Homework 4: Data separation and multinomial regression

your name

Due: March 3rd at 11:59 PM

Problem 1: Do the following regarding the Sabermetrics dataset (bball.csv),

(a) Fit the `nnet` model and comment on the similarities and differences between the `nnet` and `VGAM` fits in the Sabermetrics example in the ordinal and multinomial regression notes. Report interesting conclusions using either implementation.

(b) Provide recommendations on how an aspiring baseball player should approach hitting. You may want to consider success metrics like hits where hits = 1B + 2B + 3B + HR, or weighted hits where weighted hits = 1B + 2×2B + 3×3B + 4×HR. Note these metrics are conditional on a ball being put into play in the context of this analysis.

Problem 2: Comment on the differences between the `vglm` and `polr` implementations in the happiness and trauma example.

Problem 3: A study of factors affecting alcohol consumption measures the response variable with the scale (abstinence, a drink a day or less, more than one drink a day). For a comparison of two groups while adjusting for relevant covariates, the researchers hypothesize that the two groups will have about the same prevalence of abstinence, but that one group will have a considerably higher proportion who have more than one drink a day. Even though the response variable is ordinal, explain why a cumulative logit model with proportional odds structure may be inappropriate for this study.

Problem 4: Refer to the table below:

		Belief in Heaven		
Race	Gender	Yes	Unsure	No
Black	Female	88	16	2
	Male	54	17	5
White	Female	397	141	24
	Male	235	189	39

(a) Fit the model

$$\log(\pi_j/\pi_3) = \alpha_j + \beta_j^G x_1 + \beta_j^R x_2, \qquad j = 1, 2.$$

(b) Find the prediction equation for $\log(\pi_1/\pi_2)$.

(c) Treating belief in heaven as ordinal fit and interpret a cumulative logit model and a cumulative probit model. Compare results and state interpretations in each case.

Problem 5: Suppose that you have a coin that when flipped has a probability $0 < p < 1$ of landing heads, and that we know nothing about p. Suppose that you flip the coin four times and all four flips resulted in heads. Derive the MLE of p and the MLE of $\text{Var}(Y_i)$ under the standard Bernoulli model. Now, for some error tolerance $0 < \alpha < 1$, derive a valid one-sided confidence interval for p making use of the statement $\mathbb{P}\left(\sum_{i=1}^{4} y_i = 4\right)$.

Problem 6: Complete the following with respect to the `endometrial` example:

(a) Write your own Fisher scoring algorithm for this example. Argue that $\hat{\beta}$ diverges in some sense as the iterations of your algorithm increase.

(b) Show that the log likelihood has an asymptote in $\|\beta\|$.

(c) Code the likelihood function for this dataset, pick a value of $\tilde{\beta}$ that is in the LCM, find an eigenvector of estimated Fisher information η such that the likelihood asymptotes, and then show that the likelihood asymptotes in $\tilde{\beta} + s\eta$ as $s \to \infty$.

(d) Explain why the likelihood asymptotes in $\tilde{\beta} + s\eta$ as $s \to \infty$.

Problem 7: Summarise the Firth approach mentioned in Section 7.4.7 and 7.4.8 of Agresti. Compare and contrast the Firth approach with the direct MLE approach outlined in the complete separation notes. What are the strengths and weaknesses of each approach?

Problem 8: Use `glmdr` software to analyze the `catrec.txt` data using Poisson regression. Specifically, fit a third order model and provide confidence intervals for all mean-value parameter estimates, both one-sided intervals for responses that are constrained on the boundary and two-sided intervals for responses that are unconstrained. Also verify that the third order model is appropriate.

Problem 9: Do you think that the `glm` function can be used to provide an appropriate value of the Akaike information criterion (AIC) when complete separation or quasi-complete separation exists? Why or why not?