MATH 3190 Homework 6

akhilbandhu

April 2022

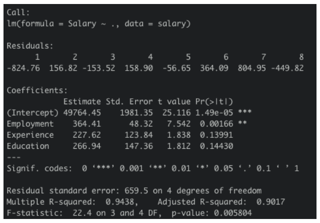
In this homework you will practice using cross-validation to fit data using LASSO and K-nearest neighbor models. Please upload to your GitHub an R Markdown document answering the following:

1. (20 points) A researcher wants to determine how employee salaries at a certain company are related to the length of employment, previous experience, and education. The researcher selects eight employees from the company and obtains the data shown below (the dataset is available as a tibble in the .Rmd).

Salary	Employment	Experience	Education
$57,310	10	2	16
$57,380	5	6	16
$54,135	3	1	12
$56,985	6	5	14
$58,715	8	8	16
$60,620	20	0	12
$59,200	8	4	18
$60,320	14	6	17

(a) Fit a standard least squares regression model to these data and interpret the results. After looking at the statistical significance of the βs, which covariates would you include in a final model?

i. After fitting a standard least squares model, the intercept and employment are the only two variables that are significantly significant. So, only Employment will be add into the final model.



```
Call:
lm(formula = Salary ~ ., data = salary)

Residuals:
      1       2       3       4       5       6       7       8 
-824.76  156.82 -153.52  158.90  -56.65  364.09  804.95 -449.82 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 49764.45    1981.35  25.116 1.49e-05 ***
Employment    364.41      48.32   7.542  0.00166 ** 
Experience    227.62     123.84   1.838  0.13991    
Education     266.94     147.36   1.812  0.14430    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 659.5 on 4 degrees of freedom
Multiple R-squared:  0.9438,	Adjusted R-squared:  0.9017 
F-statistic:  22.4 on 3 and 4 DF,  p-value: 0.005804
```

(b) Use `glmnet` to fit a LASSO model to these covariates. Try λ=1000, 800, 500, and 1. How do the results compare to each other and the least squares model?

(c) Which LASSO model (i.e. λ) would you select? (note you are not just restricted to λ values of 1000, 800, 500, and 1). Justify your answer.

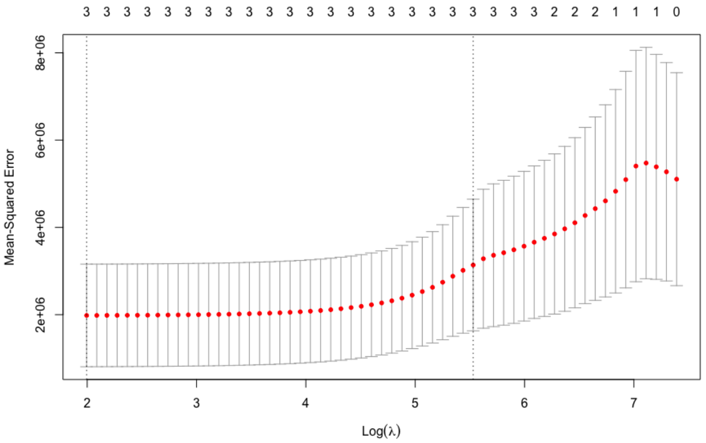
i. Using λ=1000, the model suggests to use only intercept and employment in the fitting a model.

ii. Using λ=800, it suggests to use employment and education in fitting the final model.

iii. Using λ=500, it suggests to use employment and education in fitting the final model.

iv. Using λ=1, it suggests to use employment, education, and experience in fitting the final model.

v. I went a step further and tried cross validation to pick the best lambda, it turned out to be 7.35. Here is an image showing how MSE decreases as we perform cross validation. This best lambda model suggests to use all variables availabe to build the last model (best fit model).



(d) Use `glmnet` to fit a Ridge regression model to these data. Try λ=1000, 800, 500, and 1. How do these results differ from the least squares and LASSO models?

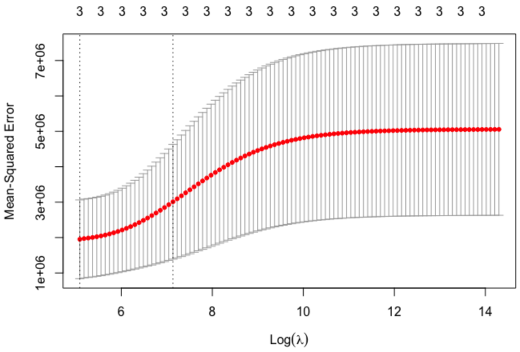
i. Using λ=1000, the model suggests to use all three variables in the fitting the final model.

ii. Using λ=800, it suggests a similar model to the one above.

iii. Using λ=500, it suggests a similar model stated with λ=1000.

iv. Using λ=1, it suggest using all three variables

v. I went a step further and tried cross validation to pick the best lambda, it turned out to be 162.104. Here is an image showing how MSE decreases as we perform cross validation. This best lambda model suggests to use all variables availabe to build the last model (best fit model).



2. (20 points) The `cereal.csv` dataset provides nutritional information on nearly 80 common breakfast cereals. The ’rating’ column provides an overall rating for each cereal (possibly from Consumer Reports?). Use a LASSO regression model to identify the best predictors of cereal rating. Evaluate the model for λ values of 8, 5, 3, and 1 (among others). Which λ would you choose and why? Which covariates best explain the rating?

   (a) First, some cleaning needs to be done on the creating the models, like creating indicator variables. All the is shown in the R code file.

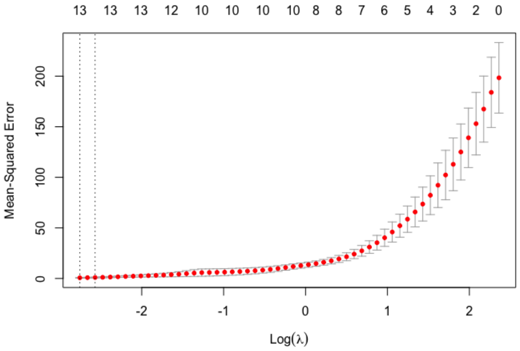
   (b) Fitting a model with λ=8, it suggests that we choose calories and sugars as the two variables in the model.

   (c) Fitting a model with λ=5, it suggests that we use calories, fiber, and sugars as the three variables in the final model.

   (d) Fitting a model with λ=3, it suggests that we use calories, fat, sodium, fiber, sugars, and manufacturer type ’N’ in building the final model.

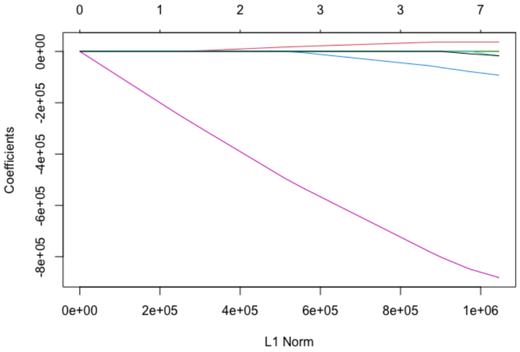
   (e) Fitting a model with λ=1, it suggests that we use calories, protein, fat, sodium, fiber, sugars, manufacturer type ’G’, and manufacturer type ’N’ in building the model.

   (f) After taking a little further and performing cross-validation, the minimum lambda came out to be 0.06356, the next plot shows how MSE reduces. This suggests to use calories, protein, dat, sodium, fiber, carbo, sugars, potass, vitamins, maufacturer type ’G’, maufacturer type ’K’, maufacturer type ’N’, maufacturer type ’Q’.



3. (20 points) An automobile consulting company wants to understand the factors on which the pricing of cars depends. Use an Elastic Net model and the `car_price_prediction.csv` dataset to determine which variables are significant in predicting the price of a car. Use cross-validation to find an optimal value for λ. Interpret your final model.

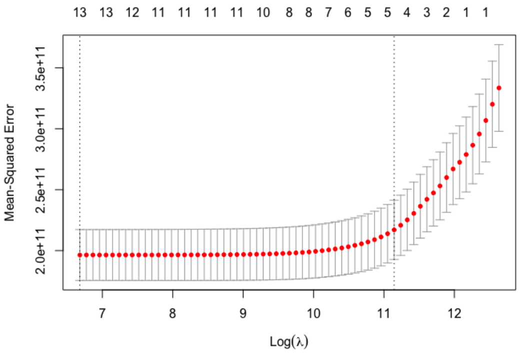
   (a) This following plot is what the elastic net depicts.



   (b) After using cross-validation, the best lambda value turns out to be 795.931

   (c) Model interpretation, this model suggests to use km driven, transmission, CNG type fuel, diesel fuel, electric fuel, petrol fuel, individual, dealer type sellers, first owner, fourth and above owners, test drive car owners, and third owners as variables in predicting selling price of a car.

   (d) As the number of kilometers increases by 1 unit, the selling price of the vehicle goes down by 2.089, holding all others constant.



(e) I can give interpretations of each variable but I wasn't sure if that is what you wanted.