**ANÁLISE PREDITIVA E DATA MINING**

**Turma 41BDT**

Trabalho em Grupo 2

**Parte 1) Modelo de regressão linear**

1. Modelo de Regressão Linear Simples

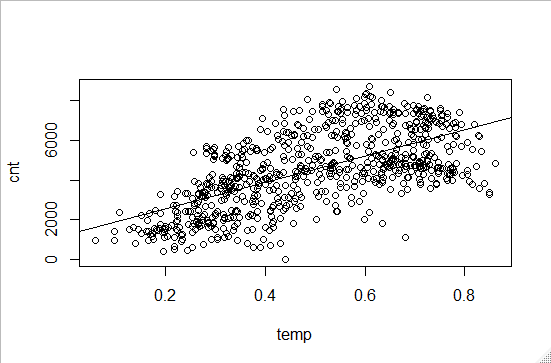
modelo1 <- lm(cnt ~ temp)

summary(modelo1)

Interprete os resultados

|  |
| --- |
| 1. Escreva o modelo matemático teórico  Y = a + bx 2. Escreva o modelo matemático ajustado  cnt = 1214.6 + 6640.7 \* temp 3. As estimativas dos coeficientes são significativas?   Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 1214.6 161.2 7.537 1.43e-13 \*\*\*  temp 6640.7 305.2 21.759 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Sim, o p-value do modelo é de 2.2e-16, menor do que 0.05, portanto é significativo.   1. Qual o coeficiente de determinação do modelo ajustado?   Multiple R-squared: 0.3937, Adjusted R-squared: 0.3929  R2 múltiplo -> regressão linear simples (uma variável preditora)  R2 ajustado -> regressão linear simples (duas ou mais variáveis) |
|  |

plot(cnt~temp)  
abline(modelo1)



É necessário mais variáveis para explicar a regressão.

## valores estimados pelo modelo

fitted(modelo1)

1 2 3 4 5 6 7 8 9 10

3500.155 3628.394 2518.638 2542.784 2721.798 2571.658 2519.688 2310.359 2133.271 2216.280

11 12 13 14 15 16 17 18 19 20

2337.526 2361.672 2310.359 2282.933 2764.139 2753.075 2382.298 2653.465 3154.885 2952.297

21 22 23 24 25 26 27 28 29 30

2393.368 1607.310 1855.615 1861.389 2698.695 2658.997 2509.581 2565.881 2519.688 2652.502

31 32 33 34 35 36 37 38 39 40

2415.502 2490.814 2941.227 2456.169 2617.851 2764.139 3112.776 3018.704 2681.130 2109.697

41 42 43 44 45 46 47 48 49 50

2173.215 2470.341 2692.200 3316.573 3970.537 2981.649 3328.121 4108.883 4678.881 3865.394

Valores estimados pelo modelo.

## resíduos do modelo

residuals(modelo1)

1 2 3 4 5 6

-2515.1553571 -2827.3941079 -1169.6384972 -980.7841187 -1121.7977382 -965.6579258

7 8 9 10 11 12

-1009.6877294 -1351.3592688 -1311.2714553 -895.2803302 -1074.5264134 -1199.6720349

13 14 15 16 17 18

-904.3592688 -861.9331365 -1516.1389051 -1549.0754823 -1382.2980802 -1970.4648323

19 20 21 22 23 24

-1504.8849221 -1025.2967822 -850.3681438 -626.3099575 -869.6147373 -445.3894987

25 26 27 28 29 30

-713.6947081 -2152.9965437 -2078.5805687 -1398.8805081 -1421.6877294 -1556.5019293

31 32 33 34 35 36

-914.5016302 -1130.8139223 -1415.2267187 -906.1693382 -909.8507046 -1759.1389051

37 38 39 40 41 42

-1489.7761800 -1306.7038822 -1151.1300301 -504.6969348 -635.2153259 -724.3406134

43 44 45 46 47 48

-1220.2000937 -1727.5729292 -2057.5367684 -1166.6487204 -1213.1211239 -1633.8826798

Margem de erro da predição.

## resíduos padronizados do modelo

rstandard(modelo1)

1 2 3 4 5 6

-1.6682626336 -1.8751556031 -0.7768625676 -0.6513983952 -0.7448216075 -0.6413185801

7 8 9 10 11 12

-0.6706235145 -0.8979239785 -0.8716158273 -0.5949948223 -0.7139402262 -0.7970512106

13 14 15 16 17 18

-0.6009104250 -0.5727523026 -1.0065759389 -1.0284612703 -0.9183486418 -1.3084483949

19 20 21 22 23 24

-0.9985433610 -0.6805059986 -0.5649413716 -0.4168540493 -0.5784182831 -0.2962435434

25 26 27 28 29 30

-0.4738784341 -1.4296413747 -1.3805962826 -0.9290426588 -0.9442693952 -1.0335661916

Valor de desvio padrão do modelo para cada item. Acima ou abaixo de 2 são os outliers.

1. Regressão linear múltipla

#Variáveis preditoras qualitativas

# criar variáveis dicotômicas

arquivo$season1 =season

arquivo$season1 =ifelse(season==1,"1","0")

arquivo$season2 =season

arquivo$season2 =ifelse(season==2,"1","0")

arquivo$season3 =season

arquivo$season3 =ifelse(season==3,"1","0")

arquivo$season4 =season

arquivo$season4 =ifelse(season==4,"1","0")

arquivo$weathersit1 =weathersit

arquivo$weathersit1 =ifelse(weathersit==1,"1","0")

arquivo$weathersit2 =weathersit

arquivo$weathersit2 =ifelse(weathersit==2,"1","0")

arquivo$weathersit3 =weathersit

arquivo$weathersit3 =ifelse(weathersit==3,"1","0")

modelo2 <- lm(cnt ~temp+ arquivo$season2+arquivo$season3+arquivo$season4+mnth+arquivo$weathersit1+arquivo$weathersit2)

summary(modelo2)

Interprete os resultados

|  |
| --- |
| 1. Escreva o modelo matemático teórico  Y = a + b1\*x1 + b2\*x2 + ... + bp\*xp 2. Escreva o modelo matemático ajustado  cnt = -1605,57 + 5807,22\*temp + 973,84\*season2 + 688,88\*season3 + 1647,20\*season4 + -20,13\*mnth + 2835,33\*weather1 + 2205,18\*weather2 3. As estimativas dos coeficientes são significativas?   Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -1605.57 342.76 -4.684 3.36e-06 \*\*\*  temp 5807.22 484.89 11.976 < 2e-16 \*\*\*  season21 973.84 187.76 5.187 2.78e-07 \*\*\*  season31 688.88 268.18 2.569 0.0104 \*  season41 1647.20 251.35 6.553 1.07e-10 \*\*\*  mnth -20.13 26.38 -0.763 0.4457  weathersit11 2835.33 300.30 9.442 < 2e-16 \*\*\*  weathersit21 2205.18 305.07 7.228 1.25e-12 \*\*\*  Sim, o p-value do modelo é de 2.2e-16, menor do que 0.05, portanto é significativo.   1. Qual o coeficiente de determinação do modelo ajustado?   Multiple R-squared: 0.5294, Adjusted R-squared: 0.5248  R2 múltiplo -> regressão linear simples (uma variável preditora)  R2 ajustado -> regressão linear simples (duas ou mais variáveis) |
|  |

## valores estimados pelo modelo

fitted(modelo2)

1 2 3 4 5 6 7 8

2578.1356 2690.2789 2349.9614 2371.0764 2527.6217 2396.3263 1720.7283 1537.6730

9 10 11 12 13 14 15 16

2012.9625 2085.5528 1561.4304 2212.6961 2167.8237 2143.8398 1934.4980 2554.9738

17 18 19 20 21 22 23 24

1600.5827 1837.7148 2276.2007 2099.0398 2240.4139 1553.0154 1770.1550 1775.2049

25 26 27 28 29 30 31 32

1877.2678 -362.6264 2342.0403 1761.1233 2350.8789 2467.0234 1629.6188 1675.3499

33 34 35 36 37 38 39 40

2069.2306 2275.2043 1786.4421 1914.3694 2849.3992 2767.1341 2471.9297 1342.0676

## resíduos do modelo

1 2 3 4 5 6 7 8

2578.1356 2690.2789 2349.9614 2371.0764 2527.6217 2396.3263 1720.7283 1537.6730

9 10 11 12 13 14 15 16

2012.9625 2085.5528 1561.4304 2212.6961 2167.8237 2143.8398 1934.4980 2554.9738

17 18 19 20 21 22 23 24

1600.5827 1837.7148 2276.2007 2099.0398 2240.4139 1553.0154 1770.1550 1775.2049

25 26 27 28 29 30 31 32

1877.2678 -362.6264 2342.0403 1761.1233 2350.8789 2467.0234 1629.6188 1675.3499

33 34 35 36 37 38 39 40

2069.2306 2275.2043 1786.4421 1914.3694 2849.3992 2767.1341 2471.9297 1342.0676

## resíduos padronizados do modelo

rstandard(modelo2)

1 2 3 4 5 6

-1.199198214 -1.422358659 -0.753140678 -0.608732947 -0.697705941 -0.594591799

7 8 9 10 11 12

-0.158672700 -0.435919507 -0.897024714 -0.575706097 -0.224795542 -0.790850799

13 14 15 16 17 18

-0.573496557 -0.544190985 -0.516732001 -1.016082599 -0.452348106 -0.869281604

19 20 21 22 23 24

-0.471256797 -0.129476370 -0.524896193 -0.431759876 -0.591224441 -0.270821090

25 26 27 28 29 30

0.081097046 0.669213176 -1.437927231 -0.447322749 -0.942685731 -1.031321136

1. Apresente o modelo final ajustado

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 469.1 435.3 1.078 0.281558

temp 14221.1 2265.8 6.276 5.98e-10 \*\*\*

atemp 8422.3 2417.3 3.484 0.000524 \*\*\*

season21 614.7 172.1 3.572 0.000377 \*\*\*

season31 756.8 225.5 3.356 0.000831 \*\*\*

season41 1082.0 150.8 7.174 1.82e-12 \*\*\*

hum -3248.1 441.3 -7.361 5.01e-13 \*\*\*

windspeed -3861.4 650.6 -5.935 4.56e-09 \*\*\*

weathersit21 -228.0 120.6 -1.891 0.059059 .

weathersit31 -1904.3 309.1 -6.161 1.20e-09 \*\*\*

temp:atemp -17808.5 1888.1 -9.432 < 2e-16 \*\*\*

cnt = 469,1 + 14221,1\*temp + 8422,3\*atemp + 614,7\*season2 + 756,8\*season3 + 1082,0\*season4 -  
 3248,1\*hum - 3861,4\*windspeed - 228,0\*weather2 - 1904,3\*weather3

1. Construir o simulador do modelo ajustado usando o Excel

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Coefficients** | **Estimate** | **Variáveis** | **Valores** | **Normalizada** | **Modelo** |
| (Intercept) | 469,1 |  |  |  | 469,10 |
| temp | 14221,1 | Temperatura Cº | 20 | 0,5957 | 8472,14 |
| atemp | 8422,3 | Sensação térmica | 33 | 0,7424 | 6252,92 |
| hum | -3248,1 | Humidade | 52 | 0,5200 | -1689,01 |
| windspeed | -3861,4 | Velocidade do vento | 9 | 0,1343 | -518,70 |
| season2 | 614,7 | Estação do ano | 1 | 0 | 0,00 |
| season3 | 756,8 |  |  | 0 | 0,00 |
| season4 | 1082,0 |  |  | 0 | 0,00 |
| weathersit2 | -228,0 | Clima | 1 | 0 | 0,00 |
| weathersit3 | -1904,3 |  |  | 0 | 0,00 |
| temp:atemp | -17808,5 |  |  | 0,4423 | -7876,62 |
|  |  |  |  | **predição cnt:** | **5109,84** |

1. Utilize o modelo ajustado para pontuar a base de dados **previsao\_bike\_sharing.xlsx**

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| id | season | yr | mnth | holiday | weekday | workingday | weathersit | temp | atemp | hum | windspeed | cnt |
| 1 | 2 | 0 | 6 | 0 | 4 | 1 | 1 | 42 | 58 | 31 | 20 | 2254,644727 |
| 2 | 3 | 0 | 7 | 0 | 6 | 0 | 1 | 43 | 60 | 44 | 8 | 2213,472217 |
| 3 | 3 | 1 | 9 | 0 | 5 | 1 | 1 | 36 | 54 | 69 | 10 | 2972,296697 |
| 4 | 1 | 1 | 12 | 0 | 0 | 0 | 1 | 20 | 33 | 52 | 9 | 5109,841124 |

1. Utilize o modelo de séries temporais para previsão **mensal** de bikes alugadas (aggr\_bike\_sharing\_mensal.xlsx)

|  |  |
| --- | --- |
| *Regression Statistics* | |
| Multiple R | 0,737712079 |
| R Square | 0,544219112 |
| Adjusted R Square | 0,523501799 |
| Standard Error | 36079,40407 |
| Observations | 24 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* |
| Intercept | 69033,22464 | 15202,07443 | 4,541039774 | 0,000160912 |
| tempo | 5452,938696 | 1063,923583 | 5,125310488 | 3,89157E-05 |

1. A série apresenta tendência? Use a regressão linear simples para detectar a tendência.

Sim, a série apresenta uma tendencia, pois o p-value é menor do que 0,05.

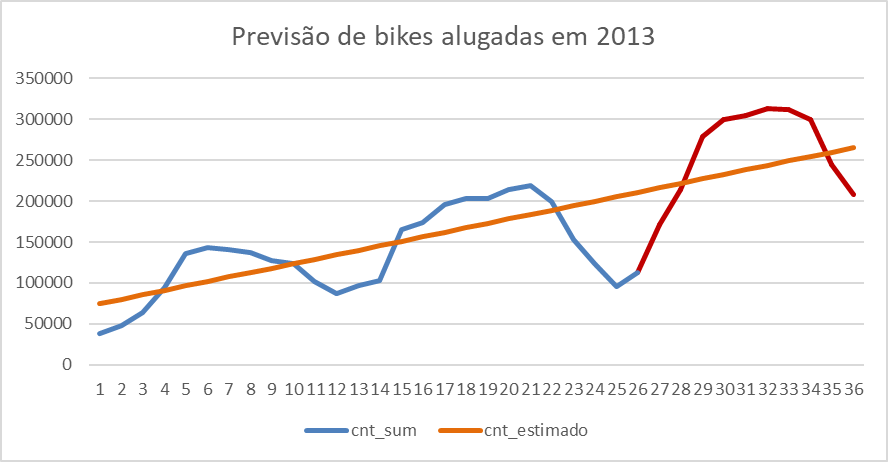
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ANOVA |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |
| Regression | 1 | 34194721481 | 34194721481 | 26,2688076 | 3,89157E-05 |
| Residual | 22 | 28637914754 | 1301723398 |  |  |
| Total | 23 | 62832636235 |  |  |  |

1. A série apresenta sazonalidade?

Sim, a série apresenta sazonalidade de acordo com os meses. Nos meses de janeiro e fevereiro o número de alugueis é bem menor do que nos outros meses.  
  


|  |  |  |  |
| --- | --- | --- | --- |
| **mês** | **projeção com  sazonalidade** | **projeção** | **sazonalidade** |
| 25 | 96012 | 205357 | 46,75% |
| 26 | 112708 | 210810 | 53,46% |
| 27 | 171233 | 216263 | 79,18% |
| 28 | 214606 | 221716 | 96,79% |
| 29 | 279176 | 227168 | 122,89% |
| 30 | 299256 | 232621 | 128,65% |
| 31 | 304318 | 238074 | 127,82% |
| 32 | 313590 | 243527 | 128,77% |
| 33 | 312435 | 248980 | 125,49% |
| 34 | 299782 | 254433 | 117,82% |
| 35 | 244302 | 259886 | 94,00% |
| 36 | 207930 | 265339 | 78,36% |

1. Qual a previsão de bikes alugadas a para janeiro de 2013?   
     
   Previsão de bikes alugadas em 2013 com projeção com sazonalidade:

****

**Parte 2)**

1) Escolha a melhor resposta a questão.

a) Selecione a opção que lista as variáveis apenas quantitativas contínuas.

( ) a temperatura do corpo, o número de crianças, o sexo, a área de um campo de futebol

( x ) idade, a temperatura corporal, o consumo de combustível, a renda

( ) número de filhos, sexo, consumo de combustível, a renda

( ) gênero, consumo de combustível, área de um campo de futebol, renda

b) Para uma distribuição assimétrica, qual das seguintes estatísticas é uma boa medida para resumir a tendência central dos dados?

( ) média

( x ) mediana

( ) ou média ou mediana

c) Classifique cada estatística a seguir em uma das opções: (a) tendência central; (b) dispersão; (c) forma;

( a ) média

( c ) curtose

( a ) mediana

( b ) desvio padrão

( b ) variância

( a ) moda

( c ) assimetria

( b ) intervalo interquartil

d) Numa distribuição de frequência os valores que se afastam das demais observações são denominados de:

( ) valor extremo

( x ) valor atípico

( ) ambos

e) O coeficiente de variação informa:

( x ) a variabilidade em torno da média

( ) a variabilidade em torno da mediana

( ) a medida relativa da variabilidade em torno da média

Data de entrega 16/03/2020

Arquivo formato Word.

Regina Bernal

07/03/2020