Homework 3
Statistics W4240: Data Mining (Section 01)
Columbia University Due Friday, March 4

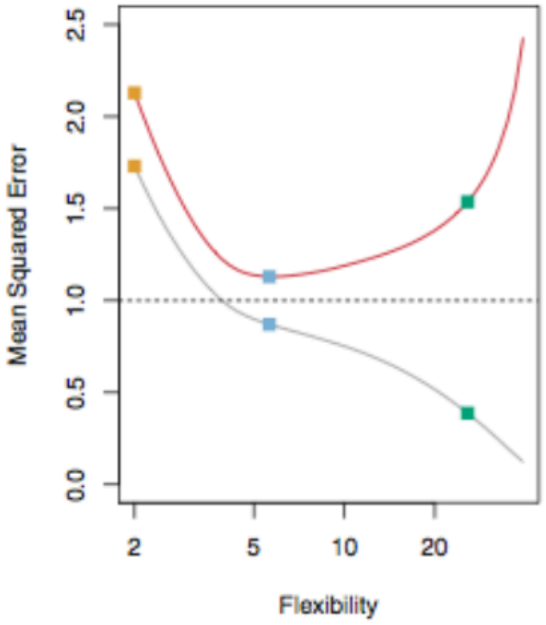
Zhang Yunyan UNI: yz2861

Q1
James 3.7.3
The fitted model is $y=50+20X_1+0.07X_2+35X_3+0.01X_4-10X_5$
For female, $y=50+20X_1+0.07X_2+35+0.01X_4-10X_1$
$=85+10X_1+0.07X_2+0.01X_4$
For male, $y=50+20X_1+0.07X_2+0.01X_4$
(a)
For a fixed value of IQ and GPA,
$y_{male}-y_{female}=-35+10X_1$
if X_1 (GPA)>3.5, $y_{male}-y_{female}>0$;
if X_1 (GPA)<3.5, $y_{male}-y_{female}<0$.
So, iii. statement is correct: for a fixed value of IQ and GPA, males earn more on average than females provided that the GPA is high enough.

(b)
$\hat{y}=85+10*4.0+0.07*110+0.01*4.0*110=137.1$
Her estimated salary is 137.1.

(c)
False.
Whether the GPA/IQ interaction term β is zero relates to the coefficient's standard error σ. According to the hypothesis testing, even if β is small, when σ is small enough, t=(β-0)/σ could be significant and we could reject H_0 that β=0. And the coefficient term for X_2 is 0.07, which is quite small as well. We could guess that the small coefficient may relate to the large number of IQ, as people's IQ always fall in the interval [80, 200]. Compared to that, GPA always falls in the interval [1, 4], which is relatively small, resulting in the large coefficient (20 and 10).

Q2
James 3.7.4
(a)
There is no enough information to tell. The relationship between RSS_{linear} and RSS_{cubic} depend on training data. Although cubic regression is more flexible than linear regression ($Y = \beta_0 + \beta_1 X + 0X^2 + 0X^3 + \epsilon$ actually), which suggests a lower RSS, the true relationship is linear. Therefore, it is likely that the RSS of linear regression is lower than cubic regression.
(b)
I would expect RSS_{linear} lower than RSS_{cubic}, because linear is the true relationship. Due to the overfitting property of cubic regression, it may show lower training RSS than linear regression, but a higher flexibility may result in a higher RSS in testing data.



picture taken from Lecture 8 notes

(c)

I would expect RSS_{cubic} lower than $RSS_{linear.}$ According to the picture above, when the true relationship is not linear, the RSS of training data is decreasing with the increase of flexibility.

(d)

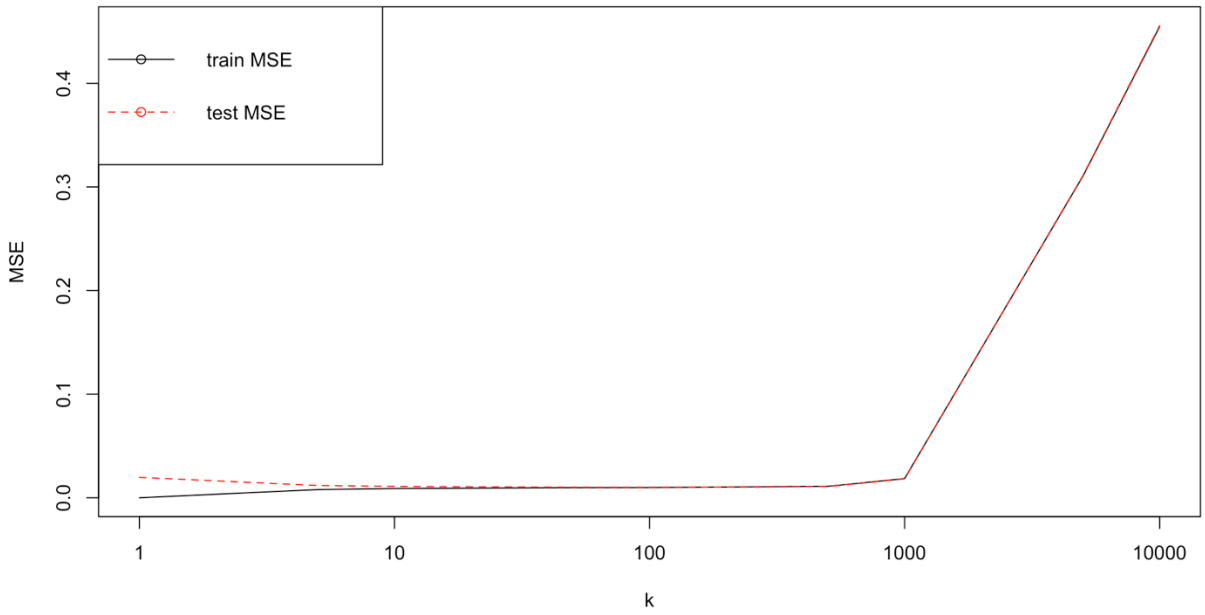
There is no enough information to tell, testing RSS depends on the true relationship. If the true polynomial degree is much closer to linear, then RSS_{linear} would be lower; if it is closer to cubic, then RSS_{cubic} would be lower.

Q3

(a)

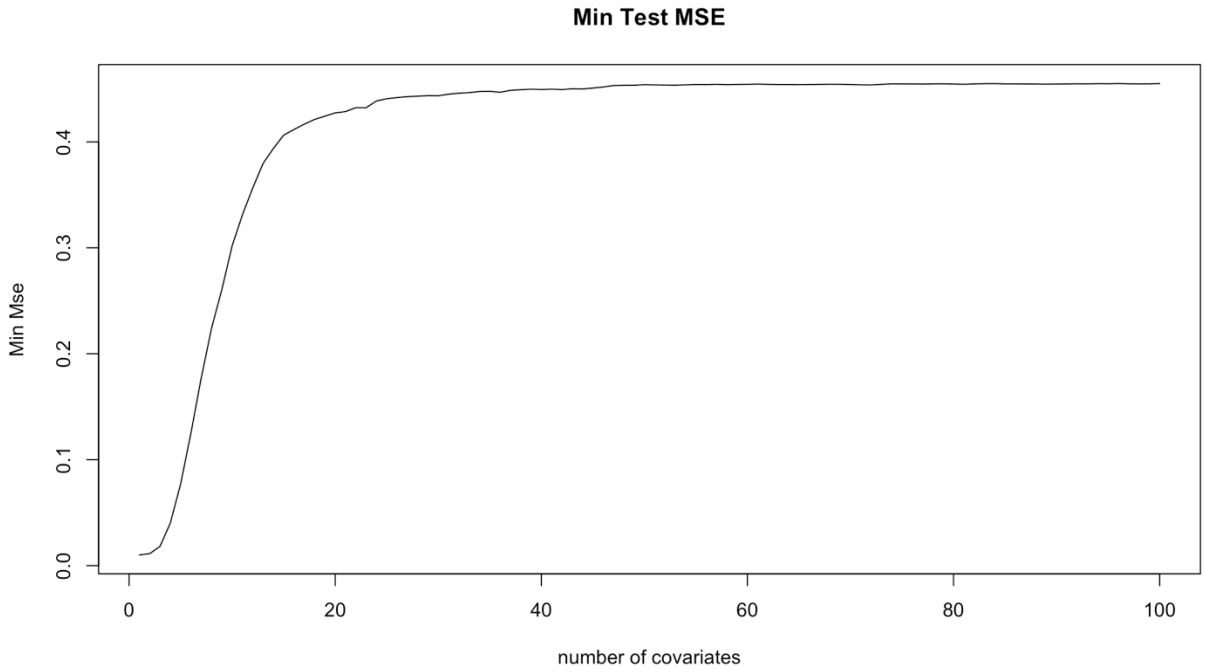
$train_mse
[1] 0.000000000 0.007928919 0.008939936 0.009739450 0.009870889
[6] 0.011015481 0.018507606 0.310306946 0.454719456

$test_mse
[1] 0.01967759 0.01186080 0.01093954 0.01016357 0.01005514
[6] 0.01082189 0.01830097 0.31077500 0.45568067



(b) The training errors are monotonically increasing, while the testing errors first decrease and then increase, indicating there is a minimum number. The best value of k should be 100 as it reaches the minimum of testing error 1.01%.

(c)

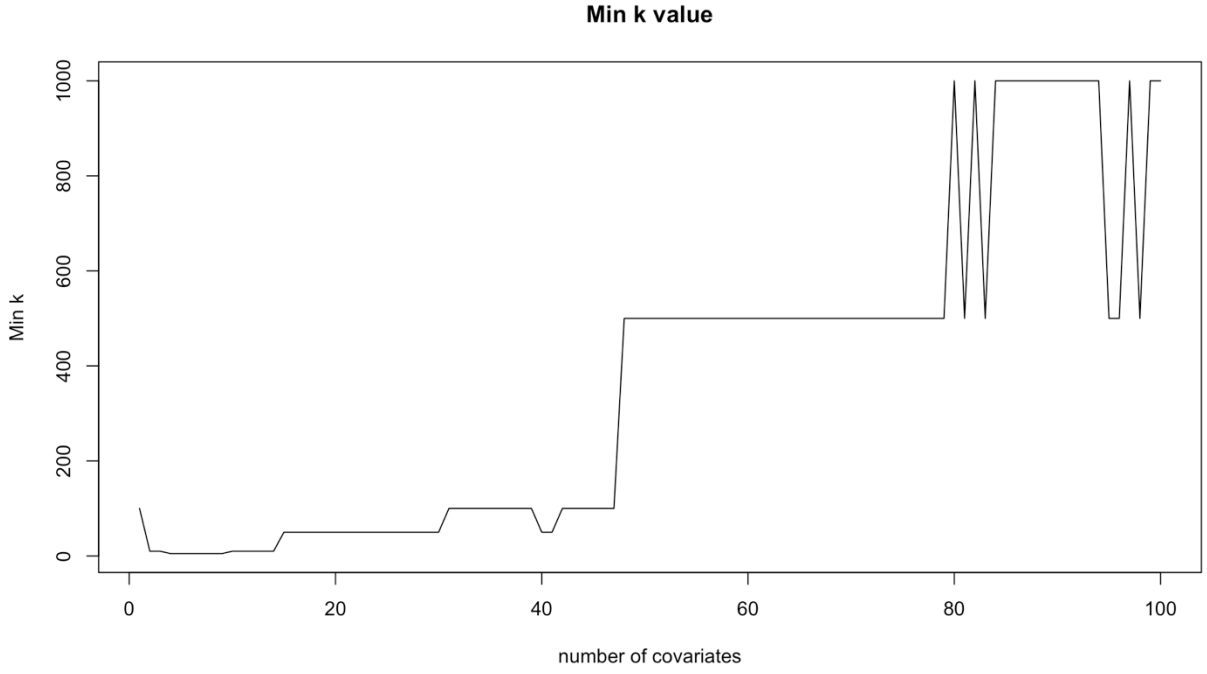


min_mse

[1] 0.01005514 0.01130003 0.01820532 0.04021780 0.07706236 0.12510489 0.17724651
[8] 0.22415452 0.26068703 0.30197980 0.33132724 0.35656586 0.37965790 0.39383175
[15] 0.40638609 0.41171385 0.41673295 0.42122625 0.42428089 0.42740587 0.42861842
[22] 0.43232119 0.43223841 0.43865668 0.44072340 0.44184324 0.44268722 0.44316714

[29] 0.44367193 0.44355710 0.44504280 0.44598745 0.44647229 0.44759583 0.44778120
[36] 0.44683965 0.44866800 0.44929589 0.44977329 0.44950068 0.44979067 0.44942871
[43] 0.45015150 0.44995287 0.45085374 0.45179207 0.45319191 0.45344298 0.45351983
[50] 0.45397929 0.45380444 0.45366497 0.45356723 0.45383605 0.45414955 0.45405130
[57] 0.45432048 0.45400037 0.45422377 0.45430691 0.45455659 0.45428848 0.45409820
[64] 0.45409548 0.45399411 0.45413679 0.45425581 0.45440011 0.45432846 0.45411804
[71] 0.45387317 0.45378268 0.45426842 0.45479570 0.45472210 0.45471336 0.45458111
[78] 0.45476211 0.45483550 0.45461915 0.45437117 0.45473467 0.45497837 0.45505031
[85] 0.45474924 0.45475722 0.45469794 0.45460968 0.45451254 0.45461025 0.45469241
[92] 0.45484953 0.45476878 0.45501141 0.45488032 0.45514000 0.45483172 0.45476949
[99] 0.45479187 0.45507988

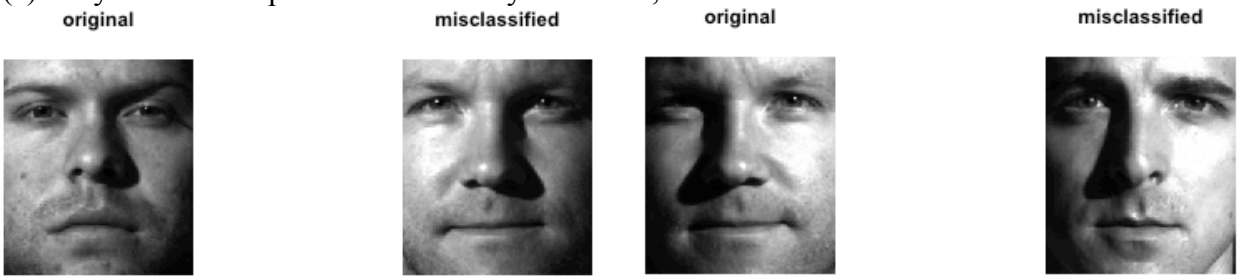
min_k
[1] 100 10 10 5 5 5 5 5 5 10 10 10 10 10 50 50 50
[18] 50 50 50 50 50 50 50 50 50 50 50 50 50 100 100 100 100
[35] 100 100 100 100 100 50 50 100 100 100 100 100 100 500 500 500 500
[52] 500 500 500 500 500 500 500 500 500 500 500 500 500 500 500 500 500
[69] 500 500 500 500 500 500 500 500 500 500 500 1000 500 1000 500 1000 1000
[86] 1000 1000 1000 1000 1000 1000 1000 1000 1000 500 500 1000 500 1000 1000

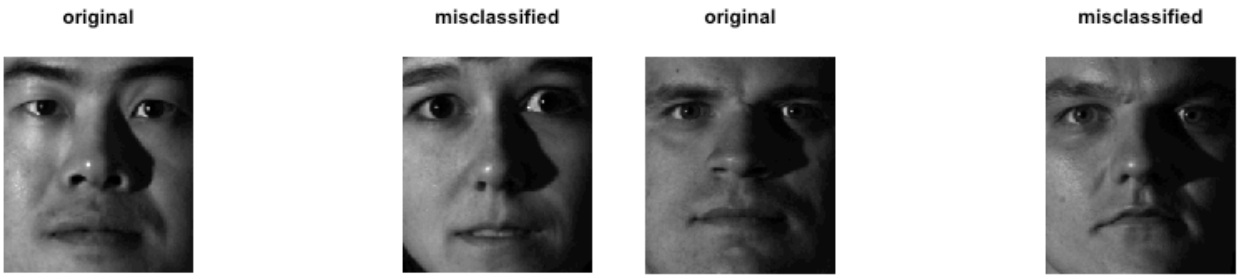


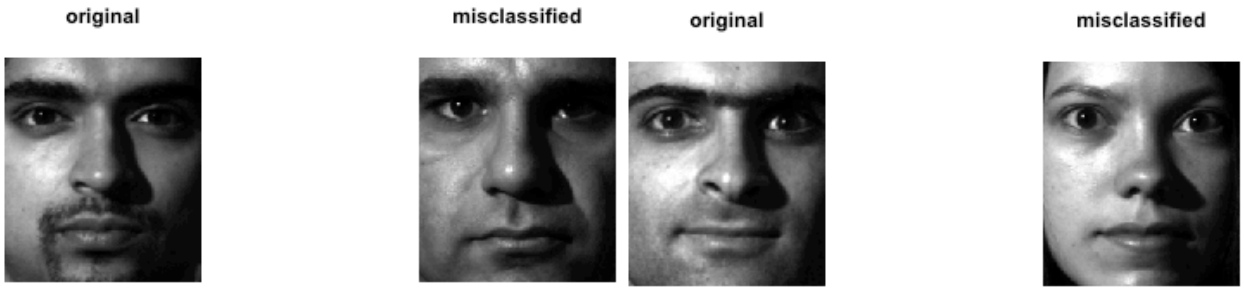
(d)
As the number of covariates increases, testing errors increase, and stay at the same value (0.44) after including 40 covariates.
The best neighbor size increases as the number of covariates increase though may fluctuate a bit.

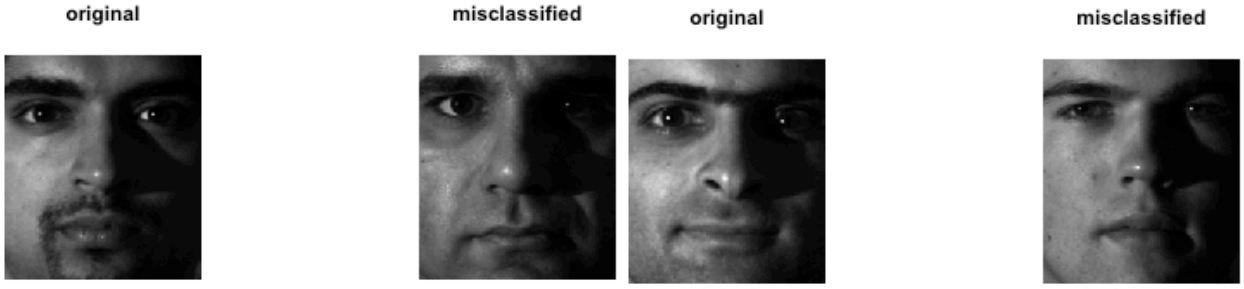
Q4
(a) The first 5 files in the training set:
41 "CroppedYale/yaleB11/yaleB11_P00A+000E+00.pgm"

57 "CroppedYale/yaleB16/yaleB16_P00A+000E+00.pgm"
86 "CroppedYale/yaleB23/yaleB23_P00A+005E+10.pgm"
136"CroppedYale/yaleB35/yaleB35_P00A+010E+00.pgm"
30 “CroppedYale/yaleB08/yaleB08_P00A+005E+10.pgm”
The first 5 files in the testing set:
5 "CroppedYale/yaleB02/yaleB02_P00A+000E+00.pgm"
12 "CroppedYale/yaleB03/yaleB03_P00A+010E+00.pgm"
18 "CroppedYale/yaleB05/yaleB05_P00A+005E+10.pgm"
20 "CroppedYale/yaleB05/yaleB05_P00A+010E+00.pgm"
21 "CroppedYale/yaleB06/yaleB06_P00A+000E+00.pgm"
(b) Mis-classification rate is 0, so no photos can be shown.
(c) Only 4 of 31 test photos are correctly classified, while 27 are mis-classified.

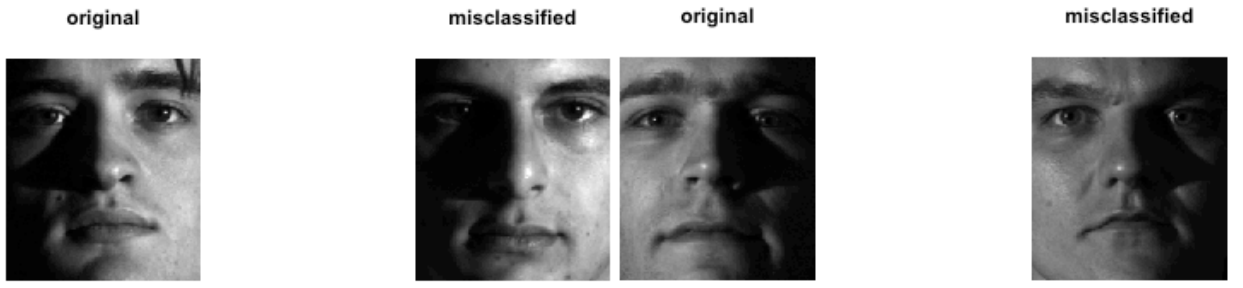





original
misclassified
original
misclassified


original
misclassified
original
misclassified


original
misclassified
original
misclassified


original
misclassified
original
misclassified


original
misclassified
original
misclassified
original
misclassified
original
misclassified
original
misclassified
original
misclassified

(d)

When setting seed from 10 to 19, mis-classification numbers of photos are 21 23 27 23 24 28 24 25 20 27 respectively.

(e)

The testing errors from b) and c) are different, 0 and 27/31 respectively. The first 25 dimensions to describe faces in (a) is enough, while doing PCA in dimension reduction in (c) is not good as the first 25 dimensions are not enough. Carefully examining the mis-classified photos, I find that some are taken in very dark places and some are taken in very light places. Looking into photos in (a), the lightness is almost the same. Maybe it is the reason for the mis-classified rates.

As photo is converted to pixels, extremely light and dark photos may have extreme data points, while correctly-classified photos may have more centered data. So it may suggest that PCA

performs well on centered data.
(f)
Using uncropped photos won't help as they may include more noise. As the environment may have features quite different from faces.