Tarefa 1 - Estimação de modelo preditivo com regressão linear

Tais Bellini - 205650

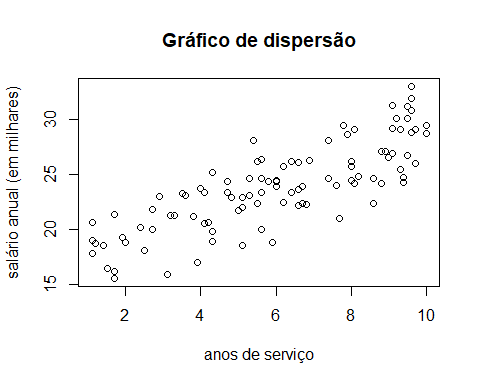
Janeiro, 2020

## Introdução

Neste relatório, iremos aplicar dois métodos para estimar um modelo de predição de salário baseado no tempo de serviço: regressão linear simples utilizando MQO e regressão linear simples minimizando a perda L1. Para realizar este estudo, usaremos uma amostra de 100 empregados de uma determinada empresa que contém o salário anual (em milhares de reais), denotado pela variável **Y**, e tempo de serviço em anos, denotado pela variável **X**.

## Dados

Primeiramente, vamos observar o gráfico de dispersão dos dados:



Observa-se que há uma relação linear positiva entre os dados e que o salário aumenta a medida que aumentam-se os anos de serviço. Portanto, consideramos apropriado utilizar a regressão linear.

## Estimação do modelo

Para a estimação do modelo que faça a predição do salário baseado no tempo de serviço, utilizamos dois métodos: regressão linear minimizando o erro quadrático (MQO) e regressão linear minimizando o erro absoluto.

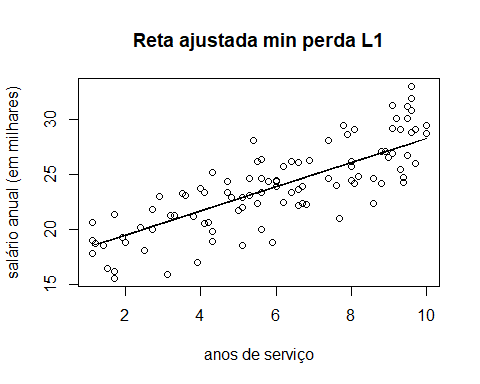
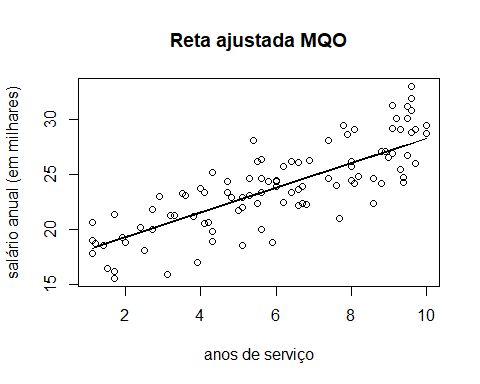
O experimento realizado consistiu em:

*(i)* dividir aleatoriamente a amostra em dois subconjuntos: treinamento (85% da amostra) e teste (15% da amostra); *(ii)* encontrar os coeficientes **b0** e **b1** que melhor ajustam a reta aos dados utilizando o MQO; *(iii)* calcular o erro absoluto médio do método MQO no subconjunto de treinamento (denotado por **MAEin**) e erro absoluto médio do método MQO no subconjunto de teste (denotado por **MAEout**); *(iv)* encontrar os coeficientes **b0** e **b1** que melhor ajustam a reta aos dados minimizando a perda L1; *(v)* calcular o **MAEin** e **MAEout** do método de minimização da perda L1 no subconjunto de teste.

Executando uma única vez o experimento, obtivemos os seguintes resultados:

### Resultados para uma repetição

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Método** | **Coeficiente b0** | **Coeficiente b1** | **MAEin** | **MAEout** |
| Regressão | 17.1135306 | 1.1174297 | 1.7708691 | 2.084269 |
| Min Perda L1 | 17.3367572 | 1.0947518 | 1.7656969 | 2.0799904 |



Observa-se que os valores ficaram muito próximos, não havendo diferença significativa entre os resultados dos dois métodos.

### Resultados para 100 repetições

Em seguida, executamos os passos *(i)* ao *(v)* 100 vezes e computamos a média dos coeficientes **b0** e **b1**, do **MAEin** e do **MAEout** calculados em cada repetição.

Observa-se que a média do coeficiente **b0** no método utilizando MQO é menor do que utilizando o método de minimização da perda L1. E o contrário se observa para **b1**.

**Estatísticas do coeficiente b0 utilizando MQO**

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.32 16.60 16.75 16.76 16.90 17.25

**Estatísticas do coeficiente b0 utilizando minimização da perda L1**

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 16.85 16.92 17.18 17.16 17.33 17.54

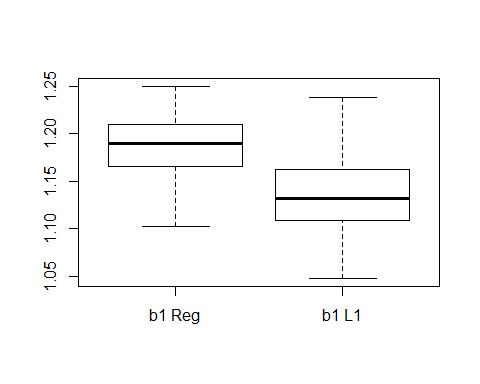
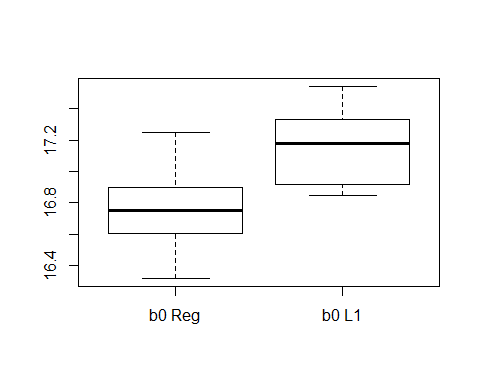
**Estatísticas do coeficiente b1 utilizando MQO**

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.102 1.166 1.190 1.187 1.209 1.250

**Estatísticas do coeficiente b1 utilizando minimização da perda L1**

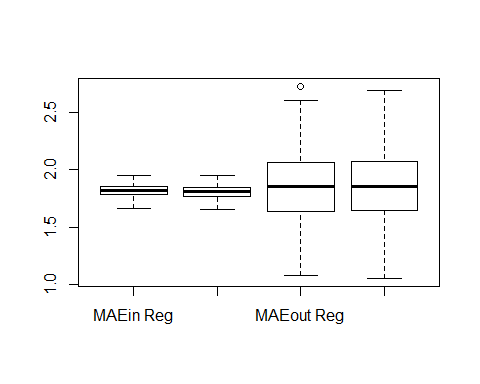
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1.048 1.109 1.132 1.135 1.162 1.238

**Gráficos comparando os coeficientes b0 e b1 para cada método de regressão**



Apesar de algumas diferenças na média dos coeficientes **b0** e **b1**, observamos que a média entre os erros absolutos médios fica muito próxima comparando ambos os métodos. Observamos também que o **MAEin** possui muito menos variabilidade do que o **MAEout** o que é esperado pois os coeficientes estão ajustados para o subconjunto de treinamento.

**Gráfico comparando o erro absoluto médio para cada método de regressão**



**Tabela com as médias dos coeficientes, MAEin e MAEout**

|  |  |  |
| --- | --- | --- |
|  | **Regressão** | **Opt L1** |
| b0 Médio | 16.762581 | 17.159672 |
| b1 Médio | 1.186750 | 1.134518 |
| MAEin Médio | 1.820388 | 1.808993 |
| MAEout Médio | 1.848896 | 1.847583 |

## Conclusões

Neste estudo, utilizamos uma amostra de 100 empregados para estimar um modelo de predição de salário baseado no tempo de serviço. Utilizamos dois métodos de regressão para esta estimativa: minimização do erro quadrático e minimização do erro absoluto. Após a execução do experimento uma vez, pudemos observar que a média dos erros absolutos da predição de ambos os métodos foram muito parecidas. Ao executarmos o experimento 100 vezes, também observamos o mesmo comportamento, apesar de a média dos coeficientes b0 e b1 ter alguma variação. Para estes dados, não há muita diferença na aplicação dos dois métodos. Se avaliarmos que a empresa em questão pode ter empregados com salários muito altos ou muito baixos, pode ser mais interessante utilizar método de minimização da perda L1, pois é mais robusto para valores discrepantes.