Modelos aditivos generalizados aplicados na eficiência energética de processadores

Universidade Federal da Paraíba - CCEN

Paulo Ricardo Seganfredo Campana

29 de abril de 2024

Índice

Introdução

Desde a década de 60 cientistas e engenheiros que trabalhavam na indústria de chips e circuitos integrados sabiam de uma coisa sobre o futuro (MOORE, 1965): o processo de fabricação dos mesmos só tem a melhorar, era observado que o número de componentes por circuito dobrava comparado com o ano anterior, isso ficou conhecido como a Lei de Moore.

Estes componentes se tornavam menores em tamanho físico e utilizavam menos energia. Isso combinado com a tendência de diminuição do custo de fabricação dos circuitos e aumento do rendimento do processo estar aumentando só indicava o crescimento da indústria, até mesmo nessa época já se imaginava conceitos do último século como carros autônomos e celulares.

Esta tendência exponencial durou muitas décadas porém era inevitável sua continuação perpétua, desvios da Lei de Moore foram avistados na década de 2010, em 2022 CEO da NVIDIA Jensen Huang disse que a lei já está morta (2022).

Ainda é possível hoje em dia a fabricação de processadores com cada vez mais transistores porém outra tendência vista na última década é o aumento do consumo de energia dos mesmos, em 2012 na segunda geração de processadores Intel Core, o processador mais poderoso lançado foi projetado para emitir 95 Watts de calor, em 2021 na década segunda geração o processador equivalente emite 240W.

O foco na diminuição do consumo de energia nos processadores será essencial para o avanço no poder de processamento das CPUs, computadores de mesa conseguem lidar com algumas centenas de Watts de calor emitido porém essa quantidade é inviável para notebooks e celulares e assim vemos disparidade de performance entre os dispositivos.

Assim este trabalho visa modelar quais características de um processador influenciam em sua eficiência energética e como influenciam, para isso será utilizado modelos de regressão aditivos generalizados, com um conjunto de dados de centenas de processadores lançados entre 2010 a 2020.

As características do modelo proposto serão então apresentadas em tabelas e gráficos como as medidas de acurácia do modelo, influência dos regressores na eficiência energética e teste de hipótese para significância do modelo e suas suposições.

Metodologia

Dados

O conjunto de dados provêm do site de entusiastas de hardware *TechPowerUp* e está disponível no pacote `openintro` do software R pelo nome `cpu`, ele contêm informações sobre 875 processadores lançados entre 2010 e 2020, é composto por 12 colunas, nas quais 4 são nominais, 7 são numéricas e uma temporal: a data do lançamento, as variáveis são as seguintes:

Empresa: Responsável pela tecnologia, design e lançamento do produto, são apenas duas empresas que nas últimas décadas competem na venda de processadores para o consumidor comum, Intel e AMD.

Nome: Nome do processador, é único para cada observação do banco e diz a respeito sobre o uso recomendado, a classe de performance e geração do processador.

Codinome: Nome de uso interno, único para cada geração de processadores de cada empresa.

Cores: Núcleos físicos de processamento contidos na CPU.

Threads: Núcleos virtuais de processamento, é comum ser o dobro do número de cores pela técnica de "hyper-threading", aumentando a capacidade de processamento paralelo.

Base Clock: Frequência de operação padrão do processador, medida em gigahertz (GHz), todas as instruções do processador seguem esse ritmo para sincronização de operações.

Boost Clock: Frequência de operação máxima do processador, é atingida temporariamente durante operações intensivas, está apenas disponível em alguns processadores que contém essa tecnologia.

Encaixe: Tipo de conexão física do processador, indica quais placas-mães são compatíveis com o produto.

Processo: Termo de marketing usado para designar a geração da tecnologia de fabricação do chip, é medido em nanômetros (nm) porém essa quantidade não possui relação certeira com as dimensões físicas das menores partes do chip.

Cache L3: Tamanho de armazenamento da terceira camada de cache do processador, medido em megabytes (MB), é usado para acelerar a transferência de dados entre o processador e a memória RAM.

TDP: Do inglês "Thermal Design Power" é a quantidade máxima de energia dissipada pelo processador como calor, medida em Watts (W), é usado como indicação do tipo de sistema de refrigeração recomendado.

Lançamento: Data de lançamento do processador.

A seguir está a estrutura das primeiras 8 observações do conjunto de dados em ordem decrescente de lançamento:

Lançamento	Empresa	Nome	Codinome	Encaixe	Processo
2020-05-27	Intel	Core i3-10100	Comet Lake	1200	14
2020-05-27	Intel	Core i3-10300	Comet Lake	1200	14
2020-05-27	Intel	Core i3-10320	Comet Lake	1200	14
2020-05-27	Intel	Core i3-10350K	Comet Lake	1200	14
2020-05-27	Intel	Core i5-10400	Comet Lake	1200	14
2020-05-27	Intel	Core i5-10400F	Comet Lake	1200	14
2020-05-27	Intel	Core i5-10500	Comet Lake	1200	14
2020-05-27	Intel	Core i5-10600	Comet Lake	1200	14

Cores	Threads	Base Clock	Boost Clock	Cache L3	TDP
4	8	3,6	4,3	6	65
4	8	3,7	4,4	8	62
4	8	3,8	4,6	8	91
4	8	4,1	4,8	8	91
6	12	2,9	4,0	12	65
6	12	2,9	4,3	12	65
6	12	3,1	4,6	12	65
6	12	3,3	4,8	12	65

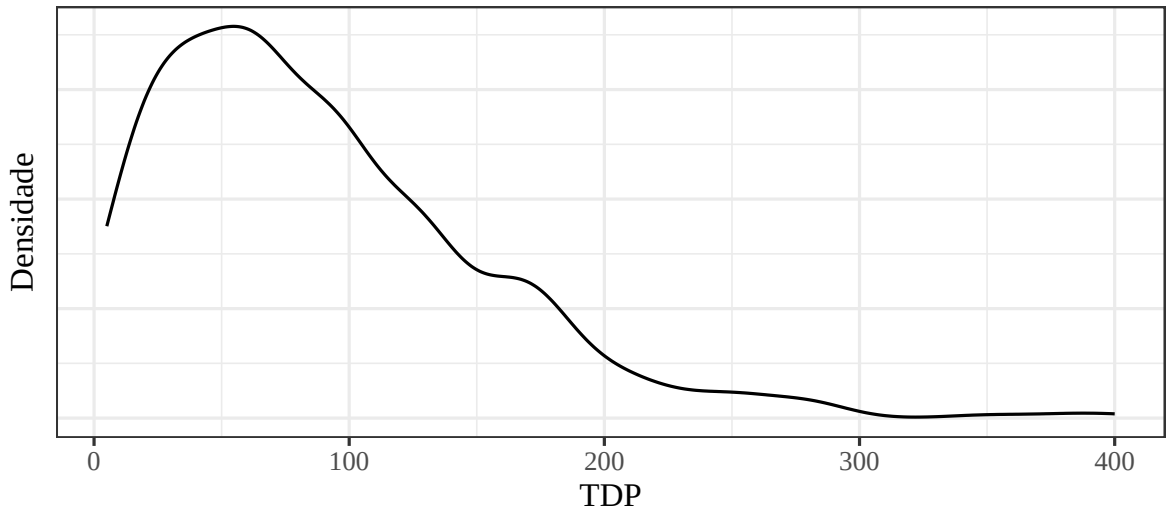
Todas as variáveis foram testadas para inclusão no modelo final, exceto o nome, codinome, encaixe e data de lançamento do processados pois além de não exercerem influência direta na eficiência energética do processador, são variáveis nominais que possuem muitos valores únicos, inviabilizando a utilização de métodos como variáveis *dummy* para serem inclusas no modelo de regressão.

Além disso, alguns processadores mais antigos não possuem as tecnologias de *Boost Clock* e terceira camada de cache, para o uso destes regressores no modelo, foram excluídos do banco de dados os processadores sem estas informações, assim o banco de dados usado para modelagem terá 472 observações.

Modelo aditivo generalizado

Quando a variável resposta do modelo de regressão linear não tem distribuição normal, podemos partir para modelos lineares generalizados, que produzem melhores inferências do que transformações na variável resposta. A eficiência energética dos processadores não possui distribuição normal como pode ser visto no gráfico da Figura 1:

Figura 1: Densidade da distribuição da variável resposta do modelo, a eficiência energética do processador.



Porém quando a contribuição de cada regressor na variável resposta não tem relação linear, este modelo não é valido, uma boa alternativa é o uso dos modelos aditivos generalizados, principalmente quando não conhecemos a formula da relação exata para o uso de um modelo de regressão não linear.

O modelo aditivo generalizado tem a seguinte forma:

$$g(\mu) = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + \varepsilon$$

Onde $g(\mu)$ representa a função de ligação entre a média da variável resposta e os preditores, β_0 representa um intercepto e as funções $f_1, f_2, \cdots$ são funções não paramétricas que podem ser estimadas ou especificadas como lineares, e atuam em sua variável regressora correspondente.

Na prática, estas funções são suavizações baseadas nos dados que podem assumir formas como suavização por regressão local, *b-splines*, *splines* cúbicas ou polinômios ortogonais.

Validação

Foram usados três métricas para avaliação da performance do modelo: raiz do erro quadrático médio (REQM), erro médio absoluto (EMA) e coeficiente de determinação (R^2) com suas formas dadas a seguir, onde $\hat{y}$ são as previsões do modelo e $\hat{\rho}$ é o coeficiente de correlação de Pearson.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \qquad \text{EMA} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \qquad R^2 = \hat{\rho}^2_{y\hat{y}}$$

As previsões usadas no cálculo das métricas provem da separação do conjunto de dados em uma parte usada para ajuste do modelo e outra para estimação das métricas, utilizando a abordagem de validação cruzada com 10 *folds* e 10 repetições, assim obtemos não só uma estimativa pontual para as métricas como também erro padrão e intervalo de confiança.

Também é necessário que o modelo cumpra algumas suposições como a normalidade dos resíduos, ausência de autocorrelação nos resíduos além de que a contribuição de cada regressor seja significante para obtermos um modelo válido. Tudo isso será verificado com testes de hipótese. Os resíduos utilizados para os testes são os resíduos de desvio definidos por

$$r_{D_i} = \text{sng}(y_i - \hat{\mu}_i)\sqrt{2}\sqrt{y_i(\hat{\theta}_i^0 - \hat{\theta}_i) + b(\hat{\theta}_i) - b(\hat{\theta}_i^0)}$$

Os resultados abaixo foram obtidos no software estatístico e linguagem de programação R (R CORE TEAM, 2024) e estão disponíveis no apêndice deste documento, utilizando os pacotes `tidymodels` (KUHN; WICKHAM, 2020) para validação cruzada e estimação das métricas de performance, e `mgcv` (WOOD, 2011) que implementa os modelos aditivos generalizados. O relatório foi feito no sistema de escrita e publicação científica Quarto (ALLAIRE *et al.*, 2022).

Resultados

Especificação

Pelo gráfico de densidade da variável resposta na Figura 1, vemos que está é uma variável com distribuição positiva e assimétrica, portanto escolhi a distribuição Gama para a parte generalizada do modelo aditivo, e foi usado a função de ligação logarítmica pois apresentou melhores resultados de performance do modelo, assim temos um modelo em que o logaritmo da esperança condicional da eficiência energética é explicada por uma combinação aditiva entre regressores, o intercepto e um erro com distribuição normal.

$$\log \mu = \beta_0 + f_1(x_1) + f_2(x_2) + \cdots + \varepsilon$$

O tipo de função suavizadora usada foi *splines* isotrópicas de baixo posto, o padrão da função `s()` do pacote `mgcv` pois de certa forma é a melhor suavização segundo WOOD (2003). Essa suavização é aplicada nos regressores *cores*, *base clock*, *boost clock* e *cache L3*, não foi aplicada no regressor *processo* pois esta apresenta uma relação com a variável resposta que reduz a uma estimação linear nas *splines*.

O regressor *empresa* foi transformado em variável *dummy* sem interação para uso no modelo, pois há diferença significativa na eficiência energética de processadores Intel e AMD nesta década.

Performance

Fazendo o ajuste do modelo com a validação cruzada, temos as estimativas pontuais, os erros padrões das estimativas e intervalos de confiança para as medidas de desempenho do modelo aditivo generalizado dadas na Tabela 1.

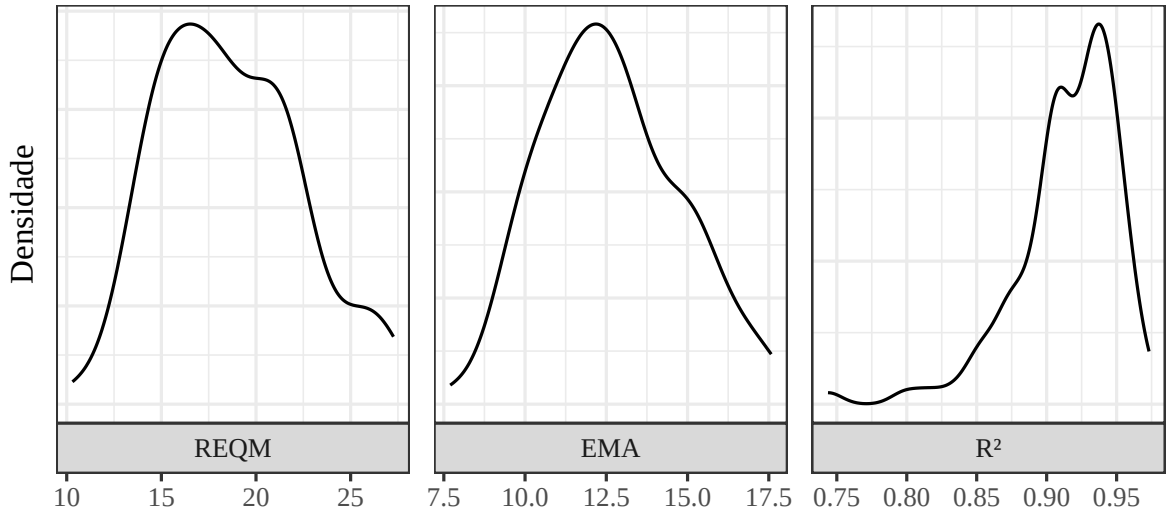
Tabela 1: Métricas de performance do modelo por validação cruzada.

			Intervalo de Confiança	
Métrica	Média	Erro padrão	Inferior	Superior
REQM	18,604	0,371	17,867	19,341
EMA	12,630	0,212	12,209	13,051
R^2	0,914	0,004	0,907	0,922

Assim vemos que o modelo utilizado se adéqua bem ao modelar a eficiência energética dos processadores pois o erro médio absoluto estimado de 12,6 é pequeno comparado com a escala de consumo de energia dos processadores, além do coeficiente de determinação R^2 ser bastante alto, acima de 90%.

Vemos também pelo gráfico de densidade das medidas na Figura 1 que não há grandes outliers na estimativa das métricas, isso indica que o modelo é capaz de fazer previsões boas para diferentes tipos de processadores e também que não ocorre *overfitting* para certos tipos.

Figura 1: Densidade da distribuição amostral das métricas de performance do modelo.



Modelo final

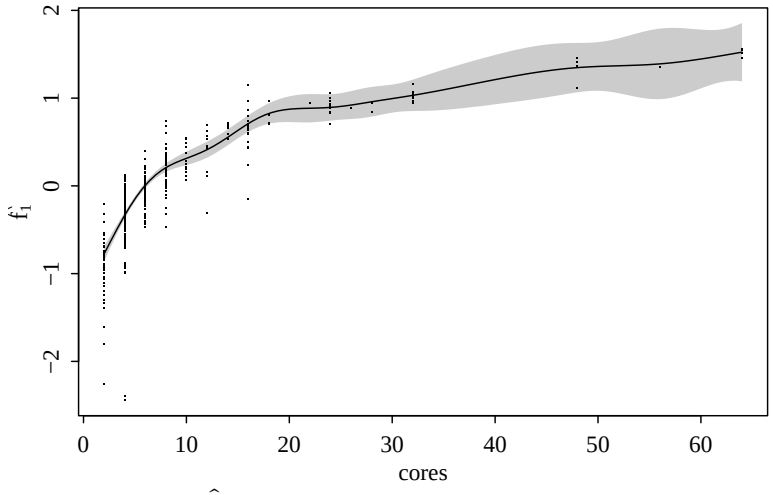
Ajustando o modelo novamente, agora com todas as observações, obtemos as equações da relação entre a eficiência energética dos processadores e as variáveis regressoras utilizadas, as duas equações se referem a processadores Intel e AMD devido ao uso de transformação em variável *dummy*.

$$\log \hat{\mu}_{\text{Intel}} = 3{,}965 + 0{,}021\text{processo} + \hat{f}_1(\text{cores}) + \hat{f}_2(\text{base clock}) + \hat{f}_3(\text{boost clock}) + \hat{f}_4(\text{cache L3})$$
$$\log \hat{\mu}_{\text{AMD}} = 3{,}653 + 0{,}021\text{processo} + \hat{f}_1(\text{cores}) + \hat{f}_2(\text{base clock}) + \hat{f}_3(\text{boost clock}) + \hat{f}_4(\text{cache L3})$$

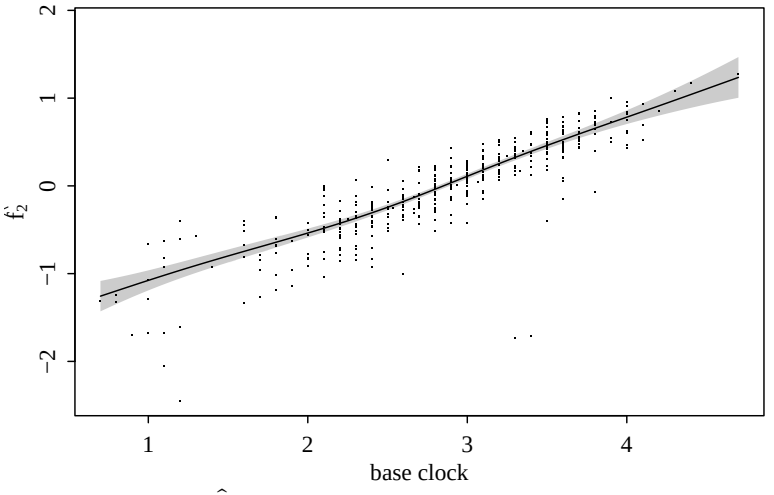
Onde as funções $\hat{f}_1$, $\hat{f}_2$, $\hat{f}_3$ e $\hat{f}_4$ foram estimadas e estão apresentadas na Figura 2, a área sombreada representa a região de confiança da função e os pontos do gráfico são os resíduos parciais de cada observação.

Figura 2: Funções não paramétricas estimadas

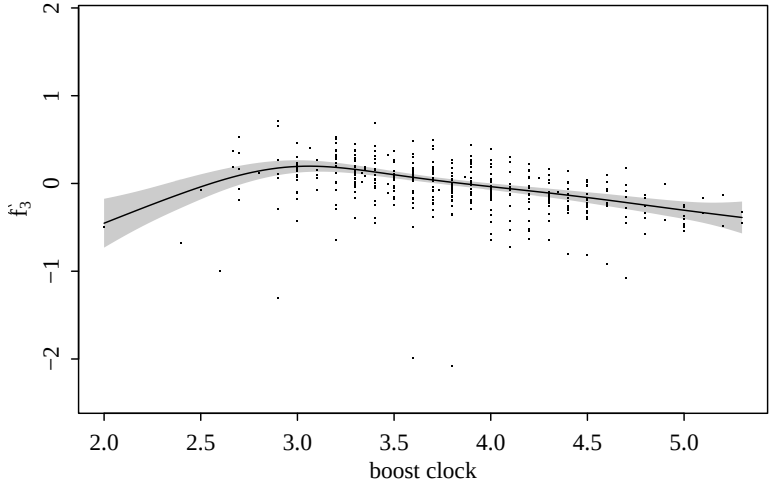
(a) $\widehat{f}_1$ associada ao regressor *cores*.



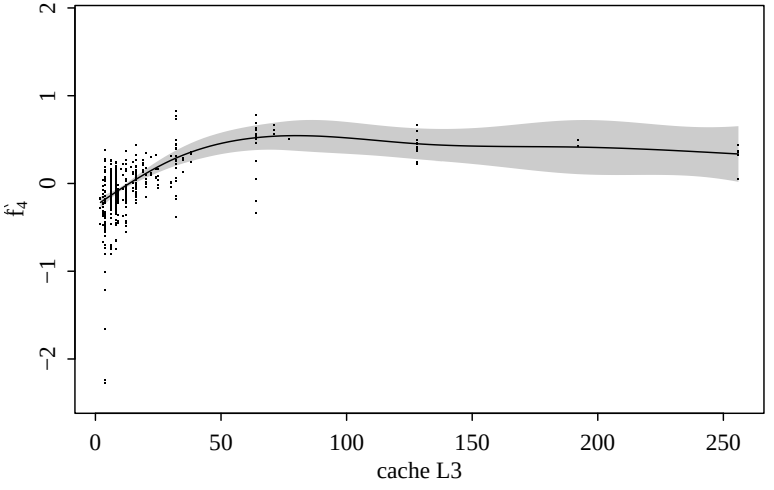
(b) $\widehat{f}_2$ associada ao regressor *base clock*.



(c) $\widehat{f}_3$ associada ao regressor *boost clock*.



(d) $\widehat{f}_4$ associada ao regressor *cache L3*.



Interpretação

Pelo mesmo gráfico da Figura 2 é possível interpretar a contribuição de cada regressor pois representa o efeito parcial da variável na eficiência energética do processador, ou seja, o efeito quando é fixado o valor dos outros regressores.

- Na Figura 2a vemos que quanto mais núcleos físicos de processamento o CPU tem, maior é a demanda energética do processador, porém a relação não é linear e se estabiliza para processadores com grande quantidade de núcleos.
- Na Figura 2b a relação é quase linear entre a frequência de operação padrão e a demanda energética do processador pois para rodar um processador em maior frequência, é necessário uma maior voltagem para mantê-lo estável e isso aumenta o consumo de energia.
- Na Figura 2c o efeito é inverso: processadores com maior frequência máxima exibem melhor eficiência energética, isso acontece pois esta tecnologia permite que o processador funcione em uma velocidade mais baixa para uso comum, salvando as frequências mais altas com maiores demandas energéticas para momentos oportunos, diminuindo o consumo geral.
- Na Figura 2d a relação também é não linear, para valores baixos da capacidade do cache L3, a relação entre a capacidade e a demanda energética é crescente, porém processadores com capacidade muito alta (maior que 64MB) não observam esse aumento no consumo de energia.

O coeficiente para o regressor *processo* foi de 0,021, isso indica que os chips mais novos fabricados em processos menores em tamanho exibem menor consumo máximo de energia, este coeficiente nos diz que em média um processador fabricado no processo de 32nm consome 70% mais energia do que outro de processo 7nm.

Adicionalmente, o intercepto de 3,965 pra processadores Intel e 3,653 para a marca AMD significa que, pela transformação inversa da função de ligação, os processadores Intel emitem cerca de 36% mais energia na forma de calor comparados com a marca AMD segundo o modelo.

Significância

A Tabela 2 e Tabela 3 abaixo mostram os resultados dos testes de hipótese para significância de cada regressor do modelo, para testar os coeficientes da parte linear do modelo foi usado o teste t comparando o coeficiente com o valor 0, para a significância das funções não paramétricas estimadas é usado os precedimentos descritos em MARRA; WOOD (2012).

Tabela 2: Testes t para significância dos coeficientes lineares

Coeficiente	Estimativa	Erro padrão	Estatística t	p-valor
Intercepto Intel	3,965	0,0345	106,0	$< 0{,}001$
Intercepto AMD	3,653	0,0325	122,1	$< 0{,}001$
Processo	0,021	0,0016	13,2	$< 0{,}001$

Tabela 3: Testes para significância dos regressores suavizados

Coeficiente	Graus de liberdade efetivo	Estatística F	p-valor
Cores	7,47	52,8	$< 0{,}001$
Base Clock	3,51	209,0	$< 0{,}001$
Boost Clock	5,18	13,1	$< 0{,}001$
Cache L3	4,11	13,2	$< 0{,}001$

Assim, todos os regressores do modelo são altamente significantes, os graus de liberdade utilizados pela estimativa das suavizações *spline* são relativamente baixos, o suficiente para não ocorrer *overfitting* no modelo como visto nos gráficos da Figura 2.

Suposições

Com os resíduos de desvios obtidos do modelo, os resultados de diferentes testes de normalidade e autocorrelação dos mesmos estão dados na Tabela 4 e Tabela 5 respectivamente.

Tabela 4: Testes para normalidade dos resíduos de desvio

Teste	Hipótese nula	Estatística	p-valor
Anderson-Darling	Normalidade	$A = 4,398$	$< 0{,}001$
Lilliefors	Normalidade	$D = 0,075$	$< 0{,}001$
Shapiro-Wilk	Normalidade	$W = 49,492$	$< 0{,}001$
χ^2 de Pearson	Normalidade	$P = 0,962$	$< 0{,}001$

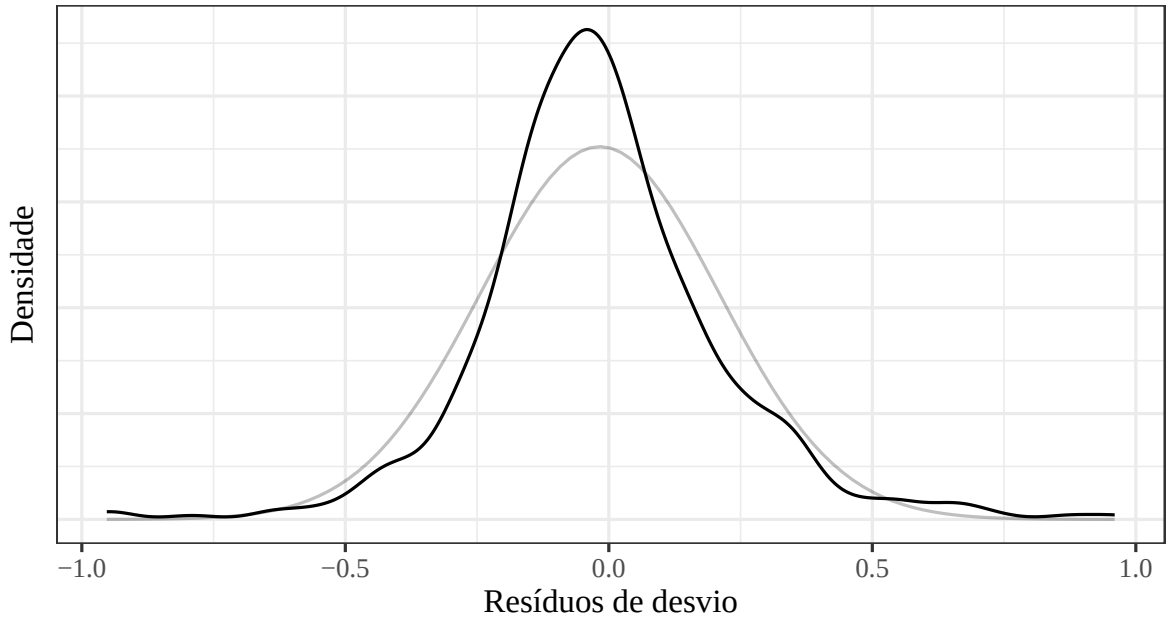
Tabela 5: Testes para autocorrelação dos resíduos de desvio

Teste	Hipótese nula	Estatística	p-valor
Breusch-Godfrey	sem autocorrelação	$BG = 155,862$	$< 0{,}001$
Durbin-Watson	sem autocorrelação	$DW = 0,862$	$< 0{,}001$

Os testes da Tabela 5 rejeitam a hipótese nula indicando que os resíduos do modelo são sim autocorrelacionados, isso é esperado pois os processadores do banco de dados estão ordenados por data de lançamento e através dos anos houve mudança estatisticamente significativa na eficiência energética dos processadores.

O mesmo acontece com a normalidade, todos os testes rejeitam a hipótese de normalidade dos resíduos de desvio, a distribuição verdadeira parece ter uma curtose maior como visto no gráfico de densidade dos resíduos da figura Figura 3.

Figura 3: Densidade da distribuição dos resíduos de desvio do modelo.



O impacto do não cumprimento destas duas suposições é que a estimativa dos coeficientes lineares não é a mais eficiente e também que os testes de hipóteses pela distribuição t agora requerem um tamanho de amostra suficiente para que os coeficientes tenham distribuição normal pelo teorema central do limite.

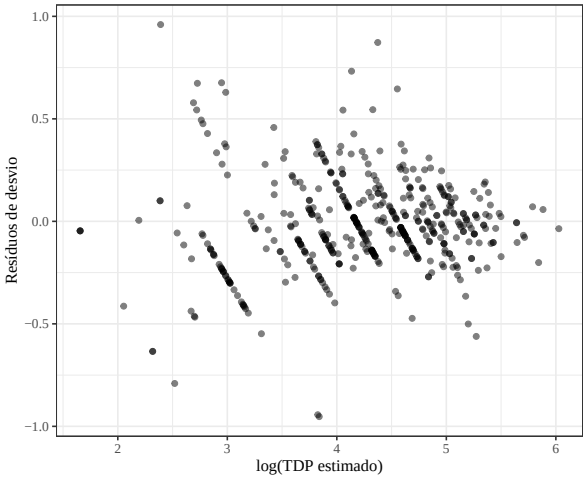
Diagnóstico

Para verificar se a escolha da distribuição utilizada no modelo foi correta, olhamos para o gráfico de resíduos versus valores estimados $\widehat{\mu}_i$ ou versus a componente sistemática do modelo $g(\widehat{\mu}_i) = \log \widehat{\mu}_i$, esperamos que os resíduos não sejam influenciados pelos valores estimados nem em média ou variância para um modelo com distribuição adequada.

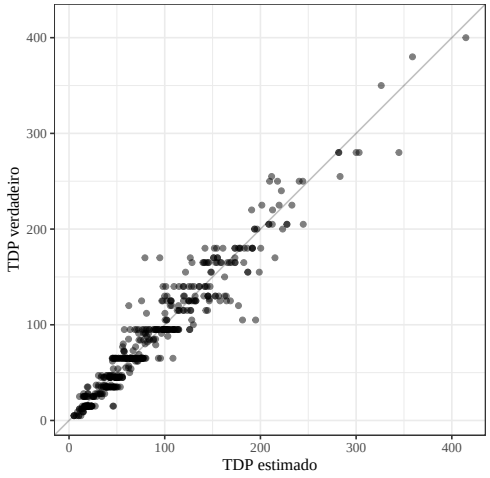
Já para validação da função de ligação é possível ver para o R^2 do modelo ou pelo gráfico de valores estimados $\hat{\mu}_i$ versus valores observados y_i, um modelo com função de ligação adequada apresenta R^2 significante e uma disposição de pontos no gráfico próximas da reta $y = x$.

Figura 4: Gráficos de diagnóstico de especificação do modelo.

(a) Relação entre os resíduos e a componente sistemática.



(b) Relação entre os valores estimados e observados.



É difícil notar alguma relação entre as medidas na Figura 4a, as extremidades laterais da componente sistemática aparentam ter menor variância nos resíduos de desvio porém isso pode ser devido a menor quantidade de observações nestas extremidades.

O coeficiente de determinação R^2 do modelo é bastante alto como visto na Tabela 1 e pelo gráfico da Figura 4b temos uma relação linear entre a eficiência energética verdadeira e a estimada pelo modelo. Assim a escolha da distribuição Gama e função de ligação logarítmica é adequada.

Conclusão

Com essas informações, vimos que a especificação do modelo é adequada em relação ao modelo aditivo pois os regressores contribuem não linearmente na eficiência energética e em relação a escolha da distribuição, função de ligação e seleção de variáveis do modelo visto que os resíduos estão bem distribuídos para diferentes tipos de processadores e todas os regressores do modelo são altamente significativos.

O modelo também apresenta boas medidas de acurácia, com erro médio absoluto estimado de 12,6 comparado com a média de 100W no TDP dos processadores e também correlação acima de 90% entre os valores verdadeiros e estimados pelo modelo.

Infelizmente o modelo não cumpre as suposições de normalidade e falta de autocorrelação dos resíduos de desvio e com isso as estimativas dos coeficientes não são as mais eficientes.

Pelos gráficos da Figura 2 onde vemos a contribuição de cada regressor no consumo de energia nos leva a pensar quais as características que lavariam a um processador de alta eficiência energética, uma possível abordagem seria um processador fabricado em um processo moderno, com baixa frequência base porém alta frequência *boost*, usando o modelo para prever a eficiência energética de um processador com as seguintes características:

- Empresa: AMD,
- Processo: 3nm
- *Cores*: 6
- *Base Clock*: 2GHz
- *Boost Clock*: 5GHz
- *Cache L3*: 32MB

É estimado pelo modelo que este processador atinga um TDP de apenas 24W, bastante abaixo da média porém mantendo uma performance muito boa para uso geral do processador, principalmente em notebooks onde um baixo consumo de energia é mais importante.

Apêndice

```r
library(tidyverse)
library(tidymodels)
library(kableExtra)

options(digits = 3)
set.seed(0)

theme_kek <- theme_bw() +
    theme(text = element_text(family = "Times"))

data <- openintro::cpu |>
    as_tibble() |>
    drop_na(l3_cache, boost_clock) |>
    arrange(released)

folds <- vfold_cv(data, v = 10, repeats = 10)

model <- gen_additive_mod() |>
    set_engine("mgcv", family = Gamma(link = "log")) |>
    set_mode("regression")

recipe <- recipe(tdp ~ ., data)

workflow <- workflow() |>
    add_model(
        model,
        formula = tdp ~ company + process - 1 +
        s(cores) + s(base_clock) + s(boost_clock) + s(l3_cache)
    ) |>
    add_recipe(recipe)

fit <- workflow |>
    fit_resamples(
        folds,
        metrics = metric_set(rmse, mae, rsq)
```

```
    )

collect_metrics(fit)

mgcv <- workflow |>
    fit(data) |>
    extract_fit_engine()

mgcv::plot.gam(
    mgcv,
    residuals = TRUE, shade = TRUE, rug = FALSE
)

# diferença multiplicativa entre processo 32nm e 7nm
exp(0.021 * (32 - 7))
# diferença multiplicativa entre processadores intel e amd
exp(3.965 - 3.653)

summary(mgcv)
r <- residuals(mgcv)

nortest::ad.test(r)
nortest::lillie.test(r)
shapiro.test(r)
nortest::pearson.test(r)

lmtest::bgtest(mgcv)
lmtest::dwtest(mgcv)

# tdp do processador sugerido
predict(mgcv, newdata = data.frame(
    company = "AMD", process = 3,
    cores = 6, l3_cache = 32,
    base_clock = 2, boost_clock = 5
)) |> exp()
```

Referências

ALLAIRE, J. J. *et al.* Quarto. 2022. Disponível em: <https://quarto.org>.

KUHN, M.; WICKHAM, H. **Tidymodels: a collection of packages for modeling and machine learning using tidyverse principles.** [S.l.]: [s.n.], 2020.

MARRA, G.; WOOD, S. N. Coverage Properties of Confidence Intervals for Generalized Additive Model Components. **Scandinavian Journal of Statistics**, 2012. v. 39, n. 1, p. 53–74.

MOORE, G. E. Cramming more components onto integrated circuits. **Electronics Magazine**, 1965. v. 38, n. 8, p. 114.

R CORE TEAM. **R: A Language and Environment for Statistical Computing**. Vienna, Austria: R Foundation for Statistical Computing, 2024.

WITKOWSKI, W. "Moore's Law's dead" Nvidia CEO Jensen Huang says in justifying gaming-card price hike. 2022. Disponível em: <https://www.marketwatch.com/story/moores-laws-dead-nvidia-ceo-jensen-says-in-justifying-gaming-card-price-hike-11663798618>.

WOOD, S. N. Thin Plate Regression Splines. **Journal of the Royal Statistical Society (B)**, 2003. v. 65, n. 1, p. 95–114. Disponível em: <https://doi.org/10.1111/1467-9868.00374>.

_____. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. **Journal of the Royal Statistical Society (B)**, 2011. v. 73, n. 1, p. 3–36.