Homework 5 Answers

Jessica Kraker

Possible Solutions to Selected Questions

Important: This document contains R code solutions and example answers for problems posed in the DS740: Data Mining homework. We are intentionally sharing this document with learners *who have already completed the associated homework*. We want you to be able to review and troubleshoot your code, as well as ask questions to prepare you for later assessments. *By using this document, you are accepting responsibility* **not** *to share it with anyone else, including other students in the course or the program who might not have completed the homework yet*. By upholding this agreement, you are helping us use this tool with learners in future terms.

From Problem 1: Fitting and Selecting Methods

We will be fitting the below linear model using multiple linear regression and penalized regression methods: $Volume = \beta_0 + \beta_1 \cdot Girth + \beta_2 \cdot Height + \beta_3 \cdot Girth \cdot Height + \beta_4 \cdot Girth^2 + \beta_5 \cdot Girth^2 \cdot Height$

Question

Fit the above model, using each of the following methods: Multiple linear regression, Ridge Regression, LASSO (α = 1), and Elastic net (with α = 0.7) and λ = 0.01, 0.02, ..., 0.99, 1.00.

Possible Code Answer:

```
Trees <- read.csv("TreesTransformed.csv")
x = model.matrix(Volume~.,data=Trees)[,-1]
y = Trees$Volume

lmfit = lm(Volume ~ ., data=Trees)

lambdalist = (1:100)/100
library(glmnet)
RRfit = glmnet(x, y, alpha = 0,lambda=lambdalist)
LASSOfit = glmnet(x, y, alpha = 1,lambda=lambdalist)
ENETfit = glmnet(x, y, alpha = 0.7,lambda=lambdalist)
```

Possible Answer regarding linear regression insignificant terms: *None of the t statistics are significant, due to high collinearity of the predictors (some are transformations of others).*

Obtain the values for the coefficients of the LASSO model fit with $\lambda = 0.1$.

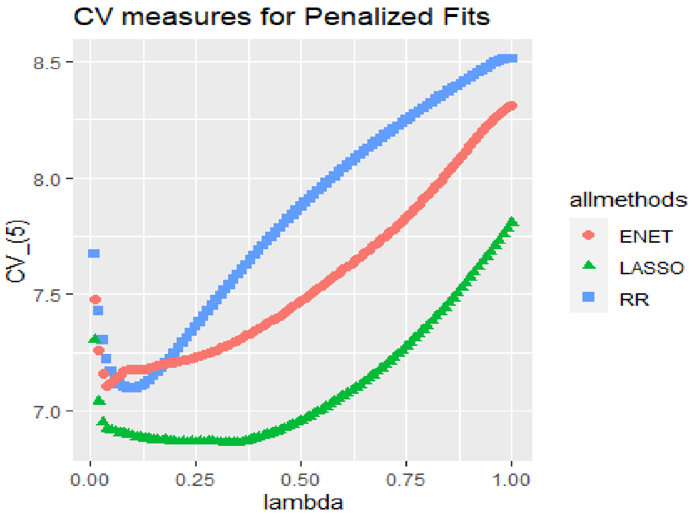
Possible Code Answer:

```
LASSOlambdaused = 0.1
LASSOcoef = coef(LASSOfit,s=LASSOlambdaused); LASSOcoef
```

Question

The image shows a plot of the $CV_{(5)}$ values for Ridge Regression, LASSO, and Elastic net models, plotted against the value of λ. The $CV_{(5)}$ value for multiple linear regression is a constant, $CV_{(5)}$ = 9.59, since no penalty is applied.

Which model is optimal?



Note about Answers: *Pick model and λ value corresponding to lowest $CV_{(5)}$ measure.*

From Problem 2: Motivation for Penalized Regression

Question

Each of the five variables *Enroll, Apps, Accept, F.Undergrad*, and *P.Undergrad* is related to the size of the college and has strongly right-skewed distribution. Explain why the skewness makes sense, in terms of the variety of colleges covered in this dataset.

Possible Answer: *Most colleges would tend to have smaller sizes, but there are a few relatively very-large colleges that would in turn have higher values for all these variables (all associated with size); these large colleges are generally public (versus private).*

Question

To make linear relationships more reasonable, log transformation of these five variables work well. Define the new variables *log.Enroll, log.Apps, log.Accept, log.F.Undergrad*, and *log.P.Undergrad* as the (natural) log transformation of the corresponding variables.

Add these new variables to the data frame; Remove the original-scale variables from the data frame; and Make the variable *Private* into a factor.

Possible Code Answer:

```
library(ISLR)
CollegeT <- College %>%
  mutate(log.Apps = log(Apps), log.Accept = log(Accept), log.F.Undergrad =
log(F.Undergrad),
         log.P.Undergrad = log(P.Undergrad), log.Enroll = log(Enroll)) %>%
  mutate(Private = factor(Private)) %>%
  select(-Apps, -Accept, -F.Undergrad, -P.Undergrad, -Enroll)
```

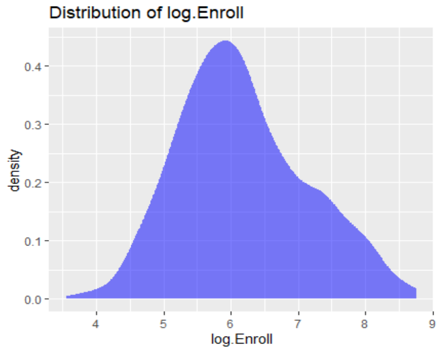
Question

Make an appropriate plot for describing the distribution of the response *log.Enroll*:

Describe the distribution of the (transformed) response *log.Enroll.*

Possible Answer:

```
gf_density( ~ log.Enroll,  fill = "blue", data = CollegeT) %>%
  gf_labs(title = paste("Distribution of log.Enroll"), y = "density", x =
"log.Enroll")
```



The distribution appears to be not too far from symmetric (much less skewed than original scale).

Possible Code:

```
CollegeTemp <- CollegeT %>%
  select(Expend,  log.Accept,  log.P.Undergrad,  perc.alumni,  Personal,
log.Enroll)
```

Question

Provide a reason that the predictor (most highly correlated with the response *log.Enroll*) you chose in the previous question makes sense, based on the description of the data.

Possible Answer: *We would anticipate that the (log of) the number who Enroll is very close to the (log of) the number who Accept spots (i.e., most people who Accept also end up enrolling). We would anticipate high values of both for large colleges and low values of both for small colleges.*

Question

Describe features of this data set that support using a penalized regression model (versus a basic multiple linear regression model).

Possible Answer: There are many highly correlated predictors, primarily those related to size: *Top10perc*, *Top25perc*, *log.Apps*, *log.Accept*, *log.F.Undergrad*, *log.P.Undergrad.* Penalized regression can select the more useful of these, and help reduce some of the variance.

From Problem 3: Applying Methods

Using the data **College** data set from the **ISLR** package, with the new variables as defined in Problem 2, fit the response *log.Enroll* on the remaining variables: *Private, Top10perc, Top25perc, Outstate, Room.Board, Books, Personal, PhD, Terminal, S.F.Ratio, perc.alumni, Expend, Grad.Rate, log.Apps, log.Accept, log.F.Undergrad, log.P.Undergrad*.

For the following questions, fit the LASSO ($\alpha = 1$) for possible values $\lambda = 0.001, 0.002, \ldots, 0.999, 1.000$.

```
# fit models
x = model.matrix(log.Enroll~.,data=CollegeT)[,-1]
y = CollegeT$log.Enroll
lambdalist = 1:1000/1000
LASSOfit = glmnet(x, y, alpha = 1, lambda=lambdalist)
```

Observe coefficients for LASSO model fits with $\lambda = 0.02$, $\lambda = 0.03$, $\lambda = 0.05$, and $\lambda = 0.50$

Possible Code:

```
coef(LASSOfit,s=0.02)

coef(LASSOfit,s=0.03)

coef(LASSOfit,s=0.05)

coef(LASSOfit,s=0.50)
```

Elastic Net model

For the following questions, fit the Elastic net model, with α = 0.75 and possible values λ = 0.001, 0.002, …, 0.999, 1.000.

Question

Using `set.seed(5)`, make groups for 10-fold cross-validation:

```
ncollege = dim(CollegeT)[1]; nfolds = 10
groups = rep(1:nfolds, length = ncollege)
set.seed(5)
cvgroups = sample(groups, ncollege)
```

Use the `cv.glmnet` command along with these cross-validation groups to perform cross-validation; note that $CV_{(10)}$ and λ values are contained, respectively, in the `cvm` and `lambda` values of the output. For the Elastic net model with α = 0.75, make a plot of $CV_{(10)}$ vs λ. Use the "Embed Image" button to submit your plot on Canvas

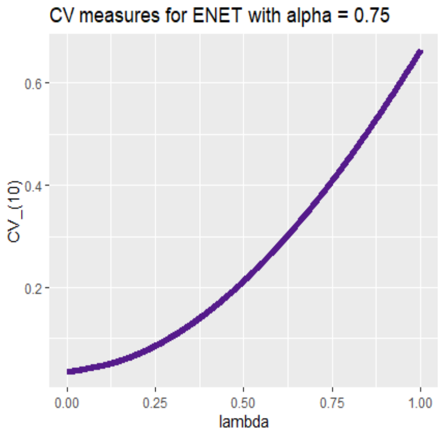
Possible Answer:

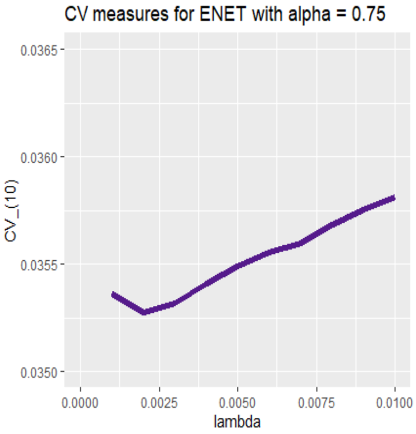
```
cvENET75 = cv.glmnet(x, y, lambda=lambdalist, alpha = 0.75, nfolds=nfolds,
foldid=cvgroups)
gf_line(cvENET75$cvm ~ cvENET75$lambda,size=2,col="purple4") %>%
  gf_labs(title = paste("CV measures for ENET with alpha = 0.75"),
          y = "CV_(10)", x = "lambda")
```

or

```
gf_line(cvENET75$cvm ~ cvENET75$lambda,size=2,col="purple4") %>%
  gf_labs(title = paste("CV measures for ENET with alpha = 0.75"),
          y = "CV_(10)", x = "lambda") %>%
  gf_lims(x = c(0, 0.01), y = c(0.035,0.0365))

## Warning: Removed 990 row(s) containing missing values (geom_path).
```





```
cvENET75$lambda[which.min(cvENET75$cvm)]
```