Problem Set 2

Group members: Xuchen Liu, Binqian Chai, He Jiang, Ella Lin

1.) Recall that lasso selects a sparse model by zeroing out covariates. We will run a bootstrap experiment to see whether this model selection procedure is stable across different samples. In general, one should be cautious when applying bootstrap with the lasso (bootstrap CIs can fail to cover the true parameter with high probability), but we're just using it to get a sense of the stability of lasso model selection.

- a.) Run a lasso regression of log(yspend) on xweb, using 5-fold cross-validation to pick λ. Report the indices of the nonzero coefficients (you don't have to report their names).

```
library(gamlr)
set.seed(0)
cv.lasso_model <- cv.gamlr(xweb, log(yspend), verb=TRUE)
```

```
## fold 1,2,3,4,5,done.
```

```
nonzero_coef<-which(coef(cv.lasso_model, select="min")!=0)
cat("The indices of the nonzero coefficients are", nonzero_coef, ".", sep=" ")
```

```
## The indices of the nonzero coefficients are 1 3 5 8 9 10 12 17 27 28 31 38 40 46 47 50 51 57 63 64 73 75 83 87 89 98 102 106 117 120 127 131 135 145 152 153 155 163 165 169 176 177 183 189 190 191 196 206 207 208 209 214 215 219 226 237 249 254 259 261 275 276 279 280 281 282 286 288 289 292 296 302 304 308 309 317 329 340 344 346 347 350 352 361 363 366 368 370 373 376 382 389 397 398 404 429 444 445 446 454 460 461 462 465 466 476 477 483 484 488 502 505 506 508 509 518 522 526 530 531 537 551 553 560 562 570 573 585 587 588 595 601 606 608 609 618 620 621 623 627 638 643 656 660 669 670 678 680 691 692 695 697 700 706 712 725 730 731 745 753 755 756 762 767 770 772 773 779 782 783 792 803 808 814 816 822 826 827 828 838 840 843 845 846 847 848 853 854 856 865 873 884 892 894 895 897 900 905 906 907 910 916 917 919 920 926 929 935 936 940 943 945 963 964 968 971 973 980 982 983 984 987 989 993 995 1001 .
```

- b.) Redraw a single "bootstrap" sample (same sample size as the original sample) by sampling from yspend and xweb with replacement. Run the lasso regression from part (a) on this bootstrap sample.

```
set.seed(0)
boot_sample_indices <- sample.int(length(yspend), replace=TRUE)
bootsample_yspend <- yspend[boot_sample_indices]
bootsample_xweb <- xweb[boot_sample_indices,]
cv.boot_lasso_model <- cv.gamlr(bootsample_xweb, log(bootsample_yspend), verb=TRUE)
```

```
## fold 1,2,3,4,5,done.
```

```
boot_nonzero_coef<-which(coef(cv.boot_lasso_model, select="min")!=0)
```

- i.) Report the indices of the nonzero coefficients.

```
cat("The indices of the nonzero coefficients are", boot_nonzero_coef, ".", sep=" ")
```

```
## The indices of the nonzero coefficients are 1 3 5 8 9 10 12 14 17 20 22 23 26 27 31 35 36 37 38 40 42 46 47 48 56 57 61 62 63 64 67 70 73 78 85 88 92 95 96 98 99 105 106 111 117 127 131 133 135 136 137 139 145 152 153 155 160 163 165 166 167 168 169 171 173 174 177 178 179 180 181 183 185 186 189 190 191 199 201 206 207 208 212 213 215 216 219 222 233 240 245 251 253 254 267 276 280 281 286 288 289 292 294 304 306 308 309 310 312 326 327 333 336 337 339 340 342 345 347 349 350 351 352 359 361 362 363 365 366 368 370 373 376 378 380 381 382 383 384 385 389 391 394 397 398 401 404 405 408 409 412 413 414 418 421 424 425 428 429 431 432 434 435 441 444 445 446 449 450 454 455 459 460 461 464 465 466 470 471 476 477 480 481 483 484 488 490 492 494 497 498 500 502 505 506 507 508 511 512 516 517 518 520 521 529 530 531 537 539 540 541 542 543 544 549 551 552 553 559 560 562 565 570 571 572 578 583 584 588 591 595 598 600 601 602 606 607 608 609 612 614 616 620 621 622 623 624 626 627 631 633 638 641 642 647 651 656 660 661 662 665 669 670 672 677 678 679 680 682 686 688 691 692 693 697 700 705 706 709 712 715 721 723 725 730 731 732 738 740 745 747 748 750 752 753 756 763 764 765 767 770 772 773 774 777 779 782 791 792 793 801 803 808 812 813 814 816 819 822 824 825 827 828 829 833 834 835 838 840 843 845 846 848 849 850 851 852 853 854 858 859 860 865 870 873 879 884 892 895 897 900 906 907 908 909 910 911 914 915 916 917 919 924 926 928 929 934 935 936 938 940 943 944 945 950 954 957 961 963 965 968 970 971 973 974 982 983 984 985 986 987 989 995 1001 .
```

- ii.) Report the indices of the coefficients that are nonzero only for the bootstrap sample.

```
bootonly_nonzero_coef<-setdiff(boot_nonzero_coef, nonzero_coef)
cat("The indices of the coefficients that are nonzero only for the bootstrap sample are", bootonly_nonzero_coef, ".", sep=" ")
```

```
## The indices of the coefficients that are nonzero only for the bootstrap sample are 14 20 22 23 26 35 36 37 42 48 56 61 62 67 70 78 85 88 92 95 96 99 105 111 133 136 137 139 160 166 167 168 171 173 174 178 179 180 181 185 186 199 201 212 213 216 222 233 240 245 251 253 267 294 306 310 312 326 327 333 336 337 339 342 345 349 351 359 362 365 378 380 381 383 384 385 391 394 401 405 408 409 412 413 414 418 421 424 425 428 431 432 434 435 441 449 450 455 459 464 470 471 480 481 483 484 488 490 492 494 497 498 500 502 505 506 507 508 511 512 516 517 520 521 529 539 540 541 542 543 544 549 552 553 559 560 562 565 570 571 572 578 583 584 591 598 600 601 602 606 607 608 609 612 614 616 622 624 626 631 633 641 642 647 651 661 662 665 672 677 679 682 686 688 693 705 709 715 721 723 732 738 740 747 748 750 752 763 764 765 774 777 791 793 801 812 813 819 824 825 829 833 834 835 849 850 851 852 858 859 860 870 879 908 909 911 914 915 924 928 934 938 944 950 954 957 961 965 970 974 985 986 .
```

- iii.) Report the indices of the coefficients that are nonzero only for the original sample.

```
originalonly_nonzero_coef<-setdiff(nonzero_coef, boot_nonzero_coef)
cat("The indices of the coefficients that are nonzero only for the original sample are", originalonly_nonzero_coef, ".", sep=" ")
```

```
## The indices of the coefficients that are nonzero only for the original sample are 28 50 51 75 83 87 89 102 120 176 196 209 214 226 237 249 259 261 275 279 282 296 302 317 329 344 346 462 509 522 526 573 585 587 618 643 695 755 762 783 808 826 847 856 894 905 920 964 980 993 .
```

- iv.) Report the indices of the coefficients that are nonzero for both samples.

```
both_nonzero_coef<-intersect(nonzero_coef, boot_nonzero_coef)
cat("The indices of the coefficients that are nonzero for both samples are", both_nonzero_coef, ".", sep=" ")
```

```
## The indices of the coefficients that are nonzero for both samples are 1 3 5 8 9 10 12 17 27 31 38 40 46 47 57 63 64 73 98 106 117 127 131 135 145 152 153 155 163 165 169 177 183 189 190 191 206 207 208 215 219 254 276 280 281 286 288 289 292 304 308 309 340 347 350 352 361 363 366 368 370 373 376 382 389 397 398 404 429 444 445 446 454 460 461 465 466 476 477 483 484 488 502 505 506 508 518 530 531 537 551 553 560 562 570 588 595 601 606 608 609 620 621 623 627 638 656 660 669 670 678 680 691 692 697 700 706 712 725 730 731 745 753 756 767 770 772 773 779 782 792 803 814 816 822 827 828 838 840 843 845 846 848 853 854 865 873 884 892 895 897 900 906 907 910 916 917 919 926 929 935 936 940 943 945 963 968 971 973 982 983 984 987 989 995 1001 .
```

- c.) Based on these results, since there are differences within the set of nonzero_coef and that of boot_nonzero_coef, the set of nonzero coefficients selected by the lasso seem to be unstable across random draws of the data.

2.) We will run an out-of-sample experiment to see how well cross-validated lasso performs. First, draw a random sample of size n = 8,000 from the original 10,000 observations, without replacement. We will refer to these observations X1,…,Xn and Y1,…,Yn as the "estimation" sample. We will refer to the remaining m = 2,000 observations Xn+1,…,Xn+m and Yn+1,…,Yn+m as the "holdout" sample.

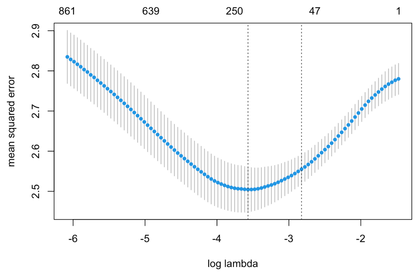
```
set.seed(0)
est_sample_ind <- sample.int(n = length(yspend), size = 8000, replace = FALSE)
```

- a.) Run 5-fold cross-validated lasso of log(yspend) on xweb on the estimation sample of n = 8,000 observations. Report a plot of the out-of-sample cross validation error as a function of λ (you can use the plot command on the cv.gamlr object as in the lecture notes).

```
library(gamlr)
est_sample_xweb <- xweb[est_sample_ind,]
est_sample_yspend <- yspend[est_sample_ind]
cv.oos_lasso_model <- cv.gamlr(est_sample_xweb, log(est_sample_yspend), verb=TRUE)
```

```
## fold 1,2,3,4,5,done.
```

```
plot(cv.oos_lasso_model)
```



- b.) Let λmin denote the optimal λ from part (a). Using the output of the lasso model with λ = λmin, compute the predicted value of log(yspend) at each X1,…, Xn in the estimation sample. Compute the in-sample prediction error of the cross-validated lasso:

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where Yi denotes log(yspend) for observation i.

```
predict_yspend <- predict(cv.oos_lasso_model, est_sample_xweb, select="min")
sum_of_square = 0
for (i in 1:8000) {
  a = (log(est_sample_yspend[i]) - predict_yspend[i])**2
  sum_of_square = sum_of_square + a
}
is_prediction_error = sum_of_square/8000
cat("The in-sample prediction error of the cross-validated lasso is ", is_prediction_error, ".", sep="")
```

```
## The in-sample prediction error of the cross-validated lasso is 2.304756.
```

- c.) Now compute the out-of-sample prediction error, using the holdout sample. Using the output of the lasso model with λ = λmin computed using the estimation sample i=1,…,n,compute the predicted value of log(yspend) at each Xn+1,…,Xn+m in the holdout sample. Call these predicted values Y'n+1, . . . , Y'n+m. Compute the out-of-sample prediction error of the cross-validated lasso model on the holdout sample:

$$\frac{1}{m}\sum_{i=1}^{m}(Y_{n+i} - \hat{Y}_{n+i})^2$$

where Yn+i denotes log(yspend) for observation n+i.

```
oos_ind <- c()
for (i in 1:10000) {
  if (!(i %in% est_sample_ind)) {
    oos_ind <- append(oos_ind, i)
  }
}
holdout_sample_xweb <- xweb[oos_ind,]
holdout_sample_yspend <- yspend[oos_ind]
predict_holdout_yspend <- predict(cv.oos_lasso_model, holdout_sample_xweb, select="min")
sum_of_square2 = 0
for (i in 1:2000) {
  b = (log(holdout_sample_yspend[i]) - predict_holdout_yspend[i])**2
  sum_of_square2 = sum_of_square2 + b
}
oos_prediction_error = sum_of_square2/2000
cat("The out-of-sample prediction error of the cross-validated lasso model on the holdout sample is ", oos_prediction_error, ".", sep="")
```

```
## The out-of-sample prediction error of the cross-validated lasso model on the holdout sample is 2.442623.
```

How does this compare to the results from part (b)?

The out-of-sample prediction error is pretty close to in-sample prediction error in part (b). Based on the prediction errors, the model performs slightly better on the in-sample (trained data) as the in-sample prediction error is slightly lower than out-of-sample one. But since the in-sample and out-of-sample prediction errors are pretty close, the model fits data well overall.