

EAD5917 – Modelagem e Métodos para Inferência e Predição aplicados à Administração

Adriana Backx Noronha Viana (backx@usp.br)

Cesar Alexandre de Souza (calesou@usp.br)

Área de MQI
Depto de Administração
FEA/USP

Métodos Regressivos

- Parte 1 – Regressão Linear Simples
 - Ferramentas de Exploração dos dados
 - Coeficientes de Regressão
 - Análise de Resíduos
- Parte 2 – Regressão Linear Múltipla
 - Ajuste do Modelo (coeficientes de regressão)
 - Análise de Resíduos
 - Multicolinearidade



FEAUSP

Objetivos da Técnica

Prever o comportamento de uma variável em função de uma outra variável

Qual deve ser o **limite de crédito** de um cliente dada a sua **renda mensal**, a sua **idade**, o seu **estado civil**, o seu **patrimônio declarado**?

Qual será a **receita de vendas** de um produto dado um determinado investimento em **ações de marketing**?





Determinar a influência de uma variável sobre outra

O índice de **satisfação** do cliente influencia seu nível de **lealdade** ao banco?

O **desempenho individual** da empresa influencia o **desempenho da cadeia de suprimentos**?





FEAUSP

Regressão Linear Simples

**Uma variável
dependente
métrica
(y)**

**Uma variável
independente
métrica ou
binária
(x)**



**y → variável que
desejamos prever**

**x → variável que se usa
para fazer a previsão**

**Receita de Vendas de
um novo produto**

**Investimento total em
ações de MKT**



Classificação das variáveis

Variável Métrica

Variáveis Numéricas

Ex: Faturamento, Valor de Contrato, Retorno sobre Investimento, Lucro

Variável Dummy

Variável Qualitativa

- Possui categorias, não é numérica
- Codificação 0 ou 1 (binária)

Ex: Gênero, Status do Cliente (Adimplente/Inadimplente), Contrato Regular/Irregular

Variável Dummy (binária)

Cliente	Gênero	Gên. Dummy	Status	Xstatus1	Xstatus2
1	F	0	Adimp	0	0
2	F	0	Inadimp	1	0
3	M	1	Insolv	0	1
4	M	1	Adimp	0	0
5	F	0	Inadimp	1	0
6	M	1	Adimp	0	0

**Quantidade de
Dummies**
**Número de
Categorias - 1**



FEAUSP

Regressão Simples

Variável
Independente (x)



Variável
Dependente (y)



Como **os anos de experiência** podem influenciar **o salário**?

Pode ser que o salário seja influenciado fortemente pelos anos de experiência e pode ser desmotivador para os funcionários.





FEAUSP

Modelo de Regressão

Previsões sobre y , baseiam-se nos valores correspondentes de x .

Com isso se cria uma **Equação de Regressão**

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

Diagram illustrating the components of the regression equation:

- Intercepto vertical** → onde a reta corta o eixo y (points to β_0)
- Variável que influencia y** → VI (points to x_1)
- Termo de erro** → quanto da variação de y não consegue ser explicada pela variável x (points to e)
- O que queremos prever** → VD (points to y)
- Coefficiente Angular** → inclinação da reta (points to β_1)

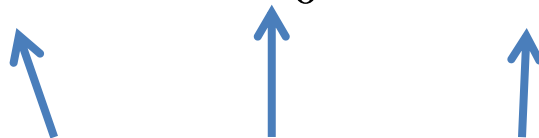




FEAUSP

Equação de Regressão

Como os valores reais de β_0 e β_1 são desconhecidos, faremos estimativas desses parâmetros com base nos dados amostrais

$$\hat{y} = b_0 + b_1 x_1$$


Parâmetros estimados
com base na amostra



Passos da Análise

- 1) Verificar se o tamanho da amostra é adequado
 - No mínimo de 5 a 10 observações por variável independente
 - Número recomendado de 20 observações por variável independente

Exemplo: Havendo 5 variáveis independentes, a amostra mínima deve conter de 25 a 50 casos

*No exemplo temos 20 observações (linhas) → tamanho adequado



Passos da Análise

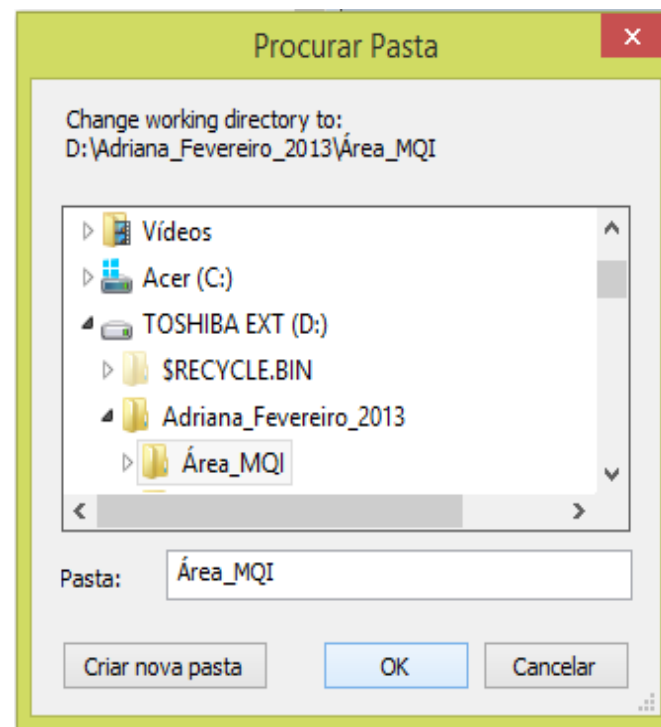
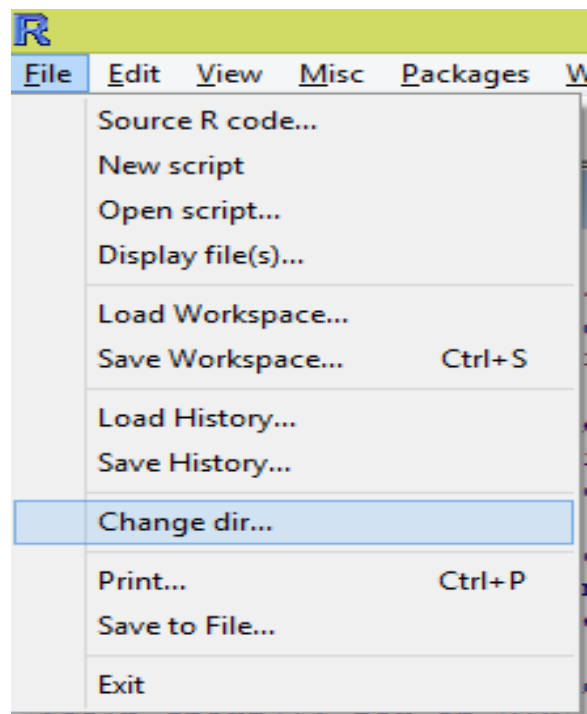
2) Examinar a relação entre as variáveis dependente e independente

- Regressão Simples → Diagrama de Dispersão e Correlação de Pearson
- Regressão Múltipla → Matriz de Correlação de Pearson entre as VI

Regressão Linear Simples

- Verificar diretório
- Carregar dados
 - *`dados <- read.csv("Base Dados EnANPAD.csv", sep=";", dec=",")`*
- Instalar pacote
 - *`install.packages("UsingR")`*
- Ferramentas para exploração de dados
 - *`plot(dados$Anos_de_Experiencia, dados$Salario, pch=19, col="blue")`*
 - *`cor(dados$Anos_de_Experiencia, dados$Salario)`*
 - *`cor.test(dados$Anos_de_Experiencia, dados$Salario)`*

Verificar diretório



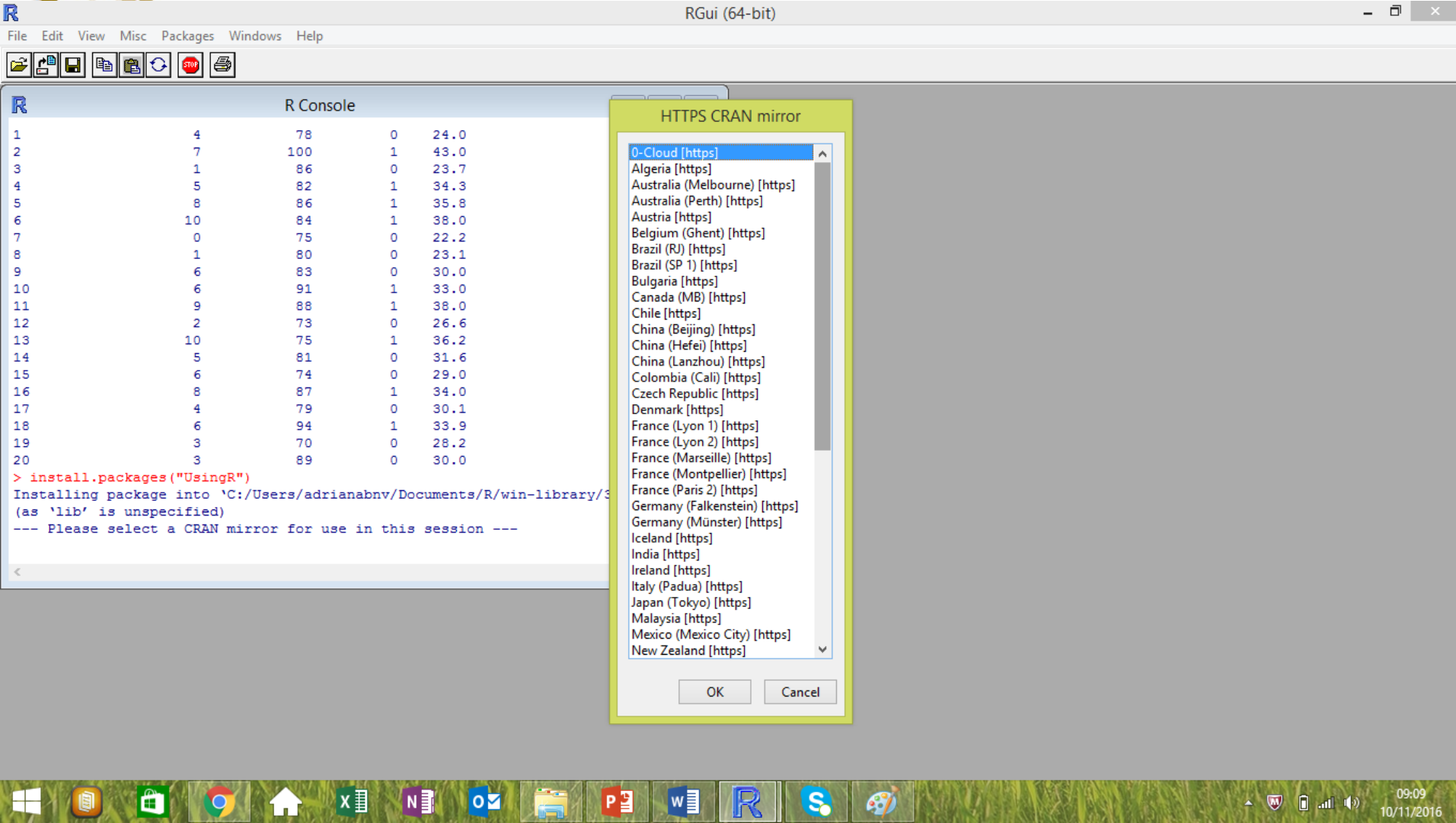
```
> dados <- read.csv("Base Dados EnANPAD.csv", sep=";", dec=",")
```

```
> dados
```

	Anos_de_Experiencia	Escore_teste	Graduacao	Salario
1	4	78	0	24.0
2	7	100	1	43.0
3	1	86	0	23.7
4	5	82	1	34.3
5	8	86	1	35.8
6	10	84	1	38.0
7	0	75	0	22.2
8	1	80	0	23.1
9	6	83	0	30.0
10	6	91	1	33.0
11	9	88	1	38.0
12	2	73	0	26.6
13	10	75	1	36.2
14	5	81	0	31.6
15	6	74	0	29.0
16	8	87	1	34.0
17	4	79	0	30.1
18	6	94	1	33.9
19	3	70	0	28.2
20	3	89	0	30.0

Carregar dados

```
> |  
< dados <- read.csv("Escore.csv", sep=";", dec=",")
```



The screenshot shows the RGui (64-bit) window. The R Console displays the output of the `install.packages("UsingR")` command, indicating the installation path and the need to select a CRAN mirror. The HTTPS CRAN mirror dialog is open, showing a list of mirrors with "0-Cloud [https]" selected.

R Console Output:

```
> install.packages("UsingR")
Installing package into 'C:/Users/adrianabnv/Documents/R/win-library/3
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
```

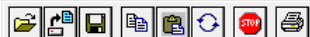
HTTPS CRAN mirror Dialog:

- 0-Cloud [https]
- Algeria [https]
- Australia (Melbourne) [https]
- Australia (Perth) [https]
- Austria [https]
- Belgium (Ghent) [https]
- Brazil (RJ) [https]
- Brazil (SP 1) [https]
- Bulgaria [https]
- Canada (MB) [https]
- Chile [https]
- China (Beijing) [https]
- China (Hefei) [https]
- China (Lanzhou) [https]
- Colombia (Cali) [https]
- Czech Republic [https]
- Denmark [https]
- France (Lyon 1) [https]
- France (Lyon 2) [https]
- France (Marseille) [https]
- France (Montpellier) [https]
- France (Paris 2) [https]
- Germany (Falkenstein) [https]
- Germany (Münster) [https]
- Iceland [https]
- India [https]
- Ireland [https]
- Italy (Padua) [https]
- Japan (Tokyo) [https]
- Malaysia [https]
- Mexico (Mexico City) [https]
- New Zealand [https]

The taskbar at the bottom shows various application icons, including Windows Explorer, Google Chrome, and the R logo.

Instalar pacote com a Regressão





R Console

1	4	78	0	24.0
2	7	100	1	43.0
3	1	86	0	23.7
4	5	82	1	34.3
5	8	86	1	35.8
6	10	84	1	38.0
7	0	75	0	22.2
8	1	80	0	23.1
9	6	83	0	30.0
10	6	91	1	33.0
11	9	88	1	38.0
12	2	73	0	26.6
13	10	75	1	36.2
14	5	81	0	31.6
15	6	74	0	29.0
16	8	87	1	34.0
17	4	79	0	30.1
18	6	94	1	33.9
19	3	70	0	28.2
20	3	89	0	30.0

```

> install.packages("UsingR")
Installing package into 'C:/Users/adrianabnv/Documents/R/win-library/3
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---

```

HTTPS CRAN mirror

- France (Paris 2) [https]
- Germany (Falkenstein) [https]
- Germany (Münster) [https]
- Iceland [https]
- India [https]
- Ireland [https]
- Italy (Padua) [https]
- Japan (Tokyo) [https]
- Malaysia [https]
- Mexico (Mexico City) [https]
- New Zealand [https]
- Norway [https]
- Philippines [https]
- Russia (Moscow) [https]
- Serbia [https]
- Spain (A Coruña) [https]
- Spain (Madrid) [https]
- Switzerland [https]
- Taiwan (Chungli) [https]
- Turkey (Denizli) [https]
- UK (Bristol) [https]
- UK (Cambridge) [https]
- UK (London 1) [https]
- USA (CA 1) [https]
- USA (IA) [https]
- USA (IN) [https]
- USA (KS) [https]
- USA (MI 1) [https]
- USA (TN) [https]
- USA (TX) [https]
- USA (WA) [https]**
- (HTTP mirrors)

OK Cancel



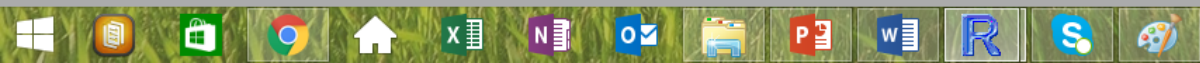
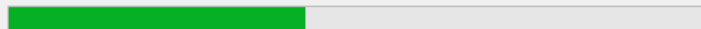
R Console

3	1	86	0	23.7
4	5	82	1	34.3
5	8	86	1	35.8
6	10	84	1	38.0
7	0	75	0	22.2
8	1	80	0	23.1
9	6	83	0	30.0
10	6	91	1	33.0
11	9	88	1	38.0
12	2	73	0	26.6
13	10	75	1	36.2
14	5	81	0	31.6
15	6	74	0	28.0
16	8	87	1	34.3
17	4	79	0	22.2
18	6	94	1	38.0
19	3	70	0	23.1
20	3	89	0	30.0

```
> install.packages("UsingR")
Installing package into 'C:/Users/adrianabnv/Docum
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
trying URL 'https://cran.fhcrc.org/bin/windows/contrib/3.2/UsingR_2.0-5.zip'
Content type 'application/zip' length 2079941 bytes (2.0 MB)
```

46% downloaded

URL: ... ps://cran.fhcrc.org/bin/windows/contrib/3.2/UsingR_2.0-5.zip





```
R Console

Installing package into 'C:/Users/adrianabnv/Documents/R/win-library/3.2'
(as 'lib' is unspecified)
--- Please select a CRAN mirror for use in this session ---
trying URL 'https://cran.fhcrc.org/bin/windows/contrib/3.2/UsingR_2.0-5.zip'
Content type 'application/zip' length 2079941 bytes (2.0 MB)
downloaded 2.0 MB

package 'UsingR' successfully unpacked and MD5 sums checked

The downloaded binary packages are in
      C:\Users\adrianabnv\AppData\Local\Temp\RtmpsJJiSw\downloaded_packages
> library(UsingR)
Loading required package: MASS
Loading required package: HistData
Loading required package: Hmisc
Loading required package: lattice
Loading required package: survival
Loading required package: Formula
Loading required package: ggplot2

Attaching package: 'Hmisc'

The following objects are masked from 'package:base':

    format.pval, round.POSIXt, trunc.POSIXt, units

Attaching package: 'UsingR'

The following object is masked from 'package:survival':

    cancer

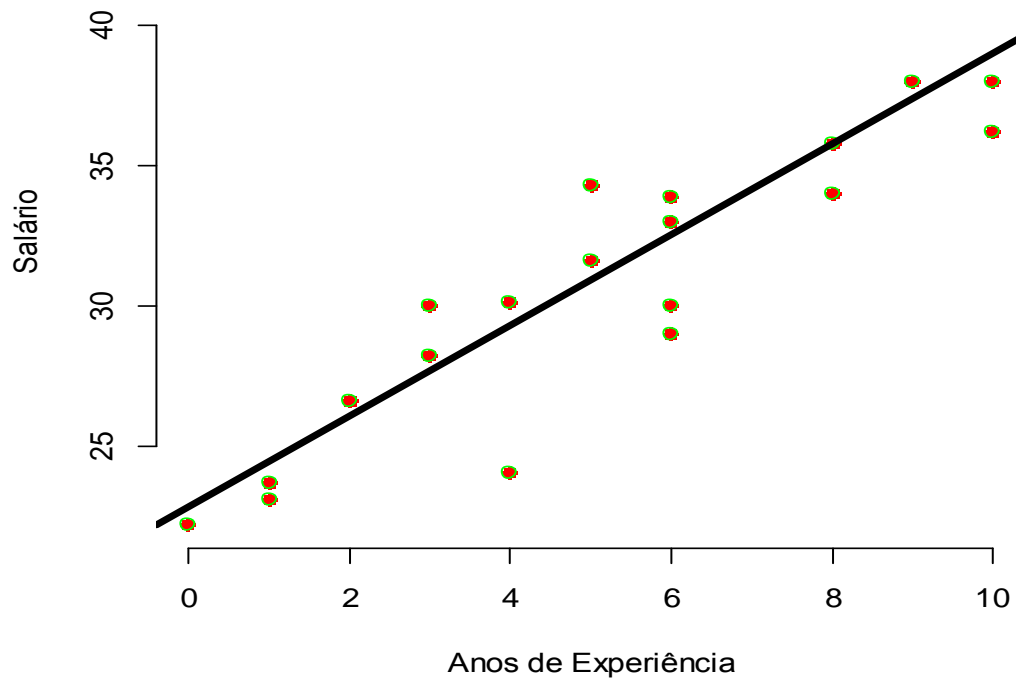
> |
```





FEAUSP

Exemplo – Gráfico de Dispersão



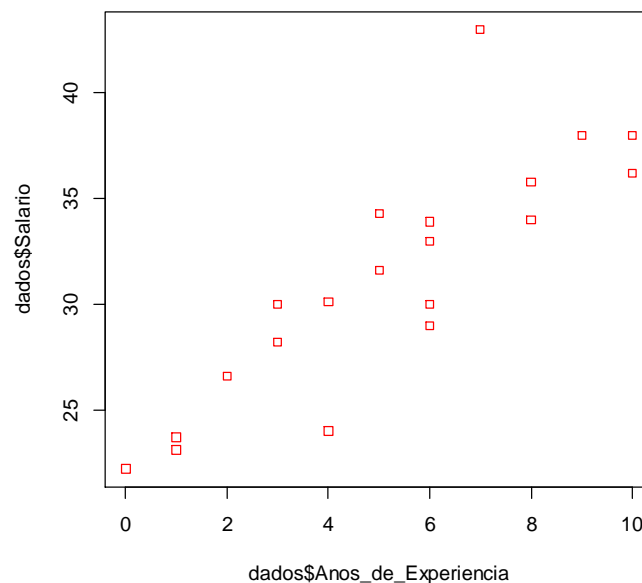
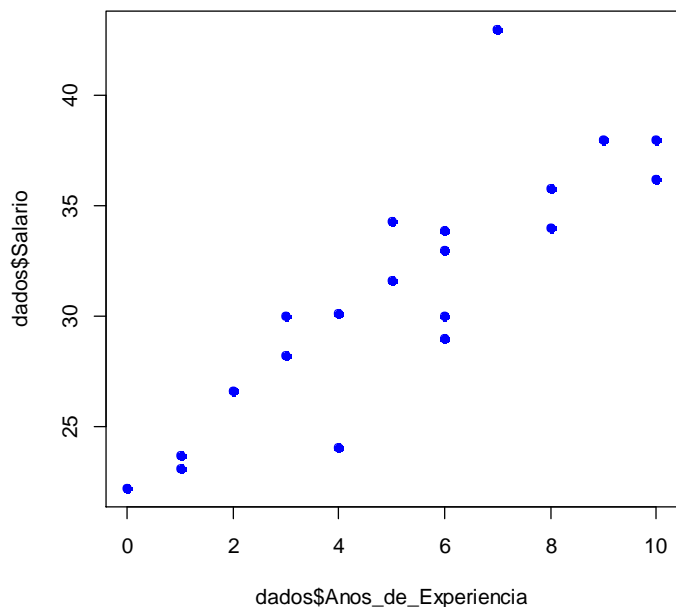


FEAUSP

Exemplo – Gráfico de Dispersão

Ferramentas para exploração de dados

```
plot(dados$Anos_de_Experiencia,dados$Salario,pch=19,col="blue")
```





FEAUSP

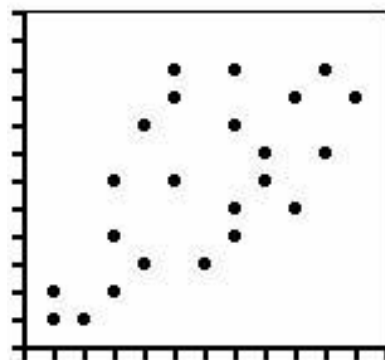
Aspectos das definições das funções

- Possibilidades do pch:
 - pch = 19: solid circle,
 - pch = 20: bullet (smaller solid circle, 2/3 the size of 19),
 - pch = 21: filled circle,
 - pch = 22: filled square,
 - pch = 23: filled diamond,
 - pch = 24: filled triangle point-up,
 - pch = 25: filled triangle point down.
- Definições em plot
 - bg – cor dos pontos
 - col – cor do envoltório dos pontos
 - cex – tamanho dos pontos
 - pch – tipo dos pontos
 - frame = FALSE significa que os eixos não vão se cruzar (fica aberto)
 - lwd – espessura da reta de regressão

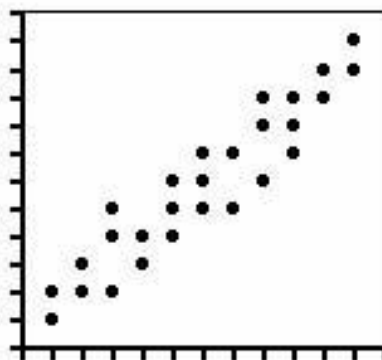




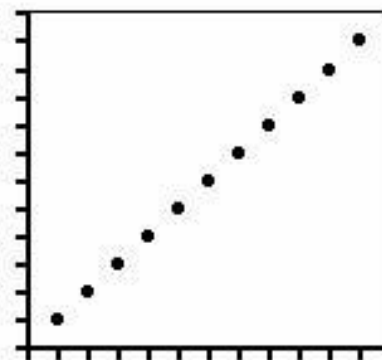
Diagramas de dispersão que mostram correlação positiva entre as variáveis



Correlação fraca

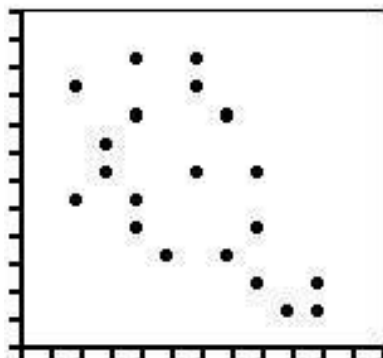


Correlação forte

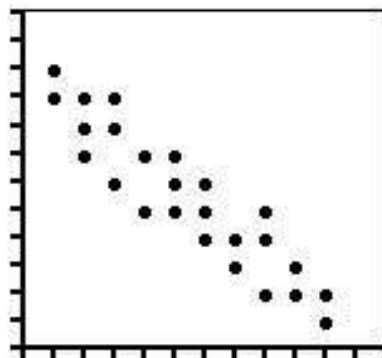


Correlação perfeita

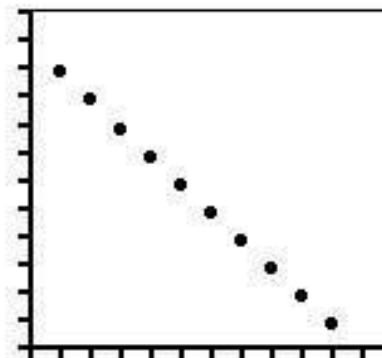
Diagramas de dispersão que mostram correlação negativa entre as variáveis



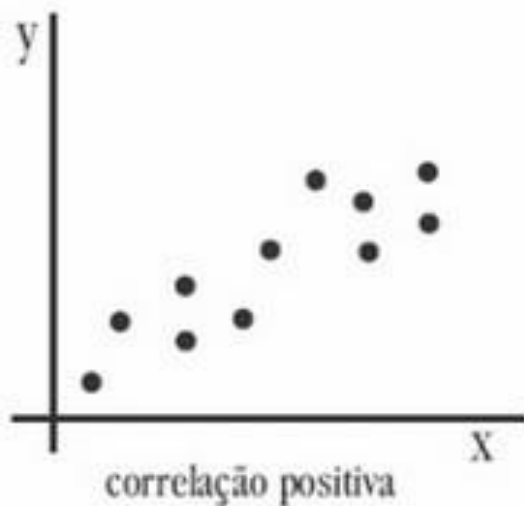
Correlação fraca



Correlação forte



Correlação perfeita




```
#Colocando a linha de regressão no gráfico  
plot(dados$Anos_de_Experiencia, dados$Salario,  
xlab = "Anos de Experiência",  
ylab = "Salário",  
bg = "red", col = "green", cex = 1.1, pch = 21, frame = FALSE)  
abline(lm(Salario ~ Anos_de_Experiencia, data = dados), lwd = 3)
```

observar que:

bg – cor dos pontos

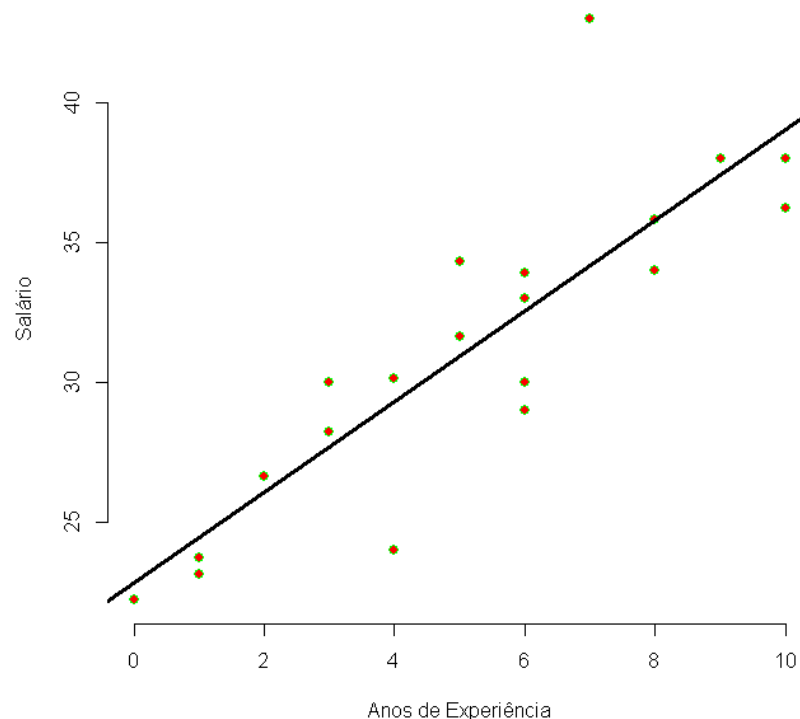
col – cor do envoltório dos pontos

cex – tamanho dos pontos

pch – tipo dos pontos

frame = FALSE significa que os eixos não
vão se cruzar (fica aberto)

lwd – espessura da reta de regressão





Correlação de Pearson

$$r_{xy} = \frac{n \sum XY - \sum X \sum Y}{\sqrt{[n \sum X^2 - (\sum X)^2] * [n \sum Y^2 - (\sum Y)^2]}}$$

Coeficiente de Correlação varia de -1 a +1

- (+) variáveis se movem na mesma direção
- (-) variáveis se movem em direções opostas
- Quanto mais próximo de $|+1|$ mais forte a relação

$r = 0,8553 \rightarrow$ indica correlação alta e positiva entre as variáveis



FEAUSP

```
> cor(dados$Anos_de_Experiencia,dados$Salario)
[1] 0.8553203
```

```
> cor.test(dados$Anos_de_Experiencia,dados$Salario)
```

Pearson's product-moment correlation

```
data: dados$Anos_de_Experiencia and dados$Salario
t = 7.0041, df = 18, p-value = 1.541e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.6641954 0.9414899
sample estimates:
   cor
0.8553203
```

Teste de hipótese em
correlação;
Hipótese Nula e
Hipótese Alternativa





Passos da Análise

3) Estimação dos parâmetros b_0 e b_1

- Método dos Mínimos Quadrados Ordinários → minimiza a soma dos quadrados dos desvios entre o valor estimado de y e o valor observado de y .



Passos da Análise

$$b_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$b_0 = \bar{y} - b_1 \bar{x}$$

x_i = valor da VI para i – ésima observação

y_i = valor da VD para i – ésima observação

\bar{x} = valor médio da VI

\bar{y} = valor médio da VD

n = tamanho amostra

Regressão Linear Simples

- Ajuste
 - `equation<-lm(dados$Salario~dados$Anos_de_Experiencia)`
 - `coef(equation)` # valor dos coeficientes
- Função `lm` – primeiro a variável dependente (y) e depois a variável independente(x)
- Para visualizar resumo das informações
 - `summary(equation)`
- Para obter os intervalos de confiança
 - `confint(equation)`

Calculando os coeficientes de Regressão

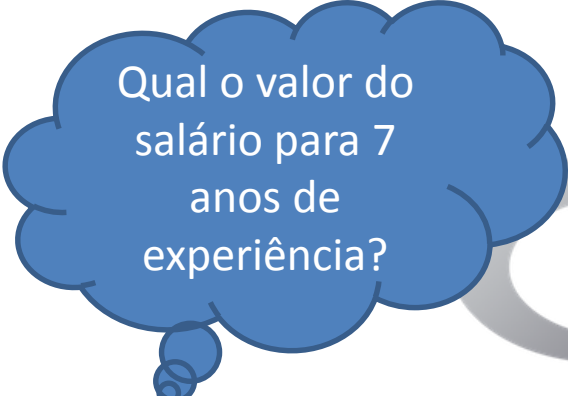
```
equation <- lm(dados$Salario ~ dados$Anos_de_Experiencia)  
coef(equation)
```

(Intercept) dados\$Anos_de_Experiencia
22.811124 1.619976

Significa que a cada ano de experiência aumenta 1,619976 o salário.

Função lm – primeiro a variável dependente (y) e depois a variável independente(x); ou ainda, primeiro a variável resposta e depois a preditora.

$$\hat{y} = 22,8111 + 1,6199x_1$$



Qual o valor do
salário para 7
anos de
experiência?



Passos da Análise

4) Avaliação da Qualidade do Modelo de Regressão

- R-Quadrado (coeficiente de determinação) → quanto da variação da VD pode ser explicado pela VI (Quanto maior, melhor)
- Significância da correlação entre as variáveis (ANOVA)
- Significância estatística de b_0 e b_1 (teste t)
- Distribuição dos resíduos da análise → devem seguir distribuição normal



FEAUSP

Para visualizar resumo das informações da análise de regressão:

summary(equation)

Call:

lm(formula = dados\$Salario ~ dados\$Anos_de_Experiencia)

Residuals:

Min	1Q	Median	3Q	Max
-5.291	-1.441	0.249	0.719	8.849

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.8111	1.3761	16.576	2.39e-12 ***
dados\$Anos_de_Experiencia	1.6200	0.2313	7.004	1.54e-06 ***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 2.991 on 18 degrees of freedom

Multiple R-squared: 0.7316, Adjusted R-squared: 0.7167

F-statistic: 49.06 on 1 and 18 DF, p-value: 1.541e-06



ANOVA

Testa se existe relação linear significativa entre as variáveis

$$H_0 : \beta_1 = 0 (\text{não há regressão})$$

$$H_1 : \beta_1 \neq 0 (\text{há regressão})$$

Quanto maior o valor da estatística F, melhor o ajuste do modelo



Significância de b_0 e b_1

Teste t-student → avalia se os parâmetros são significantes

- Avalia se o intercepto deve ser mantido na equação de regressão
- Avalia se a relação entre a VI e a VD é significativa

$$H_0 : \beta_0 = 0$$

$$H_0 : \beta_1 = 0$$

$$H_1 : \beta_0 \neq 0$$

$$H_1 : \beta_1 \neq 0$$



FEAUSP

Para visualizar resumo das informações da análise de regressão:

summary(equation)

Call:

lm(formula = dados\$Salario ~ dados\$Anos_de_Experiencia)

Residuals:

Min 1Q Median 3Q Max
-5.291 -1.441 0.249 0.719 8.849

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	22.8111	1.3761	16.576	2.39e-12 ***
dados\$Anos_de_Experiencia	1.6200	0.2313	7.004	1.54e-06 ***

Signif. codes: 0 '*' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1**

Residual standard error: 2.991 on 18 degrees of freedom

Multiple R-squared: 0.7316, Adjusted R-squared: 0.7167

F-statistic: 49.06 on 1 and 18 DF, p-value: 1.541e-06



Para obter os intervalos de confiança:

confint(lm(y~x))

considerando que y e x já foram definidos anteriormente.

Existem várias formas de fazer isso, por exemplo, definindo:

equation = lm(dados\$Salario~dados\$Anos_de_Experiencia)
confint(equation)

	2.5 %	97.5 %
(Intercept)	19.919989	25.702260
dados\$Anos_de_Experiencia	1.134054	2.105898



FEAUSP

Regressão Linear Simples

Análise de Resíduos

- Resíduo: diferença entre valor observado e valor previsto ($Y_i - \hat{Y}_i$)
 - *y <- dados\$Salario*
 - *x <- dados\$Anos_de_Experiencia*
 - *n <- length(y)*
 - *equation <- lm(y ~ x)*
 - *e <- resid(equation)*
 - *ychapeu <- predict(equation)*
 - *y[1]*
 - *ychapeu[1]*
 - *e[1]*

Análise de Resíduos



```
y <- dados$Salario
x <- dados$Anos_de_Experiencia
n <- length(y)
fit <- lm(y ~ x)
e <- resid(fit)
ychapeu <- predict(fit)
y[1]
[1] 24
ychapeu[1]
1
29.29103
e[1]
1
-5.291029
>
```



FEAUSP

Regressão Linear Simples

Análise de Resíduos

- Análise de Resíduos: verificar ajuste do modelo: suposição de que os erros possuem variância constante e não são correlacionados entre si;
- Gráfico de resíduos:
 - *plot(x,e,xlab="Anos de Experiência",ylab="Resíduos")*
 - *abline(h=0)*





FEAUSP

Regressão Linear Simples

Análise de Resíduos

- Gráficos:
 - `par(mfrow=c(1,2))`
 - `plot(fitted(equation),residuals(equation),xlab="Valores Esperados Salários",ylab="Resíduos")`
 - `abline(h=0)`
 - `plot(dados$Anos_de_Experiencia,residuals(equation),xlab="Anos de Experiência", ylab="Resíduos")`
 - `abline(h=0)`

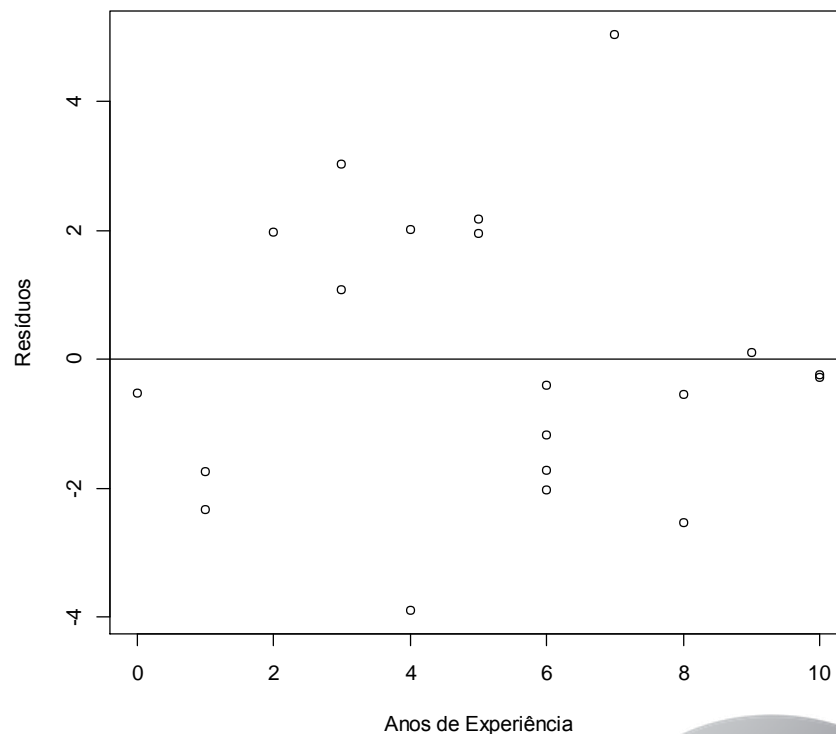
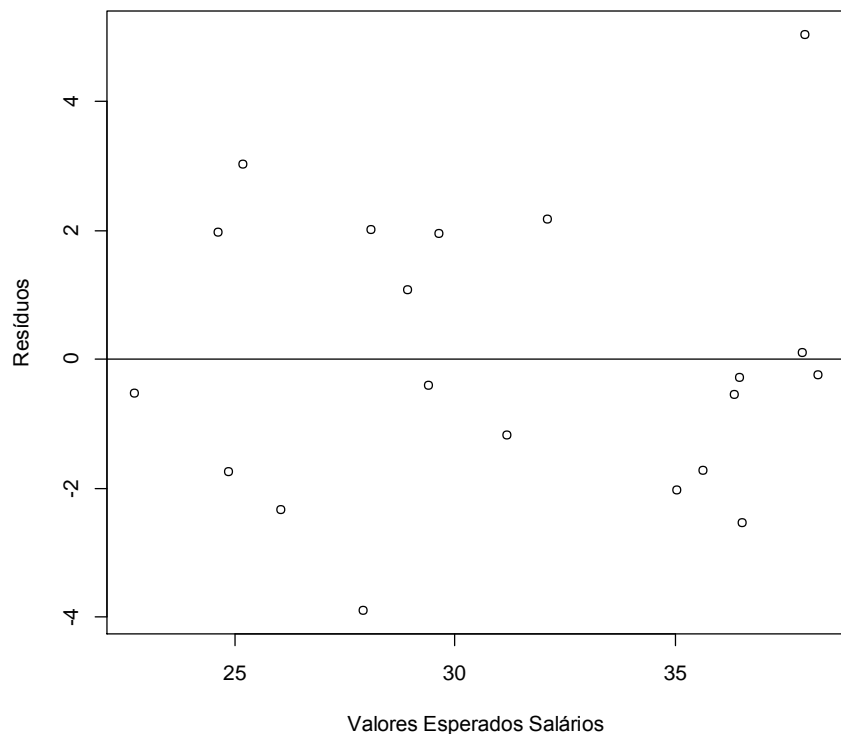




FEAUSP

Regressão Linear Simples

Análise de Resíduos



Será que houve violação da suposição de que os erros possuem variância constante, ou seja, a homocedasticidade?





FEAUSP

Regressão Linear Simples

Análise de Resíduos

- Dividir os dados em dois subgrupos e aplicar um teste para comparar as variâncias de cada subconjunto.
- Considere o valor da mediana
 - *median(dados\$Anos_de_Experiencia)*
- Teste de comparação
 - *var.test(residuals(equation)[dados\$Anos_de_Experiencia>5.5],residuals(equation)[dados\$Anos_de_Experiencia<5.5])*
 - Se nível de significância for 5%? Qual a conclusão?
 - E se for 1%?





FEAUSP

Regressão Linear Simples

Análise de Resíduos

- Gráfico de Probabilidade Normal dos Resíduos: utilizado para analisar a suposição de normalidade dos erros
 - `qqnorm(residuals(equation),
ylab="Resíduos",xlab="Quantis teóricos",main="")`
 - `qqline(residuals(equation))`
- Dificuldade de analisar o gráfico; utilização do teste de shapiro
 - `shapiro.test(residuals(equation))`





FEAUSP

Regressão Linear Simples

Análise de Resíduos

- Construindo gráfico com intervalo de confiança
 - require(ggplot2)
 - dados1 <- data.frame(dados\$Anos_de_Experiencia,dados\$Salario)
 - # O ggplot2 exige que os dados estejam em um data.frame
 - p <- ggplot(dados1, aes(x=x, y=y)) + # Informa os dados a serem utilizados
 - geom_point() + # Informa que eu quero um gráfico de dispersão
 - xlab("Anos de Experiência") +
 - ylab("Salário")
 - p
 - p1 <- p + geom_smooth(method=lm) # Acrescenta a linha de tendência e o intervalo de confiança de predição
 - p1

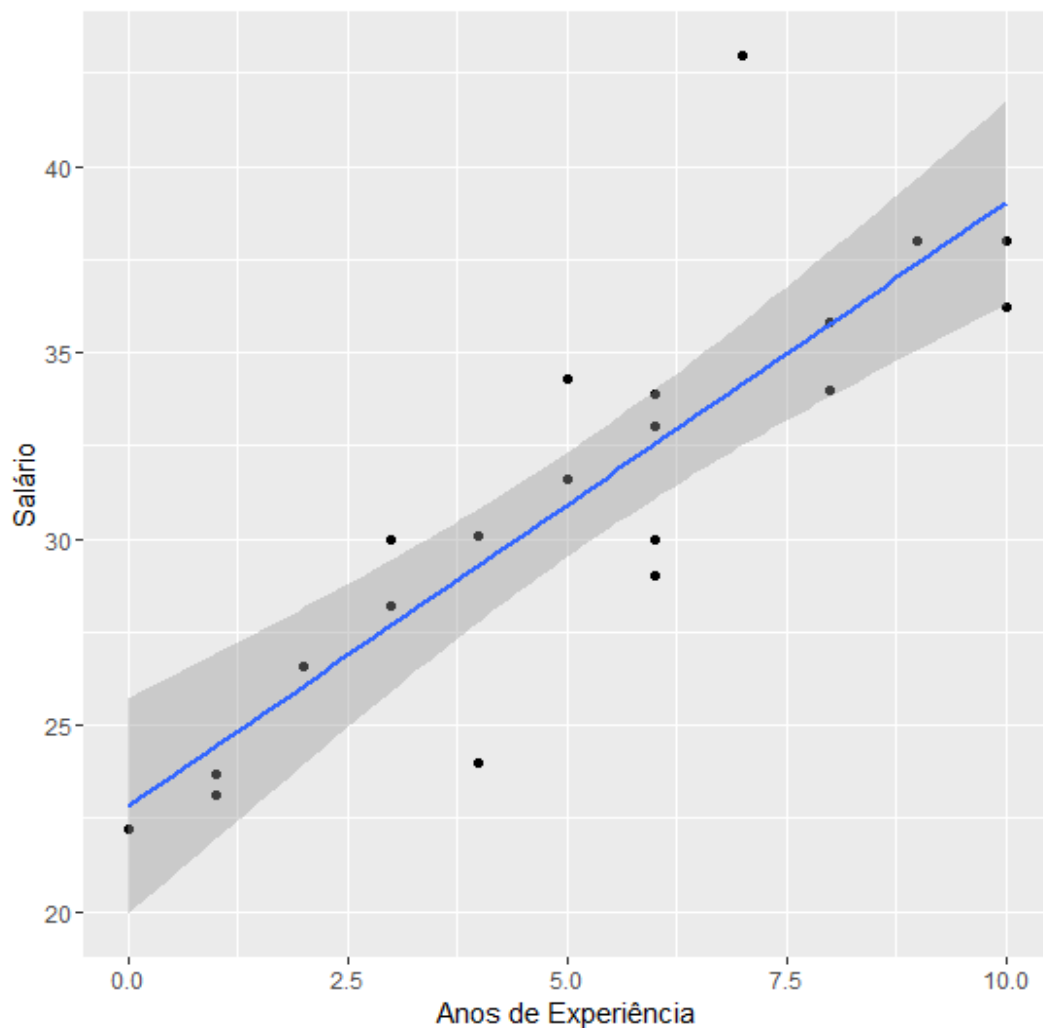




FEAUSP

Regressão Linear Simples

Análise de Resíduos





FEAUSP

Regressão Múltipla

Problemas reais são complexos e dificilmente podem ser resolvidos com modelos de Regressão Simples

Raramente uma variável dependente pode ser explicada por apenas uma variável independente

Quando incluimos mais de uma variável independente no modelo, temos um modelo de Regressão Múltipla

Princípios da Regressão Simples são mantidos



Modelo de Regressão Múltipla

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + e$$

Equação de Regressão Múltipla

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_n x_n$$

b_0, b_1, b_2, b_n são estimadores de $\beta_0, \beta_1, \beta_2, \beta_n$

\hat{y} = valor estimado de VD



FEAUSP

Regressão Linear Múltipla

- Exemplo estudado
- As variáveis a serem consideradas são:
 - Anos de Experiência
 - Escore em um teste (conhecimento/habilidades)~
 - Se tem graduação ou não
 - Valor do Salário
- Variável categórica
 - `dados$Graduacao=factor(dados$Graduacao)`





Regressão Linear Múltipla

- Construção de gráfico
 - `par(mfrow=c(1,1))`
 - `plot(dados$Escore_teste[dados$Graduacao=="Não"], dados$Salario[dados$Graduacao=="Não"], xlab="Escore Teste", ylab="Salário")`
 - `points(dados$Escore_teste[dados$Graduacao=="Sim"], dados$Salario[dados$Graduacao=="Sim"], pch=19)`
- Observar quantos dados apareceram. Qual origem dos erros?
 - Ver dados usando: `summary(dados)`

Observar as
aspas no R
(diferente do
word)





FEAUSP

Regressão Linear Múltipla

- Construção de gráfico (Salário, Escore em teste e se tem graduação ou não)
 - `par(mfrow=c(1,2))`
 - `plot(c(65,100), c(20,45), type="n", xlab="Escore Teste", ylab="Salário")`
 - `points(dados$Escore_teste[dados$Graduacao=="Não"], dados$Salario[dados$Graduacao=="Não"])`
 - `points(dados$Escore_teste[dados$Graduacao=="Sim"], dados$Salario[dados$Graduacao=="Sim"], pch=19)`





Regressão Linear Múltipla

- Construção de gráfico (Salário, Escore em teste e se tem graduação ou não)
 - `plot(c(0,11), c(20,45), type="n", xlab="Anos de Experiência", ylab="Salário")`
 - `points(dados$Anos_de_Experiencia[dados$Graduacao == "Não"], dados$Salario[dados$Graduacao == "Não"])`
 - `points(dados$Anos_de_Experiencia[dados$Graduacao == "Sim"], dados$Salario[dados$Graduacao == "Sim"], pch=19)`
- Observação: `par(mfrow=c(1,2))`





Regressão Linear Múltipla

- Ajuste do Modelo
 - $\text{Salário} = \beta_0 + \beta_1 * \text{Anos_de_Experiencia} + \beta_2 * \text{Escore_teste} + \alpha_1 * \text{Graduação} + \text{erro}$
 - *equation = lm(dados\$Salario ~ dados\$Anos_de_Experiencia + dados\$Escore_teste + dados\$Graduacao)*
 - *equation*
- $\text{Salário} = 7.9448 + 1.1476 * \text{Anos_de_Experiencia} + 0.1969 * \text{Escore_teste} + 2.2804 * \text{Graduação}$
 - *summary(equation)*
 - *anova(equation)*
- Quais variáveis foram significativas e quais não foram? Tem algo errado?





Regressão Linear Múltipla

- Gráficos para analisar dados
 - `par(mfrow=c(2,2))`
 - `boxplot(dados$Salario~ dados$Graduacao, xlab="Graduação", ylab="Salário")`
 - `boxplot(dados$Anos_de_Experiencia~ dados$Graduacao, xlab="Graduação", ylab="Anos de Experiencia")`
 - `boxplot(dados$Escore_teste~ dados$Graduacao, xlab="Graduação", ylab="Escore Teste")`
 - `plot(dados$Anos_de_Experiencia, dados$Escore_teste, xlab="Anos de Experiencia", ylab="Escore Teste")`



Regressão Linear Múltipla

- Multicolinearidade
- Formas de observação
 - `explicativas = dados[,1:3]`
 - `explicativas #mostra as variáveis`
 - `cor(explicativas)`
 - `pairs(explicativas)`
- Métodos gráficos são limitados
- Análise do VIF (Variance Factor Inflation)
 - `equation = lm(dados$Salario ~ dados$Anos_de_Experiencia + dados$Escore_teste + dados$Graduacao)`
 - `vif(equation)`
 - `sqrt(vif(equation)) > 2`



Multicolinearidade

A multicolinearidade aumenta o termo de erro, pois inflaciona a relevância da VI

- Estatística VIF
- Medida de quanto a variância de cada coeficiente de regressão estimado aumenta devido à multicolinearidade
- Os valores de VIF devem ser até 5

`dados$Anos_de_Experiencia`
2.578462

`dados$Escore_teste`
1.550574

`dados$Graduacao`
3.401613

Não há problema de multicolinearidade, $VIF < 5$



Regressão Linear Múltipla

- Gráficos – Análise de Resíduos
 - `windows()`
 - `par(mfrow=c(2,3))`
 - `plot(fitted(equation),residuals(equation),xlab="Valores Ajustados (Esperados)", ylab="Resíduos")`
 - `abline(h=0)`
 - `plot(dados$Anos_de_Experiencia,residuals(equation),xlab="Anos de Experiencia",ylab="Resíduos")`
 - `abline(h=0)`
 - `plot(dados$Escore_teste,residuals(equation),xlab="Escore Teste",ylab="Resíduos")`
 - `abline(h=0)`
 - `boxplot(residuals(equation)~ dados$Graduacao)`
 - `qqnorm(residuals(equation), ylab="Resíduos")`
 - `qqline(residuals(equation))`



FEAUSP

Regressão Linear Múltipla

- Verificação de Distribuição Normal (aplicados aos erros)
 - `shapiro.test(residuals(equation))`
- Verificação de Homocedasticidade
 - `var.test(residuals(equation)[dados$Graduacao==0], residuals(equation)[dados$Graduacao==1])`
 - `var.test(residuals(equation)[dados$Escore_teste<82.5], residuals(equation)[dados$Escore_teste>82.5])`



Observações

Os modelos de regressão só acomodam previsões para valores de x incluídos no intervalo de dados de entrada

O valor dos coeficientes angulares na equação de regressão devem ter os mesmos sinais do coeficiente de correlação entre VI e VD

Amostras muito grandes (com dezenas de milhares de casos) podem provocar viés na análise, pois as correlações tendem a ter significância estatística, ainda que a correlação realmente não exista

Observações

Quando a ANOVA não rejeita H_0 , o modelo de regressão apresenta problemas, deve-se verificar quais são as variáveis com problemas e retirá-las ou realizar transformações

Os valores de sig do teste t-student devem ser verificados individualmente. Havendo coeficientes não significantes, deve-se retirá-los um a um, rodando novamente o modelo

Havendo multicolinearidade, deve-se retirar uma das variáveis, de acordo com o julgamento do analista.



FEAUSP

Exercício Regressão Múltipla

- Somos gestores de uma fábrica de chocolates e queremos avaliar se a elevação da produção da fábrica (ton) e o aumento na produção de embalagens para chocolates (emb) possuem influência sobre os custos indiretos da empresa (ci) – Base de Dados FabricaChocolates
- Rodar a regressão múltipla e avaliar os resultados
- Rodar o modelo novamente com as variáveis dummy ferias, ano 2007 e ano 2008

$$y = b_0 + b_1x_1 + gD_1 + e$$

g= coeficiente angular da variável dummy e representa o acréscimo da influência do critério definido com o valor 1 em relação à influência da categoria definida como 0

