L5P2 - Regressão múltipla com Cultura

Rodolfo Bolconte

06/06/21

Nessa etapa do laboratório, estamos interessados em entender a relação da atuação de pessoas de diferentes países em responder perguntas no stackoverflow com características socioeconômicas e culturais dos países.

Especificamente, queremos entender a relação da proporção de pessoas que responderam em um país com:

- fluência em inglês da população (EPI);
- produto interno bruto do país;
- disponibilidade de internet no país;
- grau de individualismo na cultura do país (individualismo está explicado aqui: https://www.hofstede-insights.com/models/national-culture/, e é a coluna IDV nos dados).

Estamos interessados em inferir a partir dos dados desse estudo o que acontece na população em geral.

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")) %>%
  select(site, country, IDV, EPI, GNI, responderam_prop, Internet,
eight_regions) %>%
  filter(EPI != "NA", GNI != "NA", Internet != "NA")
```

Análise Exploratória dos Dados

```
sumarios_eda = dataset %>%
  group_by(eight_regions) %>%
  summarize(min_idv=min(IDV),
            med_idv=mean(IDV),
            max_idv=max(IDV),
            min_internet=min(Internet),
            med_internet=mean(Internet),
            max_internet=max(Internet))

graf_eda_idv = sumarios_eda %>%
  ggplot(aes(y=reorder(eight_regions, min_idv)), size=4) +
  geom_point(aes(min_idv), color='red', alpha=.5, size=4) +
  geom_point(aes(med_idv), color='green', alpha=.5, size=4) +
```

```
  geom_point(aes(max_idv), color='blue', alpha=.5, size=4) +
  labs(x='\nIDV', y='Regiões Geográficas\n') +
  theme(text=element_text(size=16))

graf_eda_internet = sumarios_eda %>%
  ggplot(aes(y=reorder(eight_regions, min_idv))) +
  geom_point(aes(min_internet, color='min'), alpha=.5, size=4) +
  geom_point(aes(med_internet, color='med'), alpha=.5, size=4) +
  geom_point(aes(max_internet, color='max'), alpha=.5, size=4) +
  scale_color_manual(name='Valores:',
                     values=c('min'='red', 'med'='green', 'max'='blue'),
                     labels=c('Máximo', 'Média', 'Mínimo')) +
  labs(x='\nInternet', y=NULL) +
  scale_y_discrete(labels=c(NULL, NULL,NULL, NULL,NULL, NULL,NULL, NULL)) +
  scale_x_continuous(breaks=seq(0,100,20)) +
  theme(text=element_text(size=16))

grid.arrange(graf_eda_idv, graf_eda_internet, ncol=2, widths=c(1.2, 1))
```

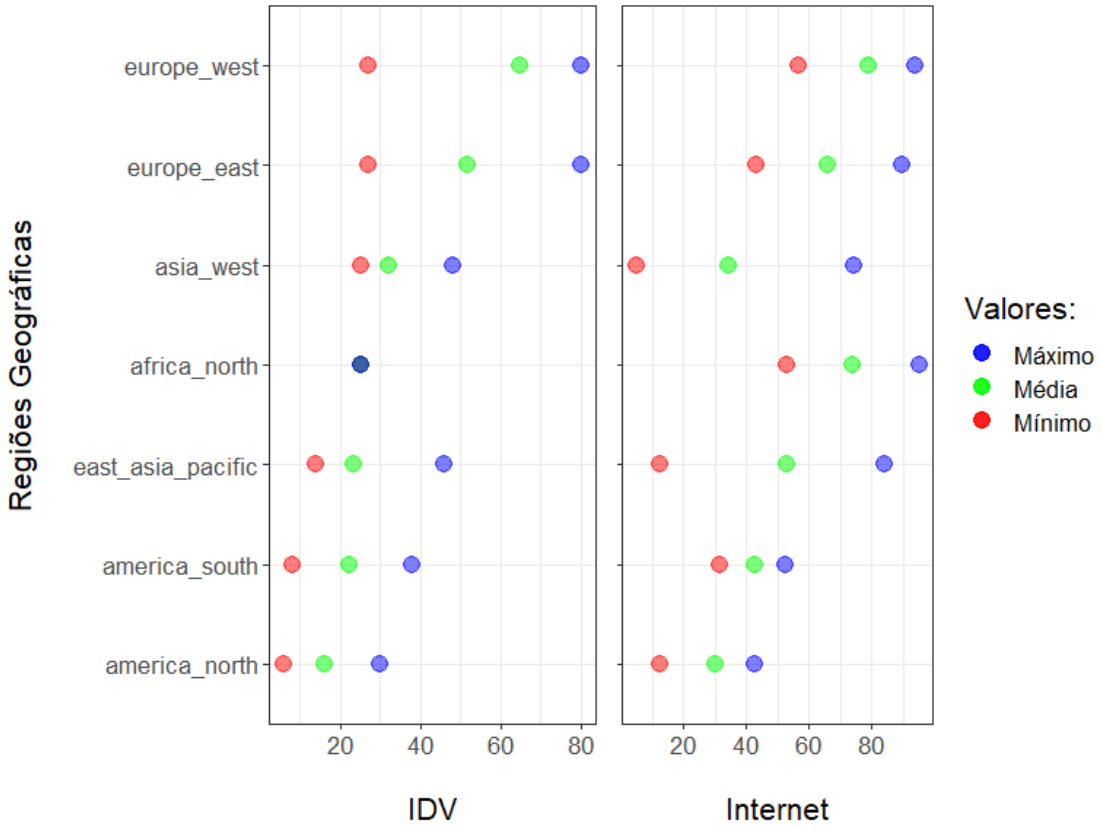

Gráfico 1.1: Visualização dos Sumários Mínimo (Vermelho), Média (Verde) e Máximo (Azul) referentes aos valores de IDV e Internet dos países separados por 7 regiões geográficas.

No Gráfico 1.1 tem-se os valores mínimo, médio e máximo de países para 7 regiões geográficas, ordenados pelo menor valor de IDV. O país que atingiu o menor valor de IDV se

encontra na América do Norte, atingindo menos de 10, enquanto o maior valor é atingido por um país do Oeste Europeu, assim como a maior média de países. Já para os dados de disponibilidade de internet, o menor valor se dá por um país do Oeste Asiático, com menos de 10, enquanto o maior é alcançado por um país do Norte da África.

Dispersão dos Dados

```
dados_stackoverflow = dataset %>% filter(site=='StackOverflow')

graf_m1_epi = dados_stackoverflow %>%
  ggplot(aes(responderam_prop, EPI)) +
  geom_point(size=3, alpha=.6) +
  geom_smooth(method=lm, se=F) +
  labs(x=NULL) +
  theme(text=element_text(size=16))

graf_m1_gni = dados_stackoverflow %>%
  ggplot(aes(responderam_prop, log10(GNI))) +
  geom_point(size=3, alpha=.6) +
  geom_smooth(method=lm, se=F) +
  labs(x=NULL) +
  theme(text=element_text(size=16))

graf_m1_int = dados_stackoverflow %>%
  ggplot(aes(responderam_prop, Internet)) +
  geom_point(size=3, alpha=.6) +
  geom_smooth(method=lm, se=F) +
  labs(x=NULL) +
  theme(text=element_text(size=16))

graf_m1_idv = dados_stackoverflow %>%
  ggplot(aes(responderam_prop, log10(IDV))) +
  geom_point(size=3, alpha=.6) +
  geom_smooth(method=lm, se=F) +
  labs(x=NULL) +
  theme(text=element_text(size=16))

grid.arrange(graf_m1_epi, graf_m1_gni, graf_m1_int, graf_m1_idv,
             bottom = textGrob("responderam_prop (Stack Overflow)",
                               gp = gpar(fontsize=16)))
```

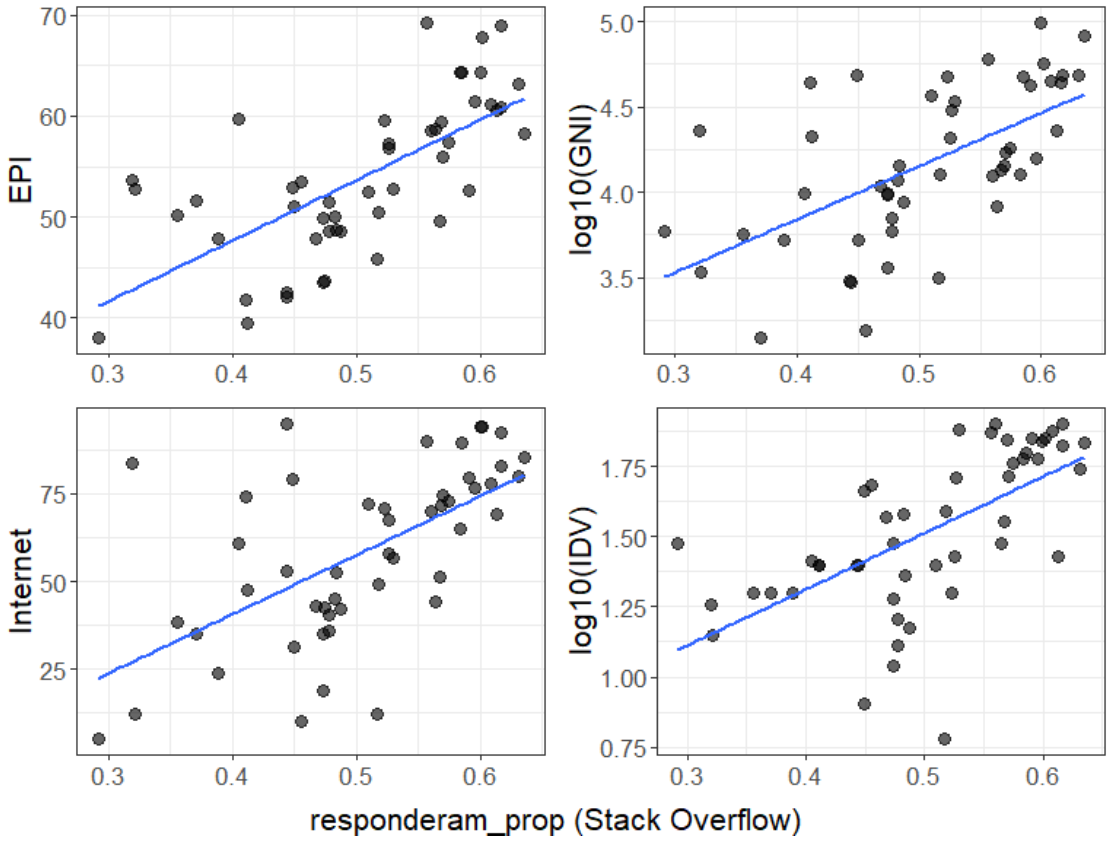

Gráfico 1.2: Dispersão dos valores das variáveis EPI, GNI, Disponibilização de Internet e IDV, com modelo linear traçado em relação à variável do índice de respostas das questões. Dados do Stack Overflow.

No Gráfico 1.2 tem-se a dispersão dos valores das variáveis de explicação (EPI, GNI, Internet e IDV) em relação à variável de resposta (responderam_prop). Cada subgráfico representa a relação de uma variável com responderam_prop, além de um modelo traçado a fim de encontração uma relação linear com cada variável, que para isto, foi feita a utilização do Log na base 10 para as variáveis GNI e IDV. Estas dispersões são relativas ao site do Stack Overflow.

```
dados_superuser = dataset %>% filter(site=='SuperUser')

graf_m3_epi = dados_superuser %>%
  ggplot(aes(responderam_prop, EPI)) +
  geom_point(size=3, alpha=.6) +
  geom_smooth(method=lm, se=F) +
  labs(x=NULL) +
  theme(text=element_text(size=16))

graf_m3_gni = dados_superuser %>%
  ggplot(aes(responderam_prop, log10(GNI))) +
  geom_point(size=3, alpha=.6) +
  geom_smooth(method=lm, se=F) +
```

```
  labs(x=NULL) +
  theme(text=element_text(size=16))

graf_m3_int = dados_superuser %>%
  ggplot(aes(responderam_prop, Internet)) +
  geom_point(size=3, alpha=.6) +
  geom_smooth(method=lm, se=F) +
  labs(x=NULL) +
  theme(text=element_text(size=16))

graf_m3_idv = dados_superuser %>%
  ggplot(aes(responderam_prop, log10(IDV))) +
  geom_point(size=3, alpha=.6) +
  geom_smooth(method=lm, se=F) +
  labs(x=NULL) +
  theme(text=element_text(size=16))

grid.arrange(graf_m3_epi, graf_m3_gni, graf_m3_int, graf_m3_idv,
             bottom = textGrob("responderam_prop (Super User)",
                               gp = gpar(fontsize=16)))
```

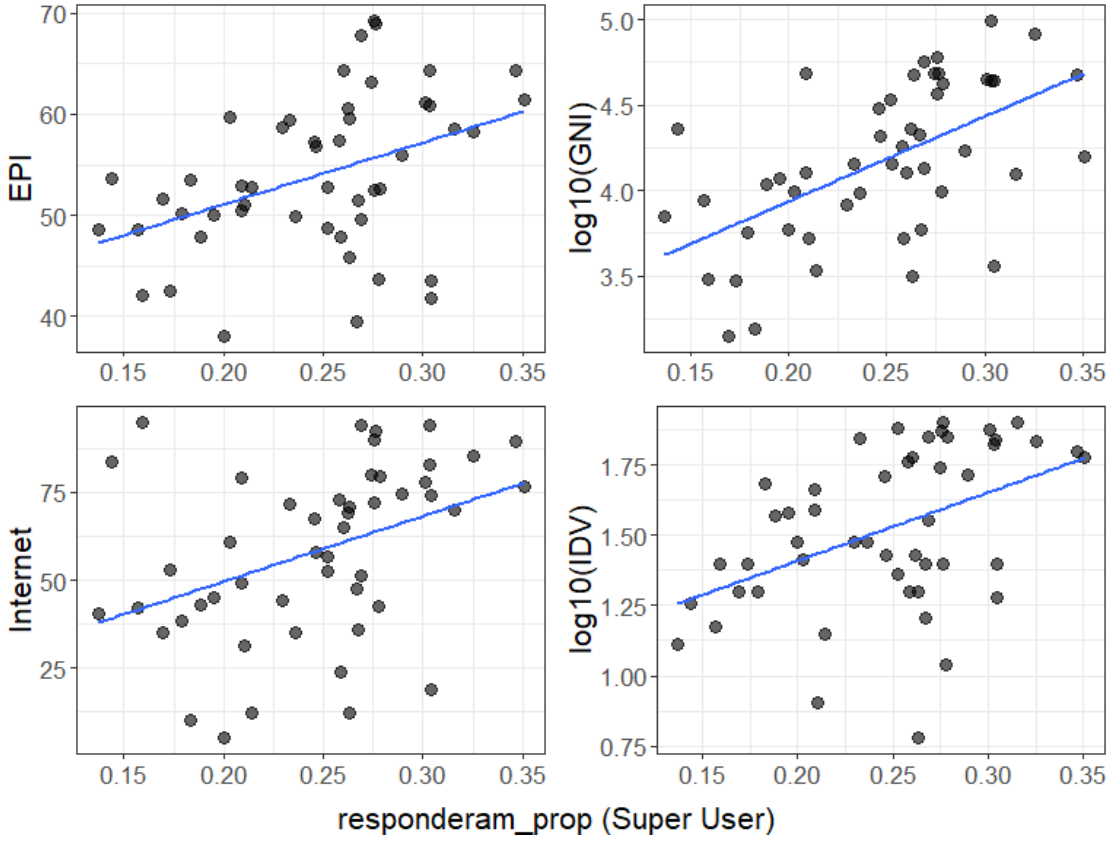

Gráfico 1.3: Dispersão dos valores das variáveis EPI, GNI, Disponibilização de Internet e IDV, com modelo linear traçado em relação à variável do índice de respostas das questões. Dados do Super User.

Assim como no gráfico anterior, o Gráfico 1.3 tem-se a dispersão dos valores das mesmas variáveis e também modelo linear traçado, porém os dados são provenientes do site Super User. Para ambos os sites, a dispersão dos valores para cada variável se mostra positiva, com a presença de pontos extremos, estes tendo destaque nas variáveis de Internet e IDV para o site Stack Overflow e Super User, além de se destacarem em EPI e um pouco em GNI para este último site.

Modelo 1

Focando apenas no StackOverflow, construa um modelo 1 com a variável responderam_prop com variável de resposta e fluência em inglês da população (EPI), produto interno bruto do país (GNI) e disponibilidade de internet no país como variáveis de explicação. Comente esse modelo em termos dos coeficientes e do ajuste. Estamos interessados em fazer inferência sobre os coeficiente.

```
modelo1 = lm(responderam_prop ~ EPI + log10(GNI) + Internet,
data=dados_stackoverflow)

tidy(modelo1, conf.int=T)

## # A tibble: 4 x 7
##   term        estimate std.error statistic  p.value  conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>    <dbl>     <dbl>     <dbl>
## 1 (Intercept) 0.0111    0.116       0.0958 0.924    -0.222      0.245
## 2 EPI         0.00535   0.00150     3.56   0.000880  0.00233    0.00838
## 3 log10(GNI)  0.0423    0.0320      1.32   0.193    -0.0221     0.107
## 4 Internet    0.000484  0.000622    0.779  0.440    -0.000769   0.00174

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     0.535         0.504 0.0621      17.3 0.000000133     3   68.7 -127.
-118.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

O modelo linear construído com as 3 variáveis de explicação é mostrado abaixo, além do valor de ajuste para o modelo:

responderam_prop = 0,0111
*+(0,0053*EPI)*
*+(0,0423*log10(GNI))*
*+(0,0005*Internet)*

R2=0,535

Para todos os coeficientes do modelo, os valores são positivos, sendo o maior valor alcançado pelo coeficiente b_2 que é multiplicado com o valor de log na base 10 da variável GNI, em torno de 0,0423, seguido de b_1 que multiplica 0,0053 com a variável EPI, e por último o coeficiente b_3 que multiplica 0,0005 com a variável Internet. Isto significa que levando em consideração todas as 3 variáveis para explicar o valor do índice de respostas no site do Stack Overflow, o log na base 10 do produto interno bruto é a variável que mais explica o índice "cuidado aqui. o coeficiente é o tamanho do efeito da variável, mas isso não é o mesmo que "o quanto a variável explica".", em que o aumento de uma unidade, equivale um aumento de 0,0423.

Quanto ao ajuste do modelo, foi obtido o valor de 0,535 para o R^2, ou seja, o modelo é capaz de explicar cerca de 53,5% o índice de resposta das perguntas do Stack Overflow, utilizando as 3 variáveis de explicação.

```
m1_graf_ic1 = tidy(modelo1, conf.int=T) %>% filter(term != '(Intercept)') %>%
  ggplot(aes(estimate, term)) +
  geom_errorbar(aes(xmin=conf.low, xmax=conf.high), width=.15) +
  geom_point(size=3) +
  geom_vline(xintercept = 0) +
  scale_x_continuous(breaks=seq(-.04, .1, .02)) +
  theme(text=element_text(size=16)) +
  labs(x=NULL, y=NULL)

m1_graf_ic2 = tidy(modelo1, conf.int=T) %>% filter(term == '(Intercept)') %>%
  ggplot(aes(estimate, term)) +
  geom_errorbar(aes(xmin=conf.low, xmax=conf.high), width=.15) +
  geom_point(size=3) +
  geom_vline(xintercept = 0) +
  theme(text=element_text(size=16)) +
  labs(x=NULL, y=NULL)

grid.arrange(m1_graf_ic1, m1_graf_ic2, ncol=1, heights=c(2.5,1),
             left = textGrob("Coeficientes do Modelo 1\n",
                             gp=gpar(fontsize=16), r=90))
```

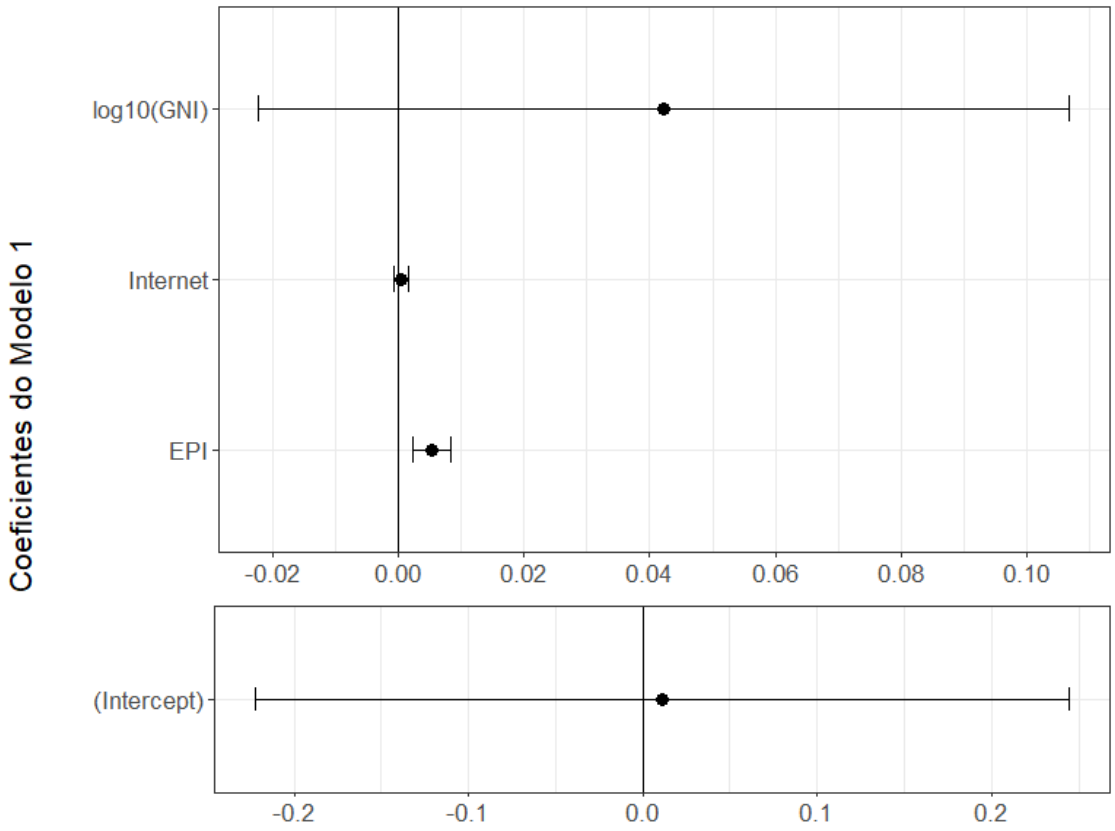

Gráfico 2.1: Intervalos de Confiança dos valores estimados para as variáveis explicativas do Modelo Linear 1, sendo as variáveis EPI, Internet e o log na base 10 de GNI.

No Gráfico 2.1 tem-se os intervalos de confiança dos coeficientes utilizados no Modelo Linear. Exceto para o coeficiente de EPI, não é possível afirmar qual seria o efeito em relação a explicação do índice de respostas, pois de acordo com os intervalos, podem apresentar valores tanto positivos quanto negativos, com o maior intervalo de confiança para o coeficiente b_0 que se encontra entre -0,22 e 0,244, enquanto o menor é o coeficiente da variável Internet, com um intervalo entre -0,0007 e 0,0017. O coeficiente de EPI é o único que pode apresentar valores apenas positivos, comprovando a certeza da ocorrência de um efeito para a explicação do índice de respostas, em que seu intervalo se encontre entre 0,0023 e 0,0083.

Modelo 2

Focando apenas no StackOverflow, construa um modelo 2 que além das variáveis do modelo 1 tem também o IDV. Esse é um modelo que considera uma variável de cultura. Comparando o modelo 2 com o modelo 1, o que podemos afirmar sobre o efeito do individualismo no comportamento das pessoas de diferentes países no stackoverflow? Há um efeito relevante (lembre de considerar a inferência para a população de onde vem os dados)? O modelo é mais explicativo do que sem a variável relacionada a cultura?

```
modelo2 = lm(responderam_prop ~ EPI + log10(GNI) + Internet + log10(IDV),
data=dados_stackoverflow)

tidy(modelo2, conf.int=T)
```

```
## # A tibble: 5 x 7
##   term        estimate std.error statistic p.value  conf.low conf.high
##   <chr>          <dbl>     <dbl>     <dbl>   <dbl>     <dbl>     <dbl>
## 1 (Intercept) 0.000329  0.115      0.00287 0.998   -0.231      0.231
## 2 EPI         0.00451   0.00159    2.84    0.00677  0.00131    0.00770
## 3 log10(GNI)  0.0344    0.0320     1.08    0.288   -0.0301     0.0989
## 4 Internet    0.000272  0.000630   0.432   0.668   -0.000997   0.00154
## 5 log10(IDV)  0.0669    0.0445     1.50    0.140   -0.0228     0.157
```

```
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     0.558         0.518 0.0612      13.9 0.000000213     4   70.0 -128.
-117.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

O modelo linear construído com as 4 variáveis de explicação para o site do Stack Overflow é mostrado abaixo, além do valor de ajuste para o modelo:

responderam_prop= 0,0003
*+(0,0045*EPI)*
*+(0,0344*log10(GNI))*
*+(0,0003*Internet)*
*+(0,0668*log10(IDV))*

R2=0,558

```
m2_graf_ic1 = tidy(modelo2, conf.int=T) %>% filter(term != '(Intercept)') %>%
  ggplot(aes(estimate, term)) +
  geom_errorbar(aes(xmin=conf.low, xmax=conf.high), width=.2) +
  geom_point(size=3) +
  geom_vline(xintercept = 0) +
  scale_x_continuous(breaks=seq(-.04, .16, .02)) +
  theme(text=element_text(size=16)) +
  labs(x=NULL, y=NULL)

m2_graf_ic2 = tidy(modelo2, conf.int=T) %>% filter(term == '(Intercept)') %>%
  ggplot(aes(estimate, term)) +
  geom_errorbar(aes(xmin=conf.low, xmax=conf.high), width=.2) +
  geom_point(size=3) +
  geom_vline(xintercept = 0) +
  theme(text=element_text(size=16)) +
```

```
  labs(x=NULL, y=NULL)

grid.arrange(m2_graf_ic1, m2_graf_ic2, ncol=1, heights=c(2.5,1),
             left = textGrob("Coeficientes do Modelo 2\n",
                             gp=gpar(fontsize=16), r=90))
```

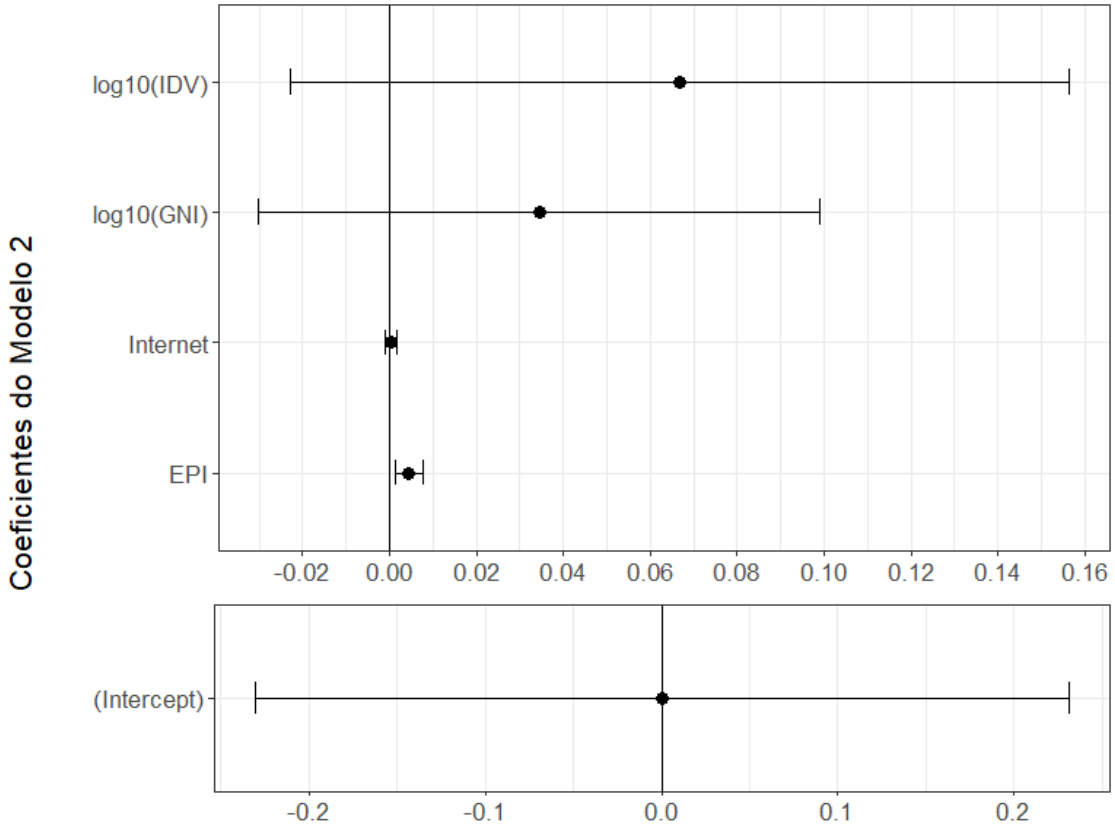

Gráfico 3.1: Intervalos de Confiança dos valores estimados para as variáveis explicativas do Modelo Linear 2, sendo as variáveis EPI, Internet e o log na base 10 de GNI e IDV. Dados do site Stack Overflow.

O segundo modelo utiliza as mesmas variáveis explicativas que o primeiro, com a adição da variável de individualismo para os países. O coeficiente de maior valor encontrado no modelo foi justamente para a variável de individualismo, porém sendo esta quando realizado o cálculo de log na base 10. O valor do coeficiente da variável ficou em torno de 0,0668, ou seja, num acréscimo de uma unidade do log na base 10 de IDV, o índice de resposta de questões no Stack Overflow aumenta em 0,0668. O menor coeficiente do segundo modelo tem como valor de 0,0002, que é multiplicado pela variável Internet assim como no primeiro modelo.

Analisando os Intervalos de Confiança dos coeficientes do segundo modelo no Gráfico 3.1, tem-se disposições parecidas para as variáveis utilizadas também no primeiro modelo, sendo o log na base 10 de IDV o maior intervalo dos variáveis, entre -0,0228 e 0,157.

Com relação ao ajuste do modelo, apresentou um valor R^2 de 0,558, ou seja, o segundo modelo é capaz de explicar 55,8% do valor para o índice de resposta. Comparando com o ajuste do primeiro modelo que não utiliza a variável IDV, tem-se uma diferença de apenas 2,3%, porém como existe tal diferença, é plausível afirmar que o segundo modelo é mais explicativo que o primeiro para o índice de respostas das questões do Stack Overflow, sendo assim, a variável de individualidade quando calculado o log na base 10 e também quando utilizada em conjunto com as demais variávies, pode explicar melhor o índice de respostas.

Modelo 3

Construa uma outra versão do modelo 2 usando agora os dados do SuperUser. Os resultados são consistentes com os do StackOverflow? Comente e mostre evidência que embase sua conclusão.

```
modelo3 = lm(responderam_prop ~ EPI + log10(GNI) + Internet + log10(IDV),
data=dados_superuser)

tidy(modelo3, conf.int=T)
```

```
## # A tibble: 5 x 7
##   term         estimate std.error statistic p.value conf.low conf.high
##   <chr>           <dbl>     <dbl>     <dbl>   <dbl>    <dbl>     <dbl>
## 1 (Intercept) -0.0990    0.0818      -1.21  0.233   -0.264    0.0658
## 2 EPI          0.000824  0.00113      0.729 0.470   -0.00146  0.00310
## 3 log10(GNI)   0.0678    0.0228       2.97  0.00479  0.0218   0.114
## 4 Internet    -0.000554  0.000449    -1.23  0.224   -0.00146  0.000351
## 5 log10(IDV)   0.0332    0.0317       1.05  0.302   -0.0308   0.0971
```

```
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     0.365         0.307 0.0437      6.32 0.000416     4   86.5 -161.
-150.
## # ... with 3 more variables: deviance <dbl>, df.residual <int>, nobs <int>
```

O modelo linear construído com as 4 variáveis de explicação para o site do Super User é mostrado abaixo, além do valor de ajuste para o modelo:

responderam_prop= −0,0989
*+(0,0008*EPI)*
*+(0,0678*log10(GNI))*
*+(−0,0005*Internet)*
*+(0,0331*log10(IDV))*

R2=0,365

```
m3_graf_ic1 = tidy(modelo3, conf.int=T) %>% filter(term != '(Intercept)') %>%
  ggplot(aes(estimate, term)) +
  geom_errorbar(aes(xmin=conf.low, xmax=conf.high), width=.2) +
  geom_point(size=3) +
  geom_vline(xintercept = 0) +
  scale_x_continuous(breaks=seq(-.04, .1, .02)) +
  theme(text=element_text(size=16)) +
  labs(x=NULL, y=NULL)

m3_graf_ic2 = tidy(modelo3, conf.int=T) %>% filter(term == '(Intercept)') %>%
  ggplot(aes(estimate, term)) +
  geom_errorbar(aes(xmin=conf.low, xmax=conf.high), width=.2) +
  geom_point(size=3) +
  geom_vline(xintercept = 0) +
  theme(text=element_text(size=16)) +
  labs(x=NULL, y=NULL)

grid.arrange(m3_graf_ic1, m3_graf_ic2, ncol=1, heights=c(2.5,1),
             left = textGrob("Coeficientes do Modelo 3\n",
                             gp=gpar(fontsize=16), r=90))
```

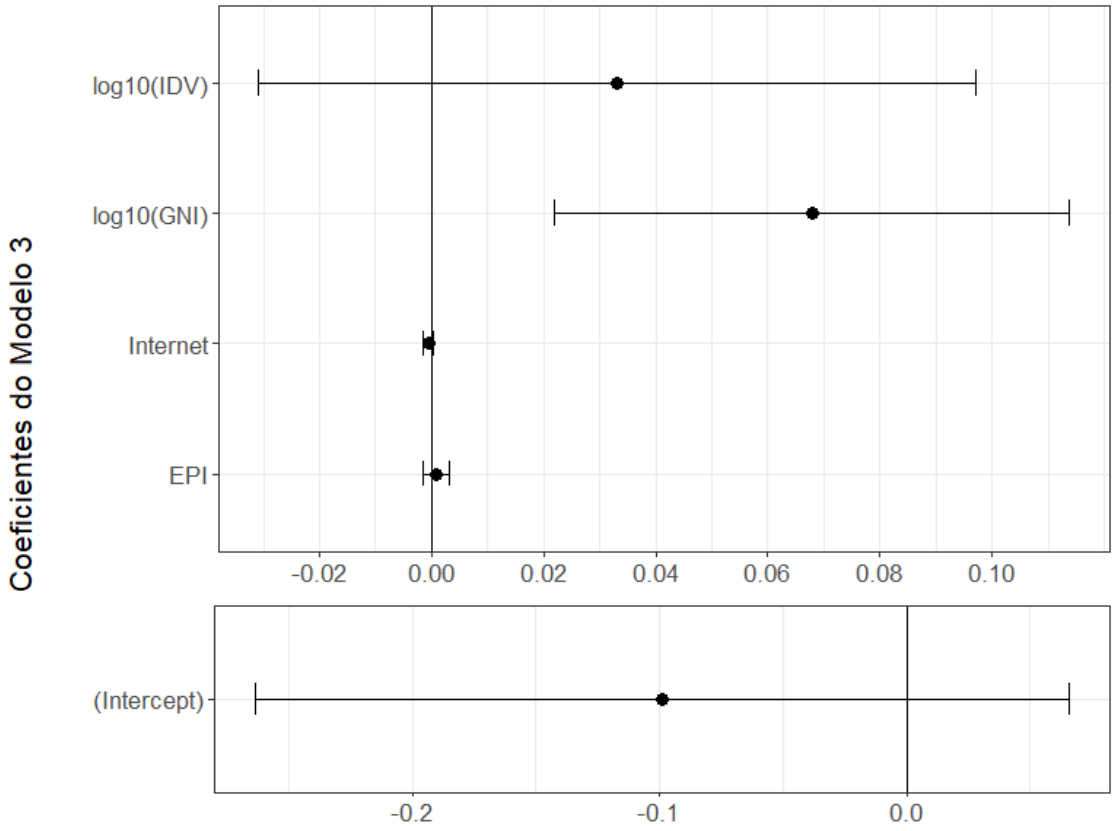

Gráfico 4.1: Intervalos de Confiança dos valores estimados para as variáveis explicativas do Modelo Linear 2, sendo as variáveis EPI, Internet e o log na base 10 de GNI e IDV. Dados do site Super User.

Utilizando as mesmas variáveis do segundo modelo para um modelo com dados do site Super User, tem-se algumas diferenças entre ambos. Com relação aos coeficientes, são apresentados dois com valores negativos, $b_0 = -\ 0,0989$ e $b_2 = -\ 0,0005$, este último indicando que no aumento de um índice da variável Internet, tem o índice de respostas diminuído em 0,0005. O coeficiente de maior índice é o log na base 10 de GNI, em que o aumento do seu valor em 1, o índice de respostas aumenta cerca de 0,0678.

De acordo com os Intervalos de Confiança visualizados no Gráfico 4.1, tem-se uma disposição diferente dos coeficientes em relação ao site Stack Overflow. O log na base 10 de IDV continua sendo o maior intervalo das 4 variáveis, de -0,0308 a 0,0971, porém o intervalo de $log_{10}(GNI)$ apresenta apenas valores positivos, de 0,0218 a 0,114, enquanto EPI que apresenta um intervalo inteiramente positivo no segundo modelo, no terceiro modelo pode apresentar valores negativos ou quase nulos também. Sendo assim, a única variável que pode apresentar qualquer valor que tenha um efeito em relação ao índice de respostas é o log na base 10 de GNI. O intervalo de b_0 se mostrou com um valor menor que pode ser atingido e não mais um intervalo balanceado de certa forma como no segundo modelo, o intervalo ficou em torno de -0,264 a 0,0658.

O modelo apresentou um ajuste de $R^2 = 0,365$, ou seja, o modelo é capaz de explicar cerca de 36,5% do valor do índice de resposta, valor menor que o segundo modelo para o site do Stack Overflow, que atingiu 55,8%, sendo assim há uma diferença de cerca de 19,3% entre os modelos. Ou seja, de acordo com o ajuste dos dois modelos, as 4 variáveis de explicação utilizadas explicam melhor o índice de respostas do site Stack Overflow em relação ao site Super User.