Modelos de regressão para dados correlacionados

Cibele Russo

cibele@icmc.usp.br

ICMC USP

Mini-curso oferecido no
Workshop on Probabilistic and Statistical Methods

28 a 30 de janeiro de 2013

Conteúdo da aula

- Modelos de regressão:
  - Modelo de regressão linear simples
  - Modelo de regressão linear múltipla
  - A distribuição normal multivariada
  - Modelo de regressão linear multivariada

Modelos de regressão:

Modelo de regressão linear simples

Modelo de regressão linear múltipla

(sem considerar a correlação entre as observações)

Modelos de regressão

Modelos de regressão são ferramentas estatísticas que buscam explicar a relação entre duas ou mais variáveis.

> **Essentially, all models are wrong, but some are useful.**
>
> George Box

(Box, G. e Draper, N. R. 1987, Empirical Model-Building and Response Surfaces, Wiley Series in Probability and Statistics.)

Modelos de regressão: modelo teórico

- Y: variável resposta (v. aleatória)
- $x_1, \ldots x_p$: p variáveis preditoras (v. não aleatórias)
- ϵ: erro aleatório

Modelo matemático:

$$Y = f(x_1, \ldots, x_p) + \epsilon$$

Suposição mais comum:

$\epsilon \sim (0, \sigma^2)$, $\sigma^2 > 0$ um parâmetro desconhecido.

($E(\epsilon) = 0$ e $\text{Var}(\epsilon_i) = \sigma^2$).

Modelos de regressão: modelo amostral

- Y_i: observação da variável resposta Y na i-ésima unidade experimental (v. aleatória),
- $x_{i1}, \ldots x_{ip}$: valores de $x_1, \ldots, x_p$ variáveis preditoras (v. não aleatórias)
- ϵ_i: erro aleatório e $i = 1, \ldots, n$.

Modelo matemático (amostral):

$$Y_i = f(x_{i1}, \ldots, x_{ip}) + \epsilon_i, \text{ para } i = 1, \ldots, n.$$

Suposições mais comuns:

- ϵ_i é independente de ϵ_j para $i \neq j$ e $i, j = 1, \ldots, n$ e
- $\epsilon_i \sim (0, \sigma^2)$, $\sigma^2 > 0$ um parâmetro desconhecido.

$(E(\epsilon_i) = 0$, $\text{Var}(\epsilon_i) = \sigma^2$, $\text{Cov}(\epsilon_i, \epsilon_j) = 0$ para todo $i, j = 1, \ldots, n$ e $i \neq j.)$

Modelo de regressão linear simples

- $Y_1, \ldots, Y_n$: n observações da variável resposta Y(v. aleatória)
- $x_1, \ldots x_n$: n observações da variável preditora x (v. não aleatória)
- ϵ: erro aleatório

Modelo matemático:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \; i = 1, \ldots, n$$

Suposições $\epsilon_i \overset{i.i.d}{\sim} (0, \sigma^2)$, $i = 1, \ldots, n$.

- β_0 é um parâmetro, chamado intercepto ou coeficiente linear da reta. É o valor esperado de Y quando $x = 0$.
- β_1 é um parâmetro, chamado coeficiente angular da reta. É o aumento (diminuição) médio (a) em Y quando aumentamos uma unidade em x.

Ao assumir um modelo de regressão linear simples para Y em x, assumimos que a relação entre essas duas variáveis é linear e que os erros $\epsilon_1, \ldots, \epsilon_n$ são independentes, o que não acontece sempre.

Suponha inicialmente que x e Y estão relacionadas de forma linear e que $\epsilon_1, \ldots, \epsilon_n$ são independentes, ou seja, $Y_1, \ldots, Y_n$ são independentes.

Temos interesse em estimar os parâmetros β_0, β_1 e σ^2.
Para isso, usamos estimadores, que são funções das observações.

Vamos denotar os valores observados (observações) de $Y_1, \ldots, Y_n$ por $y_1, \ldots, y_n$, respectivamente.

Modelo de regressão linear simples

Queremos encontrar a melhor reta para explicar a relação (suposta linear) entre x e Y. Assim, teremos o

Modelo ajustado:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i, \; i = 1, \ldots, n$$

- $\widehat{\beta}_0$ é uma estimativa de β_0
- $\widehat{\beta}_1$ é uma estimativa de β_1
- $\widehat{Y}_i$ é o valor ajustado de Y para $x = x_i$.

Resíduo: $e_i = y_i - \widehat{Y}_i$

Modelo de regressão linear simples

No modelo de regressão linear simples, o método de mínimos quadrados busca minimizar $\sum_{i=1}^{n} e_i^2$ e leva aos

Estimadores de mínimos quadrados (EMQ)

$$\widehat{\beta}_0 = \overline{y} - \widehat{\beta}_1 \overline{x} \quad \text{e} \quad \widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

com $\overline{x} = \dfrac{\sum_{i=1}^{n} x_i}{n}$ e $\overline{y} = \dfrac{\sum_{i=1}^{n} y_i}{n}$.

Modelo de regressão linear simples

Se no modelo

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \; i = 1, \ldots, n$$

assumirmos que $\epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, então os estimadores de máxima verossimilhança (EMV) de β_0 e β_1 são dados por

$$\widehat{\beta}_{0_{MV}} = \overline{y} - \widehat{\beta}_{1_{MV}}\overline{x} \;\; \text{e} \;\; \widehat{\beta}_{1_{MV}} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$$

com $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ e $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ (coincidem com os EMQs $\widehat{\beta}_0$ e $\widehat{\beta}_1$).

Modelo de regressão linear simples

No modelo

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \;\; i = 1, \ldots, n, \quad \text{com} \quad \epsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$$

o estimador de máxima verossimilhança de σ^2

$$\widehat{\sigma}^2_{MV} = \frac{\sum_{i=1}^{n}(y_i - \widehat{Y}_i)^2}{n} \text{ é viesado.}$$

Um estimador não viesado para σ^2 seria

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{n}(y_i - \widehat{Y}_i)^2}{n-2}.$$

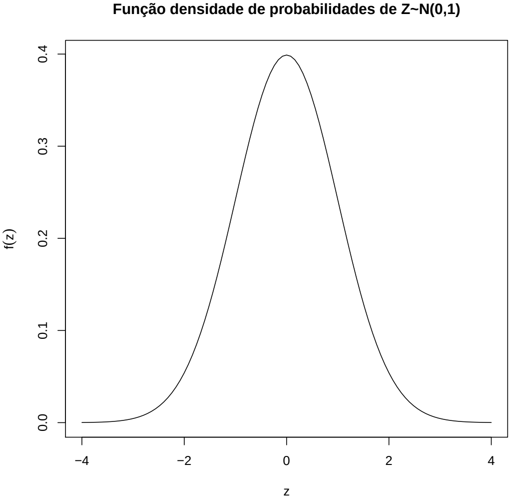
Observação: Distribuição normal

A notação $W \sim N(\mu, \sigma^2)$ indica que W é uma variável aleatória contínua com função densidade de probabilidades (f. d. p.) dada por

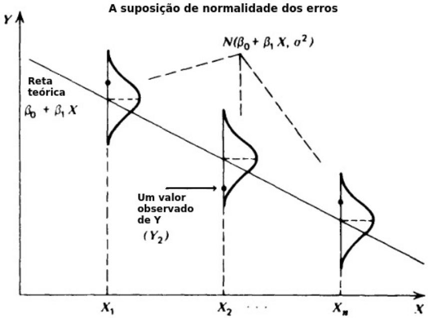
$$f(w) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \text{ com } \mu \in \mathbb{R} \text{ e } \sigma^2 > 0.$$

Observação: Distribuição normal

Função densidade de probabilidades de $Z \sim N(0, 1)$.



A suposição de normalidade para os erros



(Figura adaptada de Draper, N. A e Smith, H. 1998, Applied Regression Analysis, Wiley Series in Probability and Statistics, Wiley.)

Modelo de regressão linear múltipla

Modelo matemático:

$$Y_i = \beta_0 + x_{1i}\beta_1 + x_{2i}\beta_2 + \ldots + x_{ip}\beta_p + \epsilon_i, \;\; i = 1, \ldots, n$$

Matricialmente: $\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$

- $\mathbf{Y}$: vetor de respostas ou variáveis dependentes
- X: matriz do modelo de regressão
- $\boldsymbol{\beta}$: vetor de parâmetros
- $\boldsymbol{\epsilon}$: vetor de erros aleatórios

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & x_{21} & \ldots & x_{p1} \\ 1 & x_{12} & x_{22} & \ldots & x_{p2} \\ \vdots & \vdots & \ddots & \ldots & \vdots \\ 1 & x_{n2} & x_{n2} & \ldots & x_{n2} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

Estimação em um modelo de regressão linear múltipla

Assumindo $\boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n)$, o estimador de máxima verossimilhança (ou de mínimos quadrados) de $\boldsymbol{\beta}$ é dado por

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}$$

em que $\mathbf{y}$ é o vetor observado de $\mathbf{Y}$.

Obs 1: I_n é a matriz identidade $n \times n$

Obs 2: Falaremos mais adiante da distribuição normal multivariada.

Estimação em um modelo de regressão linear múltipla

Modelo ajustado:

$$\widehat{\mathbf{Y}} = X\widehat{\beta}$$

Vetor de resíduos:

$$\mathbf{e}_i = \mathbf{y} - \widehat{\mathbf{Y}}$$

Estimador não viesado de σ^2:

$$\widehat{\sigma^2} = \frac{(\mathbf{y} - \widehat{\mathbf{Y}})'(\mathbf{y} - \widehat{\mathbf{Y}})}{n - p - 1}$$

Propriedades

1. $\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} \sim N_p(\boldsymbol{\beta}, \sigma^2(X'X)^{-1})$

2. $\widehat{\mathbf{Y}} = X\widehat{\boldsymbol{\beta}} = X(X'X)^{-1}X'\mathbf{y} \sim N_n(X\boldsymbol{\beta}, X(X'X)^{-1}X'\sigma^2)$

Propriedades

Predição de novas observações

Dado $\mathbf{X} = \mathbf{x}$, podemos prever o valor de $Y|\mathbf{X} = \mathbf{x}$ fazendo $\widehat{Y|\mathbf{X} = \mathbf{x}} = \mathbf{x}\widehat{\beta}$

Exemplo: Um primeiro modelo

Dados ortodônticos: Suponha que não soubéssemos que os dados são correlacionados e queremos ajustar um modelo de regressão linear simples para cada grupo determinado pelo gênero.

```
> install.packages(c("stats","nlme","Hmisc","lattice"))
> library(nlme)
> attach(Orthodont)
> fit.lm<-lm(distance~ age)
> summary(fit.lm)
> plot(age,distance,pch=16)
> abline(fit.lm)
```

⚠ Certifique-se que os comandos foram passados corretamente para o R.

Exemplo de ajuste: modelo de regressão linear simples

Dados ortodônticos

```
>  summary(fit.lm)

Call:
lm(formula = distance ~ age)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.5037 -1.5778 -0.1833  1.3519  6.3167 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  16.7611     1.2256  13.676  < 2e-16 ***
age           0.6602     0.1092   6.047 2.25e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.537 on 106 degrees of freedom
Multiple R-squared: 0.2565,	Adjusted R-squared: 0.2495 
F-statistic: 36.56 on 1 and 106 DF,  p-value: 2.248e-08

> 
```

Exercício: como interpretar o modelo ajustado?

Exemplo de ajuste: modelo de regressão linear simples

O modelo ajustado para Y (distância) usando como covariável x (idade) é dado por

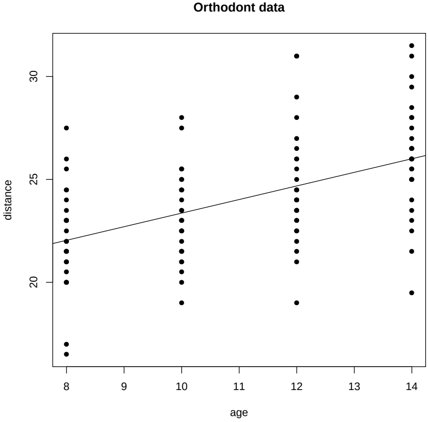
$$\widehat{Y}_i = 16,7611 + 0,6602x_i, \; 1, \ldots, n.$$

A interpretação das estimativas dos parâmetros é

- A estimativa da distância em uma criança com idade 0 é 16,7611 $(\widehat{\beta_0})^{\star}$.
- O aumento estimado na distância é de 0,6602 quando se aumenta 1 ano de idade.

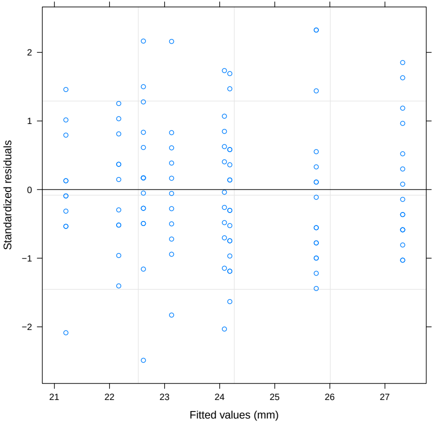
⚠* **Importante:** Deve-se tomar cuidado com a interpretação de $\widehat{\beta_0}$, pois não se deve extrapolar o modelo de regressão para intervalos distantes dos valores da covariável. A relação entre as duas variáveis pode mudar de comportamento longe dos valores de x.

Exemplo de ajuste: modelo de regressão linear simples



Modelo de regressão linear simples ajustado aos dados ortodônticos

Exemplo de ajuste: modelo de regressão linear simples



Resíduos do modelo de regressão linear simples ajustado aos dados ortodônticos

Exercício: um modelo para cada gênero

Dados ortodônticos: Suponha que não soubéssemos que os dados são correlacionados mas quiséssemos ajustar um modelo de regressão linear simples para os dados de cada grupo.

```
> attach(Orthodont)
> fit.Sex<-lmList(distance~ age|Sex,Orthodont)
> coef(fit.Sex)
> coef(fit.Sex)[1,]
> coef(fit.Sex)[2,]
> intervals(fit.Sex)
> plot(intervals(fit.Sex))
> plot(fit.Sex)
```

⚠ Certifique-se que os comandos foram passados corretamente para o R.

O que é correlação?

Correlação ou coeficiente de correlação é uma ferramenta utilizada em Probabilidade e Estatística para medir o grau de relacionamento linear entre duas variáveis aleatórias sem implicar relação de causalidade.

Correlação

Seja W uma variável aleatória contínua (unidimensional) com f. d. p. $f(w)$.

Momentos populacionais: O k-ésimo momento de $g(W)$ é definido como

$$E(g(W)^k) = \int_{-\infty}^{\infty} [g(w)]^k f(w) dw$$

desde que a integral acima exista.

Valor esperado de W**:** $E(W) = \mu^1 = \mu = \int_{-\infty}^{\infty} wf(w)dw.$

Variância de W**:** $Var(W) = E(W - \mu)^2 = \int_{-\infty}^{\infty} (w - \mu)^2 f(w) dw.$

Correlação

Seja W uma variável aleatória contínua com f. d. p. $f_w(w)$ e Z uma variável aleatória contínua com f. d. p. $f_z(z)$, e suponha que a f. d. p. conjunta de (W, Z) seja $f(w, z)$. Define-se

Valor esperado de W: $E(W) = \mu_w$
Valor esperado de W: $E(Z) = \mu_z$

Variância de W: $Var(W) = \sigma^2_W$.
Variância de W: $Var(Z) = \sigma^2_Z$.

Covariância entre W e Z: $Cov(W, Z) = E((W - \mu_w)(Z - \mu_z)) = \sigma_{WZ}$.

Coeficiente de correlação entre W e Z: $\rho_{W,Z} = \dfrac{\sigma_{WZ}}{\sqrt{\sigma^2_W}\sqrt{\sigma^2_Z}}$.

Correlação - caso multidimensional

Seja vetor aleatório

$$\mathbf{W} = \begin{bmatrix} W_1 \\ \vdots \\ W_n \end{bmatrix},$$

ou seja, cada elemento W_i é uma variável aleatória, $i = 1, \ldots, n$ com f. d. p. $f_i(w_i)$.

Correlação - caso multidimensional

Valor esperado de W:

$$E(\mathbf{W}) = \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_n \end{bmatrix}$$

em que $\mu_i = E(W_i) = \int_{-\infty}^{\infty} w_i f_i(w_i) dw_i.$

Correlação - caso multidimensional

Variância de W:

$$Var(\mathbf{W}) = E[(\mathbf{W} - \boldsymbol{\mu})(\mathbf{W} - \boldsymbol{\mu})']$$

$$= E\left(\begin{bmatrix} W_1 - \mu_1 \\ \vdots \\ W_n - \mu_n \end{bmatrix}\begin{bmatrix} W_1 - \mu_1 & \dots & W_n - \mu_n \end{bmatrix}\right) =$$

$$= E\begin{pmatrix} (W_1 - \mu_1)^2 & (W_1 - \mu_1)(W_2 - \mu_2) & \dots & (W_1 - \mu_1)(W_n - \mu_n) \\ (W_2 - \mu_2)(W_1 - \mu_1) & (W_2 - \mu_2)^2 & \dots & (W_2 - \mu_2)(W_n - \mu_n) \\ \vdots & \vdots & \ddots & \vdots \\ (W_n - \mu_n)(W_1 - \mu_1) & (W_n - \mu_n)(W_2 - \mu_2) & \dots & (W_n - \mu_n)^2 \end{pmatrix}$$

Matriz de variâncias e covariâncias

Variância de W:

$$\Sigma = Var(\mathbf{W}) = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1n} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{n1} & \sigma_{n2} & \dots & \sigma_{nn} \end{pmatrix}.$$

Matriz de correlação

Matriz de correlação de W:

$$R = Cor(\mathbf{W}) = \begin{pmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1n} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{n1} & \rho_{n2} & \cdots & \rho_{nn} \end{pmatrix}$$

com $\rho_{ij} = \dfrac{\sigma_{ij}}{\sqrt{\sigma_{ii}}\sqrt{\sigma_{jj}}}$ para $i, j = 1, \ldots, n.$

Propriedades:

- $-1 \leq \rho_{ij} \leq 1,$ para $i, j = 1, \ldots, n.$
- $\rho_{jj} = 1,$ para $j = 1, \ldots, n.$

Matrizes de variâncias e covariâncias e correlações amostrais

Em geral Σ e R são desconhecidas, mas podemos estimá-las obtendo a matriz de variância e covariâncias amostral e a matriz de correlações amostrais.

Matriz de variâncias e covariâncias amostrais (não viesada para Σ)

$$s = var(\mathbf{W}) = \begin{pmatrix} s_{11} & s_{12} & \dots & s_{1n} \\ s_{21} & s_{22} & \dots & s_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ s_{n1} & s_{n2} & \dots & s_{nn} \end{pmatrix}$$

com $s_{ij} = \dfrac{\sum_{k=1}^{n}(w_{ki} - \overline{w_i})(w_{kj} - \overline{w_j})}{n-1}$

Matrizes de variâncias e covariâncias e correlações amostrais

Matriz de correlações amostrais

$$r = cor(\mathbf{W}) = \begin{pmatrix} 1 & r_{12} & \dots & r_{1n} \\ r_{21} & 1 & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \dots & 1 \end{pmatrix}$$

com $r_{ij} = \dfrac{s_{ij}}{\sqrt{s_{ii}}\sqrt{s_{jj}}}$ para $i, j = 1, \dots, n$.

Distribuição normal multivariada

Definição

O vetor p-dimensional $\mathbf{U}$ *tem distribuição normal multivariada se, e somente se, toda combinação linear de* $\mathbf{U}$ *tem distribuição normal univariada.*

$$\mathbf{U} \sim N_p \iff \mathbf{t}'\mathbf{U} \sim N_1 \quad \forall \mathbf{t} \in \mathbb{R}^p.$$

Observação: Distribuição normal multivariada

Propriedades

1. A média $E(U) = \boldsymbol{\mu}$ e $Var(U) = \Sigma$ existem. Notação: $U \sim N_p(\mu, \Sigma)$.
2. Fixado $\mathbf{t} \in R^p$ um vetor de constantes, $\mathbf{t}'U \sim N(\mathbf{t}'U, \mathbf{t}'\Sigma\mathbf{t})$.
3. Se $\mathbf{U} \sim N_p(\boldsymbol{\mu}, \Sigma)$ e $r(\Sigma) = p$ então a função densidade de probabilidades de $\mathbf{U}$ é dada por

$$f(\mathbf{u}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{(\mathbf{u}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{u}-\boldsymbol{\mu})}{2}\right\}, \quad \boldsymbol{\mu} \in R^p$$

Observação: Distribuição normal multivariada

Consequências

1. $U \sim N_p(\boldsymbol{\mu}, \Sigma)$ é tal que $\mathbf{U} = \begin{bmatrix} U_1 \\ U_2 \\ \vdots \\ U_p \end{bmatrix}$.

   Se Σ é diagonal, então $U_1, \ldots, U_p$ são v.a.independentes, cada uma com distribuição normal univariada.

Observação: Distribuição normal multivariada

Consequências

2 Se $\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix}$ com $Var(\mathbf{U}) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}' & \Sigma_{22} \end{bmatrix}$ em que

$\Sigma_{12} = Var(\mathbf{U}_1)$, $\Sigma_{22} = Var(\mathbf{U}_2)$, $\Sigma_{12} = Cov(\mathbf{U}_1, \mathbf{U}_2)$.

Então $\mathbf{U}_1$ é independente de $\mathbf{U}_2$ se, e somente se, $\Sigma_{12} = 0$.

Observação: Distribuição normal multivariada

Propriedades

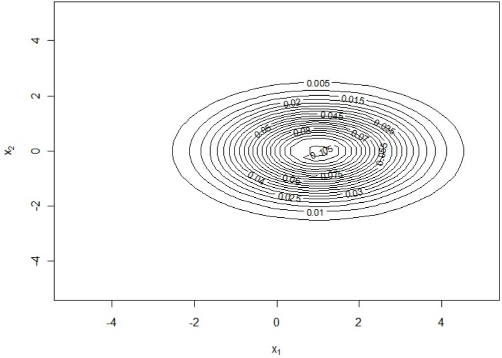
1. $\mathbf{U} \sim N_p(\boldsymbol{\mu}, \Sigma)$ é particionada como $\mathbf{U} = \begin{bmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \end{bmatrix}$,

   com a partição adequada para o vetor de médias e a matriz de variâncias e covariâncias

   $$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \text{ e } \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma'_{12} & \Sigma_{22} \end{bmatrix}.$$

   Então $\mathbf{U}_1 \sim N(\boldsymbol{\mu}_1, \Sigma_{11})$ e $\mathbf{U}_2 \sim N(\boldsymbol{\mu}_2, \Sigma_{22})$ têm distribuição normal multivariada com as dimensões dos vetores $\mathbf{U}_1$ e $\mathbf{U}_2$.

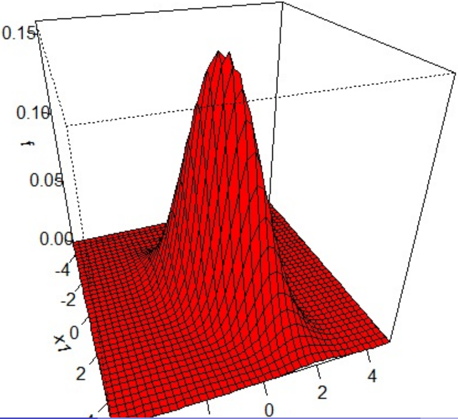
Contornos elípticos da normal bivariada (elementos não correlacionados)



Distribuição normal (elementos correlacionados)

Definindo os parâmetros da distribuição Normal ($\boldsymbol{\mu}$, Σ),

com $\boldsymbol{\mu} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ e $\Sigma = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix}$:



Modelar dados correlacionados

Qual a forma mais simples de modelar dados correlacionados?

Modelo de regressão multivariada

Um modelo de regressão linear multivariado é da forma

$$\mathbf{Y}_i = X_i\boldsymbol{\beta} + \boldsymbol{\epsilon}_i$$

onde

- $\mathbf{Y}_i = \begin{bmatrix} Y_{1i} \\ \vdots \\ Y_{ni} \end{bmatrix}$ é o vetor de respostas do i-ésimo indivíduo
- X_i é uma matriz de planejamento
- $\boldsymbol{\beta}$ é um vetor de parâmetros
- $\boldsymbol{\epsilon}_i = \begin{bmatrix} \epsilon_{1i} \\ \vdots \\ \epsilon_{ni} \end{bmatrix}$

  Suposição comum: $\boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} N(\mathbf{0}, \Sigma)$

Modelo de regressão multivariada

Uma possibilidade é reescrever o modelo de forma multivariada como

$$\mathbf{Y} = X\boldsymbol{\beta} + \epsilon$$

com matrizes adequadas $\mathbf{Y}$, X, $\boldsymbol{\beta}$ e ϵ.

Pode-se obter $\widehat{\boldsymbol{\beta}}_i = (X_i'X_i)^{-1}X_i'\mathbf{y}_i$, em que $\mathbf{y}_i$ é o valor observado de $\mathbf{Y}_i$ para $i = 1, \ldots, n$ e

$\widehat{\boldsymbol{\beta}} = [\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2, \ldots, \widehat{\boldsymbol{\beta}}_n]$, o que é equivalente a fazer

$$\widehat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y}.$$

Modelo de regressão multivariada

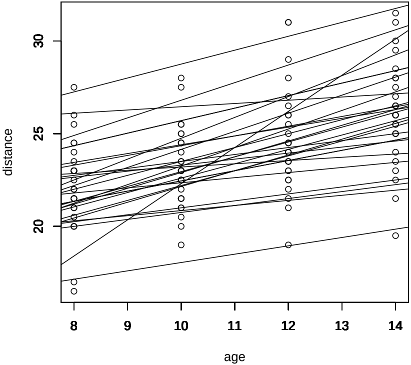
Comandos em R:

```
> library(nlme)
> attach(Orthodont)
> X<-cbind(1,matrix(age,ncol=1)[1:4])
> Y<-matrix(distance,nrow=4,byrow=F)
> Betachapeu<-solve(t(X)%*%X)%*%t(X)%*%Y
> Betachapeu
```

⚠ Certifique-se que os comandos foram passados corretamente para o R, especialmente o $*$.

Modelo multivariado

Exemplo: Dados ortodônticos



Modelo com efeitos mistos

Neste curso, vamos considerar os modelos com efeitos mistos, que permitem fazer previsões específicas para cada unidade experimental, e ao mesmo tempo identificar padrões similares entre as observações.

Próxima aula

Modelos lineares com efeitos mistos