Data Science - Análise de Regressão Linear Aplicado a Previsão de Vendas

Robson Fernandes
Professor Universitário

Meu primeiro artigo no Linkedin!

Meu nome é Robson Fernandes, sou professor universitário e aluno do programa de Mestrado em Matemática, Estatística e Computação Aplicada com ênfase em Ciência de Dados da Universidade de São Paulo - ICMC-USP.

Pretendo iniciar um ciclo de artigos relacionados a área de Ciência de Dados para incentivar profissionais, estudantes ou interessados na área a avançarem seus estudos!

Abordarei inicialmente alguns assuntos relacionados a Análise de Regressão Linear, como: Diagrama de Dispersão, Correlação Linear, Teste de Hipóteses, Teste de Normalidade, Ajuste do Modelo Linear, Previsão de Dados e Coeficiente de Determinação.

Ambos os assuntos serão aplicados a um estudo de caso, e desenvolvidos em linguagem R.

Espero que gostem!

Introdução

Análise de Regressão

A análise de regressão estuda o relacionamento entre uma variável, chamada **variável dependente,** e outras variáveis, chamadas **variáveis independentes**.

Este relacionamento é representado por um modelo matemático, isto é, por uma equação que associa a **variável dependente** com as **variáveis independentes**.

Modelo de Regressão Linear Simples

O modelo de regressão linear simples, define uma relação linear entre a **variável dependente** e uma **variável independente**. Possui esse nome porque considera que a relação entre as variáveis é descrita por uma função linear (equação da reta ou do plano).

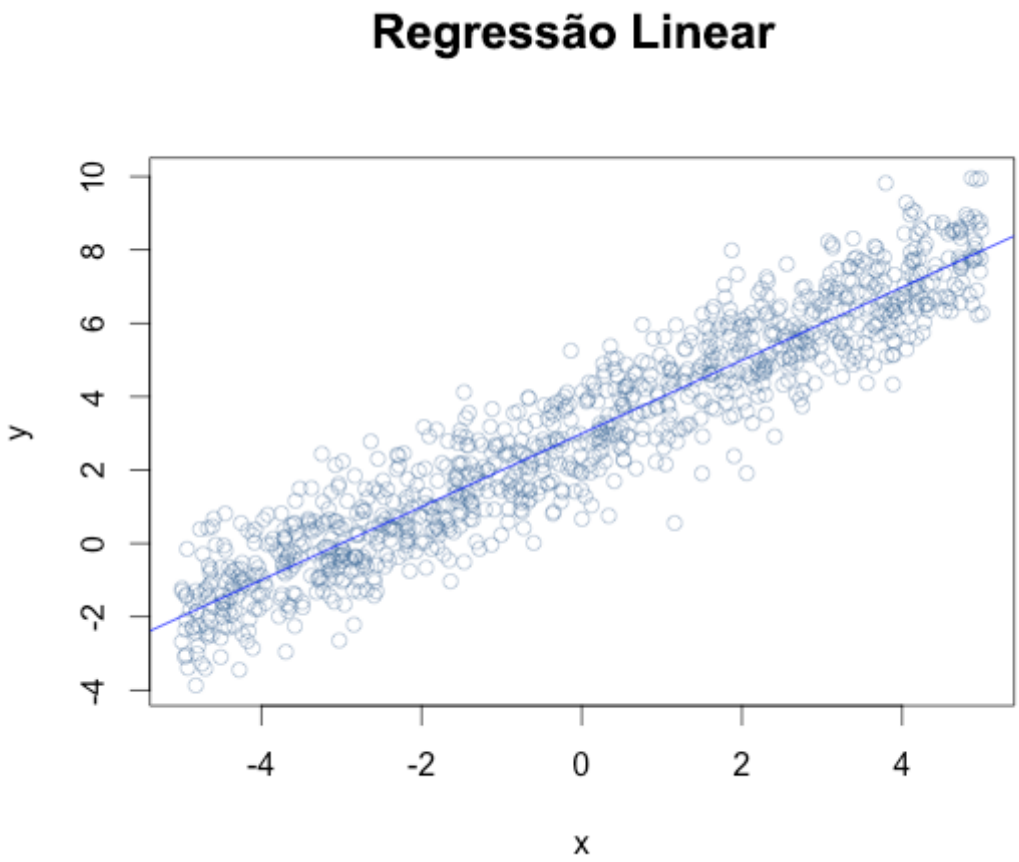
$$y = \beta_0 + \beta_1 x + \varepsilon$$

- ***y*** **-** Variável que será explicada, dependente (aleatória).

- ***βO*** **-** Parâmetro desconhecido (a estimar).
- ***β1*** - Parâmetro desconhecido (a estimar).
- ***x*** **-** Variável explicativa ou independente, medida sem erro (não aleatória).
- ***ε*** **-** Variável erro aleatório, com média 0 e variância σ^2.

Equação da Reta - Regressão Linear Simples

Como podemos observar no gráfico abaixo, o modelo de regressão linear simples é descrito por uma função linear (equação da reta ou do plano) .



Exemplos

1. Relação entre o peso e a altura de um homem adulto
   (**x**: altura; **y**: peso)

2. Relação entre o preço do vinho e o montante da colheita em cada ano
   (**x**: montante da colheita; **y**: preço do vinho)

3. Relação entre a média salarial e a experiência profissional em anos
   (**x**: experiência profissional; **y**: média salarial)

Estudo de Caso

A empresa **RX Sales (***empresa fictícia***),** está atuando no ramo de vendas de relógios. Eles perceberam as vantagens em realizar previsões de vendas sobre os seus negócios, e viram que para realizar previsões eficazes, precisam manter o controle de seus atuais e últimos números de vendas.

Atualmente, estão no processo de previsão de suas vendas para cada trimestre do ano que vem. Para fazer isso, levantaram suas vendas dos últimos 12 trimestres.

Previsão de Vendas

Vamos encontrar um modelo matemático utilizando regressão linear simples que irá nos permitir prever as vendas para os próximos trimestres. Descobriremos também, qual a capacidade do modelo matemático em descrever o conjunto de dados.

Conjunto de Dados

Trimestre	Venda
1	700,00
2	1600,00
3	1550,00
4	1500,00
5	2400,00
6	3100,00
7	2600,00
8	2950,00
9	3800,00
10	4500,00
11	4000,00
12	4900,00

O conjunto de dados possui informações das vendas que foram realizadas ao longo de cada trimestre, sendo que, a análise será efetuada sobre as vendas dos últimos 12 trimestres.

O modelo de regressão linear simples deverá analisar a relação entre as variáveis venda e trimestre (**x**: trimestre; **y**: venda)

Desenvolvimento do Algoritmo (Implementação em R)

Leitura de Dados

Primeiramente, vamos organizar os dados como *objetos de dados* no R através de um *data frame* (planilha). Para isso, é necessário que a tabela acima se encontre em uma estrutura tabular, na qual as colunas representam as variáveis e as linhas representam os valores. Logo, seja o arquivo de texto **vendas.txt (**download**)**, utiliza-se a função *read.table* para que o arquivo seja lido pelo R:

```
1  #Leitura do Arquivo vendas.txt
2  vendas = read.table('vendas.txt', header=T)
3  attach(vendas)
```

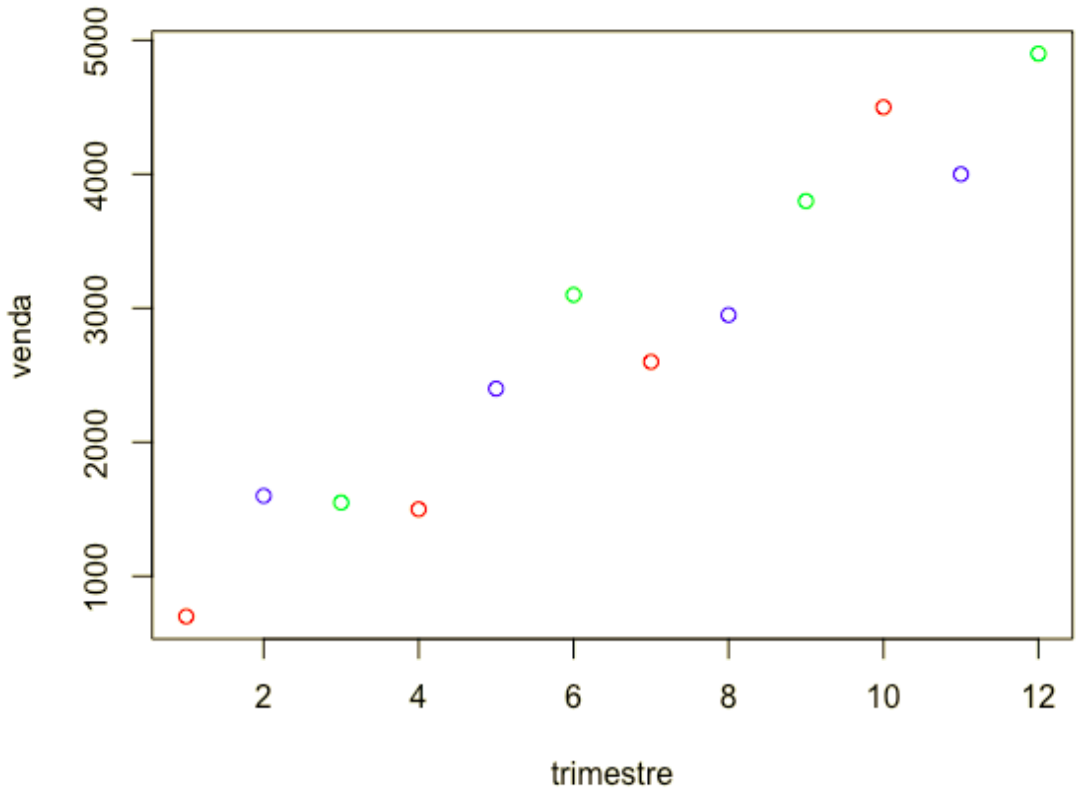
Observe que o argumento *header=T* indica que a primeira linha do arquivo contém os rótulos da planilha e que a função *attach* anexa o objeto **vendas** no ambiente de variáveis globais do R.

Diagrama de Dispersão

Para verificar a existência de alguma relação entre as variáveis ***trimestre*** e ***venda***, deve-se construir um **Diagrama de Dispersão** para as duas variáveis:

```r
#Diagrama de Dispersão
plot(trimestre, venda, col=c('red', 'blue', 'green'))
```

A figura abaixo apresenta o **Diagrama de Dispersão** produzido pelo código em R acima. É possível fazer uma interpretação da existência de uma **Correlação Positiva**, onde é observado que ao aumentar a variável **x**:**trimestre**, acarreta em um aumento da variável **y**:**venda**, assim se controlarmos **x**, **y** será também controlado.



Coeficiente de Correlação Linear

Para calcular o **Coeficiente de Correlação Linear de Pearson** entre as variáveis, utilize a função *cor*:

```r
#Coeficiente de Correlação
cor(trimestre,venda)
```

O R irá retornar o valor **0.9658553** o que evidencia uma forte correlação linear entre as variáveis. Este coeficiente, assume apenas valores entre **-1** e **1.**

Interpretação dos valores do Coeficiente de Correlação

- **1.0** positivo ou negativo. Indica uma correlação perfeita
- **0.9** positivo ou negativo. Indica uma correlação muito forte.
- **0.7 a 0.9** positivo ou negativo. Indica uma correlação forte.
- **0.5 a 0.7** positivo ou negativo. Indica uma correlação moderada.
- **0.3 a 0.5** positivo ou negativo. Indica uma correlação fraca.
- **0 a 0.3** positivo ou negativo. Indica uma correlação desprezível.

Teste de Hipóteses para o Coeficiente de Correlação

Para que possamos avaliar se esse resultado é significativo, vamos realizar um Teste de Hipóteses para o Coeficiente de Correlação.

```
11  #Teste de Hipóteses para o Coeficiente de Correlação
12  cor.test(trimestre,venda)
```

Ao executar o Teste de Hipóteses, teremos o seguinte resultado

```
        Pearson's product-moment correlation

data:  trimestre and venda
t = 11.789, df = 10, p-value = 3.451e-07
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.8794231 0.9906395
sample estimates:
      cor
0.9658553
```

Observem que o **p-value** do teste **(p-value = 3.451e-07)** é bem pequeno, conclui-se que o valor do **Coeficiente de Correlação Linear de Pearson** tem significância Estatística.

O teste de **Correlação de Pearson**, também indica que a hipótese deve ser aceita, visto o intervalo de confiança entre **0,8794231** e **0,9906395.**

Regressão Linear Simples

Ajuste do Modelo Linear

Sejam **x** e **y**, respectivamente, as variáveis **trimestre** (explicativa) e **venda** (resposta). Propõe-se um modelo de regressão linear, dado pela equação abaixo.

$$y = \beta_0 + \beta_1 x + \varepsilon$$

Onde **β0** e **β1**são parâmetros desconhecidos e **ε** é o erro aleatório.

Para ajustar um modelo de regressão linear no R utiliza-se a função **lm**, conforme o código abaixo.

```
14  #Ajuste do Modelo de Regressão Linear
15  ajuste.modeloLinear = lm(venda ~ trimestre)
16  ajuste.modeloLinear
```

Ao executar o modelo ajustado irá retornar as seguintes informações

```
Call:
lm(formula = venda ~ trimestre)

Coefficients:
(Intercept)    trimestre
      502.3        353.5
     Beta 0       Beta 1
```

Note que função **lm()** é chamada com o formato **lm(y ~ x)**, ou seja, a variável resposta é **y** e a explicativa é **x** , sempre nessa ordem.

O R retorna o valor dos coeficientes de **β0** e **β1** estimados via **Método de Mínimos Quadrados**. Logo, a equação da reta ajustada é dada por:

y = 502,3 + 353,5 * x

Onde poderíamos interpretar que, para estimar ou prever o valor da **venda** basta informar o **trimestre** na equação e realizar o cálculo.

venda = 502,3 + 353,5 * trimestre

Teste de Significância do Modelo

Realizando o teste de significância, verifica-se o *P-valor* das variáveis através da saída da função **summary**:

```
18  #Teste de Significância do Modelo
19  summary(ajuste.modeloLinear)
```

Ao analisar o teste, o *P-valor* da variável **trimestre** está bem próxima de 0, isso demonstra que esta é uma variável significante ao modelo. Conforme a saída da função **summary** abaixo.

```
Call:
lm(formula = venda ~ trimestre)

Residuals:
    Min      1Q  Median      3Q     Max
-416.26 -377.62   51.75  214.51  476.75

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   502.27     220.69   2.276   0.0461 *
trimestre     353.50      29.99  11.789 3.45e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 358.6 on 10 degrees of freedom
Multiple R-squared:  0.9329,    Adjusted R-squared:  0.9262
F-statistic:   139 on 1 and 10 DF,  p-value: 3.451e-07
```

Teste de Normalidade

A normalidade da amostra é confirmada pelo **Teste de Normalidade de Shapiro-Wilk**, cujo *P-valor* **0.1108** > **0,05**. Conforme instrução em R abaixo.

```
21  #Teste de Normalidade
22  shapiro.test(residuals(ajuste.modeloLinear))
```

Previsão de Vendas

Tendo uma função matemática significante que representa nosso conjunto de dados, vamos realizar a previsão de vendas no décimo quarto trimestre (14) do conjunto de dados utilizando a função **predict** no R.

```r
# Previsão de Vendas
predict(ajuste.modeloLinear, newdata=data.frame(trimestre=14))
```

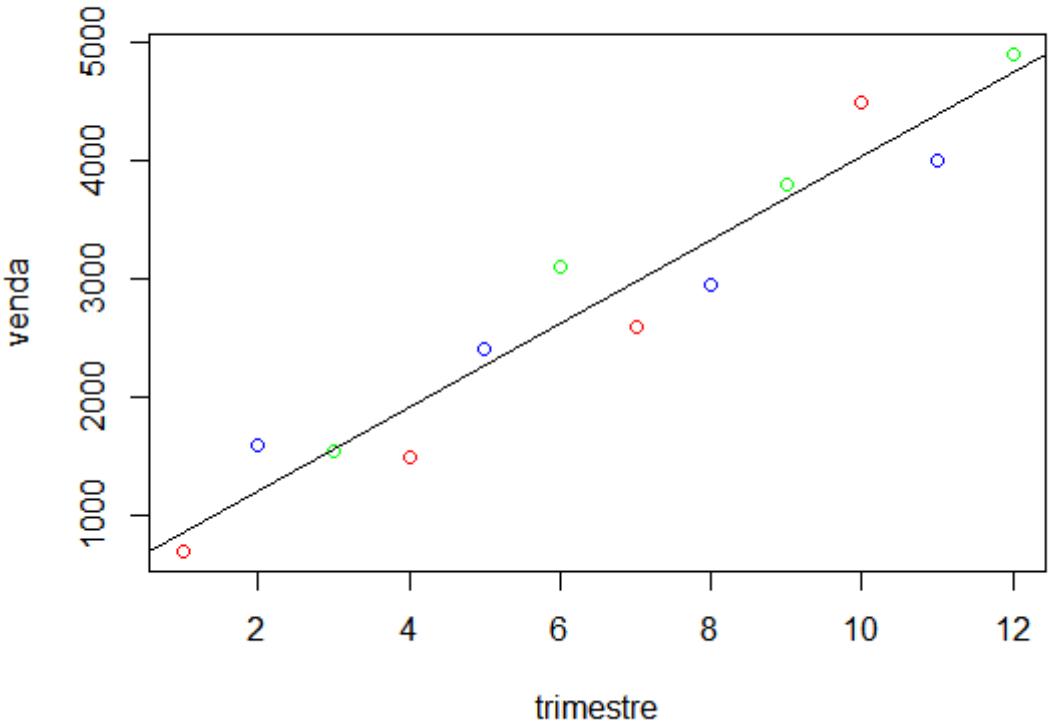
Ao executar, teremos que a previsão de vendas será de **5451,224.**

Reta Ajustada

Para esboçar a reta ajustada no diagrama de dispersão, vamos utilizar a função **abline**:

```r
#Reta ajustada no Diagrama de Dispersão
plot(trimestre, venda, col=c('red', 'blue', 'green'))
abline(ajuste.modeloLinear)
```

Ao executar o trecho em R, teremos o Diagrama de Dispersão com a reta ajustada, conforme figura abaixo.



Avaliação da Qualidade do Modelo

Coeficiente de Determinação

O coeficiente de determinação **R^2**, reflete na qualidade do ajuste do modelo, ou seja, o quanto o modelo consegue explicar o conjunto de dados.

Vamos utilizar a função **summary** do R para extrair o valor do coeficiente de determinação **R^2**.

```r
#Coeficiente de Determinação
summary(ajuste.modeloLinear)$r.squared
```

Sendo o **R^2 = 0.9328764**, logo

O modelo consegue explicar cerca de 93% dos dados observados.

Download

Link para download do algoritmo em R.

Até a próxima...

Espero que esta abordagem possa contribuir para aqueles que estão iniciando na área de Ciência de Dados, sejam, Estatísticos, Matemáticos, Cientistas da Computação ou estudantes que tenham interesse no assunto.

Um grande abraço a todos!