Nikolas Argiropoulos 20260509
STT6516 – Hiver 2024
Projet Finale
Analyzing Brain Cancer Pain Severity: Mixed-Effect Regression Comparisons

1. Introduction

This research project aims to investigate the impact of several key factors on the severity of pain experienced by brain cancer patients using two generalized linear mixed models (GLMM). Specifically, the study examines the influence of the number of tumors, tumor size, sex, cancer stage, and age on pain severity among patients. The objectives of the project are threefold: firstly, to employ logistic and ordinal mixed model regressions to analyze the relationship between the variables and pain severity in brain cancer patients; secondly, to evaluate the model fit by comparing additive, second-order, and third-order interaction models to the saturated model using Deviance, Akaike Information Criterion (AIC), and Likelihood Ratio Test (LRT); and thirdly, to qualitatively determine which model provides the best fit for the observed data. By addressing these objectives, the study aims to enhance the understanding of the factors influencing pain severity in brain cancer patients.

2. Underlying Principles of the Models

Generalized Linear Mixed Models are extensions of linear mixed models (LMMs) that offer flexibility in modeling the relationship between predictors and outcomes. GLMMs employ a link function to connect the linear predictor to the distribution of the outcome variable, enabling the modeling of various data non-continuous and non-normal distributions. For instance, GLMMs can utilize the logit link for binary outcomes, the log link for count data, or the probit link for ordinal data[1,2]. In contrast, LMMs typically use the identity link function, which is suitable for normally distributed outcomes. Moreover, GLMMs offer several advantages over generalized linear models (GLMs) when modeling multilevel, hierarchical, nested structures, or longitudinal data. GLMMs accommodate random effects to capture variability that is not explicitly accounted for by the fixed effects[1,2]. For instance, in this study where patients are nested within different doctors, accounting for the intra-correlation or nesting within each doctor necessitates coding "doctor" as a random effect. The mixed-effects logistic regression model utilized in this analysis

can be expressed as: $\text{logit}\left(\text{P}\left(Y_{ij}=1|\,X_{ij},b_j\right)\right)=X_{ij}\beta+Z_{ij}b_j$ where, $Y_{ij}$is the binary outcome for the ith observation in the jth group, $\text{P}\left(Y_{ij}=1|\,X_{ij},b_j\right)$ is the probability of the ith observation in the jth group belonging to category 1 given the predictors X_{ij} and the random effects b_j, logit(·) is the logit link function transforming probabilities to the log-odds scale, X_{ij} is the vector for fixed effects, β is the vector for fixed-effects coefficients, Z_{ij} is the design matrix for the random effects associated with the ith observation in the jth group and b_j is the vector of random effects for the jth group ($b_j \sim N(0,\Sigma)$) where $\mathbf{\Sigma}$ captures the variability in the random effects across different doctors. Similarly, the mixed-effects ordinal regression model can be represented as: $\text{logit}\big(\text{P}(\text{Y}\leq j)\big)=\alpha_j+X_{ij}\beta+Z_{ij}\gamma$ where Y is the outcome variable, $\text{logit}\big(\text{P}(\text{Y}\leq j)\big)$ represents the cumulative log-odds of being in or below category j, α_j is the intercept parameter for category j, X_{ij} is the design matrix for fixed effects, Z_{ij} is the design matrix for random effects, β is the vector for fixed-effects coefficients and γ is the vector of random-effect coefficients ($\gamma \sim N(0,\Sigma)$) where $\mathbf{\Sigma}$ captures the variability in the random effects across different doctors. These models enable the measurement of cancer pain while accounting for the variability of scores within doctors.

3. Methods

3.1. Data

The data were obtained from the *UCLA Statistical Methods and Data Analytics* website[3]. Two models were employed for the analysis: mixed-effect logistic regression and mixed-effect ordinal regression. Both models included the number of tumors, tumor size (mm), sex, cancer stage, and age as independent variables. Cancer stage and participant age were determined by the primary physician, while the number of tumors and tumor size were assessed via MRI by counting the number of tumors and measuring average tumor size, respectively. Patient pain severity was originally assessed on a 10-point scale (1-10). However, for the mixed-effect

logistic regression, pain was dichotomized using a median split: participants with scores greater than 5 were coded as experiencing pain, and those with scores less than 4 were coded as having no pain[4]. Conversely, for the mixed-effect ordinal regression, pain was recategorized into a 4-point scale: participants scoring 1-2 were categorized as experiencing "None," 3-4 as "Mild," 5-7 as "Moderate," and 8-10 as "Severe.

3.2. Data analysis

Descriptive and inferential analyses were conducted using R version 4.3.1 using the *ordinal* package[5,6]. Continuous variables were reported as the mean (*M*) with standard deviation (*SD*. A two-tailed significance criterion of 0.05 was applied for all GLMMs. To address convergence issues in the model, the number of tumors, tumor size, and age were standardized to have a mean of 0 and a standard deviation of 1. The assumptions required for the mixed-effects ordinal regression, including linearity of continuous variables and normality of random effects, and the proportional odds assumption, were not violated. Similarly, for the mixed-effects logistic regression, the assumptions of linearity of continuous variables, normality of random effects, absence of influential outliers, and multicollinearity were supported.

4. Results

4.1. Descriptive Analysis

The sample consisted of 8525 patients and 407 doctors. The majority of the sample (60%) were female. The average patient age was 50.97 years (*SD* = 6.28), with an average tumor size of 70.88 mm (*SD* = 12.07) and an average number of tumors of 3.07 (*SD* = 2.55). The average length of stay was 5.49 days (*SD* = 1.05). On a 4-point scale, the average pain intensity was categorized as "mild" (*M* = 2.61, *SD* = 0.61), and 73.63% of participants were categorized as experiencing "pain" on a binary scale. Further details of the socio-demographic characteristics partitioned by cancer stage are presented in Appendix (Table 1).

4.2. Mixed-Effects Logistic Regression Analysis

The model suggested that tumor size, age, and cancer stage were found to be nonsignificant predictors of binary cancer pain. However, the number of tumors and sex emerged as statistically significant predictors binary cancer pain ($p < 0.01$). Specifically, a one standard deviation unit increase in the number of tumors was associated with a 10% increase in the odds of experiencing pain (OR = 1.10, 95% CI [1.04, 1.16]), while holding all other variables constant. Additionally, male participants had 40% lower odds of experiencing pain compared to female participants (OR = 0.60, 95% CI [0.54, 0.66]), after adjusting for all other variables. See Appendix for additive model (Table 2).

4.3. Mixed-Effects Ordinal Regression Analysis

The analysis revealed that tumor size and age were not statistically significant predictors of cancer pain. However, the number of tumors, sex, and cancer stage were statistically significant predictors ($p < 0.05$). Specifically, a one standard deviation unit increase in the number of tumors was associated with a 10% increase in the odds of experiencing pain (OR = 1.10, 95% CI [1.04, 1.16]), holding all other variables constant. Furthermore, male participants had 39% lower odds of experiencing pain compared to female participants (OR = 0.61, 95% CI [0.55, 0.67]), after adjusting for all other variables. Additionally, participants diagnosed with stage 4 cancer had 18% lower odds of experiencing severe pain compared to those diagnosed with stage 1 cancer, while holding all other variables constant (OR = 0.82, 95% CI [0.67, 0.99]). This finding could be due to the likelihood that patients undergoing treatment for stage 4 cancer are more frequently prescribed stronger medications compared to those in stage 1. Consequently, the use of morphine may act as a confounding factor in this association. Moreover, threshold coefficients between the "None|Mild" categories (TC = -4.59) suggest that as the predictor variables increase, there is 4.59 reduced odds of moving from "None" to "Mild" category. Similarly, threshold coefficients between the "Mild|Moderate" categories (TC = -1.42) suggest that as the predictor variables

increase, there is 1.42 reduced odds of moving from “Mild” to “Moderate” category. Conversely, threshold coefficients between the “Moderate|Severe” categories (TC = 2.26) suggest that as the predictor variables increase, there is 2.26 increased odds of moving from “Moderate” to “Severe” category. In other words, patients who initially reported “None” and “Mild” pain levels are more likely to report the same score as the study progresses, whereas patients initially reporting “Moderate” pain levels are more likely to report “Severe” pain later in the study. See Appendix for additive model (Table 2).

4.4. Model-Fit Analysis

Both models indicated significance for the number of tumors and male gender. Additionally, the mixed-effects ordinal regression also identified cancer stage as a significant predictor. Evaluation of the mixed-effects logistic regression, the deviance, AIC, and LRT suggested that the additive, second-order, and third-order interaction models fit the data better than the saturated model. Specifically, when considering AIC and LRT, the additive model emerged as the preferred option, while examination of deviance and LRT favored the third-order interaction model. Similarly, for the mixed-effects ordinal regression, analysis based on deviance, AIC, and LRT indicated that the additive, second-order, and third-order interaction models outperformed the saturated model. Evaluation of the AIC and LRT favored the additive model, while assessment of deviance and LRT favored the third-order interaction model. However, given the low and nonsignificant values of AIC and LRT, respectively, and the principle of parsimony, we proceeded with the additive model. Although some third and fourth term interactions were observed, prioritizing model parsimony is warranted, as higher-order interactions tend to become more complex and less interpretable with an increasing number of variables in the model.

5. Conclusion

Based on the findings from the model-fit analysis and the results of the final model, the mixed-effects ordinal regression model appears to be superior to the mixed-effects logistic

regression model. The mixed-effects logistic regression employed a median split to construct the binary outcome variable, a method associated with potential drawbacks such as loss of information, diminished statistical power, risk of bias and misclassification[7,8,9]. Additionally, the mixed-effects ordinal regression produced lower standard errors and narrower confidence intervals for each predictor variable compared to the mixed-effects logistic regression. Furthermore, the mixed-effects ordinal regression model identified cancer stage as a statistically significant predictor, a finding that was not observed in the mixed-effects logistic regression model. Therefore, from a qualitative standpoint, the mixed-effects ordinal regression model provides a more accurate representation of the data.

6. Limitations/future considerations

The study faced limitations primarily related to sample size constraints and model convergence issues. The small sample size relative to the number of categories and predictor variables posed challenges, necessitating the reduction of categories in the outcome variable in both the logistic and ordinal mixed-effects models to facilitate convergence. Furthermore, due to the low number of quadrature points utilized in the adaptive Gauss-Hermite quadrature approximation of the likelihood function may lead to reduced estimator precision. Another limitation is not accounting for all the third- and fourth-order interactions; although their interpretation is complex and non-interpretable, it can reduce statistical power and bias estimates[10]. These limitations underscore the need for caution in interpreting the results and highlight avenues for future research. Addressing these challenges could involve increasing sample sizes, exploring alternative modeling techniques, and employing more robust estimation methods to enhance reliability.

References

1. Agresti, A. (2013). *Categorical Data Analysis*. John Wiley & Sons (etc.).
2. Dean, C., & Nielsen, J. (2007). Generalized linear mixed models: a review and some extensions. *Lifetime Data Analysis*, 13, 497-512. https://doi.org/10.1007/S10985-007-9065-X.
3. *Mixed Effects Logistic Regression*. UCLA Statistical Methods and Data Analysis. (n.d.). https://stats.oarc.ucla.edu/r/dae/mixed-effects-logistic-regression/
4. Iacobucci, D., Posavac, S., Kardes, F., Schneider, M., & Popovich, D. (2015). Toward a More Nuanced Understanding of the Statistical Properties of a Median Split. *Econometrics: Econometric & Statistical Methods - General eJournal.* https://doi.org/10.2139/ssrn.2663427.
5. Christensen, Rune Haubo B. (2016, December 15). *A tutorial on fitting cumulative link models with the ordinal ...* rdrr.io. https://rdrr.io/rforge/ordinal/f/inst/doc/clm_tutorial.pdf
6. Christensen, Rune Haubo Bojesen. (2022, November 17). *CLMM: Cumulative link mixed models in ordinal: Regression models for ordinal data*. clmm: Cumulative Link Mixed Models in ordinal: Regression Models for Ordinal Data. https://rdrr.io/cran/ordinal/man/clmm.html
7. Altman, D. G., Royston, P., & D., (2006). The cost of dichotomising continuous variables. *Bmj*, 332(7549), 1080.
8. Hedeker, D. (2003). A mixed-effects Multinomial logistic regression model. *Statistics in Medicine*, *22*(9), 1433–1446. https://doi.org/10.1002/sim.1522
9. Callum, R. C., Zhang, S., Preacher, K. J., & Rucker, D. D. (2002). On the practice of dichotomization of quantitative variables. *Psychological methods, 7*(1), 19-40.
10. Porter, M. A., Aike, L. S., & West, S. G. (1994). Multiple regression: Testing and interpreting interactions. *The Statistician*, *43*(3). https://doi.org/10.2307/2348581

Appendix

Table 1. Descriptive Statistics

Cancer Stage	**I**, N = 2,558[1]	**II**, N = 3,409[1]	**III**, N = 1,705[1]	**IV**, N = 853[1]
Age	47.55 (5.61)	50.98 (5.68)	53.38 (5.53)	56.41 (5.75)
Sex				
Female	1,542 (60%)	2,049 (60%)	1,000 (59%)	524 (61%)
Male	1,016 (40%)	1,360 (40%)	705 (41%)	329 (39%)
Tumor Size	73.07 (12.81)	70.24 (12.20)	69.03 (10.93)	70.58 (10.35)
Number of Tumors	2.30 (2.14)	2.69 (2.36)	3.59 (2.46)	5.84 (2.55)
Length of Stay	4.90 (0.95)	5.50 (0.93)	5.91 (0.93)	6.40 (0.95)
Pain Category				
None	39 (1.5%)	57 (1.7%)	35 (2.1%)	18 (2.1%)
Mild	615 (24%)	869 (25%)	415 (24%)	200 (23%)
Moderate	1,664 (65%)	2,176 (64%)	1,104 (65%)	564 (66%)
Severe	240 (9.4%)	307 (9.0%)	151 (8.9%)	71 (8.3%)
Pain Binary				
No Pain	654 (26%)	926 (27%)	450 (26%)	218 (26%)
Pain	1,904 (74%)	2,483 (73%)	1,255 (74%)	635 (74%)

[1]Mean (SD); n (%)

Table 2. Mixed-Effect Logistic Regression Additive Model

Variables	OR[1]	95%CI[1]	SE	p-value
Tumor Size	1.04	0.98, 1.11	0.03	0.2
Number of Tumors	1.10	1.04, 1.18	0.03	**0.002**
Age	1.02	0.96, 1.08	0.03	0.5
Sex				
Female	—	—		
Male	0.60	0.54, 0.66	0.05	**<0.001**
Cancer Stage				
I	—	—		
II	0.89	0.79, 1.02	0.07	0.085
III	0.90	0.77, 1.06	0.08	0.2
IV	0.85	0.68, 1.06	0.11	0.15

[1]OR = Odds Ratio, CI = Confidence Interval

Table 3. Mixed-Effect Ordinal Regression Additive Model

Variables	OR[1]	95%CI[1]	SE	p-value
Tumor Size	1.03	0.97, 1.08	0.03	0.3
Number of Tumors	1.10	1.04, 1.16	0.03	**0.001**
Age	1.01	0.96, 1.16	0.03	0.7
Sex				
Female	—	—		
Male	0.61	0.55, 0.67	0.05	**<0.001**
Cancer Stage				
I	—	—		
II	0.90	0.80, 1.01	0.06	0.074
III	0.90	0.78, 1.03	0.07	0.13
IV	0.82	0.67, 0.99	0.10	**0.041**

[1]OR = Odds Ratio, CI = Confidence Interval