Explorando a Conexão entre Educação e Renda: Análise das Condições Sociais e Padrões Sociais de Bem-Estar

Breno Marques Azevedo

09 de julho de 2024

Estudo de caso da disciplina de Modelagem Estatística, submetido como trabalho final da matéria válida pelo curso de Ciência de Dados & Inteligência Artificial da EMAp.
Professor: Luiz Max Carvalho

Resumo

O objetivo deste trabalho é analisar a relação entre escolaridade e remuneração, a fim de explorar o potencial da educação como ferramenta de ascensão social em sociedades com altos níveis de desigualdade social. Para este fim, é realizada uma breve revisão de literatura, seguida de uma análise dos dados com o objetivo de confirmar suspeitas da importância dos indicadores disponíveis e, por fim, serão exploradas técnicas de modelagem estatística, como aplicação de modelos lineares generalizados e diferentes métricas de avaliação da capacidade explicativa e estimativa dos modelos.
Palavras-chave: Educação; Remuneração; Modelos Lineares Generalizados

Conteúdo

Introdução

Nas mais diversas sociedades ao redor do mundo, incluindo o Brasil, é possível observar dados preocupantes que apontam para uma grande desigualdade social. É de enorme interesse dos pesquisadores, considerando o ramo das ciências sociais, compreender quais fatores implicam a heterogeneidade dos aspectos sociais e econômicos e é de interesse da população, principalmente indivíduos em situação de maior vulnerabilidade, entender por quais meios é possível ascender socialmente em uma realidade tão díspar. Por conseguinte, há uma robustez de estudos que visam entender a relação entre fatores como educação, gênero, segurança social e salário[6]. Nesse sentido, nos vemos compelidos a questionar: o nível educacional impacta significativamente a remuneração de um indivíduo?

Este trabalho investiga essa questão considerando o contexto social dos dados da pesquisa "Work, Family, and Well-being in the United States", conduzida por Catherine E. Ross em 1990[5]. A base de dados foi escolhida por conta de sua presença nas lições dos livros "Regression and Other Stories"[3] e "Data Analysis Using Regression and Multilevel/Hierarchical Models"[2], que foram fundamentais ao longo deste curso.

Este trabalho será estruturado da seguinte maneira: primeiramente, uma breve revisão bibliográfica sobre o uso de índices educacionais para mapear a remuneração de indivíduos, considerando o contexto americano. Em seguida, descreverei o tratamento e a análise exploratória da base de dados. Por fim, apresentarei uma explicação sobre os métodos utilizados para ajustar os dados usando modelos lineares generalizados, além de como avaliá-los. Também discutirei os resultados obtidos, avaliando se a pesquisa foi frutífera e alinhada com as sugestões da literatura.

1 Revisão de Literatura

Considerando o contexto americano, a literatura faz diversas sugestões para o porquê observa-se uma relação entre educação e o salário, como a valorização de habilidades vocacionais desenvolvidas ao longo do desenvolvimento acadêmico; manutenção das classes sociais vigentes; e, até mesmo, indicador de aptidão[1]. Note que essas sugestões não são necessariamente excludentes entre si e, também, um empregador pode recompensar um funcionário com maior nível educacional não somente por suas habilidades desenvolvidas ao longo da especialização.

Iniciativa, autonomia e cumprimento de metas são valorizadas tanto na academia quanto no mercado de trabalho. A educação, portanto, não só fornece habilidades práticas, mas também socializa indivíduos com valores e comportamentos valorizados[6], além de funcionar como um filtro para identificar talentos e motivações. Ademais, outros fatores que caracterizam um individuo, como gênero e empregabilidade formal costumam estar relacionados com a remuneração.

A fim de evidenciar a relação observada na literatura, serão utilizados os dados obtidos pela pesquisa supracitada. Os *scripts* de exploração, análise, modelagem e ajuste dos dados pode ser encontrados neste repositório no GitHub.

2 Base de Dados

Os dados que serão analisados foram coletados em uma pesquisa realizada em 1990, cujo propósito era examinar os efeitos de condições sociais na saúde mental e física e

investigar os efeitos nos padrões sociais de bem-estar nos Estados Unidos. A base de dados bruta contém 2029 entradas e dispõe de dados que vão desde características fenotípicas até sociais e financeiras. Contudo, este conjunto apresenta uma série de problemas, como valores faltantes ou sem sentido. Consequentemente, fez-se necessária uma verificação da distribuição dos dados, com o objetivo de tratá-los.

2.1 Tratamento dos Dados

Nesse sentido, houve a remoção de valores implausíveis, como altura em polegadas superiores a 11 e altura em pés superiores ou iguais a 7. Da mesma forma, pesos superiores a 500 libras são eliminados.

As variáveis categóricas como sexo e etnia são reinterpretadas para facilitar a análise. Por exemplo, o sexo é codificado para que 1 represente masculino e 0 represente feminino. A etnia é determinada com base nas variáveis de raça e hispanidade. Sobre os níveis de educação, por simplicidade, 18 representa o nível máximo de educação formal, representando por exemplo doutorado.

Além disso, a altura total é calculada combinando os valores de `height_feet` e `height_inches` em uma única variável, simplificando a análise. A renda aproximada `earn_approx` é criada a partir de `earn2` e combinada com `earn_exact` para garantir que os dados de renda estejam disponíveis, mesmo quando os valores exatos não estão presentes. O cálculo da idade é feito a partir do ano de nascimento, normalizando os dados temporais e facilitando a análise demográfica.

As variáveis como fumo, tensão e raiva são limpas substituindo valores fora dos intervalos válidos por NA, garantindo que apenas dados confiáveis sejam utilizados nas análises. O ajuste do ano de nascimento, baseado em `yearbn`, corrige possíveis erros de codificação, garantindo que a idade calculada seja precisa e útil para análises temporais.

Por fim, os dados são agrupados em um único conjunto e linhas com valores ausentes para renda e educação são removidos, assim, assegurando um conjunto de dados completo e consistente, adequado para a análise estatística que virá a seguir. As variáveis dispostas são:

- Renda: Renda exata (variável resposta)
- Altura: Polegadas.
- Peso: Libras.
- Sexo: Gênero do indivíduo.
- Raça e Etnia: Raça e indicação se a pessoa é hispânica.
- Educação: Níveis de educação do indivíduo, da mãe e do pai.
- Atividade Física: Frequência de caminhada e exercício.
- Hábitos de Fumo: Se a pessoa fuma atualmente.
- Estado Emocional: Níveis de tensão e raiva.
- Ano de Nascimento: Usado para calcular a idade

2.2 Análise Exploratória e Visualizações

Neste momento, após um tratamento inicial dos dados e anterior à modelagem em si, faz-se necessária uma exploração dos dados disponíveis, a fim de ampliar a compreensão entre a variável resposta e as possíveis covariáveis. Esta é uma ótima oportunidade

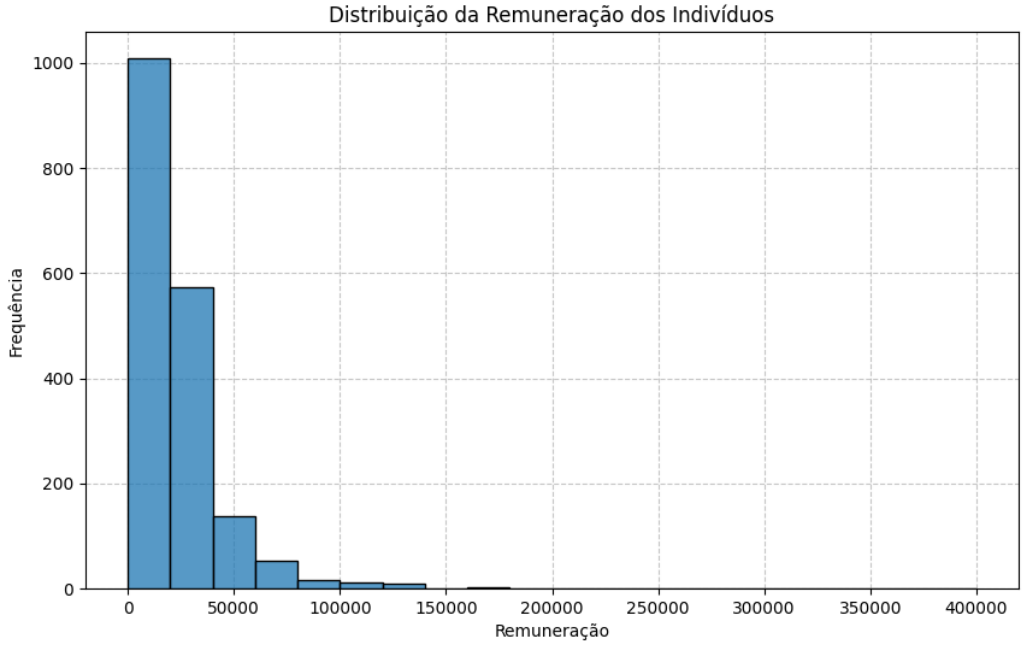

Figura 1: Histograma da variável de remuneração (earn)

também de verificar se o conjunto de dados parece estar de acordo com o que é observado na literatura. Sendo assim, vamos começar observando a distribuição da remuneração.

Ainda que de forma simples, a Figura 1 permite observar uma característica comum de sociedades com altos índices de desigualdade social, a concentração assimétrica de renda em uma pequena parcela da população, enquanto a esmagadora maioria possui salários mais baixos.

A literatura destaca a importância de fatores como educação, gênero e trabalho formal para a remuneração[6]. Infelizmente não há dados de sobre empregabilidade formal, mas podemos observar a distribuições das demais variáveis disponíveis, como educação e gênero. A variável de educação representa o nível de escolaridade dos indivíduos em um formato numérico, no qual cada valor corresponde a um número específico de anos de escolaridade completados.

- Os valores entre 1 e 12 geralmente representam o número de anos de educação formal completados, com 12 anos indicando a conclusão do ensino médio.
- Por sua vez, valores de 13 a 16 geralmente representam a educação pós-secundária, onde 13 pode indicar um ano de faculdade e 16 a conclusão de uma graduação (bacharelado).
- Finalmente, valores de 17 a 18 podem representar educação avançada ou pós-graduação. Por simplicidade, são desconsiderados níveis mais elevados de educação posterior.

Observando a Figura 2, é evidente que a grande maioria dos indivíduos concluiu o ciclo de educação básico, com relativamente poucos valores abaixo de 12 anos de escolaridade, o que sugere uma baixa evasão escolar, especialmente considerando que a maioria dos

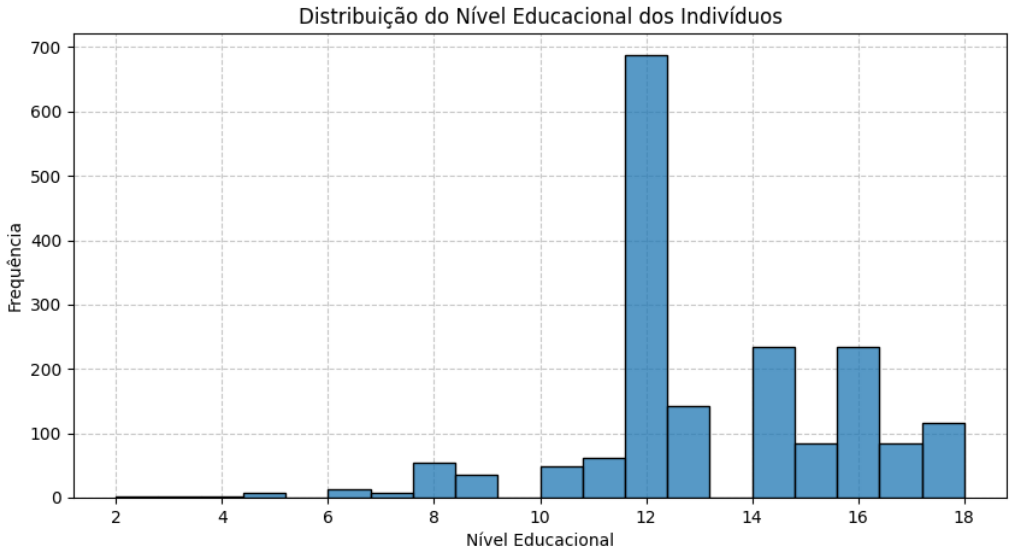

Figura 2: Histograma da variável de educação (education)

entrevistados nasceu antes da década de 1970. Além disso, há uma presença notável de pessoas com nível de ensino superior, o que pode indicar a possibilidade de maiores salários.

Ademais, a fim de testar minhas suposições iniciais, é necessário observar o comportamento das variáveis que suspeito estarem relacionadas com o salário. Como mencionado na literatura, educação e gênero costumam ser indicadores determinantes de bem-estar social [6], ainda mais considerando sociedades desiguais. Sendo assim, vamos observar este comportamento.

A Figura 3 relaciona gênero e a remuneração. Por simplicidade, o gênero é tratado como variável binária, considerando somente o masculino e o feminino. Apesar do aparente equilíbrio entre as remunerações, é possível notar uma cota superior para o salário feminino, enquanto o salário de homens apresenta maior variabilidade e pode chegar a mais que o dobro do salário de mulheres. Isso pode ser um bom indicativo de que há uma relação com a remuneração, assim, fazendo com que gênero seja um candidato à covariável nas modelagens posteriores.

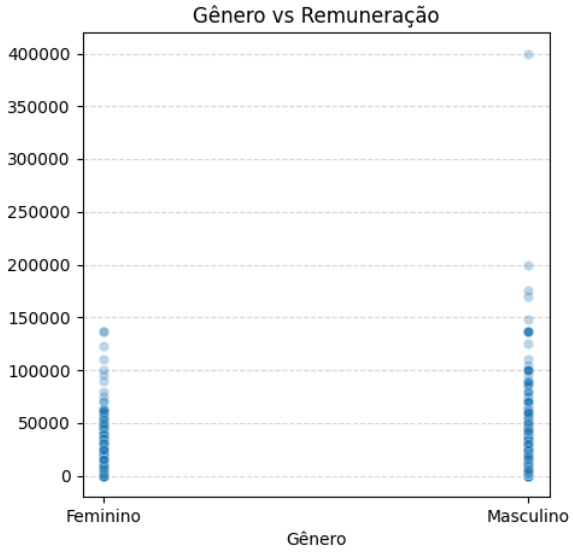

Figura 3: Relação entre gênero e salário

Por sua vez, a Figura 4 mostra que há uma tendência geral de aumento da remuneração à medida que o nível educacional aumenta. Isso sugere que, em média, pessoas com níveis educacionais mais altos estavam recebendo salários maiores. A variabilidade dos salários também aumenta com o nível educacional. Enquanto os níveis educacionais mais baixos

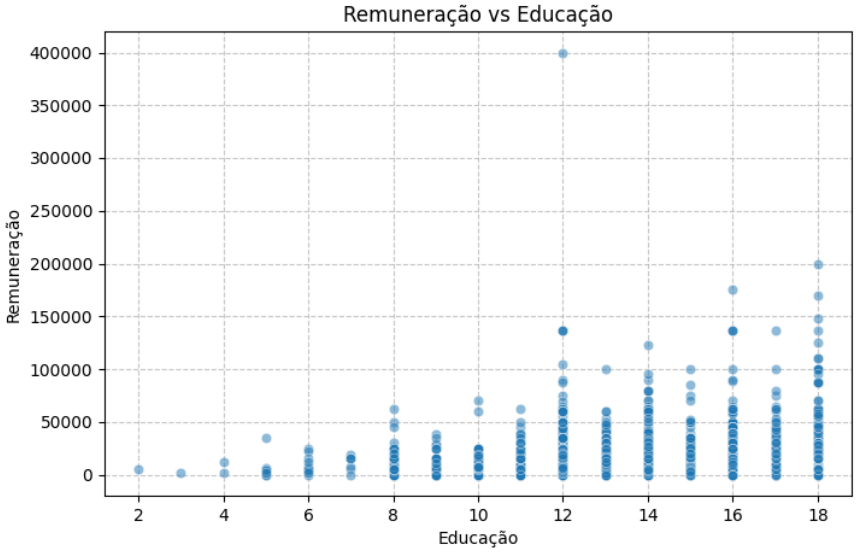

Figura 4: Relação entre educação e salário

apresentam uma faixa de salários mais restrita, os níveis educacionais mais altos, a partir dos 12 anos (ensino médio completo) mostram uma maior dispersão dos salários. Note que a investigação foi produtiva, uma vez que parece estar alinhada com o que é indicado na literatura: a educação como uma forma de ascensão social.

Por fim, é interessante também observar a correlação linear entre as variáveis disponíveis no conjunto. A figura 5, por meio da correlação de Pearson, representa uma uma matriz de correlação filtrada das covariáveis de um conjunto de dados, mantendo apenas aquelas que são maiores que 0.3 ou menores que -0.3, excluindo relações lineares fracas. Por meio da matriz, podemos observar uma relação moderada entre educação (`education`), gênero (`male`) e a variável resposta salário (`earn`).

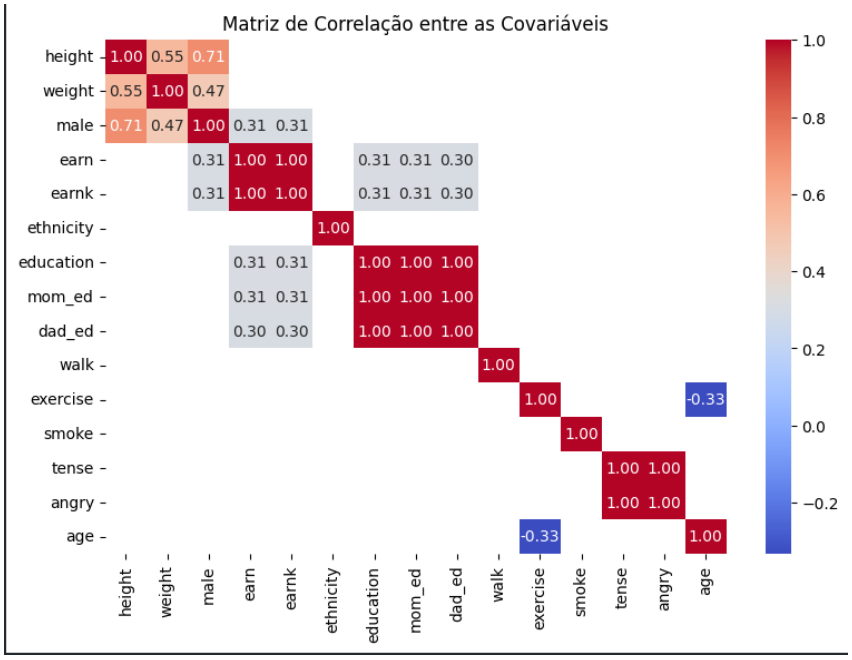

Figura 5: Correlação entre as variáveis disponíveis

3 Metodologia

Como foi possível observar na análise realizada na seção anterior, a distribuição dos dados de salário é contínua, positiva, distorcida para a direita e com aspectos assimétricos (ver Figura 1). Consequentemente, é importante que a modelagem seja condizente com as características da variável resposta, a fim evitar resultados escalafobéticos. Nesse sentido, os Modelos Lineares Generalizados (GLMs) oferecem uma abordagem flexível e robusta.

Os GLMs permitem a escolha de uma distribuição que melhor se ajusta às características dos dados. No caso da remuneração, que é contínua e positiva, a distribuição Gama é particularmente adequada, uma vez que ela é ideal para modelar dados que são assimétricos e distorcidos para a direita. Isso ocorre porque a distribuição Gama pode capturar a variabilidade e a assimetria observadas nos dados de salário.

Além disso, ao usar a função de ligação logarítmica, $\log(\mu)$, o GLM transforma a média da variável resposta em uma escala que pode lidar com a assimetria dos dados, permitindo que os efeitos dos preditores sejam modelados de maneira multiplicativa. Isso é útil, pois aumentos proporcionais nos salários são mais naturais de interpretar do que aumentos absolutos.

3.1 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados [4](GLMs) são uma extensão dos modelos lineares tradicionais que permitem a modelagem de variáveis de resposta que têm distribuições de diferentes tipos, como binária, contagem, contínua, etc. Um GLM é definido por três componentes principais:

- **Distribuição da Família Exponencial:** variável resposta Y pertence a uma família de distribuições da família exponencial.
- **Função de Ligação (Link Function):** Uma função que conecta a média da variável resposta μ com a combinação linear dos preditores η.
- **Preditores Lineares:** Uma combinação linear dos preditores, $\eta = X\beta$, na qual X são os preditores e β são os coeficientes a serem estimados.

3.1.1 Distribuição Gama

A distribuição Gama é usada quando a variável resposta é contínua e positiva, sendo adequada para modelar dados como tempo até evento, custos, ou outras variáveis onde a média é proporcional à variância. A função de densidade de probabilidade para a distribuição Gama é:

$$f(y; \alpha, \beta) = \frac{\beta^{\alpha} y^{\alpha-1} e^{-\beta y}}{\Gamma(\alpha)} \tag{1}$$

onde:

- $y > 0$
- β é o parâmetro de escala.
- α é o parâmetro de forma.
- $\Gamma(\alpha)$ é a função gama.

Para GLMs, a parametrização é feita usando a média μ e a variância ϕ.

3.1.2 Função de Ligação Logarítmica

A função de ligação mais comum para a distribuição Gama é a ligação logarítmica:

$$g(\mu) = \log(\mu) \tag{2}$$

onde μ é a média da variável resposta Y. Essa escolha garante a positividade das predições da média μ, o que é crucial para o contexto de modelagem de salários. Além disso, há muitos casos em que a relação entre os preditores e a variável resposta á multiplicativa. Ao aplicar a função logarítmica, transformamos uma relação multiplicativa em uma relação aditiva, assim, contribuindo para simplicidade da interpretação.

3.2 Modelagem

O GLM com distribuição Gama e ligação logarítmica pode ser escrito como:

$$\eta = X\beta = \log(\mu)$$

Portanto,

$$\mu = e^{X\beta}$$

A função de verossimilhança L para a distribuição Gama é:

$$\log L(\beta, \phi; y, X) = \sum_{i=1}^{n} \left(\frac{1}{\phi} \left(\log(\mu_i) - \frac{y_i}{\mu_i} \right) - \log(y_i) - \log(\Gamma(\alpha)) + \alpha \log(\beta) - \beta y_i \right)$$

onde $\alpha = \frac{1}{\phi}$ e $\beta = \frac{1}{\mu}$.

Os parâmetros β são estimados usando o método Mínimos Quadrados Ponderados Iterativamente. A log-verossimilhança é maximizada em relação a β e ϕ.

3.3 Métodos Numéricos

O método de Mínimos Quadrados Ponderados Iterativamente (IRLS) é amplamente utilizada para estimar os parâmetros de um GLM. Esse método envolve a repetição de uma série de etapas de regressão ponderada até que a convergência seja alcançada. A ideia central é reponderar os dados de forma iterativa, ajustando o modelo de maneira mais precisa à distribuição da família exponencial. Entre suas vantagens, destacam-se a eficiência no ajuste de modelos da família exponencial, a rápida convergência para muitos tipos de dados e a implementação relativamente simples e direta, facilitando sua aplicação em diversos contextos.

1. **Inicialização**: Comece com um valor inicial para os parâmetros β, geralmente zero ou pequenas perturbações.

2. **Calculando a Função de Ligação Inversa**: Calcule a estimativa da média μ usando a função de ligação inversa:

$$\mu = g^{-1}(\eta)$$

3. **Calculando os Pesos**: Calcule os pesos W baseados na variância da distribuição da variável resposta:

$$W_i = \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 (\text{Var}(Y_i))^{-1}$$

onde $\text{Var}(Y_i)$ é a variância da variável resposta Y.

4. **Atualizando os Valores**: Atualize os valores ajustados z:

$$z = \eta + (Y - \mu)\left(\frac{\partial \eta}{\partial \mu}\right)$$

5. **Reajuste do Modelo**: Reajuste o modelo linear ponderado usando os novos valores de z e W:

$$\beta = \left(X^T W X\right)^{-1} X^T W z$$

6. **Iteração**: Repita os passos 2 a 5 até que a convergência seja alcançada, ou seja, até que as mudanças nos parâmetros β sejam menores que um determinado limiar.

3.4 Avaliação e Diagnósticos

Os critérios de avaliação de modelos fornecem diferentes perspectivas sobre a qualidade e a adequação de um modelo estatístico aos dados. O coeficiente de determinação, conhecido como R-squared, mede a proporção da variabilidade na variável resposta que é explicada pelo modelo. Um R-squared alto indica que o modelo captura bem a variabilidade dos dados, sendo útil para modelos de regressão linear. Contudo, faremos uso do **Pseudo R-squared**, uma adaptação do R-squared para GLMs. Este critério fornece uma medida da proporção da variabilidade explicada pelo modelo, o que é essencial para avaliar a adequação dos GLMs.

O **Akaike Information Criterion (AIC)** mede a qualidade do ajuste do modelo, penalizando pela complexidade. Um modelo com menor AIC é preferido, pois indica um bom ajuste com menor complexidade. Esse critério é amplamente utilizado tanto para modelos de regressão linear quanto para GLMs, auxiliando na seleção de modelos que balanceiam ajuste e simplicidade. Similar ao AIC, o **Bayesian Information Criterion (BIC)** também mede a qualidade do ajuste com uma penalização pela complexidade, mas de forma mais severa. Modelos com menor BIC são preferidos, e este critério é útil na seleção de modelos, especialmente quando a penalização adicional por complexidade é desejável.

Incerteza nas estimativas refere-se à falta de confiabilidade completa sobre os valores dos parâmetros de um modelo estatístico. Em termos práticos, isso significa que as estimativas dos coeficientes de um modelo, como os parâmetros de regressão, são baseadas em dados amostrais e, portanto, estão sujeitas a variações aleatórias. O erro padrão é uma medida da variabilidade da estimativa de um coeficiente. Em um GLM, cada coeficiente estimado é acompanhado por um erro padrão, que remete à incerteza associada a essa estimativa. Dessa forma, um erro padrão menor indica uma estimativa mais precisa.

4 Resultados

A análise exploratória dos dados evidenciou que as variáveis `education` e `male` estão moderadamente correlacionadas à variável resposta `earn` (Ver Figura 5). Apesar da

correlação não parecer tão intensa, a literatura advoga a favor do uso dessas covariáveis, portanto, elas serão utilizadas nas tentativas de ajuste do modelo. Espera-se encontrar uma relação interessante, a fim de responder à pergunta proposta e explorar as alternativas de ascensão social. Como supracitado, serão ajustados modelos lineares generalizados e verificou-se que o uso da distribuição Gama e função de ligação logarítmica são apropriados para o problema.

4.1 Modelos Ajustados

Neste sentido, primeiramente, vamos ajustar o modelo usando a covariável que indica o nível educacional (`education`). Chamaremos de **modelo 1**. Em seguida, um ajuste usando a variável indicadora de gênero (`male`), que será o **modelo 2**. E, por fim, um ajuste com a interação entre as duas covariáveis, **modelo 3**.

4.1.1 Modelo 1 (Educação)

Inicialmente, a fim de responder à pergunta: "o nível educacional impacta significativamente a remuneração de um indivíduo?", tomamos uma abordagem bastante direta, ajustando um modelo com somente a educação como covariável.

Sumário Simplificado					
	Variável	**Estimação**	**Erro Padrão**	**AIC**	**Pseudo R-Squared**
Modelo 1	Intercepto	8.2655	0.1240	-19579	0.09240
	`education`	0.129	0.010		

Tabela 1: Resultados do Modelo 1

Como podemos ver na Tabela 1, o intercepto, o valor do logaritmo do salário esperado quando o nível educacional é zero, foi 8.2655. Já a educação foi 0.124, este coeficiente indica a mudança no logaritmo do salário esperado para cada unidade adicional de nível educacional.

Ambos os coeficientes têm valor-p muito baixos (< 0.000), indicando que eles são altamente significativos, o que nos leva a crer que o nível educacional tem um impacto significativo no salário. Contudo, o Pseudo R-squared (CS) foi de 0.09240, sugerindo que cerca de 9.24% da variabilidade no salário é explicada pelo nível educacional. Isso sugere que outros fatores podem ser importantes para explicar a variabilidade no salário.

4.1.2 Modelo 2 (Gênero)

Em vista ao que foi indicado pelo modelo 1, ajustamos o **modelo 2** somente usando o gênero como covariável, com o objetivo de inserir considerações sobre desigualdade de gênero na modelagem do salário.

Sumário Simplificado					
	Variável	**Estimação**	**Erro Padrão**	**AIC**	**Pseudo R-Squared**
Modelo 2	Intercepto	9.6706	0.029	-19958	0.1004
	`male`	0.6416	0.047		

Tabela 2: Resultados do Modelo 2

Observa-se na Tabela 2, um intercepto 9.6706, que representa o logaritmo do salário esperado para mulheres (já que a variável `male` é 0 para mulheres). O coeficiente que representa a diferença no logaritmo do salário esperado entre homens e mulheres foi 0.6416. Como é positivo, indica que homens têm um salário maior do que mulheres, em média.

Para interpretar na escala original, podemos tomar a exponencial do coeficiente: $\exp(0.6416) \approx 1.899$. Isso significa que, em média, o salário dos homens é cerca de 89.9% maior do que o das mulheres (Ver Figura 3).

4.1.3 Modelo 3 (Interação)

Por fim, o modelo 3 foi ajustado para incluir uma interação entre as covariáveis educação (`education`) e gênero (`male`), com o objetivo de verificar se um modelo mais complexo pode capturar melhor as propriedades da variável resposta e explorar a relação entre desigualdade de gênero e educação.

Sumário Simplificado					
	Variável	**Estimação**	**Erro Padrão**	**AIC**	**Pseudo R-Squared**
Modelo 3	Intercepto	7.9086	0.152	-19958	0.1991
	`education`	0.1300	0.011		
	`male`	0.8797	0.239		
	`education_male`	-0.0195	0.018		

Tabela 3: Resultados do Modelo 3

Em relação aos coeficientes estimados, o intercepto, que representa o logaritmo do salário esperado para mulheres com nível educacional zero, foi de 7.9086. Já `education`, a mudança no logaritmo do salário esperado para mulheres para cada unidade adicional de nível educacional, é de 0.1300. Enquanto `male`, a diferença no logaritmo do salário esperado entre homens e mulheres quando o nível educacional é zero, foi de 0.8797. Este coeficiente é positivo e significativo, indicando que, em média, homens tendem a ganhar mais do que mulheres com o mesmo nível educacional zero.

Por outro lado, o coeficiente de interação `education_male` representa a diferença na relação entre educação e salário para homens em comparação com mulheres. Este coeficiente não é significativo ($P>|z| = 0.270$), sugerindo que a inclinação da relação entre educação e salário não difere significativamente entre homens e mulheres.

4.2 Discussão

Métricas de Avaliação			
Modelo Ajustado	**Pseudo R-squared**	**AIC**	**BIC**
Modelo 1	0.09240	-19579	-5834.22
Modelo 2	0.1004	-19958	-5825.64
Modelo 3	0.1991	-20029	-5989.94

Tabela 4: Comparação dos modelos

Como podemos observar na Tabela 4, o **Modelo 3** possui o maior Pseudo R-squared (0.1991), indicando que ele explica a maior proporção da variabilidade nos salários. Isso

sugere que a combinação de educação, gênero e a interação entre ambos oferece a melhor explicação para a variabilidade nos salários, apesar da interação não ser significativa.

No que diz respeito ao AIC, o **Modelo 3** apresenta o menor valor (-20029), sugerindo que ele proporciona o melhor equilíbrio entre ajuste e complexidade. Um menor AIC indica que o modelo se ajusta melhor aos dados. Por sua vez BIC, o **Modelo 3** também apresenta o menor valor (-5989.94), indicando que ele é o mais eficiente em termos de penalização pela complexidade do modelo.

4.2.1 Análise da Incerteza nas Estimativas

Contudo, ainda é necessário considerar a incerteza nas estimativas dos modelos ajustados anteriormente. Isso porque é fundamental compreender se um modelo é capaz de refletir a variabilidade inerente aos dados e ao processo de amostragem. Por conseguinte, é do nosso interesse verificar como o **Modelo 3** se comparta com relação a essa métrica, dado que ele foi o que teve mais destaque até o momento.

Modelo	**Coeficiente**	**Erro Padrão**	**Intervalo de Confiança [0.025, 0.975]**
Modelo 3	Intercepto	0.152	[7.611, 8.206]
	`education`	0.011	[0.108, 0.152]
	`male`	0.239	[0.411, 1.349]
	`education_male`	0.018	[-0.054, 0.015]

Tabela 5: Medidas de Incerteza das Variáveis do Modelo 3

Considerando o fato de que o **modelo 3** inclui mais termos que os demais, é de se esperar que haja maior incerteza em suas estimativas, algo que pode ser observado na Tabela 5. Como supracitado, a interação das variáveis não foi significativa e os erros padrão são maiores que nos demais modelos (ver Tabelas 1 e 2), indicando maior variabilidade nas estimativas.

Modelo	**Coeficiente**	**Erro Padrão**	**Intervalo de Confiança [0.025, 0.975]**
Modelo 1	const	0.129	[8.013, 8.518]
	education	0.010	[0.105, 0.143]

Tabela 6: Medidas de Incerteza das Variáveis do Modelo 1

Modelo	**Coeficiente**	**Erro Padrão**	**Intervalo de Confiança [0.025, 0.975]**
Modelo 2	const	0.029	[9.614, 9.727]
	male	0.047	[0.549, 0.734]

Tabela 7: Medidas de Incerteza das Variáveis do Modelo 2

Por sua vez, o **Modelo 2** (apenas educação) oferece a menor incerteza com coeficientes precisos e intervalos de confiança estreitos. Se a inclusão de gênero é importante, o **Modelo 2** também é uma boa escolha devido à baixa incerteza associada ao coeficiente de male.

Por fim, o **Modelo 3**, embora tenha o maior Pseudo R-squared, apresenta maior incerteza e complexidade sem ganho claro de precisão, especialmente considerando que a interação `education_male` não é significativa.

5 Considerações Finais

Em resumo, a exploração dos modelos ajustados ao longo deste trabalho nos permitiu construir um entendimento mais profundo da relação entre remuneração e fatores sociais, frequentemente fora do controle individual. Os modelos, com diferentes graus de sucesso, indicam que uma parte significativa do salário de um indivíduo pode ser explicada pelo seu nível educacional, alinhando-se com as observações da literatura. Além disso, a questão da desigualdade de gênero foi destacada pelos modelos e dados utilizados, revelando que, em média, o salário dos homens é quase o dobro do das mulheres. Este achado destaca a necessidade de refletir sobre as diferenças de acesso a oportunidades de carreira e desenvolvimento entre os gêneros.

Finalmente, este trabalho pode servir como uma base para políticas públicas e programas educacionais que visem promover a equidade de gênero e a mobilidade social. A literatura consistentemente indica uma relação positiva entre educação e salário, sugerindo que investir em educação pode ser uma ferramenta eficaz para a ascensão social. Políticas que incentivem a educação contínua e a igualdade de oportunidades podem ajudar a reduzir disparidades salariais e promover um crescimento econômico inclusivo.

5.1 Limitações

Embora os modelos ajustados forneçam insights valiosos sobre a relação entre educação, gênero e salário, existem algumas limitações a serem consideradas. Primeiramente, a análise foi baseada em um conjunto de dados específico, o que pode limitar a generalização dos resultados. Além disso, a variabilidade não explicada pelos modelos, evidenciada pelos baixos valores de Pseudo R-squared, sugere que outros fatores não incluídos nos modelos podem desempenhar um papel significativo na determinação dos salários.

Em relação à base de dados, ficou evidente durante o desenvolvimento do trabalho o quanto ela é limitada para explorar a relação entre aspectos sociais gerais e o salário. Fatores como experiência de trabalho, indústria de atuação, localização geográfica e características pessoais possivelmente enriqueceriam a análise e os modelos trabalhados, contudo, não foram considerados devido a essas limitações.

Outra restrição está relacionada à interação entre educação e gênero no Modelo 3. Embora este modelo tenha apresentado o maior Pseudo R-squared, a interação não foi significativa, e os erros padrão foram maiores, indicando maior incerteza nas estimativas. Isso sugere que os dados disponíveis não são suficientes para capturar adequadamente essa interação.

5.2 Direções Futuras

Para abordar as limitações mencionadas e expandir a análise, futuras pesquisas podem considerar a inclusão de mais variáveis explicativas para capturar melhor a complexidade da determinação dos salários. Outros aspectos, como experiência de trabalho, tipo de emprego, setor econômico e etnia, podem ser levados em consideração ao analisar o acesso a salários maiores e possibilidades de mobilidade social. Novas pesquisas poderiam explorar a correlação entre esses fatores e a remuneração, bem como revisar a literatura relacionada à temática.

Além disso, a coleta de dados em diferentes contextos geográficos e econômicos pode ajudar a generalizar os resultados e fornecer uma compreensão mais ampla das dinâmicas

entre educação, gênero e salário. Estudos longitudinais, que acompanhem indivíduos ao longo do tempo, também poderiam oferecer insights sobre como essas relações evoluem ao longo da carreira profissional.

6 Bibliografia

[1] Mark Blaug. «The correlation between education and earnings: what does it signify?» Em: *Higher education* 1 (1972), pp. 53–76.

[2] Andrew Gelman e Jennifer Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge university press, 2006.

[3] Andrew Gelman, Jennifer Hill e Aki Vehtari. *Regression and other stories*. Cambridge University Press, 2021.

[4] John Ashworth Nelder e Robert WM Wedderburn. «Generalized linear models». Em: *Journal of the Royal Statistical Society Series A: Statistics in Society* 135.3 (1972), pp. 370–384.

[5] Catherine E. Ross. *WORK, FAMILY, AND WELL-BEING IN THE UNITED STATES, 1990*. Computer file. Ann Arbor, MI: Inter-university Consortium for Political and Social Research distributor, 1996. Champaign, IL, 1995.

[6] Valdenya Pereira da Silva. «Determinantes da desigualdade de renda no Brasil: um estudo econométrico de dados em painel para os estados do país, no período 2012-2020». Tese de mestrado. Universidade Federal do Rio Grande do Norte, 2022.