

Regressão linear múltipla bayesiana: *girl put your bayesian hat on*

Motivação

Como vimos até aqui, o modelo linear (gaussiano) é extremamente útil para modelar a relação entre possíveis variáveis explanatórias (i.e. covariáveis) e uma variável dependente/resposta contínua. Até agora vimos como fazer inferência para esse modelo sob a ótica clássica/frequentista. Vamos então nos debruçar sobre o tratamento bayesiano do problema.

O modelo

Vamos trabalhar com o mesmo modelo de antes:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \sim \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Vamos suplementar a estrutura condicional dos dados com uma estrutura probabilística para as quantidades desconhecidas do modelo, isto é, uma distribuição *a priori* $\pi_{B,S}(\boldsymbol{\beta}, \sigma^2)$.

Uma análise bayesiana não conjugada

Vamos mostrar agora uma análise do conjunto de dados ‘kid score’ usando o pacote **rstanarm**, que não utiliza prioris conjugadas (ver exercícios abaixo para a análise conjugada). Vamos preparar as coisas

```
library(ggplot2)
library(bayesplot)
theme_set(theme_bw())
library(rstanarm)
data(kidiq)
```

e agora ajustar o modelo que desenvolvemos na lição anterior (i.e. com interação) usando mínimos quadrados/máxima verossimilhança.

```
fmod <- glm(kid_score ~ mom_hs * mom_iq, data = kidiq)
summary(fmod)
```

Call:

```
glm(formula = kid_score ~ mom_hs * mom_iq, data = kidiq)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-11.4820	13.7580	-0.835	0.404422
mom_hs	51.2682	15.3376	3.343	0.000902 ***
mom_iq	0.9689	0.1483	6.531	1.84e-10 ***
mom_hs:mom_iq	-0.4843	0.1622	-2.985	0.002994 **

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 322.9736)

Null deviance: 180386 on 433 degrees of freedom
Residual deviance: 138879 on 430 degrees of freedom
AIC: 3745.1

Number of Fisher Scoring iterations: 2

Em seguida, vamos ajustar o seguinte modelo

$$\begin{aligned} \mathbf{y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim \text{Normal}(\mathbf{0}, \sigma^2 \mathbf{I}), \\ \boldsymbol{\beta} &\sim \text{Normal}(\mathbf{0}, \text{diag}(25/4)), \\ \sigma &\sim \text{Exponencial}(1). \end{aligned}$$

que corresponde às priors *default* do **rstanarm**. Note que a priori é sobre σ e não σ^2 – uma boa ideia é derivar a densidade sobre a variância dos erros. A ideia por trás desta especificação *a priori* é ter priors *fracamente informativas* (*weakly informative*); em particular, a ideia é dizer que os coeficientes são independentes *a priori* e têm desvio padrão de 2.5, permitindo a estimação de efeitos razoavelmente grandes. Além disso, a priori sobre o erro de observação (σ^2) encoraja fortemente pequenos erros – tem muita massa perto de zero.

Vamos agora ajustar o modelo usando a função `stan_glm` que utiliza um algoritmo de cadeias de Markov Monte Carlo (MCMC) chamado Hamiltonian Monte Carlo para obter amostras aproximadas da *posteriori* $p_{\mathbf{X}}(\boldsymbol{\beta}, \sigma^2 \mid \mathbf{y}) \propto f_{\mathbf{X}}(\mathbf{y} \mid \boldsymbol{\beta}, \sigma^2) \pi_{B,S}(\boldsymbol{\beta}, \sigma^2)$:

```
bmod <- stan_glm(kid_score ~ mom_hs * mom_iq,
                 data = kidiq, refresh = 0)
summary(bmod)
```

Model Info:

```
function:      stan_glm
family:        gaussian [identity]
formula:       kid_score ~ mom_hs * mom_iq
algorithm:     sampling
sample:        4000 (posterior sample size)
priors:        see help('prior_summary')
observations:  434
predictors:    4
```

Estimates:

	mean	sd	10%	50%	90%
(Intercept)	-10.5	13.7	-28.2	-10.5	7.2
mom_hs	49.9	15.3	29.9	50.0	69.5
mom_iq	1.0	0.1	0.8	1.0	1.1
mom_hs:mom_iq	-0.5	0.2	-0.7	-0.5	-0.3
sigma	18.0	0.6	17.2	18.0	18.8

Fit Diagnostics:

	mean	sd	10%	50%	90%
mean_PPD	86.8	1.2	85.2	86.8	88.4

The `mean_ppd` is the sample average posterior predictive distribution of the outcome variable

MCMC diagnostics

	mcse	Rhat	n_eff
(Intercept)	0.4	1.0	1308
mom_hs	0.4	1.0	1274
mom_iq	0.0	1.0	1287
mom_hs:mom_iq	0.0	1.0	1235
sigma	0.0	1.0	2501
mean_PPD	0.0	1.0	2913
log-posterior	0.0	1.0	1364

For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective

Depois de checar os diagnósticos do MCMC ($R_{\text{hat}} < 1.01$ e $\text{ESS} > 500$ para todos os parâmetros), vemos que as estimativas dos coeficientes não são radicalmente diferentes daquelas obtidas com o método clássico/frequentista. Isso não chega a ser surpresa porque temos uma quantidade razoável de observações ($n = 434$) em relação ao número de parâmetros (quantos são?).

Interrogando o modelo usando previsões *a posteriori*

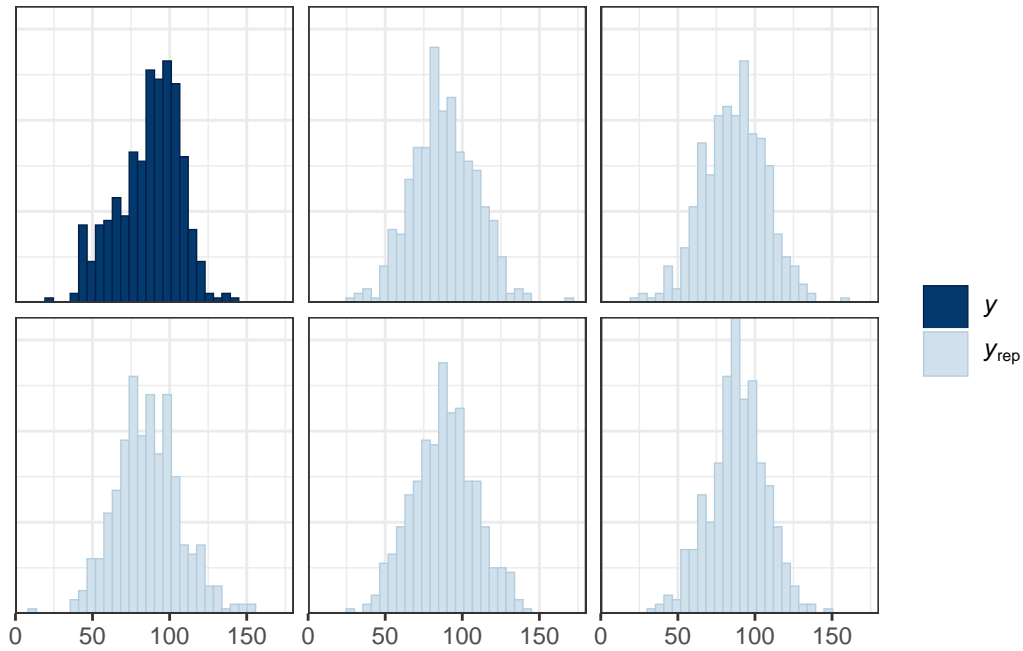
Agora que temos uma distribuição *a posteriori* para as quantidades desconhecidas do modelo, vamos utilizá-la para explorar o modelo e analisar o seu ajuste aos dados. Vamos computar a distribuição preditiva *a posteriori* de certas quantidades e comparar essas distribuições com os valores observados nos dados. Chamamos esse procedimento genérico de checagem preditiva *a posteriori* (em inglês, *posterior predictive checks* [PPC]). Vamos começar com densidade da variável dependente,

$$\tilde{p}_{\tilde{\mathbf{X}}}(\tilde{\mathbf{y}} \mid \mathbf{y}) = \int_{\Omega} f_{\tilde{\mathbf{X}}}(\tilde{\mathbf{y}} \mid \theta) p_{\tilde{\mathbf{X}}}(\theta \mid \mathbf{y}) d\theta,$$

para $\theta = (\beta, \sigma^2)$ e uma (potencialmente nova) matriz de desenho $\tilde{\mathbf{X}}$. Um bom exercício é escrever a integral acima como uma marginalização sobre a distribuição conjunta de $\tilde{\mathbf{y}}$ e as outras quantidades desconhecidas do modelo e entender que essa manipulação segue diretamente as regras do cálculo de probabilidades. Vamos olhar $\tilde{\mathbf{y}} \mid \theta$ para cinco amostras da posteriori

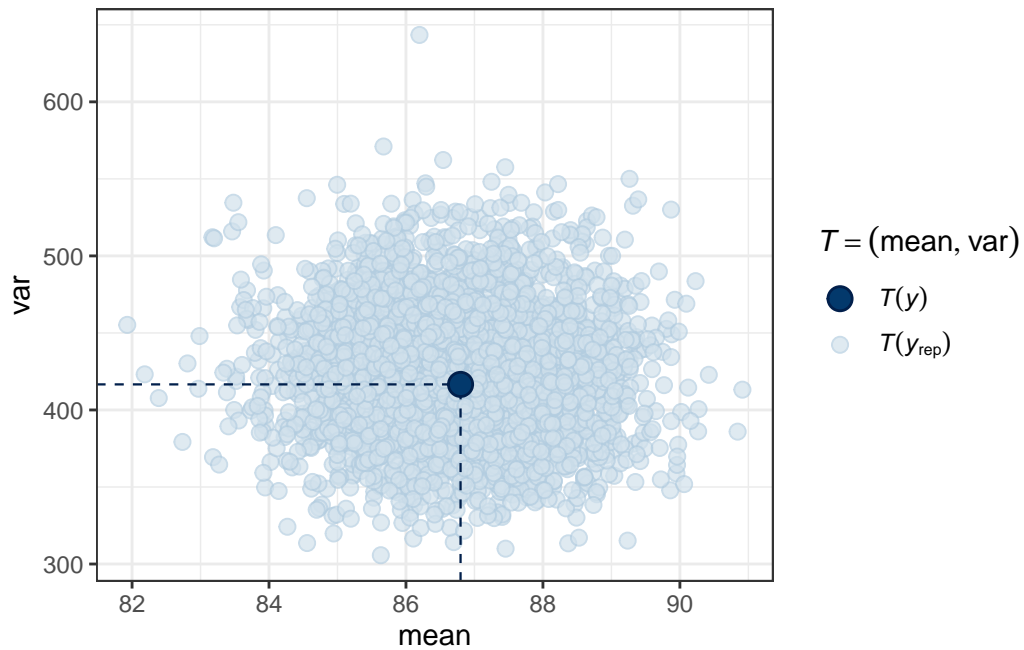
```
pp_check(bmod, plotfun = "hist", nreps = 5)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Uma aproximação de $\tilde{p}(\tilde{\mathbf{y}} \mid \mathbf{y})$ pode ser obtida tomando uma média sobre muitas amostras. Agora vamos olhar a distribuição conjunta de $\mu_{\tilde{\mathbf{X}}, \mathbf{y}} := E[\tilde{\mathbf{y}} \mid \mathbf{y}, \boldsymbol{\theta}]$ e $v_{\tilde{\mathbf{X}}, \mathbf{y}} := \text{Var}(\tilde{\mathbf{y}} \mid \mathbf{y}, \boldsymbol{\theta})$:

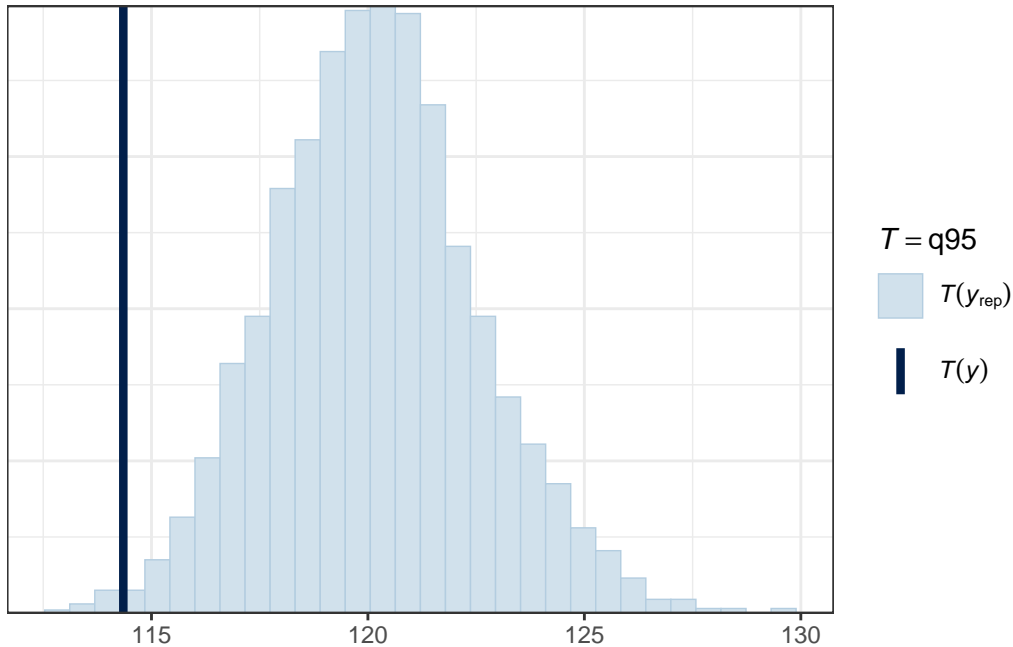
```
pp_check(bmod, plotfun = "stat_2d", stat = c("mean", "var"))
```



Onde vemos que o modelo parece produzir dados que têm os dois primeiros momentos bem parecidos com os observados. Por último, vamos estudar a capacidade do modelo de modelar a cauda da distribuição de y :

```
q95 <- function(x) quantile(x, .95)
pp_check(bmod, plotfun = "stat", stat = "q95")
```

``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.



Vemos que no que toca à modelagem do quantil 95%, nosso modelo não faz um ótimo trabalho. E tudo bem. Em um modelo de regressão linear, estamos interessados em modelar bem a média condicional e a variância dos dados. Porque será que a cauda da distribuição preditiva parece ser mais pesada que a cauda dos dados observados? Será que você consegue responder usando os resultados da análise conjugada?

Para terminar, vamos produzir previsões da variável dependente para vários valores do QI da mãe (`mom_iq`) para os dois grupos (`mom_hs = 0` e `mom_hs = 1`):

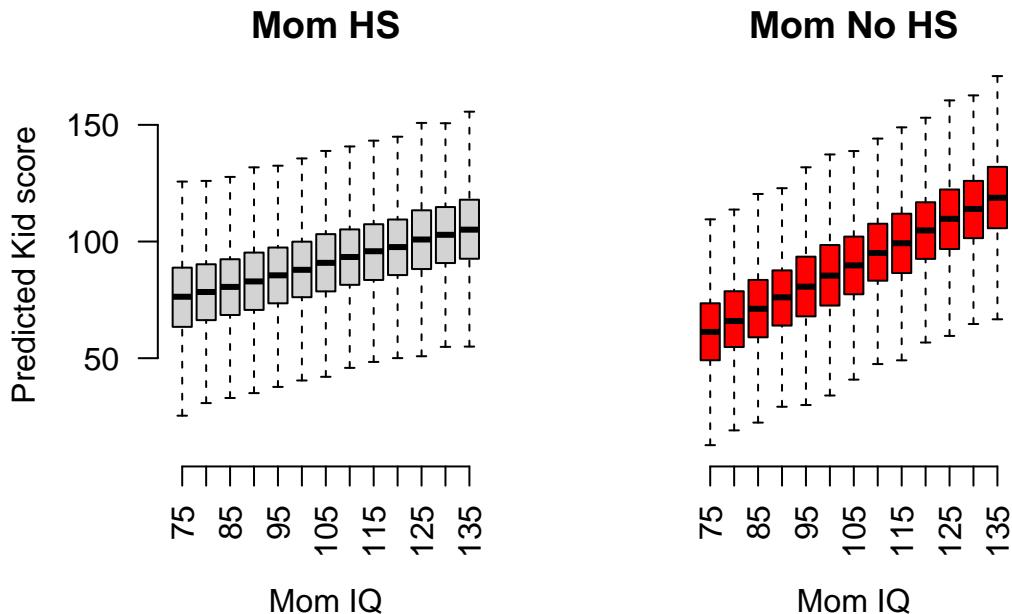
```
IQ_SEQ <- seq(from = 75, to = 135, by = 5)
y_nohs <- posterior_predict(bmod, newdata = data.frame(mom_hs = 0, mom_iq = IQ_SEQ))
y_hs <- posterior_predict(bmod, newdata = data.frame(mom_hs = 1, mom_iq = IQ_SEQ))

par(mfrow = c(1:2), mar = c(5,4,2,1))
```

```

boxplot(y_hs, axes = FALSE, outline = FALSE, ylim = c(10,170),
        xlab = "Mom IQ", ylab = "Predicted Kid score", main = "Mom HS")
axis(1, at = 1:ncol(y_hs), labels = IQ_SEQ, las = 3)
axis(2, las = 1)
boxplot(y_nohs, outline = FALSE, col = "red", axes = FALSE, ylim = c(10,170),
        xlab = "Mom IQ", ylab = NULL, main = "Mom No HS")
axis(1, at = 1:ncol(y_hs), labels = IQ_SEQ, las = 3)

```



Isto é, aqui nós construímos duas $\tilde{\mathbf{X}}$ diferentes e amostramos (aproximadamente) de $\tilde{p}_{\tilde{\mathbf{X}}}$ para produzir as nossas previsões. Note que essas previsões já levam em conta a incerteza sobre os parâmetros (veja exercício 5 abaixo).

Exercícios de fixação

Considere o modelo discutido acima. Uma escolha interessante para auxiliar no entendimento e na análise é

$$\pi_{B,S}(\boldsymbol{\beta}, \sigma^2) = \pi_{B|S}(\boldsymbol{\beta} \mid \sigma^2) \pi_S(\sigma^2),$$

isto é, uma estrutura *a priori* que modela os coeficientes de forma condicional à variância dos erros e uma distribuição marginal na variância. As consequências matemáticas dessas escolhas são bem discutidas [aqui](#) e [aqui](#) – mas você deve tentar deduzir os resultados de forma independente primeiro.

1. Mostre que a verossimilhança $f_{\tilde{\mathbf{X}}}(\mathbf{y} \mid \theta)$ pode ser escrita na forma

$$f_{\tilde{\mathbf{X}}}(\mathbf{y} \mid \theta) = g_{\tilde{\mathbf{X}}}(\mathbf{y} \mid \beta, \sigma^2) h_{\tilde{\mathbf{X}}}(\mathbf{y} \mid \sigma^2).$$

2. Utilize o resultado anterior para deduzir que a priori conjugada para este caso é da forma

$$\pi_{B,S}(\beta, \sigma^2) = \pi_{B|S}(\beta \mid \sigma^2) \pi_S(\sigma^2).$$

Em particular, mostre que

$$\pi_{B,S}(\beta, \sigma^2) \propto \left(\frac{1}{\sigma^2}\right)^{a+(P+1)/2+1} \times \exp\left(-\frac{1}{\sigma^2} \left\{b + \frac{1}{2}(\beta - \mu_\beta)^T \mathbf{V}_\beta^{-1}(\beta - \mu_\beta)\right\}\right),$$

onde $\mu_\beta \in \mathbb{R}^{P+1}$, \mathbf{V}_β é uma matriz positiva definida e $a, b \in \mathbb{R}_+$.

Dica: Que escolhas para $\pi_{B|S}$ e π_S eu preciso fazer?

3. A priori anterior chama-se **normal inversa gama** (NIG) e tem quatro parâmetros: \mathbf{m} , \mathbf{V} , a e b . Mostre que a posteriori de θ também é NIG e exiba seus hiperparâmetros.
4. **Distribuições marginais:** um objeto muito importante em qualquer análise bayesiana é a distribuição marginal de cada parâmetro, porque ela permite inferências mais interpretáveis ao mesmo tempo que acomoda a incerteza sobre as outras quantidades desconhecidas do modelo. Vamos agora calcular algumas marginais importantes.

Dica: Antes de começar os cálculos para essa seção, vale considerar a seguinte representação do nosso modelo:

$$\begin{aligned} \mathbf{y} &= \tilde{\mathbf{X}}\beta + \epsilon_1, \text{ com } \epsilon_1 \sim \text{MVN}_n(\mathbf{0}_n, \Sigma_1), \\ \beta &= \mu_\beta + \epsilon_2, \text{ com } \epsilon_2 \sim \text{MVN}_{P+1}(\mathbf{0}_{P+1}, \Sigma_2), \end{aligned}$$

onde ϵ_1 e ϵ_2 são erros independentes.

4.1 Determine Σ_1 e Σ_2 ;

4.2 Compute a verossimilhança marginal com respeito a σ^2 :

$$\tilde{f}_{\tilde{\mathbf{X}}}(\mathbf{y} \mid \sigma^2) := \int_{\mathbb{R}^{P+1}} f_{\tilde{\mathbf{X}}}(\mathbf{y} \mid \mathbf{b}, \sigma^2) \pi_{B|S}(\mathbf{b} \mid \sigma^2) d\mathbf{b}.$$

4.3 Usando o item anterior, compute a verossimilhança marginal ou *preditiva a priori*:

$$m_{\tilde{\mathbf{X}}}(\mathbf{y}) := \int_0^\infty \tilde{f}_{\tilde{\mathbf{X}}}(\mathbf{y} \mid s) \pi_S(s) ds.$$

4.4 Mostre $\bar{f}_{\tilde{\mathbf{X}}}(\beta \mid \mathbf{y})$ e comente sobre como calcular, por exemplo, $\Pr(\beta_1 > a \mid \mathbf{y})$, para $a \in \mathbb{R}$.

5. Suponha que eu coletei uma nova matriz de desenho $m \times P$, \mathbf{X}' e quero prever o valor de \mathbf{y}' a partir do que eu aprendi usando \mathbf{X} e \mathbf{y} . Compute

$$\bar{p}_{\tilde{\mathbf{X}}, \mathbf{X}'}(\mathbf{y}' \mid \mathbf{y}) := \int_{\mathbb{R}^{P+1} \times \mathbb{R}_+} p_{\tilde{\mathbf{X}}}(\mathbf{b}, s \mid \mathbf{y}) f_{\mathbf{X}'}(\mathbf{y}' \mid \mathbf{b}, s) d\mathbf{b} ds,$$

e esboce o seu gráfico para uma observação (linha de \mathbf{X}') de um conjunto de dados da sua escolha. Compare essas predições com a execução da mesma tarefa sob o ponto de vista frequentista/clássico.

Dica: use um conjunto de dados que você conheça bem. Bons exemplos são os bancos de ‘peso ao nascer’ e ‘kid score’, que já analisamos em sala.

Resultados úteis

Aqui estão enunciados alguns resultados úteis para o desenvolvimento das questões acima. Estes resultados são dados sem demonstração, que você está convidada a fazer.

- **Completando o “quadrado” em múltiplas dimensões:** tome \mathbf{A} matriz simétrica positiva definida $d \times d$ e $\boldsymbol{\alpha}, \mathbf{u} \in \mathbb{R}^d$. Vale que:

$$\mathbf{u}^T \mathbf{A} \mathbf{u} - 2\boldsymbol{\alpha}^T \mathbf{u} = (\mathbf{u} - \mathbf{A}^{-1}\boldsymbol{\alpha})^T \mathbf{A} (\mathbf{u} - \mathbf{A}^{-1}\boldsymbol{\alpha}) - \boldsymbol{\alpha}^T \mathbf{A}^{-1} \boldsymbol{\alpha}. \quad (1)$$

Dica: Expanda o produto e procure por cancelamentos de termos da forma $\mathbf{a}^T \mathbf{M}^{-1} \mathbf{a}$.

- **Sherman-Woodbury-Morrisson:** tome \mathbf{A} matriz quadrada $d \times d$ inversível, \mathbf{B} matriz $k \times d$, \mathbf{C} matriz $d \times k$ e \mathbf{D} matriz quadrada $k \times k$ inversível. Então

$$(\mathbf{A} + \mathbf{B} \mathbf{D} \mathbf{C})^{-1} = \mathbf{A}^{-1} - (\mathbf{D}^{-1} + \mathbf{C} \mathbf{A}^{-1} \mathbf{B})^{-1} \mathbf{C} \mathbf{A}^{-1}.$$

- **Determinantes:** tome \mathbf{A} , \mathbf{B} , \mathbf{C} e \mathbf{D} como antes. Então,

$$\det(\mathbf{A} + \mathbf{B} \mathbf{D} \mathbf{C}) = \det(\mathbf{A}) \det(\mathbf{D}) \det(\mathbf{D}^{-1} + \mathbf{C} \mathbf{A}^{-1} \mathbf{B}).$$

Referências

- Banerjee, S. Bayesian Linear Model: Gory Details. Pubh7440 Notes.
- Gelman, A., Hill, J., & Vehtari, A. (2020). [Regression and other stories](#). Cambridge University Press.