STA 711 Homework 4

Due: Friday, February 10, 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.

(Randomized) quantile residuals

1. In class, we talked about (randomized) quantile residuals as a method of assessing the shape assumption in logistic regression. To formally define quantile residuals, we will follow Dunn and Smyth (Section 8.3.4.2).

   Suppose we have a logistic regression model:

   $$Y_i \sim Bernoulli(p_i)$$
   $$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_k X_{i,k}.$$

   We observe data $(X_1, Y_1), ..., (X_n, Y_n)$ and fit the model, producing coefficient estimates $\widehat{\beta}$ which give estimated probabilities $\widehat{p}_i$. The *(randomized) quantile residual* $r_{Q,i}$ for the ith observation is defined by

   $$r_{Q,i} = \Phi^{-1}(u), \qquad u \sim \begin{cases} Uniform(1-\widehat{p}_i, 1) & Y_i = 1 \\ Uniform(0, 1-\widehat{p}_i) & Y_i = 0, \end{cases}$$

   where Φ is the standard normal CDF.

   (a) Show that if $\widehat{p}_i = p_i$ (our estimated probability is correct), then $r_{Q,i} \sim N(0,1)$. *Hint: treat the response Y_i as a random variable, and note that $Y_i \sim Bernoulli(\widehat{p}_i)$ if $p_i = \widehat{p}_i$.*

   (b) Show that $\mathbb{E}[r_{Q,i}] > 0$ when $\widehat{p}_i < p_i$, and $\mathbb{E}[r_{Q,i}] < 0$ when $\widehat{p}_i > p_i$.

   (c) Write your own function in R to compute randomized quantile residuals for a binary logistic regression model. (Your function may not call the `qresid` function from the `statmod` package).

   (d) Generate data for which the logistic regression shape assumption is satisfied. Then create a quantile residual plot using your R function, and show that the residuals $r_{Q,i}$ are randomly scattered around the horizontal line at 0.

   (e) Generate data for which the logistic regression shape assumption is *not* satisfied. Then create a quantile residual plot using your R function, and show that the plot shows a violation of the shape assumption.

Looking ahead: multivariate normal distributions

The multivariate normal distribution will appear frequently in 711, for example as the asymptotic distribution of our coefficient estimates $\widehat{\beta}$. The purpose of this section is to derive a basic property of the multivariate normal distribution that we use regularly, for example in constructing our Wald

test statistic.

One way to define a multivariate normal distribution is with its *moment generating function* (MGF). Let $X \in \mathbb{R}^k$ be a random vector. The (multivariate) moment generating function $M_X(t)$ of X is defined by

$$M_X(t) = \mathbb{E}[e^{t^T X}],$$

where $t \in \mathbb{R}^k$. As with univariate MGFs, if $M_X(t) = M_Y(t)$ for all t, then the two random variables X and Y have the same distribution.

We say that the random vector $X \in \mathbb{R}^k$ follows a multivariate normal distribution with mean $\mu \in \mathbb{R}^k$ and variance matrix $\Sigma \in \mathbb{R}^{k \times k}$, and write $X \sim N(\mu, \Sigma)$, if

$$M_X(t) = e^{t^T \mu} e^{\frac{1}{2} t^T \Sigma t}.$$

2. An important property of multivariate normal random variables is that if $X \sim N(\mu, \Sigma)$, then

$$\boldsymbol{a} + \boldsymbol{B}X \sim N(\boldsymbol{a} + \boldsymbol{B}\mu, \boldsymbol{B}\Sigma\boldsymbol{B}^T),$$

where $\boldsymbol{a} \in \mathbb{R}^m$ and $\boldsymbol{B} \in \mathbb{R}^{m \times k}$. Our goal is to use MGFs to prove this property.

(a) Show that for any random vector X in $\mathbb{R}^k$, the MGF of $Y = \boldsymbol{a} + \boldsymbol{B}X$ is given by

$$M_Y(t) = e^{t^T \boldsymbol{a}} M_X(\boldsymbol{B}^T t).$$

(b) Using (a), show that if $X \sim N(\mu, \Sigma)$, then $\boldsymbol{a} + \boldsymbol{B}X \sim N(\boldsymbol{a} + \boldsymbol{B}\mu, \boldsymbol{B}\Sigma\boldsymbol{B}^T)$.

Data analysis

Here we work with data from a website called ScienceForums.Net (SFN), which has been open since 2002 and hosts conversations on a range of topics from biological and physical science to religion and philosophy. Each row in the data represents one 'thread', which is comprised of a series of posts stemming from an initial post. For each thread, we have some information that SFN collects such as the number of views and the number of authors. The threads present in the data are a random sample of threads from 2002-2014, with the data collected in 2014. SFN moderators are interested in using this data to determine which threads warrant the most attention.

You can load the SFN data into R by

```
sfn <- read.csv("https://sta711-s23.github.io/homework/sfn.csv")
```

The sfn dataset contains the following columns:

- Age: the age of the thread (in days) when the data was collected in 2014, measured from the first post in the thread
- State: sometimes moderators close threads if they are inappropriate. closed indicates the thread has been closed, otherwise State is open
- Posts: the number of posts in the thread
- Views: the total number of views of the thread
- Duration: the number of days between the first and last posts in the thread

- Authors: the number of distinct authors posting in the thread
- AuthorExperience: the number of days the author of the first post in the thread had been registered on SFN when the thread began (0 indicates they registered that day)
- DeletedPosts: the number of posts in the thread that have been deleted by a moderator
- Forum: the forum in which the thread was posted (e.g., Science)
- AuthorBanned: whether the original author of the thread is currently banned from posting on SFN (at the time of data collection, not when the thread was first posted)

Research question: Suppose you have been approached by moderators at SFN. They give you the data, and ask the following question:

- Is there a relationship between the number of Posts in a thread and whether a thread will have *at least one* deleted post, after accounting for the number of Views, the number of Authors, and the Forum?

3. Here you will use logistic regression to answer the moderators' question.

   (a) Which variables should we focus on to answer the moderators' question? Which of these is our response variable, and which will be our explanatory variables, for logistic regression?

   (b) Perform univariate exploratory data analysis (EDA) for your selected variables in (a):
   - For categorical variables, present a table showing the number of observations in each category
   - For quantitative variables, present a histogram and summarize the distribution of the variable (give summary statistics and describe center, shape, spread, and any potential outliers)
   - Discuss whether there are any missing or erroneous values in the data, and if so how you will handle them

   (c) Perform multivariate EDA for your selected variables in (a):
   - Create empirical logit plots to summarize the relationship between quantitative predictors and your binary response. Details on creating empirical logit plots, with examples, can be found at `https://sta711-s23.github.io/homework/empirical_logits.html`
   - Using the empirical logit plots, discuss whether any transformations are needed on the explanatory variables.
   - Use a correlation matrix to summarize pairwise relationships between the quantitative explanatory variables. Should we be concerned with potential multicollinearity?

   (d) Based on your exploratory data analysis, write down a logistic regression model that will allow you to answer the moderators' question. Describe how you will use the model to answer their question.

   (e) Fit your model from (d), and report the equation of the fitted model. Interpret any estimated coefficients which address the moderators' question.

   (f) Assess your model assumptions:
   - 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).

(g) Address any violations to the model assumptions (transformations for shape violations; report results with and without influential points; and combine or remove columns for high multicollinearity). If you made any changes to your model from (e), report and interpret your new fitted model here.

(h) Carry out a hypothesis test to investigate the moderators' question. You should:

- State the null and alternative hypotheses in terms of one or more βs
- Calculate a test statistic and p-value
- Make a conclusion in the context of the original question