Universidade de São Paulo
Escola Superior de Agricultura "Luiz de Queiroz"

Novos modelos de regressão e algoritmos de aprendizado de máquina: teoria e aplicações

Gabriela Maria Rodrigues

Tese apresentada para obtenção do título de Doutora em Ciências.

Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2023

Gabriela Maria Rodrigues
Licenciada em Matemática

Novos modelos de regressão e algoritmos de aprendizado de máquina: teoria e aplicações

Orientador:
Prof. Dr. **EDWIN MOISES MARCOS ORTEGA**

Tese apresentada para obtenção do título de Doutora em Ciências.
Área de concentração: Estatística e Experimentação Agronômica

Piracicaba
2023

SUMÁRIO

RESUMO

Novos modelos de regressão e algoritmos de aprendizado de máquina: teoria e aplicações

Neste trabalho são definidos novos modelos de regressão, baseados na família de distribuições exponentiated odd log-logistic (EOLL-G). Esta família possui a flexibilidade de modelar dados bimodais, simétricos ou assimétricos. Utilizando a distribuição Normal como base, são propostos um modelo de regressão quantílica e um modelo de regressão parcialmente linear. Duas novas famílias bivariadas são definidas a partir da família EOLL-G e utilizando as cópulas de Clayton e de Frank. Dois modelos para dados censurados são propostos utilizando como base as distribuições Weibull e generalized Rayleigh. O desempenho preditivo do modelo parcialmente linear e de um dos modelos para dados censurados é comparado com algoritmos de aprendizado de máquinas: árvores de decisão, florestas aleatórias e florestas aleatórias de sobrevivência. Propriedades estruturais das novas distribuições foram fornecidas, que exibem a flexibilidade da família utilizada e podem ser úteis para trabalhos futuros. O método de máxima verossimilhança foi utilizado para estimação dos parâmetros e estudos de simulações para ambos os modelos são realizados, comprovando a consistência das estimativas. Diversas aplicações são realizadas ilustrando a utilidade dos novos modelos. Quanto à capacidade preditiva, eles mostraram-se competitivos aos algoritmos de aprendizado de máquina, de acordo com os estudos de simulações e com as aplicações realizadas.

Palavras-chave: Regressão quantílica, Dados censurados, Árvores de decisão, Florestas aleatórias, Florestas aleatórias de sobrevivência, Validação cruzada k-fold

ABSTRACT

New regression models and machine learning algorithms: theory and applications

In this work, new regression models are defined, based on exponentiated odd log-logistic-G (EOLL-G) family of distributions. This family has the flexibility to model bimodal, symmetric or asymmetric data. Using the Normal distribution as a basis, a quantile regression model and a partially linear regression model are proposed. Two new bivariate families are defined based on the EOLL-G family and using the Clayton and Frank copulas. Two models for censored data are proposed using the Weibull and generalized Rayleigh distributions as a basis. The predictive performance of the partially linear model and one of the models for censored data is compared with machine learning algorithms: decision trees, random forests and random survival forests. Structural properties of the new distributions were provided, which exhibit the flexibility of the family used and may be useful for future work. The maximum likelihood method was used to estimate the parameters and simulation studies for both models were carried out, proving the consistency of the estimates. Several applications are carried out illustrating the usefulness of the new models. As for predictive capacity, they proved to be competitive with machine learning algorithms, according to simulation studies and the applications carried out.

Keywords: Quantile regression, Censored data, Decision trees, Random forests, Random survival forests, Cross-validation k-fold

1 INTRODUÇÃO

A análise de regressão pode ser descrita como uma técnica para investigar e modelar o relacionamento entre duas ou mais variáveis, amplamente utilizada em diversas áreas de conhecimento como: medicina, agricultura, economia, administração, engenharia, etc. Os dados provenientes de diferentes campos de conhecimento, possuem uma variedade de características, justificando e motivando a exploração de novas metodologias que capturem de forma adequada tais particularidades.

Nos últimos anos, também decorrente do crescente volume de dados gerados, um termo que está em grande ascensão é o *Machine Learning* (ML) (Aprendizado de Máquina), uma subárea da inteligência artificial. Seu conceito consiste em alimentar as máquinas com dados para que elas aprendam com eles e sejam capazes de prever novos resultados quando novos valores forem apresentados. Este trabalho se concentra em algoritmos supervisionados de ML para problemas de regressão, ou também chamados de problemas de predição. Nesta classe, as variáveis preditoras são utilizadas para prever uma ou mais variáveis respostas quantitativas.

A análise de regressão usual, além de prever, em geral, tem como objetivo principal a inferência estatística, que envolve a estimação de parâmetros e a inferência sobre eles, como testes de hipótese e intervalos de confiança, supondo que, os dados disponíveis são provenientes de uma população. O estatístico Breiman (2003) foi muito reconhecido e prestigiado na comunidade estatística por se dedicar a trabalhos que unem a estatística e o ML, mostrando que a união entre estas duas áreas é mutuamente benéfica, uma vez que, pode-se aproveitar do melhor de cada uma delas. Efron e Hastie (2022) ressaltam que as duas características, algorítmica e inferencial, podem ser encontradas também na regressão usual.

Neste sentido, este trabalho tem como objetivo propor novos modelos de regressão para acomodar diferentes tipos de dados de diversas áreas, utilizando alguns conceitos de ML sob uma perspectiva estatística. Os principais problemas tratados nesta tese são descritos a seguir.

Em muitas situações práticas, a variável resposta apresenta assimetria e/ou multimodalidade. Nestes casos, a distribuição Normal, que é a base da teoria clássica dos modelos de regressão, pode não ser adequada. Os modelos de regressão propostos neste trabalho utilizam novas distribuições de probabilidade, baseadas na família *exponentiated odd log-logistic* (EOLL-G) (Alizadeh et al., 2020). Esta família adiciona dois parâmetros de forma extra a uma distribuição de base contínua, tornando-a muito mais flexível e possibilitando a modelagem de dados com forma bimodal, simétrica ou assimétrica, positiva ou negativa. Considerando como distribuições de base a Normal, Weibull e *generalized Rayleigh* (GR), os novos modelos são denominados de exponentiated odd log-logistic Normal (EOLLN), exponentiated odd log-logistic Weibull (EOLLW) e exponentiated odd

log-logistic generalized Rayleigh (EOLLGR).

Um novo modelo de regressão quantílica (RQ) (Koenker e Bassett Jr, 1978) é proposto, baseado em uma reparametrização da distribuição EOLLN em termos de seus quantis. Embora uma vasta literatura seja encontrada a respeito desta metodologia, verifica-se uma escassez de modelos para dados bimodais. A RQ pode ser uma alternativa para lidar com problemas como assimetria e heterogeneidade de variâncias da variável resposta, também sendo robusta à presença de pontos atípicos (outliers). Seus modelos podem caracterizar toda a distribuição da variável resposta por meio da análise de diferentes quantis, fornecendo uma avaliação mais completa sobre os efeitos das covariáveis.

Diversas áreas de conhecimento possuem duas ou mais variáveis resposta de interesse. Se estas variáveis não forem independentes e se existir uma explicação prática para isto, modelos estatísticos multivariados devem ser utilizados com o objetivo de explicar e capturar a correlação entre elas. As distribuições mais comumente utilizadas são a Normal bivariada e a T-student bivariada, que podem ser muito restritivas. Adotar estas distribuições implica em assumir uma relação linear e uma estrutura elíptica entre as variáveis, além disso, assume-se que as distribuições marginais também seguem tais distribuições univariadas, o que nem sempre é satisfeito. Neste sentido, duas novas famílias bivariadas baseadas nas cópulas arquimedianas de Clayton (Clayton, 1978) e de Frank (Frank, 1979) e na família de distribuições exponentiated odd log-logistic são definidas neste trabalho. Desta forma, são denominados os modelos bivariados de Clayton EOLL-G (BCEOLL-G) e os modelos bivariados de Frank EOLL-G (BFEOLL-G). Como caso especial destas famílias, utiliza-se a distribuição Normal e obtém-se os modelos bivariados BCEOLL-Normal e BFEOLL-Normal. A metodologia de cópulas possibilita uma modelagem de associações de forma muito flexível, uma vez que, constrói uma distribuição conjunta de duas ou mais variáveis aleatórias permitindo que cada uma delas seja individualmente modelada por uma distribuição marginal diferente (Nelsen, 1986). Além disso, as cópulas podem modelar variáveis que apresentem diversas estruturas de dependência, inclusive não-lineares. Desta forma, ao utilizar a família EOLL-G, as distribuições marginais do modelo bivariado serão muito flexíveis, permitindo a modelagem de suas particularidades.

As variáveis em estudo podem apresentar uma relação linear e ou não linear. A avaliação correta desta relação é tão importante quanto a escolha de uma distribuição adequada para a variável resposta, uma vez que, a suposição incorreta da linearidade pode comprometer a confiabilidade dos testes de hipóteses e levar a modelos mal especificados. Um modelo de regressão parcialmente linear, também chamado de semiparamétrico, é apresentado com base na distribuição EOLLN. Este tipo de modelo, pode ser uma alternativa interessante para estudar relações não lineares, pois permite estudar esta relação de forma mais flexível, por meio de funções não paramétricas. Além disso, tais modelos permitem investigar efeitos lineares e não lineares das covariáveis na variável resposta simultaneamente. O desempenho predito do novo modelo proposto é comparado com dois

algoritmos supervisionados de machine learning: *decision trees* (DTs) (Breiman, 1984) e *random forests* (RFs) (Breiman, 2001). Estes métodos são técnicas não lineares e não paramétricas, portanto, são livres de suposições e não requerem conhecimento prévio da forma funcional da relação entre a variável resposta e as covariáveis. A comparação é realizada por meio de um estudo de simulação e uma aplicação a dados reais.

A análise de sobrevivência é uma área da estatística que tem como variável resposta o tempo até a ocorrência de um evento de interesse. A principal característica deste tipo de dados é a presença de censura, ou seja, quando o evento de interesse não ocorreu e a variável resposta é parcialmente observada. Uma função importante neste tipo de análise, é a função de risco, que é muito útil para descrever a distribuição do tempo de ocorrência do evento das observações em estudo. É bem conhecido que, tal função pode assumir diferentes formas, o que desencadeou um grande número de novas distribuições, com a finalidade de obter maior flexibilidade na modelagem. Neste contexto, dois novos modelos são propostos. O primeiro é um novo modelo de regressão quantílica baseado em uma reparametrização da distribuição EOLLW em termos de seus quantis. A união da RQ e da modelagem paramétrica de dados censurados é bastante escassa na literatura, especialmente considerando distribuições bimodais. Neste contexto, a RQ pode fornecer uma interpretação direta entre a sobrevivência e as covariáveis de interesse, permitindo a identificação dos seus diferentes efeitos, além de ser flexível quanto a suposição de riscos proporcionais. O segundo modelo proposto para dados censurados é baseado na distribuição EOLLGR. O desempenho preditivo deste modelo é comparado com o algoritmo de florestas aleatórias de sobrevivência por meio de estudos de simulações e de uma aplicação a dados reais.

Desta forma, os objetivos do presente trabalho podem ser sumarizados como:

- Propor novas distribuições flexíveis para análise de dados univariados ou bivariados, de modo que, acomodem diferentes formas, como bimodalidade e/ou assimetria.
- Baseado nas novas distribuições, propor modelos de regressão flexíveis para dados censurados e não censurados, considerando conjuntos de dados que apresentem heterogeneidade de variâncias, relações não lineares e diferentes funções de risco.
- Comparar o desempenho preditivo dos novos modelos com algoritmos supervisionados de aprendizado de máquinas.

Organização do trabalho

Esta tese está organizada conforme se segue. Capítulo 2: o modelo de regressão quantílica baseado em uma reparametrização da distribuição EOLLN é apresentado. Algumas propriedades estruturais são fornecidas, um estudo de simulação avalia as estimativas de máxima verossimilhança e três aplicações a dados reais são realizadas. Capítulo 3: as duas famílias bivariadas baseadas nas cópulas de Clayton e de Frank e na família

EOLL-G são definidas. Um estudo de simulação avalia a consistência de seus estimadores e uma aplicação ilustra a sua utilidade. Capítulo 4: o modelo de regressão parcialmente linear baseado na distribuição EOLLN é proposto e seu desempenho preditivo é comparado com dois algoritmos de machine learning, por meio de um estudo de simulação e uma aplicação a dados reais. Capítulo 5: o modelo de regressão quantílica para dados censurados baseado na distribuição EOLLW é apresentado. Um estudo de simulação avalia a consistência de seus estimadores para diferentes porcentagens de censura e uma aplicação ilustra a sua utilidade. Capítulo 6: o segundo modelo para dados censurados é proposto, com base na distribuição EOLLGR. O desempenho preditivo deste modelo é comparado com o algoritmo de florestas aleatórias de sobrevivência por meio de estudos de simulações e de uma aplicação a dados reais. Capítulo 7: Algumas considerações finais e perspectivas de pesquisas futuras são apresentadas.

Referências

Alizadeh, M., Tahmasebi, S., e Haghbin, H. (2020). The exponentiated odd log-logistic family of distributions: Properties and applications. *Journal of Statistical Modelling: Theory and Applications*, 1(1):29–52.

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Breiman, L. (2003). Statistical modeling: The two cultures. *Quality control and applied statistics*, 48(1):81–82.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*, 65(1):141–151.

Efron, B. e Hastie, T. (2022). *Computer age statistical inference*. Cambridge University Press 2016, vol. 5, New York.

Frank, M. J. (1979). On the simultaneous associativity of f (x, y) and x+ y- f (x, y). *Aequationes mathematicae*, 19:194–226.

Koenker, R. e Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Nelsen, R. B. (1986). Properties of a one-parameter family of bivariate distributions with specified marginals. *Communications in statistics-Theory and methods*, 15(11):3277–3285.

2 UM NOVO MODELO DE REGRESSÃO QUANTÍLICA: PROPRIEDADES E APLICAÇÕES

2.1 Introdução

A teoria clássica dos modelos de regressão é baseada na regressão na média da distribuição da variável dependente ou variável resposta, em função de uma ou mais covariáveis. Entretanto, em muitas situações práticas, a variável resposta pode apresentar por exemplo, assimetria, multimodalidade ou variâncias heterogêneas, tornando a média uma medida inadequada para explicá-la e comprometendo as propriedades de tais estimadores. Em comparação com a regressão média convencional, a regressão quantílica (RQ), introduzida por Koenker e Bassett Jr (1978), pode caracterizar toda a distribuição condicional da variável resposta, fornecendo uma avaliação mais completa dos efeitos da covariável, por meio da análise de diferentes quantis. Essa regressão não impõe nenhuma suposição distributiva sobre o erro, exceto o requisito sobre o quantil condicional zero, sendo eficiente principalmente para modelar dados com heterocedasticidade. Nesse sentido, algumas vantagens podem ser mencionadas:

- A caracterização da variável resposta por meio dos quantis pode identificar diferentes efeitos da variável explanatória, fornecendo mais informações e identificando relações implícitas sob elas;
- Pode fornecer informações mais completas sob a resposta por meio da identificação e inferência de efeitos heterogêneos das covariáveis em diferentes quantis. Desta forma, permite o controle de forma flexível e eficiente da heterogeneidade devida a elas;
- Permite uma interpretação direta e simples dos resultados;
- São robustos a presença de pontos atípicos (outliers) na variável resposta.

Diante do exposto, diversas áreas de conhecimento tem considerado a utilização da regressão quantílica. Por exemplo, na área de medicina, Beyerlein et al. (2011) utilizaram em análises GWAS (Genome Wide Association Study) e enfatizaram as vantagens estatísticas e biológicas ao se estimar efeitos de marcadores em diferentes quantis das distribuições dos fenótipos; Rodrigues et al. (2023) analisou os tempos de vida de pacientes com câncer gástrico ao longo de diferentes quantis; Puiatti et al. (2018) estudaram o acúmulo de matéria seca em plantas de alho ao longo do tempo; Santos et al. (2018), propuseram a utilização da RQ para estimação de valores genéticos genômicos de suínos, que possuem fenótipos com distribuições assimétricas e Nascimento et al. (2019) identificaram suínos com diferentes taxas de crescimento por meio da RQ.

Os modelos de regressão quantílica são comumente baseados na distribuição de Laplace assimétrica (LA) (Yuan e Yin, 2010; Reich et al., 2010; Geraci e Bottai, 2014; Galarza et al., 2020). Esta distribuição possui a propriedade do quantil igual a zero e uma representação estocástica útil, mas pode apresentar problemas de instabilidade numérica, uma vez que, não é diferenciável em zero. Neste sentido, outras distribuições foram consideradas, por exemplo: log-extended exponential-geometric (Jodra e Jimenez-Gamero, 2020); Birnbaum-Saunders (Sánchez et al., 2020) e transmuted unit-Rayleigh (Korkmaz et al., 2021). No entanto, nenhum desses modelos pode modelar dados bimodais.

Neste contexto, este trabalho define uma nova reparametrização da distribuição exponentiated odd log-logistic normal (EOLLN) (Alizadeh et al., 2020) em termos de seus quantis. Esta distribuição tem grande flexibilidade para dados bimodais, assimétricos à direita e assimétricos à esquerda, podendo ser uma alternativa interessante aos modelos mistura. Várias de suas propriedades matemáticas são apresentadas e o modelo de regressão quantílica com dois componentes sistemáticos é definido.

2.2 Considerações finais

Um novo modelo de regressão quantílica é definido baseado na reparametrização da distribuição exponentiated odd log-logistic normal (EOLLN), a partir da qual são obtidas algumas propriedades estruturais. Esses resultados matemáticos ilustram a flexibilidade da distribuição para capturar diferentes formas. O método de máxima verossimilhança foi utilizado para estimar os parâmetros. Diversas simulações mostram que as estimativas são consistentes e que os resíduos quantílicos deste modelo convergem para uma distribuição normal padrão. A utilidade do modelo de regressão proposto é demonstrada por meio de três aplicações reais. Algumas considerações importantes podem ser úteis e motivar pesquisadores em trabalhos futuros: (i) na aplicação 1, o modelo EOLLN foi capaz de obter efeitos na variabilidade dos dados, que não foram observados em outros modelos; (ii) os modelos foram eficazes em fornecer informações sobre a relação entre a variável resposta e covariáveis contínuas ou discretas e (iii) os modelos forneceram informações sobre as caudas inferiores e superiores da distribuição da variável resposta. Assim, o modelo EOLLN e seu submodelo Exp-N mostraram-se adequados para estes conjuntos de dados, podendo ser considerados alternativas interessantes para dados assimétricos e heterogêneos. Além disso, a regressão quantílica forneceu percepções interessantes e importantes sobre as variáveis respostas em estudo. Em trabalhos futuros, este modelo pode ser estendido, por exemplo, com efeitos aleatórios, funções de suavização não paramétricas ou termos aditivos.

Referências

Alizadeh, M., Tahmasebi, S., e Haghbin, H. (2020). The exponentiated odd log-logistic family of distributions: Properties and applications. *Journal of Statistical Modelling: Theory and Applications*, 1(1):29–52.

Beyerlein, A., von Kries, R., Ness, A. R., e Ong, K. K. (2011). Genetic markers of obesity risk: stronger associations with body composition in overweight compared to normal-weight children. *Plos one*, 6(4):e19057.

Galarza, C. E., Castro, L. M., Louzada, F., e Lachos, V. H. (2020). Quantile regression for nonlinear mixed effects models: a likelihood based perspective. *Statistical Papers*, 61(3):1281–1307.

Geraci, M. e Bottai, M. (2014). Linear quantile mixed models. *Statistics and computing*, 24:461–479.

Jodra, P. e Jimenez-Gamero, M. D. (2020). A quantile regression model for bounded responses based on the exponential-geometric distribution. *REVSTAT-Statistical Journal*, 18(4):415–436.

Koenker, R. e Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Korkmaz, M. Ç., Chesneau, C., e Korkmaz, Z. S. (2021). Transmuted unit rayleigh quantile regression model: Alternative to beta and kumaraswamy quantile regression models. *Univ. Politeh. Buchar. Sci. Bull. Ser. Appl. Math. Phys*, 83:149–158.

Nascimento, M., Nascimento, A., Dekkers, J., e Serão, N. (2019). Using quantile regression methodology to evaluate changes in the shape of growth curves in pigs selected for increased feed efficiency based on residual feed intake. *animal*, 13(5):1009–1019.

Puiatti, G. A., Cecon, P. R., Nascimento, M., Nascimento, A. C. C., Carneiro, A. P. S., Silva, F. F., Puiatti, M., e Oliveira, A. C. R. d. (2018). Quantile regression of nonlinear models to describe different levels of dry matter accumulation in garlic plants. *Ciência Rural*, 48.

Reich, B. J., Bondell, H. D., e Wang, H. J. (2010). Flexible bayesian quantile regression for independent and clustered data. *Biostatistics*, 11(2):337–352.

Rodrigues, G. M., Ortega, E. M., Cordeiro, G. M., e Vila, R. (2023). Quantile regression with a new exponentiated odd log-logistic weibull distribution. *Mathematics*, 11(6):1518.

Sánchez, L., Leiva, V., Galea, M., e Saulo, H. (2020). Birnbaum-saunders quantile regression models with application to spatial data. *Mathematics*, 8(6):1000.

Santos, P. M. d., Nascimento, A. C. C., Nascimento, M., Silva, F. F., Azevedo, C. F., Mota, R. R., Guimarães, S. E. F., e Lopes, P. S. (2018). Use of regularized quantile regression to predict the genetic merit of pigs for asymmetric carcass traits. *Pesquisa Agropecuária Brasileira*, 53:1011–1017.

Yuan, Y. e Yin, G. (2010). Bayesian quantile regression for longitudinal studies with nonignorable missing data. *Biometrics*, 66(1):105–114.

3 UMA NOVA FAMÍLIA BIVARIADA BASEADA EM CÓPULAS ARQUIMEDIANAS: SIMULAÇÃO, MODELO DE REGRESSÃO E APLICAÇÃO

3.1 Introdução

Diversas áreas de conhecimento, que utilizam a modelagem estatística, podem possuir uma ou mais variáveis resposta de interesse, ou seja, dois ou mais atributos que devem ser modelados. Se estes atributos não forem independentes e se existir uma explicação prática para isto, modelos estatísticos multivariados devem ser utilizados com o objetivo de explicar e capturar a correlação entre estas variáveis.

As distribuições de probabilidade conjunta são comumente utilizadas. No contexto bivariado por exemplo, as distribuições mais utilizadas são a Normal bivariada e a T-student bivariada, que podem ser muito restritivas. Por exemplo, assumindo tais distribuições conjuntas, assume-se que as distribuições marginais também seguem tais distribuições univariadas. Além disso, adotar estas distribuições implica em uma relação linear e uma estrutura elíptica entre as variáveis. Tais premissas nem sempre são satisfeitas.

A modelagem de dados multivariados baseada em cópulas é uma alternativa interessante para superar estas desvantagens. Segundo Nelsen (1986), cópulas fornecem um meio de relacionar funções de distribuições multivariadas a partir de suas funções de distribuição marginais. As principais vantagens a respeito do interesse estatístico sob esta metodologia são:

- As cópulas permitem que, ao se construir uma distribuição conjunta de duas ou mais variáveis aleatórias, cada uma delas seja individualmente modelada por uma distribuição marginal diferente, portanto, possibilita associações mais flexíveis por meio do ajuste de diferentes distribuições marginais;

- As cópulas representam uma abordagem útil para modelar e entender o fenômeno de dependência, ressaltando sua estrutura. Em outras palavras, a dependência entre essas variáveis pode assumir estruturas diversas, até mesmo não-lineares, de acordo com o tipo de cópula utilizada;

- A escolha das distribuições marginais não depende da estrutura de associação das variáveis estudadas;

- Uma cópula é invariante sob transformações crescentes e contínuas das marginais.

Existem diversas famílias de cópulas, por exemplo: Marshall-Olkin, elípticas, vine, extremais, arquimedianas, Farlie-Gumbel-Morgenstern (FGM), entre outras. Di-

versas distribuições bivariadas foram propostas basedas em cópulas. Weibull bivariadas derivadas de funções de cópula FGM, Ali-Mikhail-Haq (AMH), Gumbel-Hougaard, Gumbel-Barnett, e Nelsen Tem (Flores, 2009). Rayleigh generalizada bivariada usando a cópula de Clayton (El-Sherpieny e Almetwally, 2019). Fréchet bivariada baseada na cópula de cópula FGM e AMH (Almetwally e Muhammed, 2020). Kumaraswamy invertida generalizada usando o método de Marshal–Olkin (Muhammed, 2020). Samanthi e Sepanski Samanthi e Sepanski (2022) propuseram famílias de cópulas Gumbel, Clayton, Frank e Galambos bivariadas com base na distribuição Kumaraswamy. Exponentiated half logistic bivariada baseada em Marshal–Olkin (Alotaibi et al., 2021). Lindley bivariada usando a cópula FGM (Vaidyanathan e Sharon Varghese, 2016).

Este trabalho se concentra nas cópulas arquimedianas, que tem como importante característica expressões com formulas fechadas, isto é, podem ser facilmente construídas a partir de funções geradoras específicas, além disso, são bastante flexíveis, permitindo a modelagem de diversas formas de dependência, incluindo assimetria e dependência extrema nas extremidades. Devido a facilidade em que podem ser construídas, estas cópulas abrangem um grande número de aplicações em diversas áreas. Por exemplo, em finanças (El-Sherpieny e Almetwally, 2019; Naifar, 2011; Yang et al., 2015), saúde (Novianti et al., 2021; Li e Lu, 2019), Hidrologia (Janga Reddy e Ganguli, 2012; Zhang e Singh, 2007; Tsakiris et al., 2016), análise de sobrevivência (He e Lawless, 2005; Wienke et al., 2006; Fachini et al., 2014). Em Nelsen (2006) estudos mais detalhados sob estas famílias e suas aplicações podem ser encontrados.

Em muitas situações práticas, a variável resposta apresenta um comportamento assimétrico e/ou bimodal. A motivação deste trabalho provem de um experimento realizado na Universidade Russa de Economia Plekhanov, com o objetivo de avaliar o crescimento da alface de carvalho (*Lactuca sativa var. crispa*). Os histogramas das variáveis respostas massa fresca (gramas) (y_1) e altura da planta (cm) (y_2) medidas no experimento são relatados nas Figuras 3.1a-b. Observa-se que estas variáveis apresentam comportamento bimodal e assimetrias positivas. Além disso, o gráfico de dispersão entre elas (Figure3.1c) indica que elas apresentam uma forte e positiva correlação.

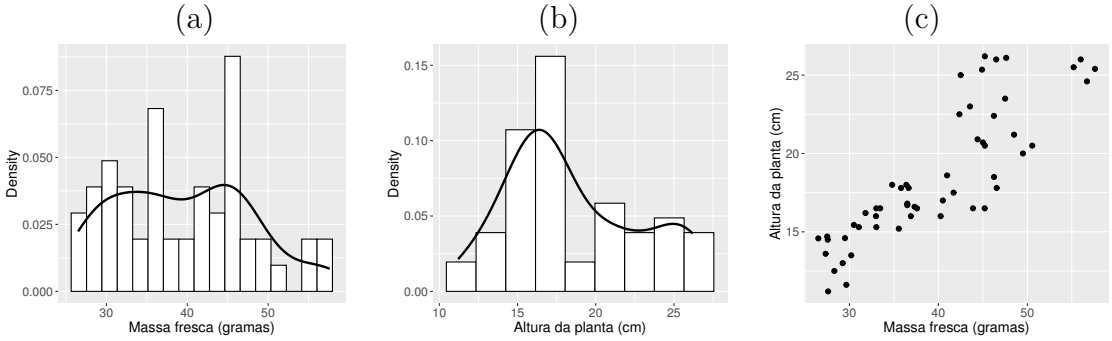

Figura 3.1. (a) Histograma da massa fresca, (b) Histograma de altura da planta e (c) Gráfico de dispersão entre massa fresca e altura da planta.

Embora uma gama de distribuições bivariadas flexíveis possa ser encontrada na literatura, nota-se uma escassez de distribuições bivariadas bimodais. Neste sentido, com o objetivo de fornecer distribuições bivariadas mais flexíveis, este trabalho propõe uma nova família bivariada baseada nas cópulas arquimedianas. Para isto, a família exponentiated odd log-logistic family (EOLL-G) (Alizadeh et al., 2020) é utilizada, cujas densidades possuem grande flexibilidade na modelagem de dados, como bimodalidade e/ou assimetria positiva ou negativa. As cópulas arquimedianas utilizadas foram as de Clayton e de Frank, portanto denomina-se as novas famílias de BCEOLL-G e BFEOLL-G respectivamente. A escolha destas cópulas se deve ao fato de que elas são consideradas adequadas para modelar dados com correlação positiva (veja por exemplo, Naifar (2011)), conforme o conjunto de dados (Figura 3.1c).

3.2 Considerações finais

Este trabalho propôs uma nova família bivariada baseada em cópulas arquimedianas, motivado por um experimento que avaliou o crescimento da alface de carvalho. Este estudo apresentou as variáveis massa fresca e altura da planta com comportamento bimodal, portando, a família exponentiated odd log-logistic foi utilizada, pois suas densidades podem modelar diferentes formas dos dados, incluindo bimodalidade. Além disso, estas variáveis apresentaram uma correlação forte e positiva, por isso as cópulas utilizadas foram as de Clayton e de Frank, que são adequadas para estudar dados com correlação positiva.

Algumas propriedades matemáticas da nova família são apresentadas e um estudo de simulação foi realizado, mostrando a consistência dos estimadores de máxima verossimilhança. O tratamento utilizado no experimento (fator com nove níveis) foi considerado como uma variável explicativa. O modelo de regressão bivariado proposto mostrou-se adequado para predizer os valores de massa fresca e altura de plantas da alface de carvalho sob efeito dos diferentes tratamentos. A escolha dessas cópulas também se mostrou

adequada devido à estrutura de dependência entre as variáveis. Para pesquisas futuras, outras distribuições de base podem ser utilizadas, bem como aplicações em outras áreas de conhecimento.

Referências

Alizadeh, M., Tahmasebi, S., e Haghbin, H. (2020). The exponentiated odd log-logistic family of distributions: Properties and applications. *Journal of Statistical Modelling: Theory and Applications*, 1(1):29–52.

Almetwally, E. M. e Muhammed, H. Z. (2020). On a bivariate fréchet distribution. *J Stat Appl Probab*, 9(1):1–21.

Alotaibi, R. M., Rezk, H. R., Ghosh, I., e Dey, S. (2021). Bivariate exponentiated half logistic distribution: Properties and application. *Communications in Statistics-Theory and Methods*, 50(24):6099–6121.

El-Sherpieny, E. e Almetwally, E. M. (2019). Bivariate generalized rayleigh distribution based on clayton copula. In *Proceedings of the annual conference on statistics (54rd), computer science and operation research, faculty of graduate studies for statistical research, Cairo University*, pages 1–19.

Fachini, J. B., Ortega, E. M., e Cordeiro, G. M. (2014). A bivariate regression model with cure fraction. *Journal of Statistical Computation and Simulation*, 84(7):1580–1595.

Flores, A. Q. (2009). Testing copula functions as a method to derive bivariate weibull distributions. In *American Political Science Association (APSA), Annual Meeting*, pages 1–19. Citeseer.

He, W. e Lawless, J. F. (2005). Bivariate location–scale models for regression analysis, with applications to lifetime data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):63–78.

Janga Reddy, M. e Ganguli, P. (2012). Risk assessment of hydroclimatic variability on groundwater levels in the manjara basin aquifer in india using archimedean copulas. *Journal of Hydrologic Engineering*, 17(12):1345–1357.

Li, H. e Lu, Y. (2019). Modeling cause-of-death mortality using hierarchical archimedean copula. *Scandinavian Actuarial Journal*, 2019(3):247–272.

Muhammed, H. Z. (2020). On a bivariate generalized inverted kumaraswamy distribution. *Physica A: Statistical Mechanics and its Applications*, 553:124281.

Naifar, N. (2011). Modelling dependence structure with archimedean copulas and applications to the itraxx cds index. *Journal of Computational and Applied Mathematics*, 235(8):2459–2466.

Nelsen, R. B. (1986). Properties of a one-parameter family of bivariate distributions with specified marginals. *Communications in statistics-Theory and methods*, 15(11):3277–3285.

Nelsen, R. B. (2006). *An introduction to copulas*. Springer.

Novianti, P., Kartiko, S., e Rosadi, D. (2021). Application of clayton copula to identify dependency structure of covid-19 outbreak and average temperature in jakarta indonesia. In *Journal of Physics: Conference Series*, page 012154. IOP Publishing.

Samanthi, R. G. M. e Sepanski, J. (2022). On bivariate kumaraswamy-distorted copulas. *Communications in Statistics-Theory and Methods*, 51(8):2477–2495.

Tsakiris, G., Kordalis, N., Tigkas, D., Tsakiris, V., e Vangelis, H. (2016). Analysing drought severity and areal extent by 2d archimedean copulas. *Water Resources Management*, 30:5723–5735.

Vaidyanathan, V. e Sharon Varghese, A. (2016). Morgenstern type bivariate lindley distribution. *Statistics, Optimization & Information Computing*, 4(2):132–146.

Wienke, A., Locatelli, I., e Yashin, A. I. (2006). The modelling of a cure fraction in bivariate time-to-event data. *Austrian Journal of Statistics*, 35(1):67–76.

Yang, L., Cai, X. J., Li, M., e Hamori, S. (2015). Modeling dependence structures among international stock markets: Evidence from hierarchical archimedean copulas. *Economic Modelling*, 51:308–314.

Zhang, L. e Singh, V. P. (2007). Bivariate rainfall frequency distributions using archimedean copulas. *Journal of Hydrology*, 332(1-2):93–109.

4 NOVOS MODELOS DE REGRESSÃO PARCIALMENTE LINEARES E ALGORITMOS DE APRENDIZADO DE MÁQUINA APLICADOS A DADOS AGRONÔMICOS

4.1 Introdução

A análise de regressão é uma importante ferramenta estatística para estudar a relação entre duas ou mais variáveis. Esta relação pode se apresentar como linear ou não linear e deve ser verificada de forma adequada. A suposição incorreta de linearidade pode comprometer a confiabilidade dos testes de hipóteses e levar a modelos mal especificados. Os modelos de regressão não lineares são frequentemente adotados em diversas áreas de conhecimento para explicar uma relação não linear, entretanto impõem uma forma rígida de dependência na modelagem das variáveis em questão. Os modelos parcialmente lineares podem ser uma alternativa interessante, pois permitem estudar esta relação de forma mais flexível, por meio de funções não paramétricas. Além disso, tais modelos permitem investigar efeitos lineares e não lineares das covariáveis na variável resposta simultaneamente.

Outro ponto a ser considerado é a distribuição da variável resposta. Os modelos mais comumente utilizados, principalmente em aplicações, tem como pressupostos a normalidade do erro e a homogeneidade de variâncias. Quando tais suposições são verdadeiras, os estimadores possuem propriedades ótimas, o que constitui a justificativa teórica para este método ser amplamente difundido. Entretanto, a violação destas suposições pode ter consequências adversas na eficiência dos estimadores.

Diante do exposto, este trabalho propõe um modelo de regressão parcialmente linear, estabelecendo relações funcionais entre as covariáveis e os parâmetros de média e variância e assumindo que a variável resposta segue a distribuição exponenciated odd log-logistic normal (EOLLN). Esta distribuição pode modelar dados bimodais e/ou dados com assimetria positiva ou negativa, sendo uma alternativa aos modelos de mistura, comumente utilizados na presença de bimodalidade

Alguns modelos parcialmente lineares ou também chamados de semiparamétricos, podem ser mencionados: Vanegas e Paula (2015) utilizam modelos semiparamétricos para modelagem da mediana e da assimetria, Xu et al. (2015) propõem o modelo semiparamétrico skew-normal, Ramires et al. (2018b) introduziram o modelo de regressão semiparamétrico exponentiated sinh Cauchy para dados bimodais, assimétricos e censurados e Ramires et al. (2018a) propuseram um modelo com fração de cura para dados censurados utilizando a distribuição log-sinh Cauchy. No contexto bayesiano, Lee e Sison-Mangus (2018) estudaram um modelo de regressão semiparamétrico aplicado em microbiologia. Dhekale et al. (2017) apresentaram uma aplicação à produtividade do chá utilizando modelos parcialmente lineares.

Outras alternativas que vêm ganhando popularidade, são os modelos de aprendizado de máquina (Alonso e Renard, 2020; Oukawa et al., 2022; Khan et al., 2022; Subeesh et al., 2022). Modelos baseados em *decision trees* (DTs) e *random forests* (RFs) que são técnicas não lineares e não paramétricas, podem ser mais flexíveis que os modelos de regressão usuais. Como métodos não paramétricos, são livres de suposições e não requerem conhecimento prévio da forma funcional da relação entre a variável resposta e as covariáveis. Além disso, também são robustos à presença de outliers e podem ser usados com dados assimétricos (Kuhn et al., 2013).

Artigos relacionados às variedades de banana são importantes para agricultores, pesquisadores e profissionais da área. Existem diversas variações de potencial agronômico entre as variedades que são pouco explorados (Subeesh et al., 2022; Ortiz et al., 1998; Depigny et al., 2017). O conhecimento destes potenciais pode auxiliar a obtenção de práticas mais sustentáveis e rentáveis, sendo os padrões de crescimento um conhecimento básico. Dada a escassez destes estudos, este trabalho define um novo modelo de regressão para a altura do pseudocaule no período de plantio-floração de variedades de bananeira.

O gráfico de dispersão entre a altura do pseudocaule e o período de plantio-floração (em dias) é exibido na Figura 4.1a, em que é possível observar que estas variáveis não têm uma relação linear. A segunda covariável relacionada à variável resposta é a variedade da banana (um fator com nove níveis). A relação desta variável com a resposta é exibida na Figura 4.1b, em que é possível observar que existe uma heterogeneidade de variâncias e diversos pontos discrepantes (outliers). Neste sentido, duas ferramentas estatísticas podem ser apropriadas.

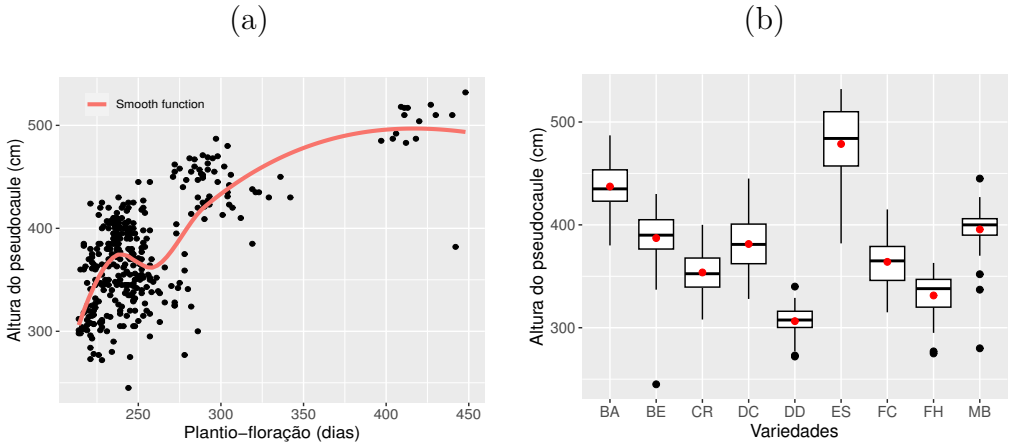

Figura 4.1. Dados de altura do pseudocaule: (a) Gráfico de dispersão entre a altura do pseudocaule e período de plantio-floração e (b) Boxplot da altura do pseudocaule por variedade.

Primeiro, um modelo parcialmente linear heterogêneo pode ser uma alternativa. Este modelo pode estudar de forma adequada a relação não linear da covariável dia

(contínua) com a resposta por meio de funções não paramétricas conjuntamente com o efeito linear da covariável variedade (fator). Além disso, o efeito da variabilidade pode ser verificado por meio de uma relação função funcional das covariáveis com o parâmetro de variância.

Em segundo lugar, os algoritmos de aprendizado de máquina DT e RF podem ser alternativas para predizer a altura do pseudocaule em termos das covariáveis. A escolha destes algoritmos baseia-se na sua simplicidade, uma vez que não requerem conhecimentos prévios da forma funcional da relação entre essas covariáveis e a variável resposta. Devido à sua fácil interpretação e implementação computacional, os modelos baseados em árvores são muito atrativos para pesquisadores de outras áreas além da estatística, embora sejam menos precisos que os modelos de regressão usuais.

4.2 Considerações finais

Este trabalho propõe um novo modelo de regressão parcialmente linear baseado na distribuição exponentiated odd log-logistic normal, motivado por um experimento agronômico de variedades de banana da terra. Neste conjunto de dados observa-se uma relação não linear entre os dias de plantio–floração e a altura do pseudocaule. Além disso, também se observa uma heterogeneidade de variâncias entre as variedades. Neste sentido, o modelo proposto mostrou-se uma alternativa adequada para estudar a relação não linear entre as variáveis por meio de funções não paramétricas e também para verificar o efeito da variabilidade das variedades na altura das plantas, através do parâmetro relacionado a variância da distribuição. São fornecidos resultados de comparações entre todas as variáveis, em especial, a variedade ES obteve diferenças significativas com relação as outras variáveis.

A capacidade preditiva do novo modelo é comparada com dois algoritmos de aprendizado de máquina: árvores de decisão e florestas aleatórias. Estes modelos podem ser alternativas interessantes para prever a variável resposta, pois não necessitavam de conhecimento prévio da forma funcional entre elas. Todos os métodos obtiveram desempenho de predição semelhante, ou seja, nenhum se destacou como melhor preditor. As florestas aleatórias alcançaram uma taxa de erro OOB estável com apenas algumas árvores, e a covariável variedade se destacou como o preditor mais importante para predizer a altura do pseudocaule da bananeira. Porém, embora as RFs sejam mais robustas que as DTs, não obteve melhores resultados quanto à capacidade preditiva de novos valores.

Nesse sentido, para quem deseja fazer predições, recomenda-se o modelo de árvore de decisão devido à sua simplicidade. Por outro lado, para pesquisadores que desejam fazer inferências, recomenda-se o novo modelo de regressão, que fornece mais informações sobre o relacionamento das variáveis em estudo, além de também ter bom desempenho preditivo. Com o modelo pode-se verificar, por exemplo, quais variedades são significativas

em relação à média e variância da resposta além de comparações entre as variedades.

Um estudo de simulação considerando dois cenários, que fornecem relações cúbicas e quadráticas entre as variáveis, mostrou que o novo modelo de regressão foi adequado para capturar diferentes formas não lineares e forneceu estimativas de máxima verossimilhança (EMVs) consistentes.

Em trabalhos futuros recomenda-se a análise de outras variáveis de resposta da bananeira devido à escassez de estudos relacionados à esta planta. Sugere-se também, a utilização deste modelo para analisar outros conjuntos de dados e problemas em outras áreas do conhecimento, em que as variáveis apresentem uma relação não linear. O modelo também pode ser interessante para casos em que a variável resposta seja bimodal e/ou assimétrica. Por fim, sugere-se a comparação do novo modelo com os métodos de aprendizado de máquina em conjuntos de dados com maior número de covariáveis.

Referências

Alonso, L. e Renard, F. (2020). A new approach for understanding urban microclimate by integrating complementary predictors at different scales in regression and machine learning models. *Remote Sensing*, 12(15):2434.

Depigny, S., Lescot, T., Achard, R., Diouf, O., Côte, F.-X., Fonbah, C., Sadom, L., e Tixier, P. (2017). Model-based benchmarking of the production potential of plantains (musa spp., aab): application to five real plantain and four plantain-like hybrid varieties in cameroon. *The Journal of Agricultural Science*, 155(6):888–901.

Dhekale, B., Sahu, P., Vishwajith, K., Mishra, P., e Narsimhaiah, L. (2017). Application of parametric and nonparametric regression models for area, production and productivity trends of tea (camellia sinensis) in india. *Indian Journal of Ecology*, 44(2):192–200.

Khan, M. A., Shah, M. I., Javed, M. F., Khan, M. I., Rasheed, S., El-Shorbagy, M., El-Zahar, E. R., e Malik, M. (2022). Application of random forest for modelling of surface water salinity. *Ain Shams Engineering Journal*, 13(4):101635.

Kuhn, M., Johnson, K., et al. (2013). *Applied predictive modeling*, volume 26. Springer.

Lee, J. e Sison-Mangus, M. (2018). A bayesian semiparametric regression model for joint analysis of microbiome data. *Frontiers in Microbiology*, 9:522.

Ortiz, R., Madsen, S., e Vuylsteke, D. (1998). Classification of african plantain landraces and banana cultivars using a phenotypic distance index of quantitative descriptors. *Theoretical and applied genetics*, 96:904–911.

Oukawa, G. Y., Krecl, P., e Targino, A. C. (2022). Fine-scale modeling of the urban heat island: A comparison of multiple linear regression and random forest approaches. *Science of the total environment*, 815:152836.

Ramires, T. G., Hens, N., Cordeiro, G. M., e Ortega, E. M. (2018a). Estimating nonlinear effects in the presence of cure fraction using a semi-parametric regression model. *Computational Statistics*, 33:709–730.

Ramires, T. G., Ortega, E. M., Hens, N., Cordeiro, G. M., e Paula, G. A. (2018b). A flexible semiparametric regression model for bimodal, asymmetric and censored data. *Journal of Applied Statistics*, 45(7):1303–1324.

Subeesh, A., Bhole, S., Singh, K., Chandel, N. S., Rajwade, Y. A., Rao, K., Kumar, S., e Jat, D. (2022). Deep convolutional neural network models for weed detection in polyhouse grown bell peppers. *Artificial Intelligence in Agriculture*, 6:47–54.

Vanegas, L. H. e Paula, G. A. (2015). A semiparametric approach for joint modeling of median and skewness. *Test*, 24:110–135.

Xu, D., Zhang, Z., e Du, J. (2015). Skew-normal semiparametric varying coefficient model and score test. *Journal of Statistical Computation and Simulation*, 85(2):216–234.

5 NOVO MODELO DE REGRESSÃO QUANTÍLICA PARA DADOS CENSURADOS

5.1 Introdução

A análise de sobrevivência é uma área da estatística que tem como variável resposta o tempo até a ocorrência de um evento de interesse. Uma característica comum deste tipo de dados é a presença de censura, ou seja, quando o evento de interesse não ocorreu e a variável resposta é parcialmente observada. Além disso, frequentemente esta variável está relacionada a uma ou mais variáveis explicativas (covariáveis), que apresentam características das unidades amostrais em estudo. Para estudar esta relação, usualmente utiliza-se os modelos de Cox ou modelos de tempo de falha acelerado (TFA). Os modelos de Cox possuem a forte suposição de riscos proporcionais, que frequentemente não é válida. Além disso, examinam os efeitos das covariáveis na função de risco, o que pode proporcionar interpretações difíceis. Alternativamente, os modelos TFA assumem uma associação entre os preditores e o tempo de sobrevivência, permitindo uma interpretação direta dos efeitos das covariáveis no tempo do evento. Entretanto, estes métodos podem falhar em capturar a heterogeneidade dos efeitos das covariáveis. Neste sentido, a regressão quantílica (RQ) (Koenker e Bassett Jr, 1978) pode ser um complemento ou uma alternativa a estes modelos permitindo avaliar os efeitos heterogêneos de preditores através da análise de diferentes quantis. Esta metodologia modela o quantil do tempo de sobrevivência e o vincula às covariáveis, deste modo, algumas vantagens de sua utilização podem ser mencionadas:

- Permite identificar e inferir sob os efeitos heterogêneos das covariáveis em diferentes quantis. Desta forma, fornecendo informações mais completadas relacionadas às covariáveis e controlando de forma flexível a heterogeneidade devida a elas;
- É flexível quanto a suposição de riscos proporcionais;
- Fornece uma interpretação direta dos resultados, isto é, entre a sobrevivência e covariáveis de interesse;
- A análise de diferentes quantis permite a identificação dos diferentes efeitos das covariáveis para indivíduos com diferentes riscos; e
- É robusta na presença de outliers.

Inicialmente, os modelos de RQ são baseados na minimização de erros absolutos ponderados (Koenker e Bassett Jr, 1978) sem qualquer distribuição de probabilidade, e a estimação dos parâmetros ocorre por meio de algoritmos de programação linear. Embora esta abordagem seja muito flexível, apresenta alguns desafios como: i) o cruzamento de

quantis, ou seja, quando duas ou mais curvas de quantis estimadas se cruzam ou se sobrepõem, causando dificuldade na interpretabilidade (Gijbels et al., 2021) e ii) a limitação em não poder explorar ferramentas da inferência paramétrica. Desta forma, distribuições de probabilidade passaram a ser relacionadas no contexto da regressão quantílica. Inicialmente, Koenker e Machado (1999) conectaram a distribuição de Laplace assimétrica a esses modelos.

Uma vasta literatura pode ser encontrada de RQ para dados censurados. Peng e Huang (2008) desenvolveram uma abordagem RQ para dados de sobrevivência com censura condicionalmente independente, Wang e Wang (2009) propuseram uma abordagem de RQ censurada ponderada localmente, Zarean et al. (2018) usou a RQ para determinar a sobrevivência e os fatores de risco no câncer de esôfago, Yang et al. (2018) apresentou uma nova abordagem para estimativa da RQ para dados censurados e Du et al. (2018) desenvolveram procedimentos de estimativa para modelos de RQ parcialmente lineares, em que algumas das respostas são censuradas por outra variável aleatória. Xue et al. (2018) ilustra a interpretação adequada do modelo de RQ censurado e as diferenças e vantagens do modelo em comparação com o modelo de Cox. Hong et al. (2019) forneceu um guia prático para usar a RQ para dados censurados à direita com covariáveis de baixa ou alta dimensionalidade e De Backer et al. (2019) estudou uma nova abordagem para a estimativa de quantis. Recentemente, muitos outros trabalhos podem ser mencionados: De Backer et al. (2020); Qiu et al. (2021); Yazdani et al. (2021); Peng (2021); Hsu et al. (2021); Wei (2022). Observe que, os trabalhos mencionados no contexto de dados censurados não utilizam modelos paramétricos ou utilizam a distribuição de Laplace assimétrica, cujos estimadores coincidem.

Posteriormente, outras distribuições foram propostas no contexto paramétrico, em que, a ideia consiste em reparametrizar uma distribuição em termos de sua função quantílica. Recentes trabalhos envolvem as seguintes distribuições: log-extended exponential-geometric (Jodra e Jimenez-Gamero, 2020); Birnbaum-Saunders (Sánchez et al., 2020, 2021); generalized half-normal discreta (Gallardo et al., 2020); transmuted unit-Rayleigh (Korkmaz et al., 2021); unit-Burr-XII (Korkmaz e Chesneau, 2021); unit-Chen (Korkmaz et al., 2022); log-symmetric (Saulo et al., 2022); arcsecant hyperbolic Weibull (Korkmaz et al., 2023) e Dagum e Singh-Maddala (Saulo et al., 2023). Contudo, há uma relativa carência na literatura de modelos para dados censurados no contexto paramétrico: generalized Gompertz (Rodrigues et al., 2021) e skew-t (Galarza Morales et al., 2021).

É bem conhecido que a função de risco pode assumir diferentes formas, o que desencadeou um grande número de novas distribuições, baseadas em extensões de distribuições comumente utilizadas, com a finalidade de obter maior flexibilidade na modelagem dos dados. Neste sentido, o presente estudo tem como objetivo apresentar um modelo de regressão quantílica baseado na reparametização da distribuição exponentiated odd

log-logistic Weibull (EOLLW).

A família que fornece está distribuição possui dois parâmetros de forma extras, permitindo a modelagem de diferentes formas de funções de risco, bem como dados com forma bimodal simétrica ou assimétrica, positiva ou negativa, tornando-se uma alternativa aos modelos de mistura comumente utilizados na presença da bimodalidade. Outra característica importante do novo modelo de RQ é que ele tem como casos especiais os modelos de RQ: exponentiated Weibull, odd log-logistic Weibull e Weibull. Adicionalmente, este trabalho mostra que o modelo pode estabelecer relações funcionais das covariáveis com outros parâmetros, incluindo escala e curtose, além do parâmetro quantílico.

5.2 Considerações finais

Um novo modelo de regressão quantílica para dados censurados é definido com base na reparametrização da distribuição exponentiated odd log-logistic Weibull (EOLLW) em termos dos quantis, considerando dois componentes sistemáticos. Algumas propriedades matemáticas da distribuição reparametrizada são apresentadas. Diversas simulações foram realizadas para diferentes configurações de parâmetros, tamanhos amostrais e porcentagens de censura, indicando a precisão das estimativas de máxima verossimilhança. A utilidade do novo modelo também foi ilustrada por meio de um conjunto de dados de câncer gástrico. O modelo de regressão quantílica proposto, permitiu avaliar de forma mais completa toda distribuição da variável resposta por meio da análise de diferentes quantis. Desta forma, pode ser uma alternativa interessante e flexível para análise de tempos de vida.

Referências

De Backer, M., El Ghouch, A., e Van Keilegom, I. (2020). Linear censored quantile regression: A novel minimum-distance approach. *Scandinavian Journal of Statistics*, 47(4):1275–1306.

De Backer, M., Ghouch, A. E., e Van Keilegom, I. (2019). An adapted loss function for censored quantile regression. *Journal of the American Statistical Association*, 114(527):1126–1137.

Du, J., Zhang, Z., e Xu, D. (2018). Estimation for the censored partially linear quantile regression models. *Communications in Statistics-Simulation and Computation*, 47(8):2393–2408.

Galarza Morales, C. E., Lachos, V. H., e Bourguignon, M. (2021). A skew-t quantile regression for censored and missing data. *Stat*, 10(1):e379.

Gallardo, D. I., Gómez-Déniz, E., e Gómez, H. W. (2020). Discrete generalized half-normal distribution and its applications in quantile regression. *SORT-Statistics and Operations Research Transactions*, 44(2):265–284.

Gijbels, I., Karim, R., e Verhasselt, A. (2021). Semiparametric quantile regression using family of quantile-based asymmetric densities. *Computational Statistics & Data Analysis*, 157:107129.

Hong, H. G., Christiani, D. C., e Li, Y. (2019). Quantile regression for survival data in modern cancer research: expanding statistical tools for precision medicine. *Precision clinical medicine*, 2(2):90–99.

Hsu, C.-Y., Wen, C.-C., e Chen, Y.-H. (2021). Quantile function regression analysis for interval censored data, with application to salary survey data. *Japanese Journal of Statistics and Data Science*, pages 1–20.

Jodra, P. e Jimenez-Gamero, M. D. (2020). A quantile regression model for bounded responses based on the exponential-geometric distribution. *REVSTAT-Statistical Journal*, 18(4):415–436.

Koenker, R. e Bassett Jr, G. (1978). Regression quantiles. *Econometrica: journal of the Econometric Society*, pages 33–50.

Koenker, R. e Machado, J. A. (1999). Goodness of fit and related inference processes for quantile regression. *Journal of the american statistical association*, 94(448):1296–1310.

Korkmaz, M. C., Altun, E., Chesneau, C., e Yousof, H. M. (2022). On the unit-chen distribution with associated quantile regression and applications. *Mathematica Slovaca*, 72(3):765–786.

Korkmaz, M. Ç. e Chesneau, C. (2021). On the unit burr-xii distribution with the quantile regression modeling and applications. *Computational and Applied Mathematics*, 40(1):29.

Korkmaz, M. Ç., Chesneau, C., e Korkmaz, Z. S. (2021). Transmuted unit rayleigh quantile regression model: Alternative to beta and kumaraswamy quantile regression models. *Univ. Politeh. Buchar. Sci. Bull. Ser. Appl. Math. Phys*, 83:149–158.

Korkmaz, M. Ç., Chesneau, C., e Korkmaz, Z. S. (2023). A new alternative quantile regression model for the bounded response with educational measurements applications of oecd countries. *Journal of Applied Statistics*, 50(1):131–154.

Peng, L. (2021). Quantile regression for survival data. *Annual review of statistics and its application*, 8:413–437.

Peng, L. e Huang, Y. (2008). Survival analysis with quantile regression models. *Journal of the American Statistical Association*, 103(482):637–649.

Qiu, Z., Ma, H., Chen, J., e Dinse, G. E. (2021). Quantile regression models for survival data with missing censoring indicators. *Statistical methods in medical research*, 30(5):1320–1331.

Rodrigues, A., Borges, P., e Santos, B. (2021). A defective cure rate quantile regression model for male breast cancer data. *arXiv preprint arXiv:2105.03699*.

Sánchez, L., Leiva, V., Galea, M., e Saulo, H. (2020). Birnbaum-saunders quantile regression models with application to spatial data. *Mathematics*, 8(6):1000.

Sánchez, L., Leiva, V., Galea, M., e Saulo, H. (2021). Birnbaum-saunders quantile regression and its diagnostics with application to economic data. *Applied Stochastic Models in Business and Industry*, 37(1):53–73.

Saulo, H., Dasilva, A., Leiva, V., Sánchez, L., e de la Fuente-Mella, H. (2022). Log-symmetric quantile regression models. *Statistica Neerlandica*, 76(2):124–163.

Saulo, H., Vila, R., Borges, G. V., Bourguignon, M., Leiva, V., e Marchant, C. (2023). Modeling income data via new parametric quantile regressions: Formulation, computational statistics, and application. *Mathematics*, 11(2):448.

Wang, H. J. e Wang, L. (2009). Locally weighted censored quantile regression. *Journal of the American Statistical Association*, 104(487):1117–1128.

Wei, B. (2022). Quantile regression for censored data in haematopoietic cell transplant research. *Bone Marrow Transplantation*, 57(6):853–856.

Xue, X., Xie, X., e Strickler, H. D. (2018). A censored quantile regression approach for the analysis of time to event data. *Statistical methods in medical research*, 27(3):955–965.

Yang, X., Narisetty, N. N., e He, X. (2018). A new approach to censored quantile regression estimation. *Journal of Computational and Graphical Statistics*, 27(2):417–425.

Yazdani, A., Yaseri, M., Haghighat, S., Kaviani, A., e Zeraati, H. (2021). The comparison of censored quantile regression methods in prognosis factors of breast cancer survival. *Scientific Reports*, 11(1):18268.

Zarean, E., Mahmoudi, M., Azimi, T., e Amini, P. (2018). Determining overall survival and risk factors in esophageal cancer using censored quantile regression. *Asian Pacific journal of cancer prevention: APJCP*, 19(11):3081.

6 FLORESTAS ALEATÓRIAS DE SOBREVIVÊNCIA E NOVO MODELO DE REGRESSÃO PARA ANÁLISE DE DADOS CENSURADOS

6.1 Introdução

A análise de sobrevivência é uma área da estatística que analisa dados que tem como variável resposta o tempo até a ocorrência de um evento, denominado de tempo de vida, tempo de falha ou tempo de sobrevivência. Este evento pode ser a morte, o aparecimento de um tumor, o desenvolvimento de uma doença, a quebra de um componente eletrônico, entre outros. A principal característica em relação aos dados de sobrevivência é a presença de censura, que é a observação parcial da resposta. Em geral, neste tipo de análise o interesse é verificar a influência de covariáveis em relação ao tempo de ocorrência de um evento. Um dos modelos amplamente utilizados é o modelo de regressão de Cox. Entretanto, este modelo tem suposições restritivas, como a de riscos proporcionais, isto é, supõe que a função de risco para quaisquer dois indivíduos é constante ao longo do tempo, o que nem sempre é verdade. Além disso, examina o efeito da covariável na função de risco, o que pode tornar a interpretação dos resultados difícil.

Neste sentido, os modelos de regressão paramétricos passaram a ser amplamente utilizados. Estes modelos relacionam as covariáveis diretamente a função de sobrevivência, levando a interpretação direta dos efeitos das covariáveis no tempo do evento. Como é bem conhecido, a função de risco é muito útil para descrever a distribuição do tempo de vida e pode assumir diferentes formas, o que levou a proposta de um grande número de novas famílias de distribuições com o objetivo de obter maior flexibilidade de modelagem de dados.

Por exemplo, Mudholkar et al. (1996) propõe a família de distribuições exponentiated-G, Marshall e Olkin (1997) introduzem a família de distribuições Marshall-Olkin-G, Eugene et al. (2002) propõe a família de distribuição beta-G, Cordeiro e de Castro (2011) propõem a família Kumaraswamy-G, Zografos e Balakrishnan (2009) definem a família gamma-G, Alzaatreh et al. (2013) apresenta a família transformer (T-X), entre outras.

Este trabalho utiliza a família de distribuições proposta recentemente em Alizadeh et al. (2020), denominada exponentiated odd log-logistic-G e a distribuição de probabilidade denominada de generalized Rayleigh (GR) proposta por Kundu e Raqab (2005) para definir a nova distribuição denominada de exponentiated odd log-logistic generalized Rayleigh (EOLLGR). Esta nova distribuição apresenta bastante flexibilidade. Por exemplo, permite a modelagem de diferentes formas da função de risco (constante, unimodal, forma de U). Assim, um dos objetivos deste trabalho é propor o modelo de regressão, baseada na distribuição EOLLGR, considerando dados censurados.

Por outro lado, em grande ascensão atualmente, o machine learning (ML) é um

campo da inteligência artificial (IA), que consiste em alimentar as máquinas com dados para elas aprendam com eles e sejam capazes de prever novos resultados quando novos dados forem apresentados. A IA recebeu grande atenção na área médica, o que motivou a sua aplicação em dados de sobrevivência. Estes algoritmos são muito flexíveis, o que pode melhorar a precisão preditiva. A obtenção de previsões precisas das probabilidades de sobrevivência, pode auxiliar profissionais da área da saúde a desenvolver planos viáveis para programas de triagem e para melhores estratégias de tratamento, a fim de ajudar a reduzir a carga de doenças e melhorar a sobrevida dos pacientes.

A metodologia das *Random survival forests* (RSF) (florestas aleatórias de sobrevivência) (Ishwaran et al., 2008) estende o método *Random forests* (RFs) (Breiman, 2001) para dados de sobrevivência censurados à direita. Este é um método estatístico não linear e não paramétrico, baseado em um conjunto de *decision trees* (DTs) (árvores de decisão) (Breiman, 1984). O autor introduz duas formas de aleatorização na construção das árvores, mostrando que o desempenho preditivo pode ser muito melhor. Como não requer suposições distribucionais sobre a relação das covariáveis com a variável resposta, pode ser uma alternativa interessante aos modelos de regressão usuais. Com o ajuste de várias árvores, as estimativas do modelo são estabilizadas, otimizando a acurácia de predição, além disso, os efeitos não lineares ou interações de ordem superior para os preditores não precisam ser definidos previamente, como em regressões usuais.

Neste contexto, este trabalho se concentra na comparação de duas ferramentas estatísticas: o modelo flexível de regressão baseado na distribuição EOLLGR e as florestas aleatórias de sobrevivência. Esta comparação é feita realizando diferentes estudos de simulações com diferentes cenários e analisando um conjunto de dados epidemiológicos de pacientes internados com Síndrome Respiratória Aguda Grave (SRAG) por Covid-19, no Brasil.

6.2 Considerações finais

Este trabalho propõe uma nova distribuição de probabilidade com base na família exponentiated odd log-logistic family e na distribuição generalized Rayleigh, chamada exponentiated odd log-logistic generalized Rayleigh. Algumas propriedades matemáticas são fornecidas, úteis para pesquisadores da área estatística. Um modelo de regressão para dados censurados é proposto com base na nova distribuição e sua capacidade preditiva é comparada ao modelo de florestas aleatórias de sobrevivência. Um estudo de simulação mostra a consistência das estimativas de máxima verossimilhança do novo modelo para diferentes porcentagens de censura. Além disso, mostrou que ele pode ser um modelo competitivo com as florestas aleatórias com relação ao desempenho preditivo. Uma aplicação a dados de sobrevivência de pacientes com Covid-19 é realizada para verificar a utilidade do modelo. Em geral, entre os modelos de regressão, o modelo EOLLGR mostrou-se o

mais adequado. Quanto a capacidade preditiva dos modelos, todos obtiveram desempenhos de predição semelhantes, incluindo o modelo de floresta aleatória. Além disso, ambas as metodologias fornecerem probabilidades de sobrevivência similares.

Os modelos de florestas aleatória selecionaram como variáveis mais importantes para predizer o tempo de vida: idade, doença renal, doença neurológica e doença cardiovascular. Por outro lado, os modelos de regressão selecionaram idade, doença renal e doença neurológica, sendo que, a variável doença cardiovascular não mostrou melhorar seus desempenhos, portanto, não foi selecionada. A partir do modelo de regressão EOLLGR, pode ser observado que os tempos de sobrevivência diminuem com a presença de doença neurológica e doença renal e também com o aumento da variável idade. Com base nas duas metodologias utilizadas, o sexo não foi determinante no tempo de vida dos pacientes.

Pode-se concluir que o novo modelo de regressão foi adequado para o conjunto de dados analisado, podendo ser considerado uma alternativa flexível para análise de tempos de vida. Além disso, o novo modelo mostrou-se competitivo ao algoritmo de machine learning para predizer novos valores. Futuramente, o novo modelo pode ser comparado a outros algoritmos ou considerando outras situações, como maior número de covariáveis.

Referências

Alizadeh, M., Tahmasebi, S., e Haghbin, H. (2020). The exponentiated odd log-logistic family of distributions: Properties and applications. *Journal of Statistical Modelling: Theory and Applications*, 1(1):29–52.

Alzaatreh, A., Lee, C., e Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron*, 71(1):63–79.

Breiman, L. (1984). *Classification and regression trees*. Routledge.

Breiman, L. (2001). Random forests. *Machine learning*, 45:5–32.

Cordeiro, G. M. e de Castro, M. (2011). A new family of generalized distributions. *Journal of statistical computation and simulation*, 81(7):883–898.

Eugene, N., Lee, C., e Famoye, F. (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and methods*, 31(4):497–512.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., e Lauer, M. S. (2008). Random survival forests. *Ann. Appl. Stat.*, 2:841–860.

Kundu, D. e Raqab, M. Z. (2005). Generalized rayleigh distribution: different methods of estimations. *Computational statistics & data analysis*, 49(1):187–200.

Marshall, A. W. e Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and weibull families. *Biometrika*, 84(3):641–652.

Mudholkar, G. S., Srivastava, D. K., e Kollia, G. D. (1996). A generalization of the weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association*, 91(436):1575–1583.

Zografos, K. e Balakrishnan, N. (2009). On families of beta-and generalized gamma-generated distributions and associated inference. *Statistical methodology*, 6(4):344–362.

7 CONSIDERAÇÕES FINAIS

Novos modelos de regressão são definidos neste trabalho para acomodar diferentes tipos de dados, baseados na família de distribuições exponentiated odd log-logistic (EOLL-G), que possui a flexibilidade de modelar dados bimodais, simétricos ou assimétricos. Os novos modelos são denominados de exponentiated odd log-logistic Normal (EOLLN), exponentiated odd log-logistic Weibull (EOLLW) e exponentiated odd log-logistic generalized Rayleigh (EOLLGR). Propriedades estruturais das novas distribuições foram fornecidas, que exibem a flexibilidade da família utilizada e podem ser úteis para trabalhos futuros. O método de máxima verossimilhança foi utilizado para estimação os parâmetros e estudos de simulações para ambos os modelos são realizados, comprovando a consistência das estimativas.

Dois modelos de regressão quantílica são propostos baseados em reparametrizações das distribuições EOLLN e EOLLW, para dados não censurados e censurados, respectivamente. Três aplicações são realizadas com o modelo EOLLN, sendo duas delas considerando dados de experimentação agronômica. Uma aplicação a dados de pacientes com câncer gástrico foi realizada com o modelo EOLLW. Os novos modelos mostraram-se adequados para os conjuntos de dados utilizados e forneceram informações a respeito da variável resposta em estudo por meio da análise de vários quantis.

Duas novas famílias bivariadas são definidas a partir da família EOLL-G e das cópulas de Clayton e de Frank. Uma aplicação a dados de um experimento que avaliou o crescimento da alface de carvalho, mostrou que as cópulas foram adequadas para modelar o comportamento bimodal das variáveis em estudo, para explicar a estrutura de dependência positiva entre as variáveis e para predizer os valores de massa fresca e altura de plantas da alface de carvalho sob efeito dos diferentes tratamentos.

Um modelo parcialmente linear, também baseado na distribuição EOLLN, foi proposto para estudar relações não lineares entre as variáveis utilizando funções não paramétricas. A capacidade preditiva do novo modelo foi comparada com dois algoritmos de machine learning (ML): árvores de decisão e florestas aleatórias. Esta comparação foi realizada por meio de um estudo de simulação e uma aplicação a dados de um experimento agronômico de variedades de banana da terra. O modelo proposto mostrou-se uma alternativa adequada para estudar a relação não linear entre as variáveis por meio de funções não paramétricas e também para verificar os efeitos das covariáveis na variabilidade da variável resposta, através da modelagem do parâmetro relacionado a variância da distribuição. Quanto a capacidade preditiva, o novo modelo mostrou-se competitivo aos algoritmos de ML, de acordo com o estudo de simulação e a aplicação realizada.

Um segundo modelo de regressão para dados censurados é definido com base na distribuição EOLLGR. Este modelo teve seu desempenho preditivo comparado ao algoritmo de ML: florestas aleatórias de sobrevivência. A comparação foi realizada por

meio de um estudo de simulação e de uma aplicação a pacientes com Covid-19. O novo modelo mostrou-se adequado para descrever este conjunto de dados em comparação com seus submodelos e mostrou-se competitivo ao algoritmo de ML na predição de novos dados, considerando o estudo de simulação realizado e o conjunto de dados analisado. Desta forma, pode ser uma alternativa interessante e flexível para análise de tempos de vida.

Pesquisas futuras

Como perspectiva futura deste trabalho, os seguintes temas podem ser abordados:

- Modelos de regressão quantílica com efeitos aleatórios, funções de suavização não paramétricas ou termos aditivos;
- Modelos de regressão bivariados com efeitos aleatórios, funções de suavização não paramétricas ou termos aditivos. Além disso, pode-se considerar modelos bivariados de regressão quantílica;
- Outras distribuições de base podem ser utilizadas com as novas famílias bivariadas;
- Para dados censurados, pode-se considerar extensões dos modelos propostos considerando a presença de fração de cura, variáveis respostas bivariadas, efeitos aleatórios, funções de suavização não paramétricas ou termos aditivos. Além disso, pode-se considerar a regressão quantílica para todos estes contextos;
- Os modelos propostos podem ser utilizados para aplicações em outras áreas, cujos conjuntos de dados apresentem as particularidades estudadas nesta tese;
- Comparações dos modelos propostos com outros métodos de machine learning ou considerando outras situações, como por exemplo, maior número de covariáveis.