Análises dos Investimentos sobre a População e efeito da CFEM

Shayane Dos Santos Cordeiro

07/10/2021

Introdução

Conhecer a finalidade dos recursos arrecadados é de grande importância para o desenvolvimento de uma nação, principalmente do ponto de vista econômico. Diversas são as fontes dos recursos de um ente da federação (Estado, Distrito Federal e Municípios), seja em impostos, investimentos ou repasses da União. O objetivo do presente estudo é utilizar os dados de uma dessas fontes e verificar correlações utilizando um Modelo Linear Generalizado (MLG).

Metodologia

Os dados utilizados são públicos e contém informações acerca de investimentos na saúde (SIOPS), educação (SIOPE), infraestrutura (FINBRA), Compensação Financeira pela Exploração de Minerais (CFEM), receita corrente líquida (FINBRA) e população residente (IBGE) a nível municipal.

Para formar a base de dados utilizada no presente estudo foram realizados alguns ajustes. Os dados investimentos correspondem ao total investido em saúde, educação e infraestrutura. As bases foram unidas através dos códigos dos municípios (IBGE) e ano do investimento utilizando o comando *full_join*, a fim de considerar todos os valores independente de ter correspondências entre os códigos dos munícipios e anos.

Os dados da CFEM correspondem a distribuições financeiras mensais para municípios que produzem ou são afetados pela exploração de recursos minerais. Para esses dados foi feito o agrupamento anual considerando a soma do valor distribuído por mês para o município correspondente. Foi realizada a união das bases receita corrente líquida, investimentos (saúde, educação e infraestrutura) e CFEM utilizando o código do município (IBGE) e Ano como referência, utilizando o comando *full_join* para assegurar que todos os municípios, independente de correspondência com outra base, fossem adicionados. Assim a base contemplaria tanto municípios que distribuem quanto municípios que não distribuem CFEM , porém declararam o valor investido em pelo menos uma das áreas em estudo ou a receita corrente líquida do município.

A base ajustada anteriormente foi unida a segunda base que contém tamanho da população utilizando um \textit{lef_join} , porém considerando apenas os municípios presentes na primeira base.

Para a modelagem dos dados serão consideradas as variáveis explicativas Unidade da Federação (UF), Ano e se o município é distribuidor sim/não de CFEM e como variável resposta a soma dos investimentos em educação, saúde e públicos sobre a receita líquida do respectivo município, ou sobre o tamanho da população (investimento per capita).

Para tratamento dos valores ausentes a princípio será utilizada considerada a média da respectiva variável em anos anteriores para município.

Base de dados

Variável	Descrição
Ano	Ano da observação do dado.
UF	Unidade da Federação.
Receita Total	Receitas diversas. Fonte: IPEA.2015-2019
Receita	Receita corrente líquida. Fonte: FINBRA. 2015-2020
Educação	Investimentos em educação: Fonte: SIOPE.2015-2029
Saúde	Investimentos em saúde Fonte: SIOPS. 2015-2020
Público	Investimento em infraestrutura. Fonte: FINBRA.2015-2020
CFEM	Valor distribuído de CFEM (Município Produtor e Afetado). Fonte: ANM.2015-2020
Distribuidor de CFEM	0 - não; 1 - sim.
População	Tamanho da população. Fonte: IBGE. 2015-2020
PIB	Produto Interno Bruto. Fonte: IBGE. 2015-2018

Medidas Descritivas

```
summary(base[,8:12])
```

```
##     Receita                Saude              Publico
##  Min.   :   -5591336   Min.   : -4174292   Min.   :      1389
##  1st Qu.:   17858572   1st Qu.:   143637   1st Qu.:    948808
##  Median :   30515292   Median :   345732   Median :   1863989
##  Mean   :  105494418   Mean   :   803431   Mean   :   6508696
##  3rd Qu.:   63823550   3rd Qu.:   717731   3rd Qu.:   3953813
##  Max.   :48830405884   Max.   :451368274   Max.   :6367758075
##  NA's   :16677         NA's   :1190        NA's   :5200
##     Educacao              Cfem
##  Min.   :        0   Min.   :        0
##  1st Qu.:    97612   1st Qu.:     1705
##  Median :   345955   Median :    11182
##  Mean   :  1122464   Mean   :   762161
##  3rd Qu.:   919451   3rd Qu.:    63544
##  Max.   :679164904   Max.   :878348158
##  NA's   :9802        NA's   :17396
```

```
tab_01 = data.frame(
Variável = c("Receita", "Saúde", "Público","Educação"),
Mínimo = c(comma(min(base$Receita,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
           comma(min(base$Saude,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
           comma(min(base$Publico,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
           comma(min(base$Educacao,na.rm = TRUE), digits =2L, format = "f", big.mark = ",")),

Máximo = c(comma(max(base$Receita,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
           comma(max(base$Saude,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
           comma(max(base$Publico,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
           comma(max(base$Educacao,na.rm = TRUE), digits =2L, format = "f", big.mark = ",")),

Média = c(comma(mean(base$Receita,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
          comma(mean(base$Saude,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
          comma(mean(base$Publico,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
          comma(mean(base$Educacao,na.rm = TRUE), digits =2L, format = "f", big.mark = ",")),

Mediana = c(comma(median(base$Receita,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
          comma(median(base$Saude,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
          comma(median(base$Publico,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
          comma(median(base$Educacao,na.rm = TRUE), digits =2L, format = "f", big.mark = ",")),

Desvio = c(comma(sd(base$Receita,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
          comma(sd(base$Saude,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
          comma(sd(base$Publico,na.rm = TRUE), digits =2L, format = "f", big.mark = ",") ,
          comma(sd(base$Educacao,na.rm = TRUE), digits =2L, format = "f", big.mark = ","))
)

kbl(tab_01, booktabs = T) %>%
kable_styling(font_size = 10,latex_options = c("striped", "hold_position"))
```

Variável	Mínimo	Máximo	Média	Mediana	Desvio
Receita	-5,591,336.08	48,830,405,884.01	105,494,417.67	30,515,292.03	702,896,892.28
Saúde	-4,174,291.61	451,368,273.80	803,431.03	345,731.51	4,340,750.21
Público	1,389.08	6,367,758,075.42	6,508,695.74	1,863,989.10	61,929,449.43
Educação	0.01	679,164,904.12	1,122,464.37	345,955.19	6,935,095.04

Tratamento dos Valores Ausentes

Considerando o percentual investido sobre a receita corrente líquida

```
dados1 <- base  %>% filter((is.na(Receita) | Receita > 0)&(is.na(Saude) | Saude> 0 ))

dados1 <- dados1 %>% group_by(Cod.IBGE)%>%
          mutate(., Receita  = replace_na(Receita, mean(Receita, na.rm = TRUE)),
                    Saude    = replace_na(Saude, mean(Saude, na.rm = TRUE)),
                    Publico  = replace_na(Publico, mean(Publico, na.rm = TRUE)),
                    Educacao = replace_na(Educacao, mean(Educacao, na.rm = TRUE)))%>%
          mutate( Distribuidor = ifelse(Cfem>0, 1)) %>%
          mutate(., Distribuidor = replace_na(Distribuidor,0)) %>%
          filter((Receita != "NaN")&(Educacao != "NaN")&
                 (Saude != "NaN")&(Publico != "NaN"))%>%
          mutate(Invest = Educacao + Saude + Publico) %>%
          mutate(Invest_per = (Invest/Receita)*100, Saude_per = (Saude/Receita)*100,
                 Publico_per = (Publico/Receita)*100, Educacao_per = (Educacao/Receita)*100)
nrow(base)
```

```
## [1] 33414
```

```
nrow(base) - nrow(dados1)
```

```
## [1] 891
```

```
nrow(dados1)/nrow(base)
```

```
## [1] 0.9733345
```

Considerando o investimento per capita.

```
base1 <- base  %>% filter((is.na(Saude) | Saude> 0 ))

base1 <- base1 %>% group_by(Cod.IBGE)%>%
          mutate(., Saude    = replace_na(Saude, mean(Saude, na.rm = TRUE)),
                    Publico  = replace_na(Publico, mean(Publico, na.rm = TRUE)),
                    Educacao = replace_na(Educacao, mean(Educacao, na.rm = TRUE)))%>%
          mutate( Distribuidor = ifelse(Cfem>0, 1)) %>%
          mutate(., Distribuidor = replace_na(Distribuidor,0)) %>%
          filter((Educacao != "NaN")&(Saude != "NaN")&(Publico != "NaN"))%>%
          mutate(Invest = Educacao + Saude + Publico) %>%
          mutate(Invest_per = (Invest/População), Saude_per = (Saude/População),
                 Publico_per = (Publico/População), Educacao_per = (Educacao/População))
nrow(base)
```

```
## [1] 33414
```

```
nrow(base1)
```

```
## [1] 33233
```

```
nrow(base) - nrow(base1)
```

```
## [1] 181
```

```
nrow(base1)/nrow(base)
```

```
## [1] 0.9945831
```

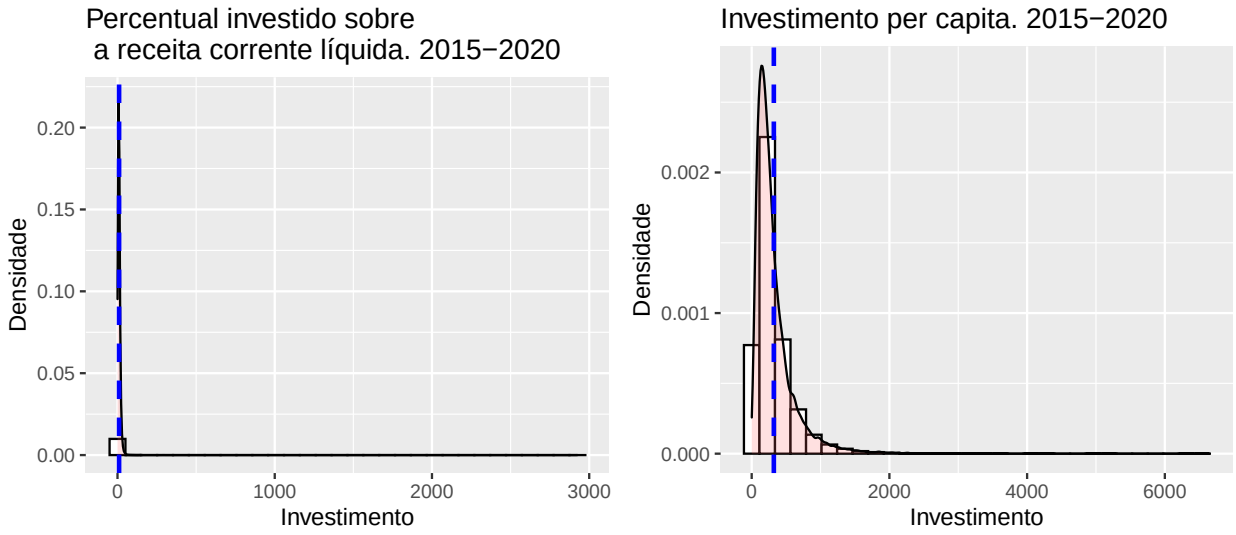
Comportamento da Variável Resposta

```
g1 <- ggplot(dados1,aes(x = Invest_per,  y = ..density..,)) +
      geom_histogram( fill = 'white', color = 'black') +
      labs(title = "Percentual investido sobre \n a receita corrente líquida. 2015-2020",
      x = "Investimento", y = "Densidade")+ geom_density(alpha=.2, fill="#FF6666") +
     geom_vline(aes(xintercept=mean(Invest_per)), color="blue", linetype="dashed", size=1)

g2 <- ggplot(base1,aes(x = Invest_per,  y = ..density..,)) +
      geom_histogram( fill = 'white', color = 'black') +
      labs(title = "Investimento per capita. 2015-2020", x = "Investimento", y = "Densidade")+
      geom_density(alpha=.2, fill="#FF6666") +
      geom_vline(aes(xintercept=mean(Invest_per)), color="blue", linetype="dashed", size=1)

grid.arrange(g1,g2, ncol=2)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Os histogramas apresentam o comportamento da variável resposta que apresenta assimetria e cauda pesada.

Percentual investido sobre a receita corrente líquida

```
summary(dados1$Invest_per)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    0.059    5.130    8.173    9.980   12.531 2924.933
```

Investimento per capita

```
summary(base1$Invest_per)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##    1.102  140.155  236.251  321.449  396.438 6541.962
```

A vantagem de utilizar a população no lugar da receita corrente líquida é problemas que podem ocorrer ao substituir o NA pela média. Caso fossem substituídos pela média apenas os dados dos investimentos (educação, saúde e infraestrutura), a perda de observações seria de aproximadamente 50%, para contornar este problema, também foram substituídos os valores ausentes da receita corrente líquida por sua média a nível do município.

Ao se utilizar o tamanho da população, além de não ocorrer perda expressiva dos dados, cerca de 1% foram perdidos, o viés de substituir um valor não observado pela média é muito menor. Outro ponto a destacar é a possibilidade de um intervalo de tempo maior e consequentemente mais observações, além da interpretação da variável investimento per capita ser mais clara.

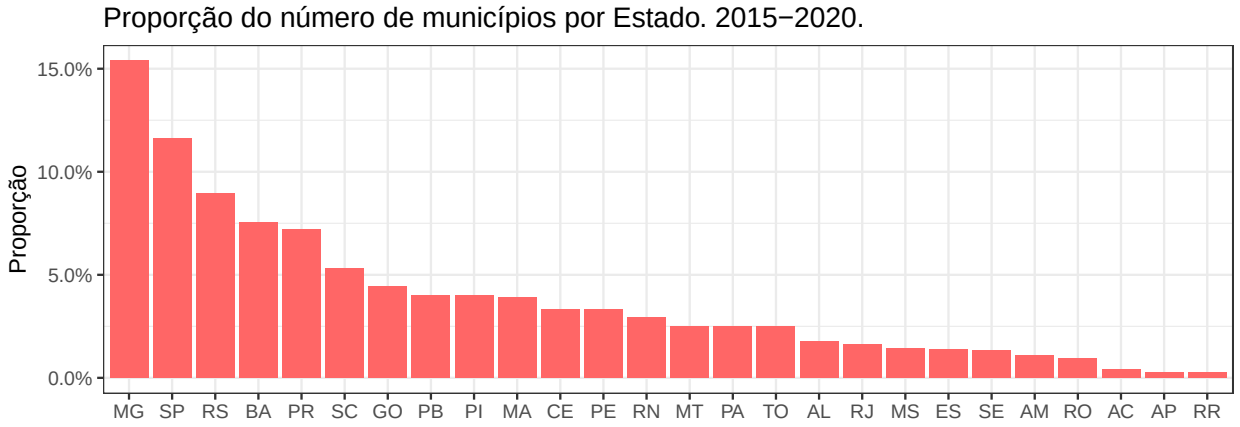
Análise Univariada

Para identificar se os estados que seriam utilizados na modelagem conteriam o minímo de observações necessárias para a análise foi realizado o gráfico a seguir.

```
tabela1    <- base1 %>% group_by(UF)   %>% summarise(total=n( ))
data_prop1 <- tabela1 %>% group_by(UF) %>%  mutate(prop=(total/sum(tabela1$total)))
#Forma de ordenar
#idx         <-  order(tabela1$prop , decreasing = TRUE)
#levels      <- tabela1$UF[idx]
#tabela1$UF <- factor(tabela1$UF, levels=levels, ordered=TRUE)

# Forma simples
tabela2    <- prop.table(table(base1$Distribuidor)) %>% as.data.frame
tabela3    <- prop.table(table(base1$Ano)) %>% as.data.frame
```

```
ggplot(data_prop1
       , aes(x = fct_reorder(UF, prop, .desc = TRUE), y = prop))+
geom_col(fill="#FF6666",position = "dodge") +
scale_y_continuous(labels = scales::percent)+theme_bw()+
labs(title = "Proporção do número de municípios por Estado. 2015-2020.", x = " ",y = "Proporção")
```

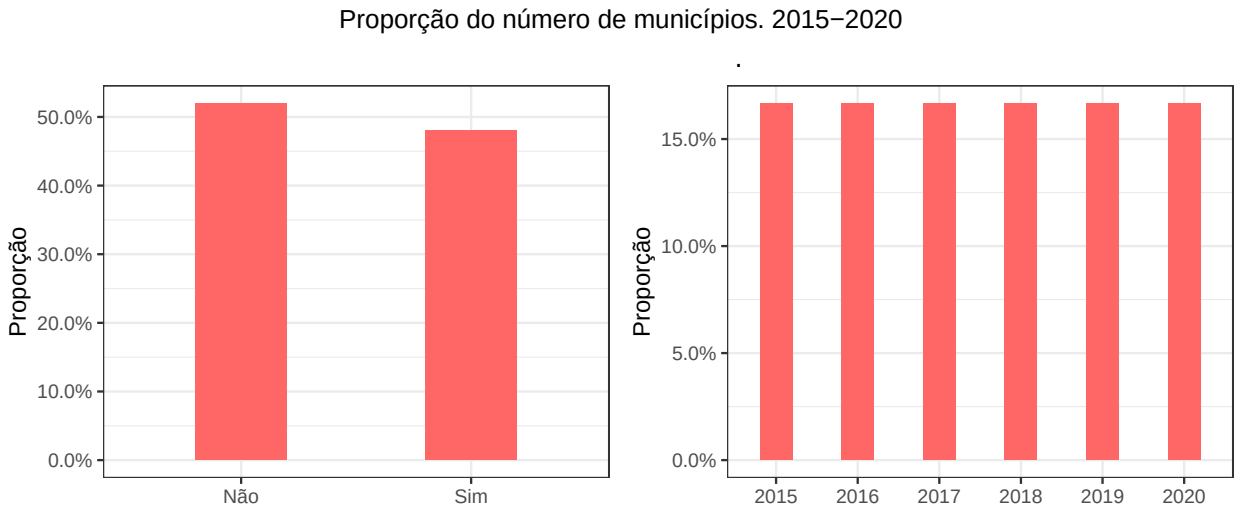


```r
tabela2$Var1 <- as.character(tabela2$Var1)
tabela2$Var1[tabela2$Var1=="1"] <-"Sim"
tabela2$Var1[tabela2$Var1=="0"] <-"Não"

g1<- ggplot(tabela2, aes(x =factor(Var1), y = Freq)) +
     geom_col(fill="#FF6666",position = "dodge",  width = .4) +
     scale_y_continuous(labels = scales::percent)+ theme_bw()+
     labs(title = " ", x = " ",y = "Proporção")

g2<- ggplot(tabela3, aes(x =Var1, y =Freq)) +
     geom_col(fill="#FF6666",position = "dodge",  width = .4) +
     scale_y_continuous(labels = scales::percent)+theme_bw()+
     labs(title = " .", x = " ",y = "Proporção")

grid.arrange(g1,g2, ncol=2,top ="Proporção do número de municípios. 2015-2020")
```



Para efeitos de análise será considerado:

- 0: quem não arrecada/distribui CFEM
- 1: quem arrecada/distribui CFEM

```
tabela1 <- base1 %>% filter(Distribuidor == "1") %>%
                  group_by(UF) %>% summarise(Sim = n())
tabela2 <- base1 %>% filter(Distribuidor == "0") %>%
                  group_by(UF) %>% summarise(Nao = n())
tab     <- merge(tabela1,tabela2)

tab1_2015 <- base1 %>% filter(Distribuidor == "1", Ano== "2015") %>%
                  group_by(UF) %>% summarise(Sim_2015 = n())
tab2_2015 <- base1 %>% filter(Distribuidor == "0", Ano== "2015") %>%
                  group_by(UF) %>% summarise(Nao_2015 = n())

tab1_2016 <- base1 %>% filter(Distribuidor == "1", Ano== "2016") %>%
                  group_by(UF) %>% summarise(Sim_2016 = n())
tab2_2016 <- base1 %>% filter(Distribuidor == "0", Ano== "2016") %>%
                  group_by(UF) %>% summarise(Nao_2016 = n())

tab1_2017 <- base1 %>% filter(Distribuidor == "1", Ano== "2017") %>%
                  group_by(UF) %>% summarise(Sim_2017 = n())
tab2_2017 <- base1 %>% filter(Distribuidor == "0", Ano== "2017") %>%
                  group_by(UF) %>% summarise(Nao_2017 = n())

tab1_2018 <- base1 %>% filter(Distribuidor == "1", Ano== "2018") %>%
                  group_by(UF) %>% summarise(Sim_2018 = n())
tab2_2018 <- base1 %>% filter(Distribuidor == "0", Ano== "2018") %>%
                  group_by(UF) %>% summarise(Nao_2018 = n())

tab1_2019 <- base1 %>% filter(Distribuidor == "1", Ano== "2019") %>%
                  group_by(UF) %>% summarise(Sim_2019 = n())
tab2_2019 <- base1 %>% filter(Distribuidor == "0", Ano== "2019") %>%
                  group_by(UF) %>% summarise(Nao_2019 = n())

tab1_2020 <- base1 %>% filter(Distribuidor == "1", Ano== "2020") %>%
                  group_by(UF) %>% summarise(Sim_2020 = n())
tab2_2020 <- base1 %>% filter(Distribuidor == "0", Ano== "2020") %>%
                  group_by(UF) %>% summarise(Nao_2020 = n())


tab1 <- full_join(tab1_2015, tab2_2015, by="UF")
tab2 <- full_join(tab1_2016, tab2_2016, by="UF")
tab3 <- full_join(tab1_2017, tab2_2017, by="UF")
tab4 <- full_join(tab1_2018, tab2_2018, by="UF")
tab5 <- full_join(tab1_2019, tab2_2019, by="UF")
tab6 <- full_join(tab1_2020, tab2_2020, by="UF")

table <- tab1 %>% full_join(tab2, by="UF") %>% full_join(tab3, by="UF") %>%
full_join(tab4,by="UF") %>% full_join(tab5,by="UF") %>%full_join(tab6,by="UF")
```

```
kbl(table, booktabs = T) %>%
kable_styling(font_size = 10,
              position = "left",latex_options = c("striped", "scale_down"))
```

UF	Sim_2015	Nao_2015	Sim_2016	Nao_2016	Sim_2017	Nao_2017	Sim_2018	Nao_2018	Sim_2019	Nao_2019	Sim_2020	Nao_2020
AC	7	15	4	18	6	16	5	17	6	16	6	16
AL	34	64	30	68	27	71	25	73	27	71	30	68
AM	24	37	19	42	20	41	23	38	20	41	24	37
AP	10	5	10	5	10	5	10	5	10	5	10	5
BA	169	248	165	252	179	238	178	239	212	205	223	194
CE	90	93	89	94	84	99	83	100	85	98	99	84
ES	62	15	64	13	64	13	65	12	65	12	70	7
GO	157	88	152	93	143	102	148	97	147	98	169	76
MA	38	178	40	176	42	174	45	170	64	152	69	147
MG	474	378	485	367	483	369	482	370	525	327	603	249
MS	48	31	50	29	56	23	57	22	54	25	62	17
MT	68	71	72	67	78	61	82	57	83	56	90	49
PA	60	79	62	77	59	80	63	76	67	72	73	66
PB	64	157	55	166	58	163	53	168	54	167	80	141
PE	70	113	67	116	65	118	58	125	67	116	66	117
PI	51	170	43	178	50	171	48	173	51	170	59	162
PR	177	222	175	224	179	220	183	216	189	210	201	198
RJ	68	23	66	25	66	25	65	26	71	20	76	15
RN	52	111	46	117	50	113	51	112	51	112	62	101
RO	33	19	34	18	32	20	32	20	34	18	38	14
RR	5	10	4	11	5	10	5	10	6	9	6	9
RS	204	293	213	284	209	288	208	289	217	280	237	260
SC	188	107	183	112	184	111	188	107	186	109	192	103
SE	34	41	35	40	33	42	31	44	31	44	28	47
SP	342	303	335	310	347	298	336	309	359	286	376	269
TO	37	102	39	100	36	103	41	98	41	98	55	84

```
tb_2015<-base1 %>% filter( Ano == "2015") %>% with(table(UF, Distribuidor))%>%
         prop.table(.,1) %>% round(2) %>%  as.data.frame.matrix( )
tb_2016<-base1 %>% filter( Ano == "2016") %>% with(table(UF, Distribuidor))%>%
         prop.table(.,1) %>% round(2) %>% as.data.frame.matrix( )
tb_2017<-base1 %>% filter( Ano == "2017") %>% with(table(UF, Distribuidor))%>%
         prop.table(.,1) %>% round(2) %>% as.data.frame.matrix( )
tb_2018<-base1 %>% filter( Ano == "2018") %>% with(table(UF, Distribuidor))%>%
         prop.table(.,1) %>% round(2) %>% as.data.frame.matrix( )
tb_2019<-base1 %>% filter( Ano == "2019") %>% with(table(UF, Distribuidor))%>%
         prop.table(.,1) %>% round(2) %>% as.data.frame.matrix( )
tb_2020<-base1 %>% filter( Ano == "2020") %>%with(table(UF, Distribuidor))%>%
         prop.table(.,1) %>% round(2) %>% as.data.frame.matrix( )

tb1 <- cbind(tb_2015,tb_2016,tb_2017,tb_2018,tb_2019,tb_2020)
kbl(tb1, longtable = T, booktabs = T) %>%
    add_header_above(c(" ", "2015" = 2, "2016" = 2,"2017" = 2,
                       "2018" = 2, "2019" =2, "2020" = 2)) %>%
    kable_styling(latex_options = c("repeat_header"))
```

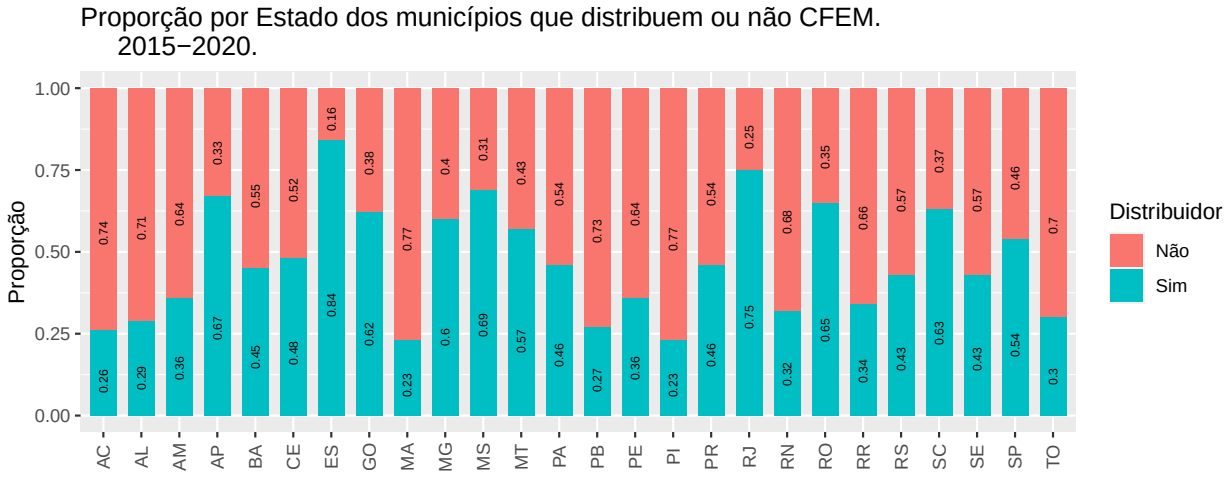
	2015		2016		2017		2018		2019		2020	
	0	1	0	1	0	1	0	1	0	1	0	1
AC	0.68	0.32	0.82	0.18	0.73	0.27	0.77	0.23	0.73	0.27	0.73	0.27
AL	0.65	0.35	0.69	0.31	0.72	0.28	0.74	0.26	0.72	0.28	0.69	0.31
AM	0.61	0.39	0.69	0.31	0.67	0.33	0.62	0.38	0.67	0.33	0.61	0.39
AP	0.33	0.67	0.33	0.67	0.33	0.67	0.33	0.67	0.33	0.67	0.33	0.67
BA	0.59	0.41	0.60	0.40	0.57	0.43	0.57	0.43	0.49	0.51	0.47	0.53
CE	0.51	0.49	0.51	0.49	0.54	0.46	0.55	0.45	0.54	0.46	0.46	0.54
ES	0.19	0.81	0.17	0.83	0.17	0.83	0.16	0.84	0.16	0.84	0.09	0.91
GO	0.36	0.64	0.38	0.62	0.42	0.58	0.40	0.60	0.40	0.60	0.31	0.69
MA	0.82	0.18	0.81	0.19	0.81	0.19	0.79	0.21	0.70	0.30	0.68	0.32
MG	0.44	0.56	0.43	0.57	0.43	0.57	0.43	0.57	0.38	0.62	0.29	0.71
MS	0.39	0.61	0.37	0.63	0.29	0.71	0.28	0.72	0.32	0.68	0.22	0.78
MT	0.51	0.49	0.48	0.52	0.44	0.56	0.41	0.59	0.40	0.60	0.35	0.65
PA	0.57	0.43	0.55	0.45	0.58	0.42	0.55	0.45	0.52	0.48	0.47	0.53
PB	0.71	0.29	0.75	0.25	0.74	0.26	0.76	0.24	0.76	0.24	0.64	0.36
PE	0.62	0.38	0.63	0.37	0.64	0.36	0.68	0.32	0.63	0.37	0.64	0.36
PI	0.77	0.23	0.81	0.19	0.77	0.23	0.78	0.22	0.77	0.23	0.73	0.27
PR	0.56	0.44	0.56	0.44	0.55	0.45	0.54	0.46	0.53	0.47	0.50	0.50
RJ	0.25	0.75	0.27	0.73	0.27	0.73	0.29	0.71	0.22	0.78	0.16	0.84
RN	0.68	0.32	0.72	0.28	0.69	0.31	0.69	0.31	0.69	0.31	0.62	0.38
RO	0.37	0.63	0.35	0.65	0.38	0.62	0.38	0.62	0.35	0.65	0.27	0.73
RR	0.67	0.33	0.73	0.27	0.67	0.33	0.67	0.33	0.60	0.40	0.60	0.40
RS	0.59	0.41	0.57	0.43	0.58	0.42	0.58	0.42	0.56	0.44	0.52	0.48
SC	0.36	0.64	0.38	0.62	0.38	0.62	0.36	0.64	0.37	0.63	0.35	0.65
SE	0.55	0.45	0.53	0.47	0.56	0.44	0.59	0.41	0.59	0.41	0.63	0.37
SP	0.47	0.53	0.48	0.52	0.46	0.54	0.48	0.52	0.44	0.56	0.42	0.58
TO	0.73	0.27	0.72	0.28	0.74	0.26	0.71	0.29	0.71	0.29	0.60	0.40

```
tabela1 <- base1 %>% filter(Distribuidor == "1") %>%
                  group_by(UF) %>% summarise(Total = n()) %>% mutate(Distribuidor = "Sim")
tabela2 <- base1 %>% filter(Distribuidor == "0") %>%
                  group_by(UF) %>% summarise(Total = n()) %>% mutate(Distribuidor = "Não")
tab     <- rbind(tabela1,tabela2)

prop <- round(prop.table(table(base1$UF,base1$Distribuidor),1),2) %>% as.data.frame

prop <- prop %>% rename(UF = "Var1")
prop <- prop %>% rename(Distribuidor = "Var2")
prop$Distribuidor <- as.character(prop$Distribuidor)
prop$Distribuidor[prop$Distribuidor == "0" ] <- "Não"
prop$Distribuidor[prop$Distribuidor == "1"] <- "Sim"

ggplot(prop, aes(x = UF, y = Freq , fill = Distribuidor)) +
geom_bar(stat = "identity",  width = .7) +
theme(axis.text.x=element_text(angle = 90, vjust = 0.5)) +
geom_text(aes(label = Freq),position = position_stack(0.4),angle = 90,size = 2)+
labs(title = "Proporção por Estado dos municípios que distribuem ou não CFEM.
     2015-2020.", x = " ",y = "Proporção")
```



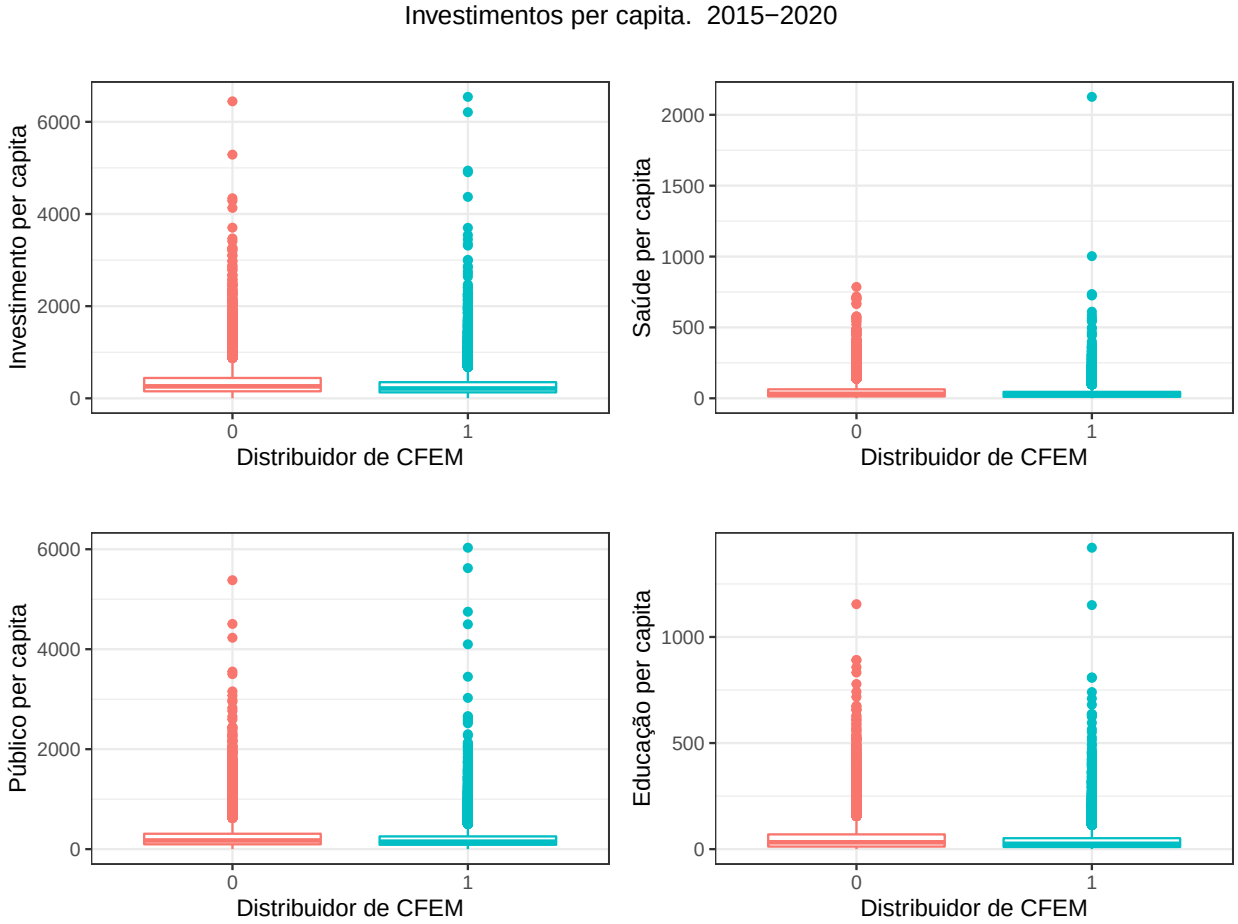
Análise bivariada: variáveis explicativas versus resposta.

```
p1<-qplot(factor(Distribuidor), Invest_per, data=base1,
          geom="boxplot",color=factor(Distribuidor),  show.legend = FALSE) +
theme_bw ( ) + labs(title = " ",x = "Distribuidor de CFEM", y = "Investimento per capita")

p2<-qplot(factor(Distribuidor), Saude_per, data=base1,
          geom="boxplot",color=factor(Distribuidor),  show.legend = FALSE) +
theme_bw(  ) + labs(title = " ",x = "Distribuidor de CFEM", y = "Saúde per capita")

p3<-qplot(factor(Distribuidor), Publico_per, data=base1,
          geom="boxplot",color=factor(Distribuidor),  show.legend = FALSE) +
theme_bw ( ) + labs(title = " ",x = "Distribuidor de CFEM", y = "Público per capita")

p4<-qplot(factor(Distribuidor), Educacao_per, data=base1,
          geom="boxplot",color=factor(Distribuidor),  show.legend = FALSE) +
theme_bw ( ) + labs(title = " ",x = "Distribuidor de CFEM", y = "Educação per capita")
grid.arrange(p1,p2, p3, p4,top ="Investimentos per capita.  2015-2020")
```

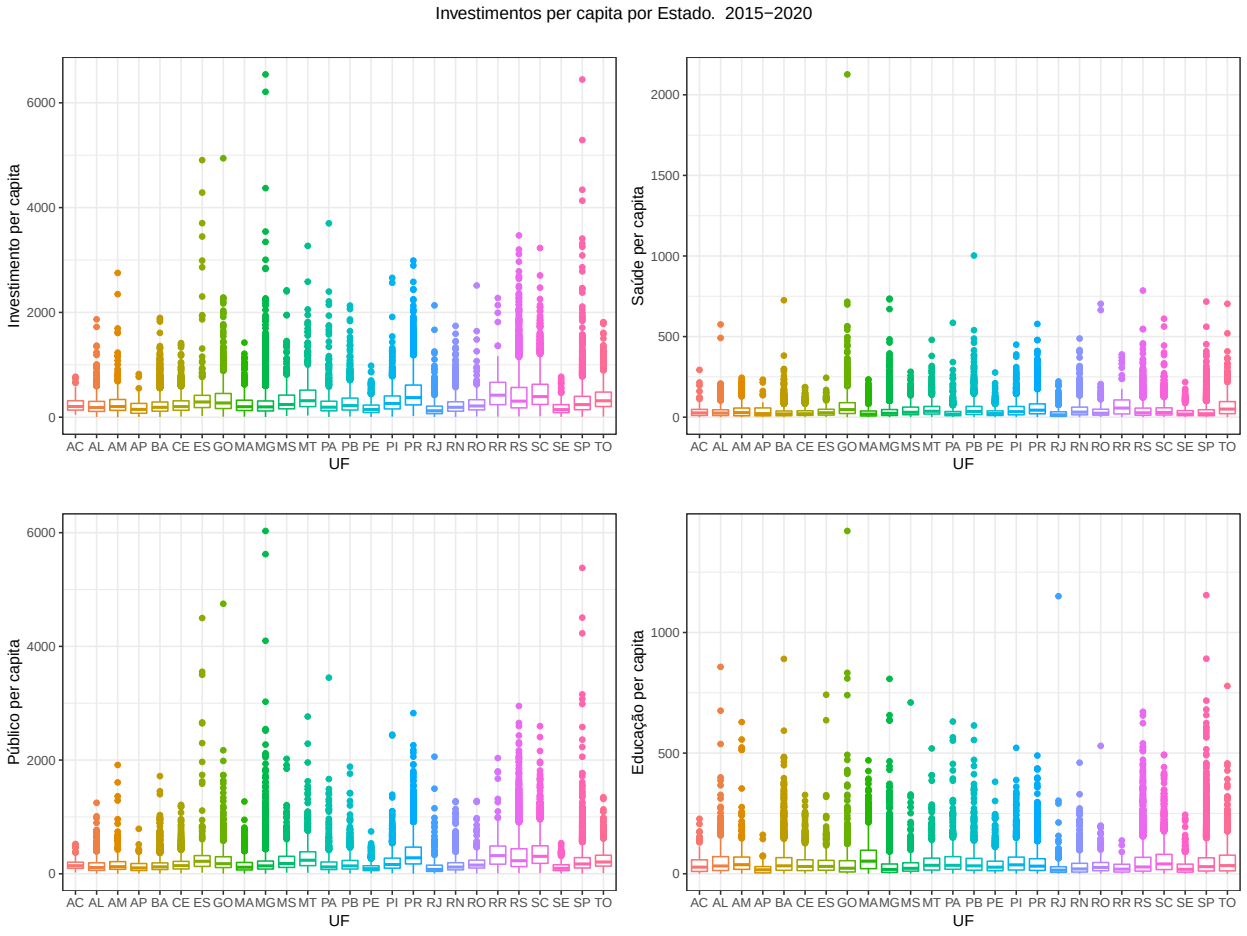


```
p1<-qplot(factor(UF), Invest_per, data=base1,
          geom="boxplot",color=factor(UF),  show.legend = FALSE) +
theme_bw ( ) + labs(title = " ",x = "UF", y = "Investimento per capita")

p2<-qplot(factor(UF), Saude_per, data=base1,
          geom="boxplot",color=factor(UF),  show.legend = FALSE) +
theme_bw(  ) + labs(title = " ",x = "UF", y = "Saúde per capita")

p3<-qplot(factor(UF), Publico_per, data=base1,
          geom="boxplot",color=factor(UF),  show.legend = FALSE) +
theme_bw ( ) + labs(title = " ",x = "UF", y = "Público per capita")

p4<-qplot(factor(UF), Educacao_per, data=base1,
          geom="boxplot",color=factor(UF),  show.legend = FALSE) +
theme_bw ( ) + labs(title = " ",x = "UF", y = "Educação per capita")
grid.arrange(p1,p2, p3, p4, top="Investimentos per capita por Estado.  2015-2020" )
```



Modelagem considerando dados de cauda pesada

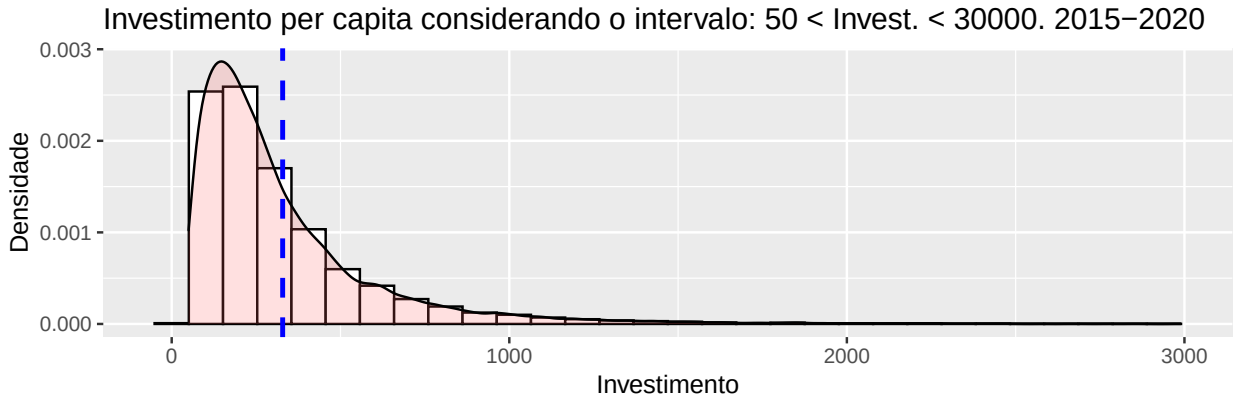
Considerando ajustes para distribuições de cauda pesada gama, lognormal e normal com diferentes funções de ligação: inversa, log e identidade, para o conjunto de dados com 33817 observações a análise do gráfico envelope, que avalia os quantis dos resíduos ajustados versus os empirícos, indicou que acima de 90% dos pontos ficaram fora da região de banda com 95% de confiança. Suspeitando de valores extremos, foi realizado um ajuste para valores acima de 50 e abaixo de 3000. A distribuição que melhor se ajustou foi a lognomral com ligação identidade.

```
base1$UF[base1$UF     == "MG"]    <- "A_MG"
base2 <- base1 %>% filter((Invest_per < 3000)&(Invest_per > 50) )
nrow(base1) - nrow(base2)
```

```
## [1] 1156
```

```
ggplot(base2,aes(x = Invest_per,  y = ..density..,)) +
geom_histogram( fill = 'white', color = 'black') +
labs(title = "Investimento per capita considerando o intervalo: 50 < Invest. < 30000. 2015-2020",
x = "Investimento", y = "Densidade")+geom_density(alpha=.2, fill="#FF6666") +
geom_vline(aes(xintercept=mean(Invest_per)), color="blue", linetype="dashed", size=1)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```




```
lnmod3 <- glm(log(Invest_per) ~ UF + Distribuidor + factor(Ano) +
              UF*Distribuidor, data = base2, family="gaussian"(link = "identity"))
tabela <- dust(lnmod3) %>%  sprinkle(col = 2:4, round = 3) %>%
          sprinkle(col = 5, fn = quote(pvalString(value))) %>%
          sprinkle_colnames(term = "Termos", estimate = "Estimativas",
                            std.error = "SE",statistic = "T-Estatística",
          p.value = "P-Valor") %>% kbl( booktabs = T, longtable = TRUE) %>%
          kable_styling(font_size = 7,latex_options = c("striped", "hold_position"))
```

```
tabela
```

Termos	Estimativas	SE	T-Estatística	P-Valor
(Intercept)	5.412	0.017	319.019	< 0.001
UFAC	-0.04	0.069	-0.576	0.56
UFAL	-0.061	0.037	-1.659	0.097
UFAM	0.072	0.047	1.538	0.12
UFAP	-0.161	0.131	-1.231	0.22
UFBA	-0.093	0.024	-3.948	< 0.001
UFCE	-0.005	0.032	-0.161	0.87
UFES	0.432	0.081	5.317	< 0.001
UFGO	0.325	0.032	10.139	< 0.001
UFMA	0.007	0.026	0.253	0.8
UFMS	0.349	0.057	6.089	< 0.001
UFMT	0.538	0.038	14.096	< 0.001
UFPA	-0.062	0.035	-1.749	0.08
UFPB	0.139	0.026	5.283	< 0.001
UFPE	-0.297	0.03	-9.952	< 0.001
UFPI	0.207	0.026	8.06	< 0.001
UFPR	0.688	0.024	28.942	< 0.001
UFRJ	-0.283	0.066	-4.257	< 0.001
UFRN	-0.042	0.03	-1.38	0.17
UFRO	0.206	0.066	3.13	0.002
UFRR	0.663	0.089	7.428	< 0.001
UFRS	0.57	0.022	25.891	< 0.001
UFSC	0.819	0.03	27.239	< 0.001
UFSE	-0.31	0.046	-6.755	< 0.001
UFSP	0.234	0.022	10.719	< 0.001
UFTO	0.4	0.031	12.702	< 0.001
Distribuidor	-0.129	0.019	-6.674	< 0.001
factor(Ano)2016	0.035	0.013	2.722	0.007
factor(Ano)2017	-0.275	0.013	-21.148	< 0.001
factor(Ano)2018	0.065	0.013	5.082	< 0.001
factor(Ano)2019	0.01	0.013	0.804	0.42
factor(Ano)2020	0.135	0.013	10.584	< 0.001
UFAC:Distribuidor	0.047	0.135	0.347	0.73
UFAL:Distribuidor	0.031	0.066	0.472	0.64
UFAM:Distribuidor	-0.059	0.077	-0.766	0.44
UFAP:Distribuidor	0.006	0.161	0.036	0.97
UFBA:Distribuidor	0.05	0.033	1.504	0.13
UFCE:Distribuidor	0.04	0.045	0.888	0.37
UFES:Distribuidor	-0.109	0.089	-1.221	0.22
UFGO:Distribuidor	-0.03	0.041	-0.733	0.46
UFMA:Distribuidor	-0.156	0.049	-3.206	0.001
UFMS:Distribuidor	-0.135	0.07	-1.939	0.052
UFMT:Distribuidor	-0.143	0.05	-2.836	0.005
UFPA:Distribuidor	0.099	0.051	1.938	0.053
UFPB:Distribuidor	-0.196	0.046	-4.25	< 0.001
UFPE:Distribuidor	0	0.048	0.007	> 0.99
UFPI:Distribuidor	-0.164	0.048	-3.409	< 0.001
UFPR:Distribuidor	-0.225	0.034	-6.712	< 0.001
UFRJ:Distribuidor	0.004	0.076	0.053	0.96
UFRN:Distribuidor	-0.109	0.051	-2.155	0.031
UFRO:Distribuidor	-0.176	0.082	-2.141	0.032
UFRR:Distribuidor	0.002	0.149	0.014	0.99
UFRS:Distribuidor	-0.34	0.031	-10.798	< 0.001
UFSC:Distribuidor	-0.259	0.038	-6.8	< 0.001
UFSE:Distribuidor	0.076	0.069	1.102	0.27
UFSP:Distribuidor	-0.1	0.029	-3.447	< 0.001
UFTO:Distribuidor	-0.051	0.054	-0.94	0.35

Interpretação dos parâmetros

- Equação do modelo: $\beta_0 + \beta_1 UF + \beta_2 Distribuidor + \beta_3 Ano + \beta_4 UF * Distribuidor$
- Quando $\beta_2 = 0$, ou seja, o município não distribui CFEM
  - $$\beta_0 + \beta_1 UF + \beta_2 0 + \beta_3 Ano + \beta_4 UF * 0 = \beta_0 + \beta_1 UF + \beta_3 Ano$$

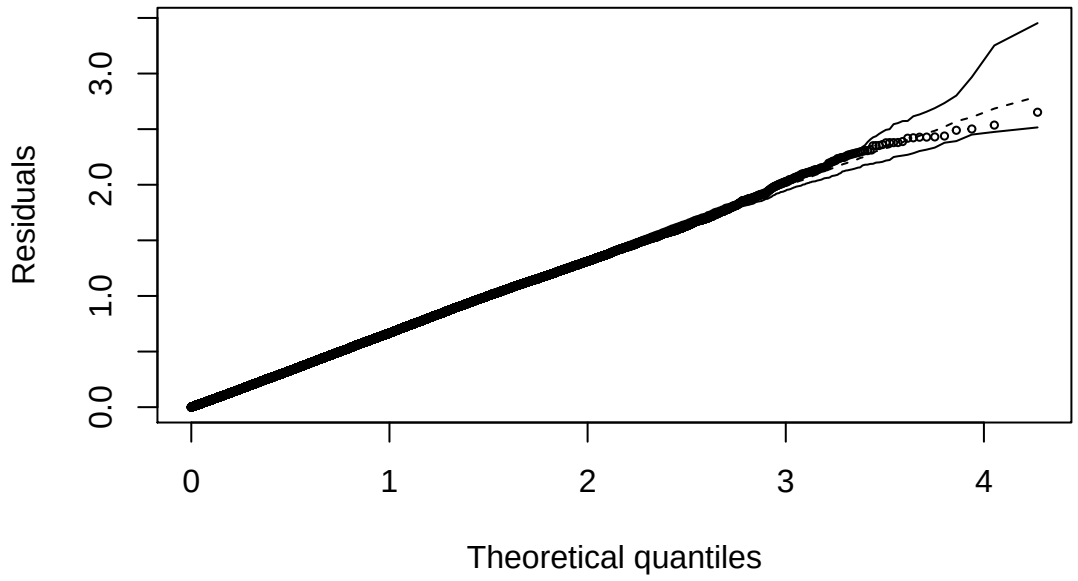
  1) O valor esperado para investimentos per capita no RJ para municípios não distribuidores de CFEM é MAIOR em 0.283 comparado com municípios não distribuidores de CFEM em Minas Gerais.
  2) O valor esperado para investimentos per capita no PR para municípios não distribuidores de CFEM é MAIOR em 0.688 comparado com municípios não distribuidores de CFEM em Minas Gerais.
  3) O valor esperado para investimentos per capita no PA para municípios não distribuidores de CFEM é MENOR em -0.062 comparado com municípios não distribuidores de CFEM em Minas Gerais.
  3) O valor esperado para investimentos per capita em SP para municípios não distribuidores de CFEM é MAIOR em 0.234 comparado com municípios não distribuidores de CFEM em Minas Gerais.

- Quando $\beta_2 = 1$, ou seja, o município distribui CFEM
  - $$\beta_0 + \beta_1 UF + \beta_2 1 + \beta_3 Ano + \beta_4 UF * 1 = \beta_0 + \beta_1 UF + \beta_2 + \beta_3 Ano + \beta_4 * UF$$
  - $$= (\beta_0 + \beta_2) + (\beta_1 + \beta_4) UF + \beta_3 Ano$$

  1) O valor esperado para investimentos per capita no PR para municípios distribuidores de CFEM é MAIOR em (-0.225 + 0.688 = 0.463) comparado com municípios distribuidores de CFEM em Minas Gerais.
  2) O valor esperado para investimentos per capita no PA para municípios distribuidores de CFEM é MAIOR em (-0.062 + 0.099 = 0.037) comparado com municípios distribuidores de CFEM em Minas Gerais.
  3) O valor esperado para investimentos per capita em SP para municípios distribuidores de CFEM é MAIOR em (-0.1 + 0.234= 0.134) comparado com municípios distribuidores de CFEM em Minas Gerais.

```r
#Ligação identidade
hnp( lnmod3$residuals , sim = 99 , resid.type =" deviance " ,
     how.many.out=T , conf = 0.99 , scale = T)
```

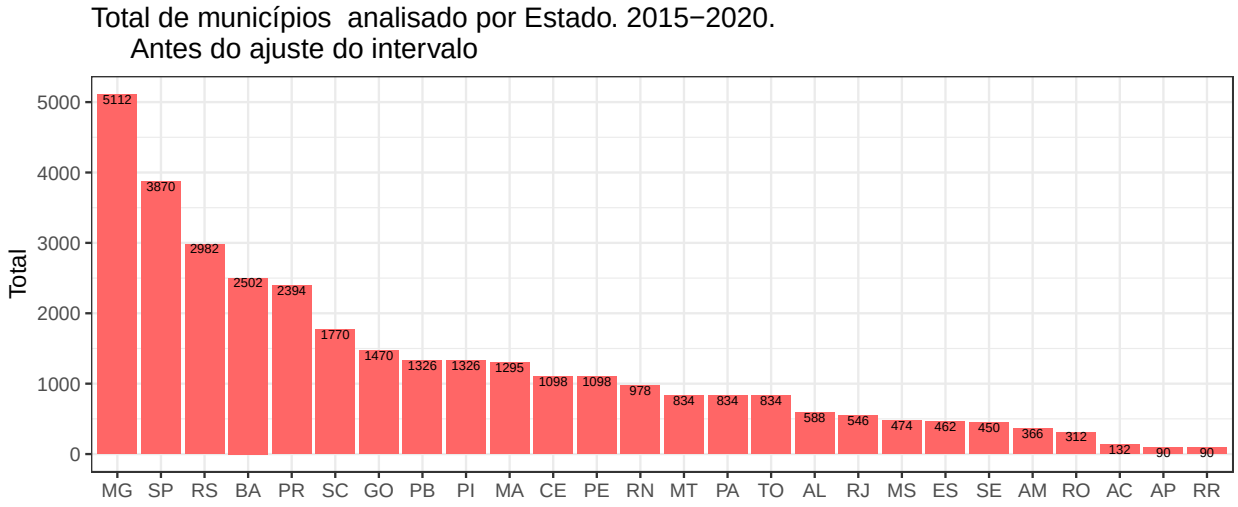
```
## Half-normal plot with simulated envelope generated assuming the residuals are
##         normally distributed under the null hypothesis.
## Estimated mean: 0.00000000000001411681
## Estimated variance: 0.4400898
```



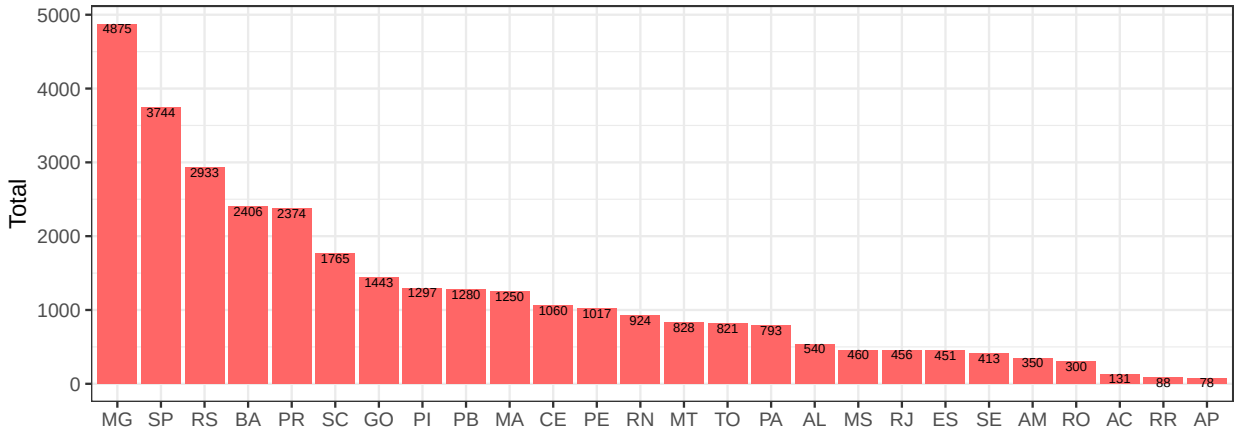

```
## Total points: 32077
## Points out of envelope: 28 ( 0.09 %)
```

O interesse é saber qual o percentual para as categorias da variável unidade da federação e distribuidor foram afetados.

```
base1$UF[base1$UF == "A_MG"] <- "MG"
base1 %>% group_by(UF) %>% summarise(Total = n()) %>%
ggplot(aes(x = fct_reorder(UF, Total, .desc = TRUE), y = Total)) + geom_col(fill="#FF6666",position = "
geom_text(aes(label = Total), vjust = 1,size = 2) + theme_bw() +
labs(title = "Total de municípios  analisado por Estado. 2015-2020.
     Antes do ajuste do intervalo ", x = " ",y = "Total")
```




```
base2$UF[base2$UF == "A_MG"] <- "MG"
base2 %>% group_by(UF) %>% summarise(Total = n()) %>%
ggplot(aes(x = fct_reorder(UF, Total, .desc = TRUE), y = Total)) + geom_col(fill="#FF6666",position = "
geom_text(aes(label = Total), vjust = 1,size = 2) + theme_bw() +
labs(title = "Total de municípios  analisado por Estado. 2015-2020.
     Após o ajuste do intervalo ", x = " ",y = "Total")
```



```r
tabela7 <- base1 %>% group_by(UF) %>% summarise(Total1 = n())
tabela8 <- base2 %>% group_by(UF) %>% summarise(Total2 = n())
tab1 <- merge(tabela7, tabela8)
tab1 <- tab1 %>% mutate(A_Completa = round(Total1/(Total1 + Total2),3),
                        A_Intervalo= round(Total2/(Total1 + Total2),3))

untidy <- tab1%>%
  pivot_longer(
    cols = A_Completa:A_Intervalo, # as colunas desse intervalo
    names_to = "Base", # terão seus nomes armazenados nessa nova coluna
    names_prefix = "A_", # pegar apenas os nomes que vem depois de 'ano_'
    values_to = "Prop") # e os seus valores armazenados nessa nova coluna



ggplot(untidy, aes(x = UF, y = Prop , fill = Base)) +
geom_bar(stat = "identity",  width = .7) + theme( ) +
geom_text(aes(label = Prop),position = position_stack(0.4),angle = 90,size = 2)+
labs(title = "Proporção por Estado dos municípios na base completa e na intervalar.
     2015-2020.", x = " ",y = "Proporção")
```