STA 5207: Homework 4

Due: Wednesday, October 11 by 11:59 PM

Include your R code as an appendix at the end of your homework. Do not include your code in your answers unless the question explicitly tells you to include your code. Your answers to each exercise should be self-contained without code so that the grader can determine your solution without reading your code or deciphering its output.

Exercise 1 (Using `step`) [40 points]

For this exercise we will use the `prostate` data set from the `faraway` package. You can also find the data in `prostate.csv` on Canvas. The data set comes from a study on 97 men with prostate cancer who were due to receive a radical prostatectomy. The variables in teh data set are

- `lcavol`: log(cancer volume).
- `lweight`: log(prostate weight).
- `age`: The patient's age in years.
- `lbph`: log(benign prostatic hyperplasia amount).
- `svi`: Seminal vesicle invasion.
- `lcp`: log(capsular penetration).
- `gleason`: Gleason score.
- `pgg45`: percentage Gleason score 4 or 5.
- `lpsa`: log(prostate specific antigen).

In the following exercises, use `lpsa` as the response and the other variables as predictors.

1. (6 points) Identify the best model based on AIC and BIC using forward selection. Create a table listing each quality criterion (AIC, BIC) and the subset of variables chosen by the method.

Ans:-

Criterion	Selected_Variables	Value
AIC	lcavol, lweight,svi,lbph,age	-61.37
BIC	lcavol, lweight,svi	-50.38

2. (6 points) Identify the best model based on AIC and BIC using backward selection. Create a table listing each quality criterion (AIC, BIC) and the subset of variables chosen by the method.

Ans:-

Criterion	Selected_Variables	Value
AIC	lweight, age, lbph, svi, lcavol	-61.37
BIC	lweight, lcavol,svi	-50.38

3. (6 points) Identify the best model based on AIC and BIC using stepwise selection. Create a table listing each quality criterion (AIC, BIC) and the subset of variables chosen by the method.

Ans:-

Criterion	Selected_Variables	Value
AIC	lcavol, lweight, svi, lbph, age	-61.37
BIC	lcavol, lweight, svi	-50.38

4. (12 points) Identify the best model based on R_a^2, AIC, and BIC using best subset selection. Create a table listing each quality criterion (R_a^2, AIC, BIC) and the subset of variables chosen by the method.

Ans:-

Criterion	Subset	Value
R_a^2	lcavol,lweight,age,lbph,svi,lcp,pgg45	0.6272521
AIC	lcavol,lweight,age,lbph,svi	-61.37
BIC	lcavol,lweight,svi	-50.38

5. (10 points) For each unique candidate model chosen in parts 1 - 4, report their $\text{RMSE}_{\text{LOOCV}}$. Which model do you prefer based on this criteria?

Ans:-

Model	$\text{RMSE}_{\text{LOOCV}}$
AIC Forward	0.736896
BIC Forward	0.7381178
AIC Backward	0.736896
BIC Backward	0.7381178
AIC Stepwise	0.736896
BIC Stepwise	0.7381178
R^2 Subset	0.7410915
AIC Subset	0.736896
BIC Subset	0.7381178

Based on $\text{RMSE}_{\text{LOOCV}}$, AIC Backward/Forward/stepwise/Subset model is preferred which is lpsa ~ lcavol + lweight + age + lbph + svi

Exercise 2 (Boston Housing Data) [40 points]

For this exercise we will use the Boston data set from the ISLR2 package. You can also find the data in Boston.csv on Canvas. The data set contains housing values in 506 suburbs of Boston. There are a total of 12 predictors. You can type ?ISLR2::Boston in R to read about the data set and the meaning of the predictors. In the following exercises, use crim (the per capita crime rate) as the response and the other variables as predictors.

1. (6 points) Identify the best model based on AIC and BIC using forward selection. Create a table listing each quality criterion (AIC, BIC) and the subset of variables chosen by the method.

   Ans:-

Criterion	Variables	Values
AIC	rad, lstat, medv, ptratio	1903.797
BIC	rad, lstat	1914.161

2. (6 points) Identify the best model based on AIC and BIC using backward selection. Create a table listing each quality criterion (AIC, BIC) and the subset of variables chosen by the method.

   Ans:-

Criterion	Variables	Values
AIC	zn, nox, dis, rad, ptratio, lstat, medv	1894.7
AIC	zn, dis, rad, medv	1920.358

3. (6 points) Identify the best model based on AIC and BIC using stepwise selection. Create a table listing each quality criterion (AIC, BIC) and the subset of variables chosen by the method.

   Ans:-

Criterion	Variables	Values
AIC	rad, lstat, medv, ptratio	1903.797
BIC	rad, lstat	1919.116

4. (12 points) Identify the best model based on R_a^2, AIC, and BIC using best subset selection. Note that you have to set nvmax = 12 when calling regsubsets, since there are 12 predictors. Create a table listing each quality criterion (R_a^2, AIC, and BIC) and the subset of the variables chosen by the method.

Ans:-

Criterion	Variables	Values
R^2	zn, indus, nox, rm, dis, rad, ptratio, lstat, medv	0.438
AIC	zn, nox, dis, rad, ptratio, lstat, medv	1894.7
BIC	rad, lstat	1919.116

5. (10 points) For each unique candidate model chosen in parts 1 - 4, report their RMSE$_{\text{LOOCV}}$. Which model do you prefer based on this criteria?

   Ans:-

Model	RMSELOOCV
AIC Forward	6.576221
BIC Forward	6.601046
AIC Backward	6.497268
BIC Backward	6.53288
AIC Stepwise	6.576221
BIC Stepwise	6.601046
R^2 Subset	6.509403
AIC Subset	6.497268
BIC Subset	6.601046

Exercise 3 (Post-Selection Inference and Data Splitting) [20 points]

For this exercise, we will use the `prostate_fake_train.csv` and `prostate_fake_test.csv` data sets on Canvas. These data sets are subsets of the `prostate` data set you analyzed in Exercise 1; however, I replaced the `lpsa` column with a column of noise drawn from a uniform distribution on $[-1,1]$. I then split the data set into a training subset and a testing subset. I ran the following code:

For this exercise, use `noise` as the response and the remaining variables as predictors. Note that by design there is no relationship between `noise` and any of the predictors.

1. (6 points) Identify the best model using AIC and backward selection based on the data in `prostate_fake_train.csv`. Report the subset of the variables chosen by this method.

   Ans:-

   AIC=-60.54

   model = noise ~ lweight + gleason + pgg45

   variables chosen = lweight,gleason,pgg45

2. (7 points) Using your model from part 1, perform a t-test at the $\alpha = 0.05$ significance level for each predictor. Report the predictors that are significant according to this test. Should we trust the results of this test? Why or why not?

   Ans:-

   gleason & pgg45 are significant according to t-test at the α = 0.05 significance level.

   p-values obtained after variable selection is much smaller than the true values.So We can't trust the above results.

3. (7 points) Using the predictors you selected in part 1, fit a multiple linear regression model on the data in `prostate_fake_test.csv`. Perform a t-test at the $\alpha = 0.05$ significance level for each predictor. Report the predictors that are significant according to this test. Do the results match the results from part 2? Should we trust these results? Why or why not?

   Ans:-

   In part 2, the t-test conducted at the α = 0.05 significance level indicates that none of the predictors (lweight, gleason, pgg45) are statistically significant, as their p-values exceed 0.05. This outcome is inconsistent with the findings from part 2.

   We can trust this result as the inference is made on testing dataset.

Code Appendix

```
# Code for Problem 1, Question 1
library(faraway)
data("prostate")
ms = lm(lpsa ~ 1, data = prostate)
mfaic = step(ms,
scope = lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45,
direction = 'forward')
n = nrow(prostate)
mfbic = step(
ms,
scope = lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45,
direction = 'forward',
k = log(n))
extractAIC(mod_forwd_aic)
extractAIC(mod_forwd_bic, k = log(n))

# Code for Problem 1, Question 2
library(faraway)
data("prostate")
mpreds = lm(lpsa ~ ., data = prostate)
mbaic = step(mpreds, direction = 'backward')
n = nrow(prostate)
mbbic = step(mpreds, direction = 'backward', k = log(n))
extractAIC(mbaic)
extractAIC(mbbic, k = log(n))


# Code for Problem 1, Question 3
library(faraway)
data("prostate")
ms = lm(lpsa ~ 1, data = prostate)
msaic = step(ms,
scope = lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45,
direction = 'both')
n = nrow(prostate)
msbic = step(
ms,
scope = lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45,
direction = 'both',
k = log(n))
extractAIC(msaic)
extractAIC(msbic, k = log(n))

# Code for Problem 1, Question 4
library(leaps)
me = summary(regsubsets(lpsa ~ ., data = prostate, nvmax = 8))
max(me$adjr2)
bestr2ind = which.max(me$adjr2)
me$which[bestr2ind,]
p = ncol(me$which)
maic = n * log(me$rss / n) + 2 * (2:p)
min(maic)
bestaicind = which.min(maic)
me$which[bestaicind, ]
n = nrow(prostate)
p = ncol(prostate)
mbic = n * log(me$rss / n) + log(n) * (2:p)
min(mbic)
bestbicind = which.min(mbic)
me$which[bestbicind, ]


# Code for Problem 1, Question 5
model_formulas <- list(
  mod_forwd_aic = "lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45",
  mod_forwd_bic = "lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45",
  mod_back_aic = "lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45",
  mod_back_bic = "lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45",
  mod_stepwise_aic = "lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
pgg45",
  mod_stepwise_bic = "lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
pgg45",
  mod_exhaust_r2 = "lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
```

```r
pgg45",
  mod_exhaust_aic = "lpsa ~ lcavol + lweight + age + lbph + svi",
  mod_exhaust_bic = "lpsa ~ lcavol + lweight + svi"
)
calc_loocv_rmse = function(formula) {
  model <- lm(formula, data = prostate)
  sqrt(mean((resid(model) / (1 - hatvalues(model)))^2))
}
loocv_rmse_results <- sapply(model_formulas, calc_loocv_rmse)
loocv_rmse_results



# Code for Problem 2, Question 1
data(Boston,package='ISLR2')
ms = lm(crim ~ 1, data = Boston)
mfaic = step(ms,
scope = crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio
+ lstat + medv,
direction = 'forward')
n = nrow(prostate)
mfbic = step(
ms,
scope = crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio
+ lstat + medv,
direction = 'forward',
k = log(n))
extractAIC(mfaic)
extractAIC(mfbic, k = log(n))


# Code for Problem 2, Question 2
data(Boston,package='ISLR2')
mallpreds = lm(crim ~ ., data = Boston)
mbackaic = step(mod_all_preds, direction = 'backward')
n = nrow(Boston)
mbackbic = step(mallpreds, direction = 'backward', k = log(n))
extractAIC(mbackaic)
extractAIC(mbackbic, k = log(n))


# Code for Problem 2, Question 3
data(Boston,package='ISLR2')
ms = lm(crim ~ 1, data = Boston)
mstepwiseaic = step(ms,
scope = crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio
+ lstat + medv,
direction = 'both')
n = nrow(Boston)
```

```r
mstepwisebic = step(ms,
scope = crim ~ zn + indus + chas + nox + rm + age + dis + rad + tax + ptratio
+ lstat + medv,
direction = 'both',
k = log(n))
extractAIC(mstepwiseaic)
extractAIC(mstepwisebic, k = log(n))


# Code for Problem 2, Question 4
library(leaps)
me = summary(regsubsets(crim ~ ., data = Boston, nvmax = 12))
max(me$adjr2)
bestr2ind = which.max(me$adjr2)
me$which[bestr2ind,]
p = ncol(me$which)
maic = n * log(mod_exhaustive$rss / n) + 2 * (2:p)
min(maic)
bestaicind = which.min(maic)
me$which[bestaicind, ]
n = nrow(Boston)
p = ncol(Boston)
mbic = n * log(me$rss / n) + log(n) * (2:p)
min(mbic)
bestbicind = which.min(mbic)
me$which[bestbicind, ]

# Code for Problem 2, Question 5
model_formulas <- list(
  mod_forwd_aic = "crim ~ zn + indus + nox + rm + dis + rad + ptratio + lstat
+ medv",
  mod_forwd_bic = "crim ~ zn + nox + dis + rad + ptratio + lstat + medv",
  mod_back_aic = "crim ~ zn + indus + nox + rm + dis + rad + ptratio + lstat
+ medv",
  mod_back_bic = "crim ~ zn + indus + nox + rm + dis + rad + ptratio + lstat
+ medv",
  mod_stepwise_aic = "crim ~ zn + indus + nox + rm + dis + rad + ptratio +
lstat + medv",
  mod_stepwise_bic = "crim ~ zn + nox + dis + rad + ptratio + lstat + medv",
  mod_exhaust_r2 = "crim ~ zn + indus + nox + rm + dis + rad + ptratio +
lstat + medv",
  mod_exhaust_aic = "crim ~ zn + nox + dis + rad + ptratio + lstat + medv",
  mod_exhaust_bic = "crim ~ rad + lstat"
)
calc_loocv_rmse = function(formula, data) {
  model <- lm(formula, data = data)
  sqrt(mean((resid(model) / (1 - hatvalues(model)))^2))
}
loocv_rmse_results <- sapply(model_formulas, calc_loocv_rmse, data = Boston)
loocv_rmse_results
```

```r
# Code for Problem 3, Question 1
library(readr)
train=read_csv("prostate_fake_train.csv",show_col_types = FALSE)
mallpredstr = lm(noise ~ ., data = train)
mbackaic = step(mallpredstr, direction = 'backward')
extractAIC(mbackaic)

# Code for Problem 3, Question 2
library(readr)
train=read_csv("prostate_fake_train.csv",show_col_types = FALSE)
model=lm(noise ~ lweight + gleason + pgg45, data=train)
summary(model)

# Code for Problem 3, Question 3
library(readr)
test=read_csv("prostate_fake_test.csv",show_col_types = FALSE)
model=lm(noise ~ lweight + gleason + pgg45, data=test)
summary(model)
```