Estatística com Apoio Computacional Regressão Linear

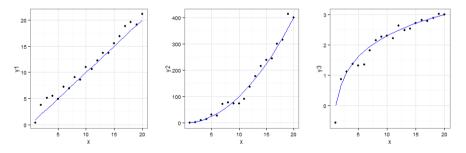
Universidade Estadual Vale do Acaraú – UVA

Paulo Regis Menezes Sousa paulo_regis@uvanet.br



- A análise de regressão consiste na realização de uma análise estatística com o objetivo de verificar se há uma relação funcional entre uma variável dependente e uma ou mais variáveis independentes.
- Em alguns textos essa variável independente também é chamada de covariável ou variável explicativa.

 Em geral quando existe apenas uma variável independente, para tentar estabelecer uma equação que representa o fenômeno em estudo pode-se fazer um gráfico, chamado diagrama de dispersão.

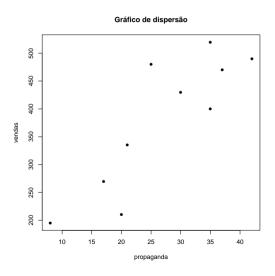


Exemplo

O objetivo do diretor de vendas de uma rede de varejo é analisar a relação entre o investimento realizado em propaganda e as vendas das lojas da rede, para realizar projeções de vendas de futuros investimentos em propaganda. A tabela a seguir registra uma amostra representativa extraída dos registros históricos das lojas de tamanho equivalente, com os valores de Propaganda e Vendas em milhões. Analise a possibilidade de definir um modelo que represente a relação entre as duas variáveis ou amostras.

```
propaganda <- c(30, 21, 35, 42, 37, 20, 8, 17, 35, 25)
vendas <- c(430, 335, 520, 490, 470, 210, 195, 270, 400, 480)

pdf("grafico-dispersao.pdf")
plot(propaganda, vendas, main = "Gráfico de dispersão", pch = 16)
dev.off()</pre>
```



 Quando o diagrama de dispersão apresenta os pontos agrupados em torno de uma reta imaginária, provavelmente existe uma relação linear entre as variáveis envolvidas

Exemplo

Um engenheiro civil coleta dados em um laboratório, a fim de estudar a dilatação de um pilar de concreto segundo a temperatura ambiente no local onde o pilar se encontra. Veja os dados:

T (°C)	18	16	25	22	20	21	23	19	27
Dilat. linear (mm)	5	3	10	8	6	7	9	6	5

- Posso realizar um estudo de regressão nestes dados?
- Qual modelo usar?
- Como montar a equação que relaciona a temperatura com a dilatação neste estudo?
- A temperatura realmente exerce influência na dilatação do pilar?
- É possível quantificar essa relação?

Diagrama de dispersão

```
temperatura <- c(18, 16, 25, 22, 20, 21, 23, 19, 17)
dilatacao <- c(5, 3, 10, 8, 6, 7, 9, 6, 5)
plot(temperatura, dilatacao, pch = 16, col = "#de5d83", cex = 2)
```

Modelo de regressão linear simples

```
dados <- data.frame(dilatacao, temperatura)
reg <- lm(dilatacao~temperatura, data = dados)
reg
```

```
Call:
lm(formula = dilatacao ~ temperatura, data = dados)

Coefficients:
(Intercept) temperatura
-8.1710 0.7323
```

- Com base no modelo ajustado temos duas informações:
 - I. o valor do intercepto (valor em que a reta de regressão intercepta o eixo das ordenadas),
 - II. e o valor que representa o coeficiente de inclinação da reta (o quanto a dilatação varia para cada variação unitária da temperatura).
- Logo podemos concluir que o modelo de regressão ajustado seria:

$$y = a + bx$$

 $y = -8.1710 + 0.7323x$

 Com o comando predict() podemos obter os valores calculados de dilatacao, de acordo com o modelo ajustado para os valores observados em temperatura:

```
preditos <- predict(reg)
preditos
```

```
1 2 3 4 5 6 7
5.009677 3.545161 10.135484 7.938710 6.474194 7.206452 8.670968
8 9
5.741935 4.277419
```

- Podemos também obter os resíduos associados a cada observação.
- Esses resíduos seriam simplesmente a diferença entre os valores observados e os calculados:

```
residuos <- resid(reg)
residuos
```

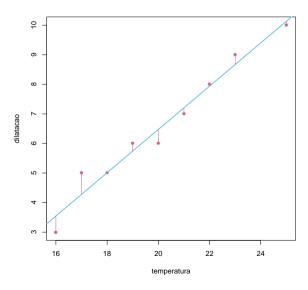
```
1 2 3 4 5

-0.009677419 -0.545161290 -0.135483871 0.061290323 -0.474193548

6 7 8 9

-0.206451613 0.329032258 0.258064516 0.722580645
```

 Podemos agora acrescentar ao diagrama de dispersão a reta de regressão ajustada e segmentos de reta representando os resíduos.



• Podemos usar a função anova apresentar alguns dados estatísticos da regressão:

```
1 anova(reg) #ANOVA da regressão
```

- Por meio dessa análise podemos verificar que o coeficiente b é significativo (p-value encontrado é da ordem de 10^{-6}), ou seja, a temperatura influencia significativamente a dilatação.
- Com o comando summary() podemos obter outras informações importantes:

| | summary(reg) |

```
Call:
lm(formula = dilatacao ~ temperatura, data = dados)
Residuals:
    Min 10 Median 30
                                    Max
-0.54516 -0.20645 -0.00968 0.25806 0.72258
Coefficients:
           Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.1710 1.0475 -7.801 0.000107 ***
temperatura 0.7323 0.0516 14.191 2.05e-06 ***
_ _ _
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0 4283 on 7 degrees of freedom
Multiple R-squared: 0.9664. Adjusted R-squared: 0.9616
F-statistic: 201.4 on 1 and 7 DF. p-value: 2.048e-06
```

Regressão Linear

- O valor do coeficiente de determinação (R²) é apresentado em Multiple R-squared: 0.9664 e representa o quanto a variação da dilatação pode ser explicada pela variação da temperatura neste experimento.
- Uma vez que o valor encontrado foi quase 97%, há indicação de que o modelo escolhido (linear) se ajusta bem aos dados.

Problema 1

Construa a reta de regressão linear dos retornos das ações PN em função dos retornos das ações ON, a partir das amostras registradas na tabela seguinte.

ON%												
PN%	20,9	5,4	49,4	31,1	30	28	367,1	6,9	45,4	27,8	43,1	13,4

Regressão Linear 20/24

Problema 2

Continuando com o problema anterior. Verifique se a regressão deve ser aceita aplicando o nível de significância de 5%.

Regressão Linear 21/24

Problema 3

Continuando com o Problema anterior. Identifique e remova possíveis *outliers*, e obtenha a nova reta de regressão linear.

Regressão Linear 22/24

Problema 4

Continuando com o problema anterior. Verifique novamente se a regressão deve ser aceita aplicando o nível de significância de 5%.

Problema 5

O gerente de vendas está sempre insistindo com os vendedores que a venda dos seus produtos tem forte relação com as visitas realizadas pelos vendedores aos seus clientes. A empresa tem onze vendedores e, como regra, eles visitam seus clientes uma vez por mês. Para confirmar a crença do gerente de vendas, foi preparada a tabela a seguir com as visitas realizadas e as vendas de cada vendedor durante o mês passado. Obtenha a reta de regressão das vendas em função das visitas.

	Visitas do mês	Vendas do mês
Samuel	42	140
Ricardo	105	330
Suely	66	190
Manoel	87	350
Ivany	50	110
Rafaela	55	135
Carlos	51	140
João	60	235
Susana	40	70
Marcos	87	320
Andréa	78	220

Regressão Linear

Problema 6

Calcule o coeficiente de determinação (R^2) e verifique se a regressão deve ser aceita considerando o nível de significância de 5%.