Lista 9 [AD-UFPE-2019]

Antonio Fernandes

11 de maio de 2019

Conteúdo

Apresentação

Este documento apresenta as respostas da lista de exercícios 9 da disciplina de Análise de dados.

O link está disponível no GitHub: https://github.com/alvesat/AD_9

Questão 1

a)

Caso a correlação entre Zi e Xi é igual a 0, não acontece nenhum efeito em β, mas a conclusão acerca da variável dependente estará distorcida.

b)

Se Zi e Xi estão correlacionados, então Zi é uma variável omitida (viés de variável omitida) e deve ser incluída no modelo. Dado que a correlação entre Zi e Xi é positiva, significa que ambas as variáveis explicam uma mesma parcela da variação em Y.

c)

Do mesmo modo, caso Zi e Xi estejam negativamente correlacionadas, então Zi é uma variável omitida (viés de variável omitida) e deve ser incluída no modelo. Dado que a correlação é negativa, significa que a inclusão de Zi no modelo irá reduzir o efeito de Xi em β.

Questão 2

A tabela 9.4 apresenta o resultado de três modelos de regressão de salários de professores nos estados americanos e no distrito de colúmbia.

O modelo A contém apenas a variável *porcentagem de residentes no estado que possuem ensino superior*. O efeito dessa variável na VD é de 704.02 (com p-valor < 0.05). O R^2 do modelo é de 0.34.

Já o modelo B apresenta apenas a variável *Renda per capita*. O efeito dessa variável na VD é de 0.68 (com p-valor < 0.05). O R^2 do modelo é de 0.47.

Questão 3

O modelo C da tabela 9.4 contém as duas variáveis independentes (do modelo A e do modelo B). O efeito da variável *porcentagem de residentes no estado que possuem ensino superior* é de 24.56 mas não é significativo (p-valor > 0.05) enquanto que o efeito da variável *Renda per capita* é de 0.66 e significativo (p-valor > 0.05). Nesse caso, podemos concluir que as VIs estão correlacionadas, ocorrendo o viés de variável omitida tanto no modelo A quanto no modelo B.

Questão 4.1

a)

Modelo com as variáveis do banco *worldrecall.txt*

```
# lendo banco worldrecall

worldrecall <- read.delim("~/Dados/Listas/AD_9/AD_9/worldrecall.txt")

  # modelo linear

Linear <- lm(prop ~ time, data = worldrecall)

  # resumo do modelo

summary(Linear)
```

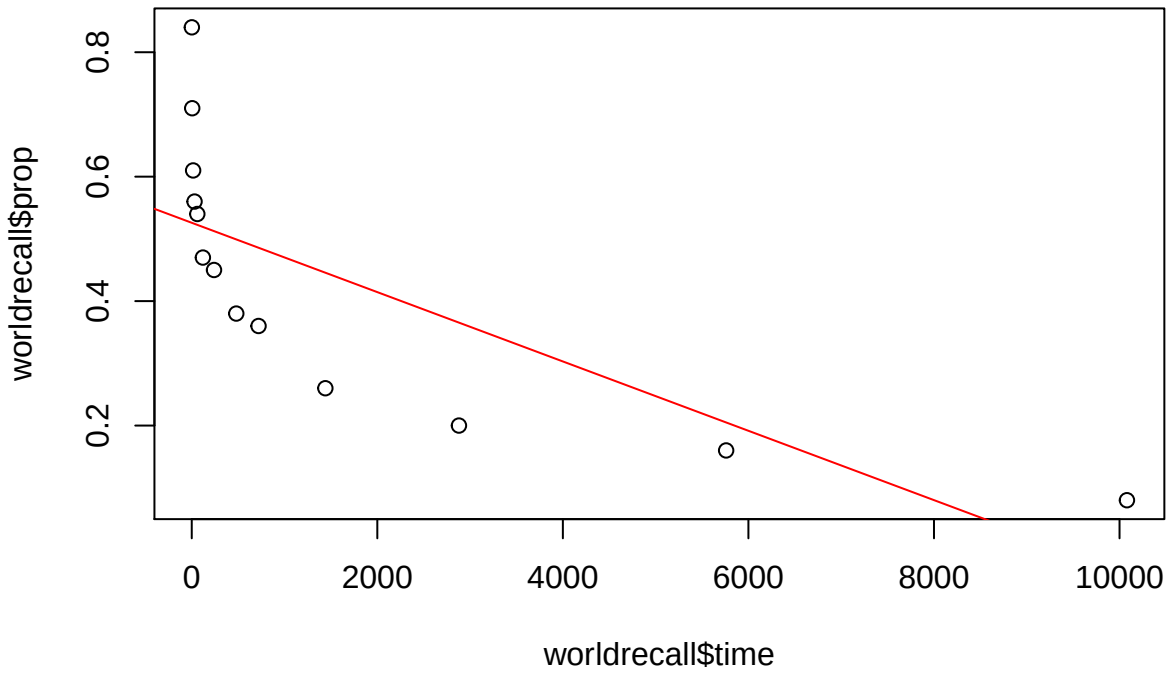
```
##
## Call:
## lm(formula = prop ~ time, data = worldrecall)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.18564 -0.11913 -0.04495  0.08496  0.31418
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.259e-01  4.881e-02  10.774 3.49e-07 ***
## time        -5.571e-05  1.457e-05  -3.825  0.00282 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1523 on 11 degrees of freedom
```

```
## Multiple R-squared:  0.5709, Adjusted R-squared:  0.5318
## F-statistic: 14.63 on 1 and 11 DF,  p-value: 0.002817
```

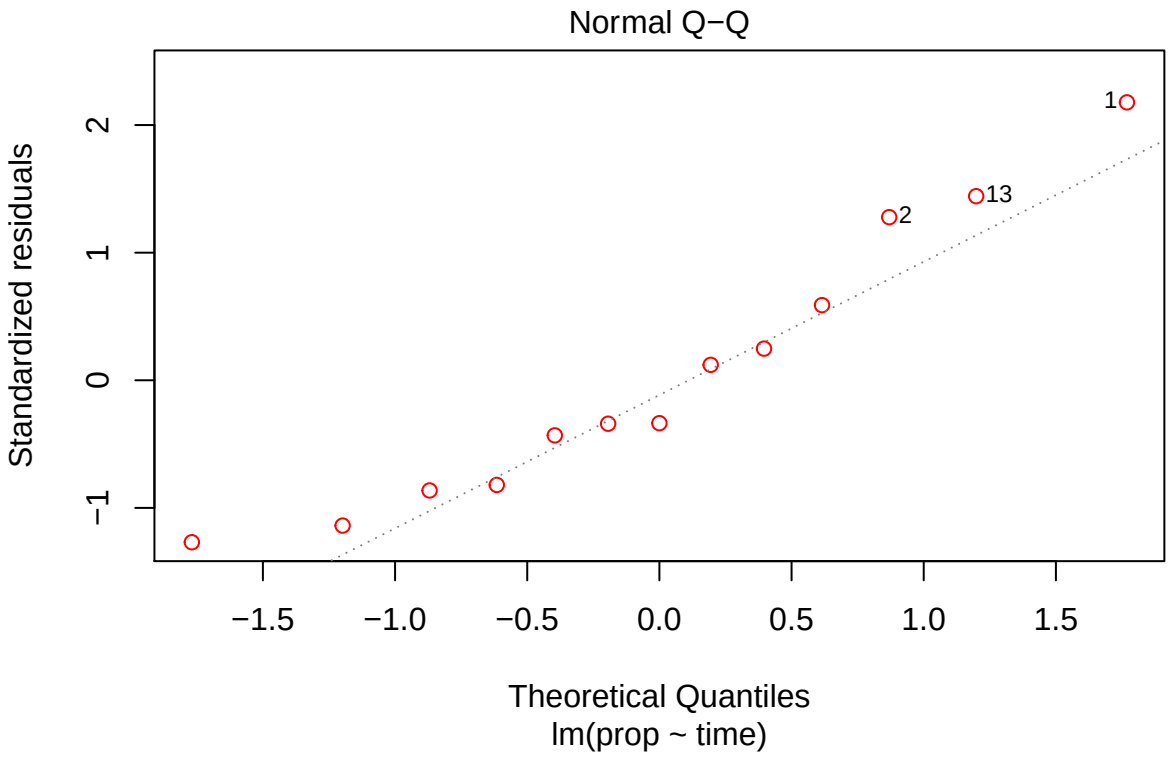
O resutado do modelo mostra um efeito de -0.00005571 da VI na VD. Entretanto o resultado apenas se mantém significativo com o P-valor < 0.1. O R^2 do modelo é de 0.53.

Entretanto, é necessário verificar se o modelo está devidamente ajustado.

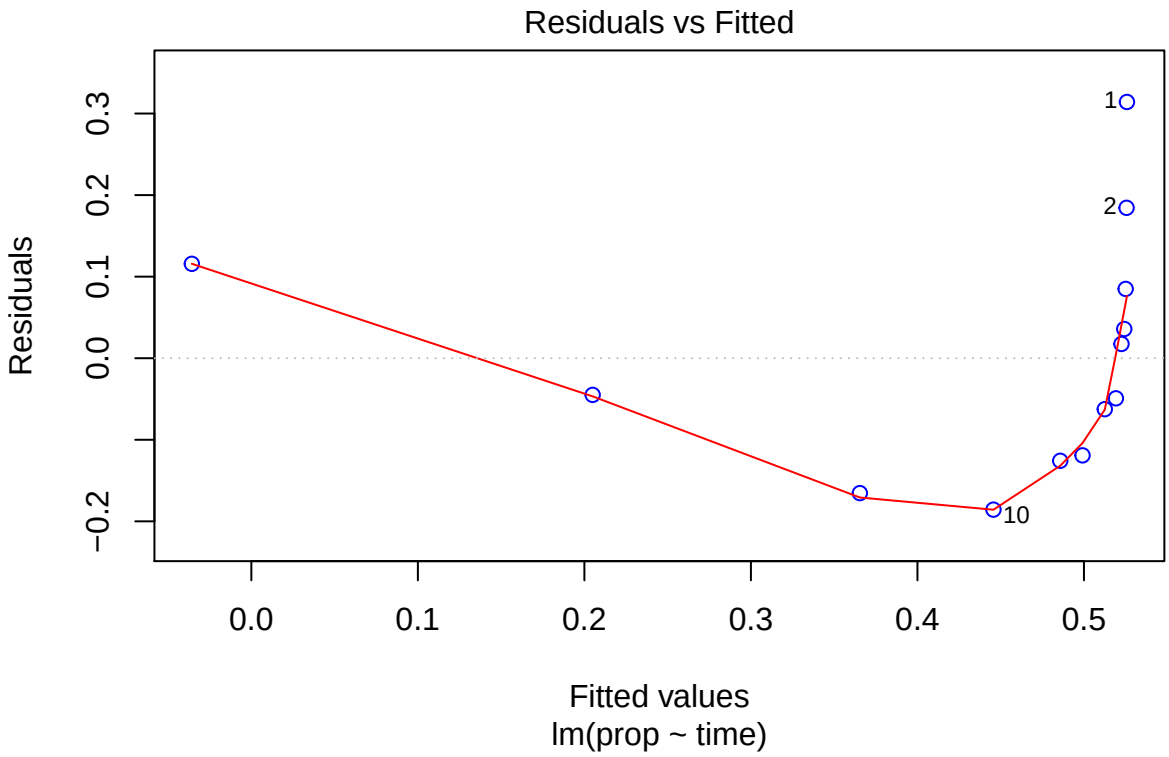
```
  # ajuste do modelo

plot(worldrecall$time, worldrecall$prop)
abline(lm(prop ~ time, data = worldrecall), col = 'red')
```




```
plot(Linear, which=2, col=c("red")) # Residuals vs Fitted Plot
```




```
plot(Linear, which=1, col=c("blue")) # Q-Q plot
```



Com base nos gráficos acima, podemos concluir que a relação entre as variáveis do modelo não é linear. A linha de regressão do gráfico mostra que as observações da variável independente não possuem um comportamento linear, do mesmo modo que o gráfico de residuals vs fitted.

Por isso, a variável independente do modelo será transformada em log

```
  # colocando vi em log

worldrecall$timeln <- log(worldrecall$time)

  # modelo linear

Linear_ln <- lm(prop ~ timeln, data = worldrecall)

  # resumo do modelo

summary(Linear_ln)
```

```
##
## Call:
## lm(formula = prop ~ timeln, data = worldrecall)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.036077 -0.015330 -0.006415  0.017967  0.037799
##
```

```
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.846415   0.014195   59.63 3.65e-15 ***
## timeln      -0.079227   0.002416  -32.80 2.53e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.02339 on 11 degrees of freedom
## Multiple R-squared:  0.9899, Adjusted R-squared:  0.989
## F-statistic:  1076 on 1 and 11 DF,  p-value: 2.525e-12
```

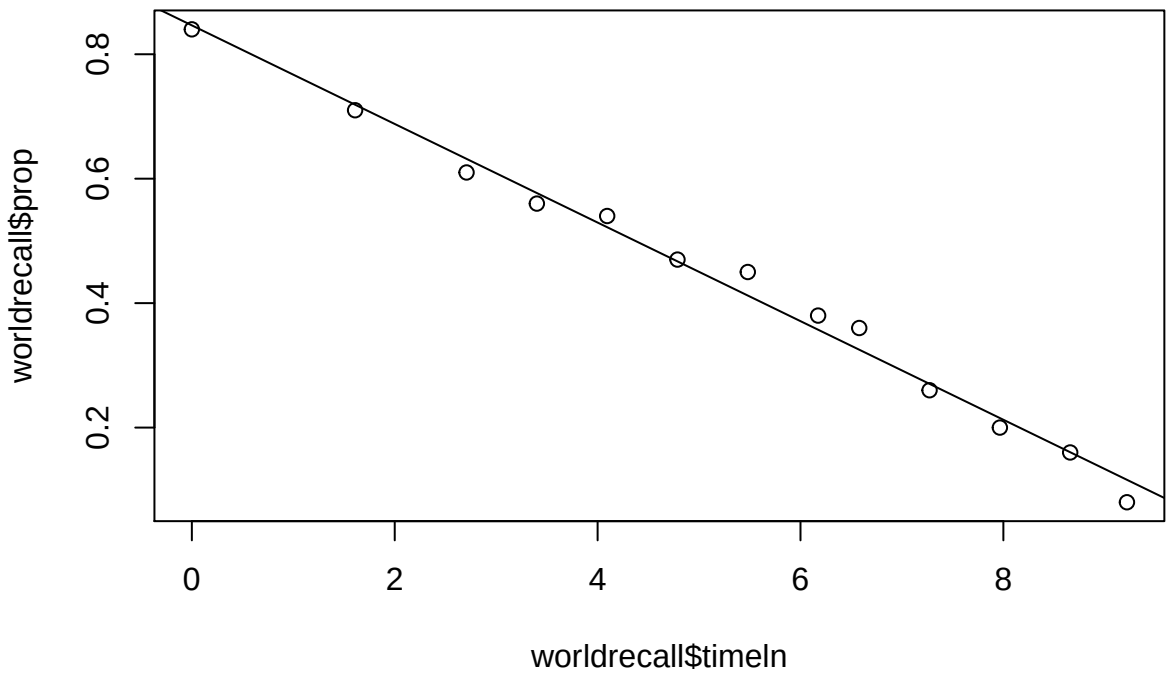
O modelo Level-log mostra que a variável independente tem um efeito negativo e estatisticamente significativo na VD (P-valor < 0.05). O R^2 do modelo é de 0.989, mostrando que a VI explica quase que completamente a variância da VD. Para interpretar a VI, é necessário efetuar uma transformação:

```
(-0.079227/100)
```

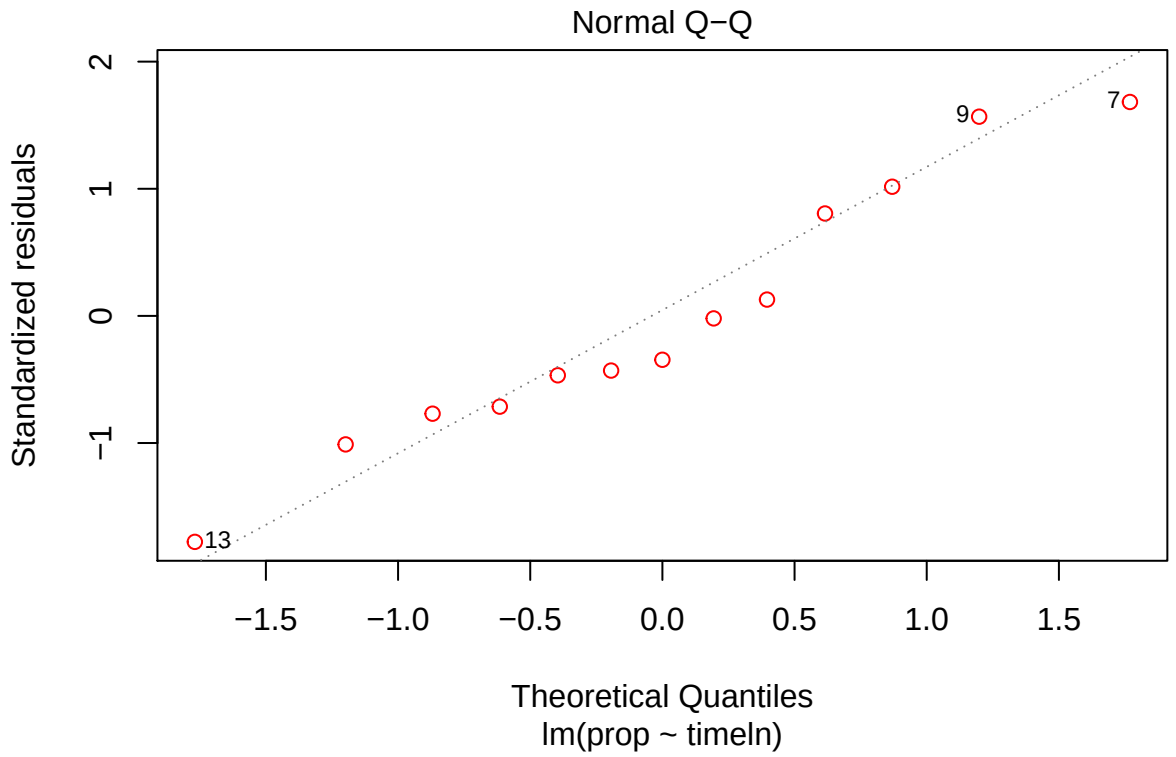
```
## [1] -0.00079227
```

Isso significa que 1% de variação em Xi implica em -0.00079227 de variação em Yi. Agora vamos verificar o ajuste do modelo

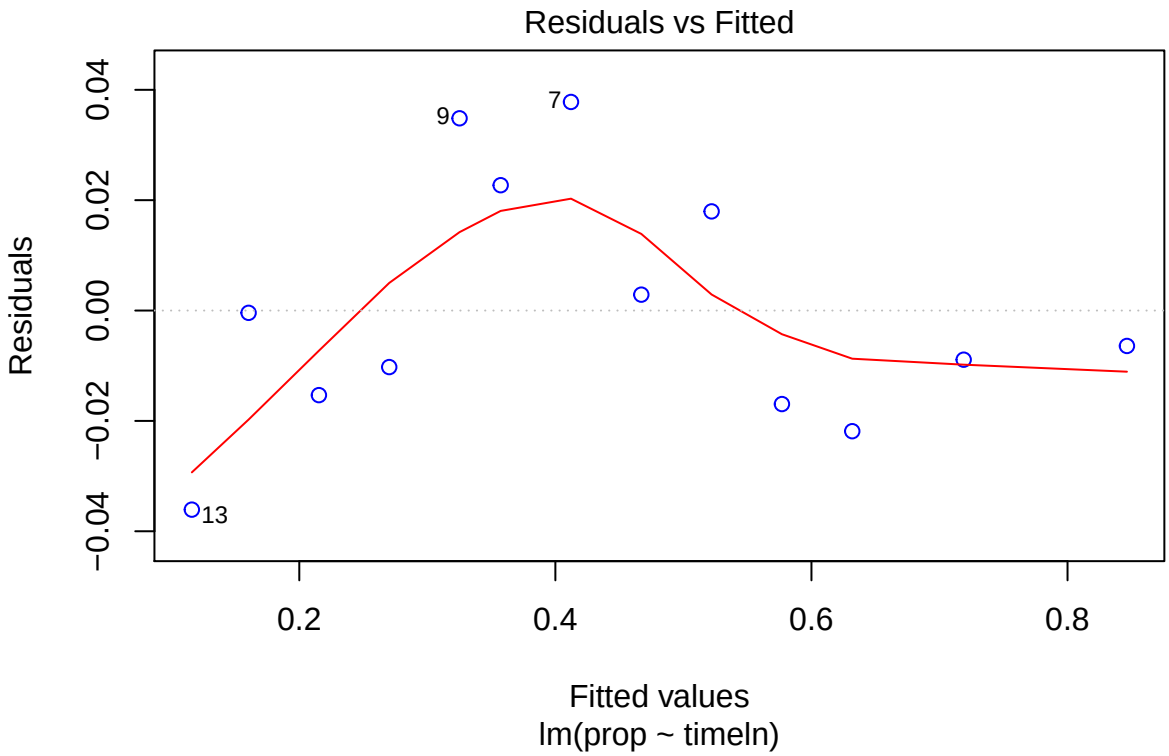
```
# ajuste do modelo

plot(worldrecall$timeln, worldrecall$prop)
abline(lm(prop ~ timeln, data = worldrecall))
```



```r
plot(Linear_ln, which=2, col=c("red")) # Residuals vs Fitted Plot
```




```r
plot(Linear_ln, which=1, col=c("blue")) # Q-Q plot
```



Ao observar os gráficos, percebemos que, com a transformação da variável independente em log, o modelo apresenta um ajuste adequado e uma relação linear entre as variáveis.