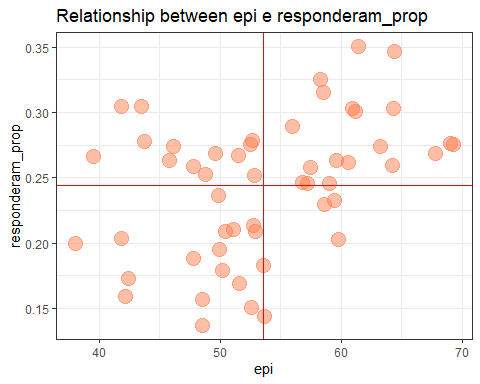
L5P1-Exercício 2

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\_SuperUser = dados %>%  
 filter (is.na(epi) == FALSE, site == "SuperUser") %>%  
 select(site, epi, responderam\_prop)

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



O gráfico de dispersão mostra comportamentos das variáveis respoderam\_prop e EPI na amostra SuperUser. Considera-se respoderam\_prop como variável dependente e EPI 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, porém com bastante disperção dos pontos amostrais. 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.

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 = 0.09851818 + 0.00271697 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 0.16018 ou de 16,01%, isto é, 16,01% das variações na taxa de pessoas que responderam são explicadas pelas variações na fluência de inglês. Não houve um bom ajuste do modelo ao dados, a variável dependente parece ser explicada por outras variáveis que não estão no modelo. Um aumento de 1% na fluência de inglês (X) provoca um aumento de 0.0027 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 [0.1751 0.9965] com probabilidade de 95%IC.

mod1 = lm (responderam\_prop ~ epi, data = relacao\_dados\_SuperUser)  
tidy(mod1)

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.0985 0.0472 2.09 0.0417   
## 2 epi 0.00272 0.000871 3.12 0.00298

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 0.160 0.144 0.0486 9.73 0.00298 1 86.1 -166. -160.  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

mod1

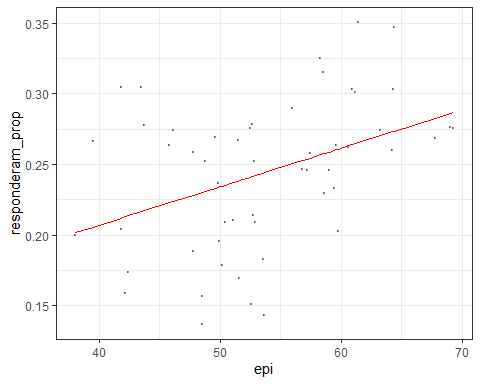
##   
## Call:  
## lm(formula = responderam\_prop ~ epi, data = relacao\_dados\_SuperUser)  
##   
## Coefficients:  
## (Intercept) epi   
## 0.098518 0.002717

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

## # A tibble: 53 x 9  
## site epi responderam\_prop .fitted .resid .hat .sigma .cooksd  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 SuperUser 59.0 0.246 0.259 -0.0129 0.0283 0.0491 0.00105   
## 2 SuperUser 63.2 0.274 0.270 0.00410 0.0485 0.0491 0.000191   
## 3 SuperUser 61.2 0.301 0.265 0.0361 0.0375 0.0488 0.0111   
## 4 SuperUser 50.0 0.195 0.234 -0.0389 0.0231 0.0488 0.00775   
## 5 SuperUser 48.8 0.252 0.231 0.0214 0.0264 0.0490 0.00269   
## 6 SuperUser 50.2 0.179 0.235 -0.0560 0.0227 0.0485 0.0157   
## 7 SuperUser 48.5 0.137 0.230 -0.0934 0.0271 0.0473 0.0527   
## 8 SuperUser 48.5 0.157 0.230 -0.0735 0.0271 0.0480 0.0327   
## 9 SuperUser 57.4 0.258 0.255 0.00335 0.0236 0.0491 0.0000587  
## 10 SuperUser 69.3 0.276 0.287 -0.0111 0.0980 0.0491 0.00311   
## # ... with 43 more rows, and 1 more variable: .std.resid <dbl>

GRÁFICO COM O Y real e y ESTIMADO

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



NULL

## NULL

Inferência

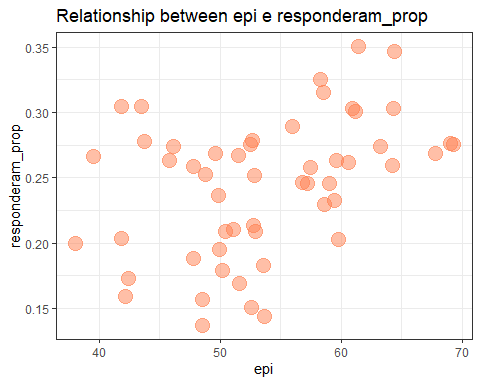
mod1 %>%  
 tidy(conf.int = T, conf.level = 0.95) %>%  
 select(-p.value)

## # A tibble: 2 x 6  
## term estimate std.error statistic conf.low conf.high  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.0985 0.0472 2.09 0.00383 0.193   
## 2 epi 0.00272 0.000871 3.12 0.000968 0.00447

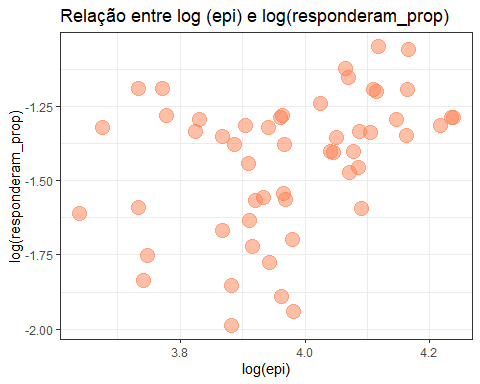
Modelo utlizando o log das variaveis

O gráfico de dispersão mostra comportamentos das variáveis respoderam\_prop e EPI . Considera-se respoderam\_prop como variável dependente e EPI 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\_SuperUser %>%  
ggplot(aes(x = epi, y = responderam\_prop)) +  
 geom\_point(size = 5, alpha = .5, color = "coral") +  
  
 labs(x = "epi", y = "responderam\_prop",  
 title = "Relationship between epi e responderam\_prop")



relacao\_dados\_SuperUser2 = relacao\_dados\_SuperUser %>%  
 mutate (epi = log (epi), responderam\_prop = log(responderam\_prop))  
  
  
relacao\_dados\_SuperUser2 %>%  
ggplot(aes(x = epi, y = responderam\_prop)) +  
 geom\_point(size = 5, alpha = .5, color = "coral") +  
  
 labs(x = "log(epi)", y = "log(responderam\_prop)",  
 title = "Relação entre log (epi) 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 fluência de inglês na população do país (X) é dada por: logY = -3.7613143 + 0.5858449 logX (logYi=b0 +b1.logX). O coeficiente de determinação (R^2 ou (R-squared) foi de 0.1385248 ou de 13.85%, isto é, 13.85% das variações na taxa de pessoas que responderam são explicadas pelas variações na fluência de inglês.Não indicando um bom ajuste da reta de regressão ou modelo aos dados. Um aumento de 1% na fluência de inglês (X) provoca um aumento de 0,58% 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 [0.1751409 0.9965489] com probabilidade de 95%IC.

mod1 = lm (responderam\_prop ~ epi, data = relacao\_dados\_SuperUser2)  
tidy(mod1)

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -3.76 0.813 -4.63 0.0000258  
## 2 epi 0.586 0.205 2.86 0.00606

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 0.139 0.122 0.217 8.20 0.00606 1 6.90 -7.80 -1.89  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

mod1

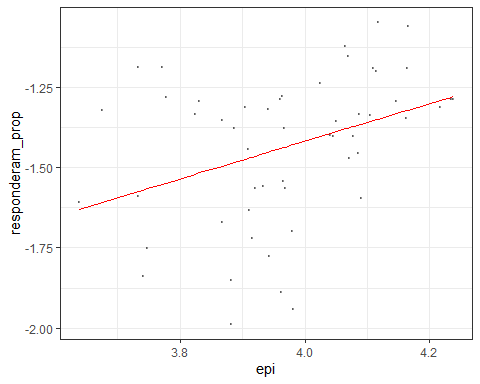
##   
## Call:  
## lm(formula = responderam\_prop ~ epi, data = relacao\_dados\_SuperUser2)  
##   
## Coefficients:  
## (Intercept) epi   
## -3.7613 0.5858

mod1 %>%   
 augment(relacao\_dados\_SuperUser2)

## # A tibble: 53 x 9  
## site epi responderam\_prop .fitted .resid .hat .sigma .cooksd .std.resid  
## <chr> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 Supe~ 4.08 -1.40 -1.37 -0.0302 0.0291 0.219 2.99e-4 -0.141   
## 2 Supe~ 4.15 -1.29 -1.33 0.0388 0.0463 0.219 8.19e-4 0.184   
## 3 Supe~ 4.11 -1.20 -1.35 0.150 0.0372 0.218 9.63e-3 0.706   
## 4 Supe~ 3.91 -1.63 -1.47 -0.163 0.0221 0.217 6.54e-3 -0.761   
## 5 Supe~ 3.89 -1.38 -1.48 0.107 0.0252 0.218 3.26e-3 0.502   
## 6 Supe~ 3.92 -1.72 -1.47 -0.254 0.0217 0.216 1.55e-2 -1.18   
## 7 Supe~ 3.88 -1.99 -1.49 -0.501 0.0259 0.207 7.30e-2 -2.34   
## 8 Supe~ 3.88 -1.85 -1.49 -0.365 0.0259 0.212 3.89e-2 -1.71   
## 9 Supe~ 4.05 -1.36 -1.39 0.0332 0.0245 0.219 3.01e-4 0.155   
## 10 Supe~ 4.24 -1.29 -1.28 -0.0100 0.0827 0.219 1.05e-4 -0.0483  
## # ... with 43 more rows

GRÁFICO COM O Y real e y ESTIMADO

mod1 %>%  
 augment(relacao\_dados\_SuperUser2) %>%  
 ggplot(mapping = aes(x = epi)) +  
 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) -3.76 0.813 -4.63 -5.39 -2.13   
## 2 epi 0.586 0.205 2.86 0.175 0.997

**Os modelos tentam explicar o comportamento da taxa de resposta das pessoas em diferentes países por meio da fluência em inglês dos paises em duas em dois grupos (sites: “StackOverflow” e “SuperUser”). Comparando o R^2 (R-squared) dos dois modelos, verificou-se um R-squared maior usando o site StackOverflow (0.4029 ou de 40,29%) relativo ao R-squared do modelo usando o site SuperUser (0.1385248 ou de 13.85%). Portanto, o modelo usando o site StackOverflow teve um melhor ajuste aos dados.**