Uma síntese do saneamento básico brasileiro com uma perspectiva estatística.

Vítor Pereira

Resumo

De acordo com a Agência Senado, no Brasil temos quase 35 milhões de pessoas sem água tratada e cerca de 100 milhões não têm acesso à coleta de esgoto. É um dos maiores causadores de doenças, segundo o IBGE, cerca de 11 mil pessoas morrem por ano por falta de saneamento. O presente trabalho é uma tentativa de conscientizar e aumentar o entendimento sobre essa questão imprescindível para o desenvolvimento brasileiro. Para esse fim busca-se mensurar a qualidade das políticas públicas e o desenvolvimento de cada um 26 estados brasileiros e o Distrito Federal. Para isso serão utilizados variáveis sociais, econômicas, educacionais e da saúde com o ajuste de um modelo de Regressão Beta para avaliar a significância das variáveis e o ajuste da variável porcentagem da população urbana residente em domicílios ligados à rede de esgotamento sanitário no ano de 2017.

Sumário

1 Introdução

A ideia do presente trabalho é estudar, conscientizar e divulgar com abordagem estatística rigorosa a situação do saneamento básico brasileiro por meio de índices e taxas dos 26 estados e do Distrito Federal.

Segundo o Instituto Trata Brasil, 96% da população urbana brasileira possui acesso à água potável, mas apenas 61% têm coleta de esgoto. No entanto, apenas 36% do esgoto é tratado corretamente. Posto isso, teremos como objetivo analisar a variável a porcentagem da população urbana residente em domicílios ligados à rede de esgotamento sanitário no ano de 2017, devido ao seu enorme impacto ambiental e de saúde pública.

Infelizmente, o saneamento básico no Brasil ainda é insuficiente e ineficiente em muitas áreas do país e para melhorar a situação, o governo brasileiro tem implementado diversas políticas e programas de saneamento básico, como o Plano Nacional de Saneamento Básico (Plansab) e o Programa de Aceleração do Crescimento (PAC). Esses programas visam

ampliar o acesso à água tratada e ao esgoto tratado, bem como aumentar a eficiência e a qualidade dos serviços de saneamento.

Utilizaremos índices e taxas de variáveis econômicas, educacionais, de saúde e sociais para melhorar a compreensão da situação da variável objetivo. É importante ressaltar que as variáveis não necessariamente causam a precariedade ou a qualidade do saneamento, mas servem para estimar outras variáveis: Qualidade das políticas públicas e desenvolvimento dos estados que são as variáveis que mais interferem nos níveis de higiene do sistema cloacal brasileiro. As variáveis que utilizaremos são:

- **EDUCACIONAIS**
  - % dos ocupados com ensino fundamental completo 2010
  - % dos ocupados com ensino médio completo 2010
  - IDEB anos iniciais do ensino fundamental 2015
  - % de docentes na rede privada do fundamental com formação adequada 2016
  - Taxa de analfabetismo - 15 anos ou mais de idade 2016
- **ECONÔMICAS**
  - Produto Interno Bruto per capita 2016
  - Participação da Agropecuária no Valor Adicionado 2016
  - Participação da Indústria no Valor Adicionado 2016
  - % de pobres 2016
- **SAÚDE**
  - % de nascidos vivos com pelo menos sete consultas de pré-natal 2016
  - % de nascidos vivos com baixo peso ao nascer 2016
  - Mortalidade infantil 2016
- **SOCIAIS**
  - % de cobertura vegetal natural 2016
  - Índice de Gini 2016

2 Análise Exploratória

Nesta seção veremos um breve resumo das variáveis de estudo, com medidas descritivas, medidas de dispersão e gráficos de dispersão. Para simplificar a análise gráfica da Taxa do Saneamento Básico, iremos classificá-lo:

- Taxa do Saneamento < 0.2 -> Saneamento Péssimo
- 0.2 <= Taxa do Saneamento < 0.4 -> Saneamento Ruim
- 0.4 <= Taxa do Saneamento < 0.6 -> Saneamento Médio
- Taxa do Saneamento >= 0.6 -> Saneamento Bom

2.1 Variáveis Educacionais

Na análise das variáveis educacionais, três são taxas e apenas o IDEB é o índice, assim percebemos que tem-se uma proximidade entre a média e a mediana. Com todos os valores medianos das taxas ficando entre 40 e 60, assim mostrando que essas métricas indicam uma precariedade na educação brasileira, conforme podemos ver pela Tabela 1.

Tabela 1: Medidas Resumo das Variáveis Educacionais

variable	mean	median	sd	min	max	na_count
% dos ocupados com ensino fundamental completo 2010	59.18	59.7	7.250	48.0	76.4	0
% dos ocupados com ensino médio completo 2010	42.64	42.3	6.610	33.4	61.0	0
IDEB anos iniciais do ensino fundamental 2015	5.11	5.2	0.649	4.1	6.2	0
% de docentes na rede privada do fundamental com formação adequada 2016	51.22	51.5	12.309	29.8	67.0	0

Observa-se que a maioria dos estados tem entre 30% e 50% dos trabalhadores com ensino médio ou fundamental completos e pelas Figuras 1 e 2 percebe-se que existe tendência de crescimento da Avaliação do Saneamento concomitantemente com a porcentagem dos trabalhadores ocupados com ensino fundamental ou médio completo.

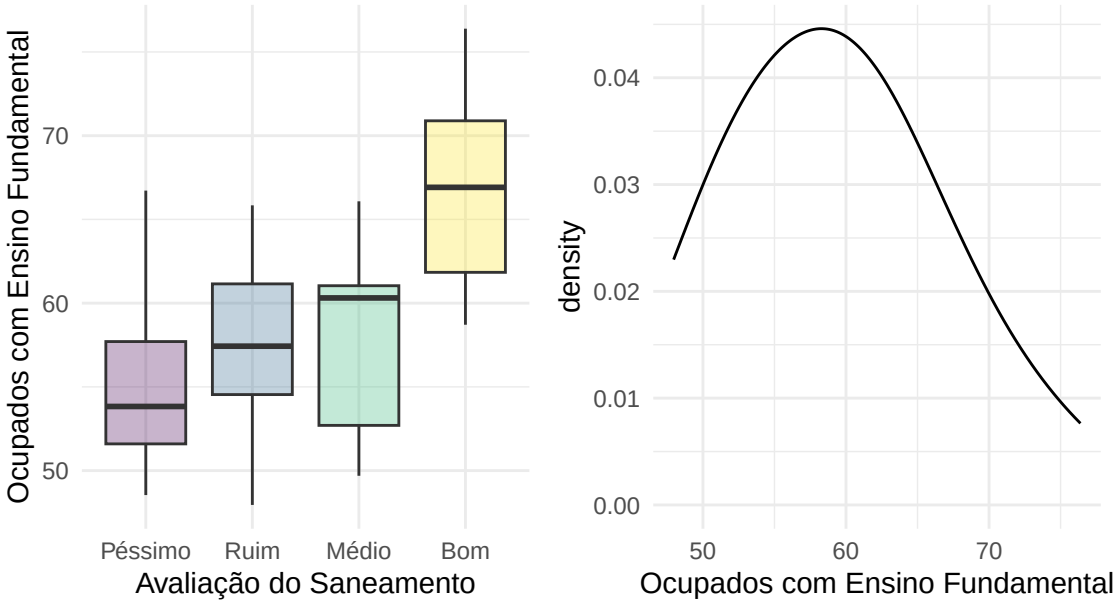

Figura 1: Comportamento da variável: Porcentagem de trabalhadores com ensino fundamental.

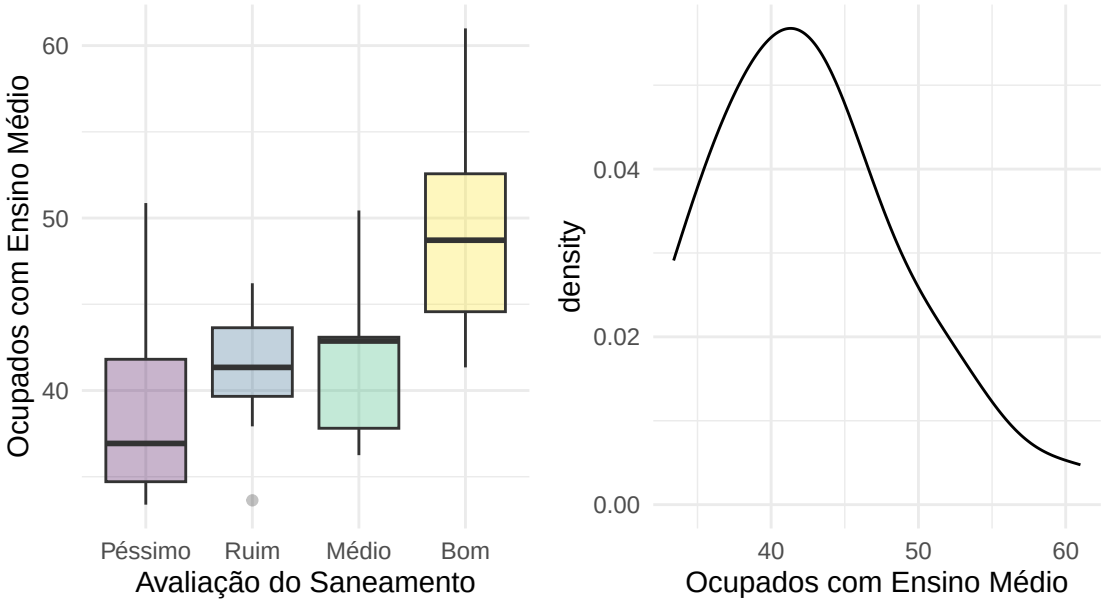

Figura 2: Comportamento da variável: Porcentagem de trabalhadores com ensino médio.

Conforme as Figuras 3 e 4, o IDEB tem valores semelhantes para as três menores classificações da Avaliação do Saneamento, com o nível bom se sobressaindo, assim como, para a porcentagem de docentes com formação adequado no ensino privado, em que as avaliações de Péssimo a Médio são parecidos e com a qualificação do saneamento como boa, sendo pouco superior.

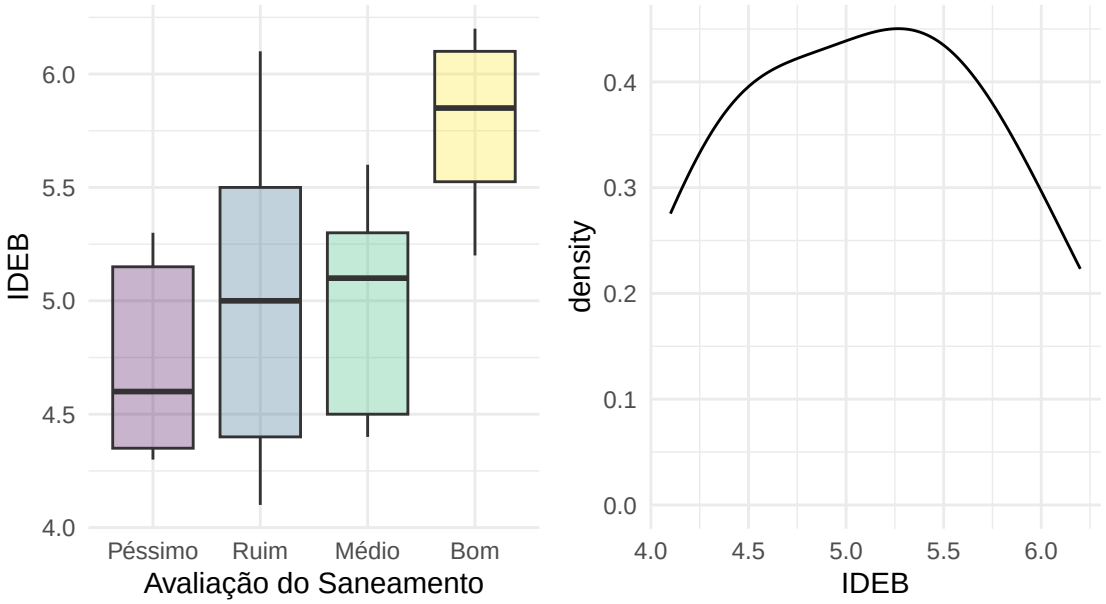

Figura 3: Comportamento da variável: IDEB do ensino fundamental.

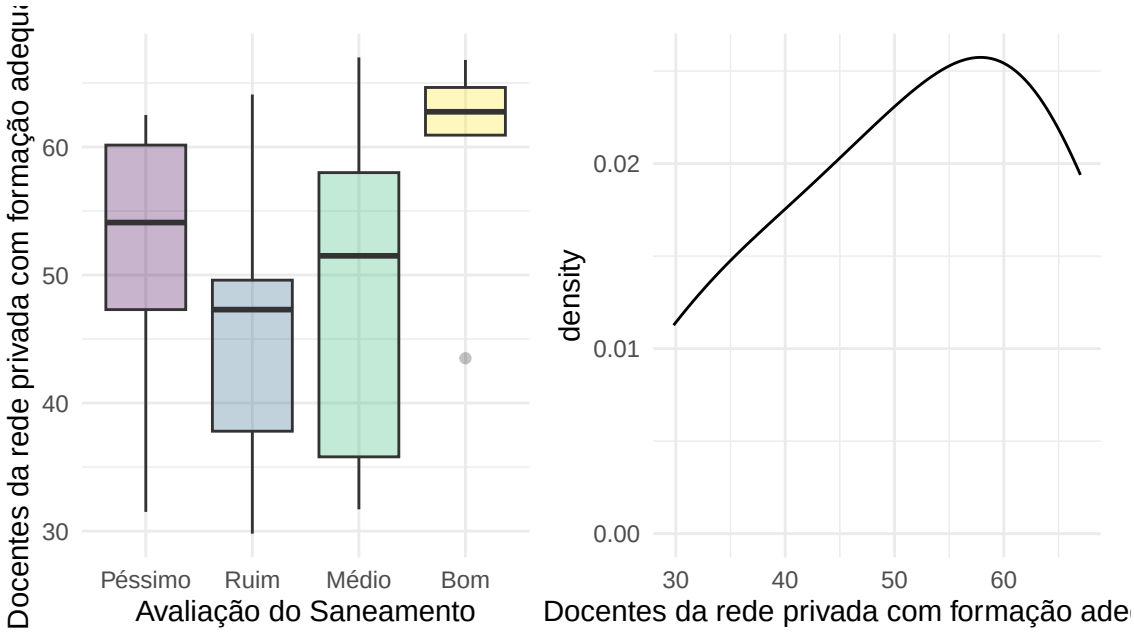

Figura 4: Comportamento da variável: Porcentagem de docentes com formação adequada no ensino privado.

2.2 Variáveis econômicas

Outro tipo de variável que deve impactar diretamente a classificação do saneamento básico são as variáveis econômicas, em que pela Tabela 2, notamos um desvio padrão grande para o PIB per capita e a % de pobres e em todas as medidas

de resumo a participação da agropecuária é inferior a participação da indústria na economia do estado.

Tabela 2: Medidas Resumo das Variáveis Econômicas

variable	mean	median	sd	min	max	na_count
Produto Interno Bruto per capita 2016	17.26	14.22	9.41	8.14	52.5	0
Participação da Agropecuária no Valor Adicionado 2016	9.38	8.57	5.17	0.40	18.3	0
Participação da Indústria no Valor Adicionado 2016	15.69	16.01	5.31	4.68	25.8	0
% de pobres 2016	14.04	13.01	8.42	2.86	28.8	0

É perceptível que a conforme diminui a % de pobres existe uma tendência de aumento na avaliação do saneamento, no entanto com poucas diferenças para a classificação Ruim e Médio. Pela Figura 5, temos que as classificações Péssimo, Ruim e Médio estão com valores muito próximos.

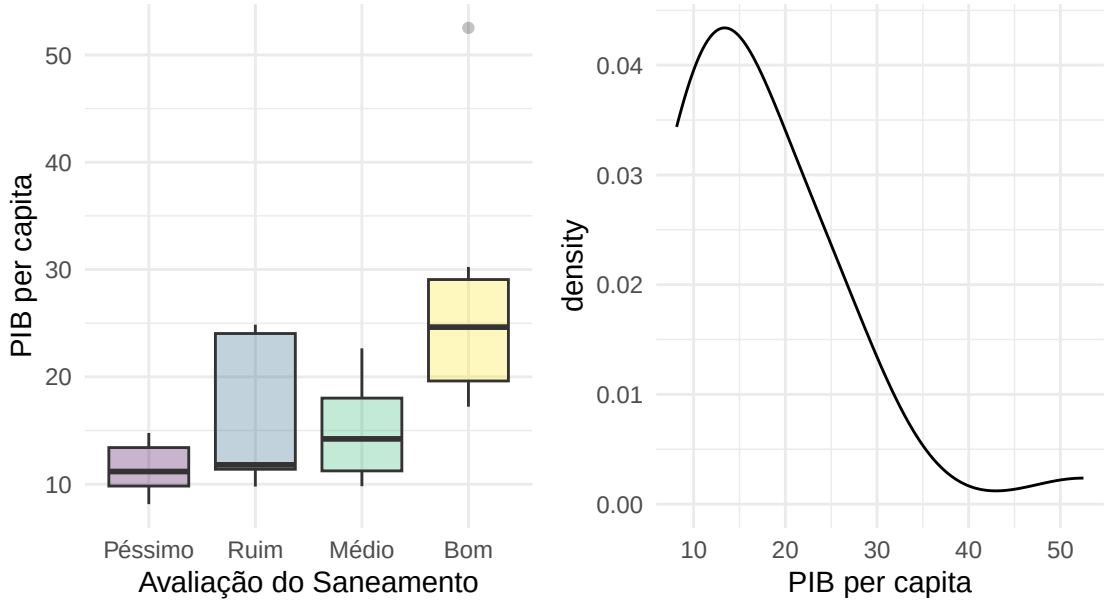

Figura 5: Comportamento da variável: PIB per capita.

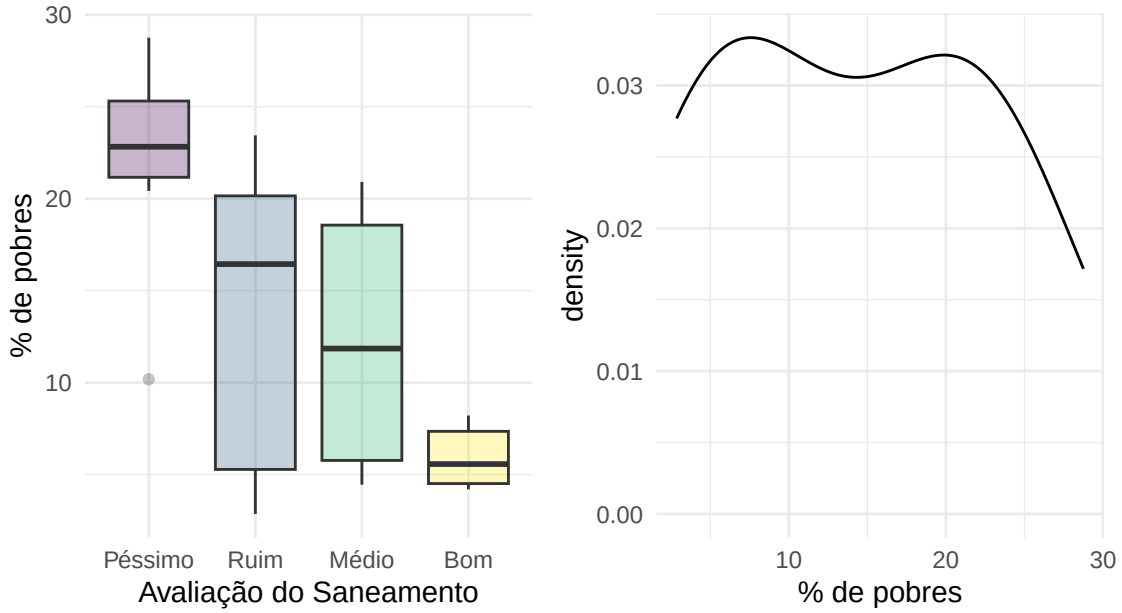

Figura 6: Comportamento da variável: Porcentagem de pessoas em situação de pobreza.

Também nota-se uma tendência de crescimento da participação da indústria conjuntamente com o crescimento da mediana da classificação do saneamento, de acordo com a Figura 7. Entretanto na Figura 8 temos o exato oposto em que podemos notar que conforme cresce a classificação diminui a participação da agropecuária.

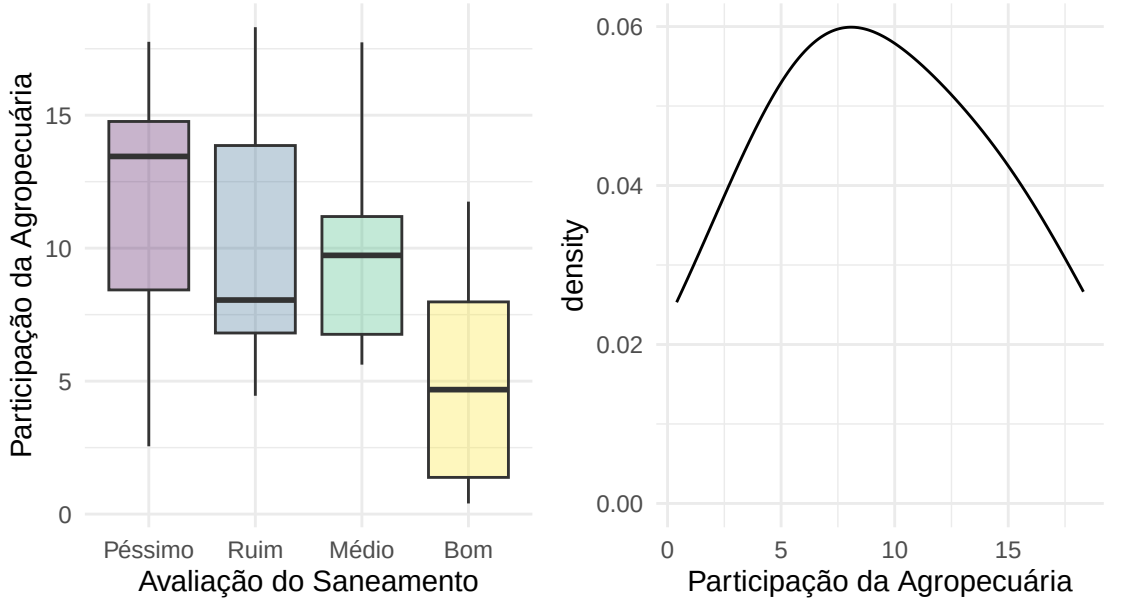

Figura 7: Comportamento da variável: Participação da agropecuária na economia.

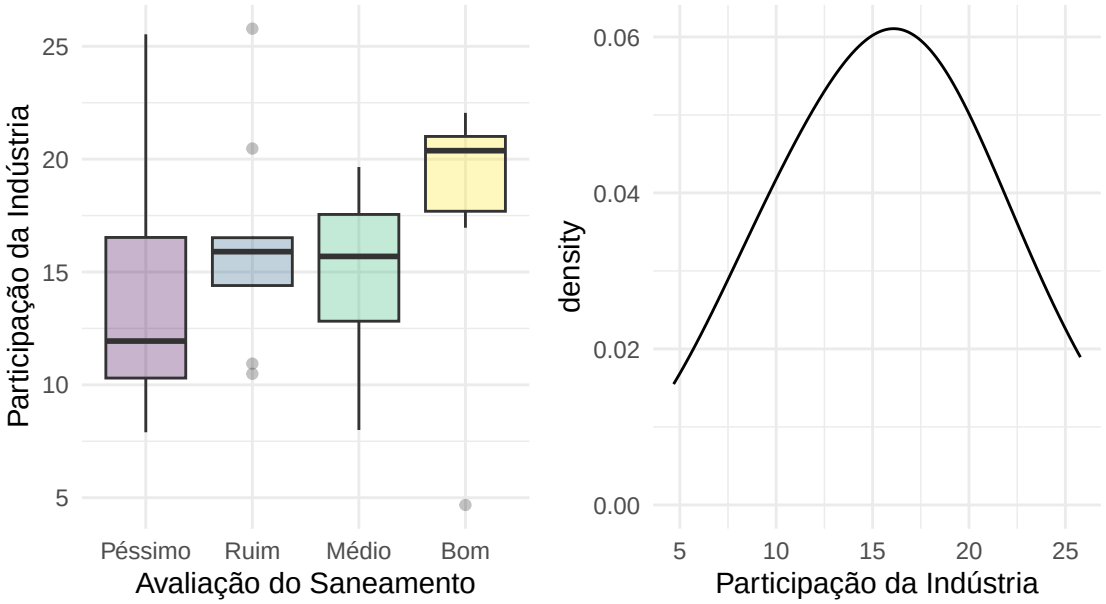

Figura 8: Comportamento da variável: Participação da indústria na economia.

2.3 Variáveis da saúde

Na Tabela 3, tem-se bons valores de nascidos com acompanhamento médico e com bons valores dos nascido com baixo peso. Para o desvio padrão a Mortalidade Infantil e os nascidos com baixo peso tem valores baixos.

Tabela 3: Medidas Resumo das Variáveis da Saúde

variable	mean	median	sd	min	max	na_count
% de nascidos vivos com pelo menos sete consultas de pré-natal 2016	62.31	64.21	11.844	38.60	83.21	0
% de nascidos vivos com baixo peso ao nascer 2016	8.11	7.94	0.716	6.88	9.41	0
Mortalidade infantil 2016	15.09	15.91	4.065	8.81	23.44	0

Nas Figuras 10 e 11, é facilmente notado uma tendência crescente e uma tendência decrescente, respectivamente com os níveis do saneamento. No entanto, a Figura 9 os níveis Ruim e Médio são muito próximos, com a maioria da porcentagem dos nascidos vivos ficando entre 55 e 80%.

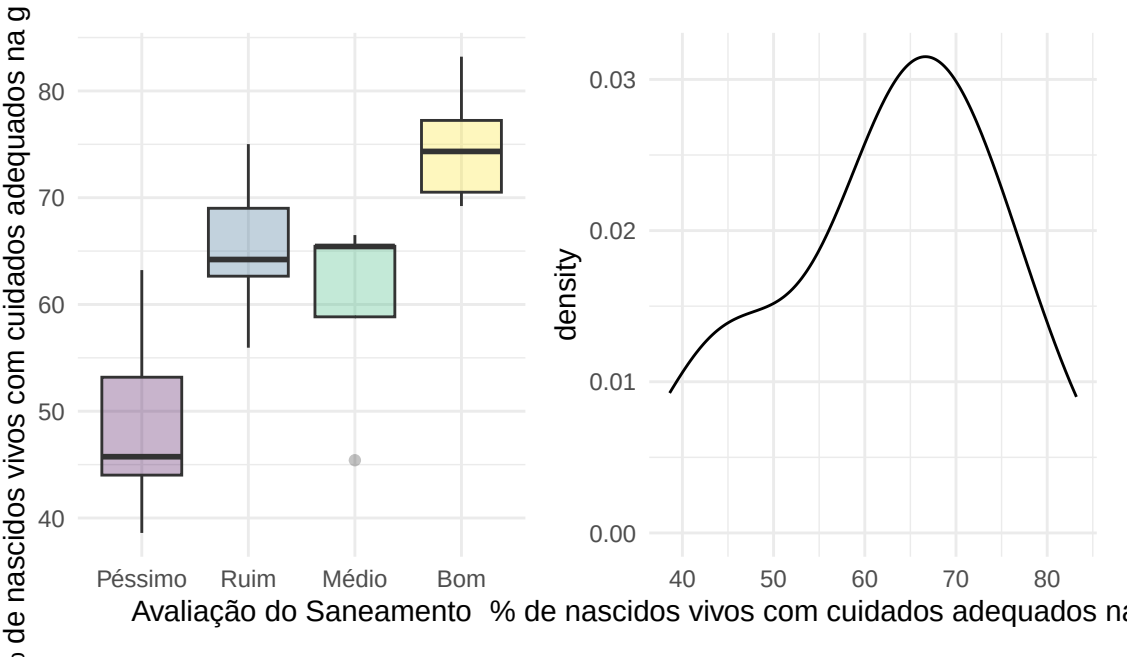

Figura 9: Comportamento da variável: Porcentagem de nascidos vivos com cuidado na gravidez adequado.

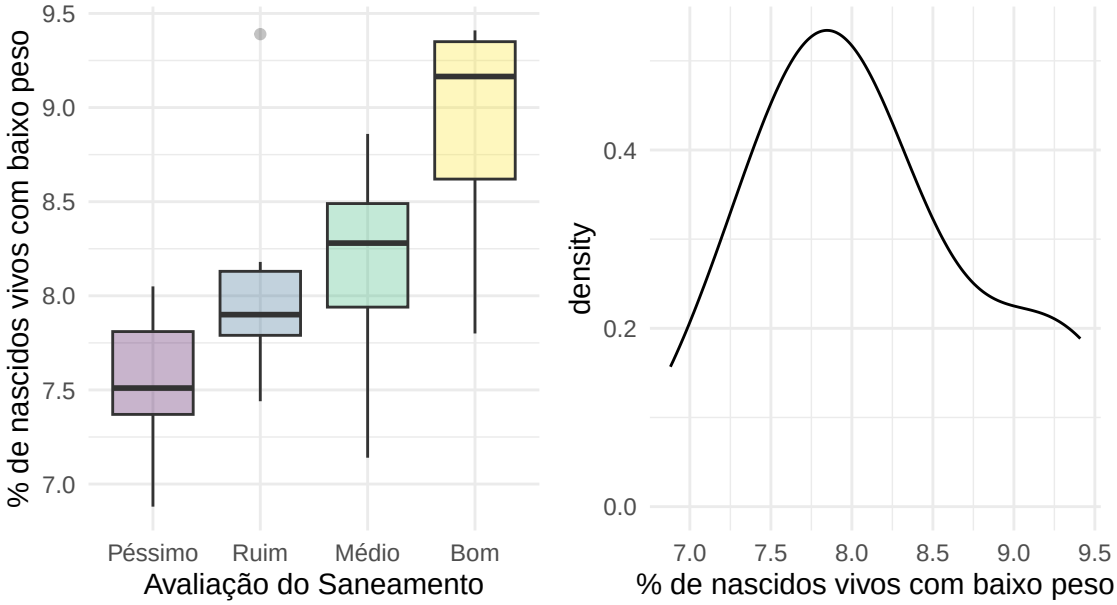

Figura 10: Comportamento da variável: Porcentagem de nascidos vivos com com baixo peso.

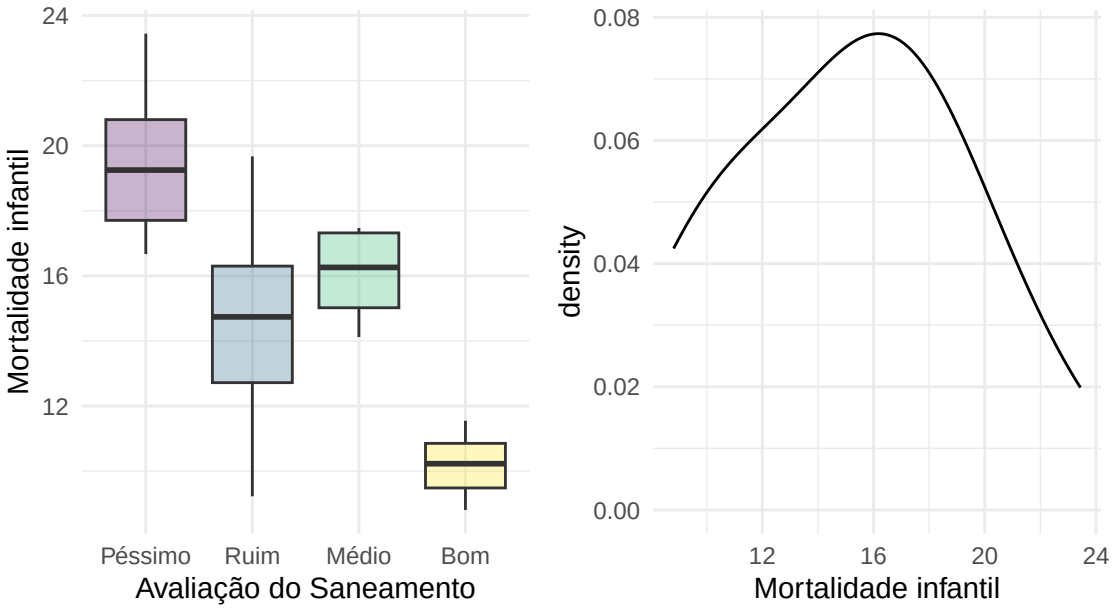

Figura 11: Comportamento da variável: Mortalidade infantil.

2.4 Variáveis sociais

As variáveis sociais são variáveis que medem a desigualdade social e o impacto do desmatamento de vegetação natural. Conforme demonstrado pela Tabela 4 percebemos que as variáveis são completamente opostas quanto ao desvio padrão, com a cobertura vegetal brasileira ficando com valores intermediários de desmatamento, no entanto com alto desvio padrão, ao contrário do índice de gini, com desvio padrão baixo, mas também com média razoável.

Tabela 4: Medidas Resumo das Variáveis Sociais

variable	mean	median	sd	min	max	na_count
% de cobertura vegetal natural 2016	54.650	52.170	24.72	14.680	95.600	0
Índice de Gini 2016	0.516	0.524	0.04	0.421	0.578	0

Percebe-se uma característica peculiar na Figura 12, em que conforme cresce o desmatamento cresce a classificação do saneamento básico, com o intervalo dos valores sendo amplo de 25% a 75%. No entanto, na Figura 13 não é perceptível nenhuma tendência do Índice de Gini com a Avaliação do Saneamento.

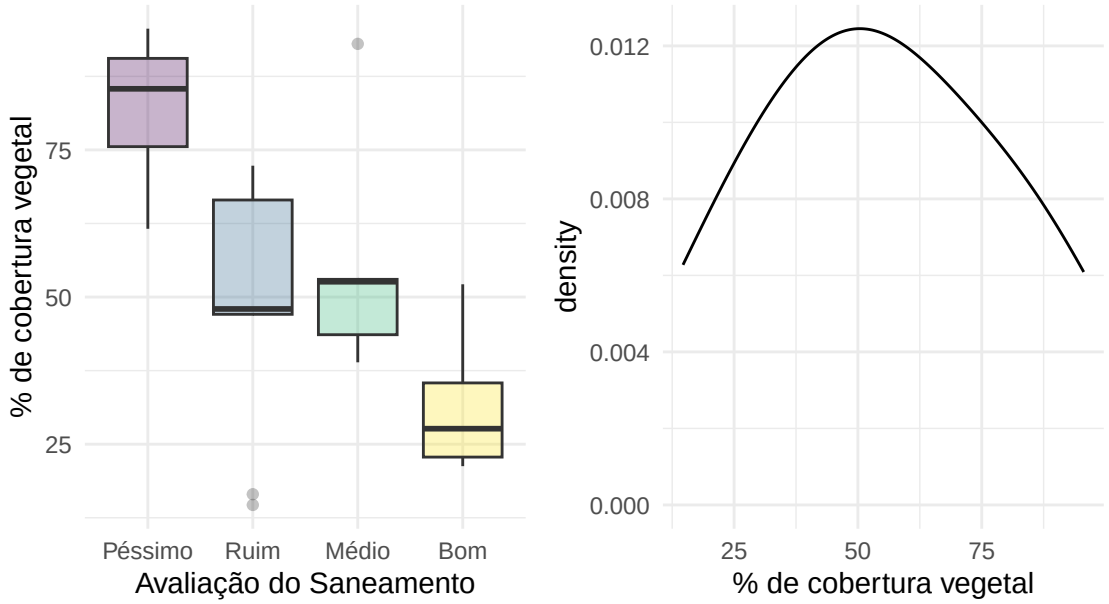

Figura 12: Comportamento da variável: Porcentagem de cobertura vegetal natural.

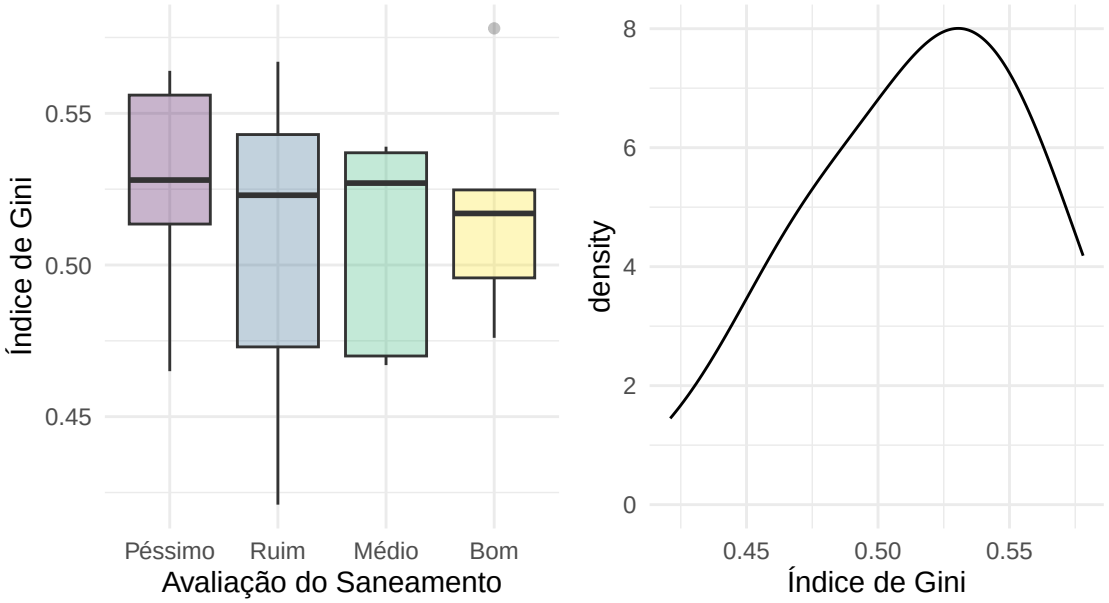

Figura 13: Comportamento da variável: Índice de Gini.

3 Análise Inferencial

Nessa seção realizaremos e verificaremos o ajuste do modelo de regressão beta. Conjuntamente com a análise de diagnóstico, em que busca encontrar possíveis distorções das suposições do modelo, principalmente observações discrepantes e mal especificação do modelo. Para finalizar a análise inferencial é realizado de testes de hipóteses, apresentação de coeficientes e seleção do modelo.

3.1 Análise de Diagnóstico

Nessa subseção realizaremos a investigação de pontos influentes, pois avaliando a existência de observações aberrantes, isto é, pontos que exercem peso desproporcional nas estimativas dos parâmetros do modelo de Regressão Beta, conseguiremos avaliar a qualidade do modelo ajustado e a análise de resíduos, para examinar a adequação da distribuição

3.1.1 Distância de cook e Alavancagem

A Distância de Cook, mede essencialmente a influência das observações sobre os parâmetros e o ajuste, avaliando a influência de o que pequenas pertubarções nas variâncias das observações causam nas estimativas dos parâmetros. Ou de forma simplificada, temos a influência da observação i sobre todos os n valores ajustados.

No entanto a medida alavancagem, que informam se uma observação é discrepante em termos de covariável, ou seja, utilizando os resíduos busca medir a discrepância entre o valor observado e o valor ajustado.

Assim na Figura 14, é notável que não tem nenhuma observação candidata a ponto influente, pois não tem influência desproporcional nas covariáveis e não achata nenhum dos gráficos.

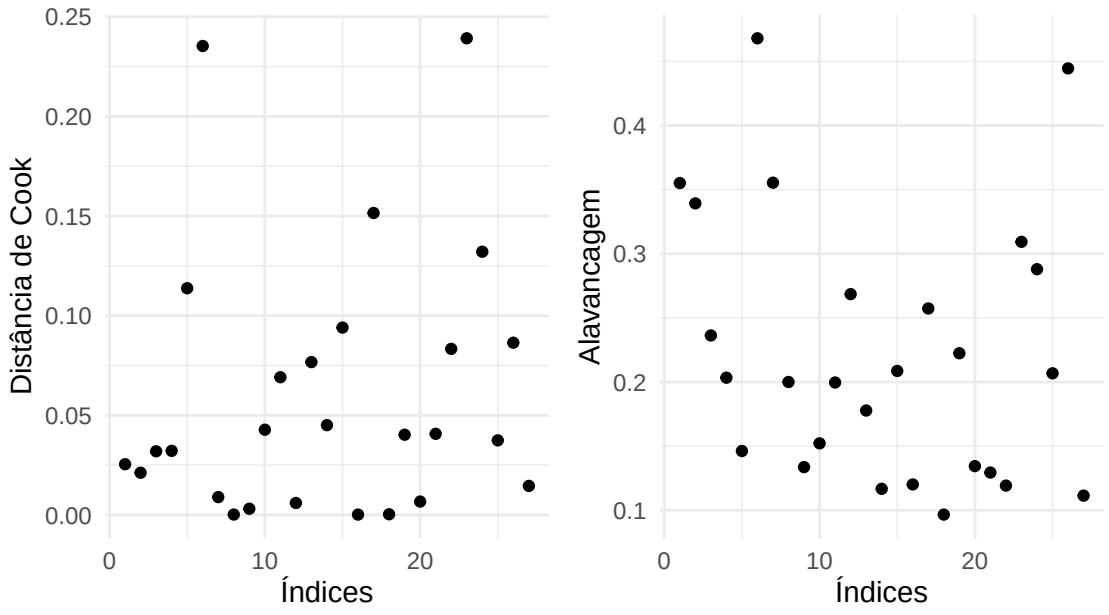

Figura 14: Distância de Cook e Alavancagem para a Regressão Beta do Saneamento Brasileiro

3.1.2 Índices vs Resíduos e valor ajustado vs Resíduos

Os gráficos de resíduos versus índices ou valor ajustado versus resíduos, devem possuem comportamento aleatório e poucos valores acima de -2 e 2 (ou -3 e 3), para que possamos garantir que não há nenhuma evidência de variância não constante.

Portanto percebemos que na Figura 15, não tem nenhum ponto fora de -3 e 3, assim como não conseguimos notar nenhum comportamento de não aleatório.

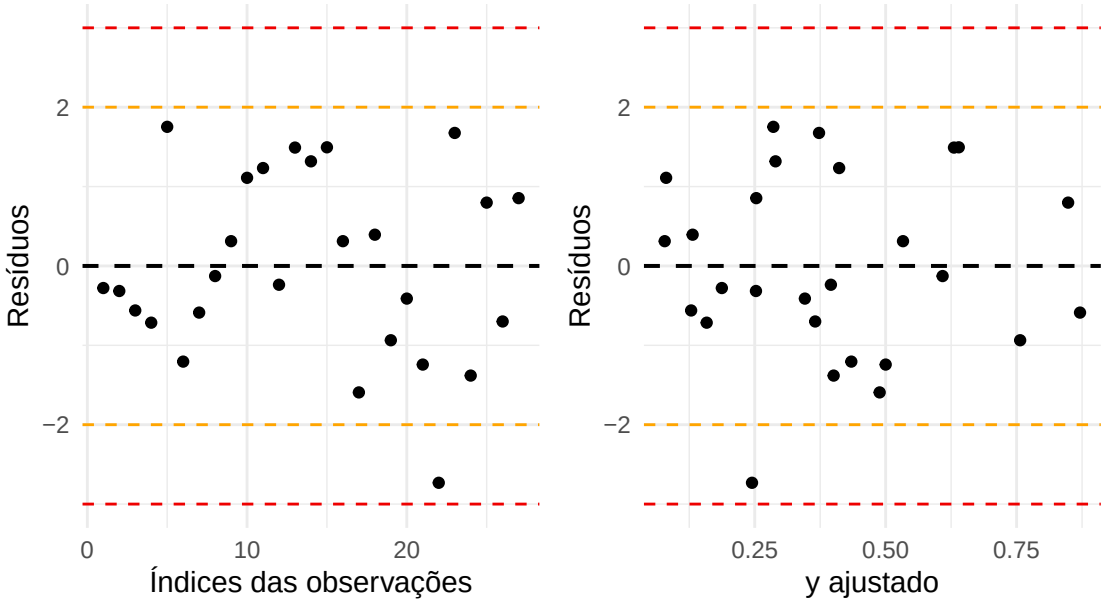

Figura 15: Análise de Resíduo para a Regressão Beta do Saneamento Brasileiro

3.1.3 Envelope Simulado

O envolope simulado fornece a comparação entre os resíduos e os percentis da distribuição, nos dando a ideia se a distribuição é adequada para o ajuste, como percebemos na Figura 16, em que todas as observações estão dentro das bandas de confiança.

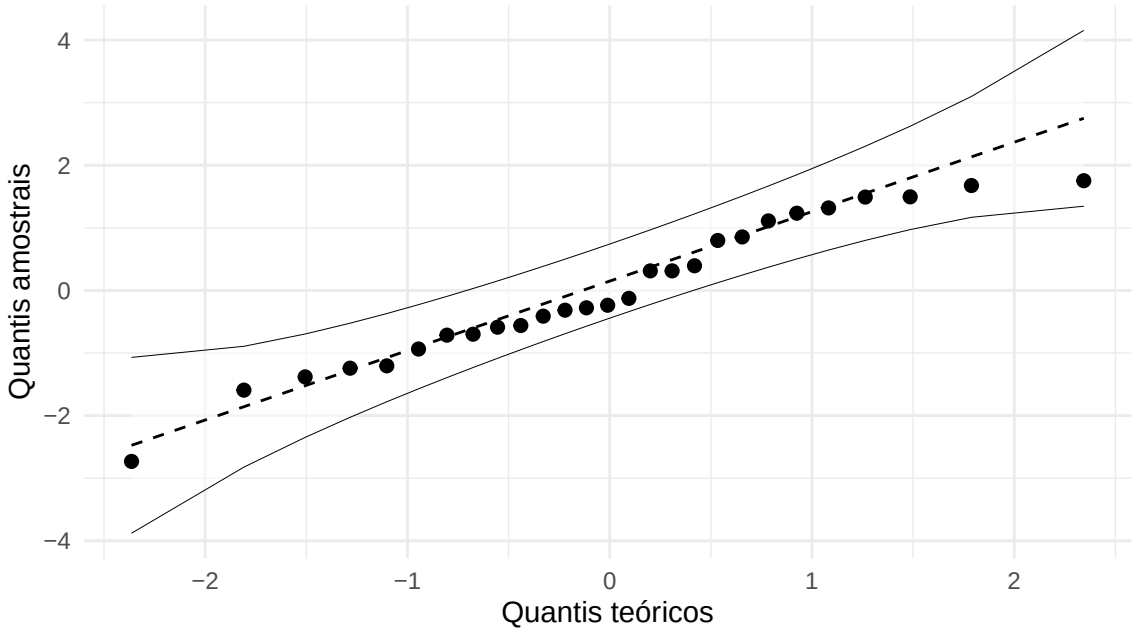

Figura 16: Envelope Simulado da Regressão Beta do Saneamento Brasileiro

3.2 Critérios de Seleção de Modelos

Na análise inferencial temos inúmeros procedimentos para a seleção de modelos. Porém os os critérios que se destacam são AIC e BIC, no entanto nessa análise também utilizaremos o BIC corrigido, Hannan-Quinnn e Hannan-Quinnn corrigido conforme podemos ver na Tabela 5. Em uma visão geral podemos dizer que esses são processos de minimização que não envolvem testes, com a ideia de buscar um modelo que seja parcimonioso.

Tabela 5: Resumo dos critérios de seleção de molelos para função de ligação

	Logit	Probit	Cloglog	Cauchit	Loglog
AIC	-34.62	-34.06	-34.52	-35.55	-30.91
BIC	-25.55	-24.99	-25.45	-26.48	-21.84
BICc	-15.84	-15.28	-15.73	-16.76	-12.13
HQ	-2.48	-1.92	-2.38	-3.41	1.23
HQc	-3.41	-47.19	-47.64	-48.67	-44.03

A função de ligação com melhor resultado em todos os critérios é a Cauchit. Todavia, essa mesma função de ligação possui a menor explicação da variável resposta, de acordo com a Tabela 6. Assim, prosseguiremos a analise com a função de ligação Logit, que possui valores maiores que a Cauchit nos critérios de seleção (quanto menor, melhor) e apenas fica atrás da Probit no pseudo R^2, sendo a função de ligação mais equilibrada.

Tabela 6: Pseudo r quadrado para cada função de ligação

	Pseudo.R.2
Logit	0.798
Probit	0.802
Cloglog	0.770
Cauchit	0.679
Loglog	0.793

3.3 Testes

Terminamos a análise inferencial com a avaliação do modelo de Regressão Beta ajustado, a análise de significância das variáveis dadas pelo seguinte teste de hipótese:

$$H_0 : \beta_i = 0 \text{ (Covariável não-significativa)}$$

$$H_0 : \beta_i \neq 0 \text{ (Covariável significante)}$$

O qual podemos analisar com a Tabela 7.

Tabela 7: Estatísticas do Modelo Ajustado

	Estimativa	Desvio padrão	Estatística z	P.valor
(Intercept)	-10.504	2.345	-4.48	<0.001*
'% de docentes na rede privada do fundamental com formação adequada 2016'	-0.026	0.012	-2.07	0.039*
'IDEB anos iniciais do ensino fundamental 2015'	0.813	0.256	3.17	0.001*
'% de cobertura vegetal natural 2016'	-0.012	0.005	-2.40	0.016*
'% de pobres 2016'	-0.114	0.027	-4.24	<0.001*
'Índice de Gini 2016'	18.274	3.651	5.00	<0.001*

A Tabela 7 detalha, algumas estatísticas muito importantes sobre as covariáveis, mas principalmente informa que todas as variáveis preditivas são significativas e que a melhor combinação para a estimação da Porcentagem da população urbana residente em domicílios ligados à rede de esgotamento sanitário no ano de 2017 tem variáveis de praticamente todos os tipos: Sociais (Índice de Gini e porcentagem de cobertura vegetal natural), Educacional (Porcentagem de docentes na rede privada do fundamental com formação adequada e IDEB nos anos iniciais do ensino fundamental) e Econômica (Porcentagem de pobres).

4 Conclusão

Constatamos que as variáveis propostas para o modelo promovem conjuntamente uma boa explicação para o Saneamento brasileiro, quanto maior o investimento em políticas públicas para educação e econômica, o estado tende a focar no desenvolvimento de um sistema cloacal mais amplo, ou seja, basicamente o que define a porcentagem de esgoto tratato é o desenvolvimento do estado.

5 Apêndice

```
options(digits = 3)
options(scipen = 999)
ggplot2::theme_set(ggplot2::theme_minimal())
knitr::opts_chunk$set(echo=F, message=F, warning=F, fig.pos = 'H',
                      fig.align = 'center', fig.width = 6, fig.height= 3.4)
scale_fill_discrete = \(...) ggplot2::scale_fill_brewer(... , palette = "Set2")
library(dplyr)
library(magrittr)
library(betareg)
library(zoo)
library(ggplot2)
library(qqplotr)
library(patchwork)

fit2df<-function(fit) {
  summary(fit) |>
    (\(x) x$coefficients)() |>
    data.frame() %>%
    .[,0:4]|>
    round(3) |>
    mutate(P.valor = ifelse(
      `mean.Pr...z..` < 0.001,"<0.001*",
      ifelse(`mean.Pr...z..` < 0.05, paste0(`mean.Pr...z..`, '*', sep = ''), `mean.Pr...z..`))) |>
    select(-`mean.Pr...z..`,
      "Estimativa" = "mean.Estimate",
      "Desvio padrão" = "mean.Std..Error",
      "Estatística z" = "mean.z.value"
    )
}

boxplot <- function(v1, y){
  ggplot(df, aes(x=Avaliacao_Saneamento, y={{v1}}, fill = Avaliacao_Saneamento)) +
    geom_boxplot(alpha=0.3) +
    theme(legend.position="none") + labs(x = "Avaliação do Saneamento", y = y)
}
density <- function(v1, x){
  ggplot(data=df, aes(x={{v1}})) +
    geom_density(adjust=1.5, alpha=.4) +
    labs(x = x)
}
dot <- function(v1){
  ggplot(df, aes(x={{v1}}, y=Saneamento)) +
  geom_point() +
  geom_smooth(method=lm , color="red", fill="#69b3a2", se=TRUE)
}
data = readr::read_csv('./data/data_modificado.csv')
df = data[2:28, 2:ncol(data)]
df = df %>%
  mutate(Saneamento =
```

```
          `% da população urbana residente em domicílios ligados à rede de esgotamento sanitário 2017`/100)
df = df %>%
  mutate(Avaliacao_Saneamento = case_when(
                                          (Saneamento < 0.2) ~ 'Péssimo',
                                          (Saneamento >= 0.2 & Saneamento < 0.4) ~ 'Ruim',
                                          (Saneamento >= 0.4 & Saneamento < 0.6) ~ 'Médio',
                                          (Saneamento >= 0.6) ~ 'Bom')) %>%
  mutate(Avaliacao_Saneamento = ordered(Avaliacao_Saneamento,
                                        levels = c("Péssimo", "Ruim", "Médio",
                                                   "Bom")))
fastrep::describe(df %>%
                    select(`% dos ocupados com ensino fundamental completo 2010`,
                           `% dos ocupados com ensino médio completo 2010`,
                           `IDEB anos iniciais do ensino fundamental 2015`,
                           `% de docentes na rede privada do fundamental com formação adequada 2016`)) %>%
  fastrep::tbl("Medidas Resumo das Variáveis Educacionais")
boxplot(`% dos ocupados com ensino fundamental completo 2010`, "Ocupados com Ensino Fundamental") +
  density(`% dos ocupados com ensino fundamental completo 2010`, "Ocupados com Ensino Fundamental")
boxplot(`% dos ocupados com ensino médio completo 2010`, "Ocupados com Ensino Médio") +
  density(`% dos ocupados com ensino médio completo 2010`, "Ocupados com Ensino Médio")
boxplot(`IDEB anos iniciais do ensino fundamental 2015`, "IDEB") +
  density(`IDEB anos iniciais do ensino fundamental 2015`, "IDEB")
boxplot(`% de docentes na rede privada do fundamental com formação adequada 2016`, "Docentes da rede privada
  density(`% de docentes na rede privada do fundamental com formação adequada 2016`, "Docentes da rede priva
fastrep::describe(df %>%
                    select(`Produto Interno Bruto per capita 2016`,
                           `Participação da Agropecuária no Valor Adicionado 2016`,
                           `Participação da Indústria no Valor Adicionado 2016`,
                           `% de pobres 2016`)) %>%
  fastrep::tbl("Medidas Resumo das Variáveis Econômicas")
boxplot(`Produto Interno Bruto per capita 2016`, "PIB per capita") +
  density(`Produto Interno Bruto per capita 2016`, "PIB per capita")
boxplot(`% de pobres 2016`, "% de pobres") +
  density(`% de pobres 2016`, "% de pobres")
boxplot(`Participação da Agropecuária no Valor Adicionado 2016`, "Participação da Agropecuária") +
  density(`Participação da Agropecuária no Valor Adicionado 2016`, "Participação da Agropecuária")
boxplot(`Participação da Indústria no Valor Adicionado 2016`, "Participação da Indústria") +
  density(`Participação da Indústria no Valor Adicionado 2016`, "Participação da Indústria")
fastrep::describe(df %>%
                    select(`% de nascidos vivos com pelo menos sete consultas de pré-natal 2016`,
                           `% de nascidos vivos com baixo peso ao nascer 2016`,
                           `Mortalidade infantil 2016`)) %>%
  fastrep::tbl("Medidas Resumo das Variáveis da Saúde")
boxplot(`% de nascidos vivos com pelo menos sete consultas de pré-natal 2016`, "% de nascidos vivos com cuid
  density(`% de nascidos vivos com pelo menos sete consultas de pré-natal 2016`, "% de nascidos vivos com cu
boxplot(`% de nascidos vivos com baixo peso ao nascer 2016`, "% de nascidos vivos com baixo peso") +
  density(`% de nascidos vivos com baixo peso ao nascer 2016`, "% de nascidos vivos com baixo peso")
boxplot(`Mortalidade infantil 2016`, "Mortalidade infantil") +
  density(`Mortalidade infantil 2016`, "Mortalidade infantil")
fastrep::describe(df %>%
                    select(`% de cobertura vegetal natural 2016`,
```

```
                          `Índice de Gini 2016`)) %>%
  fastrep::tbl("Medidas Resumo das Variáveis Sociais")
boxplot(`% de cobertura vegetal natural 2016`, "% de cobertura vegetal") +
  density(`% de cobertura vegetal natural 2016`, "% de cobertura vegetal")
boxplot(`Índice de Gini 2016`, "Índice de Gini") +
  density(`Índice de Gini 2016`, "Índice de Gini")
fit1 <- betareg(Saneamento ~
                  `% de docentes na rede privada do fundamental com formação adequada 2016` +
                  `IDEB anos iniciais do ensino fundamental 2015`+
                  `% de cobertura vegetal natural 2016` +
                  `% de pobres 2016` +
                  `Índice de Gini 2016`,
                data = df, x=TRUE)
residuot2<- residuals(fit1, type= "sweighted2")
yajust<-fitted.values(fit1)
yhat=hatvalues(fit1)
dcook<- cooks.distance(fit1)
residuot2_df<- data.frame(residuot2)
deviance<- sum(residuals(fit1, tipe= "deviance")^2)
ggplot(df, aes(x=index(Saneamento), y= dcook))+
  geom_point(size=1.5)+
  labs(x = "Índices", y = "Distância de Cook") +
  ggplot(df, aes(x=index(Saneamento), y=yhat))+
  geom_point(size=1.5) +
  labs(x = "Índices", y = "Alavancagem")
ggplot(df, aes(x=index(Saneamento),y=residuot2))+
  geom_point(size=1.5) +
  geom_hline(yintercept=3, colour="red2",
             linewidth=0.5,  linetype="dashed") +
  geom_hline(yintercept=2, colour="orange",
             linewidth=0.5,  linetype="dashed") +
  geom_hline(yintercept=-2, colour="orange",
             linewidth=0.5,  linetype="dashed") +
  geom_hline(yintercept=0, colour="black",
             linewidth=0.7, linetype="dashed") +
  geom_hline(yintercept=-3, colour="red2",
             linewidth=0.5,  linetype="dashed")+
  labs(x = "Índices das observações", y = "Resíduos") +
ggplot(df, aes(x=yajust, y=residuot2))+
  geom_point(size=1.5) +
  geom_hline(yintercept=3, colour="red2", size=0.5,
             linetype="dashed") +
  geom_hline(yintercept=2, colour="orange",
             linewidth=0.5,  linetype="dashed") +
  geom_hline(yintercept=-2, colour="orange",
             linewidth=0.5,  linetype="dashed") +
  geom_hline(yintercept=0, colour="black", size=0.7,
             linetype="dashed") +
  geom_hline(yintercept=-3, colour="red2", size=0.5,
             linetype="dashed")+
  labs(x = "y ajustado", y = "Resíduos")
```

```r
ggplot(data = residuot2_df,
       mapping = aes(sample = residuot2))+
  geom_qq_band( alpha = 0.5, fill="white", col="black") +

  stat_qq_line(size=0.5, linetype="dashed") +
  stat_qq_point(size=2) +
  scale_fill_discrete("Bandtype")  +
  labs(x = "Quantis teóricos", y = "Quantis amostrais")
fit1 <- betareg(Saneamento ~
                  `% de docentes na rede privada do fundamental com formação adequada 2016` +
                  `IDEB anos iniciais do ensino fundamental 2015`+
                  `% de cobertura vegetal natural 2016` +
                  `% de pobres 2016` +
                  `Índice de Gini 2016`,
                data = df, x=TRUE)
fit1a <- betareg(Saneamento ~
                  `% de docentes na rede privada do fundamental com formação adequada 2016` +
                  `IDEB anos iniciais do ensino fundamental 2015`+
                  `% de cobertura vegetal natural 2016` +
                  `% de pobres 2016` +
                  `Índice de Gini 2016`,
                data = df, x=TRUE,
                 link = "probit")
fit1b <- betareg(Saneamento ~
                  `% de docentes na rede privada do fundamental com formação adequada 2016` +
                  `IDEB anos iniciais do ensino fundamental 2015`+
                  `% de cobertura vegetal natural 2016` +
                  `% de pobres 2016` +
                  `Índice de Gini 2016`,
                data = df, x=TRUE,
                 link = "cloglog")
fit1c <- betareg(Saneamento ~
                  `% de docentes na rede privada do fundamental com formação adequada 2016` +
                  `IDEB anos iniciais do ensino fundamental 2015`+
                  `% de cobertura vegetal natural 2016` +
                  `% de pobres 2016` +
                  `Índice de Gini 2016`,
                data = df, x=TRUE,
                 link = "cauchit")
fit1d <- betareg(Saneamento ~
                  `% de docentes na rede privada do fundamental com formação adequada 2016` +
                  `IDEB anos iniciais do ensino fundamental 2015`+
                  `% de cobertura vegetal natural 2016` +
                  `% de pobres 2016` +
                  `Índice de Gini 2016`,
                data = df, x=TRUE,
                 link = "loglog")
n=length(fit1$y)
k=1+length(fit1$coefficients$mean)
BICc=-2*fit1$loglik+(n*k*log(n)/(n-k-1))
HQ=-2*fit1$loglik+(2*k*log(n))
```

```
HQc=-2*fit1$loglik+(2*k*log(log(n))/(n-k-1))

ka=1+length(fit1a$coefficients$mean)
BICca=-2*fit1a$loglik+(n*ka*log(n)/(n-ka-1))
HQa=-2*fit1a$loglik+(2*ka*log(n))
HQca=-2*fit1a$loglik+(2*ka*log(log(n))/(n-ka-1))

kb=1+length(fit1b$coefficients$mean)
BICcb=-2*fit1b$loglik+(n*kb*log(n)/(n-kb-1))
HQb=-2*fit1b$loglik+(2*kb*log(n))
HQcb=-2*fit1b$loglik+(2*kb*log(log(n))/(n-kb-1))

kc=1+length(fit1c$coefficients$mean)
BICcc=-2*fit1c$loglik+(n*kc*log(n)/(n-kc-1))
HQc=-2*fit1c$loglik+(2*kc*log(n))
HQcc=-2*fit1c$loglik+(2*kc*log(log(n))/(n-kc-1))

kd=1+length(fit1d$coefficients$mean)
BICcd=-2*fit1d$loglik+(n*kd*log(n)/(n-kd-1))
HQd=-2*fit1d$loglik+(2*kd*log(n))
HQcd=-2*fit1d$loglik+(2*kd*log(log(n))/(n-kd-1))

metrics = data.frame("Logit" = c(AIC(fit1), BIC(fit1), BICc, HQ, HQc),
           "Probit" = c(AIC(fit1a), BIC(fit1a), BICca, HQa, HQca),
           "Cloglog" = c(AIC(fit1b), BIC(fit1b), BICcb, HQb, HQcb),
           "Cauchit" = c(AIC(fit1c), BIC(fit1c), BICcc, HQc, HQcc),
           "Loglog" = c(AIC(fit1d), BIC(fit1d), BICcd, HQd, HQcd))
rownames(metrics) = c('AIC', 'BIC', 'BICc', 'HQ', 'HQc')

metrics %>% fastrep::tbl('Resumo dos critérios de seleção de molelos para função de ligação')

r2 = data.frame("Pseudo R^2" = c(fit1$pseudo.r.squared, fit1a$pseudo.r.squared,
                            fit1b$pseudo.r.squared, fit1c$pseudo.r.squared,
                            fit1d$pseudo.r.squared))
rownames(r2) = c('Logit', 'Probit', 'Cloglog', 'Cauchit', 'Loglog')
r2 %>%
  fastrep::tbl('Pseudo r quadrado para cada função de ligação')

fit2df(fit1) %>%
  fastrep::tbl('Estatísticas do Modelo Ajustado')
```