COMS 4771-2 Fall 2018 Homework 2

Zhuangyu Ren(zr2209)

October 10, 2018

Problem 1

@inproceedingsseabold2010statsmodels, title=Statsmodels: Econometric and statistical modeling with python, author=Seabold, Skipper and Perktold, Josef, booktitle=9th Python in Science Conference, year=2010,

(a) Test risks of the ordinary least squares estimator is 0.53978372.
Test risks of the sparse linear predictor is 0.5565456.

(b) Names of the variables with non-zero coefficients in the sparse linear predictor, along with the coefficient values are:
$\widehat{\beta}_0$: 5.82610034
volatile acidity : -0.24474264
sulphates : 0.10847122
alcohol : 0.37362602

(c) volatile acidity :
total sulfur dioxide : -0.4220072736318842
citric acid : -0.3755630209692122

sulphates :
chlorides : 0.4071152150059981
fixed acidity : 0.29686322762159284

alcohol :
density : -0.6795934959123543
residual sugar : -0.34626243022889536

Problem 2

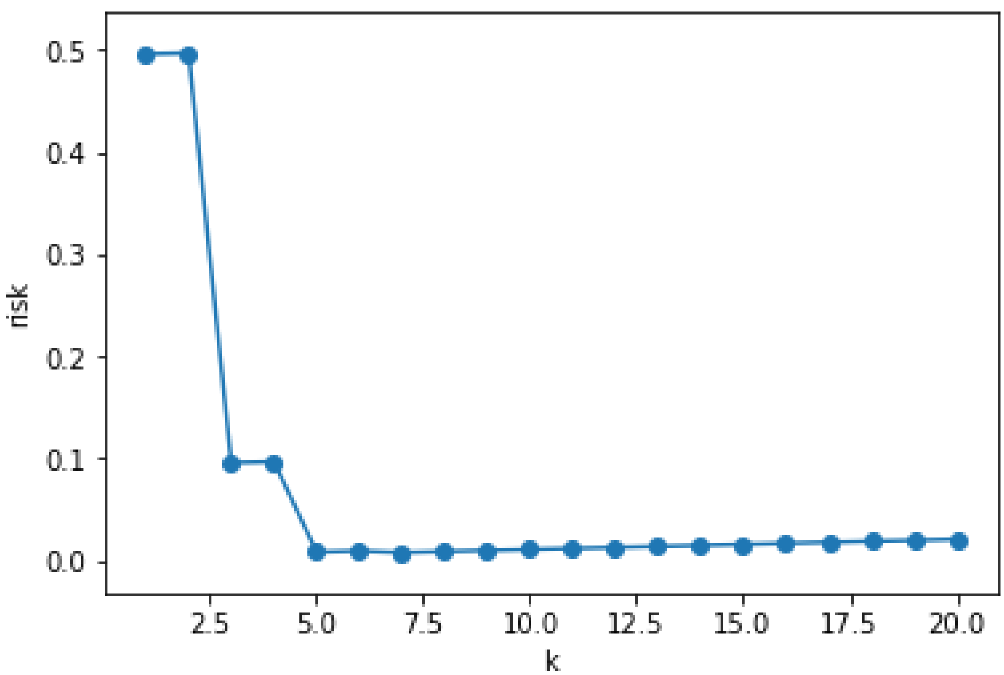
@inproceedingsseabold2010statsmodels, title=Statsmodels: Econometric and statistical modeling with python, author=Seabold, Skipper and Perktold, Josef, booktitle=9th Python in Science Conference, year=2010,

(a) As the conclusion in handout, we know that $E[R(\widehat{\beta})] \leq R(\beta^*) + \frac{\bar{\sigma}^2 r}{n}$
so, $E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i{}^T\widehat{\beta} - E(Y_i))^2\right] \leq R\left[\frac{1}{n}\sum_{i=1}^{n}(x_i{}^T\beta^* - E(Y_i))^2\right] + \frac{\bar{\sigma}^2 r}{n}$
$A^T A\widehat{\beta}^* = A^T b$
E(Y) using the Taylor's (remainder) theorem, so only the (k+1)-th will remain.
var(Y)=1
rank of A is (k+1)
so, $E\left[\frac{1}{n}\sum_{i=1}^{n}(x_i{}^T\widehat{\beta} - E(Y_i))^2\right] \leq \left(\frac{\pi^{k+1}}{(k+1)!}\right)^2 + \frac{k+1}{n}$

(b)

(c) The average fixed design risks for each k are shown in the table and plot as below.

k	risk	k	risk	k	risk	k	risk
1	0.496164	6	0.009518	11	0.012068	16	0.017014
2	0.497105	7	0.008000	12	0.012997	17	0.018002
3	0.096005	8	0.009021	13	0.014052	18	0.019102
4	0.097012	9	0.010011	14	0.015061	19	0.020058
5	0.008510	10	0.011038	15	0.016037	20	0.021035



Problem 3

@inproceedingsseabold2010statsmodels, title=Statsmodels: Econometric and statistical modeling with python, author=Seabold, Skipper and Perktold, Josef, booktitle=9th Python in Science Conference, year=2010,

(a) Features remain after the screening is 73

(b) The empirical risk of $\widehat{\beta}(\widehat{J})$ is 0.5184631137998037

(c) The empirical risk of $\widehat{\beta}(I)$ is 0.838553272038375

(d) I randomly generated 10 pairs of data and labels, computed their $\widehat{\beta}(\widehat{J})$ and $\widehat{\beta}(I)$. Below are the outcomes:
n is 236 ; d is 753 ; random risk is 0.410838458083448 ; Irisk is 0.7323342590472934
n is 217 ; d is 447 ; random risk is 0.5105407622750471 ; Irisk is 0.8575392192518303
n is 269 ; d is 772 ; random risk is 0.4170041318711085 ; Irisk is 0.7552813920061883
n is 235 ; d is 560 ; random risk is 0.5092864102832999 ; Irisk is 0.7832188161174825
n is 217 ; d is 812 ; random risk is 0.27949908586100098 ; Irisk is 0.6756825703025061
n is 765 ; d is 903 ; random risk is 0.7074898586441014 ; Irisk is 0.9530342963972447
n is 224 ; d is 637 ; random risk is 0.451797462546454553 ; Irisk is 0.7863053297575778
n is 339 ; d is 342 ; random risk is 0.750291984275502 ; Irisk is 0.9355944079878
n is 485 ; d is 469 ; random risk is 0.6744905836310303 ; Irisk is 0.9382904951357046
n is 332 ; d is 311 ; random risk is 0.7060903365984399 ; Irisk is 0.9407575809911612
I find that when n is much larger than k, the difference between two risks will decrease, while when n is equal or smaller than k, the difference will grow larger.
This may be caused by the screening procedure. We add the relationship to x and y by counting their correlations. So the risk we compute by that way is smaller than other methods. But with the number of n grows, this relationship can be reduced, so the difference between two risks gets smaller.

(e) I think the $\widehat{\beta}(\widehat{J})$ will become smaller but the empirical risk will be larger.
Because when we choose the same $\widehat{\beta}$ on a different data-label set, the $\widehat{\beta}$ cannot fully reflect the new relationships between data and labels. So the ordinary least squares estimate $\widehat{\beta}(\widehat{J})$ will be smaller as the correlation decreases. At the same time, square loss risk will become larger because of the inaccurate parameters.