

# Aplicação

Patrícia Iwagaki

25 de novembro de 2016

Os dados foram coletados com os resultados, em segundos, da final dos 100 m nas olimpíadas, desde 1900 até 2012, para homens e mulheres, incluindo informações sobre o primeiro, segundo e terceiro colocados. As mulheres só começaram a participar dessa prova (e das provas de atletismo de um modo geral) em 1928.

Foram analisados as seguintes variáveis:

ano - ano dos jogos

cid - cidade onde ocorreram os jogos

elev - elevação em relação ao nível do mar em m da cidade dos jogos

mod - modalidade da prova (100m, não varia no data frame)

sexo - sexo do atleta (M - masculino, F - feminino)

nome - nome do atleta

ord - ordem de classificação na prova (1, 2, 3)

pais - código do país do atleta (3 letras)

t - tempo em segundos para completar a prova

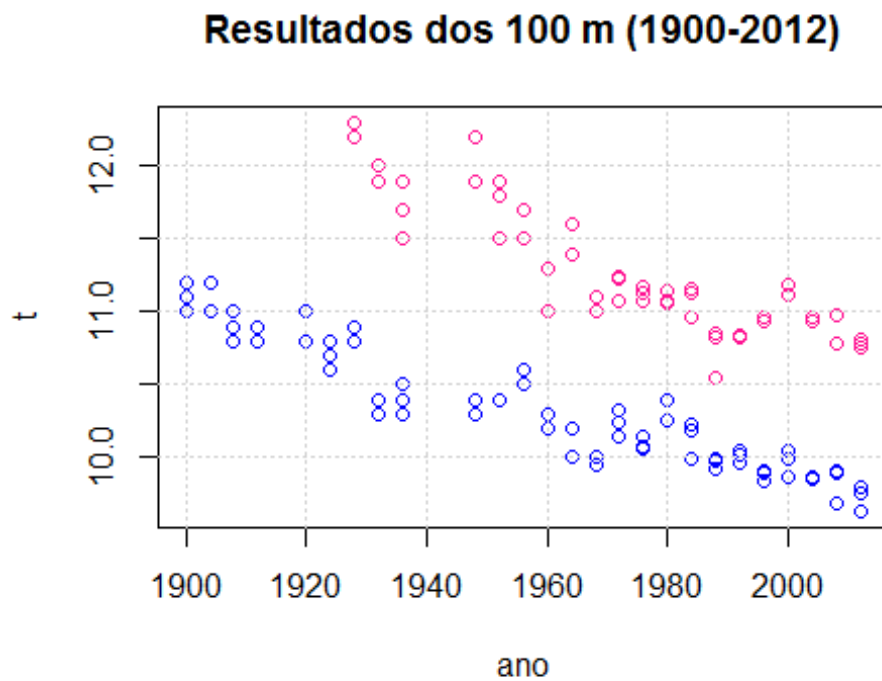
Primeiramente vamos carregar os dados, utilizando um pacote do próprio R

```
olimp<-read.csv2("http://ihbs.com.br/html/dados100m.csv")
head(olimp)
```

| ##   | ano  | cid         | elev | mod  | sexo | nome               | ord | pais | t    |
|------|------|-------------|------|------|------|--------------------|-----|------|------|
| ## 1 | 1928 | Amsterdam   | 2    | 100m | F    | Elizabeth Robinson | 1   | USA  | 12.2 |
| ## 2 | 1928 | Amsterdam   | 2    | 100m | F    | Fanny Rosenfeld    | 2   | CAN  | 12.3 |
| ## 3 | 1928 | Amsterdam   | 2    | 100m | F    | Ethel Smith        | 3   | CAN  | 12.3 |
| ## 4 | 1932 | Los Angeles | 113  | 100m | F    | S Walasiewicz      | 1   | POL  | 11.9 |
| ## 5 | 1932 | Los Angeles | 113  | 100m | F    | Hilda Strike       | 2   | CAN  | 11.9 |
| ## 6 | 1932 | Los Angeles | 113  | 100m | F    | Wilhelmina Bremen  | 3   | USA  | 12.0 |

Faremos um diagrama de dispersão para visualizar os dados dos homens e mulheres, ao longo dos anos, colorindo os pontos com cores rosa (mulheres) e azul (homens).

```
colvec<-ifelse(olimp$sexo=="M","blue","deeppink")
plot(olimp$ano,olimp$t,col=colvec,xlab="ano",ylab="t")
title(main="Resultados dos 100 m (1900-2012)")
grid()
```



Porém para nosso exemplo consideremos apenas os resultados dos 100 m das mulheres no data frame olimp.

```
olimpf<-olimp[olimp$sexo=="F",]
olimpf<-olimp[olimp$sexo=="F",]
y<-olimpf$t
x<-olimpf$ano
plot(x,y,xlab="ano",ylab="t",col="deeppink")
title(main="Dados da prova 100 m - Mulheres (1928-2012)")
```

A scatter plot showing the relationship between 'ano' (year) on the x-axis and 't' on the y-axis. The x-axis ranges from approximately 1930 to 2010, with major ticks at 1940, 1960, 1980, and 2000. The y-axis ranges from 10.5 to 12.0, with major ticks at 10.5, 11.0, 11.5, and 12.0. The data points are represented by open circles. The plot shows a general downward trend, with 't' values starting around 12.2 in the early 1930s and decreasing to around 10.7 by the early 2010s. There is a notable outlier at approximately (1988, 10.5).

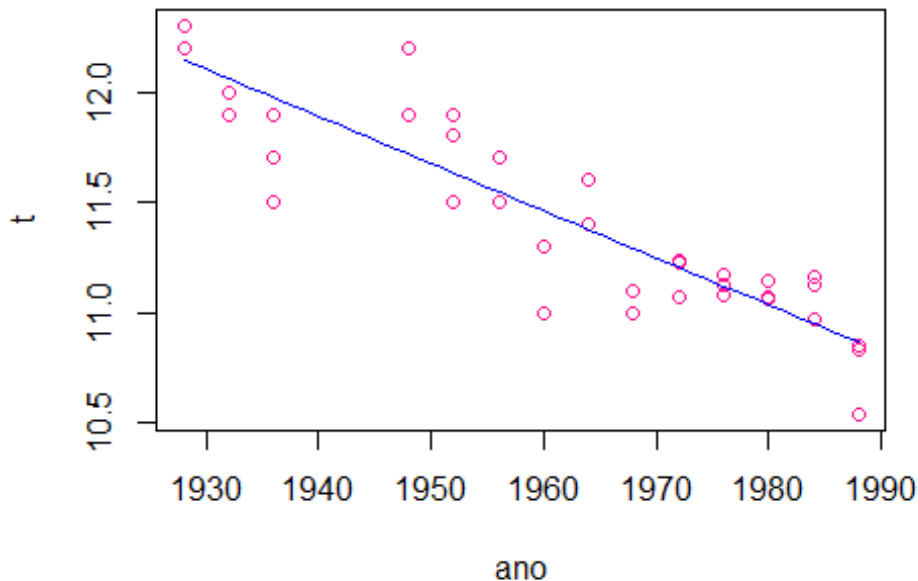
Suponha que, na situação dos dados relativos aos resultados das mulheres nos 100 m nos jogos olímpicos, desejamos incluir a possibilidade de haver uma mudança de regime no modelo, visando capturar o possível efeito do maior rigor nos exames antidoping após 1988. Desejamos, inclusive, verificar se esse efeito foi estatisticamente significativo.

[illegible]

Uma análise menos elaborada, poderia considerar a estimativa de 2 regressões separadas, uma para o período 1928-1988 e outra para o período 1989-2012 e tentar

construir o argumento sobre o resultado da análise. Essa primeira alternativa será examinada a seguir. As categorias da variável contdop estariam representadas em cada uma das regressões:

```
y<-olimpf$t[contdop=="A"]
x<-olimpf$ano[contdop=="A"]
modeloA<-lm(y~x)
novox<-1928:1988
yest<-predict(modeloA,newdata=data.frame(x=novox))
plot(x,y,xlab="ano",ylab="t",col="deeppink")
lines(novox,yest,col="blue")
```



```
summary(modeloA)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47472 -0.15417  0.02125  0.15394  0.48138
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.291804   3.467552   15.37  < 2e-16 ***
## x           -0.021341   0.001769  -12.06 6.59e-15 ***
## ---
```

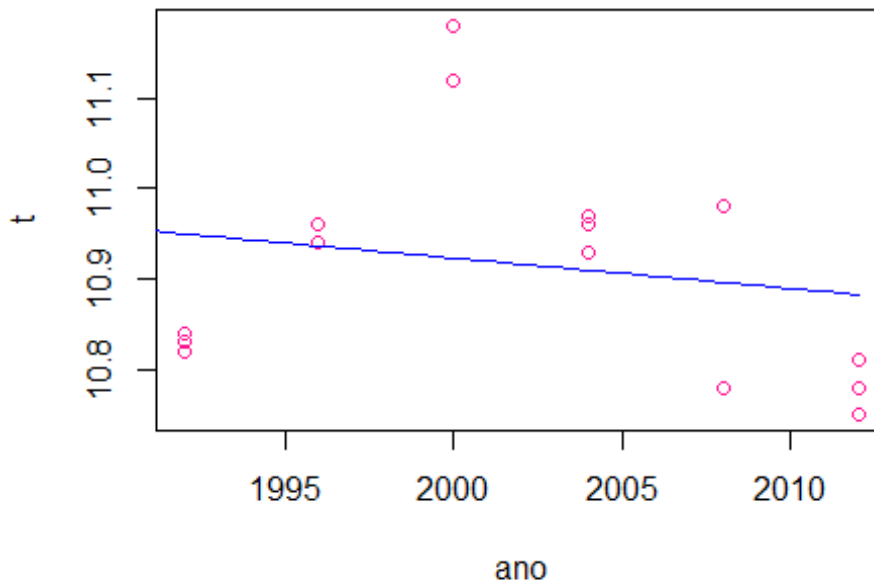
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2133 on 40 degrees of freedom
## Multiple R-squared:  0.7845, Adjusted R-squared:  0.7791
## F-statistic: 145.6 on 1 and 40 DF,  p-value: 6.592e-15
```

Estimando modelo para o período depois de 1988:

```
y<-olimpf$t[contdop=="D"]
x<-olimpf$ano[contdop=="D"]
modeloD<-lm(y~x)
summary(modeloD)

##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.132635 -0.109921  0.003536  0.060450  0.256993
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.651610   8.317354   2.122  0.0509 .
## x           -0.003364   0.004154  -0.810  0.4307
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1201 on 15 degrees of freedom
## Multiple R-squared:  0.04189,    Adjusted R-squared:  -0.02198
## F-statistic: 0.6558 on 1 and 15 DF,  p-value: 0.4307

novox<-1989:2012
yest<-predict(modeloD,newdata=data.frame(x=novox))
plot(x,y,xlab="ano",ylab="t",col="deeppink")
lines(novox,yest,col="blue")
```



Note que, no caso da segunda regressão estimada (período 1989-2012), os valores-p associados aos 2 parâmetros estimados foram superiores a 5%, sugerindo a não rejeição das hipóteses  $H_0: b_0=0$  e  $H_0: b_1=0$ . Isso ocorre pois ao analisar o gráfico de dispersão percebemos que não há características de linearidade.

Agora Utilizaremos as variáveis Dummy:

O procedimento ideal, desenvolvido a seguir, não separa os dados em 2 regressões, mas considera um único modelo que comporta as 2 regressões de uma só vez, com o apoio de 1 variável "dummy" que denominaremos  $z$ .

$z$ : assume valor 0 para o período A (1928-1988) e valor 1 para o período D (1989-2012)

$$y_i = b_0 + b_1x_i + b_2z + b_3z_ix_i + \epsilon_i$$

$$\epsilon_i \sim N(0, \sigma),$$

No modelo,  $y_i$  é a variável resposta que representa  $t$  (tempo em seg),  $x_i$  representa ano e  $z_i$  é a variável "dummy" que assume valor 0 para o período A (1928-1988) e valor 1 para o período D (1989-2012).

$$z_i = 0 \text{ (período A 1928-1988)} \quad y_i = b_0 + b_1x_i + \epsilon_i$$

$$z_i = 1 \text{ (período D 1989-2012)} \quad y_i = (b_0 + b_1) + (b_1 + b_3)x_i + \epsilon_i$$

Observamos então que os parâmetros no modelo geral representam:

$b_0$ : intercepto da curva de regressão no período A

$b_1$ : inclinação (efeito marginal) da curva de regressão no período A

$b_0 + b_2$ : intercepto da curva de regressão no período D

$b_1 + b_3$ : inclinação (efeito marginal) da curva de regressão no período D

Logo, nesse modelo, os parâmetros  $b_2$  e  $b_3$  representam os "efeitos", respectivamente, no intercepto e inclinação da regressão observados ao se passar do período A (1928-1988) para o período D (1989-2012).

A vantagem dessa formulação, com relação à formulação anterior, é a possibilidade de testarmos, diretamente, hipóteses envolvendo  $b_2$  e  $b_3$  utilizando os valores-p associados a esses novos parâmetros, considerando todas as observações disponíveis.

Só efeito no intercepto: para estimar somente o efeito no intercepto, sem considerar o efeito na inclinação da curva, o modelo seria definido somente por:

$$Y_i = b_0 + b_1 X_i + b_2 z + \varepsilon$$

Inicialmente, vamos criar a variável  $z$  no conjunto de dados, a partir da variável `contdop` criada anteriormente.

```
z<-ifelse(contdop=="D",1,0)
z

## [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
## [36] 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

Para estimar o modelo, note que precisamos de uma nova variável, associada ao parâmetro  $b_3$ , que considera o produto de  $z$  e  $x$ . Essa nova variável é indicada na fórmula do comando `lm`, dado que `ano` está associada à variável  $x$ . Essa é a notação utilizada para criação de novas variáveis a partir das variáveis existentes:

```
modelcomp<-lm(t~ano+z+I(ano*z),data=olimpf)
summary(modelcomp)

##
## Call:
## lm(formula = t ~ ano + z + I(ano * z), data = olimpf)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.47472 -0.12492  0.01967  0.12821  0.48138
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  53.291804   3.127992  17.037  <2e-16 ***
## ano         -0.021341   0.001596 -13.375  <2e-16 ***
## z          -35.640193  13.687233  -2.604   0.0118 *
##
```

```
## I(ano * z)    0.017977    0.006844    2.627    0.0111 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1924 on 55 degrees of freedom
## Multiple R-squared:  0.8332, Adjusted R-squared:  0.8241
## F-statistic: 91.56 on 3 and 55 DF,  p-value: < 2.2e-16
```

Os resultados mostram que os estimadores dos parâmetros  $b_2$  e  $b_3$ , são, respectivamente,  $b^2 = 35.64019$  e  $b^3 = 0.01798$ . Os valores-p associados às hipóteses da nulidade  $H_0: b_2 = 0$  e  $H_0: b_3 = 0$  são próximos de 1, o que suporta a rejeição dessas hipóteses.

Pelo modelo, as regressões para os períodos A e D seriam dadas por:

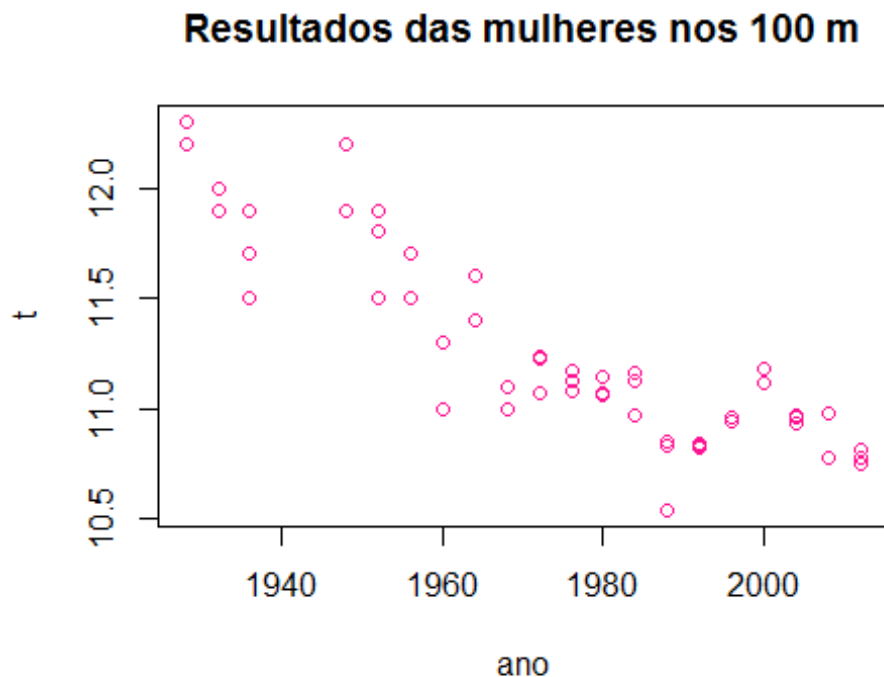
Período A (1929-1988):  $Y = 53.29180 - 0.02134x$

Período B(1988-2012):  $\hat{Y} = (53.29180 - 35.64019) + (-0.02134 + 0.01798)x = 17.65161 - 0.00336x$

A próxima figura mostra essas duas retas dentro do diagrama de dispersão, usando o modelo `modelcomp` estimado nos parágrafos anteriores:

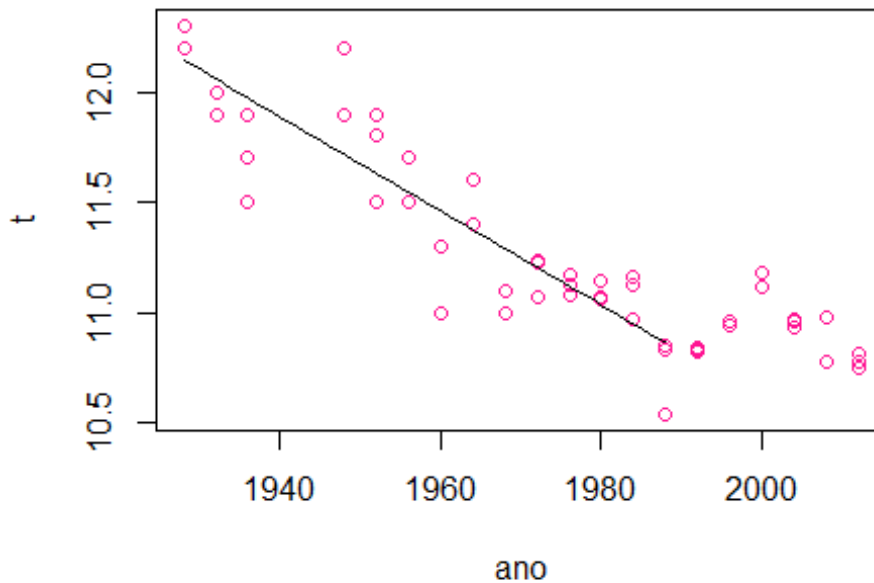
Diagrama de dispersão:

```
plot(olimpf$ano,olimpf$t,xlab="ano",ylab="t",col="deeppink")
title(main="Resultados das mulheres nos 100 m")
```



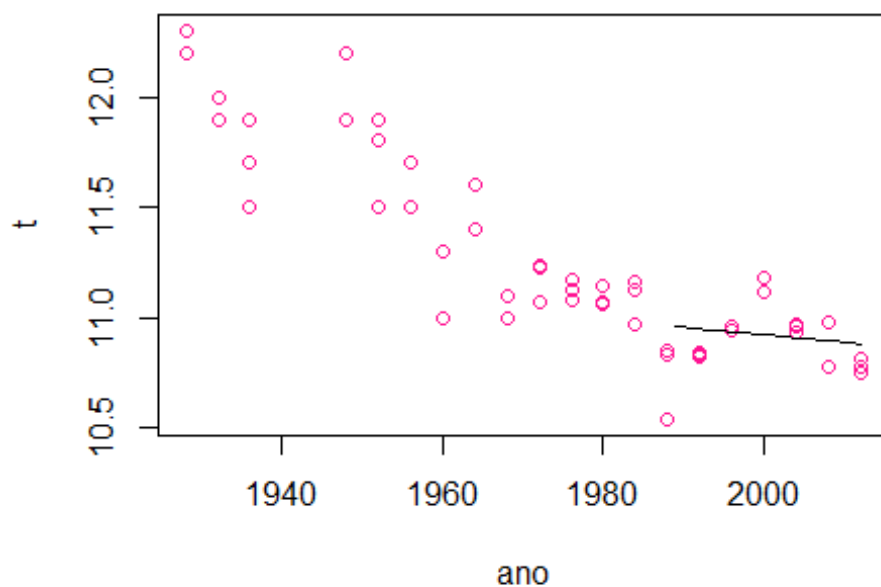
Regressão para o período (1928-1988)

```
novoano=1928:1988  
novoz=0  
novoy=predict(modelocomp,newdata=data.frame(ano=novoano,z=novoz))  
plot(olimpf$ano,olimpf$t,xlab="ano",ylab="t",col="deeppink")  
lines(novoano,novoy,type="l")
```



Regressão para o período (1988-2012)

```
novoano=1989:2012  
novoz=1 ## valor de z  
novoy=predict(modelocomp,newdata=data.frame(ano=novoano,z=novoz))  
plot(olimpf$ano,olimpf$t,xlab="ano",ylab="t",col="deeppink")  
lines(novoano,novoy,type="l")
```



Após 1988 (ex. esteróides anabolizantes, etc.) houve uma diminuição na utilização de anabolizantes. Isso indica que, possivelmente, essas substâncias eram utilizadas com maior frequência no período anterior a 1989. Essa evidência também suporta a conclusão de que a própria tendência de redução dos tempos das mulheres ao longo dos anos parece ter sido reduzida substancialmente após 1988.