Universidade Federal do Paraná
Setor de Ciências Exatas
Departamento de Estatística

Caio Gomes Alves

Daniel Krügel

Trabalho 1 - Uma modelagem matrimonial

Curitiba
2023

Caio Gomes Alves
Daniel Krügel

Trabalho 1 - Uma modelagem matrimonial

Trabalho de Conclusão de Curso apresentado à disciplina Laboratório B do Curso de Graduação em Estatística da Universidade Federal do Paraná, como exigência parcial para obtenção do grau de Bacharel em Estatística.

Orientador(a): Cesar Augusto Taconeli

Curitiba
2023

Agradecimentos

Resumo

O trabalho consiste em adaptar um modelo com resposta binária com uma base de dados retirada de um repositório de aprendizado de máquina. O principal problema dessa base é que ele foi concebido para utilizar a correlação das respostas para predizer se o casal irá ou não se divorciar ao final do estudo. Porém com a correlação perfeita entre variáveis explicativas criam problemas em modelos de regressão, em especial nas iterativas como os modelos generalizados.

Palavras-chave: Regressão linear generalizada. Penalização Lasso. multicolinearidade.

Sumário

1 Introdução

O trabalho consiste em ajustar um modelo de regressão linear generalizada com resposta binária em um conjunto de dados retirado do site UCI Machine Learning Repository, no caso deste projeto foi escolhido o data set Divorce Predictors.

Este dataset foi utilizado para predizer caso um casal irá ou não se divorciar ao final da terapia de casal aplicando um questionário de 54 questões, que variam desde conhecimento de hábitos e gostos até como eles se comportam durante as brigas.

Inicialmente a idéia era utilizar algoritmos de máquina como redes neurais, a utilização do método de regressão linear com família binomial e link logito não convergiu para um modelo conciso portanto foi necessário utilizar algum método de penalização, portanto a regressão Lasso foi escolhida como método de regularização.

2 Material e Métodos

2.1 Material

O software R foi utilizado junto com a tabela de dados retirados do site UCI, os pacotes **glmnet** , **tidyverse**, **GGally** e **highcharter**.

Os dados foram resumidos e apresentados na tabela 1 para o leitor ter uma idéia de como estão dispostos a formatação dos dados. A variável Class é a nossa resposta enquanto as variáveis Atr representam cada uma das perguntas feitas.

Tabela 1 – Uma tabela resumida dos dados utilizados: **xtable**.

Atr1	Atr5	Atr30	Class
2	0	1	1
4	4	1	1
2	1	2	1
3	3	3	1
2	1	1	1

2.2 Métodos

Os métodos utilizados neste trabalho foram os modelos de regressão linear generalizada (**??**), visualizações de correlação e métodos de regularização como a regressão lasso (**??**).

$$-(1/n)\sum |(y_i\beta_0 + x_i'\beta) + \lambda[(1-\alpha)||\beta||_2 + \alpha||\beta||_i]$$

Onde neste contexto

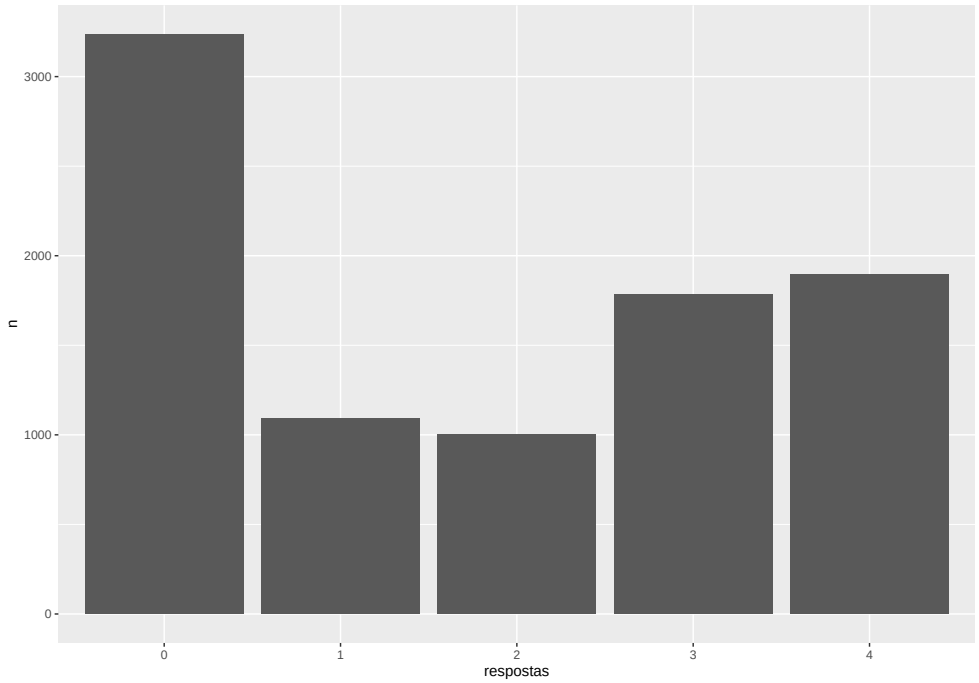
$$\alpha = 1$$

para intensificarmos a penalização e selecionarmos algumas das variáveis mais importantes .

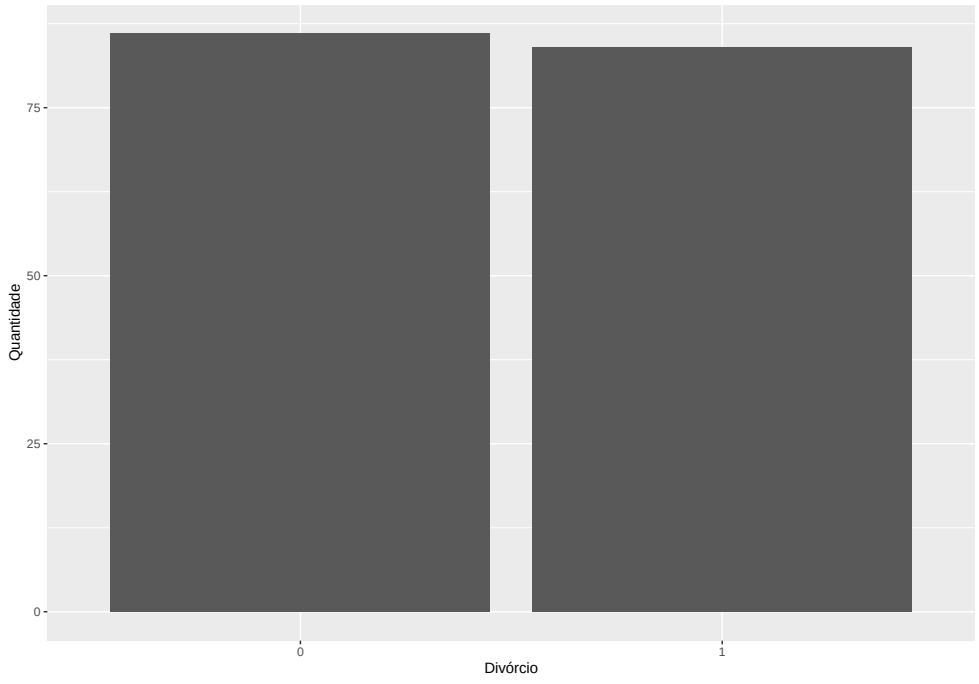
3 Resultados e Discussão

3.1 Análise descritiva

No gráfico a seguir podemos ver a distribuição das respostas dadas ao questionário aplicado aos casais

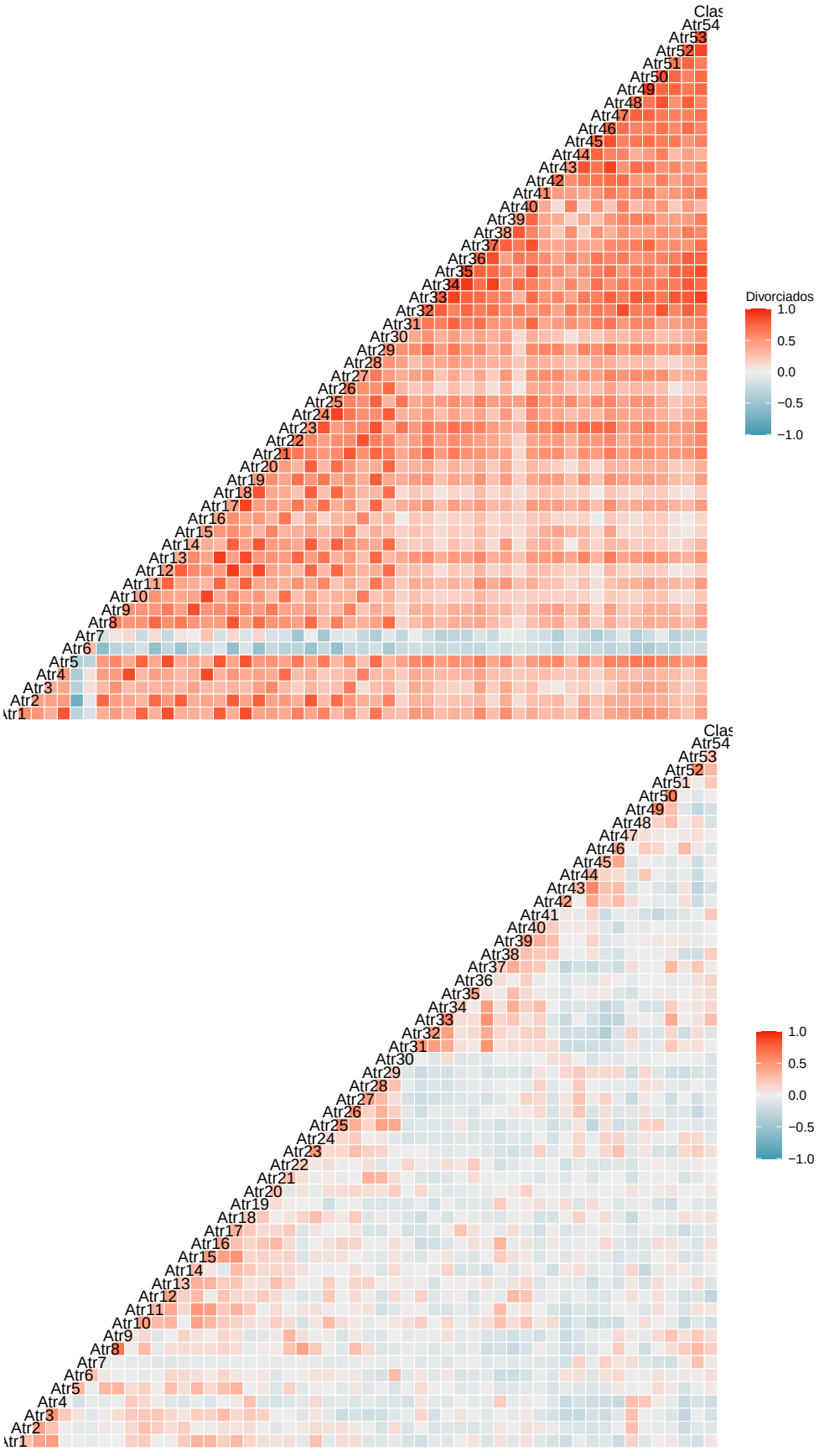


E no seguinte podemos ver o balanceamento das respostas

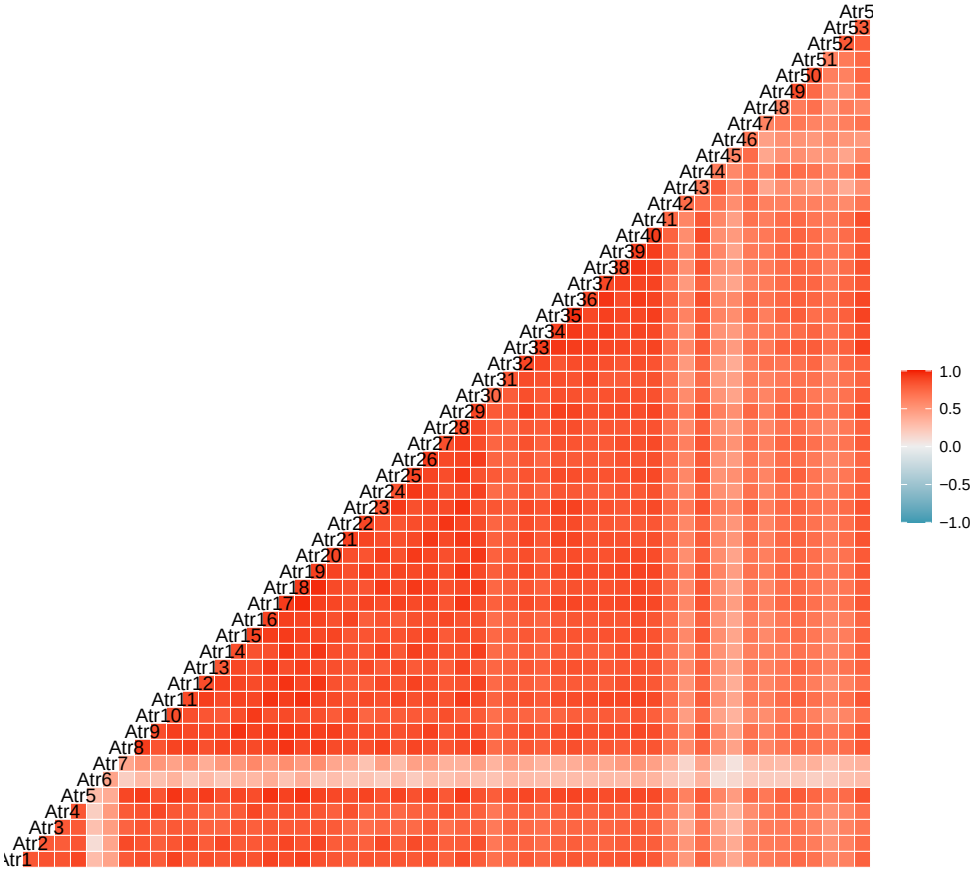


Como já descrito anteriormente a base de dados sofre de um problema crônico de multicolinearidade, portando separamos os dados em duas categorias, com a resposta

1 (ou seja, se divorciou ao final do estudo), resposta 0 (Continuou casado) para criar os correlogramas e identificar se há uma diferença entre as correlações das variáveis explicativas dependendo da resposta.



Como podemos ver quando o casal ao final do estudo apresenta uma correlação predominantemente positiva, possívelmente foi utilizado este fato para a predição utilizando um modelo de aprendizado de máquina. Porém como estamos tratando da base como um todo vamos fazer também o correlograma de todos os dados em conjunto.



Aqui podemos ver claramente que teremos problemas ao tentar ajustar um modelo linear com todas as variáveis explicativas.

3.2 Modelagem

Para questões ilustrativas o ajuste inicial com todas as covariáveis fica da seguinte forma:

```
fitZero <- glm(formula = Class ~.,
               family = binomial(link = 'logit'),
               data = dados)
```

Como a criação do modelo criou vários avisos diversos separamos o aviso com maior perigo, no caso:

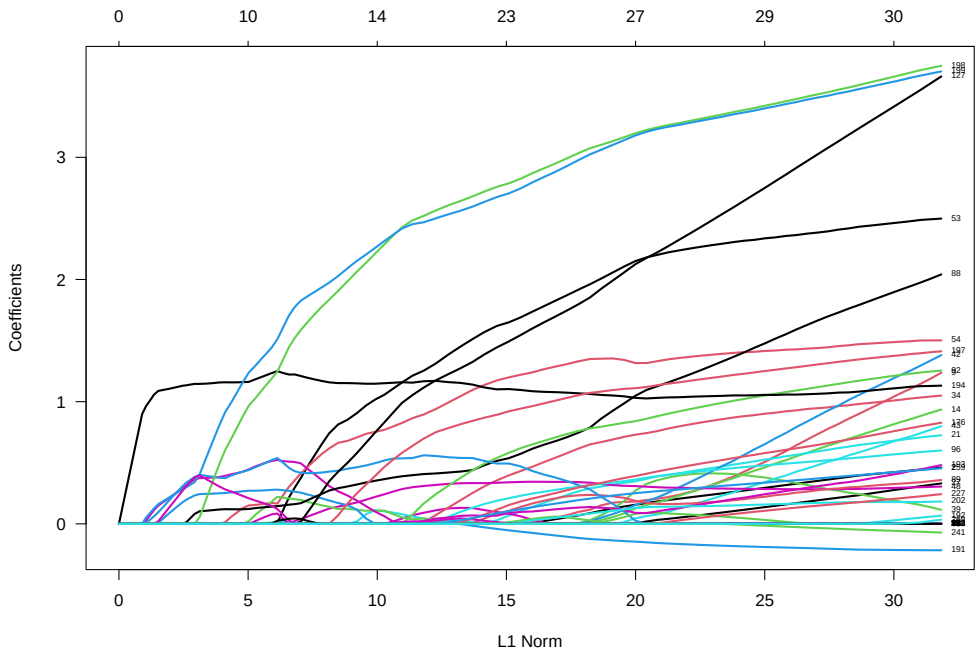
```
fitZero$converged
```

```
## [1] FALSE
```

Resultando em além de vários Betas estimados iguais a NA vários outros com probabilidade de ocorrência igual a 1, não tendo nenhum coeficiente sendo estatísticamente significativo ao modelo.

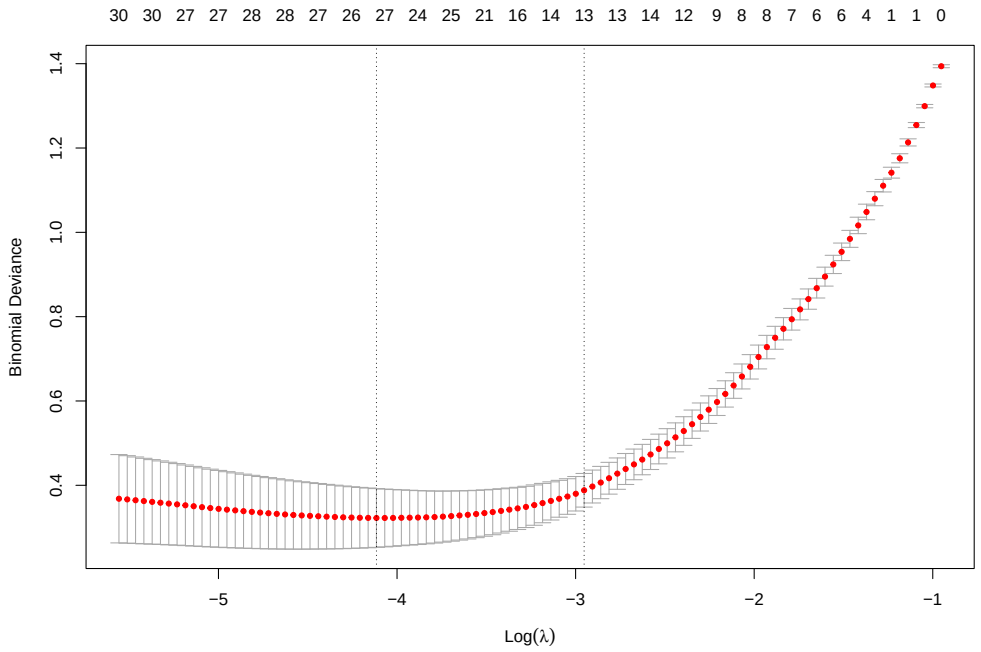
Portanto seguiremos com a técnica de regressão Lasso, para separar as variáveis significativas e punir o modelo pela incrementação de correlação. As regressões Ridge e Elastic-net foram ajustadas porém como não apresentaram melhora significativa no ajuste, não serão incluídas neste trabalho.

```
x <- model.matrix(Class ~., data = dados)[,-1]; y <- dados$Class

fit.lasso <- glmnet(x, y,
                    family = binomial(link = 'logit'),
                    alpha = 1)

plot(fit.lasso,
     las = 1,
     lwd = 2, label=TRUE)
```



Com uma quantidade tão grande de variáveis fica extremamente difícil de retirar uma informação deste gráfico poluído então em seguida vamos escolher o lambda ótimo para minimizar a soma de quadrados utilizando a função “cv.glmnet”.

```
cvfit <- cv.glmnet(x, y, family = 'binomial', alpha = 1, nfolds = 17)
plot(cvfit)
```



A seleção de folds para a validação cruzada foi de 17, para ser múltiplo da quantidade de amostras disponíveis mas não tão grande para que se reduzisse muito o número de iterações possíveis. O valor retornado pelo objeto cvfit$lamda.min foi de 0.01557972 e será utilizado dentro do glmnet para uma melhor adaptação do modelo.

```
fit.lasso.final <- glmnet(x, y,
                          family = binomial(link = 'logit'),
                          alpha = 1,
                          lambda = cvfit$lambda.min)
```

Os coeficientes finais do ajuste ficaram como:

```
## (Intercept)       Atr12       Atr24       Atr34       Atr51       Atr74
## -3.24827648  0.12193167  0.38698120  0.10541674  0.30825781  0.67433931
##       Atr84       Atr92       Atr93      Atr103      Atr113      Atr114
##  0.20929996  0.10093715  0.04338780  0.31795723  2.07709680  1.25492171
##      Atr183      Atr192      Atr193      Atr201      Atr262      Atr361
##  1.00200691  0.79484213  0.03292979  0.33537569  2.02803013  0.34902493
##      Atr373      Atr391      Atr393      Atr394      Atr402      Atr403
##  0.07931621 -0.13032455  0.06709633  1.01809968  1.07417525  3.14162322
##      Atr404      Atr412      Atr524
##  3.11630063  0.11828345  0.23784413
```

Como a qualidade de ajuste para regressões Lasso requerem habilidades computacionais fora do escopo da matéria encerraremos com este modelo o trabalho.

4 Considerações Finais

O primeiro trabalho de regressão linear generalizada nos ajudou a desenvolver melhor nossas percepções quanto a aplicações desses modelos e a explorar os limites da regressão quanto a efeitos de multicolinearidade e como encontrar métodos para contornar o problema.