Atividade 1

Deborah Pereira

22/04/2021

```
library(modelr)
library(tidyverse)
library(gapminder)
library(caret)
```

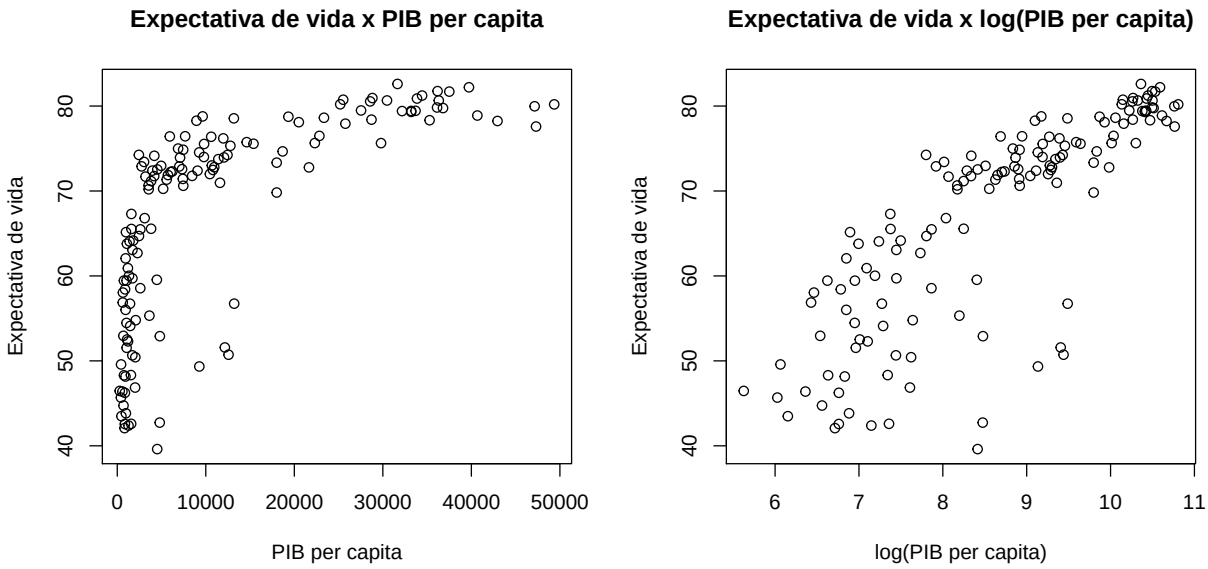
(a)

```
dados = gapminder[gapminder$year ==2007,]
dados["ln_gdpPercap"] = log(dados$gdpPercap)

par(mfrow=c(1,2))

plot(dados$gdpPercap
     ,dados$lifeExp
     ,main = "Expectativa de vida x PIB per capita"
     ,ylab = "Expectativa de vida"
     ,xlab = "PIB per capita")

plot(dados$ln_gdpPercap
     ,dados$lifeExp
     ,main = "Expectativa de vida x log(PIB per capita)"
     ,ylab = "Expectativa de vida"
     ,xlab = "log(PIB per capita)")
```



Inicialmente, não parece ter uma relação linear entre Expectativa de vida e PIB per capita. Sim, após a transformação do PIB per capita a relação parece ser mais linear.

(b)

Divisão entre dados teste e dados treinamento:

```
set.seed(123)

dados["Aux"]=rbinom(n=nrow(dados),size=1,1/3)

dados_teste = dados[dados$Aux == 0,1:7]
dados_treino = dados[dados$Aux == 1,1:7]
```

- Modelo 1: dados x e y na escala original;

```
modelo1= lm(lifeExp ~ gdpPercap,  data = dados_treino)

summary(modelo1)
```

```
##
## Call:
## lm(formula = lifeExp ~ gdpPercap, data = dados_treino)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.709  -4.483   1.446   5.881  12.385
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.023e+01  1.916e+00  31.435  < 2e-16 ***
```

```
## gdpPercap   6.691e-04  1.098e-04   6.096 2.08e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.966 on 46 degrees of freedom
## Multiple R-squared:  0.4469, Adjusted R-squared:  0.4349
## F-statistic: 37.16 on 1 and 46 DF,  p-value: 2.077e-07
```

- Modelo 2: log(x) como covariável, y na escala original;

```
modelo2= lm(lifeExp ~ ln_gdpPercap, data = dados_treino)

summary(modelo2)
```

```
##
## Call:
## lm(formula = lifeExp ~ ln_gdpPercap, data = dados_treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.7938  -2.0704   0.6554   4.5224  14.0399
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -1.1845     7.7100  -0.154    0.879
## ln_gdpPercap   7.8706     0.8587   9.166 6.03e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.171 on 46 degrees of freedom
## Multiple R-squared:  0.6462, Adjusted R-squared:  0.6385
## F-statistic: 84.01 on 1 and 46 DF,  p-value: 6.026e-12
```

O R^2 está indicando que a variabilidade da expectativa de vida é explicada em 63,85% pelo modelo 2, já pelo modelo 1, apenas 43,49%.

(c)

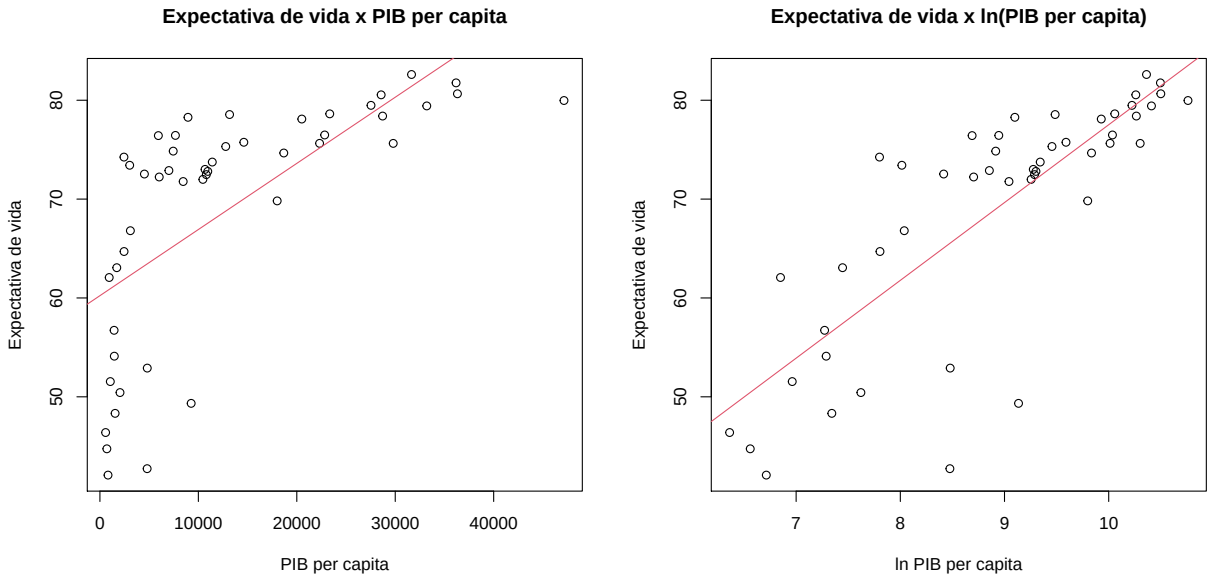
Scatter plot para os dois ajustes:

```
par(mfrow=c(1,2))

plot(dados_treino$lifeExp ~ dados_treino$gdpPercap
     ,data = dados_treino
     ,main ="Expectativa de vida x PIB per capita"
     ,ylab="Expectativa de vida"
     ,xlab="PIB per capita")
abline(modelo1, col =2)

plot(dados_treino$lifeExp ~ dados_treino$ln_gdpPercap
     ,data = dados_treino
```

```
     ,main ="Expectativa de vida x ln(PIB per capita)"
     ,ylab="Expectativa de vida"
     ,xlab="ln PIB per capita")
abline(modelo2, col =2)
```

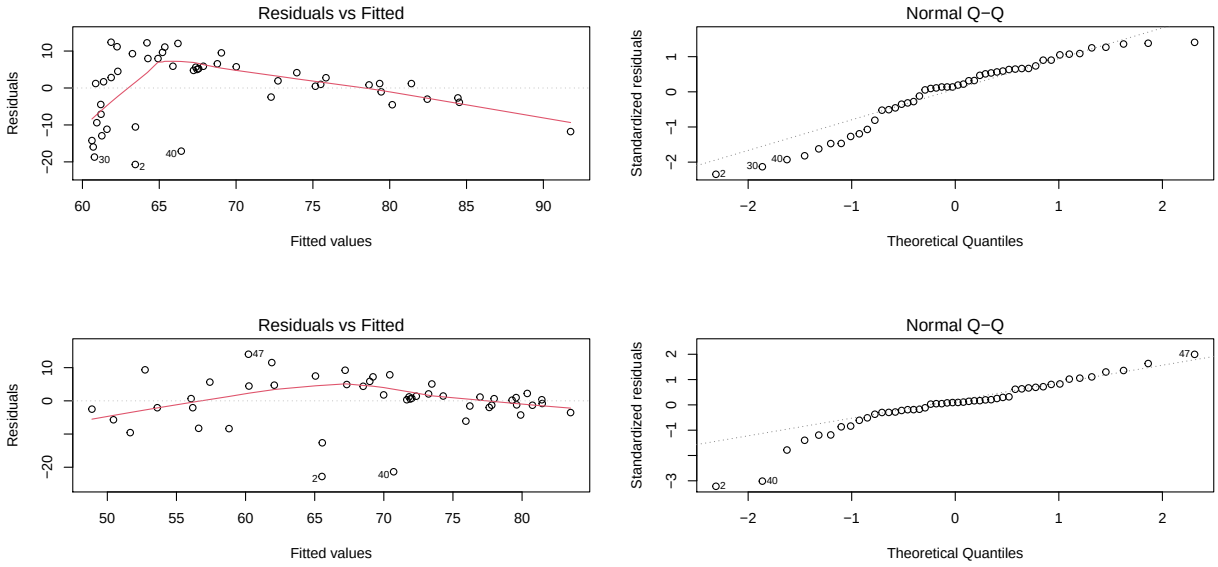


Utilizando a transformação logarítmica do PIB per capita, melhorou para que utilizássemos uma regressão linear para previsão da expectativa de vida. Mesmo assim, ainda não parece ser o melhor modelo, pois é possível notar que ainda há simetria nos erros.

```
par(mfrow=c(2,2))

plot(modelo1, which = 1:2)

plot(modelo2, which = 1:2)
```



(d)

Previsão no conjunto de teste para os 2 modelos:

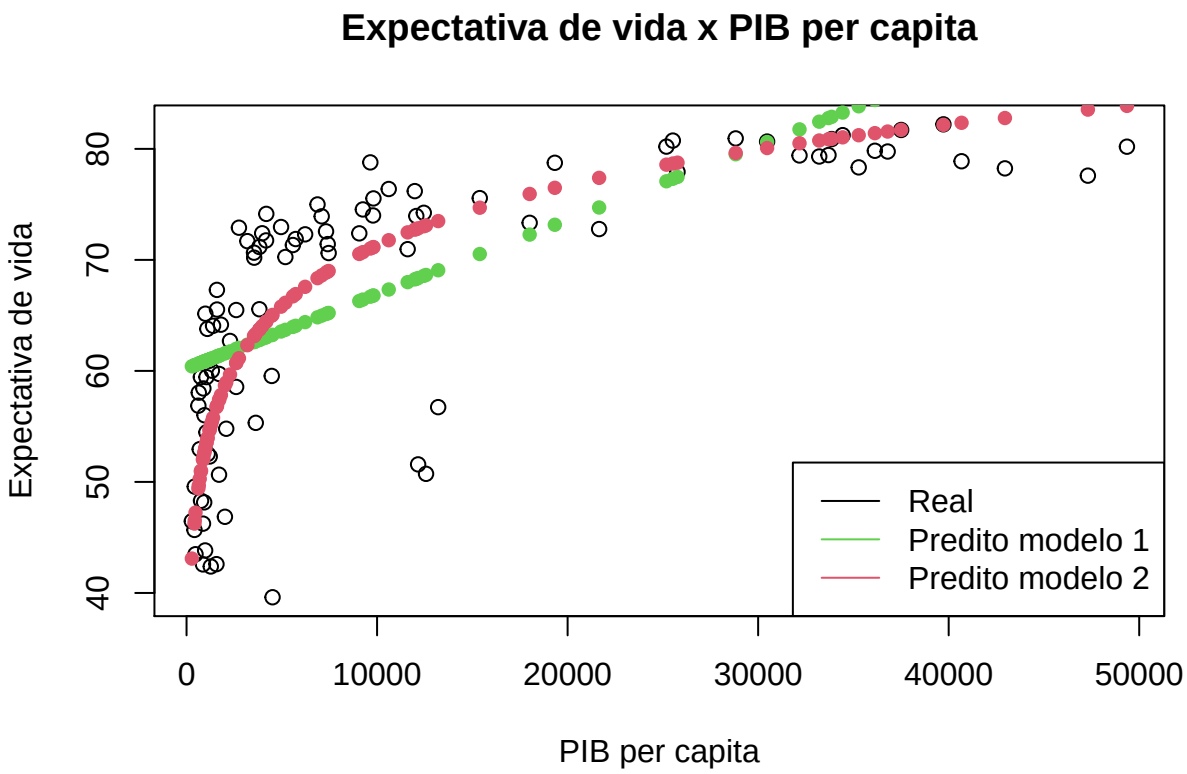
```
previsao_modelo1= predict(modelo1,dados_teste)
previsao_modelo2= predict(modelo2,dados_teste)

plot(dados_teste$lifeExp ~ dados_teste$gdpPercap
     ,data = dados_teste
     ,main ="Expectativa de vida x PIB per capita"
     ,ylab="Expectativa de vida"
     ,xlab="PIB per capita")

# Adicionando o valor previsto pelo modelo1
points(x = dados_teste$gdpPercap,
       y = previsao_modelo1,
       pch = 16,
       col = 3)

# Adicionando o valor previsto pelo modelo2
points(x = dados_teste$gdpPercap,
       y = previsao_modelo2,
       pch = 16,
       col = 2)

legend("bottomright",
       legend=c("Real","Predito modelo 1", "Predito modelo 2"),
       lty=c(1,1,1),
       col=c(1,3,2))
```



É possível ver que o modelo 2 prevê melhor os dados.

Medidas de acurácia para comparar os modelos:

```
rmse(model = modelo1, data =  dados_teste)
```

```
## [1] 8.970874
```

```
rmse(model = modelo2, data =  dados_teste)
```

```
## [1] 7.199603
```

Pelo RMSE, o modelo 2 parece ser um melhor modelo em relação ao modelo 1.

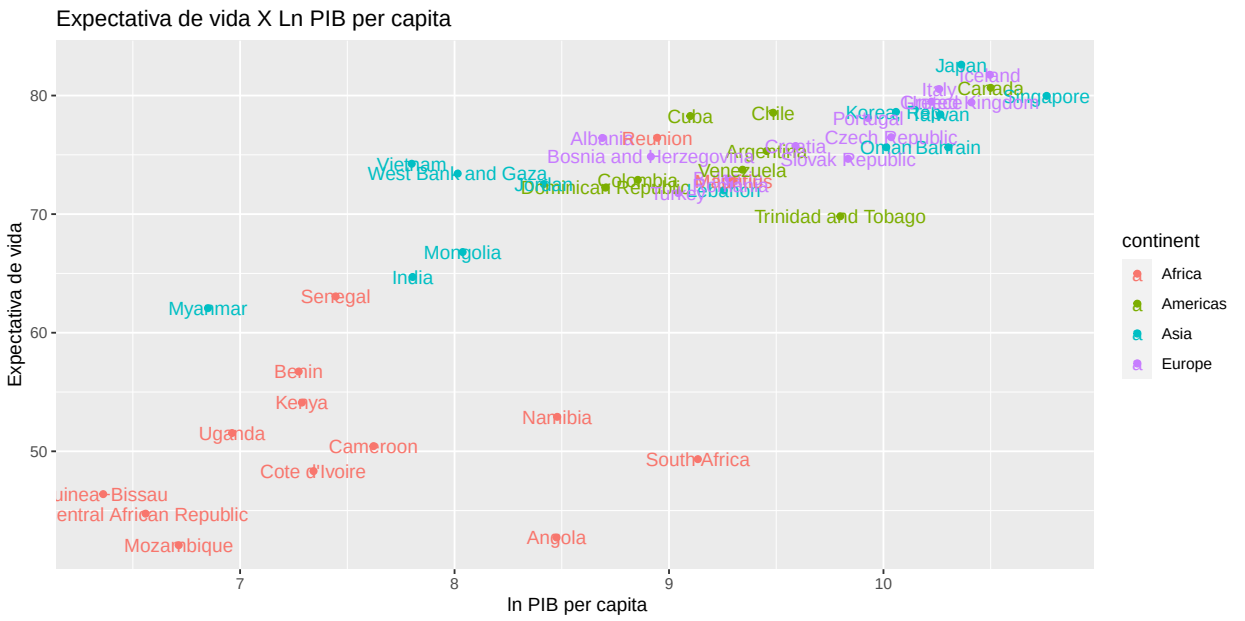
(e)

Verificando se existem países com comportamentos que destoem dos demais, na amostra de treinamento, com o uso de scatter:

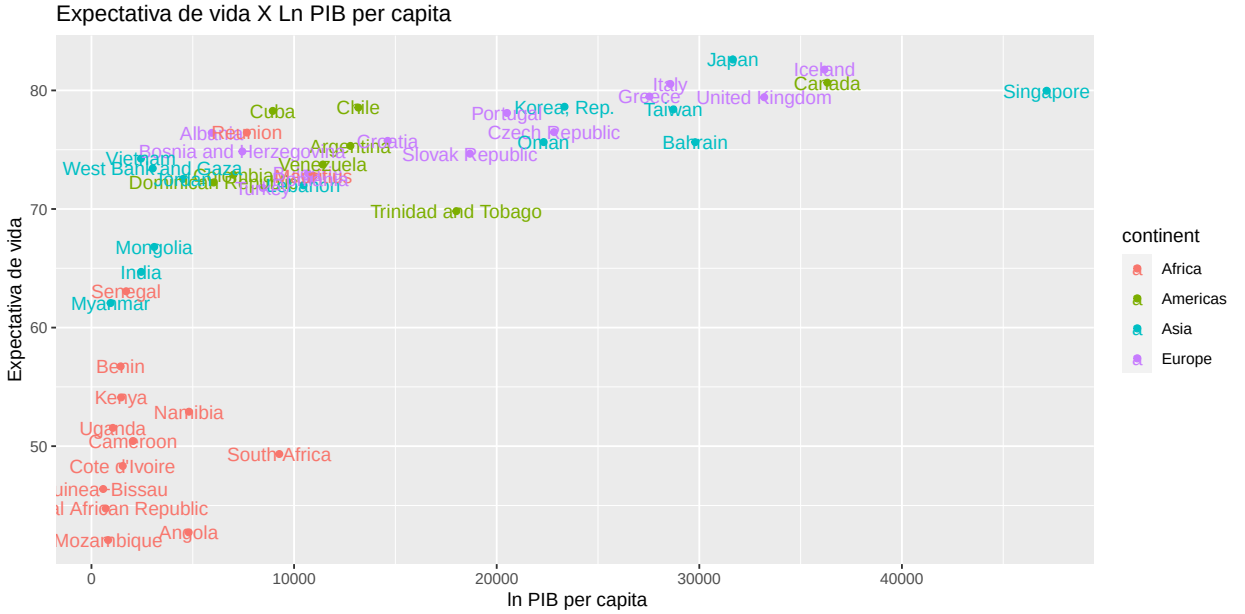
```
par(mfrow=c(1,2))

ggplot(dados_treino, aes(x=ln_gdpPercap, y=lifeExp, color= continent)) +
geom_point() +
```

```
geom_text(label=dados_treino$country)+
labs(title="Expectativa de vida X Ln PIB per capita",
x=" ln PIB per capita", y = "Expectativa de vida")
```




```
ggplot(dados_treino, aes(x= gdpPercap, y= lifeExp, color= continent)) +
geom_point() +
geom_text(label=dados_treino$country)+
labs(title="Expectativa de vida X Ln PIB per capita",
x=" ln PIB per capita", y = "Expectativa de vida")
```



Angola, África do Sul, Namíbia e Trinidad and Tobago parecem se destoar. Retirando esses países da análise:

```r
dados_treino2 = subset(dados_treino,
                       country != "Angola" & country != "South Africa" & country != "Namibia" & country
```

Refazendo o ajuste:

```r
modelo2_2= lm(lifeExp ~ ln_gdpPercap, data = dados_treino2)

summary(modelo2_2)
```

```
##
## Call:
## lm(formula = lifeExp ~ ln_gdpPercap, data = dados_treino2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.1289  -2.7466  -0.5028   3.3040  12.5477
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.7520     5.5552   0.135    0.893
## ln_gdpPercap   7.8136     0.6188  12.628 6.86e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.122 on 42 degrees of freedom
## Multiple R-squared:  0.7915, Adjusted R-squared:  0.7866
## F-statistic: 159.5 on 1 and 42 DF,  p-value: 6.859e-16
```

O R^2 é agora de 78,66%, antes da retirada dos países era de 63,85%. Houve uma melhora.