

Regressão Linear

Interpretando os dados após a transformação logarítmica

O que é Regressão Linear?

É uma ferramenta de análise quantitativa de dados, partindo do pressuposto que temos duas grandezas com relação linear, que variam proporcionalmente, podemos estimar a melhor reta que representa nossos dados. Melhor no sentido da minimização de alguma função de erro. Ela permite prever a partir dos dados iniciais, fazer previsões de valores futuros.

“A análise de regressão é a parte da estatística que investiga a relação entre duas ou mais variáveis relacionadas de maneira não determinística” (DEVORE, 2008, p. 455)

Notação matemática

$$y = \beta_0 + \beta_1 x_1$$

onde:

y = variável aleatória ou dependente

x_1 = variável preditora ou variável independente

β_0 = intercepto

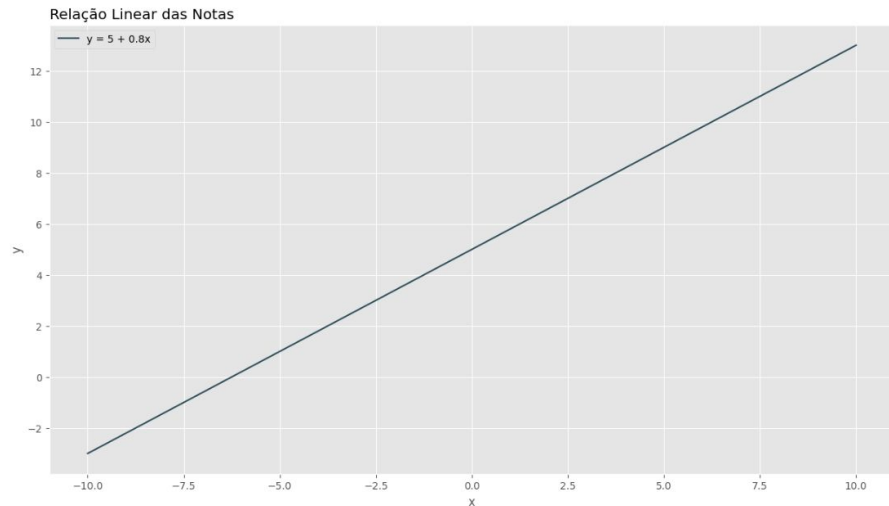
β_1 = inclinação

- Intercepto: é o ponto onde a reta atravessa y.
- Inclinação: em caso de inclinação positiva os dados vão ter valores crescentes, e em caso de inclinação negativa terão valores decrescentes. **Quando os valores de x aumentam, os de y também aumentam e vice-versa.**
- A variável dependente é aquela que buscamos prever valores, ou entender sua relação com as outras. Apesar de ser usada mais frequentemente para prever valores futuros a regressão linear também nos permite entender como as variáveis independentes vão afetar de forma quantitativa a variável dependente.

Visualizando um exemplo

Vamos supor que encontramos a relação linear das notas de uma turma da graduação e as horas de estudo sendo ela: $y = 5 + 0,8x$.

O que isso significa? O intercepto (5) é a nota média do aluno caso ele não estude ($x = 0$), já para cada hora estudada a nota desse aluno irá aumentar 0,8. Em exemplos com mais variáveis entendemos que estes valores correspondem ao caso das mesmas permanecerem constantes.



O modelo probabilístico

$$\text{✏ } Y = \beta_0 + \beta_1 x + \varepsilon$$

onde:

ε = desvio aleatório ou termo do erro aleatório

- Nos modelos que utilizamos em Machine Learning, temos o desvio aleatório. Esta é a diferença entre os valores previstos e reais, quando sua distribuição se afasta da normal indica a necessidade de ajustes no modelo.
- A adequação do modelo depende de alguns conceitos de linearidade, sendo eles a linearidade na relação de x e y, a homocedasticidade, a independência e a normalidade.

Mas porque fazer a transformação logarítmica?

Vamos antes explicar melhor os conceitos de linearidade citados:

- **Linearidade:** as variáveis x e y devem ter uma relação linear;
- **Homocedasticidade:** assume que os resíduos têm uma variância constante;
- **Independência:** assume que os termos de erro são independentes;
- **Normalidade:** assume que os resíduos são normalmente distribuídos.

Em muitos casos reais, os dados que iremos trabalhar não obedecem totalmente estes conceitos e devemos buscar formas de realizar ajustes para melhor adequação do modelo preditivo.

Portanto a transformação logarítmica permite:

- **Linearizar as relações das variáveis:** Quando a relação entre as variáveis é não linear, um aumento constante na variável independente não leva a um aumento proporcional na variável dependente;
- **Estabilizar a variância:** A variância desigual dos resíduos, heterocedasticidade, é um problema recorrente nas regressões lineares, ela gera aumento na variabilidade dos erros de predição.
- **Normalizar as distribuições:** Um modelo de regressão linear assume que os resíduos seguem uma distribuição normal, caso contrário a precisão das inferências e a robustez do modelo será afetada.


Usar a transformação logarítmica pode levar a um melhor ajuste do modelo. Se traduzindo em menores erros de predição e maior confiabilidade nas estimativas dos parâmetros do modelo.

Mas como interpretar os resultados agora?

	Coeficientes
Intercepto	7.713988
Sex_Female	0.038399
Sex_Male	-0.038399
Smoker_No	-0.768754
Smoker_Yes	0.768754
Region_Northeast	0.068377
Region_Northwest	0.024719
Region_Southeast	-0.060221
Region_Southwest	-0.032875
Age	0.033997
BMI	0.013150
Children	0.099169

Após a transformação vamos ter resultados como o da imagem (vinda de um projeto de regressão linear baseado em dados de seguro saúde). O intercepto e os coeficientes estão log e a forma da notação matemática também será alterada. Seguiremos para um exemplo utilizando a variável Age e considerando os restante das variáveis constantes.

Para interpretar o resultado aplicamos a função exponencial. De maneira simplificada para cada aumento unitário de idade (Age), os custos aumentaram aproximadamente 3,4% já que o resultado de resultado de $e^{0.033997}$ é aproximadamente 1,034. E da mesma forma é possível interpretar as relações dos outros coeficientes.


$$\log(\text{Charges}) = \beta_0 + \beta_{\text{Age}} \times \text{Age}$$

$$\log(\text{Charges}) = 7.713988 + 0.33997 \times \text{Age}$$

$$\text{Charges} = e^{\log(\text{Charges})}$$

$$\text{Charges} = e^{7.713988 + 0.033997 \times \text{Age}}$$

$$\text{Charges} = e^{7.713988} \times e^{+0.033997 \times \text{Age}}$$

Referências

- Curso de Modelagem e Inferência Estatística da UNIVESP
 - <https://youtu.be/dJrn2Uww254?si=W8fufCzBs8ig9AiU>
 - https://youtu.be/aeIDTQ31Jaw?si=T_m9TSIFurj1--rK
- Universidade dos Dados
 - <https://www.universidadedados.com/>
- Meu projeto de Regressão Linear
 - https://github.com/datalopes1/medical_cost/
 - <https://www.kaggle.com/code/andreluizls1/previs-o-de-pre-os-de-seguro-com-regress-o-linear>