Homework 4

Due Monday 10/25

Turn in R code and output – try to use `R markdown`.

Dexiu Ma, Nuo Ma and Chen Tao will lead HW discussion.

1. (60 points) *Modeling COVID-19 Low Transmission Levels*

   The U.S. Center for Disease Prevention and Control (CDC) has provides guidance on prevention strategies for safe in-person learning at K-12 schools. The implementation of these prevention measures depend on risk factors, such as level of community transmission. For example, screening testing is recommended for K-12 students if community transmission is not *low*: see Table 1 on the website `https://www.cdc.gov/coronavirus/2019-ncov/community/schools-childcare/k-12-guidance.html`. The CDC defines *low* transmission as having the number of new cases per 100,000 persons within the last 7 days less than 10.

   In this question you will carry out a logistic regression analysis modeling *low* COVID-19 transmission status. This question uses the same predictors you already "got to know" in HW2 Q2 and used for linear regression in HW3 Q1, but now with a new response variable. The binary response variable is the indicator of whether the 7-day COVID-19 case count in the county is less than 10 per 100,000 people:

$$y = \mathrm{I}\left[100,000 \times \left(\frac{covid\_count\_sep17 - covid\_count\_sep10}{Tot\_Population\_ACS\_14\_18}\right) < \ 10\right]$$

   where the denominator is the total population of the county estimated from the 5-year (2014-2018) American Community Survey from the U.S. Census Bureau. County-level COVID-19 case counts (as of 9/10/21 and 9/17/21) are obtained from `https://usafacts.org/visualizations/coronavirus-covid-19-spread-map/`; and geographic and demographic characteristics of each county are obtained from the Census Bureau's 2020 Planning Database (PDB). The dataset for this HW assignment *covid_data_pdb_2.csv* is posted on Blackboard under HW4. Complete the following:

   (a) Perform EDA to understand the binary response variable and its association with each predictor;

   (b) Fit a logistic regression model with all the predictors listed in HW2 and HW3. Assess goodness-of-fit via deviance comparisons – make sure to state null/alternative hypothesis, test statistic, p-value and conclusion;

   (c) Refine your model via variable selection (e.g. stepwise selection, check multicollinearity). Compare your reduced model to the model fit with all predictors via deviance comparisons – make sure to state null/alternative hypothesis, test statistic, p-value and conclusion;

   (d) Consider interactions (e.g. pairwise between continuous variables, and continuous variables interacted with `division`). Select a final logistic regression model (include interactions or not?) based on deviance tests;

(e) Check for influential points and explain your findings;

(f) Provide an interpretation for one of the predictors that is statistically significant at $\alpha = .05$ and calculate the 95% Wald confidence interval for that predictor effect;

(g) Investigate binary predictions using these step-by-step instructions comparing two cutoffs/thresholds:

i. Use 0.5 as a cutoff. This corresponds to the prediction rule:

$$\hat{y}_i = \begin{cases} 1 \text{ if } \hat{\pi}(x_i) > .5 \\ 0 \text{ otherwise} \end{cases}$$

Provide a cross-tabulation of y and $\hat{y}$ using the `table` command in R. How many of the observations have incorrect predictions (i.e. the predicted value is not equal to the observed value)?

ii. Use the observed proportion of *low* transmission as a cutoff. This corresponds to the prediction rule:

$$\hat{y}_i = \begin{cases} 1 \text{ if } \hat{\pi}(x_i) > \frac{\sum_{i=1}^n y_i}{n} \\ 0 \text{ otherwise} \end{cases}$$

Provide a cross-tabulation of y and $\hat{y}$ using the `table` command in R. How many of the observations have incorrect predictions (i.e. the predicted value is not equal to the observed value)?

iii. Comparing the proportion of incorrect predictions is one way determine the cutoff. For the two rules used above, calculate the proportion of incorrect predictions. Based on the proportion of incorrect predictions, which of the two cutoffs would you choose to use?

iv. Calculate the false positive rate (FPR) and false negative rate (FNR) for each of the two cutoffs, where:

$$FPR = P(\hat{y} = 1 | y = 0)$$
$$FNR = P(\hat{y} = 0 | y = 1)$$

Note that probabilities can be estimated by proportions observed in the data in this case.

v. Based on the false positive and false negative rates, which of the two cutoffs would you choose to use to define binary predictions?

vi. Plot the ROC curve for your final model. Describe the predictive power of the model based on the ROC plot. Mark the points on the curve that correspond to the chosen cutoff π_0 in part (i) and (ii) (i.e. .5 and $\overline{y}$). *You can write your own function or use an R package for ROC curves.*

2. (40 points) *Modeling COVID-19 Transmission Levels*

Using the same data as Q1, now consider modeling the four-level color-coded COVID-19 transmission categorical variable. You will need to create this response variable by assigning each county's value

$$100,000 \times \left(\frac{covid\_count\_sep17 - covid\_count\_sep10}{Tot\_Population\_ACS\_14\_18} \right)$$

into one of the four categories: blue/low (< 10), yellow/moderate ($10 - 49$), orange/substantial ($50 - 99$), and red/high (≥ 100).

(a) Perform EDA to understand this categorical response variable and its association with the predictors;

(b) Fit a multinomial logistic regression model (baseline category logit) with all the predictors listed in HW2 and HW3. What baseline category did you use and why?

(c) Assess goodness-of-fit via deviance comparisons – make sure to state null/alternative hypothesis, test statistic, p-value and conclusion;

(d) Refine your model via variable selection (e.g. stepwise selection – see `step4vglm` in `VGAM` package, check multicollinearity). Compare your reduced model to the model fit with all predictors via deviance comparisons – make sure to state null/alternative hypothesis, test statistic, p-value and conclusion;

(e) Provide an interpretation for one of the predictors that is statistically significant at $\alpha = .05$ – make sure to determine statistical significance by the deviance "joint" test of the predictor across all category levels;

(f) Discrete categorical predictions are usually defined as the category that has the largest predicted probability for an observation. That is,

$$\hat{y}_i = j \text{ for which } \hat{\pi}_{ij} = \max(\hat{\pi}_{i1}, \ldots, \hat{\pi}_{iJ}).$$

Calculate categorical predictions for each county and provide a cross-tabulation of y and $\hat{y}$. What proportion of the observations have incorrect predictions?

(g) Calculate false positive rate and false negative rate within each category, $P(\hat{y}_i = j | y_i \neq j)$ and $P(\hat{y}_i \neq j | y_i = j)$, respectively. Compare these predictive accuracy measures across categories and comment on any interesting findings.

(h) Fit the cumulative logit (a.k.a. proportional odds) model to this categorical COVID-19 transmission level response using only the predictors from your "final" baseline category logit model from part (d). Compare the number of regression parameters $\boldsymbol{\beta}$ in this model to the number of regression parameters in the baseline category logit model. Explain the difference.

(i) Discrete categorical predictions for the cumulative logit model are usually defined in the same way as the baseline category logit noted in part (f). Calculate categorical predictions for each county and provide a cross-tabulation of y and $\hat{y}$. What proportion of the observations have incorrect predictions?

(j) Compare the correct classification rate and FPR/FNR for cumulative logit model to those calculated from the baseline category logit model. Based on these predictive accuracy measures, which model would you choose? Why do you think your chosen model fits better?

3. (0 points) *Group Request for Final Project*

If you have a group preference for the final data analysis project, please submit the list of students that you would like in your group in the ”comments” section when you submit your HW on Blackboard. Only one member of each group needs to submit the request. Groups will be two to three students. If you submit a group of two, it is possible you will be randomly assigned another group member depending on how the numbers work out. If you do not have a group preference and would like to be randomly assigned, then you do not need to submit anything.