Regressão linear múltipla: *the madness continues*

Motivação

O modelo de regressão é extremamente útil como ferramenta explanatória e preditiva, mas até agora nos limitamos à situação ao em que temos apenas uma variavel independente. Na vida real é comum estudarmos fenômenos com múltiplas causas. Em aplicações reais, em geral dispomos de muitas covariaveis que podem, em princípio, estar relacionadas à variável dependente (desfecho/resposta).

Nesta lição vamos fazer o nosso modelo de regressão ficar mais realista incluindo múltiplas covariáveis (ou variáveis explanatórias) de uma vez. Vamos entender como estimar quantidades-chaves do modelo e também como diagnosticar o seu ajuste.

O modelo

Vamos escrever

$$y = X\beta + \varepsilon,$$

com

$$\varepsilon \sim \text{Normal}(0, \sigma^2 I),$$

onde V é uma matriz positiva semi-definida **conhecida**.

Estimação

É possível mostrar que os estimadores de máxima verossimilhança das quantidades desconhecidas são

$$
\begin{aligned}
\hat{\beta} &= \left(X^T X\right)^{-1} X^T y, && (1)\\
\hat{\sigma^2} &= \frac{1}{n-2}\left(y - X^T\beta\right)^T \left(y - X^T\beta\right). && (2)
\end{aligned}
$$

Nos exercícios abaixo, você vai mostrar que estes estimadores têm boas propriedades, como não-viesamento. Em particular,

$$
\hat{\beta} \sim \text{Normal}(\beta, \sigma^2 \left(X^T X\right)^{-1}) \tag{3}
$$

Além disso, a informação de Fisher vale

$$
I(\beta) = \frac{X^T X}{\sigma^2}. \tag{4}
$$

Diagnóstico

Parte integral de qualquer análise é diagnosticar o ajuste do modelo aos dados. Em uma análise de regressão, um diagnóstico importante é a *análise de resíduos*. Defina

$$
\hat{e}_i = y_i - \hat{\theta}_i.
$$

Como o resíduo da i-ésima observação em relação ao seu valor ajustado. Podemos então escrever

$$
E[\hat{e}\hat{e}^T] = \sigma^2 \left[I - X(X^T X)^{-1} X^T\right],
$$

para a matriz de covariância dos resíduos. Um passo importante é *padronizar* os resíduos:

$$
r_i = \frac{\hat{e}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}}.
$$

Uma boa medida para a *influência* de uma observação é a distância de Cook:

$$
D_i := \frac{1}{p}\left(\frac{h_{ii}}{1 - h_{ii}}\right) r_i^2. \tag{5}
$$

Os dados

Os dados que vamos analisar aqui são as notas (*scores*) em um teste padronizado de $n =$ crianças, para os quais foram medidos o quociente de inteligência (QI) da mãe e também se a mãe completou o ensino médio (*high school*, `hs`). Os dados estão aqui e são discutidos no capítulo 10 de Gelman, Hill & Vehtari (2020).

Nosso objetivo é entender como a habilidade inata da criança (predita presumivelmente pelo QI da mãe) e sua condição socioeconômica (sinalizado pelo `hs` da mãe) influenciam no desempenho.

Vamos olhar as principais estatísticas descritivas

```
head(kidiq)
```

```
# A tibble: 6 x 5
  kid_score mom_hs mom_iq mom_work mom_age
      <dbl>  <dbl>  <dbl>    <dbl>   <dbl>
1        65      1  121.         4      27
2        98      1   89.4        4      25
3        85      1  115.         4      27
4        83      1   99.4        3      25
5       115      1   92.7        4      27
6        98      0  108.         1      18
```

```
summary(kidiq)
```

```
   kid_score         mom_hs           mom_iq          mom_work
 Min.   : 20.0   Min.   :0.0000   Min.   : 71.04   Min.   :1.000
 1st Qu.: 74.0   1st Qu.:1.0000   1st Qu.: 88.66   1st Qu.:2.000
 Median : 90.0   Median :1.0000   Median : 97.92   Median :3.000
 Mean   : 86.8   Mean   :0.7857   Mean   :100.00   Mean   :2.896
 3rd Qu.:102.0   3rd Qu.:1.0000   3rd Qu.:110.27   3rd Qu.:4.000
 Max.   :144.0   Max.   :1.0000   Max.   :138.89   Max.   :4.000
    mom_age
 Min.   :17.00
 1st Qu.:21.00
 Median :23.00
 Mean   :22.79
 3rd Qu.:25.00
 Max.   :29.00
```

As perguntas

De posse desses dados, podemos nos fazer várias perguntas sobre *associações* nos dados. Por exemplo, queremos saber se o QI da mãe tem alguma associação (leia-se: capacidade preditiva) com as notas (*scores*) da criança. Além disso, essa associação é mediada pelo nível educacional da mãe? Como o fato de que a mãe trabalha fora impacta a variável resposta (na presença das outras covariáveis relevantes)?

Análise

Exercícios de fixação

Tome X uma matriz real $n \times P$ e $Y = \{Y_1, \dots, Y_n\}^T \in \mathbb{R}^n$ um vetor contendo os valores da variável dependente.

Nosso modelo (um pouco menos geral que o dado acima) é

$$E[Y_i] =: \mu_i(\beta) = \tilde{X}_i^T \beta,$$

onde $\beta \in \mathbb{R}^{P+1}$ é o vetor de coeficientes e parâmetro de interesse e $\tilde{X}$ é uma matriz obtida adicionando uma coluna de uns, $X_0 = \{1, \dots, 1\}^T$, a X. Para completar a especificação do modelo, vamos assumir que os erros em torno do preditor linear são normalmente distribuídos com variância comum:

$$\begin{aligned} Y_i &= \mu_i(\beta) + \epsilon_i \\ \epsilon_i &\overset{\text{i.i.d.}}{\sim} \text{Normal}(0, \sigma^2), \end{aligned}$$

com $\sigma^2 \in \mathbb{R}_+$ desconhecida.

1. Escreva a log-verossimilhança e deduza seu gradiente e a sua derivada segunda (hessiana);
2. Com base nos cálculos do item anterior, mostre a forma do estimador de máxima verossimilhança para β, $\hat{\beta}$;
3. Mostre que $\hat{\beta}$ é não-viesado;
4. Considere um outro estimador <u>não-viesado</u> de β: $\tilde{\beta} = My$, onde

   $$M = (X^T X)^{-1} X^T + D,$$

   e D é uma matriz $P \times n$ cujas entradas são não-zero. Mostre que $R := \text{Var}(\tilde{\beta}) - \text{Var}(\hat{\beta})$ é positiva-definida.

   Dica: Compute $E[\tilde{\beta}]$ e considere o que deve valer para D sob a premissa de que $\tilde{\beta}$ é não-viesado.

Comentário: Ao resolver o último item, você terá mostrado que o estimador de máxima verossimilhança (e também o estimador de mínimos quadrados) é o melhor estimador linear não-viesado (*best linear unbiased linear estimator, BLUE*). Em particular esta é a versão de Gauss[1] do famoso teorema de Gauss-Markov.

[1] Carl Friedrich Gauss (1777-1855), matemático alemão conhecido como o Príncipe dos Matemáticos.

5. Considere o seguinte modelo alternativo:

$$\varepsilon \sim \text{Normal}(0, \sigma^2 V),$$

Deduza $\widehat{\beta}_{\text{EMV}}$ e sua distribuição amostral, além de $I(\beta)$. Discuta como este modelo viola as premissas de Gauss-Markov e quais os efeitos desta viola

Referências

- Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press. (Cap 6)
- Gelman, A., Hill, J., & Vehtari, A. (2020). Regression and other stories. Cambridge University Press.