L5P1: Modelos de Regressão Linear - Yure Campos

```
dados = read_csv(
    here::here("data/participation-per-country.csv"),
    col_types = cols(
        .default = col_double(),
        site = col_character(),
        country = col_character(),
        geo = col_character(),
        four_regions = col_character(),
        eight_regions = col_character(),
        six_regions = col_character(),
        `World bank income group 2017` = col_character()
    )
) %>%
    filter(usuarios > 200)
glimpse(dados)
```

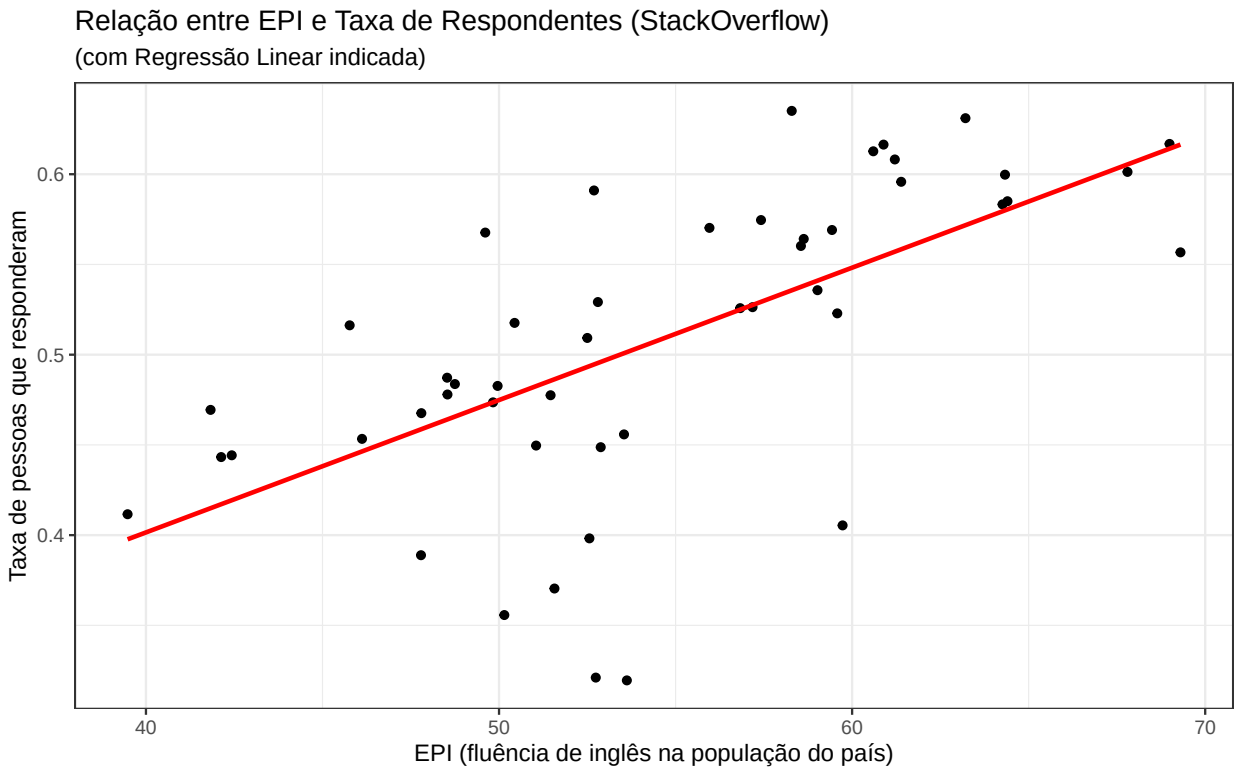
```
## Rows: 121
## Columns: 21
## $ site                          <chr> "StackOverflow", "StackOverflow", "Stac~
## $ country                       <chr> "Argentina", "Australia", "Austria", "B~
## $ PDI                           <dbl> 49, 36, 11, 80, 65, 69, 70, 39, 63, 80,~
## $ IDV                           <dbl> 46, 90, 55, 20, 75, 38, 30, 80, 23, 20,~
## $ MAS                           <dbl> 56, 61, 79, 55, 54, 49, 40, 52, 28, 66,~
## $ UAI                           <dbl> 86, 51, 70, 60, 94, 76, 85, 48, 86, 30,~
## $ usuarios                      <dbl> 2798, 12313, 2518, 2558, 4275, 10717, 1~
## $ responderam_prop              <dbl> 0.5357398, 0.6133355, 0.6310564, 0.3928~
## $ perguntaram_prop              <dbl> 0.5210865, 0.5897832, 0.5933280, 0.4757~
## $ editaram_prop                 <dbl> 0.09256612, 0.14699911, 0.14932486, 0.0~
## $ comentaram_prop               <dbl> 0.25339528, 0.33395598, 0.35027800, 0.1~
## $ GNI                           <dbl> NA, 59570, 48160, 840, 44990, 11630, 68~
## $ Internet                      <dbl> 51.0, 79.5, 79.8, 5.0, 78.0, 45.0, 51.0~
## $ EPI                           <dbl> 59.02, NA, 63.21, NA, 61.21, 49.96, NA,~
## $ geo                           <chr> "arg", "aus", "aut", "bgd", "bel", "bra~
## $ four_regions                  <chr> "americas", "asia", "europe", "asia", "~
## $ eight_regions                 <chr> "america_south", "east_asia_pacific", "~
## $ six_regions                   <chr> "america", "east_asia_pacific", "europe~
## $ Latitude                      <dbl> -34.00000, -25.00000, 47.33333, 24.0000~
## $ Longitude                     <dbl> -64.00000, 135.00000, 13.33333, 90.0000~
## $ `World bank income group 2017` <chr> "Upper middle income", "High income", "~
```

1. Descreva a relação entre EPI (fluência de inglês na população do país) e a taxa de pessoas daquele país que responderam alguma pergunta no StackOverflow.

```r
dados_stack = dados %>%
        filter(site == "StackOverflow") %>%
        filter(!is.na(responderam_prop)) %>%
        filter(!is.na(EPI))


dados_stack %>%
    ggplot(aes(x = EPI, y = responderam_prop)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(title = "Relação entre EPI e Taxa de Respondentes (StackOverflow)",
         subtitle = "(com Regressão Linear indicada)",
         x = "EPI (fluência de inglês na população do país)",
         y = "Taxa de pessoas que responderam"
         )
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
cor(dados_stack$responderam_prop, dados_stack$EPI)
```

```
## [1] 0.6345309
```

```r
modelo1 = lm(dados_stack$responderam_prop ~ dados_stack$EPI)

summary(modelo1)
```

```
##
## Call:
## lm(formula = dados_stack$responderam_prop ~ dados_stack$EPI)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.181903 -0.008037  0.013796  0.037407  0.099443
##
## Coefficients:
##                 Estimate Std. Error t value    Pr(>|t|)
## (Intercept)     0.108082   0.071716   1.507       0.138
## dados_stack$EPI 0.007335   0.001303   5.628 0.000000978 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06502 on 47 degrees of freedom
## Multiple R-squared:  0.4026, Adjusted R-squared:  0.3899
## F-statistic: 31.68 on 1 and 47 DF,  p-value: 0.0000009785
```

```
tidy(modelo1)
```

```
## # A tibble: 2 x 5
##   term            estimate std.error statistic     p.value
##   <chr>              <dbl>     <dbl>     <dbl>       <dbl>
## 1 (Intercept)      0.108     0.0717       1.51 0.138
## 2 dados_stack$EPI  0.00734   0.00130      5.63 0.000000978
```

```
glance(modelo1)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared  sigma statistic     p.value    df logLik   AIC   BIC
##       <dbl>         <dbl>  <dbl>     <dbl>       <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.403         0.390 0.0650      31.7 0.000000978     1   65.4 -125. -119.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Existe uma correlação entre as variáveis EPI (fluência de inglês na população do país) e a taxa de pessoas daquele país que responderam (responderam_prop) de média para alta (0.6345309).

Equação da regressão:

$responderam_prop = 0.108082 + 0.007335 * EPI$

A cada uma unidade de EPI, aumenta, em média, 0.734% das pessoas que responderam (IC 95%[0.697;0.771]), com 0,1% de significância.

E o R_Quadrado (coeficiente de determinação) é de 40.3% (R2 = 0.4026295).

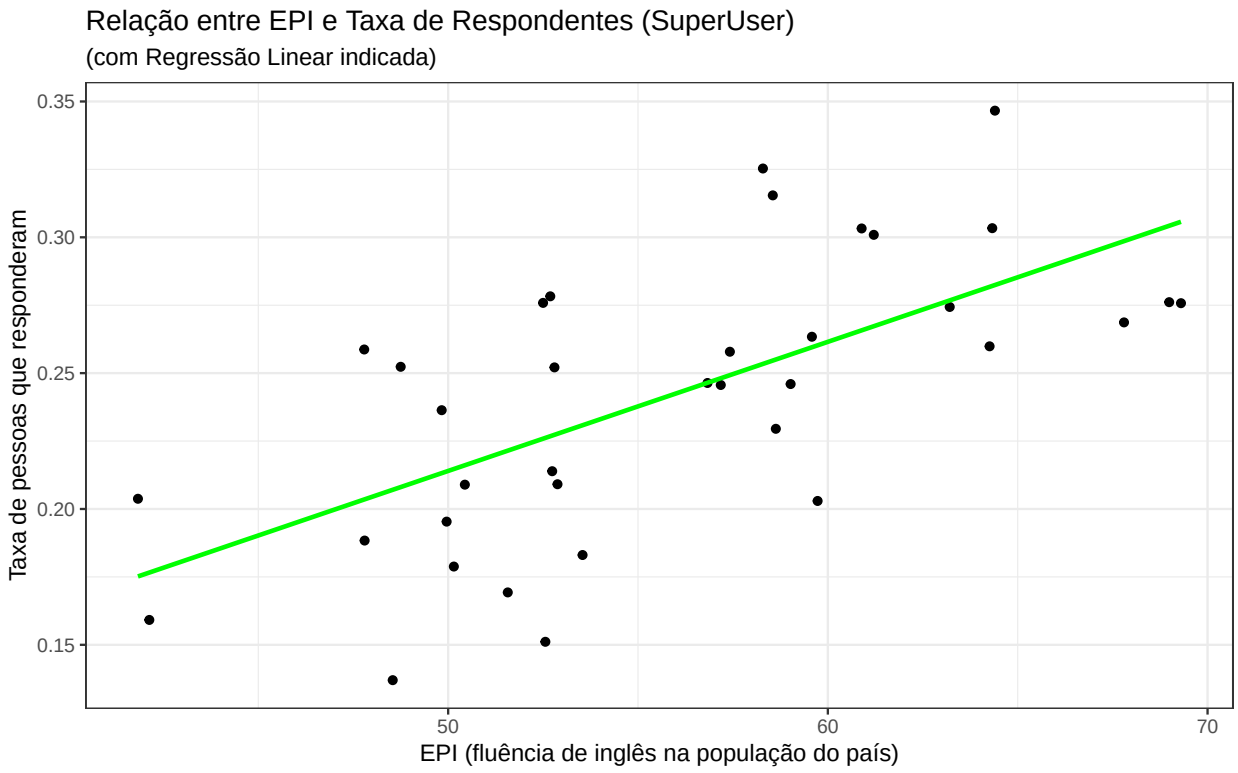
---

2. Descreva a relação entre as mesmas duas variáveis no SuperUser e compare o comportamento das pessoas de diferentes países nos dois sites comparando os resultados dos dois modelos.

```
dados_superuser = dados %>%
        filter(site == "SuperUser") %>%
        filter(!is.na(responderam_prop)) %>%
        filter(!is.na(EPI))

dados_superuser %>%
    ggplot(aes(x = EPI, y = responderam_prop)) +
    geom_point()+
    geom_smooth(method = "lm", se = FALSE, color = "green") +
    labs(title = "Relação entre EPI e Taxa de Respondentes (SuperUser)",
         subtitle = "(com Regressão Linear indicada)",
         x = "EPI (fluência de inglês na população do país)",
         y = "Taxa de pessoas que responderam"
         )
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(dados_superuser$responderam_prop, dados_superuser$EPI)
```

```
## [1] 0.6482049
```

```
modelo2 = lm(dados_superuser$responderam_prop ~ dados_superuser$EPI)

summary(modelo2)
```

```
##
```

```
## Call:
## lm(formula = dados_superuser$responderam_prop ~ dados_superuser$EPI)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.075052 -0.026155 -0.004809  0.029849  0.071932
##
## Coefficients:
##                       Estimate Std. Error t value  Pr(>|t|)
## (Intercept)         -0.0236147  0.0538137  -0.439     0.664
## dados_superuser$EPI  0.0047522  0.0009574   4.964 0.0000191 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03971 on 34 degrees of freedom
## Multiple R-squared:  0.4202, Adjusted R-squared:  0.4031
## F-statistic: 24.64 on 1 and 34 DF,  p-value: 0.00001914
```

```
tidy(modelo2)
```

```
## # A tibble: 2 x 5
##   term                estimate std.error statistic   p.value
##   <chr>                  <dbl>     <dbl>     <dbl>     <dbl>
## 1 (Intercept)         -0.0236   0.0538      -0.439 0.664
## 2 dados_superuser$EPI  0.00475  0.000957     4.96  0.0000191
```

```
glance(modelo2)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared  sigma statistic   p.value    df logLik   AIC   BIC
##       <dbl>         <dbl>  <dbl>     <dbl>     <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.420         0.403 0.0397      24.6 0.0000191     1   66.1 -126. -121.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Para o site SuperUser, também existe uma correlação entre as variáveis EPI (fluência de inglês na população do país) e a taxa de pessoas daquele país que responderam (responderam_prop) de média para alta (0.6482049).

Equação da regressão:

$responderam_prop = -0.0236147 + 0.0047522 * EPI$

A cada uma unidade de EPI, aumenta, em média, 0.475% das pessoas que responderam (IC 95%[0.451;0.499]), com 0,1% de significância.

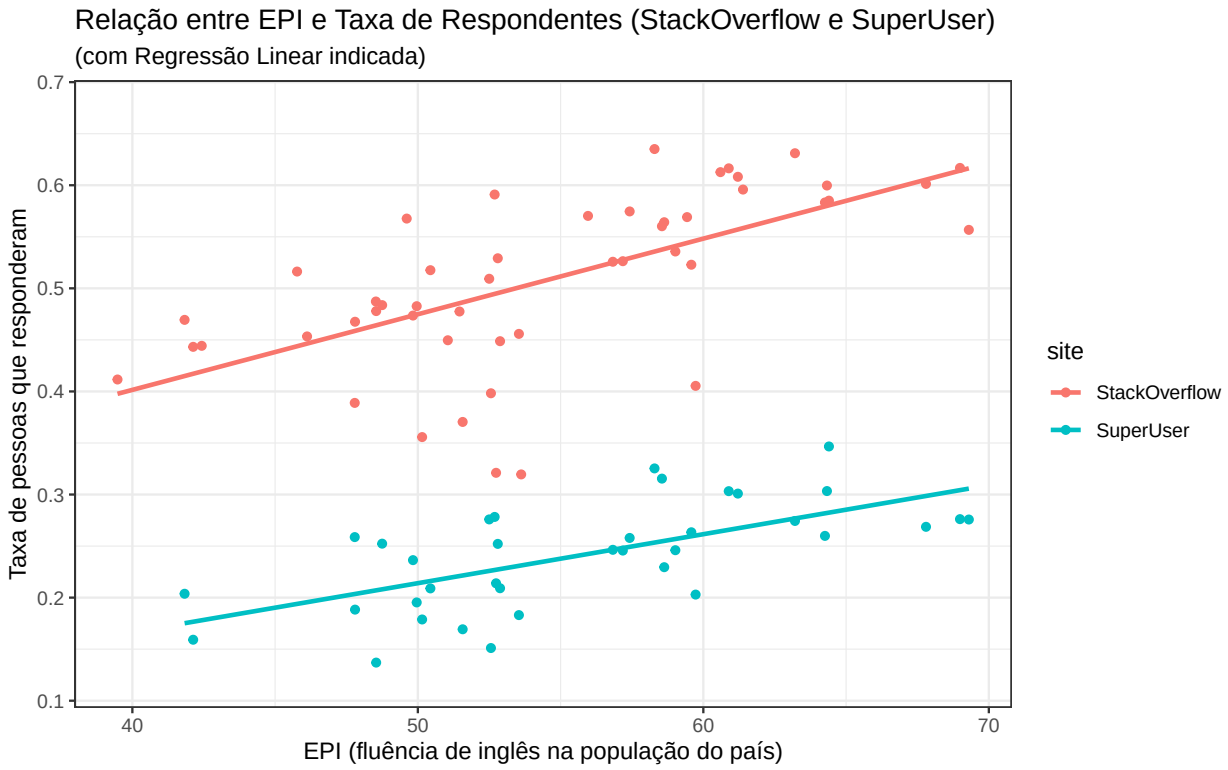
E o R_Quadrado (coeficiente de determinação) é de 42% (R2 = 0.4201696).

```
dados %>%
    ggplot(aes(x = EPI, y = responderam_prop, color = site)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE) +
    labs(title = "Relação entre EPI e Taxa de Respondentes (StackOverflow e SuperUser)",
         subtitle = "(com Regressão Linear indicada)",
         x = "EPI (fluência de inglês na população do país)",
         y = "Taxa de pessoas que responderam"
         )
```

```
## `geom_smooth()` using formula 'y ~ x'

## Warning: Removed 36 rows containing non-finite values (stat_smooth).

## Warning: Removed 36 rows containing missing values (geom_point).
```



Coeficientes StackOverflow: 0.007335 SuperUser: 0.0047522

P Valor StackOverflow: 0.000000978 SuperUser: 0.0000191

Multiple R-squared StackOverflow: 0.4026 SuperUser: 0.4202

De uma maneira simplificada, o efeito da variável EPI na variável responderam_prop é maior em StackOverflow do que o site SuperUser.

O RQuadrado do modelo SuperUser é maior do que o StackOverflow.

```
t.test(dados_stack$EPI,dados_superuser$EPI)
```

```
##
##  Welch Two Sample t-test
##
## data:  dados_stack$EPI and dados_superuser$EPI
## t = -0.7827, df = 76.694, p-value = 0.4362
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -4.318506  1.881590
## sample estimates:
## mean of x mean of y
##  54.56265  55.78111
```

Por intervalo de confiança:

```
cor(dados_stack$responderam_prop, dados_stack$EPI)
```

```
## [1] 0.6345309
```

```
cor(dados_superuser$responderam_prop, dados_superuser$EPI)
```

```
## [1] 0.6482049
```

```
theta_stack <- function(d, i) {
    r = d %>%
        slice(i) %>%
        summarise(r = cor(responderam_prop, EPI, method = "pearson")) %>%
        pull(r)
    r
}
```

```
ci_stack = boot(data = dados_stack,
           statistic = theta_stack,
           R = 2000) %>%
    tidy(conf.level = .95,
         conf.method = "bca",
         conf.int = TRUE)
```

```
ci_stack
```

```
## # A tibble: 1 x 5
##   statistic    bias std.error conf.low conf.high
##       <dbl>   <dbl>     <dbl>    <dbl>     <dbl>
## 1     0.635 0.00686    0.0701    0.454     0.745
```

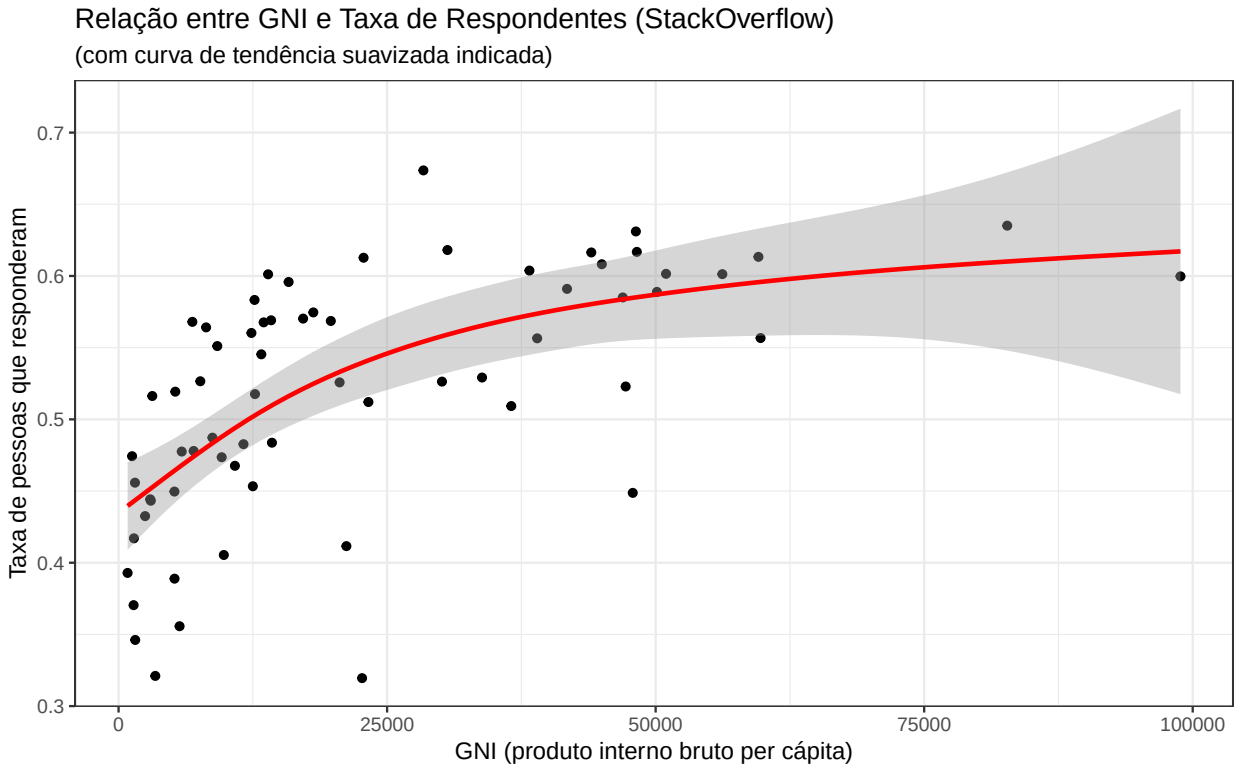
---

3. Descreva a relação entre GNI (produto interno bruto per cápita) dos países e a taxa de pessoas daquele país que responderam alguma pergunta no StackOverflow.

```
dados_stack = dados %>%
        filter(site == "StackOverflow") %>%
        filter(!is.na(responderam_prop)) %>%
        filter(!is.na(GNI))

dados_stack %>%
    ggplot(aes(x = GNI, y = responderam_prop)) +
    geom_point() +
    geom_smooth(method = "gam", color = "red") +
#    geom_smooth(method = "loess") +
    labs(title = "Relação entre GNI e Taxa de Respondentes (StackOverflow)",
         subtitle = "(com curva de tendência suavizada indicada)",
         x = "GNI (produto interno bruto per cápita)",
         y = "Taxa de pessoas que responderam"
         )
```

```
## `geom_smooth()` using formula 'y ~ s(x, bs = "cs")'
```




```
cor(dados_stack$responderam_prop, dados_stack$GNI)
```

```
## [1] 0.5857416
```

```
modelo3 = lm(dados_stack$responderam_prop ~ dados_stack$GNI)

summary(modelo3)
```

```
##
## Call:
## lm(formula = dados_stack$responderam_prop ~ dados_stack$GNI)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.198612 -0.036133  0.003288  0.049280  0.142212
##
## Coefficients:
##                     Estimate   Std. Error t value             Pr(>|t|)
## (Intercept)     0.4652685491 0.0125099164  37.192 < 0.0000000000000002 ***
## dados_stack$GNI 0.0000023313 0.0000004032   5.782          0.000000239 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06893 on 64 degrees of freedom
## Multiple R-squared:  0.3431, Adjusted R-squared:  0.3328
## F-statistic: 33.43 on 1 and 64 DF,  p-value: 0.0000002386
```

```
tidy(modelo3)
```

```
## # A tibble: 2 x 5
##   term              estimate   std.error statistic  p.value
##   <chr>                <dbl>       <dbl>     <dbl>    <dbl>
## 1 (Intercept)     0.465      0.0125          37.2  4.65e-45
## 2 dados_stack$GNI 0.00000233 0.000000403      5.78 2.39e- 7
```

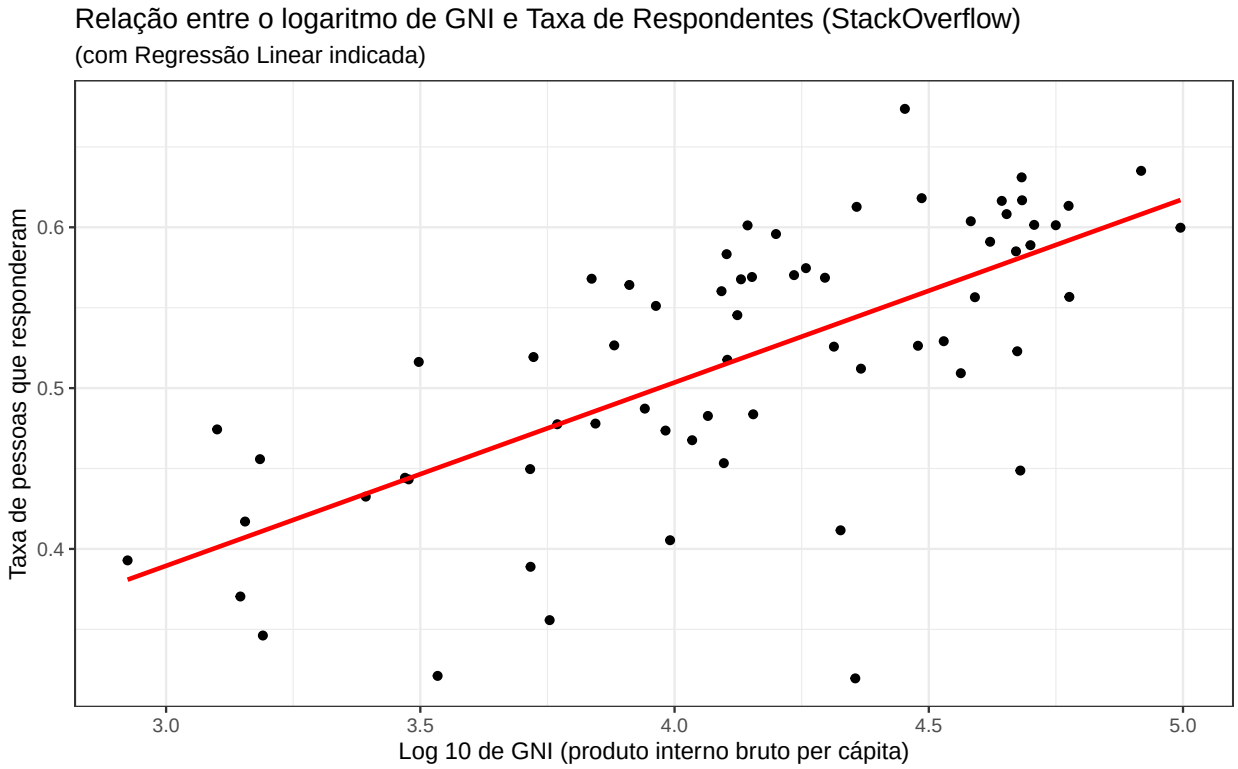
```
glance(modelo3)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared  sigma statistic     p.value    df logLik   AIC   BIC
##       <dbl>         <dbl>  <dbl>     <dbl>       <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.343         0.333 0.0689      33.4 0.000000239     1   83.9 -162. -155.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

A relação não é linear, e uma melhor visualização é alterando a escala para logarítmica do eixo X (GNI).

```
dados_stack %>%
    ggplot(aes(x = log10(GNI), y = responderam_prop)) +
    geom_point() +
    geom_smooth(method = "lm", se = FALSE, color = "red") +
    labs(title = "Relação entre o logaritmo de GNI e Taxa de Respondentes (StackOverflow)",
         subtitle = "(com Regressão Linear indicada)",
         x = "Log 10 de GNI (produto interno bruto per cápita)",
         y = "Taxa de pessoas que responderam"
         )
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
cor(dados_stack$responderam_prop, log10(dados_stack$GNI))
```

```
## [1] 0.6803346
```

```
modelo4 = lm(dados_stack$responderam_prop ~ log10(dados_stack$GNI))

summary(modelo4)
```

```
##
## Call:
## lm(formula = dados_stack$responderam_prop ~ log10(dados_stack$GNI))
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.224533 -0.032858  0.007685  0.044357  0.118479
##
## Coefficients:
##                         Estimate Std. Error t value       Pr(>|t|)
## (Intercept)              0.04747    0.06388   0.743           0.46
## log10(dados_stack$GNI)   0.11401    0.01535   7.426 0.000000000331 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.06233 on 64 degrees of freedom
## Multiple R-squared:  0.4629, Adjusted R-squared:  0.4545
## F-statistic: 55.15 on 1 and 64 DF,  p-value: 0.0000000003311
```

```
tidy(modelo4)
```

```
## # A tibble: 2 x 5
##   term                   estimate std.error statistic  p.value
##   <chr>                     <dbl>     <dbl>     <dbl>    <dbl>
## 1 (Intercept)              0.0475    0.0639     0.743 4.60e- 1
## 2 log10(dados_stack$GNI)   0.114     0.0154     7.43  3.31e-10
```

```
glance(modelo4)
```

```
## # A tibble: 1 x 12
##   r.squared adj.r.squared  sigma statistic  p.value    df logLik   AIC   BIC
##       <dbl>         <dbl>  <dbl>     <dbl>    <dbl> <dbl>  <dbl> <dbl> <dbl>
## 1     0.463         0.454 0.0623      55.1 3.31e-10     1   90.5 -175. -168.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

Refuta-se a hipótese nula de não existir relação (efeito) entre GNI e a taxa de pessoas que responderam.

A cada uma unidade do logaritmo de GNI aumenta 11,4% das pessoas que responderam (IC 95%[11.97;10.83]).

Equação da regressão:

$responderam_prop = 0.04746886 + 0.11401133 * LOG10(GNI)$

A cada uma unidade de EPI, aumenta, em média, 0.475% das pessoas que responderam (IC 95%[0.451;0.499]), com 0,1% de significância.

E o R_Quadrado (coeficiente de determinação) é de 46.3% (R2 = 0.4628552).