STAT 6543 Sample Exam

July 03, 2024

Name: _____

ABC123: _____

Note: **Please PRINT your name and ABC123 number** above. You may use your computer/laptop to access the class notes and textbook, but you are NOT allowed to use the internet to google and/or search answers for each problem. Any violation of Academic Conduct (including looking other people's work) will result a 0 on the exam and subsequential treatment. This exam will be due on July 11, 2024 by 11:59 pm central time, and no submission will be accepted after midnight. You need to submit a pdf, html, or word file and your R code or rmd in a separate file to the Canvas. Other format such as qmd may NOT be graded. The submission link is available in the Midterm folder. You will have had three days to complete the exam. If you have any questions or concerns, please contact me as soon as possible. **Good luck!**

I True (T) or False (F). (20 Points)

For these problems, **if false, briefly justify your answer**. Each problem worth 2 points.

1. __F__ In general the more flexible a method is, the ~~*lower*~~ *lower or higher* its RMSE of the test data will be.

2. __F__ When we fit the linear regression model, the collinearity between predictors will ~~*improve*~~ *deteriorate* the coefficient estimates.

3. __F__ ~~*All*~~ *Many* types of statistical models discussed in this course are beneficial from data pre-processing.

4. __T__ One advantage of Principal component analysis (PCA) is that it is a data reduction technique which creates uncorrelated components.

5. __T__ The bias-variance trade-off means that as a method gets more flexible the bias will decrease and the variance will increase but expected RMSE of the testing data may go up or down.

6. __T__ The trade-off between prediction accuracy and interpretability means that a predictive model that is most powerful is usually the least interpretable.

7. __T__ When the sample size n is extremely large, and the number of predictors p is small, we do not expect the performance of a flexible statistical learning method to be better than an inflexible method.

8. __F__ Elastic net, ~~*OLS*~~ *cannot be used*, Ridge regression, Lasso regression can all be used and implemented in situations where the number of predictors is larger than the sample size.

9. __F__ The bootstrap is a widely applicable and extremely powerful statistical tool that can be used to quantify the uncertainty associated with a given estimator. Each "bootstrap set" is created by sampling ~~*without replacement*~~ *with replacement*, and the size is smaller than our original dataset.

10. __F__ The last name of the instructor of this course is ~~*Min*~~ *Wang*.

II Free Response Questions (40 Points)

Problem 1 (Total: 10 Points)
You think of some real-life applications for statistical learning and predictive modelling.

(a) Describe a real-life application in which classification might be useful. Describe the response, as well as the predictors. Is the goal of this application inference or prediction? Clearly explain your answer. (5 Points)

The stock market data that we discussed in Chapters 12, 13, and 14. The response variable is up or down. The predictors are Volume, Today, the percentage returns for each of the five previous trading days, Lag1 through Lag5. This is a prediction problem, since we are interested in predicting if the market is up or down.

(b) Describe a real-life application in which regression might be useful. Describe the response, as well as the predictors. Is the goal of this application inference or prediction? Clearly explain your answer. (5 Points)

We are interested in determining which predictor plays an important role for annual income: The response variable is annual income; the predictors could be age, education level, gender, race, etc. This is a application inference problem, since the goal of this problem is to investigate the impacts of predictors to annual income.

Problem 2 (Total: 10 Points)
During the class time, we learned k-fold cross-validation.

(a) (5 points) Explain how k-fold cross-validation is implemented.

For a data with n observations, we can implement the k-fold cross-validation by randomly splitting the data set into k non-overlapping sets. Thereafter, each of these groups can be viewed as a validation (like the test data set) set and the others treated as a training set. Through the k-fold cross-validation, we could estimate the test error by averaging the k RMSE estimates.

(b) (5 points) What are the advantages and disadvantages of k-fold cross-validation relative to the bootstrap sample.

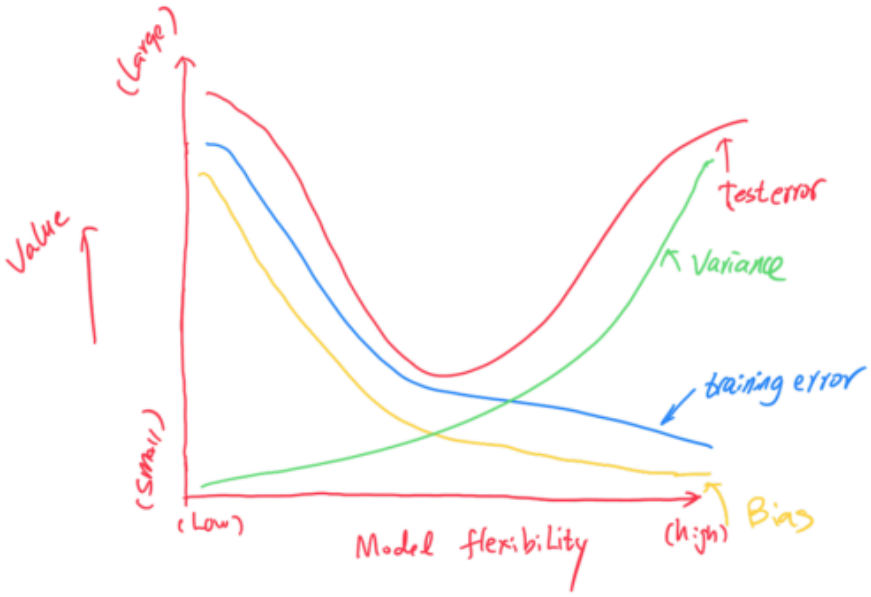
Advantages: conceptually simple and easily implemented. Disadvantages: (i): the estimate of the test error rate can be highly variable depending on which observations are included in the training and test sets and (ii) it could overestimate the test error rate for the model fitting on the whole data.

Problem 3 (Total: 10 Points)
What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

- Advantage: a very flexible (e.g., nonlinear) approach could provide a better fit with a decreasing bias.
- Disadvantages: a very flexible approach often requires to estimate a number of tuning parameters, which could result in overfiting and/or increasing variance.
- When we are interested in prediction, rather than interpretability, we should consider a more flexible approach.
- When we are interested in making statistical inference and interpreting the results, rather than prediction, we should consider a less flexible approach.

Problem 4 (Total: 10 Points)
In this class, we discussed the bias-variance trade-off. Answer the following questions.

(a) Provide a sketch of typical (squared) bias, variance, training error, test error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be four curves. Make sure to label each one.



(b) Briefly explain why each of the four curves has the shape displayed in part (a)

- (squared) bias - decreases monotonically since increases in flexibility result in a closer fit.
- variance - increases monotonically since increases in flexibility result in overfit.
- training error - decreases monotonically since increases in flexibility result in a closer fit.
- test error - concave up curve since increases in flexibility result in a closer fit before it overfits.

III Coding Questions (40 Points)

Problem 5 (Total: 16 Points)
Suppose we are interested in examining the relationship between the response variable sales and the amount of money spent advertising on the TV, radio, and newspapers (i.e, there are three predictors: TV, radio, and newspapers). We fit a multiple linear regression with four predictors (TV, radio, newspaper, and the TV and radio interaction term, denoted by TV:radio) and obtain the following results:

```
> summary(fit)

Call:
lm(formula = sales ~ TV + radio + newspaper + radio * TV, data = dat

Residuals:
   Min     1Q Median     3Q    Max
-6.293 -0.398  0.181  0.596  1.501

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.73e+00   2.53e-01   26.56   <2e-16 ***
TV          1.91e-02   1.51e-03   12.63   <2e-16 ***
radio       2.80e-02   9.14e-03    3.06   0.0025 **
newspaper   1.44e-03   3.30e-03    0.44   0.6617
TV:radio    1.09e-03   5.26e-05   20.69   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.945 on 195 degrees of freedom
Multiple R-squared:  0.968,     Adjusted R-squared:  0.967
F-statistic: 1.47e+03 on 4 and 195 DF,  p-value: <2e-16
```

(a) (4 points) Provide an appropriate interpretation for the coefficient $1.91e-02$.

For every one unit increase in TV, the sales will increase by $1.91e-02$ units.

(b) (4 points) True or false: Since the coefficient for the TV and radio interaction term "TV:radio" is quite small, there is very little evidence that this interaction term is important in predicting the response variable "sales". Justify your answer.

False, since the p-value for the interaction term "TV:radio" is $< 2e-16$, which is less than $\alpha = 0.05$. We thus conclude that this interaction term is statistically significant.

(c) (4 points) Suppose that the company has two options to split $\$12,000$ for the three types of advertising: (i) invest equally $\$4,000$ for each type of advertising, (ii)invest $\$6,000$ for TV, $\$3,000$ for radio, and $\$3,000$ for newspapers. Which option should be recommended for the company. Justify your answer.

- Option 1: $\hat{y} = 6.73 + 1.91e-2*4000 + 2.80e-2*4000 + 1.44e-3*4000 + 1.09e-3*4000*4000 = \$17640.89$

- Option 2: $\hat{y} = 6.73 + 1.91e-2 * 6000 + 2.80e-2 * 3000 + 1.44e-3 * 3000 + 1.09e-3 * 4000 * 3000 = \$19829.65$
- Thus, Option 2 is recommended for the company. (You will receive 0 if you do not show the reason!)

(d) (4 points) Based on this model fit, which predictors are important in predicting the sales? In other words, explain what conclusions you can draw based on the p-values. Your explanation should be phrased in terms of sales, TV, radio, newspaper, and TV:radio, rather than in terms of the coefficients of the linear model.

At $\alpha = 5\%$ significance level, we may conclude that the predictors TV, Radio, and the interaction term "TV:radio" are statistically significant, since their p-values are less than α.

Problem 6 (Total: 24 Points)
we will predict the number of applications received using the other variables in the College data set available in the R package **ISLR**, which can be accessed as follows.

```
library(ISLR)
data(College)
#data basic information
head(College)
dim(College)
# The column Apps is the response variable, and others may be treated as predictors.
#For instance, for linear regression model in R, you may use
lm(Apps~.,data=College)
```

(a) Appropriately split the data set into a training set (80%) and a test set (20%). [4 points]

(b) Fit a linear model using least squares on the training set, and report the test error obtained. [5 points]

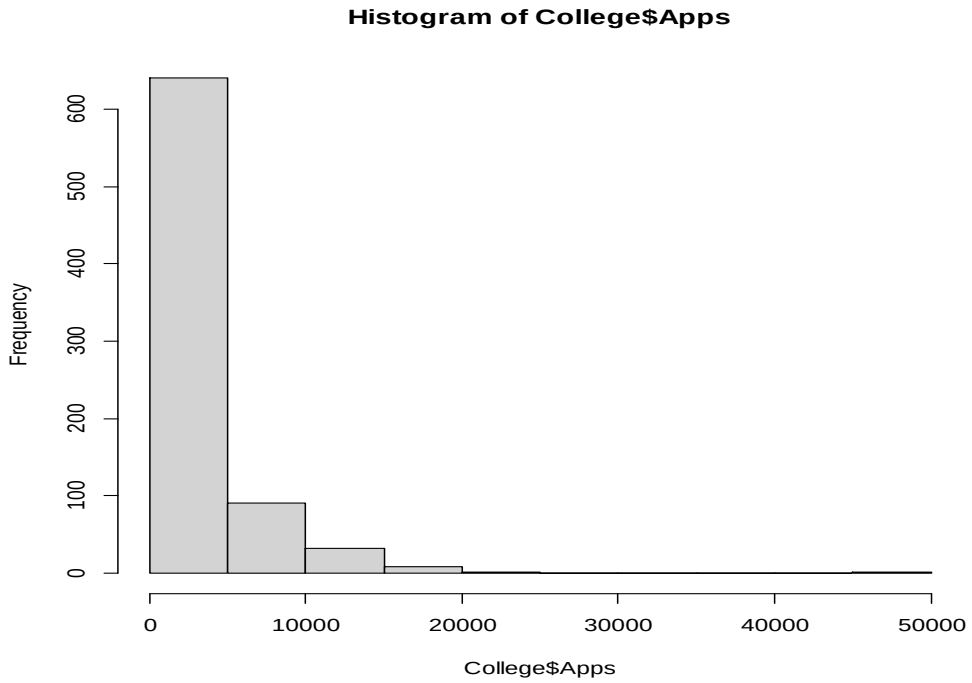
(c) Fit a ridge regression model on the training set, with λ chosen by cross-validation. Report the test error obtained. [5 points]

(d) Fit an ENET model on the training set with tuning parameters chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates. [5 points]

(e) Comment on the results obtained. How accurately can we predict the number of college applications received? Is there much difference among the test errors resulting from these three approaches? [5 points]

(a) Appropriately split the data set into a training set (80%) and a test set (20%).

We first need to check if the response variable Apps is symmetric using the histogram given by



We observe that the response is heavily skewed, so we may need to consider *box-cox* transformation or the logarithm transformation for the response in our data preprocessing stage. In addition, to keep the data structure when we split the data into training and test data, we need to use the *createDataPartition* to keep the structure of the original response variable.

(b) Fit a linear model using least square

```
> summary(lmTune)

Call:
lm(formula = .outcome ~ ., data = dat)

Residuals:
   Min     1Q Median     3Q    Max
 -4189  -1159   -401    724  35382

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  2964.78      92.66  31.997  < 2e-16 ***
PrivateYes   -372.55     167.23  -2.228 0.026262 *
```

```
Accept       2312.16     349.67   6.612 8.31e-11 ***
Enroll       -494.32     476.85  -1.037 0.300317
Top10perc     225.88     247.53   0.913 0.361852
Top25perc     126.72     233.73   0.542 0.587894
F.Undergrad   945.77     399.41   2.368 0.018202 *
P.Undergrad   182.31     140.29   1.300 0.194257
Outstate     -371.62     193.43  -1.921 0.055178 .
Room.Board    171.15     140.71   1.216 0.224355
Books          26.64      99.05   0.269 0.788101
Personal      -28.11     106.84  -0.263 0.792569
PhD          -225.62     205.62  -1.097 0.272950
Terminal     -103.55     197.35  -0.525 0.599981
S.F.Ratio    -129.80     140.39  -0.925 0.355568
perc.alumni  -177.04     127.51  -1.389 0.165491
Expend        683.53     188.52   3.626 0.000312 ***
Grad.Rate     275.35     124.56   2.211 0.027436 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2315 on 606 degrees of freedom
```

The test RMSE of linear model is 1962.887.

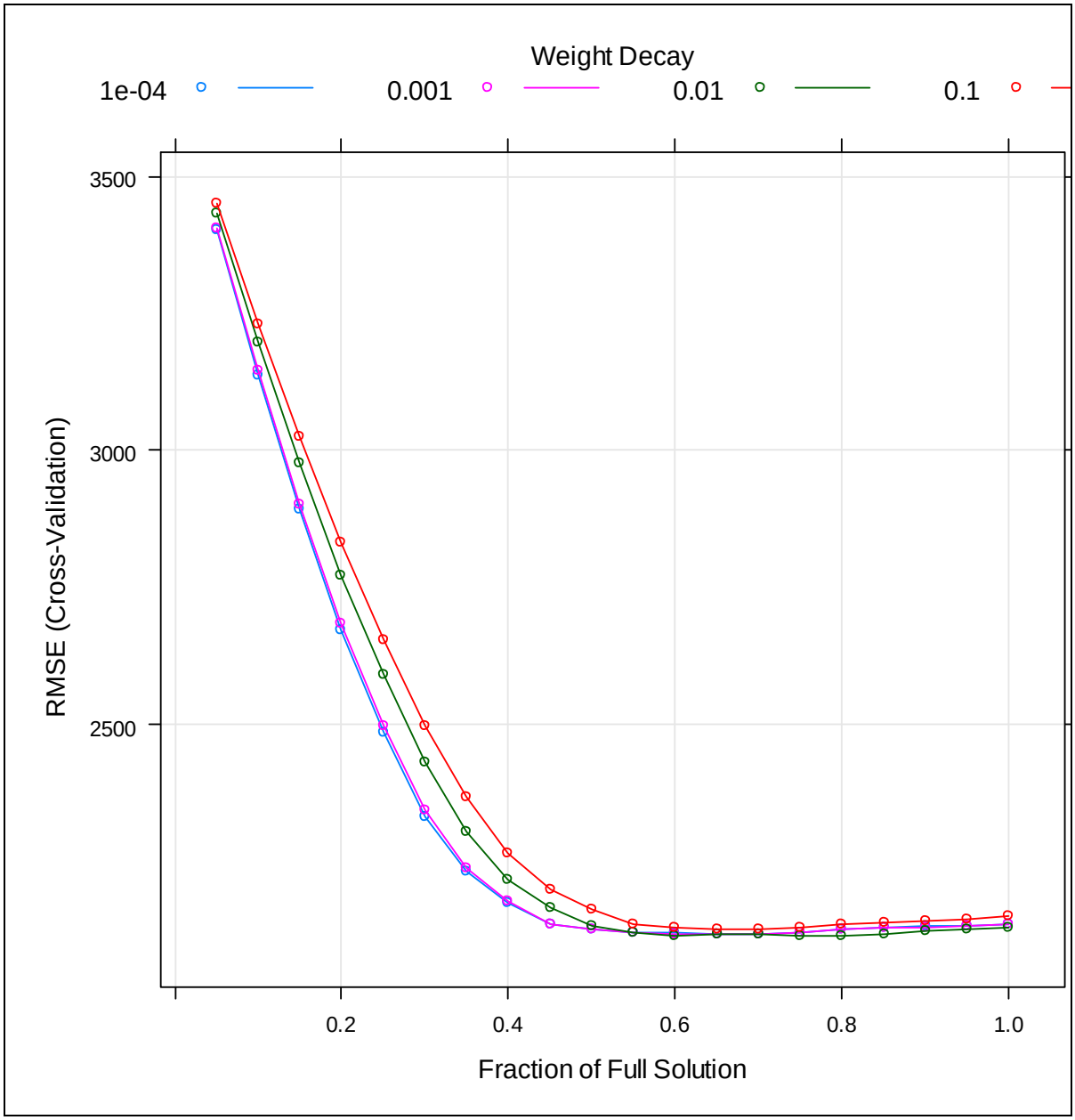
(c) Fit a ridge regression model

```
> ridgeTune
Ridge Regression

624 samples
 17 predictor

Pre-processing: centered (17), scaled (17), Box-Cox transformation (15)
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 560, 561, 563, 561, 562, 562, ...
Resampling results across tuning parameters:

  lambda      RMSE      Rsquared   MAE
  0.00000000  2137.180  0.7106780  1401.926
  0.01111111  2129.928  0.7126297  1399.719
  0.02222222  2128.649  0.7133962  1402.018
  0.03333333  2129.574  0.7137901  1405.858
  0.04444444  2131.673  0.7140163  1410.563
  0.05555556  2134.547  0.7141490  1415.884
  0.06666667  2138.004  0.7142216  1421.721
```

```
0.07777778  2141.938  0.7142516  1428.022
0.08888889  2146.282  0.7142495  1434.487
0.10000000  2150.991  0.7142218  1441.452

RMSE was used to select the optimal model using the smallest value.
The final value used for the model was lambda = 0.02222222.
```

The test RMSE of the ridge model is 1939.691.

(d) Fit an enet model given by



The test RMSE of the enet model is 1960.351.

(e) We observe that all errors are comparable. In order to see if there exist the mean differences of the predicted value from the three considered methods, we conduct the analysis of variance (ANOVA) for the predicted from the three different models under consideration. The results of ANOVA are shown below

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
methods	2	3.272e+04	16360	0.002	0.998
Residuals	456	4.713e+09	10336315		

Thus, we may conclude that all methods are comparable, since the p-value = 0.998>0.05, indicating that there is statistically significant difference among the mean differences of the three models under consideration.

R codes for repeating the reported results above

```
library(caret)
library(ISLR)
library(elasticnet)
library(Metrics) #For calculating RMSE, MAE, and R2

data(College)
#data basic information
head(College)
dim(College)


#(a) Appropriately split the data set into a training set (80%) and a test set (20%).
#Check the distribution of the response varible Apps to
#decide if we need to use sample or createDataPartition
hist(College$Apps) #right skewed, we need to use createDataPartition

set.seed(1)
index <- createDataPartition(College$Apps, p=0.8)[[1]]

#Training data
traindata <- College[index,]
#Test data
testdata <- College[-index,]

#Since we will fit various models in the following, we create a control function first
set.seed(100)
#Create a series of test/training partitions
```

```
#default is 10, the funtion below creates 10 folder
indx <- createFolds(traindata$Apps, returnTrain = TRUE)
#control the computational nuances of the train function
ctrl <- trainControl(method = "cv", index = indx)


#(b) Fit a linear model
set.seed(1)
lmTune <- train(Apps ~ ., data = traindata, method = "lm",
trControl = ctrl, preProcess = c("center","scale", "nzv","BoxCox"))
summary(lmTune)

#Test RMSE
#Save the predicted value based on the linear model
testdata$lm = predict(lmTune, testdata)
rmse(testdata$lm, testdata$Apps)

#(b) Fit a ridge model
#Fit a ridge regression
set.seed(1) #it may take a while to get results
ridgeGrid <- expand.grid(lambda = seq(0, .1, length = 10))
ridgeTune <- train(Apps ~ ., data = traindata,
              method = "ridge",
              tuneGrid = ridgeGrid,
              trControl = ctrl,
              preProc = c("center", "scale","BoxCox"))

ridgeTune
#Test RMSE
#Save the predicted value based on the ridge model
testdata$ridge = predict(ridgeTune, testdata)
rmse(testdata$ridge, testdata$Apps)

#(b) Fit an enet model
#Fit an ENET regression
enetGrid <- expand.grid(lambda = c(0.0001, .001, 0.01, 0.1),
                  fraction = seq(.05, 1, length = 20))
set.seed(100)
enetTune <- train(Apps ~ ., data = traindata,
             method = "enet",
             tuneGrid = enetGrid,
             trControl = ctrl,
             preProc = c("center", "scale","BoxCox"))
```

```
enetTune

#Test RMSE
#Save the predicted value based on the enet model
testdata$enet = predict(enetTune, testdata)
rmse(testdata$enet, testdata$Apps)

#(e) Comments on your results
# Conduct analysis of variance to see if the mean difference of the predicted value from the
three considered methods
methods = c(rep('lm', 153), rep('ridge', 153), rep('enet', 153))
values = c(testdata$lm, testdata$ridge, testdata$enet)
aov = aov(values~ methods)
summary(aov)
#> summary(aov)
#            Df    Sum Sq  Mean Sq F value Pr(>F)
#methods      2 3.272e+04    16360   0.002  0.998
#Residuals   456 4.713e+09 10336315

#All methods are comparable, since the p-value = 0.998>0.05, indicating that there
# is statistically significant difference among the mean differences of the
# three models under consideration.
```