Trabalho MLG

Eliana Cardoso Gonçalves e Sophia Araujo de Moraes

26/08/2024

Índice

1 Dados da Penn World Table de 2020

Dois tópicos desse estudo (MODELO LINEAR NORMAL e MODELO GAMA LOG-LINEAR) se concentram em uma amostra de 183 países distintos no ano de 2014. Os dois modelos foram utilizados para explorar como a força de trabalho e o estoque de capital influenciam o PIB, desse modo por meio de regressões lineares múltiplas, será estimado os coeficientes que quantificam essas relações, permitindo-nos compreender melhor os determinantes do crescimento econômico.

Essa relação pode ser expressa pela seguinte equação:

$$Y = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta pXp + \varepsilon$$

Onde:

- Y é a variável dependente (PIB Nacional (Y) em dólares PPP de 2014).
- $X1, X2, ..., Xp$ são as variáveis independentes (Força de Trabalho (L)e o Estoque de Capital (K) em dólares PPP de 2014).
- $\beta 1, \beta 2, ..., \beta p$ são os coeficientes de regressão que representam os efeitos das variáveis independentes na variável dependente.
- ε é o termo de erro, que representa a variação não explicada pelo modelo.

Para complementar a análise, é importante entender como os dois modelos — o Modelo Linear Normal e o Modelo Gama Log-Linear — foram utilizados para explorar as relações entre o PIB nacional, a força de trabalho e o estoque de capital. Esses modelos oferecem diferentes abordagens para analisar os dados e podem fornecer insights valiosos sobre os fatores que impulsionam o crescimento econômico:

Modelo Linear Normal: O Modelo Linear Normal é uma aplicação clássica da regressão linear múltipla. Neste contexto, assume-se que a relação entre as variáveis independentes (força de trabalho e estoque de capital) e a variável dependente (PIB) é linear. Isso significa que cada incremento em uma variável independente resulta em um aumento ou diminuição constante na variável dependente, dependendo do sinal e do valor dos coeficientes de regressão.

Este modelo é baseado em algumas suposições, como a normalidade dos resíduos, homocedasticidade (ou seja, a variância constante dos resíduos), e a ausência de multicolinearidade entre as variáveis independentes. Quando essas suposições são satisfeitas, o Modelo Linear Normal fornece estimativas consistentes e eficientes dos coeficientes de regressão, permitindo uma compreensão clara de como cada fator contribui para o PIB.

Modelo Gama Log-Linear: Por outro lado, o Modelo Gama Log-Linear oferece uma abordagem alternativa, especialmente útil quando os dados não atendem às suposições de normalidade e homocedasticidade do modelo linear. Este modelo é particularmente apropriado para situações

onde a variável dependente (PIB) assume valores positivos e pode apresentar uma distribuição assimétrica, o que é comum em dados econômicos.

No Modelo Gama Log-Linear, a transformação logarítmica é aplicada à variável dependente. Isso permite capturar relações não lineares entre o PIB e as variáveis independentes, como a força de trabalho e o estoque de capital. A distribuição Gama, utilizada nesse modelo, é adequada para modelar a variável dependente em casos onde há heterocedasticidade — ou seja, quando a variabilidade dos dados aumenta com o valor previsto.

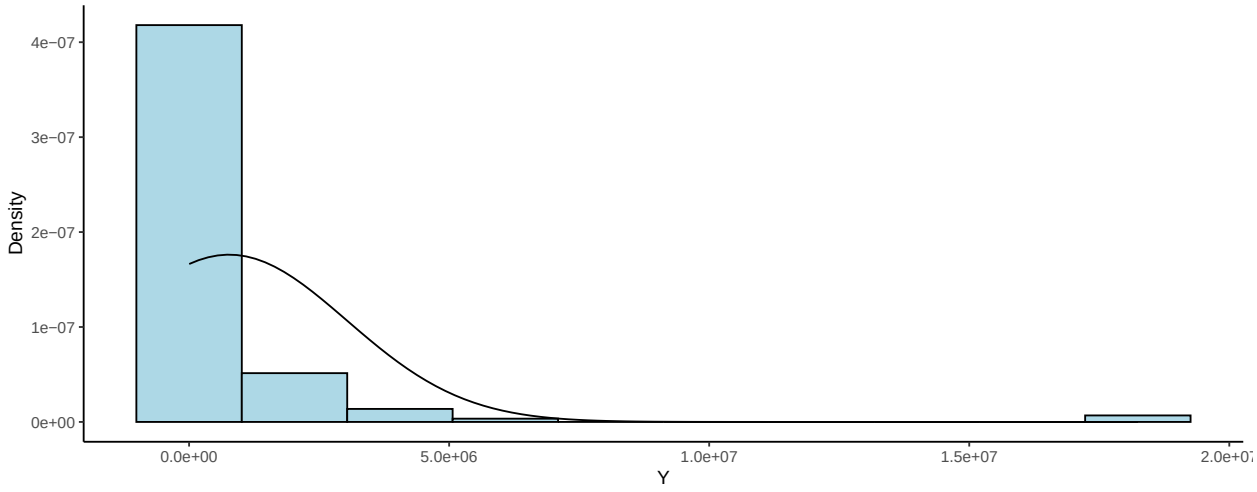
Essa abordagem é especialmente útil para capturar efeitos multiplicativos, onde as variáveis independentes influenciam a variável dependente de forma proporcional, em vez de aditiva. O Modelo Gama Log-Linear pode, portanto, revelar nuances e complexidades na relação entre os fatores estudados e o crescimento econômico que o modelo linear normal pode não captar.

1.1 Modelo Linear Normal

Tabela 1.1: Tabela Resumo das Variáveis Y, L e K.

	K	L	Y
Mean	3058761.263	54.6094443	769941.411
Std.Dev	8413105.576	196.3877082	2264112.549
Min	7345.099	0.2993725	2569.155
Median	394840.953	11.6562212	123419.395
Max	64118472.000	2045.9122435	18244220.000

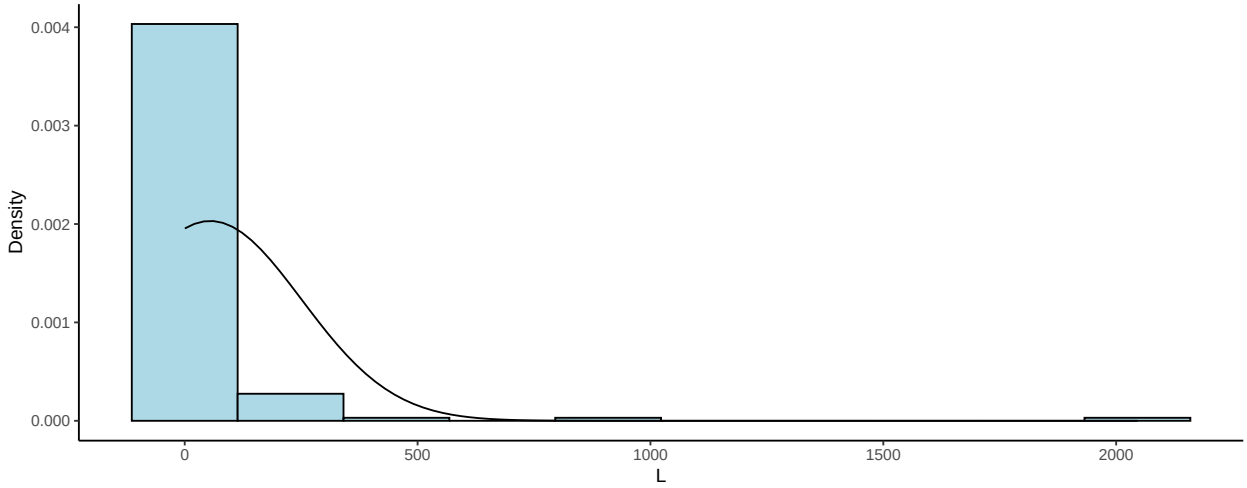
Figura 1.1: Histograma PIB Nacional (Y) em dólares PPP de 2014



Observando o Figura **??** e Tabela **??** , o valor mínimo do PIB é $2.569, enquanto o valor máximo é significativamente maior, chegando a $18.244.220, indicando uma grande variação no PIB entre os países da amostra.A média do PIB é $769.941, o que é substancialmente maior que a mediana de $123.419, sugerindo que alguns países com PIB muito alto estão puxando a média para cima. A

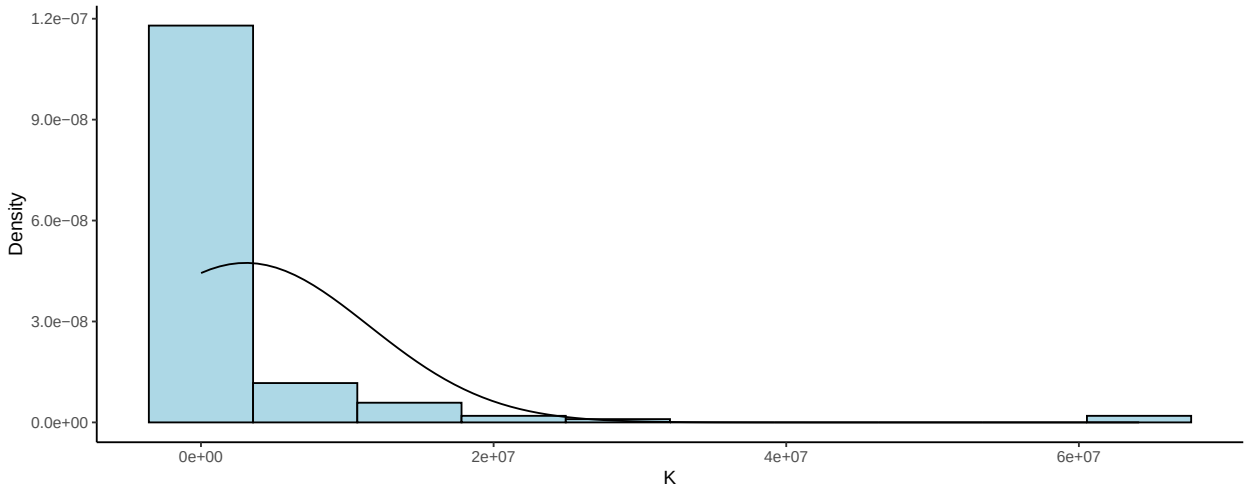
distribuição do PIB é bastante assimétrica, com muitos países tendo PIBs relativamente baixos e poucos países com PIBs muito altos.

Figura 1.2: Histograma Força de Trabalho (L)



Analisando o Figura **??** e a Tabela **??**, a variação da força de trabalho de 0,2994 a 2.045,91 reflete uma grande disparidade na população economicamente ativa entre os países. Com uma média de 54,60 e uma mediana de 11,65, a distribuição mostra-se assimétrica, sugerindo a presença de países com forças de trabalho extremas, seja pela sua grandeza ou pequenez em relação à média da amostrm alguns países com países com a força de trabalho outlier.

Figura 1.3: Histograma Estoque de Capital (K) em dólares PPP de 2014)



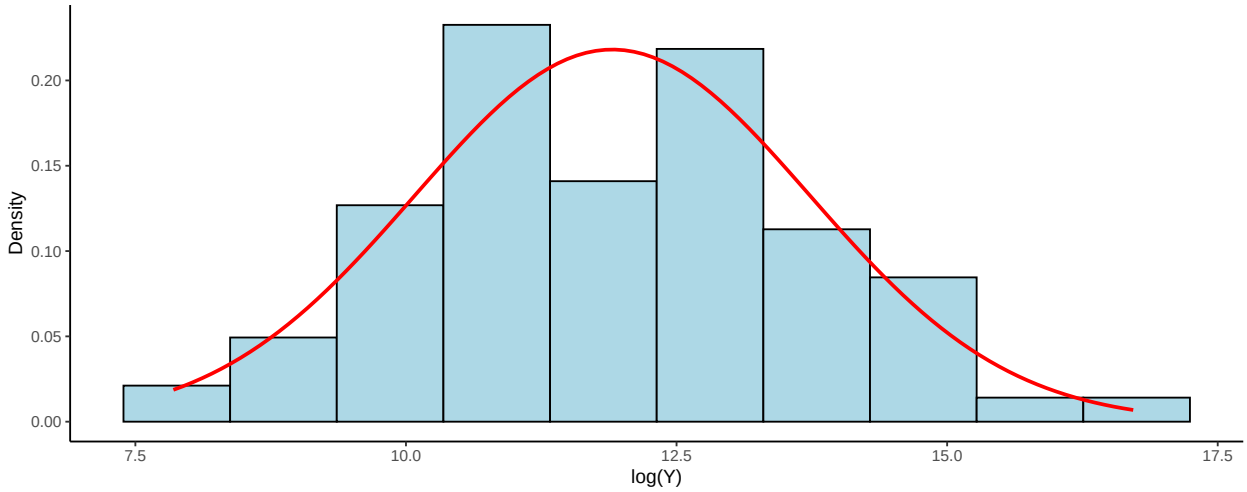
Analisando os valores da Tabela **??** e Figura **??** , observamos que o estoque de capital varia amplamente, de $7.345 a $64.118.472, indicando diferenças significativas no nível de investimento em capital produtivo entre os países.

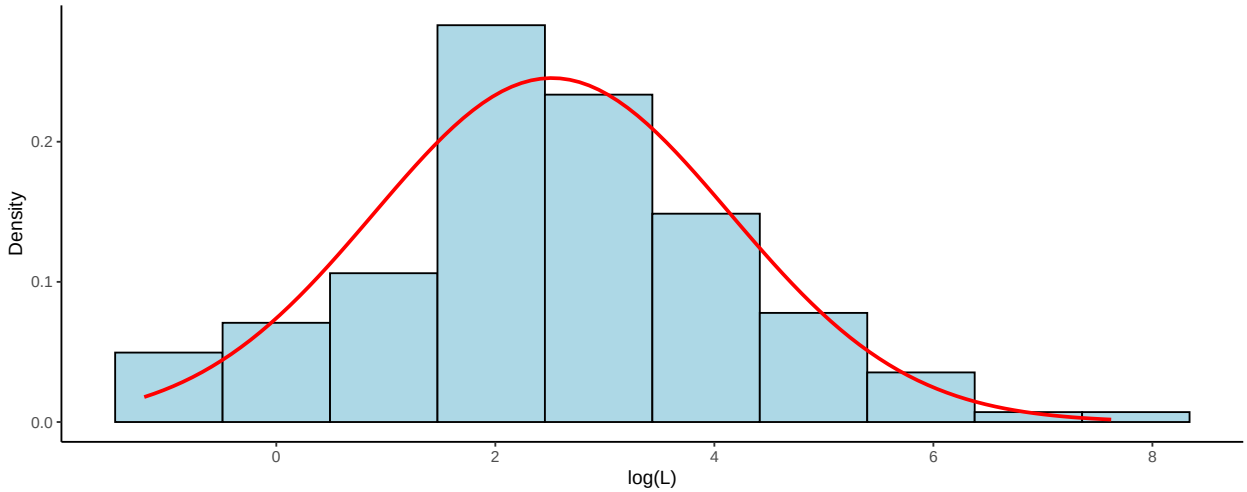
A média do estoque de capital é $3.058.761, consideravelmente maior que a mediana de $394.841, o que sugere que alguns países têm estoques de capital extremamente altos que estão elevando a média.

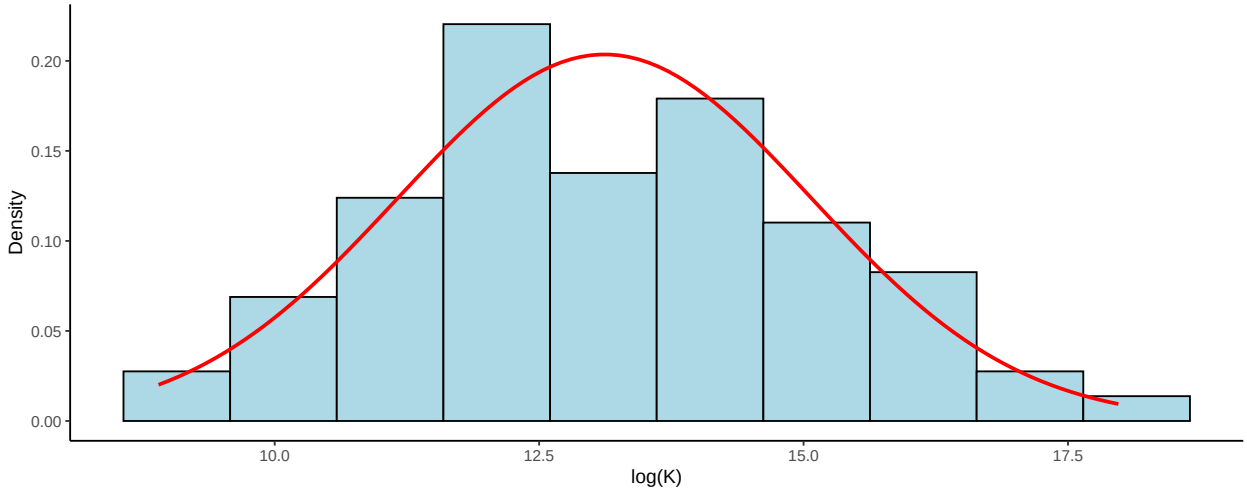
1.1.1 Resultados

Para garantir a robustez e confiabilidade dos resultados, optamos por realizar uma transformação log-log nas variáveis, considerando que o PIB Nacional (Y) em dólares PPP de 2014, juntamente com as variáveis independentes "força de trabalho" e "estoque de capital", não apresentam uma distribuição normal. Essa abordagem vai permitir que aos pressupostos do modelo normal sejam atendidos e obter resultados mais consistentes.

Ao aplicar a transformação log-log, estamos ajustando a distribuição das variáveis para se adequarem melhor ao modelo, mitigando quaisquer distorções ou viés que possam surgir devido à falta de normalidade. Isso nos permite realizar inferências estatísticas mais confiáveis e interpretar os efeitos das variáveis independentes sobre o PIB Nacional de forma mais precisa.







Essa estratégia de transformação aumenta a robustez da análise, pois reduz a influência de valores extremos e torna os resultados menos sensíveis a distribuições não normais. Portanto, podemos ter maior confiança nas conclusões derivadas do modelo, garantindo uma abordagem metodológica sólida e resultados mais confiáveis para tomada de decisão.

A análise macroeconômica é fundamental para entender o desenvolvimento econômico e social dos países. No contexto da contabilidade nacional, variáveis como o Produto Interno Bruto (PIB), a força de trabalho e o estoque de capital são essenciais para avaliar a produtividade e o crescimento econômico. Este estudo utiliza dados da Penn World Table de 2020 para investigar a relação entre essas variáveis.

O PIB nacional (Y), medido em dólares PPC de 2014, é uma medida abrangente da atividade econômica de um país. A força de trabalho (L) representa o total de pessoas empregadas ou em busca de emprego, refletindo a capacidade produtiva humana. O estoque de capital (K), também medido em dólares PPC de 2014, indica o valor total dos ativos produtivos de um país, como máquinas, edifícios e infraestrutura.

Ao analisar essas variáveis, o objetivo é fornecer insights sobre as políticas econômicas que podem fomentar o crescimento e a produtividade. Esta investigação pode ajudar formuladores de políticas, economistas e pesquisadores a identificar áreas-chave para intervenção e investimento, promovendo um desenvolvimento econômico sustentável e inclusivo.

Tabela 1.2: Ajuste do Modelo de Regressão Log-Normal

	Estimação	Pvalor	sig
(Intercept)	1.8881930	0	<0.001***
log(L)	0.2940893	0	<0.001***
log(K)	0.7072816	0	<0.001***

Todos os p-valores associados aos coeficientes são muito pequenos (<0.001), o que significa que podemos rejeitar a hipótese nula para todos os coeficientes, ao nível de 5% de confiança. Isso indica que tanto a força de trabalho quanto o estoque de capital têm efeitos significativos no PIB, conforme medido pelo logaritmo.

O coeficiente estimado para log(K) é 0,70728. Isso significa que, se o estoque de capital (K) aumentar em 1%, o PIB (Y) aumentará em aproximadamente $0.70728 \times 100 = 70.728$, mantendo todas as outras variáveis constantes.

O coeficiente estimado para log (L) é0.29. Isso significa que, se a força de trabalho (L) aumentar em 1%, o PIB (Y) aumentará em aproximadamente $0{,}29 \times 100 = 29{,}409$, mantendo todas as outras variáveis constantes.

Além disso, o modelo tem um R-quadrado ajustado de aproximadamente 0,96, o que significa que aproximadamente 95,92% da variabilidade no logaritmo do PIB pode ser explicada pelas variáveis independentes incluídas no modelo.

Esses resultados sugerem que tanto a força de trabalho quanto o estoque de capital têm um impacto significativo no PIB, conforme medido pelo logaritmo.

- **Hipótese Nula** (H_0): $\alpha + \beta = 1$ A soma dos coeficientes é igual a 1, sugerindo retornos constantes à escala.
- **Hipótese Alternativa** (H_1): $\alpha + \beta \neq 1$ A soma dos coeficientes não é igual a 1, sugerindo que não há retornos constantes à escala.

Tabela 1.3: Teste

Estatística	Valor
Estimativa de alpha	0.2941
Estimativa de beta	0.7073
Soma de alpha e beta	1.0014
Estatística t	0.0721
Valor-p	0.9427

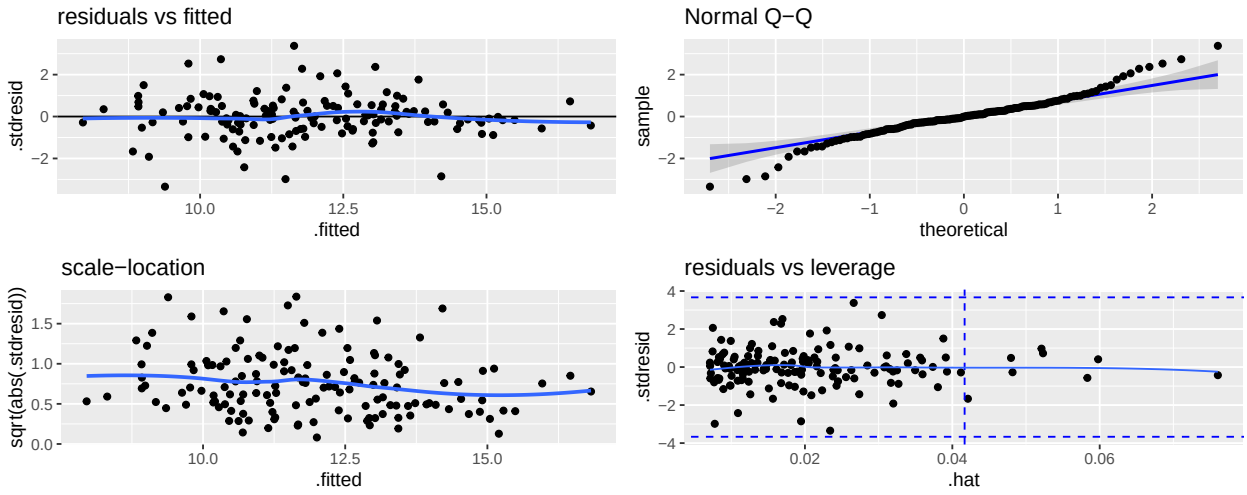
Na Tabela **??**, com teste da Teste $H0 : \alpha + \beta = 1$ contra $H1 : \alpha + \beta \neq 1$, obtivemos um p-valor de aproximadamente 0,943. Não rejeitamos a hipótese nula H_0 ao nível de 5% de significância. Isso indica que os dados não fornecem evidências suficientes para concluir que a soma de α e β é diferente de 1. Em outras palavras, a suposição de retornos constantes à escala $\alpha + \beta = 1$ é razoável para o ano analisado. Isso sugere que, com base nos dados, um aumento proporcional igual na força de trabalho (L) e no capital (K) resulta em um aumento proporcional na produção (Y), confirmando a hipótese de retornos constantes à escala na função de produção Cobb-Douglas.

1.1.2 Análise dos Resíduos

Tabela 1.4: Resultados dos Testes de Resíduos

Teste	Estatística	p.valor
Shapiro-Wilk Normality Test	W = 0.96	0.0003379
Non-constant Variance Score Test	Chisquare = 5.338458	0.0208600
Durbin-Watson Test for Autocorrelated Errors	D-W Statistic = 2.133801	0.4200000

Figura 1.4: Grafico de Análise de resíduos



Com base na análise da tabela Figura **??**, podemos concluir o seguinte sobre o ajuste do modelo de regressão:

Homocedasticidade: - A variância dos resíduos parece ser constante em toda a faixa dos valores ajustados, o que é consistente com a suposição de homocedasticidade na regressão linear.

Normalidade: - Embora o teste de Shapiro-Wilk sugira que os resíduos não sigam uma distribuição normal, desconsiderando caudas pesadas tanto nos extremos inferiores quanto nos superiores, podemos aproximar que os resíduos estão próximos de uma distribuição normal.

Independência dos Resíduos: - Tanto o teste de Durbin-Watson quanto a inspeção do gráfico de resíduos versus valores ajustados na tabela Figura **??** não fornecem evidências significativas para rejeitar a hipótese nula de ausência de autocorrelação positiva ou negativa nos resíduos. Assim, parece que os resíduos são independentes entre si.

Linearidade: - Não há um padrão claro nos resíduos plotados em relação aos valores ajustados, indicando linearidade entre as variáveis independentes e dependentes.

Portanto, concluirmos que o modelo de regressão parece atender às suposições de homocedasticidade, normalidade aproximada dos resíduos, independência dos resíduos e linearidade entre as variáveis.

1.2 Modelo Gama Log-Linear

1.2.1 Resultados

Utilizamos o mesmo banco de dados do Modelo de Log-Normal.Por meio da tabela abaixo é possível verificar como foi o ajuste das variáveis ao modelo de gama log-linear:

Tabela 1.5: Modelo Gamma Log-linear

	Estimacao	ErroPadrao	Pvalor	sig
(Intercept)	1.6175	0.0286	0	<0.001***
log(L)	0.0237	0.0032	0	<0.001***
log(K)	0.0601	0.0026	0	<0.001***

O intercepto estimado de 1.61754583 (Tabela **??**) representa o logaritmo da produção quando tanto o trabalho (L) quanto o capital (K) são iguais a 1. Com um p-valor extremamente pequeno, o intercepto é estatisticamente significativo, indicando que ele desempenha um papel importante na modelagem da variável dependente , ao nível de 5% de significância.

O coeficiente para log(L) é 0.02371, indicando que um aumento de 1% na força de trabalho (L) leva a um aumento de aproximadamente 0.024% na produção, mantendo o capital constante. O p-valor extremamente pequeno indica que este coeficiente é altamente significativo, o que valida a importância da força de trabalho na determinação da produção no modelo.

O coeficiente para log(K) é 0.06012, sugerindo que um aumento de 1% no capital (K) resulta em um aumento de aproximadamente 0.060% na produção, mantendo a força de trabalho constante. Com um p-valor extremamente baixo, esse coeficiente é altamente significativo, demonstrando que o capital tem uma influência importante e estatisticamente significativa na produção.

O parâmetro de dispersão estimado é 0.00135, o que indica que há uma baixa variabilidade dos dados em torno da média ajustada pelo modelo. Isso sugere que o modelo gamma log-linear está capturando bem a dispersão dos dados, com pouca heterogeneidade residual não explicada pelo modelo.

Tabela 1.6: Teste

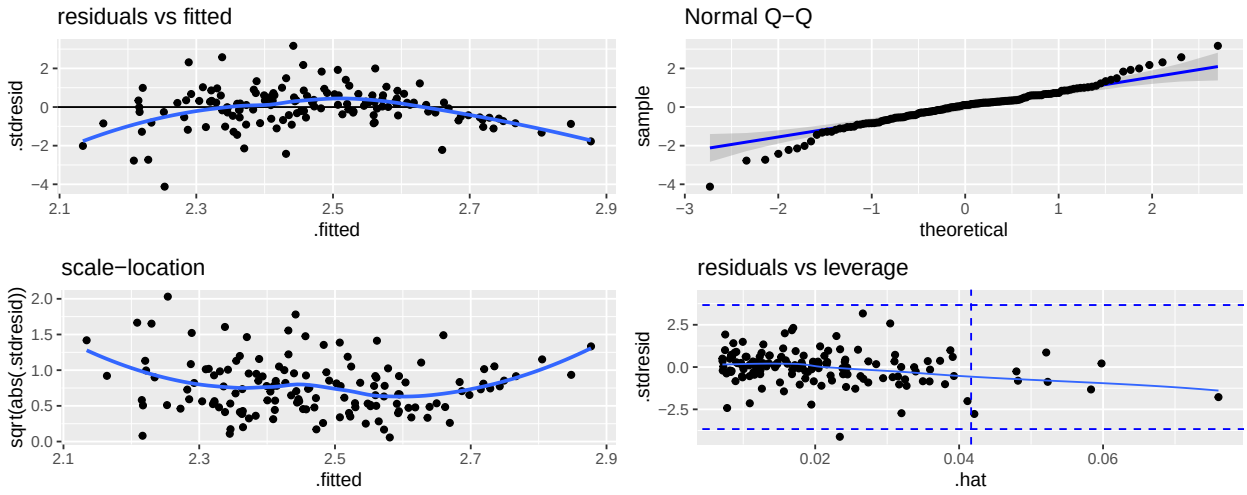
Descrição	Valor
Soma de alpha e beta	0.0838
Erro padrão	0.0019
Estatística t	-483.1625
Valor-p	7.4303651624535e-229
Decisão	Rejeitamos H0

Decisão sobre Teste $H0 : \alpha + \beta = 1$ contra $H1 : \alpha + \beta \neq 1$: (Tabela **??**) Como o valor-p é extremamente pequeno, rejeitamos a hipótese nula H_0 indicando que a soma de α e β é significativamente diferente de 1. Isso significa que a hipótese de retornos constantes à escala não é válida para os dados analisados.

Retornos constantes à escala implicam que, se ambos os insumos (L e K) aumentarem na mesma proporção, a produção Y_i também aumentará na mesma proporção. No entanto, dado que a soma de α e β é significativamente diferente de 1, isso sugere que os retornos à escala não são constantes para este modelo. Dependendo do valor específico da soma ($\alpha + \beta < 1$ ou $\alpha + \beta > 1$), os retornos à escala podem ser decrescentes ou crescentes, respectivamente.

Assim, a soma de α e β é 0.08383678, o que está longe de 1, indicando retornos decrescentes à escala. Isso significa que aumentar proporcionalmente a força de trabalho e o capital levará a um

Tabela 1.7: Gráfico de Análise de resíduos



aumento menos que proporcional na produção. Em outras palavras, conforme mais insumos são adicionados, a produtividade marginal desses insumos diminui, levando a um crescimento menos eficiente da produção.

1.2.2 Análise dos Resíduos

Tabela 1.8: Resultados dos Testes de Breusch-Pagan e Durbin-Watson

Teste	Estatística	Valor.p	Decisão
Breusch-Pagan	2.6679	0.2634	Não rejeitamos H0
Durbin-Watson	2.1338	0.7801	Não rejeitamos H0

Com base na análise dos resíduos, incluindo os testes de Breusch-Pagan e Durbin-Watson, podemos concluir o seguinte sobre o ajuste do modelo de regressão:

Homocedasticidade: - A variância dos resíduos parece ser constante em toda a faixa dos valores ajustados. Com o teste de Teste de Breusch-Paga, com um valor-p de 0.2634, não rejeitamos a hipótese nula de homocedasticidade. Isso significa que não há evidências suficientes para sugerir a presença de heterocedasticidade nos resíduos do modelo gamma.

Normalidade: - Embora o teste de Shapiro-Wilk sugira que os resíduos não sigam uma distribuição normal, desconsiderando caudas pesadas tanto nos extremos inferiores quanto nos superiores, podemos aproximar que os resíduos estão próximos de uma distribuição normal.

Independência dos Resíduos: - Tanto o teste de Durbin-Watson quanto a inspeção do gráfico de resíduos versus valores ajustados na tabela Tabela **??** não fornecem evidências significativas para rejeitar a hipótese nula de ausência de autocorrelação positiva ou negativa nos resíduos. Assim, parece que os resíduos são independentes entre si, o que é uma boa indicação de que o modelo está adequadamente ajustado quanto a essa suposição.

Linearidade: - Há um padrão claro nos resíduos plotados em relação aos valores ajustados, indicando não linearidade entre as variáveis independentes e dependentes.

1.3 Comparação dos Modelos: Gamma Log-Linear vs. Modelo Normal

Tabela 1.9: Comparação dos Modelos Gamma Log-Linear e Normal

Critério	Modelo.Gamma.Log.Linear	Modelo.Normal
Log-Likelihood	-83.07791	-59.39108
AIC	174.15580	126.78220
BIC	186.03510	138.66140

Log-Likelihood (Log-Verossimilhança) - **Modelo Gamma Log-Linear**: -83.07791 - **Modelo Normal**: -59.39108

O valor de log-likelihood mais alto (menos negativo) indica um melhor ajuste do modelo aos dados. O modelo normal tem um valor de log-likelihood mais alto, sugerindo que se ajusta melhor aos dados em comparação com o modelo Gamma log-linear.

Critério de Informação de Akaike (AIC) - **Modelo Gamma Log-Linear**: 174.1558 - **Modelo Normal**: 126.7822

Valores menores de AIC indicam um modelo mais eficiente em termos de ajuste aos dados com penalização pela complexidade. O modelo normal apresenta um AIC significativamente menor, indicando que é mais eficiente e se ajusta melhor aos dados do que o modelo Gamma log-linear.

Critério de Informação de Bayes (BIC) - **Modelo Gamma Log-Linear**: 186.0351 - **Modelo Normal**: 138.6614

O BIC também penaliza a complexidade do modelo e valores menores indicam um ajuste melhor com menor penalização por complexidade. Novamente, o modelo normal apresenta um BIC significativamente menor, reforçando que é mais adequado para os dados em comparação ao modelo Gamma log-linear.

Com base nos critérios de log-likelihood, AIC e BIC, o **modelo normal** é a melhor escolha. Ele não só se ajusta melhor aos dados (conforme indicado pelo log-likelihood), mas também é mais eficiente e simples em termos de complexidade (conforme indicado pelos valores de AIC e BIC).

Portanto, o modelo normal é preferível ao modelo Gamma log-linear para a análise dos seus dados, oferecendo um melhor equilíbrio entre ajuste e complexidade.

2 Dados Kaggle - análise de clssificação de crédito

2.1 Modelo Logistico

Para este estudo, utilizamos uma base de dados obtida do Kaggle, que contém informações detalhadas sobre clientes de uma instituição financeira. O objetivo principal é desenvolver um modelo de machine learning para classificar o crédito dos clientes em diferentes faixas de risco, otimizando os processos de decisão e reduzindo esforços manuais. A base de dados inclui 100.000 observações e 28 variáveis, como idade, ocupação, renda anual, número de contas bancárias e histórico de pagamentos.

A análise foca na construção e avaliação de um modelo de **regressão logística**, adequado para situações em que a variável dependente é categórica, como a classificação de crédito ("Poor" e "Good"). A **regressão logística** estima a probabilidade de um evento binário ocorrer e classifica as observações em categorias. A equação do modelo logístico é expressa como:

$$log(p/(1-p)) = \beta 0 + \beta 1 * X_1 + \beta 2 * X_2 + ... + \beta p * X_p$$

Onde:

- p é a probabilidade de um cliente ser classificado como "Good" (cliente ser um bom pagador).
- $1-p$ é a probabilidade de um cliente ser classificado como "Poor" (cliente possui um histórico problemático).
- $X_1, X_2, ..., X_p$ são as variáveis independentes (como histórico de pagamentos, dívida pendente, etc.)
- $\beta_0, \beta_1, ..., \beta_p$ são os coeficientes que medem o impacto de cada variável independente na probabilidade de ser classificado como "Good".

Como parte da análise, será explorado o ajuste do modelo, avaliando os resíduos para verificar a adequação do modelo.

Tabela 2.1: Resumo Descritivo das Variáveis

Variável	Estatística
Credit_Score	Good: 12192, Poor: 19884
Credit_Mix	Good: 13518, Standard: 8509, Bad: 10049

Outstanding_Debt	Min.: 0.23, 1st Qu.: 702.54, Median: 1360.45, Mean: 1593.26, 3rd Qu.: 2258.30, Max.: 4998.07
Payment_of_Min_Amount	No: 14432, Yes: 17644
Changed_Credit_Limit	Min.: -6.480, 1st Qu.: 4.570, Median: 8.560, Mean: 9.573, 3rd Qu.: 13.130, Max.: 35.820

A análise descritiva das variáveis na base de dados dta revela que, dos 32.076 clientes analisados, 12.192 foram classificados como "Good" e 19.884 como "Poor", indicando que a maioria dos clientes tem uma classificação de crédito ruim. A variável Credit_Mix mostra que a maioria dos clientes tem uma mistura de crédito considerada "Good", seguida por "Standard" e "Bad", o que pode influenciar a classificação final de risco. Em relação à dívida pendente, os valores variam significativamente, com uma mediana de 1.360,45 e uma média de 1.593,26, sugerindo que alguns clientes têm dívidas elevadas, o que pode aumentar o risco de crédito. Quanto ao pagamento do valor mínimo, 17.644 clientes o fizeram, enquanto 14.432 não, indicando que pagar apenas o valor mínimo pode ser um indicador de problemas financeiros. A variável Changed_Credit_Limit mostra variações entre -6,48 e 35,82, com uma mediana de 8,56, sugerindo mudanças no comportamento de crédito ao longo do tempo. Essas observações iniciais são fundamentais para a construção do modelo de regressão logística e devem orientar a escolha de variáveis e a interpretação dos resultados.