Week 6 Assignment: Predictive Modeling Using Generalized Linear Models

Jonathan Ibifubara Pollyn

College of Science, Engineering and Technology, Grand Canyon University

DSC-530: Predictive Modeling

Filippo Posta

October 13, 2021

Part 1: Operational Tasks

Construct a table of statistics summarizing your clusters. Describe what these two clusters consist of and test validation.

One benefit of modeling a logistic model is determining the outcome of the response variable, giving a set of predictors. In this report, the adult dataset is used to create a logistic regression model; from the dataset, the age, education-num, and hours-per-week are predictors and the income as a response. After building the model, the p-values from the summary below show that all the variables belong to the model because they are less than the cut of the value of 0.05, so we do not need to remove any variable.

	Coef.	Std.Err.	z	P>|z|	[0.025	0.975]
const	-8.4611	0.1252	-67.5880	0.0000	-8.7064	-8.2157
age	0.0459	0.0013	34.9566	0.0000	0.0434	0.0485
education-num	0.3449	0.0074	46.5794	0.0000	0.3304	0.3595
hours-per-week	0.0423	0.0014	29.1656	0.0000	0.0394	0.0451

After evaluating the above data, we can easily extract the descriptive form of the final regression model can be expressed as follows.

$$\text{yhat(income)}=\frac{\exp\left(-8.461+0.046(age)+0.345(education-num)+0.042(hours-per-week)\right)}{1+exp(-8.461+0.046(age)+0.345(education-num)+0.042(hours-per-week))}$$

The age coefficient explains that income is likely to increase for every increase in age by approximately 105% or 1.05. The probability of having a higher income every ten years is likely to increase by 160%. The coefficient of the education num explains that income is likely to increase with an increase in education-num by approximately 140% or 1.4. Then if someone has

four or more years of education, their income is likely to increase more than double about 397% or 3.97. The coefficient per week stipulates that the income would grow at a rate of 1.2 hours for every additional hour per week.

The predicted value is obtained from the prediction of the predictors, while the target value if the actual response variable did not tell any difference in information. The Poisson regression with the same adult data shows that all variables belong to the model, and we do not have to remove anyone. The summary below shows that the p-values for all three variables are below the cutoff value of 0.05.

	coef	std err	z	P>|z|	[0.025	0.975]
const	-5.8308	0.080	-73.288	0.000	-5.987	-5.675
age	0.0275	0.001	28.924	0.000	0.026	0.029
education-num	0.2100	0.005	39.866	0.000	0.200	0.220
hours-per-week	0.0230	0.001	24.289	0.000	0.021	0.025

We can retrieve the descriptive form of the Poisson regression from the above data as $yhat = e^{(-5.831+0.028(age)+0.210(education-num)+0.023(hours-per-week))}$

PART 2: Mathematical and Statistical Basis

Evaluate the generalized linear mixed model (GLMM) presented in the statistical analysis section. Is this negative binomial regression fit appropriate for this situation? Compare this model to the generalized additive mixed model (GAMM) based on the Poisson distribution. Which model is more appropriate for this situation, as discussed in the results section?

Kamiyama, M. T., Bradford, B. Z., Groves, R. L., & Guédot, C. (2020). Degree day models to forecast the seasonal phenology of drosophila suzukii in tart cherry orchards in the midwest u.s. *PLOS ONE*, *15*(4), e0227726. https://doi.org/10.1371/journal.pone.0227726

According to the authors, the spotted-wing drosophila (Drosophilidae) pest targets soft-skinned and stone fruit, Drosophila suzukii (Matsumura), and based on four years of adult monitoring trap data from Wisconsin tart cherry orchards gathered during the growing season, the authors work to constructs predictive generalized linear mixed models (GLMM) and generalized additive mixed models (GAMM) of D. suzukii's dynamic seasonal floristics. The author asserted that relative humidity and degree days are incorporated into the models and relate to trapping catch. The authors also stipulated that the GLMM calculated a coefficient of 2.21 for DD/1000, indicating that the trap catch rises by around nine flies for every 1000 degree of day increase. The degree of days smoothing function approximates the important points of the first adult D. suzukii detection at 1276 DD, above-average field populations beginning at 2019 degree of days, and peak activity in the 3180 degrees of days' timeframe and the peak activity timeframe is 3180 degree of days. The author's incorporated seasonal floristics data spanning four

years from the same sites to introduce strong models capable of forecasting altering adult D. suzukii population trends in the field, which makes it possible to implement more effective management techniques earlier and with more success. The authors created a generalized linear mixed model (GLMM) and a generalized additive mixed model (GAMM) better to comprehend the complicated population dynamics of D. suzukii. The authors stipulated that previous studies have used linear and additive models to better understand insect phenology through population modeling; researchers may monitor changes in species phenology, investigate the role abiotic factors play in population dynamics, and get a basic idea of insect population density. When it comes to phenological events like initial detection or peak activity, the authors asserted that GAM models could estimate different degree days. In contrast, GLM models can approximate trap catches if they are given a specific detection date. Compatible models can emphasize different aspects of phenology to understand better how abiotic factors influence the pest's seasonal population dynamics. The authors confirmed that that action enables making more precise predictions about the seasonal population trend of the pest, which could be helpful for preemptive management in areas where D. suzukii is prevalent and overlaps with susceptible crops. The authors stated that in Wisconsin tart cherry orchards, four years of adult monitoring trap data gathered during the growing season connected to degree days and relative humidity helped develop the GLMM and GAMM used in their work, which can accurately predict the seasonal phenology of D. suzukii based on its dynamic seasonality. The data collected from the D. suzukii adult trap catch, degree days, relative humidity, year, and site were used to create a generalized linear mixed model (GLMM). The authors used the glmer function in the lme4 package in R to maximum

likelihood to fit a negative binomial regression model. That model included year, location, and site interaction as random factors while degree and relative humidity are fixed effects. To keep the scale in line with the other factors, the authors multiplied the number of degree days by 1000. The authors showed an estimate of weekly adult D. suzukii trap catch as a function of nonlinear time degree days and relative humidity in a graph. The authors asserted that using the GLMM account for random effects before calculating the regression coefficients, as the coefficient estimate will only rise or decrease, isolating random from mixed-effects results in a more understandable model. Still, this method limits the model's adaptability. The GLMM is ideal for predicting D. suzukii trap counts at any given point in the field season using historical data. D. suzukii seasonal phenology is predicted by this model simply by adjusting degree days. Still, it can also forecast a specific trap catch given an individual degree days estimate. For important occurrences in the field, such as the first identification of an adult D. suzukii or a surge in activity, the authors stipulated that the GLMM pairs well with the GAMM during the 2015–2018 trap catch was tracked using scatter plots. The author creates a GAMM due to the identical D. suzukii adult trap catch and data from the degree days, relative humidity, and year and site. The predict.gam function in the lme4 and mgcv packages in R was used in the model, which is based on negative Poisson regression with a log-link. Relative humidity was introduced as a fixed variable, while the number of degree days, year, site interactions were included as random effects in the model. The authors used graphs to represent their models. They stated that the estimated population trends in D. suzukii using the GAMM throughout the growing season, there is less interpretability in the GAMM than the GLMM. Still, it is more flexible when attributing

specific degree days to crucial points in the seasonal phenology of D. suzukii's phenology, such as the first adult detection, above-average trap catches, and the peak of activity. Using the GAMM in conjunction with the GLMM, trap catches of D. suzukii throughout the field season may be predicted for any given data point distribution.

The author did an extensive amount of work to get the comparison between the GLMM and GAMM. However, it wasn't easy to connect the result as the graph in question did not come up in the article. The result shows that the GLMM produced a more favorable outcome than the GAMM because, with the GLMM, the trap count of D. Suzukii can be predicted at any time with historical data.

References List

Kamiyama, M. T., Bradford, B. Z., Groves, R. L., & Guédot, C. (2020). Degree day models to forecast the seasonal phenology of drosophila suzukii in tart cherry orchards in the midwest u.s. *PLOS ONE*, *15*(4), e0227726. https://doi.org/10.1371/journal.pone.0227726

Larose, C. D., Larose, D. T., & Larose, Chantal D., Author. (2019). *Data science using python and r*. John Wiley & Sons,inc,.