Regressões Ridge, Lasso e ElasticNet

Antes de iniciarmos os testes de regressão pedidos pelo trabalho realizamos um pré teste com uma regressão linear. Aplicando a função *<-lm(lwage~., data=trabalhosalarios)* na base *trabalhosalarios.Rdata* e em seguida listando os resultados com a função *summary* obtivemos o seguinte:

```
> resultadoRegr <-lm(lwage~., data=trabalhosalarios)
> summary(resultadoRegr)

Call:
lm(formula = lwage ~ ., data = trabalhosalarios)

Residuals:
    Min      1Q  Median      3Q     Max
-4.9480 -0.1218  0.0033  0.1230  1.7260

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7.572e-01  1.749e+00  -0.433 0.665135
husage       1.053e-03  1.308e-03   0.805 0.420637
husunion     5.850e-03  1.418e-02   0.412 0.680064
husearns     1.148e-04  2.000e-05   5.738 1.07e-08 ***
huseduc      3.993e-03  2.791e-03   1.431 0.152612
husblck      3.238e-03  7.133e-02   0.045 0.963801
hushisp     -1.033e-02  4.076e-02  -0.253 0.799913
hushrs      -9.905e-04  4.567e-04  -2.169 0.030166 *
kidge6       1.896e-02  1.357e-02   1.397 0.162457
earns        1.611e-03  2.781e-05  57.948  < 2e-16 ***
age          3.146e-01  2.915e-01   1.079 0.280532
black       -3.658e-02  7.229e-02  -0.506 0.612947
educ        -2.919e-01  2.915e-01  -1.001 0.316757
hispanic    -5.939e-02  3.889e-02  -1.527 0.126827
union        6.013e-02  1.703e-02   3.531 0.000421 ***
exper       -3.140e-01  2.915e-01  -1.077 0.281451
kidlt6       7.148e-02  1.637e-02   4.365 1.32e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2909 on 2557 degrees of freedom
Multiple R-squared:  0.6914,    Adjusted R-squared:  0.6894
F-statistic:   358 on 16 and 2557 DF,  p-value: < 2.2e-16
```

Dessa forma entendemos de antemão quais eram as variáveis de maior influência no valor Lwage (logaritmo do salário da esposa). Com essa pré-análise encontramos que as variáveis que influenciam na lwage com significância de 99.99% são: *husearns; earns; union; kidlt6;* E também que a variável *hushrs* influência na lwage com significância de 95%. Assim temos uma ideia do que esperar dos outros modelos de regressão.

Tabela de influência baseada na magnitude

Utilizaremos essa tabela de magnitude para explicar o grau de influência das variáveis em cada um dos modelos de regressão.

Magnitude do Coeficiente	Classificação
≥ 0.80	Muito Alta Influência
0.30 – 0.79	Alta Influência
0.10 – 0.29	Média Influência
0.01 – 0.09	Baixa Influência
< 0.01	Neutra ou Nula

Metodologia:

Para obtenção dos resultados de cada modelo de regressão proposto, executamos rotinas específicas para os três métodos avaliados: Ridge, Lasso e ElasticNet. Cada uma dessas rotinas foi estruturada separadamente em arquivos no formato .R, primeiramente buscamos encontrar a influência de cada variável para cada modelo, após isso elaboramos e calculamos as predições, os resultados foram organizados em uma tabela com destaque para os principais indicadores de desempenho sendo o coeficiente de determinação (R^2) e o erro quadrático médio da raiz (RMSE).

Modelo Ridge

De acordo com o modelo de predição Ridge, há apenas variáveis de média e baixa influência, como **husearns, black e kidlt6**.

Variavel	Ridge Matrix	Impacto na predição da var lwage
husage	0.0254412729	Baixa +
husunion	0.0308134832	Baixa +
husearns	0.1019435461	Media +
huseduc	- 0.0103491439	Baixa -
husblck	0.0494568551	Baixa +
hushisp	- 0.0001701853	Neutro
kidge6	0.0619025327	Baixa +
age	- 0.0109434839	Baixa -
black	- 0.1004324182	Media -
educ	0.0656719287	Baixa +
hispanic	- 0.0448851549	Baixa -
union	0.0173058698	Baixa +
exper	- 0.0020554000	Neutra
kidlt6	0.2008039146	Media +

Modelo Lasso

No modelo Lasso, ao fazermos uma avaliação das variáveis que seriam usadas para predizer a variável **Lwage**, notamos que algumas variáveis foram excluídas do modelo, sendo as variáveis mais relevantes a **union, educ**. E as **husage, husunion, husblck, hispanic, exper** são valores que não influenciam na predição e por isso foram excluídos.

Variavel	Lasso Matrix	Impacto na predição da var lwage
husage	.	Excluido
husunion	.	Excluido
husearns	0.231587248	Media +
huseduc	0.032330293	Baixa +
husblck	.	Excluido
hushisp	0.003433219	Neutro
hushrs	-0.062850796	Baixa -
kidge6	-0.151430086	Media -
age	0.047766633	Baixa +
black	-0.040958740	Baixa -
educ	0.340455604	Alta +
hispanic	.	Excluido
union	0.349678832	Alta +
exper	-	Excluido
kidlt6	-0.007916586	Neutro

Modelo ElasticNet

Nesse modelo temos uma regressão que, diferentemente dos modelos Ridge e Lasso, oscila o valor de alpha entre 0.0 até 1.0 a fim de encontrar o melhor “tuning” para a regressão. Ao rodar o código o melhor alpha encontrado, em todo o treino de regressão, foi 0.378 e o lambda foi 0.0127.

Resultados:

Métricas da Avaliação no conjunto de treino

Nome da Regressão	Índice RMSE	Índice R^2
Regressão Ridge:	0.8411446	0.2921320
Regressão Lasso:	0.8420298	0.2906413
Regressão ElasticNet:	0.4261596	0.2904916

Métricas da Avaliação no conjunto de teste

Nome da Regressão	Índice RMSE	Índice R^2
Regressão Ridge:	0.9893328	0.259084
Regressão Lasso:	0.9894877	0.258852
Regressão ElasticNet:	0.500716	0.2589132

Valores preditos por cada modelo

Modelo	Valor Predito do Salário/ Hora	Intervalo Inferior	Intervalo Sup
Ridge	$9.71	$9.50	$9.92
Lasso	$8.65	$8.46	$8.84
Elastic	$8,02	$7,87	$8,16

Com base nos indicadores apresentados, o modelo que apresentou o melhor desempenho para este problema foi o modelo “inserir”, evidenciado pelo menor RMSE e maior R^2.
Na sequência do trabalho, foi proposto apresentar os resultados das predições com intervalos de confiança para determinados valores. No entanto, surgiu a dúvida se essa análise deveria ser realizada apenas com o modelo que apresentou melhor desempenho ou com todos os modelos, a fim de comparar o comportamento preditivo entre eles. Optamos por aplicar essa etapa nos três modelos, de modo a confirmar que, além de apresentar os melhores indicadores de ajuste, o modelo selecionado também fornece as predições mais precisas e confiáveis

Obs: Nota-se que o código foi rodado em diferentes computadores então houve uma pequena variação nos indices analisados em cada computador rodado, mas a conclusão foi a mesma.

Conclusão

Observamos que o Elasticnet apresentou o melhor desempenho no modelo de regressão, por possuir o menor RMSE e maior R2. É interessante observar que o mesmo código rodado em diferentes computadores apresentou diferentes RMSE e R2 mas neles foi observado que o RMSE , no modelo ElasticNet obteve o menor erro dentre os 3. É importante considerar que uma possível razão para os modelos não terem desempenhado tão bem deve-se a exclusão da variável earns no treinamento dos modelos de regressão. Que no próprio pré treinamento apresentou maior influência na variável Lwage.