MSiA 401/IEMS404: Homework 6
Due: Nov 13, 3:00pm
Professor Malthouse

1. This is a variation of problem 11 on page 264 of JWHT (section 6.8). Hint: see the college problem that we did in class. You will build various predictive models for the Boston data set.

   (a) Load the data and create a training/test split as follows. Submit a frequency distribution of the `train` variable.

   ```
   library(MASS)
   dim(Boston)
   Boston$logcrim = log(Boston$crim) # create log transform of crim
   summary(Boston)
   set.seed(12345)
   train = runif(nrow(Boston))<.5    # pick train/test split
   ```

   (b) Regress `logcrim` on all variables (except for `crim`) using only the training data. Apply the model to the test set and report the test set MSE. Examine the residual plot and comment.

   (c) Apply backward selection to the model from the previous part. Report test set MSE.

   (d) Fit a ridge regression using `cv.glmnet` to choose the optimal λ value. Report test set MSE.

   (e) Fit a lasso regression using `cv.glmnet` to pick λ. Report test set MSE.

   (f) Add transformations to improve your model (work hard on this). Report test set MSE for stepwise, ridge and lasso. Which transformations are important, by coming into the stepwise and/or lasso models?

2. In an experiment testing the effect of a toxic substance, 1,500 experimental insects were divided at random into six groups of 250 each. The insects in each group were exposed to a fixed dose of the toxic substance. A day later, each insect was observed. Death from exposure was scored 1 and survival was scored 0. The results are in the data frame below, where x_j is the dose level (on a logarithmic scale), administered to the insects in group j and y_j denotes the number of insects that dies out of the $n_j = 250$ in the group. As a hint, study the bottle return problem we worked in class. Each student should submit answers on Canvas.

   ```
   toxicity = data.frame(x=1:6, n=rep(250,6), s=c(28,53,93,126,172,197))
   fit = glm(s/n ~ x, binomial, toxicity, weight=n)
   ```

(a) Plot the estimated proportions $f_j = s_j/n_j$ against x_j. Does the plot support the analyst's belief that the logistic response function is appropriate?

(b) Find the MLEs of the slope and intercept, e.g., using `glm` in R. State the fitted response function and superimpose it on the scatterplot from part (a).

(c) Obtain $\exp(b_1)$ and interpret this number.

(d) What is the estimated probability that an insect dies when the dose level is $x = 3.5$?

(e) What is the estimated median lethal dose—that is, the dose for which 50% of the experimental insects are expected to die?

(f) Find a 99% confidence interval for β_1. Convert it into ones for the odds ratio.

3. Problem ACT 7.3 (Deviance for grouped data). For part b, instead of using the art museum data, use the toxicity data from the previous problem.

4. 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 $\hat{\pi}_x$ be the predicted probability for x using the logistic regression model.

(a) Complete the following table, showing work:

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

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

5. Consider a subscription-based service such as Netflix or a cell phone. Assume that customers join at some point and pay some monthly fee until they decide to cancel the service. Assume that (1) after canceling they never return; (2) the retention rate π (i.e., the probability that a customer is retained in any period) is constant over time and that all customers have the same retention rate; (3) the event that a customer cancels in one period is independent of the event that the customer cancels in any other period. Let T be the time of cancelation, which has geometric PMF $P(T = t) = \pi^{t-1}(1 - \pi)$. Suppose you have a sample of n customers and that customer $i = 1, \ldots, n$ canceled at time t_i. You also have m customers who joined but have not yet canceled (they are said to be *censored*). Among the censored customers, c_i is the time of censoring for customer $i = 1, \ldots, m$, i.e., customer i has been retained c_i months. Estimate the retention rate π using maximum likelihood.

6. Problem ACT 7.7 (Simpson's paradox) See here for more discussion.

```
dat = data.frame(
  female = c(rep(0,6), rep(1,6)),
  dept = rep(LETTERS[1:6],2),
  apps = c(825,560,325,417,191,373,108,25,593,375,393,341),
  admits = c(512,353,120,138,53,22,89,17,202,131,94,24))
```

7. The defaulting customer data set is from a fitness club. The club is for at *adults*. All customers join for three years, but some cancel early. The company would like to understand **which factors are associated with defaulting.** The `default` variable equals 1 if a customer has defaulted within the three months of the membership and 0 otherwise. Customers pay an initial down payment amount (`downpmt`) and 36 monthly payments (`monthdue`) over the next three years. Members use one of four monthly payment methods (`pmttype`): 1=Book, 3=Statement, 4=Checking, and 5=Credit Card. You also know the date of enrollment (`enrolldt`), price of the membership (`price`), a scale indicating how much the customer used the service during the first week of the membership (`use`), and the `age` and `gender` of the member (1=male, 2=female). **Ignore the `enrolldt` variable.**

   (a) (3 points) Study the `age` variable and list things that don't make sense.

   (b) (3 points) What relationship do you expect between `downpmt`, `monthdue` and `price`? Does the relationship hold approximately?

   (c) (3 points) Generate histograms of the `downpmt` variable, varying the number of bins. What pattern do you see? Why do you think this is?

   (d) (5 points) Examine the other variables. Based on your exploratory analyses, state which variables you think are not trustworthy and should be omitted, and which variables have some problems but are otherwise trustworthy. For those in the second class, tell what you will do to fix them.

   (e) (15 points) Analyze the data to understand how the predictors variables are associated with defaulting. Consider only the variables you think are trustworthy from `age`, `price`, `downpmt`, `pmttype`, `gender`, and `monthdue`. Submit your final model and write a short summary telling which variables are most predictive. You should think about all of the techniques and issues discussed in the class, such as dummies, transformations, interactions, multicollinearity, etc. I want you to apply everything you have learned.