



UNIVERSIDADE FEDERAL DE MINAS GERAIS

**MÍNIMOS QUADRADOS ORDINÁRIOS: UMA APLICAÇÃO NA ANÁLISE DAS
QUESTÕES INSTITUCIONAIS DE MUNICÍPIOS BRASILEIROS**

Flávio Hugo Pangrácio Silva 
flaviopangracio@cedeplar.ufmg.br
Cedeplar - UFMG

Guilherme Gomes Ferreira 
guilhermegf2019@cedeplar.ufmg.br
Cedeplar - UFMG

Belo Horizonte - MG

Abril - 2024

LISTA DE FIGURAS

| | | |
|----|---|----|
| 1 | O Modelo Clássico de Regressão Linear | 4 |
| 2 | Histograma do PIB per capita dos municípios de Minas Gerais (2018) | 9 |
| 3 | Histograma do Coeficiente de Intensidade da Gestão Empresarial (CIGE) dos municípios de Minas Gerais (2018) | 10 |
| 4 | Histograma da Centralidade da Gestão Pública (CGP) dos municípios de Minas Gerais (2018) | 10 |
| 5 | Mapa de calor da matriz de correlação. | 11 |
| 6 | Boxplot da Centralidade da Gestão Pública (CGP) dos municípios de Minas Gerais (2018) | 12 |
| 7 | Boxplot do Coeficiente de Intensidade da Gestão Empresarial (CIGE) dos municípios de Minas Gerais (2018) | 12 |
| 8 | Boxplot do PIB per capita dos municípios de Minas Gerais (2018) | 13 |
| 9 | PIB per capita dos municípios de Minas Gerais - REGIC 2018 | 14 |
| 10 | CIGE dos municípios de Minas Gerais - REGIC 2018 | 15 |
| 11 | CGP dos municípios de Minas Gerais - REGIC 2018 | 16 |
| 12 | PIB per capita x CIGE (Municípios de Minas Gerais - REGIC 2018) | 17 |
| 13 | PIB per capita x CGP (Municípios de Minas Gerais - REGIC 2018) | 17 |

SUMÁRIO

| | | |
|----------|---|-----------|
| 1 | INTRODUÇÃO | 1 |
| 2 | O MODELO CLÁSSICO DE REGRESSÃO LINEAR | 1 |
| 2.1 | Linearidade do modelo | 2 |
| 2.2 | Posto Completo | 2 |
| 2.3 | Exogeneidade | 2 |
| 2.4 | Homocedasticidade e não autocorrelação residual | 3 |
| 2.5 | Processo Gerador dos dados para a regressão | 3 |
| 2.6 | Normalidade dos erros | 4 |
| 3 | REGRESSÃO POR MÍNIMOS QUADRADOS | 4 |
| 3.1 | Ajuste do modelo | 6 |
| 3.1.1 | Grau de Ajuste | 6 |
| 4 | APLICAÇÃO | 7 |
| 4.1 | Análise Descritiva | 8 |
| 4.2 | Análise de Regressão | 16 |
| 4.3 | Análise de Regressão Múltipla | 19 |
| 5 | REFERÊNCIAS | 21 |

1. INTRODUÇÃO

O presente trabalho se propõe a explorar de maneira detalhada o método de mínimos quadrados ordinários (MQO), apresentando uma aplicação na análise das questões institucionais presentes nos municípios brasileiros. Este método estatístico é amplamente utilizado na análise econômica, sendo fundamental para compreender as relações entre variáveis e realizar previsões.

A escolha desse enfoque se justifica pela relevância crescente do estudo das instituições no contexto municipal brasileiro, visto que as políticas públicas e a gestão eficiente dessas instituições desempenham um papel fundamental no desenvolvimento socioeconômico local. Nesse sentido, compreender como diferentes variáveis institucionais estão relacionadas entre si e como influenciam indicadores de crescimento e desenvolvimento municipal torna-se uma questão de interesse.

Por meio deste trabalho, pretendemos não apenas apresentar a aplicação prática do modelo de MQO, mas também fornecer uma base sólida de compreensão teórica, destacando os fundamentos matemáticos e estatísticos subjacentes a esse método. Para isso, organizaremos o conteúdo em várias seções, nas quais abordaremos desde os princípios básicos da regressão linear até aspectos mais avançados, passando pela discussão sobre a formulação teórica do modelo de MQO.

Inicialmente, abordaremos os principais conceitos e definições relacionados à regressão linear, discutindo os pressupostos e as limitações desse modelo estatístico. Posteriormente, dedicaremos atenção especial à formulação teórica do modelo de MQO, descrevendo o processo de estimativa dos parâmetros e apresentando as principais propriedades estatísticas dos estimadores obtidos por esse método. Além disso, discutiremos técnicas de diagnóstico e avaliação da qualidade do modelo, destacando a importância da interpretação correta dos resultados obtidos.

Por fim, demonstraremos a aplicação do modelo de MQO na análise das questões institucionais de municípios brasileiros, utilizando dados da REGIC 2018 para ilustrar o processo de formulação, estimação e interpretação do modelo. Espera-se que este trabalho contribua para ampliar o entendimento sobre o método de MQO e sua aplicação.

2. O MODELO CLÁSSICO DE REGRESSÃO LINEAR

A priori, antes de adentrar em detalhes do estimador de MQO, é preciso explicar o modelo clássico de regressão linear, bem como suas hipóteses subjacentes. Nesse sentido, deve se salientar que o modelo clássico de regressão linear admite a forma simples e a forma múltipla. No modelo simples, também conhecido como modelo de regressão bivariada, temos apenas uma variável explicada e uma variável explicativa, além de um intercepto e dos resíduos do modelo.

Um problema fundamental do modelo de regressão simples, no entanto, é a dificuldade de fazer uma análise parcial com apenas uma variável explicativa, ignorando todas outras variáveis que afetam a variável explicada, Y , e são não correlacionadas com a variável independente, X . É nesse sentido que existe o modelo de regressão linear múltipla, o qual permite explicar uma variável através de uma junção de mais variáveis independentes e não correlacionadas uma com a outra. Doravante, este trabalho focará no modelo de regressão linear múltipla, com a justificativa de que os pressupostos são análogos aos pressupostos do modelo simples e que com mais variáveis, o que só é permitido neste modelo, é possível fazer uma análise mais robusta.

Nesta perspectiva, para a definição do modelo clássico de regressão linear, são necessárias algumas hipóteses:

2.1. Linearidade do modelo

A primeira hipótese implica que o modelo deve ser linear nos parâmetros estimados. Disso decorre que as variáveis explicativas podem ser não lineares. Essa hipótese basicamente indica que a relação das variáveis independentes com o parâmetro estimado é linear (1), ou seja, uma variação marginal nas variáveis independentes resultará em uma variação constante na variável explicada.

$$y = x_1\beta_1 + x_2\beta_2 + \dots + x_k\beta_k + \varepsilon \quad (1)$$

2.2. Posto Completo

Essa hipótese é uma condição necessária do MCRL, haja vista que, se não satisfeita, é impossível estimar os parâmetros do modelo. Em termos matriciais, implica que a matriz das variáveis independentes deve ser não singular o que, por sua vez, exige que essas variáveis não sejam combinações lineares perfeitas umas das outras. Também é conhecida como condição de identificação.

2.3. Exogeneidade

Tal condição garante que a média condicional do erro dadas as variáveis explicativas é igual a zero. Também conhecida como exogeneidade estrita, seu significado é de que as variáveis explicativas não possuem relação com o termo de perturbação (2). Além disso, é importante ressaltar que, como a média condicional do erro é zero, sua média incondicional também é zero, o que é garantido pela lei das expectativas iteradas (3). Essa é uma forte implicação que garante que uma estimação pelo MCRL sempre acerta na média. Ademais, o MCRL garante a aleatoriedade dos resíduos, isto é, a média condicional do erro i , dado um erro j qualquer é zero.

$$E[\boldsymbol{\varepsilon}|\mathbf{X}] = \begin{bmatrix} E[\varepsilon_1|\mathbf{X}] \\ E[\varepsilon_2|\mathbf{X}] \\ \vdots \\ E[\varepsilon_n|\mathbf{X}] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \quad (2)$$

$$E[\varepsilon_i] = E_{\mathbf{X}}[E[\varepsilon|\mathbf{X}]] = E_{\mathbf{X}}[0] = 0 \quad (3)$$

2.4. Homocedasticidade e não autocorrelação residual

Essa quarta hipótese define que a variância condicional do erro é constante (4) e que a covariância condicional dos erros é zero (3). A variância constante é conhecida como homocedasticidade, o que significa que para qualquer ponto da amostra, a variância sempre será a mesma. Quando isso não ocorre, dizemos que a variância é heterocedástica.

$$Var[\varepsilon_i|\mathbf{X}] = \sigma^2, \quad \forall i \in \{1, \dots, n\}. \quad (4)$$

$$Cov[\varepsilon_i, \varepsilon_j|\mathbf{X}] = 0, \quad \forall i \neq j. \quad (5)$$

Já o fato da covariância condicional dos erros ser igual a zero define a não autocorrelação entre os termos de perturbação. Em termos matriciais, temos que a matriz de erros vezes a sua transposta é igual a matriz identidade vezes a variância dos resíduos (6). Vale ressaltar que isso não implica que as observações não são autocorrelacionadas.

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}'|\mathbf{X}] = \begin{bmatrix} E[\varepsilon_1\varepsilon_1|\mathbf{X}] & E[\varepsilon_1\varepsilon_2|\mathbf{X}] & \cdots & E[\varepsilon_1\varepsilon_n|\mathbf{X}] \\ E[\varepsilon_2\varepsilon_1|\mathbf{X}] & E[\varepsilon_2\varepsilon_2|\mathbf{X}] & \cdots & E[\varepsilon_2\varepsilon_n|\mathbf{X}] \\ \vdots & \vdots & \ddots & \vdots \\ E[\varepsilon_n\varepsilon_1|\mathbf{X}] & E[\varepsilon_n\varepsilon_2|\mathbf{X}] & \cdots & E[\varepsilon_n\varepsilon_n|\mathbf{X}] \end{bmatrix} = \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 \\ 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma^2 \end{bmatrix} \quad (6)$$

2.5. Processo Gerador dos dados para a regressão

A quinta premissa se refere a não aleatoriedade do vetor de variáveis explicativas, em outras palavras, ele é não estocástico. Isso quer dizer que o vetor de variáveis explicativas é gerado exogenamente. No entanto, usualmente isso é de difícil aplicação, haja vista que o vetor \mathbf{X} tende a ser aleatório, tal qual o vetor \mathbf{Y} . Desse modo, uma forma alternativa é assumir \mathbf{X} como um vetor aleatório e tratar da distribuição conjunta de \mathbf{X} e \mathbf{Y} . Desse modo, essa premissa firma que \mathbf{X} pode ser fixo ou aleatório.

2.6. Normalidade dos erros

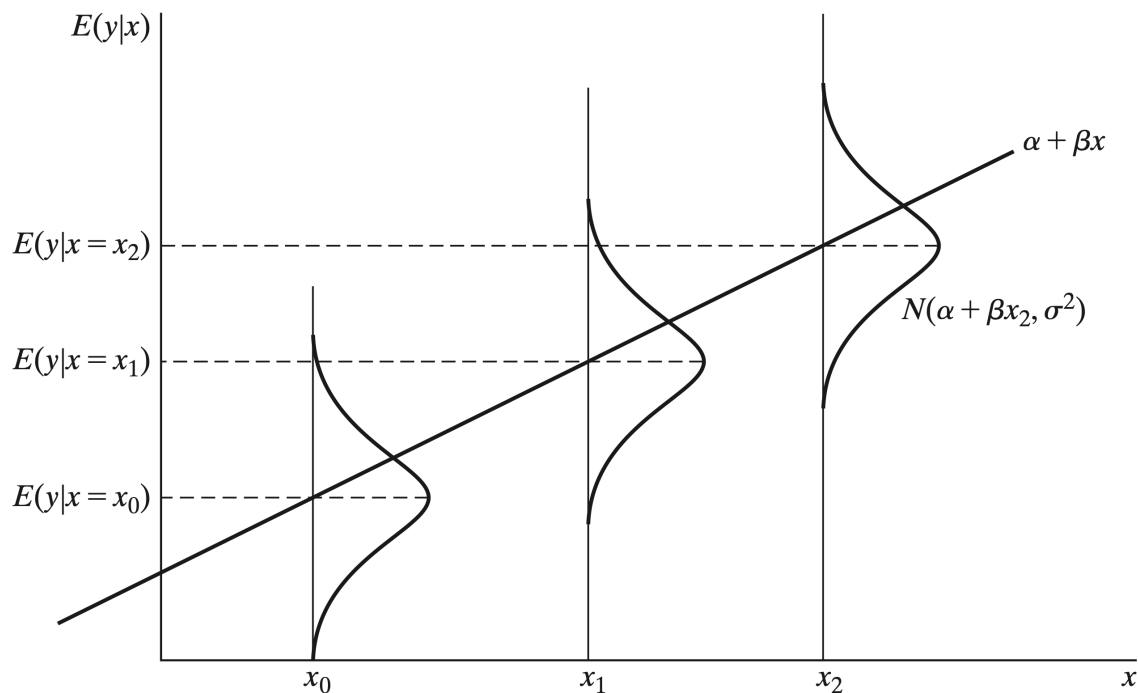
Implica que os termos de perturbação são normalmente distribuídos, possuindo média zero e variância constante (7). Essa premissa é bastante razoável, haja vista que o teorema do limite central garante essa normalidade, pelo menos, assintoticamente. Todavia, essa suposição geralmente não é necessária para obter a maioria dos resultados em uma regressão linear.

Finalizada esta parte, apresentou-se as premissas do MCRL, as quais servem como base para a construção de um modelo econométrico. O objetivo seguinte será descrever métodos de estimação de modelos, dentre eles, o famoso e amplamente utilizado, método de mínimos quadrados ordinários.

$$\varepsilon|\mathbf{X} \sim N[\mathbf{0}, \sigma^2\mathbf{I}] \quad (7)$$

A figura 1 representa bem o Modelo Clássico de Regressão Linear, com os pressupostos definidos acima:

Figura 1: O Modelo Clássico de Regressão Linear



Fonte: Greene (2019)

3. REGRESSÃO POR MÍNIMOS QUADRADOS

O método de mínimos quadrados ordinários consiste em minimizar a soma do quadrado dos resíduos, a fim de encontrar os parâmetros do modelo. O primeiro passo é distinguir entre as quantidades populacionais não observadas e os parâmetros amostrais. Em outras palavras,

existem os parâmetros verdadeiros e os parâmetros calculados no modelo agem como uma estimativa desses parâmetros populacionais, desde que sejam satisfeitas as condições que tornem o MQO aplicável. Em termos matriciais, haja vista que estamos tratando de uma regressão linear múltipla, podemos escrever o modelo da seguinte forma:

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon \quad (8)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix}_{n \times k} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{k \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1} \quad (9)$$

$$CPO : \quad \frac{\partial S(\beta)}{\partial \beta} = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\beta = \mathbf{0} \quad (10)$$

$$\Rightarrow \mathbf{X}'\mathbf{X}\beta = \mathbf{X}'\mathbf{Y} \quad (11)$$

$$\Rightarrow \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \quad (12)$$

Essa estimação nada mais é que as condições de primeira ordem do modelo. Nesse sentido, a partir dos valores estimados após encontrar os parâmetros amostrais, existem algumas relações importantes, sobretudo entre o termo de erro e os valores preditos da variável dependente: i) o MQO garante que a média dos resíduos é zero; ii) como não há covariância amostral entre o termo de erro e as variáveis independentes, não há covariância amostral entre os valores estimados e os resíduos; iii) os pontos médios das variáveis estão sempre sobre a reta de regressão.

Além disso, sob a hipótese de homocedasticidade, tratada anteriormente no MCRL, existe um teorema, conhecido como teorema de Gauss-Markov, o qual garante que os estimadores de mínimos quadrados são os melhores estimadores não viesados da classe dos lineares. Todavia, esse teorema é muito restrito, haja vista as limitações impostas, como homocedasticidade e exogeneidade estrita, haja vista a ausência de vies. Nesse sentido, é mais vantajoso analisar as propriedades assintóticas dos estimadores, que tratam de convergência em probabilidade e flexibilizam mais o modelo estimado.

Conforme Wooldrige (2010), para um estimador ser consistente, são necessárias duas premissas. A primeira implica que a covariância entre o resíduo e o vetor de variáveis explicativas seja igual a zero e essa é uma versão mais fraca da exogeneidade (3). Por outro lado, a segunda premissa diz que a multiplicação matricial de \mathbf{X} e sua transposta tem que ser

igual a ordem de \mathbf{X} (3), implicando independência linear.

$$E[\mathbf{X}'\boldsymbol{\varepsilon}] = \mathbf{0} \quad (13)$$

$$E[\mathbf{X}'\mathbf{X}] = k \quad (14)$$

3.1. Ajuste do modelo

Se chamarmos de $\hat{\beta}$ o vetor de parâmetros estimados por MQO, podemos obter o vetor $\hat{\mathbf{Y}}$ de valores estimados de \mathbf{Y} . Todos esses valores estimados estarão necessariamente sobre a reta de regressão do MQO. A diferença entre $\hat{\mathbf{Y}}$ e \mathbf{Y} será justamente o vetor de resíduos $\hat{\boldsymbol{\varepsilon}}$. Ou seja, $\mathbf{Y} - \hat{\mathbf{Y}} = \hat{\boldsymbol{\varepsilon}}$. Logo, se pensarmos em cada observação i , $\hat{\varepsilon}_i$ é a diferença entre y_i e \hat{y}_i . Se $\hat{\varepsilon}_i > 0$, a reta subestima y_i ; se $\hat{\varepsilon}_i < 0$ a reta superestima y_i ; se $\hat{\varepsilon}_i = 0$, a reta passa exatamente sobre y_i .

Pela propriedade da equação (3), temos que a média dos resíduos é zero. De forma equivalente, a média dos valores estimados é a mesma média amostral dos valores observados, $\bar{\hat{y}} = \bar{y}$. Dada essa característica, podemos definir medidas de perturbação em relação à média amostral \hat{y} :

$$SQT = \sum_{i=1}^n (y_i - \bar{y})^2 \quad (15)$$

$$SQE = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (16)$$

$$SQR = \sum_{i=1}^n \hat{\varepsilon}_i^2 \quad (17)$$

SQT (Soma dos quadrados totais) mede a dispersão das i observações (y_i), tendo a média amostral como centro. SQE (Soma dos quadrados estimados) mede a dispersão das i estimativas de y (\hat{y}_i). SQR (Soma dos quadrados dos resíduos) mede a variação dos erros estimados ($\hat{\varepsilon}_i$). A variação total em y pode ser sempre expressada como a soma da variação explicada e da variação não explicada (ou dos resíduos):

$$SQT = SQE + SQR \quad (18)$$

3.1.1. Grau de Ajuste

As definições acima nos ajudarão a definir uma medida de ajuste do modelo, ou seja, dizer quão bem nossas variáveis independentes (\mathbf{X}) explicam as observações da variável dependente

(Y).

Se assumirmos que $SQT \neq 0$, podemos definir uma medida conhecida como R^2 (coeficiente de determinação), que basicamente é uma proporção entre a variação explicada (SQE) e a variação total (SQT). Em outras palavras, é uma medida que diz quanto da variação total em y foi explicada pelas variáveis x .

$$R^2 = \frac{SQE}{SQT} = 1 - \frac{SQR}{SQT} \quad (19)$$

$$R^2 \in [0, 1]$$

Como $SQE \leq SQT$, R^2 sempre estará entre 0 e 1. Graficamente falando, quanto mais próximo de 1, mais próximas as observações estarão da reta de regressão.

4. APLICAÇÃO

Para aplicar as questões tratadas nas últimas seções, vamos estimar um modelo linear por MQO, utilizando a linguagem R e os dados da REGIC 2018 para municípios de Minas Gerais.

Carregando pacotes:

```
library(readxl)
library(dplyr)
library(ggplot2)
library(stargazer)
library(geobr)
library(ggspatial)
library(ggrepel)
library(scatterplot3d)
```

Carregando a base da REGIC 2018 e ajustando as variáveis de interesse:

```
df_regic <- readxl::read_xlsx(
  path = "data/REGIC2018 Cidades v2.xlsx",
  sheet="Base de dados por Cidades"
) |>
dplyr::filter(
  UF == "MG"
) |>
dplyr::select(
  "COD_CIDADE",
  "NOME_CIDADE",
  "VAR01",
```

```

    "VAR03",
    "VAR23",
    "VAR29",
    "VAR85",
    "VAR89"
  ) |>
  dplyr::rename(
    "populacao" = "VAR01",
    "pib" = "VAR03",
    "cige" = "VAR23",
    "cgp" = "VAR29"
  ) |>
  dplyr::mutate(
    "populacao" = as.numeric(populacao),
    "pib" = as.numeric(pib),
    "cige" = as.numeric(cige),
    "cgp" = as.numeric(cgp),
    "banco_publico" = ifelse(VAR85 | VAR89, 1, 0),
    "log_cige" = ifelse(as.numeric(cige) < 1, 0, log(as.numeric(cige))),
    "log_cgp" = ifelse(as.numeric(cgp) < 1, 0, log(as.numeric(cgp)))
  )

df_regic <- na.omit(df_regic)

df_regic$pib_pc <- df_regic$pib / df_regic$populacao

```

Foi necessário excluir os registros que tinham alguma variável de interesse NULA (NA). Optou-se por não substituir os *NA* por valores arbitrários, como zero ou qualquer outro valor, pois causaria um grande viés na regressão, visto que se tratam municípios que não tiveram essas variáveis mensuradas.

4.1. Análise Descritiva

Nesta seção, uma análise descritiva abrangente das variáveis em questão será conduzida, visando fornecer uma compreensão profunda de seus padrões, distribuições e relações. Esta abordagem permite não apenas a caracterização detalhada de cada variável individualmente, mas também a identificação de tendências e padrões globais dentro do conjunto de dados.

A Tabela Tabela 1 contém as principais estatísticas descritivas das variáveis analisadas (PIB per capita, Centralidade da Gestão Pública (CGP) e Coeficiente de Intensidade da Gestão Empresarial (CIGE)) para as 168 observações (municípios) restantes na base:

Abaixo, plotamos os histogramas dessas variáveis:

Podemos ver a correlação entre as variáveis por meio de um mapa de calor obtido da matriz de correlação:

Tabela 1: Estatísticas descritivas das variáveis

| Variável | Min. | 1st Qu. | Median | Mean | 3rd Qu. | Max. |
|----------|------|---------|--------|--------|---------|--------|
| CIGE | 36 | 93.50 | 147.50 | 355.32 | 306.75 | 15,174 |
| CGP | 1 | 2 | 2.50 | 4.12 | 6 | 50 |
| PIBpc | 8.53 | 16.83 | 21.11 | 24.09 | 27.57 | 174.21 |

Fonte: Elaboração própria.

Figura 2: Histograma do PIB per capita dos municípios de Minas Gerais (2018)

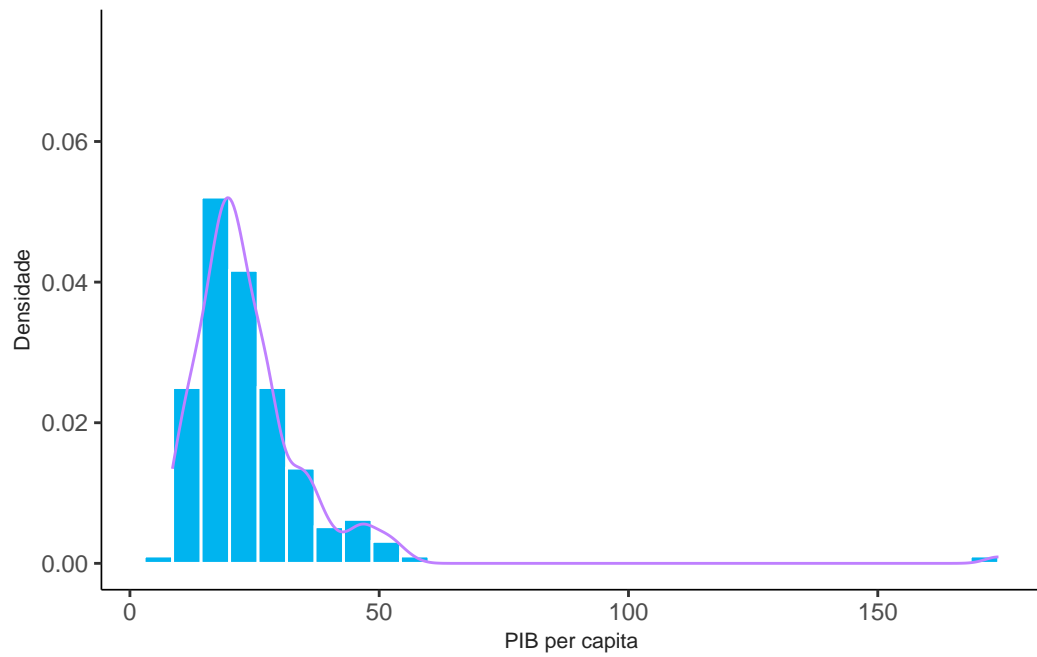


Figura 3: Histograma do Coeficiente de Intensidade da Gestão Empresarial (CIGE) dos municípios de Minas Gerais (2018)

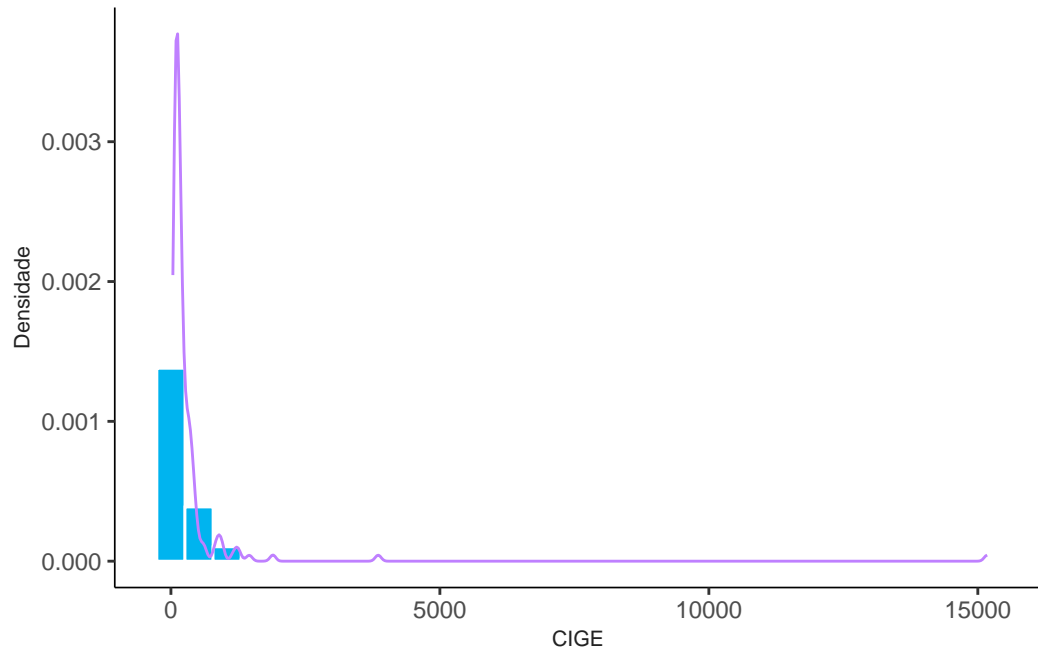
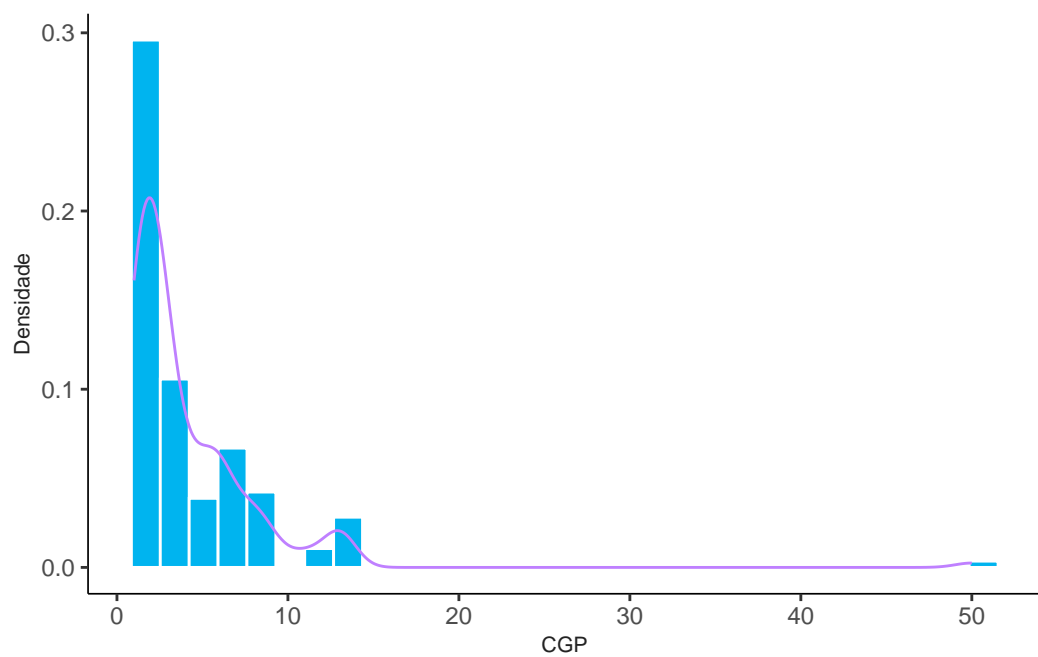


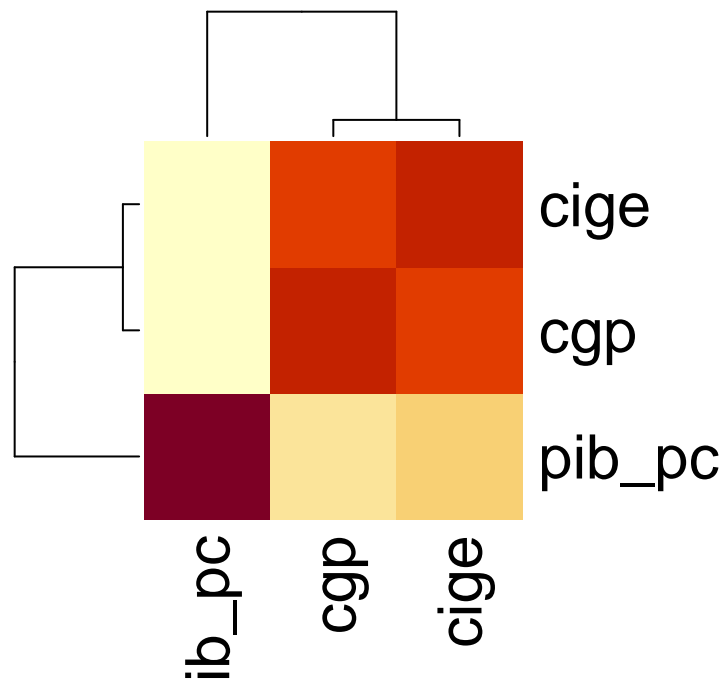
Figura 4: Histograma da Centralidade da Gestão Pública (CGP) dos municípios de Minas Gerais (2018)



```
# Matriz com as variáveis [Y/X]
matrix <- df_regic |>
  dplyr::select("pib_pc", "cige", "cgp") |>
  as.matrix.data.frame()

matrix |> cor() |> heatmap()
```

Figura 5: Mapa de calor da matriz de correlação.



Outra medida importante é a de covariância entre as variáveis, descrita na matriz de covariância abaixo (variância na diagonal principal):

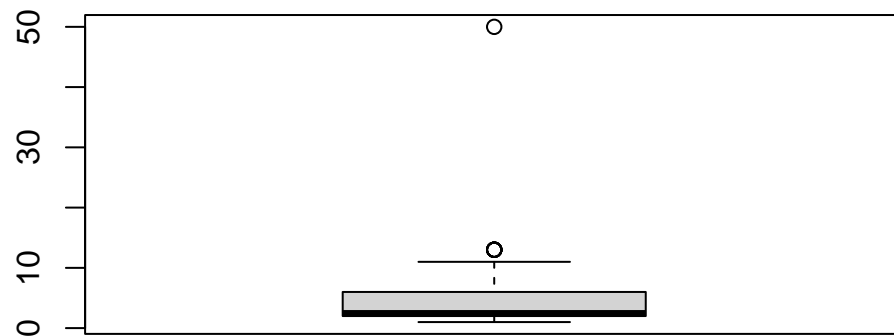
```
matrix |> cov() |> round(2)
```

| | pib_pc | cige | cgp |
|--------|---------|------------|---------|
| pib_pc | 230.22 | 2298.78 | 8.27 |
| cige | 2298.78 | 1474769.11 | 4994.06 |
| cgp | 8.27 | 4994.06 | 22.55 |

Abaixo plotamos os *boxplots* das variáveis utilizadas:

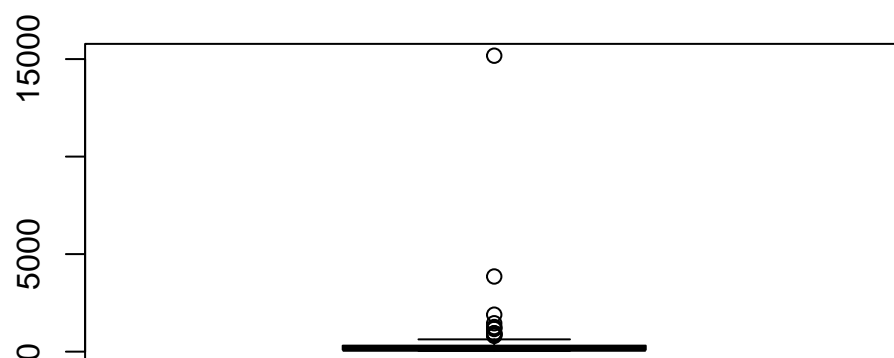
```
boxplot(df_regic$cgp)
```

Figura 6: Boxplot da Centralidade da Gestão Pública (CGP) dos municípios de Minas Gerais (2018)



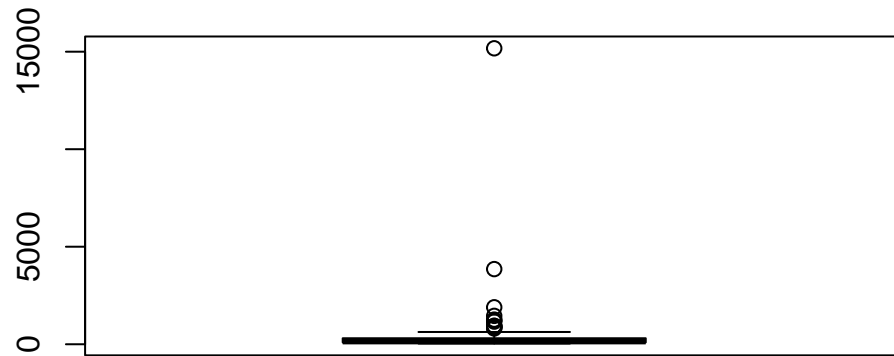
```
boxplot(df_regic$cige)
```

Figura 7: Boxplot do Coeficiente de Intensidade da Gestão Empresarial (CIGE) dos municípios de Minas Gerais (2018)



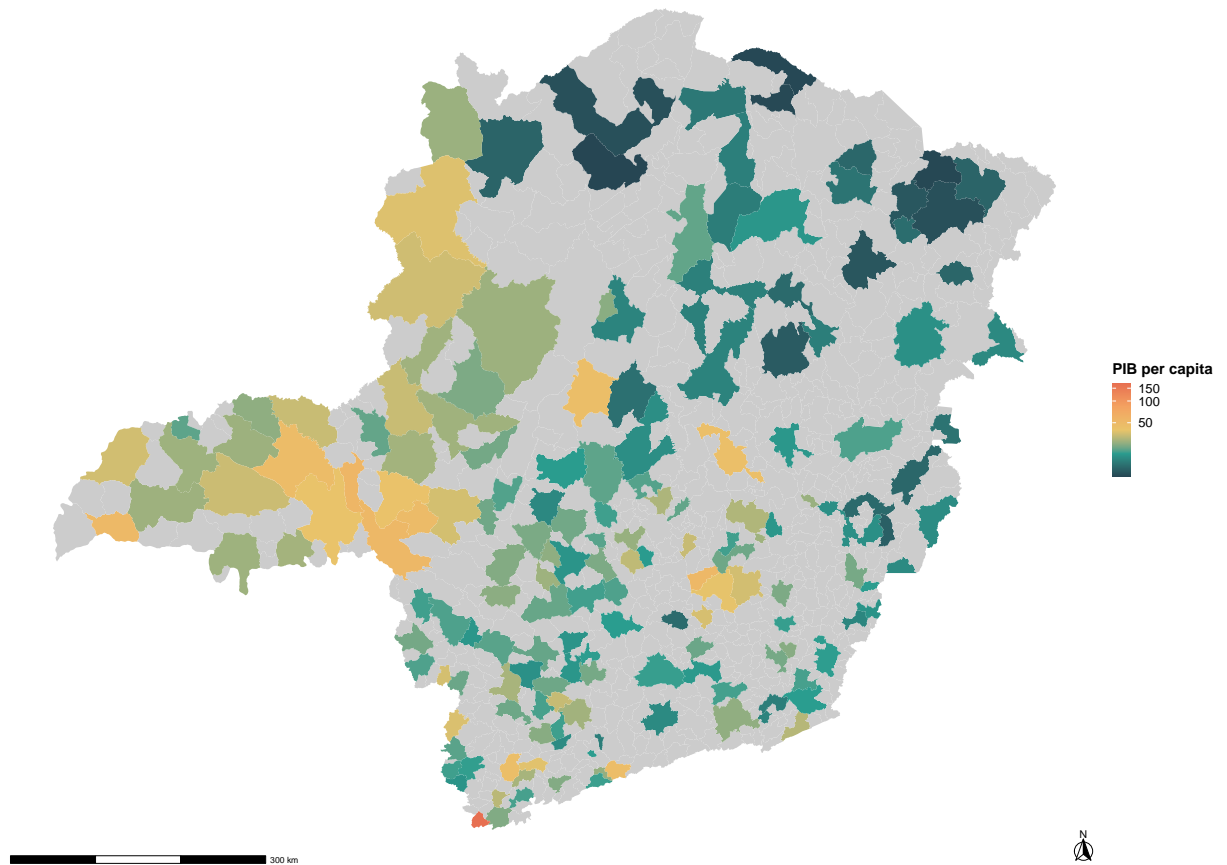
```
boxplot(df_regic$cige)
```

Figura 8: Boxplot do PIB per capita dos municípios de Minas Gerais (2018)



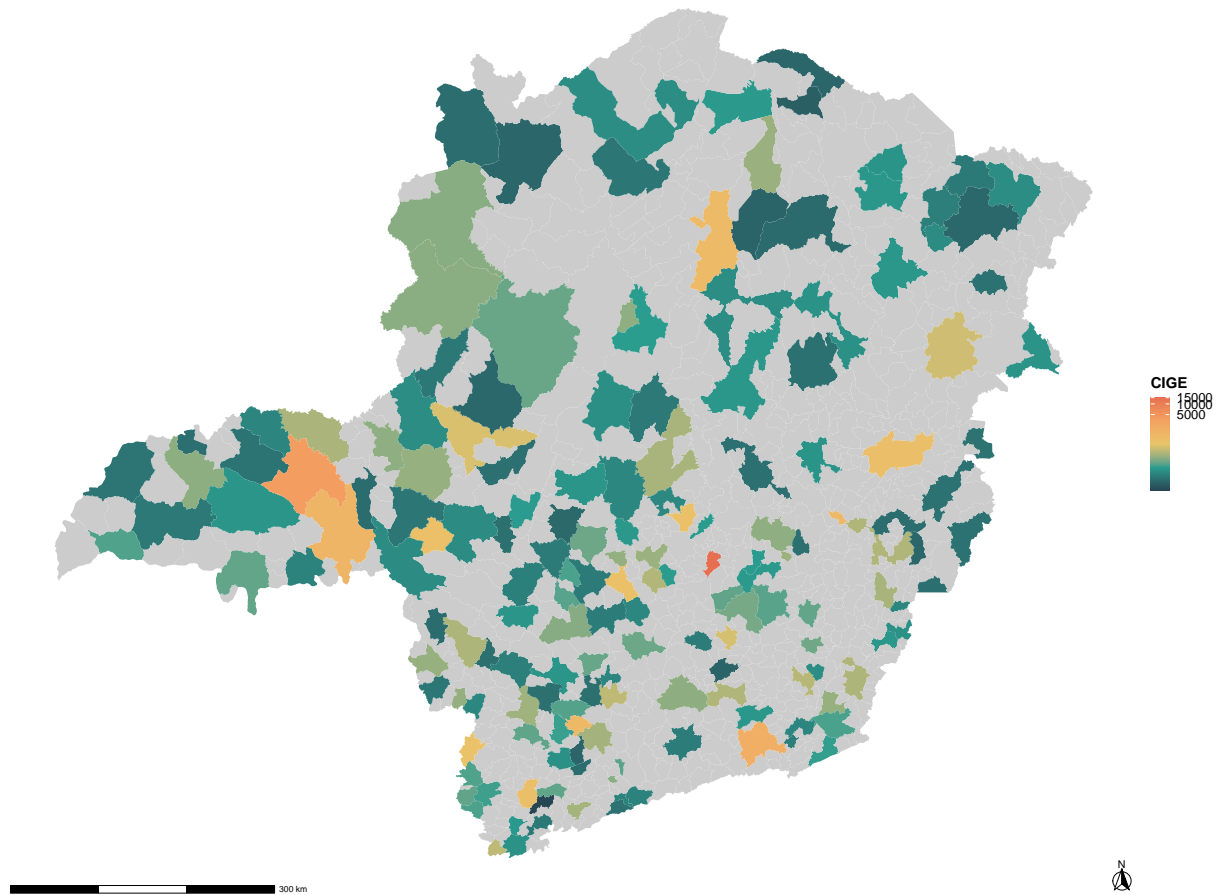
A visualização geográfica das variáveis do modelo (para o caso de municípios) é essencial para compreender, com clareza, onde os processos estão acontecendo:

Figura 9: PIB per capita dos municípios de Minas Gerais - REGIC 2018



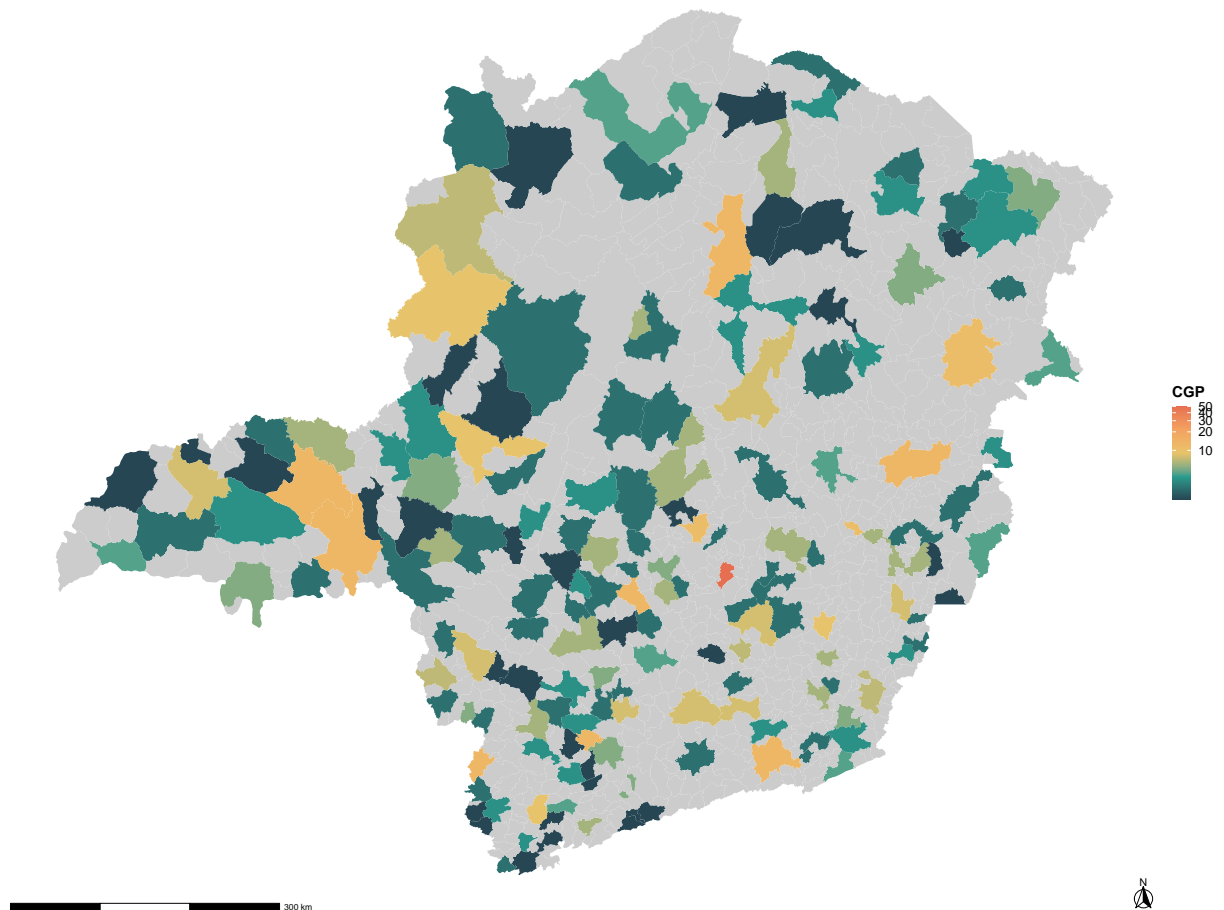
Fonte: Elaboração própria.

Figura 10: CIGE dos municípios de Minas Gerais - REGIC 2018



Fonte: Elaboração própria.

Figura 11: CGP dos municípios de Minas Gerais - REGIC 2018



Fonte: Elaboração própria.

Por fim, podemos plotar alguns gráficos de dispersão para esboçar a relação entre as variáveis do modelo:

4.2. Análise de Regressão

Com as mesmas operações definidas na seção sobre Regressão Por Mínimos Quadrados, vamos estimar um modelo linear que tenta explicar o $\ln(PIBpc)$ (logaritmo natural do PIB per capita) de municípios mineiros, com variáveis que medem intensidade de gestão empresarial (CIGE) e nível de centralidade da gestão pública (CGP).

Figura 12: PIB per capita x CIGE (Municípios de Minas Gerais - REGIC 2018)

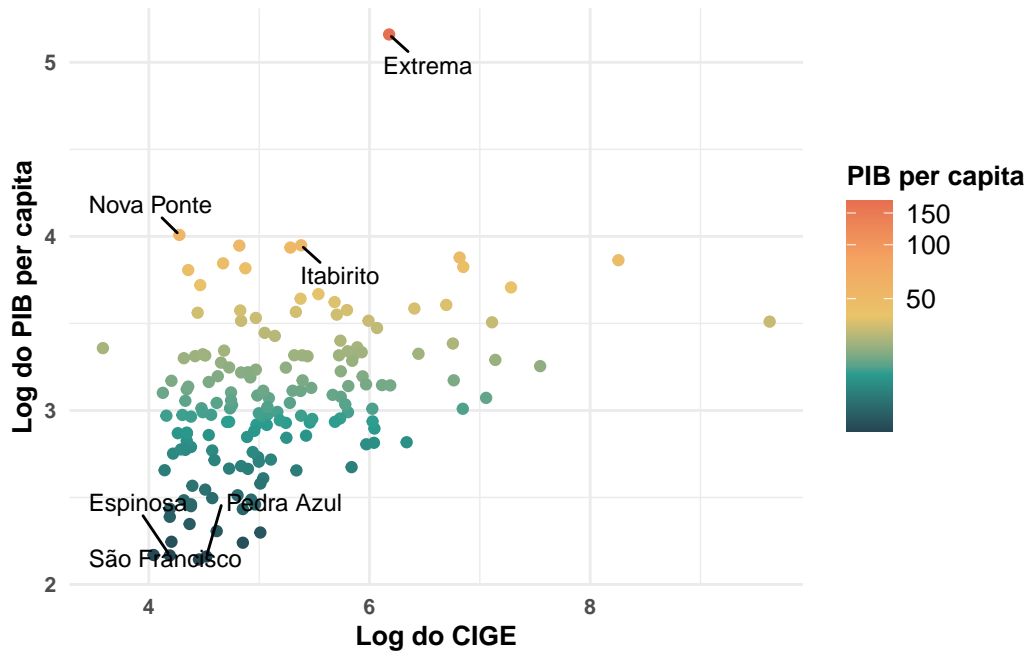
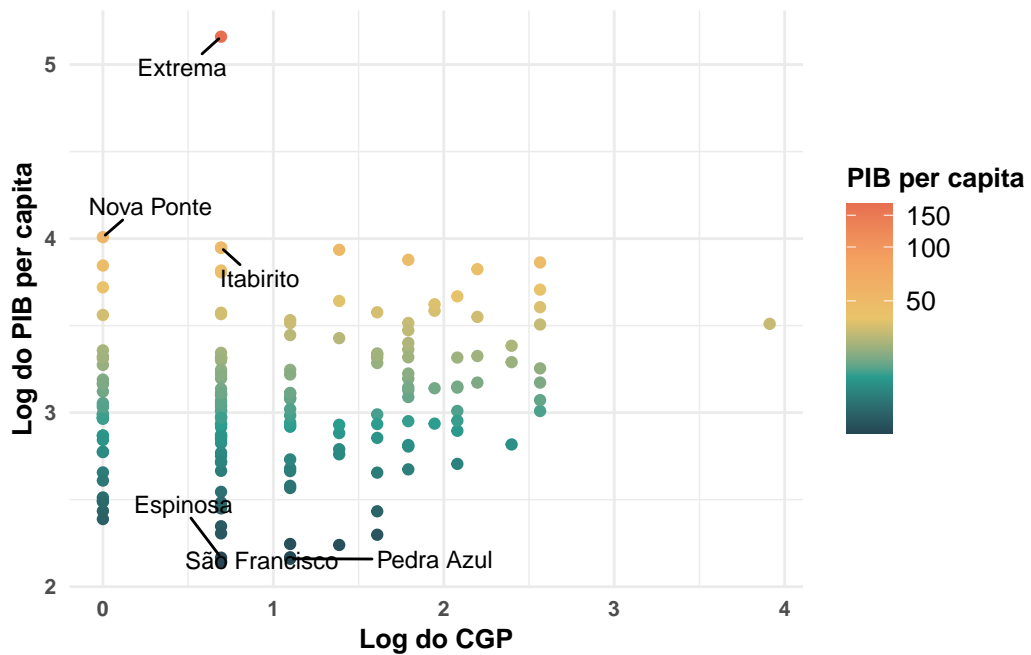


Figura 13: PIB per capita x CGP (Municípios de Minas Gerais - REGIC 2018)



```
# Regressão simples
covxy <- cov(df_regic$log_cige, log(df_regic$pib_pc))
varx <- var(df_regic$log_cige)
mediay <- mean(log(df_regic$pib_pc))
mediax <- mean(df_regic$log_cige)
b1 <- covxy/varx

b0 <- mediay - b1*mediax

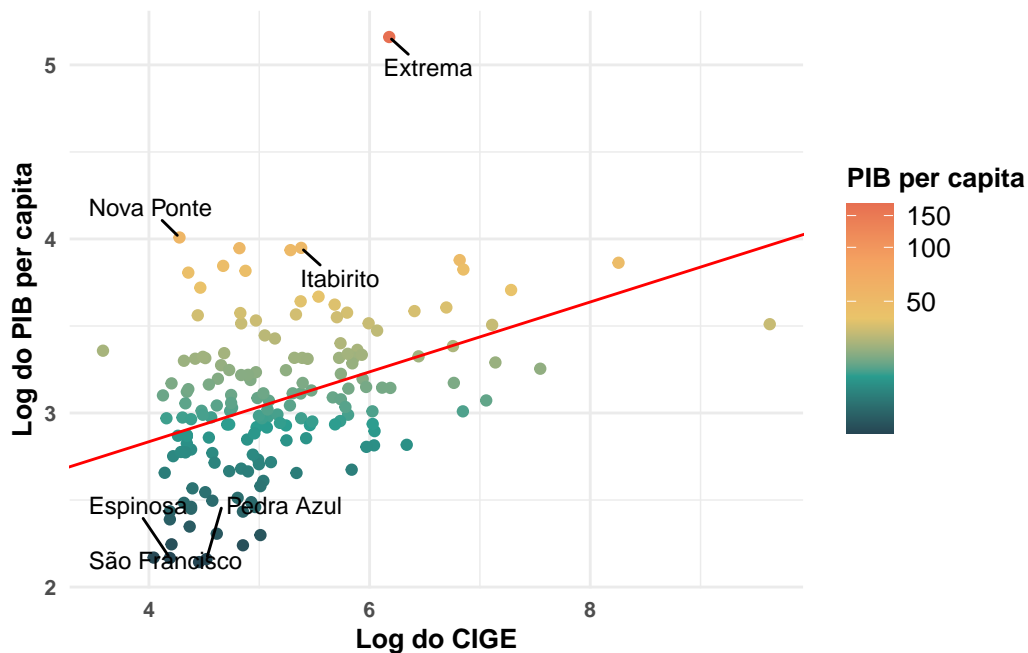
print(paste0("Intercepto: ", round(b0, 2)))
```

```
[1] "Intercepto: 2.03"
```

```
print(paste0("Coeficiente estimado: ", round(b1, 2)))
```

```
[1] "Coeficiente estimado: 0.2"
```

Com esses coeficientes já é possível traçar uma reta de regressão no gráfico:



Para verificar o ajuste do modelo, podemos calcular o R^2 , dado que $R^2 = Var(\hat{y})/Var(y)$:

```
# Cálculo do R-quadrado
y <- log(df_regic$pib_pc)
x <- log(df_regic$cige)
y.hat <- b0 + b1*x
r2 <- var(y.hat)/var(y)
r2
```

```
[1] 0.1609252
```

Esse resultado indica que cerca de 16,10% da variação do log do PIB per capita dos

municípios da amostra é explicado pelo log do Coeficiente de Interação da Gestão Empresarial (CIGE) desses municípios.

4.3. Análise de Regressão Múltipla

Vamos acrescentar mais uma variável no modelo anterior e estimar novamente os parâmetros por meio de álgebra matricial:

```
# Número de observações da amostra
n <- nrow(df_regic)

# Número de variáveis independentes
k <- 2

# Matriz de variáveis independentes
X <- matrix(1, nrow = n, ncol = 3) # Coluna de 1 (intercepto)
X[,2] <- log(df_regic$cige) # CIGE
X[,3] <- log(df_regic$cgp) # CGP

# Vetor de variável observada
Y <- as.matrix(log(df_regic$pib_pc))

# Vetor de parâmetros estimados
bhat <- solve(t(X) %*% X) %*% t(X) %*% Y

# Vetor de resíduos estimados
uhat <- y - X %*% bhat

# Variância dos resíduos
sigma2hat <- as.numeric (t(uhat) %*% uhat / (nrow(df_regic)-k-1))

# Matriz var-cov dos coeficientes
varbetahat <- sigma2hat * solve(t(X) %*% X)
erropadraobeta <- sqrt (diag(varbetahat))

bhat
```

```
[,1]
```

```
[1,] 1.3763741
```

```
[2,] 0.3748468
```

```
[3,] -0.2316740
```

```
sigma2hat
```

```
[1] 0.1515339
```

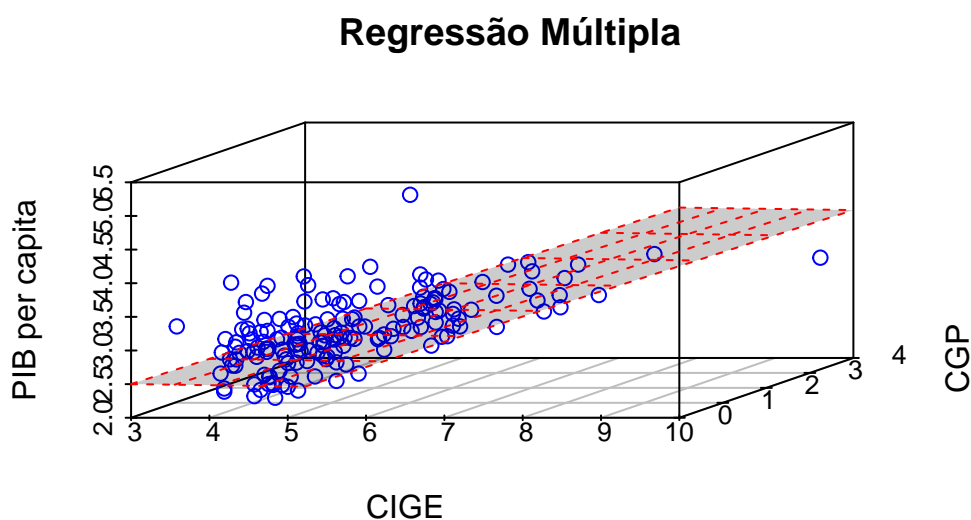
```
varbetahat
```

```
      [,1]      [,2]      [,3]
[1,] 0.07041639 -0.01612038 0.013226893
[2,] -0.01612038 0.003830140 -0.003512181
[3,] 0.01322689 -0.003512181 0.004671999
```

```
erropadraobeta
```

```
[1] 0.26536086 0.06188812 0.06835202
```

Com apenas 2 variáveis independentes ainda é possível ver graficamente essa regressão, e agora, ao invés de uma reta de regressão, temos um plano de regressão:



Como fizemos para a regressão simples, podemos calcular o ajuste R^2 para o modelo de regressão múltipla:

```
r2 <- var(y.hat)/var(y)
r2
```

```
      [,1]
[1,] 0.2155433
```

Esse resultado indica que cerca de 21,55% da variação do log do PIB per capita dos municípios da amostra é explicada pela combinação linear entre o log do Coeficiente de Interação da Gestão Empresarial(CIGE) e o log do nível de Centralidade da Gestão Pública (CGP) desses municípios.

5. REFERÊNCIAS

GREENE, WILLIAM. H. **Econometric Analysis Global Edition**. 8. ed. [s.l.] Pearson-prentice Hall, 2019.

HANSEN, B. **Econometric**. 1. ed. [s.l.] Princeton University Press, 2022.

HEISS, F. **Using R for Introductory Econometrics**. 2. ed. [s.l.: s.n.].

WOOLDRIDGE, J. M. **Econometric Analysis of Cross-Section and Panel Data**. 2. ed. [s.l.] MIT Press, 2010.