STA 711 Homework 9

Due: Tuesday, April 30, 11am 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

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.

1. 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 $p_i = \widehat{p}_i$.*
2. Show that $\mathbb{E}[r_{Q,i}] > 0$ when $\widehat{p}_i < p_i$, and $\mathbb{E}[r_{Q,i}] < 0$ when $\widehat{p}_i > p_i$.
3. 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). Demonstrate your function works by creating an example quantile residual plot (it may help to modify the code from the class activity on April 10).

Putting it all together: predicting artist from Spotify data

Now that we're at the end of the semester, let's look back on some of the topics we have discussed with logistic regression models, and start to look ahead to future courses. In this assignment, you will work with data from Spotify containing 138 songs total, from two different artists: Manchester Orchestra and The Front Bottoms. Each row represents one song, and data has been extracted from Spotify that shows information about the musical qualities of each song (danceability, energy, key, etc).

Research questions: We are interested in whether we can determine the artist of a song using information about the musical qualities of the song. To investigate, we will look at two related questions: the first is an *inference/association* question which asks about the presence of a relationship, while the second is a *prediction* question that asks about the quality of our predictions.

- Do the different artists (Manchester Orchestra and the Front Bottoms) have different styles of music? That is, is there a relationship between artist and variables like speechiness, energy, etc.?
- How well can we distinguish between the two artists in the data, using available variables?

Getting the data: The data can be loaded into R with the following:

```
spotify <- read.csv("https://sta711-s24.github.io/homework/spotify.csv")
```

EDA

To begin, let's explore some of the available variables

3. The response variable of interest is artist. Create a table or figure summarizing the distribution of artist in the available data.
4. Explain why, if we are trying to predict the artist for a song, Album is *not* a valid explanatory variable to include in our model.
5. There are two categorical explanatory variables in the Spotify data: key and mode. Recall that when using categorical variables in a model, it is important that we have enough observations at each level of the variable (otherwise, our coefficient estimates will be based on only a few observations!).
    (a) Create two-way tables showing the relationship between each of these variables and artist.
    (b) Which of these variables could be included in a model to predict artist?
6. Now explore the univariate distributions of the quantitative explanatory variables (loudness, energy, etc.) with histograms or density plots.
    (a) Show a couple example plots and describe the distributions (you do not need to include every plot in your homework submission).
    (b) As in linear regression, we may consider transforming (e.g. with a log, square root, etc.) explanatory variables which are highly skewed. Do you think any of the quantitative explanatory variables in the Spotify data warrant a transformation?
7. As in a linear regression model, multicollinearity between the variables is a potential concern. In EDA, we can explore potential multicollinearity by creating a correlation matrix for the quantitative explanatory variables.
    (a) Create a correlation matrix; are there any pairs of variables which are highly correlated (correlation around 0.80 or more)? Are these variables you would expect to be correlated?
    (b) To handle multicollinearity, we can omit some of the highly correlated variables. If you choose to omit any variables, list them here.
8. Finally, let's examine the relationship between some of the quantitative explanatory variables and the binary response, using empirical logit plots. (Note: I would view empirical logit plots as an EDA tool to use *before* building a model, and quantile residual plots as a diagnostic tool to use *after* a model is built). The following link contains code and instructions on creating an empirical logit plot in R:

    `https://sta214-s23.github.io/resources/codebook.html#empirical-logit-plots`

(a) Show a couple example plots and describe the relationships (you do not need to include every plot in your homework submission).

(b) Are there any quantitative explanatory variables for which a transformation may be needed to satisfy the shape assumption?

Modeling and diagnostics

After our initial EDA, we are ready to build a model and check diagnostics.

9. Using your EDA from above, fit a logistic regression model to predict the artist of each song using the musical qualities of the song. Provide a table of the estimated coefficients for the fitted model, and interpret some of the results.

10. Now let's assess the shape assumption.

    (a) Assess the shape assumption for your fitted model by plotting quantile residuals (y-axis) against the fitted values (x-axis). You may use the `qresid` function from the `statmod` package, or your own function from the first part of the assignment.

    (b) Discuss your plot: does the shape assumption appear reasonable?

    (c) If the shape assumption does not appear reasonable, investigate further and propose some transformations to address violations. Then re-fit the model and report your new model here.

11. Next, check for potential multicollinearity using variance inflation factors (VIFs). The `vif` function in the `car` package is a good option. As a rule of thumb, we are concerned if a variance inflation factor is above 5, and we may consider removing that variable. If you make any changes to the model because of multicollinearity concerns, report the changes and the updated model here.

12. Finally, use Cook's distance to check for any influential points. If you identify any influential points, report the results in the next sections with and without those influential points.

Inference

Now that we have a reasonable model and we have checked the model assumptions, we can address the first research question.

13. Carry out a hypothesis test to address the first research question. Clearly state the hypotheses you are testing (i.e., the full and reduced models you are comparing), the test statistic, and the p-value. Then make a conclusion in context of the first research question.

Prediction

Finally, let's begin to assess the predictive ability of our model. You will see more about this in STA 712 (and you may already have seen it in STA 663), but the following questions will give you a brief introduction.

14. To assess the predictive ability of our logistic regression model, we need to convert the predicted probabilities $\widehat{p}_i \in [0,1]$ into binary predictions $\widehat{Y}_i \in \{0,1\}$. Typically, we do this by thresholding the probabilities, and a threshold of 0.5 is common:

$$\widehat{Y}_i = \begin{cases} 1 & \widehat{p}_i > 0.5 \\ 0 & \widehat{p}_i \leq 0.5 \end{cases}$$

(a) Create binary predictions from your fitted model for each song in the dataset.

(b) Make a 2×2 table comparing the binary predictions with the true song labels (Manchester Orchestra or the Front Bottoms). This is called a *confusion matrix.*

(c) To summarize performance, we can calculate several statistics from the confusion matrix. To define these, consider the following confusion matrix:

	$Y = 0$	$Y = 1$
$\widehat{Y} = 0$	TN (true negatives)	FN (false negatives)
$\widehat{Y} = 1$	FP (false positives)	TP (true positives)

- $Accuracy = \frac{TN+TP}{TN+TP+FN+FP}$
- $Sensitivity = \frac{TP}{TP+FN}$
- $Specificity = \frac{TN}{TN+FP}$

Calculate accuracy, sensitivity, and specificity for your fitted model. How well is your model performing? (*Note:* we should really assess performance on a held-out test dataset, and other performance metrics are available, but those are topics for another course)