Entendendo as relações entre indicadores econômicos, sociais e educacionais nos municípios brasileiros

Zuilho Rodrigues Castro Segundo

[1]FGV EMAp

segundozuilho@gmail.com

1. Introdução

O projeto em questão visa investigar os fatores econômicos e sociais que exercem influência sobre o Índice de Desenvolvimento da Educação Básica (IDEB) no Brasil. O IDEB é uma importante métrica que combina informações sobre desempenho em exames padronizados de estudantes (como Prova Brasil) e taxas de aprovação, fornecendo uma avaliação abrangente da qualidade da educação nas escolas brasileiras. Entender esse tipo de relação é importante para conseguir pensar políticas públicas que possam alavancar o ensino brasileiro, principalmente nos Ensino Fundamental e Médio, que são a base da educação.

Os dados uilizados visam abranger essas diferentes área, são públicos e podem ser encontrados na Base dos Dados. Para os dados econômico-sociais, utilizei o Atlas do Desenvolvimento Humano (ADH), que são dados do censo e IDH a nível municipal. Para os dados referentes ao PIB, utilizei a base Produto Interno Bruto do Brasil (PIB), o PIB é um indicador crucial para avaliar o desenvolvimento econômico das regiões e pode influenciar diretamente os recursos disponíveis para investimentos em educação. Já para os indicadores educacionais, utilizei as bases Indicadores Educacionais e Índice de Desenvolvimento da Educação Básica (Ideb), a última sendo a base que servirá como aquilo que queremos prever.

2. Limpeza dos dados e Análise Exploratória

Após avaliar cada uma das tabelas, que possuem centenas de dados, selecionei algumas variáveis que pareciam fazer sentido para o problema que queria responder. No repositório é possível encontrar as colunas que foram mantidas. Em seguida, fiz a junção das tabelas, repetindo os dados do censo (IDH) para os outros anos, já que esses dados são tomados de 10 em 10 anos. Além disso, me desfiz das linhas onde não possuíam a nota do IDEB, e por fim, coloquei a média para as taxas de aprovação e reprovação nos valores faltantes.

Em seguida, comecei uma análise exploratória. Comecei plotando a matriz de correlação:

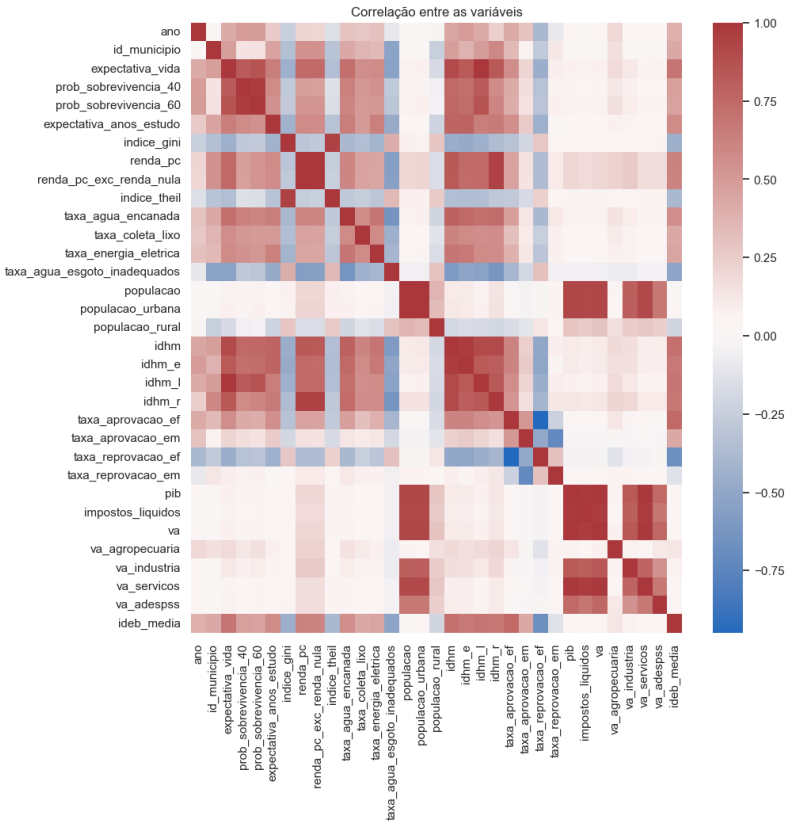

Figura 1. Correlação

Como podemos ver pela última colunas, várias linhas tem correlação com a variável target. Assim, vamos analisar cada uma das correlações.

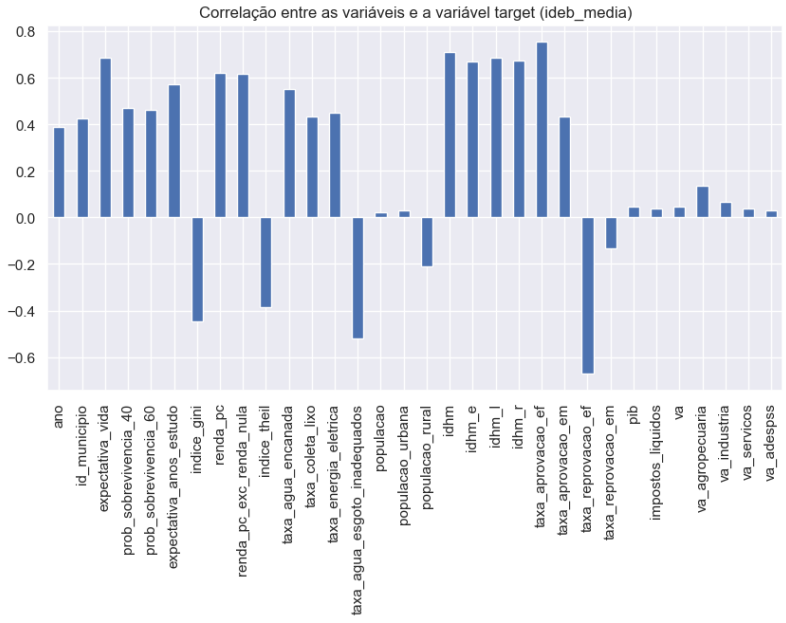

Figura 2. Correlação com a target

A partir daqui, selecionamos apenas as variáveis com correlação cujo módulo é maior ou igual a 4, porém, removendo os valores do IDH e seus subsequentes, que estão altamente relacionados. Uma coisa que é interessante perceber é que os valores de PIB e demais serviços não tem correlação alta, mas a renda per capita sim. Isso acontece pois o PIB é um indicador muito geral e não consegue indicar os valores aplicados por famílias que estarão disponíveis para educação. Por fim, ficamos com 16 variáveis.

Ainda analisando, temos a distribuição do IDEB e os scatters da correlação.

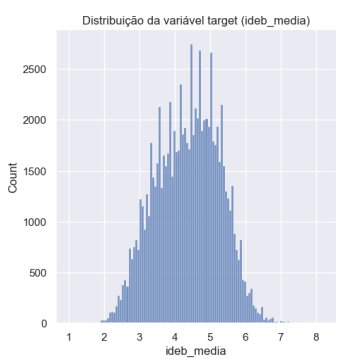

Figura 3. Target

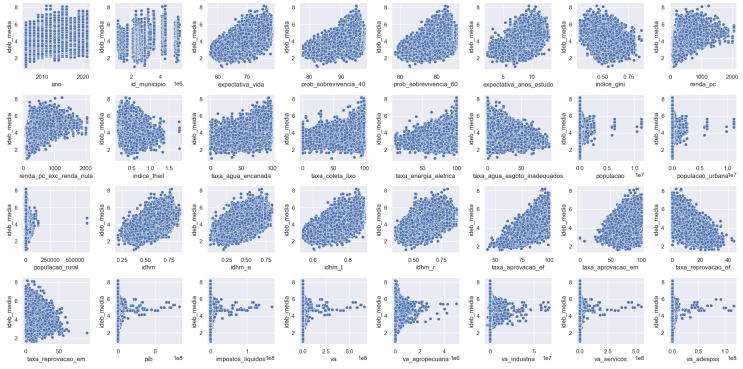

Figura 4. Scatter com a target

Como podemos ver, temos correlações significativas nas variáveis, e a distribuição parece algo normal.

3. Modelos Lineares Generalizados (GLMs)

A seguir, aprsento os modelos GLM utilizados para analisar o IDEB:

3.1. Modelo M1: GLM com Distribuição Gaussiana e função de ligação identidade

Este modelo assume uma distribuição Gaussiana (Normal) para os dados de resposta e utiliza a função de ligação identidade.

$$\begin{aligned} &\text{Modelo: } Y_i|X_i \sim N(\mu_i, \sigma^2), \\ &\mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \end{aligned}$$

onde Y_i representa a variável de resposta (no caso, `ideb_media`), X_{ij} são as variáveis explicativas selecionadas, $\beta_0, \beta_1, \ldots, \beta_p$ são os coeficientes a serem estimados, e σ^2 é a variância dos erros.

3.2. Modelo M2: GLM com Distribuição Gamma e função de ligação log

Neste modelo, a variável de resposta é assumida seguir uma distribuição Gamma com a função de ligação log.

$$\begin{aligned} &\text{Modelo: } Y_i|X_i \sim \text{Gamma}(\alpha, \beta), \\ &g(\mu_i) = \log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \end{aligned}$$

onde Y_i é a variável de resposta, μ_i é o valor esperado da distribuição Gamma, $g()$ é a função de ligação log, e α e β são os parâmetros da distribuição Gamma.

3.3. Modelo M3: GLM com Distribuição Inversa Gaussiana e função de ligação log

Aqui, a variável de resposta segue uma distribuição Inversa Gaussiana, e a função de ligação é log.

$$\begin{aligned} &\text{Modelo: } Y_i|X_i \sim \text{InverseGaussian}(\mu_i, \lambda), \\ &\quad g(\mu_i) = \log(\mu_i) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \end{aligned}$$

onde Y_i é a variável de resposta, μ_i é o valor esperado da distribuição Inversa Gaussiana, $g()$ é a função de ligação log, e λ é o parâmetro da distribuição Inversa Gaussiana.

3.4. Modelo M4: GLM com Transformação Logarítmica e Distribuição Gaussiana

Neste modelo, a variável de resposta `ideb_media` é transformada utilizando o logaritmo natural e segue uma distribuição Gaussiana.

$$\begin{aligned} &\text{Modelo: } \log(Y_i)|X_i \sim N(\mu_i, \sigma^2), \\ &\quad \mu_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} \end{aligned}$$

onde Y_i é a variável de resposta original `ideb_media`, e $\log(Y_i)$ é a variável transformada. Os parâmetros $\beta_0, \beta_1, \ldots, \beta_p$ são estimados utilizando mínimos quadrados ordinários (OLS).

3.5. Modelo M5: GLMM com Distribuição Gaussiana e função de ligação identidade

Este é um Modelo Linear Generalizado Misto (GLMM) que assume uma distribuição Gaussiana para os dados originais e utiliza a função de ligação identidade.

$$\text{Modelo: } Y_{ij}|X_i = \beta_0 + \beta_1 X_{1ij} + \beta_2 X_{2ij} + \ldots + \beta_p X_{pij} + b_j + \epsilon_{ij}$$

onde:

- Y_{ij} é a variável resposta (IDEb médio) para a observação i no grupo j.

- $X_{1ij}, X_{2ij}, \ldots, X_{pij}$ são os preditores fixos (ou variáveis explicativas) para a observação i no grupo j.

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes associados aos preditores fixos.

- b_j é o efeito aleatório para o grupo j, que captura a variabilidade não explicada pelos preditores fixos.

- ϵ_{ij} é o erro aleatório associado à observação i no grupo j, assumindo uma distribuição normal com média zero e variância σ^2.

3.6. Comentários sobre as escolhas

Modelo M1: É a escolha padrão quando os dados de resposta seguem uma distribuição normal e não há necessidade de transformação. Foi escolhido pela cara da distribuição do IDEB.

Modelo M2 e M3: São úteis quando os dados de resposta são positivos e assimétricos, como é o caso de dados de contagem ou tempo positivo. A distribuição Gamma e Inversa Gaussiana são apropriadas para lidar com essas características dos dados.

Modelo M4: Transformar a variável de resposta pode melhorar a adequação do modelo aos pressupostos da distribuição normal, especialmente se os dados originais não são normalmente distribuídos.

Modelo M5: É adequado quando há efeitos aleatórios para os municípios que não são modelados diretamente pelas variáveis explicativas fixas. Isso ajuda a capturar a variabilidade não explicada pelas variáveis fixas e a considerar a estrutura hierárquica dos dados. Um exemplo comum de aplicação de GLMMs é nesses casos, onde temos medições repetidas ao longo do tempom, permitindo assim que se modelem os efeitos fixos e as tendências médias ao longo do tempo.

4. Fittando os modelos

Para o modelo 1, as variáveis relacionadas à saúde e expectativa de vida, assim como indicadores de infraestrutura e desenvolvimento humano, têm impactos significativos sobre o ideb_media. Há uma combinação de efeitos positivos e negativos, destacando a complexidade das relações entre esses fatores e o desempenho educacional. A desigualdade (índice de Gini) tem um impacto fortemente negativo no desempenho educacional, sugerindo que políticas para reduzir a desigualdade podem melhorar os resultados educacionais.

No segundo modelo, usando a família Gamma com função de ligação Inverse-Power, encontramos relações semelhantes, mas com sinais opostos para muitas variáveis, refletindo a natureza da transformação. Por exemplo, a expectativa de vida e a probabilidade de sobrevivência aos 40 anos têm coeficientes negativos, enquanto no primeiro modelo eram positivos, e a probabilidade de sobrevivência aos 60 anos tem um coeficiente positivo em ambos, mas de magnitudes diferentes. Mas fora issa, revelam quase a mesma coisa.

O modelo 3 mostrou uma mistura de resultados dos dois modelos anteriores, mas com suas próprias nuances. A expectativa de vida e a probabilidade de sobrevivência aos 40 anos tiveram coeficientes negativos, semelhantes ao modelo Gamma. Para o modelo 4, temos analises muito parecidas com os demais.

Para o modelo 5 a expectativa de vida, probabilidade de sobrevivência até 40 anos, expectativa de anos de estudo, IDH e subíndices de educação e renda têm impactos positivos no IDEB, enquanto a probabilidade de sobrevivência até 60 anos, renda per capita, taxa de água encanada inadequada e índice de Gini têm impactos negativos. A taxa de aprovação no ensino fundamental e médio também contribui positivamente para o IDEB. O modelo explica uma parte significativa da variação no IDEB, como evidenciado pelos coeficientes das variáveis e seus valores p, que indicam significância estatística alta.

Generalized Linear Model Regression Results

Dep. Variable:	ideb_media	No. Observations:	87998
Model:	GLM	Df Residuals:	87981
Model Family:	Gaussian	Df Model:	16
Link Function:	Identity	Scale:	0.24774
Method:	IRLS	Log-Likelihood:	-63460.
Date:	Sun, 23 Jun 2024	Deviance:	21796.
Time:	23:07:12	Pearson chi2:	2.18e+04
No. Iterations:	3	Pseudo R-squ. (CS):	0.8954
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-10.1545	0.165	-61.692	0.000	-10.477	-9.832
id_municipio	-8.465e-08	2.62e-09	-32.306	0.000	-8.98e-08	-7.95e-08
expectativa_vida	0.1398	0.002	80.689	0.000	0.136	0.143
prob_sobrevivencia_40	0.1018	0.003	35.893	0.000	0.096	0.107
prob_sobrevivencia_60	-0.1211	0.002	-65.465	0.000	-0.125	-0.118
expectativa_anos_estudo	0.0175	0.002	8.004	0.000	0.013	0.022
renda_pc	-0.0017	0.000	-8.241	0.000	-0.002	-0.001
renda_pc_exc_renda_nula	0.0021	0.000	10.278	0.000	0.002	0.003
taxa_agua_encanada	-0.0024	0.000	-15.521	0.000	-0.003	-0.002
taxa_coleta_lixo	0.0015	0.000	11.485	0.000	0.001	0.002
taxa_energia_eletrica	0.0005	0.000	1.891	0.059	-1.73e-05	0.001
idhm	1.0076	0.073	13.866	0.000	0.865	1.150
taxa_aprovacao_ef	0.0446	0.001	52.680	0.000	0.043	0.046
taxa_aprovacao_em	0.0088	0.000	40.304	0.000	0.008	0.009
indice_gini	-1.0588	0.032	-33.202	0.000	-1.121	-0.996
taxa_agua_esgoto_inadequados	-0.0047	0.000	-24.768	0.000	-0.005	-0.004
taxa_reprovacao_ef	0.0050	0.001	4.813	0.000	0.003	0.007

Figura 5. Modelo 1

Dep. Variable:	ideb_media	No. Observations:	87998
Model:	GLM	Df Residuals:	87981
Model Family:	Gamma	Df Model:	16
Link Function:	InversePower	Scale:	0.013839
Method:	IRLS	Log-Likelihood:	-64812.
Date:	Sun, 23 Jun 2024	Deviance:	1227.0
Time:	23:07:16	Pearson chi2:	1.22e+03
No. Iterations:	7	Pseudo R-squ. (CS):	0.8953
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	1.3864	0.010	132.894	0.000	1.366	1.407
id_municipio	2.807e-09	1.39e-10	20.214	0.000	2.53e-09	3.08e-09
expectativa_vida	-0.0070	9.36e-05	-74.557	0.000	-0.007	-0.007
prob_sobrevivencia_40	-0.0084	0.000	-49.068	0.000	-0.009	-0.008
prob_sobrevivencia_60	0.0071	0.000	67.298	0.000	0.007	0.007
expectativa_anos_estudo	-0.0006	0.000	-5.487	0.000	-0.001	-0.000
renda_pc	0.0001	1.21e-05	8.842	0.000	8.35e-05	0.000
renda_pc_exc_renda_nula	-0.0001	1.22e-05	-9.484	0.000	-0.000	-9.2e-05
taxa_agua_encanada	6.966e-05	9.11e-06	7.644	0.000	5.18e-05	8.75e-05
taxa_coleta_lixo	-0.0001	8.79e-06	-12.381	0.000	-0.000	-9.16e-05
taxa_energia_eletrica	-0.0002	1.64e-05	-10.488	0.000	-0.000	-0.000
idhm	-0.0436	0.004	-10.879	0.000	-0.051	-0.036
taxa_aprovacao_ef	-0.0041	5.62e-05	-73.535	0.000	-0.004	-0.004
taxa_aprovacao_em	-0.0003	1.22e-05	-27.428	0.000	-0.000	-0.000
indice_gini	0.0291	0.002	17.132	0.000	0.026	0.032
taxa_agua_esgoto_inadequados	0.0003	1.21e-05	27.490	0.000	0.000	0.000
taxa_reprovacao_ef	-0.0019	6.63e-05	-29.137	0.000	-0.002	-0.002

Figura 6. Modelo 2

Generalized Linear Model Regression Results

Dep. Variable:	ideb_media	No. Observations:	87998
Model:	GLM	Df Residuals:	87981
Model Family:	InverseGaussian	Df Model:	16
Link Function:	InverseSquared	Scale:	0.0035327
Method:	IRLS	Log-Likelihood:	-69529.
Date:	Sun, 23 Jun 2024	Deviance:	320.11
Time:	23:07:23	Pearson chi2:	311.
No. Iterations:	7	Pseudo R-squ. (CS):	0.8768
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	0.6878	0.005	127.953	0.000	0.677	0.698
id_municipio	8.693e-10	6.58e-11	13.203	0.000	7.4e-10	9.98e-10
expectativa_vida	-0.0032	4.5e-05	-70.131	0.000	-0.003	-0.003
prob_sobrevivencia_40	-0.0047	8.7e-05	-53.935	0.000	-0.005	-0.005
prob_sobrevivencia_60	0.0035	5.2e-05	67.028	0.000	0.003	0.004
expectativa_anos_estudo	-0.0002	5.42e-05	-3.773	0.000	-0.000	-9.82e-05
renda_pc	5.463e-05	5.94e-06	9.194	0.000	4.3e-05	6.63e-05
renda_pc_exc_renda_nula	-5.552e-05	5.98e-06	-9.277	0.000	-6.72e-05	-4.38e-05
taxa_agua_encanada	1.387e-05	4.49e-06	3.088	0.002	5.07e-06	2.27e-05
taxa_coleta_lixo	-6e-05	4.62e-06	-12.997	0.000	-6.91e-05	-5.1e-05
taxa_energia_eletrica	-0.0001	8.61e-06	-15.661	0.000	-0.000	-0.000
idhm	-0.0166	0.002	-8.656	0.000	-0.020	-0.013
taxa_aprovacao_ef	-0.0023	2.93e-05	-79.981	0.000	-0.002	-0.002
taxa_aprovacao_em	-0.0001	5.81e-06	-19.608	0.000	-0.000	-0.000
indice_gini	0.0065	0.001	8.221	0.000	0.005	0.008
taxa_agua_esgoto_inadequados	0.0002	6.25e-06	27.656	0.000	0.000	0.000
taxa_reprovacao_ef	-0.0013	3.41e-05	-39.436	0.000	-0.001	-0.001

Figura 7. Modelo 3

Generalized Linear Model Regression Results

Dep. Variable:	np.log(ideb_media)	No. Observations:	87998
Model:	GLM	Df Residuals:	87978
Model Family:	Gaussian	Df Model:	19
Link Function:	Identity	Scale:	0.013891
Method:	IRLS	Log-Likelihood:	63308.
Date:	Sun, 23 Jun 2024	Deviance:	1222.1
Time:	23:38:37	Pearson chi2:	1.22e+03
No. Iterations:	3	Pseudo R-squ. (CS):	0.9086
Covariance Type:	nonrobust		

	coef	std err	z	P>|z|	[0.025	0.975]
Intercept	-3.6265	0.581	-6.241	0.000	-4.765	-2.488
id_municipio	-1.922e-08	6.3e-10	-30.519	0.000	-2.05e-08	-1.8e-08
expectativa_vida	0.0713	0.023	3.070	0.002	0.026	0.117
prob_sobrevivencia_40	0.0288	0.001	42.732	0.000	0.028	0.030
prob_sobrevivencia_60	-0.0294	0.000	-66.190	0.000	-0.030	-0.028
expectativa_anos_estudo	0.0052	0.001	10.049	0.000	0.004	0.006
renda_pc	-0.0005	4.88e-05	-10.451	0.000	-0.001	-0.000
renda_pc_exc_renda_nula	0.0005	4.89e-05	9.983	0.000	0.000	0.001
taxa_agua_encanada	-0.0006	3.82e-05	-15.063	0.000	-0.001	-0.001
taxa_coleta_lixo	0.0002	3.22e-05	5.550	0.000	0.000	0.000
taxa_energia_eletrica	-4.27e-05	6.33e-05	-0.675	0.500	-0.000	8.13e-05
idhm	0.0473	0.096	8.867	0.000	0.660	1.035
idhm_e	-0.3015	0.042	-7.199	0.000	-0.384	-0.219
idhm_l	-2.5830	1.393	-1.854	0.064	-5.314	0.148
idhm_r	0.2264	0.039	5.768	0.000	0.149	0.303
taxa_aprovacao_ef	0.0128	0.000	63.771	0.000	0.012	0.013
taxa_aprovacao_em	0.0020	5.21e-05	38.273	0.000	0.002	0.002
indice_gini	-0.2037	0.008	-26.439	0.000	-0.219	-0.189
taxa_agua_esgoto_inadequados	-0.0011	4.49e-05	-24.260	0.000	-0.001	-0.001
taxa_reprovacao_ef	0.0035	0.000	14.332	0.000	0.003	0.004

Figura 8. Modelo 4

Mixed Linear Model Regression Results

Model:	MixedLM	Dependent Variable:	ideb_media
No. Observations:	1064393	Method:	REML
No. Groups:	27	Scale:	0.2545
Min. group size:	239	Log-Likelihood:	-782272.9703
Max. group size:	159374	Converged:	Yes
Mean group size:	39422.0		

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
Intercept	-5.876	0.726	-8.094	0.000	-7.298	-4.453
id_municipio	-0.000	0.000	-4.736	0.000	-0.000	-0.000
expectativa_vida	0.089	0.029	3.098	0.002	0.033	0.145
prob_sobrevivencia_40	0.086	0.001	86.551	0.000	0.084	0.088
prob_sobrevivencia_60	-0.078	0.001	-90.684	0.000	-0.079	-0.076
expectativa_anos_estudo	0.021	0.001	28.419	0.000	0.020	0.023
renda_pc	-0.003	0.000	-44.711	0.000	-0.003	-0.003
renda_pc_exc_renda_nula	0.003	0.000	49.222	0.000	0.003	0.003
taxa_agua_encanada	-0.004	0.000	-69.893	0.000	-0.004	-0.004
taxa_coleta_lixo	0.002	0.000	53.986	0.000	0.002	0.003
taxa_energia_eletrica	-0.001	0.000	-13.516	0.000	-0.001	-0.001
idhm	0.639	0.123	5.182	0.000	0.398	0.881
idhm_e	0.937	0.054	17.242	0.000	0.831	1.044
idhm_l	-1.026	1.722	-0.596	0.551	-4.401	2.349
idhm_r	0.996	0.051	19.487	0.000	0.896	1.097
taxa_aprovacao_ef	0.018	0.000	93.927	0.000	0.018	0.018
taxa_aprovacao_em	0.006	0.000	83.670	0.000	0.006	0.006
indice_gini	-1.455	0.010	-144.348	0.000	-1.475	-1.436
taxa_agua_esgoto_inadequados	-0.003	0.000	-44.054	0.000	-0.003	-0.003
taxa_reprovacao_ef	-0.003	0.000	-12.893	0.000	-0.004	-0.003
1 | code	0.020	0.005	3.601	0.000	0.009	0.030
Group Var	0.042	0.018				

Figura 9. Modelo 5

5. Comparando os Modelos

Tabela 1. Comparação de Modelos

Modelo	MSE	R2	AIC	BIC
M1	0.299330	0.625552	126953.4	-979873.6
M2	0.307918	0.614808	129657.5	-1000443.0
M3	0.327857	0.589865	139091.8	-1001350.0
M4	0.293677	0.632624	-126575.0	-1000414.0
M5	0.307287	0.615598	1564310.0	1564583.0

Como podemos ver, os modelos tem aproximações muito parecidas. No entanto o modelo 4 consegue, de alguma forma, ter valores menores. Isso provavelmente se deve ao fato de estarmos escalando os dados com uma função log, o que ajuda a entender melhor.

Observando os resíduos, podemos perceber que, apesar das diferenças, as predições são bem parecidas, tanto entre modelos, quanto em comparação entre anos.

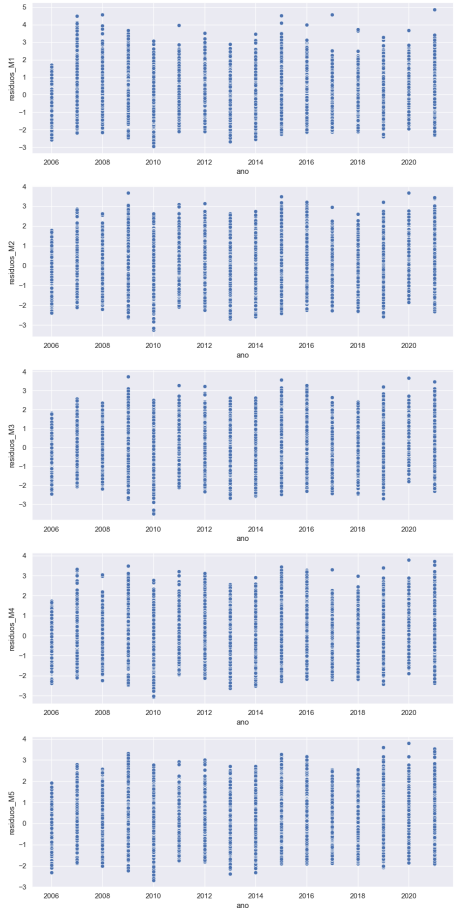

Figura 10. Resíduos

6. Conclusao

6.1. Resumo

Os principais achados podem ser resumidos da seguinte forma:

- Variáveis Socioeconômicas e Educacionais: Identificamos que várias variáveis socioeconômicas, como expectativa de vida, IDH, e renda per capita, têm correlações significativas com o IDEB. Por outro lado, a desigualdade (medida pelo índice de Gini) demonstrou um impacto negativo significativo no desempenho educacional.
- Modelos Utilizados: Comparando os cinco modelos aplicados (M1 a M5), todos eles forneceram insights valiosos sobre os fatores que afetam o IDEB. O Modelo M1, que utiliza uma distribuição Gaussiana com função de ligação identidade, e o Modelo M4, que aplica uma transformação logarítmica, apresentaram os melhores desempenhos em termos de erro médio quadrático (MSE) e coeficiente de determinação (R^2).
- Complexidade das Relações: Os resultados dos modelos revelam que as relações entre os fatores socioeconômicos e o IDEB são complexas. Por exemplo, enquanto a renda per capita mostrou correlação positiva, o PIB não apresentou uma correlação significativa, destacando a necessidade de considerar variáveis mais específicas e menos agregadas ao analisar o desempenho educacional.
- Efeitos Aleatórios: O Modelo M5, que incorpora efeitos aleatórios para os municípios, mostrou que considerar a estrutura hierárquica dos dados pode ajudar a capturar variabilidade não explicada pelos preditores fixos. Este modelo é particularmente útil para analisar dados com medições repetidas ao longo do tempo.
- Políticas Públicas: Os resultados sugerem que políticas focadas na redução da desigualdade e no aumento da expectativa de vida, assim como no aprimoramento das condições socioeconômicas gerais, podem ter um impacto positivo no IDEB.

Em conclusão, este estudo oferece uma análise detalhada das relações entre indicadores socioeconômicos e educacionais e o desempenho educacional nos municípios brasileiros. Os modelos utilizados forneceram insights importantes que podem orientar a formulação de políticas públicas mais eficazes para melhorar a qualidade da educação básica no Brasil. Futuros estudos podem expandir essa análise, incluindo mais variáveis e diferentes metodologias, para obter uma compreensão ainda mais profunda das dinâmicas envolvidas.

6.2. Limitações

Dentre as limitações temos o fato dos dados do Índice de Desenvolvimento Humano (IDH) serem coletados a cada dez anos, enquanto outros indicadores são anuais. A repetição dos dados do censo para os anos intermediários pode não refletir mudanças rápidas nas condições socioeconômicas dos municípios, levando a uma possível defasagem na análise.

A utilização do Produto Interno Bruto (PIB) como um indicador econômico pode ser limitado, pois o PIB é uma métrica geral que não captura adequadamente as nuances dos investimentos em educação em nível municipal. Isso pode explicar a baixa correlação encontrada entre o PIB e o IDEB. Talves utilizar dados de investimentos na educação fossem mais efetivos.

Embora o Modelo M5 tenha considerado efeitos aleatórios para os municípios, a complexidade da estrutura hierárquica dos dados pode não ter sido completamente capturada. Fatores contextuais e regionais específicos podem influenciar os resultados educacionais de maneiras não modeladas explicitamente.

Além disso, cada modelo tem suas próprias suposições e limitações. Por exemplo, os Modelos GLM assumem uma relação linear entre os preditores e a variável de resposta, o que pode não ser adequado para todos os tipos de dados. Além disso, modelos com transformações logarítmicas (como o Modelo M4) podem ser sensíveis a outliers.

Por fim, as análises foram conduzidas em uma base de dados transversal, considerando múltiplos anos, mas sem um modelo dinâmico que capture adequadamente as mudanças ao longo do tempo. Estudos futuros podem se beneficiar de abordagens de séries temporais ou modelos longitudinais para explorar melhor as tendências temporais.

6.3. Trabalhos Futuros

Nos trabalhos futuros, seria interessante aprofundar a análise dos dados temporais utilizando modelos de painéis dinâmicos que considerem a endogeneidade e a autocorrelação, oferecendo uma compreensão mais detalhada das variações do IDEB ao longo do tempo. Além disso, a inclusão de novas variáveis explicativas, como indicadores de infraestrutura escolar e dados mais detalhados de economia voltada para educação podem proporcionar uma visão mais completa dos fatores que influenciam o desempenho educacional. Finalmente, desenvolver estudos comparativos entre diferentes regiões e períodos pode ajudar a identificar padrões e diferenças regionais, contribuindo para a formulação de políticas públicas mais eficazes.

Referências

[1] Base dos Dados. *Base dos Dados*. Disponível em: `https://basedosdados.org`.

[2] Atlas do Desenvolvimento Humano (ADH). *Atlas do Desenvolvimento Humano no Brasil*. Disponível em: `https://basedosdados.org/dataset/cbfc7253-089b-44e2-8825-755e1419efc8?table=65639055-2408-46b4-8f82-ecae3d04b800`.

[3] Produto Interno Bruto do Brasil (PIB). *Produto Interno Bruto dos Municípios Brasileiros*. Disponível em: `https://basedosdados.org/dataset/fcf025ca-8b19-4131-8e2d-5ddb12492347?table=93007431-7ce9-42ee-8740-8c2274d345ad`.

[4] Indicadores Educacionais. *Indicadores Educacionais*. Disponível em: `https://basedosdados.org/dataset/63f1218f-c446-4835-b746-f109a338e3a1?table=95f49a8d-fb99-416c-ab92-10bcb523b3a3`.

[5] Índice de Desenvolvimento da Educação Básica (Ideb). *Índice de Desenvolvimento da Educação Básica*. Disponível em: `https://basedosdados.org/dataset/96eab476-5d30-459b-82be-f888d4d0d6b9?table=bc84dea9-1126-4423-86d2-8835e6b19a72`.

[6] McCullagh, P., & Nelder, J. A. (1989). *Generalized Linear Models*. 2nd Edition, Chapman and Hall/CRC, Boca Raton.

[7] *statsmodels Documentation*. Available at: https://www.statsmodels.org/stable/index.html