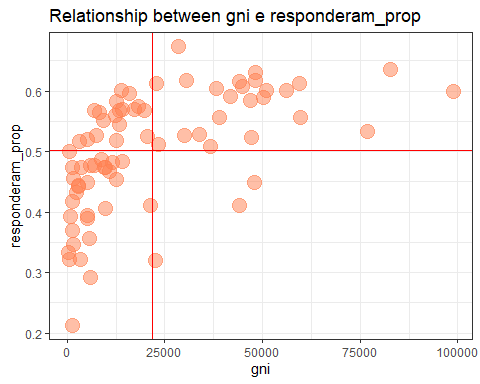
L5P1-Exercício 3

dados\_raw = 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()  
 )  
) %>% janitor::clean\_names()  
  
  
dados = dados\_raw %>%   
 filter(site == "StackOverflow" || site == "SuperUser")  
  
glimpse(dados)

## Rows: 157  
## Columns: 21  
## $ site <chr> "StackOverflow", "StackOverflow", "StackO~  
## $ country <chr> "Argentina", "Australia", "Austria", "Ban~  
## $ pdi <dbl> 49, 36, 11, 80, 65, 69, 70, 39, 63, 80, 6~  
## $ idv <dbl> 46, 90, 55, 20, 75, 38, 30, 80, 23, 20, 1~  
## $ mas <dbl> 56, 61, 79, 55, 54, 49, 40, 52, 28, 66, 6~  
## $ uai <dbl> 86, 51, 70, 60, 94, 76, 85, 48, 86, 30, 8~  
## $ usuarios <dbl> 2798, 12313, 2518, 2558, 4275, 10717, 146~  
## $ responderam\_prop <dbl> 0.5357398, 0.6133355, 0.6310564, 0.392885~  
## $ perguntaram\_prop <dbl> 0.5210865, 0.5897832, 0.5933280, 0.475762~  
## $ editaram\_prop <dbl> 0.09256612, 0.14699911, 0.14932486, 0.080~  
## $ comentaram\_prop <dbl> 0.25339528, 0.33395598, 0.35027800, 0.159~  
## $ gni <dbl> NA, 59570, 48160, 840, 44990, 11630, 6870~  
## $ 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, N~  
## $ geo <chr> "arg", "aus", "aut", "bgd", "bel", "bra",~  
## $ four\_regions <chr> "americas", "asia", "europe", "asia", "eu~  
## $ eight\_regions <chr> "america\_south", "east\_asia\_pacific", "eu~  
## $ six\_regions <chr> "america", "east\_asia\_pacific", "europe\_c~  
## $ latitude <dbl> -34.00000, -25.00000, 47.33333, 24.00000,~  
## $ longitude <dbl> -64.00000, 135.00000, 13.33333, 90.00000,~  
## $ world\_bank\_income\_group\_2017 <chr> "Upper middle income", "High income", "Hi~

relacao\_dados\_StackOverflow = dados %>%  
 filter (is.na(gni) == FALSE, site == "StackOverflow") %>%  
 select(country, site, gni, responderam\_prop)

relacao\_dados\_StackOverflow %>%  
  
ggplot(aes(x = gni, y = responderam\_prop)) +  
 geom\_point(size = 5, alpha = .5, color = "coral") +  
 geom\_hline(yintercept = mean(relacao\_dados\_StackOverflow$responderam\_prop), color = "red") +   
 geom\_vline(xintercept = mean(relacao\_dados\_StackOverflow$gni), color = "red") +  
   
   
 labs(x = "gni", y = "responderam\_prop",  
 title = "Relationship between gni e responderam\_prop")



O Gráfico de dispersão parece indicar uma relação positiva entre GNI e taxa de pessoas daquele país que responderam alguma pergunta no StackOverflow. Entretanto, parece não ser uma relação linear nas variáveis. Para encontrar a reta que melhor se ajusta aos dados, emprega-se o método de Mínimos Quadrados Ordinários (MQO). Esse método minimiza a função que é o somatório dos erros aleatórios ao quadrado. O erro aleatório do modelo é dado por: ei= Yi - Yiestim, onde ei é erro aleatório, Yi é o valor da variável dependente na amostra/real/efetivo e Yiestim é o valor do Y estimado pela reta de regressão. O primeiro modelo estimado é dado abaixo. A reta estimada, que mostra a associação entre as variáveis taxa de pessoas que responderam (Y) e fluência de inglês na população do país (X) é dada por: Y = 4.482229e-01 + 2.478583e-06 X (Yi=b0 +b1.X). O coeficiente de determinação (R^2) mostra quanto da variação do Y é explicada pela variação de X. Nesse primeiro modelo, o R^2 (R-squared) foi de 0.3162905 ou de 31,62%, isto é, 31,62% das variações na taxa de pessoas que responderam são explicadas pelas variações no produto interno bruto per cápita. Um aumento de $1 no produto interno bruto per cápita (X) provoca um aumento de 2.478583e-06 na proporção de pessoas que responderam. Esse impacto na população pode-se ser visto por meio do intervalo de confiança para (B1), que é o parâmetro populacional. Isto é, o impacto na fluência de inglês sobre a taxa de pessoas que responderam está no intervalo de [1.634494e-06 3.322671e-06] com probabilidade de 95%IC.

mod1 = lm (responderam\_prop ~ gni, data = relacao\_dados\_StackOverflow)  
tidy(mod1)

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.448 0.0130 34.5 2.46e-47  
## 2 gni 0.00000248 0.000000424 5.85 1.24e- 7

glance(mod1)

## # 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 0.316 0.307 0.0799 34.2 0.000000124 1 85.2 -164. -157.  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

mod1

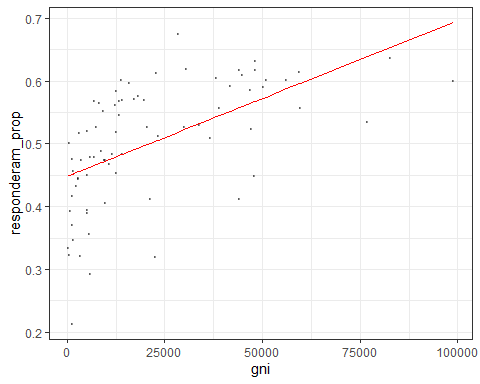
##   
## Call:  
## lm(formula = responderam\_prop ~ gni, data = relacao\_dados\_StackOverflow)  
##   
## Coefficients:  
## (Intercept) gni   
## 4.482e-01 2.479e-06

mod1 %>%   
 augment(relacao\_dados\_StackOverflow)

## # A tibble: 76 x 10  
## country site gni responderam\_prop .fitted .resid .hat .sigma .cooksd  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Austral~ Stack~ 59570 0.613 0.596 1.75e-2 0.0533 0.0805 1.42e-3  
## 2 Austria Stack~ 48160 0.631 0.568 6.35e-2 0.0327 0.0801 1.10e-2  
## 3 Banglad~ Stack~ 840 0.393 0.450 -5.74e-2 0.0254 0.0802 6.91e-3  
## 4 Belgium Stack~ 44990 0.608 0.560 4.85e-2 0.0283 0.0803 5.51e-3  
## 5 Brazil Stack~ 11630 0.483 0.477 5.64e-3 0.0160 0.0805 4.12e-5  
## 6 Bulgaria Stack~ 6870 0.568 0.465 1.03e-1 0.0194 0.0796 1.66e-2  
## 7 Canada Stack~ 50970 0.602 0.575 2.69e-2 0.0371 0.0804 2.27e-3  
## 8 Chile Stack~ 14280 0.484 0.484 1.04e-4 0.0147 0.0805 1.28e-8  
## 9 China Stack~ 5680 0.356 0.462 -1.07e-1 0.0204 0.0795 1.89e-2  
## 10 Colombia Stack~ 6990 0.478 0.466 1.24e-2 0.0193 0.0805 2.41e-4  
## # ... with 66 more rows, and 1 more variable: .std.resid <dbl>

GRÁFICO COM O Y real e y ESTIMADO

mod1 %>%  
 augment(relacao\_dados\_StackOverflow) %>%  
 ggplot(mapping = aes(x = gni)) +  
 geom\_point(aes(y = responderam\_prop), alpha = 0.4, size = .1) +  
 geom\_line(aes(y = .fitted), colour = "red")



NULL

## NULL

Inferência

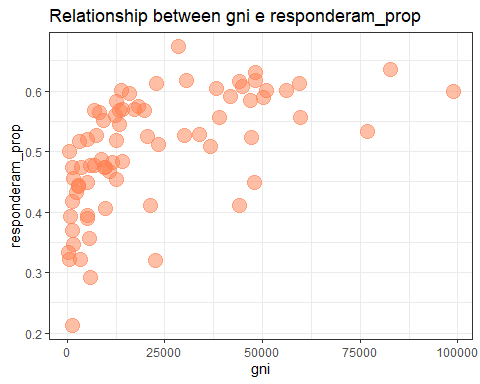
mod1 %>%  
 tidy(conf.int = T, conf.level = 0.95) %>%  
 select(-p.value)

## # A tibble: 2 x 6  
## term estimate std.error statistic conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.448 0.0130 34.5 0.422 0.474   
## 2 gni 0.00000248 0.000000424 5.85 0.00000163 0.00000332

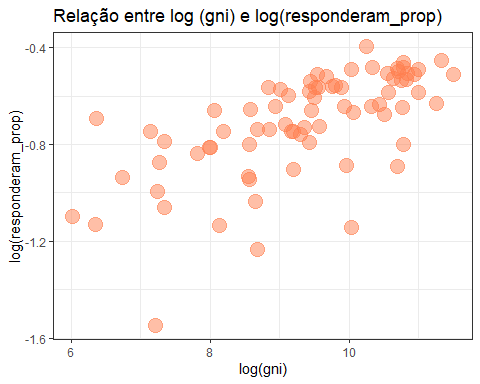
Modelo utlizando o log das variaveis

O gráfico de dispersão mostra comportamentos das variáveis respoderam\_prop e GNI . Considera-se respoderam\_prop como variável dependente e GNI como variáveil independente. O gráfico de dispersão mostra uma possível relação positiva entre essas variáveis por visualizar que seus pontos amostrais indicam uma reta positivamente inclinada. Também, é possível que a melhor função matemática que descreva a relação entre essas duas variáveis possa não ser um modelo linear. Inicialmente, estima-se um modelo linear com as variáveis em nível, ou seja, com seus valores originais. Em seguida, estima-se um modelo transformando as variáveis (dependente e independete) em logarítimo para linearizar essa relação.

relacao\_dados\_StackOverflow %>%  
ggplot(aes(x = gni, y = responderam\_prop)) +  
 geom\_point(size = 5, alpha = .5, color = "coral") +  
  
 labs(x = "gni", y = "responderam\_prop",  
 title = "Relationship between gni e responderam\_prop")



relacao\_dados\_StackOverflow2 = relacao\_dados\_StackOverflow %>%  
 mutate (gni = log (gni), responderam\_prop = log(responderam\_prop))  
  
  
relacao\_dados\_StackOverflow2 %>%  
ggplot(aes(x = gni, y = responderam\_prop)) +  
 geom\_point(size = 5, alpha = .5, color = "coral") +  
  
 labs(x = "log(gni)", y = "log(responderam\_prop)",  
 title = "Relação entre log (gni) e log(responderam\_prop)")



O segundo modelo foi obtido estimando a reta de regressão transformando as variáveis em log. A reta estimada, que mostra a associação entre as variáveis taxa de pessoas que responderam (Y) e GNI na população do país (X) é dada por: logY = -1.7067554 + 0.1066706 logX (logYi=b0 +b1.logX). O coeficiente de determinação (R^2 ou (R-squared) foi de 0.4261386 ou de 42,61%, isto é, 42,61% das variações na taxa de pessoas que responderam são explicadas pelas variações no GNI.Não indicando um bom ajuste da reta de regressão ou modelo aos dados. Um aumento de 1% no GNI (X) provoca um aumento de 0,1066% na proporção de pessoas que responderam. Esse impacto na população pode-se ser visto por meio do intervalo de confiança para (B1), que é o parâmetro populacional. Isto é, o impacto na fluência de inglês sobre a taxa de pessoas que responderam está no intervalo de [0.07799815 0.135343] com probabilidade de 95%IC.

mod1 = lm (responderam\_prop ~ gni, data = relacao\_dados\_StackOverflow2)  
tidy(mod1)

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -1.71 0.136 -12.6 4.86e-20  
## 2 gni 0.107 0.0144 7.41 1.66e-10

glance(mod1)

## # 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 0.426 0.418 0.164 55.0 1.66e-10 1 30.3 -54.7 -47.7  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

mod1

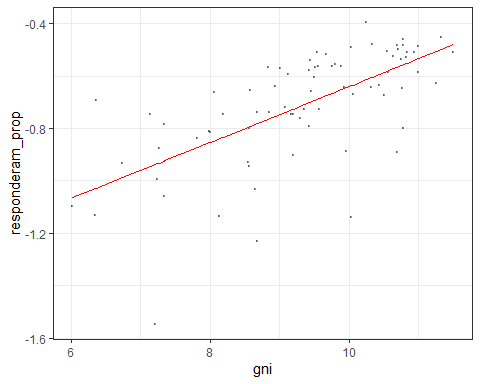
##   
## Call:  
## lm(formula = responderam\_prop ~ gni, data = relacao\_dados\_StackOverflow2)  
##   
## Coefficients:  
## (Intercept) gni   
## -1.7068 0.1067

mod1 %>%   
 augment(relacao\_dados\_StackOverflow2)

## # A tibble: 76 x 10  
## country site gni responderam\_prop .fitted .resid .hat .sigma .cooksd  
## <chr> <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Austral~ StackO~ 11.0 -0.489 -0.534 0.0451 0.0340 0.166 1.37e-3  
## 2 Austria StackO~ 10.8 -0.460 -0.557 0.0962 0.0290 0.165 5.26e-3  
## 3 Banglad~ StackO~ 6.73 -0.934 -0.988 0.0543 0.0654 0.165 4.07e-3  
## 4 Belgium StackO~ 10.7 -0.497 -0.564 0.0666 0.0275 0.165 2.38e-3  
## 5 Brazil StackO~ 9.36 -0.728 -0.708 -0.0202 0.0132 0.166 1.02e-4  
## 6 Bulgaria StackO~ 8.83 -0.566 -0.764 0.199 0.0151 0.164 1.14e-2  
## 7 Canada StackO~ 10.8 -0.508 -0.551 0.0422 0.0302 0.166 1.06e-3  
## 8 Chile StackO~ 9.57 -0.726 -0.686 -0.0400 0.0135 0.166 4.11e-4  
## 9 China StackO~ 8.64 -1.03 -0.785 -0.249 0.0169 0.163 2.00e-2  
## 10 Colombia StackO~ 8.85 -0.738 -0.762 0.0242 0.0150 0.166 1.68e-4  
## # ... with 66 more rows, and 1 more variable: .std.resid <dbl>

GRÁFICO COM O Y real e y ESTIMADO

mod1 %>%  
 augment(relacao\_dados\_StackOverflow2) %>%  
 ggplot(mapping = aes(x = gni)) +  
 geom\_point(aes(y = responderam\_prop), alpha = 0.4, size = .1) +  
 geom\_line(aes(y = .fitted), colour = "red")



NULL

## NULL

Inferência

mod1 %>%  
 tidy(conf.int = T, conf.level = 0.95) %>%  
 select(-p.value)

## # A tibble: 2 x 6  
## term estimate std.error statistic conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -1.71 0.136 -12.6 -1.98 -1.44   
## 2 gni 0.107 0.0144 7.41 0.0780 0.135