UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
CURSO DE LICENCIATURA EM MATEMÁTICA

RODOLFO PINHEIRO CORREA

AJUSTE DE UMA CURVA LOGÍSTICA A PARTIR DE DADOS CENSITÁRIOS

TRABALHO DE CONCLUSÃO DE CURSO

CURITIBA

2018

RODOLFO PINHEIRO CORREA

AJUSTE DE UMA CURVA LOGÍSTICA A PARTIR DE DADOS CENSITÁRIOS

Trabalho de conclusão de curso apresentada ao Curso de Licenciatura em Matemática da Universidade Tecnológica Federal do Paraná como requisito parcial da disciplina Trabalho de Conclusão de Curso 1.

Orientador: Professor Dr. Mateus Bernardes

CURITIBA

Ministério da Educação
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
Câmpus Curitiba
Diretoria de Graduação e Educação Profissional
Departamento Acadêmico de Matemática
Coordenação do Curso de Licenciatura em Matemática

TERMO DE APROVAÇÃO

"AJUSTE DE UMA CURVA LOGÍSTICA A PARTIR DE DADOS CENSITÁRIOS"

por

"RODOLFO PINHEIRO CORREA"

Este Trabalho de Conclusão de Curso foi apresentado às 13 horas e 30 minutos do dia 28 de junho de 2018 na sala E201 como requisito parcial à obtenção do grau de Licenciado em Matemática na Universidade Tecnológica Federal do Paraná - Câmpus Curitiba. O aluno foi arguido pela Banca de Avaliação de Defesa abaixo assinados. Após deliberação, de acordo com o parágrafo 1º do art. 21 do Regulamento Específico do Trabalho de Conclusão de Curso para o Curso de Licenciatura em Matemática da UTFPR do Câmpus Curitiba, a Banca de Avaliação considerou o trabalho APROVADO.

_____ Prof. Dr. Mateus Bernardes **(Presidente - UTFPR/Curitiba)**	_____ Profa. Dra. Nara Bobko **(Avaliadora 1 – UTFPR/Curitiba)**
_____ Profa. Ms. Angélica Maria Tortola Ribeiro **(Avaliadora 2 – UTFPR/Curitiba)**	_____ Profa. Ms. Violeta Maria Estephan **(Professora Responsável pelo TCC – UTFPR/Curitiba)**
_____ Profa. Dra. Neusa Nogas Tocha (**Coordenadora do Curso de Licenciatura em Matemática – UTFPR/Curitiba**)	

"A Folha de Aprovação assinada encontra-se na Coordenação do Curso."

RESUMO

CORREA, Rodolfo Pinheiro. AJUSTE DE UMA CURVA LOGÍSTICA A PARTIR DE DADOS CENSITÁRIOS. 62 f. Trabalho de conclusão de curso – Departamento Acadêmico de Matemática – DAMAT, Universidade Tecnológica Federal do Paraná. UTFPR, Curitiba, 2018.

Neste trabalho realizamos um estudo das equações diferenciais autônomas e dos principais modelos de crescimento populacional. A partir de dados censitários referentes ao crescimento da população do município de Curitiba, procuramos modelar uma curva de crescimento populacional através do método de otimização numérica chamado de mínimos quadrados, e para tanto, implementamos no processo de modelagem a técnica das diferenças finitas. Nesta pesquisa ocorreu uma superestimação dos parâmetros nos modelos gerados, portanto realizamos um estudo estatístico, com o objetivo de compreender o comportamento destes modelos e também entender a natureza dos dados disponíveis.

Palavra-chaves: Crescimento populacional, equações diferencias autônomas, mínimos quadrados.

ABSTRACT

CORREA, Rodolfo Pinheiro. **ADJUSTMENT OF A LOGISTIC CURVE FROM CENSITIVE DATA**. 62 f. Completion of course work - Academic Department of Mathematics - DAMAT, Universidade Tecnológica Federal do Paraná. UTFPR, Curitiba, 2018.

In this work we perform a study of the autonomous differential equations and the main models of population growth. Based on census data regarding population growth in the city of Curitiba, we tried to model a population growth curve using the numerical optimization method called least squares, and for that, we implemented the finite difference technique in the modeling process. In this research, an overestimation of the parameters in the generated models was carried out, so we performed a statistical study, with the objective of understanding the behavior of these models and also understanding the nature of the available data.

Key words: Population growth, autonomous equations, least square equations.

SUMÁRIO

1 INTRODUÇÃO

Ao tanto faz observar as populações de modo em geral pode-se retirar varias informações e traduzi-las para modelos matemáticos. O processo de ajustar os parâmetros que melhor definem uma curva pode levar a muitos erros de cálculo. Em vista disto obter caminhos que tornem menor a chance de erro parece razoável para este projeto. A escolha do tema para este trabalho teve como origem compreender a natureza dos dados censitários e para tanto parece plausível buscar por análises estatísticas que ajudem a entender as discrepâncias entre os modelos propostos e os dados reais.

Iremos analisar o crescimento populacional de acordo com o tempo, mas para tornar o modelo mais próximo do caso real é possível incorporar outras variáveis explicativas como por exemplo renda, acesso dos indivíduos a saúde e acesso a educação no processo de modelagem, em contrapartida o processo de modelagem pode se tornar mais pesaroso e suscetível ao erro, já que existirão mais elementos a serem correlacionados. Sendo assim não temos como objetivo implementar mais de uma variável independente nesta pesquisa. Ao tratar de taxa de crescimento se supõe que esta taxa está relacionada com o tamanho da população. Se tomarmos P como o tamanho da população dependente da variável t, que representa o tempo, chamamos o modelo a seguir como *densidade-dependente*

$$\frac{dP}{dt} = f(P) \tag{3}$$

Abordaremos algumas formas funcionais para f e suas principais características, realizaremos testes estatísticos sobre os possíveis modelos e em seguida analisaremos os resultados obtidos.

Temos como objetivo modelar uma curva logística de acordo com dados populacionais e compreender que o processo de modelagem matemática impõe limites na descrição de uma situação real, mas que dentro destes limites é possível obter descrições qualitativas suficientes para traçar projeções e conjecturas para futuras situações reais, uma vez que a curva modelada não descreve exatamente a situação real, porem através de técnicas numéricas pode-se chegar

numa representação matemática mais adequada.

Veremos ao longo deste trabalho que, no intuito de alcançar este objetivo, necessitamos discutir o conceito de equações diferencias de primeira ordem autônomas, identificando os pontos críticos de cada modelo, bem como o seu comportamento assintótico, assinalando as diferenças entre os casos assintoticamente estáveis e os instáveis e relacionando esses objetos matemáticos com a situação real. Também se faz necessário conhecer o método de ajuste de curvas dos mínimos quadrados o suficientemente para estimar os parâmetros da taxa de variação populacional de cada modelo e se necessário implementar modificações no método, como por exemplo, escolher um peso para cada elemento da amostra. A realização de testes que verifiquem a qualidade dos dados disponíveis e a confiabilidade dos modelos ajustados, estas análises foram feita através do *software R*.

Este trabalho de conclusão de curso este divido em 7 capítulos, onde no Capítulo 2, apresentamos sobre dois dos modelos clássicos em dinâmica populacional e discutimos alguns aspectos importantes a respeito de estabilidade de soluções de equilíbrio

No Capítulo 3 discorremos sobre o método dos mínimos quadrados, onde primeiro falamos sobre o método ordinário e sua respectiva forma matricial e, em seguida, como contraponto, quando os dados disponíveis são altamente dispersos apresentamos o método ponderado como alternativa para modelagens.

Já no Capítulo 4 falamos sobre a análise de regressão, onde verificamos certas premissas a respeito do método dos mínimos quadrados nas quais quando satisfeitas garantem um modelo consistente. Neste mesmo capítulo apresentamos alguns testes que são capazes de verificar estas premissas.

No Capítulo 5 comentamos dois trabalhos cuja metodologia e resultado serviu de inspiração para este, e utilizamos a estratégia usada pelos respectivos autores no processo de modelagem. Ainda neste capitulo apresentamos os modelos gerados tanto pelo método ordinário quanto pelo ponderado.

Já no Capítulo 6 expomos os resultados dos testes descritos no quarto capítulo, apresentamos alguns modelos gerados no *Software R* e em seguida discutimos sobre as respectivas respostas dos testes. Por fim no Capítulo 7 contem as considerações finais sobre os estudos descritos neste trabalho.

2 MODELOS CLÁSSICOS EM DINÂMICA POPULACIONAL

2.1 EQUAÇÕES AUTÔNOMAS

Segundo (ZILL, 2003) uma Equação Diferencial Ordinária (EDO) de primeira ordem onde não aparece explicitamente a variável independente chama-se de autônoma. Se x representa uma variável independente de uma equação diferencial de primeira ordem *autônoma*, a EDO pode ser escrita como $f(y,y') = 0$ ou na forma

$$\frac{dy}{dx} = f(y). \tag{5}$$

2.1.1 SOLUÇÃO ESTACIONÁRIA OU DE EQUILÍBRIO

Um número real c é um *ponto crítico* (chamado também de *ponto de equilíbrio ou estacionário*) de uma EDO autônoma (5) se for um zero de f, ou seja $f(c) = 0$.

Dada um problema do valor inicial:

$$\begin{cases} \dfrac{dy}{dx} = f(y) \\ y(0) = y_0. \end{cases} \tag{6}$$

Segundo (FIGUEIREDO; NEVES, 2010), um ponto crítico pode ser classificado como estável se:

- Um ponto de equilíbrio c é estável se dado $\varepsilon > 0$ existe $\delta > 0$, tal que para $|y_0 - c| < \delta \Rightarrow |y(x) - c| < \varepsilon$, para todo $x \geq 0$.
- Um ponto de equilíbrio c é *assintoticamente estável* se for estável e se existir $\eta > 0$ tal que $\lim_{x\to\infty} y(x) = c$ quando $|y_0 - c| < \eta$.
- Um ponto de equilíbrio que não cumpre estas condições que foram descritas acima é dito instável.

O teorema 1 (FIGUEIREDO; NEVES, 2010) apresenta um conceito mais simples do que este que foi descrito acima e que na prática é o que se aplica na análise de estabilidade, em especifico analisamos a derivada de f.

Teorema 1: Seja c um ponto de equilíbrio do PVI com f de classe C^1. Então $f'(c) < 0$ implica que c é assintoticamente estável, e $f'(c) > 0$ implica que c é assintoticamente instável.

2.1.2 CURVAS INTEGRAIS

Sabe-se que a solução de (6) representa uma curva. Segundo (ZILL, 2003) a curva integral pode ser classificada de acordo com o seu comportamento.

Teorema 2: Seja R uma região retangular do plano xy definida por $a \leq x \leq b, c \leq y \leq d$ que contém o ponto (x_0, y_0). Se $f(y)$ e $\dfrac{\partial f}{\partial y}$ da Equação 6 são continuas em R, então existe algum intervalo $I_0 : x_0 - h < x < x_0 + h, h > 0$, contido em $a \leq x \leq b$, e uma única solução $y(x)$ definida em I_0, que é a solução do problema do valor inicial.

Este teorema estabelece em que condições existe a (única) solução de uma EDO deste tipo, alem do mais temos que este teorema foi definido para equações autônomas, porém também é valido para uma equação diferencial não autônoma dentro das mesmas condições. Este tipo de teorema é chamado de Teorema da existência e unicidade.

Pode-se tomar (6) para todo intervalo $-\infty < x < \infty$, ou para um intervalo em específico $0 \leq x < \infty$. Como f e sua derivada f' são funções continuas de x no eixo y, pelo Teorema 1, dado um ponto (x_0, y_0), então existe uma única curva integral em xy que passa por este ponto, ou seja, para (6) existe solução e é única.

Dados dois pontos críticos de uma EDO autônoma c_1, c_2, com $c_1 < c_2$, os gráficos $r(x) = c_1, r(x) = c_2$ são retas horizontais que dividem a região R em três sub-regiões. Dada um solução $y(x)$ qualquer, podemos classificar esta solução da seguinte maneira:

1. $y(x)$ é dita *limitada acima* por um ponto crítico c_1 quando $\lim\limits_{x \to -\infty} y(x) < c_1$ e $\lim\limits_{x \to \infty} y(x) < c_1$.
2. $y(x)$ é dita *limitada* pelos pontos críticos c_1 e c_2 quando $c_1 < \lim\limits_{x \to -\infty} y(x) < c_2$ e $c_1 < \lim\limits_{x \to \infty} y(x) < c_2$.
3. $y(x)$ é dita *limitada abaixo* por um ponto crítico c_2 quando $c_2 < \lim\limits_{x \to -\infty} y(x)$ e $c_2 < \lim\limits_{x \to \infty} y(x)$.

2.2 DINÂMICA DE UMA POPULAÇÃO

2.2.1 MODELO MALTHUSIANO

Thomas Malthus (1766-1834) graduou-se em matemática pela universidade de Cambridge, foi professor de historia e economia política. A partir de suas pesquisas desenvolveu o trabalho seminal *An Essay on the Principle of Population* onde descreveu, utilizando dados da população inglesa, um modelo de crescimento populacional. Malthus defendeu a ideia de que taxa para este modelo de crescimento populacional é constante e escrevemos da seguinte maneira a equação diferencial com condição inicial $P(t_0) = P_0$:

$$\begin{cases} \frac{dP}{dt} = \lambda P \\ P(t_0) = P_0 \end{cases} \tag{7}$$

onde λ é uma constante positiva. Sendo assim temos a solução única garantida pelo teorema de existência e unicidade:

$$P(t) = P_0 e^{\lambda t} \tag{8}$$

Este modelo tem como característica o tempo de duplicação populacional igual a:

$$\frac{\ln(2)}{\lambda} \tag{9}$$

a duplicação da quantidade da população independentemente do tamanho de P_0.

De acordo com (ZILL; CULLEN, 2001) Malthus declarou em suas pesquisas que as populações crescem numa progressão geométrica. Esta afirmação se verifica sem dificuldade tomando $a = P_0$, $c = e^{\lambda}$ e substituindo em (8) nos instantes $t = 1,\ t = 2, \ldots$ podemos observar a seguinte sequência

$$a, ac, ac^2, ac^3 \ldots$$

O pesquisador também examinou com cuidado dados dos últimos censos dos Estados Unidos daquela época e comparou com a de outros países onde concluiu que o crescimento natural de populações era de natureza exponencial e com tempo de duplicação de vinte e cinco anos para humanos.

O modelo malthusiano é um dos modelos de crescimento populacional mais simples,

onde supomos que a taxa de crescimento constante. Geralmente podemos obter essa taxa a partir da diferençam ou seja $\lambda = \lambda_n - \lambda_m$, onde λ_n é a taxa de natalidade e λ_m é a taxa de mortalidade.

Este modelo parece suficiente para descrever pequenas populações em intervalos curtos de tempo, como no caso do estudo de populações de microrganismo, que se dispõem de mecanismos simples de sobrevivência de ele se adequar. Uma crítica que pode ser feita a este modelo é o fato de que, apesar dele se adequou aos dados disponíveis na época, o mesmo não era realístico no que diz respeito a população máxima, já que o crescimento exponencial cresce indefinidamente quando a razão positiva e maior que um, portanto um modelo inviável para populações com estruturas mais complexas.

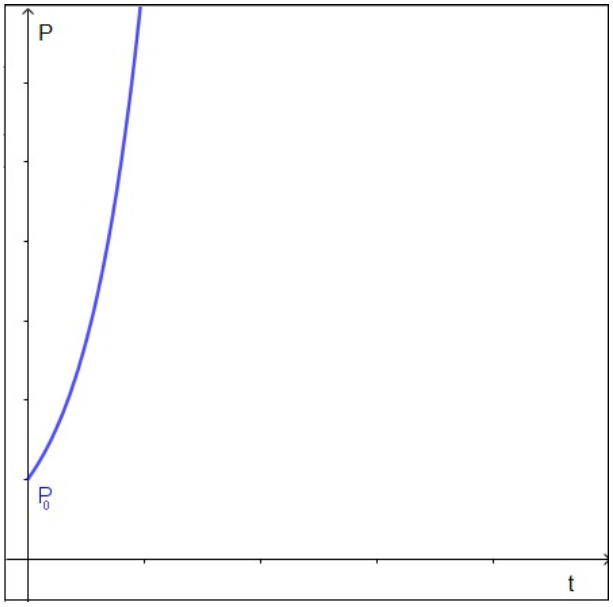

Figura 4: A curva vermelha representa o modelo malthusiano

O gráfico na Figura 4 exemplifica a dinâmica populacional descrita pelo modelo de Malthus. Repare que a curva inicia na população inicial P_0 e cresce exponencialmente e este tipo de desenvolvimento não se encaixa com casos reais em longos intervalos de tempo.

2.2.2 MODELO DE VERHULST OU MODELO LOGÍSTICO

O modelo logístico foi desenvolvido pelo matemático belga Pierre-François Verhulst (1804-1849). Apesar de ter tentado testar o seu modelo, Verhulst se frustou com os dados populacionais da época que não eram suficientes para um teste efetivo do modelo logístico. O modelo ficou esquecido até ser descoberto por dois cientistas americanos que trabalhavam na Universidade Johns Hopkins, sendo eles Raymond Pearl e Lowell J. Reed (por este fato o

modelo logístico é conhecido também como modelo de Verhulst-Pearl (ZILL; CULLEN, 2001). Em 1920 Pearl e Reed analisaram a proximidade da curva de crescimento populacional dos Estados Unidos com a curva logística e se surpreenderam com a precisão do modelo descoberto pelo matemático belga.

Sabemos que o modelo malthusiano considera a taxa de crescimento como constante. Segundo (ZILL; CULLEN, 2001) este tipo de hipótese parece razoável quando o estudo de população se dá em intervalos curtos de tempo e para pequenas populações que sofrem poucas interferências no seu crescimento. Porém esta hipótese não parece ser logicamente plausível para outros tipo de populações, pois quando uma população cresce, em certo ponto começam a surgir mecanismos que diminuem a taxa de crescimento, como a superpopulação, nesta condição os recursos para manter a sobrevivência dos indivíduos começam a se tornar escassos. Além dos mais, dependendo da espécie estudada, a superpopulação pode alterar o comportamento fisiológico, tal como seus hábitos reprodutivos.

Apesar do modelo logístico não ser ideal por não considerar algumas variáveis como, por exemplo, a distribuição espacial da população e a maturação dos sujeitos, dependendo da espécie, certos indivíduos não produzem novos membros antes de certo período, ainda sim podemos tirar informações, como por exemplo, o modelo nos fornece a população máxima de uma população.

No modelo de Verhulst a taxa de crescimento decresce linearmente com o aumento do tamanho da população $\lambda(P) = a - bP$, onde a e b são constantes positivas. A partir desta hipótese pode-se escrever a equação diferencial com condição inicial $P(t_0) = P_0$:

$$\begin{cases} \dfrac{dP}{dt} = (a - bP)P, \\ \quad P(t_0) = P_0 \end{cases} \tag{10}$$

Para $P_0 \neq 0$ e $P_{max} \neq \dfrac{a}{b}$

A equação (10) é uma equação separável

$$\frac{dP}{(a - bP)P} = dt. \tag{11}$$

O lado esquerdo da equação 11 pode ser decomposto em frações parciais, como descrito a seguir

$$\frac{dP}{(a - bP)P} = \left[\frac{1}{aP} + \frac{b}{a(a - bP)}\right] dP \tag{12}$$

substituindo (12) na equação (10) e integrando, tem-se

$$\frac{1}{a}\ln|P| - \frac{1}{a}\ln|a-bP| = t + c, \text{ onde } c \text{ constante,} \tag{13}$$

usando as propriedades de logaritmo temos:

$$\ln\left|\frac{P}{a-bP}\right| = a(t+c). \tag{14}$$

utilizando as propriedades de logaritmos novamente

$$\frac{P}{a-bP} = e^{a(t+c)} \tag{15}$$

e realizando operações algébricas

$$P = \frac{ae^{at}c}{1+be^{at}c} \tag{16}$$

Aplicando a condição inicial $P(t_0) = P_0$ a constante c assume a forma

$$c = \frac{P_0}{ae^{at} - bP_0e^{at}} \tag{17}$$

Substituindo a equação 17 em 16

$$P = \frac{ae^{at}P_0}{ae^{at_0} - bP_0e^{at_0} + bP_0e^{at}} \tag{18}$$

Dividindo a Equação 18 no numerador e no denominador por e^{at} a Equação10 tem como solução

$$P(t) = \frac{aP_0}{bP_0 + (a-bP_0)e^{-a(t-t_0)}} \tag{19}$$

Para facilitar a formulação dos modelos reescrevemos a solução dividindo o numerador e o denominador por $bP0$, da onde temos a seguinte função

$$P(t) = \frac{\dfrac{a}{b}}{1 + \left(\dfrac{a}{bP_0} - 1\right)e^{-a(t-t_0)}}. \tag{20}$$

Substituindo $P_{max} = \dfrac{a}{b}$ a função adquire a forma:

$$P(t) = \frac{P_{max}}{1 + \left(\dfrac{P_{max} - P_0}{P_0}\right) e^{-a(t-t_0)}} \tag{21}$$

Fazendo um estudo dos pontos críticos desta equação diferencial de primeira ordem autônoma tem-se que $P(t) = 0$ e $P(t) = \frac{a}{b} = P_{max}$ são soluções. Alem do mais, estas soluções demonstram comportamento assintótico, onde $P(t) = 0$ é a solução de equilíbrio instável e P_{max} é o assintoticamente estável.

Chama-se de *capacidade suporte* de uma população solução estável da curva integral, desta maneira $\frac{a}{b}$ é *capacidade suporte* da população em questão estudada. Sendo assim independente da população inicial P_0 tem-se $\lim\limits_{t\to\infty} P(t) = \dfrac{a}{b}$

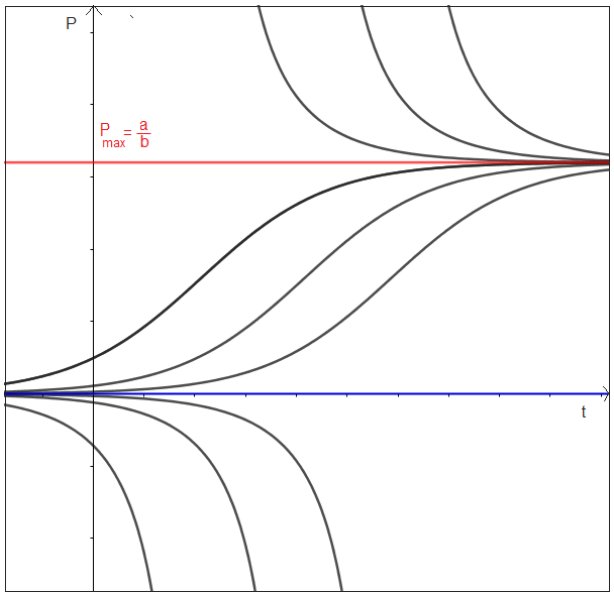

Figura 5: A reta vermelha corresponde a capacidade suporte (solução estável), já a reta azul refere-se a solução instável dos modelos ilustrados

Por exemplo a partir da figura 5 podemos perceber o comportamento de tais soluções, note que a reta vermelha correspondente a solução de equilíbrio estável atrai as curvas soluções, já para a solução instável (reta azul) as curvas soluções se afastam da reta.

Ainda mais, podemos interpretar as soluções estacionárias do ponto de vista populacional, a reta vermelha corresponde a capacidade suporte. Note que a quantidade de indivíduos da população decai naturalmente se está acima da reta mencionada e se está a baixo cresce até atingir o valor da sua população máxima. Sendo assim esta solução estacionária se comporta como um limitante dentro de uma dinâmica populacional. A reta azul representa a solução

$P = 0$, porém as curvas abaixo são desconsideradas, ja que não existe quantidade de indivíduos negativa. Por fim esta solução instável pode representar o *limiar*, onde a quantidade populacional decai se não atinge o valor mínimo para desenvolvimento ou cresce até se aproximar da capacidade máxima de indivíduos. Não lidaremos com modelos que consideram o limiar de uma população, já que se o limiar é igual a zero o modelo não fará sentido biológico, já que não existe quantidade populacional negativa.

2.2.3 MODELO LOGÍSTICO GENERALIZADO

No trabalho de (VLADAR, 2005) nos foi apresentado o *Modelo logístico generalizado*, levemente diferente do proposto por Verhulst e tem a seguinte forma:

$$\frac{dP}{dt} = rP\left[1 - \left(\frac{P}{K}\right)^{\theta}\right] \tag{22}$$

Este modelo tem como parâmetros a taxa de crescimento intrínseca, r, capacidade suporte da população K e o expoente que não depende nem do tamanho da população P e nem de K. Para este modelo P_0 e P_{max} são os mesmos do logístico, porém o expoente θ dá uma nova interpretação para a curva de crescimento, se $\theta > 1$ a competição intraespecifica[1] é alta o crescimento desta população demora mais para se aproximar da capacidade suporte, caso $0 < \theta < 1$, assim a competição entre os indivíduos é baixa e o crescimento é mais acelerado até alcançar a quantidade máxima. O comportamento das curvas do modelo logístico generalizado quando θ varia pode ser observado na figura 6.

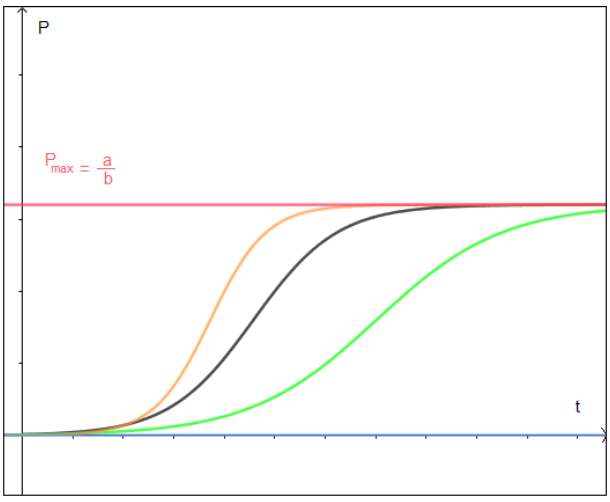

Figura 6: A curva em laranja representa o crescimento logístico quando $0 < \theta < 1$, a curva em preto quando $\theta = 1$ e a curva em verde quando $\theta > 1$

[1] Quando membros da mesma espécie que competem por recursos limitados

3 MÉTODO DE AJUSTE DE CURVAS: MÍNIMOS QUADRADOS ORDINÁRIOS

O problema de ajustar uma curva a um conjunto de dados tabelados expressos por $(x_i, f(x_i))$ com $i = 1, ..., n$, no sentido de quadrados mínimos, consiste em aproximar este conjunto por uma função g específica, cuja forma funcional depende de um conjunto de parâmetros fixados (CUNHA, 2013). Ou seja, estamos supondo que os dados serão aproximados por uma função do tipo:

$$f(x_i) \cong g(x_i) = c_1\alpha_1(x_i) + c_2\alpha_2(x_i) + ... + c_n\alpha_n(x_i), \tag{24}$$

com $c_1, c_2, ..., c_n \in \mathbb{R}$ e $\alpha_1, \alpha_2, ..., \alpha_n$ são funções reais preestabelecidas.

Para uma melhor aproximação da curva deve-se minimizar as diferenças dos resíduos entre f e g. A diferença dos valores numéricos destas funções podem assumir tanto valores positivos como negativos e a fim de evitar uma soma nula das diferenças toma-se o quadrado dos resíduos neste método, onde o resíduo é dado pela expressão $r(x_i) = f(x_i) - g(x_i)$.

Definição: O produto interno entre duas funções f e g é definida por:

$$< f, g >= \sum_{i=1}^{n} f(x_i)g(x_i) \tag{25}$$

A soma do quadrado dos resíduos pode ser escrita na forma:

$$< r, r >= \sum_{i=1}^{n} (f(x_i) - g(x_i))^2 \tag{26}$$

No método dos mínimos quadrados o critério é a minimização da soma dos quadrados dos resíduos, ou seja, deve-se derivar $< r, r >$ relação a cada um dos parâmetros c_i impondo a condição do ponto critico sobre cada derivada parcial na variável c_i.

Para determinar as constantes $c_1, c_2, ..., c_n$ da aproximação (24) deve-se minimizar a função:

$$< r,r >=< f(x) - c_1\alpha_1(x) - ... - c_n\alpha_n(x)), f(x) - c_1\alpha_1(x) - ... - c_n\alpha_n(x) > \quad (27)$$

Usando a linearidade do produto escalar, deriva-se esta equação em relação a cada um dos parâmetros c_i e iguala-se a zero:

$$\frac{\partial < r,r >}{\partial c_i} = -2 < f(x) - c_1\alpha_1(x) - c_2\alpha_2(x) - ... - c_n\alpha_n(x), \alpha_i(x) >= 0,\ i = 1,...,n. \quad (28)$$

Distribuindo o produto escalar obtemos a equação:

$$c_1 < \alpha_1, \alpha_i > + c_2 < \alpha_2, \alpha_i > + ... + c_n < \alpha_n, \alpha_i >=< f, \alpha_i >\ \ i = 1,...,n \quad (29)$$

Tomando $i = 1,\ i = 2,...,i = n$ chegamos no sistema:

$$\begin{cases} < \alpha_1, \alpha_1 > c_1 + < \alpha_2, \alpha_1 > c_2 + ... + < \alpha_n, \alpha_1 > c_n =< f, \alpha_1 > \\ < \alpha_1, \alpha_2 > c_1 + < \alpha_2, \alpha_2 > c_2 + ... + < \alpha_n, \alpha_2 > c_n =< f, \alpha_2 > \\ \qquad\qquad\qquad\qquad ... \\ < \alpha_1, \alpha_n > c_1 + < \alpha_2, \alpha_n > c_2 + ... + < \alpha_n, \alpha_n > c_n =< f, \alpha_n > \end{cases} \quad (30)$$

Se o determinante desta matriz for diferente de zero podemos obter os coeficientes de (24) que definem a melhor aproximação para f.

3.1 FORMA MATRICIAL PARA O MÉTODO DOS MÍNIMOS QUADRADOS

Segundo (CUNHA, 2013) o processo de calcular os parâmetros que melhor se ajustam por uma reta de mínimos quadrados leva a um sistema $m \times n$, o que pode ser reescrito como uma matriz. Dada as medidas $y_1, y_2, ..., y_m$ nos pontos distintos $x_1, x_2, ..., x_m$, tem-se o sistema linear:

$$\begin{cases} x_1 a + b = y_1 \\ x_2 a + b = y_2 \\ \qquad ... \\ x_m a + b = y_m. \end{cases} \quad (31)$$

pode-se reescrever este sistema na equação matricial

$$\begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \dots & \\ x_m & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \dots \\ y_m \end{bmatrix} \tag{32}$$

Seja A uma matriz $m \times n$ com $m > n$ e b um vetor que projetamos no espaço-coluna da matriz A, tal que o sistema $Ax = b$ seja inconsistente. Deve-se escolher x (matriz dos parâmetros) para minimizar a matriz dos resíduos $E = b - Ax$, sendo $||E|$ exatamente a distância de b até o ponto Ax no espaço-coluna da Matriz A. O que desejamos é minimizar E e isto equivale a encontra $\bar{x}$ tal que $\overline{E} = b - A\bar{x}$ seja perpendicular ao espaço de Ax.

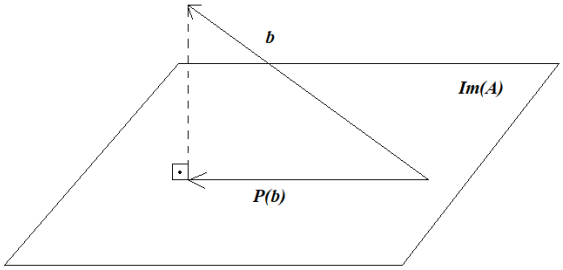

Figura 8: Projeção $P(b)$ do vetor b sobre o espaço de Ax, denotado por $Im(A)$

Devemos verificar que todos os vetores perpendiculares ao espaço-coluna estão no espaço-nulo a esquerda, deste modo o vetor $E = b - Ax$ deve estar no espaço-nulo de A^T. Em termos algébricos tem-se o sistema:

$$\begin{cases} a_1^T e = 0 \\ a_2^T e = 0 \\ \quad \dots \\ a_n^T e = 0 \end{cases} \tag{33}$$

Cuja combinação resulta em $A^T E = 0$

Deve-se multiplicar a equação inconsistente $Ax = b$ por A^T, deste modo obtêm-se a equação $A^T Ax = A^T b$ conhecida como equação normal, onde a $A^T A$ será uma matriz quadrada e simétrica. Como a matriz $A^T A$ é uma matriz positiva definida, logo o $det(A^T A) \neq 0$ e admite uma matriz inversa, assim a solução da equação normal será $x = (A^T A)^{-1} A^T b$. Portanto a projeção de b no espaço coluna é o ponto Ax e tem-se $Ax = A(A^T A)^{-1} A^T b$.

3.1.1 MATRIZ DE PROJEÇÃO

Vamos chamar $P = A(A^TA)^{-1}A^T$ de matriz de projeção, onde esta matriz projeta qualquer vetor b no espaço coluna de A. Em outras palavras Pb é o componente de b no espaço-coluna e a matriz dos resíduos, como $Ax = Pb$, temos que $E = b - Pb = (I - P)b$ é o componente do complemento onde $(I - P)$ é a matriz que projeta qualquer b no espaço nulo a esquerda de A^T (ortogonal ao espaço-coluna de A).

A matriz de projeção possui duas propriedades interessantes:

1. $P^2 = P$
2. $P^T = P$

A propriedade 1 é verificada na linha abaixo:

$$\begin{aligned} P^2 = PP &= (A(A^TA)^{-1}A^T)(A(A^TA)^{-1}A^T) \\ &= (A(A^TA)^{-1}A^TA(A^TA)^{-1}A^T) \\ &= (A(A^TA)^{-1}IA^T) = (A(A^TA)^{-1}A^T) = P \end{aligned} \tag{34}$$

A propriedade 2 é verificada na linha abaixo:

$$\begin{aligned} P^T &= (A(A^TA)^{-1}A^T)^T = ((A^T)^T((A^TA)^{-1})^TA^T) \\ &= (A((A^T(A^T)^T)^{-1})A^T) = (A(A^TA)^{-1}A^T) = P \end{aligned} \tag{35}$$

Como Pb é a projeção no espaço coluna de P a matriz dos resíduos $b - Pb$ é ortogonal para qualquer vetor a_i^T no espaço-coluna, logo o produto escalar $< a_i^T, b - Pb >= 0$.

3.2 MÍNIMOS QUADRADOS PONDERADOS

O método dos mínimos quadrados ordinários estima uma reta no qual os resíduos são identicamente distribuídos com a *distribuição normal*, ou seja, a variância é constante para todo erro gerado. Existem situações onde a variância não é constante para todas as amostras, e este método pode gerar um modelo que não explique o caso real com a precisão esperada, já que durante o procedimento de regressão todos os pontos são considerados com a mesma influência, mesmo aqueles mais distantes da média das amostras e que podem aumentar a imprecisão da regressora.

O método dos mínimos quadrados ponderados implementa um peso para cada ponto no

processo de regressão. Esta modificação tem como objetivo corrigir a imprecisão causada pelas amostras mais distantes acrescentando um peso maior para os pontos que consideramos mais significativos e um peso menor para os pontos mais distantes. Por outro lado, este método não tem como finalidade diminuir os resíduos e sim adequar a reta regressora próxima aos pontos mais relevantes da amostra. Este método pode ser exemplificado, comparando os valores dos resíduos do modelo gerado pelo mínimos quadrados ordinários (MQO) com os resíduos do modelo gerado pelo método dos mínimos quadrados ponderado (MQP) de uma mesma amostra aleatória gerada pelo *Software R*, através da tabela 5. Note que não houve resíduos menores, na coluna do método ordinário em comparação com o método ponderado. Graficamente fica visível na imagem 9 o deslocamento da reta regressora em direção ao pontos mais consideráveis.

Tabela 5: Resíduos gerados pelo mínimo quadrado ordinário e pelo mínimo quadrado ponderado

Resíduos MQO (Em módulo)	Resíduos MQP (Em módulo)
0,029495381	0,063618137
0,030066353	0,059959590
0,062499118	0,044140933
0,103493409	0,092825257
0,092290867	0,085467732
0,070456754	0,067094135
0,053101970	0,053584369
0,004779831	0,003528429
0,022655991	0,017559573
0,047409812	0,038852878
0,114880431	0,105938995
0,087337653	0,070706183
0,074905774	0,051353273
0,142693727	0,118372223
0,172382834	0,227079973

Fonte: Dados aleatórios gerados pelo *Software R*

Portanto o uso do método ponderado tem como intuito dar ênfase a certos pontos da amostra disponível e este foco depende do peso escolhido. Na pesquisa realizada para este trabalho optou-se por dar maior significância para os pontos mais próximos da média amostral.

No caso do método dos mínimos quadrados ordinário consideramos os valores de y_i igualmente confiáveis, já no caso do método dos mínimos quadrados ponderados isto não ocorre para todos os y_i da amostra, pois são colocados pesos w_i^2 para cada valor. Deste modo tem-se:

$$E^2 = w_1^2(y_1 - ax_1 - b)^2 + w_2^2(y_2 - ax_2 - b)^2 + ... + w_m^2(y_m - ax_m - b)^2 \tag{36}$$

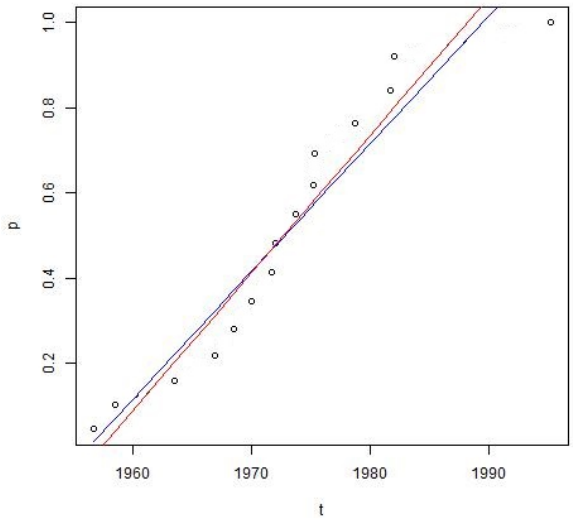

Figura 9: Gráfico tempo (t) $\times$ população (P): reta gerada através dos mínimos quadrados ordinário em azul e reta gerada através dos mínimos quadrados ponderado em vermelho

Onde se $w_j > w_i$ então será atribuído maior importância a b_j do que a b_i.

Deve-se então fazer uma mudança no sistema $Ax = b$ para $WAx = Wb$ e isto muda a solução para x_w e a matriz W^TW aparece em ambos os lados da equação

$$(WA)^TWA = (WA)^TWb = A^T(W^TW)A = A^T(W^TW)b \tag{37}$$

Vamos chamar a combinação simétrica de W como $C = W^TW$ o produto escalar de x e y por y^TCx denotado por $< x, y >_w$, onde

$$< x, y >_w = (Wy)^T(Wx) = y^TW^TWx = y^TCx \tag{38}$$

Existe várias meios de escolher C, um destes meios é escolher a matriz que tem na diagonal principal a média ao quadrado entre o erro e a media dos valores do experimento, sendo esta medida a *variância* da amostra.

4 ESTUDO ESTATÍSTICO

O processo de modelagem ocorre quando temos uma situação real e a partir disto geramos um modelo matemático capaz de explicar em termos numéricos a problemática, porém existem situações onde o modelo criado não é suficientemente capaz de produzir respostas significativas. Realizamos os estudos estatísticos através do *Software R* com a finalidade de encontrar justificativas para as dificuldades encontradas nas pesquisas desempenhadas neste trabalho.

4.1 ANALISE DE REGRESSÃO

Dado um modelo de regressão linear da forma

$$y = ax + b + e \tag{40}$$

com erro e e parâmetros a e b, a análise de regressão tem como finalidade tecer inferências a respeito de a e b, e saber o quão distante estes parâmetros estão de suas contrapartes na população. Para isto precisamos especificar a forma funcional do modelo e também verificar certas premissas, segundo (GUJARATI, 2006) tais como a homoscedasticidade, a normalidade e independência dos resíduos.

4.2 PREMISSAS PARA A UTILIZAÇÃO DOS MÍNIMOS QUADROS

4.2.1 OS RESÍDUOS SEGUEM DISTRIBUIÇÃO NORMAL

O modelo de regressão linear clássico pressupõe que os resíduos seguem uma distribuição normal $e_i \sim N(0, \sigma^2)$, independentes entre si. Neste caso os estimadores de mínimos quadrados ordinários apresentam algumas propriedades como: não são tendenciosos, tem variância mínima e são consistentes (isto é a medida que o tamanho da amostra aumenta os estimadores convergem para os verdadeiros valores da população).

4.2.2 AUSÊNCIA DE AUTOCORRELAÇÃO

Dados dois resíduos quaisquer $e_i\, e_j$ $(i \neq j)$, desde que sejam ambas variáveis aleatórias independentes, isto é a $cov(e_i, e_j) = 0$. Se os termos do erro seguem assim y_i, depende não somente de x_i, mas também de e_{i-1}, pois e_{i-1} determina e_i.

4.2.3 HOMOSCEDASTICIDADE

Dado o valor de x_i, a variância de e_i é a mesma para todas as observações. Isto é as variâncias condicionais de ε_i são idênticas, ou seja, $var(e_i|x_i) = \sigma^2$ para todas observações. Tecnicamente estamos representando as premissas da *homoscedasticidade*, basicamente a presença da homoscedasticidade nos revela que a variação em torno da linha de regressão é a mesma para todos os x_i. A ausência de homoscedasticidade chamamos heteroscedasticidade e isto pode levar a uma estimação maior do erro-padrão e consequentemente a subestimação dos parâmetros.

4.2.4 MÍNIMOS QUADRADOS PONDERADOS E A HETEROCEDASTICIDADE

Segundo (SHALIZI, 2009) o uso do método dos mínimos quadrados ponderados se faz por dois motivos, o primeiro é para uma melhorar a precisão nos pontos nos quais procuramos priorizar, consequentemente a curva gerada pela regressão será puxada para os respectivos pontos, sendo assim pesos maiores para uma região e peso menores para outras regiões geraram um modelo com ajuste distinto do gerado pelo método ordinário. O segundo motivo seria para descontar a imprecisão no processo de modelagem, já que dados com altas variações tendem, por exemplo, causar resíduos maiores em partes da regressora,, isto significa que a variância dos resíduos deve ser constante, deste modo a regressão é mensurada com a mesma precisão em todos os pontos, conhecemos esta situação como *Homoscedasticidade*. Se os resíduos não forem constantes, ou seja dispersos, o método dos mínimos quadrados ponderados no caso de os pesos serem escolhidos de maneira coerente podem corrigir a heterocedasticidade.

Nesta ultima condição mencionada o método dos mínimos ordinários pode perde a veracidade com o caso real, como exemplo pode-se gerar uma capacidade suporte superestimada no modelo logístico, sendo assim a regressão não se torna eficiente o suficiente para ser considerada. Em outros termos, não conseguimos estimar uma curva com precisão se o erros são grandes. Outra observação que deve ser levantada é que os resíduos e a variância também dependem do número de entradas que são inseridas no processo de modelagem, portanto, quanto mais dados disponíveis menor será a variância. Como em muitas ocasiões a disponibilidade de dados pode ser restrita, busca-se tecnicamente reduzir a *Heterocedasticidade* gerada por poucas

amostras a partir do método ponderado.

4.3 TESTE DE SHAPIRO WILK

O teste de Shapiro-Wilk (W) (ROYSTON, 1982) foi proposto com a finalidade de averiguar se a distribuição de probabilidade de um conjunto de dados se aproxima pela distribuição normal. A estatística W é dada por:

$$W = \frac{b^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{41}$$

Onde o valores de x_i da amostra com tamanho n estão ordenados em ordem crescente. A constante b é definida da forma:

$$b = \begin{cases} \sum_{i=1}^{n/2} a_{n-i+1}(x_{n-i+1} - x_i) \text{ se n é par} \\ \sum_{i=1}^{(n+1)/2} a_{n-i+1}(x_{n-i+1} - x_i) \text{ se n é impar} \end{cases} \tag{42}$$

em que a_{n-i+1} é uma constante e os valores são tabelados[1]. Rejeitamos a hipótese da amostra prover de uma população normal se o $W_{calculado} < w_\alpha$, onde W_α, α nível de significância, provem da tabela[2] dos valores críticos da estatística W de Shapiro-Wilk

4.4 TESTE DE GOLDFELD-QUANDT

O teste de Goldfeld-Quandt (GQ) (GOLDFELD; QUANDT, 1965) tem como finalidade verificar a homoscedasticidade dos resíduos, comparar a variância de dois submodelos divididos por um ponto de quebra especifico. O teste é realizado sobre o conjunto das variáveis explicativas com tamanho n e é dividido em três etapas:

1. Os valores de cada observação x_i são ordenados em ordem crescente
2. c observações centrais são retiradas, sendo que c é especificado a priori e em seguida os dados restantes são divididos em dois grupos com $(n-c)/2$ elementos.
3. Para cada conjunto com $(n-c)/2$ elementos é gerado uma regressão por mínimos quadrados ordinários e posteriormente devemos obter a soma dos quadrados dos resíduos de

[1] A tabela com os valores dos a_i estão no Anexo 1

[2] A tabela com os níveis de significância de W_α esta no Anexo 2

cada modelo (SQR_1 e SQR_2), sendo que o primeiro conjunto corresponde ao menores valores x_i e o segundo aos maiores valores x_i.

Após as etapas anteriores do teste a estatística F_{GQ} é dada pela forma:

$$F_{GQ} = \frac{SQR_2/gl}{SQR_1/gl} \tag{43}$$

com graus de liberdade (gl), $gl = \dfrac{(n-c-2k)}{2}$ onde k é a quantidade de parâmetros a serem estimados. Temos que F_{GQ} segue uma distribuição de $F - Snedecor$ com gl como parâmetro. Cabe ressaltar que são retiradas c observações do conjunto de dados justamente para acentuar a diferença entre as variâncias entre o grupos, isto é acentuar as diferenças entres SQR_1 SQR_2

4.5 TESTE DE DURBIN-WATSON

O teste de Durbin-Watson (DW) (DURBIN; WATSON, 1950) tem a finalidade de identificar a presença de autocorrelação nos erros de um modelo de regressão. Se caso os erros estiverem correlacionados, a regressora pode insuficientemente estimar o erro padrão dos coeficientes o que pode torna-los aparentemente significativos. A estatística de Durbin-Watson (DW) é dada por:

$$DW = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} \tag{44}$$

com e_i sendo o i-ésimo resíduo da observação. Pode-se verificar a presença de autocorrelação comparando o valor de DW com os valores críticos inferior e superior tabelados[3], respectivamente denominados como dL e dU.

- se $0 \leq DW < dL$ os erros são dependentes;
- se $dU \leq DW \leq dL$ o teste é inconclusivo;
- se $dU < DW < 4 - dU$ os erros são independentes;
- se $4 - dU \leq DW \leq 4 - dL$ o teste é inconclusivo;
- se $4 - dL < DW \leq 4$ os erros são dependentes.

Como neste trabalho de conclusão de curso estaremos lidando somente com amostras relativamente pequenas e com apenas uma variável explicativa escolhemos $dL = 0,81$ e $dU = 1,07$ para todos os testes de Durbin-Watson realizados.

[3] Tabela disponível no Anexo 3

Já o teste das hipóteses se baseia na suposição de que os erros da regressora são gerados por um processo auto-regressivo e tem a forma

$$e_i = \rho e_{i-1} + a_i \tag{45}$$

onde ε_i é o termo do erro do modelo na i-ésima observação e $a_i \sim N(0, \sigma_a^2)$ e ρ $(|\rho| < 1)$ é o parâmetro de autocorrelação. Pode-se verificar a presença de autocorrelação tomando a hipótese nula igual a zero e a hipótese alternativa sendo qualquer valor diferente de zero.

4.6 CRITÉRIO DE SELEÇÃO DE MODELOS DE AKAIKE

Um modelo é uma representação matemática simplificada de um problema ou de uma situação real e muitas vezes pode existir mais de um modelo capaz de ilustrar o mesmo fenômeno, sendo assim é necessário escolher qual é o mais adequado. O critério de informação de Akaike (AIC) (SAKAMOTO; ISHIGURO, 1986) é um estimador da qualidade de modelos estatísticos. Portanto buscamos através do teste AIC o modelo mais parcimonioso, que envolva o mínimo de parâmetros possíveis e que ainda explique bem o comportamento do conjunto de dados. O teste proposto por Akaike utiliza a função de máxima verossimilhança.

Dado um conjunto de dados $x_i (i = 1, .., n)$, um conjunto de parâmetros θ e um modelo estatístico. A função de verossimilhança é definida da forma:

$$L(\theta; x_1, ..., x_n) = f(x_1; \theta) \times, ..., \times (x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta) = L(\theta) \tag{46}$$

O estimador de máxima verossimilhança é o θ que maximiza $L(\theta)$ e estes valores podem ser calculados resolvendo o sistema de equação:

$$U(\theta) = \frac{\partial log(L(\theta))}{\partial \theta} = 0 \tag{47}$$

Deste modo o Critério de informação Akaike (AIC) é definido como:

$$AIC = -2log(L(\theta)) + 2p \tag{48}$$

em que $L(\theta)$ é a função de máxima verossimilhança de um modelo e p é a quantidade de variáveis explicativas no qual o modelo se baseou. Quanto menor for o valor obtido no teste AIC melhor será o ajuste da regressora em questão.

5 PROBLEMAS E PREMISSAS

Algumas dificuldades podem surgir no ajuste linear de curvas de crescimento populacional, como por exemplo, estimações imprecisas de parâmetros, resultados tendenciosos e alta variação dos erros. O ajuste dos mínimos quadrados toma cada ponto dentro de um conjunto discreto de igual peso e fornece uma curva que pode levar a valores numéricos distantes em comparação com a previsão esperada, notadamente quanto à capacidade suporte que pode ficar superestimada (ou subestimada) em excesso. Existem vários caminhos a se seguir, desde restringir o conjunto de informações ou até mesmo atribuir pesos a cada ponto do conjunto de dados e as respostas razoáveis podem somente ser alcançadas por testes que envolvem experimentações numéricas e analise de dados.

No trabalho *Crescimento Logístico da População do Brasil* de (SANTOS, 2011) foi utilizada a técnica dos mínimos quadrados e o autor toma outro caminho para calcular os valores numéricos que melhor se ajustam aos dados dispostos. Sabe-se que a Equação (10) é não-linear, porem fazendo uma manipulação algébrica tem-se:

$$\frac{1}{dP}\frac{dP}{dt} = a + bP. \tag{50}$$

Desta maneira o lado direito da equação é linear e o lado esquerdo pode-se aproximar através da técnica das diferenças finitas avançada, dada pela aproximação:

$$\frac{1}{P_i}\frac{dP}{dt}(t_i) \approx g_i = \frac{1}{P_i}\frac{P(t_{i+1}) - P(t_i)}{t_{i+1} - t_i}, \tag{51}$$

e pela diferença finita retrógrada:

$$\frac{1}{P_i}\frac{dP}{dt}(t_i) \approx h_i = \frac{1}{P_i}\frac{P(t_{i-1}) - P(t_i)}{t_{i-1} - t_i} \tag{52}$$

Neste caso, (SANTOS, 2011) primeiramente calculou as diferenças finitas para cada ponto t_i e tomou a média aritmética de g_i e h_i. Assim:

$$\frac{1}{P_i}\frac{dP}{dt}(t_i) = a + bP(t_i) \approx \frac{g_i + h_i}{2}, \tag{53}$$

para cada t_i. Ao utilizar o método dos mínimos quadrados foram encontrados os parâmetros a e b da reta que melhor se ajustavam ao conjunto de ponto.

Seguindo esta estratégia o autor realizou modelagem utilizando o dados oficiais do crescimento populacional do Brasil de 1950 a 2000. Os valores referentes de recenseamentos estão disponíveis na Tabela 8.

Tabela 8: Recenseamentos realizados no Brasil.

Ano	População (em milhares)
1950	52
1960	70
1970	93
1980	119
1991	147
2000	170

Fonte: IBGE

O autor utilizou as diferença finitas sobre as amostras populacionais do Brasil e com isto chegou aos valores presentes na Tabela 9

Tabela 9: Diferenças finitas referentes as amostras populacionais brasileira.

Variável explicativa (em milhares)	Variável resposta
141	0,0293
181	0,0263
361	0,0216
624	0,0174

Os parâmetros obtidos foram $a = 0,04$ e $b = -1,58 \times 10^{-10}$, ao substituir $P_0 = 170$ milhões e $t_0 = 2000$ na equação logística o modelo gerado tem a forma:

$$P(t) = \frac{257 \cdot 10^6}{1 + 0,51e^{-0,04(t-2000)}} \tag{54}$$

Ao aplicar a função sobre alguns pontos tem-se $P(2010) = 190,7$ milhões, $P(2037) = 231$ milhões e $P(2096) = 254$ milhões, alem do mais segundo o autor o erro destas aproximações sobre os pontos das amostras são de $0,5\%$, o que nos indica que o modelo calculado por Santos é uma boa preditora.

Na dissertação de mestrado *Equações de Diferenças na Projeção de Populações* de (NOVAKI, 2017) trabalhou-se com a variação populacional em Curitiba, usando dados oficiais, de 1920 a 2010. Pela Tabela 10 é possível visualizar estes dados.

Tabela 10: EVOLUÇÃO POPULACIONAL DE CURITIBA.

Ano	População (Dados em Milhares)
1920	79
1940	141
1950	181
1960	361
1970	624
1980	1,025
1991	1,315
2000	1,587
2010	1,752

Fonte: IBGE

Neste trabalho a autora procura modelar a curva logística que melhor representa a dinâmica populacional da capital. Na Figura 12 é possível verificar o *diagrama de dispersão* da Tabela 10.

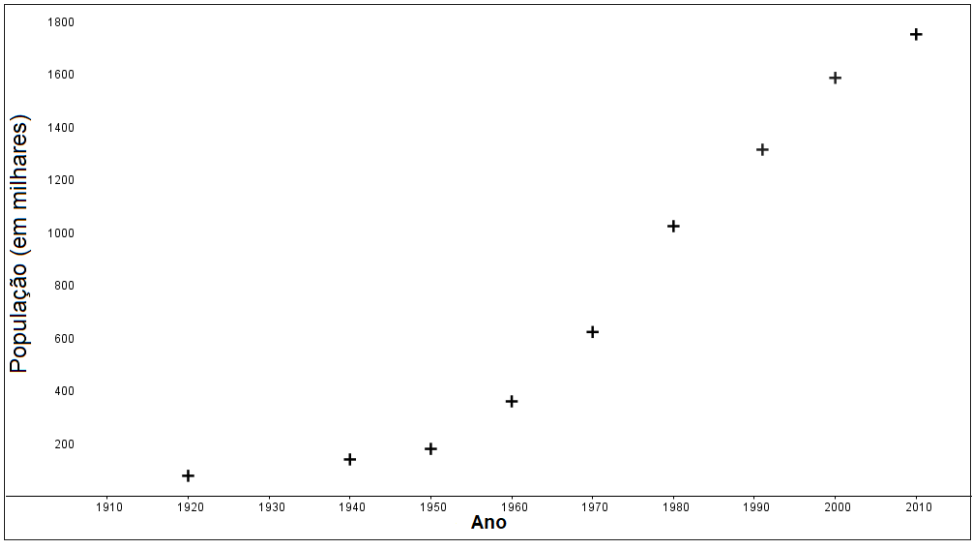

Figura 12: Diagrama de dispersão: dinâmica populacional de Curitiba

Novaki segue a mesma estratégia de calcular a media aritmética das diferenças finitas para alcançar os parâmetros a e b da função logística. Na Tabela 11 podemos observar os valores das médias das diferenças finitas

Os parâmetros obtidos foram $a = 0,056$ e $b = -2,3746 \times 10^{-5}$. Ao substituir $P_0 = 1,587$ milhões e $t_0 = 2000$ na equação logística o modelo gerado tem a forma:

Tabela 11: Diferenças finitas presentes na tese de Novaki.

Variável explicativa (em milhares)	Variável resposta
141	0,025
181	0,061
361	0,061
624	0,053
1,025	0,032
1,315	0,022
1,587	0,015

$$P(t) = \frac{2367 \cdot 10^5}{1 + 0,5e^{-5,6(t-2000)}} \tag{55}$$

Mesmo seguindo o mesmo método utilizado por Santos, em comparação ao dados fornecido pelo IPARDES (Instituto Paranaense de Desenvolvimento Econômico e Social) as previsões para 2020 e 2030 calculadas por Novaki de 2,041 milhões e de 2,170 milhões de habitantes, respectivamente, está muito acima das previsões de 2020 e 2030 de 1,945 milhões e 2,031 milhões de habitantes fornecidas pelo Instituto.

Segundo a autora isto se deve pois o modelo calculado por ela parece razoável para reproduzir a dinâmica da população desde 1920, porém a função não acaba se encaixando para projetar populações futuras. Segundo Novaki isto se deve porque a capacidade suporte do meio também se modifica com o passar dos anos, por vários fatores como por exemplo, epidemias e mudanças nas taxas de natalidade e mortalidade. Para evitar a influência de dados passados a autora restringiu o intervalo do censo, dispensando valores mais antigos e calculou as respectivas previsões. Esta restrição pode ser observada na tabela 12.

Tabela 12: PREVISÃO POPULACIONAL DE CURITIBA.

Intervalo	Previsão 2020 (em milhões)	Previsão 2030 (em milhões)
[1920,2010]	2,041	2,170
[1940,2010]	1,857	1,909
[1950,2010]	1,819	1,860
[1960,2010]	1,817	1,857
[1970,2010]	1,887	1,954
[1980,2010]	1,934	2,029

Então foi gerado o modelo descrito abaixo com população máxima de 2,176 milhões

$$P(t) = \frac{2176 \cdot 10^5}{1 + 0,37e^{-0,54(t-2000)}}. \tag{56}$$

Nota-se que as previsões mais recentes tem valores mais próximos a do IPARDES para 2020 e 2030.

Pode-se observar neste dois exemplos que no processo de calcular os parâmetros da curva logística podem surgir certas dificuldades, tal como valores que não condizem com as previsões esperadas. Os autores exemplificados tomaram estratégias numéricas com o propósito de diminuir estas disparidades de valores, porém para uma pesquisa deve-se sempre buscar alternativas e métodos eficazes para se alcançar bons resultados e este projeto de conclusão de curso também alcança resultados precisos, porém tomando rumos diferentes dos trabalhos descritos anteriormente.

5.1 MODELOS GERADOS ATRAVÉS DOS MÍNIMOS QUADRADOS

Implementamos as diferenças finitas (Deve-se ter em mente quando aplicamos as diferenças finitas no conjunto de dados, a quantidade populacional se torna a variável explicativa e os valores obtidos pela diferenças a variável resposta.) e aplicamos mínimos quadrados ordinários nos dados da tabela 10 a fim de obter um modelo que correspondesse as previsões estipuladas pelo IPARDES e também as do trabalho da Novaki. Na tabela 13 temos que a primeira coluna são valores da variável explicativa, a segunda coluna contém os valores da variável resposta, já a terceira e quarta coluna correspondem aos pesos utilizados no processo de regressão. Os pesos foram gerados e atribuídos respectivamente a partir da distância de cada ponto da média do conjunto da variável dependente, na figura 13 podemos observar graficamente a dimensão de cada peso. A partir deste dados foram gerados modelos através do método ordinário e também do ponderado.

Tabela 13: Diferenças finitas aplicadas aos dados do município de Curitiba

Variável explicativa	Variável Resposta	Peso w_i	Peso $1/w_i$
141	0,02517730	0,0132756	75,32613
181	0,06077348	0,0223205	44,80172
361	0,06135734	0,0229044	43,65968
624	0,05320513	0,0147522	67,78642
1025	0,03242129	0,0060316	165,79282
1315	0,02151554	0,0169373	59,04103
1587	0,01472030	0,0237326	42,13611

Para o primeiro modelo realizamos a regressão sem peso e obtivemos o coeficiente angular $b = -0,002375$, o coeficiente linear $a = 5,620832$ e $P_{max} = 2,366$ milhões. Usando

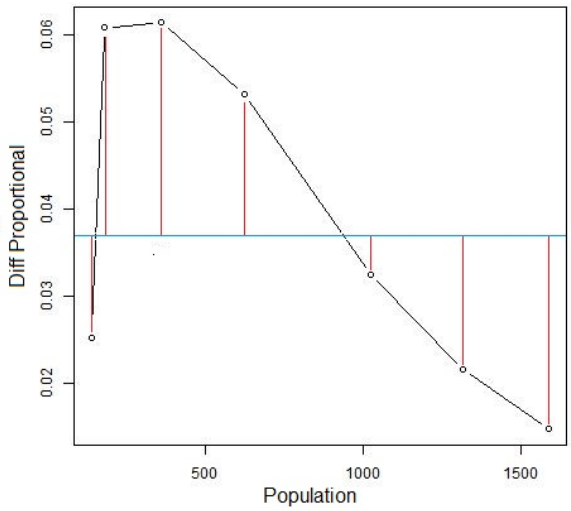

Figura 13: A reta em azul representa a posição da média do conjunto da variável explicativa e as retas em vermelho correspondem a distancia de cada ponto da média

$t_0 = 2010$ e $P_0 = 1752$ milhões obtivemos

$$P(t) = \frac{2366 \cdot 10^5}{1 + 0,3508e^{-5,62(t-2010)}} \tag{57}$$

Note que neste ajuste ocorreu uma superestimação de parâmetros, já que a população máxima superou a aproximação proposta por Novaki, em sua dissertação de mestrado, logo o modelo também não se encaixou nas estimativas do IPARDES. Para fins comparativos, realizamos previsões e ocorreu que o modelo também cresce rapidamente, $P(2010) = 1,751$milhões, $P(2020) = 2,365$ milhões e $P(2030) = 2,365$. Observe que em 10 anos ja atinge valores próximos da capacidade suporte. Portanto rejeitamos esta regressora como função de previsão para a população de Curitiba.

Para o segundo modelo implementamos o peso $\frac{1}{w_i}$, a escolha deste peso teve como finalidade atribuir um foco maior aos ponto mais próximos da média da amostra e descontar a precisão dos mais distantes, observe que na escolha deste peso estamos focando em dados mais centrais, ou seja os mais próximos da reta azul como pode ser verificado na imagem 13. Geramos pela regressão linear o coeficiente angular $b = -0,001917$, o coeficiente linear $a = 5,149338$ e $P_{max} = 2,686$ milhões. Usando $t_0 = 2010$ e $P_0 = 1752$ milhões obtivemos

$$P(t) = \frac{2686 \cdot 10^5}{1 + 0,533e^{-5,14(t-2010)}} \tag{58}$$

Constamos que houve uma superestimação ainda maior da população máxima do que

a anterior e ainda mais para $P(2010) = 1,752$ milhões e $P(2020) = 2,685$ milhões evidência que este modelo também cresce rapidamente, consequente não o aceitamos como previsor.

No terceiro modelo atribuímos os pesos w_i. A escolha deste peso teve como finalidade acrescentar maior focagem aos ponto ao extremo do gráfico, em específico os mais próximos a direita e diminuir a importância dos pontos centrais (aqueles mais próximos da reta que representa a média da amostra), ou seja, o objetivo nesta ponderação são os valores censitários mais recentes, porém podemos reparar na imagem 13 que os pontos mais a esquerda do gráfico também ganham maior foco. Para fins de teste geramos o coeficiente angular $b = -0,00274$, o coeficiente linear $a = 6,03795$ e $P_{max} = 2,203$ milhões. Usando $t_0 = 2010$ e $P_0 = 1752$ milhões elaboramos a regressora

$$P(t) = \frac{2203 \cdot 10^5}{1 + 0,257e^{-6,03(t-2010)}} \tag{59}$$

Novamente, obtivemos valores insatisfatórios com a população máxima acima da esperada e se aplicarmos a função em $f(2010) = 1752$ milhões e em $f(2020) = 2,202$ milhões, o modelo apresenta desenvolvimento acelerado na variável dependente.

Por fim para a quarta regressora atribuímos $w = (1,2,3,4,5,6,7)$, apesar de ser uma escolha simplória ainda sim está dentro dos objetivos, já que optamos em aplicar os mínimos quadrados ponderados justamente para dar foco maior aos recenseamentos mais recentes (Imaginamos que deste modo o modelo gerado pode explicar com qualidade o crescimento atual da população de Curitiba.) e assim atingir pelo menos uma capacidade suporte satisfatória, com isto foi produzido o coeficiente angular $b = -0,003206$, o coeficiente linear $a = 6,578779$ e $P_{max} = 2,052$ milhões. Usando $t_0 = 2010$ e $P_0 = 1752$ milhões foi gerado o modelo:

$$P(t) = \frac{2052 \cdot 10^5}{1 + 0,1,71e^{-6,58(t-2010)}} \tag{60}$$

Ao contrário dos exemplos anteriores este modelo em específico atingiu uma capacidade suporte aceitável, porém ao aplicarmos $P(2010) = 1752$ milhões e $P(2020) = 2,051$ milhões, os resultados apontam novamente para um desenvolvimento acelerado da função.

A partir dos quatros modelos descritos acima podemos reparar que apesar da capacidade suporte variar em cada regressão ainda assim o crescimento acelerado continua a ocorrer, o que não condiz com o caso real. Então, acreditamos que exista um fator desconhecido no modelo logístico que limite o desenvolvimento exagerado.

6 ANÁLISE ATRAVÉS DO R

O *Software R* é um *programa de desenvolvimento integrado*, ou seja, um *Software* que reúne características e ferramentas com o objetivo de agilizar processos. Este programa é usado por estatísticos e analistas de dados que buscam obter resultados através de análise de dados. Utilizamos o teste de *Shapiro-Wilk*, *Goldfeld-Quandt*, *Durbin-Watson* e o *Critério de Informação Akaike* através do *R* para obter resultados sobre a natureza dos modelos presentes nesta pesquisa.

6.1 DADOS DO MUNICÍPIO DE CURITIBA

Para o conjunto de dados da Tabela 10, referentes ao crescimento populacional do município de Curitiba, foi verificado se tais análises satisfazem as premissas do modelo de regressão linear padrão. Foi realizado o teste de Goldfeld-Quandt, onde $GQ = 1,1376$ e como $p - valor = 0,4993$ rejeitamos assim a presença heterocedasticidade nos dados. Construímos pelo *R* os possíveis modelos lineares generalizados, e para fins comparativos, calculamos o critério de seleção Akaike sobre os modelos gerados, cujos resultados podem ser observados na Tabela 15.

Tabela 15: Modelos construídos pelo R.

Modelos	AIC
MLG Binomial com FDL logito	6,216151
MLG Gamma com FDL inversa	-27,566071
MLG Gaussiana com FDL identidade	-35,069237

O conjunto da variável resposta foi ajustado em escala de maneira que o somatório deste valores seja igual a um. Esta conversão para quantidades proporcionais foi necessária para gerar o modelo generalizado linear binomial. Note que o modelo binomial com função de ligação logito teve um dos piores ajuste, isto evidencia que nem toda amostra originada de um determinado experimento ou fenômeno pode se encaixar em modelos esperados ou preestabe-

lecidos. As regressoras da família gaussiana e gamma, respectivamente, obtiveram os melhores resultados em comparação pelo teste Akaike.

Deve-se ter em mente que o teste de Shapiro-Wilk não faz sentido nos modelos lineares generalizados com distribuição binomial, gamma, possion, quasi, quasibinomial e quasipossoin, pois são modelos nos quais não seguem a distribuição normal, sendo assim aplicamos o teste de normalidade no modelo gaussiano e com isto obtivemos $W = 0,94643$ e o $p-valor = 0,651$ e como $p-valor > 0,05$ do nível de significância temos que o modelo gaussiano realmente tem distribuição normal, isto significa que os resultados não são tendenciosos e o resíduo gerado tem variância mínima.

Através do teste de Durbin-Watson verificamos se os modelos apresentavam a correlação serial entre seus erros e como pode ser observado para todos os modelos o resultado foi o mesmo com $DW = 0,77453$ e $p-valor = 0,001824$ e como $p-valor < 0,05$ do nível de significância, portanto rejeitamos a hipótese de presença de autocorrelação nos erros, sendo assim podemos dizer que a covariância de erros é nula, ou seja os resíduos não dependem entre sí e são aleatórios.

Para fins de teste resolvemos mudar o polinômio de ajuste de segunda ordem do modelo gaussiano, com ajuste:

$$P_2 = 2.858e^{-5}t^2 - 1.094e^{-1}t + 1.046e^2 \tag{62}$$

e verificamos pelo teste Akaike $AIC = -45,556356$, ou seja este modelo se ajustou melhor que os demais e ainda satisfaz as premissas do modelo padrão proposto por Gauss, pois os testes de Shapiro-Wilk e de Durbin-Watson apresentaram os seguintes valores $W = 0,93928$ e $p-valor = 0,5743$, $DW = 1,202$ e $p-valor = 0,002812$, ou seja obtivemos um modelo mais preciso do que os anteriormente gerados, além do mais está afirmação pode ser verificada pela Tabela 16 dos valores numéricos dos modelos ajustados. Se comparamos a coluna da população (proporcional) com a da ultima coluna percebe-se que que os valores ajustados do modelo parabólico estão mais próximos dos valores reais do que os das demais colunas. Pode-se visualizar o comportamento dos modelos mencionados pela Figura 15.

Ao incorporamos as diferenças finitas no processo da obtenção das possíveis regressoras estamos mudando a configuração dos dados disponíveis, consequentemente novos testes devem ser realizados. Os modelos analisados são regressões baseadas nos dados da tabela 13. O gráfico com a dispersão dos pontos podem ser observados na Figura 16.

Aplicamos o teste de Goldfeld-Quandt neste conjunto de dados obtivemos $GQ = 0,01988$

Tabela 16: Comparação entre os valores reais com os ajustados

Ano	População (proporcional)	Binomial	Gaussiano	Gaussiano de segunda ordem
1920	0,01118188	0,01278990	0,03194286	0,03617530
1940	0,01995754	0,02710020	0,04013184	0,02394160
1950	0,02561925	0,03924119	0,04603235	0,05400005
1960	0,05109696	0,05650549	0,05396702	0,08405851
1970	0,08832272	0,08072707	0,06520683	0,11411696
1980	0,14508139	0,11407608	0,08236011	0,14417541
1991	0,18612880	0,16400201	0,11589659	0,17723970
2000	0,22462845	0,21682285	0,17379919	0,20429231
2010	0,24798301	0,28873523	0,39066321	0,23435076

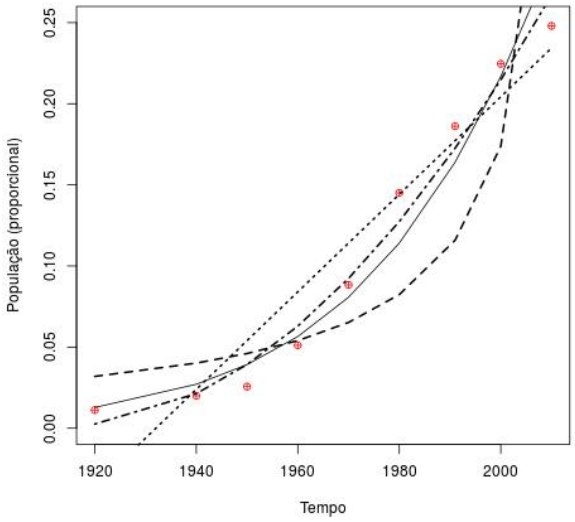

Figura 15: população $\times$ diferenças finitas (proporcional) - dados (pontos), modelo binomial (linha), modelo gamma (tracejado), modelo gaussiano (pontilhado) e modelo gaussiano de segunda ordem (pontilhado e tracejado)

com $p - valor = 0,987$ o que indica a ausência de heterocedasticidade, sendo assim possível aplicar o critério de AIC sobre os modelos gerados. A comparação pode ser observada na tabela 17, porém deve-se ter em mente que não podemos comparar com os valores da Tabela 15, pois estamos tratando de conjuntos de dados diferentes.

O teste AIC nos revela que o modelo com distribuição gamma apresentou o melhor ajuste dentre os demais. O Gráfico 17 nos mostra os modelos com os melhores desempenhos.

Aplicamos o teste de normalidade somente no modelo gaussiano e obtivemos como resultado $W = 0,86822$ e com $p - valor = 0,1791$, portanto os resíduos do modelo seguem a distribuição normal, consequentemente a regressora usufrui das propriedades do modelo linear

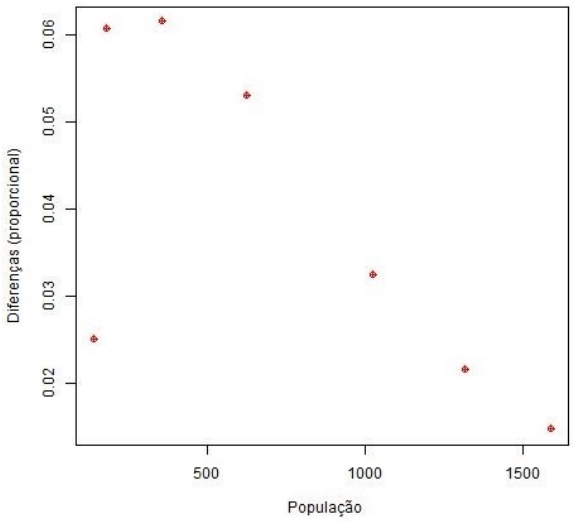

Figura 16: população $\times$ diferenças finitas (proporcional) - dados (pontos)

Tabela 17: Modelos construídos pelo R.

Modelos com diferenças finitas	AIC
MLG Binomial com FDL logito	6,179023
MLG Gamma com FDL inversa	-17,027603
MLG Gaussiana com FDL identidade	-16,4632206

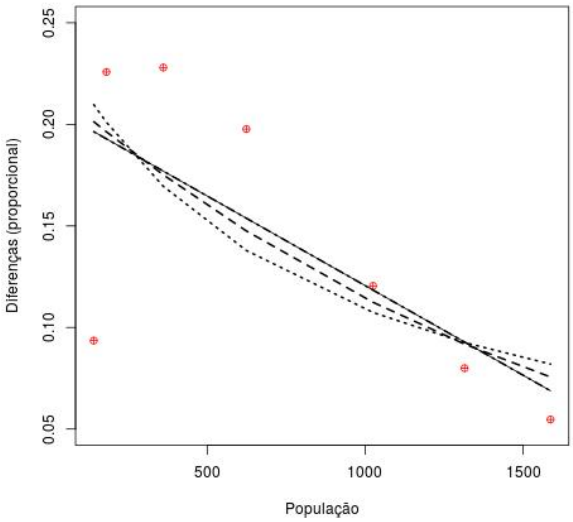

Figura 17: população $\times$ diferenças finitas (proporcional) - dados (pontos), modelo gaussiano (linha), modelo binomial (tracejado), modelo gamma (pontilhado)

gaussiano.

Pelo teste de Durbin-Watson foi verificado que modelos com diferenças finitas não apresentavam a autocorrelação entre seus erros, ja que para todas as funções geradas o resultados foram os mesmos com $DW = 1,2554$ e $p - valor = 0,03481$, portanto podemos dizer que não há presença de autocorrelação nos erros.

Ao alteramos o polinômio de ajuste do modelo gaussiano para o de segunda ordem

$$-1,247e^{-7}p^2 + 1,193e^{-4}p + 1,582e^{-1} \tag{63}$$

o de quarta ordem

$$-6,866e^{-13}p^4 + 2,723e^{-9}p^3 - 3,707e^{-6}p^2 + 1,839e^{-3}p - 5,693e^{-2} \tag{64}$$

e o de quinta ordem

$$1,664e^{15}p^5 - 7,815e^{-12}p^4 + 1,381e^{-8}p^3 - 1,130e^{-5}p^2 + 4,017e^{-3}p - 2,522e^{-1} \tag{65}$$

Pode ser verificado pela Tabela 18 que os modelos das duas ultimas linhas da tabela obtiveram melhor desempenho do que o caso linear pelo AIC.

Tabela 18: Modelos construídos pelo R.

Modelos com diferenças finitas	AIC
Modelo gaussiano	-16,463226
Modelo gaussiano ajuste de segunda ordem	-16,458417
Modelo gaussiano ajuste do quarto grau	-18,321437
Modelo gaussiano ajuste do quinto grau	-18,246283

Na Figura 18 podemos observar o comportamento de cada modelo, repare que quanto mais aumentamos a ordem do polinômio de ajuste, melhor a curva se ajusta aos pontos dispostos no gráfico.

Além do mais se observarmos os valores d da Tabela 19 percebe-se que os resultados mais a direta da tabela com os valores das diferenças finitas, notamos que os resultados destes modelos se aproximam melhor dos valores reais, isto indica que as curvas se ajustaram consideravelmente aos pontos em relação aos modelos que não tiveram o seu polinômio de ajuste alterado. Repare também que os últimos resultados da ultima coluna, correspondente ao modelo de quinta ordem, se aproximam melhor aos últimos resultados da coluna referente as diferenças finitas, isto indica que apesar de ser o segundo segundo melhor modelo pelo critério de informação de Akaike ainda sim é o modelo capaz de oferecer as melhores previsões, além do mais este comportamento também pode ser visualizado no Gráfico 18.

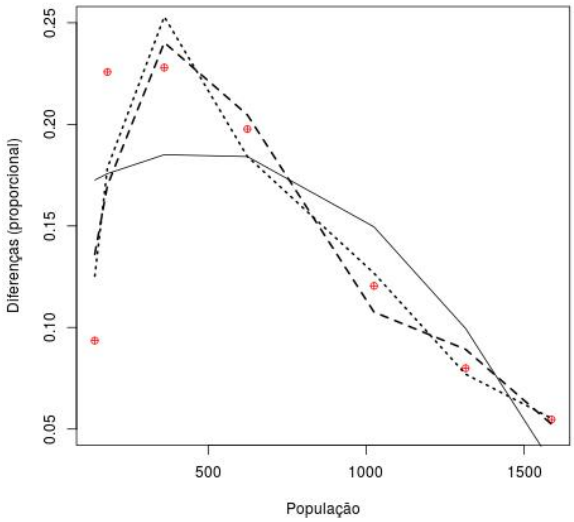

Figura 18: população $\times$ diferenças finitas (proporcional) - dados (pontos), modelo gaussiano quadrático (linha), modelo gaussiano de quarta ordem (tracejado), modelo gaussiano de quinta ordem (pontilhado).

Tabela 19: Comparação entre os valores das diferenças finitas com os dos modelos ajustados

Diferenças finitas	gaussiano	se segunda ordem	Quarta ordem	Quinta ordem
0,09353669	0,19638157	0,1725875	0,13602774	0,12529247
0,22578072	0,19285277	0,1757533	0,16988967	0,17858917
0,22794983	0,17697314	0,1850607	0,24028685	0,25299266
0,19766339	0,15377124	0,1841302	0,20465113	0,18417420
0,12044894	0,11839495	0,1495059	0,10756777	0,12670498
0,07993278	0,09281111	0,0994755	0,08921803	0,07700683
0,05468766	0,06881522	0,0334869	0,05235881	0,05523969

As premissas do modelo padrão também são verificadas para os modelos gaussiano de segunda, de quarta e de quinta ordem. A Tabela 20 contêm os resultados do teste de Shapiro-Wilk para as respectivas preditoras. Pelos $p-valor$ temos que as regressoras gaussianas presentes na tabela não são tendenciosas e a variância dos seus resíduos são mínimas.

Já a Tabela 21 contem os resultados do teste de Durbin-Watson para os modelos gaussianos e pelo $p-valor$ verificamos que não há presença de autocorrelação nos resíduos de cada modelo, portanto temos que os coeficientes que descrevem as preditoras são verdadeiramente significativos.

Repare que o teste AIC nos informou quais modelos são os mais parcimoniosos e os demais nos dizem que estes três últimos modelos se encaixam dentro das condições do modelo padrão, com es isto podemos dizer que as modificações na ordem dos polinômios geraram

Tabela 20: Resultados do teste de normalidade para os modelos gaussianos.

Modelos	W	p-valor
Modelo gaussiano de segunda ordem	0,93652	0.6076
Modelo gaussiano de quarta ordem	0,92466	0,5065
Modelo gaussiano de quinta ordem	0,94425	0,6772

Tabela 21: Resultados do teste de Durbin-Watson para os modelos lineares.

Modelos	DW	p-valor
Modelo gaussiano de segunda ordem	1,6978	0,02652
Modelo gaussiano de quarta ordem	2,8552	0,1661
Modelo gaussiano de quinta ordem	3,2795	$p-valor < 2,2e^{-16}$

modelos melhores dos quais não tiveram esta modificação implementada.

7 CONSIDERAÇÕES FINAIS

O estudo da dinâmica populacional envolve a escolha de modelos matemáticos adequados para o conjunto de dados e conhecimentos estatísticos para realização de testes que averiguem a qualidade das previsoras geradas. O processo de estimar uma regressora que explique com precisão o desenvolvimento de uma população pode levar a dificuldades. Escolhemos o método de regressão dos mínimos quadrados, e suas variações, para gerar as curvas com os dados que nos foram disponíveis. Percebemos que para uma boa estimação é fundamental a escolha de um modelo de crescimento adequado, porém este procedimento não foi simples e testes estatísticos comparativos foram necessários.

Discutimos sobre alguns modelos de crescimento populacional, como o de Malthus e de Verhulst, sobre as respectivas soluções estacionárias e sua contextualização biológica, como a capacidade suporte de uma população que o modelo logístico nos fornece.

Ainda discorremos sobre os trabalhos de Reginaldo dos Santos e da Cristiane Novaki onde estes autores utilizaram as diferenças finitas para estimar linearmente seus respectivos modelos e em especifico cita o trabalho da Novaki, no qual restringiu seu conjunto de dados, dando enfase aos recenseamentos mais recentes, e o resultado foi uma regressora com população máxima de 2,176 milhões compatível com a da projeção do IPARDES.

Devemos salientar que existem situações que estes modelos de previsão podem não ser suficientes para descrever o caso real, um exemplo disto são as capacidades suportes superestimadas que obtivemos de 2,366 milhões para o método dos mínimos quadrados ordinários e 2,868 milhões, 2,203 milhões, 2,052 milhões para o método ponderado com diferentes tipos de pesos, além do mais todas as funções de crescimento citadas apresentaram desenvolvimento acelerado e uma possível solução para este comportamento seria adotar o modelo logístico generalizado, pois quando o parâmetro $\theta > 1$, a previsora apresenta crescimento mais lento, porém esta estratégia tomaremos como uma possível pesquisa futura.

Durante o estudo da nossa modelagem implementamos alguns métodos de análise, utilizamos os teste de *Shapiro-Wilk*, *Goldfeld-Quandt*, *Durbin-Watson* e o *Critério de informação*

Akaike para verificar se os modelos gerados se encaixavam no modelo padrão proposto por Gauss, percebemos que mesmo satisfazendo as premissas as regressoras ainda não obtiveram a qualidade que esperávamos. Sendo assim foi necessário mudar o polinômio de ajuste na regressão.

REFERÊNCIAS

CUNHA, M. C. C. **Métodos numéricos**. São Paulo: Editora da Unicamp, 2013.

DURBIN, J.; WATSON, G. Testing for serial correlation in least squares regression i. **Biometrika 37**, 1950.

FIGUEIREDO, D. G.; NEVES, A. F. **Equações Diferencias Aplicadas**. Rio de Janeiro: IMPA, 2010.

GOLDFELD, S.; QUANDT, R. Some tests for homoskedasticity. **Journal of the American Statistical Association 60**, 1965.

GUJARATI, D. **Econometria Básica**. Rio de Janeiro: CAMPUS, 2006.

NOVAKI, C. **Equações de Diferenças na Projeção de populações**. Dissertação (Mestrado) — Universidade Tecnologica Federal do Parana, 2017.

ROYSTON, P. An extension of shapiro and wilk's w test for normality to large samples. **Applied Statistics**, 1982.

SAKAMOTO, Y.; ISHIGURO, M. Akaike information criterion statistics. **D. Reidel Publishing Company**, 1986.

SANTOS, R. J. **Crescimento Logístico da População do Brasil**. 2011. Disponível em: <http://www.mat.ufmg.br/ regi>. Acesso em: 8 de novembro de 2009.

SHALIZI, C. **Extending Linear Regression: Weighted Least Squares, Heteroskedasticity, Local Polynomial Regression**. 2009. Disponível em: <http://www.stat.cmu.edu/ cshalizi/350/lectures/18/lecture-18i>. Acesso em: 25 de maio de 2018.

VLADAR, H. P. d. Density-dependence as a size-independent regulatory mechanism. **Elsevier**, 2005.

ZILL, D. G. **Equações Diferenciais com Aplicações em Modelagem**. São Paulo: Cengage Learning, 2003.

ZILL, D. G.; CULLEN, M. R. **Equações Diferenciais**. São Paulo: Makron Books, 2001.

8 ANEXOS

8.1 ANEXO 1

n =	15	16	17	18	19	20	21	22	23	24	25	26
a1	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407
a2	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043
a3	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533
a4	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151
a5	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836
a6	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563
a7	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316
a8		0.0196	0.0359	0.0496	0.0612	0.0711	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089
a9				0.0163	0.0303	0.0422	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876
a10						0.0140	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672
a11								0.0122	0.0228	0.0321	0.0403	0.0476
a12									0.0000	0.0107	0.0200	0.0284
a13											0.0000	0.0094

8.2 ANEXO 2

n \ p	0.01	0.02	0.05	0.1	0.5	0.9	0.95	0.98	0.99
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	0.687	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
5	0.686	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
10	0.781	0.806	0.842	0.869	0.938	0.972	0.978	0.983	0.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	0.805	0.828	0.859	0.883	0.943	0.973	0.979	0.984	0.986
13	0.814	0.837	0.866	0.889	0.945	0.974	0.979	0.984	0.986
14	0.825	0.846	0.874	0.895	0.947	0.975	0.980	0.984	0.986
15	0.835	0.855	0.881	0.901	0.950	0.975	0.980	0.984	0.987
16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
17	0.851	0.869	0.892	0.910	0.954	0.977	0.981	0.985	0.987
18	0.858	0.874	0.897	0.914	0.956	0.978	0.982	0.986	0.988
19	0.863	0.879	0.901	0.917	0.957	0.978	0.982	0.986	0.988
20	0.868	0.884	0.905	0.920	0.959	0.979	0.983	0.986	0.988
21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
22	0.878	0.892	0.911	0.926	0.961	0.980	0.984	0.987	0.989
23	0.881	0.895	0.914	0.928	0.962	0.981	0.984	0.987	0.989
24	0.884	0.898	0.916	0.930	0.963	0.981	0.984	0.987	0.989
25	0.888	0.901	0.918	0.931	0.964	0.981	0.985	0.988	0.989
26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
27	0.894	0.906	0.923	0.935	0.965	0.982	0.985	0.988	0.990
28	0.896	0.908	0.924	0.936	0.966	0.982	0.985	0.988	0.990
29	0.898	0.910	0.926	0.937	0.966	0.982	0.985	0.988	0.990
30	0.900	0.912	0.927	0.939	0.967	0.983	0.985	0.988	0.990
31	0.902	0.914	0.929	0.940	0.967	0.983	0.986	0.988	0.990
32	0.904	0.915	0.930	0.941	0.968	0.983	0.986	0.988	0.990
33	0.906	0.917	0.931	0.942	0.968	0.983	0.986	0.989	0.990
34	0.908	0.919	0.933	0.943	0.969	0.983	0.986	0.989	0.990
35	0.910	0.920	0.934	0.944	0.969	0.984	0.986	0.989	0.990
36	0.912	0.922	0.935	0.945	0.970	0.984	0.986	0.989	0.990
37	0.914	0.924	0.936	0.946	0.970	0.984	0.987	0.989	0.990
38	0.916	0.925	0.938	0.947	0.971	0.984	0.987	0.989	0.990
39	0.917	0.927	0.939	0.948	0.971	0.984	0.987	0.989	0.991
40	0.919	0.928	0.940	0.949	0.972	0.985	0.987	0.989	0.991
41	0.920	0.929	0.941	0.950	0.972	0.985	0.987	0.989	0.991
42	0.922	0.930	0.942	0.951	0.972	0.985	0.987	0.989	0.991
43	0.923	0.932	0.943	0.951	0.973	0.985	0.987	0.990	0.991
44	0.924	0.933	0.944	0.952	0.973	0.985	0.987	0.990	0.991
45	0.926	0.934	0.945	0.953	0.973	0.985	0.988	0.990	0.991
46	0.927	0.935	0.945	0.953	0.974	0.985	0.988	0.990	0.991
47	0.928	0.936	0.946	0.954	0.974	0.985	0.988	0.990	0.991
48	0.929	0.937	0.947	0.954	0.974	0.985	0.988	0.990	0.991
49	0.929	0.939	0.947	0.955	0.974	0.985	0.988	0.990	0.991
50	0.930	0.938	0.947	0.955	0.974	0.985	0.988	0.990	0.991

8.3 ANEXO 3

	Nível de significância	Número de variáveis explicativas									
		1		2		3		4		5	
n		d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
	0,01	0,81	1,07	0,7	1,25	0,59	1,46	0,49	1,7	0,39	1,96
15	0,025	0,95	1,23	0,83	1,4	0,71	1,61	0,59	1,84	0,48	2,09
	0,05	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
	0,01	0,95	1,15	0,86	1,27	0,77	1,41	0,63	1,57	0,6	1,74
20	0,025	1,08	1,28	0,99	1,41	0,89	1,55	0,79	1,7	0,7	1,87
	0,05	1,2	1,41	1,1	1,54	1	1,68	0,9	1,83	0,79	1,99
	0,01	1,05	1,21	0,98	1,3	0,9	1,41	0,83	1,52	0,75	1,65
25	0,025	1,13	1,34	1,1	1,43	1,02	1,54	0,94	1,65	0,86	1,77
	0,05	1,2	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
	0,01	1,13	1,26	1,07	1,34	1,01	1,42	0,94	1,51	0,88	1,61
30	0,025	1,25	1,38	1,18	1,46	1,12	1,54	1,05	1,63	0,98	1,73
	0,05	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83
	0,01	1,25	1,34	1,2	1,4	1,15	1,46	1,1	1,52	1,05	1,58
40	0,025	1,35	1,45	1,3	1,51	1,25	1,57	1,2	1,63	1,15	1,69
	0,05	1,44	1,54	1,39	1,6	1,34	1,66	1,29	1,72	1,23	1,79
	0,01	1,32	1,4	1,28	1,45	1,24	1,49	1,2	1,54	1,16	1,59
50	0,025	1,42	1,5	1,38	1,54	1,34	1,59	1,3	1,64	1,26	1,69
	0,05	1,5	1,59	1,46	1,63	1,42	1,67	1,38	1,72	1,34	1,7

8.4 TESTES REALIZADOS NO R

```
> ########################################################################
#####
> #       Recarregando pacotes
> ########################################################################
#####
> rm(list=ls())
> require(scatterplot3d)
> require(gdata)
> require(stringr)
> require(broom)
> require(geepack)
> require(hnp)
> require(lmtest)
> require(ggplot2) #graficos
> rm(dados)
>
> ########################################################################
#####
> #       Entrada dos dados
> ########################################################################
#####
> t<-c(1920, 1940, 1950, 1960, 1970, 1980, 1991, 2000, 2010)
> p<-c(79, 141, 181, 361, 624, 1025, 1315, 1587, 1752)
# populacao em T
> tt<-t
> pp<-p
> dados<-data.frame(t,p)
>
> ########################################################################
#####
> #       Teste de Goldfeld-Quandt
> ########################################################################
#####
>
> gqtest(p~t)

        Goldfeld-Quandt test

data:  p ~ t
GQ = 1.1376, df1 = 3, df2 = 2, p-value = 0.4993
alternative hypothesis: variance increases from segment 1 to 2

>
> ########################################################################
#####
> #       Teste de Durbin-Watson
> ########################################################################
#####
>
> dwtest(p~t)

        Durbin-Watson test

data:  p ~ t
DW = 0.77453, p-value = 0.001824
alternative hypothesis: true autocorrelation is greater than 0

>
> ########################################################################
#####
> #       Graficos
```

```
> ########################################################################
#####
>
> #X11()
> jpeg(file = "C:/Users/Rodolfo.RODOLFO-PC/Desktop/Pessoal/rodolfoTCC1
/Imagens curitiba/figura1.jpeg") # FIGURA 1
> plot(dados,type="b")
> dev.off()
null device
          1
>
> ########################################################################
#####
> #       Mudança de escala
> ########################################################################
#####
>
> p_prop<-prop.table(p)
> dadosP<-data.frame(t,p_prop)
>
> ########################################################################
#####
> #       Modelos
> ########################################################################
#####
>
> modeloL_P<-lm(p~t)
> modelo1_P<- glm(p_prop~t, family=binomial(link = "logit"))
> modelo2_P<- glm(p_prop~t, family=Gamma(link = "inverse"))
> modelo3_P<- glm(p_prop~t, family=gaussian(link = "identity"))
> modelo4_P<- glm(p_prop~t, family=poisson(link = "log"))
> modelo5_P<- glm(p_prop~t, family=quasi(link = "identity", variance =
"constant"))
> modelo6_P<- glm(p_prop~t, family=quasibinomial(link = "logit"))
> modelo7_P<- glm(p_prop~t, family=quasipoisson(link = "log"))
>
> ########################################################################
#####
> #       Critério de informação Akaike
> ########################################################################
#####
>
> AIC(modeloL_P,modelo1_P,modelo2_P,modelo3_P,modelo4_P,modelo5_P,mode
lo6_P,modelo7_P)
          df         AIC
modeloL_P  3 124.463112
modelo1_P  2   6.216151
modelo2_P  3 -27.566021
modelo3_P  3 -35.069237
modelo4_P  2         Inf
modelo5_P  2          NA
modelo6_P  2          NA
modelo7_P  2          NA
>
> ########################################################################
#####
> #       Tabela de comparações entre valores ajustados e valores reai
s
> ########################################################################
#####
>
```

```
> valores<-cbind(p_prop,modelo1_P$fitted.values,modelo2_P$fitted.value
s,modelo3_P$fitted.values)
> valores
      p_prop
1 0.01118188 0.01278990 0.03194286 -0.03617530
2 0.01995754 0.02710020 0.04013184  0.02394160
3 0.02561925 0.03924119 0.04603235  0.05400005
4 0.05109696 0.05650549 0.05396702  0.08405851
5 0.08832272 0.08072707 0.06520683  0.11411696
6 0.14508139 0.11407608 0.08236011  0.14417541
7 0.18612880 0.16400201 0.11589659  0.17723970
8 0.22462845 0.21682285 0.17379919  0.20429231
9 0.24798301 0.28873523 0.39066321  0.23435076
> ##########################################################################
#####
> #       Graficos
> ##########################################################################
#####
>
>   #X11()
>  jpeg(file = "C:/Users/Rodolfo.RODOLFO-PC/Desktop/Pessoal/rodolfoTCC
1/Imagens curitiba/figura3.jpeg") # FIGURA 3
>  plot(dadosP,type="b")
>  lines(modelo3_P$fitted.values~t,col="blue")
>  lines(modelo1_P$fitted.values~t,col="red")
>  lines(modelo2_P$fitted.values~t,col="yellow")
>  dev.off()
null device
          1
>
>
>  ##########################################################################
######
>  #       Teste de Shapiro-Wilk
>  ##########################################################################
######
>
>  shapiro.test(modelo3_P$residuals) # OK

        Shapiro-Wilk normality test

data:  modelo3_P$residuals
W = 0.94643, p-value = 0.651

>
>  ##########################################################################
######
>  #       Mudança para o modelo gaussiano parabolico e testes
>  ##########################################################################
######
>
>  modelo3_P_par <- glm(p_prop~I(t^2)+t, family=gaussian(link = "ident
ity"))
>  dwtest(modelo3_P_par)

        Durbin-Watson test

data:  modelo3_P_par
DW = 1.202, p-value = 0.002812
alternative hypothesis: true autocorrelation is greater than 0

>  shapiro.test(modelo3_P_par$residuals)
```

```
        Shapiro-Wilk normality test

data:  modelo3_P_par$residuals
W = 0.93928, p-value = 0.5743

>
>  ##########################################################################
######
>  #       Graficos
>  ##########################################################################
######
>
>  #X11()
>  jpeg(file = "C:/Users/Rodolfo.RODOLFO-PC/Desktop/Pessoal/rodolfoTCC
1/Imagens curitiba/figura5.jpeg") # FIGURA 5
>  par(mfrow=c(2,2))
>  plot(modelo3_P_par)
>  dev.off()
null device
          1
>
>  ##########################################################################
######
>  #       Graficos
>  ##########################################################################
######
>
>  #X11()
>  jpeg(file = "C:/Users/Rodolfo.RODOLFO-PC/Desktop/Pessoal/rodolfoTCC
1/Imagens curitiba/figura6.jpeg") # FIGURA 6
>  plot(dadosP,type="p",ylim=c(0,0.25),pch=10,xlab="Tempo",ylab="Popul
ação (proporcional)",col="red")
>  lines(modelo1_P$fitted.values~t,lty=1,lwd=1)
>  lines(modelo2_P$fitted.values~t,lty=2,lwd=2)
>  lines(modelo3_P$fitted.values~t,lty=3,lwd=2)
>  lines(modelo3_P_par$fitted.values~t,lty=4,lwd=2)
>  dev.off()
null device
          1
>
>  ##########################################################################
######
>  #       Critério de informação Akaike
>  ##########################################################################
######
>
>  AIC(modelo1_P,modelo3_P,modelo2_P,modelo3_P_par)
             df        AIC
modelo1_P     2   6.216151
modelo3_P     3 -35.069237
modelo2_P     3 -27.566021
modelo3_P_par  4 -45.556356
>
>  ##########################################################################
######
>  #       Tabela de comparações entre valores ajustados e valores rea
is
>  ##########################################################################
######
>
```

```
>  valores_par<-cbind(p_prop,modelo1_P$fitted.values,modelo3_P$fitted.
values,modelo3_P_par$fitted.values);
>  valores
      p_prop
1 0.01118188 0.01278990 0.03194286 -0.03617530
2 0.01995754 0.02710020 0.04013184  0.02394160
3 0.02561925 0.03924119 0.04603235  0.05400005
4 0.05109696 0.05650549 0.05396702  0.08405851
5 0.08832272 0.08072707 0.06520683  0.11411696
6 0.14508139 0.11407608 0.08236011  0.14417541
7 0.18612880 0.16400201 0.11589659  0.17723970
8 0.22462845 0.21682285 0.17379919  0.20429231
9 0.24798301 0.28873523 0.39066321  0.23435076
>
>  ##################################################################
##
>  #                          DIFERENCAS FINITAS
>  ##################################################################
##
>
>  t<-tt
>  p<-pp
>
> g<-(diff(p,lag=1)/diff(t,lag=1))/p[1:length(p)-1]
> h<-(diff(p,lag=1)/diff(t,lag=1))/p[2:length(p)]
> p2<-p[3:length(p)-1]
>
> dados<-((h[1:length(h)-1]+g[2:length(g)])/2)
> ##################################################################
######################
> t<-p2
# Tempo T
> p<-dados
> dados<-data.frame(t,p)
> ##################################################################
######################
>
> ##################################################################
#####
> #       Graficos
> ##################################################################
#####
>
> #X11()
> jpeg(file = "C:/Users/Rodolfo.RODOLFO-PC/Desktop/Pessoal/rodolfoTCC1
/Imagens curitiba/figura7.jpeg") # FIGURA 7
> plot(dadosP,type="p",ylim=c(0,0.25),pch=10,xlab="Tempo",ylab="Popula
ção (proporcional)",col="red")
> dev.off()
null device
          1
>
> ##################################################################
#####
> #       Teste de Goldfeld-Quandt
> ##################################################################
#####
>
> gqtest(p~t)

        Goldfeld-Quandt test
```

```
data:  p ~ t
GQ = 0.019888, df1 = 2, df2 = 1, p-value = 0.9807
alternative hypothesis: variance increases from segment 1 to 2

>
> ##########################################################################
#####
> #       Mudança de escala
> ##########################################################################
#####
>
> p_prop<-prop.table(p)
> dadosP<-data.frame(t,p_prop)
>
> ##########################################################################
#####
> #       Modelos
> ##########################################################################
#####
>
> modeloL_P<-lm(p_prop~t)
> modelo1_P <- glm(p_prop~t, family=binomial(link = "logit"))
> modelo2_P <- glm(p_prop~t, family=Gamma(link = "inverse"))
> modelo3_P <- glm(p_prop~t, family=gaussian(link = "identity"))
> modelo4_P <- glm(p_prop~t, family=poisson(link = "log"))
> modelo5_P <- glm(p_prop~t, family=quasi(link = "identity", variance
= "constant"))
> modelo6_P <- glm(p_prop~t, family=quasibinomial(link = "logit"))
> modelo7_P <- glm(p_prop~t, family=quasipoisson(link = "log"))
>
> ##########################################################################
#####
> #       Critério de informação Akaike
> ##########################################################################
#####
>
> AIC(modeloL_P,modelo1_P,modelo2_P,modelo3_P,modelo4_P,modelo5_P,mode
lo6_P,modelo7_P)
           df        AIC
modeloL_P   3 -16.463226
modelo1_P   2   6.179023
modelo2_P   3 -17.027603
modelo3_P   3 -16.463226
modelo4_P   2        Inf
modelo5_P   2         NA
modelo6_P   2         NA
modelo7_P   2         NA
>
> ##########################################################################
#####
> #       Teste de Durbin-Watson
> ##########################################################################
#####
>
> dwtest(modeloL_P)

        Durbin-Watson test

data:  modeloL_P
DW = 1.2554, p-value = 0.03481
alternative hypothesis: true autocorrelation is greater than 0
```

```
>
> ##########################################################################
#####
> #       Graficos
> ##########################################################################
#####
>
> #X11()
> jpeg(file = "C:/Users/Rodolfo.RODOLFO-PC/Desktop/Pessoal/rodolfoTCC1
/Imagens curitiba/figura8.jpeg") # FIGURA 8
> par(mfrow=c(3,1))
> hnp(modelo1_P)
Binomial model
> hnp(modelo3_P)
Gaussian model (glm object)
> hnp(modeloL_P)
Gaussian model (lm object)
> dev.off()
null device
          1
>
> ##########################################################################
#####
> #       Tabela de comparações entre valores ajustados e valores reai
s
> ##########################################################################
#####
>
> valores<-matrix(cbind(p_prop,modelo1_P$fitted.values,modelo3_P$fitte
d.values,modeloL_P$fitted.values),nc=4)
> colnames(valores)<-c("prop","logit","gaussian-glm","gaussian-lm")
> valores<-data.frame(valores)
> valores
        prop      logit gaussian.glm gaussian.lm
1 0.09353669 0.20126808   0.19638157  0.19638157
2 0.22578072 0.19630984   0.19285277  0.19285277
3 0.22794983 0.17514208   0.17697314  0.17697314
4 0.19766339 0.14750484   0.15377124  0.15377124
5 0.12044894 0.11240523   0.11839495  0.11839495
6 0.07993278 0.09177837   0.09281111  0.09281111
7 0.05468766 0.07559155   0.06881522  0.06881522
>
> ##########################################################################
#####
> #       Graficos
> ##########################################################################
#####
>
> #X11()
> jpeg(file = "C:/Users/Rodolfo.RODOLFO-PC/Desktop/Pessoal/rodolfoTCC1
/Imagens curitiba/figura9.jpeg") # FIGURA 9
> plot(dadosP,type="p",ylim=c(0.05,0.25),pch=10,xlab="População",ylab=
"Diferenças (proporcional)",col="red")
> lines(modeloL_P$fitted.values~t,lty=1,lwd=1)
> lines(modelo1_P$fitted.values~t,lty=2,lwd=2)
> lines(modelo2_P$fitted.values~t,lty=3,lwd=2)
> lines(modelo3_P$fitted.values~t,lty=4,lwd=2)
> dev.off()
null device
          1
>
```

```
> ##########################################################################
#####
> #       Teste de Shapiro-Wilk para o modelo gaussiano
> ##########################################################################
#####
>
> shapiro.test(modelo3_P$residuals) # OK

        Shapiro-Wilk normality test

data:  modelo3_P$residuals
W = 0.86822, p-value = 0.1791

>
> ##########################################################################
#####
> #       Mudança para o modelo linear parabolico
> ##########################################################################
#####
>
> modeloL_par <- lm(p_prop~I(t^2)+t)
> summary(modeloL_par)

Call:
lm(formula = p_prop ~ I(t^2) + t)

Residuals:
       1        2        3        4        5        6        7
-0.07905  0.05003  0.04289  0.01353 -0.02906 -0.01954  0.02120

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.582e-01  5.721e-02   2.766   0.0505 .
I(t^2)      -1.247e-07  1.086e-07  -1.149   0.3148
t            1.193e-04  1.850e-04   0.645   0.5542
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05579 on 4 degrees of freedom
Multiple R-squared:  0.6083,   Adjusted R-squared:  0.4125
F-statistic: 3.107 on 2 and 4 DF,  p-value: 0.1534

>
> ##########################################################################
#####
> #       Mudança para o modelo linear de quarta ordem
> ##########################################################################
#####
>
> modeloL_quar <- lm(p_prop~I(t^4)+I(t^3)+I(t^2)+t)
> summary(modeloL_quar)

Call:
lm(formula = p_prop ~ I(t^4) + I(t^3) + I(t^2) + t)

Residuals:
        1         2         3         4         5         6         7
-0.042491  0.055891 -0.012337 -0.006988  0.012881 -0.009285  0.002329

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5.693e-02  1.659e-01  -0.343    0.764
```

```
I(t^4)      -6.866e-13  8.184e-13  -0.839    0.490
I(t^3)       2.723e-09  2.829e-09   0.962    0.437
I(t^2)      -3.707e-06  3.237e-06  -1.145    0.371
t            1.839e-03  1.374e-03   1.338    0.313

Residual standard error: 0.0519 on 2 degrees of freedom
Multiple R-squared:  0.8305,   Adjusted R-squared:  0.4915
F-statistic:  2.45 on 4 and 2 DF,  p-value: 0.3103

>
> ##########################################################################
#####
> #       Mudança para o modelo linear de quinta ordem
> ##########################################################################
#####
>
> modeloL_quin <- lm(p_prop~I(t^5)+I(t^4)+I(t^3)+I(t^2)+t)
> summary(modeloL_quin)

Call:
lm(formula = p_prop ~ I(t^5) + I(t^4) + I(t^3) + I(t^2) + t)

Residuals:
        1         2         3         4         5         6         7
-0.031756  0.047192 -0.025043  0.013489 -0.006256  0.002926 -0.000552

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.522e-01  4.029e-01  -0.626    0.644
I(t^5)       1.664e-15  2.958e-15   0.563    0.674
I(t^4)      -7.815e-12  1.271e-11  -0.615    0.649
I(t^3)       1.381e-08  2.002e-08   0.690    0.615
I(t^2)      -1.130e-05  1.407e-05  -0.803    0.569
t            4.017e-03  4.226e-03   0.951    0.516

Residual standard error: 0.06397 on 1 degrees of freedom
Multiple R-squared:  0.8713,   Adjusted R-squared:  0.2275
F-statistic: 1.353 on 5 and 1 DF,  p-value: 0.5707

>
> ##########################################################################
#####
> #       Teste de Shapiro-Wilk
> ##########################################################################
#####
>
> shapiro.test(modeloL_par$residuals)

        Shapiro-Wilk normality test

data:  modeloL_par$residuals
W = 0.93652, p-value = 0.6076

>
> shapiro.test(modeloL_quar$residuals)

        Shapiro-Wilk normality test

data:  modeloL_quar$residuals
W = 0.92466, p-value = 0.5065

>
```

```
> shapiro.test(modeloL_quin$residuals)

        Shapiro-Wilk normality test

data:  modeloL_quin$residuals
W = 0.94425, p-value = 0.6772

>
> ##########################################################################
#####
> #       Graficos
> ##########################################################################
#####
>
> #X11()
> jpeg(file = "C:/Users/Rodolfo.RODOLFO-PC/Desktop/Pessoal/rodolfoTCC1
/Imagens curitiba/figura14.jpeg") # FIGURA 14
> plot(dadosP,type="p",ylim=c(0.05,0.25),pch=10,xlab="População",ylab=
"Diferenças (proporcional)",col="red")
> lines(modeloL_par$fitted.values~t,lty=1,lwd=1)
> lines(modeloL_quar$fitted.values~t,lty=2,lwd=2)
> lines(modeloL_quin$fitted.values~t,lty=3,lwd=2)
> dev.off()
null device
          1
>
> ##########################################################################
#####
> #       Critério de informação Akaike
> ##########################################################################
#####
>
> AIC(modeloL_P,modelo1_P,modelo3_P,modeloL_par,modeloL_quar,modeloL_q
uin)
             df        AIC
modeloL_P     3 -16.463226
modelo1_P     2   6.179023
modelo3_P     3 -16.463226
modeloL_par   4 -16.458417
modeloL_quar  6 -18.321437
modeloL_quin  7 -18.246293
>
> ##########################################################################
#####
> #       Teste de Durbin-Watson
> ##########################################################################
#####
>
> dwtest(modeloL_par)

        Durbin-Watson test

data:  modeloL_par
DW = 1.6978, p-value = 0.02652
alternative hypothesis: true autocorrelation is greater than 0

> dwtest(modeloL_quar)

        Durbin-Watson test

data:  modeloL_quar
DW = 2.8552, p-value = 0.1661
```

```
alternative hypothesis: true autocorrelation is greater than 0

> dwtest(modeloL_quin)

        Durbin-Watson test

data:  modeloL_quin
DW = 3.2795, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0

>
> ##################################################################
#####
> #       Tabela de comparações entre valores ajustados e valores reai
s
> ##################################################################
#####
>
> valores_par<-cbind(p_prop,modelo1_P$fitted.values,modelo3_P$fitted.v
alues,modeloL_par$fitted.values,modeloL_quar$fitted.values,modeloL_qui
n$fitted.values);
> valores
        prop      logit gaussian.glm gaussian.lm
1 0.09353669 0.20126808   0.19638157  0.19638157
2 0.22578072 0.19630984   0.19285277  0.19285277
3 0.22794983 0.17514208   0.17697314  0.17697314
4 0.19766339 0.14750484   0.15377124  0.15377124
5 0.12044894 0.11240523   0.11839495  0.11839495
6 0.07993278 0.09177837   0.09281111  0.09281111
7 0.05468766 0.07559155   0.06881522  0.06881522
>
> valores<-data.frame(valores)
> valores
        prop      logit gaussian.glm gaussian.lm
1 0.09353669 0.20126808   0.19638157  0.19638157
2 0.22578072 0.19630984   0.19285277  0.19285277
3 0.22794983 0.17514208   0.17697314  0.17697314
4 0.19766339 0.14750484   0.15377124  0.15377124
5 0.12044894 0.11240523   0.11839495  0.11839495
6 0.07993278 0.09177837   0.09281111  0.09281111
7 0.05468766 0.07559155   0.06881522  0.06881522
```