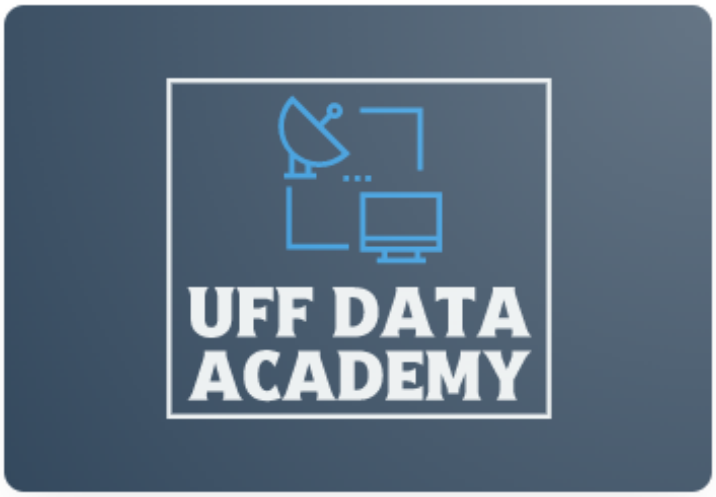
Universidade Federal Fluminense



Gerência de Redes e Engenharia de Tráfego

Professor Cledson de Sousa

Versão: 0.2.340– 13 de maio/2024 – Até MPLS.

Conteúdo

PREFÁCIO

Caros futuros engenheiros,

Bem-vindos ao curso de Gerência de Redes e Engenharia de Tráfego, uma disciplina criada para introduzir os conceitos fundamentais e práticas avançadas da controle de tráfego. administração e gerência de modernas infraestruturas de rede. Este curso, estruturado com uma mistura equilibrada de aulas em classe e extra-classe, cobre uma carga horária total de 60 horas, das quais 40 horas são dedicadas a conceitos teóricos e 20 horas a aplicações fora de sala. Então o objetivo é proporcionar aos alunos uma compreensão abrangente dos modelos de gerência de redes, monitoramento, auditoria e a engenharia de tráfego necessária para lidar com o encaminhamento do tráfego.

Esta apostila, embora tenha um caráter acadêmico, não possui a pretensão de ser extremamente rigorosa, especialmente na forma. O autor empenhou-se em dar o devido crédito a todas as fontes utilizadas; no entanto,por vezes, estende-se o texto sem as devidas citações específicas dos autores originais. Então peço que aceitem as escusas do autor, pois muitas das fontes são os textos originais presentes nos locais onde as figuras foram extraídas. Além disso, diversas explicações são fruto do esforço do autor em dissecar os diagramas, que frequentemente simplificam o verdadeiro encaminhamento dos pacotes.

Ao longo do texto há diversos sinais de parada  , em notas de margem, gráficos, figuras e outros sinais. Estes sinais lá estão para o leitor desopilar, se você chegou em um desses sinais parabéns! É por que você já leu o suficiente, nesse momento é bom parar, descontrair, visitar os sites e sair um pouco do texto, naturalmente deve-se retornar ao texto tão logo possível. Essa estratégia

funciona comigo! Um curso leve, produtivo, enriquecedor e sem reprovações. É o que desejo!

Sobre o autor:

O autor recebeu seu título de Doutor em Computação pela Universidade Federal Fluminense (UFF) em 2019, e seus títulos de graduação e mestrado em Engenharia de Telecomunicações pela mesma universidade, em 1997 e 2013. Com mais de 30 anos de experiência na indústria de telecomunicações, desde 2021 é Professor Adjunto do Departamento de Engenharia de Telecomunicações da Universidade Federal Fluminense. Seus interesses de pesquisa atuais incluem redes de sensores sem fio, SDN, rádios cognitivos e CSI.

Plataformas:

- LinkedIn: `https://www.linkedin.com/in/cledsonsousa`
- Página pessoal: `https://cledsonsousa.github.io`
- Currículo Lattes: `http://lattes.cnpq.br/7195080748145566`

1 Introdução

1.1 O CONCEITO DE CAMADAS

O conceito de camadas (*layering*) é fundamental para a visualização de dados, Vamos explorar como esse conceito facilita essas representações:

Separação de Elementos Visuais:
Camadas permitem separar diferentes elementos visuais em uma visualização de dados. Por exemplo, ao representar um conjunto de dados, uma camada pode ser usada para mostrar os valores médios, enquanto outra camada pode ser usada para mostrar a variação desses valores (como barras de erro ou intervalos de confiança). Isso ajuda a distinguir claramente entre os dados principais e as informações auxiliares que podem por exemplo representar incerteza.

Flexibilidade na Visualização:
O uso de camadas oferece flexibilidade na criação de gráficos complexos. Diferentes tipos de representações visuais, como pontos de dados, linhas de tendência, e intervalos de confiança, podem ser adicionados ou removidos conforme necessário. Isso permite que os visualizadores ajustem a complexidade da visualização com base em suas necessidades específicas.

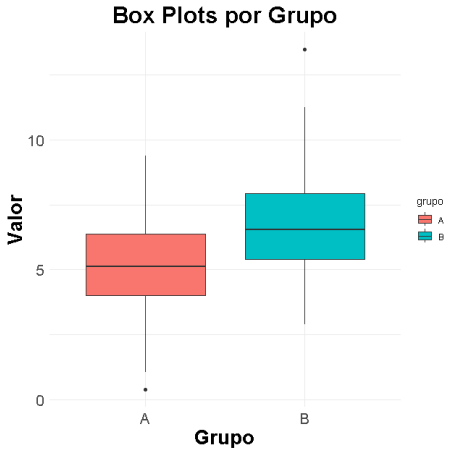


Figura 1

Exemplo Prático:
Box plots como o apresentados na Figura 1, por exemplo, utilizam camadas. A camada do *box plot* mostra a mediana e os quartis, enquanto camadas adicionais indicam *outliers*.

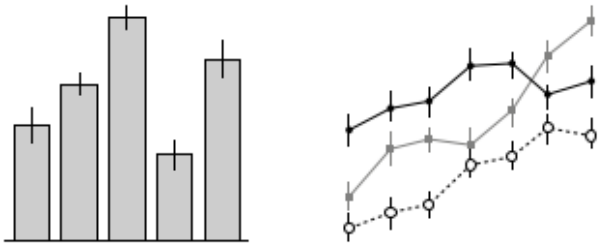
1.2 VARIAÇÃO E INCERTEZAS

O conceito de camadas também facilita a representação da variação estatística e da incerteza. É importante se lembrar que dados experimentais tipicamente apresentam alguma incerteza e a visualização e representação destes dados deve "caracterizar a magnitude dessa incerteza em relação aos dados reais [1].

Os autores infelizmente mostram resultados que em muitos trabalhos as figuras publicadas não atendem a esse padrão, especialmente à medida que a dimensionalidade dos dados aumenta [2]. Quando o objetivo de uma visualização é comparar uma quantidade medida ou derivada entre categorias ou condições, deve-se incluir um elemento (são as *geom* do `ggplot`) gráfico retratando a quantidade e um segunda elemento retratando a incerteza da quantidade. Note que, sem a representação da incerteza, uma comparação visual precisa não é possível, os leitores podem tirar conclusões incorretas ou mal informadas.

A variação e a incerteza podem ser retratadas com uma variedade de *geoms*, mas são mais comumente exibidas com barras de erro. Infelizmente, não há um padrão único para o que a barra de erro deve representar já que há uma pluralidade esmagadora de significados possíveis, como desvio padrão (DP) da amostra, erro padrão da média (EPM) ou simplesmente erro padrão, Intervalo de Confiança (IC) paramétrico de $100(1-\alpha)\%$, intervalo de probabilidade bayesiano, um intervalo de previsão, etc. Cada quantidade tem sua própria interpretação estatística

Portanto, ao usar barras de erro, certifique-se de que (1) a quantidade codificada pela barra é consistente com o objetivo da visualização e (2) a quantidade é definida de forma inequívoca. Em relação ao primeiro ponto, oferecemos as seguintes diretrizes ao usar barras de erro para retratar a variação de uma estimativa de parâmetro ou a variação dos dados.



(a) Gráficos de barra com barra de erros.

(b) Gráficos de linhas com barra de erros.

Figura 2: as barras normalmente para comparar categorias e as linhas para comparar grupos e tendências.

1.2.1 Intervalo de Confiança e Nível de Significância

Se o interesse for estimar um parâmetro populacional, como a média ou a variância, então a variação da estimativa (isto é, a distribuição amostral da estatística) é desejada. Exemplos de barras de erro adequadas incluem o EPM ou um IC paramétrico ou *bootstrap* de 95%, como visto em visualizações que enfatizam comparações (Figura 2). ICs paramétricos devem ser usados apenas se os dados atenderem às suposições do modelo subjacente, caso contrário, um *bootstrap* (ou outra estratégia para aproximar a distribuição amostral) deve ser usado[1].

[1]E o que é α? o nível de significância: α é a probabilidade de cometer um erro tipo I, que ocorre quando rejeitamos a hipótese nula (H_0) quando ela é verdadeira. **Em outras palavras, é a taxa de falso positivo permitida.**

IC um conceito muitas vezes mal entendido [3] e [4]
Um intervalo de confiança de $100(1-\alpha)\%$ para um parâmetro populacional é um intervalo calculado a partir dos dados amostrais que, em $100(1-\alpha)\%$ das amostras possíveis, conteria o verdadeiro valor do parâmetro.
Isto é: Suponha que você calcule a média de alturas de uma amostra de 100 pessoas e obtenha uma média de 170 cm com um desvio padrão de 10 cm. Um intervalo de confiança de 95% pode ser calculado, e você pode obter algo como (168 cm, 172 cm). Isso significa que, se você repetisse esse experimento várias vezes, 95% dos intervalos de confiança calculados conteriam a verdadeira média populacional (nesse caso 170 cm).

Um outro exemplo de interpretação geométrica dos intervalos de confiança são a sua sobreposição ou não entre duas amostras.

A sobreposição dos ICs [5] e [6]
Um intervalo de confiança de $100(1-\alpha)\%$ para um parâmetro populacional é um intervalo calculado a partir dos dados amostrais que, em $100(1-\alpha)\%$ das amostras possíveis, conteria o verdadeiro valor do parâmetro.
Isto é: Suponha que você calcule a média de alturas de uma amostra de 100 pessoas e obtenha uma média de 170 cm com um desvio padrão de 10 cm. Um intervalo de confiança de 95% pode ser calculado, e você pode obter algo como (168 cm, 172 cm). Isso significa que, se você repetisse esse experimento várias vezes, 95% dos intervalos de confiança calculados conteriam a verdadeira média populacional (nesse caso 170 cm).

Enquanto ICs de 95% não sobrepostos indicam uma diferença significativa (sob um modelo de probabilidade normal), o inverso não é verdadeiro — dependendo do tamanho da amostra, as barras de IC podem se sobrepor em até 50% e ainda atender aos critérios de significância [5].

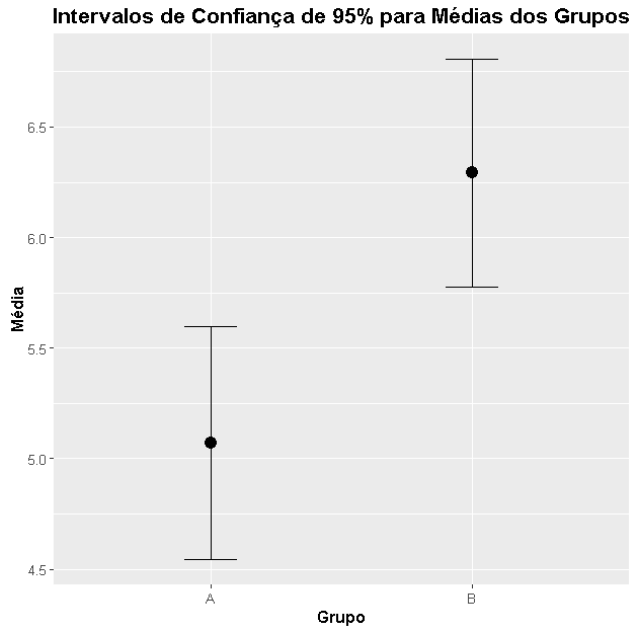


Figura 3: ICs de 95% não sobrepostos indicam uma diferença significativa (sob um modelo de probabilidade normal).

As barras de erro são destinadas a indicar a faixa de valores prováveis para alguma estimativa ou medição. Elas se estendem horizontalmente e/ou verticalmente a partir de um ponto de referência que representa a estimativa ou medição. No entanto o erro ou o IC podem ser mostrado de várias maneiras, como por pontos barras ou faixas. Barras de erro graduadas mostram múltiplas faixas ao mesmo tempo, onde cada faixa corresponde a um grau diferente de confiança. Elas são, na prática, múltiplas barras de erro com diferentes espessuras de linha plotadas umas sobre as outras [7].

Os gráficos apresentados na Figura 4 são três diferentes visualizações que utilizam intervalos de confiança para representar a incerteza e variabilidade dos dados. a Figura 4a) mostra um Gráfico de Dispersão com Banda de Confiança: As bandas de confiança servem para mostrar a incerteza na relação entre duas variáveis. Já a Figura 4b) do Gráfico de Meio-Violino mostra a metade da distribuição dos dados para diferentes categorias, destacando a densidade dos dados e os intervalos interquartis. E o da Figura 4c), o Gráfico de Violino Completo, apresenta uma visão completa da distribuição dos dados, facilitando uma comparação detalhada entre as categorias.

Esses gráficos são úteis para entender não apenas os valores médios, mas também a variabilidade, a distribuição e a incerteza dos dados, proporcionando uma compreensão mais completa e detalhada das características dos dados analisados.

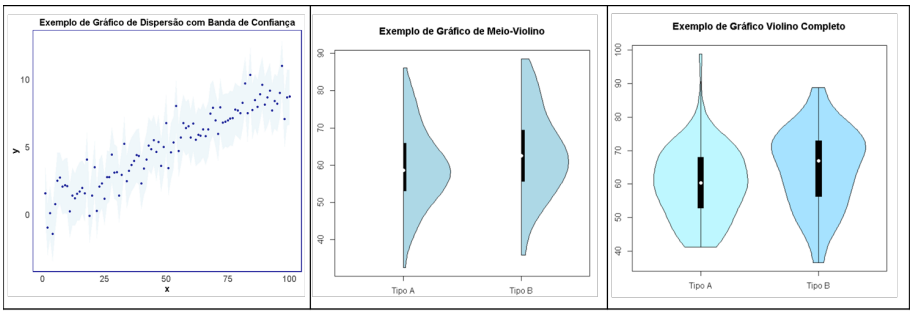


Figura 4: três diferentes visualizações que utilizam intervalos de confiança para representar a incerteza e variabilidade dos dados.

2 Introdução a Análise de Dados

Experimentos na nossa área (Engenharia) tipicamente nos inundam com dados muitas vezes difíceis de entender e quiçá explicar. O objetivo aqui é escrutinar o método, a lógica, a arte e a prática da análise e representação de dados, fornecendo as habilidades e ferramentas essenciais para examinar dados e resolver problemas. Concordamos com a filosofia de "aprender fazendo"para uma melhor compreensão da análise de dados[2].

[2]Este capítulo foi fortemente baseado em [8] de David Brown (ver Figura 5) 👇 .

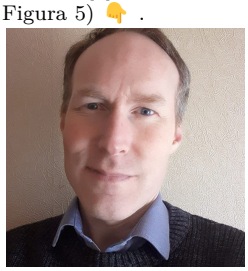


Figura 5: Mr. Brown Tem mais de 20 anos de experiência como avaliador estatístico na MHRA e, antes do Brexit, foi membro do grupo de trabalho de bioestatística e do grupo de trabalho de aconselhamento científico da EMA. Ele fez parte do grupo que formulou a orientação recentemente publicada da MHRA sobre dados do mundo real.

2.1 CASOS REAIS

Ao perceber um aumento na quantidade de propagandas de fraldas, fórmulas infantis e macacões da Target, o destinatário liga para a Target perguntando por que estão enviando tanta publicidade voltada para bebês (já fazia anos que não havia um bebê em casa). A Target explicou que os dados recentes de compras indicavam que havia uma mulher grávida na residência. Uma semana após ligar para a Target, ele descobre que sua filha está grávida.

Em um cenário como os dos surtos de H1N1, Ebola, Zika ou COVID-19, o peíodo comum de 2 semanas de análise é longo demais. Uma equipe do Google permitiu que a Internet descobrisse onde os surtos ocorriam. Para desenvolver seu modelo, eles rastrearam a disseminação do H1N1 e correlacionaram com termos de busca na Internet: febre alta, tosse e dores. Informando as autoridades quando e onde exatamente o novo surto de gripe estava ocorrendo.

Para ajudar a reduzir o crime em Chicago. Equipados com sensores que localizam disparos de armas pela cidade, junto com mapas de lojas de bebidas alcoólicas e acessos de rodovias, os analistas identificam áreas onde o crime provavelmente ocorrerá. Essa informação, combinada com dados sobre eventos esportivos televisionados, aumenta a precisão na localização de possíveis problemas[3].

[3]Esses e outros casos podem ser encontrados em [9].

2.2 COMPONENTES DA ANÁLISE DE DADOS

O foco está no processo de formação de hipóteses, teste de teorias e obtenção de inferências. Vamos abordar essas bases de investigação por meio da visualização de dados e trabalhando em problemas reais com dados reais.

São quatro os componentes da análise de dados nessa ordem:

1. Descrição de dados e formulação de hipóteses.
2. Construção e estimativa de modelos.
3. Diagnósticos.
4. Próxima pergunta.

Existem múltiplos conceitos e técnicas (ou seja, construção e estimativa de modelos, transformação de variáveis, diagnósticos etc.). O propósito deste capítulo é introduzir as linhas gerais estratégias de uma boa análise de dados com um exemplo.

Descrição de Dados e Formulação de Hipóteses:

Descrever dados significa, a princípio, identificar o caso típico (**tendência central**) e entender quão típico é esse caso típico (dispersão). No entanto com as novas ferramentas, deve-se ir muito além disso. Significa, entender, correlacionar e encontrar padões. E as hipóteses? Para nós uma hipótese se referirá a uma suposição específica sobre como duas coisas estão relacionadas (por exemplo, probabilidade de acesso ao meio e vazão total).

Construção e Estimativa de Modelos:

Modelos são versões simplificadas da realidade que nos ajudam a entender nosso mundo complexo. São argumentos para explicar um problema empírico. Por exemplo, se queremos explicar por que alguns países têm altas taxas de homicídio, construímos um modelo que pode incluir renda, idade da população, número de policiais e eficácia do sistema judiciário. Há uma infinidade de outras possíveis causas que poderíamos incluir, mas é útil manter as coisas simples, mas em engenharia não queremos recriar a realidade; queremos apenas aproximá-la.

Diagnósticos:
Depois de construirmos modelos e obtermos algumas estimativas, passamos para os diagnósticos. Diagnósticos são um conjunto de ferramentas que usamos para determinar se estamos usando o tipo certo de modelo. Para verificar se nosso modelo é apropriado, examinamos quão bem as previsões do nosso modelo correspondem à realidade. A diferença entre nossa previsão e a realidade é chamada de erro residual. Por exemplo, se nosso modelo faz um bom

trabalho ao prever a vazão total (em bps) em todas as redes, exceto nas redes sem fio em áreas urbanas densas, os diagnósticos resultantes dirão isso. Ou seja, os resíduos para esses casos serão relativamente grandes. Talvez nossas estimativas de modelo estejam sendo excessivamente influenciadas por essas redes sem fio urbanas. Diagnósticos nos ajudam a determinar se nossas estimativas fornecem uma boa noção de como a vazão de rede realmente funciona, são o produto de alguns casos atípicos ou são o resultado de um modelo mal escolhido.

É importante lembrar que diagnósticos podem tanto detectar problemas quanto ajudar a descobrir relações interessantes, gerando explicações adicionais ou hipóteses.

As Próximas Perguntas:
se as estimativas que obtivemos estão corretas, então esperaríamos ver nossa variável acompanhar certo comportamento. Seguir cada conjunto de estimativas com essa declaração ajuda a descobrir explicações possíveis e hipóteses adicionais a serem testadas. Como é impossível provar qualquer coisa com total certeza, o exercício de gerar hipóteses adicionais para testar é extremamente importante.

2.3 DESCREVENDO E FORMULANDO HIPÓTESES

Os testes de hipóteses constituem uma pedra angular, o teste de hipótese trata de determinar a probabilidade de que uma determinada premissa sobre um conjunto de dados seja verdadeira. É um método usado para validar ou refutar suposições, muitas vezes levando a novos *insights* e entendimentos. Na sua essência, envolve a formulação de duas hipóteses concorrentes: a hipótese nula (H_0) e a hipótese alternativa (H_1).

A hipótese nula, H_0, representa uma crença básica. É uma afirmação de nenhum efeito ou nenhuma diferença, como "Não há diferença nas alturas médias entre duas espécies de plantas". Em contrapartida, a hipótese alternativa, H1, representa o que procuramos estabelecer. É uma afirmação de efeito ou diferença, como "Há uma diferença significativa nas alturas médias entre estas duas espécies".

Para decidir entre essas hipóteses, usamos um valor p, uma estatística crucial no teste de hipóteses. O valor p nos diz a probabilidade de observar nossos dados, ou algo mais extremo, se a hipótese nula fosse verdadeira. Um valor p (*p-value*) baixo (geralmente abaixo de 0,05) sugere que os dados observados são improváveis sob a hipótese nula, levando-nos a considerar a hipótese alternativa.

No contexto do teste de correlação de Pearson, o valor p desempenha um papel fundamental na determinação da significância estatística da correlação observada entre duas variáveis. O teste de Pearson avalia a força e a direção da relação linear entre duas variáveis contínuas. Ao realizar este teste, calculamos o coeficiente de correlação de Pearson (r), que pode variar de -1 a 1. No

entanto, para inferir se a correlação observada é estatisticamente significativa, analisamos o valor p associado.

Quando o valor p é baixo, indica que a probabilidade de obter um coeficiente de correlação tão extremo quanto o observado, se a hipótese nula de correlação zero fosse verdadeira, é pequena. Por exemplo, um valor p menor que 0,05 sugere que há menos de 5% (0.05) de chance de a correlação observada ser devida ao acaso, fornecendo evidências contra a hipótese nula e a favor de uma correlação verdadeira entre as variáveis. Portanto, a interpretação do valor p no teste de Pearson nos ajuda a decidir se podemos rejeitar a hipótese nula e aceitar a hipótese alternativa de que existe uma correlação significativa.

No entanto, o teste de hipóteses não está isento de riscos, nomeadamente erros do Tipo I e do Tipo II. Um erro Tipo I, ou falso positivo, ocorre quando rejeitamos incorretamente uma hipótese nula verdadeira. Por exemplo, concluir que um novo medicamento é eficaz quando não o é, seria um erro do Tipo I. Este tipo de erro pode levar a uma falsa confiança em tratamentos ou intervenções ineficazes.

Por outro lado, um erro Tipo II, ou falso negativo, ocorre quando não conseguimos rejeitar uma hipótese nula falsa. Isto seria como não reconhecer a eficácia de um medicamento benéfico. Os erros do tipo II podem levar à perda de oportunidades de intervenções ou tratamentos benéficos.

O valor crítico é o ponto de corte que determina a fronteira entre a região onde rejeitamos a hipótese nula (H_0) e a região onde não a rejeitamos, com base no nível de significância (α) do teste. É calculado de modo que a probabilidade de cometer um erro do Tipo I (rejeitar H_0 quando H_0 é verdadeira) seja igual a α. Por exemplo, em um teste unilateral com nível de significância de 5% ($\alpha = 0.05$), o valor crítico na distribuição normal padrão seria aproximadamente 1.645. Se a estatística de teste exceder este valor crítico, rejeitamos H_0, caso contrário, não a rejeitamos.

O equilíbrio entre esses erros é crucial. O nível de significância, muitas vezes fixado em 0,05, ajuda a controlar a taxa de erros do Tipo I. No entanto, a redução dos erros do Tipo I pode aumentar a probabilidade de erros do Tipo II. Assim, a análise estatística não consiste apenas na aplicação de uma fórmula; requer uma consideração cuidadosa do contexto, dos dados e das implicações potenciais de ambos os tipos de erros.

A programação R, com seu conjunto abrangente de ferramentas estatísticas, simplifica a aplicação de testes de hipóteses. Ele não apenas realiza os cálculos necessários, mas também auxilia na visualização dos dados, o que pode fornecer *insights* adicionais. Através do R, podemos executar com eficiência vários testes de hipóteses, desde testes t simples até análises mais complexas, tornando-o uma ferramenta inestimável tanto para estatísticos quanto para analistas de dados.

Em resumo, o teste de hipóteses é um método, que se observado com rigor e correção, suporta a decisão dentro de fatores repetíveis. Requer uma compreensão de conceitos estatísticos como hipóteses nulas e alternativas, valores p e os tipos de erros que podem ocorrer.

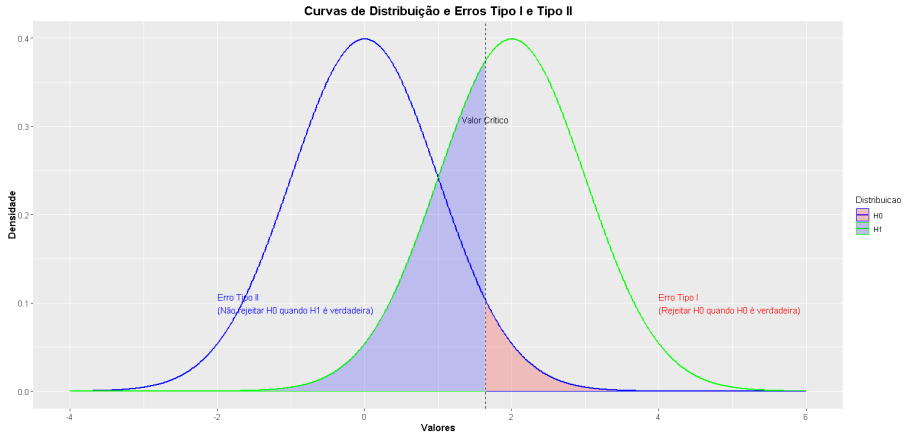


Figura 6: duas curvas de distribuição de probabilidades, uma para a hipótese nula (H_0) e outra para a hipótese alternativa (H_1). As áreas sombreadas em vermelho e azul ilustram os erros do Tipo I e do Tipo II, respectivamente.

2.3.1 Teste de Hipótese – Um Exemplo Prático

Nesta seção, demonstraremos como conduzir um teste de hipóteses[4] no R usando um conjunto de dados do mundo real. Exploraremos o conjunto de dados `PlantGrowth`,[5] incluído no R, que contém dados sobre o peso das plantas sob diferentes condições de crescimento. Nosso objetivo será determinar se há uma diferença estatisticamente significativa no crescimento das plantas entre dois grupos de tratamento.

[4] O teste apresentado na Seção 2.3.1 foi extraído de [10].

[5] O R possui pacotes *libraries* e *datasets* para auxiliar a comunidade a aprender e ensinar. o `PlantGrowth` é um dentre tantos.

Formulação da Hipótese:

A hipótese nula (H_0) é a manutenção do *status quo*, ou seja, afirma que não há diferença no crescimento médio das plantas entre os dois grupos. A hipótese alternativa (H_1) postula que existe uma diferença significativa.

Conduzindo o Teste de Hipótese:
Precisamos descobrir o números de grupos do *dataset* e em seguida aplicar de um teste t para comparar os pesos médios das plantas entre dois dos grupos que se quer comparar. Este teste é apropriado para comparar as médias de dois grupos independentes.

Passo 1: instalar os pacotes apropriados ver Listagem 2.1.

```
1 >install.packages("easystats")
2 >library(easystats)
```

Código 2.1: comando para instalar e carregar a bibliotrca easystat.

Passo 2: exibição da estrutura e sumário dos dados, ver Listagem 2.2.

```
> head (PlantGrowth)
weight group
1   4.17  ctrl
2   5.58  ctrl
3   5.18  ctrl
4   6.11  ctrl
5   4.50  ctrl
6   4.61  ctrl
> summary (PlantGrowth)
weight       group
Min.   :3.590   ctrl:10
1st Qu.:4.550   trt1:10
Median :5.155   trt2:10
Mean   :5.073
3rd Qu.:5.530
Max.   :6.310
```

Código 2.2: comando para exibição da estrutura e sumário do *dataframe* no R.

Passo 4: formulando as hipóteses e conduzindo o teste da hipótese.

Nossa hipótese nula (H_0) afirma que não há diferença no crescimento médio das plantas entre os dois grupos (*status quo*). A hipótese alternativa (H_1) postula que existe uma diferença significativa. E aplicação do teste [6] t, para comparar os pesos médios das plantas entre dois dos grupos. Este teste é apropriado para comparar as médias de dois grupos independentes.

[6] O teste t é empregado para comparar as médias de dois grupos e determinar se elas são significativamente diferentes uma da outra. O teste t é baseado na distribuição t de Student e é especialmente útil quando as amostras são pequenas e a variância é desconhecida. E Amostras independentes são aquelas em que as observações de uma amostra não têm qualquer relação com as observações da outra amostra.

```
> result <- t.test(weight ~ group,
> data = PlantGrowth,
> subset = group %in% c("ctrl", "trt1"))
```

Código 2.3: aplicação do test t para comparação do subgrupo (ctrl e trt1) no R.

Na Listagem 2.3, `weight` é a variável de interesse (dependente) e `group` é a variável que define os grupos (a variável independente). E a comparação deve se dar entre o grupo de controle e um segundo que queremos avaliar.

Passo 5: avaliação dos resultados, ver Listagem 2.4.

```
> report(result)
Effect sizes were labelled following Cohen (1988) recommendations.

The Welch Two Sample t-test testing the difference of weight by group
(mean in group ctrl = 5.03, mean in group trt1 = 4.66) suggests that
the effect is positive, statistically not significant, and
medium (difference = 0.37, 95% CI [-0.29, 1.03], t(16.52) = 1.19,
p = 0.250; Cohen d = 0.59, 95% CI [-0.41,1.56])
```

Código 2.4: apresentação dos resultados.

A função `result` gera um relatório teste t, incluindo a estimativa, o intervalo de confiança e o valor p.

Interpretando e visualizando os Resultados A saída da função de relatório nos dirá se a diferença nas médias é estatisticamente significativa. Um valor p menor que 0,05 geralmente indica que a diferença é significativa, e podemos

rejeitar a hipótese nula em favor da alternativa. No entanto, se o valor p for maior que 0,05, não temos evidências suficientes para rejeitar a hipótese nula.

De modo que, olhando para nossos resultados, podemos avaliar que há uma certa diferença na medida que estamos verificando, mas de acordo com o valor de p alto, nos habilita a dizer que essa diferença pode ser simplesmente uma questão de acaso, não sendo estatisticamente significativa [7]. Acompanhar a tradução gráfica do resultado[8] na Figura 7. O código está na Listagem 2.5.

```
>library(ggplot2)
>ggplot(PlantGrowth, aes(x = group, y = weight)) +
geom_boxplot() +
theme_minimal() +
labs(title = "Crescimento das Plantas por Grupo de  Tratamento",
x = "Grupo",
y = "Peso")
```

Código 2.5: apresenta os resultados em forma gráfica.

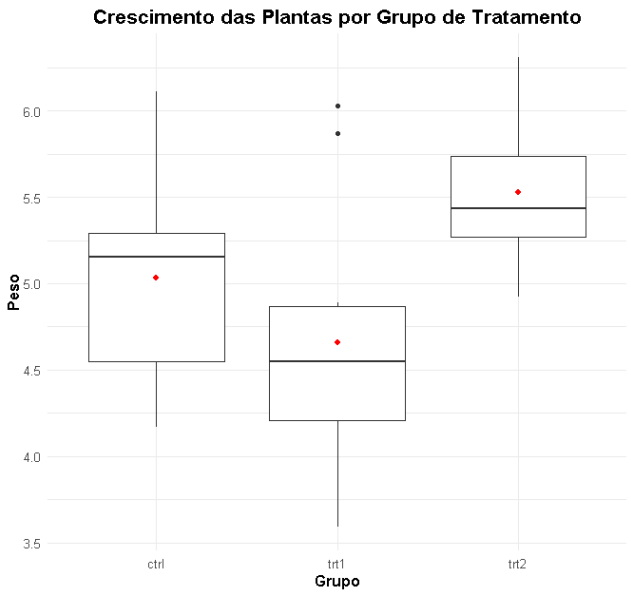


Figura 7: o *boxplot* exibe a média, a mediana, os quartis e possíveis *outliers*.

É possível concluir que os grupos ctrl e trt1 realmente não têm grande diferença, seus intervalos superam um ao outro [9].

2.4 CONSTRUÇÃO E ESTIMATIVA DE MODELOS

Um objetivo comum para o desenvolvimento de um modelo é prever qual será o valor de saída de um sistema para conjunto de valores de entrada. A ideia é discutir como desenvolver o modelo, como avaliar até que ponto o modelo criado se ajusta aos dados e como interpretar os resultados.

[7] **Significância Estatística:** se o valor p é menor que o nível de significância escolhido (por exemplo, 0,05), rejeitamos a hipótese nula em favor da hipótese alternativa, indicando que a diferença observada é estatisticamente significativa.

Não Significância Estatística: se o valor p é maior que o nível de significância, não rejeitamos a hipótese nula, indicando que não há evidências suficientes para afirmar que a diferença observada é estatisticamente significativa.

[8] **Explore Antes de Testar:** Familiarize-se com seu conjunto de dados antes de realizar testes de hipótese.
Verifique as Assunções: Cada teste estatístico tem suposições (como normalidade, independência ou variância igual).
Escolha o Teste Correto: diferentes testes são projetados para diferentes tipos de dados e objetivos. Por exemplo, use um teste t para comparar médias, testes qui-quadrado para dados categóricos e ANOVA para comparar mais de dois grupos.
Considere Opções Não Paramétricas: Se seus dados não atenderem às suposições dos testes paramétricos (são normais?).

[9] Que tal como exercício tentar verificar o par ctrl e trt2?

Suponha que medimos o desempenho de vários dispositivos computacionais. Podemos organizar as $n \times k$ medições, mostradas na Tabela 1 Como medimos o desempenho de n dispositivos diferentes, obteremos n linhas na tabela. Cada linha é chamada de "observação"única.

Tabela 1: Um exemplo em que queremos prever o desempenho de novos sistemas $n+1, n+2$ e $n+3$ usando o medido anteriormente resultados dos outros n sistemas [11].

System	**Clock (MHz)**	**Cache (kB)**	**Transistors (M)**	**Output Performance**
1	1500	64	2	98
2	2000	128	2.5	134
⋮	⋮	⋮	⋮	⋮
i	...	...	...	...
⋮	⋮	⋮	⋮	⋮
n	1750	32	4.5	113
$n+1$	2500	256	2.8	?
$n+2$	1560	128	1.8	?
$n+3$	900	64	1.5	?

O objetivo da modelagem é usar essas k medidas independentes para determinar uma função f, que descreva as relações entre os parâmetros de entrada e a saída, por exemplo $desempenho = f(clock, cache, transistors)$. Um dito modelo de regressão pode assumir qualquer forma. Nos restringiremos a uma função que é uma combinação linear (regressão linear) dos parâmetros de entrada. Mas note que, embora a função seja linear, os parâmetros em si não precisam ser lineares.

2.5 A REGRESSÃO LINEAR - SLR

A regressão linear é um método estatístico utilizado para modelar a relação entre uma variável dependente (**resposta**) e uma ou mais variáveis independentes (**preditores**). A equação da regressão linear múltipla pode ser expressa da seguinte forma[10]:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{1}$$

onde:

- Y é a variável resposta (dependente).
- $X_1, X_2, \ldots, X_p$ são as variáveis preditoras (independentes).
- β_0 é o intercepto.
- $\beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes de regressão associados às variáveis preditoras.

[10]Em notação matemática, letras maiúsculas denotam variáveis aleatórias ou conjuntos de dados:
– Y representa a variável resposta como um conjunto de observações.
– $X_1, X_2, \ldots, X_p$ representam as variáveis preditoras como conjuntos de observações.
Na prática, o intercepto (β_0) representa o ponto onde a linha de regressão cruza o eixo y no gráfico de dispersão.

- ϵ é o termo de erro, que captura a variação não explicada pelo modelo.

Os coeficientes de regressão (β_i) são estimados de forma a minimizar a soma dos quadrados dos resíduos (diferença entre os valores observados e os valores preditos pelo modelo).

$$\min_{\beta_0,\beta_1,\ldots,\beta_p} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \tag{2}$$

onde $\hat{Y}_i$ é o valor predito pelo modelo para a i-ésima observação.

O primeiro passo no processo de modelagem com um único preditor é determinar se parece haver uma relação entre o preditor e o valor de saída. Com base no conhecimento sobre projeto de dispositivos computacionais, sabemos que a frequência do *clock* influencia fortemente o desempenho do sistema.De forma que é esperada uma correlação entre o desempenho do processador[11] e sua frequência de *clock*.

[11]O *Integer Component of* SPEC CPU2000 (Componente Inteiro do SPEC CPU2000), cujo gráfico está apresentado na Figura 9 é uma parte do conjunto de *benchmarks* desenvolvido pela *Standard Performance Evaluation Corporation* (SPEC) para avaliar o desempenho de CPUs em tarefas que envolvem cálculos inteiros. O SPEC CPU2000 é um conjunto de testes que mede o desempenho de sistemas de computação, e é dividido em dois componentes principais: o *integer* (inteiro) e o floating-point (ponto flutuante).

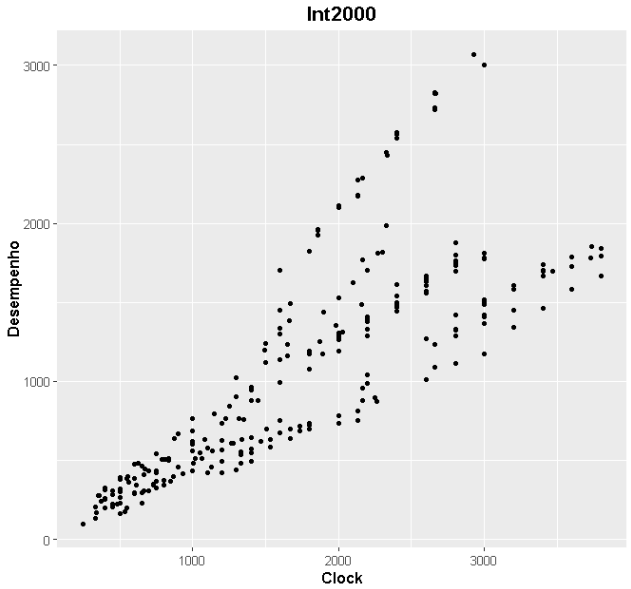


Figura 8: Um gráfico de dispersão do desempenho dos processadores que testamos usando o *benchmark* Int2000 versus a frequência do *clock*.

[12]

A varável independente neste caso é o *clock* e a variável dependente é o desempenho. Se sobrepusermos uma linha reta a este gráfico de dispersão, vemos que há uma aparente relação entre o preditor (a frequência do *clock*) e a saída (o desempenho). O gráfico ainda mostra que esta relação não é perfeitamente linear. À medida que a frequência do *clock* aumenta, vemos uma maior espalhamento em valores de desempenho. Nosso próximo passo

[12]O *Integer Component of* SPEC CPU2000 (Componente Inteiro do SPEC CPU2000) é uma parte do conjunto de *benchmarks* desenvolvido pela *Standard Performance Evaluation Corporation* (SPEC) para avaliar o desempenho (Dhrystones por segundo[12])de CPUs em tarefas que envolvem cálculos inteiros. O SPEC CPU2000 é um conjunto de testes que mede o desempenho de sistemas de computação, e é dividido em dois componentes principais: o *integer* (inteiro) e o *floating-point* (ponto flutuante).

é desenvolver uma regressão modelo que nos ajudará a quantificar o grau de linearidade na relação entre a saída e o preditor.

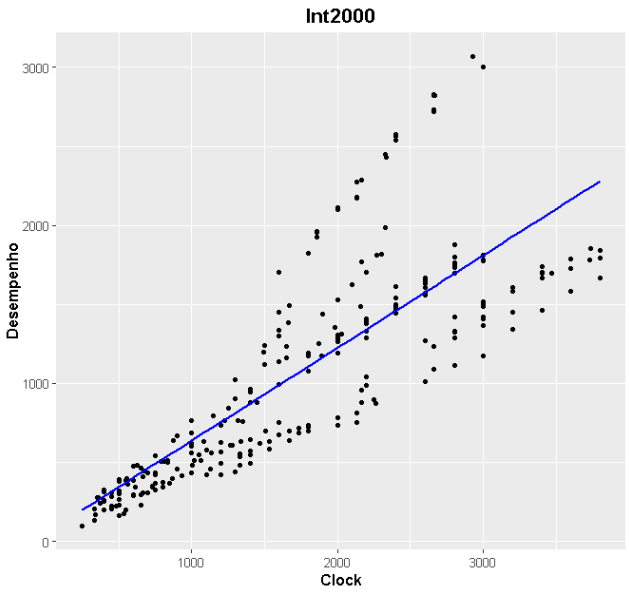


Figura 9: o modelo de regressão linear simples sobreposto aos dados da Figura 9.

As informações que obtemos digitando o comando `int00.lm` alguns valores básicos do modelo, mas não nos diz nada sobre as qualidades do modelo.

```
> summary(int00.lm)

Call:
lm(formula = perf ~ clock, data = int00.dat)

Residuals:
Min      1Q  Median      3Q     Max
-634.61 -276.17  -30.83   75.38 1299.52

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 51.78709   53.31513   0.971    0.332
clock        0.58635    0.02697  21.741   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘’. 0.1 ‘ ’ 1

Residual standard error: 396.1 on 254 degrees of freedom
Multiple R-squared:  0.6505,    Adjusted R-squared:  0.6491
F-statistic: 472.7 on 1 and 254 DF,  p-value: < 2.2e-16
```

Código 2.6: apresentação dos resultados do comando `summary(int00.lm)`.

2.5.1 Análise dos Resultados da Regressão Linear

Primeiramente vamos dissecar os valores dos resíduos.

Resíduos:

- **Min:** -634.61
- **1Q:** -276.17
- **Mediana:** -30.83
- **3Q:** 75.38
- **Max:** 1299.52

Esses valores indicam a distribuição dos resíduos. A mediana (-30.83) próxima de de zero (em comparação ao valor máximo 1299.52) sugere que os resíduos estão aproximadamente balanceados em torno de zero, o que é um bom sinal. No entanto, a diferença entre os valores mínimo e máximo é bastante grande, indicando a presença de *outliers*.

Coeficientes:

- ***Residual standard error*:** 396.1 on 254 *degrees of freedom*
- ***Multiple R-squared*:** 0.6505,
- ***Adjusted R-squared*:** 0.6491
- ***F-statistic*:** 472.7 on 1 and 254 DF,
- ***p-value*** $<$ 2.2e-16
- **R-quadrado múltiplo:** 0.6505
- **R-quadrado ajustado:** 0.6491
- ***F-value*:** 472.7 on 1 and 254 DF, p-valor: $<$ 2.2e-16

O valor de R-quadrado indica que aproximadamente 65.05% da variação na performance pode ser explicada pela variação no *clock*. Isso sugere um bom ajuste do modelo, mas ainda há 34.95% da variação que não é explicada pelo modelo linear.

No contexto de um modelo de regressão linear, o F-valor é uma medida estatística que testa a significância global do modelo. Ele é usado para avaliar se existe uma relação linear entre a variável dependente e as variáveis independentes no modelo.

Os graus de liberdade (DF, do inglês *Degrees of Freedom*) são importantes para o cálculo do *F-value*. No caso "1 and 254 DF", o primeiro número (1) representa os graus de liberdade do numerador, que correspondem ao número de variáveis independentes no modelo (neste caso, apenas uma variável independente: clock). O segundo número (254) representa os graus de liberdade do denominador, que são baseados no número de observações menos o número de parâmetros estimados (neste caso, 256 observações menos 2 parâmetros: o intercepto e a inclinação).

Um *F-value* alto e um *p-value*[13]. muito baixo indicam que o modelo de regressão linear é estatisticamente significativo e que a variável independente (*clock*) está fortemente associada à variável dependente (desempenho).

[13]Não confunda *p-value* com *p-value*...
O p-value do teste de correlação de Pearson refere-se à significância da relação linear entre duas variáveis contínuas, enquanto o p-valor em regressão linear refere-se à significância de um coeficiente específico no modelo de regressão.

Em resumo:

- O modelo de regressão linear sugere uma forte relação entre a frequência de *clock* e a performance do processador.
- O coeficiente do *clock* é altamente significativo, indicando que aumentos na frequência de clock estão fortemente associados a aumentos na performance.
- A mediana dos resíduos próxima de zero sugere um bom ajuste, embora a presença de outliers (valores mínimos e máximos distantes) possa indicar a necessidade de verificar a presença de pontos atípicos ou considerar modelos mais complexos.
- O R-quadrado de 65.05% indica que o modelo explica uma parte significativa da variação na performance, mas há espaço para melhorias.

2.5.2 Análise dos Resíduos

A função `summary()` fornece uma quantidade substancial de informações para nos ajudar a avaliar o ajuste de um modelo de regressão aos dados utilizados para desenvolvê-lo. Para aprofundar a análise da qualidade do modelo, precisamos examinar informações adicionais sobre os valores observados em comparação com os valores previstos pelo modelo. Em particular, a análise de resíduos examina esses valores residuais para entender melhor a qualidade do modelo.

Lembre-se de que o valor residual é a diferença entre o valor medido real e o valor que a linha de regressão ajustada prevê para aquele ponto de dados correspondente. Valores residuais maiores que zero significam que o modelo de regressão previu um valor muito pequeno em comparação com o valor medido real, e valores negativos indicam que o modelo previu um valor muito grande. Um modelo que se ajusta bem aos dados tenderia a superestimar e subestimar os valores com a mesma frequência. Assim, ao que parece a primeira vista, dado a conformação geométrica dos pontos em torno da reta ajustada, se plotarmos os valores residuais, esperaríamos vê-los distribuídos normalmente em torno de zero para um modelo bem ajustado.

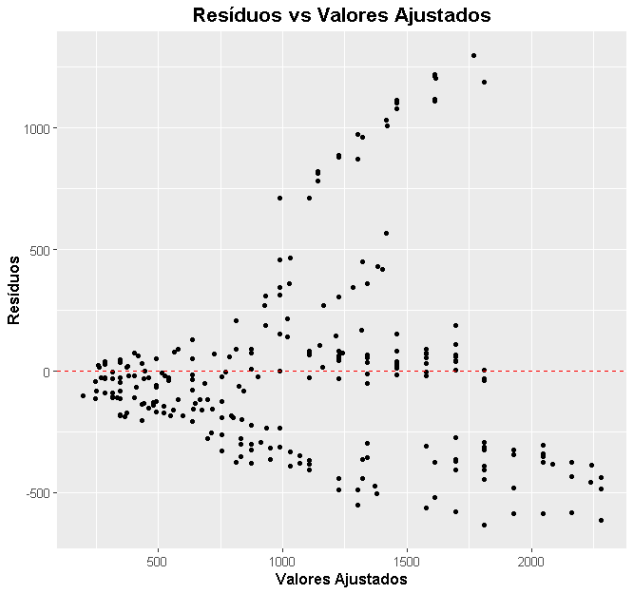


Figura 10: Os valores residuais versus os valores de saída do modelo SLR desenvolvido usando os dados do Int2000.

Neste gráfico da Figura 10, vemos que os resíduos tendem a aumentar conforme nos movemos para a direita. Além disso, os resíduos não estão uniformemente dispersos acima e abaixo de zero. No geral, esse gráfico nos diz que usar o *clock* como único preditor no modelo de regressão não explica suficientemente ou totalmente os dados. Em geral, se você observar qualquer tipo de tendência ou padrão claro nos resíduos, provavelmente precisará gerar um modelo melhor. Isso não significa que nosso modelo de regressão linear simples seja inútil. Significa apenas que podemos construir um modelo que produza valores residuais mais ajustados e melhores previsões.

Outro teste dos resíduos utiliza o gráfico quantil-quantil[14], ou Q-Q plot. Anteriormente, dissemos que, se o modelo se ajustasse bem aos dados, esperaríamos que os resíduos fossem distribuídos normalmente (Gaussianamente) em torno de uma média de zero. O Q-Q plot fornece uma boa indicação visual de se os resíduos do modelo são distribuídos normalmente. As chamadas de função a seguir geram o Q-Q plot mostrado na Figura

Se os resíduos fossem normalmente distribuídos, esperaríamos que os pontos plotados nesta figura seguissem uma linha reta. No entanto, com nosso modelo, vemos que as extremidades divergem consideravelmente dessa linha. Esse comportamento indica que os resíduos não são normalmente distribuídos. A forma como as caudas se desviam da linha de referência pode indicar como os resíduos observados se desviam do esperado caso fossem normalmente distribuídos.

Na verdade, este gráfico sugere que a cauda direita da distribuição é "mais pesada"do que o esperado de uma distribuição normal e que a cauda esquerda

[14] O gráfico Q-Q (Quantil-Quantil) é uma ferramenta diagnóstica usada para comparar a distribuição dos resíduos de um modelo de regressão com uma distribuição teórica, normalmente a distribuição normal

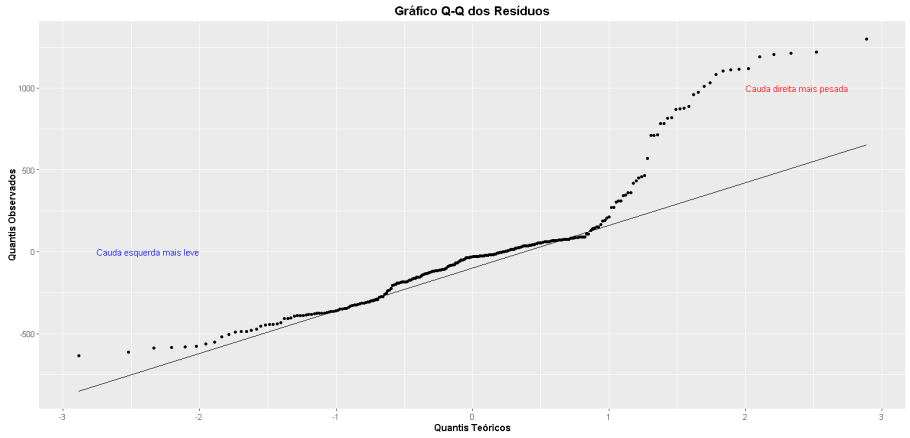


Figura 11: gráfico Q-Q para o modelo de regressão linear com os dados do Int2000.

é "mais leve"do que o esperado. Esse padrão é indicativo de uma distribuição enviesada à direita. Este teste confirma ainda mais que usar apenas o *clock* como preditor no modelo é insuficiente para explicar os dados.

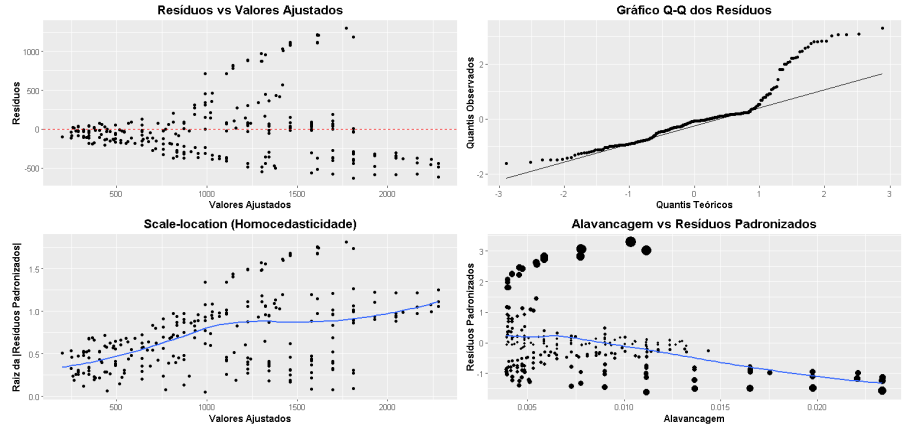


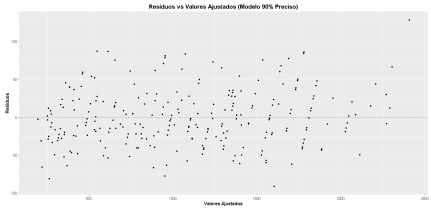
Figura 12: gráficos de análise mais profunda dos resíduos e sua possível normalidade.

Dos 4 gráficos apresentados na Figura 12 os dois primeiros já foram explicados anteriormente. Os outros serão explicados a seguir.

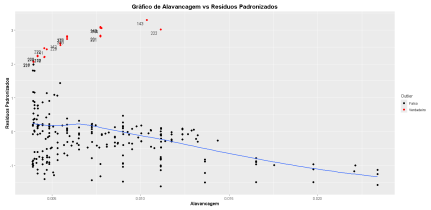
O gráfico "*Scale-location*"é uma maneira alternativa de visualizar os resíduos em relação aos valores ajustados do modelo de regressão linear. Nesse gráfico, os resíduos são padronizados e depois transformados pela raiz quadrada. Isso essencialmente dobra os resíduos e pode ajudar a encontrar padrões nos resíduos.

O gráfico de Resíduos vs Alavancagem pode ser usado para identificar possíveis *outliers*. Neste cenário, não há *outliers*.

Já o gráfico apresentado na Figura serve de comparação com a situação estudada até aqui, já que ele mostra um cenário onde o modelo se ajusta bem ao comportamento da variável dependente.



(a) Um modelo de regressão linear básico ajustado para uma precisão próxima de 90%, para comparação com a Figura 10.



(b) Um gráfico de alavancagem para evidenciar os *outliers* em um cenário em que realmente há *outliers*, para comparação Figura 12.

Figura 13: Dois gráficos que devem ser usados como comparação em cenários opostos as Figuras 10 e 12.

2.6 MULTIPLE LINEAR REGRESSION [11]

Chapter 2 • An Introduction to Data Analysis
Chapter 3 • Describing Data
Chapter 4 • Central Tendency and Dispersion
Chapter 5 • Univariate and Bivariate Descriptions of Data
Chapter 6 • Transforming Data
Chapter 7 • Some Principles of Displaying Data
Chapter 8 • The Essentials of Probability Theory
Chapter 9 • Confidence Intervals and Testing Hypotheses
Chapter 10 • Making Comparisons
Chapter 11 • Controlled Comparisons

Bibliografia

[1] Howard Wainer. Depicting error. *The American Statistician*, 50(2):101–111, 1996.

[2] Elena A Allen, Erik B Erhardt, and Vince D Calhoun. Data visualization in the neurosciences: overcoming the curse of dimensionality. *Neuron*, 74(4):603–608, 2012.

[3] S. Belia, F. Fidler, J. Williams, and G. Cumming. Researchers misunderstand confidence intervals and standard error bars. *Psychological Methods*, 10(4):389–396, 2005.

[4] R. Hoekstra, R. D. Morey, J. N. Rouder, and E. J. Wagenmakers. Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5):1157–1164, 2014.

[5] G. Cumming, F. Fidler, and D. L. Vaux. Error bars in experimental biology. *The Journal of Cell Biology*, 177(1):7–11, 2007.

[6] M. Krzywinski. Points of view: Elements of visual style. *Nature Methods*, 10(5):371–371, 2013.

[7] Claus O Wilke. *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. 2019.

[8] David S Brown. *Statistics and data visualization using R: the art and practice of data analysis*. SAGE Publications, 2021.

[9] Viktor Mayer-Schönberger and Kenneth Cukier. *Big data: A revolution that will transform how we live, work, and think*. Houghton Mifflin Harcourt, 2013.

[10] Number Around Us. Hypothesis testing in r: Elevating your data analysis skills, 2024. Acesso: 31-05-2024.

[11] David J. Lilja and Greta M. Linse. *Linear Regression Using R: An Introduction to Data Modeling, 2nd Edition*. 2022. Retrieved from the University Digital Conservancy.

[12] Jim Gray. Chapter 9: Performance measurement and benchmarking, 2000. Accessed: 2024-06-02.