Tarefa 2 - LASSO

Tais Bellini - 205650

Janeiro, 2020

Introdução

Neste relatório, serão realizados experimnentos para avaliar a *Regressão Lasso* e sua capacidade de escolher corretamente as variáveis adequadas para o modelo em diferentes cenários. Para isto, definiremos um modelo que será o verdadeiro e do qual os resultados da *Regressão Lasso* deverão se aproximar. Os cenários incluem: avaliar amostras com distribuição Normal e Exponencial, variações nos parâmetros de *Cross Validation* para definição do λ, e algumas variações de tamanho de amostra e modelo definido.

LASSO para amostra com distribuição Normal

Para este experimento, executamos 1000 vezes os seguinte passos:

(i) gerar uma amostra aleatória de tamanho $n = 50$ de $X_1, ..., X_{100}$, onde $X_i \sim N(0, \sqrt(i))$

(ii) gerar o modelo $Y = 20 + 5X_1 + 5X_{10} + 5X_{20} + 5X_{50} + 5X_{90} + \varepsilon$ onde $\varepsilon \sim N(0, 1)$

(iii) ajustar um modelo LASSO utilizando *k-fold Cross Validation* para $k = 5$

(iv) idem ao item (iii) para $k = 10$

Resultados

Número de acertos

A tabela abaixo descreve quantas vezes o modelo LASSO acertou exatamente quais eram as variáveis corretas do modelo original e quantas vezes o modelo acertou as variáveis, mas incluiu outras que não eram corretas:

Table 1: Acertos exatos e parciais para cada tamanho de k

	k=5	k=10
Exato	542	542
Inclui	458	458

Podemos observar que em todas as repetições o modelo LASSO acertou quais eram a variáveis corretas e que houve um equilíbrio entre as vezes em que o acerto foi exato ou apenas incluiu as variáveis certas. Na maioria das vezes não foi um acerto exato e o número de *folds* não promoveu diferença no resultado.

Coeficientes ajustados

Para avaliar se os valores escolhido para $\beta 0$ (Intercept), quando o LASSO acertou exatamente o modelo correto foram próximos do verdadeiro, calculamos a média entre eles e depois fizemos um arredondamento utilizando a função

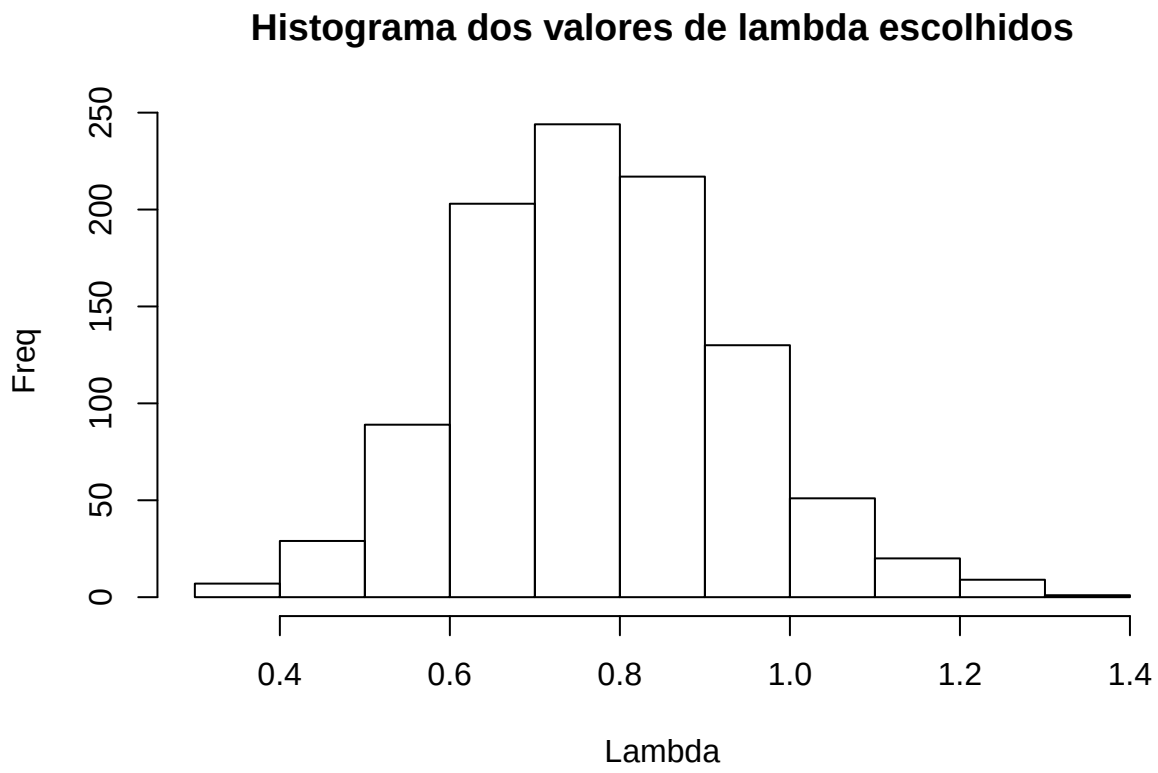
.

Observamos que o valor foi 20 para ambos os tamanhos de k. Ou seja, o modelo LASSO estimou $\beta 0$ próximo do valor verdadeiro.

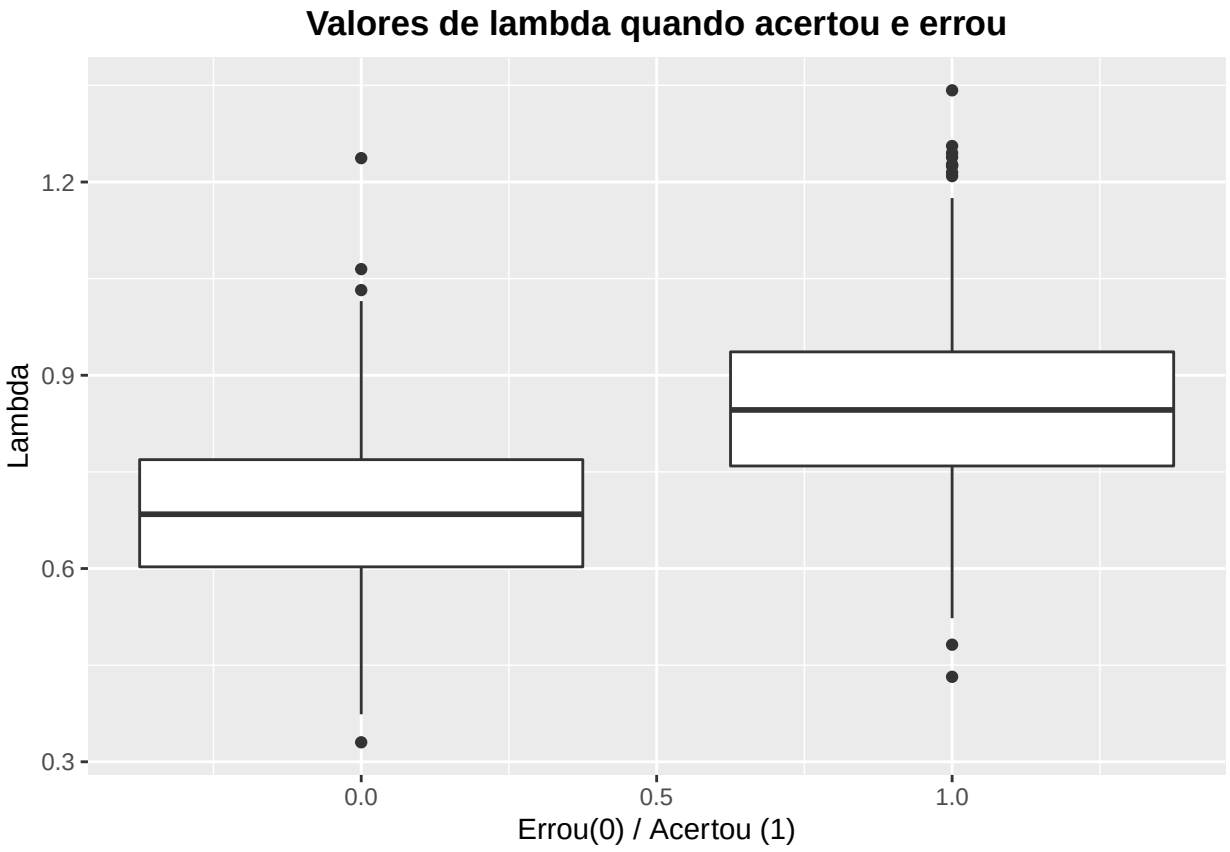
Similarmente, para os coeficientes, calculamos a média de todos os coeficientes escolhidos para cada vez em que o modelo acertou exatamente, depois a média entre todos eles e, por fim, um arredondamento. O resultado tanto para $k = 5$ e $k = 10$ foi 5, indicando que a escolha dos coeficientes também foi próxima dos verdadeiros valores.

Escolha do lambda otimo

Observamos que os valores de λ escolhidos para $k = 5$ e $k = 10$ foram os mesmos e a média foi 0.7769204. A distribuição ficou aproximada da normal.



Podemos ver no gráfico abaixo que o modelo acertou mais para valores de λ maiores:



LASSO para amostra com distribuição Exponencial

Para este experimento, executamos 1000 vezes os seguinte passos:

(i) gerar uma amostra aleatória de tamanho 50 de $X_1, ..., X_{100}$, onde $X_i \sim Exp(\sqrt(i))$

(ii) gerar o modelo $Y = 20 + 5X_1 + 5X_{10} + 5X_{20} + 5X_{90} + \varepsilon$ onde $\varepsilon \sim Exp(1)$

(iii) ajustar um modelo LASSO utilizando *k-fold Cross Validation* para $k = 5$

(iv) idem ao item (iii) para $k = 10$

Além disso, executamos os mesmos passos acima, mas gerando amostras de tamanho 200 e apenas para $k = 10$.

Resultados

Número de acertos

Podemos observar na tabela abaixo que o modelo LASSO já não obteve tantos acertos quando a amostra possui distribuição exponencial:

Table 2: Acertos exatos e parciais para cada tamanho de k

	k=5	k=10
Exato	1	1
Inclui	833	833

Podemos ver que em apenas uma das repetições o modelo LASSO acertou exatamente as variáveis corretas e que na maioria das vezes acertou parcialmente, incluindo todas as variáveis certas mas outras incorretas. Neste caso, houveram repetições em que o modelo errou completamente, não incluindo todas as variáveis verdadeiras. O número de *folds* também não promoveu diferença no resultado.

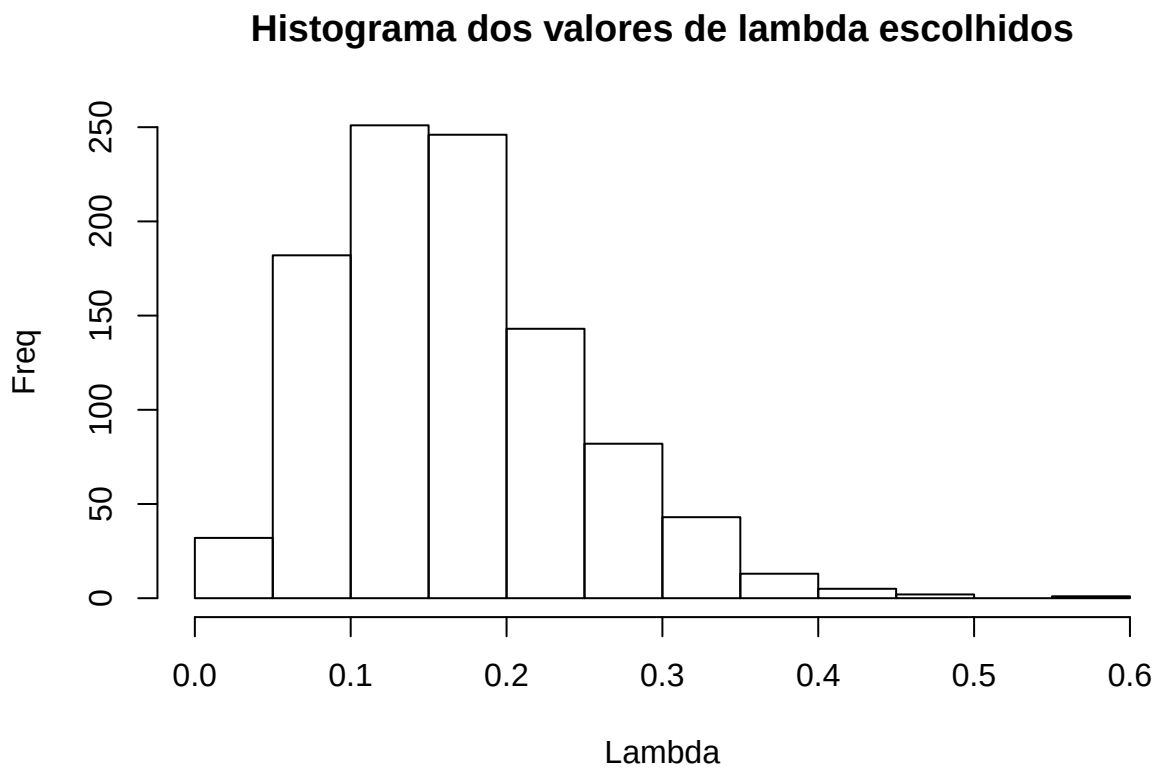
Coeficientes ajustados

Utilizamos os mesmos cálculos do experimento anterior para avaliar se os coeficientes escolhidos pelo modelo LASSO foram próximos dos verdadeiros definidos no modelo original. Observamos que a média foi 22 para ambos os tamanhos de k. Ou seja, o modelo LASSO **não** estimou $\beta 0$ próximo do valor verdadeiro.

Já a média dos coeficientes tanto para $k = 5$ e $k = 10$ foi 4, indicando que a escolha dos coeficientes também **não** se aproximou dos verdadeiros valores.

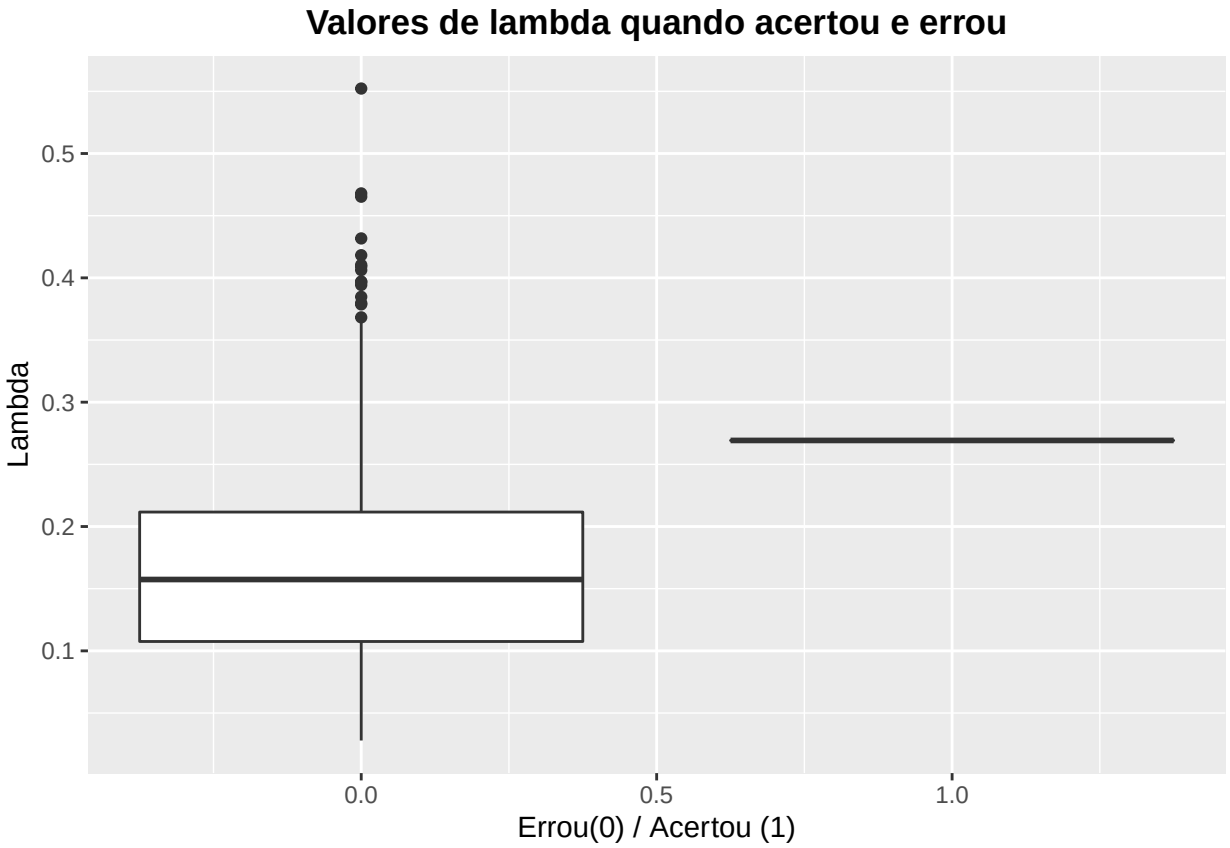
Escolha do lambda otimo

Observa-se que os valores de λ escolhidos para $k = 5$ e $k = 10$ foram os mesmos e a média foi 0.1662805.



Percebemos que a distribuição do λ neste caso já não se aproxima de uma Normal.

A única vez que o modelo LASSO acertou foi quando o valor de λ foi mais elevado do que a média dos valores escolhidos, como pode ser observado no gráfico abaixo:



O valor de λ quando o modelo acertou foi 0.2692271 e a média do λ escolhido quando o modelo errou foi 0.1661775.

Tamanho de amostra 200

Vemos na tabela abaixo que o modelo LASSO não apresentou melhora mesmo com o tamanho de amostra $n = 200$.

Table 3: Acertos exatos e parciais para tamanho de amostra 200

	Número de acertos
Exato	0
Inclui	1000

Alterando as variáveis resposta

Agora, vamos repetir o experimento mas alterando o modelo base para: $Y = 20 + 5X_1 + 5X_2 + 5X_3 + 5X_4 + \varepsilon$ onde $\varepsilon \sim N(0, 1)$

Table 4: Acertos exatos e parciais para cada tamanho de k

	k=5	k=10
Exato	0	0
Inclui	1000	1000

Observamos que com as variáveis corretas sendo X_1, X_2, X_3, X_4, X_5 o modelo LASSO não acerta as variáveis verdadeiras, apesar de incluí-las em todas as repetições. Acredito que a natureza Normal da distribuição das variáveis e do ϵ possam ter influência neste resultado.