2a aula prática de modelos lineares

Leo Bastos (PROCC)

Dados simulados

Nessa sequencia de comandos vamos simular um conjunto de dados, para verificarmos a performance da estimação e avaliarmos os resíduos quando as suposições do modelo são verificadas.

Vamos gerar algumas covariáveis, uma discreta uma contínua e uma categórica. E vamos gerar um desfecho com média definida por uma combinação linear dessas covariáveis.

```r
library(tidyverse)
# Aqui definimos nosso tamanho de amostra
n <- 250

# Aqui vamos gerar as variáveis categóricas:

# Uma variavel binária
dados <- tibble(X1 = rbinom(n, 1, 0.5))

# Uma variável contínua
dados$X2 <- rnorm(n)

# E eu uma variável categórica (letras A,B,C,D)
dados$X3 <- sample(x = LETTERS[1:4],
                   size = n, replace = T)
dados$X3 <- factor(dados$X3, levels = LETTERS[1:4])
```

Agora vamos gerar gerar um valor médio para nosso desfecho, que vai ser dado por

$$\mathbb{E}[Y_i] = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2i} + \beta_3^{(B)} X_{3i}^{(B)} + \beta_3^{(C)} X_{3i}^{(C)} + \beta_3^{(D)} X_{3i}^{(D)}$$

Onde fixaremos valores para α, para os βs, e incluiremos uma variabilidade σ^2 permitindo uma certa aleatoriedade em torno da média.

```r
# Gerando a média

# Gerando a matrix de dados
XX <- model.matrix( ~ X1 + X2 + X3, data = dados)

# Se quiser visualizar a matriz de dados
# View(XX)

# Vetor de coeficientes
parametros <- c(50, 15, 5, -10, 0, 40)

# Gerando a media usando produto matricial
mediaY <- XX %*% parametros
```

```r
# Desvio padrao do desfecho
sigma <- 5

# Gerando Y
dados$Y <- rnorm(n = n, mean = mediaY, sd = sigma)

#   Vamos adicionar também ao banco uma
# nova variável que nao tem nada a ver
# com o desfecho
dados$Z <- rnorm(n,30,10)
```

Exercício 1

Faça as estatísticas descritivas das variaveis do banco dados, avalie os cruzamentos entre as variáveis explicativas $(X1, X2, X3, Z)$ com o desfecho (Y).

Exercício 2

Vimos que o desfecho foi gerado direto da distribuição, avalie a normalidade do desfecho visualmente e usando algum teste formal. Qual a conclusão? Há alguma contradição no resultado?

O modelo linear

Vamos agora para o modelo linear. Queremos recuperar os parâmetros do modelo linear usados para gerar o desfecho.

Vamos supor que a variável X_1 é a principal exposição de interesse, e queremos avaliar o efeito de X_1 em Y. Seja o modelo $M1$ o modelo em questão, que é dado por:

$$Y_i \sim N(\mu_i, \sigma^2), \qquad i = 1, 2, \ldots, n,$$

onde $\mu_i = \alpha + \beta_1 X_1$.

Vamos chamar esse modelo de modelo $M1$, que é dado por

```r
M1 <- lm( formula = Y ~ X1, data = dados)

# A saída resumo
summary(M1)

# Os coeficientes e seu intervalo de confianca
cbind(Est = coef(M1), confint(M1))
```

O valor do coeficiente de $X1$ significa que a exposição $X1 = 1$ aumenta (ou diminui a depender do sinal) o valor esperado de Y em β_1 unidades.

Lembrando que os testes de hipóteses e os intervalos de confiança para os coeficiente só tem validade se as suposições do modelo linear $M1$ não são violadas.

Podemos avaliar os resíduos ordinários e/ou os resíduos padronizados. Vamos nesse exemplo avaliar os resíduos padronizados.

```r
resid.M1 <- rstandard(M1)

# Residuos versus ordem dos dados
plot(resid.M1)
```

```r
# Normalidade
hist(resid.M1)
```

```r
qqnorm(resid.M1)
qqline(resid.M1)
```

```r
shapiro.test(resid.M1)
```

O que podemos dizer a respeito das suposições do modelo?

Podemos avaliar os resíduos versus variáveis não inclusas no modelo na busca de algum padrão.

```r
# Residuos versus outras variaveis
# X2
plot(dados$X2, resid.M1)
```

```r
# X3
plot(dados$X3, resid.M1)
```

```r
# Z
plot(dados$Z, resid.M1)
```

O que podemos dizer?

Exercício 3

Ajuste os modelos $M2$ incluindo a variável $X2$, $M3$ incluindo as variáveis $X2$ e $X3$, e $M4$ incluindo $X2$, $X3$, e Z. Ajuste os modelos e avalie os resíduos.

Exercício 4

Preencha a tabela abaixo, e escolha o melhor modelo segundo o AIC.

Model	AIC
$M1$	
$M2$	
$M3$	
$M4$	

Exercício 5

O modelo $M3$ é o modelo que, de fato, representa o processo que gerou o desfecho. Os coeficientes estimados estão próximos dos parâmetros verdadeiros ($\alpha = 50, \beta_1 = 15, \beta_2 = 5, \beta_3^{(B)} = -10, \beta_3^{(C)} = 0, \beta_3^{(D)} = 40$)? Construa o intervalo de confiança de 95% para os coeficientes.

Avaliando o efeito do tamanho da amostra

Para avaliar o efeito do tamanho da amostra na estimação do coeficiente de $X1$ (nossa exposição de interesse) no modelo $M3$, geramos bancos com diferentes tamanhos amostrais, e para cada banco calcularemos a estimativa do efeito de $X1$ e seu IC de 95%. Depois disso vamos incluir em um gráfico com o tamanho de amostra no eixo X e o coeficiente com seu intervalo no eixo Y.

```r
# Vamos criar uma funcao que vai gerar um bando de tamanho n
geraDados <- function(n){
  parametros <- c(50, 15, 5, -10, 0, 40)
  dados <- tibble(X1 = rbinom(n, 1, 0.5))
```

```r
  dados$X2 <- rnorm(n)
  dados$X3 <- sample(x = LETTERS[1:4],
                     size = n, replace = T)
  dados$X3 <- factor(dados$X3,
                       levels = LETTERS[1:4])
  XX <- model.matrix( ~ X1 + X2 + X3, data = dados)
  mediaY <- XX %*% parametros
  sigma <- 5
  dados$Y <- rnorm(n = n, mean = mediaY, sd = sigma)
  return(dados)
}

# teste
geraDados(10)

# Gerando uma funcao que retorna o coeficiente e o IC da variavel X1
estimaCoefX1 <- function(dados){
  out <- lm(formula = Y ~ X1 + X2 + X3,
            data = dados)
  aux <- cbind(Est = coef(out), confint(out))
  # A segunda linha eh o coeficiente de X1
  aux[2,]
}

estimaCoefX1(geraDados(100))

# Criando uma sequencia de valores de n a serem testados.
n.seq <- c(seq(30,1000, by = 10))

dados.plot <- c(n = n.seq[1], estimaCoefX1(geraDados(n.seq[1])))


# Fazendo um loop
for(k in 2:length(n.seq)){
  dados.plot <- bind_rows(dados.plot, c(n = n.seq[k], estimaCoefX1(geraDados(n.seq[k]))))
}


ggplot(dados.plot,
       aes(x=n, y=Est, ymin=`2.5 %`, ymax=`97.5 %`)) +
  geom_point() + geom_linerange() +
  theme_bw(base_size = 18) +ylab("Coeficiente de X1")
```

Exercício 6

O que podemos dizer da Figura gera com os comandos acima?