

**PSYC 560**

**Rita Qifan Yang (260893989)**

**February 23, 2023**

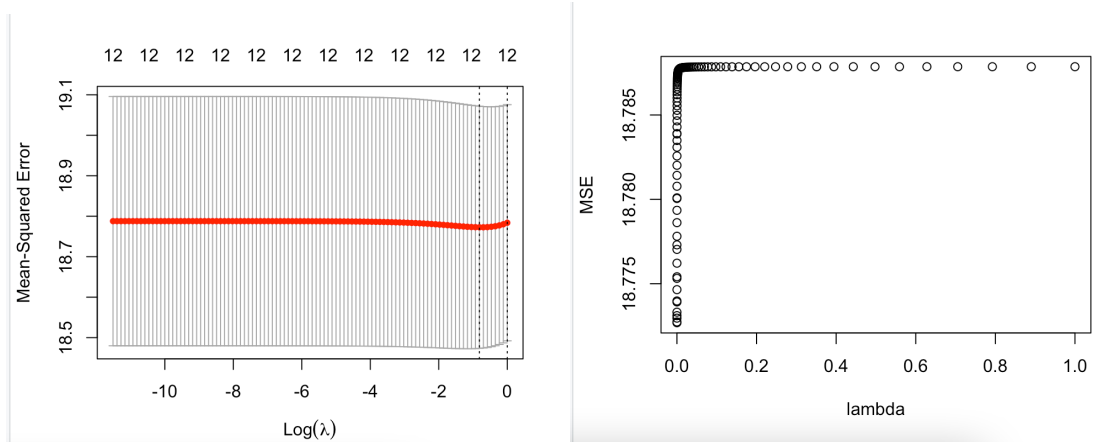
**Assignment #3: Shrinkage and Data Reduction Methods**

[Q1] The data file **mobilephone2\_training.csv** is part of large European survey data for mobile phone customers. It includes the following variables:

Suppose that you are interested in predicting the duration of remaining a customer (dura24) using 12 predictors (variables 2 – 12).

1. dura24: the duration of remaining a customer in the past 24 months
2. age
3. Ssound: satisfaction with sound quality (1 - 10)
4. Sglob: satisfaction with overall network functioning (1 - 10)
5. Svmail: satisfaction with voicemail (1 - 10)
6. Sinfor: satisfaction with information (1 - 10)
7. Sprice: satisfaction on price (1 - 10)
8. Spromo: satisfaction on promotion (1 - 10)
9. Sadver: satisfaction with the advertisement (1 - 10)
10. Sperson: satisfaction with the salesperson (1 - 10)
11. Simage: satisfaction on company image (1 - 10)
12. Sglobal: global satisfaction (1 – 10)
13. Lintent: intention of loyalty (1 - 10)

(1) Apply a series of ridge regression to the data, considering 100 candidate values of the tuning parameter  $\lambda$  within the interval  $[10^{-5}, 1]$ . Based on 10-fold cross validation, plot all MSE values against the candidate values of  $\lambda$  and choose the optimal  $\lambda$  value. Given the chosen  $\lambda$  value, re-run a ridge regression for the data and report its coefficient estimates. (2 points)



At seed = 42, we generate a sequence of 100 values from  $10^{-5}$  to 1, and we find the optimal  $\lambda = 0.4430621$  based on 10-fold cross validation, then we re-run the ridge regression using  $\lambda = 0.4430621$ , then we get the coefficient estimates as follows:

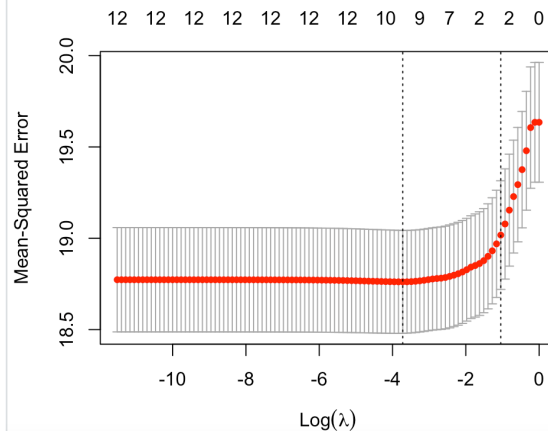
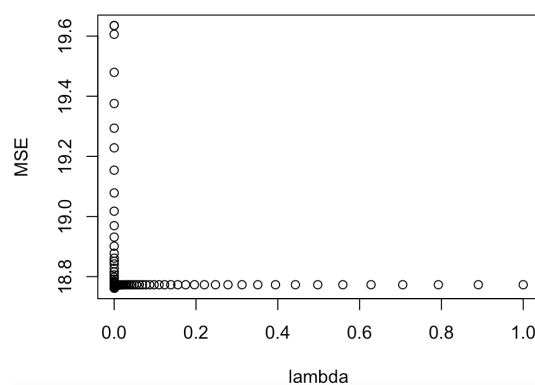
```

s0
(Intercept) 15.343174382
age          0.066382545
ssound       0.055948175
sglob        0.080414500
svmail       -0.005857497
sinfor       0.061046521
sprice       0.024152055
spromo       0.048993454
sadver       -0.106313829
sperson      -0.073973560
simage       -0.153642246
sglobal      -0.023394492
lintent      0.198328487

```

(2) Apply a series of the lasso to the data, considering 100 candidate values of the tuning parameter  $\lambda$  within the interval  $[10^{-5}, 1]$ . Based on 10-fold cross validation, plot all MSE values against the candidate values of  $\lambda$  and choose the optimal  $\lambda$  value. Given the chosen  $\lambda$  value,

re-run the lasso for the data and report its coefficient estimates. (2 points)



At seed = 42, we generate a sequence of 100 values from  $10^{-5}$  to 1, and we find the optimal  $\lambda = 0.2420128$  based on 10-fold cross validation, then we re-run the lasso regression using  $\lambda = 0.2420128$ , then we get the coefficient estimates as follows:

```

              s0
(Intercept) 15.299091281
age          0.070929659
ssound       0.043658908
sglob        0.070234621
svmail       .
sinfor       0.041688576
sprice       0.002682897
spromo       0.040784117
sadver       -0.088932598
sperson      -0.068532178
simage       -0.162160281
sglobal      .
lintent      0.209719283

```

(3) Apply a principal component regression to the data and choose the number of components that minimizes the cross-validated RMSE. Plot the RMSE values against the number of components. Also, describe how much variance of the predictors and the response the chosen number of components explain (2 points)

```
> summary(pcr.fit)
```

```
Data: X dimension: 3000 12
```

```
      Y dimension: 3000 1
```

```
Fit method: svdpc
```

```
Number of components considered: 12
```

```
VALIDATION: RMSEP
```

```
Cross-validated using 10 random segments.
```

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps
CV	4.432	4.431	4.399	4.366	4.364	4.344	4.345	4.348	4.349	4.335
adjCV	4.432	4.431	4.398	4.365	4.364	4.343	4.344	4.347	4.349	4.334

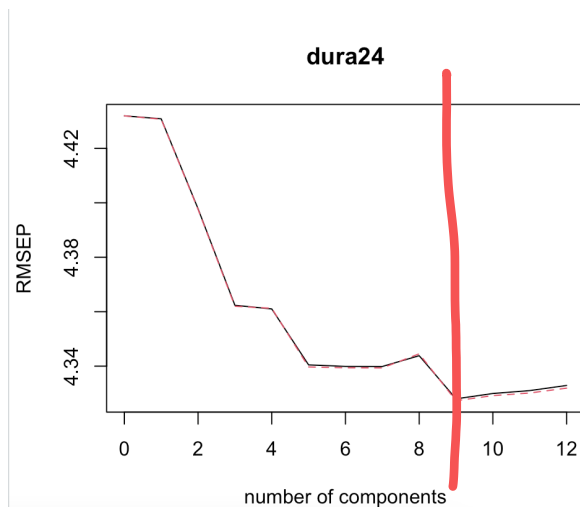
	10 comps	11 comps	12 comps
CV	4.336	4.337	4.338
adjCV	4.335	4.336	4.337

```
TRAINING: % variance explained
```

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps	10 comps
X	38.2731	47.603	56.262	63.684	70.815	77.615	83.543	87.970	92.276	95.328
dura24	0.1165	1.674	3.311	3.399	4.386	4.452	4.475	4.493	5.139	5.142

	11 comps	12 comps
X	98.214	100.000
dura24	5.175	5.181



From the plot above, we can see that choosing 9 components will minimize the cross-validated RMSE. 92.276% variance of the predictors are explained by the 9 chosen components, 5.139% of variance of the response is explained by the 9 chosen components.

(4) Apply a partial least squares regression to the data and choose the number of components that minimizes the cross-validated RMSE. Plot the RMSE values against the number of components. Also, describe how much variance of the predictors and the response the chosen number of components explain (2 points)

```
> summary(plsr.fit)
```

Data: X dimension: 3000 12  
Y dimension: 3000 1  
Fit method: kernelpls  
Number of components considered: 12

VALIDATION: RMSEP  
Cross-validated using 10 random segments.

	(Intercept)	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps
CV	4.432	4.353	4.331	4.331	4.331	4.332	4.333	4.333	4.333	4.333
adjCV	4.432	4.352	4.330	4.330	4.330	4.331	4.332	4.332	4.332	4.332

	10 comps	11 comps	12 comps
CV	4.333	4.333	4.333
adjCV	4.332	4.332	4.332

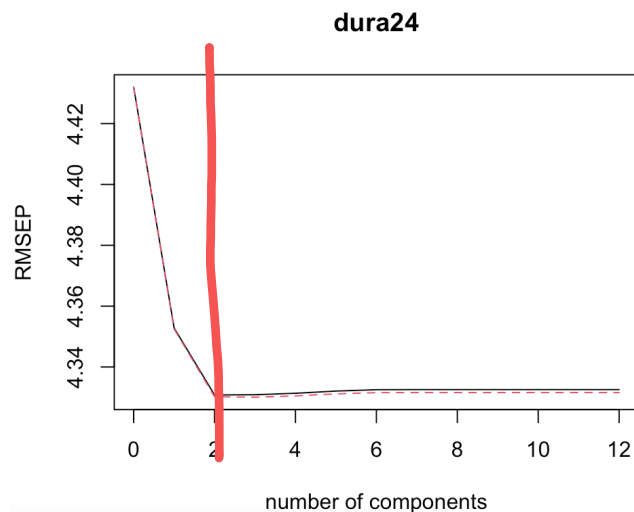
  

TRAINING: % variance explained

	1 comps	2 comps	3 comps	4 comps	5 comps	6 comps	7 comps	8 comps	9 comps	10 comps
X	18.646	46.580	52.682	59.12	65.309	70.085	73.691	78.849	85.415	89.894
dura24	3.942	5.014	5.167	5.18	5.181	5.181	5.181	5.181	5.181	5.181

	11 comps	12 comps
X	95.466	100.000
dura24	5.181	5.181



We choose 2 components to minimize the cross-validated RMSE. 46.580% variance of the predictors are explained by the 2 chosen components, 5.014% of variance of the response is explained by the 2 chosen components.

[Q2] Using the estimated solutions obtained from the above four methods, report their MSE values in the test sample (**mobilephone2\_test.csv**) and conclude which method appears to perform best in terms of the test MSE values (2 points).

Using the solutions obtained from ridge regression, lasso regression, principal component regression, and partial least squares regression, we find the MSE values in the test sample(**mobilephone2\_test.csv**) as follows:

MSE\_RIDGE: 19.74827

MSE\_LASSO: 19.76836

MSE\_PCR: 19.76372

MSE\_PLSR: 19.81273

We conclude that the ridge regression appears to perform the best in terms of having the lowest test MSE value.