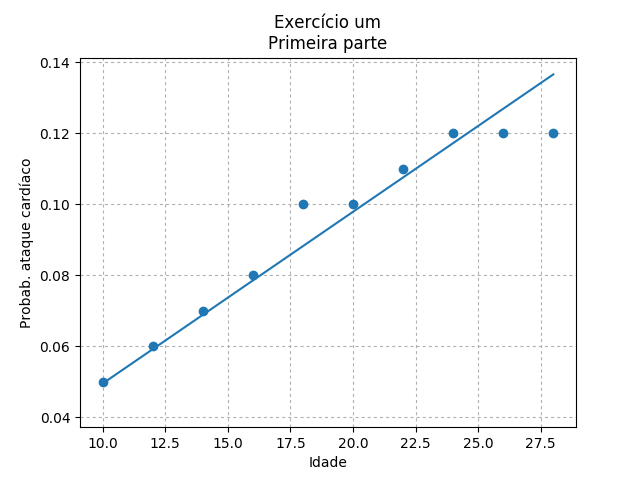
|  |  |  |
| --- | --- | --- |
|  | **UNIVERSIDADE FEDERAL DE GOIÁS**  **INSTITUTO DE INFORMÁTICA**  **CURSO DE MESTRADO/DOUTORADO EM CIÊNCIA DA COMPUTAÇÃO** |  |

# Lista de exercícios 1 – Regressão

**Os exercícios abaixo referem-se a base de dados “Risco de ataque cardíaco”**

1. Obtenha o modelo utilizando os 10 primeiros exemplos da base de dados. Calcule e apresente o erro quadrático médio aplicando o modelo de regressão nos mesmos 10 primeiros exemplos da base de dados. Depois calcule e apresente o erro quadrático médio do modelo de regressão obtido nos demais exemplos. Argumente se o modelo tem ou não uma boa capacidade de predição em novos exemplos.

**Resposta:** Ao utilizar os 10 primeiros exemplos da base, temos as seguintes informações:

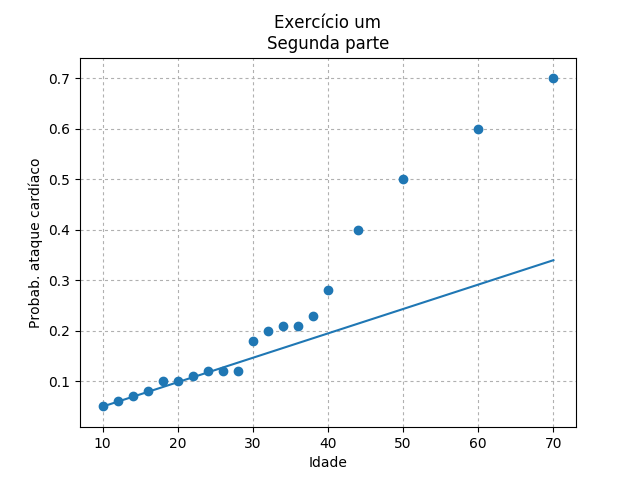


Sendo a função geradora do gráfico igual a:

Y = 0,00109 + 0,00484 \* X

Além disso, também temos que o Erro Quadrático Médio (ou EQM) aproximadamente de 3.08148791102e-34, e o coeficiente de determinação (ou R²) aproximadamente de 99,46%.

Porém, ao aplicar o mesmo modelo em todos os exemplos, teremos o seguinte resultado:



Nesse caso, teremos um EQM de 0.0979488391662 e um R² de 93,32%. Como é possível perceber, o erro quadrático médio aumentou consideravelmente e, por mais que o R² ainda permaneça acima dos 90%, o modelo começa a não ser tão eficaz para prever os futuros resultados. Isso se dá pois a probabilidade de ataque cardíaco não aumenta de maneira linear com o aumento da idade, e sim de maneira exponencial.

1. Agora obtenha o modelo utilizando os 5 primeiros exemplos da base de dados e também os 5 últimos. Calcule e apresente o erro quadrático médio aplicando o modelo de regressão nos 10 exemplos utilizados para obter o modelo de regressão. Depois calcule e apresente o erro quadrático médio do modelo de regressão obtido nos demais exemplos. Argumente se o modelo tem ou não uma boa capacidade de predição em novos exemplos. Compare com os resultados do exercício anterior e argumente as possíveis diferenças de resultados.

**Os exercícios 3, 4 e 5 referem-se a base de dados “Preços de apartamentos”**

1. Qual o coeficiente de correlação entre cada uma das variáveis com o preço de apartamento? Qual a variável mais importante para explicar o preço de apartamento? Justifique sua resposta.

**Resposta:** Os coeficientes de correlação são:

- Tamanho do apartamento: 0.60310357816718896

- Idade do prédio: -0.37834857467465671

- Andar do prédio: 0.28134239935629496

- Número de quartos: 0.46296723429664738

- Garagens: 0.41558722742498921

1. O banco de dados contém informações de 40 apartamentos vendidos no mês passado. Cada linha do banco de dados é um apartamento. Ajuste o seguinte modelo de regressão múltipla para os dados:

Y=b0+b1x1+b2x2+b3x3+erro, em que:

Y=preço do apartamento;

X1=tamanho do apartamento, em metros quadrados;

X2=idade do prédio, em anos;

X3=andar em que o apartamento está.

Obs.: Note que não usaremos todas as variáveis independentes.

Qual é o preço previsto de um imóvel com 80m2, 10 anos e que está no 9º andar?

**Resposta:** Ao utilizar apenas as variáveis indicadas, temos os seguintes resultados:

Y = 324728.19 + 6920.21 \* (tamanho) - 17590.17 \* (idade) + 11216.95 \* (andar)

Preço do novo imóvel: R$ 803.396,11

1. Ajuste o modelo de regressão múltipla fazendo uso de todas as variáveis. Qual deve ser o preço de um imóvel com 100m2, 3 anos, andar de número 5, 3 quartos e 2 vagas de garagem?

**Resposta:** Ao calcular o modelo de regressão utilizando todas as variáveis, temos os seguintes resultados:

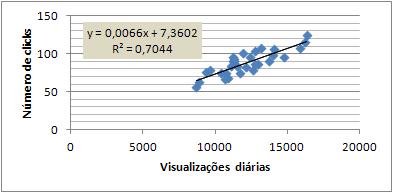
Y = 136719.21 + 7442.84 \* (tamanho) - 13814.20 \* (idade) + 11419.89 \* (andar) - 21884.64 + (quartos) \* 99843.12 + (vagas)

Preço do novo imóvel: R$ 1.030.691,91

1. Explique com suas palavras a importância do uso da regressão no exemplo deste artigo: <https://www.linkedin.com/pulse/using-regression-predict-baseball-salaries-nate-reed>

**Resposta:** Nesse artigo, fora proposto calcular e utilizar um modelo de regressão para tentar sistematizar a relação entre o salário de um atleta de baseball e suas habilidades dentro do jogo. A regressão nesse caso foi importante para, além de determinar uma fórmula para calcular os salários de acordo com as características dos jogadores, também serviu para indicar quais habilidades são mais ou menos importantes na hora de calcular o salário de um atleta (ou seja, aquelas variáveis que mais aumentam ou diminuem o salário final a ser pago), além de verificar que certas características observáveis são inter-dependentes entre si, ao ponto de ser necessário apenas considerar algumas, e as outras podem ser vistas como consequências dessas iniciais.

1. Um portal da internet cobra anúncios na página principal de acordo com o número de visualizações diários da página. Um anunciante diz que o mais importante para ele é o número de “clicks” diários no seu anúncio. O portal preparou um estudo com dados dos últimos 30 dias. Foi observado o número de visitas únicas diárias do portal e o número de clicks diários do anúncio. O resultado do ajuste da regressão de Y(número de clicks) por X(número de visualizações) é mostrado no gráfico de dispersão a seguir.



Assinale V ou F justificando sua resposta para as falsas:

* + - * 1. Apenas 0,66% das visualizações resultam em clicks. Portanto, a regressão não está boa. **Falso, ainda que 0,66% das visualizações realmente resultem em clicks, a regressão está em um bom status, haja visto o R² que está aproximadamente em 0,7**
        2. O número de visualizações é um bom preditor do número de clicks porque o R2 é alto e a reta parece bem ajustada. **Verdadeiro**
        3. 0,7044% das visualizações são convertidas em clicks. **Falso, 0,7044% representa o quão os dados informados determinam uma tendência**
        4. As duas variáveis não estão linearmente correlacionadas porque o R2 é menor que 1. **Falso, a correlação nunca tem como ficar maior que 1**
        5. Se num determinado dia o site tiver 10000 visualizações, o número estimado de clicks é 66. **Falso, na verdade ele irá ter aproximadamente 73 clicks**

1. Ilustre e explique uma aplicação em que a regressão logística se apresente como mais adequada do que uma regressão linear.

**Resposta:** A regressão logística é aquela onde o resultado Yi apresenta um valor binário, ou seja, um status de “fracasso” ou “sucesso”. Um exemplo onde poderia ser usado uma regressão logística seria no departamento de vendas, como forma de encontrar quais segmentos de clientes responderiam de maneira bem-sucedida a abordagem padrão de uma equipe de vendas, indicando para quais grupos a abordagem de vendas apresenta “fracasso” ou “sucesso”.

**Opcionais:**

* Brincando com os coeficientes:

http://students.brown.edu/seeing-theory/regression/index.html#first

* Notebook de regressão linear para predição de preços de casas na cidade de Boston.

http://facweb.cs.depaul.edu/mobasher/classes/csc478/Notes/IPython%20Notebook%20-%20Regression.html

* Notebook de regressão logística para predição de relacionamento extra-matrimoniais de mulheres:

http://nbviewer.jupyter.org/gist/justmarkham/6d5c061ca5aee67c4316471f8c2ae976