Homework 8

Alex Soupir

March 27, 2020

Packages: ISLR, glmnet, pls, MASS, leaps

Collaborators:

1. Question 6.8.4 pg 260

Question 4: Suppose we estimate the regression coefficients in a linear regression model by minimizing the equation for a particular value of λ. For parts (a) through (e), indicate which of i. through v. is correct. Justify your answer.

(a) As we increase λ from 0, the training RSS will:

```
i. Increase initially, and then eventually start decreasing in an inverted `U` shape.

ii. Decrease initially, and then eventually start increasing in a `U` shape.

iii. Steadily increase.

iv. Steadily decrease.

v. Remain constant.
```

iii: The training RSS will steadily increase because the beta in the error term decreases and approaches zero. When λ is 0 then we get the oringary least squares.

(b) Repeat (a) for test RSS.

ii: The test RSS will decrease initially, and then eventually start increasing in a "U" shape. At some combination of coefficients and λ where the error is the lowest would provide the optimal λ.

(c) Repeat (a) for variance.

iv: As λ increases from zero, the variance in the model starts to decrease. Since λ is a bias, the larger the bias the smaller the variance there will be. This is the bias-variance tradeoff that Dr. Saunders talked about in the lecture.

(d) Repeat (a) for (squared) bias.

iii: When λ is 0, the beta values are their natural values, and as λ increases the betas approach 0 and bias steadily increases.

(e) Repeat (a) for the irreducible error.

v: Since the irreducible error is coming from the data itself, it remains constant when λ changes. The only way to decrease it is to clean up the data in pre-processing (TowardsDataScience).

2. Question 6.8.9 pg 263

Question 9: In this exercise, we will predict the number of application received using the other variables in the `College` data set.

(a) Split the data set into a training set and a test set.

Setting seed to 702 for reproducibility, and creating an 0.8/0.2 train/test split.

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

```
## MSE of linear model for validation data using all variables:
```

```
## [1] 758588.8
```

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

```
## Best lambda by CV for ridge regression:
```

```
## [1] 378.0387
```

```
## MSE of ridge regression for validation data:
```

```
## [1] 922799.7
```

(d) Fit a lasso model on the training set, with λ chosen by cross-validation. Report the test error obtained, along with the number of non-zero coefficient estimates.

```
## Best lambda by CV for LASSO:
```

```
## [1] 18.81508
```

```
## MSE of LASSO for validation data:
```

```
## [1] 744638.2
```

```
##
## LASSO non-zero coefficients:
```

```
##   (Intercept)    PrivateYes        Accept        Enroll     Top10perc
## -5.790230e+02 -4.364438e+02  1.472023e+00 -2.537644e-01  3.574120e+01
##     Top25perc   P.Undergrad      Outstate    Room.Board      Personal
## -3.848376e+00  2.573436e-02 -6.185067e-02  1.282848e-01  2.405581e-03
##           PhD      Terminal     S.F.Ratio   perc.alumni        Expend
## -5.992569e+00 -3.255498e+00  5.878945e+00 -8.776550e-01  7.070084e-02
##     Grad.Rate
##  5.568614e+00
```

16 of the 18 coefficients are not zero, with only `F.Undergrad` (number of full-time undergrads) and `Books` (estimated book costs) having a coefficient of *0*.

(e) Fit a PCR model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

```
## Data:    X dimension: 619 17
##  Y dimension: 619 1
## Fit method: svdpc
## Number of components considered: 17
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV            4023     3932     2107     2120     1790     1641     1640
## adjCV         4023     3932     2104     2123     1728     1632     1636
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
```

```
## CV        1622     1618     1594      1590      1599      1599      1611
## adjCV     1620     1610     1590      1586      1595      1595      1606
##        14 comps  15 comps  16 comps  17 comps
## CV         1611      1542      1264      1232
## adjCV      1607      1512      1254      1223
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X      31.797    56.76    63.89    69.60    75.04    80.02    83.76    87.36
## Apps    5.156    73.31    73.31    84.14    84.51    84.52    84.90    85.22
##       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
## X       90.54     93.03     95.11     96.74     97.84     98.74     99.35
## Apps    85.51     85.75     85.75     85.82     85.82     85.88     91.07
##       16 comps  17 comps
## X        99.83    100.00
## Apps     92.52     92.87
```

```
## MSE of PCR for validation data:
```

```
## [1] 758588.8
```

The book states to pick the M with the lowest cross-validation error which was acheived with an M of 17.

(f) Fit a PLS model on the training set, with M chosen by cross-validation. Report the test error obtained, along with the value of M selected by cross-validation.

```
## Data:    X dimension: 619 17
##  Y dimension: 619 1
## Fit method: kernelpls
## Number of components considered: 17
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV            4023     1927     1701     1526     1440     1318     1259
## adjCV         4023     1924     1698     1520     1417     1292     1246
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV        1246     1239     1237      1235      1233      1232      1231
## adjCV     1235     1228     1227      1226      1223      1223      1222
##        14 comps  15 comps  16 comps  17 comps
## CV         1232      1232      1232      1232
## adjCV      1223      1223      1223      1223
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X       25.55    46.29    62.26    64.22     66.6    70.12    73.94     76.0
## Apps    78.09    83.68    87.54    91.14     92.6    92.70    92.74     92.8
##       9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
## X       80.27     84.38     87.12     90.34     92.44     95.04     96.83
## Apps    92.83     92.84     92.85     92.86     92.87     92.87     92.87
##       16 comps  17 comps
## X        98.56    100.00
## Apps     92.87     92.87
```

```
## MSE of PLS for validation data:
```

```
## [1] 760622.5
```

(g) 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 five approaches?

```
##                        [,1]
## linear regression 758588.8
## ridge regression  922799.7
## LASSO             744638.2
## PCR               758588.8
## PLS               760622.5
```

Linear regression, LASSO, PCR, and PLS are all fairly comparable in their ability to accurately predict the number of applications received based on the other variables. The ridge regression performed the worst out of all the model methods tested on an 80/20 split of the College data.

3. Question 6.8.11 pg 26

Question 11: We will now try to predict per capita crime rate in the `Boston` data set.

Seed was set to 702 for reproducibility

(a) Try out some of the regression methods explored in this chapter, such as best selection, the lasso, ridge regression, and PCR. Present and discuss results for the approaches that you consider.

```
## Best Exhaustive Selection (12) coefficient for minimum error:
```

```
##   (Intercept)            zn         indus          chas           nox
##  18.274924264   0.045738206  -0.063664843  -0.744077456 -12.317940564
##            rm           dis           rad           tax       ptratio
##   0.510142725  -1.075362634   0.600178320  -0.003456042  -0.291818465
##         black         lstat          medv
##  -0.006406766   0.138043116  -0.223603768
```

```
## Best exhaustive selection error:
```

```
## [1] 14.28401
```

```
## Best lambda by CV for ridge regression:
```

```
## [1] 0.5602717
```

```
## Ridge regression error:
```

```
## [1] 13.90234
```

```
## Best lambda by CV for LASSO:
```

```
## [1] 0.09345895
```

```
## LASSO error:
```

```
## [1] 13.85281
```

```
## Data:    X dimension: 409 13
##  Y dimension: 409 1
## Fit method: svdpc
## Number of components considered: 13
##
## VALIDATION: RMSEP
## Cross-validated using 10 random segments.
##        (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
## CV           9.142    7.651    7.649    7.209    7.221    7.221    7.237
## adjCV        9.142    7.650    7.648    7.206    7.218    7.219    7.233
```

```
##        7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
## CV       7.232    7.144    7.144     7.148     7.130     7.096     7.032
## adjCV    7.227    7.136    7.137     7.139     7.121     7.085     7.021
##
## TRAINING: % variance explained
##       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
## X       48.78    61.08    70.14    76.93    83.32    88.28    91.34    93.61
## crim    30.01    30.17    37.97    38.17    38.19    38.56    38.84    40.57
##       9 comps  10 comps  11 comps  12 comps  13 comps
## X       95.56     97.15     98.53     99.53       100
## crim    41.02     41.26     41.78     42.79        44
```

```
##
## PCR error:
```

```
## [1] 14.29468
```

```
##                                         [,1]
## Best Selection Error (exhaustive)   14.28401
## Ridge Regression Error              13.90234
## Lasso Error                         13.85281
## Pricipal Component Regression Error 14.29468
```

The error of the principal component regression performed the worst out of the 4 that were tested with an error of 14.29. Slightly less than the PCR model was the best selection model using the exhaustive method which, in this case with only having 13 predictors, doesn't take a long time. The error of the best selection method was 14.28. All of the models are similar, but both ridge regression and LASSO were in the high 13's for error, with ridge regression having a test error of 13.9 and LASSO producing the lowest error of 13.85.

(b) 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 validation set error, cross-validation, or some other reasonable alternative, as opposed to using trianing error.

All models above use validation set, and ridge regression, LASSO, and PCR use a cross-validation on the training split of the data for model selection before moving to the testing error calculation. The model which I would propose is the LASSO model, since it uses cv.glmnet on the validation split and produces the lowest error on the testing data. However, they all perform similar to each other, just that LASSO and Ridge Regression perform better than PCR and best selection method.

(c) Does your chosen model involve all of the features in the data set? Why or why not?

```
##
## LASSO model non-zero coefficients:
```

```
##  (Intercept)           zn        indus         chas          nox           rm
##  9.787151305  0.032017111 -0.053739888 -0.522943332 -4.171779026  0.052469831
##          dis          rad      ptratio        black        lstat         medv
## -0.625420312  0.497500013 -0.118022257 -0.007568194  0.118962490 -0.129744927
```

Using the example in the lab code to find the coefficients that are 0, `age` and `tax` result in a coefficient of 0. From the book, this means that the LASSO with the λ chosen by cross-validation contains only 11 of the variables (`zn`, `indus`, `chas`, `nox`, `rm`, `dis`, `rad`, `ptratio`, `black`, `lstat`, and `medv`) out of the 13 predictors in the `Boston` data set.

References:

- Towards Data Science
- ISLR Book

- Stack Exchange