MSiA 401: Homework 6
Due: Nov 8, 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
```

   Answer: Here is the distribution:

```
> table(train)
train
FALSE  TRUE
  282   224
```

   (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.

   Answer: No major problems with the residuals plot

	No Trans		With Trans
Part	Model	MSE	MSE
b	Full	0.7083	
c	Backward	0.7033	
d	Ridge	0.7823	0.6178
e	Lasso	0.6954	**0.5342**

```
# Question b
fit = lm(logcrim ~ ., Boston[,-1], subset=train)
plot(fit, pch=16, cex=.8, which=1)
yhat = predict(fit, Boston[!train,])
mean((Boston$logcrim[!train] - yhat)^2)       # compute test set MSE
summary(fit)

# Question c
```

```
fit2 = step(fit)
yhat = predict(fit2, Boston[!train,])
mean((Boston$logcrim[!train] - yhat)^2)       # compute test set MSE
summary(fit)

# Question d
x = model.matrix(logcrim ~ ., Boston[,-1])
fit.ridge = glmnet(x[train,], Boston$logcrim[train], alpha=0)
fit.cv = cv.glmnet(x[train,], Boston$logcrim[train], alpha=0) # find optimal lambda
yhat = predict(fit.ridge, s=fit.cv$lambda.min, newx=x[!train,])  # find yhat for best
mean((Boston$logcrim[!train] - yhat)^2)      # compute test set MSE

# Question e
fit.lasso = glmnet(x[train,], Boston$logcrim[train], alpha=1)
fit.cv = cv.glmnet(x[train,], Boston$logcrim[train], alpha=1)
yhat = predict(fit.lasso, s=fit.cv$lambda.min, newx=x[!train,])
mean((Boston$logcrim[!train] - yhat)^2)       # compute test set MSE

# forward stepwise model with trans
fit= lm(logcrim ~ 1, Boston, subset=train)
fit2 = step(fit, scope=~zn + indus + I(indus^2) + nox + rm + age + dis + log(dis) + s
yhat = predict(fit2, Boston[!train,])
mean((Boston$logcrim[!train] - yhat)^2)

# ridge model with trans
x = model.matrix(logcrim ~ zn + sqrt(zn)
                 + indus + I(indus^2)
                 + nox + log(nox) + I(nox^2)
                 + rm + age + I(age^2) + log(age)
                 + dis + I(1/dis)+I(dis^2)+ log(dis) + sqrt(dis)
                 + rad+log(rad) + tax
                 + ptratio + I(ptratio^2) + log(ptratio)
                 + black + log(black) + I(black^2)
                 + lstat + sqrt(lstat) + log(lstat)
                 +medv + log(medv) + sqrt(medv), Boston)
fit.ridge = glmnet(x[train,], Boston$logcrim[train], alpha=0)
plot(fit.ridge, xvar="lambda")
fit.cv = cv.glmnet(x[train,], Boston$logcrim[train], alpha=0)
yhat = predict(fit.ridge, s=fit.cv$lambda.min, newx=x[!train,])
mean((Boston$logcrim[!train] - yhat)^2)

# lasso model with trans
fit.lasso = glmnet(x[train,], Boston$logcrim[train], alpha=1)
plot(fit.lasso, xvar="lambda")
fit.cv = cv.glmnet(x[train,], Boston$logcrim[train], alpha=1)
```

```
yhat = predict(fit.lasso, s=fit.cv$lambda.min, newx=x[!train,])
mean((Boston$logcrim[!train] - yhat)^2)       # compute test set MSE

# try tree
library(tree)
fit = tree(logcrim ~ ., Boston[train,-1])
yhat = predict(fit, newdata=Boston[!train,])
mean((Boston$logcrim[!train] - yhat)^2)

# overgrow the tree
fit = tree(logcrim ~ ., Boston[train,-1], mindev= .0001)
plot(cv.tree(fit))
yhat = predict(prune.tree(fit, best=10), newdata=Boston[!train,])
mean((Boston$logcrim[!train] - yhat)^2) #
yhat = predict(prune.tree(fit, best=20), newdata=Boston[!train,])
mean((Boston$logcrim[!train] - yhat)^2) #
yhat = predict(prune.tree(fit, best=30), newdata=Boston[!train,])
mean((Boston$logcrim[!train] - yhat)^2) #
```

(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 *Answer: Lasso with transformations wins, retaining zn, sqrt(zn), indus, indus squared, log(nox), nox squared, rm, age squared, log(age), 1/dis, rad, ptratio squared, log(ptratio), log(black), black squared, lstat, sqrt(lstat), log(lstat), medv, and log(medv).* 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.

```
> library(dplyr)
> toxicity=data.frame(x=1:6, n=rep(250,6), s=c(28,53,93,126,172,197)) %>%
  mutate(f = s/n) # used in the next problem
```

```
> fit = glm(y/n ~ x, binomial, toxicity, weight=n)
> summary(fit)
Call: glm(formula = s/n ~ x, family = binomial, data = toxicity, weights = n)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.64367    0.15610  -16.93   <2e-16 ***
x            0.67399    0.03911   17.23   <2e-16 ***

    Null deviance: 383.0695  on 5  degrees of freedom
Residual deviance:   1.4491  on 4  degrees of freedom
AIC: 39.358

> toxicity$phat = fit$fitted.values # used in next problem
> toxicity
  x   n   s     f      phat
1 1 250  28 0.112 0.1224230
2 2 250  53 0.212 0.2148914
3 3 250  93 0.372 0.3493957
4 4 250 126 0.504 0.5130710
5 5 250 172 0.688 0.6739903
6 6 250 197 0.788 0.8022286

> plot(f~x, data=toxicity)   # part a
> x=seq(1, 6, by=.1)     # part b
> lines(x, predict(fit, data.frame(x=x), type="resp"), col=2)
> predict(fit, data.frame(x=3.5), type="resp") # part d
        1
0.4293018
> 2.6436750/0.6739928         # part e
[1] 3.922408
> confint(fit, level=.99)     # part f
                 0.5 %     99.5 %
(Intercept) -3.0559585 -2.2509822
x            0.5753782  0.7770383
> exp(confint(fit, level=.99))     # part f
                 0.5 %    99.5 %
(Intercept) 0.04707758 0.1052958
x           1.77780273 2.1750210
```

(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? *Answer: The plot looks like the logistic response 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). *Answer:* $\log[\pi/(1-\pi)] = -2.64 + 0.674x$

(c) Obtain $\exp(b_1)$ and interpret this number. *Answer: We find* $e^{0.674} = 1.96$*, so for every additional log step in toxicity the odds of dying are roughly doubled.*

(d) What is the estimated probability that an insect dies when the dose level is $x = 3.5$? *Answer:* $= .4293$.

(e) What is the estimated median lethal dose—that is, the dose for which 50% of the experimental insects are expected to die? *Answer:* $2.6437/0.6740 = 3.92$

(f) Find a 99% confidence interval for β_1. Convert it into ones for the odds ratio. *Answer:* $(.575, .777)$*. Exponentiating it we find* $(1.78, 2.18)$.

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. *Answer: I will first restate the logistic regression model. We observe a sample of* n *ordered pairs* (x_i, y_i)*, where* $y_i \in \{0, 1\}$*. Let* $\pi_i = P(Y_i = 1)$ *and* $\ln\left(\frac{\pi_i}{1-\pi_i}\right) = \mathbf{x}_i^\mathsf{T}\hat{\boldsymbol{\beta}} = \eta_i$*, where* $\mathbf{x}_i$ *is a* $p+1$ *vector with* p *predictors and a 1 for the intercept. Then*

$$\hat{\pi}_i = \frac{e^{\eta_i}}{1 + e^{\eta_i}}.$$

Now suppose there are only g *unique values of* $\mathbf{x}$ *(grouped data). I will use* h *to index groups. Suppose there are* n_h *obserations for group* h*, where* $n_1 + \cdots + n_g = n$*. Let* s_h *be the number of successes (sum of* y_i *values) in group* h*, and* $f_h = s_h/n_h$ *be the fraction of successes. I will assume ACT's equation (7.10), which is (1) below, and substitute the two equations that follow it for the MLE and saturated models giving (2) below. Equation (3) converts to group notation. . Equation (4) gathers terms using properties of logs.*

$$\begin{aligned}
D^2 &= -2[\ln L_{\max}(\mathrm{M}) - \ln L_{\max}(\mathrm{SM})] && (1)\\
&= -2\sum_{i=1}^{n}[y_i \ln \hat{\pi}_i + (1-y_i)\ln(1-\hat{\pi}_i) - y_i \ln y_i - (1-y_i)\ln(1-y_i)] && (2)\\
&= -2\sum_{h=1}^{g}[s_h \ln \hat{\pi}_h + (n_h - s_h)\ln(1-\hat{\pi}_h) - s_h \ln f_h - (n_h - s_h)\ln(1-f_h)] && (3)\\
&= -2\sum_{h=1}^{g}\left[s_h \ln\left(\frac{\hat{\pi}_h}{f_h}\right) + (n_h - s_h)\ln\left(\frac{1-\hat{\pi}_h}{1-f_h}\right)\right] && (4)
\end{aligned}$$

The computation matches the deviance reported in the R output for the toxicity data above. Note that R report 4df for the residual deviance.

```
> with(toxicity, -2*sum( s*log(phat/f) + (n-s)*log((1-phat)/(1-f)) ) )
[1] 1.449093
```

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." Here are hints:

   (a) 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.

   (b) Complete the following table, showing work: *Answer: I have computed it exactly below. Note that there are 669 positives and* $1500 - 669 = 831$ *negatives, which provide the denominators for the TPR and FPR columns.*

Cut value	# yes (Cum)	TPR	FPR	Area
$0 \le c < 0.123$	0 (0)	$1 - \frac{0}{669} = 1$	$1 - \frac{0}{831} = 1$	
$0.123 \le c < 0.215$	28 (28)	$1 - \frac{28}{669} = 0.958$	$1 - \frac{222}{831} = 0.733$	0.2616
$0.215 \le c < 0.349$	53 (81)	$1 - \frac{81}{669} = 0.879$	$1 - \frac{419}{831} = 0.496$	0.2178
$0.349 \le c < 0.512$	93 (174)	$1 - \frac{174}{669} = 0.740$	$1 - \frac{576}{831} = 0.307$	0.1529
$0.512 \le c < 0.673$	126 (300)	$1 - \frac{300}{669} = 0.552$	$1 - \frac{700}{831} = 0.158$	0.0964
$0.673 \le c < 0.801$	172 (472)	$1 - \frac{472}{669} = 0.295$	$1 - \frac{778}{831} = 0.064$	0.0397
$0.801 \le c \le 1$	197 (197)	$1 - \frac{669}{669} = 0$	$1 - \frac{831}{831} = 0$	0.0094
Total	669			0.77768

   (c) Plot TPR against FPR and find the area assuming a trapezoid between successive values. *Answer: See table above for .77768. For example,* $.2616 = \frac{1}{2}(1+.9581)(1-.7329)$. *Here is my R code.*

```
# this is used to set things up for the plot.roc function
toxlong = data.frame(
  x = c(rep(1,250), rep(2,250), rep(3,250), rep(4,250),
    rep(5,250), rep(6,250)),
  y = c(
    rep(1, 28), rep(0, 250-28), rep(1, 53), rep(0, 250-53),
    rep(1, 93), rep(0, 250-93), rep(1, 126), rep(0, 250-126),
    rep(1, 172), rep(0, 250-172), rep(1, 197), rep(0, 250-197)
  )
)
fit2 = glm(y~x, binomial, toxlong) # estimates and SE match prob 1
summary(fit2)
library(pROC)
plot.roc(toxlong$y, fit2$fitted.values, print.auc=T)

myroc = data.frame(
  tpr = c(1, 1-28/669, 1-89/669, 1-174/669, 1-300/669, 1-472/669, 0),
```

```
  fpr = c(1, 1-222/831, 1-419/831, 1-576/831, 1-700/831, 1-778/831, 0)
)
plot(myroc$fpr, myroc$tpr, type="l")
```

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. *Answer: Suppose that n_1 customers cancel and n_0 are censored. Let t_i be the time of cancelation for those who cancel, and c_i be the time of censoring. We now derive the maximum likelihood estimate for the retention rate (r).*

$$L(r) = \prod_{i=1}^{n_1} P(T = t_i) \prod_{i=1}^{n_0} S(c_i + 1) = \prod_{i=1}^{n_1} (1 - r) r^{t_i - 1} \prod_{i=1}^{n_0} r^{c_i}.$$

This is maximized by taking logs, differentiating with respect to r, equating the derivative to 0, and solving for r:

$$\hat{r} = \frac{\sum t_t + \sum c_t - n_1}{\sum t_t + \sum c_t} = 1 - \frac{n_1}{\sum t_t + \sum c_t}.$$

This equation has an intuitive interpretation. The denominator of the second term gives the total number of periods in which customers can cancel (opportunities to cancel) and numerator is number of cancelations. Thus, the second term estimates the default rate. The cancelation time was observed for $n_1 = 5$ customers and $n_0 = 3$ were censored. The sum of defection times is $\sum t_i = 3 + 11 + 5 + 8 + 3 = 30$ and the sum of censoring times is $\sum c_i = 12 + 12 + 8 = 32$. The retention rate is estimated as

$$\hat{r} = 1 - \frac{5}{30 + 32} \approx 92\%.$$

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))
```

Answer: (a) A large number of women (the most frequent) apply to department C, which has the lowest admission rate for both sexes. The two most frequent departments for men are A and B, which also have the highest admission rates for both sexes. When we average over the departments, it appears that men have a higher rate. (b) The log odds ratio is $\log[(1493/1198)/(1278/557)] = -0.6103524$*, which equals the logistic regression coefficient for female below.*

```
glm(admits/apps ~ female, binomial, dat, apps) # part b
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.22013    0.03879  -5.675 1.38e-08 ***
female      -0.61035    0.06389  -9.553  < 2e-16 ***


glm(admits/apps ~ dept + female, binomial, dat, apps) # part c
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.58205    0.06899   8.436   <2e-16 ***
deptB       -0.04340    0.10984  -0.395    0.693
deptC       -1.26260    0.10663 -11.841   <2e-16 ***
deptD       -1.29461    0.10582 -12.234   <2e-16 ***
deptE       -1.73931    0.12611 -13.792   <2e-16 ***
deptF       -3.30648    0.16998 -19.452   <2e-16 ***
female       0.09987    0.08085   1.235    0.217
```

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. *Answer: The histogram and frequency distribution show that there is a large group of customers with 0 age, which cannot be. The group of customers who are 99 years old and those between the ages of 1 and 10 are also suspicious. This service is for adults, so it is odd to have anyone younger than 18. The erroneous values, especially the 0's and 99's should be set to missing.*

   (b) (3 points) What relationship do you expect between `downpmt`, `monthdue` and `price`? Does the relationship hold approximately? *Answer: Compute a new variable* `diff = price - downpmt - 36*monthdue`*. One would expect this variable*

to be close to 0. We can evaluate how often it is close to 0 by constructing either a box plot or looking at the percentiles. The tenth percentile is −288 and the 25th percentile is 275. If we define "close to 0" as being within $250, then we are close to 0 less than 15% of the time. Often we are way off. The median is 583, indicating that for half the cases the price is at least $583 higher than the down payment-monthly due combination. Something is wrong with at least one of these variables. We should seek clarification. The price variable is mostly reasonable, but there are some extreme values. Some of you ran a regression, but this is no the best way to see the problems. Others gave a scatterplot, which was helpful. The median monthly due is only $6, which should seem low for a health club membership, unless they are putting a lot of money down. The most common price is 999, followed by 899. There are outliers on price that should bother you.

(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? *Answer: The distribution is highly right skewed with outliers. There are large spikes, which correspond to common amounts such as $100, 150, 50, 75, 25, 5, 19, 250, 175 and 200. This makes me trust the variable more. Beyond this, the max is $9371, which should seem very high for a three-year health club membership.*

(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. *Answer: (1) I don't use monthdue price for reasons given above. (2) There are extreme outliers for downpmt, but I trust the variable otherwise. I don't know why the values are so extreme, but logging the variable will reduce the influence of the extreme values. (3) We need to fix the age values. A value of 0 or 99 likely means that the person refused to give an age, so age should be missing, or there should be an indicator that age is missing. (4) Price is mostly reasonable, but there are some small numbers and some numbers that are too large.*

(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. *Answer: There is a lot to find for this problem. Below I give a rough model, but there is much more than this in the data. Price has a strong association: those who pay more are more likely to churn. The second most important term is the interaction between payment type and down-payment. Down payment has the strongest effect for the base category of payment type (book): of those using book, larger down payments are associated with lower churn rates. Down payment has less effect for other payment methods. Females*

are less likely to churn then males. Age has an inverted U shaped relationship with a max around age 25. Use has a similar, but weaker, interaction with payment type: for those with book or check, using the club is associated with lower churn rates. The effect of use is not as strong for statement, and maybe credit (borderline significant). I assigned partial credit as follows: 4 points for a basic model, 2 for logging down payment, 2 for finding the inverted U for age, 3 for the interaction between payment type and use, and 4 for the interaction between payment type and down payment.

```
> drop1(fit, test="Chisq")
Single term deletions

Model: default ~ log(price+1) + pmttype*log(downpmt+1) + pmttype*use
    + age2 + I(age2^2) + gender
                         Df Deviance     AIC    LRT  Pr(>Chi)
<none>                        9695.4  9727.4
log(price + 1)            1  10224.7 10254.7 529.31 < 2.2e-16 ***
age2                      1   9707.2  9737.2  11.80 0.0005938 ***
I(age2^2)                 1   9718.1  9748.1  22.74 1.851e-06 ***
gender                    1   9735.3  9765.3  39.85 2.743e-10 ***
pmttype:log(downpmt + 1)  3   9787.8  9813.8  92.41 < 2.2e-16 ***
pmttype:use               3   9707.0  9733.0  11.59 0.0089358 **

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                   -5.6181585  0.5629966  -9.979  < 2e-16 ***
log(price+1)                   1.3389743  0.0712462  18.794  < 2e-16 ***
pmttypeState                  -2.9106881  0.2086400 -13.951  < 2e-16 ***
pmttypeCheck                  -5.8985839  0.4173019 -14.135  < 2e-16 ***
pmttypeCredit                 -4.5941545  0.2775906 -16.550  < 2e-16 ***
log(downpmt+1)                -0.9880801  0.0333787 -29.602  < 2e-16 ***
use                           -0.8241268  0.0414088 -19.902  < 2e-16 ***
age2                           0.0590708  0.0176992   3.337 0.000845 ***
I(age2^2)                     -0.0011538  0.0002557  -4.512 6.42e-06 ***
genderFemale                  -0.3282107  0.0521575  -6.293 3.12e-10 ***
pmttypeState:log(downpmt+1)    0.4360913  0.0495837   8.795  < 2e-16 ***
pmttypeCheck:log(downpmt+1)    0.4766803  0.1031775   4.620 3.84e-06 ***
pmttypeCredit:log(downpmt+1)   0.3246730  0.0676247   4.801 1.58e-06 ***
pmttypeState:use               0.1907871  0.0583851   3.268 0.001084 **
pmttypeCheck:use               0.0075860  0.1938405   0.039 0.968783
pmttypeCredit:use              0.1796885  0.1051771   1.708 0.087555 .
```