Module 2 Assignment on Linear Regression - 2 - V1

First Name + Last Name // Undergraduate/Graduate Student

Today's date

Read and Delete This Part When Typing

- Give a name to this rmd file: `ModuleNumber_ModuleName_HWSubmission_FirstName_LastName` (for example, `Module0_Reviews_HWSubmission_Yusuf_Bilgic.rmd`).
- First, read the slides, review the notes, and run the lab codes. Then do the assignment, type the solution here. Knit (generate the pdf) the file. Check if it looks good.
- **Especially this Module 2, find a pair, work together, split parts among each of you, explain each other, make sure you understand all pair solutions, combine solutions, and submit separately. It is fine if your codes and results are the same. I expect comments will be your own.**
- You will then submit two files to Blackboard:
  1) `ModuleNumber_ModuleName_HWSubmission_FirstName_LastName.pdf` and
  2) `ModuleNumber_ModuleName_HWSubmission_FirstName_LastName.rmd`.
- Grading will be based on the pdf file uploaded (avoid uploading extra docs). Make it easy and readable. Grader or me may take a look at the rmd file.
- Unless otherwise specified, use a 5% level for statistical significance.
- Always include your comments on results: don't just leave the numbers without explanations. Use full sentences, structured paragraphs if needed, correct grammar, and proofreading.
- Don't include irrelevant and uncommented outputs. Don't include all codes: use `echo=False, results='hide'` for most of time. You can include the codes when your solution becomes easier to follow. Also, include useful results. Try to call the outputs from $'r\ xyz'$.
- Show your knowledge with detailed work in `consistency` with course materials though tons of other ways may exist.
- Each part is 1 pt, so the the total is 20 pt (4 pt is baseline score). If the response is not full or not reflecting the correct answer as expected, you may still earn 0.5 or just get 0.0 pt. Your TA will grade your work. Any questions, you can write directly to your TA and cc me. Visit my office hours on TWR. Thanks!

Module Assignment Questions

In this assignment, you will use the `Auto` data set with 7 variables (one response `mpg` and six numerical) and $n = 392$ vehicles. For sake of simplicity, categorical variables were excluded. Before each randomization used, use `set.seed(99)` so the test results are comparable.

Q1) (*Forward and Backward Selection*)

In `Module 1 Assignment`, `Q2`, you fitted `Model 3` with `mpg` as the response and the six numerical variables as predictors. This question involves the use of `forward` and `backward` selection methods on the same data set.

a. Using `OLS`, fit the model with all predictors on `mpg`. Report the predictors' coefficient estimates, R_{adj}, and MSE. Note: The method in `lm()` is called ordinary least squares (OLS).

```
#This is setup to start
library(ISLR)
Model_3 = mpg ~ horsepower+year+cylinders+displacement+weight+acceleration
Model_3.fit = lm(Model_3, data=Auto)
summary(Model_3.fit)
# Or, prefer this restructuring way
# by excluding categorical variables:
# Make sure AutoNum is a data.frame
AutoNum = Auto[, !(colnames(Auto) %in% c("origin", "name"))]
Model_Full = mpg ~ . #you can write models in this way to call later
Model_Full.fit = lm(Model_Full, data=AutoNum)
summary(Model_Full.fit)
```

b. Using `forward selection method` from `regsubsets()` and `method="forward"`, fit MLR models and select the `best` subset of predictors. Report the best model obtained from the default setting by including the predictors' coefficient estimates, R_{adj}, and MSE.

```
# helpful code from the r lab: review it
Model_Full = mpg ~ .
regfit.m1=regsubsets(Model_Full, data=AutoNum, nbest=1,
                     nvmax=6, method="forward")
reg.summary=summary(regfit.m1)
reg.summary
names(reg.summary)
reg.summary$adjr2
coef(regfit.m2, 1:6) #coefficients of all models built
```

c. What criterion had been employed to find the best subset? What other criteria exist? Explain.

d. Using `backward selection method` from `regsubsets()` and `method="backward"`, fit MLR models and select the `best` subset of predictors. Report the best model obtained from the default setting by including predictors, their coefficient estimates, R_{adj}, and MSE.

e. Compare the results obtained from `OLS`, `forward` and `backward` selection methods (parts a, b and d): What changed? Which one(s) is better? Comment and justify.

Q2) (*Cross-Validated with k-Fold*)

What changes in model selection results and the coefficient estimates when cross-validated set approach is employed? Specifically, we will use k-fold cross-validation (`k-fold CV`) here.

a. Using the 5-fold CV approach, fit the OLS MLR model on `mpg` including all the predictors. Report the all predictors' coefficient estimates, MSE_{train}, and MSE_{test}.

b. Using the 5-fold CV approach and `forward selection method`, fit MLR models on `mpg` and select the `best` subset of predictors. Report the best model obtained from the default setting by including the predictors' coefficient estimates, the averaged MSE_{train}, and the averaged MSE_{test}.

c. Compare the MSE_{test}'s. Explain.

d. Using the 5-fold CV approach and `backward selection method`, fit MLR models on `mpg` and select the `best` subset of predictors. Report the best model obtained from the default setting by including the predictors' coefficient estimates, the averaged MSE_{train}, MSE_{test}.

e. Did you come up with a different model on parts b and d? Are the predictors and their coefficient estimates same? Compare and explain.

f. Which fitted model is better among parts a, b, and d? Why? Justify.

Q3) (*Shrinkage Methods*)

Results for `OLS`, `lasso`, and `ridge` regression methods can be comparable. Now, you are expected to observe that ridge and lasso regression methods may reduce some coefficients to zero (so in this way, these features are eliminated) and shrink coefficients of other variables to low values.

In this exercise, you will analyze theses estimation and prediction methods (OLS, ridge, lasso) on the `mpg` in the Auto data set using $k-fold$ cross-validation test approach.

a. Fit a ridge regression model on the entire data set (including all six predictors, don't use yet any validation approach), with the optimal λ chosen by `cv.glmnet()`. Report $\hat{\lambda}$, the predictors' coefficient estimates, and MSE.

b. Fit a lasso regression model on the entire data set (including all six predictors, don't use yet any validation approach), with the optimal λ chosen by `cv.glmnet()`. Report $\hat{\lambda}$, the predictors' coefficient estimates, and MSE.

c. Compare the parts a and b in Q3 to part a in Q1. What changed? Comment.

d. How accurately can we predict `mpg`? Using the three methods (OLS, ridge and lasso) with all predictors, you will fit and test using 5-fold cross-validation approach with the optimal λ chosen by `cv.glmnet()`. For each, report the averaged train and test errors (MSE_{train}, MSE_{test}):

   1) Fit an `OLS` model.
   2) Fit a `ridge` regression model.
   3) Fit a `lasso` regression model.

e. Write an overall report on part d by addressing the inquiry, `how accurately can we predict mpg?`. Is there much difference among the test errors resulting from these three approaches? Show your comprehension.

f. (BONUS) Propose a different model (or set of models) that seem to perform well on this data set, and justify your answer.

g. (BONUS) Include categorical variables to the models you built in part d, Q3. Report.

h. (GOLDEN BONUS) Propose a model (or set of models) that seem to perform well on this data set, and justify your answer. Make sure that you are evaluating model performance using 5-fold cross-validation approach. You can transform the data, scale and try any methods. When MSE_{test} is the lowest (under the setting of Q3, part d) in the class, your HW assignment score will be 100% (20 pts).

i. (BONUS) You can make a hybrid design in model selection using all the methods here in a way that yields better results. Show your work, justify and obtain better results in part d, Q3.

Your Solutions

Q1)

Part a:

Part b:

Part c:

Part d:

Part e:

Q2)

Part a:

Part b:

Part c:

Part d:

Part e:

Part f:

Q3)

Part a:

Part b:

Part c:

Part d:

Part e:

Write comments, questions: ...

I hereby write and submit my solutions without violating the academic honesty and integrity. If not, I accept the consequences.

List the fiends you worked with (name, last name): ...

Disclose the resources or persons if you get any help: ...

How long did the assignment work take?: ...

References

...