STA207 homework6

Juanjuan Hu, Zhen Zhang

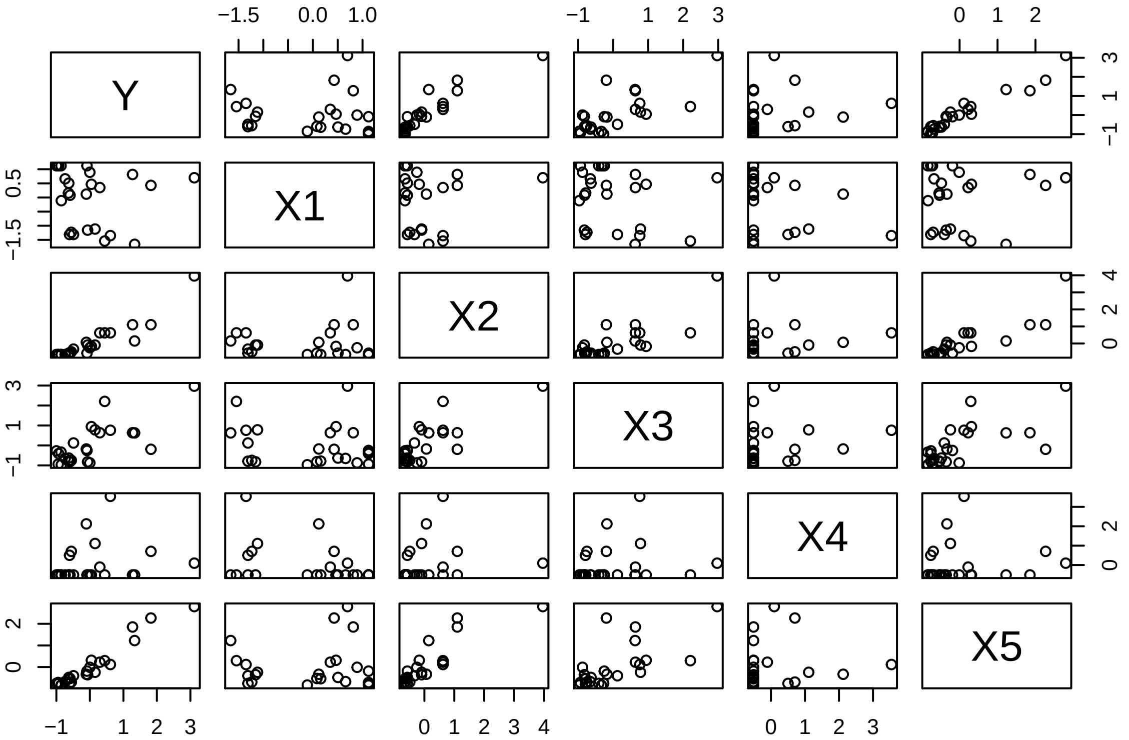
March 5, 2016

1

```
## Loading required package: rJava
## Loading required package: xlsxjars
```

(a)

```
apartments=data.frame(sapply(apartment, function(x) scale(as.numeric(as.character(x)))))
pairs(apartments)
```



```
cor(apartments) # x2 and x5 shows linear pattern
```

```
##              Y          X1          X2         X3          X4         X5
## Y    1.0000000 -0.11453460  0.92345442  0.7413715  0.22497528 0.96813120
## X1 -0.1145346  1.00000000 -0.01415504 -0.1885895 -0.36265249 0.02700832
## X2  0.9234544 -0.01415504  1.00000000  0.8000696  0.22412565 0.87786360
## X3  0.7413715 -0.18858951  0.80006959  1.0000000  0.16609137 0.67269398
## X4  0.2249753 -0.36265249  0.22412565  0.1660914  1.00000000 0.08929658
## X5  0.9681312  0.02700832  0.87786360  0.6726940  0.08929658 1.00000000
```

From the matrix plot of data and the correlation matrix, we can find that some independent variables seem to be correlated, like X2 and X5. It seems that multicollinearity is present.

(b)

```r
X = as.matrix(apartments[,2:6])
eigen(t(X)%*%X)$values
```

```
## [1] 63.219483 31.924456 15.769011  7.098047  1.989003
```

We can see that the maximum eigenvalue is 63.22 and the smallest eigenvalue is 1.99. The condition index (maximum divided by the minumum) is 31.78, which is relatively large. It indicates the existance of multicollinearity.

(c)

```r
fit1=lm(apartments$Y~0+X)
summ.fit1=summary(fit1)
summ.fit1
```

```
##
## Call:
## lm(formula = apartments$Y ~ 0 + X)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.25674 -0.08137 -0.01993  0.07042  0.34636
##
## Coefficients:
##     Estimate Std. Error t value Pr(>|t|)
## XX1 -0.10461    0.03556  -2.942  0.00807 **
## XX2  0.24656    0.08636   2.855  0.00979 **
## XX3  0.01854    0.05545   0.334  0.74159
## XX4  0.06294    0.03581   1.758  0.09410 .
## XX5  0.73642    0.06744  10.920 7.06e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1529 on 20 degrees of freedom
## Multiple R-squared:  0.9805, Adjusted R-squared:  0.9757
## F-statistic: 201.4 on 5 and 20 DF,  p-value: < 2.2e-16
```

```r
anova(fit1)
```

```
## Analysis of Variance Table
##
## Response: apartments$Y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## X          5 23.5325  4.7065  201.37 < 2.2e-16 ***
## Residuals 20  0.4675  0.0234
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the regression result, we can see that the the R-square is 0.98, meaning that the five variables could explain the response variable well. The p-value of the F test is 2.2e-16, indicating that the not all the coefficients are equal to zero.

(d)

```
sapply(1:5, function(i) 1/(1-summary(lm(X[,i]~0+X[,-i]))$r.squared) )
```

```
## [1] 1.298654 7.657888 3.157590 1.316618 4.670186
```

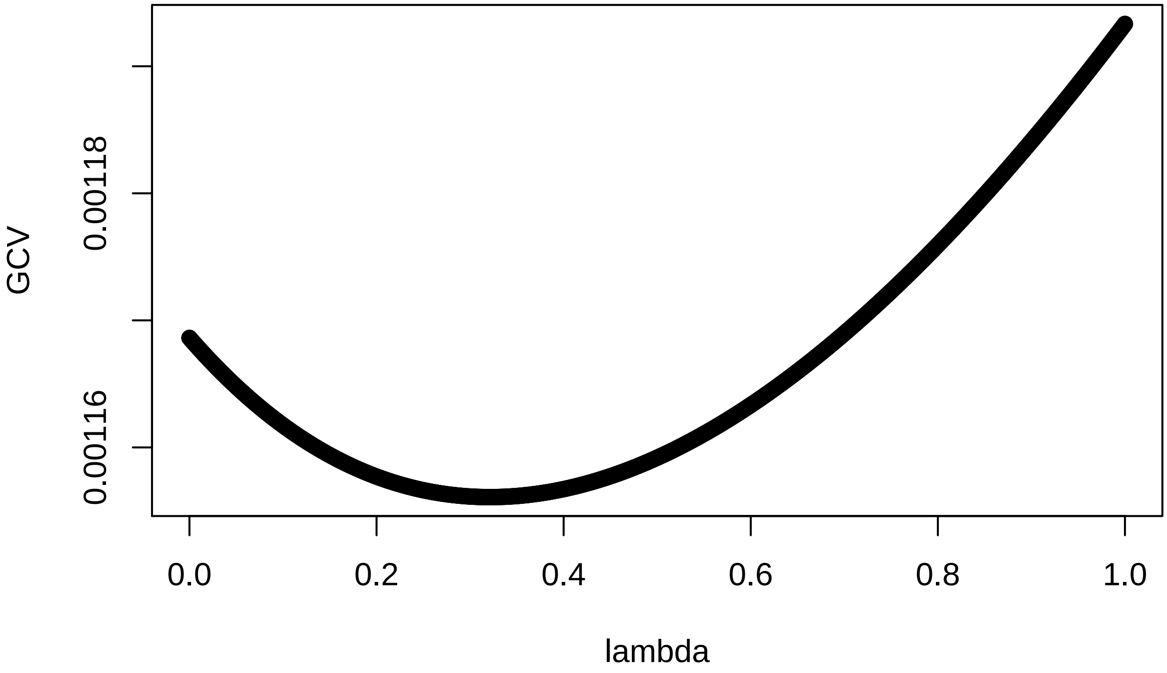
The largest VIF is the VIF for variable X2, which is 7.66. It means that X2 can be partially linear represented by other variables, indicating presence of multicollinearity.

(e)

```
fit.ridge = lm.ridge(apartments$Y~0+X, lambda=seq(0,1,by=0.001))
lambda=lambda=seq(0,1,by=0.001)
GCV = fit.ridge$GCV
lambda.opt=lambda[which.min(GCV)] # lambda=0.321
lambda.opt
```

```
## [1] 0.321
```

```
plot(x=lambda, y=GCV)
```



The

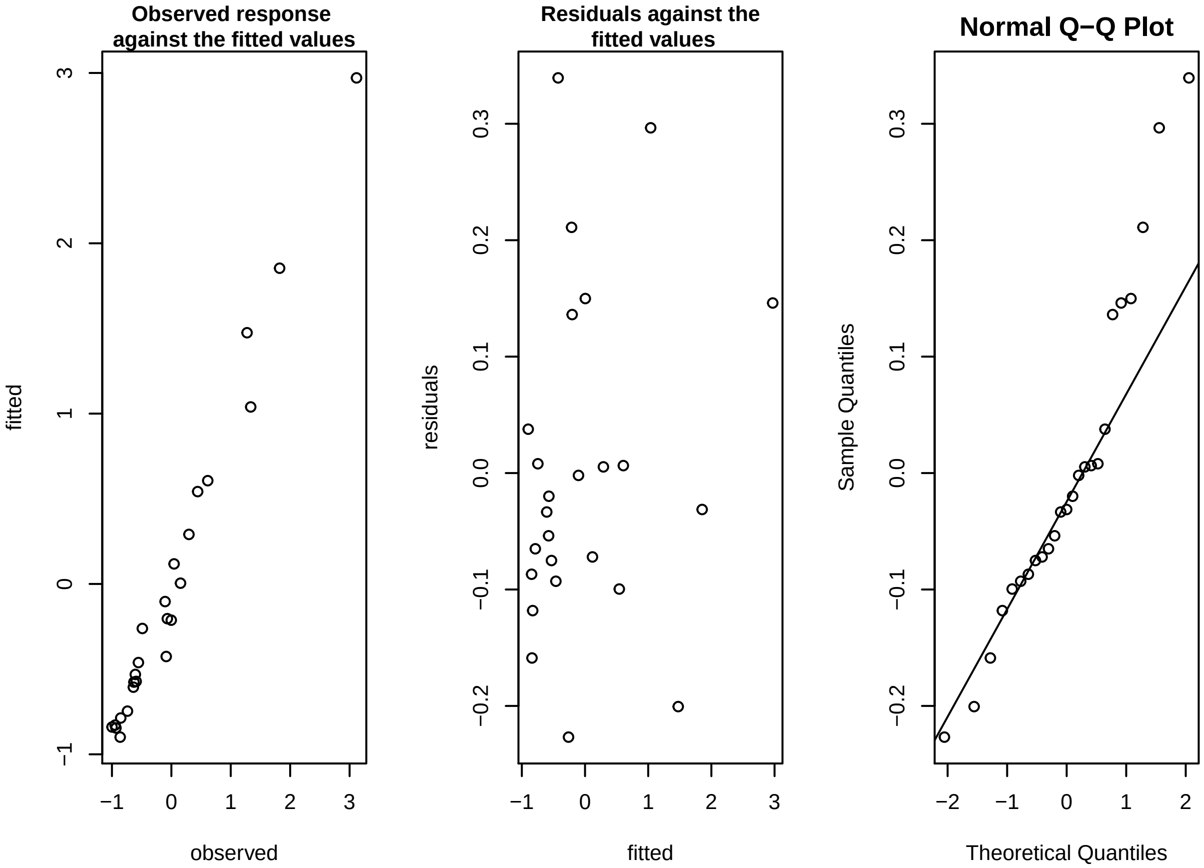
optimal penalty parameter is 0.321.

```
fit2 = lm.ridge(apartments$Y~0+X,lambda=lambda.opt)
beta.hat=fit2$coef
y.hat=X%*%beta.hat
summ.fit2=summary(linearRidge(apartments$Y~0+X,lambda=lambda.opt,scaling="scale"))
summ.fit2
```

```
##
## Call:
## linearRidge(formula = apartments$Y ~ 0 + X, lambda = lambda.opt,
##     scaling = "scale")
##
##
## Coefficients:
##     Estimate Scaled estimate Std. Error (scaled) t value (scaled) Pr(>|t|)
## XX1 -0.10191        -0.10191             0.03475            2.933 0.003358
## XX2  0.26388         0.26388             0.07500            3.518 0.000434
## XX3  0.02439         0.02439             0.05164            0.472 0.636752
## XX4  0.06080         0.06080             0.03482            1.746 0.080791
## XX5  0.70793         0.70793             0.06051           11.700  < 2e-16
##
## XX1 **
## XX2 ***
## XX3
## XX4 .
## XX5 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Ridge parameter: 0.321
##
## Degrees of freedom: model 4.783 , variance 4.587 , residual 4.978
```

(f)

```
par(mfrow=c(1,3),mar=c(4,4,2,2))
plot(x=apartments$Y, y=y.hat, main = "Observed response \nagainst the fitted values", cex.main=1, xlab=
res.fit2 = apartments$Y-y.hat
plot(x=y.hat,y = res.fit2, main = "Residuals against the\n fitted values", xlab="fitted", ylab="residual
qqnorm(res.fit2)
qqline(res.fit2)
```



From the residuals plot, we can see that there is no obvious pattern within the residuals. And the residuals do not show serious departure from normality.

(g)

```
R=t(X)%*%X
I=diag(nrow=5)
M=solve(R+lambda.opt*I)
A=M%*%R%*%M
diag(solve(A))*diag(A)
```

```
##       X1       X2       X3       X4       X5
## 1.260866 5.881164 2.785603 1.266044 3.825674
```

Compared to the VIF obtained in part(d), the VIF for the estimated ridge regression is smaller. It indicates that the problem of multicollinearity is remedied a bit.