STA 712 Homework 6

Due: Monday, November 7, 12:00pm (noon) on Canvas.

Instructions: Submit your work as a single PDF. For this assignment, you may include written work by scanning it and incorporating it into the PDF. Include all R code needed to reproduce your results in your submission.

Practice with the Negative Binomial

1. If $Y_i \sim NB(r, p)$, then Y_i takes values $y = 0, 1, 2, 3, ...$ with probabilities

$$P(Y_i = y) = \frac{\Gamma(y + r)}{\Gamma(y + 1)\Gamma(r)}(1 - p)^r p^y,$$

where $r > 0$ and $p \in [0, 1]$.

   (a) Show that if r is known, then the negative binomial is an EDM by identifying $\theta, \kappa(\theta)$, and ϕ.

   (b) Use (a) to show that $\mathbb{E}[Y] = \frac{pr}{(1-p)}$ and $Var(Y) = \frac{pr}{(1-p)^2}$.

   (c) Deduce the canonical link function for the negative binomial distribution (it is *not* the log link that we use in practice!).

   (d) Derive the unit deviance $d(y, \mu)$ for the negative binomial distribution (assuming r is known).

2. The negative binomial is one distribution we can use when there is more variability in the response variable than accounted for in a Poisson model. It turns out there is a very cool relationship between the Poisson and the negative binomial.

   Suppose that λ_i is treated as a random variable, instead of a fixed parameter, and that

$$Y_i|\lambda_i \sim Poisson(\lambda_i) \qquad \lambda_i \sim Gamma\left(\frac{1}{\psi}, \mu_i\psi\right).$$

   (This means that $\mathbb{E}[\lambda_i] = \mu_i$).

   Show that the marginal distribution of Y_i is

$$Y_i \sim NB\left(\frac{1}{\psi}, \frac{\mu_i}{\mu_i + 1/\psi}\right).$$

Data Analysis

3. In this question, you will model the number of purchases of different kinds of books on Amazon. We have a random sample of data from a particular book seller on how many books were purchase from their Amazon store in the last 30 days. Your report will be given to this (imaginary) bookseller who wants you to tell them what kinds of books they might want to stock, or not stock, in their Amazon store for next month. What characteristics might be related to a book that sells a lot of copies? One that sells very few? And so on.

   Note: You may assume there are no season buying patterns we need to be aware of (no holiday spending or extraordinary sales). In other words, you can assume there is nothing special about the month of data you have, nor the month you are predicting for, that would impact book buying habits.

   Your variables include:

   - title: The title of the Book
   - author: The author of the book
   - rating: An average score the book has received on Amazon.
   - purchases: The number of copies of the book purchased in the last 30 days.
   - price: The price of the book in US. Dollars.
   - publisher: The company that published the book.
   - page_count: The number of pages in the book.
   - ISBN: a unique numeric identifier for the book.
   - published_date: The date the book was published.
   - Year: the year in which the book was published
   - genre: the book's genre (Fiction, Fantasy, Mystery, Business, General Interest, Comics and Graphic Novels, or Other.

   You can load the data into R by

   ```
   books <- read.csv("https://sta712-f22.github.io/homework/books.csv")
   ```

   (a) Create empirical log means plots (see `https://sta712-f22.github.io/class_activities/ca_lecture_22.html`) to explore the relationships between quantitative explanatory variables and the number of purchases. Do you think any transformations are necessary?

   (b) Do you think we need to include an offset in our model? If so, what would the offset be?

   (c) Using your exploratory data analysis, fit a Poisson regression model for the number of purchases in the last 30 days.

   (d) Perform model diagnostics:

   - Create quantile residual plots to check the shape assumption for quantitative variables (you may use the `qresid` function in the `statmod` package)
   - Calculate Cook's distance to check for any influential points (use a threshold of 0.5 or 1 to identify influential points)
   - Calculate variance inflation factors to check for multicollinearity (see the `vif` function in the `car` package, and use a threshold of 5 or 10 to identify high multicollinearity).

(e) Perform a χ^2 goodness of fit test for your fitted Poisson regression model. Do you think there might be overdispersion in the data?

(f) Fit a quasi-Poisson model instead, and reported the estimated dispersion $\widehat{\phi}$.

(g) Now fit a negative binomial model, and report the estimate $\widehat{r}$.

(h) As in (d), perform model diagnostics for your negative binomial model. Do you think the quasi-Poisson or the negative binomial does a better job at modeling the mean-variance relationship? (Or do they look equivalent?)

(i) Using either your quasi-Poisson or negative binomial model, carry out at least one hypothesis test to address the client's question: which book characteristics are related to the number of copies sold?

(Optional) Practice with EDMs

4. Determine which of these functions are suitable link functions for a GLM. For those that are not suitable, explain why not. (*Hint: what is true about the relationship between the mean μ and the canonical parameter θ in an EDM?*)

   (a) $g(\mu) = -1/\mu^2$ when $\mu > 0$

   (b) $g(\mu) = |\mu|$ when $-\infty < \mu < \infty$

   (c) $g(\mu) = \log(\mu)$ when $\mu > 0$

   (d) $g(\mu) = \mu^2$ when $-\infty < \mu < \infty$

   (e) $g(\mu) = \mu^2$ when $0 < \mu < \infty$

5. Suppose we are told that $Y \sim EDM(\mu, \phi)$, and we know that $V(\mu) = \mu + \dfrac{\mu^2}{k}$, where k is known. In this problem, we will work backwards from $V(\mu)$ to show that Y must follow a negative binomial distribution. *Note: some of the integrals in this question are quite tricky! You are welcome to use software, like WolframAlpha or Mathematica, to help with the integration.*

   (a) Using the fact that $V(\mu) = \partial\mu/\partial\theta$, find θ as a function of μ.

   (b) Using the fact that $\mu = \partial\kappa(\theta)/\partial\theta$, find $\kappa(\theta)$.

   (c) Conclude, using Question 1, that Y must follow a negative binomial distribution.