Regressão Horas Trabalhadas

Bruno Mesquita dos Santos

2023-11-06

Regressão Horas Trabalhadas

Abrindo base:

```
horas_trabalhadas <- read.delim("C:/Users/onurb/Downloads/horas_trabalhadas.txt")
```

Separando Variaveis

```
y  = horas_trabalhadas$y
x1 = horas_trabalhadas$XI
x2 = horas_trabalhadas$x2
x3 = horas_trabalhadas$x3
x4 = horas_trabalhadas$x4
x5 = horas_trabalhadas$x5
x6 = horas_trabalhadas$x6

rm(horas_trabalhadas)
```

Passo 1: Escolhendo a base (MRLS)

Modelo x1

```
mx1 <- lm(y~x1)
summary(mx1)
```

```
##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.707 -11.250  -2.624   6.829  46.524
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.23690   10.40733   8.671 2.03e-09 ***
## x1           0.05097    0.01663   3.064  0.00479 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 14.68 on 28 degrees of freedom
## Multiple R-squared:  0.2511, Adjusted R-squared:  0.2244
## F-statistic: 9.388 on 1 and 28 DF,  p-value: 0.004792
```

Modelo x2

```
mx2 <- lm(y~x2)
summary(mx2)
```

```
##
## Call:
## lm(formula = y ~ x2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.682 -11.656  -1.451   9.744  56.562
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.193e+02  7.788e+00  15.321 3.84e-15 ***
## x2          2.698e-03  1.119e-02   0.241    0.811
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.94 on 28 degrees of freedom
## Multiple R-squared:  0.002073,   Adjusted R-squared:  -0.03357
## F-statistic: 0.05816 on 1 and 28 DF,  p-value: 0.8112
```

Modelo x3

```
mx3 <- lm(y~x3)
summary(mx3)
```

```
##
## Call:
## lm(formula = y ~ x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.552 -11.472  -2.754  10.875  57.145
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.69924    7.66677  14.961 6.97e-15 ***
## x3            0.02915    0.03229   0.903    0.374
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.72 on 28 degrees of freedom
## Multiple R-squared:  0.0283, Adjusted R-squared:  -0.006405
## F-statistic: 0.8154 on 1 and 28 DF,  p-value: 0.3742
```

Modelo x4

```
mx4 <- lm(y~x4)
summary(mx4)
```

```
##
## Call:
## lm(formula = y ~ x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -32.624 -10.119  -2.046  10.632  52.508
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 114.83680    8.53094   13.46 9.46e-14 ***
## x4            0.06423    0.08231    0.78    0.442
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.78 on 28 degrees of freedom
## Multiple R-squared:  0.02129,    Adjusted R-squared:  -0.01367
## F-statistic: 0.609 on 1 and 28 DF,  p-value: 0.4417
```

Modelo x5

```
mx5 <- lm(y~x5)
summary(mx5)
```

```
##
## Call:
## lm(formula = y ~ x5)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.210 -10.356  -1.141   7.594  54.797
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 123.427216   2.933519  42.075   <2e-16 ***
## x5           -0.003996   0.001552  -2.575   0.0156 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.25 on 28 degrees of freedom
## Multiple R-squared:  0.1915, Adjusted R-squared:  0.1626
## F-statistic:  6.63 on 1 and 28 DF,  p-value: 0.0156
```

Modelo x6

```
mx6 <- lm(y~x6)
summary(mx6)
```

```
##
## Call:
## lm(formula = y ~ x6)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.964 -11.242  -2.612  10.176  55.082
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.157e+02  6.697e+00  17.274   <2e-16 ***
## x6          9.346e-03  1.037e-02   0.901    0.375
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.72 on 28 degrees of freedom
## Multiple R-squared:  0.02819,    Adjusted R-squared:  -0.006522
## F-statistic: 0.8121 on 1 and 28 DF,  p-value: 0.3752
```

Conclusão:

Como melhor modelo MRLS podemos escolher a variavel x1, com o maior *F: 9.388* e menor *p-value: 0.004792*, também possui o maior R quadrado apesar de não ser muito alto.

Passo 2: Etapa de forward

Adicionar x2

```
mx1x2 <- lm(y~x1+x2)
summary(aov(mx1x2))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6   9.514 0.00467 **
## x2           1    292   292.4   1.375 0.25112
## Residuals   27   5740   212.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mx1, mx1x2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 6032.3
## 2     27 5739.9  1     292.4 1.3755 0.2511
```

Adicionar x3

```
mx1x3 <- lm(y~x1+x3)
summary(aov(mx1x3))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6   9.161 0.00538 **
## x3           1     71    71.2   0.323 0.57473
## Residuals   27   5961   220.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mx1, mx1x3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 6032.3
## 2     27 5961.0  1    71.229 0.3226 0.5747
```

Adicionar x4

```
mx1x4 <- lm(y~x1+x4)
summary(aov(mx1x4))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6   9.647 0.00442 **
## x4           1    372   371.6   1.772 0.19422
## Residuals   27   5661   209.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mx1, mx1x4)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x4
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     28 6032.3
## 2     27 5660.7  1    371.59 1.7724 0.1942
```

Adicionar x5

```r
mx1x5 <- lm(y~x1+x5)
summary(aov(mx1x5))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6  10.432 0.00325 **
## x5           1    798   797.5   4.113 0.05252 .
## Residuals   27   5235   193.9
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mx1, mx1x5)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x5
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     28 6032.3
## 2     27 5234.8  1     797.5 4.1134 0.05252 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adicionar x6

```r
mx1x6 <- lm(y~x1+x6)
summary(aov(mx1x6))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6  10.203 0.00355 **
## x6           1    680   679.7   3.429 0.07504 .
## Residuals   27   5353   198.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(mx1, mx1x6)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x6
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     28 6032.3
## 2     27 5352.6  1    679.69 3.4286 0.07504 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusão:

O Fmax é igual *4.113* e é o modelo mais provável de ser adicionado.

Passo 3: Validando a variavel

F tabelado com 1 gl e o gl do residuo do modelo completo (y~x1+x5) a 90% de confiança

```r
qf(0.9, 1 , 27)
```

```
## [1] 2.901192
```

Conclusão:

Fmax(4.113) > Ftab(2.901192) aceitamos o modelo completo como *y~x1+x5*

Passo 3: Etapa de backward

Tentar remover x1

```r
mx5x1 <- lm(y~x5+x1)
anova(mx5x1, mx1)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x5 + x1
## Model 2: y ~ x1
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     27 5234.8
## 2     28 6032.3 -1    -797.5 4.1134 0.05252 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusão:

A 90% de confiança, não é significativo voltar ao modelo com apenas x5.

Passo 4: Etapa de forward

Adicionar x2

```
mx1x5x2 <- lm(y~x1+x5+x2)
summary(aov(mx1x5x2))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6  10.528 0.00322 **
## x5           1    798   797.5   4.151 0.05191 .
## x2           1    240   239.6   1.247 0.27431
## Residuals   26   4995   192.1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mx1x5, mx1x5x2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x5
## Model 2: y ~ x1 + x5 + x2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 5234.8
## 2     26 4995.2  1     239.6 1.2471 0.2743
```

Adicionar x3

```
mx1x5x3 <- lm(y~x1+x5+x3)
summary(aov(mx1x5x3))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6  10.396 0.00339 **
## x5           1    798   797.5   4.099 0.05329 .
## x3           1    177   176.5   0.907 0.34959
## Residuals   26   5058   194.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mx1x5, mx1x5x3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x5
## Model 2: y ~ x1 + x5 + x3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 5234.8
## 2     26 5058.2  1    176.52 0.9073 0.3496
```

Adicionar x4

```
mx1x5x4 <- lm(y~x1+x5+x4)
summary(aov(mx1x5x4))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6  12.022 0.00184 **
## x5           1    798   797.5   4.740 0.03873 *
## x4           1    861   860.6   5.115 0.03231 *
## Residuals   26   4374   168.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mx1x5, mx1x5x4)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x5
## Model 2: y ~ x1 + x5 + x4
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     27 5234.8
## 2     26 4374.2  1    860.58 5.1152 0.03231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Adicionar x6

```
mx1x5x6 <- lm(y~x1+x5+x6)
summary(aov(mx1x5x6))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6  10.725 0.00299 **
## x5           1    798   797.5   4.229 0.04991 *
## x6           1    331   331.3   1.757 0.19655
## Residuals   26   4903   188.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mx1x5, mx1x5x6)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x5
## Model 2: y ~ x1 + x5 + x6
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 5234.8
## 2     26 4903.4  1    331.33 1.7568 0.1965
```

Conclusão:

O Fmax é igual *5.115* e é o modelo mais provável de ser adicionado.

Passo 5: Validando a variavel

F tabelado com 1 gl e o gl do residuo do modelo completo (y~x1+x5+x4) a 90% de confiança

```
qf(0.9, 1 , 26)
```

```
## [1] 2.909132
```

Conclusão:

Fmax(5.115) > Ftab(2.909132) aceitamos o modelo completo como *y~x1+x5+x4*

Passo 6: Etapa de backward

Tentar remover x1

```
mrx4x5 <- lm(y~x4+x5)
anova(mrx4x5 , mx1x5x4)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x4 + x5
## Model 2: y ~ x1 + x5 + x4
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     27 5756.0
## 2     26 4374.2  1    1381.8 8.2132 0.00813 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tentar remover x5

```
mrx4x1 <- lm(y~x4+x1)
anova(mrx4x1 , mx1x5x4)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x4 + x1
## Model 2: y ~ x1 + x5 + x4
```

```
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1     27 5660.7
## 2     26 4374.2  1    1286.5 7.6468 0.01032 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Conclusão:

A 90% de confiança, escolhemos o Fcalc mínimo para comparar com Ftab, Fmin(7.6468) > Ftab(2.901192[1 gl 27]), rejeita-se Ho, e conclui-se que não se pode tirar a variável x5. Então mantém o modelo com modelo com *y~x1+x5+x4*

Passo 7: Etapa de forward

Adicionar x2

```
mx1x5x4x2 <- lm(y~x1+x5+x4+x2)
summary(aov(mx1x5x4x2))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6  12.659 0.00153 **
## x5           1    798   797.5   4.992 0.03465 *
## x4           1    861   860.6   5.386 0.02874 *
## x2           1    380   379.9   2.378 0.13564
## Residuals   25   3994   159.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mx1x5x4, mx1x5x4x2)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x5 + x4
## Model 2: y ~ x1 + x5 + x4 + x2
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 4374.2
## 2     25 3994.3  1    379.89 2.3777 0.1356
```

Adicionar x3

```
mx1x5x4x3 <- lm(y~x1+x5+x4+x3)
summary(aov(mx1x5x4x3))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6  12.287 0.00174 **
```

```
## x5           1    798   797.5   4.845 0.03719 *
## x4           1    861   860.6   5.228 0.03097 *
## x3           1    259   258.8   1.572 0.22154
## Residuals   25   4115   164.6
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mx1x5x4, mx1x5x4x3)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x5 + x4
## Model 2: y ~ x1 + x5 + x4 + x3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 4374.2
## 2     25 4115.4  1    258.75 1.5719 0.2215
```

Adicionar x6

```
mx1x5x4x6 <- lm(y~x1+x5+x4+x6)
summary(aov(mx1x5x4x6))
```

```
##             Df Sum Sq Mean Sq F value Pr(>F)
## x1           1   2023  2022.6  12.042 0.0019 **
## x5           1    798   797.5   4.748 0.0390 *
## x4           1    861   860.6   5.124 0.0325 *
## x6           1    175   175.1   1.042 0.3170
## Residuals   25   4199   168.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(mx1x5x4, mx1x5x4x6)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x5 + x4
## Model 2: y ~ x1 + x5 + x4 + x6
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     26 4374.2
## 2     25 4199.1  1    175.08 1.0424  0.317
```

Conclusão:

O Fmax é igual *2.378* e é o modelo mais provável de ser adicionado.

Passo 8: Validando a variavel

F tabelado com 1 gl e o gl do residuo do modelo completo (y~x1+x5+x4+x2) a 90% de confiança

```
qf(0.9, 1 , 25)
```

```
## [1] 2.917745
```

Conclusão:

Fmax(2.378) < Ftab(2.917745) rejeitamos o modelo completo como *y~x1+x5+x4+x2*

Conclusão final

O modelo final é y~x1+x5+x4 e possuí os seguintes resultados:

```
summary(mx1x5x4)
```

```
##
## Call:
## lm(formula = y ~ x1 + x5 + x4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.935  -7.524  -1.792   6.233  34.655
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 82.192065  11.858322   6.931 2.33e-07 ***
## x1           0.043787   0.015279   2.866  0.00813 **
## x5          -0.003979   0.001439  -2.765  0.01032 *
## x4           0.152531   0.067441   2.262  0.03231 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 12.97 on 26 degrees of freedom
## Multiple R-squared:  0.457,  Adjusted R-squared:  0.3943
## F-statistic: 7.293 on 3 and 26 DF,  p-value: 0.001052
```

Possuí um R-quadrado de 0.3943 o que não é muito bom já que identifica a porcentagem de variância no campo Y que é explicada pela variaveis independentes(Xs). Também tem um p-value: 0.001052 o que é possitivo pois está a baixo de 0.1 que é nosso nivel de significância.

```
summary(aov(mx1x5x4))
```

```
##             Df Sum Sq Mean Sq F value  Pr(>F)
## x1           1   2023  2022.6  12.022 0.00184 **
## x5           1    798   797.5   4.740 0.03873 *
## x4           1    861   860.6   5.115 0.03231 *
## Residuals   26   4374   168.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Todos os Pr(>F) são significativos em até 0.01.