MLDS 401/IEMS404: Homework 7
Due: November 30, 15:00
Professor Malthouse

1. Use the estimates from the toxicity problem. Generate an ROC curve and find the area under the curve. You have summarized data and I would like for you to generate the ROC curve "by hand." Hint: there are $g = 6$ values of $x = 1, \ldots, 6$. Let $\hat{p}_x$ be the predicted probability for x using the logistic regression model.

   (a) Complete the following table, showing work:

   (b) Plot TPR against FPR and find the area assuming a trapezoid between successive values.

Cut value	TPR	FPR
$0 \le c < \hat{p}_1$		
$\hat{p}_1 \le c < \hat{p}_2$		
$\hat{p}_2 \le c < \hat{p}_3$		
$\hat{p}_3 \le c < \hat{p}_4$		
$\hat{p}_4 \le c < \hat{p}_5$		
$\hat{p}_5 \le c < \hat{p}_6$		
$\hat{p}_6 \le c \le 1$		

2. Suppose we have a sample of size n where observation i consists of dependent variable Y_i, a multinomial RV taking values $\{1, \ldots, K\}$, and $(p+1)$-vector of predictor variables $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{ip})^\mathsf{T}$. Let $\boldsymbol{\alpha}_k$ be a $(p+1)$-vector of regression coefficients. Let $\pi_{ik} = \mathsf{P}(Y_i = k)$ for $k = 1, \ldots, K$ and

$$\log \pi_{ik} = \boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i - \log Z, \qquad (k = 1, \ldots, K)$$

   where log is the natural log function and the term $\log Z$ ensures that the probabilities sum to one, i.e., $\sum_{k=1}^K \pi_{ik} = 1$.

   (a) Show that $Z = \sum_{k=1}^K \exp(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i)$.

   (b) Show that $\pi_{ik} = \exp(\boldsymbol{\alpha}_k^\mathsf{T} \mathbf{x}_i)/Z$. This is called the softmax function.

   (c) The usual formulation of the multinomial logit from class picks a base category (WLOG class 1) and assumes:

$$\log\left(\frac{\pi_{ik}}{\pi_{i1}}\right) = \boldsymbol{\beta}_k^\mathsf{T} \mathbf{x}_i, \qquad (k = 2, \ldots, K)$$

   How is $\boldsymbol{\beta}_k$ related to $\boldsymbol{\alpha}_k$? You will see that multinomial and softmax are just reparameterizations of each other.

3. This problem studies news deserts. You have data for (nearly) every county in the US:

   - `numPub23`: number of newspapers published for the county in 2023. This count is the **dependent variable**.

- `numPub18`: number of newspapers published for the county in 2018. With a five-year period this is a lagged version of the dependent variable.
- `age`: average age in county in 2021
- `SES21`: socioeconomic status (average of income and education)
- `Lpopdens2021`: population density of the county in 2021
- `Lblack2021`: percent of county that is black in 2021
- `Lhisp2021`: percent of county that is Hispanic in 2021

The goal is to build a predictive model forecasting which counties are likely to be news deserts in five years. We will consider two models

- **demographic** use `age, SES21, Lpopdens2021, Lblack2021 and Lhisp2021` as predictors
- **AR1+** Use the `log(numPub18+1)` and the demographics as predictors.

(a) Create a new variable `atrisk` that equals 1 if the county is at risk (`numPub23` ≤ 1) and 0 otherwise.

(b) Use a logistic regression to predict `atrisk` from the demographics only. Which variables increase the probability of being at risk? Which decrease the probability?

(c) Use a logistic regression to predict `atrisk` from the AR1+ variables. Interpret the model. How do you explain the difference in significant variables?

(d) Create an ROC curve showing the predicted values from the two models on the same plot. Find AUC for each of the two models.

4. This problem also uses the news desert data.

   (a) Create a variable `pub3.2023` that takes three values: 0 newspapers, 1 newspaper, or 2+ newspapers. Submit a frequency distribution (`table`).

   (b) Use a multinomial regression to predict `pub3.2023` from the demographics. Find the missing logit. Interpret all three logits (0 vs. 1, 1 vs. 2+ and 0 vs. 2+).

   (c) Use a multinomial regression to predict `atrisk` from the AR1+ variables. Find the missing logit and interpret the model.

   (d) For your two models, find accuracy; per-class precision, recall and F_1; and macro precision, recall and F_1. What do you conclude about which classes can be easily distinguished versus those that are more difficult to predict?

5. Return to problem 4 from homework 5 using data from the German book company.

   (a) You estimated a model in part d using the logs of tof, r, f and m+1, and in part e you applied it to the test set. Compute a gains table using the test-set data.

(b) How much money do you expect to make per customer if you used this model to select 40% of the names to be contacted?

(c) What fraction of customers will respond if you use this model to select 40% of the names?

(d) The next two parts estimate a two-step model using the training data only. This part estimates the **response model**. Create a variable `buy` that equals 1 if the customer bought (i.e., `target`> 0). Estimate a logistic regression predicting `buy` from any variables you wish. This estimates **conversion probabilities**, $\hat{\pi}_i$. What variables are predictive in this model?

(e) Now estimate a **conditional demand model** using the training data only. To do so, regress `logtarg` on some predictor variables using only buyers in the training set. This estimates the log spending amount of buyers, $\hat{y}_i$. What variables are predictive?

(f) Apply the response and conditional demand models to the test set and multiply $\hat{\pi}_1 e^{\hat{y}_i}$ and use this score to create a gains table. Which model is better @40%? The one from homework 5 or the twostep?