ECON 490 HW4

1. su sat

```
. su sat

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
         sat |      4,137    1030.331    139.4014        470       1540
```

   There are 4137 Observations.

2. reg colgpa athlete, rob

```
. //q2
. reg colgpa athlete, rob

Linear regression                               Number of obs     =      4,137
                                                F(1, 4135)        =      41.64
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0083
                                                Root MSE          =     .65596

------------------------------------------------------------------------------
             |               Robust
      colgpa | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     athlete |  -.2845336   .0440938    -6.45   0.000    -.3709811    -.198086
       _cons |   2.666028   .0104907   254.13   0.000     2.645461    2.686596
------------------------------------------------------------------------------
```

3. colgpa=2.666028-0.2845336athlete
   From the model inference, if the person is not an athlete, his/her gpa is 2.666028. For 1 unit increase in athlete, there will be 0.2845336 unit of College GPA decrease, and since the athlete is a logical variable, thus we could also say that when study is an athlete, the estimated colgpa will decrease for 0.2845336.

4. The error term has to be uncorrelated with the explanatory variable, which is athlete. Besides, this regression also needs to include the correct variable and additional variables, like athlete.

5. It is normally impossible for me to estimate the effect

of athlete is an unbiased estimate since there are many factors that are included in error term is related to athlete, such as the female, shrank and etc. terms.

6. ge satsq = sat^2
ge satcub = sat^3
ge verbmathsq = verbmath^2
ge verbmathcub = verbmath^3

```
local x "hsize hsrank sat satsq satcub female verbmath verbmathsq verbmathcub i.tothrs"
reg colgpa athlete `x', rob
```

```
. ge satcub = sat^3

. ge verbmathsq = verbmath^2

. ge verbmathcub = verbmath^3

.
. local x "hsize hsrank sat satsq satcub female verbmath verbmathsq verbmathcub i.tothrs"

. reg colgpa athlete `x', rob

Linear regression                               Number of obs     =      4,137
                                                F(130, 4002)      =          .
                                                Prob > F          =          .
                                                R-squared         =     0.3335
                                                Root MSE          =     .54663

------------------------------------------------------------------------------
             |               Robust
      colgpa | Coefficient  std. err.      t    P>|t|     [95% conf. interval]
-------------+----------------------------------------------------------------
     athlete |   .1262987   .0386025     3.27   0.001     .0506162    .2019812
       hsize |   .0579356   .0064037     9.05   0.000     .0453809    .0704904
      hsrank |  -.0034132   .0001755   -19.45   0.000    -.0037573   -.0030692
         sat |  -.0043144   .0029291    -1.47   0.141     -.010057    .0014282
       satsq |   4.32e-06   2.91e-06     1.48   0.138    -1.39e-06      .00001
      satcub |  -9.02e-10   9.54e-10    -0.94   0.345    -2.77e-09    9.69e-10
      female |   .1676872   .0186113     9.01   0.000     .1311987    .2041757
    verbmath |   -.564694   2.530723    -0.22   0.823    -5.526321    4.396933
  verbmathsq |   .5349674    2.58299     0.21   0.836    -4.529131    5.599066
 verbmathcub |   -.189332   .8553011    -0.22   0.825    -1.866199    1.487534
             |
      tothrs |
          9  |   1.166759   .0333818    34.95   0.000     1.101312    1.232206
         10  |  -.6082065   .1114951    -5.46   0.000     -.826799    -.389614
         11  |  -.2672486   .1333036    -2.00   0.045    -.5285978   -.0058993
         12  |  -.3602308   .0950212    -3.79   0.000    -.5465252   -.1739364
         13  |  -.3626754   .0659108    -5.50   0.000    -.4918972   -.2334535
         14  |  -.2184005    .056066    -3.90   0.000    -.3283211   -.1084799
         15  |  -.1654739   .0476254    -3.47   0.001    -.2588461   -.0721017
         16  |  -.0295001   .0410759    -0.72   0.473    -.1100316    .0510315
         17  |  -.0169598   .0384945    -0.44   0.660    -.0924305    .0585108
         18  |  -.0282522   .0549501    -0.51   0.607     -.135985    .0794806
         19  |   .0518595   .1106177     0.47   0.639    -.1650128    .2687319
         20  |   .0719334   .1134775     0.63   0.526    -.1505457    .2944126
         21  |   .2188149   .2303115     0.95   0.342    -.2327238    .6703536
```

100	.1658921	.1154027	1.44	0.151	-.0603615	.3921456
101	-.1704272	.0956693	-1.78	0.075	-.3579923	.0171379
102	-.115947	.1024073	-1.13	0.258	-.3167224	.0848285
103	.0933916	.1388618	0.67	0.501	-.1788549	.3656381
104	.0159774	.1206981	0.13	0.895	-.2206582	.252613
105	.0610474	.1386461	0.44	0.660	-.2107761	.332871
106	-.1163251	.0907494	-1.28	0.200	-.2942445	.0615943
107	-.1383102	.105389	-1.31	0.189	-.3449313	.0683109
108	.0567691	.0713835	0.80	0.427	-.0831823	.1967205
109	-.1651594	.1036198	-1.59	0.111	-.3683119	.0379931
110	.1423876	.0713432	2.00	0.046	.0025152	.2822601
111	.1001366	.0706931	1.42	0.157	-.0384611	.2387344
112	.2538851	.0768791	3.30	0.001	.1031592	.4046111
113	.1263419	.0968201	1.30	0.192	-.0634793	.3161632
114	.1507692	.0741414	2.03	0.042	.0054107	.2961277
115	.0368089	.0966089	0.38	0.703	-.1525983	.2262161
116	.1468627	.0666597	2.20	0.028	.0161726	.2775528
117	.0667038	.0792124	0.84	0.400	-.0885965	.2220041
118	.2140968	.0998143	2.14	0.032	.0184051	.4097885
119	.0807495	.0994175	0.81	0.417	-.1141642	.2756631
120	.0077251	.1108863	0.07	0.944	-.2096737	.2251239
121	.1382967	.0933334	1.48	0.138	-.0446888	.3212822
122	-.0993638	.0968533	-1.03	0.305	-.2892503	.0905227
123	-.0089179	.1374937	-0.06	0.948	-.2784821	.2606462
124	.2962014	.1153153	2.57	0.010	.0701191	.5222837
125	.1162994	.1604508	0.72	0.469	-.1982734	.4308722
126	.0237637	.1325793	0.18	0.858	-.2361655	.283693
127	-.231197	.1652303	-1.40	0.162	-.5551404	.0927465
128	-.0452173	.2655734	-0.17	0.865	-.5658891	.4754545
129	-.4997246	.1768836	-2.83	0.005	-.846515	-.1529343
130	-.4152458	.7082421	-0.59	0.558	-1.803795	.9733032
131	.3287013	.1187478	2.77	0.006	.0958896	.5615131
132	.1290691	.1713426	0.75	0.451	-.2068578	.464996
133	.1815349	.1635762	1.11	0.267	-.1391656	.5022354
134	-.647872	.0198939	-32.57	0.000	-.6868751	-.6088688
136	-.8524644	.005155	-165.37	0.000	-.862571	-.8423578
137	.5507171	.2025983	2.72	0.007	.1535115	.9479227
_cons	3.680433	1.270118	2.90	0.004	1.190293	6.170572

```
.
end of do-file
```

7. the RMSE is 0.54663.
   the MSE is 0. 54663^2=0.29880.

8. set seed 1234
   local x hsize hsrank sat satsq satcub female verbmath verbmathsq verbmathcub i.tothrs
   lasso linear colgpa athlete `x'

```
Lasso linear model                          No. of obs        =      4,137
                                            No. of covariates =        131
Selection: Cross-validation                 No. of CV folds   =         10
```

ID	Description	lambda	No. of nonzero coef.	Out-of-sample R-squared	CV mean prediction error
1	first lambda	.2733989	0	0.0006	.4334285
39	lambda before	.0079697	70	0.2991	.3039925
* 40	selected lambda	.0072617	74	0.2992	.3039314
41	lambda after	.0066166	79	0.2992	.3039468
46	last lambda	.0041554	100	0.2975	.304667

* lambda selected by cross-validation.

9. So basically, as we are using Lasso for prediction, we are trying to reach our assumption that there are few variables related to the number of observations in the sample in the unknown true model, and as lambda increase, Lasso will use the penalty regression that shrinkage those variables that has 0 coefficients, since their penalty is larger than contribution, and they are not contributing enough to the model, and this process select the most important covariates out of the potential explanatory variables list, and preventing the overfitting for the model.
   When lambda is small, Lasso will have almost same solution to OLS, and we are choosing the lambda by cross validation, with the highest out-of-sample R-square in this case.

10. lassoselect lambda=0.0072617
    lassoinfo

```
. lassoinfo

    Estimate: active
     Command: lasso
```

Dependent variable	Model	Selection method	Selection criterion	lambda	No. of selected variables
colgpa	linear	user	user	.0072617	74

There are 74 non-zero variables are retained at the selected lambda. (also can obtain from last question, ID=40)

11. lassoselect lambda=0.0072617
lassocoef

```
. do "C:\Users\suisx\AppData\Local\Temp\STD97d4_000000.tmp"

. lassoselect lambda=0.0072617
ID = 40  lambda = .0072617 selected

. lassocoef
```

	active
athlete	x
hsize	x
hsrank	x
satsq	x
satcub	x
female	x
verbmathsq	x

The variable athlete is still **retained** at the selected value of lambda.

12.

```
Lasso linear model                          No. of obs        =      4,137
                                            No. of covariates =        131
Selection: Cross-validation                 No. of CV folds   =         10
```

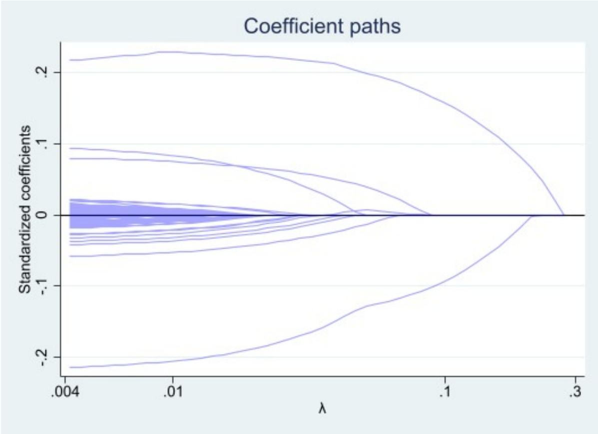
ID	Description	lambda	No. of nonzero coef.	Out-of-sample R-squared	CV mean prediction error
1	first lambda	.2733989	0	0.0006	.4334285
39	lambda before	.0079697	70	0.2991	.3039925
* 40	selected lambda	.0072617	74	0.2992	.3039314
41	lambda after	.0066166	79	0.2992	.3039468
46	last lambda	.0041554	100	0.2975	.304667

* lambda selected by cross-validation.

The Mean prediction Error of the Lasso at the lambda=0.0072617 is: 0.3039314.

From the Question 7, the MSE of the OLS regression (full set of explanatory variables) is 0. 54663^2=0.29880.

13. coefpath, xunit(lnlambda) minmax



14. The Lasso algorithm will iterates in descending order through the lambda, while lambda is decreasing from lambda-Max, there are more and more variables coefficients are shrinking, reached a 0 coefficients, determined and then drop out from the Lasso algorithm. During the process, lambda will keep decreasing and until all of the nonzero coefficient are presented at this lambda.
The shrinking of the graph represents the higher penalty of the lasso algorithm while lambda is increasing. The selection effect of the Lasso will depend on the cross-validation, which could also be plotted by: cvplot, graphregion(color(white)), and we could see it will select a lambda with best fitting, which has not-too-big explain power for coefficients and has the intermediate value for model complexity.

15. set seed 1234
dsregress colgpa athlete, controls(hsize hsrank sat satsq satcub female verbmath verbmathsq verbmathcub i.tothrs) select(cv)
//------------(I tried 2 methods, and if I do not set seed,

the result will be different)
ds colgpa athlete, not
set seed 1234
local x hsize hsrank sat satsq satcub female verbmath verbmathsq verbmathcub i.tothrs
local controls `x'
dsregress colgpa athlete, controls(`x') select(cv)

```
. //15 double selection
. set seed 1234

. dsregress colgpa athlete, controls(hsize hsrank sat satsq satcub female verbmath verbmat
> hsq verbmathcub i.tothrs) select(cv)

Estimating lasso for colgpa using cv
Estimating lasso for athlete using cv

Double-selection linear model         Number of obs               =      4,137
                                      Number of controls          =        134
                                      Number of selected controls =        107
                                      Wald chi2(1)                =      12.19
                                      Prob > chi2                 =     0.0005

------------------------------------------------------------------------------
             |               Robust
      colgpa | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     athlete |   .1326493   .0379883     3.49   0.000     .0581936     .207105
------------------------------------------------------------------------------
Note: Chi-squared test is a Wald test of the coefficients of the variables
      of interest jointly equal to zero. Lassos select controls for model
      estimation. Type lassoinfo to see number of selected variables in each
      lasso.
```

```
. ds colgpa athlete, not
sat         verbmath    hsrank      female      black       satsq       verbmathsq
tothrs      hsize       hsperc      white       hsizesq     satcub      verbmathcub

. set seed 1234

. local x hsize hsrank sat satsq satcub female verbmath verbmathsq verbmathcub i.tothrs

. local controls `x'

. dsregress colgpa athlete, controls(`x') select(cv)

Estimating lasso for colgpa using cv
Estimating lasso for athlete using cv

Double-selection linear model         Number of obs               =      4,137
                                      Number of controls          =        134
                                      Number of selected controls =        107
                                      Wald chi2(1)                =      12.19
                                      Prob > chi2                 =     0.0005

------------------------------------------------------------------------------
             |               Robust
      colgpa | Coefficient  std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
     athlete |   .1326493   .0379883     3.49   0.000     .0581936     .207105
------------------------------------------------------------------------------
Note: Chi-squared test is a Wald test of the coefficients of the variables
      of interest jointly equal to zero. Lassos select controls for model
      estimation. Type lassoinfo to see number of selected variables in each
      lasso.
```

16.

```
Number of obs               =      4,137
Number of controls          =        134
Number of selected controls =        107
Wald chi2(1)                =      12.19
Prob > chi2                 =     0.0005
```

There are 107 selected control variables retained.
The reason that more controls are retained is because there are Double selection Lasso, it run a lasso of z of x

which the original lasso does not have, and then it run the original lasso which is y on x, and then there is a last regress for y on z and the union of the selected covariate from 2 previous lasso. Therefore, the union set might result in a larger number of selected variables.

17. H0: the effect of athlete on college gpa is 0.
    Ha: the effect of athlete on college gpa is not 0
    The P value that we had from previous question is 0, and 0<0.01.
    Therefore, we should reject the Null hypotheses at the 1% significance level.