L5P1 - Regressão com uso do Stack Overflow e Super User

Rodolfo Bolconte

31/05/21

Nessa parte do laboratório, queremos que você crie 3 modelos de regressão linear simples usando os dados de atividade de pessaos de diferentes países nos sits StackOverflow e SuperUser.

Submeta um google doc como resultado. Use o parágrafo modelo que mostro no 2o vídeo de regressão e visualizações. Cuidado com a necessidade de transformações para que as relações que você modelará sejam lineares. Opcionalmente, faça inferência e comente os ICs dos coeficientes encontrados (basta o b\_1, não precisa do b\_0).

## Carregamento dos Dados

O repositório com os dados: <https://github.com/cienciadedados-ufcg/vis-cultura-stackoverflow>

dataset = read\_csv(here::here("data/participation-per-country.csv"))

## Questão 1

**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.**

dados\_epi = dataset %>%   
 select(site, country, eight\_regions, responderam\_prop, EPI) %>%   
 filter(EPI != "NA")  
  
q1\_dados = dados\_epi %>% filter(site=="StackOverflow")  
  
q1\_sumarios = q1\_dados %>%   
 group\_by(eight\_regions) %>%   
 summarise(resp\_min=min(responderam\_prop),  
 resp\_med=mean(responderam\_prop),  
 resp\_max=max(responderam\_prop),  
 epi\_min=min(EPI),  
 epi\_med=mean(EPI),  
 epi\_max=max(EPI))

q1\_graf\_resp = q1\_sumarios %>%   
 ggplot(aes(y=reorder(eight\_regions, resp\_min))) +  
 geom\_point(aes(resp\_min), color='red', alpha=.5, size=4) +  
 geom\_point(aes(resp\_med), color='green', alpha=.5, size=4) +  
 geom\_point(aes(resp\_max), color='blue', alpha=.5, size=4) +  
 labs(y='Regiões Mundiais\n', x='\nÍndice de Respostas', subtitle='(a)') +  
 theme(text=element\_text(size=16),  
 axis.text.x=element\_text(angle=90, vjust=.5))  
  
q1\_graf\_epi = q1\_sumarios %>%   
 ggplot(aes(y=reorder(eight\_regions, resp\_min))) +  
 geom\_point(aes(epi\_min, color='epi\_min'), alpha=.5, size=4) +  
 geom\_point(aes(epi\_med, color='epi\_med'), alpha=.5, size=4) +  
 geom\_point(aes(epi\_max, color='epi\_max'), alpha=.5, size=4) +  
 scale\_color\_manual(name='EPI:',  
 values=c('epi\_min'='red', 'epi\_med'='green', 'epi\_max'='blue'),  
 labels=c('Máximo', 'Médio', 'Mínimo')) +  
 labs(y=NULL, x='\nEPI', subtitle='(b)') +  
 theme(text=element\_text(size=16),  
 axis.text.x=element\_text(angle=90, vjust=.5)) +  
 scale\_y\_discrete(labels=c(NULL, NULL,NULL, NULL,NULL, NULL,NULL, NULL))  
   
  
grid.arrange(q1\_graf\_resp, q1\_graf\_epi, ncol=2, widths=c(1.3,1))

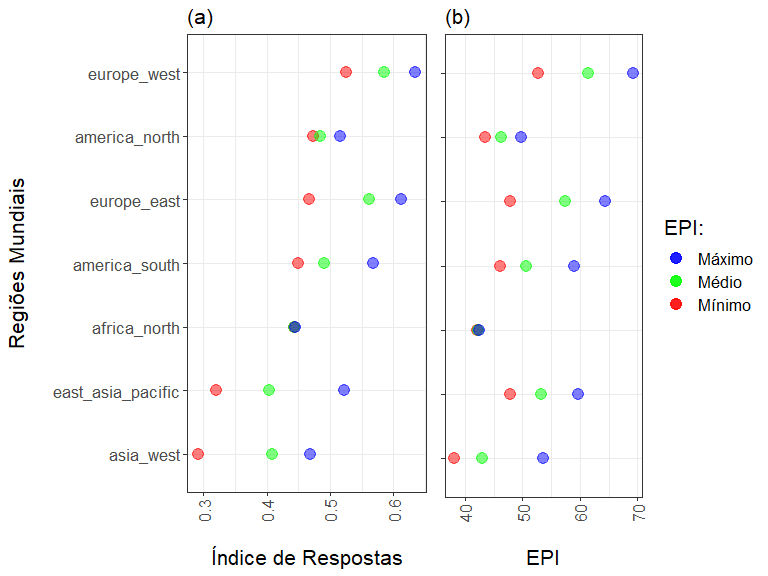


Gráfico 1.1: Visualização dos valores mínimos, médios e máximos do (a) Índice de Respostas das Perguntas e (b) English Proficiency Index (EPI), sendo valores referentes à países, porém agrupados em 7 regiões geográficas mundiais. Dados referentes ao site Stack Overflow.

No Gráfico 1.1 são feitas análises dos valores dos Índices de Respostas no site do Stack Overflow e de English Proficiency Index (EPI), ambas variáveis de países sendo agrupadas por 7 regiões geográficas mundiais. Os valores foram sumarizados para apresentar os valores mínimos (Ponto Vermelho), máximos (Ponto Azul) e a média (Ponto Verde).

A partir do gráfico é possível perceber que países do Oeste Europeu apresentam os maiores valores para os 3 sumários tanto no índice de respostas quanto em EPI, enquanto os menores valores são mais diversificados: para o índice de respostas, o menor valor mínimo se concentra em um país do Oeste Asiático, porém a menor média é alcançada por países do Leste Asiático do Pacífico. Para os valores EPI, o menor valor mínimo é encontrado no Oeste Asiático também, porém a menor média e concentra nos países do Norte Africano.

Vale destacar os valores bem próximos dos 3 sumários nas duas variáveis para os países do Norte da África, enquanto nas demais regiões os valores dos sumários são mais distantes uns dos outros.

q1\_modelo = lm(responderam\_prop ~ EPI, data=q1\_dados) # responderam\_prop = b0 + b1 \* EPI  
  
tidy(q1\_modelo)

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.0975 0.0625 1.56 0.125   
## 2 EPI 0.00753 0.00115 6.52 0.0000000310

glance(q1\_modelo)

## # 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 0.455 0.444 0.0645 42.5 0.0000000310 1 71.1 -136. -130.  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

q1\_aug = q1\_modelo %>% augment(q1\_dados)  
  
q1\_dados %>%  
 ggplot(aes(responderam\_prop, EPI)) +  
 geom\_point(size=3, alpha=.5) +  
 geom\_line(aes(x=q1\_aug$.fitted), color='red') +  
 labs(x='\nÍndice de Resposta de Questões', y='EPI\n') +  
 theme(text=element\_text(size=16))

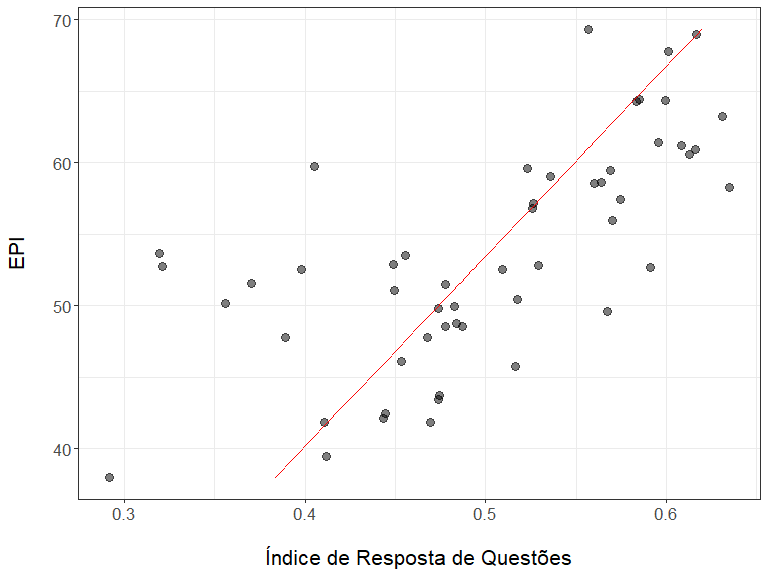


Gráfico 1.2: Visualização da dispersão de valores para o Índice de Respostas no site Stack Overflow e o valor de EPI, com um modelo linear calculado e visualizado pela linha na cor vermelha.

Na questão foi utilizada a Regressão Linear Simples para amostrar e analisar a associação entre duas variáveis: Índice de Resposta de Questões e EPI a partir de usuários de diversos países que respondem à perguntas no site Stack Overflow.

No Gráfico 1.2, foi criado um modelo linear resultando na função:

,

que apresenta 45,5% da variância da variável, sendo .

A partir do modelo calculado, é esperado que o aumento do valor de EPI gere um acréscimo em torno de 0,0075 do valor do índice de respostas.

## Questão 2

**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.**

q2\_dados = dados\_epi %>% filter(site=="SuperUser")  
  
q2\_sumarios = q2\_dados %>%   
 group\_by(eight\_regions) %>%   
 summarise(resp\_min=min(responderam\_prop),  
 resp\_med=mean(responderam\_prop),  
 resp\_max=max(responderam\_prop),  
 epi\_min=min(EPI),  
 epi\_med=mean(EPI),  
 epi\_max=max(EPI))

q2\_graf\_resp = q2\_sumarios %>%   
 ggplot(aes(y=reorder(eight\_regions, resp\_min))) +  
 geom\_point(aes(resp\_min), color='red', alpha=.5, size=4) +  
 geom\_point(aes(resp\_med), color='green', alpha=.5, size=4) +  
 geom\_point(aes(resp\_max), color='blue', alpha=.5, size=4) +  
 labs(y='Regiões Mundiais\n', x='\nÍndice de Respostas', subtitle='(a)') +  
 theme(text=element\_text(size=16),  
 axis.text.x=element\_text(angle=90, vjust=.5))  
  
q2\_graf\_epi = q2\_sumarios %>%   
 ggplot(aes(y=reorder(eight\_regions, resp\_min))) +  
 geom\_point(aes(epi\_min, color='epi\_min'), alpha=.5, size=4) +  
 geom\_point(aes(epi\_med, color='epi\_med'), alpha=.5, size=4) +  
 geom\_point(aes(epi\_max, color='epi\_max'), alpha=.5, size=4) +  
 scale\_color\_manual(name='EPI:',  
 values=c('epi\_min'='red', 'epi\_med'='green', 'epi\_max'='blue'),  
 labels=c('Máximo', 'Médio', 'Mínimo')) +  
 labs(y=NULL, x='\nEPI', subtitle='(b)') +  
 theme(text=element\_text(size=16),  
 axis.text.x=element\_text(angle=90, vjust=.5)) +  
 scale\_y\_discrete(labels=c(NULL, NULL,NULL, NULL,NULL, NULL,NULL, NULL)) +  
 scale\_x\_continuous(breaks=seq(40.5,70.5,10))  
   
  
grid.arrange(q2\_graf\_resp, q2\_graf\_epi, ncol=2, widths=c(1.3,1))

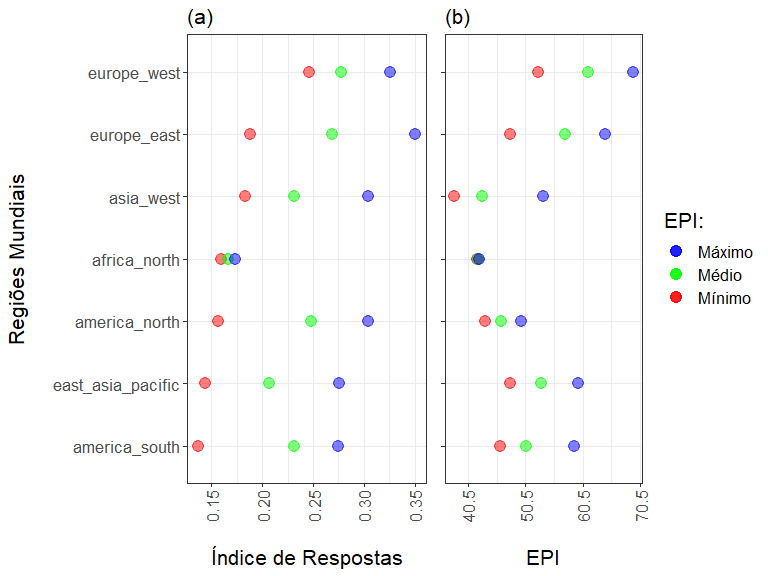


Gráfico 2.1: Visualização dos valores mínimos, médios e máximos do (a) Índice de Respostas das Perguntas e (b) English Proficiency Index (EPI), sendo valores referentes à países, porém agrupados em 7 regiões geográficas mundiais. Dados referentes ao site Super User.

No Gráfico 2.1 são feitas análises dos valores dos Índices de Respostas e de EPI, assim como na Questão 1 (com mesma disposição de amostragem), porém os dados são referentes ao site Super User.

Assim como os valores do site Stack Overflow, países do Oeste Europeu apresentam os maiores valores para o valor mínimo e médio no índice de respostas, porém o maior valor é atingido por país do Leste Europeu. Já os menores valores do índice se diverificam, sendo menor valor mínimo atingido por um país da América do Sul e a menor média um por país do Norte Africano.

Assim como no site do Stack Overflow, os países do Norte da África apresentam valores bem próximos dos 3 sumários quanto ao índice de respostas.

q2\_modelo = lm(responderam\_prop ~ EPI, data=q2\_dados) # responderam\_prop = b0 + b1 \* EPI  
  
tidy(q2\_modelo)

## # 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(q2\_modelo)

## # 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>

q2\_aug = q2\_modelo %>% augment(q2\_dados)  
  
q2\_dados %>%  
 ggplot(aes(responderam\_prop, EPI)) +  
 geom\_point(size=3, alpha=.5) +  
 geom\_line(aes(x=q2\_aug$.fitted), color='red') +  
 labs(x='\nÍndice de Resposta de Questões', y='EPI\n') +  
 theme(text=element\_text(size=16))

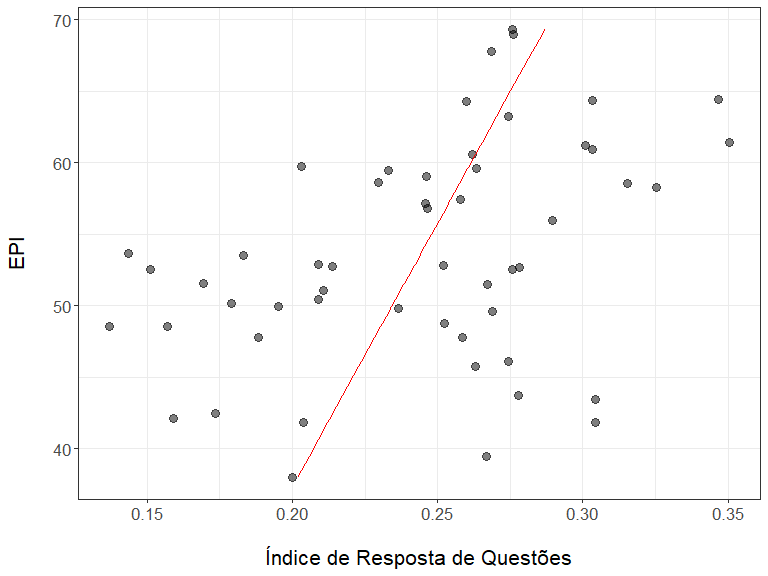


Gráfico 2.2: Visualização da dispersão de valores para o Índice de Respostas no site Super User e o valor de EPI, com um modelo linear calculado e visualizado pela linha na cor vermelha.

Na questão foi utilizada a Regressão Linear Simples para a análise da associação das variáveis: Índice de Resposta de Questões e EPI a partir de usuários de diversos países que respondem à perguntas no Super User.

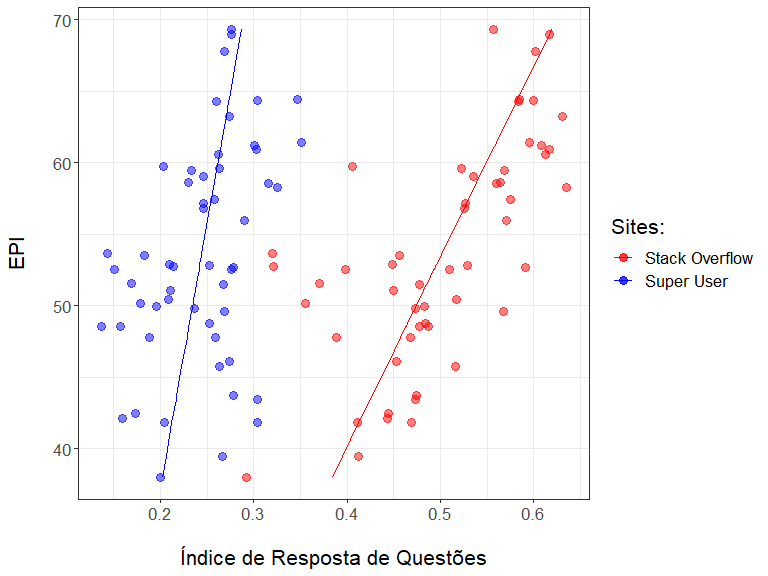
No Gráfico 2.2, foi criado um modelo linear resultando na função:

,

que apresenta 16% da variância da variável, sendo .

A partir do modelo calculado, é esperado que o aumento do valor de EPI gere um acréscimo em torno de 0,0027 do valor do índice de respostas.

ggplot() +  
 geom\_point(data=q1\_dados, aes(responderam\_prop, EPI, color='so'), size=3, alpha=.5) +  
 geom\_line(data=q1\_aug, aes(x=.fitted, y=EPI, color='so')) +  
 geom\_point(data=q2\_dados, aes(responderam\_prop, EPI, color='su'), size=3, alpha=.5) +  
 geom\_line(data=q2\_aug, aes(x=.fitted, y=EPI, color='su')) +  
 labs(x='\nÍndice de Resposta de Questões', y='EPI\n') +  
 scale\_color\_manual(name='Sites:',  
 values=c('so'='red', 'su'='blue'),  
 labels=c('Stack Overflow', 'Super User')) +  
 theme(text=element\_text(size=16))

 Os valores

Para o índice de respostas do Stack Overflow tem-se os seguintes valores:

Para o índice de respostas do Super User tem-se:

Comparando os valores dos coeficientes e (Visualização no Gráfico 2.3) dos modelos lineares dos dois sites, ambos são positivos, embora visualmente se mostrem bem verticais. É importante destacar uma pequena diferença entre os valores de , uma vez que existem valores EPI para países no site Super User que não são contabilizados no site Stack Overflow e vice-versa. Já com relação à , é perceptível uma maior variação do valor do índice de respostas em relação ao valor de EPI para o site do Stack Overflow, visto que o valor de se mostra mais alto em relação ao do Super User.

Comparando os valores de é ainda mais evidente a diferença entre os modelos lineares para os sites. O Stack Overflow atingiu 0,455 de variação, o que implica que o modelo linear é capaz de explicar em torno de 45,5% o valor do índice de resposta a partir do valor de EPI, o que pode ser uma relação média se levarmos em conta uma porcentagem de até 50% para isto. No Super User, é verificada uma variação de 0,160, ou seja, o modelo é 16% capaz de explicar o índice de respostas utilizando a variável EPI, sendo uma relação pequena se for levado em conta o valor de até 20% para tal.

## Questão 3

**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.**

q3\_dados = dataset %>%   
 select(site, country, eight\_regions, responderam\_prop, GNI) %>%   
 filter(GNI != "NA" & responderam\_prop != "NA" & site=="StackOverflow")  
  
q3\_sumarios = q3\_dados %>%   
 group\_by(eight\_regions) %>%   
 summarise(resp\_min=min(responderam\_prop),  
 resp\_med=mean(responderam\_prop),  
 resp\_max=max(responderam\_prop),  
 gni\_min=min(GNI),  
 gni\_med=mean(GNI),  
 gni\_max=max(GNI))

q3\_graf\_resp = q3\_sumarios %>%   
 ggplot(aes(y=reorder(eight\_regions, gni\_min))) +  
 geom\_point(aes(resp\_min), color='red', alpha=.5, size=4) +  
 geom\_point(aes(resp\_med), color='green', alpha=.5, size=4) +  
 geom\_point(aes(resp\_max), color='blue', alpha=.5, size=4) +  
 labs(y='Regiões Mundiais\n', x='\nÍndice de\nRespostas', subtitle='(a)') +  
 theme(text=element\_text(size=16),  
 axis.text.x=element\_text(angle=90, vjust=.5)) +  
 scale\_x\_continuous(breaks=seq(.25,.75,.1))  
  
q3\_graf\_gni = q3\_sumarios %>%   
 ggplot(aes(y=reorder(eight\_regions, gni\_min))) +  
 geom\_point(aes(gni\_min, color='gni\_min'), alpha=.5, size=4) +  
 geom\_point(aes(gni\_med, color='gni\_med'), alpha=.5, size=4) +  
 geom\_point(aes(gni\_max, color='gni\_max'), alpha=.5, size=4) +  
 scale\_color\_manual(name='GNI:',  
 values=c('gni\_min'='red', 'gni\_med'='green', 'gni\_max'='blue'),  
 labels=c('Máximo', 'Médio', 'Mínimo')) +  
 labs(y=NULL, x='\nGNI', subtitle='(b)') +  
 theme(text=element\_text(size=16),  
 axis.text.x=element\_text(angle=90, vjust=.5)) +  
 scale\_y\_discrete(labels=c(NULL, NULL,NULL, NULL,NULL, NULL,NULL, NULL))  
  
grid.arrange(q3\_graf\_resp, q3\_graf\_gni, ncol=2, widths=c(1.3,1))

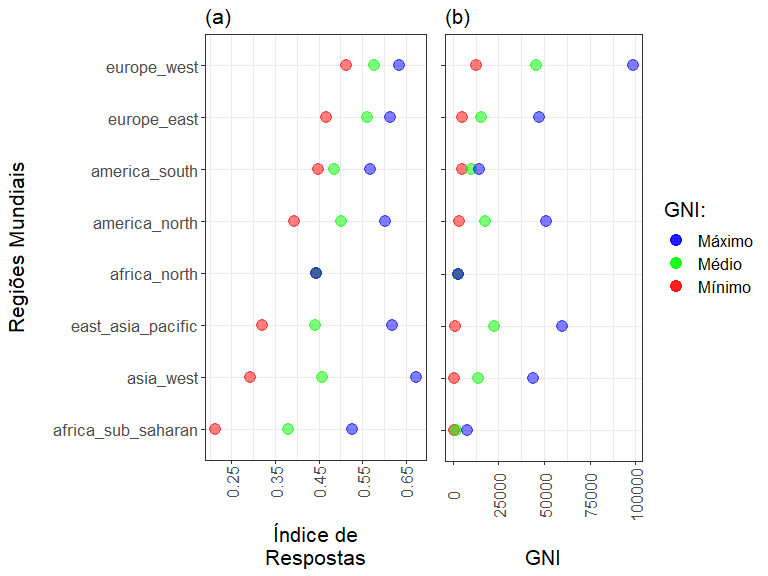


Gráfico 3.1: Visualização dos valores mínimos, médios e máximos do (a) Índice de Respostas das Perguntas e (b) Gross National Income (GNI), sendo valores referentes à países, porém agrupados em 8 regiões geográficas mundiais. Dados referentes ao site Stack Overflow.

No Gráfico 3.1 são feitas análises dos valores dos Índices de Respostas do site Stack Overflow e de Gross National Income (GNI), com a mesma disposição visual das questões anteriores.

Como os índices de respostas do Stack Overflow já foram discutidos na Questão 1, o foco desta discussão são os valores de GNI, em que os menores valores mínimo e médio são apresentados por países da África Subsariana, já o menor valor máximo é atingido por país do Leste Asiático do Pacífico. Já os maiores valores dos sumários de GNI são atingidos por países do Oeste Europeu, curiosamente com um país atingindo um valor próximo de 100000 de GNI.

Assim como para os valores de EPI, os países do Norte da África apresentam valores bem próximos dos 3 sumários quanto aos valores de GNI.

q3\_modelo = lm(responderam\_prop ~ log10(GNI), data=q3\_dados) # responderam\_prop = b0 + b1 \* GNI  
  
tidy(q3\_modelo)

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) 0.0422 0.0588 0.718 4.75e- 1  
## 2 log10(GNI) 0.113 0.0143 7.90 1.99e-11

glance(q3\_modelo)

## # 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 0.458 0.450 0.0712 62.4 1.99e-11 1 94.0 -182. -175.  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

q3\_aug = q3\_modelo %>% augment(q3\_dados)  
  
q3\_dados %>%  
 ggplot(aes(responderam\_prop, GNI)) +  
 geom\_point(size=3, alpha=.5) +  
 scale\_y\_log10() +  
 geom\_line(aes(x=q3\_aug$.fitted), color='red') +  
 labs(x='\nÍndice de Resposta de Questões', y='GNI\n') +  
 theme(text=element\_text(size=16))

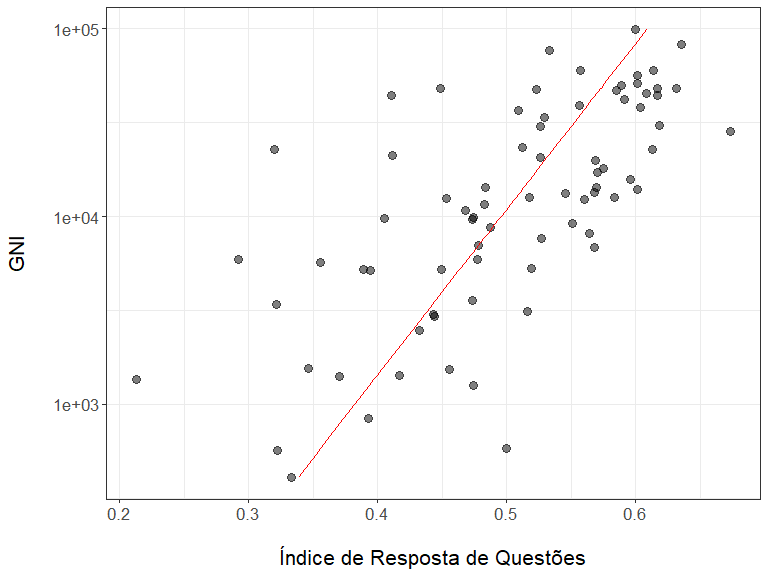


Gráfico 3.2: Visualização da dispersão de valores para o Índice de Respostas no site Stack Overflow e o valor de , com um modelo linear calculado e visualizado pela linha na cor vermelha.

Na questão também foi utilizada a Regressão Linear Simples para a análise de variáveis, sendo elas: Índice de Resposta de Questões e GNI a partir de usuários de diversos países que respondem à perguntas no Stack Overflow.

No Gráfico 3.2, foi criado um modelo linear resultando na função:

,

que apresenta 16% da variância da variável, sendo . Foi realizado o cálculo de log na base 10 para o valor GNI como forma de uma melhor visualização do modelo linear a partir do gráfico de dispersão dos valores.

A partir do modelo calculado, é esperado que o aumento do valor de GNI gere um acréscimo em torno de 0,1133 do valor do índice de respostas.

## Questão 4

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

q4\_dados = dataset %>%   
 select(site, country, eight\_regions, responderam\_prop, GNI) %>%   
 filter(GNI != "NA" & responderam\_prop != "NA" & site=="SuperUser")  
  
q4\_sumarios = q4\_dados %>%   
 group\_by(eight\_regions) %>%   
 summarise(resp\_min=min(responderam\_prop),  
 resp\_med=mean(responderam\_prop),  
 resp\_max=max(responderam\_prop),  
 gni\_min=min(GNI),  
 gni\_med=mean(GNI),  
 gni\_max=max(GNI))

q4\_graf\_resp = q4\_sumarios %>%   
 ggplot(aes(y=reorder(eight\_regions, gni\_min))) +  
 geom\_point(aes(resp\_min), color='red', alpha=.5, size=4) +  
 geom\_point(aes(resp\_med), color='green', alpha=.5, size=4) +  
 geom\_point(aes(resp\_max), color='blue', alpha=.5, size=4) +  
 labs(y='Regiões Mundiais\n', x='\nÍndice de\nRespostas', subtitle='(a)') +  
 theme(text=element\_text(size=16),  
 axis.text.x=element\_text(angle=90, vjust=.5)) +  
 scale\_x\_continuous(breaks=seq(.25,.75,.1))  
  
q4\_graf\_gni = q4\_sumarios %>%   
 ggplot(aes(y=reorder(eight\_regions, gni\_min))) +  
 geom\_point(aes(gni\_min, color='gni\_min'), alpha=.5, size=4) +  
 geom\_point(aes(gni\_med, color='gni\_med'), alpha=.5, size=4) +  
 geom\_point(aes(gni\_max, color='gni\_max'), alpha=.5, size=4) +  
 scale\_color\_manual(name='GNI:',  
 values=c('gni\_min'='red', 'gni\_med'='green', 'gni\_max'='blue'),  
 labels=c('Máximo', 'Médio', 'Mínimo')) +  
 labs(y=NULL, x='\nGNI', subtitle='(b)') +  
 theme(text=element\_text(size=16),  
 axis.text.x=element\_text(angle=90, vjust=.5)) +  
 scale\_y\_discrete(labels=c(NULL, NULL,NULL, NULL,NULL, NULL,NULL, NULL))  
  
grid.arrange(q4\_graf\_resp, q4\_graf\_gni, ncol=2, widths=c(1.3,1))

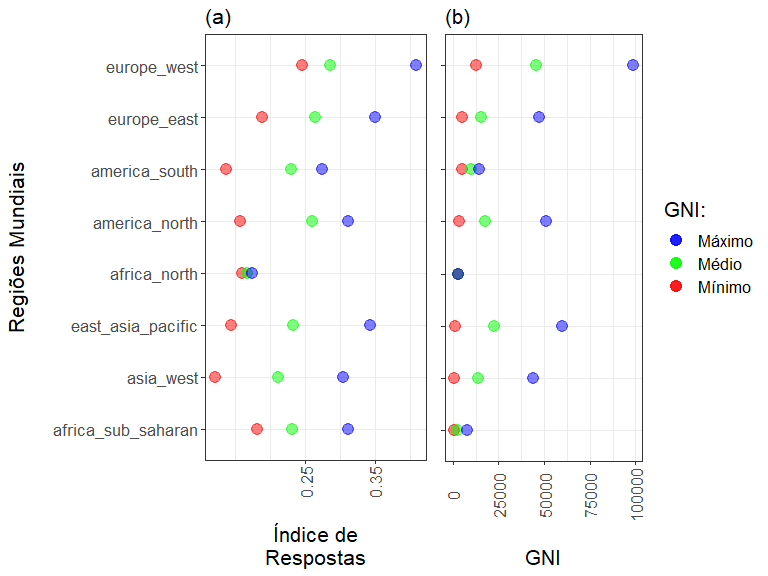


Gráfico 4.1: Visualização dos valores mínimos, médios e máximos do (a) Índice de Respostas das Perguntas e (b) Gross National Income (GNI), sendo valores referentes à países, porém agrupados em 8 regiões geográficas mundiais. Dados referentes ao site Super User.

No Gráfico 4.1 são feitas análises dos valores dos Índices de Respostas do site Super User e de Gross National Income (GNI).

Como os índices de GNI já foram discutidos na Questão 3, o foco desta discussão são os valores de Índice de Respostas do Super User, em que o menor valor mínimo é apresentado por um país do Oeste Asiático, já os menores valores médio e máximo são apresentados pelos países do Norte da África. Já os maiores valores dos sumários de índice de resposta são atingidos por países do Oeste Europeu, curiosamente com um país atingindo um valor acima de 0,40.

q4\_modelo = lm(responderam\_prop ~ log10(GNI), data=q4\_dados) # responderam\_prop = b0 + b1 \* GNI  
  
tidy(q4\_modelo)

## # A tibble: 2 x 5  
## term estimate std.error statistic p.value  
## <chr> <dbl> <dbl> <dbl> <dbl>  
## 1 (Intercept) -0.00637 0.0385 -0.166 0.869   
## 2 log10(GNI) 0.0629 0.00936 6.72 0.00000000345

glance(q4\_modelo)

## # 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 0.382 0.374 0.0448 45.1 3.45e-9 1 127. -249. -242.  
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>

q4\_aug = q4\_modelo %>% augment(q4\_dados)  
  
q4\_dados %>%  
 ggplot(aes(responderam\_prop, GNI)) +  
 geom\_point(size=3, alpha=.5) +  
 scale\_y\_log10() +  
 geom\_line(aes(x=q4\_aug$.fitted), color='red') +  
 labs(x='\nÍndice de Resposta de Questões', y='GNI\n') +  
 theme(text=element\_text(size=16))

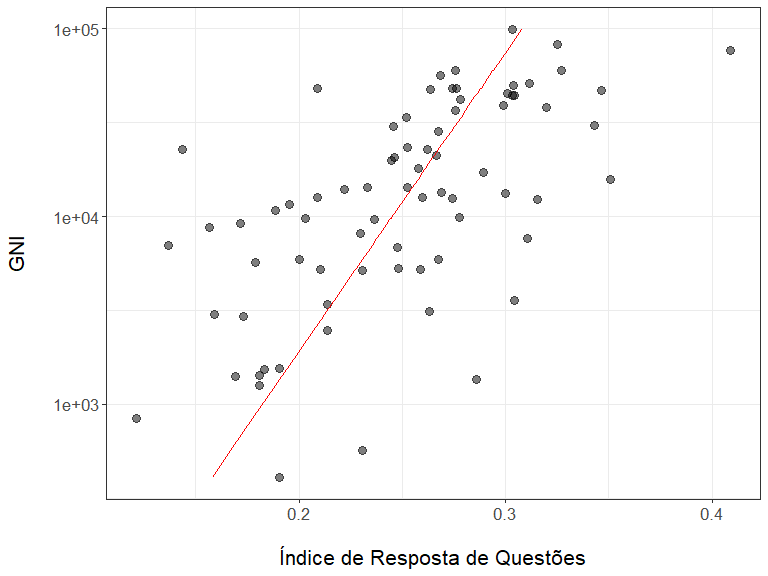


Gráfico 4.2: Visualização da dispersão de valores para o Índice de Respostas no site Super User e o valor de , com um modelo linear calculado e visualizado pela linha na cor vermelha.

Na questão também foi utilizada a Regressão Linear Simples para a análise de variáveis, sendo elas: Índice de Resposta de Questões e GNI a partir de usuários de diversos países que respondem à perguntas no Stack Overflow.

No Gráfico 4.2, foi criado um modelo linear resultando na função:

,

que apresenta 38% da variância da variável, sendo . Foi realizado o cálculo de log na base 10 para o valor GNI como forma de uma melhor visualização do modelo linear a partir do gráfico de dispersão dos valores.

A partir do modelo calculado, é esperado que o aumento do valor de GNI gere um acréscimo em torno de 0,0628 do valor do índice de respostas.

ggplot() +  
 geom\_point(data=q3\_dados, aes(responderam\_prop, GNI, color='so'), size=3, alpha=.5) +  
 geom\_line(data=q3\_aug, aes(.fitted, GNI, color='so')) +  
 geom\_point(data=q4\_dados, aes(responderam\_prop, GNI, color='su'), size=3, alpha=.5) +  
 geom\_line(data=q4\_aug, aes(.fitted, GNI, color='su')) +  
 labs(x='\nÍndice de Resposta de Questões', y='EPI\n') +  
 scale\_color\_manual(name='Sites:',  
 values=c('so'='red', 'su'='blue'),  
 labels=c('Stack Overflow', 'Super User')) +  
 scale\_y\_log10() +  
 theme(text=element\_text(size=16))

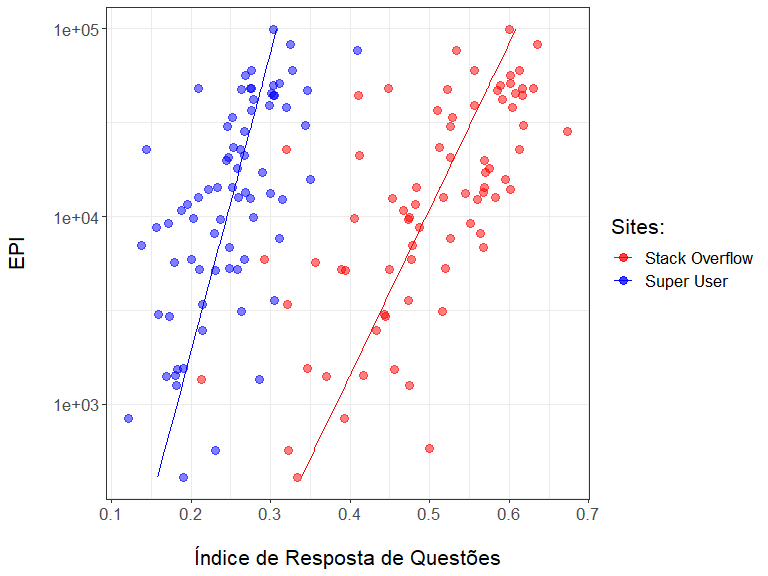


Gráfico 4.3: Visualização da dispersão dos valores e modelos lineares do Índice de Respostas e dos sites Stack Overflow (Cor Vermelha) e Super (Cor Azul).

Para o índice de respostas do Stack Overflow tem-se os seguintes valores:

Para o índice de respostas do Super User tem-se:

Antes da análise, é importante destacar que para os modelos lineares das Questões 3 e 4 é utilizado o cálculo de log na base 10 para a variável GNI, devida a uma melhor visualização de relação linear entre variáveis.

Comparando os valores dos coeficientes e (Visualização no Gráfico 4.3) dos modelos lineares dos dois sites, apenas do site Super User é negativo. Com relação à , é perceptível uma maior variação do valor do índice de respostas em relação ao valor de para o site do Stack Overflow, visto que o valor de se mostra mais alto em relação ao do Super User.

Ao comparar os valores de , a variação no site Super User se mostra mais alta que no site Stack Overflow, em que atinge 0,382 de variação, o que implica que o modelo linear é capaz de explicar em torno de 38,2% o valor do índice de resposta a partir do valor de , que pode ser uma relação média se for levado em conta uma porcentagem acima de 20% até até 50% para isto. No Stack Overflow, é verificada uma variação de 0,160, ou seja, o modelo é 16% capaz de explicar o índice de respostas utilizando , sendo uma relação pequena se for levado em conta o valor de até 20% para tal.