



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)	19
3	Histograma do Coeficiente de Intensidade da Gestão Empresarial (CIGE) dos municípios de Minas Gerais (2018)	19
4	Histograma da Centralidade da Gestão Pública (CGP) dos municípios de Minas Gerais (2018)	20
5	Mapa de calor da matriz de correlação.	20
6	Boxplot da Centralidade da Gestão Pública (CGP) dos municípios de Minas Gerais (2018)	21
7	Boxplot do Coeficiente de Intensidade da Gestão Empresarial (CIGE) dos municípios de Minas Gerais (2018)	22
8	Boxplot do PIB per capita dos municípios de Minas Gerais (2018)	22
9	PIB per capita dos municípios de Minas Gerais - REGIC 2018	23
10	CIGE dos municípios de Minas Gerais - REGIC 2018	24
11	CGP dos municípios de Minas Gerais - REGIC 2018	25
12	PIB per capita x CIGE (Municípios de Minas Gerais - REGIC 2018)	26
13	PIB per capita x CGP (Municípios de Minas Gerais - REGIC 2018)	26
14	PIB per capita x Valor adicionado da Administração Pública (Municípios de Minas Gerais - REGIC 2018)	34
15	PIB per capita x Índice de Atração Geral (Municípios de Minas Gerais - REGIC 2018)	35
16	PIB x Índice de Atração Geral (Municípios de Minas Gerais - REGIC 2018)	35
17	PIB per capita x Distância ao quadrado da agência (banco público) mais próxima (Municípios de Minas Gerais - REGIC 2018)	36
18	PIB per capita x Distância ao quadrado da agência (banco privado) mais próxima (Municípios de Minas Gerais - REGIC 2018)	37
19	PIB per capita estimado x Resíduos (Modelo com dummies bancárias)	38
20	PIB per capita estimado x Resíduos (Modelo com distância)	39

SUMÁRIO

1	INTRODUÇÃO	1
2	O MODELO CLÁSSICO DE REGRESSÃO LINEAR	1
3	REGRESSÃO POR MÍNIMOS QUADRADOS	4
4	Tópicos adicionais	7
5	APLICAÇÃO	16
5.1	Análise Descritiva	18
5.2	Análise de Regressão	25
5.3	Análise de Regressão Múltipla	28
5.4	Análise de Regressão com diferentes formas funcionais	30
5.5	Testes	36
6	REFERÊNCIAS	41

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:

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.

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

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.

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.

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

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

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.

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

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.

Normalidade dos erros

Implica que os termos de perturbação são normalmente distribuídos, possuindo média

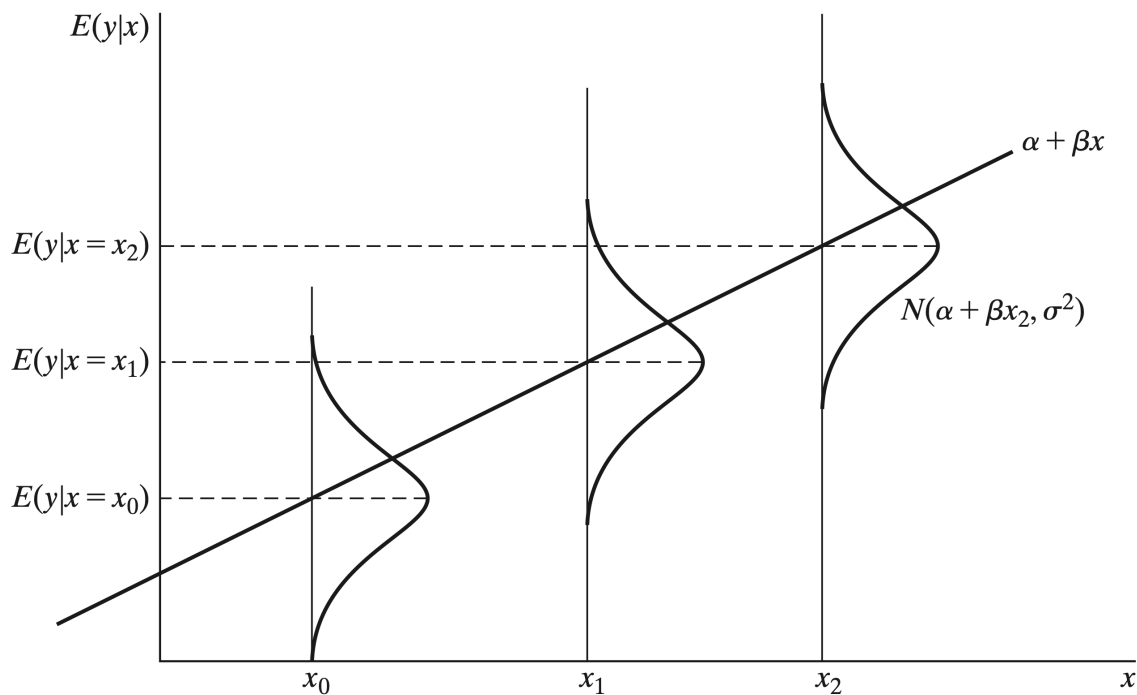
zero e variância constante (7). Essa premissa é bastante razoável, haja vista que o teorema do limite central garante essa normalidade, pelo menos, assintoticamente. Todavia, essa suposição geralmente não é necessária para obter a maioria dos resultados em uma regressão linear.

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.

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

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

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):

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

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 (\mathbf{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. Gráficamente falando, quanto mais próximo de 1, mais próximas as observações estarão da reta de regressão.

4. Tópicos adicionais

Multicolinearidade

A multicolinearidade ou colinearidade é definida como uma correlação entre variáveis dispostas em um modelo. Em síntese, podemos separá-la em dois casos: multicolinearidade exata e multicolinearidade prejudicial.

A multicolinearidade exata, também conhecida como multicolinearidade perfeita, é quando as variáveis são exatamente correlacionadas entre si, ou seja, a correlação é de 100%. Nesse caso específico, é ferido um dos pressupostos do MCRL e esta falha é tão grave que, nesse caso, não é possível estimar os parâmetros dos modelos, uma vez que a matriz de variáveis independentes é singular. Destaca-se, no entanto, que esta correlação exata só é problemática quando é linear, ou seja, uma variável pode ter uma correlação exata quadrática, e.g., que não causará danos aos pressupostos do MCRL.

Felizmente, casos de multicolinearidade perfeita, na prática, são raros, mas a multicolinearidade prejudicial, com frequência, acontece. Nesse caso, por sua vez, a correlação não é exata, mas alta o suficiente para gerar problemas. Não há falha nos pressupostos do modelo, mas pode ser que surjam problemas estatísticos potencialmente graves, sobretudo no que diz respeito à eficiência do modelo. Algumas problemáticas que podemos citar são: grandes oscilações nas estimativas dos parâmetros; os erros padrão dos coeficientes podem ser muito altos, fazendo com que as estatísticas t , por exemplo, não sejam confiáveis, dado os baixos níveis de significância; pode ser que haja alteração no sinal do parâmetro ou magnitudes irreais.

Como a variância de um estimador β_k qualquer é dada por:

$$Var[\beta_k|X] = \frac{\sigma^2}{(1 - R_k^2) \sum_{i=1}^n (x_{ik} - \bar{x}_k)^2}$$

Onde R_k^2 é o R^2 na regressão de X_k em relação a todas as outras variáveis. Assim, de X_k é

altamente correlacionada com as variáveis, R_k^2 tende a 1, de modo que a variância de b_k tende a infinito. Caso R_k^2 seja igual a zero, então a condição de ortogonalidade é satisfeita e não há multicolinearidade no modelo. Assim, uma forma de detectar a multicolinearidade é analisar o fator $1/(1 - R_k^2)$, conhecido como fator de inflacionamento da variância (VIF), pois, se ele for muito alto, geralmente, valores superiores a 20, há multicolinearidade prejudicial, que trará os problemas estatísticos mencionados acima.

Alguns métodos para tentar solucionar este problema são: i) obter mais dados, ou seja, aumentar o tamanho da amostra, haja vista que isso aumentará a variabilidade de X ; ii) transformar as variáveis, como, por exemplo, utilizar uma variável instrumental; iii) retirar do modelo as variáveis suspeitas de causar o problema. Este terceiro método é o mais óbvio e, com frequência, o mais utilizado, todavia, pode ser problemático, uma vez que, se essas variáveis retiradas pertencem ao modelo populacional, suas informações irão para o termo de perturbação, causando um viés que será explicado posteriormente, conhecido como viés de variável omitida.

Inferência

São três as principais funções do modelo de regressão linear. Em síntese, ele pode ser usado para estimação, testes de hipótese e previsão. O objetivo desta parte do trabalho, no entanto, é focar em algumas aplicações dos testes de hipóteses, ressaltando a importância da inferência estatística. Uma abordagem comumente vista para testar hipóteses são modelos cujos parâmetros possuem restrições. Em modelos restritos, por imposição teórica, por exemplo, temos que dentro do modelo irrestrito, apenas alguns serão válidos, isto é, apenas os modelos cuja restrição seja satisfeita. Nesse sentido, dizemos que os modelos são aninhados, ou seja, os restritos estão contidos dentro do irrestrito.

Os testes de hipóteses podem ser abordados a partir de dois pontos de vista. Em primeiro lugar, tendo calculado um conjunto de estimativas de parâmetros, é possível verificar se o fracasso das estimativas em satisfazer as restrições é apenas um erro de amostragem ou é algo sistemático. Por outro lado, uma outra abordagem seria considerar o modelo de mínimos quadrados irrestritos, com restrições tacitamente teóricas, de modo que, tal fato poderia levar a uma perda de ajuste do modelo, visto que ele não é adequado. Desta forma, é possível verificar se essa perda de ajuste é, novamente, um erro de amostragem ou se é tão relevante que levantaria dúvidas sobre a validade das restrições e, por consequência, do modelo em si. Apesar de serem consideradas separadamente, essas suas abordagens são equivalentes e servem para o mesmo propósito. Um importante e central ponto é que, para realizar as inferências, serão assumidos que os erros são normalmente distribuídos.

O teste F

Em primeiro lugar, é necessário definir uma hipótese nula e uma hipótese alternativa, a fim de poder fazer conclusões sobre o teste. Essas hipóteses são comumente tratadas como H_0

e H_1 , respectivamente. O objetivo do teste, portanto, é aceitar ou não (esse caso, na maioria das vezes, implica aceitar a hipótese alternativa) a hipótese nula. Um método bem conhecido de testar H_0 é o chamado Critério de Wald:

$$\begin{aligned} W &= \mathbf{m}' Var[\mathbf{m}|\mathbf{X}]^{-1} \mathbf{m} \\ &= (\mathbf{Rb} - \mathbf{q})' [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{Rb} - \mathbf{q}) \\ &= \frac{(\mathbf{Rb} - \mathbf{q})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{Rb} - \mathbf{q})}{\sigma^2} \sim \chi^2[J] \end{aligned}$$

Essa estatística consiste em uma distribuição χ^2 com J graus de liberdade se a hipótese estiver correta. Sua interpretação intuitiva é de que quanto maior for o fracasso dos MQ em satisfazer as restrições, maior será o valor da estatística, i.e, maior o valor χ^2 . Desse modo, um alto valor diminuirá a probabilidade de se aceitar H_0 . Entretanto, empiricamente este teste não é aplicável, uma vez que ele trabalha com a variância populacional, desconhecida. Todavia, existe uma forma de contornar este problema, transformando a estatística W em um teste F . Para isso, é necessário dividir W por seus graus de liberdade e utilizar a variância estimada, que é conhecida, em vez da variância populacional:

$$\begin{aligned} F &= \frac{W}{J} \frac{\sigma^2}{s^2} \\ &= \left(\frac{(\mathbf{Rb} - \mathbf{q})' [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{Rb} - \mathbf{q})}{\sigma^2} \right) \left(\frac{1}{J} \right) \left(\frac{\sigma^2}{s^2} \right) \left(\frac{n - K}{n - K} \right) \\ &= \frac{(\mathbf{Rb} - \mathbf{q})' [\sigma^2 \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} \mathbf{R}']^{-1} (\mathbf{Rb} - \mathbf{q}) / J}{[(n - K) s^2 / \sigma^2] / (n - K)} \end{aligned}$$

Após as devidas transformações, chegamos ao teste F , de fato, que é uma estatística na qual ocorre a divisão de duas qui-quadrado, independentes, e ambas divididas pelos seus respectivos graus de liberdade.

$$F[J, n - K] = \frac{(\mathbf{Rb} - \mathbf{q})' \{ \mathbf{R} [s^2 (\mathbf{X}'\mathbf{X})^{-1}] \mathbf{R}' \}^{-1} (\mathbf{Rb} - \mathbf{q})}{J}$$

Ademais, deve-se salientar que, para testar uma única restrição, a estatística F se equivale ao quadrado da estatística t , dada por:

$$t^2 = \frac{(\hat{q} - q)^2}{\text{Var}(\hat{q} - q|\mathbf{X})}$$

$$= \frac{(\mathbf{r}'\mathbf{b} - q)\{\mathbf{r}'[s^2(\mathbf{X}'\mathbf{X})^{-1}]\mathbf{r}\}^{-1}(\mathbf{r}'\mathbf{b} - q)}{1}$$

O estimador de mínimos quadrados restritos

O estimador de mínimos quadrados restritos, neste caso, é igual ao estimador de mínimos quadrados irrestritos, mais um termo que explica a ineficiência da solução irrestrita em satisfazer as condições de restrição. Em outras palavras, o estimador irrestrito, caso haja imposição de restrições, é viesado e seu viés é muito parecido com viés causado por uma variável relevante omitida, como será mostrado mais a frente.

Nesse sentido, o teste F permite fazer a comparação entre os modelos, de modo a verificar qual é o mais adequado entre o restrito e o irrestrito. Essa comparação, de modo geral, consiste em analisar o R^2 dos modelos, ponderados pelos seus graus de liberdade, da forma que se segue:

$$F[J, n - K] = \frac{(R^2 - R_*^2)/J}{(1 - R^2)/(n - K)}$$

Formas funcionais e variáveis dummies

Um modelo de regressão linear pode ser adaptado para situações particulares e específicas da natureza de um fenômeno. Essa adaptação pode ser feita em sua forma funcional, que define a relação entre a variável observada Y e suas explicativas X_i em um modelo de regressão. A forma funcional adequada é a que captura com precisão a natureza da relação entre as variáveis.

A forma funcional mais básica é a linear, em que a variável explicada Y depende das demais variáveis X sem nenhuma transformação, como apresentado nas seções anteriores:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Além do caso linear, temos uma forma funcional que utiliza variável binária ou dummy. Uma variável dummy assume o valor 1 para algumas observações para indicar a presença de um efeito ou participação em um grupo e 0 para as observações restantes. As variáveis binárias são um meio conveniente de construir mudanças discretas da função em um modelo de regressão. Variáveis dummy são geralmente usadas em equações de regressão que também contêm outras variáveis quantitativas:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{D}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}$$

Expandindo a equação anterior:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}_{n \times 1} = \begin{bmatrix} 1 & x_{12} & \cdots & x_{1k} \\ 1 & x_{22} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n2} & \cdots & x_{nk} \end{bmatrix}_{n \times k} \cdot \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix}_{k \times 1} + \begin{bmatrix} 0 & \cdots & 1 \\ 1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}_{n \times \theta} \cdot \begin{bmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_\theta \end{bmatrix}_{\theta \times 1} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}_{n \times 1} \quad (20)$$

Sendo \mathbf{D} uma matriz formada por zeros (0) e uns (1) que indicam que uma dada observação possui alguma característica ou pertence a um grupo específico. Essa matriz tem dimensão $n \times \theta$, em que n é o tamanho da amostra e θ o número de dummies. O vetor coluna γ são os coeficientes de regressão associados as θ dummies.

Quando uma dummy recebe valor 1, então o valor de γ associado é incorporado na estimação. Caso contrário, a dummy será 0 e não incluirá o valor de γ .

No caso de variáveis categóricas com múltiplas categorias, podemos representá-las por dummies e omitir uma das categorias na forma funcional, para que não tenhamos multicolinearidade perfeita:

$$y = \beta_0 + \beta_1 x_1 + \gamma_1 \text{primavera} + \gamma_2 \text{outuno} + \gamma_3 \text{verao}$$

No exemplo acima, a variável categórica tenta captar o efeito da estação do ano na variável observada y . Note que não criamos uma dummy para inverno, pois essa foi omitida para ser a base de comparação, no caso em que todas assumam valor 0.

Outros exemplos de formas funcionais são as que aplicam transformações não lineares em variáveis numéricas. As mais comuns são:

Regressão quadrática, em que a uma ou mais variáveis dependentes tem uma relação quadrática com a variável explicada:

$$y = X\beta + X^2\gamma + \varepsilon$$

Regressão Log-linear, em que a variável explicada e as variáveis dependentes crescem de forma exponencial e se deseja suavizar ou linearizar a relação:

$$\ln(y) = \ln(X)\beta + \varepsilon$$

Regressão Semi-logarítima, em que apenas a variável explicada é exponencial e se deseja linearizar:

$$\ln(y) = X\beta + \varepsilon$$

Regressão de efeitos interativos, utilizada quando se deseja mensurar a interação entre duas variáveis, adicionando-se um coeficiente entre o produto dessas variáveis:

$$y = X\beta + \beta_\tau(x_\alpha \cdot x_\delta) + \varepsilon$$

β_τ mede a interação entre as variáveis dependentes na estimação de y .

Análise da qualidade de ajuste e escolha de modelos

Quando há mais de um candidato para a escolha de um modelo, deve-se seguir alguns testes com objetivo de verificar qual o mais adequado. Nesse sentido, nesta parte serão discutidos os impactos de uma má especificação, bem como os métodos estatísticos para realizar a escolha quando têm-se dois ou mais modelos.

Um dos erros mais comuns de se ver em especificação de modelos são a omissão de variáveis relevantes e a inclusão de variáveis irrelevantes. Suas implicações, contudo, são bem diversas. Quando omitimos uma variável relevante de um modelo, automaticamente geramos um viés no estimador, haja vista que os efeitos dessa variável vão para o termo de perturbação. Um problema ainda mais inquietante é quando o pesquisador se depara com um modelo com mais de uma variável explicativa e duas ou mais dessas variáveis possuem uma alta correlação entre si. Ocorrendo isso, há o dilema entre retirar uma variável que pode ser relevante ou deixá-la e viesar o estimador, haja vista a alta colinearidade. Quando tal problema ocorre, é comum a criação de um outro estimador, conhecido como estimador pré teste. Todavia, este estimador também se revela viesado e a experiência mostra que ele costuma ser o menos preciso entre o estimador com variável (eis) omitida (s) e o que considera as variáveis que gera colinearidade prejudicial.

Totalmente distinta a omissão de variáveis relevantes está a inclusão de variáveis irrelevantes, haja vista que essa não gera um viés de estimador, uma vez que o que ocorre é que, se a variável é supérflua, seu parâmetro estimado será igual a zero. Um problema empírico só ocorre do ponto de vista da variável irrelevante ao modelo ser altamente correlacionada com variáveis relevantes, pois, como dito anteriormente, isso poderia causar imprecisão das estimativas.

Construindo um modelo - Estratégias gerais

Com os avanços de hardwares e softwares, houve uma mudança na construção dos modelos durante os últimos vinte anos. Dado o estágio atual da tecnologia, os pesquisadores se mostram mais confortáveis em iniciar suas pesquisas com modelos grandes e elaborados, o que é justificável, visto que, como vimos, omitir uma variável relevante é bem mais problemático que acrescentar uma variável irrelevante. Todavia, ainda são necessários testes para verificar qual o modelo mais correto para determinada abordagem.

Tratando-se de modelos não aninhados, ou seja, modelos que não são mutuamente excludentes, os testes clássicos tornam-se mais complicados de serem aplicados. A partir

disso, uma atenção geral para essa problemática foi dada por teóricos e empiristas. Uma abordagem bastante chamativa, nesse sentido, é a abordagem bayesiana, uma vez que permite a comparação de duas hipóteses (dois modelos), em vez da validade de uma sobre a outra. Em outras palavras, não há verdade absoluta, o método gera uma razão de probabilidade de um modelo ser melhor, em detrimento do outro.

Entretanto, ao focar nos testes clássicos deve-se salientar que estes, em geral, buscam evidências para refutar a hipótese nula, isto é, rejeitar H_0 . Mas, como considerar qual é H_0 e qual é H_1 ? Para isso, existe a metodologia Neyman-Pearson, a qual sugere que a hipótese nula seja o modelo mais estreito do conjunto considerado. Há um grande problema, no entanto: os modelos clássicos nunca chegam a uma conclusão certa. Isso porque sempre existem possibilidades do erro tipo 1. Em outras palavras, existe um intervalo de confiança que permite rejeitar H_0 , mas este intervalo é sempre menor que um.

Nesse sentido, o primeiro trabalho sobre a escolha de modelos não aninhados foi o de Cox (1961, 1962), baseado em princípios de máxima verossimilhança. Uma abordagem mais recente é baseada no chamado princípio da abrangência. Basicamente, esse princípio procura determinar se o modelo escolhido é capaz de abranger as características de seus concorrentes, i.e, se o modelo é capaz de explicar a hipótese alternativa.

Mais especificamente, um modelo abrangente é um modelo que explica características especiais do seu concorrente. Assim, para realizar o teste, uma possibilidade é o aninhamento artificial dos dois modelos, isto é, juntar suas variáveis próprias e as variáveis que possuem em comum em um só modelo.

$$Y = \bar{\mathbf{X}}\boldsymbol{\beta} + \bar{\mathbf{Z}}\boldsymbol{\gamma} + \mathbf{W}\boldsymbol{\delta} + \varepsilon$$

Todavia, essa possibilidade pode acarretar em uma série de problemas, como alto número de regressores, o que seria problemático para um cenário de série temporal, de modo que poderia causar problemas de colinearidade. Além disso, como um regressor é composto por partes dos regressores dos dois modelos, um teste F seria impreciso, logo não seria possível rejeitar H_0 ou H_1 de fato. Uma outra forma mais interessante é supor que, digamos H_0 seja a hipótese correta, pois, nesse sentido, a variável explicada dependeria apenas de um dos regressores mais o termo de erro. Para tanto, podemos gerar um modelo contando com as variáveis singulares de cada um e testar se o regressor do modelo que corresponde a H_1 é igual a zero, o que indicaria aceitar a hipótese nula. Em outros termos, seria como se estivéssemos adicionando uma variável irrelevante ao modelo, haja vista que o modelo de H_0 já daria conta de explicar.

Uma proposição alternativa é o teste J , proposto por Davidson and MacKinnon (1981), que traz as implicações deste modelo para a regressão linear:

$$\mathbf{y} = (1 - \lambda)\mathbf{X}\beta + \lambda(\mathbf{Z}\gamma) + \varepsilon$$

Assim, se λ é igual a zero, devemos rejeitar H_1 . O problema é que não é possível estimar λ separadamente, assim o teste J consiste em estimar um parâmetro por uma regressão de MQ de y em Z seguida por uma regressão de y em X . Outra possibilidade é testar λ assintoticamente, mas isso não garante um resultado exatamente robusto para amostras finitas.

Outros testes comumente utilizados são os testes de razão de verossimilhança. Os testes de razão de verossimilhança são fundamentados em três características importantes da densidade da variável aleatória de interesse. Primeiramente, sob a hipótese nula, a densidade logarítmica média é esperada ser menor do que na alternativa, o que decorre do fato de que o modelo nulo está contido dentro da alternativa. Em segundo lugar, os graus de liberdade para a estatística qui-quadrado são determinados pela redução na dimensão do espaço de parâmetros especificado pela hipótese nula, comparado com a alternativa. Terceiro, para que o teste seja válido, sob a hipótese nula, a estatística de teste deve seguir uma distribuição conhecida que seja independente dos parâmetros do modelo sob a hipótese alternativa. Quando os modelos não são aninhados, nenhum desses requisitos é completamente satisfeito. Primeiramente, a diferença na densidade logarítmica média pode não ser mantida. Em segundo lugar, o espaço de parâmetros no modelo nulo pode ser maior ou do mesmo tamanho que na alternativa, o que complica a interpretação dos graus de liberdade. Terceiro, as distribuições das estatísticas de teste baseadas em probabilidade geralmente se tornam funções dos parâmetros do modelo alternativo, devido à simetria das hipóteses nula e alternativa.

Para resolver essas questões, a análise de Cox produziu uma reformulação estatística de teste baseada na distribuição normal padrão e centrada em zero. Essa abordagem foi posteriormente adaptada por Pesaran e Deaton para modelos de regressão linear e não linear. Posteriormente, Evans e Deaton estenderam essa estatística de teste para modelos lineares versus modelos log-lineares. É interessante notar que, como nos modelos de regressão clássicos, o estimador de mínimos quadrados é também o estimador de máxima verossimilhança, o que sugere uma relação entre esses diferentes métodos de teste. Por exemplo, Davidson e MacKinnon descobriram que sua estatística de teste é assintoticamente igual ao negativo da estatística Cox-Pesaran e Deaton.

$$c_{01} = \frac{n}{2} \ln \left[\frac{s_Z^2}{s_X^2 + (1/n)\mathbf{b}'\mathbf{X}'\mathbf{M}_Z\mathbf{X}\mathbf{b}} \right] = \frac{n}{2} \ln \left[\frac{s_Z^2}{s_{ZX}^2} \right]$$

sendo:

$$\mathbf{M}_Z = \mathbf{I} - \mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}',$$

$$\mathbf{M}_X = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}',$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y},$$

$$s_Z^2 = \mathbf{e}'_Z \mathbf{e}_Z / n,$$

$$s_X^2 = \mathbf{e}'_X \mathbf{e}_X / n,$$

$$s_{ZX}^2 = s_X^2 + \mathbf{b}'\mathbf{X}'\mathbf{M}_Z\mathbf{X}\mathbf{b},$$

A hipótese é testada ao comparar o valor crítico com a tabela normal. Um alto valor de q é uma evidência para rejeitar a hipótese nula.

$$q = \frac{c_{01}}{\sqrt{\frac{s_X^2}{s_{ZX}^2} \mathbf{b}'\mathbf{X}'\mathbf{M}_Z\mathbf{M}_X\mathbf{M}_Z\mathbf{X}\mathbf{b}}}$$

Heterocedasticidade

A heterocedasticidade, um fenômeno estatístico comum em diversos tipos de dados, se caracteriza pela desigualdade na variância dos erros entre as diferentes observações. Essa variação pode ocorrer tanto em dados transversais quanto em séries temporais.

Em séries temporais de alta frequência, como as cotações diárias no mercado financeiro, a volatilidade dos dados é um indicativo da heterocedasticidade. Já nos dados transversais, a heterocedasticidade se manifesta quando a escala da variável dependente e o poder explicativo do modelo variam entre as observações. Um exemplo clássico são os dados microeconômicos, como pesquisas sobre gastos de consumo. Mesmo considerando o tamanho das empresas, espera-se observar uma maior variabilidade nos lucros de grandes empresas em comparação com as menores. Essa variabilidade também pode ser influenciada por fatores como diversificação de produtos, investimentos em P&D, características do setor de atuação, entre outros, gerando diferenças inclusive entre empresas de porte similar.

Ao analisarmos os padrões de despesa familiar, notamos que a variabilidade das despesas em determinados grupos de bens é maior entre famílias de alta renda do que entre as de baixa renda. Essa diferença se deve à maior discricionariedade proporcionada por uma renda mais elevada.

É importante salientar que, mesmo com a presença de heterocedasticidade, as perturbações

(ou erros) ainda são consideradas não correlacionadas entre si.

Se a matriz de covariância é escrita na forma $E[\varepsilon\varepsilon'|\mathbf{X}] = \sigma^2\Omega$, podemos dizer que no caso homocedástico $\Omega = I$. Se $\Omega \neq I$, então os resíduos são heterocedásticos ou autocorrelacionados ou ambos.

Portanto, o estimador de mínimos quadrados é:

$$\mathbf{b} = \beta + (\mathbf{X}'\mathbf{X})^{-1} \sum_{i=1}^n \mathbf{x}_i \varepsilon_i$$

e sua matriz de covariância é:

$$Var[\mathbf{b}|\mathbf{X}] = \frac{1}{n} \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1} \left(\frac{\mathbf{X}'(\sigma^2\Omega)\mathbf{X}}{n} \right) \left(\frac{\mathbf{X}'\mathbf{X}}{n} \right)^{-1}$$

Logo, vemos que $s^2(\mathbf{X}'\mathbf{X})^{-1}$ não é um estimador apropriado para a matriz de covariância assintótica para o estimador de MQ. Para resolver essa questão, White (1980, 2001) mostrou que, sob condições gerais, o estimador:

$$S_0 = \frac{1}{n} \sum_{i=1}^n e_i^2 \mathbf{x}_i \mathbf{x}_i'$$

converge assintoticamente para a matriz Q_* de soma de quadrados e produtos vetoriais que envolvem σ_{ij} e as linhas de X :

$$Q = \left(\frac{\mathbf{X}'(\sigma^2\Omega)\mathbf{X}}{n} \right)$$

Dessa forma, obtemos o estimador consistente de White:

$$Asy.Var[\mathbf{b}|\mathbf{X}] = n(\mathbf{X}'\mathbf{X})^{-1} S_0 (\mathbf{X}'\mathbf{X})^{-1}$$

que ser usado para estimar a matriz de covariância assintótica de \mathbf{b} . Este resultado implica que, sem realmente especificar o tipo de heterocedasticidade, ainda podemos fazer inferências apropriadas com base no estimador de mínimos quadrados. Esta implicação é útil se não tivermos certeza da natureza da heterocedasticidade.

5. 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$piib_pc <- df_regic$piib / 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.

5.1. 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 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:

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.

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:

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

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

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

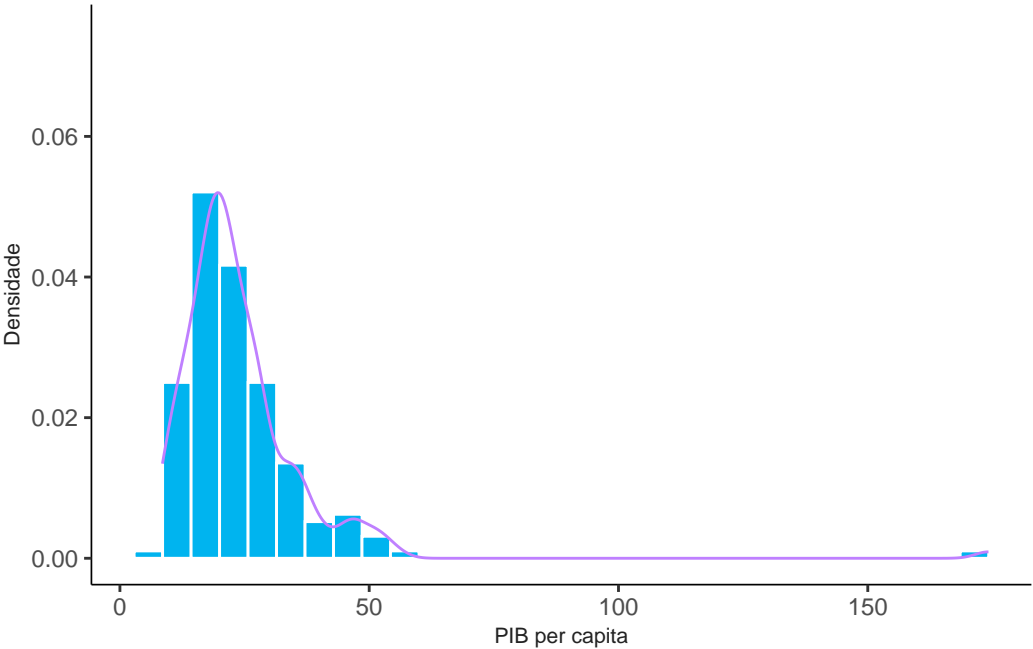


Figura 3: Histograma do Coeficiente de Intensidade da Gestão Empresarial (CIGE) dos municípios de Minas Gerais (2018)

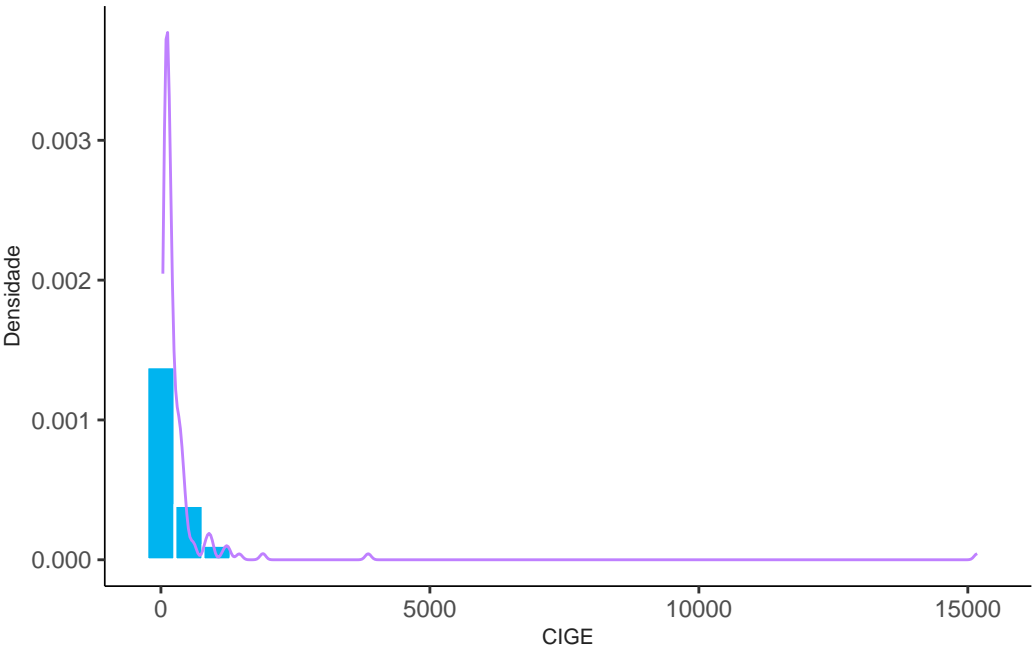


Figura 4: Histograma da Centralidade da Gestão Pública (CGP) dos municípios de Minas Gerais (2018)

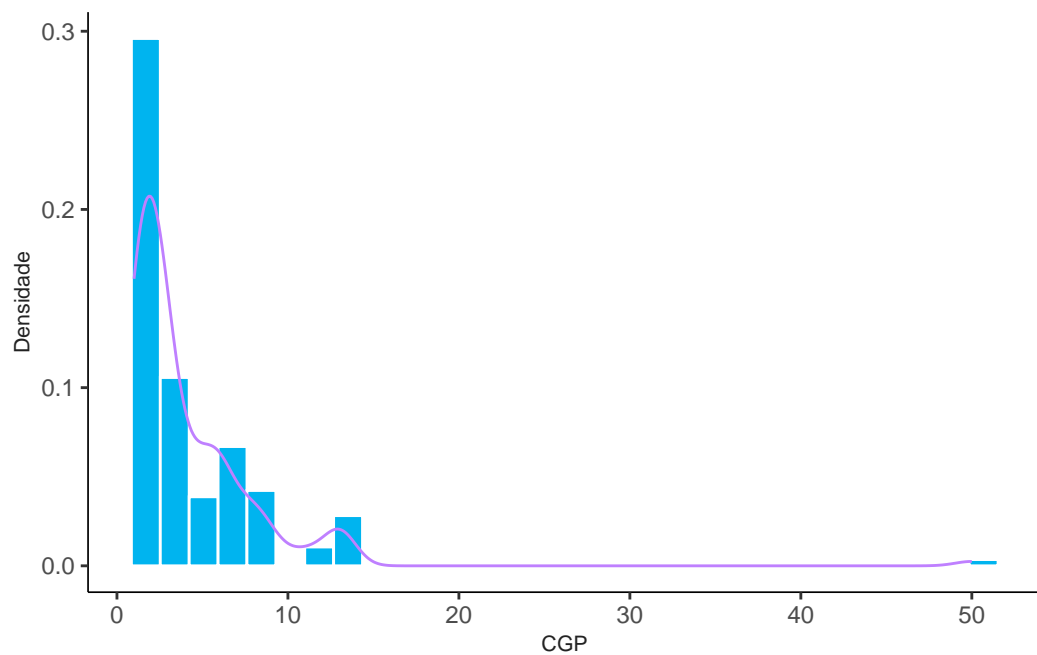
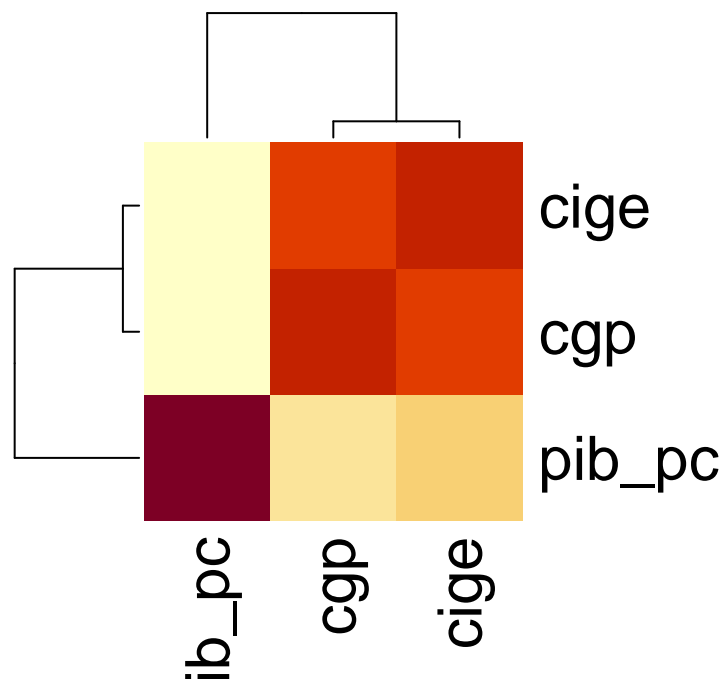


Figura 5: Mapa de calor da matriz de correlação.



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
```

```

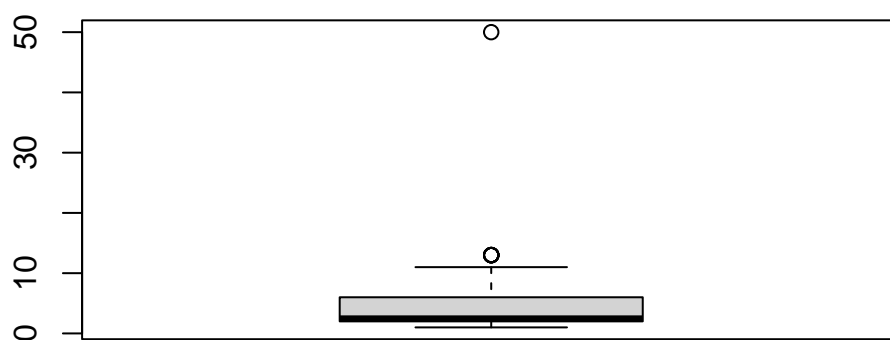
pib_pc  230.22    2298.78    8.27
cige    2298.78 1474769.11 4994.06
cgp      8.27    4994.06    22.55

```

Abaixo plotamos os *boxplots* das variáveis utilizadas:

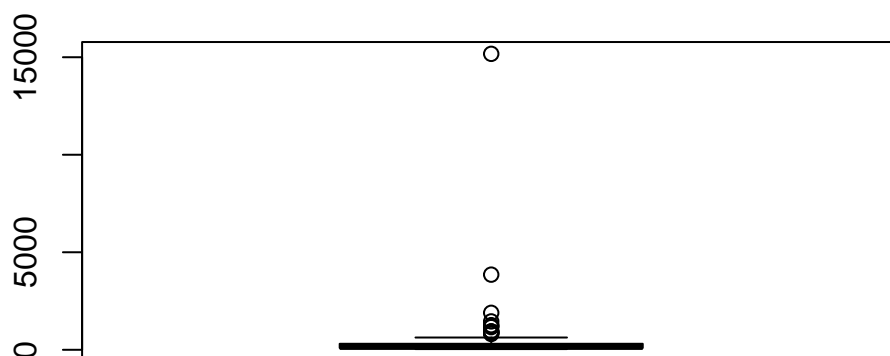
```
boxplot(df_regic$cgp)
```

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



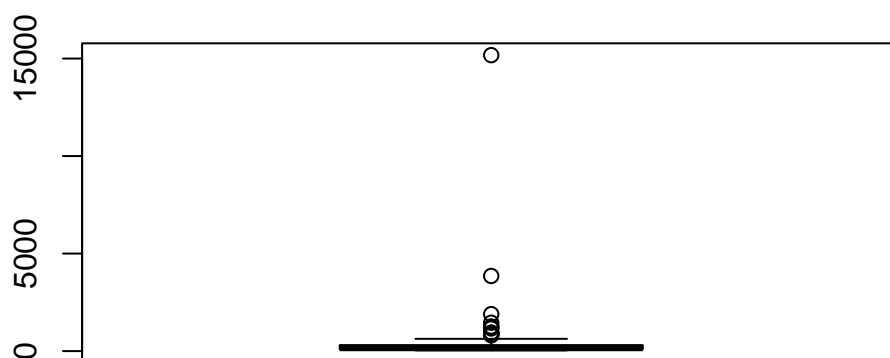
```
boxplot(df_regic$cige)
```


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



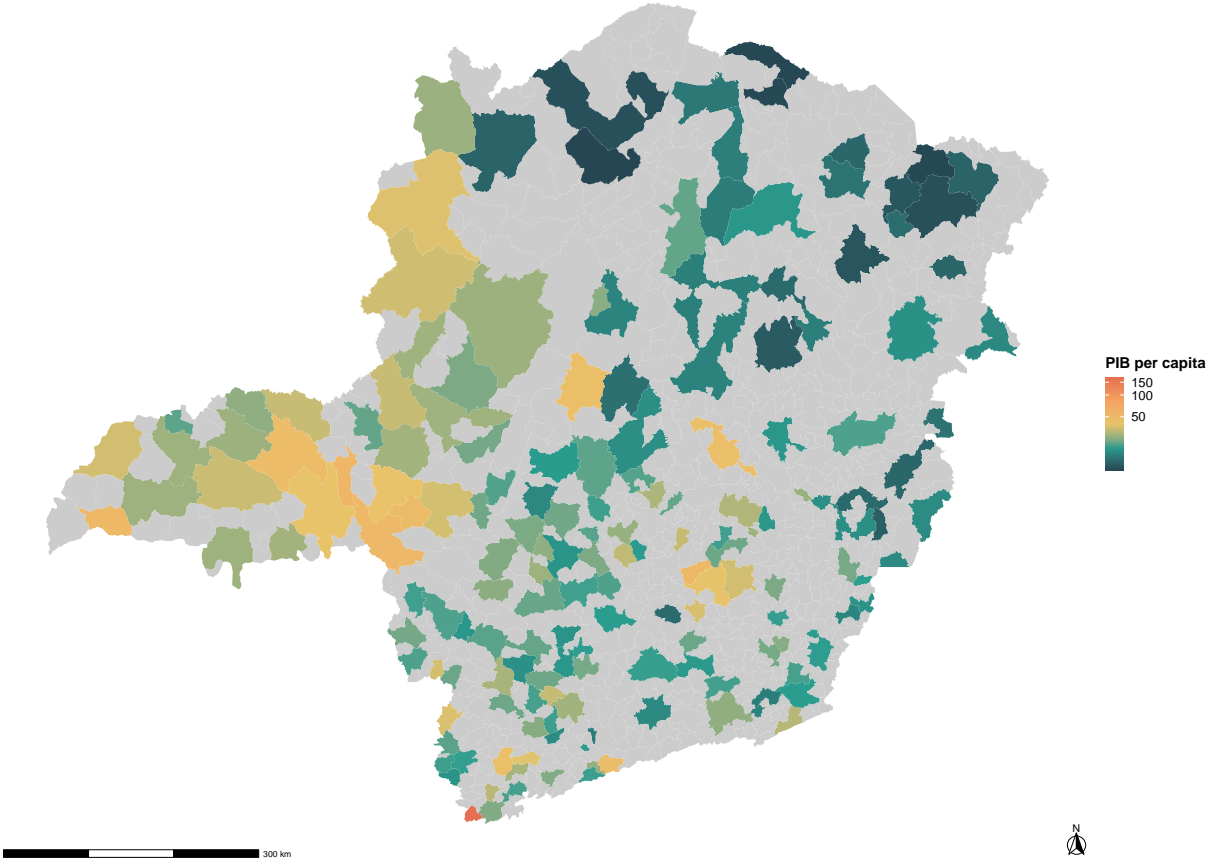
```
boxplot(df_regic$cige)
```

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



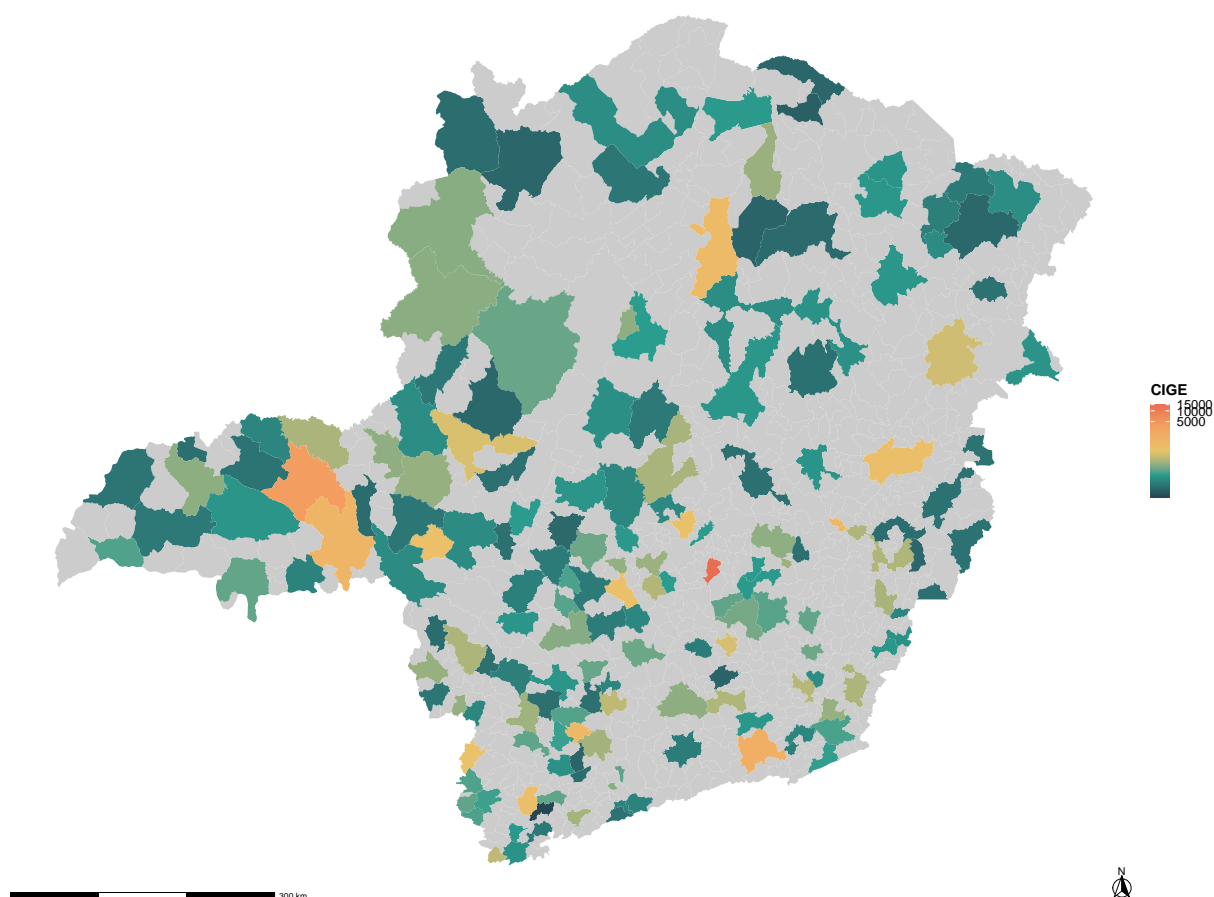
A visualização geográfica das variáveis do modelo (para o caso de municípios) é essencial para compreender, com clareza, onde os processos estão acontecendo:

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



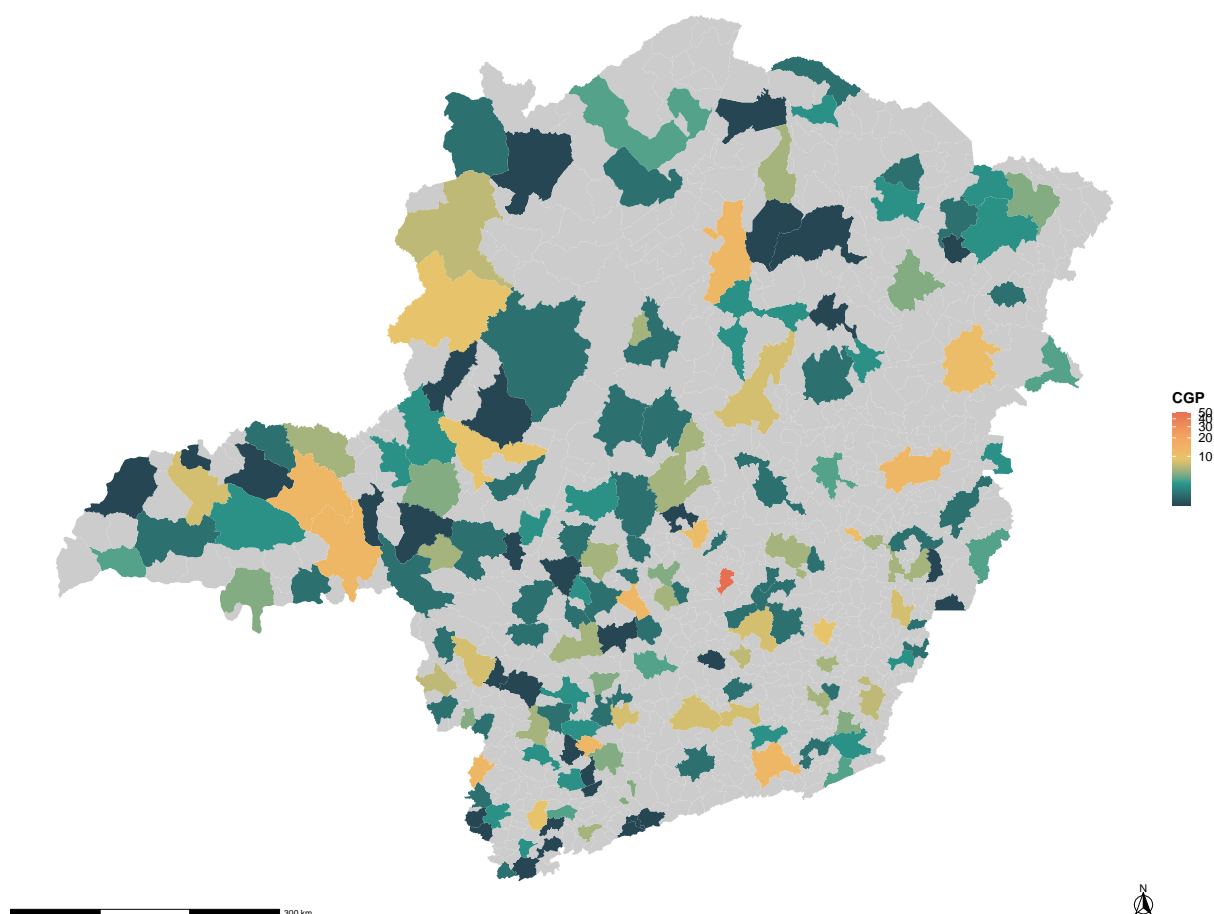
Fonte: Elaboração própria.

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:

5.2. Análise de Regressão

Com as mesmas operações definidas na seção sobre Regressão Por Mínimos Quadrados, vamos estimar um modelo linear que tenta explicar o $\ln(PIB_{pc})$ (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)

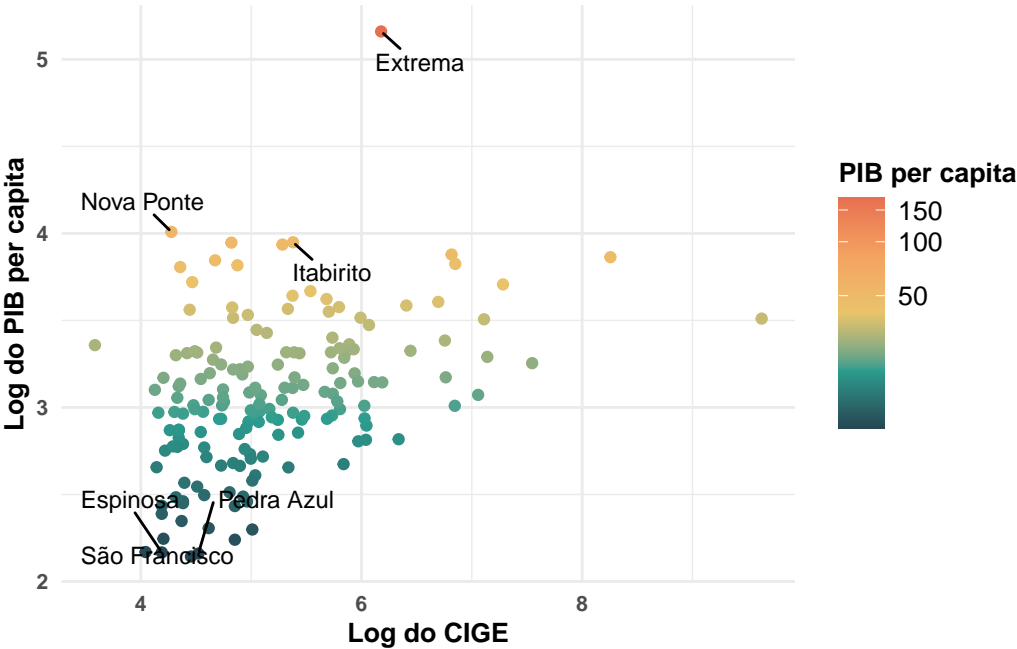
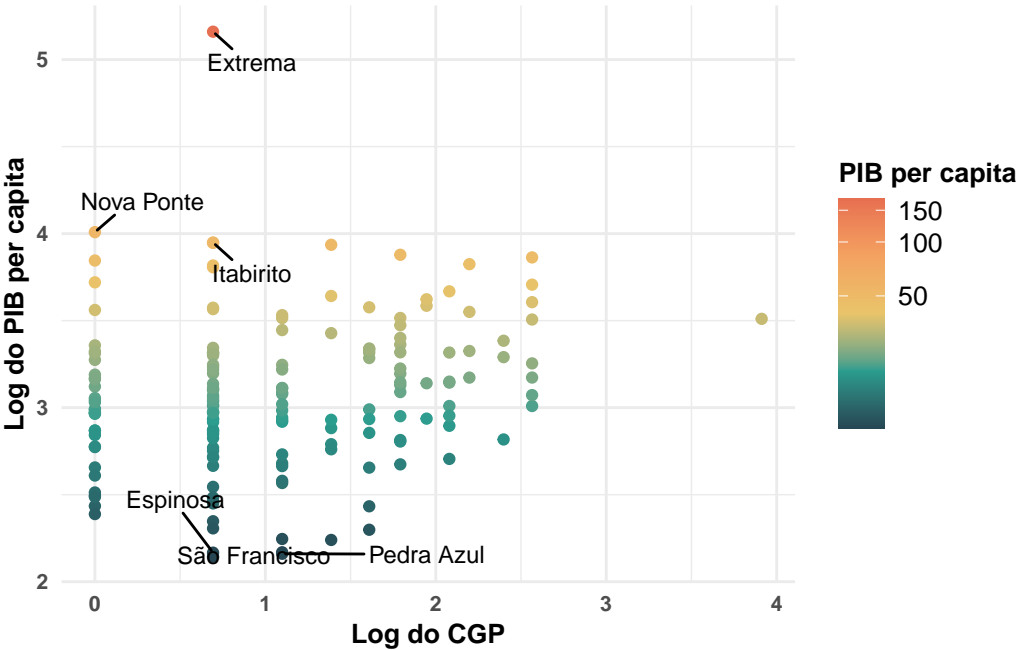


Figura 13: PIB per capita x CGP (Municípios de Minas Gerais - REGIC 2018)



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

b0 <- mediay - b1*mediax

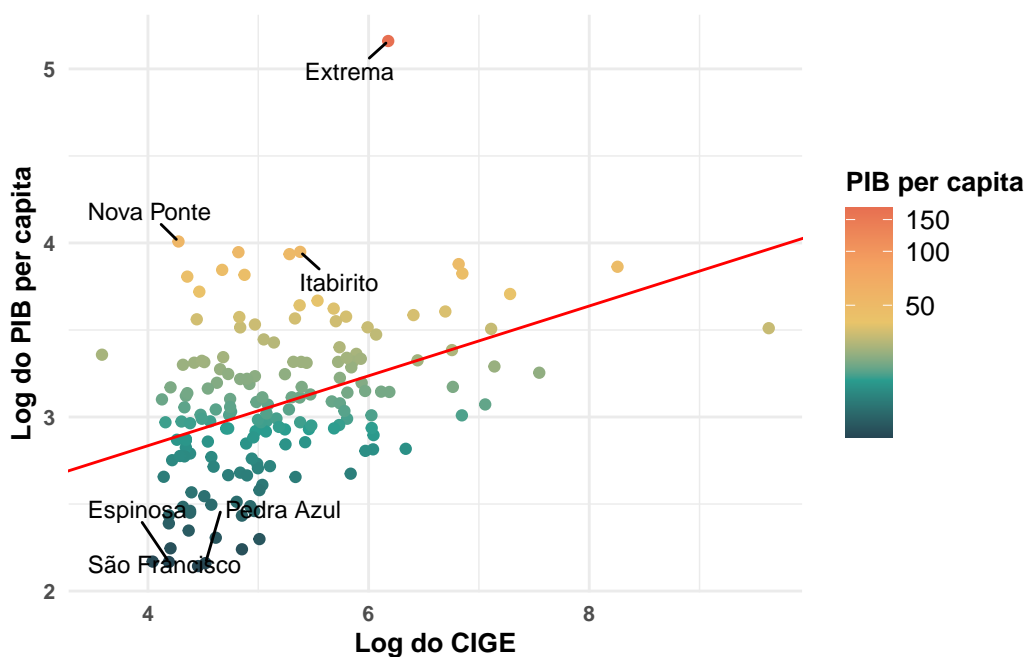
print(paste0("Intercepto: ", round(b0, 2)))
```

```
[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$piib_pc)
x <- log(df_regic$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.

5.3. 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
```

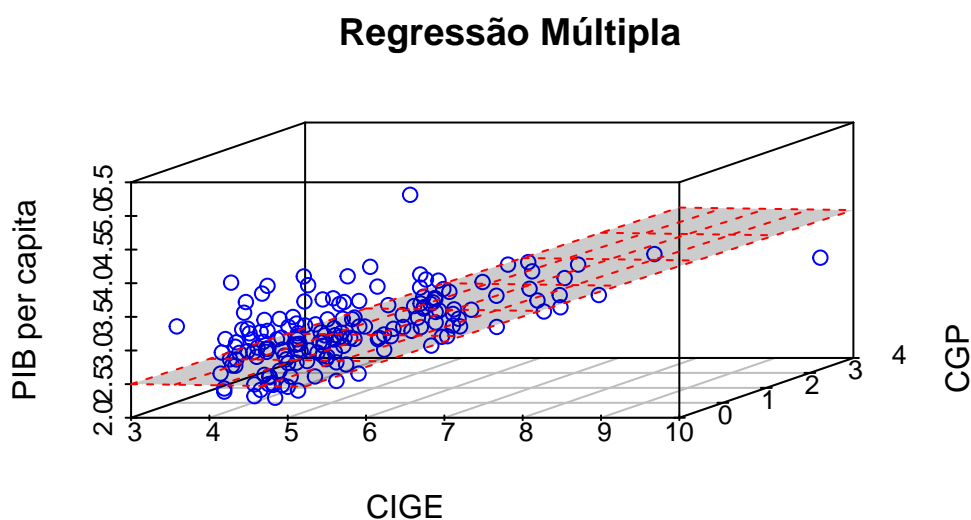
```
varbetahat
```

```
      [,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.

5.4. Análise de Regressão com diferentes formas funcionais

Optamos por substituir as variáveis utilizadas anteriormente para aumentar o tamanho da amostra e inferir melhor sobre os impactos de variáveis institucionais sobre os municípios mineiros. No lugar do Índice de Centralidade da Gestão Pública, incluímos uma variável de valor agregado da administração pública, que está completa na REGIC 2018. Além disso, trocamos o coeficiente de Gestão Empresarial pelo Índice de Atração Geral, que também está completo. Essas alterações mudaram o tamanho da amostra de 168 municípios (arranjos populacionais) para 770.

Para melhorar a análise sobre a presença de bancos nos municípios, adicionamos variáveis de uma base geolocalizada da ESTBAN para completar os dados faltantes na REGIC 2018.

Estimamos 2 modelos, incorporando diferentes formas funcionais. Ambos linearizam as variáveis contínuas com características exponenciais por meio de logaritmo. O primeiro, utiliza dummies da presença de bancos públicos e privados no município, dando relevância à presença de bancos na estimação do PIB per capita. O segundo, utiliza variáveis que mensuram a distância do centróide de um município até o centróide do município com agência bancária (separando por público e privado) mais próximo (municípios com agências bancárias recebem valor 0).

```
df_regic2 <- 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",
  "VAR07",
  "VAR56",
) |>
dplyr::rename(
  "populacao" = "VAR01",
  "pib" = "VAR03",
  "va_adm_publica" = "VAR07",
  "ia" = "VAR56"
) |>
dplyr::mutate(
  "populacao" = as.numeric(populacao),
  "pib" = as.numeric(pib),
  "va_adm_publica" = as.numeric(va_adm_publica),
```

```

    "log_ia" = ifelse(as.numeric(ia) < 1, 0, log(as.numeric(ia)))
  )

df_regic2 <- na.omit(df_regic2)

df_regic2$pib_pc <- df_regic2$pib / df_regic2$populacao

project_id <- "cloud-learning-doing"

sql <- "SELECT * FROM estban.estban_agencias_geolocalizadas WHERE data_base = '2018-12-01'"

query <- bigrquery::bq_project_query(
  project_id,
  sql,
)

agencias_2018 <- bigrquery::bq_table_download(query)

minas_gerais <- geobr::read_state(code_state = 31, year = 2010, showProgress = FALSE, simplified = FALSE)

# Agências bancárias dentro do território mineiro
agencias_2018_mg <- agencias_2018 |>
  dplyr::filter(!is.na(lat) & !is.na(lng)) |>
  sf::st_as_sf(coords = c("lng", "lat"), crs = 4674) |>
  sf::st_filter(minas_gerais) |>
  dplyr::filter(uf == "MG")

# CNPJ dos bancos públicos (BB e CEF)
cnpj_banco_publico <- c("00000000", "00360305")

# Classificando as agências
agencias_2018_mg <- agencias_2018_mg |> mutate(
  tipo_banco = case_when(
    cnpj %in% cnpj_banco_publico ~ "Banco público",
    TRUE ~ "Banco privado"
  )
)

# Enriquecendo municípios com os dados da REGIC

municipios <- geobr::read_municipality(code_muni = 31, showProgress = FALSE, simplified = FALSE)

municipios <- municipios |> dplyr::left_join(
  df_regic2,
  by = c("code_muni" = "COD_CIDADE")
)

```

```

)

# Contando os bancos públicos e privados por cidade
bancos_municipios <- agencias_2018_mg |>
  sf::st_drop_geometry() |>
  dplyr::group_by(cod_mun_ibge, tipo_banco) |>
  dplyr::summarise(agencias = n()) |>
  tidyr::pivot_wider(
    names_from = tipo_banco,
    values_from = agencias,
    values_fill = 0
  ) |>
  dplyr::mutate(
    "cod_mun_ibge" = as.numeric(cod_mun_ibge),
    "banco_publico" = ifelse(`Banco público` > 0, 1, 0),
    "banco_privado" = ifelse(`Banco privado` > 0, 1, 0)
  ) |>
  dplyr::select(
    "cod_mun_ibge",
    "banco_publico",
    "banco_privado"
  ) # Dummies para presença de banco público e/ou banco privado

# Enriquecendo com as informações sobre banco
municipios <- municipios |> dplyr::left_join(
  bancos_municipios,
  by = c("code_muni" = "cod_mun_ibge")
) |> dplyr::mutate(
  "banco_publico" = ifelse(is.na(banco_publico), 0, banco_publico),
  "banco_privado" = ifelse(is.na(banco_privado), 0, banco_privado)
)

# Adicionando os centróides dos municípios
municipios <- sf::st_make_valid(municipios)
municipios$centroid <- sf::st_centroid(municipios)

coords <- sf::st_coordinates(municipios$centroid)
mat.dist <- as.matrix(dist(coords, method = "euclidean"))
dist2 <- mat.dist^2
diag(dist2) <- 0

rownames(dist2) <- municipios$code_muni
colnames(dist2) <- municipios$code_muni

municipios_banco_publico <- municipios |> dplyr::filter(banco_publico == 1)

```

```

municipios_banco_privado <- municipios |> dplyr::filter(banco_privado == 1)

resultados <- data.frame(city = character(nrow(dist2)), dist2.min.banco.publico = numeric(nrow(

for(i in 1:nrow(dist2)){
  city <- names(dist2[, 1])[i]

  if (city %in% as.character(municipios_banco_publico$code_muni)){
    dist.pub <- 0
  } else {
    line <- dist2[i,]
    dist.pub <- 100000000000
    for (j in 1:length(line)) {
      if (j != i & line[j] < dist.pub){
        if (names(line)[j] %in% as.character(municipios_banco_publico$code_muni))
          dist.pub <- line[j]
      }
    }
  }
}

if (city %in% as.character(municipios_banco_privado$code_muni)){
  dist.priv <- 0
} else {
  line <- dist2[i,]
  dist.priv <- 100000000000
  for (j in 1:length(line)) {
    if (j != i & line[j] < dist.priv){
      if (names(line)[j] %in% as.character(municipios_banco_privado$code_muni))
        dist.priv <- line[j]
    }
  }
}

resultados[i, 1] <- city
resultados[i, 2] <- dist.pub
resultados[i, 3] <- dist.priv
}

resultados <- resultados |> mutate("cod" = as.numeric(city))

municipios <- municipios |> dplyr::left_join(
  resultados,
  by = c("code_muni" = "cod")
)

```

```

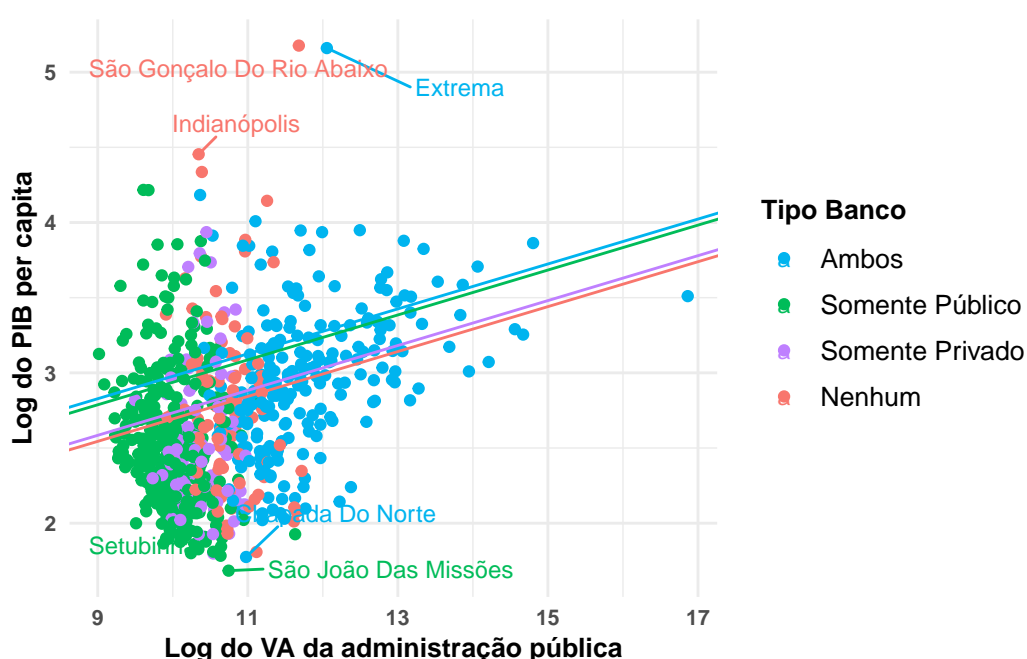
modelo_dummies <- lm(log(pib_pc) ~ log_ia + log(va_adm_publica) + banco_publico + banco_privado)

modelo_dist <- lm(log(pib_pc) ~ log_ia + log(va_adm_publica) + dist2.min.banco.publico + dist2.

```

A Figura 14 mostra como o valor adicionado da Administração Pública está correlacionada com o PIB per capita, controlando pelos tipos de bancos presentes nos municípios. As diferentes linhas de regressão mostram como a presença de um banco público eleva o nível de renda per capita dos municípios. O efeito dos bancos privados não foi significativo.

Figura 14: PIB per capita x Valor adicionado da Administração Pública (Municípios de Minas Gerais - REGIC 2018)



A Figura 15 mostra que o Índice de Atratividade não está tão relacionado com PIB per capita. Apesar de ter um efeito estatisticamente diferente de zero, nem conseguimos ver as retas de regressão no gráfico, ou seja, nesse modelo em específico, não há tanto ajuste para essa variável. Os municípios com população pequena em Minas Gerais e que não são tão atrativos comercialmente, podem estar sobrevalorizando o PIB per capita e distorcendo a análise. A Figura 16 mostra que o Índice de Atratividade está muito mais correlacionado ao tamanho do PIB, das grandes cidades.

A Figura 17 mostra que quanto mais distantes os municípios estão de um banco público, menos renda per capita eles detêm, indicando menor desenvolvimento à medida que se dificulta o acesso às agências, dado que há maiores custos de deslocamento e menor interação entre agentes mais distantes no espaço.

O mesmo acontece para a distância que os municípios estão de bancos privados (Figura 18), mas numa intensidade menor. No modelo com ambas as variáveis, apenas a distância para o

Figura 15: PIB per capita x Índice de Atração Geral (Municípios de Minas Gerais - REGIC 2018)

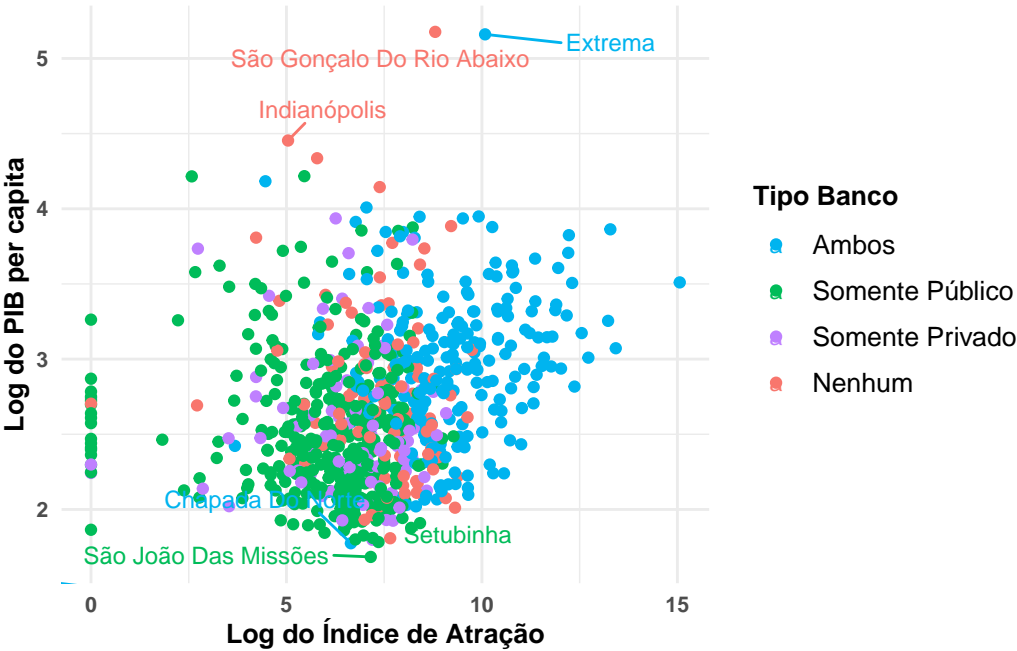


Figura 16: PIB x Índice de Atração Geral (Municípios de Minas Gerais - REGIC 2018)

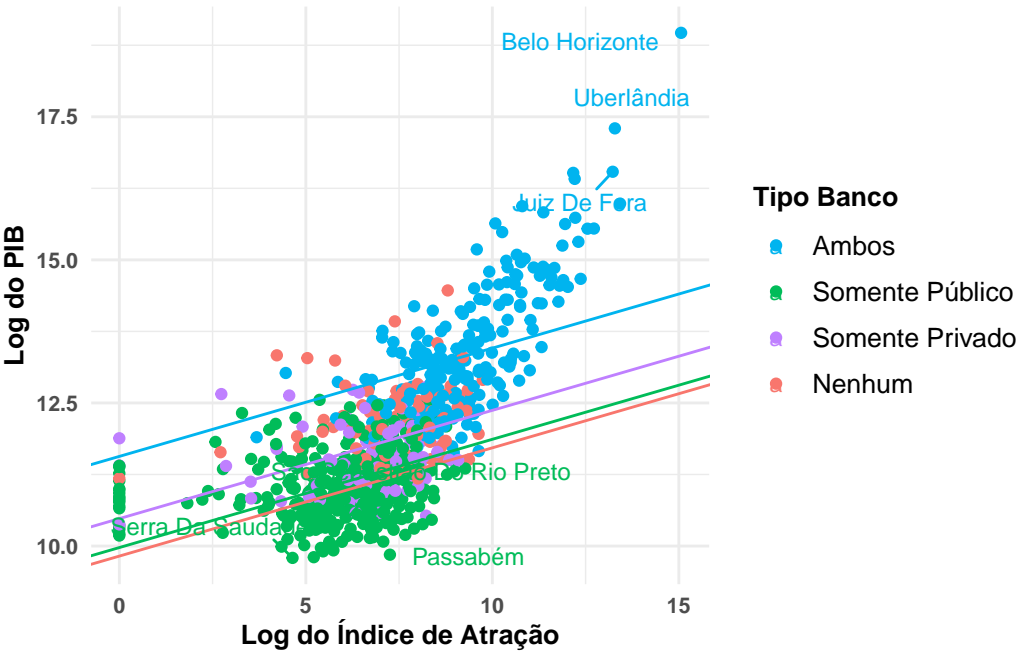
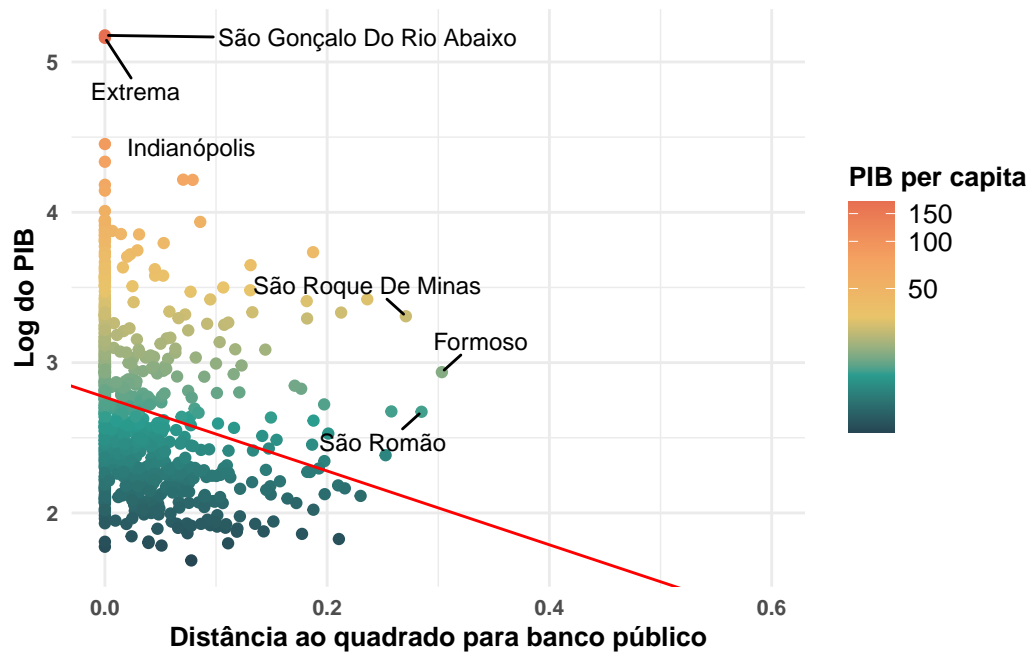


Figura 17: PIB per capita x Distância ao quadrado da agência (banco público) mais próxima (Municípios de Minas Gerais - REGIC 2018)



banco público foi significativa na estimação.

A Tabela 2 mostra o resultado da estimação e serve de comparativo entre os modelos com dummy bancária e o que incorpora o quadrado da distância.

5.5. Testes

Para testar multicolinearidade, podemos utilizar o fator de inflacionamento da variância (VIF), definido na seção sobre multicolinearidade:

```
library(car)
```

```
# Modelo de dummies
```

```
car::vif(modelo_dummies)
```

log_ia	log(va_adm_publica)	banco_publico	banco_privado
2.276871	3.553922	2.107366	1.684299

```
# Modelo de distância
```

```
car::vif(modelo_dist)
```

log_ia	log(va_adm_publica)	dist2.min.banco.publico	dist2.min.banco.privado
2.258555	2.381648	1.484573	
			1.378272

Figura 18: PIB per capita x Distância ao quadrado da agência (banco privado) mais próxima (Municípios de Minas Gerais - REGIC 2018)

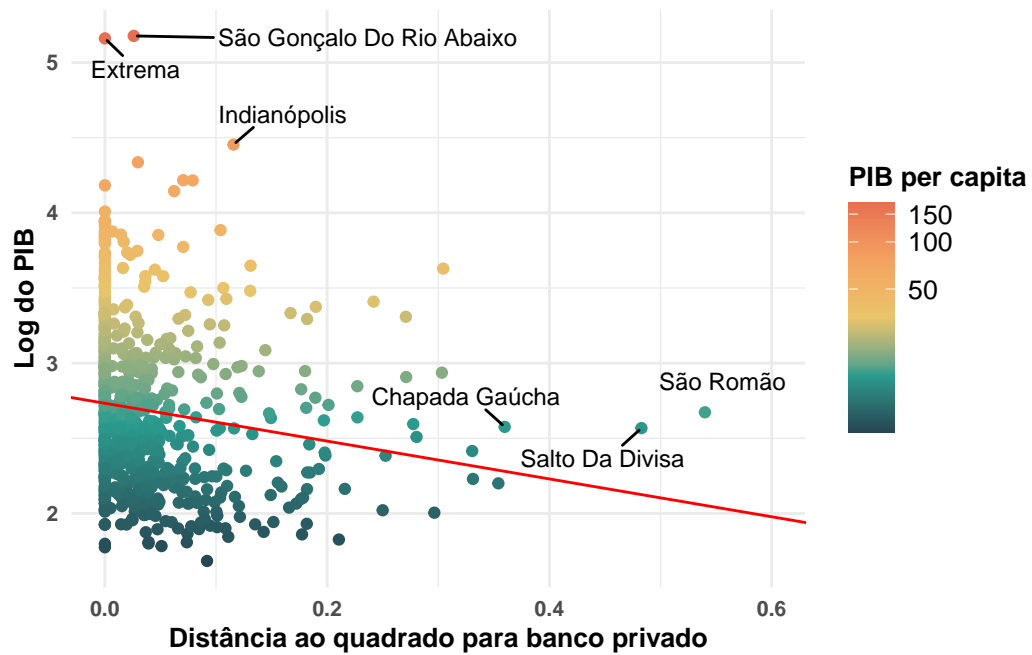


Tabela 2: Modelos estimados

	Dummies	
	(1)	(2)
log_ia	−0.031*** (0.011)	−0.027** (0.012)
log(va_adm_publica)	0.149*** (0.033)	0.221*** (0.027)
banco_publico	0.243*** (0.050)	
banco_privado	0.040 (0.045)	
dist2.min.banco_publico		−1.227*** (0.419)
dist2.min.banco_privado		0.017 (0.323)
Constant	1.200*** (0.300)	0.566** (0.245)
Observations	770	770
R ²	0.171	0.155
Adjusted R ²	0.167	0.151
Residual Std. Error (df = 765)	0.469	0.474
F Statistic (df = 4; 765)	39.522***	35.098***

Notes:

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

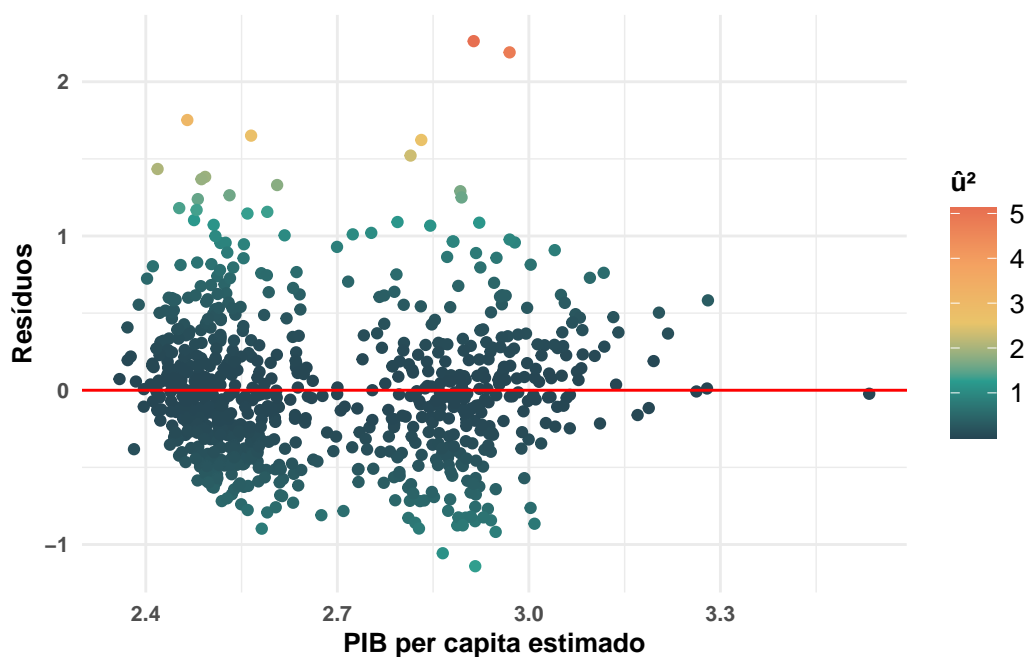
* Significant at the 10 percent level.

Fonte: Elaboração própria.

Como os valores do VIF são baixos, podemos dizer que não há multicolinearidade prejudicial em nenhum dos modelos estimados. Não sendo necessário realizar ajustes nesse sentido.

No que tange a heterocedasticidade, podemos tentar visualizar graficamente a interação entre os resíduos e os valores estimados:

Figura 19: PIB per capita estimado x Resíduos (Modelo com dummies bancárias)



A princípio, graficamente, não há correlação entre os resíduos e os valores estimados, indicando que não há heterocedasticidade nos modelos estimados. Apesar disso, faremos o teste de Breusch-Pagan para verificar se os erros são normalmente distribuídos. A hipótese nula é de que a variância dos resíduos é igual à σ^2 . Rejeitaremos a hipótese nula, caso $p\text{-value} < 0,05$:

```
library(lmtest)
```

```
lmtest::bptest(modelo_dummies)
```

studentized Breusch-Pagan test

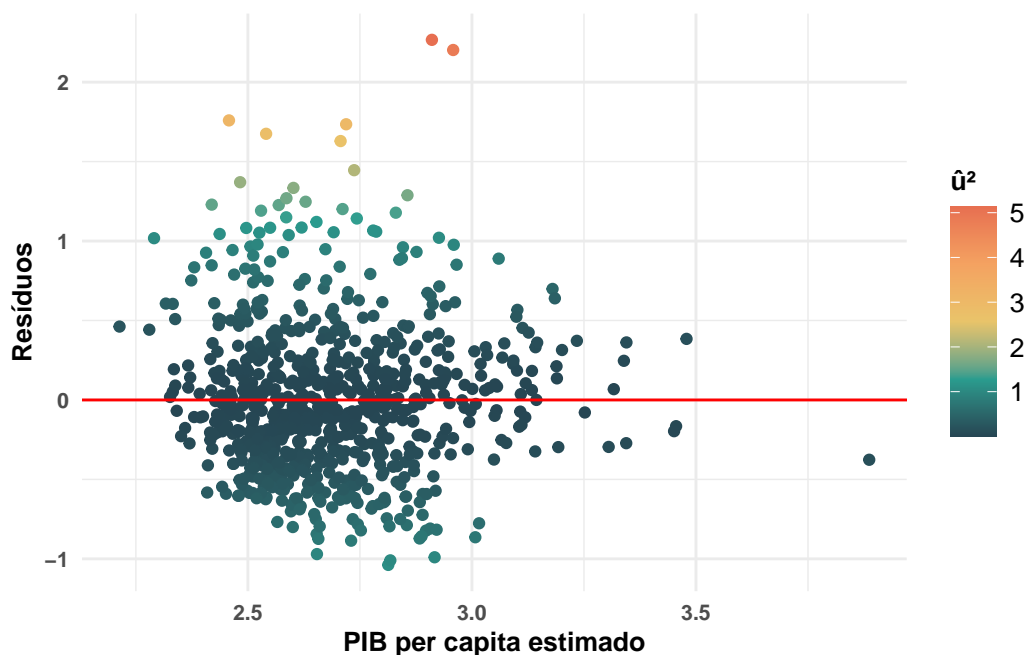
```
data: modelo_dummies
```

```
BP = 9.589, df = 4, p-value = 0.04795
```

```
lmtest::bptest(modelo_dummies)
```

studentized Breusch-Pagan test

Figura 20: PIB per capita estimado x Resíduos (Modelo com distância)



```
data: modelo_dummies
```

```
BP = 9.589, df = 4, p-value = 0.04795
```

Como em ambos os modelos, $p - value < 0,05$ rejeitamos a hipótese nula e tomamos os modelos como heterocedásticos.

Na presença de heterocedasticidade, os erros padrões não são confiáveis, e portanto devemos fazer a correção de White para que inclua erros robustos aos modelos.

```
library(sandwich)
library(lmtest)

# Erros padrões robustos após correção pela matrix de White
erros_robustos <- coeftest(modelo_dummies, vcov=sandwich::vcovHC(modelo_dummies, type="HC0"))
modelo_dummies_robusto <- modelo_dummies
modelo_dummies_robusto$coefficients <- erros_robustos[,1]

# Erros padrões robustos após correção pela matrix de White
erros_robustos <- coeftest(modelo_dist, vcov=sandwich::vcovHC(modelo_dist, type="HC0"))
modelo_dist_robusto <- modelo_dist
modelo_dist_robusto$coefficients <- erros_robustos[,1]
```

A Tabela 3 mostra a diferença dos coeficientes após as correções de heterocedasticidade nos modelos.

Tabela 3: Comparação entre os modelos padrão e com erros robustos

	Dummies			
	(1)	(2)	(3)	(4)
log_ia	−0.031*** (0.011)	−0.031*** (0.011)	−0.027** (0.012)	−0.027** (0.012)
log(va_adm_publica)	0.149*** (0.033)	0.149*** (0.033)	0.221*** (0.027)	0.221*** (0.027)
banco_publico	0.243*** (0.050)	0.243*** (0.050)		
banco_privado	0.040 (0.045)	0.040 (0.045)		
dist2.min.banco.publico			−1.227*** (0.419)	−1.227*** (0.419)
dist2.min.banco.privado			0.017 (0.323)	0.017 (0.323)
Constant	1.200*** (0.300)	1.200*** (0.300)	0.566** (0.245)	0.566** (0.245)
Observations	770	770	770	770
R ²	0.171	0.171	0.155	0.155
Adjusted R ²	0.167	0.167	0.151	0.151
Residual Std. Error (df = 765)	0.469	0.469	0.474	0.474
F Statistic (df = 4; 765)	39.522***	39.522***	35.098***	35.098***

Notes:

*** Significant at the 1 percent level.

** Significant at the 5 percent level.

* Significant at the 10 percent level.

Fonte: Elaboração própria.

6. 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.