Sumário

1 Introdução

O objetivo deste relatório é identificar os principais fatores que influenciam a inflação nos anos enre 2002 e 2022 e oferecer subsídios para a formulação de políticas econômicas mais assertivas, por meio de análises que utilizam métodos estatísticos para obter gráficos e valores.

Este projeto é baseado em um conjunto de dados sobre o índice de inflação e diversas variáveis econômicas no Brasil. Onde as variáveis de interesse a serem utilizadas no estudo são IPCA acumulado por ano, IPCA acumulado por doze meses, salário mínimo, meta Selic e INCC; que são variáveis quantiaivas conínuas.

2 Referencial Teórico

2.1 Média

A média se constitui como a razão entre a soma do valor das observações e o número total delas, dada pela fórmula:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Com:

- $i = 1, 2, ..., n$
- $n =$ número total de observações

2.2 Mediana

Sejam as n observações de um conjunto de dados $X = X_{(1)}, X_{(2)}, \dots, X_{(n)}$ de determinada variável ordenadas de forma crescente. A mediana do conjunto de dados X é a medida que divide o valor das observações ao meio, de modo que metade delas tenham valor menor que a mediana e a outra metade, maior.
Com isso, pode-se calcular a mediana da seguinte forma:

$$med(X) = \begin{cases} X_{\frac{n+1}{2}}, \text{ para n ímpar;} \\ \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2}, \text{para n par.} \end{cases}$$

2.3 Variância

A variância é uma medida que avalia o quanto que os dados estão dispersos em relação à média, em uma escala ao quadrado da escala dos dados que dificulta a interpretação dessa medida.

2.3.1 Variância Populacional

Para uma população, a variância é dada por:

$$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$$

Com:

- $X_i = i$-ésima observação da população
- $\mu =$ média populacional
- $N =$ tamanho da população

2.3.2 Variância Amostral

Para uma amostra, a variância é dada por:

$$S^2 = \frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $n =$ tamanho da amostra

2.4 Desvio Padrão

O desvio padrão é a raiz quadrada da variância, haja vista a ideia de retirar a escala ao quadrado da variância para que se tenha uma medida mais facilmente interpretável. Avalia o quanto os dados estão dispersos em relação à média.

2.4.1 Desvio Padrão Populacional

Para uma população, o desvio padrão é dado por:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \mu\right)^2}{N}}$$

Com:

- $X_i =$ i-ésima observação da população
- $\mu =$ média populacional
- $N =$ tamanho da população

2.4.2 Desvio Padrão Amostral

Para uma amostra, o desvio padrão é dado por:

$$S = \sqrt{\frac{\sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2}{n-1}}$$

Com:

- $X_i =$ i-ésima observação da amostra
- $\bar{X} =$ média amostral
- $n =$ tamanho da amostra

2.5 Boxplot

O boxplot é uma representação gráfica na qual se pode perceber de forma mais clara como uma variável quantitativa está distribuída. A figura abaixo ilustra um exemplo de boxplot.

A parte inferior do retângulo corresponde ao primeiro quartil, enquanto a parte superior representa o terceiro quartil. O traço dentro do retângulo indica a mediana, que divide o conjunto de dados em duas partes de tamanhos iguais. A média é ilustrada por um losango branco, e os pontos representam os outliers. Outliers são valores discrepantes da série de dados, ou seja, valores que não refletem a realidade do conjunto.

2.6 Gráfico de Dispersão

O gráfico de dispersão é uma representação gráfica utilizada para ilustrar o comportamento conjunto de duas variáveis quantitativas. A figura abaixo ilustra um exemplo de gráfico de dispersão, onde cada ponto representa uma observação do banco de dados.

2.7 Tipos de Variáveis

2.7.1 Qualitativas

As variáveis qualitativas são as variáveis não numéricas, que representam categorias ou características da população. Estas subdividem-se em:

- Nominais: quando não existe uma ordem entre as categorias da variável (exemplos: sexo, cor dos olhos, fumante ou não, etc)

- Ordinais: quando existe uma ordem entre as categorias da variável (exemplos: nível de escolaridade, mês, estágio de doença, etc)

2.7.2 Quantitativas

As variáveis quantitativas são as variáveis numéricas, que representam características numéricas da população, ou seja, quantidades. Estas subdividem-se em:

- Discretas: quando os possíveis valores são enumeráveis (exemplos: número de filhos, número de cigarros fumados, etc)
- Contínuas: quando os possíveis valores são resultado de medições (exemplos: massa, altura, tempo, etc)

2.8 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida que verifica o grau de relação linear entre duas variáveis quantitativas. Este coeficiente varia entre os valores -1 e 1. O valor zero significa que não há relação linear entre as variáveis. Quando o valor do coeficiente r é negativo, diz-se existir uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando r é positivo, diz-se que as duas variáveis são diretamente proporcionais.

O coeficiente de correlação de Pearson é normalmente representado pela letra r e a sua fórmula de cálculo é:

$$r_{Pearson} = \frac{\sum_{i=1}^{n} \left[(x_i - \bar{x}) \, (y_i - \bar{y}) \right]}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2} \times \sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

Onde:

- x_i = i-ésimo valor da variável X
- y_i = i-ésimo valor da variável Y
- $\bar{x}$ = média dos valores da variável X
- $\bar{y}$ = média dos valores da variável Y

Vale ressaltar que o coeficiente de Pearson é paramétrico e, portanto, sensível quanto à normalidade (simetria) dos dados.

2.9 Coeficiente de Correlação de Spearman

O coeficiente de correlação de Spearman é uma medida não paramétrica que verifica através de postos de variáveis quantitativas ou qualitativas ordinais o grau de relação monótona entre duas variáveis. Este coeficiente varia entre os valores -1 e 1. O valor zero significa que não há relação monótona entre as variáveis. Quando o valor do coeficiente ρ é negativo, diz-se ter uma relação de grandeza inversamente proporcional entre as variáveis. Analogamente, quando ρ é positivo, diz-se que as duas variáveis são diretamente proporcionais. O coeficiente é calculado da seguinte maneira:

$$\rho_{Spearman} = \frac{\sum_{i=1}^{n}\left[\left(R(x_i) - \frac{n+1}{2}\right)\left(R(y_i) - \frac{n+1}{2}\right)\right]}{\sqrt{\sum_{i=1}^{n}\left(R(x_i)^2\right) - n\left(\frac{n+1}{2}\right)^2} \times \sqrt{\sum_{i=1}^{n}\left(R(y_i)^2\right) - n\left(\frac{n+1}{2}\right)^2}}$$

No qual:

- $x_i =$ i-ésimo valor da variável X
- $y_i =$ i-ésimo valor da variável Y
- $R(x_i) =$ posto relativo a observação i de X
- $R(y_i) =$ posto relativo a observação i de Y
- $n =$ número total de observações na amostra

Observação: Ao ordenar de forma crescente a amostra de observações $x_1, x_2, ..., x_n$, $y_1, y_2, ..., y_n$, diz-se que $R(x_1) = 1$ se x_1 é o menor valor dentre as duas amostras, $R(x_3) = 2$ se x_3 é o segundo menor valor dentre as duas amostras, $R(y_4) = 3$ se y_4 é o terceiro menor valor dentre as duas amostras, e assim sucessivamente.

2.10 Coeficiente de Determinação (R^2)

O coeficiente R^2 de determinação utiliza a variância dentro de cada grupo como insumo para explicar a variância global dos dados. Uma forma de quantificar essa medida é utilizar a média das variâncias em cada categoria, dada por:

$$\overline{var(S)} = \frac{\sum_{i=1}^{k} n_i \times var_i(S)}{n}$$

No qual:

- $n =$ tamanho total da amostra
- $var_i(S) =$ variância dentro da categoria i
- $n_i =$ tamanho da amostra i

Assim, o coeficiente de determinação é dado por:

$$R^2 = 1 - \frac{\overline{var(S)}}{var(S)}$$

Com $0 \leq R^2 \leq 1$. Além disso, 1 indica que a variável categórica explica 100% da variação da variável quantitativa e 0 indica que a variável categórica não impacta na variância da variável quantitativa.

2.11 Nível de significância (α)

Nível de significância do teste é o nome dado à probabilidade de se rejeitar a hipótese nula quando essa é verdadeira; essa rejeição é chamada de *erro do tipo I*. O valor de α é fixado antes da extração da amostra e, usualmente, assume 5%, 1% ou 0,1%.

Por exemplo, um nível de significância de $\alpha = 0,05$ (5%) significa que, se for tomada uma grande quantidade de amostras, em 5% delas a hipótese nula será rejeitada quando não havia evidências para essa rejeição, isto é, a probabilidade de se tomar a decisão correta é de 95%.

2.12 P-valor

P-valor, ou nível descritivo, é uma medida utilizada para sintetizar o resultado de um teste de hipóteses. Ele pode ser chamado também de *probabilidade de significância* do teste e indica a probabilidade de se obter um resultado da estatística de teste mais extremo do que o observado na presente amostra, considerando que a hipótese nula é verdadeira. Dessa forma, rejeita-se H_0 para P-valor $< \alpha$, porque a chance de uma nova amostra possuir valores tão extremos quanto o encontrado é baixa, ou seja, há evidências para a rejeição da hipótese nula.

2.13 Análise de Variância (ANOVA)

A Análise de Variância, mais conhecida por ANOVA, consiste em um teste de hipótese, em que é testado se as médias dos tratamentos (ou grupos) são iguais. Os dados são descritos pelo seguinte modelo:

$$y_{ij} = \mu + \alpha_i + e_{ij},\ i = 1, ..., a\ e\ j = 1, ..., N$$

Em que:

- i é o número de tratamentos
- j é o número de observações
- y_{ij} é a j-ésima observação do i-ésimo tratamento

No modelo, μ é a média geral dos dados e α_i é o efeito do tratamento i na variável resposta. Já e_{ij} é a variável aleatória correspondente ao erro. Supõe-se que tal variável tem distribuição de probabilidade Normal com média zero e variância σ^2. Mais precisamente, $e_{ij} \sim N(0, \sigma^2)$.

A variabilidade total pode ser decomposta na variabilidade devida aos diferentes tratamentos somada à variabilidade dentro de cada tratamento:

$$\begin{aligned} Soma\ de\ Quadrados\ Total\ (SQTOT) &= Soma\ de\ Quadrados\ de\ Tratamento\ (SQ \\ &+ Soma\ de\ Quadrados\ de\ Residuos\ (SQR \end{aligned}$$

Sendo o estudo não balanceado, ou seja, quando os tratamentos possuem tamanhos de amostra distintos:

$$SQTOT = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \frac{y_{..}^2}{N}$$

$$SQTRAT = \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i} - \frac{y_{..}^2}{N}$$

$$SQRES = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}^2 - \sum_{i=1}^{a} \frac{y_{i.}^2}{n_i}$$

Em que:

- n_i é o número de observações do i-ésimo tratamento
- N é o número total de observações
- $y_{..} = \sum_{i=1}^{a}\sum_{j=1}^{n_i} y_{ij}$
- $y_{i.} = \sum_{j=1}^{n_i} y_{ij}$

As hipóteses do teste são:

$$\begin{cases} H_0 : \text{As médias dos } a \text{ tratamentos são iguais} \\ H_1 : \text{Existe pelo menos um par de médias diferente} \end{cases}$$

A estatística do teste é composta pelo Quadrado Médio de Tratamento (QMTRAT) e Quadrado Médio de Resíduos (QMRES), sendo a definição de Quadrado Médio a divisão da Soma de Quadrados pelos seus graus de liberdade. Por conta da suposição de Normalidade dos erros no modelo, a estatística do teste, F, tem distribuição F de Snedecor com $(a-1)$ e $(\sum_{i=1}^{a} n_i - a)$ graus de liberdade.

$$F_{obs} = \frac{QMTRAT}{QMRES} = \frac{\frac{SQTRAT}{(a-1)}}{\frac{SQRES}{(\sum_{i=1}^{a} n_i - a)}}$$

A hipótese nula é rejeitada caso o p-valor seja menor que o nível de significância pré-fixado. A tabela abaixo resume as informações anteriores:

Fonte de Variação	Graus de Liberdade	Soma de Quadrados	Quadrado Médio	Estatística F	P-valor
Tratamento	$(a-1)$	SQTRAT	$\frac{SQTRAT}{(a-1)}$	$\frac{QMTRAT}{QMRES}$	$P(F > F_{obs})$
Resíduos	$(\sum_{i=1}^{a} n_i - a)$	SQRES	$\frac{SQRES}{(\sum_{i=1}^{a} n_i - a)}$		
Total	$(\sum_{i=1}^{a} n_i - 1)$	SQTOT			

2.14 Análise de Regressão Linear

A análise de regressão é um instrumento eficaz para verificar a relação entre uma variável resposta quantitativa e uma ou mais variáveis explicativas, as quais podem ser tanto qualitativas quanto quantitativas. Essa análise é feita por meio do estudo de uma função de regressão entre as variáveis estudadas. A equação abaixo exemplifica como essa função pode ser escrita:

$$Y = \alpha + \beta X + \varepsilon$$

Esta equação mostra a regressão linear simples. Nela, é evidenciado o comportamento de uma variável dependente ou resposta Y em função de uma variável X, chamada de variável independente ou explicativa. O termo β indica o quanto espera-se que Y varie se X tiver um acréscimo de uma unidade e o coeficiente α mostra o

valor esperado da variável Y se X fosse nulo. Além disso, o termo ε indica o erro aleatório associado à equação em estudo.

Uma generalização do modelo de regressão simples é o modelo de regressão múltipla, no qual são consideradas mais de uma variável independente na equação. Dessa forma, a função será dada por:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

Os coeficientes são interpretados de maneira semelhante: β_0 indica o valor esperado de Y se todas as variáveis X_i $(i = 1, 2, ..., k)$ forem nulas; β_i mostra a variação esperada de Y para um aumento de uma unidade na variável X_i quando todas as outras variáveis são mantidas constantes; e ε informa o erro aleatório associado à equação em estudo.

É necessário assumir as seguintes suposições para o modelo:

- Os erros seguem distribuição normal com média igual a zero
- A variância do erros é constante
- Os erros são independentes

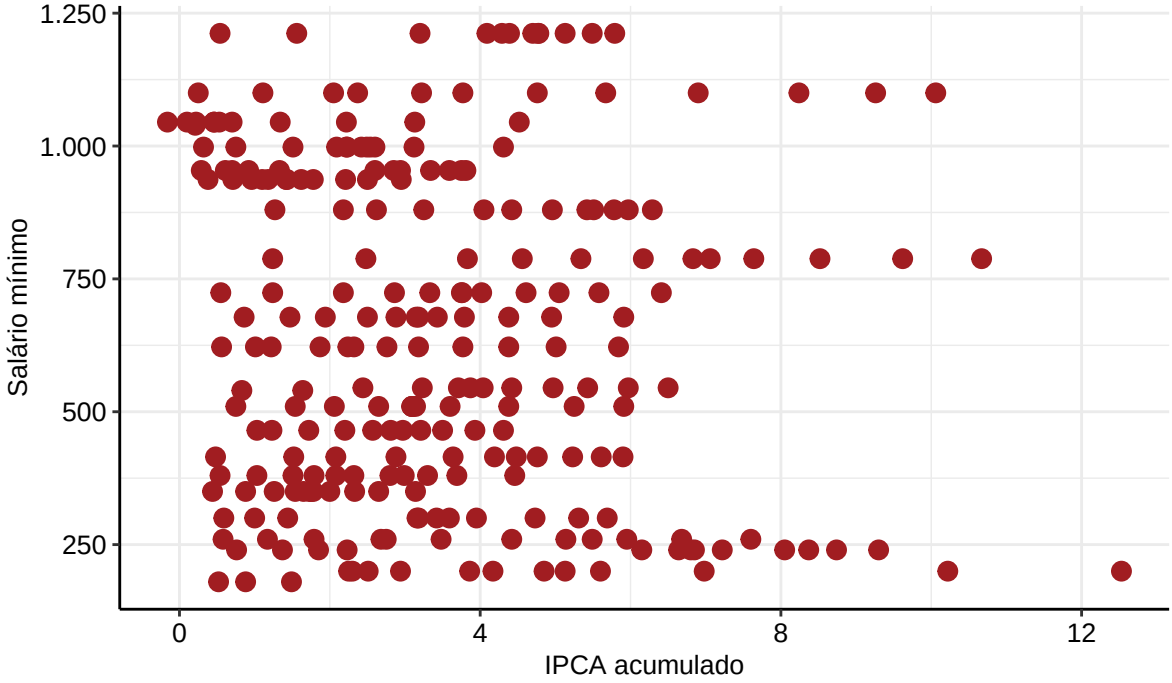
3 Análises

3.1 Análise 1

Para esse estudo será observado o impacto da inflação acumulada ao ano(IPCA) no salário mínimo entre 2002 e 2022, que são variáveis quantitativas contínuas, e veremos se eles se relacionam.

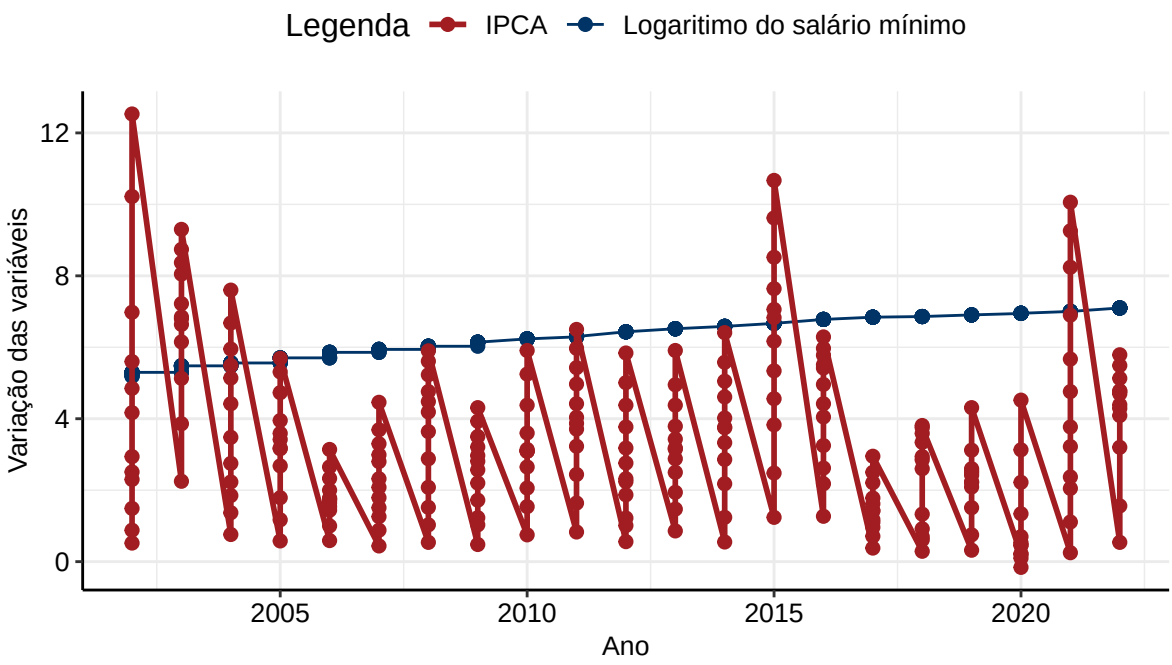
Primeiro, é utilizado um gráfico de dispersão conjunta para observar o comportamento individual do salário mínimo e da inflação para observar se os resíduos do gráfico seguirão algum padrão.

Figura 1: Gráfico de dispersão Salário mínimo x IPCA acumulado



Observa-se no gráfico, que os resíduos de salário mínimo e IPCA parecem não seguir um padrão, o que significaria independência entre as variáveis. A fim de confirmar essa hipótese, é construído um gráfico de linhas utilizando a transformação logaritimica no salário mínimo(aproximando os valores com o IPCA) para observar o comportamento das variáveis ao longo do tempo:

Figura 2: Salário mínimo e IPCA ao longo do tempo



No gráfico, é possível ver que o salário mínimo não possuí grandes variações, enquanto o IPCA varia bastante. O que, deve significar que o comportamento de uma não afeta a outra.

Para comprovar as conclusões iniciais a respeito das interação, observaremos os coeficientes dos testes de correlação de Pearson e Spearman para a hipótese nula de não correlação entre as variáveis e 95% de confiança. Os coeficientes são valores entre -1 e 1, onde um valor próximo de 0 significa não correlação, um valor próximo de -1 significa forte correlação inversamente proporcional e um valor próximo de 1 significa forte correlação diretamente proporcional. A diferença entre os testes é que enquanto o de Pearson assume que variáveis têm uma distribuição normal e que a relação é linear, Spearman avalia a relação monotônica sem assumir normalidade.

Os coeficientes obtidos nos testes foram:

Tabela 1: Testes de hipóteses prara correlação

Testes	Coeficiente obtido
Pearson	-0,1072948
Spearman	-0,1067281

Ambos os coeficientes obtidos estão próximos de -0.1, o que significa que a correlação entre eles é muito fraca e portanto não se pode dizer que há uma correlação linear.

Após as análises, tanto os gráficos, quanto os teste levam a concluir que as variáveis salário mínimo e inflação não estão relacionadas. O que significa que o aumento da

inflação não impacta em mudanças salário mínimo.

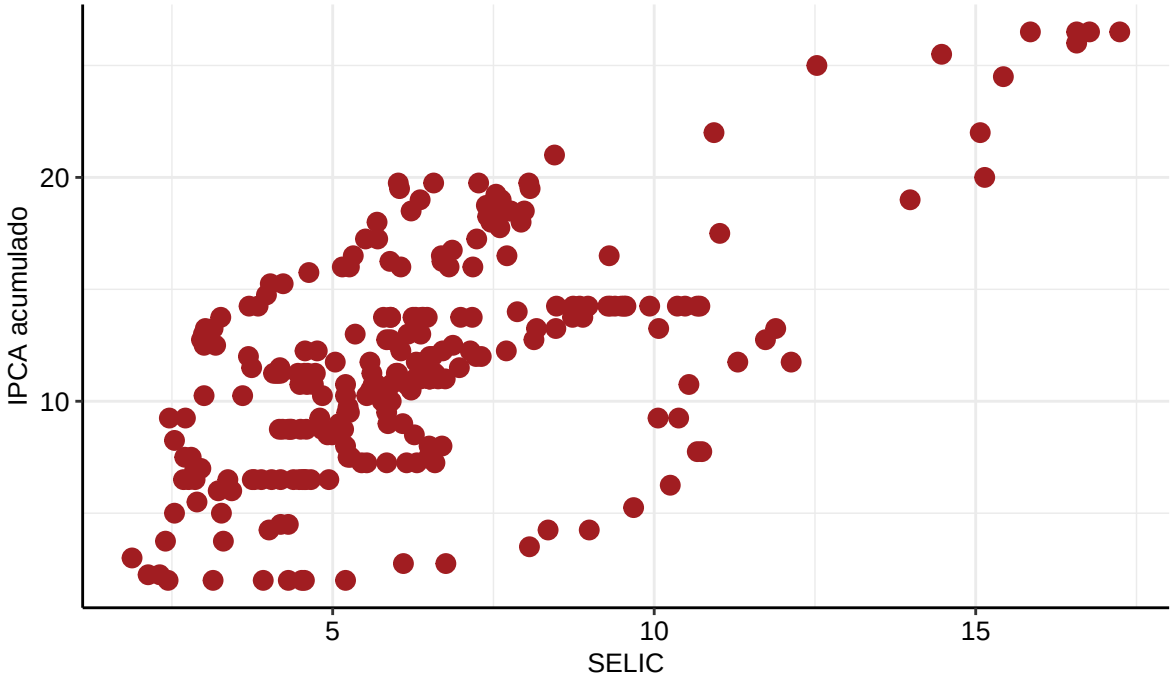
Este resultado indica que o aumento da inflação resulta na perda de poder de compra do consumidor, já que seu salário não deverá sofrer alterações e a moeda perdera valor. E portanto, o salário mínimo não deve ser usado para explicar a inflação.

3.2 Análise 2

O objetivo desse estudo é explorar relações heterogêneas entre taxa de juros e o imposto. para isso serão utilizados dados do IPCA (Índice Nacional de Preços ao Consumidor Amplo) e da taxa Selic (taxa básica de juros da economia). Que são variáveis quantitativas contínuas.

Para iniciar a análise das variáveis, pode-se analisar um gráfico de dispersão conjunta usando as transformações logarítimicas das variáveis, por causa da diferença entre os valores de ambas. Então, observar-se a aleátoriedade dos resíduosnno gráfico para concluir se são relacionadas.

Figura 3: Gráfico de dispersão IPCA acmlado x taxa Selic



Os resíduos no gráfico parecem estar seguindo um padrão, onde o valor do IPCA acumulado cresce junto com a taxa Selic. O que leva a conclusão de que não seriam independentes e crescem .

Para comprovar as conclusões iniciais a respeito das interação, são observados os coeficientes dos testes de correlação de Pearson e Spearman para a hipótese nula de não correlação entre as variáveis e 95% de confiança, como foi feito na análise 1.

Os coeficientes obtidos nos testes foram:

Tabela 2: Testes de hipótese

Testes	Coeficiente obtido
Pearson	0,6353118
Spearman	0,5743377

Os coeficientes encontrados foram próximos de 0,6 , o que significa que a meta Selic e o IPCA possuem uma correlação direta moderada. O aumento de uma variável deve influenciar a outra não na mesma proporção.

Para colaborar com a análise, realiza-se um modelo de regressão utilizando o IPCA como variável resposta e a meta da taxa Selic como variável explicativa. As medidas resumo que encontramos são:

Tabela 3: Tabela ANOVA da regressão

	Estimativa	Erro padrão	Valor t	P-Valor
Intercepto	2,12891	0,35705	5,963	0,00000000842
meta Selic	0,36145	0,02779	13,008	0,0000000000000002

Tabela 4: Medidas resumo da regressão

R2	R2 ajustado	p-valor
0,4036	0,4012	0,00000000000000022

Os p-valores da tabela indicam que a meta Selic é significativa para o IPCA. Já na segunda, o p-valor baixo indica que o modelo é estatisticamente significativo e o R2 indica que 40% da variabilidade é explicada pelo modelo.

Para verificar que este resultado é valido, o modelo deve ter normalidade, resíduos aleatórios e homoscedasticidade. Para testar a normalidade, faz-se um gráfico quantil-quantil com envelope para analisar os resíduos studentizados sob os quantis da normal e o teste de Jarque-Bera para 0,95 de confiança e é rejeitada a normalidade se os p-valores forem menores que 0,05. Para correlação dos resíduos, realiza-se o teste de Durbin-Watson e um p-valor menor que 0,05 indicará autocorrelação nos resíduos. Com respeito à homoscedasticidade, analisa um gráficos de disperção do valor ajustado e da covariável pelo resíduo studentzado e o teste de Goldfeld-Quandt para 0,95 de confiança e um p-valor maior que 0,05 indicará homoscedasticidade.

Figura 4: Gráfico quantil-quantil

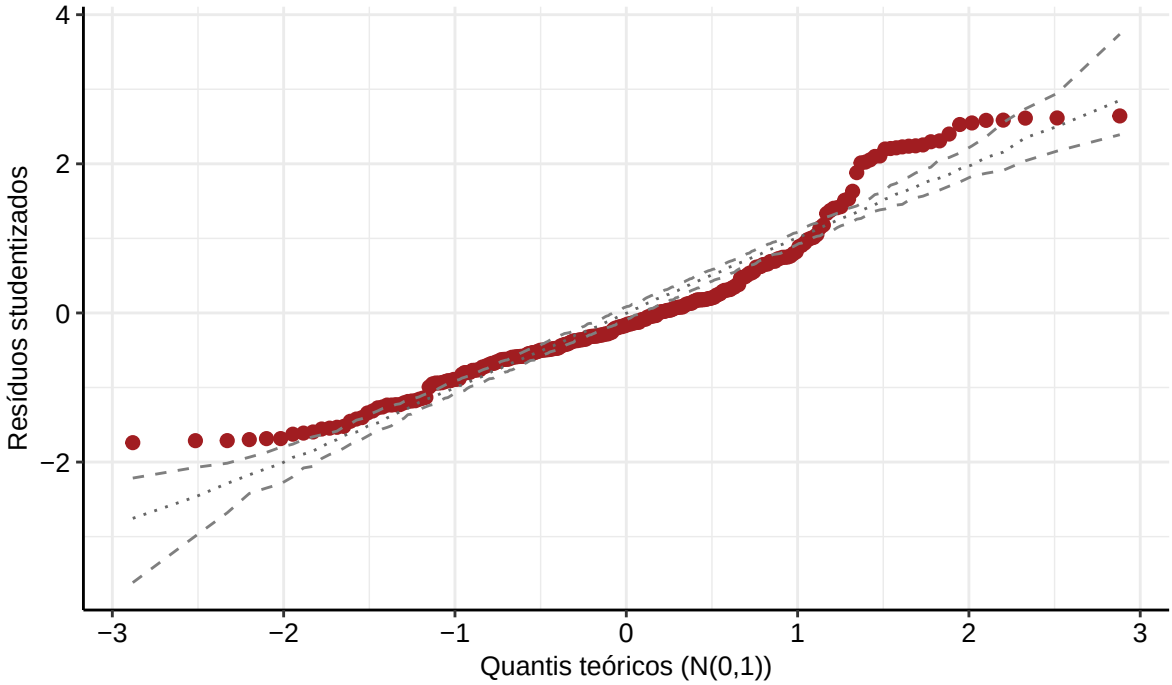
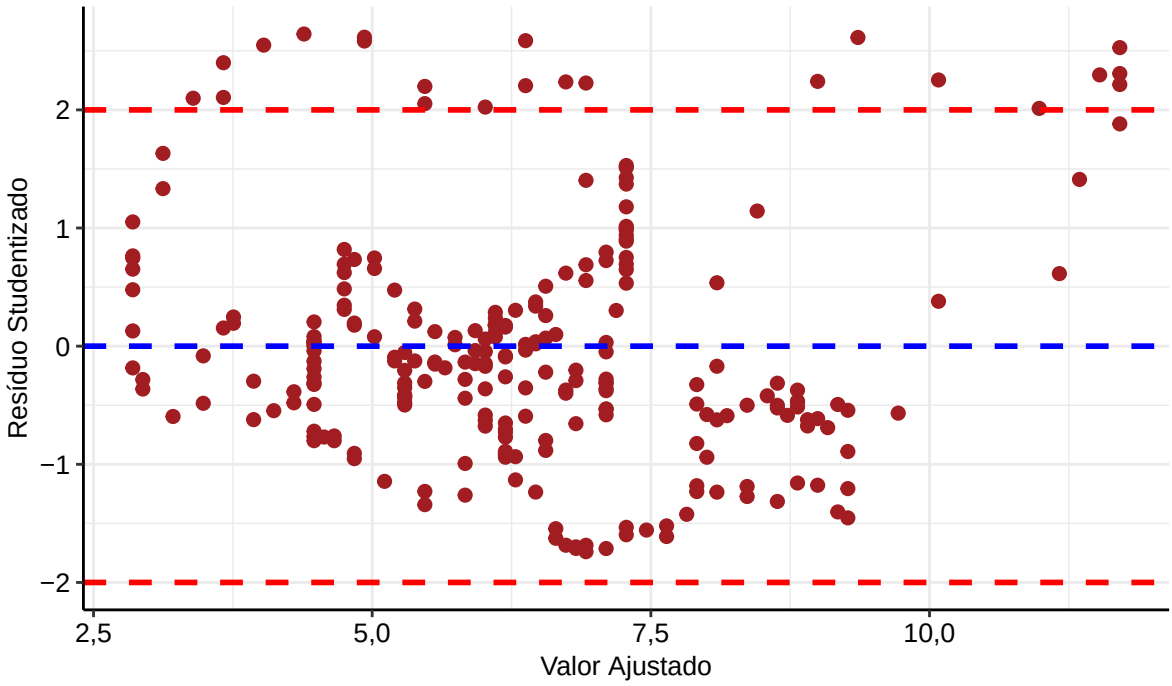


Figura 5: Gráfico de dispersão: Resíduos Studentizados vs Valores Ajustados



Pelo gráfico quantil-quantil, é possível identificar um padrão, já que o esperado era que os resíduos estivessem completamente aleatório e o p-valor de Jarque-Bera confirma a rejeição da normalidade. O p-valor de Durbin-Watson indica autocorrelação nos resíduos e o gráfico de dispersão mostra que a quantidade de resíduos aumenta

Tabela 5: Testes das suposições

Teste	P-valor
Jarque-Bera	0,000000246
Durbin-Watson	0,00000000000000022
Goldfeld-Quandt	0,1573

para os valores ajustados médios e o p-valor de Goldfeld-Quandt rejeita homoscedasticidade. Assim, tendo rejeitado os pressupostos, modelo não é significativo.

Tanto os resultados obtidos pelos gráficos, quanto pelos coeficientes dão indícios de que os dados do IPCA e da taxa Selic são relacionados, apesar de o modelo de regressão proposto não ter sido significativo.

Portanto, conclui-se que um movimento de aumento ou queda da taxa de juros, deve causar o mesmo movimento na inflação, mas sem a mesma proporção. Ou seja, o juros pode ser usado para explicar a inflação, mas não deve ser o principal fator.

3.3 Análise 3

Este estudo tem como objetivo, analisar o INCC (Índice Nacional de Custo da Construção) entre os anos de 2002 e 2022 a fim de ver como se ela se comporta ao longo do tempo. O INCC é uma variável quantitaiva conínua.

Para analisar este comportamento, observa-se um gráfico de barras ao longo dos anos, boxplot por ano e algumas medidas resumo:

Figura 6: INCC ao longo do tempo

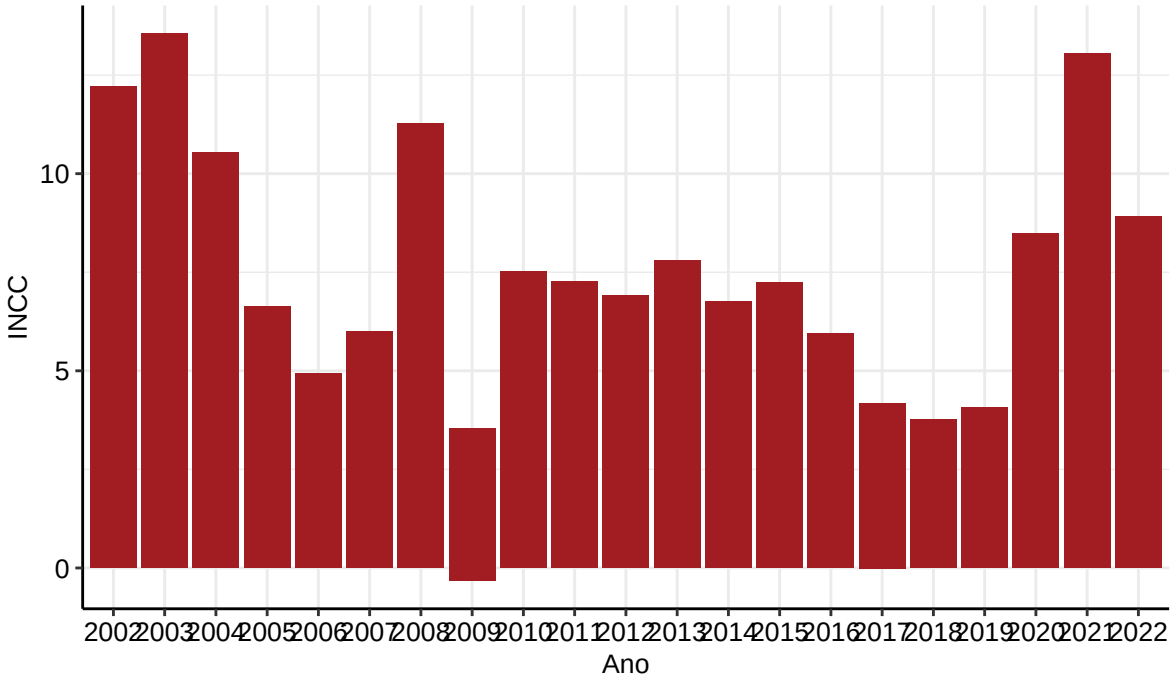


Figura 7: Boxplot do INCC por ano

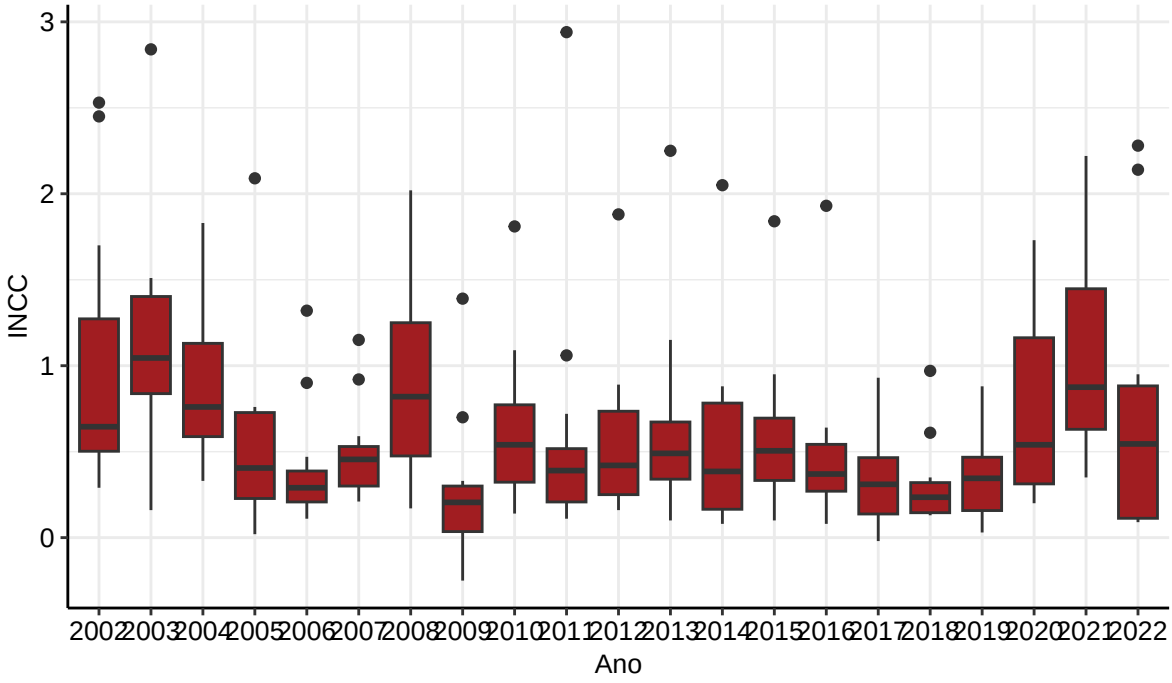


Tabela 6: tabela para medidas resumo

Medida	Valor
Média	0,6362698
Mediana	0,455
Variância	0,3220267
Desvio padrão	0,5674739
Mínimo	-0,25
Máximo	2,94

Nesses gráficos é possível ver que o INCC varia muito entre os anos e também dentro de um mesmo ano. Os pontos que se destacam são os picos em 2002, 2003, 2004 e 2008, o baixo valor de 2009 e volta a subir em 2020 e 2021 provavelmente por conta da pandemia de COVID-19 os preços de construção aumentaram.

Pelas medidas resumo, percebe-se que tanto os altos valores de variância e desvio padrão comparados com a média, quanto a diferença entre os valores extremos, indicam que o INCC varia muito entre os anos.

4 Conclusão

Após terem sido analisados o impacto da inflação no salário mínimo, a relação entre taxa de juros e inflação anual e a distribuição do INCC por ano, conclui-se que:

O salário mínimo não pode ser usado para explicar o comportamento da inflação, ou seja o IPCA não tem impacto sobre o salário mínimo. Enquanto a taxa de juros tem uma correlação direta moderada com a inflação anual, e portanto pode ser usada para ajudar a explicar o comportamento da inflação.

Sobre a distribuição do INCC por ano, foi encontrado que há muita variação entre os anos, que provavelmene devem ser causados or motivos externos que não foram estudados neste relatório, como foi teorizado à respeito da pandemia de 2020.