MATH-661: Generalized Linear Models
Midterm Exam
Due Tuesday March 27, 2018

INSTRUCTIONS

No collaboration or discussion is permitted on the midterm exam. If you need clarifications, you can contact me but you are not allowed to ask anyone else. Please note that clarifications are limited to ambiguities in the wording of questions. This exam is intended to demonstrate your grasp of the material, so no help will be provided.

Please fill your name and sign the following honor pledge:
I, _____, pledge that I have not violated the Georgetown University honor code (see http://gervaseprograms.georgetown.edu/honor/). The work I am submitting for this exam is completely my own. I have not communicated with anyone and have not allowed any other student to use or borrow portions of my work. I understand that if I violate this honesty pledge, I will be reported for academic dishonesty to the Honor Council.
Signature : _____

- Show the details of your work in order to get full credit for correct answers, and partial credit for incorrect answers if you are on the right track.
- Provide interpretations and conclusions in the context of the problem.
- Include the relevant R code and output for each question, when applicable.
- The exam must be e-mailed to `mgt26@georgetown.edu` by 11:59 pm on Tuesday, March 27, 2018.

Part I: Aspirin and heart attack [20 points]

A study investigating the association between heart attacks and the use of aspirin is conducted. Age is a potential confounder and is also considered. The following indicator variables are defined:

$$Y = \begin{cases} 1 & \text{if heart attack} \\ 0 & \text{if no heart attack} \end{cases} \qquad \text{Aspirin} = \begin{cases} 1 & \text{if aspirin} \\ 0 & \text{if placebo} \end{cases}$$

$$\text{Age1} = \begin{cases} 1 & \text{if age is } 40-50 \\ 0 & \text{otherwise} \end{cases} \qquad \text{Age2} = \begin{cases} 1 & \text{if age is } > 50 \\ 0 & \text{otherwise} \end{cases}$$

The following table shows the results of fitting logistic regression models for $P(Y = 1)$:

Model	Covariates	Estimate $\hat{\beta}$	Standard Error	log-likelihood
1	None	-2.99	0.19	-116.54
2	Aspirin	-0.82	0.41	-114.41
3	Age1	-0.19	0.47	-116.27
	Age2	0.17	0.45	
4	Aspirin	-0.82	0.41	-114.14
	Age1	-0.18	0.47	
	Age2	0.19	0.45	
5	Aspirin	-0.65	0.63	-113.83
	Age1	-0.22	0.59	
	Age2	0.39	0.54	
	(Age1)*Aspirin	0.10	0.97	
	(Age2)*Aspirin	-0.68	1.03	

(a) Test the null hypothesis of constant aspirin effect on the risk of heart attack across age groups (i.e., no interaction between aspirin and age).

(b) Based on the model with additive/main effects for age and aspirin (Model 4)

i. Calculate the MLE of the odds ratio of aspirin use on heart attack, adjusting for age. Provide a 95% confidence interval for this odds ratio and interpret it in context.

ii. Perform a Wald test of the null hypothesis that there is no effect of aspirin on the risk of heart attack, controlling for age. What do you conclude?

iii. Perform a likelihood ratio test of the null hypothesis that there is no effect of age on the risk of heart attack, controlling for aspirin use. State your conclusion.

(c) Evaluate the deviance of each model provided in the table and assess its goodness-of fit. Which models do not provide adequate fit to the data?

(d) Perform model selection using analysis-of-deviance. Make sure you describe all the steps to arrive at your final model.

Part II: Credit risks for bank loan [30 points]

Banks want to reduce the rate of loan defaults. Loan officers want to be able to identify characteristics that are indicative of people who are likely to default on loans, and then use those characteristics to identify good and bad credit risks.

Financial and demographic information are collected on 850 past and prospective customers. Of these, 700 are customers who were previously given loans and 150 are prospective customers that the bank needs to classify as good of bad credit risks. The data are saved in `Bank_loan.txt` and contain the following variables:

age	age in years
ed	highest level of education
	1: did not complete high school; 2: high school degree
	3: some college; 4: college degree; 5: post-bachelor degree
employ	years with current employer
address	years at current address
income	household income in thousands
debtinc	debt to income ratio ($\times 100$)
creddebt	credit card debt in thousands
othdebt	other debt in thousands
default	previously defaulted – 0: No, 1: Yes, NA: prospective customers

1. **Exploratory data analysis & data processing**

   (a) Provide appropriate summary statistics and graphical displays for the variables in the data. Discuss their distributions.

   (b) Since there are few observations with post-bachelor degree (`ed` $= 5$), combine these with the group with college degree (`ed` $= 4$). You will be using education with these 4 levels in subsequent analyses.

   (c) Separate the 150 prospective customers for whom credit risk is to be predicted from the 700 past customers.

2. **Model building & diagnostics** – use the 700 past customers for this task.

   (a) Perform stepwise selection.

      i. Provide the equation of the selected model.

      ii. Interpret the effect of each of the covariates in the selected model.

      iii. Assess the goodness-of-fit of the selected model.

   (b) Perform a lasso variable selection using the misclassification error as criterion for choosing λ.

      i. Compare the models selected using `lambda.1se` to the stepwise selected model in (a). Perform a likelihood ratio test to choose the preferred model between the two at $\alpha = 0.05$.

   (c) For the model selected based on lasso

      i. Identify observations with unusual/outlying standardized residuals. How do the predictions for these individuals, based on their fitted values $\hat{\pi}_i$, compare to their observed default status?

      ii. Using a cut-off of 0.3 for predicting whether a person defaults or not on a loan

         - what proportion of the 700 customers would have been predicted as defaulting (and thus would have been denied a loan)?
         - what would be the misclassification rate?
         - what would be the misclassification rate among the defaulters?
         - what would be the misclassification rate among the non-defaulters?

      iii. Provide the ROC curve and the area under the ROC curve for the selected model.

3. **Prediction for future customers** – consider the model selected based on lasso.

   (a) Calculate the predicted probabilities of loan default for the 150 prospective customers.

   (b) Provide a histogram and a boxplot of the predicted probabilities.

   (c) Using a cut-off of 0.3, how many of the 150 prospective customers would be expected to default on a loan?