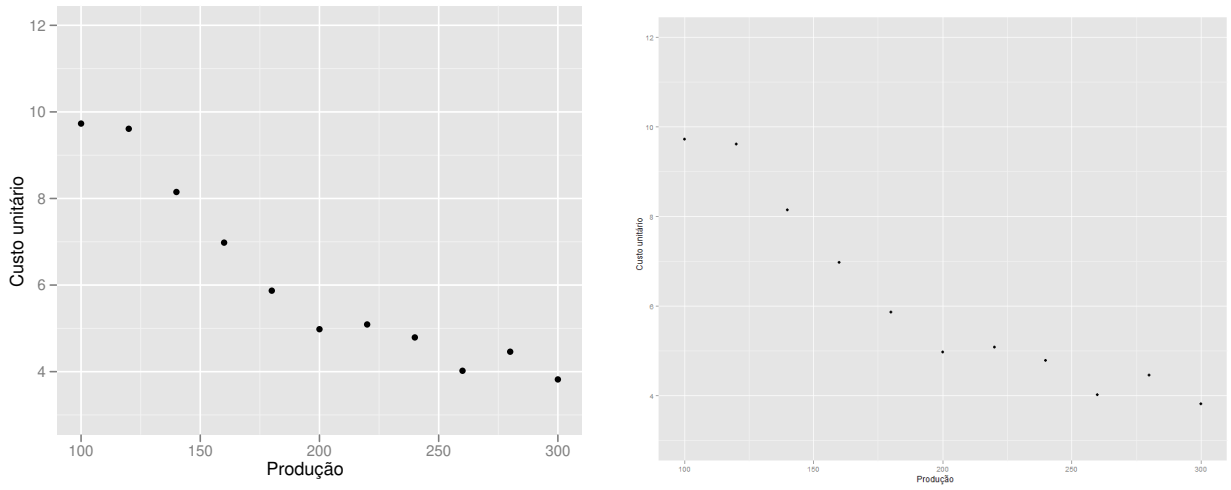

Figura 2.2: Gráficos de dispersão.

A Figura 2.3 mostra o chamado gráfico de dispersão, nele podemos perceber que deve existir relação linear negativa entre o tamanho do lote produzido (Produção) e o custo unitário (Custo). Isso significa que, a maior o lote de produção menor será o custo de produção. Este fato implica que

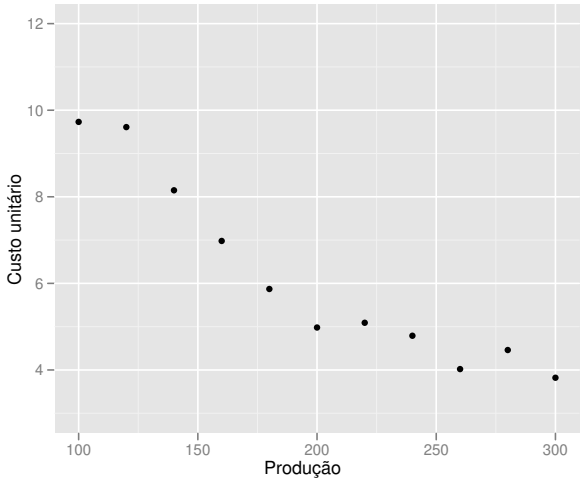

Figura 2.3: Dispersão dos dados no Exemplo **??**.

devemos esperar um valor negativo, se significativo for, da estimativa do coeficiente que acompanha a variável Produção. O modelo proposto é

$$\text{Custo}_i = \beta_0 + \beta_1 \text{Produção}_i + \epsilon_i,$$

$i = 1, \ldots, 11$.

Assumiremos que as suposições para o modelo de regressão linear, apresentadas na Seção 1.1, são satisfeitas do qual temos que

$$\text{E}(\text{Custo}_i) = \beta_0 + \beta_1 \text{Produção}_i.$$

As seguintes linhas de comando `R` nos permitem encontrar as estimativas dos parâmetros da regressão:

```
> Custo = c(9.73,9.61,8.15,6.98,5.87,4.98,5.09,
              4.79,4.02,4.46,3.82)
> Producao = c(100,120,140,160,180,200,220,240,
              260,280,300)
> n = length(Custo)
> y.media=mean(Custo)
> x.media=mean(Producao)
> sum.xy=sum(Custo*Producao)
> sum.x.quadrado=sum(producao^2)
> beta.0=y.media-x.media*beta.1
> beta.1=(sum.xy-n*y.media*x.media)
               /(sum.x.quadrado-n*x.media^2)
> beta.0
[1] 12.29091
> beta.1
[1] -0.03077273
```

Como resposta as linhas de comando acima nos retornam os valores $\widehat{\beta}_0 = 12.29091$ e $\widehat{\beta}_1 = -0.03077273$. Uma outra maneira seria utilizar a função `lm` (Fitting Linear Models) da seguinte forma:

```
> ajuste=lm(Custo~Producao)
> summary(ajuste)

Call:
lm(formula = Custo ~ Producao)

Residuals:
Min      1Q  Median      3Q     Max
-1.1564 -0.4091 -0.1154  0.6386  1.0118

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.290909   0.749146  16.407 5.17e-08 ***
Producao    -0.030773   0.003571  -8.616 1.22e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7491 on 9 degrees of freedom
Multiple R-squared:  0.8919,Adjusted R-squared:  0.8799
F-statistic: 74.24 on 1 and 9 DF,  p-value: 1.218e-05
```

Aqui a função `summary` retorna o resumo da estimação dos parâmetros da regressão e outras informações, que serão estudadas posteriormente. Interessa-nos agora o valor em `Estimate`, isto é, as estimativas sendo `(Intercept)` a correspondente a $\widehat{\beta}_0$. A significância dos parâmetros da regressão sera objeto de estudo na Seção 2.3. Assumindo então que o modelo estimado é significativo então a média estimada do custo unitário é a seguinte função linear do lote de produção

$$\widehat{\mathrm{E}(\mathrm{Custo})} = 12.290909 - 0.030773\,\mathrm{Produção},$$

independentemente do método de estimação escolhido.

Agora podemos dizer que, considerando influente o tamanho do lote produzido no custo, cada unidade a mais produzida reduz o custo em 0.03 unidades aproximadamente. Assim se o lote produzido é de tamanho 101 unidades o custo estimado é de $12.290909 - 0.030773 \times 101 = 9.182864$.

Devemos observar que os nomes das variáveis no `R` somente podem conter caracteres específicos da língua inglesa, assim acentos, tildes e outros caracteres específicos do português não podem ser utilizados durante a programação nessa linguagem. Por isso, a variável Produção se escreve Producao. As linhas de comando mostradas aqui e outras estão disponíveis no arquivo `Custo.R`.

Estimadores do modelo de regressão linear múltipla

Caso o modelo de regressão não seja simples, isto é, contenha mais do que uma variável explicativa recorremos à notação matricial para encontrar os estimadores do vetor de parâmetros da regressão $\widehat{\beta}$ e da variância $\widehat{\sigma}^2$.

Se o vetor de variáveis aleatórias $Y = (Y_1, \cdots, Y_n)^\top$ tem função de densidade normal multivariada, o logaritmo da função de verossimilhança é da forma

$$\ell(\beta, \Sigma; \underset{\sim}{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\underset{\sim}{y} - X\beta)^\top \Sigma^{-1}(\underset{\sim}{y} - X\beta),$$

onde $\underset{\sim}{y} = (y_1, \cdots, y_n)^\top$ é o vetor de observações. Acontece que no modelo de regressão linear múltipla, $\Sigma = \sigma^2\,\mathrm{I}$, devemos lembrar que $\mathrm{Var}\{Y\} = \sigma^2$ e que as componentes $Y_1, \cdots, Y_n$ do vetor Y são independentes. Então

$$\ell(\beta, \sigma^2; \underset{\sim}{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\underset{\sim}{y} - X\beta)^\top(\underset{\sim}{y} - X\beta)\cdot \tag{2.11}$$

Devemos agora encontrar $\widehat{\beta}$ e $\widehat{\sigma}^2$ que maximizem (2.11).

Teorema 2.3. *Seja Y um vetor aleatório satisfazendo um modelo de regressão linear múltipla. Então o estimador de máxima verossimilhança do vetor β é solução do sistema de equações normais*

$$X^\top X\widehat{\beta} = X^\top Y\cdot$$

Demonstração. Ver Teorema 2.2. □

No caso de $\widehat{\sigma}^2$ devemos derivar (2.11) em relação a σ^2, do qual obtemos

$$-\frac{n}{2\sigma^2} - \frac{1}{2\sigma^4}(\underset{\sim}{y} - X\widehat{\beta})^\top(\underset{\sim}{y} - X\widehat{\beta}),$$

uma vez avaliado o logaritmo da função de verossimilhança em $\widehat{\beta}$. O estimador $\widehat{\sigma}^2$ se obtém da solução da equação de verossimilhança (ver 2.7) para σ^2, da qual chegamos a que

$$\widehat{\sigma}^2 = \frac{1}{n}(\underset{\sim}{y} - X\widehat{\beta})^\top(\underset{\sim}{y} - X\widehat{\beta})\cdot$$

Estimadores dos parâmetros da regressão

Para completar o estudo da estimação dos parâmetros de um modelo de regressão faltaria-nos encontrar soluções para as equações normais. Sabemos dos teoremas 2.2 e 2.3 que $\widehat{\beta}$ seria uma solução das equações normais. Vejamos qual.

Teorema 2.4. *Seja Y um vetor satisfazendo o modelo de regressão linear $Y = X\beta + \epsilon$. Se X é tal que* posto$(X) = p$*, onde p é o número de parâmetros da regressão. Então, o sistema de equações normais*

$$X^\top X\widehat{\beta} = X^\top Y,$$

tem como solução única

$$\widehat{\beta} = (X^\top X)^{-1}X^\top Y\cdot \tag{2.12}$$

Demonstração. O exercício 11, item (a) do Capítulo 1 garante que $X^\top X$ é uma matriz simétrica e do item (b) do mesmo exercício sabemos que posto$(X^\top X) = p + 1$. Sabemos também que toda matriz quadrada de posto completo[2] tem inversa única (Rao, 1973, Pg.16). □

Devemos mencionar que a suposição de posto$(X^\top X) = p + 1$ no teorema anterior é muito importante. Na demonstração foi utilizada esta suposição para justificar a existência e unicidade da inversa da matriz $X^\top X$. Agora, o que aconteceria se isso não fosse cumprido? isto é, e se posto$(X^\top X) < p + 1$ o sistema de equações normais teria solução? se sim, como encontrá-la?

Acontece que a resposta a estas perguntas é positiva, ou seja, caso a matriz X não seja de posto completo o sistema de equações normais ainda teria solução mas não única. Para encontrarmos uma solução do sistema de equações normais temos dois caminhos: (a) impor restrições nos parâmetros da regressão ou (b) utilizar a chamada inversa generalizada.

Uma aplicação clássica de modelos de regressão com certas restrições nos parâmetros de regressão são os chamados modelos de Análise de Variância. Estes modelos são utilizados, dentre muitas outras situações, em planejamento de experimentos. Em geral, impor restrições nos parâmetros não faz sentido em modelos de regressão, por isso restrições nos parâmetros de regressão são impostas em situações especiais (Dykstra, 1983).

[2]Uma matriz quadrada A de ordem $p + 1$ é dita de posto completo se posto$(A) = p + 1$.

Uma outra solução, como mencionado, porém geral seria utilizar as chamadas matrizes inversas generalizadas. Uma matriz quadrada A^- é dita ser inversa generalizada de A se $AA^-A = A$. A definição e propriedades destas matrizes podem ser encontradas nos artigos de Moore (1986) e Penrose (1955). Uma deficiência neste caso é que a inversa generalizada não é única; isso implica que desta maneira teríamos diversas soluções para o sistema de equações normais. Uma outra dificuldade é que, na tentativa de providenciar solução única para o sistema de equações normais, podemos definir de diversas maneiras matrizes que podem ser consideradas como inversas generalizadas. Excelentes discussões sobre diversas definições e propriedades das matrizes inversas generalizadas pode ser encontrado em Rao (1973) e Wetherill (1986).

Como podemos perceber, em situações práticas, se a matriz $X^\top X$ for singular e, portanto, de determinante zero não podendo calcular a matriz inversa da forma tradicional, enfrentamos um problema sério de difícil solução. Uma forma de resolver esse problema é utilizar soluções para a multicolinearidade, como é conhecido na literatura estatística essa dificuldade. Enquanto isso, consideraremos que a solução do sistema de equações normais existe e é única.

No R (Team, 2024) a função `lm` permite-nos ajustar o modelo proposto, isto é, encontrar os estimadores dos parâmetros da regressão e da variância assim como verificar a influência das variáveis regressoras e analisar a qualidade do ajuste.

A função `lm` é a mais simples das diversas funções R disponíveis para ajustar estes modelos, outras que podem ser igualmente utilizadas são: `glm`, `gamlss`, `lme`, `nlme`, `gam`, dentre muitas outras. A diferença básica entre estas funções é que em `lm` os estimadores se obtém por mínimos quadrados e em todas as outras funções o procedimento de estimação é por máxima verossimilhança. As estimativas podem, eventualmente, serem diferentes embora os estimadores são iguais.

Exemplo 2.5. *O conjunto de dados apresentado em `Health.csv` foi retirado dos registros de saúde de 30 funcionários que eram membros regulares do clube de saúde de uma empresa (Chatterjee & Hadi, 1988).*

As variáveis medidas neste estudo foram:

- Peso: peso em Kg;
- Pulso: pulso em repouso;
- Pernas: força de braço e perna (Kg que cada empregado foi capaz de levantar);
- Treino: tempo (em segundos) de um treinamento de 250mts;
- Tempo: tempo (em segundos) em uma corrida de 1,6Km.

O objetivo do estudo é tentar prever o tempo numa prova de corrida a partir de informações simples do condicionamento físico e do tempo numa corrida de treino. Então o modelo proposto é:

$$\text{Tempo} = \beta_0 + \beta_2\text{Peso} + \beta_2\text{Pulso} + \beta_3\text{Pernas} + \beta_4\text{Treino} + \epsilon,$$

satisfazendo as condições de um modelo de regressão linear.

A Figura 2.4 mostra o gráfico de dispersão entre cada uma das variáveis incluídas no estudo. Isto na diagonal inferior do gráfico. A diagonal superior mostra os valores estimados do coeficiente de correlação linear assim, por exemplo, a correlação linear entre `Peso` e `Pernas` é de 0.74 a qual consideramos forte e a correlação entre `Pulso` e `Perna` é de 0.06, aproximadamente zero. A diagonal principal mostra os nomes e os histogramas das variáveis. Podemos perceber que a resposta `Tempo` está mais correlacionada com o valor observado no `Treino` e o `Peso` do participante no estudo. A referida figura foi construída fazendo uso dos comandos listados abaixo.

```
> library(psych)
> par(mar=c(5,4,1,1),pch=19,cex.axis=0.6)
> pairs.panels(cbind(Peso,Pulso,Pernas,Treino,Tempo),smooth=F,cex=1.5,ellipses=F,pch=20)
```

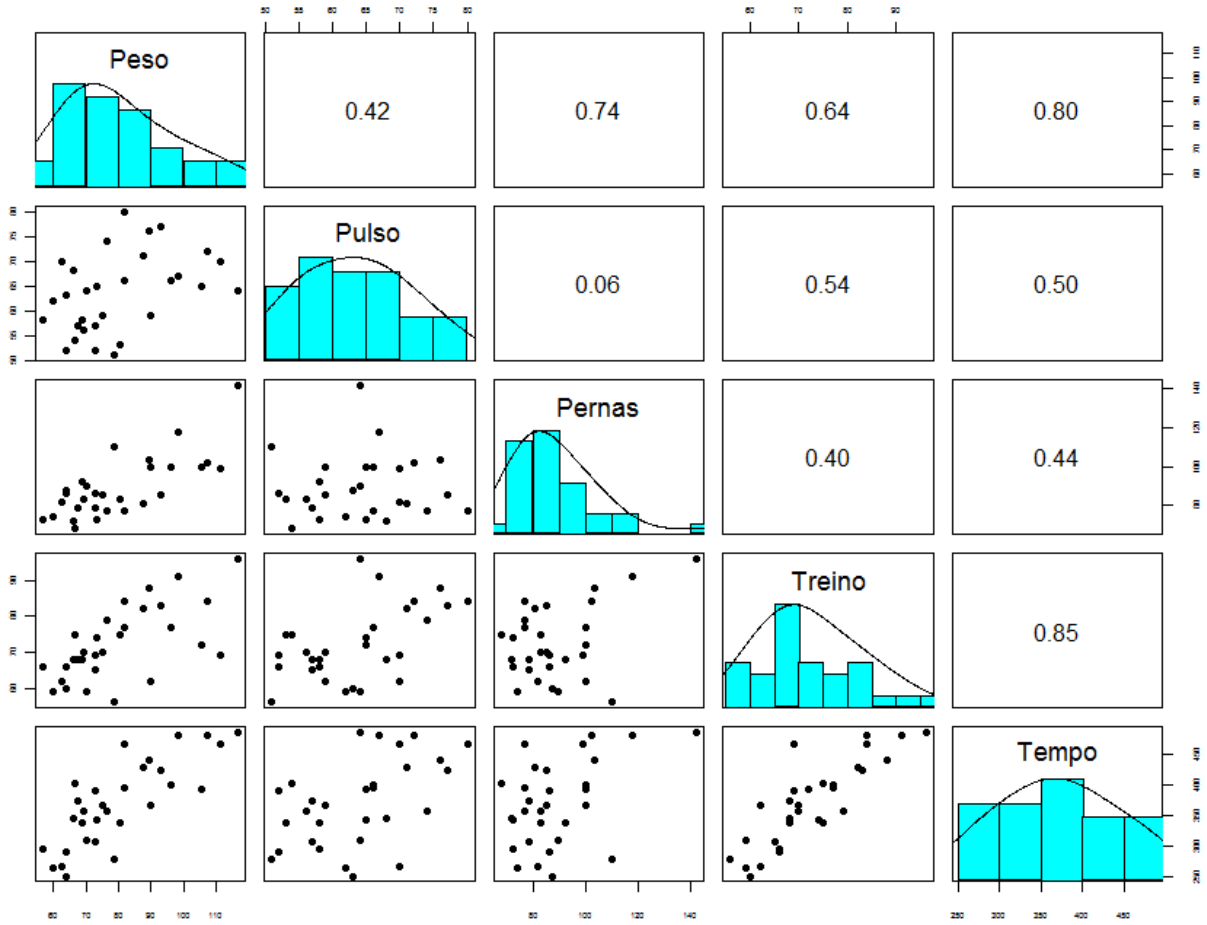

Figura 2.4: Gráficos de dispersão na parte triangular inferior da figura. Histograma de cada variável na diagonal principal e na parte triangular superior os coeficientes de correlação entre as diferentes variáveis observadas.

As linhas de comando

```
> ajuste=lm(Tempo~.,data=Health)
> summary(ajuste)
```

fornecem o resultado do ajuste do modelo de regressão proposto que, resumidamente, apresentamos a seguir.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.6186    56.1027  -0.064 0.949086
Peso          2.7947     0.6324   4.419 0.000168 ***
Pulso        -0.5252     0.8628  -0.609 0.548194
Pernas       -1.1134     0.5422  -2.054 0.050614 .
Treino        3.9030     0.7477   5.220 2.11e-05 ***
```

Até o momento podemos dizer que o modelo ajustado é:

$$\widehat{\mathrm{E}(\mathrm{Tempo})} = -3.6186 + 2.7947\,\mathrm{Peso} - 0.5252\,\mathrm{Pulso} - 1.1134\,\mathrm{Pernas} + 3.9030\,\mathrm{Treino}.$$

2.1.4 Propriedades dos estimadores

Propriedades do estimador do vetor de parâmetros da regressão $\widehat{\beta}$ e do estimador da variância $\widehat{\sigma}^2$ serão estudadas aqui. Sabemos que os estimadores dos parâmetros da regressão são soluções das equações normais, únicas ou não, não importando o método de estimação escolhido.

Assim, as estimativas do vetor de parâmetros da regressão são obtidas como $(X^\top X)^{-1}X^\top y$ e os valores apresentados em `Estimate`, no resumo da estimação dos parâmetros da regressão, obtido pela função R `summary`.

Consideraremos $Y_1, \cdots, Y_n$ variáveis aleatórias satisfazendo um modelo de regressão linear, isto é, tais que

$$Y = X\beta + \epsilon,$$

onde $\epsilon \sim N_n(0, \sigma^2 \mathrm{I})$ com as suposições resumidas em 1.2.1 assumidas certas. Os teoremas apresentados a seguir podem ser encontrados em referências clássicas deste tema, como os livros de Searle (1971) e Graybill (1976).

Teorema 2.5. *Seja $Y = (Y_1, \cdots, Y_n)^\top$ um vetor aleatório satisfazendo um modelo de regressão linear. Então o vetor $\widehat{\beta}$ é não enviesado, ou seja,*

$$\mathrm{E}(\widehat{\beta}) = \beta\cdot$$

Demonstração. Sabemos que $\widehat{\beta} = (X^\top X)^{-1}X^\top Y$, então

$$\mathrm{E}(\widehat{\beta}) = (X^\top X)^{-1}X^\top \mathrm{E}(Y),$$

dado que $\mathrm{E}(Y) = X\beta$, concluímos que o vetor $\widehat{\beta}$ é não enviesado. □

O Teorema de Gauss-Markov, em homenagem a Carl Friedrich Gauss e Andrey Markov[3], afirma que em um modelo de regressão linear o melhor estimador linear imparcial dos coeficientes da regressão são dados pelo estimador por $\widehat{\beta}$. Aqui o termo melhor estimador significa que, tanto por mínimos quadrados quanto por máxima verossimilhança, a solução das equações normais é o estimador não enviesado de mínima variância de β. Interessante é o fato que os erros não precisam ser normais, nem independentes e identicamente distribuídos bastam serem não correlacionadas e homocedásticos, ou seja, possuírem mesma variância.

Teorema 2.6 (Gauss-Markov)**.** *Se $Y_1, \cdots, Y_n$ satisfazem um modelo de regressão linear então $\widehat{\beta}$ é o melhor estimador linear não enviesado para β, ou seja, dentre todos os estimadores lineares não enviesados possíveis de β, $\widehat{\beta}$ é o de menor variância. A variância de $\widehat{\beta}$ é*

$$\mathrm{Var}(\widehat{\beta}) = \sigma^2(X^\top X)^{-1}\cdot$$

Demonstração. Consideremos $c^\top\beta$ uma combinação linear arbitrária dos coeficientes da regressão. Seja $a^\top Y$ um estimador não enviesado de $c^\top\beta$. Significa que

$$\mathrm{E}(a^\top Y) = a^\top X\beta = c^\top\beta,$$

para todo β. Isto implica que

$$a^\top X = c^\top\cdot$$

[3]Andrey Andreyevich Markov (1856 - 1922) foi um matemático russo. Ele é mais conhecido por seu trabalho sobre a teoria de processos estocásticos. Um objeto principal de sua pesquisa mais tarde se tornou conhecida como cadeias de Markov.

Dado que $X^\top X$ é de posto completo (ver exercício 7 do Capítulo 1), c é uma combinação linear de vetores coluna de $X^\top X$. Então

$$c = X^\top X\lambda$$

e

$$c^\top\widehat{\beta} = \lambda^\top X^\top X\widehat{\beta} = \lambda^\top X^\top X(X^\top X)^{-1}X^\top Y = \lambda^\top X^\top Y\cdot$$

Calculemos agora a variância de $a^\top Y$.

$$\begin{aligned}\operatorname{Var}(a^\top Y) &= \operatorname{Var}(a^\top Y - c^\top\widehat{\beta} + c^\top\widehat{\beta}) \\ &= \operatorname{Var}(a^\top Y - c^\top\widehat{\beta}) + \operatorname{Var}(c^\top\widehat{\beta}) + 2\operatorname{Cov}(a^\top Y - c^\top\widehat{\beta}; c^\top\widehat{\beta})\cdot\end{aligned}$$

Por outro lado

$$\begin{aligned}\operatorname{Cov}(a^\top Y - c^\top\widehat{\beta}; c^\top\widehat{\beta}) &= \operatorname{Cov}(a^\top Y - \lambda^\top X^\top Y; \lambda^\top X^\top Y) \\ &= (a^\top - \lambda^\top X^\top)\operatorname{Var}(Y)(\lambda^\top X^\top)^\top \\ &= (a^\top - \lambda^\top X^\top)\sigma^2\,\mathrm{I}X\lambda \\ &= (a^\top X - \lambda^\top X^\top X)\sigma^2 \\ &= (c^\top - c^\top)\sigma^2\lambda = 0\cdot\end{aligned}$$

Então

$$\operatorname{Var}(a^\top Y) = \operatorname{Var}(a^\top Y - c^\top\widehat{\beta}) + \operatorname{Var}(c^\top\widehat{\beta})$$

e como não existe variância negativa

$$\operatorname{Var}(a^\top Y) \geq \operatorname{Var}(c^\top\widehat{\beta})\cdot$$

Em outras palavras, a variância de $c^\top\widehat{\beta}$ é a menor dentre todos os estimadores não enviesados de $c^\top\widehat{\beta}$. Mais ainda, $\operatorname{Var}(a^\top Y) \geq \operatorname{Var}(c^\top\widehat{\beta})$ se, e somente se, $\operatorname{Var}(a^\top Y - c^\top\widehat{\beta}) = 0$, o qual implica $a^\top = c^\top\widehat{\beta}$. Portanto, o estimador linear de mínima variância de β é único. Calculemos a variância mínima agora.

$$\operatorname{Var}(\widehat{\beta}) = (X^\top X)^{-1}X^\top\operatorname{Var}(Y)\left[(X^\top X)^{-1}X^\top\right]^\top,$$

dado que $\operatorname{Var}(Y) = \sigma^2\,\mathrm{I}$, então

$$\operatorname{Var}(\widehat{\beta}) = \sigma^2(X^\top X)^{-1}X^\top X(X^\top X)^{-1},$$

e desta relação obtemos $\operatorname{Var}(c^\top\widehat{\beta}) = \sigma^2 c^\top(X^\top X)^{-1}c$. □

Teorema 2.7. *Sejam $Y_1, \cdots, Y_n$ variáveis aleatórias que satisfazem um modelo de regressão linear. Então*

$$\widehat{\beta} \sim N_p\left[\beta, \sigma^2(X^\top X)^{-1}\right], \tag{2.13}$$

onde p é a dimensão do vetor de parâmetros da regressão.

Demonstração. Considerar $C = (X^\top X)^{-1}X^\top$ e aplicar o Teorema 1.9. □

Definição 2.4. *Seja Y um vetor aleatório que satisfaz um modelo de regressão linear e seja $\widehat{\beta}$ o vetor de estimadores dos parâmetros da regressão. Chamamos de valores estimados ou valores preditos pelo modelo a estimadores da média da variável resposta, dados por*

$$\widehat{\mu} = X\widehat{\beta}\cdot \tag{2.14}$$

Os valores estimados ou preditos de um modelo de regressão linear podem ser escritos como

$$\widehat{\mu} = HY,$$

onde

$$H = X(X^\top X)^{-1}X^\top. \tag{2.15}$$

Isto devido a que

$$\widehat{\mu} = X\widehat{\beta} = \underbrace{X(X^\top X)^{-1}X^\top}_{H} Y.$$

A matriz H é muito importante, por exemplo, permite-nos provar propriedades dos estimadores dos parâmetros do modelo de regressão.

Exemplo 2.6. *Continuando o Exemplo* **??**, *temos que*

$$\widehat{\mu} = 12.29091 - 0.03077273\,\textit{Produção},$$

é a expressão matemática que define os valores preditos do custo de produção segundo o tamanho do lote.

Teorema 2.8. *Seja Y um vetor aleatório que satisfaz um modelo de regressão linear. O vetor valores preditos ou vetor de estimadores da média da variável resposta tem como distribuição de probabilidade*

$$\widehat{\mu} \sim N_n(X\beta, \sigma^2 H). \tag{2.16}$$

Demonstração. Considerar $C = X(X^\top X)^{-1}X^\top$ e aplicar o Teorema 1.9. □

O teorema anterior fornece suporte teórico à afirmação de que

$$\widehat{\mathrm{E}(\mathrm{Custo})} = 12.290909 - 0.030773\,\mathrm{Produção},$$

no Exemplo **??** e demais exemplos.

Vejamos agora como fazer para encontrar um estimador para a variância.

Teorema 2.9. *Seja Y um vetor aleatório que satisfaz um modelo de regressão linear. Um estimador não enviesado de σ^2 é dado por*

$$\widehat{\sigma}^2 = \frac{1}{n-p-1} Y^\top (\mathrm{I} - H)Y.$$

Demonstração. Pelo Teorema 1.4

$$\mathrm{E}\left[Y^\top(\mathrm{I} - H)Y\right] = \sigma^2 \,\mathrm{tr}(\mathrm{I} - H) + \beta^\top X^\top (\mathrm{I} - H)X\beta. \tag{2.17}$$

Observemos que

$$\begin{aligned} \beta^\top X^\top(\mathrm{I} - H)X\beta &= \beta^\top X^\top X\beta - \beta^\top X^\top HX\beta \\ &= \beta^\top X^\top X\beta - \beta^\top X^\top X(X^\top X)^{-1}X^\top X\beta \\ &= \beta^\top X^\top X\beta - \beta^\top X^\top X\beta = 0. \end{aligned}$$

No primeiro termo da expressão em (2.17) devemos identificar o valor de $\mathrm{tr}(\mathrm{I} - H)$. No Capítulo 1, Exercício 7 (c) encontra-se o valor $\mathrm{tr}(\mathrm{I} - H)$, desde que $\mathrm{I} - H$ seja uma matriz idempotente. Verifiquemos que $H = HH$,

$$X(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = X(X^\top X)^{-1}X^\top = H,$$

ou seja, a matriz H é idempotente. Pelo Exercício 6 (b), temos que

$$\text{tr}(\text{I} - H) = \text{tr}(\text{I}) - \text{tr}(H) = \text{posto}(\text{I}) - \text{posto}(H) = n - \text{posto}(H)\cdot$$

Pelo resultado do Exercício 7 (b), Capítulo 1 temos que, escolhendo $A = (X^\top X)^{-1}X^\top$ e $S = X$, concluímos que $\text{posto}(H) = p + 1$. Concluindo assim que $\text{E}(\widehat{\sigma}^2) = \sigma^2$. □

Utilizaremos as seguintes linhas de comando para comprovarmos resultados aplicados na demonstração do Teorema 2.9. Primeiro verificamos que $\text{posto}(H) = 2$, Exemplo **??**, com os comandos **R**:

```
> X = model.matrix(ajuste)
> H = X %*% solve(t(X) %*% X) %*% t(X)
> sum(diag(H))
[1] 2
```

Também estamos agora em condições de entender a frase

```
Residual standard error: 0.7491 on 9 degrees of freedom
```

na saída **R** no Exemplo **??**. Os graus de liberdade (degrees of freedom) correspondem ao denominador do estimador da variância $n-p-1$, que neste caso corresponde a $11-2 = 9$. Isso foi obtido no Teorema 2.9, ao mesmo tempo que o valor de $\widehat{\sigma}^2$ pode ser encontrado com os comandos **R**:

```
> sigma2 = (t(Custo) %*% (diag(n) - H) %*% Custo)/(n-2)
> sqrt(sigma2)
          [,1]
[1,] 0.7491463
```

correspondentes aos dados do Exemplo **??**.

2.2 Estimação por intervalos de confiança

Como todo processo de estimação podemos não somente apresentar os resultados pontualmente, podemos mostrar o intervalo ou banda de confiança de cada estimador dos parâmetros da regressão, assim como do estimador da resposta média quanto um chamado intervalo de predição para novas observações.

Os Teoremas 2.7 e 2.8 permitem obter a distribuição de probabilidades tanto de $\widehat{\beta}$ quanto $\widehat{\mu}$. Esses resultados poderiam permitir a construção de intervalos de confiança para esses estimadores não fosse um problema: ambos resultados dependem do conhecimento do valor do parâmetro σ^2.

Em situações práticas não conhecemos o valor de σ^2 e devemos então utilizar o estimador não enviesado demonstrado no Teorema 2.9. Temos que deduzir novos resultados para chegar a nosso objetivo. Nesse sentido é crucial deduzir a distribuição de probabilidades de $\widehat{\sigma}^2$ a qual, por sua vez, depende da distribuição de formas quadráticas.

2.2.1 Formas quadráticas

Formas quadráticas de vetores aleatórios normais são de grande importância em muitos ramos da estatística, tais como no método de mínimos quadrados, análise de variância, análise de regressão e planejamento de experimentos. Começamos investigando uma função de distribuição importante, a chamada distribuição qui-quadrado.

Definição 2.5. *Dizemos que a variável aleatória Y tem como distribuição de probabilidade qui-quadrado, com m graus de liberdade, se Y tem como função de probabilidade*

$$f(y) = \frac{e^{-y/2} y^{m/2-1}}{\Gamma(m/2) 2^{m/2}} \quad se \quad 0 \leq y < \infty,$$

e $f(y) = 0$ se $y \leq 0$, onde $\Gamma(\cdot)$ representa a função gama. Escrevemos $Y \sim \chi^2_{(m)}$.

Nesta definição fizemos uso da função gama, a qual, para $\alpha > 0$, define-se como $\Gamma(\alpha) = \int_0^\infty y^{\alpha-1} e^{-y} \, \mathrm{d}y$. Satisfaz que $\Gamma(1) = 1$, $\Gamma(n) = (n-1)!$ para n inteiro e $\Gamma(1/2) = \sqrt{\pi}$, por estas relações esta função é conhecida como uma generalização do fatorial de um número inteiro.

Existe uma relação importante entre as distribuições normal e qui-quadrado, objetivo do seguinte resultado. Este constitui o começo do processo que nos permitirá encontrar a distribuição de probabilidades de determinadas formas quadráticas.

Teorema 2.10. *Seja Z uma variável aleatória com distribuição normal padrão. Então $Y = Z^2 \sim \chi^2_{(1)}$.*

Demonstração. Exercício. □

Mais interessante é saber que soma de variáveis aleatórias χ^2 independentes, também tem como distribuição χ^2. Ainda mais, o parâmetro da distribuição da soma de variáveis aleatórias independentes χ^2 é a soma dos parâmetros da distribuição de cada somando. Este resultado é apresentado e demonstrado no seguinte teorema.

Teorema 2.11. *Sejam $Y_1, \cdots, Y_n$ variáveis aleatórias independentes com distribuição $Y_i \sim \chi^2_{(m_i)}$. Então, a soma destas também possui como distribuição de probabilidades $\chi^2_{(m)}$, onde $m = \sum_{i=1}^n m_i$, isto é,*

$$\sum_{i=1}^n Y_i \sim \chi^2_{(m)}.$$

Demonstração. A demonstração utiliza a função geradora de momentos[4] de cada Y_i, dada por $M_{Y_i}(t) = (1-2t)^{-n/2}$. A função geradora de momentos da soma de variáveis aleatórias independentes é o produto das funções geradoras de momentos respectivas. Assim

$$\begin{aligned} M_{\sum_{i=1}^n Y_i}(t) &= \prod_{i=1}^n M(t)_{Y_i}, = \prod_{i=1}^n (1-2t)^{-m_i/2} \\ &= (1-2t)^{-\sum_{i=1}^n m_i/2} = (1-2t)^{-m/2}. \end{aligned}$$

Esta última expressão corresponde à função geradora de momentos de uma variável aleatória qui-quadrado com m graus de liberdade. □

[4] A função geradora de momentos $M_X(t)$ da variável aleatória X é definida como $M_X(t) = \mathrm{E}\left(e^{tX}\right)$ caso esta esperança exista numa vizinhança da origem.

Definição 2.6. *Seja $Y = (Y_1, \cdots, Y_n)^\top$ um vetor aleatório e seja $A = (a_{ij})$ uma matriz quadrada de ordem n. Dizemos que a função real $Q(\underset{\sim}{y}) : \mathbb{R}_n \to \mathbb{R}$ é uma forma quadrática se*

$$Q(\underset{\sim}{y}) = \sum_{i=1}^{n}\sum_{j=1}^{n} a_{ij} y_i y_j.$$

Qualquer forma quadrática pode ser escrita de maneira matricial como $Y^\top AY$ e não é difícil demonstrar que a matriz A que define qualquer forma quadrática é simétrica. Observemos que $Y^\top AY$ é um número real, logo $(Y^\top AY)^\top = Y^\top A^\top Y$ então,

$$Y^\top AY = \frac{Y^\top AY + Y^\top A^\top Y}{2},$$

do qual concluímos que $A = (A + A^\top)/2$ e isto implica que A é simétrica. Uma das formas quadráticas mais simples é estudada no seguinte teorema. Este resultado é base para a demonstração de resultados mais interessantes do ponto de vista prático.

Teorema 2.12. *Seja Y um vetor aleatório tal que $Y \sim N_n(0, \mathrm{I})$. Então, a distribuição da forma quadrática $Y^\top Y$ e qui-quadrado com n graus de liberdade, isto é, $Y^\top Y$ satisfaz*

$$Y^\top Y \sim \chi^2_{(n)}.$$

Demonstração. Utilizamos o Teorema 2.11. Cada Y_i, $i = 1, \cdots, n$, é uma variável aleatória normal padrão independente das outras. Podemos escrever então a forma quadrática como soma de variáveis aleatórias independentes ao quadrado, isto é,

$$Y^\top Y = \sum_{i=1}^{n} Y_i^2 \sim \chi^2_{(n)}.$$

□

A seguir mostramos que isto também se satisfaz no caso mais geral.

Teorema 2.13. *Seja Y um vetor aleatório tal que $Y \sim N_n(\mu, \Sigma)$, com $|\Sigma| > 0$. Então*

$$(Y - \mu)^\top \Sigma^{-1} (Y - \mu) \sim \chi^2_{(n)},$$

onde n é a dimensão de Y.

Demonstração. Seja $Z = \Sigma^{-1/2}(Y - \mu)$. Então

$$\mathrm{E}(Z) = 0 \qquad \text{e} \qquad \mathrm{Cov}(Z) = \Sigma^{-1/2}\Sigma\Sigma^{-1/2} = \mathrm{I},$$

isto significa que $Z \sim N_n(0, \mathrm{I})$. Segue então que

$$(Y - \mu)^\top \Sigma^{-1} (Y - \mu) = [\Sigma^{-1/2}(Y - \mu)]^\top [\Sigma^{-1/2}(Y - \mu)] = Y^\top Y \sim \chi^2_{(n)}.$$

□

Nesta demonstração utilizamos indiretamente a exigência de $|\Sigma| > 0$, isto garante a existência de $\Sigma^{-1/2}$. No início desta seção comentamos que nosso objetivo seria encontrar a distribuição de probabilidades de determinadas formas quadráticas. A continuação apresentamos e demonstramos um resultado que nos disse quais formas quadráticas tem distribuição χ^2. Como consequência, depois deduzimos a distribuição do estimador não enviesado da variância do erro em modelos de regressão.

Teorema 2.14. *Seja Y um vetor aleatório tal que $Y \sim N_n(0, \sigma^2 \mathrm{I})$ e M uma matriz simétrica idempotente tal que* $\mathrm{posto}(M) = m$. *Então*

$$\frac{Y^\top M Y}{\sigma^2} \sim \chi^2_{(m)}.$$

Demonstração. Dado que M é uma matriz simétrica existe uma matriz ortogonal Q (Searle, 1971; Rao, 1973) de forma que $Q^\top M Q = \Lambda$, onde

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

Os elementos da matriz diagonal Λ são os valores próprios de M. Mais ainda, devido a M ser idempotente todo autovalor ou é 0 ou é 1. Então podemos escolher Q de maneira que Λ seja de forma

$$Q^\top M Q = \Lambda = \begin{pmatrix} \mathrm{I}_m & 0 \\ 0 & 0 \end{pmatrix}.$$

A dimensão de I_m é m porque este é o número de autovalores de M diferentes de zero. Lembremos que $\mathrm{tr}(M) = \mathrm{posto}(M)$ (Ver exercício 7).

Seja $V = Q^\top Y$, então $\mathrm{E}(V) = 0$ e

$$\mathrm{Var}(V) = Q^\top \sigma^2 \mathrm{I} Q = \sigma^2 Q^\top Q = \sigma^2 \mathrm{I},$$

isto devido a Q ser uma matriz ortogonal. Concluímos que $V \sim N(0, \sigma^2 \mathrm{I})$. Podemos então fazer $Y = (Q^\top)^{-1} V = QV$ e

$$\frac{Y^\top M Y}{\sigma^2} = \frac{V^\top Q^\top M Q V}{\sigma^2} = \frac{1}{\sigma^2} V^\top \begin{pmatrix} \mathrm{I}_m & 0 \\ 0 & 0 \end{pmatrix} V = \frac{1}{\sigma^2} \sum_{i=1}^{m} V_i^2.$$

Isto é a soma de m variáveis aleatórias normal padrão independentes. Pelo Teorema 2.12 concluímos que $\dfrac{Y^\top M Y}{\sigma^2} \sim \chi^2_{(m)}$. □

Teorema 2.15. *Seja Y um vetor aleatório que satisfaz um modelo de regressão linear. Então,*

$$\frac{(n-p-1)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-p-1)},$$

onde $\widehat{\sigma}^2$ é o estimador não enviesado de σ^2 dado por

$$\widehat{\sigma}^2 = \frac{1}{n-p-1} Y^\top (\mathrm{I} - H) Y.$$

Demonstração. A matriz $\mathrm{I} - H$ é idempotente e de posto $n - p - 1$. Essas propriedades serão demonstradas no Teorema 3.5. □

2.2.2 Predição em modelos de regressão

Uma das utilidades dos valores preditos do modelo de regressão é fazer previsões, respeitando sempre o intervalo de observação das variáveis explicativas. Embora a estimativa para o valor esperado de Y, dado um conjunto de valores das variáveis explicativas, seja obtida de maneira simples a construção de intervalos de confiança não é tão simples assim.

Intervalos de confiança para a resposta média

Em muitas situações é desejável apresentar os resultados em intervalos de confiança para a resposta média, sejam estes num único ponto ou para um conjunto de pontos e é nosso objetivo encontrar as expressões teóricas e mostrar como proceder em situações práticas. Primeiro, uma discussão da teoria envolvida nos cálculos.

Definição 2.7. *Sejam $Z \sim N(0,1)$ e $W \sim \chi^2_{(k)}$ duas variáveis aleatórias independentes. Então a variável aleatória*

$$T = \frac{Z}{\sqrt{W/k}}, \tag{2.18}$$

tem distribuição t-Student com k graus de liberdade. Escrevemos $T \sim t_{(k)}$.

A distribuição t-Student é simétrica e para n grande aproxima-se da distribuição normal. No entanto, para n pequeno, T difere consideravelmente de uma variável normal. De fato,

$$P(|T| \geq t_0) \geq P(|Z| \geq t_0),$$

onde $Z \sim N(0,1)$ e t_0 um número real positivo qualquer. Isto significa que nas caudas da distribuição t-Student as probabilidade são maiores do que nas caudas da distribuição normal padrão. Não é difícil deduzir que, para $k > 2$, $\mathrm{E}(T) = 0$ e $\mathrm{Var}(T) = k/(k-2)$. Numa primeira leitura pode parecer que esta definição nada tem a ver com o tema, porém depois de poucos resultados perceberemos a importância desta distribuição para encontrar intervalos de confiança para a resposta média.

Em uma grande variedade de problemas de inferência o interesse não é estimar os parâmetros ou testar algumas hipóteses sobre ele. Pelo contrário, se desejam estabelecer limites inferior e superior para o valor real de algum parâmetro. Problemas deste tipo são chamados problemas de estimação por intervalos de confiança os quais são intervalos reais aleatórios que podem incluir o verdadeiro valor do parâmetro com determinada probabilidade conhecida próxima de 1.

Definição 2.8. *Seja $f(y)$ a função de densidade da variável aleatória Y com $\Theta \subset \mathbb{R}$ o espaço paramétrico. A família de subconjuntos $S(Y)$ de $\Theta \subset \mathbb{R}$, é dito que constituem uma família de conjuntos de confiança a um nível de confiança $1-\alpha$ se*

$$P\big(S(Y) \in \theta\big) \geq 1-\alpha, \qquad \textit{para todo } \theta \in \Theta,$$

isto é, o conjunto aleatório $S(Y)$ cobre o verdadeiro valor do parâmetro θ com probabilidade $\geq 1-\alpha$.

Se $S(Y)$ é da forma

$$S(Y) = \big(\underline{\theta}(Y), \overline{\theta}(Y)\big),$$

vamos chamar-lhe um intervalo de confiança a um nível de confiança $1-\alpha$, desde que

$$P\big(\underline{\theta}(Y) < \theta < \overline{\theta}(Y)\big) \geq 1-\alpha, \qquad \text{para todo } \theta \in \Theta,$$

e a quantidade

$$\left\{ \theta \in \Theta \,:\, P\big(\underline{\theta}(Y) < \theta < \overline{\theta}(Y)\big) \geq 1-\alpha \right\},$$

define o coeficiente de confiança associado com o intervalo aleatório[5].

Observemos que as funções $\underline{\theta}(Y)$ e $\overline{\theta}(Y)$ dependem somente de Y e não de θ, isto implica que ambas funções são variáveis aleatórias. Chamamos então $S(Y)$ de um intervalo aleatório com $\underline{\theta}(Y)$ e $\overline{\theta}(Y)$ como os limites inferior e superior, respectivamente.

[5]Um elemento $\theta_0 \in \Theta$ é dito ser ínfimo do conjunto Θ, resumidamente $\inf \Theta = \inf_\theta \Theta = \theta_0$, se $\theta_0 \leq \theta$, $\forall \theta \in \Theta$.

Definição 2.9. *A família de subconjuntos* $S(Y) = \big(\underline{\theta}(Y), \overline{\theta}(Y)\big)$ *de* $\Theta \subset \mathbb{R}$ *constitui uma família de conjuntos de confiança, a um nível de confiança* $1-\alpha$*, simétricos em probabilidade se*

$$P\big(\underline{\theta}(Y) \leq \theta\big) = P\big(\overline{\theta}(Y) \geq \theta\big) \leq 1-\alpha,$$

para todo $\theta \in \Theta$*.*

O seguinte resultado proporciona um método geral de encontrar os intervalos de confiança e abrange a maioria dos casos na prática.

Teorema 2.16. *Seja* Y *um vetor aleatório com função de densidade* $f(\underset{\sim}{y};\theta)$*,* $\theta \in \Theta$ *e* $\Theta \subset \mathbb{R}_k$*. Seja* $T(Y,\theta)$ *uma função real tal que, como função de* θ*, seja crescente ou decrescente em cada ponto* $\underset{\sim}{y} = (y_1, \cdots, y_n) \in \mathbb{R}_n$*. Seja* $\Lambda \subset \mathbb{R}$ *o intervalo de variação de* T *de maneira que, para todo* $\lambda \in \Lambda$ *e* $\underset{\sim}{y} \in \mathbb{R}_n$*, a equação* $\lambda = T(\underset{\sim}{y},\theta)$ *tenha solução. Se a distribuição de probabilidades de* $T(Y,\theta)$ *não depende de* θ*, podemos construir intervalos de confiança para* θ *qualquer seja o nível de confiança escolhido.*

Demonstração. Seja $0 < \alpha < 1$. Podemos então escolher números reais $\lambda_1(\alpha)$ e $\lambda_2(\alpha)$ em Λ não necessariamente únicos, tais que

$$P\big(\lambda_1(\alpha) < T(Y,\theta) < \lambda_2(\alpha)\big) \geq 1-\alpha, \qquad \text{para todo } \theta \in \Theta\cdot$$

Dado que a distribuição de probabilidades de $T(Y,\theta)$ não depende de θ, os números $\lambda_1(\alpha)$ e $\lambda_2(\alpha)$ também não dependem de θ. Uma vez que, além disso, T é monótona em θ, podemos resolver as equações

$$T(\underset{\sim}{y},\theta) = \lambda_1(\alpha) \qquad \text{e} \qquad T(\underset{\sim}{y},\theta) = \lambda_2(\alpha)$$

para todo $\underset{\sim}{y}$ unicamente para θ. Temos então que

$$P\big(\underline{\theta}(Y) < \theta < \overline{\theta}(Y)big\big) \geq 1-\alpha, \qquad \text{para todo } \theta \in \Theta,$$

onde $\underline{\theta}(Y) < \overline{\theta}(Y)$ são variáveis aleatórias. □

A prova deste resultado nos fornece um procedimento de trabalho para encontrar intervalos de confiança partindo de funções da amostra e do parâmetro de interesse. Estamos interessados agora na resposta média. Devemos, portanto, encontrar uma função da amostra e da resposta média num modelo de regressão cuja distribuição de probabilidades não dependa desta, ou seja, que não dependa da própria resposta média. A resposta a este questionamento será fornecida no Teorema 2.18. Antes, devemos apresentar o seguinte resulto auxiliar.

Teorema 2.17. *Seja* Y *um vetor aleatório com distribuição* $N_n(0, \sigma^2\,\mathrm{I})$*,* M *uma matriz simétrica e idempotente de ordem* n *e* L *uma matriz de ordem* $k \times n$*. Então* LY *e* $Y^\top MY$ *são independentes se* $LM = 0$*.*

Demonstração. Definamos Q como no Teorema 2.14, então

$$Q^\top MQ = \Lambda = \begin{pmatrix} \mathrm{I}_m & 0 \\ 0 & 0 \end{pmatrix}\cdot$$

Lembremos que $\mathrm{tr}(M) = m$. Seja $\nu = Q^\top Y$, o qual pode ser escrito como

$$\nu = (\nu_1, \nu_2) = (\nu_{11}, \cdots, \nu_{1m}, \nu_{21}, \cdots, \nu_{2(n-m)})\cdot$$

Os vetores ν_1 e ν_2 são independentes uma vez que tem distribuição normal padrão independentes. O que vamos mostrar agora é que $Y^\top MY$ depende apenas de ν_1 e que LY depende apenas de ν_2 e dado que os ν_i são independentes, $Y^\top MY$ e LY seriam independentes.

Pelo Teorema 2.14 podemos escrever

$$\begin{aligned} Y^\top MY &= \nu^\top Q^\top MQ\nu \\ &= \nu^\top \begin{pmatrix} \mathrm{I}_m & 0 \\ 0 & 0 \end{pmatrix} \nu \\ &= \nu_1^\top \nu_1 \cdot \end{aligned}$$

Considere agora $C = LQ$, logo $C = (C_1, C_2)$ onde C_1 é de ordem $k \times m$ e C_2 é de ordem $k \times (n-m)$. Agora, considere o seguinte produto

$$C(Q^\top MQ) = LQQ^\top MQ \cdot$$

Dado que Q é uma matriz ortogonal[6] temos que $QQ^\top = \mathrm{I}$, então $C(Q^\top MQ) = LMQ = 0$, devido a que por hipóteses $LM = 0$. Por outro lado,

$$\begin{aligned} C(Q^\top MQ) &= (C_1, C_2) \begin{pmatrix} \mathrm{I}_m & 0 \\ 0 & 0 \end{pmatrix} \\ &= 0 \cdot \end{aligned}$$

Isto implica logicamente que $C_1 = 0$. Portanto, $LQ = C = (0, C_2)$. Podemos, agora, escrever

$$LY = LQQ^\top Y,$$

devido a ser Q uma matriz ortogonal, mas $C = LQ$ e $\nu = Q^\top Y$, então $LY = C\nu = (0, C_2)\nu = C_2\nu_2$. Podemos perceber que LY depende somente de ν_2 e foi provado que $Y^\top MY$ depende somente de ν_1. Desde que ν_1 e ν_2 são variáveis aleatórias independentes, se $LM = 0$ então LY e $Y^\top MY$ são independentes. □

Estamos agora em condições de encontrar a forma do intervalo de confiança para a resposta média. Para isso, seja $\widehat{\mu}_{\underset{\sim}{x}}$ o estimador do valor esperado da resposta, onde $\underset{\sim}{x}$ é um vetor qualquer de dimensão p definido por $\underset{\sim}{x} = (1, x_1, \cdots, x_p)$. O vetor $\underset{\sim}{x}$ é qualquer no sentido de que cada x_j deve pertencer ao intervalo de observação da variável explicativa correspondente X_j, $j = 1, \cdots, p$.

De acordo com (2.14) a previsão para esse vetor é dada por

$$\widehat{\mu}_{\underset{\sim}{x}} = \underset{\sim}{x}(X^\top X)^{-1} X^\top Y \tag{2.19}$$

com variância na forma

$$\mathrm{Var}(\widehat{\mu}_{\underset{\sim}{x}}) = \sigma^2 \underset{\sim}{x}(X^\top X)^{-1} \underset{\sim}{x}^\top \cdot \tag{2.20}$$

Segundo o Teorema 2.16 devemos encontrar uma função $T(\widehat{\mu}_{\underset{\sim}{x}}, \mu_{\underset{\sim}{x}})$ cuja distribuição de probabilidades não depende de $\mu_{\underset{\sim}{x}}$, a resposta média esperada, sendo

$$\mu_{\underset{\sim}{x}} = \underset{\sim}{x}(X^\top X)^{-1} X^\top X\beta,$$

ou

$$\mu_{\underset{\sim}{x}} = \underset{\sim}{x}\beta \cdot \tag{2.21}$$

O seguinte resultado permite encontrar a forma da função T procurada.

[6]Uma matriz Q é ortogonal se sua inversa coincide com sua transposta, isto é, $QQ^\top = Q^\top Q = Q^{-1}Q = QQ^{-1} = \mathrm{I}$.

Teorema 2.18. *Seja Y um vetor aleatório que satisfaz um modelo de regressão linear. Então*

$$\frac{\widehat{\mu}_{\underset{\sim}{x}} - \underset{\sim}{x}\beta}{\sqrt{\widehat{\sigma}^2 \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top}} \sim t_{(n-p-1)},$$

onde $\mathrm{E}(\widehat{\mu}_{\underset{\sim}{x}}) = \underset{\sim}{x}\beta$.

Demonstração. Utilizaremos o Teorema 2.17. O numerador é

$$\frac{\widehat{\mu}_{\underset{\sim}{x}} - \underset{\sim}{x}\beta}{\sqrt{\mathrm{Var}\{\widehat{\mu}_{\underset{\sim}{x}}\}}} \sim N(0,1)$$

e o denominador

$$\frac{(n-p-1)\widehat{\sigma}^2}{\sigma^2} \Big/ (n-p-1) \sim \chi^2_{(n-p-1)}.$$

Bastaria provar que numerador e denominador são variáveis aleatórias independentes. Observemos que se no Teorema 2.17 definimos $L = \underset{\sim}{x}(X^\top X)^{-1}$ e $M = \mathrm{I} - H$, as quais satisfazem que $LM = 0$, podemos então utilizar a Definição 2.7, a definição da distribuição t-Student. Observando que se Z e W forem variáveis aleatórias independentes, então $Z + c$ e dW também são independentes, sendo $c = -\underset{\sim}{x}\beta$ e $d = \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top$ constantes. □

Este resultado nos permite chegar à forma que deve ter o intervalo de confiança para a média, em estatística as funções de estimadores cuja distribuição não dependa de parâmetros desconhecidos são chamadas de quantidades pivotais.

A forma do intervalo de confiança para a média ou reta de regressão estimada é objetivo do seguinte teorema e do Exemplo 2.7 logo em seguida, o qual mostra este intervalo de uma maneira alternativa.

Teorema 2.19 (Intervalo de confiança para a média)**.** *Sejam Y um vetor aleatório que satisfaz um modelo de regressão linear e $\widehat{\mu}_{\underset{\sim}{x}}$ o estimador do valor esperado da resposta, onde $\underset{\sim}{x}$ é um vetor qualquer de dimensão $p+1$ definido por $\underset{\sim}{x} = (1, x_1, \cdots, x_p)$ respeitando o intervalo de observação de cada variável explicativa. Então o intervalo de confiança simétrico em probabilidade para a resposta média estimada de nível de significância $1-\alpha$ ou o intervalo de $100(1-\alpha)\%$ de confiança para $\widehat{\mu}_{\underset{\sim}{x}}$ é dado por*

$$\left(\widehat{\mu}_{\underset{\sim}{x}} - t_{\alpha/2(n-p-1)}\sqrt{\widehat{\sigma}^2 \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top}, \widehat{\mu}_{\underset{\sim}{x}} + t_{\alpha/2(n-p-1)}\sqrt{\widehat{\sigma}^2 \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top}\right),$$

onde $t_{\alpha/2(n-p-1)}$ é o percentil da distribuição t-Student com $n-p-1$ graus de liberdade.

Demonstração. Definamos

$$T(\widehat{\mu}_{\underset{\sim}{x}}, \mu_{\underset{\sim}{x}}) = (\widehat{\mu}_{\underset{\sim}{x}} - \underset{\sim}{x}\beta) \Big/ \sqrt{\widehat{\sigma}^2 \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top}.$$

Sabemos que $T(\widehat{\mu}_{\underset{\sim}{x}}, \mu_{\underset{\sim}{x}})$ tem como distribuição t-Student com $n-p-1$ graus de liberdade, segundo o Teorema 2.18 e ainda que não depende de $\mu_{\underset{\sim}{x}}$. A demonstração do Teorema 2.16 nos disse que, escolhendo

$$\lambda_1(\alpha) = -t\alpha/2(n-p-1) \qquad \text{e} \qquad \lambda_2(\alpha) = t\alpha/2(n-p-1)$$

e resolvendo as equações

$$T(\widehat{\mu}_{\underset{\sim}{x}}, \mu_{\underset{\sim}{x}}) = \lambda_1(\alpha) \qquad \text{e} \qquad T(\widehat{\mu}_{\underset{\sim}{x}}, \mu_{\underset{\sim}{x}}) = \lambda_2(\alpha),$$

encontramos que os limites inferior e superior, respectivamente, do intervalo de confiança para $\mu_{\underset{\sim}{x}}$ assumem a forma

$$\underline{\theta}(Y) = \widehat{\mu}_{\underset{\sim}{x}} - t_{\alpha/2(n-p-1)}\sqrt{\widehat{\sigma}^2 \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top}$$

e

$$\overline{\theta}(Y) = \widehat{\mu}_{\underset{\sim}{x}} + t_{\alpha/2(n-p-1)}\sqrt{\widehat{\sigma}^2 \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top}.$$

□

Lembremos que, se a variável aleatória T tem como distribuição de probabilidades t-Student com $n-p-1$ graus de liberdade, $t_{\alpha/2(n-p-1)}$ é percentil $\alpha/2$ desta distribuição se satisfaz a equação

$$P\big(|T| \geq t_{\alpha/2(n-p-1)}\big) = \alpha.$$

Exemplo 2.7. *Seja $Y = \beta_0 + \beta_1 X + \epsilon$, onde $\epsilon \sim N(0, \sigma^2\,\mathrm{I})$ um modelo de regressão linear simples. Neste caso, a expressão para o intervalo $100(1-\alpha)\%$ de confiança para a resposta média no ponto $(1, x)$ é da forma*

$$\overline{y} + \widehat{\beta}_1(x - \overline{x}) \pm t_{\alpha/2(n-2)}\sqrt{\widehat{\sigma}^2\left[\frac{1}{n} + \frac{(x-\overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}.$$

Este resultado é obtido aplicando o Teorema 2.19 ao modelo de regressão linear simples.

Devemos esclarecer que nestes resultados e nos próximos utilizamos o estimador não enviesado da variância, apresentado no Teorema 2.9 e depois no Corolário 2.15.

Intervalo de confiança para novas observações

Também podemos definir uma banda de confiança para a média de novas observações, esta banda é conhecida como o intervalo de confiança para novas observações. Devemos esclarecer que o modelo de regressão, pelo fato de assumir independência entre as observações da variável resposta, somente pode ser utilizado no intervalo de variação das variáveis explicativas observadas nada podendo-se afirmar fora destes intervalos.

Considere agora a situação de se fazer previsões para a média de uma nova observação Y_f no ponto $\underset{\sim}{x} = (1, x_1, \cdots, x_p)$ e obtida de maneira independente das n primeiras observações utilizadas para estimar β.

Esta nova observação cumpre o mesmo modelo de regressão assumido, isto é, Y é um vetor aleatório satisfazendo o modelo de regressão $Y = X\beta + \epsilon$, onde $\epsilon \sim N_n(0, \sigma^2\,\mathrm{I})$ e também

$$Y_f = \underset{\sim}{x}\beta + \epsilon_f,$$

onde $\epsilon_f \sim N(0, \sigma^2)$ é independente de $\epsilon_1, \cdots, \epsilon_n$.

Novamente utilizamos o Teorema 2.18 para encontrar um quantidade pivotal que dependa de novas observações e, então, definir estes intervalos no teorema a seguir.

Teorema 2.20 (Intervalo de confiança para novas observações)**.** *Seja Y um vetor aleatório que satisfaz o modelo de regressão linear $Y = X\beta + \epsilon$, onde $\epsilon \sim N_n(0, \sigma^2\,\mathrm{I})$. Seja Y_f uma nova observação satisfazendo que $Y_f = \underset{\sim}{x}\beta + \epsilon_f$, onde $\epsilon_f \sim N(0, \sigma^2)$ é independente de $\epsilon_1, \cdots, \epsilon_n$. Então $\widehat{\mu}_{\underset{\sim}{x}} = \underset{\sim}{x}\widehat{\beta}$ é o estimador do valor esperado da nova observação e o intervalo de confiança simétrico em probabilidade para uma nova observação de nível de significância $1-\alpha$ tem como limites*

$$\widehat{\mu}_{\underset{\sim}{x}} \pm t_{\alpha/2(n-p-1)}\sqrt{\widehat{\sigma}^2[1 + \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top]},$$

onde $t_{\alpha/2(n-p-1)}$ é o percentil da distribuição t-Student com $n-p-1$ graus de liberdade.

Demonstração. Sabemos que $\mathrm{E}(Y_f) = \underset{\sim}{x}\beta$, logo $\widehat{\mathrm{E}(Y_f)} = \widehat{\mu}_{\underset{\sim}{x}} = \underset{\sim}{x}\widehat{\beta}$ é o estimador do valor esperado da nova observação. Para encontrarmos o intervalo de confiança observemos que $Y_f - \underset{\sim}{x}\widehat{\beta} = \underset{\sim}{x}\beta + \epsilon_f - \underset{\sim}{x}\widehat{\beta}$, do qual obtemos que

$$\mathrm{Var}(Y_f - \underset{\sim}{x}\widehat{\beta}) = \mathrm{Var}(\epsilon_f) + \mathrm{Var}(\underset{\sim}{x}\widehat{\beta})\cdot$$

Pela definição do modelo de regressão sabemos que $\mathrm{Var}(\epsilon_f) = \sigma^2$ e pelo resultado (2.20) temos que

$$\mathrm{Var}(Y_f - \underset{\sim}{x}\widehat{\beta}) = \sigma^2 + \sigma^2\underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top,$$

de forma que

$$T(\widehat{\mu}_{\underset{\sim}{x}}, \mu_{\underset{\sim}{x}}) = \frac{\widehat{\mu}_{\underset{\sim}{x}} - \underset{\sim}{x}\beta}{\sqrt{\widehat{\sigma}^2(1 + \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top)}}$$

e continuamos similarmente à demonstração do Teorema 2.19. □

Observando os resultados dos Teoremas 2.19 e 2.20 podemos perceber que o intervalo de predição para a nova observação Y_f é mais largo do que o intervalo de confiança para a estimativa do valor da função de regressão. Isto deve-se à incerteza adicional na previsão Y_f, a qual é representada pelo termo adicional $\widehat{\sigma}^2$ na expressão $\widehat{\sigma}^2\big(1 + \underset{\sim}{x}(X^\top X)^{-1}\underset{\sim}{x}^\top\big)$; o qual por sua vez vem da presença do termo de erro ϵ_f.

Exemplo 2.8. *Seja $Y = \beta_0 + \beta_1 X + \epsilon$, onde $\epsilon \sim N(0, \sigma^2\,\mathrm{I})$ um modelo de regressão linear simples. Neste caso, a expressão dos limites do intervalo $100(1-\alpha)\%$ de confiança para uma nova observação no ponto $(1, x)$ são da forma*

$$\overline{y} + \widehat{\beta}_1(x - \overline{x}) \pm t_{\alpha/2(n-2)}\sqrt{\widehat{\sigma}^2\left[1 + \frac{1}{n} + \frac{(x-\overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2}\right]}.$$

Este exemplo serve para mostrar os cálculos manualmente. Em situações práticas os intervalos de confiança, sejam estes para a resposta média ou para novas observações, costumam-se serem apresentados de maneira gráfica. No exemplo a seguir mostramos como gerar no **R** as informações necessárias para a construção destes gráficos. Um esclarecimento, estes intervalos somente podem estão definidas caso a correspondente variável explicativa seja do tipo contínua, como é o caso dos exemplos 2.4 e 2.5.

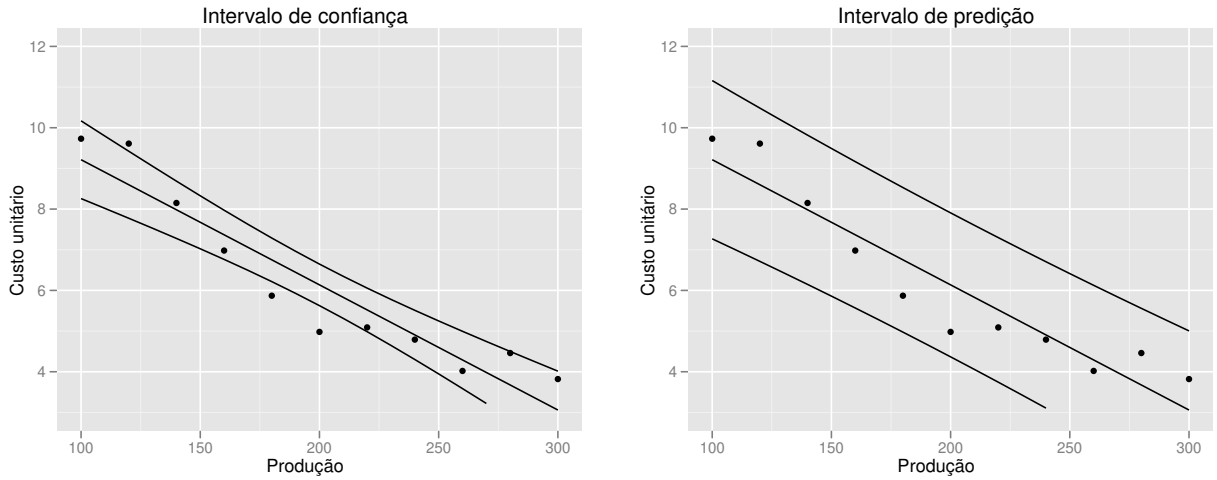

Figura 2.5: Gráficos de dispersão, média estimada e intervalos de confiança dos dados apresentados no Exemplo 2.4. A esquerda mostramos o gráfico de dispersão junto a reta estimada e o intervalo de confiança para a média. A direita mostramos o gráfico de dispersão, a média estimada junto ao intervalo de confiança para novas observações.

Exemplo 2.9 (Continuando o Exemplo 2.4)**.** *Na Figura 2.5 mostramos a dispersão dos dados, a resposta média ou estimada e o intervalo ou banda de confiança para a resposta média a esquerda. Com esse objetivo definimos qual intervalo de confiança queremos, isso utilizando a opção* `interval='confidence'` *no comando* ***R*** `predict` *ou* `predict.lm`*. Com esse comando geramos uma matriz de três colunas e tantas linhas quanto de dados independentes.*

```
> preditos.lim=predict(ajuste,newdata=novos, interval='confidence',level=0.95)
```

Ainda escolhemos o nível de confiança, sendo aqui de 0.95. Na matriz `preditos.lim`*, a primeira coluna contém os valores preditos da média; as outras colunas os limites inferior e superior, respectivamente, do intervalo no Teorema 2.19. Para construir o gráfico mostrado à esquerda da Figura* **??***, utilizamos as seguintes linhas de comando:*

```
> qplot(Producao,Custo,xlab='Produção',ylab='Custo unitário',
                                ylim=c(3,12),main='Intervalo de confiança')+
    geom_line(aes(y=preditos.lim[,1],x=novos.valores))+
    geom_line(aes(y=preditos.lim[,2],x=novos.valores))+
    geom_line(aes(y=preditos.lim[,3],x=novos.valores))
```

Exemplo 2.10 (Continuação do Exemplo 2.4)**.** *Vamos construir o gráfico a direita na Figura 2.5. Nesta caso mostramos o gráfico de dispersão dos dados, a resposta média ou resposta estimada e a banda de confiança para novas observações. Para isso primeiro consideram-se novos valores preditos, sempre dentro do intervalo de variação da variável regressora que neste caso é Producao e colocam-se num novo arquivo de dados, seguindo as linhas de comando a seguir:*

```
> novos.valores=seq(100,300,by=10)
> novos=data.frame(Producao=novos.valores)
```

Para gerar o intervalo de confiança para novos valores ou intervalo de predição escolhemos `interval='prediction'` *no comando* `predict`*, como mostrado a seguir:*

```
> preditos.lim=predict(ajuste,newdata=novos, interval='prediction',level=0.95)
```

Para construir o gráfico mostrado à direita da Figura **??***, utilizamos as linhas de comando:*

```
> qplot(Producao,Custo,xlab='Produção',ylab='Custo unitário',
                                  ylim=c(3,12),main='Intervalo de predição')+
      geom_line(aes(y=preditos.lim[,1],x=novos.valores))+
      geom_line(aes(y=preditos.lim[,2],x=novos.valores))+
      geom_line(aes(y=preditos.lim[,3],x=novos.valores))
```

Podemos observar na Figura **??** que, como foi mencionado, o intervalo de predição de novas observações é sempre mais largo do que o intervalo de confiança para a média, mantendo o mesmo nível de confiança.

2.3 Teste de hipóteses

Em muitos problemas práticos estamos interessados em testar a validade de uma afirmação sobre um parâmetro desconhecido. Um problema deste tipo é chamado como o problema de teste de hipóteses o qual é o objeto de discussão na presente seção. Referências importantes são os livros Wilks (1962), Rohatgi (1976) e Knight (1999), dentre muitos outros.

Na década de 1920, desenvolveu-se a teoria por trás do p-valor e da teoria do teste de hipóteses. Estas teorias desenvolvidas por distintos investigadores proporcionaram importantes ferramentas quantitativas para confirmar ou refutar suas hipóteses. O p-valor é a probabilidade de obter um efeito igual ou mais extremo do que o observado presumindo que a hipóteses nula não teve nenhum efeito verdadeiro, isto fornece aos pesquisadores uma forma de medir a força da evidência contra a hipóteses nula. Tal como comumente usado, os pesquisadores vão selecionar um limiar do p-valor abaixo do qual eles vão rejeitar a hipótese nula. A teoria dos testes de hipóteses permite aos pesquisadores rejeitar a hipótese nula em favor de uma hipótese alternativa de algum efeito. Tal como comumente usado, os investigadores irão escolher valores do erro tipo I, ou seja, da probabilidade de rejeitar a hipótese nula quando é verdadeira e do erro tipo II, ou seja, da probabilidade de aceitar a hipótese nula quando ela é falsa. Os níveis destes erros permitirão determinar uma chamada região crítica. Se a estatística de teste cai nessa região crítica, a hipótese nula é rejeitada em favor da hipótese alternativa.

2.3.1 Teste de hipóteses paramétrico

Fixando ideias consideremos que o parâmetro de interesse seja $\theta \in \Theta$, onde Θ é o espaço paramétrico, um subconjunto dos reais.

Definição 2.10. *Uma hipóteses paramétrica é uma afirmação sobre o vetor de parâmetros desconhecidos θ. A afirmação que queremos verificar é chamada de hipótese nula $H_0 : \theta \in \Theta_0 \subset \Theta$. A afirmação contrária $H_A : \theta \in \Theta_A = \Theta \setminus \Theta_0$ é conhecida como hipótese alternativa.*

Nesta definição a notação $A \setminus B$, significa a diferença entre dois conjuntos A e B a qual é o conjunto dos elementos que pertencem a A e que não pertencem a B. Observemos que os subconjuntos Θ_0 e Θ_A formam uma partição do espaço paramétrico, isto é, $\Theta_0 \cup \Theta_A = \Theta$ e $\Theta_0 \cap \Theta_A = \emptyset$. O problema de teste de hipóteses pode ser descrito da seguinte forma: dado uma amostra, devemos encontrar uma regra de decisão que permita decidir aceitar ou rejeitar a hipótese nula. Em outras palavras, dada a amostra $\underset{\sim}{y} = (y_1, \cdots, y_n)$ precisamos encontrar uma regra que determine se θ está em Θ_0 ou em Θ_A.

Definição 2.11 (Função teste)**.** *Uma função aleatória $\delta : \mathbb{R}_n \to \mathbb{R}$, que assume somente dois valores, tal que*

$$\delta(\underset{\sim}{y}) = \begin{cases} 0 & se \quad \theta \in \Theta_0 \\ 1 & se \quad \theta \in \Theta_A \end{cases}$$

é chamada de função teste.

Em muitos casos, $\delta(\cdot)$ dependerá apenas de alguma estatística de valores reais T, a que chamaremos a estatística de teste. É improvável que qualquer função de determinado teste será perfeita. Assim, para uma dada função de teste δ, temos que examinar a probabilidade de tomar uma decisão errônea.

Definição 2.12 (Erros de decisão)**.** *Seja δ uma função teste. Se $\theta \in \Theta_0$ ocorrerá um erro se $\delta(\underset{\sim}{y}) = 1$ e a probabilidade de acontecer este erro, chamado de erro tipo I, é*

$$P\big(\delta(\underset{\sim}{Y}) = 1\big) \;=\; \mathrm{E}\big(\delta(\underset{\sim}{Y})\big), \quad \theta \in \Theta_0.$$

Também, se $\theta \in \Theta_A$ ocorrerá um erro se $\delta(\underset{\sim}{y}) = 0$ e a probabilidade de acontecer este erro, chamado de erro tipo II, é

$$P\big(\delta(\underset{\sim}{Y}) = 0\big) \;=\; 1 - \mathrm{E}\big(\delta(\underset{\sim}{Y})\big), \quad \theta \in \Theta_A.$$

O procedimento utilizado historicamente foi o de limitar a probabilidade de erro tipo I para alguns valores pequenos pré determinados de α, geralmente 0.01, 0.05 ou inclusive 0.10, e minimizar a probabilidade de erro do tipo II.

Definição 2.13 (Nível de significância)**.** *Dizemos que a função teste δ tem nível de significância α se*

$$\mathrm{E}\big(\delta(\underset{\sim}{Y})\big) \leq \alpha,$$

para todo $\theta \in \Theta_0$.

devemos particionar o espaço $\underset{\sim}{y} \in \mathbb{R}_n$ em dois subconjuntos disjuntos Ω_0 e Ω_A de tal modo que, se

De tempos em tempos, a discussão é útil usar a descrição curta-mão de H0 como sendo possivelmente verdadeira. Agora em termos estatísticos H0 se refere a um modelo de probabilidade e "modelo"a própria palavra implica idealização. Com uma muito poucas exceções possíveis seria absurdo pensar que um modelo matemático é uma representação exata de um sistema real e, nesse sentido, todos H0 são definidos dentro de um sistema que é falso. Nós usamos o termo para significar que, no estado atual de conhecimento é razoável proceder como se a hipótese é verdadeira. Note-se que uma hipótese objecto subjacente, como que um determinado ambiente exposição tem absolutamente nenhum efeito sobre a evolução da doença em particular pode de fato ser verdade.

p-valor associado a testes de hipóteses

Até este ponto, temos assumido um nível de significância fixo α quando se discute testes de hipóteses. Isto é, dado α, definimos uma função de teste δ e então rejeitamos a hipótese nula ao nível α se $\delta(\underset{\sim}{y}) = 1$.

Uma abordagem alternativa, mais de acordo com a atual prática, é considerar uma família de funções de teste δ_α para $0 < \alpha < 1$ em que a função de teste δ_α tem nível de significância α, não mais fixo. Vamos assumir também uma família de funções teste $\{\delta_\alpha\}$ satisfazendo que se $\delta_{\alpha_1}(\underset{\sim}{y}) = 1$ implica que $\delta_{\alpha_2}(\underset{\sim}{y}) = 1$, para quaisquer $\alpha_1 < \alpha_2$.

Definição 2.14 (p-valor)**.** *Definimos p-valor ou nível de significância observado como*

$$p\text{-}valor = \inf\big(\alpha : \delta_\alpha(\underset{\sim}{y}) = 1\big)\cdot \tag{2.22}$$

Esta definição implica que o p-valor é o menor valor de nível de significância α para o qual a hipóteses nula seria rejeitada ao nível α. O p-valor é a mais utilizada medida de prova contra a hipótese nula: quanto menor o p-valor, mais provas contra a hipótese nula. Embora este uso de p-v alores é bastante comum em estatística prática, a sua utilização como um ofevidence medida é muito controversa. em em particular, é difícil de calibrar valores p como ofevidence medidas (Goodman, 1999a).

2.3.2 Testes acerca dos parâmetros da regressão

Na apresentação dos resultados do ajuste de qualquer modelo de regressão são mostrados testes de hipóteses acerca dos coeficientes da regressão. Nestes testes basicamente queremos saber se

$$H_0 : \beta_k = 0 \qquad \text{ou se} \qquad H_A : \beta_k \neq 0, \tag{2.23}$$

para $k = 0, \cdots, p$.

O objetivo destes testes é verificar se a variável explicativa correspondente ao coeficiente testado influencia ou não na resposta. A ideia do teste é que se β_k for zero, o que equivale à conclusão de aceitar a hipótese nula, então a k-ésima variável regressora pode ser retirada do modelo, indicando que esta variável não influencia na resposta.

Para testar as hipóteses propostas precisamos construir estatísticas de teste e para isso utilizamos o resultado no Teorema 2.7 acerca da distribuição de probabilidades dos estimadores do vetor de parâmetros da regressão. O teorema referido afirma que se $Y_1, \cdots, Y_n$ satisfazem um modelo de regressão linear e, portanto, as suposições na Seção 1.2.1 são cumpridas o vetor $\widehat{\beta}$ tem como distribuição de probabilidades $N_p\big(\beta, \sigma^2(X^\top X)^{-1}\big)$.

Bastaria construir uma medida entre o estimador $\widehat{\beta}_k$ e o valor teórico pretendido do parâmetro $\beta_k = 0$, de maneira que possamos estabelecer uma distribuição de probabilidades para essa mediada.

Teorema 2.21. *Seja Y um vetor aleatório satisfazendo o modelo de regressão $Y = X\beta + \epsilon$, onde $\epsilon \sim N_n(0, \sigma^2\,\mathrm{I})$. Para testar as hipóteses em (2.23) utilizamos a estatística de teste*

$$t_{obs_k} = \frac{\widehat{\beta}_k}{\sqrt{\widehat{\sigma}^2(X^\top X)^{-1}_{kk}}}, \tag{2.24}$$

onde $(X^\top X)^{-1}_{kk}$ representa o k-ésimo elemento da diagonal principal da matriz $(X^\top X)^{-1}$, para $k = 0, \cdots, p$. A estatística t_{obs_k} tem por distribuição de probabilidade marginal t-Student com $n - p - 1$ graus de liberdade e calculamos o p-valor correspondente como $2P(t_{(n-p-1)} > |t_{obs}|)$.

Demonstração. Exercício. □

Devemos observar duas coisas: a expressão no denominador de (2.24) representa o estimador da variância de $\widehat{\beta}_k$, ou seja,

$$\widehat{\mathrm{Var}(\widehat{\beta}_k)} = \widehat{\sigma}^2(X^\top X)^{-1}_{kk}\cdot \tag{2.25}$$

Devemos observar também que

$$P\left(t_{obs_k} > t\right) = P\left[\widehat{\beta}_k / \sqrt{\widehat{\sigma}^2 (X^\top X)^{-1}_{kk}} > t\right] = \alpha,$$

é a probabilidade de acontecer o erro tipo I e, desta expressão, obtemos que a função teste é

$$\underset{\sim}{\delta(y)} = \begin{cases} 0 & \text{se} \quad \widehat{\beta}_k < t_{\alpha/2(n-p)} / \sqrt{\widehat{\sigma}^2 (X^\top X)^{-1}_{kk}} \\ 1 & \text{se} \quad \widehat{\beta}_k \geq t_{\alpha/2(n-p)} / \sqrt{\widehat{\sigma}^2 (X^\top X)^{-1}_{kk}} \end{cases} .$$

Vejamos no seguinte exemplo como são apresentados os resultados das estatísticas de teste para os parâmetros da regressão.

Exemplo 2.11. *Continuando o Exemplo* **??**. *Apresentamos a tabela resumo do ajuste do modelo a qual estamos em condições de entender quase toda ela. As colunas representam as estimativas dos parâmetros da regressão obtidos por mínimos quadrados, o desvio padrão calculado como raiz quadrada positiva da expressão da variância estimada de cada $\widehat{\beta}$ mostrada em (2.25), a correspondente estatística de teste (2.24) e o p-valor associado. Por último é apresentada a estimativa do desvio padrão do erro, ou seja, apresenta-se o valor de $\widehat{\sigma}^2$ e também que $n - p - 1 = 9$.*

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.290909   0.749146  16.407 5.17e-08 ***
Producao    -0.030773   0.003571  -8.616 1.22e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.7491 on 9 degrees of freedom
```

*Ainda no exemplo, podemos ter como referência a significância do p-valor em relação a valores padrão deste. Assim, o código '***' significa que o p-valor associado à estatística de teste é aproximadamente zero ou seja, temos que $2P(t_{(9)} > |16.407|) \approx 0$ e que também $2P(t_{(9)} > |-8.616|) \approx 0$. O código '**' significa que o p-valor correspondente seria menor do que 0.001, o código '*' significa que o p-valor correspondente seria menor do que 0.05, o código '.' significa que o p-valor correspondente seria menor do que 0.1 e se não aparecer nada ao lado do p-valor significa que este é aproximadamente 1. Considerando como certas as condições do Teorema 2.21 rejeitamos as hipóteses nulas e, portanto, o tamanho do lote de Produção influencia no Custo de produção.*

É muito importante é notar que este teste é marginal, isto é, não considera a influência conjunta dos outros estimadores dos coeficientes das variáveis regressoras no modelo. Esta última afirmação significa que para confiarmos nos resultados deste teste devemos ter certeza de que os estimadores dos parâmetros da regressão são independentes ou, pelo menos, aproximadamente independentes.

Significa também que o teste acima somente pode ser utilizado quando os estimadores dos coeficientes das variáveis regressoras são independentes o qual não é comum de acontecer em modelos de regressão. Para saber se podemos ou não confiar nos resultados do teste mostrados na saída do ajuste devemos observar a matriz de correlações de $\widehat{\beta}$.

Exemplo 2.12. *Continuação do exemplo* **??**. *Utilizando a opção* `cor=T` *no sumario do ajuste, isto é, digitando*

```
> summary(ajuste, correlation = T)
```

obtemos como resposta, nas últimas linhas

```
Correlation of Coefficients:
         (Intercept)
Producao -0.95
```

o qual significa que

$$\text{Corr}(\widehat{\beta}_0, \widehat{\beta}_1) = -0.95.$$

Este valor de correlação é muito alto, portanto, não são confiáveis os resultados dos teste de hipóteses marginais acerca dos parâmetros de regressão.

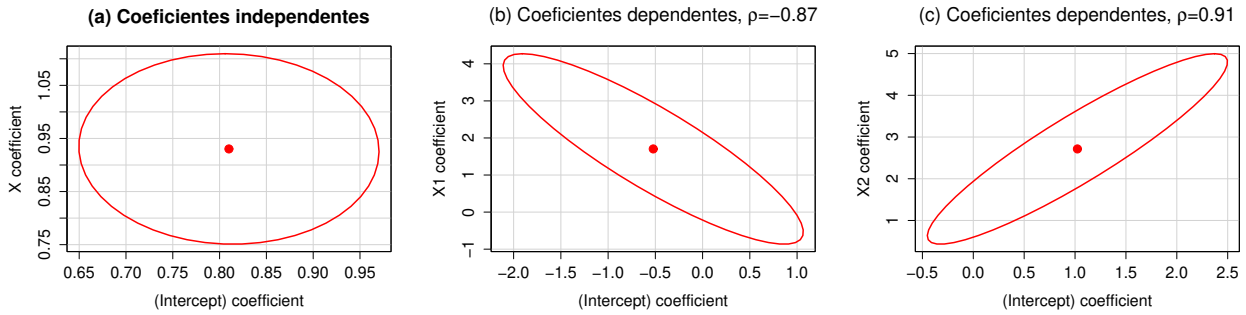

Figura 2.6: Elipses de confiança num modelo de regressão teórico: (a) coeficientes da regressão independentes, (b) coeficientes altamente dependentes negativamente e (c) altamente dependentes positivamente.

Observando as elipses de confiança entre cada par de coeficientes da regressão (Murdoch & Chow, 1996) temos uma outra forma de perceber se a correlação entre os estimadores dos parâmetros da regressão é muito alta ou não. As gráficos mostradas na Figura 2.6 correspondem aos contornos obtidos das elipses baseadas em variáveis aleatórias com distribuição normal bivariada de variância unitária e correlação ρ.

As elipses de confiança são os pontos (x, y) que satisfazem a relação

$$(x, y) = \Big(\cos(\theta + d/2), \cos(\theta - d/2) \Big), \tag{2.26}$$

onde $\theta \in (0, 2\pi)$ e $\cos(d) = \text{Corr}(\widehat{\beta}_0, \widehat{\beta}_1)$.

Exemplo 2.13. *Continuação do Exemplo* **??**. *Na Figura* **??** *mostramos a elipse de confiança entre os coeficientes* $\widehat{\beta}_0$ *e* $\widehat{\beta}_1$*, obtida pelo comando*

```
> library(car)
> confidenceEllipse(ajuste, col = "red")
```

no pacote R car. Podemos perceber a forte correlação negativa entre estes, invalidando os testes de hipóteses marginais.

O que fazer se como no Exemplo 2.4 cujos resultados estamos acompanhando, acontece que os estimadores dos coeficientes da regressão forem altamente correlacionados invalidando o teste marginal? Uma forma de verificar a significância dos coeficientes da regressão é utilizando a Análise de Variância da regressão ou ANOVA da regressão outra forma de libar com este problema será apresentada na Seção 2.3.4.

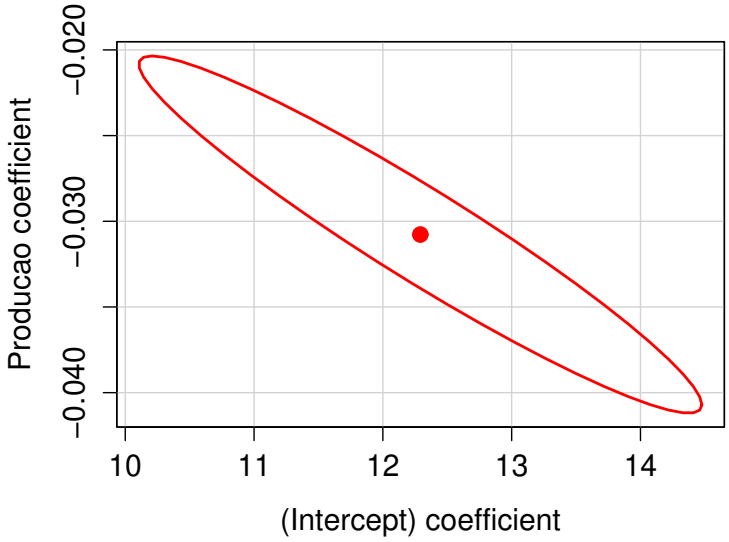

Figura 2.7: Elipse de confiança no modelo de regressão do Exemplo 2.4. Coeficientes da regressão altamente dependentes negativamente.

2.3.3 Análise de Variância da regressão

Vamos agora abordar a questão de como verificar se a maior parte da variação dos dados tenha sido explicada pelo modelo de regressão. Uma maneira de realizar isso é utilizando a tabela de Análise de Variância da regressão, um item que é frequentemente apresentado juntamente com as estimativas do modelo de regressão linear.

Neste caso o teste de hipóteses de interesse é

$$H_0 : \beta_1 = \cdots = \beta_p = 0,$$

e alternativa H_A : que algum $\beta_k \neq 0$, para $k = 1, \cdots, p$. Para construirmos uma estatística de teste considere a seguinte identidade

$$(Y_i - \widehat{\mu}_i) = (Y_i - \overline{Y}) - (\widehat{\mu}_i - \overline{Y}). \tag{2.27}$$

Esta identidade significa que $r_i = Y_i - \widehat{\mu}_i$, chamado de resíduo ordinário, é a diferença entre duas quantidades: (1) o desvio da observação Y_i a partir da média $\overline{Y}$ e (2) o desvio do valor predito $\widehat{\mu}_i$ da média $\overline{Y}$.

A diferença em (2.27) pode ser escrita como $(Y_i - \overline{Y}) = (\widehat{\mu}_i - \overline{Y}) + (Y_i - \widehat{\mu}_i)$, da qual elevando ao quadrado cada termo e somando sob todas as observações temos

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(\widehat{\mu}_i - \overline{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{\mu}_i)^2. \tag{2.28}$$

O termo a esquerda desta expressão é a soma de quadrados total que será denotada por SQT, enquanto o primeiro termo do lado direito é a soma de quadrados devida à regressão de Y sobre as p variáveis explicativas, sendo denotado por SQReg. O segundo termo é chamado de soma de quadrados do resíduo, denotada como SQRes. Observemos que $\overline{Y}$ é o valor predito do modelo de

regressão linear $Y = \beta_0 + \epsilon$, no qual não são consideradas variáveis explicativas; por isso o termo SQReg mede o efeito preditivo somente das variáveis explicativas.

Uma maneira de se medir a adequação do ajuste é comparando a soma de quadrados residual, que se espera seja pequena, com a soma de quadrados devida à regressão. Somas de quadrado divididas pelos seus graus de liberdade, sob certas condições, tem como distribuição de probabilidade χ^2 e quociente de variáveis aleatórias independentes χ^2 tem como distribuição F-Fisher.

O Teorema de Fisher-Cochran (Searle, 1971; Rao, 1973), apresentado logo em seguida, é fundamental para determinar a distribuição das somas de quadrados definidas em (2.28) e deduzir a distribuição de probabilidades da estatística de teste. A conclusão deste teorema é que, sob a suposição de normalidade, as formas quadráticas envolvidas são independentes e com distribuição χ^2.

Teorema 2.22 (Fisher-Cochran)**.** *Sejam $Y_1, \cdots, Y_n$ variáveis aleatórias independentes com distribuição $N(0, \sigma^2)$. Suponhamos que*

$$Y^\top A Y = Y^\top A_1 Y + Y^\top A_2 Y + \cdots + Y^\top A_k Y,$$

onde cada $Y^\top A_i Y$ é uma forma quadrática definida não-negativa. As matrizes A assim como cada A_i, são simétricas idempotentes tais que $\text{posto}(A) = r$ e $\text{posto}(A_i) = r_i$, $i = 1, 2, \cdots, k$. Se

$$r_1 + r_2 + \cdots + r_k = r$$

então $Y^\top A_1 Y, Y^\top A_2 Y, \cdots, Y^\top A_k Y$ são variáveis aleatórias independentes com distribuição

$$\frac{Y^\top A_i Y}{\sigma^2} \sim \chi^2_{(r_i)},$$

para $i = 1, 2, \cdots, k$.

Demonstração. Demonstraremos para o caso $k = 2$, não é difícil generalizar. Outras demonstrações podem ser encontradas em livros clássicos como Scheffé (1959), Rao (1973) e Graybill (1976), dentre outros. Provar que cada forma quadrática tem como distribuição χ^2 é uma consequência direta da aplicação do Teorema 2.14. Demonstremos então a independência. De maneira similar ao Teorema 2.17 pode-se provar que duas formas quadráticas $Y^\top A_1 Y$ e $Y^\top A_2 Y$ são independentes se $A_1 A_2 = 0$ (exercício). Provemos então que $A_1 A_2 = 0$. Dado que A é uma matriz simétrica existe uma matriz ortogonal Q de forma que $Q^\top A Q = \Lambda$, onde Λ é uma matriz diagonal com elementos os valores próprios de A. Podemos escolher Q de maneira que Λ é da forma

$$Q^\top A Q = \Lambda = \begin{pmatrix} \mathrm{I}_r & 0 \\ 0 & 0 \end{pmatrix},$$

devido a que os autovalores de A são ou 0 ou 1, pelo fato de A ser idempotente. Então

$$Q^\top A Q = Q^\top A_1 Q + Q^\top A_2 Q,$$

a qual pode ser escrita como

$$\begin{pmatrix} \mathrm{I}_r & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} J_r & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} G_r & 0 \\ 0 & 0 \end{pmatrix}. \tag{2.29}$$

Dados que A_1 é simétrica, temos que

$$(Q^\top A_1 Q)^\top = Q^\top A_1 Q = \begin{pmatrix} J_r & 0 \\ 0 & 0 \end{pmatrix}.$$

Multiplicando em (2.29) a esquerda por $Q^\top A_1 Q$ obtemos

$$\begin{pmatrix} J_r & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} J_r & 0 \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} J_r G_r & 0 \\ 0 & 0 \end{pmatrix},$$

o qual equivale a identidade de formas quadráticas

$$Q^\top A_1 Q = Q^\top A_1 Q + (Q^\top A_1 Q)(Q^\top A_2 Q),$$

da qual concluímos que $(Q^\top A_1 Q)(Q^\top A_2 Q) = Q^\top A_1 A_2 Q = 0$ e portanto $A_1 A_2 = 0$. □

Sabemos agora como provar que, se temos uma forma quadrática a qual pode ser escrita como soma de outras formas quadráticas, os termos da soma são independentes. Falta-nos saber qual a distribuição de probabilidade associada à ANOVA. Para isso definimos a chamada distribuição F-Fisher.

Definição 2.15. *Sejam Y_1 e Y_2 duas variáveis aleatórias independentes χ^2 com m_1 e m_2 graus de liberdade, respectivamente. Dizemos que a variável aleatória*

$$F = \frac{Y_1/m_1}{Y_2/m_2}, \tag{2.30}$$

tem por distribuição de probabilidades F-Fisher com (m_1, m_2) graus de liberdade. Escrevemos resumidamente $F \sim F_{(m_1,m_2)}$.

A ideia geral é a de dividir a soma dos quadrados das observações em um certo número de formas quadráticas, onde cada uma corresponde a uma causa de variação. Esta nova distribuição de probabilidades nos disse que quociente de formas quadráticas independentes pode ter distribuição de probabilidades conhecida. Nossos próximos passos são nesse sentido.

Teorema 2.23. *A função de densidade da variável aleatória definida em (2.30) é dada por*

$$g(f) = \frac{\Gamma[(m_1+m_2)/2]}{\Gamma(m_1/2)\Gamma(m_2/2)} \left(\frac{m_1}{m_2}\right) \left(\frac{m_1}{m_2} f\right)^{(m_1/2)-1} \left(1 + \frac{m_1}{m_2} f\right)^{-(m_1+m_2)/2},$$

se $f > 0$ e zero em caso contrário.

Demonstração. Exercício. □

Temos então a chamada tabela de Análise de Variancia da regressão, também chamada de ANOVA da regressão.

Tabela 2.1: Tabela de Análise de Variancia da regressão.

Efeito	g.l	Soma de Quadrados	Quadrados Médios	F-Fisher
Regressão	p	SQReg	QMReg	F_{Obs} = QMReg/QMRes
Residual	$n-p-1$	SQRes	QMRes	
Total	$n-1$	SQT		

Na Tabela 2.1 mostramos que o quadrado médio da regressão, calcula-se como QMReg=SQReg/p, sendo que SQReg=$\sum_{i=1}^n (\widehat{\mu}_i - \overline{Y})^2$. O quadrado médio dos resíduos, calcula-se como QMRes=SQRes/($n-$

$p-1$), onde SQRes=$\sum_{i=1}^{n}(Y_i-\widehat{\mu}_i)^2$. Também temos que o código g.l. significa graus de liberdade associados a cada soma de quadrados e F-Fisher é o nome da estatística de teste de valor observado F_{Obs}.

O propósito é testar a significância não de cada elemento do vetor β, testa-se a significância conjunta das componentes de β efetivamente associadas às variáveis explicativas. Queremos dizer que testamos a significância de todos os coeficientes além do intercepto.

Justificaremos os termos na Tabela 2.1 escrevendo-os matricialmente. Não é difícil perceber que podemos escrever a soma de quadrados total como

$$SQT = Y^\top \left(\mathrm{I} - \frac{1}{n}\mathbf{1} \right) Y, \tag{2.31}$$

onde $\mathbf{1}$ é uma matriz onde todos seus elementos são o número um. Também, a soma de quadrados dos resíduos pode ser escrita como

$$SQRes = Y^\top (\mathrm{I} - H) Y \tag{2.32}$$

e por último, a soma de quadrados da regressão

$$SQReg = Y^\top \left(H - \frac{1}{n}\mathbf{1} \right) Y\cdot \tag{2.33}$$

Desta forma, se as somas de quadrados envolvidas na tabela ANOVA cumprirem as condições do Teorema 2.22, teríamos justificada a utilização da distribuição F-Fisher. Neste sentido, observemos que $A = (\mathrm{I} - \frac{1}{n}\mathbf{1})$, $A_1 = (H - \frac{1}{n}\mathbf{1})$ e $A_2 = (\mathrm{I} - H)$, é claro que $A = A_1 + A_2$. Mais complicado é calcular o posto de cada matriz, o Exercício 7 item (c), Capítulo 1, facilita o trabalho.

Todas as matrizes envolvidas nas expressões em (2.31), (2.32) e (2.33) são idempotentes. A demonstração de que as matrizes I e $\frac{1}{n}\mathbf{1}$ são idempotentes é simples e é exercício para o leitor. A situação da matriz H é mais complexa. Esta matriz será estudada profundamente no Capítulo 3, no Teorema 3.6 é demonstrado que $\text{posto}(H) = p+1$. Com esta informação podemos afirmar que $\text{posto}(A) = n-1$, $\text{posto}(A_1) = p$ e que $\text{posto}(A_2) = n-p-1$ justificando-se assim as somas de quadrados na Tabela ANOVA e, pelo Teorema 2.22, são independentes.

Utilizamos agora a Definição 2.15. A densidade F-Fisher é definida como quociente de somas de quadrados independentes ponderadas pelos graus de liberdade, isso justifica o fato de F_{Obs}=QMReg/QMRes ter como distribuição de referência $F_{(p,n-p-1)}$. Assim, o p-valor correspondente à estatística de teste F_{Obs} calcula-se como $P(F_{(p,n-p-1)} > F_{obs})$.

Exemplo 2.14. *Continuando o exemplo* **??**. *Uma vez obtido o ajuste do modelo de regressão proposto, tentou-se verificar a significância da variável explicativa. Observou-se que os estimadores dos parâmetros da regressão são altamente dependentes o qual levantou suspeitas acerca da utilidade da estatística de teste marginal em (2.24). Surge como alternativa a ANOVA da regressão para verificar a significância de todas as variáveis explicativas.*

```
> anova(ajuste)
Analysis of Variance Table

Response: Custo
          Df Sum Sq Mean Sq F value    Pr(>F)
Producao   1 41.666  41.666  74.242 1.218e-05 ***
Residuals  9  5.051   0.561
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Com estes resultados concluímos que o tamanho do lote produzido, conteúdo da variável Produção, influencia fortemente no Custo. Interpretamos que `Df` *são os graus de liberdade,* `Sum Sq` *são as somas de quadrados da regressão e dos resíduos, sendo estas SQReg=41.666 e SQRes=5.051.* `Mean Sq` *simboliza os quadrados médios e o valor de* F_{Obs}*, em* `F value`*, é de 74.242.*

Dispomos também do p-valor associado à estatística F_{Obs} *em* `Pr(>F)`*, sendo este de 1.218e-05 ou 0.00001218. Ainda temos a oportunidade de decidir acerca da significância desta estatística de teste. Para isso comparamos o p-valor com valores historicamente de referência informados em* `Signif. codes` *ou códigos de significância.*

Vejamos como entender o simbolo ao lado direito do p-valor. Caso apareça um 0 significa que o p-valor associado é zero ou seja temos certeza ao aceitar o resultado do teste, isto não deve acontecer em situações práticas. Se o código for `'***'` *indica que o p-valor é menor do que 0.001 ou que temos menos do que 0.1% de probabilidade de errar ao afirmar que o lote produzido influencia no custo, esta é a interpretação neste exemplo. Se o código for* `'**'` *indica que o p-valor é menor do que 0.01, caso o código seja* `'*'` *indica que o p-valor é menor do que 0.05, se o código for* `'.'` *indica que o p-valor é menor do que 0.1 e neste caso temos, como máximo, 10% de probabilidade de errar ao aceitar o resultado do teste, em algumas situações considera-se um valor muito alto de erro. Caso não apareça nada significa que o p-valor é 1, ou seja, temos certeza de errar completamente, isto também não deve acontecer em situações práticas.*

2.3.4 Verossimilhança perfilada

Num determinado modelo estatístico podemos estar interessados em parte do vetor de parâmetros e não no vetor completo ϑ. Especificamente, se o vetor de parâmetros completo ϑ pode ser decomposto como $\vartheta = (\psi, \zeta)$ e nos interessa estimar e inferir acerca de valores de ψ, chamaremos ψ de vetor de parâmetros de interesse e ao vetor ζ de parâmetros de perturbação. Em situações como estas definimos e função de verossimilhança perfilada.

Em situações como estas é possível, por diferentes metodologias, construir uma função que dependa somente de ψ e que possamos utilizar para realizar inferências acerca de ψ. Estas funções são conhecidas como funções de pseudo-verossimilhança. Uma destas, à qual têm sido considerados mutos esforços na literatura é a chamada de função de verossimilhança perfilada.

Diversas destas funções têm sido consideradas na literatura e muitos esforços dedicados a uma delas, a função de verossimilhança perfilada. Devemos ressaltar que esta função somente tem sentido quando parte do vetor de parâmetros que define o modelo estatístico em estudo é considerado como de perturbação.

Definição 2.16. *Define-se a da função de verossimilhança perfilada para* ψ *como*

$$L_P(\psi) = \max_{\zeta} L(\psi, \zeta),$$

sendo que o máximo é obtido em todo o espaço paramétrico Ω*.*

Quando explicamos o princípio da estimação por máxima verossimilhança em (2.6), indicamos que a maximização pode ser realizada tanto na função de verossimilhança quanto no logaritmo neperiano dela. Assim, o logaritmo da função de verossimilhança perfilada é dado por

$$\ell_P(\psi; \underset{\sim}{y}) = \log L_P(\psi; \underset{\sim}{y}). \tag{2.34}$$

A função ℓ_P têm interesse pela facilidade no cálculo, ou seja, é mais fácil trabalharmos com o logaritmo da função de verossimilhança do que com a função de verossimilhança propriamente.

Em situações como estas é possível, por diferentes metodologias, construir uma função que dependa somente de ψ e que possamos utilizar para realizar inferências acerca de ψ. Estas funções são conhecidas como funções de pseudo-verossimilhança. Uma destas, à qual têm sido considerados mutos esforços na literatura é a chamada de função de verossimilhança perfilada.

Observemos que o processo de maximização ao qual faz referência a Definição 2.16 é realizado quando obtemos $\widehat{\zeta}(\psi)$. Desta forma o logaritmo da função de verossimilhança perfilada pode ser definido como

$$\ell_P(\psi) = \ell(\psi, \widehat{\zeta}(\psi)),$$

e é esta função que terá utilidade nos modelos lineares para resolver a questão da influência indesejada entre covariáveis. O valor de ψ no qual se atinge o ponto de máximo de ℓ_P será chamado de estimador de máxima verossimilhança perfilado e denotado por

$$\widetilde{\psi} \;=\; \arg\max \ell_P(\psi)\cdot \tag{2.35}$$

No seguinte exemplo mostramos a forma desta função no modelo linear simples.

Exemplo 2.15. *Seja Y um vetor aleatório satisfazendo um modelo de regressão linear simples. Na Seção 2.1.3 foi obtido que o logaritmo da função de verossimilhança neste modelo é*

$$\ell(\beta_0, \beta_1, \sigma^2; \underset{\sim}{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \sum_{i=1}^{n}\frac{(y_i - \beta_0 - \beta_1 x_i)^2}{2\sigma^2},$$

da qual obtemos que $\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1\overline{x}$,

$$\widehat{\beta}_1 \;=\; \Big(\sum_{i=1}^{n} x_i y_i - n\overline{x}\overline{y}\Big)\Big/\Big(\sum_{i=1}^{n} x_i^2 - n\overline{x}^2\Big)$$

e

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2\cdot$$

Considerando $\widehat{\beta}_0$ e $\widehat{\sigma}^2$ como funções de β_1 e substituindo no logaritmo da função de verossimilhança acima, obtemos que o logaritmo da função de verossimilhança perfilada para β_1 assuma a forma

$$\begin{aligned}
\ell_P(\beta_1) \;&=\; -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\Big(\frac{1}{n}\sum_{i=1}^{n}\big(y_i - (\overline{y} - \beta_1\overline{x}) - \widehat{\beta}_1 x_i\big)^2\Big) \\
&\qquad - \frac{2\dfrac{1}{n}\displaystyle\sum_{i=1}^{n}\big(y_i - (\overline{y} - \beta_1\overline{x}) - \widehat{\beta}_1 x_i\big)^2}{\displaystyle\sum_{i=1}^{n}\big(y_i - (\overline{y} - \beta_1\overline{x}) - \widehat{\beta}_1 x_i\big)^2} \\
&=\; -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\Big(\frac{1}{n}\sum_{i=1}^{n}\big(y_i - (\overline{y} - \beta_1\overline{x}) - \widehat{\beta}_1 x_i\big)^2\Big) \;-\; \frac{n}{2}\cdot
\end{aligned} \tag{2.36}$$

Mostramos na Figura 2.8 o comportamento da função de verossimilhança perfilada (2.36) obtida para o caso do modelo de regressão simples e aplicada aos dados do Exemplo 2.4. Essa figura pode ser obtida utilizando as funções `profilelike.lm` e `profilelike.plot`, ambas programadas no pacote de funções `ProfileLikelihood`, da maneira apresentada a seguir.

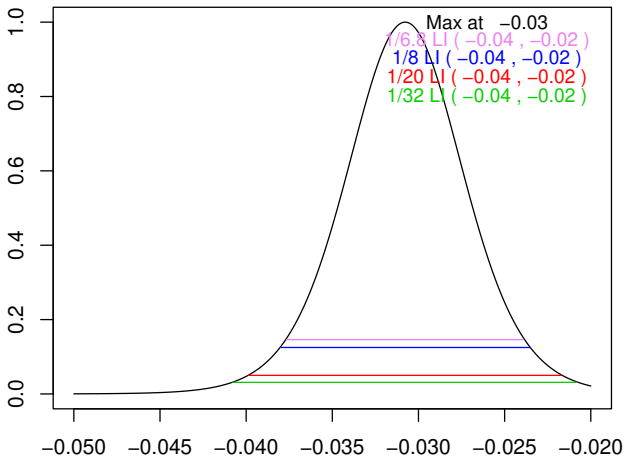

Figura 2.8: Função de verossimilhança perfilada $L_P(\beta_1)$.

```
> library(ProfileLikelihood)
> dados= data.frame(Custo, Producao)
> perfil = profilelike.lm(formula = Custo ~ 1, data = dados, profile.theta = "Producao")
> profilelike.plot(theta = perfil$theta, profile.lik.norm = perfil$profile.lik.norm)
```

As funções utilizadas para gerar o gráfico na Figura 2.8 somente estão definidas para o caso de variáveis explicativa contínuas, assim como a própria função de verossimilhança perfilada somente está definida para estas situações. Isso significa que não faz sentido procurarmos encontrar a função de verossimilhança perfilada para o intercepto β_0 nos modelos de regressão. Devemos esclarecer que o gráfico mostra a função de verossimilhança perfilada, não o logaritmo desta função como obtido em (2.36). A curva mostrada na Figura 2.8 é a exponencial ponderada de $\ell_P(\beta_1)$. O objetivo da ponderação é para sempre mostrarmos a curva no intervalo $(0, 1)$ do eixo vertical. Para isso fazemos $mx = \max(\ell_P(\beta_1))$ e mostramos a curva de $\exp(\ell_P(\beta_1) - mx)$. Na figura podemos encontrar a informação do valor de que $\widetilde{\beta}_1 = -0.03$ como `Max at. -0.03`, ou seja, o estimador de máxima verossimilhança perfilada de β_1 é -0.03, o qual é numericamente idêntico a $\widehat{\beta}_1$. Isto é uma propriedade, o estimador de máxima verossimilhança perfilada coincide com o estimador de máxima verossimilhança. As linhas no gráfico nos permitem encontrar diferentes limites de confiança para o estimador. A linha correspondente à $1/20 = 0.05$ delimita o intervalo de confiança de 95%, por isso podemos dizer que este intervalo é (-0.040,-0.022) aproximadamente. Podemos afirmar então que a `Producao` é significativa para explicar o `Custo`, isto porque o intervalo não contêm o zero. Ainda podemos dizer que esta afirmação procede, dado que isolamos desta maneira o parâmetro β_1. Construímos uma função que depende somente deste parâmetro, por isso a conclusão da influência da variável correspondente `Producao` na resposta, independe da correlação entre os estimadores de β_0 e β_1. Vejamos no seguinte exemplo como proceder no caso do modelo de regressão linear múltipla.

Continuando o Exemplo **??**. Uma vez obtido o ajuste do modelo de regressão proposto, tentou-se verificar a significância da variável explicativa. Observou-se que os estimadores dos parâmetros da regressão são altamente dependentes o qual levantou suspeitas acerca da utilidade da estatística de teste marginal em (2.24). Surge como alternativa a ANOVA da regressão para verificar a significância de todas as variáveis explicativas.

```
> library(ProfileLikelihood)
> perfil = profilelike.lm(formula = Custo ~ 1, data = dados, profile.theta = "Producao")
```

```
> Profile.beta1 = function(x,y,b1){
    n = length(y);
    b0 = mean(y)-b1*mean(x);
    lp = -n*log(2*pi)/2 - n*log(sum((y-b0-b1*x)^2)/n)/2 - n/2
    lp}
> lp.beta1 = rep(0,length(perfil$theta))
> for(i in 1:length(perfil$theta)){
> lp.beta1[i] = Profile.beta1(Producao,Custo,perfil$theta[i])}
> mm = max(lp.beta1)
> plot(perfil$theta, exp(lp.beta1-mm), type = "l", lty = 1, lwd = 1, col = "red")
```

Exemplo 2.16 (Continuação do Exemplo 2.5)**.** *Vejamos o resultado do ajuste do modelo de regressão linear múltipla incluindo a correlação entre os estimadores dos coeficientes da regressão.*

```
> ajuste = lm(Tempo~., data=Health)
> sumario = summary(ajuste, cor=T)
> sumario$correlation
            (Intercept)       Peso      Pulso      Pernas      Treino
(Intercept)   1.0000000  0.4117746 -0.6382122 -0.56087747 -0.26782766
Peso          0.4117746  1.0000000 -0.3508727 -0.72516830 -0.36261105
Pulso        -0.6382122 -0.3508727  1.0000000  0.38412754 -0.36379315
Pernas       -0.5608775 -0.7251683  0.3841275  1.00000000 -0.01759478
Treino       -0.2678277 -0.3626110 -0.3637931 -0.01759478  1.00000000
```

Podemos perceber que, em geral, a maioria das correlações entre os estimadores dos parâmetros da regressão são fracas, ou seja, são pequenas a menos das correlações entre os coeficientes das variáveis `Pulso` e `Intercepto` e `Peso` e `Pernas`. São essas correlações que sugerem desconfiar dos testes correspondes dos coeficientes da regressão. Ainda percebemos que a correlação entre os estimadores dos coeficientes das variáveis `Pernas` e `Treino` é quase zero. Uma maneira para melhor visualizar estas correlações é mostrando-as através de elipses de confiança, definidas em (2.26). As linhas de comando a seguir permitem-nos gerar estes gráficos.

```
> library(car)
> confidenceEllipse(ajuste, L = c("Peso", "Pulso"))
> confidenceEllipse(ajuste, L = c("Peso", "Pernas"))
> confidenceEllipse(ajuste, L = c("Peso", "Treino"))
> confidenceEllipse(ajuste, L = c("Pulso", "Pernas"))
> confidenceEllipse(ajuste, L = c("Pulso", "Treino"))
> confidenceEllipse(ajuste, L = c("Pernas", "Treino"))
```

Na Figura 2.9 apresentamos o resultado da utilização dos comandos acima. Percebemos que a correlação entre os estimadores dos coeficientes das variáveis `Peso` e `Pernas` é -0.072 o qual se manifesta numa elipse estreita, sugerindo assim uma relação linear negativa forte entre estes estimadores.

Reproduzimos a seguir um resumo da tabela resposta do ajuste do modelo de regressão do Exemplo 2.5, este resumo é obtido do resultado em `summary(ajuste)`.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -3.6186    56.1027  -0.064 0.949086
Peso          2.7947     0.6324   4.419 0.000168 ***
Pulso        -0.5252     0.8628  -0.609 0.548194
Pernas       -1.1134     0.5422  -2.054 0.050614 .
Treino        3.9030     0.7477   5.220 2.11e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

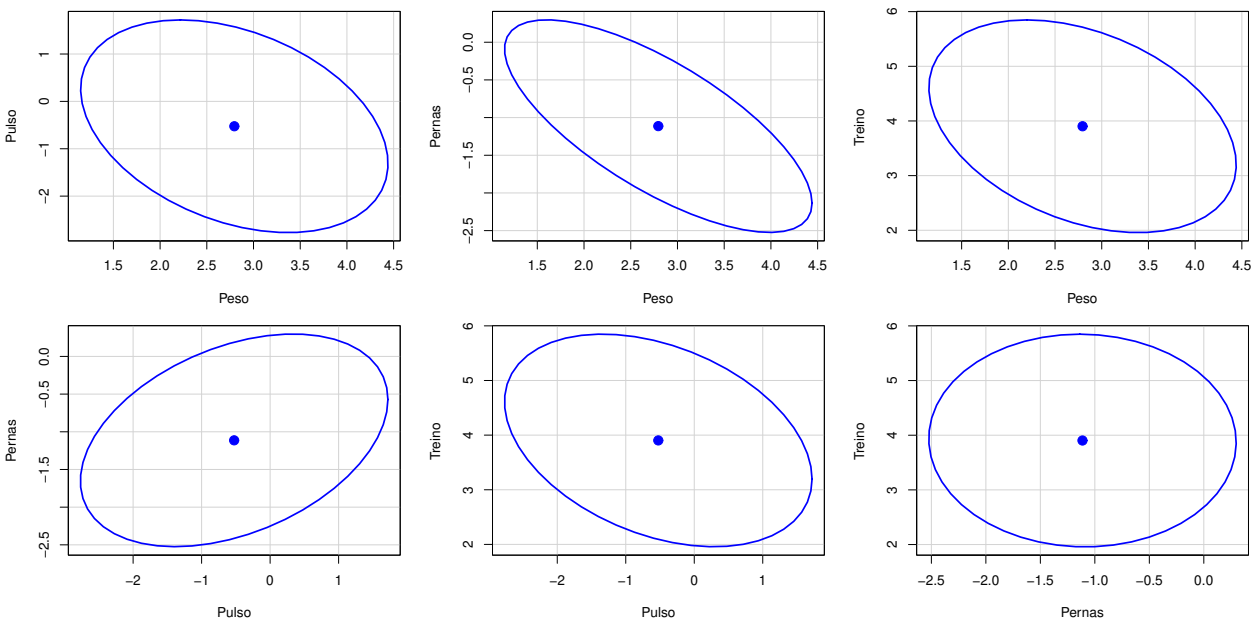

Figura 2.9: Elipses de confiança entre as variáveis explicativas contínuas no Exemplo 2.5.

Podemos perceber que as variáveis `Peso` e `Treino` são significativas na estimação da média do `Tempo` em segundos na prova de corrida, também vemos que `Pulso` não é significativo nem o `Intercepto` e que a variável `Pernas` têm influência duvidosa. A menos existam motivos teóricos que indiquem como mais adequado excluir o intercepto do modelo, este sempre será considerado na modelagem. Significa que não faremos inferências acerca da significância deste estimador. Exemplos de situações nas quais consideram-se modelos excluindo o intercepto podem ser encontrados nos Exercícios 16, 17, 18, 19 e 20 da Seção 2.1.

As linhas de comandos a seguir nos mostram como proceder para encontrarmos as funções de verossimilhança perfilada para cada um dos coeficientes associados à variáveis explicativas contínuas no Exemplo 2.5. O resultado é apresentado na Figura 2.10.

```
> library(ProfileLikelihood)
> par(mfrow = c(1,1), mar=c(2,2,1,1))
> perfil1 = profilelike.lm(formula = Tempo ~ .-Peso, data = Health, profile.theta = "Peso")
> profilelike.plot(perfil1$theta, perfil1$profile.lik.norm)
> text(4.5, 0.8, "Peso", cex = 1.2)
> perfil2 = profilelike.lm(formula = Tempo ~ .-Pulso, data = Health, profile.theta = "Pulso")
> profilelike.plot(perfil2$theta, perfil2$profile.lik.norm)
> text(-3.0, 0.8, "Pulso", cex = 1.2)
> perfil3 = profilelike.lm(formula = Tempo ~ .-Pernas, data = Health, profile.theta = "Pernas")
> profilelike.plot(perfil3$theta, perfil3$profile.lik.norm)
> text(-3.0, 0.8, "Pernas", cex = 1.2)
> perfil4 = profilelike.lm(formula = Tempo ~ .-Treino, data = Health, profile.theta = "Treino")
> profilelike.plot(perfil4$theta, perfil4$profile.lik.norm)
> text(6.0, 0.8, "Treino", cex = 1.2)
```

Fica claro dos gráficos na Figura 2.10 a significância de cada parâmetro: `Peso`, `Pernas` e `Treino` são significativas para explicar a resposta média do `Tempo` e `Pulso` claramente não. Faremos a seguir a estimação num novo modelo, desta vez eliminando a variável `Pulso`. Isto pode ser realizado de diversas maneira, uma delas seria escrevendo o modelo sem a inclusão da referida variável uma outra forma, pela qual nos decidimos proceder, solicitamos uma atualização dp modelo anterior

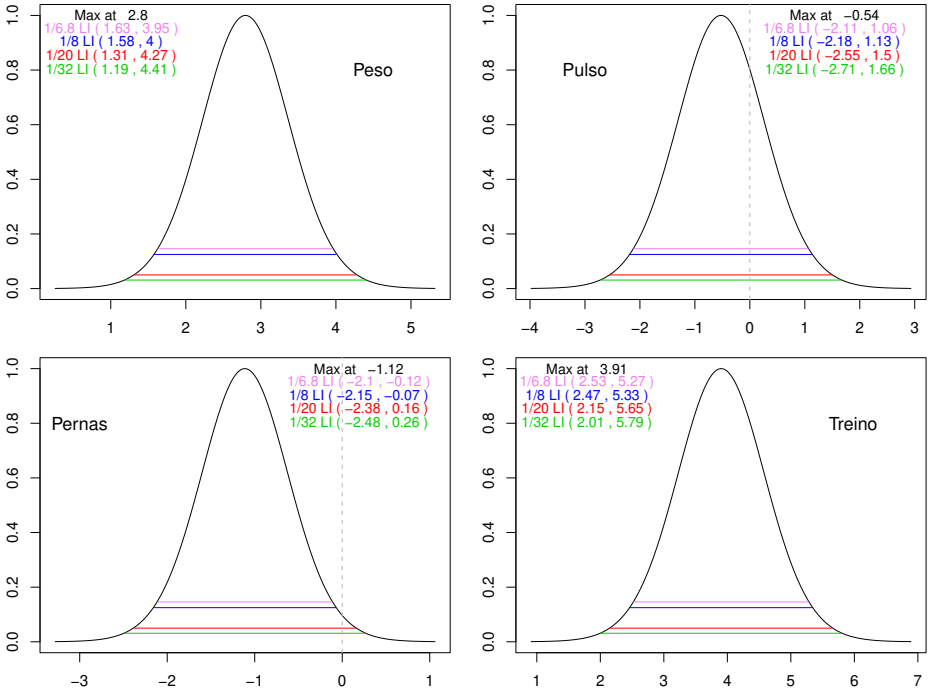

Figura 2.10: Funções de verossimilhança perfilada das variáveis explicativas contínuas do Exemplo 2.5.

indicando somente a exclusão da variável `Pulso`. Esta forma de trabalho é mostrada nos comandos a seguir.

```
> ajuste1 = update(ajuste, . ~ .-Pulso)
> summary(ajuste1, cor=T)
Call:
lm(formula = Tempo ~ Peso + Pernas + Treino, data = Health)

Residuals:
Min      1Q Median      3Q     Max
-48.52 -21.89  -5.03  17.97  49.34

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -25.4145    42.6652  -0.596 0.556544
Peso          2.6596     0.5850   4.546 0.000111 ***
Pernas       -0.9866     0.4945  -1.995 0.056590 .
Treino        3.7374     0.6880   5.432 1.08e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 28.32 on 26 degrees of freedom
Multiple R-squared:  0.8509,Adjusted R-squared:  0.8337
F-statistic: 49.47 on 3 and 26 DF,  p-value: 6.969e-11

Correlation of Coefficients:
        (Intercept) Peso  Pernas
Peso    0.26
Pernas -0.44       -0.68
Treino -0.70       -0.56  0.14
```

Podemos perceber que `Peso` e `Treino` são significativas para o `Tempo` médio estimado, mas `Pernas` está num limiar de significância. Nessas situações, junto com o detalhe da forte influência da variável `Peso` no estimador do coeficiente da variável `Pernas`, correlação linear entre ambas de -0.68, faz duvidar na influência da variável `Pernas` na resposta. Essa dúvida é resolvida, por exemplo, mostrando a função de verossimilhança perfilada correspondente. Na Figura 2.11 mostramos as funções de verossimilhança perfilada para cada uma das três variáveis explicativas: `Peso`, `Pernas` e `Treino`. Percebemos que todas influenciam significativamente na resposta.

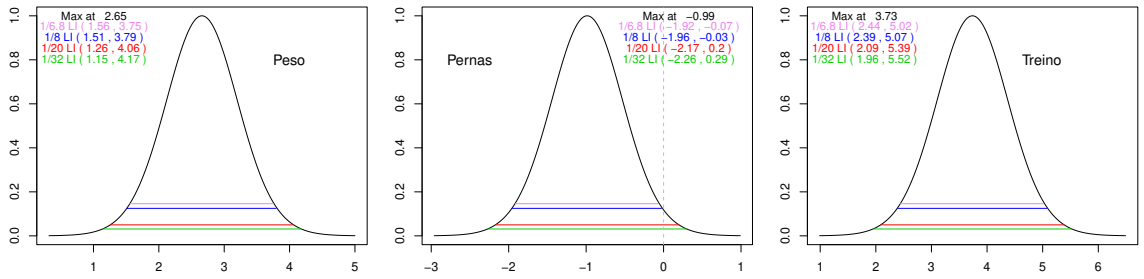

Figura 2.11: Funções de verossimilhança perfilada das variáveis explicativas contínuas do Exemplo 2.5.

Como conclusão, afirmamos que o modelo de regressão linear mais adequado é:

$$\widehat{\mathrm{E}(\mathrm{Tempo})} = -25.4145 + 2.6596\,\mathrm{Peso} - 0.9866\,\mathrm{Pernas} + 3.7374\,\mathrm{Treino}. \tag{2.37}$$

Observe que as estimativas dos parâmetros da regressão mudaram, fundamentalmente modificou o valor estimado do intercepto, os outros nem tanto. Isto porque foi eliminada uma variável não significativa para a resposta média do modelo. Realizando operações simples percebemos que as estimativas dos coeficientes das variáveis `Peso` e `Treino` mantiveram 95% do valor obtido com o modelo completo e a estimativa do coeficiente da variável `Pernas` manteve 88% do valor anterior. Por outro lado, a estimativa do intercepto (Intercept) no modelo em (2.37) é 7 vezes o valor anterior.

As linhas de comando a seguir permitem-nos construir uma tabela com a resposta estimada pelo modelo (2.37) para valores selecionados das variáveis explicativas.

```
> novos.Health = data.frame(Peso = c(60,70,80,90,100,110),
                      Pernas = c(70,80,90,100,110,120), Treino = c(60,65,70,75,80,90))
> predict.lm(ajuste1, novos.Health)
       1        2        3        4        5        6
289.3429 324.7597 360.1764 395.5932 431.0100 485.1139
```

Utilizando os resultados acima apresentamos as estimativas da média da resposta na Tabela 2.2. É claro que ainda resta verificarmos se as suposições do modelo de regressão são satisfeitas, isso faremos a través dos resíduos fundamentalmente no Capítulo 3. A qualidade do ajuste e detalhes relacionados serão considerados no Capítulo 4 e ainda no Capítulo 5 vamos estudar uma outra forma de escolher modelos, mas automática daquela utilizada até momento.

0.3cm

Variáveis	Valores selecionados					
Peso	60	70	80	90	100	110
Pernas	70	80	90	100	110	120
Treino	60	65	70	75	80	90
Tempo	289.3	324.8	360.2	395.6	431.0	485.1

Tabela 2.2: Resposta estimada do `Tempo` médio obtida para valores selecionadas das covariáveis.

2.4 Exercícios

Exercícios da Seção 2.1

1. Imagine que você tenha um conjunto de dados com quatro variáveis explicativas e $n = 42$ observações. Se o modelo de regressão linear considera incluir o intercepto, qual seria a ordem das matrizes $X^\top X$, $(X^\top X)^{-1}$, $X^\top Y$ e H?

2. De um conjunto de dados com uma variável independente e o intercepto obtemos a seguinte matriz

$$(X^\top X)^{-1} = \begin{pmatrix} \frac{31}{177} & \frac{-3}{177} \\ \frac{-3}{177} & \frac{6}{177} \end{pmatrix}.$$

   Quantas observações contém o conjunto de dados? encontre também o valor de $\sum_i x_i^2$, a soma dos quadrados da variável independente.

3. No modelo de regressão $Y = \beta_0 + \beta_1 x + \epsilon$ mostre que, se a média amostral $\overline{x}$ de x for zero, $\text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) = 0$ onde $\widehat{\beta}_0$ e $\widehat{\beta}_1$ são os estimadores de mínimos quadrados de β_0 e β_1. Prove também que $\text{Cov}(\overline{Y}, \widehat{\beta}_1) = 0$.

4. Considere o modelo de regressão $Y = \beta_0 + \epsilon$. Prove que $\widehat{\beta}_0 = \overline{Y}$ é o estimador de mínimos quadrados e de máxima verossimilhança de β_0.

5. Obtenha os estimadores dos parâmetros da regressão para o modelo

$$Y = \beta_0 + \beta_1 \widetilde{x} + \epsilon,$$

   onde $\widetilde{x} = (x - \overline{x})$. O modelo expresso desta forma é conhecido como modelo "centrado", neste caso a variável independente foi deslocada para ser simétrica ao redor do zero. Compare os resultados com os obtidos em (2.3) e (2.4).

6. Encontre os estimadores do modelo de regressão para os dados do Exemplo 2.4 utilizando o modelo centrado

$$Y = \beta_0 + \beta_1 \widetilde{x} + \epsilon,$$

   onde $\widetilde{x} = (x - \overline{x})$. Compare os resultados com aqueles apresentados no exemplo.

7. Seja $Y = \beta_0 + \beta_1 x + \epsilon$ um modelo de regressão linear simples. Considere a distância

$$d_i = |y_i - (\beta_0 + \beta_1 x_i)|,$$

   entre a reta $y_i = \beta_0 + \beta_1 x_i$ e os pontos (x_i, y_i), $i = 1, \cdots, n$. Encontre as expressões de $\widehat{\beta}_0$ e $\widehat{\beta}_1$ minimizam a soma de quadrados

$$\sum_{i=1}^{n} d_i^2.$$

8. Os dados em `Ruptura.csv` foram analisados por Cordeiro & Paula (1989). Nesse estudo peças de aço inoxidável foram submetidas a esforços e mediu-se o tempo de ruptura em horas. Observaram-se duas variáveis: Y, a variável resposta, contendo o logaritmo dos tempos de ruptura das peças de aço e x, a variável explicativa, contendo o logaritmo do esforço ao qual as peças foram submetidas. Ajuste um modelo de regressão linear simples e interprete os resultados.

9. Mostre que em um modelo de regressão linear simples, o ponto $(\overline{x}, \overline{y})$ está exatamente sobre a linea de regressão estimada.

10. (Montgomery & Runger, 2003) A resistência do papel usado na fabricação de caixas de papelão Y está relacionada à percentagem da concentração de madeira de lei na polpa original x. Sob condições controladas, uma planta piloto fabrica 16 amostras, cada uma sendo proveniente de uma batelada diferente de polpa. Mede-se a resistência à tração. Os dados estão armazenados em `Papel.csv`.

    (a) Ajuste um modelo de regressão linear simples. Verifique se o modelo proposto satisfaz as suposições.

    (c) Construa um intervalo de confiança de 95% para a resistência média para o caso da percentagem da concentração de madeira de lei na polpa original seja $x = 2.5$.

11. **Fatores como variáveis explicativas** Significa que .....

12. **Fatores como variáveis explicativas** Os dados em `Galaxias.csv` são a partir de Dressler (1984). Eles consistem de magnitudes integrados V (V26) e log da dispersão de velocidade central (log s) de uma amostra de 53 galáxias de dois aglomerados de galáxias, os aglomerados de Coma e Virgo. De acordo com a relação Faber-Jackson (Faber e Jackson, 1976), a relação entre estas duas quantidades é aproximadamente linear, tendo a forma

$$log(\sigma) = \beta_0 + \beta_1 V_{26},$$

    onde o parâmetro de um depende da distância ao cluster e b é uma constante. Uma vez que existem dois conjuntos, há duas distâncias e, portanto, dois valores independentes de um. Estes dados são apresentados graficamente na Figura 6. nesse calcular a distância entre os dois grupos é claramente evidente. Dressler identificou quatro valores extremos óbvios, dois de cada cluster, que são observados na figura de seu catálogo números.

13. **Fatores como variáveis explicativas** Os dados em .....

14. Prove que os estimadores dos parâmetros da regressão do modelo $Y = \beta_0 + \beta_1 X + \epsilon$ coincidem com os estimadores dos parâmetros da regressão do modelo $X = \beta_0 + \beta_1 Y + \epsilon$ se, e somente se, $\rho(X, Y) = 1$, o coeficiente de correlação entre X e Y.

15. Uma vez que a variância dos coeficientes de regressão $\widehat{\beta}$ varia inversamente com a variância de X, é muitas vezes sugerido que deveríamos permitir observar a maioria dos valores de X na sua faixa média e utilizar apenas as observações extremas de X no cálculo de $\widehat{\beta}$. É este um procedimento desejável?

16. **Regressão pela origem** Em algumas situações, o modelo de regressão é esperado passar através da origem. Isto é, no modelo de regressão linear simples a média da variável dependente é esperada ser zero quando o valor da variável independente é zero. Assim o modelo de regressão linear é forçado a passar através da origem definindo β_0 igual a zero. O modelo linear simples, então é da forma

$$Y_i = \beta_1 x_i + \epsilon_i, \tag{2.38}$$

    para $i = 1, \cdots, n$. Prove que

$$\widehat{\beta}_1 = \sum_{i=1}^{n} x_i Y_i / \sum_{i=1}^{n} x_i^2.$$

17. Considere o modelo de regressão pela origem em (2.38). Suponha que $x_i \geq 0$ para todo i. Defina $\widetilde{\beta}_1 = \sum_{i=1}^{n} Y_i / \sum_{i=1}^{n} x_i$ e $\widehat{\beta}_1 = \sum_{i=1}^{n} x_i Y_i / \sum_{i=1}^{n} x_i^2$.

    (a) Prove que $\widetilde{\beta}_1$ e $\widehat{\beta}_1$ são estimadores de β_1 tais que $\mathrm{E}(\widetilde{\beta}_1) = \mathrm{E}(\widehat{\beta}_1) = \beta_1$;

    (b) Compare as variâncias de $\widetilde{\beta}_1$ e $\widehat{\beta}_1$.

18. **Regressão pela origem** Dados foram obtidos a partir das emissões de CO_2 e da geração de eletricidade registrados no banco de dados eGrid, hospedado no site da Agência de Proteção Ambiental dos EUA e disponíveis no arquivo `Usinas.csv`.

    Dispomos de três variáveis: nome, o nome da usina de geração de eletricidade na Califórnia em 1997; MMBTU na qual foi registrada A tabela e gráfico exibir a entrada de calor no MMBTU (maneira de escrever milhões de BTUs do setor de energia) ea saída de CO2 (em toneladas) para a maioria das usinas de geração de eletricidade na Califórnia em 1997 . Outros dados do estado estão disponíveis no EGRID97 . Ao comparar a entrada de calor para a produção de CO2 , pode-se obter uma medida da "eficiência de poluição "de cada planta. Ao examinar o gráfico de dispersão , claramente há uma usina que produz uma quantidade excessiva

de CO2 para a sua entrada de calor . É este usina um outlier por causa de informações falsas , ou é realmente um desperdício ?

BTU Uma é a quantidade de calor necessária para elevar a temperatura de uma libra de água um grau Fahrenheit 58,5-59,5 graus sob pressão normal de 30 polegadas de mercúrio na ou próximo do seu ponto de densidade máxima . Um BTU é igual a 252 calorias, 778 libras-pé , 1055 joules ou 0,293 watt - hora .

As Emissões e Geração de Recursos Integrados de banco de dados (eGrid) é uma base de dados abrangente de atributos ambientais dos sistemas de energia elétrica. EGrid é baseado em dados específicos de fábrica disponível para todas as usinas geradoras de eletricidade dos EUA , que fornecem energia e dados para o relatório do governo dos EUA . Os dados reportados inclui a geração de eletricidade (em MWh) , mix de recursos (para as energias renováveis ??e nonrenewables), as emissões (em toneladas de NOx , SO2 e CO2) , as taxas de emissão (em libras por megawatt- hora [lbs / MWh] e libras por milhão de BTUs [lbs / MMBtu] para NOx , SO2 e CO2) , a entrada de calor (em MMBTU) e capacidade (em MW) .

19. no arquivo `Calor.csv`

20. **Regressão pela origem** Exemplo: Johnson, R. (1995). A multiple regression project. Teaching Statistics, 17(2), 64-6.

21. Seja Y uma variável aleatória que satisfaz um modelo de regressão linear. Prove que:

    (a) se $\widehat{\beta}$ é uma solução da equação normal (2.2) então $(\widehat{\beta}-\beta)^\top X^\top(Y-X\widehat{\beta})=0$;

    (b) se X é uma matriz não singular então $(\widehat{\beta}-\beta)^\top X^\top X(\widehat{\beta}-\beta)\geq 0$.

22. Seja Y uma variável aleatória que satisfaz um modelo de regressão linear. Prove que $\mu^\top(\mathrm{I}-H)\mu=0$.

23. Os dados são da n = 29 casas usados para testar a energia térmica solar. As variáveis de interesse para o nosso modelo é y = fluxo total de calor, e x1, x2 e x3, que são os pontos focais para o leste, norte, sul e direções, respectivamente.

    Há duas outras medidas deste conjunto de dados: uma outra medição dos pontos focais e da hora do dia. Não vamos utilizar esses indicadores neste momento. A Tabela 7.1 apresenta os dados utilizados para esta análise.

24. (Montgomery & Runger, 2003) Tratamento térmico é frequentemente usado para carbonizar peças metálicas, tais como engrenagens. A espessura da camada carbonizada é considerada uma característica crucial da engrenagem e contribui para a confiabilidade geral da peça. Por causa da natureza crítica dessa característica, dois testes laboratoriais diferentes são feitos em cada carga da fornalha. Um teste é feito em um pino que acompanha cada carga. O outro teste é destrutivo, em que uma peça real é secionada transversalmente. Esse teste implica correr uma análise de carbono na superfície do passo da engrenagem (topo do dente da engrenagem) e na raiz da engrenagem (entre os dentes da engrenagem). Os dados em `PASSO.csv` são os resultados dos testes da análise de carbono no passo para 32 peças.

    Os regressores são a temperatura da fornalha (TEMP), a concentração de carbono (CONCCARB), a duração do ciclo de carbonização (CICLO), a concentração de carbono (CONCCARBD) e a duração do ciclo de difusão (CICLODIF).

    (a) Ajuste um modelo de regressão linear relacionando os resultados do teste da análise de carbono no passo (PASSO) aos cinco regressores.

    (b) Use o modelo estimado para prever PASSO, quando TEMP=1650, CONCCARB=1.10, CICLO=1.00, CONCCARBD=0.80 e CICLODIF=1.10.

25. **Estimação com restrições** Suponha que desejamos encontrar o estimador dos parâmetros da regressão no modelo $Y=X\beta+\epsilon$, onde $\epsilon\sim N_n(\sigma^2\,\mathrm{I})$ e onde também o vetor β está sujeito a um conjunto de restrições de igualdade, como $T\beta=c$.

    (a) Mostre que o estimador é

    $$\beta_c=\widehat{\beta}+(X^\top X)^{-1}T[T(X^\top X)^{-1}T]^{-1}(c-T\widehat{\beta}),$$

    sendo $\widehat{\beta}=(X^\top X)^{-1}X^\top Y$ o estimador irrestrito do vetor de parâmetros da regressão.

    (b) Discuta situações em que esse modelo pode ser apropriado.

26. Exemplo de modelo de regressão com restrições nos parâmetros. Data anorexia pacote MASS, falta uma descrição melhor dos dados. Ver o exemplo em http://seattlecentral.edu/qelp/sets/073/073.html

27. No mesmo site ver os exemplos Data Set 014, 038 e 047

28. Draper & Smith (1998) Os dados na tabela consistem em treze espécimes de 90/10 ligas de Cu-Ni com diferentes teores de ferro em porcentagem. As amostras foram imersas em água do mar durante 60 dias e a perda de peso devido à corrosão foi registada em unidades de miligramas por decímetro quadrado por dia .... `corrosion.csv`

29. Os seguintes dados referem-se a produção de biomassa de soja segundo a radiação solar interceptada cumulativamente ao longo de um período de oito semanas após a emergência. A produção de biomassa é o peso seco em gramas de amostras independentes de quatro plantas. Os dados foram cortesia dos professores Virginia Lesser e Mike Unsworth da Universidade de Carolina do Norte nos Estados Unidos e publicados em Rawlings *et al.* (1998).

X	29.7	68.4	120.7	217.2	313.5	419.1	535.9	641.5
Y	16.6	49.1	121.7	219.6	375.5	570.8	648.2	755.6

Nesta tabela X representa a radiação solar recebida e Y a produção de biomassa.

(a) Calcule $\widehat{\beta}_0$? e $\widehat{\beta}_1$? para a regressão linear da produção de biomassa vegetal segundo a radiação solar, considerando o modelo com interceptado. Escreva a equação de regressão.

(b) Encontre a expressão dos intervalos de confiança de β_0 e β_1, com 95% de nível de confiança. Interprete esses ntervalos.

(c) Teste H0: ?1 = 1,0 vs Ha: ? ?1 = 1,0 utilizando um teste t com ? = 0,1 . O seu resultado para o teste t de acordo com o intervalo de confiança da parte (b)? Explicar.

(a)

(b)

(c)

(d) utilizar um teste-t para testar H0: ?0 = 0 contra Ha: ? ?0 = 0. Interpretar os resultados. Agora caber uma regressão com ?0 = 0. Submeter a análise de variância para a regressão através da origem e utilizar um teste de F para testar H0: ? ?0 = 0. Compara os resultados do teste t e o teste F . Você adotar o modelo com ou sem a interceptação ?

(e) Calcule s2 (? ?1) para a equação de regressão sem interceptação. Compare as variâncias das estimativas das encostas ? ?1 para os dois modelos . Qual o modelo oferece a maior precisão para a estimativa do declive ?

(f) Calcule o intervalo de confiança de estimativas de produção de biomassa média por 95% X = 30 e X = 600 , tanto para a interceptação e os modelos não - interceptar . Explicar as diferenças nos intervalos obtidos para os dois modelos.

Exercícios da Seção 2.2

1. Prove que se $Z \sim N(0,1)$ e $Y = Z^2$ então $Y \sim \chi^2_{(1)}$.

2. Seja $Y = \beta_0 + \beta_1 X + \epsilon$, onde $\epsilon \sim N_n(0, \sigma^2 \mathrm{I})$ um modelo de regress ao linear simples. Prove que

$$(X^\top X)^{-1} = \frac{1}{\sum_{i=1}^n (x_i - \overline{x})^2} \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n x_i^2 & -\overline{x} \\ -\overline{x} & 1 \end{pmatrix}.$$

3. Seja Y um vetor aleatório satisfazendo o modelo de regressão linear $Y = X\beta + \epsilon$, onde $\epsilon \sim N_n(\sigma^2 \mathrm{I})$. Seja $\mathbf{1} = (1, 1, \cdots, 1)^\top$ um vetor $n \times 1$ de somente números um. Mostre que $(X^\top X)^{-1} X^\top \mathbf{1} = (1, 0, \cdots, 0)^\top$ e também que $\mathbf{1}^\top X (X^\top X)^{-1} X^\top \mathbf{1} = n$.

4. Considere $\underset{\sim}{x} = (1, \overline{x}_2, \cdots, \overline{x}_p)$ onde $\overline{x}_k = \sum_{i=1}^n x_{ik}/n$, $k = 2, \cdots, p$. Este pode ser um ponto de interesse para o cálculo do intervalo de confiança para a média da resposta e observe que este ponto pode ser escrito como $X^\top \mathbf{1}/n$. Utilizando o resultado do exercício anterior prove que $\mathrm{Var}(\widehat{\mu})$ calculada no ponto $\underset{\sim}{x}$ é σ^2/n.

5. Suponha que interessa-nos ajustar um modelo de regressão simples $Y = \beta_0 + \beta_1 X + \epsilon$ em que a intercepção, β_0, seja conhecida.

   (a) Encontre os estimadores de mínimos quadrados e de máxima verossimilhança de $\widehat{\beta}_1$,

   (b) Qual a variância do estimador da inclinação?

   (c) Encontre uma expressão para um intervalo de confiança de $100(1-\alpha)\%$ para o parâmetro de inclinação β_1. Esse intervalo é maior do que o intervalo correspondente à situação em que tanto a interseção quanto a inclinação sejam desconhecidos? Justifique sua resposta.

Exercícios da Seção 2.3

1. Demonstre o Teorema 2.24.

2. Prove que a soma de quadrados da regressão no modelo de regressão linear simples pode ser escrita como

$$\text{SQReg} = \widehat{\beta}_1^2 \sum_{i=1}^{n} (x_i - \overline{x})^2.$$

3. Prove a relação (2.28), isto é, prove que

$$\sum_{i=1}^{n} (y_i - \overline{y})^2 = \sum_{i=1}^{n} (\widehat{\mu}_i - \overline{y})^2 + \sum_{i=1}^{n} (y_i - \widehat{\mu}_i)^2.$$

4. Faça a Análise de Variancia da regressão nos exercícios 8 e 12 sa Seção 2.1. Quais as conclusões em cada caso?

5. Sejam $Y_1, Y_1, \cdots, Y_n$ variáveis aleatórias independentes e igualmente distribuídas.

   (a) Prove que a seguinte relação é verdadeira

$$\sum_{i=1}^{n} Y_i^2 = Q_1 + Q_2,$$

   onde $Q_1 = \sum_{i=1}^{n-1} (Y_i - \overline{Y})^2$ e $Q_2 = (Y_n - \overline{Y})^2 + n\overline{Y}^2$.

   (b) O Teorema de Fisher-Cochran é aplicável nesta situação? Justifique?

6. Seja Y um vetor aleatório com distribuição $N_n(0, \sigma^2 \, \mathrm{I})$, A_1 e A_2 duas matrizes simétricas e idempotentes. Prove que se $A_1 A_2 = 0$ então as formas quadráticas $Y^\top A_1 Y$ e $Y^\top A_2 Y$ são independentes.

7. Seja Y um vetor aleatório satisfazendo o modelo de regressão linear $Y = X\beta + \epsilon$, onde $\epsilon \sim N_n(\sigma^2 \, \mathrm{I})$. Mostre que $\text{posto}(\mathrm{I} - \frac{1}{n}\mathbf{1}) = n - 1$, $\text{posto}(\mathrm{I} - H) = n - p$ e que $\text{posto}(H - \frac{1}{n}\mathbf{1})$.

8. Suponha que a resposta Y esteja relacionada matematicamente com a variável a segundo o modelo de regressão quadrático

$$Y_i = \beta_0 + \beta_1 a_i + \beta_2 a_i^2 + \epsilon_i,$$

   onde $\beta_0, \beta_1, \beta_2$ são os parâmetros da regressão desconhecidos, $a_1, a_2, \cdots, a_n$ são valores de a conhecidos e $\epsilon_1, \cdots, \epsilon_n$ são variáveis aleatórias não observáveis independentes e identicamente distribuídas normal de média zero e variância σ^2. Assuma que os vetores de coeficientes $(a_1^k, a_2^k, \cdots, a_n^k)$, $k = 0, 1, 2, \cdots$ são linearmente independentes.

   (a) Escreva as equações normais para estimar os coeficientes β.

   (b) Obtenha o teste da razão de verossimilhanças para testar $H_0 : \beta_2 = 0$.