FUNDAÇÃO GETÚLIO VARGAS
ESCOLA BRASILEIRA DE ECONOMIA E FINANÇAS

TRABALHO FINAL ESTATÍSTICA: ANÁLISE DE ECONOMIAS DE AGLOMERAÇÃO

OTÁVIO BOPP
LEONARDO GROSSMAN

PROFESSOR: BRUNO BARSANETTI

RIO DE JANEIRO
2024

SUMÁRIO:

CONTEXTO:

A população e a produção econômica são bastante concentradas espacialmente. Algumas poucas cidades concentram uma fração significativa da renda e da produção. Um exemplo marcante é a região metropolitana de São Paulo, que abriga 10% da população brasileira e é responsável por quase 17% do PIB do país. A principal explicação para essa alta concentração é a presença de **economias de aglomeração**: mercados de trabalho mais populosos tendem a ser locais onde os trabalhadores são mais produtivos e, consequentemente, recebem salários mais altos. Esse fenômeno cria um ciclo virtuoso: cidades maiores levam a salários mais altos, que atraem mais trabalhadores, resultando em um aumento adicional nos salários. As economias de aglomeração têm diversas causas. Uma delas é a possibilidade de maior especialização, o que melhora o funcionamento dos mercados de trabalho, especialmente para trabalhadores especializados. É importante ressaltar que essas economias de aglomeração são **estáticas**, pois descrevem uma relação contemporânea (no mesmo período) entre o tamanho do mercado e os salários. Estudos empíricos confirmam a existência dessas economias de aglomeração estáticas, demonstrando uma relação positiva entre o tamanho de um mercado de trabalho local e a produtividade dos trabalhadores. O número de trabalhadores (ou seu logaritmo) é uma medida eficaz do tamanho do mercado de trabalho, enquanto a renda é uma boa medida da produtividade, dado que, em mercados competitivos, o salário tende a se igualar ao produto marginal do trabalho.

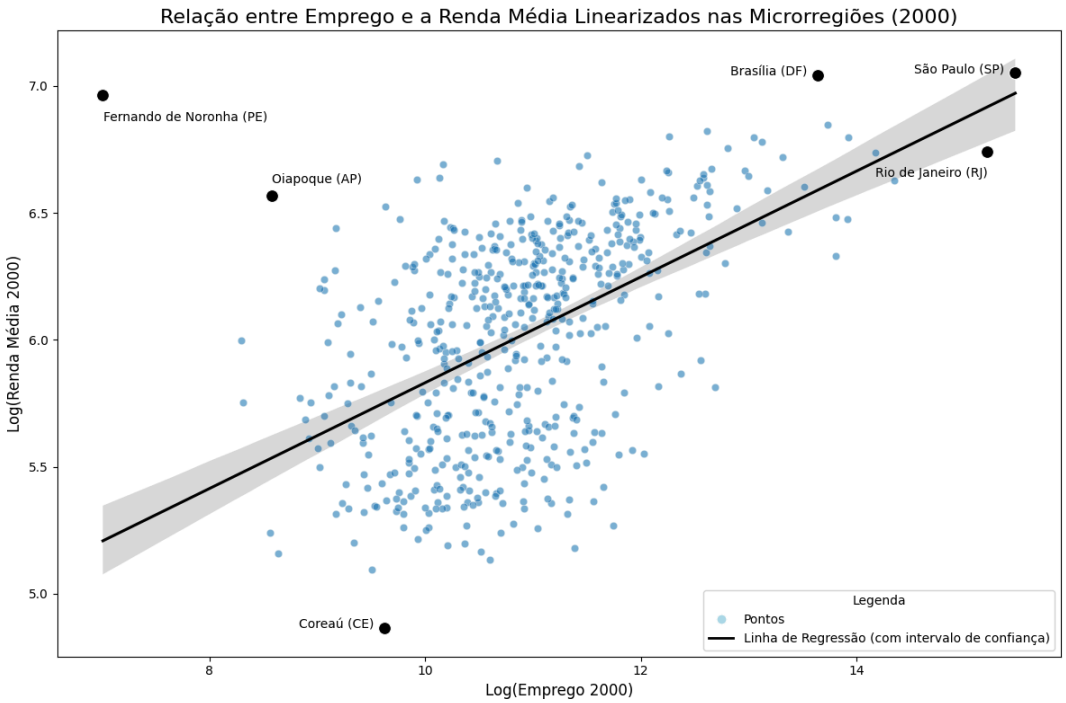
INTRODUÇÃO:

Em primeiro momento, foi feito o download dos dados no ambiente de programação de escolha (Google Colab) e uma breve análise da base de dados à procura de informações faltantes ou outros problemas comuns. Como os dados estavam em perfeito estado, foi dado início ao trabalho de fato. O link do colab é https://colab.research.google.com/drive/1TqYz4n8_7HhG5TCb2udkYZdop48m94HQ?usp=sharing ou https://tinyurl.com/trabalhoestat

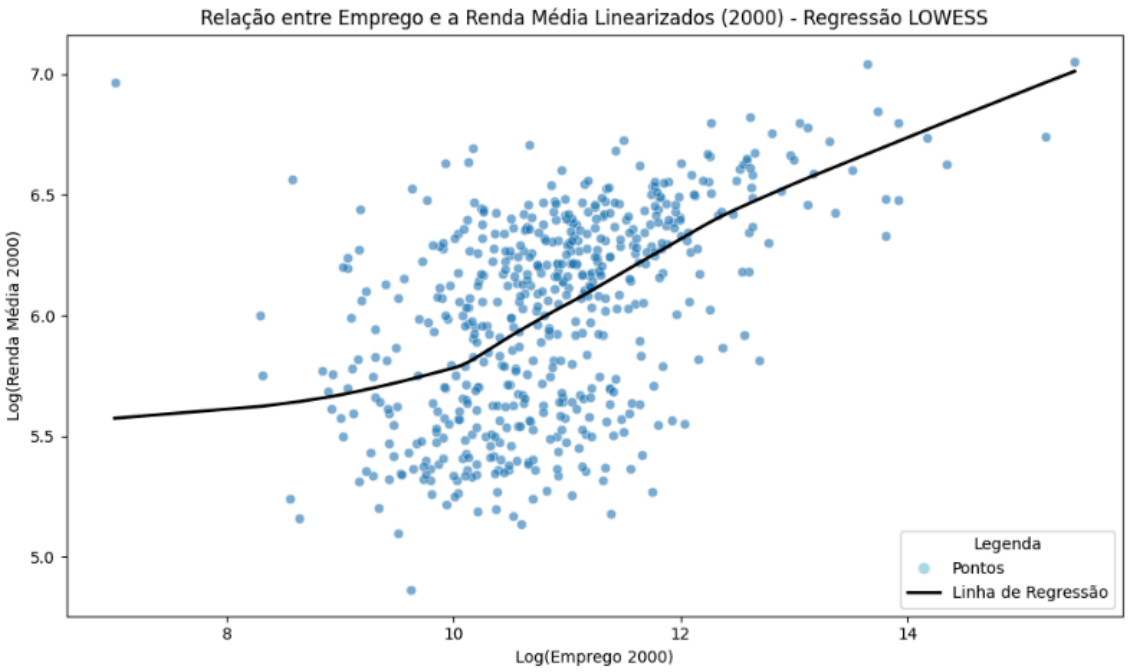
1. RELAÇÃO ENTRE EMPREGOS E RENDA MÉDIA NOS ANOS 2000

Após transformar os dados para o formato linearizado (log-linearizado), foi possível notar uma forte relação entre ambas as variáveis em questão para os dados dos anos 2000. Para isso foram usados gráficos tipo scatterplot, que são excelentes para mostrar tendências, reconhecer padrões, identificar outliers, ao mesmo tempo em que são intuitivos e fáceis de interpretar. A seguir estão os gráficos feitos com esses objetivos em mente, além das orientações do artigo seminal de Jonathan Schwabish para confecção de gráficos limpos e intuitivos:

- **Gráfico tipo scatterplot com regressão linear simples:**



- **Gráfico tipo scatterplot com regressão Lowess:**



- **Gráfico tipo scatterplot interativo:**

Nosso trabalho inclui uma série de gráficos interativos. Eles funcionam em computadores ou dispositivos móveis. Porém, no caso da visualização com um celular, é importante deixá-lo no modo horizontal. Você pode dar zoom selecionando uma área para dar zoom nela, e dar um clique/toque duplo para voltar. Você também pode clicar na legenda para remover elementos do gráfico.

Links para acesso ao gráfico e QR CODE: https://tinyurl.com/interativo2000 ou https://ottoboop.github.io/EstatisticaPopRenda/grafico_interativo2000.html

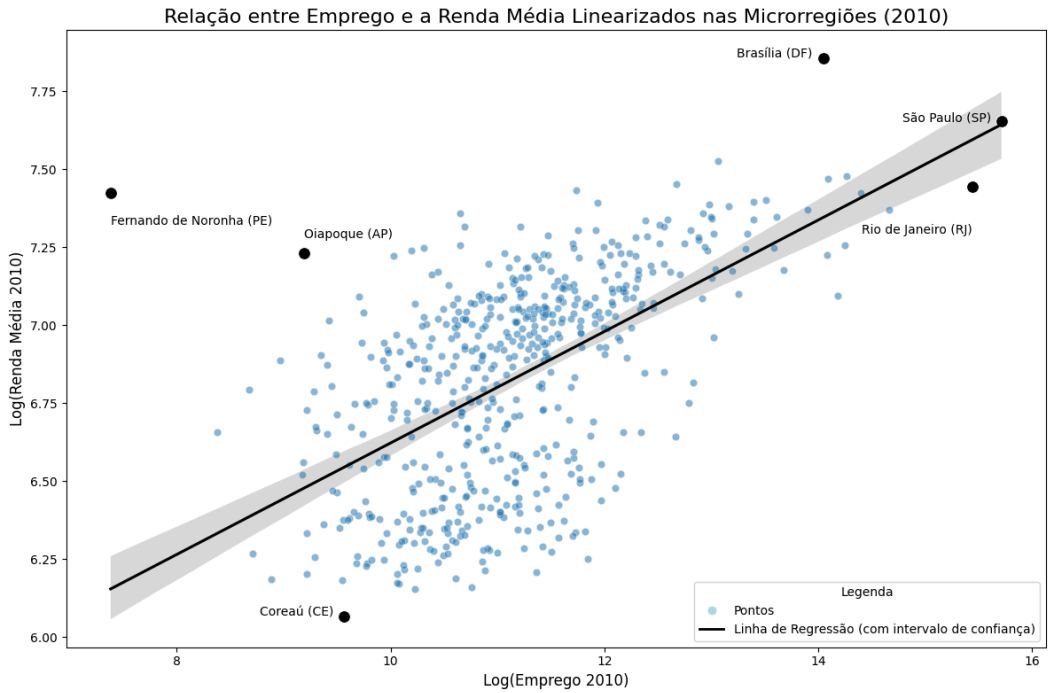
Conclusões:

Por meio das regressões iniciais, é possível notar uma tendência crescente entre o Emprego e a Renda Média no ano 2000, o que será avaliado mais profundamente ao longo do trabalho.

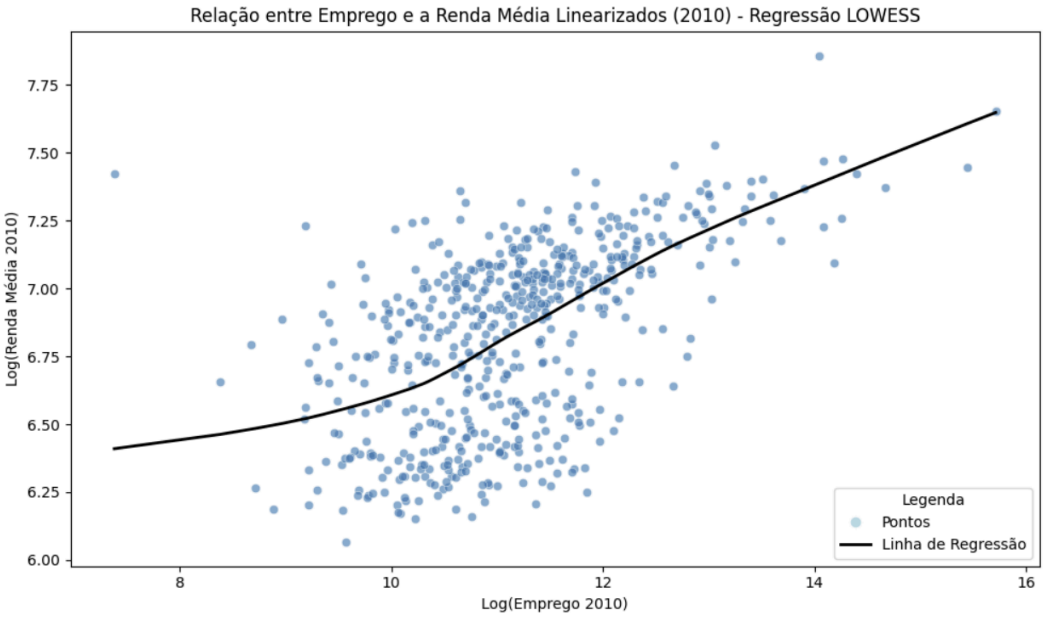
2. RELAÇÃO ENTRE EMPREGOS E RENDA MÉDIA NO ANO 2010

Assim como anteriormente, diferentes gráficos serão apresentados a fim de mostrar a relação entre o Emprego e a Renda Média - agora durante o ano de 2010

- **Gráfico tipo scatterplot com regressão linear simples:**



- **Gráfico tipo scatterplot com regressão Lowess:**



- **Gráfico tipo scatterplot interativo:**

Link para acesso ao gráfico e QR CODE:https://tinyurl.com/interativo2010 ou *https://ottoboop.github.io/EstatisticaPopRenda/grafico_interativo2010.html*

- **Conclusões:**

Novamente, nota-se uma tendência crescente entre o Emprego e a Renda Média no ano 2010. Podemos notar visualmente que os pontos parecem estar mais próximos das linhas de tendência do que os gráficos de 2000, indicando uma possível intensificação dos efeitos de aglomeração.

3. CÁLCULO DA CORRELAÇÃO

Em seguida, foi calculada a correlação entre ambas as variáveis. A correlação é um valor entre -1 e 1 que mostra a relação entre duas variáveis – caso seja um valor positivo, elas variam na mesma direção; caso seja um valor negativo, elas variam em direções opostas. Quanto mais próximo dos extremos (1 e -1), mais forte é a relação entre as duas variáveis.

Para o ano 2000, foi encontrado um valor de **0.51**, que mostra uma correlação forte positiva entre o logaritmo do emprego e o logaritmo da renda média nas microrregiões. Isso indica que microrregiões com um maior número de empregos tendem a apresentar rendas médias mais elevadas, apoiando a hipótese de ganhos de aglomeração.

Já para o ano de 2010, foi encontrado **0.56**, o que confirma a intuição anterior dos valores estarem seguindo um pouco mais a tendência do que antes. Sendo uma correlação forte e positiva, isso sugere que, ao longo da década, a relação entre o tamanho do mercado de trabalho e a renda média se intensificou nas microrregiões analisadas.

Assim, os resultados indicam que há uma relação positiva e robusta entre o tamanho do mercado de trabalho (emprego) e a renda média nas microrregiões tanto em 2000 quanto em 2010. Além disso, a elevação do coeficiente de correlação de **0.51** para **0.56** demonstra um aumento das forças de aglomeração no país.

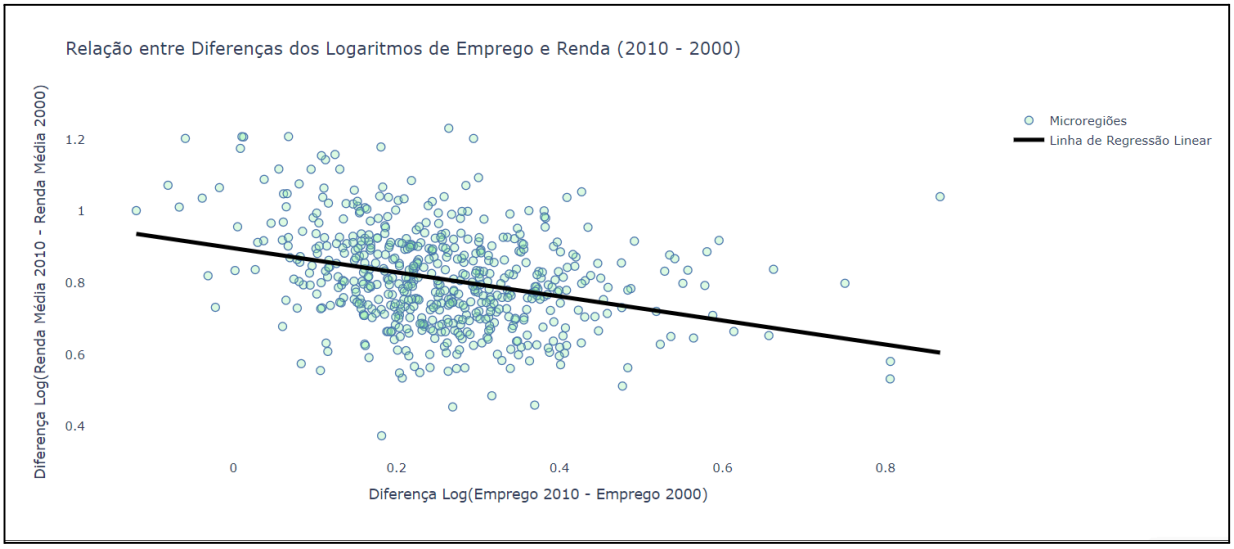
4. CORRELAÇÃO ENTRE AS DIFERENÇAS

A correlação entre duas variáveis nem sempre é um bom indicativo de uma relação causal entre elas, devido à possível presença de outros fatores que impactariam essa interação. Cidades com mais empregos podem simplesmente ter melhores recursos que atraem mais pessoas, e um aumento na população pode ter um efeito nulo, ou até negativo, na renda média. Nesse sentido, o método de diferenças visa eliminar o impacto desses outros fatores na análise da correlação, excluindo o viés das características que se mantiveram constantes ao longo do período de análise. O viés não é completamente eliminado, mas as diferenças entre uma mesma microrregião após 10 anos tendem a ser menores que as diferenças entre regiões.

Se os ganhos de aglomeração de fato existem, esperamos que as cidades onde a população empregada aumentou entre 2000 e 2010 também tivessem um aumento de sua renda per capita.

Calculamos as diferenças entre o log das rendas e dos empregos entre 2000 e 2010 e calculamos a correlação entre as diferenças. Contrariando o resultado anteriormente encontrado, a correlação foi negativa e significante, num valor de **-0.31**. Isto implica que um aumento na população entre 2000 e 2010 resultou em uma queda na renda das cidades, contradizendo a hipótese inicial, de que uma maior população estaria relacionado com maior renda. Ademais, a correlação contradiz o que foi encontrado na etapa anterior, onde a correlação entre população e renda foi ainda maior em 2010.

- **Gráfico tipo scatterplot com regressão linear simples:**



- **Gráfico tipo scatterplot interativo:**

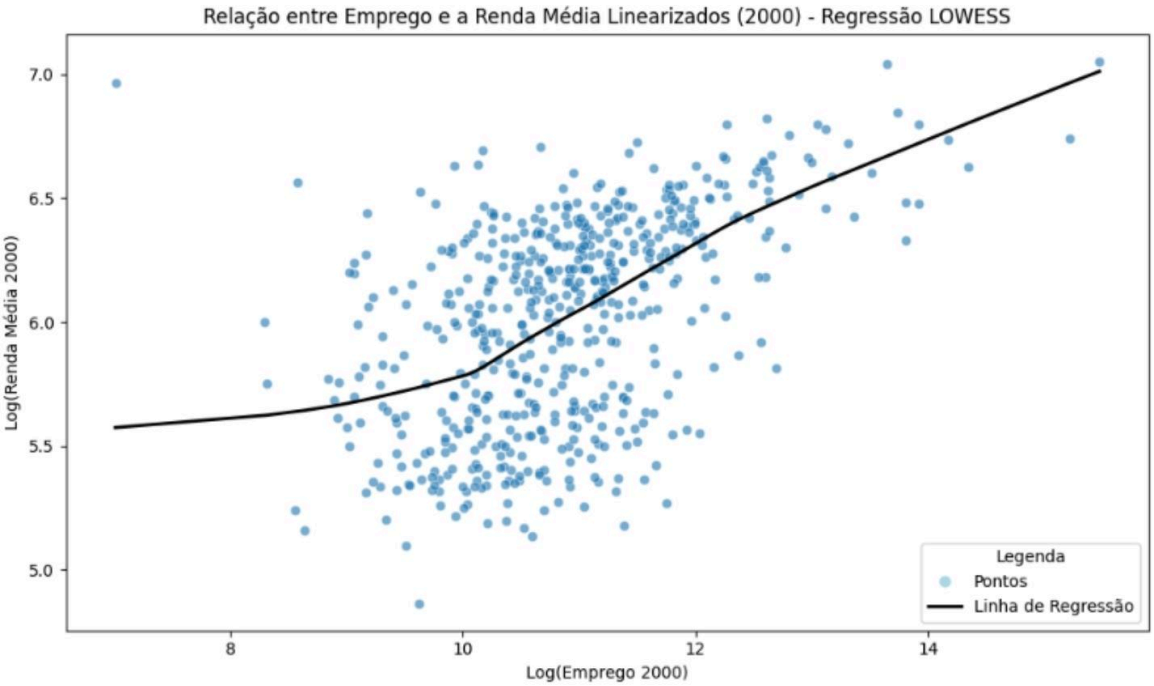
Link para acesso ao gráfico e QR CODE: https://tinyurl.com/diferencasempregorenda ou https://ottoboop.github.io/EstatisticaPopRenda/grafico_interativo_diferencas.html

5. CORRELAÇÃO ENTRE AS DIFERENÇAS

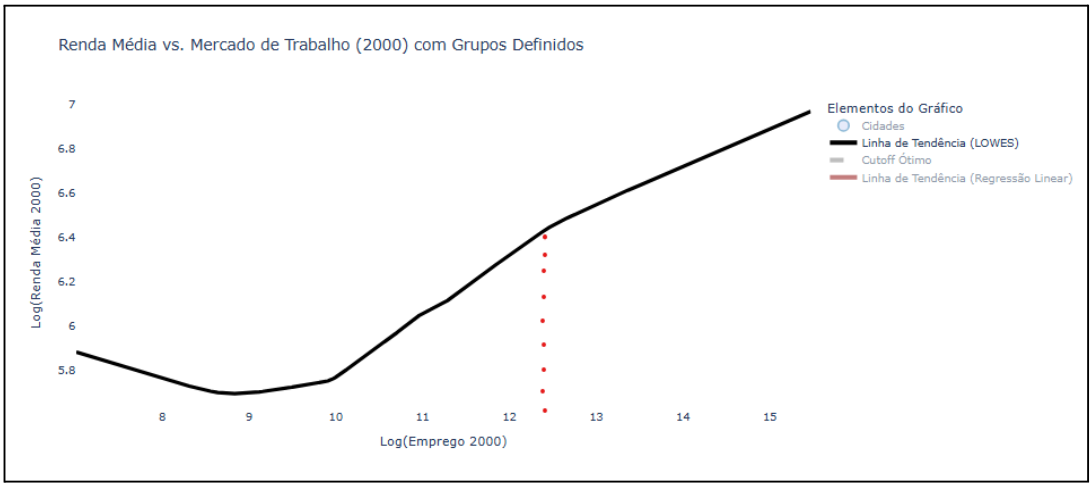
Em seguida, foi feito um teste de diferenças de média usando os dados linearizados. Para o cálculo, apesar de ser uma aproximação, foi utilizada como medida de cutoff o número de empregados a fim de representar a população.

a) Escolha do Cutoff

O cutoff entre os grupos de maior e menor população foi feito baseado em quatro pontos: a análise visual do gráfico, a presença de um variação perceptível na inclinação da regressão LOWESS em determinado ponto, os desvios padrões da média e o ponto em que a correlação entre as diferenças das duas variáveis mudava de negativa para positiva (utilizando o resultado do passo anterior como base. Além desses métodos, foi feito testes usando a média e a mediana como medidas de cutoff, o que levaram a resultados decepcionantes no sentido da hipótese inicial.



Por meio da observação do gráfico, podemos ver que os pontos ficam mais próximos da linha de regressão, tanto da regressão linear quanto da LOWESS. Porém, quando analisamos apenas a regressão LOWESS, podemos ver uma mudança em seu comportamento próximo ao ponto **12.5**. Esta mudança pode indicar um diferente comportamento dos pontos neste valor.

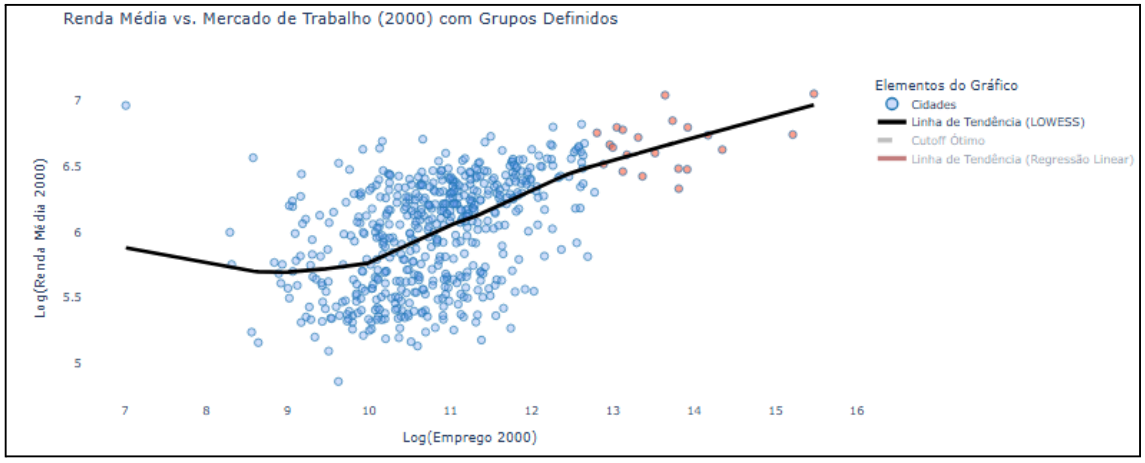


A regressão LOWESS (Locally Weighted Scatterplot Smoothing) é uma técnica de suavização não paramétrica utilizada para identificar tendências em conjuntos de dados sem assumir uma forma funcional específica para a relação entre as variáveis. Ela funciona ajustando modelos de regressão locais em diferentes pontos dos dados, ponderando mais fortemente os pontos próximos à região de interesse. Como essa regressão se ajusta aos pontos mais próximos, e não simplesmente à tendência completa dos dados, uma mudança em seu formato indica uma mudança no comportamento dos dados.

Já analisando a média e os desvios padrões de emprego, vemos que a média de população fica em aproximadamente em **10.87** no ano 2000, e um desvio padrão é aproximadamente **1.02** pontos. Então, no ponto **12.92** fica a dois desvios padrões da média, ou aproximadamente no quantil **0.967**

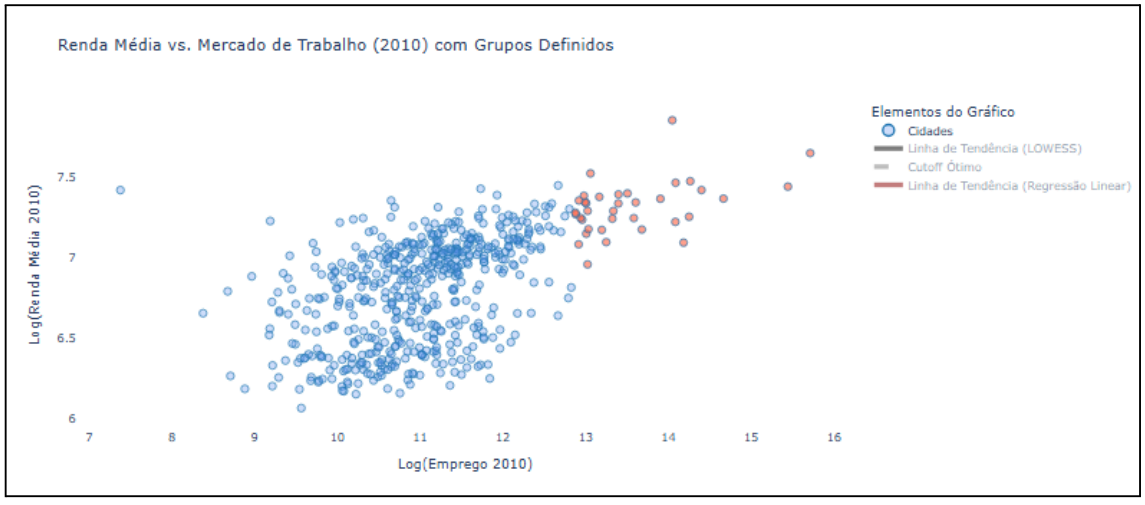
Para definir o ponto final da divisão dos dois grupos, repetimos o teste da questão 4 com os grupos separados. Aproximadamente, no log-emprego **12.6**, o grupo “maior” de cidades em 2000 passa a ter correlação positiva entre o aumento de empregados entre 2000 e 2010 e o aumento de renda no mesmo período. Quando elevamos o cutoff para **12.7**, a correlação se torna **+0.415** ao invés do **-0.31** encontrado em 4. Este foi o critério final para

definir o cutoff. O grupo “maior” possui correlação positiva forte entre o aumento da população e o aumento de sua renda.



Links para acesso ao gráfico e QR CODE: https://tinyurl.com/grupos2000
https://ottoboop.github.io/EstatisticaPopRenda/grafico_grupos_2000.html

Adotamos o cutoff de 12.85 para as cidades do tipo maior em 2010 pois isto gera uma correlação de **+0.315** da renda com a mudança no número de empregados. Com esse valor de corte, 21 cidades acabaram no grupo “maior” em 2000 e 36 cidades em 2010.



Links para acesso ao gráfico e QR CODE: https://tinyurl.com/grupos2010 ou
https://ottoboop.github.io/EstatisticaPopRenda/grafico_grupos_2010.html

b) Resultado do Teste para ano 2000

Após a escolha do cutoff, o Teste T de Welch foi feito para analisar a média. Este é utilizado para comparar as médias de dois grupos distintos, dentro de um conjunto de amostras independentes, e determinar se há uma diferença significativa entre elas. O Teste T de Welch, diferente de outro teste similar, o T de Student, assume que as variâncias são distintas, o que foi verificado após a escolha do cutoff. Para o grupo maior, foi encontrado **0.0398** e para o grupo menor, **0.1655**.

O resultado gerado pelo T de Student foi um **valor p de 9.654824e-15** que prova que os dois grupos são distintos entre si, e o cutoff usado é válido. Lembrando que é comum rejeitar hipóteses nulas quanto o valor p é inferior a 0.05

c) Resultado do Teste para ano 2010

Assim como anteriormente, o teste T de Welch foi utilizado para analisar a validade do cutoff escolhido. Analisando as variâncias, para o grupo maior foi encontrado um valor de **0.03985**, e para o grupo menor um valor de **0.1023**.

Dessa vez, o resultado gerado foi um **valor p de 1.256911e-23**, que novamente prova a validade do cutoff e a diferença entre os dois grupos.

CONCLUSÃO:

Durante o trabalho, foi feita uma análise profunda do problema de microrregiões no país. A hipótese considerada inicialmente foi a de que cidades com maior número de empregos (ou uma população maior) teriam os salários mais altos, não somente porque cidades com melhor padrão de vida atraem mais pessoas, mas também porque a proximidade de profissionais gera mais produtividade e, portanto, maiores salários médios.

Como primeiro passo, criamos os logaritmos dos empregos e salários e fizemos algumas regressões, optando pela regressão linear e a lowess. Plotamos uma scatterplot, tanto

para os anos 2000 quanto para 2010, com ambas regressões, e podemos notar que o aumento do número de empregados numa microrregião estava relacionado com o aumento da renda média.

Em seguida, a correlação entre as duas variáveis foi calculada, o que confirmou novamente a hipótese inicial. Encontramos uma relação positiva e relativamente forte para ambos os anos, sendo as duas maiores que 0.5, o que demonstra um valor explicativo significativo. E o aumento dessa correlação entre 2000 e 2010 indicou que essas forças se intensificaram durante a década.

Apesar disso, após calcular as diferenças entre a população e a renda das microrregiões, a correlação entre estas variáveis foi negativa e significativa, -0,31. Isso indica que um aumento da população empregada resultou numa queda da renda média dentro de uma microrregião, contrariando a tese dos ganhos de aglomeração ajudando no desenvolvimento econômico.

Separamos os dados em dois grupos, um com maior renda e outro com menor renda. Assim, notamos que as microrregiões menores que a média não tinham quaisquer ganhos de aglomeração, e um aumento em sua população poderia ser detrimental para sua renda média. Porém, os ganhos de aglomeração apareciam na medida que as regiões cresciam, de tal forma que em nosso cutoff havia uma correlação de 0.41 entre a mudança na população nas regiões maiores (em 2000) e na renda média. O crescimento populacional estava altamente correlacionado com o aumento da renda local!

Com os grupos definidos, utilizamos um teste de duas médias para definir se existia uma diferença nas distribuições das macrorregiões maiores e menores. Os resultados estão entre o mais próximo de certeza que a estatística pode definir – com um p valor com 23 ou 15 zeros, em casas decimais, rejeitamos a hipótese nula, em 2000 e 2010, respectivamente. Tipicamente rejeitamos uma hipótese nula quando seu p valor é inferior a 0,05. Ao menos entre as grandes microrregiões, os ganhos de aglomeração parecem existir. Porém, este efeito não é visto em regiões menores.