Trabalho de Estatística II - Turma 22

Integrantes: Gabriel Rulka Tardoski, Israel Matias, Jordhan Emmanuel Marciano da Silva

& Zaira Mendonça Amorim.

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

O enunciado pedia que fizéssemos 'as regressões Ridge, Lasso e ElasticNet com a variável dependente “lwage” (salário-hora da esposa em logaritmo neperiano) e todas as demais variáveis explicativas na base de dados (todas aquelas que tentam explicar o salário-hora da esposa).', e com esse entuito, fizemos o seguinte processo:

**Pré-processamento dos Dados:**



Importamos a biblioteca glmnet (utilizada para as regressões de Ridge, Lasso e Elastic Neet) e configuramos a seed para o número fixo de '75', para garantir a replicabilidade e consistência da amostragem dos dados entre cada modelo.  
Ainda neste passo, a coluna earns também é removida do dataset, pois não será utilizada na análise



**Rotina genérica:**

Existem variáveis e dados em comum, acessados e utilizados pelos modelos em questão, e que portanto devem ser normalizados (no caso dos valores não binários), e indexados em um dataset:





Também se instância uma matriz, que logo adiante será usada para predição (com base nos valores do trabalho e aplicando logaritmo neperiano:









Feito isso, são gravados os valores para aplicação posterior em todos os modelos aqui sendo aplciados, com exceção de **‘*lambdas’*** que só é utilizada em **Ridge e Lasso.**



**Preview de como ficou os nossos dados genéricos:**



Regressão Ridge

Calculando o lambda





Estimativa de coeficientes:





Calculando o RMSE e R² do modelo em Ridge:





Fazendo a predição em Ridge:





Valor original predito foi -0.238, após a normalização, voltando a ficar em logaritmo neperiano, foi de $3.24, por fim, após a aplicação do antilog, se tornou $7.95.

O intervalo de confiança para nosso modelo ficou em 7.93 o inferior e 7.98 o superior.





O que significa que o salário pode variar entre 7.93 e 7.97.

Regressão Lasso

Como feito na Regressão Ridge, não será necessário recriar variáveis para dataset ou para predição, pois são os mesmos valores, além do próprio lambda, porém será feito um lambda otimizado para essa regressão.





Agora será realizado a predição do modelo Lasso e encontrar o RMSE e R²:





Em comparação a Regressão Ridge, o RMSE e R² também apresentou resultados baixos, explicando que o modelo não está tão bem otimizado.

Vamos realizar a predição com os dados montados:





Valor original predito foi -0.19, após a normalização, voltando a ficar em logaritmo neperiano, foi de $2.10, por fim, após a aplicação do antilog, se tornou $8.14.

Agora calculando o intervalo de confiança:





O que significa que o salário pode variar entre 8.12 e 8.16 na Regressão Lasso.

Regressão ElasticNet

Novamente, usando os mesmos dados genéricos, iremos direto para o treinamento do modelo, não há parâmetro alpha pois a própria regressão ElasticNet irá encontrá-lo.



Melhor alpha escolhido:



Fazendo a predição e adquirindo os valores de RMSE e R²:



Novamente não saíram tão bons.



Valor original predito foi -0.19, após a normalização, voltando a ficar em logaritmo neperiano, foi de $2.10, por fim, após a aplicação do antilog, se tornou $8.13, bem semelhante ao Lasso.

Para os intervalos de confiança:





Valores entre 8.11 e 8.15.