Regressão linear: o melhor modelo ruim que você já viu

Motivação

Nesta lição vamos estudar um dos cavalos de batalha da estatística: o modelo linear. Em particular, vamos discutir um modelo que relaxa a premissa de identidade de distribuição ao propor uma estrutura linear para a média condicional. Suponha que temos o seguinte conjunto de dados:

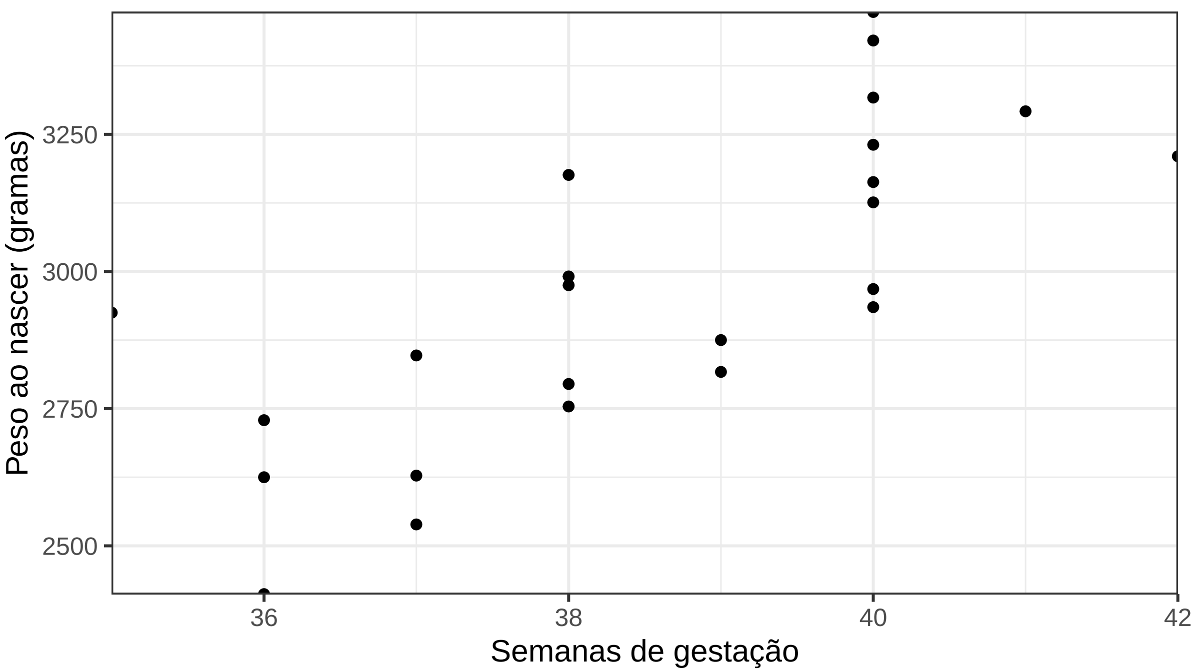

Figure 1: Peso ao nascer *vs* semanas de gestação para $n = 24$ bebês.

que apresenta o peso ao nascer (em gramas) da i-ésima criança, Y_i, contra a sua idade gestacional no parto (em semanas), X_i. Queremos estimar a probabilidade (condicional) de uma criança nascer com baixo peso, $\omega(x_i) = \Pr(Y_i \leq 2500 \mid X_i = x_i)$. Este será o nosso alvo inferencial pelas próximas lições.

Vamos assumir que o **desfecho** do i-ésimo indivíduo, Y_i, $i = 1, 2, \ldots, n$, depende do **preditor** ou **covariável** X_i através da seguinte estrutura:

$$E\left[Y_i \mid X_i\right] = \beta_0 + \beta_1 X_i =: \theta_i,$$

onde $\beta = (\beta_0, \beta_1) \in \mathbb{R}^2$ são os **coeficientes**. Em particular, chamos β_0 de **intercepto** e β_1 de **coeficiente angular**.

Se definimos

$$L(\theta, Z) := \sum_{i=1}^{n} \left(Y_i - \theta_i\right)^2, \tag{1}$$

$$= \sum_{i=1}^{n} \left(Y_i - [\beta_0 + \beta_1 X_i]\right)^2, \tag{2}$$

podemos sem muita dificuldade mostrar que a função de perda é convexa em β.

Pensando de forma estatística, vamos escrever

$$Y_i = \theta_i + \varepsilon_i, \tag{3}$$

e assumir que

1. $E[\epsilon_i] = 0$ para todo $i = 1, 2, \ldots, n$;
2. $\operatorname{Cov}(\varepsilon) = \sigma^2 I$;
3. $Y_i \perp\!\!\!\perp Y_j \mid X_i, X_j$ para todo $i \neq j$.

But are you really BLUE?

Vamos pôr nossas premissas à prova e estudar o que acontece com os estimadores dos coeficientes sob diversas distribuições para os erros.

Vamos simular dados sob três modelos[1] para os erros:

$$\mathcal{M}_1 : \varepsilon_i \sim \text{Uniforme}(-1, 1),$$
$$\mathcal{M}_2 : \varepsilon_i \sim \text{Normal}(0, (1/3)^2),$$
$$\mathcal{M}_3 : \varepsilon_i \sim \text{Gama}(1/3, 1),$$

Para começar o experimento, vamos preparar as coisas:

[1]: Esta formulação está **errada**, de propósito. Você consegue encontrar o erro? **Dica**: o que assumimos sobre $E[\varepsilon_i]$?

```r
library(ggplot2)
beta0 <- -2
beta1 <- 1.3
Nobs <- 50
X <- rnorm(Nobs) ## gerando a covariada
```

e olhar o resultado de **uma** simulação, apresentado na Figure 2.

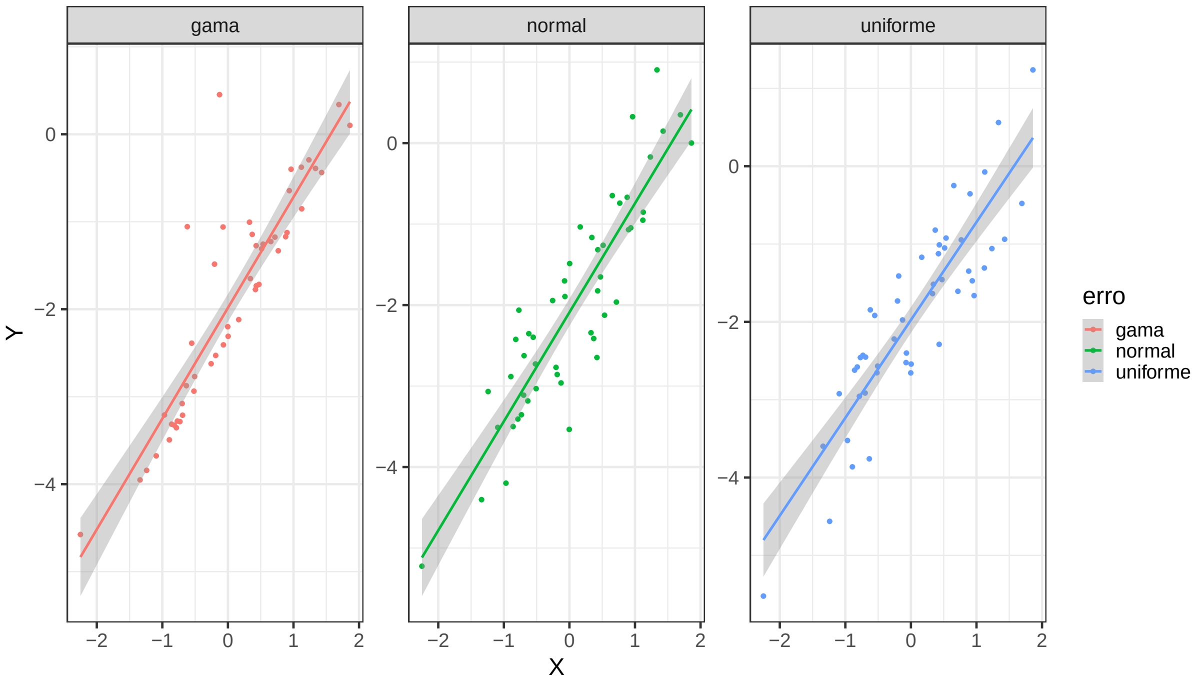

Figure 2: Regressão sob diversos modelos para erros.

Vamos agora conduzir um experimento de Monte Carlo estudar a distribuição amostral dos estimadores dos coeficientes sob cada um dos *DGPs*. Vamos rodar $M = 1000$ experimentos e apresentar os resultados na Figure 3.

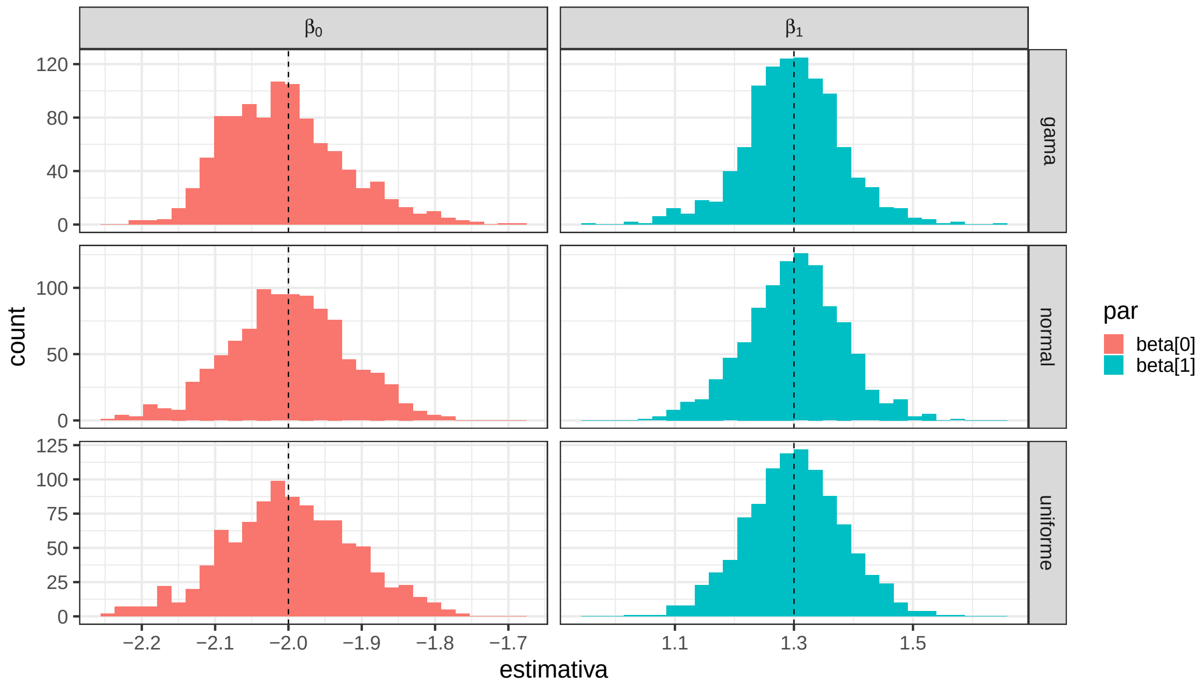

Figure 3: Distribuição amostral dos coeficientes sob diversos modelos para erros. A linha pontilhada vertical marca os valores 'verdadeiros'.

Vamos agora calcular a média de Monte Carlo dos coeficientes:

```
aggregate(estimativa~erro+par, ests, mean)
```

```
      erro     par estimativa
1     gama beta[0]  -2.005267
2   normal beta[0]  -2.000200
3 uniforme beta[0]  -1.998081
4     gama beta[1]   1.298613
5   normal beta[1]   1.300943
6 uniforme beta[1]   1.300660
```

Exercícios de fixação

Em várias aplicações científicas, temos interesse em estudar se dois grupos diferem quanto à sua composição em algum aspecto, por exemplo, a média do processo gerador dos dados. O modelo de regressão permite especificar uma relação entre duas variáveis, em particular uma estrutura condicional para a média. Podemos utilizar modelos de regressão para testar hipóteses sobre diferenças de grupos na presença de uma variável que influencia a média. Neste exercício vamos nos basear no exemplo apresentado na seção 2.2.2 de Dobson(2001) e modelar o peso ao nascer de bebês (em gramas) em relação à idade gestacional (em semanas) para bebês do sexo masculino e do sexo feminino. A pergunta principal aqui é se o sexo da criança influencia no seu peso ao nascer uma vez que ajustamos para o tempo de gestação.

Seja Y_{jk} o peso ao nascer de uma criança do sexo j e seja x_{jk} sua idade gestacional. Considere o modelo

$$E\left[Y_{jk}\right] = \alpha_j + \beta_j x_{jk} = \mu_{jk},$$

para o k-ésimo bebê no grupo j. Note que este modelo presume que as linhas de base dos sexos são diferentes e que os coeficientes angulares também são.

Aqui vamos explorar hipóteses sobre os coeficientes angulares, isto é, sobre o desenvolvimento do bebê ao longo das semanas de gestação.

1. Que outras premissas são necessárias para abordar essa questão sob o ponto de vista do modelo de regressão linear?
2. Considere a hipótese
$$H_0 : \beta_1 = \beta_2 = \beta,$$
para $\beta \in \mathbb{R}$. Elabore dois modelos, um mais geral e outro menos geral, para avaliar H_0 frente aos dados.

   Dica: Considere o que acontece com o coeficiente angular quando H_0 é verdadeira e quando ela é falsa.
3. Escreva a função de densidade de probabilidade de Y_{jk} sob cada modelo.

   Dica: Considere o logaritmo da f.d.p.
4. Mostre como obter estimativas de máxima verossimilhança sob os dois modelos considerados no item 2;
5. Liste as estatísticas suficientes necessárias para proceder à estimação no item anterior;
6. Descreva em detalhes a elaboração de um teste estatístico para testar H_0 contra H_1. Não se esqueça de descrever o procedimento para cálculo da estatística de teste, bem como deduzir a sua distribuição de probabilidade sob H_0 – veja também o próximo item.

   Dica: Considere o que esperamos a respeito da soma de erros quadráticos dos dois modelos do item 2.
7. Sobre o item anterior, mostre como se livrar do parâmetro de estorvo σ^2 no cálculo da estatística de teste.
8. Faça o teste desenvolvido utilizando os dados na Figure 1, disponíveis aqui e discuta se o sexo do bebê parece influenciar o peso ao nascer.

Referências

- Dobson, A. J., & Barnett, A. G. (2018). An introduction to generalized linear models. CRC press.