Math405 – HW22

Nick Huo

2022-11-18

Question 1

```
dat <- read.csv("HW22.csv")
attach(dat)
```

a.

The principle of hierarchy generally means that we should keep the fundamental terms when we add terms that are modifying the original terms (like an interaction term). In the case of polynomial regressions, when we add higher power terms, we should keep the lower order terms. For example, if we need to fit X_1^3, we should still keep X_1^2 and X_1 in the model.

b.

```
mod1 <- lm(Y ~ ., data=dat)
summary(mod1)
```

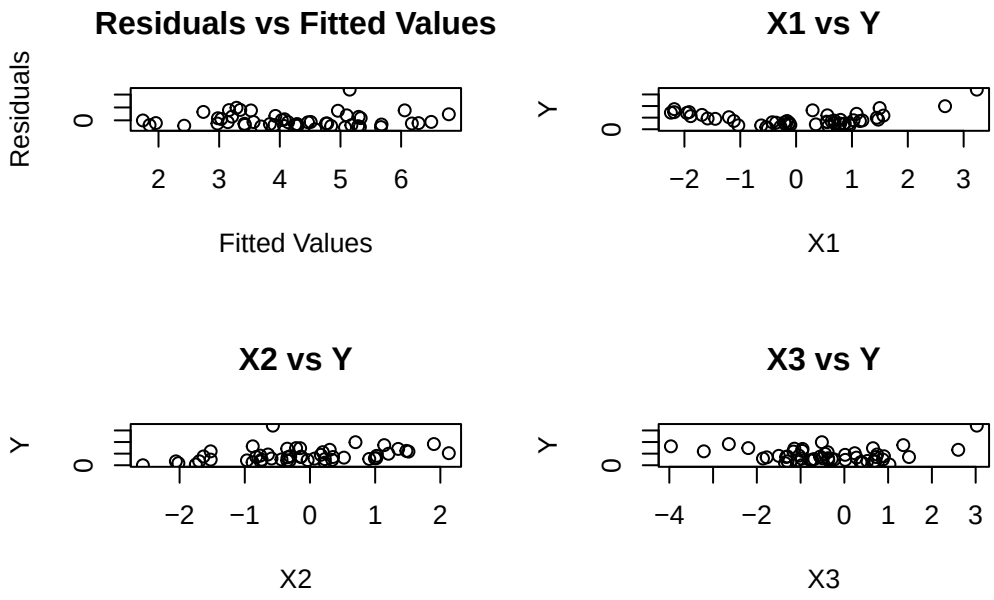
```
##
## Call:
## lm(formula = Y ~ ., data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.4646 -1.8226 -0.9138  1.2273 11.9186
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.40772    0.42308  10.418 1.09e-13 ***
## X1           0.35190    0.32440   1.085  0.28367
## X2           1.07343    0.37972   2.827  0.00693 **
## X3           0.07127    0.31246   0.228  0.82058
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.833 on 46 degrees of freedom
## Multiple R-squared:  0.1631, Adjusted R-squared:  0.1085
## F-statistic: 2.988 on 3 and 46 DF,  p-value: 0.04061
```

The Global F-test is significant with F-statistic of 2.988 on 3 and 46 DF with a p-value = 0.041. This indicates that at least one of the predictors is significant in explaining the variability in Y.

The R^2_{adj}, however, is really small – only 0.11 – meaning that only 11% of the variability in Y is being explained by the predictors.

Finally, we see that only X_2 is significant, with p-values = 0.007.

```
par(mfrow=c(2,2))
plot(fitted(mod1), resid(mod1), main="Residuals vs Fitted Values",
        xlab="Fitted Values", ylab="Residuals")
plot(X1, Y, main="X1 vs Y", xlab="X1", ylab="Y")
plot(X2, Y, main="X2 vs Y", xlab="X2", ylab="Y")
plot(X3, Y, main="X3 vs Y", xlab="X3", ylab="Y")
```



When looking at the Residuals vs. Fitted plot, we are seeing a mostly no pattern, and one possible outlier, but it shouldn't have high influence on the model.

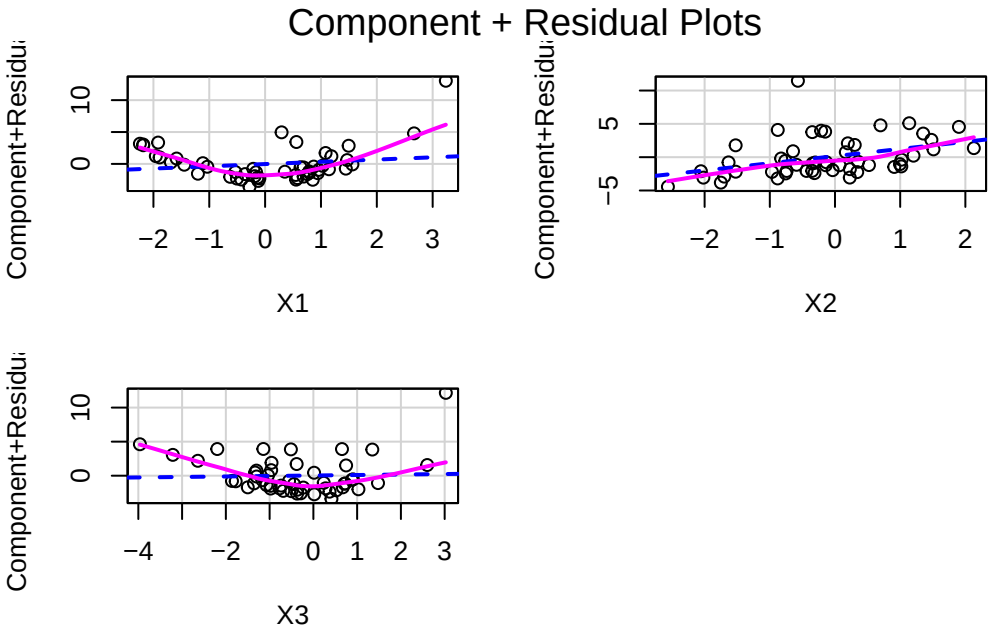
Looking at the relationships between $X1, X2, X3$ vs Y though, we see seeing clear nonlinear relationships between $X1, Y$ and $X3, Y$. This is a problem as it violates the linearity condition. The relationship between $X2, Y$ seems to be linear.

c.

```
library(car)
```

```
## Loading required package: carData
```

```
crPlots(mod1)
```



The plots are showing the relationship between the predictors and their partial residuals, that is, the relationship between specific predictors and the response while accounting for the effects of other predictors. And we would expect to see a linear relationship between predictors and the response.

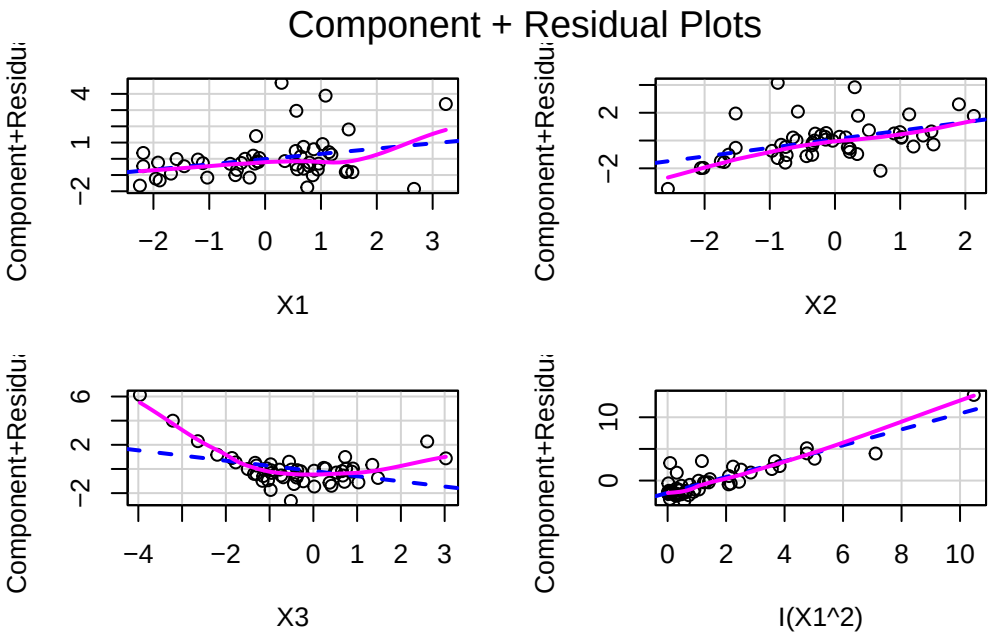
For $X1, X3$, we are seeing clear curved relationships, as indicated by the pink trend line. So we might want to explore that. For $X2$, the relationship seems to be linear and should be ok.

d.

```
mod2 <- lm(Y~.+I(X1^2), data=dat)
summary(mod2)
```

```
##
## Call:
## lm(formula = Y ~ . + I(X1^2), data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6977 -0.7779 -0.1004  0.3258  4.6044
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.19510    0.26354   8.329 1.15e-10 ***
## X1           0.32444    0.15241   2.129  0.03878 *
## X2           0.61774    0.18191   3.396  0.00144 **
## X3          -0.42761    0.15188  -2.815  0.00720 **
## I(X1^2)      1.25355    0.09805  12.784  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.331 on 45 degrees of freedom
## Multiple R-squared:  0.8193, Adjusted R-squared:  0.8033
## F-statistic: 51.02 on 4 and 45 DF,  p-value: 3.705e-16
```

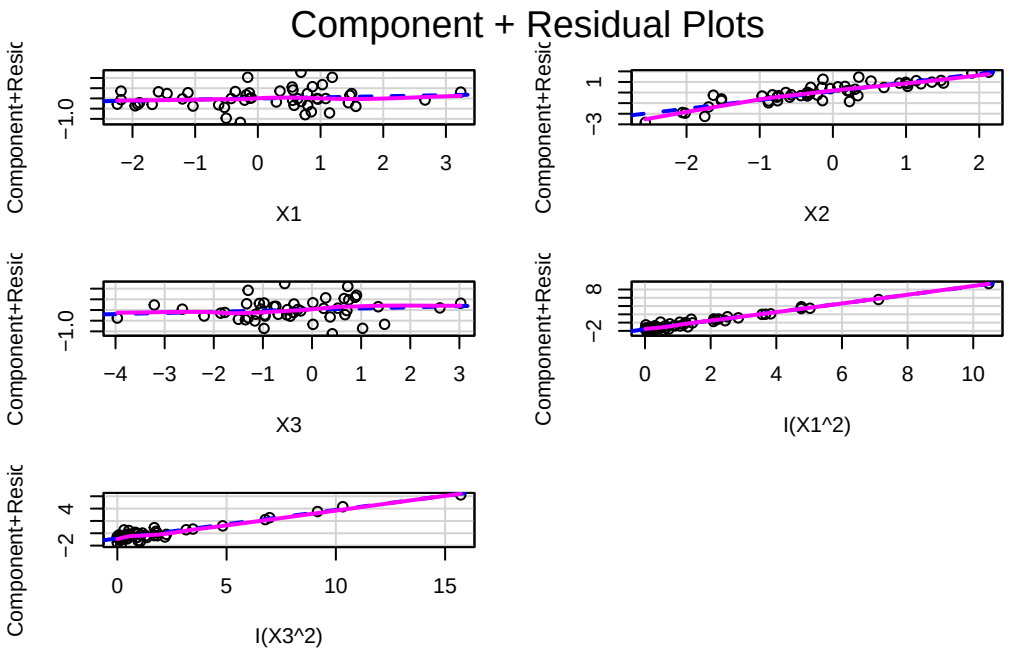
```
crPlots(mod2)
```




```
mod3 <- lm(Y~.+I(X1^2)+I(X3^2), data=dat)
summary(mod3)
```

```
##
## Call:
## lm(formula = Y ~ . + I(X1^2) + I(X3^2), data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.15479 -0.24425 -0.04772  0.30806  1.24665
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.91939    0.10030  19.137  < 2e-16 ***
## X1           0.05432    0.05947   0.913    0.366
## X2           0.82286    0.06938  11.860 2.69e-15 ***
## X3           0.05304    0.06394   0.830    0.411
## I(X1^2)      1.04263    0.03893  26.780  < 2e-16 ***
## I(X3^2)      0.46142    0.02780  16.597  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4994 on 44 degrees of freedom
## Multiple R-squared:  0.9751, Adjusted R-squared:  0.9723
## F-statistic: 344.8 on 5 and 44 DF,  p-value: < 2.2e-16
```

```
crPlots(mod3)
```



e.

In mod1, the original model is the model is $\hat{Y} = X1 + X2 + X3$, and we saw horrible performance with it: Residual SE was 2.83, the R^2_{adj} was only 0.10, and only one predictor is significant. According to the partial residual plots, $X1$ and $X3$ might have linearity problems.

In mod2, we first add a second order term for $X1$, so the model is $\hat{Y} = X1+X2+X3+(X1)^2$. Now, Residual SE was 1.33, the R^2_{adj} is 0.80 and all the predictors are significant. The partial residual plots indicate that the linearity problem in $X1$ has been largely accounted for. Now, $X3$ has the most problem.

In mod3, we add a second order term for $X3$, so now the model is $\hat{Y} = X1 + X2 + X3 + (X1)^2 + (X3)^2$. Now, Residual SE was 0.50, the R^2_{adj} is 0.97 and the all, except the low order predictors, are significant. The partial residuals plots show that all the predictors now have a linear relationship with the response, while accounting for the effects of other predictors. Thus, linearity condition is met. Also, even if the lower order predictors are not significant, by the Principle of hierarchy, we should still keep them in the model.

As we fit the problems in linearity in the predictors, we see the R^2_{adj} increasing and Residual SE decreasing drastically.

2.

```
detach(dat)
wine <- read.csv("winequality.csv")
attach(wine)
```

a.

```
mod_wine <- lm(quality~., data=wine)
summary(mod_wine)
```

```
##
## Call:
## lm(formula = quality ~ ., data = wine)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.74543 -0.26718  0.00016  0.39485  1.74384
##
## Coefficients:
##                        Estimate Std. Error t value Pr(>|t|)
## (Intercept)           33.775154  86.275642   0.391   0.6964
## fixed.acidity         -0.066238   0.105971  -0.625   0.5335
## volatile.acidity      -0.043529   0.465435  -0.094   0.9257
## citric.acid            0.104247   0.524185   0.199   0.8428
## residual.sugar         0.025261   0.058712   0.430   0.6681
## chlorides             -0.140844   1.839394  -0.077   0.9391
## free.sulfur.dioxide    0.006640   0.008254   0.804   0.4233
## total.sulfur.dioxide  -0.003833   0.002724  -1.407   0.1630
## density              -27.475495  88.155740  -0.312   0.7560
## pH                    -1.639071   0.902130  -1.817   0.0726 .
## sulphates              0.432786   0.538811   0.803   0.4240
## alcohol                0.472418   0.115313   4.097 9.29e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6177 on 88 degrees of freedom
## Multiple R-squared:  0.4301, Adjusted R-squared:  0.3588
## F-statistic: 6.037 on 11 and 88 DF,  p-value: 2.89e-07
```

The Global F-test is significant with F-statistic of 6.04 on 11 and 88 DF with a p-value < 0.0001. This indicates that at least one of the predictors is significant in explaining the variability in wine quality.

The R^2_{adj} is quite small – only 0.36 – meaning that only 36% of the variability in wine quality is being explained by the predictors.

Finally, we notice that only one predictor (alcohol) is significant. There might be multicollinearity in the predictors. We should look at the correlation matrix to check for this. We can also use Principle Components Regression to help.

b.

```
X <- scale(as.matrix(wine[,-12]))
round(cor(X),3)
```

```
##                       fixed.acidity volatile.acidity citric.acid residual.sugar
## fixed.acidity                 1.000           -0.342       0.588         -0.080
## volatile.acidity             -0.342            1.000      -0.605         -0.098
## citric.acid                   0.588           -0.605       1.000          0.257
## residual.sugar               -0.080           -0.098       0.257          1.000
## chlorides                     0.037           -0.017       0.182         -0.002
## free.sulfur.dioxide          -0.057            0.065       0.045          0.126
## total.sulfur.dioxide         -0.029            0.108       0.139          0.179
```

```
## density                     0.595           -0.089       0.385          0.233
## pH                         -0.660            0.188      -0.378          0.064
## sulphates                   0.078           -0.301       0.258         -0.133
## alcohol                     0.013           -0.191       0.111          0.028
##                      chlorides free.sulfur.dioxide total.sulfur.dioxide density
## fixed.acidity            0.037              -0.057               -0.029   0.595
## volatile.acidity        -0.017               0.065                0.108  -0.089
## citric.acid              0.182               0.045                0.139   0.385
## residual.sugar          -0.002               0.126                0.179   0.233
## chlorides                1.000              -0.097               -0.082   0.105
## free.sulfur.dioxide     -0.097               1.000                0.594  -0.025
## total.sulfur.dioxide    -0.082               0.594                1.000   0.076
## density                  0.105              -0.025                0.076   1.000
## pH                      -0.216              -0.137               -0.215  -0.338
## sulphates                0.226               0.031                0.014  -0.100
## alcohol                 -0.123              -0.067               -0.183  -0.474
##                           pH sulphates alcohol
## fixed.acidity         -0.660     0.078   0.013
## volatile.acidity       0.188    -0.301  -0.191
## citric.acid           -0.378     0.258   0.111
## residual.sugar         0.064    -0.133   0.028
## chlorides             -0.216     0.226  -0.123
## free.sulfur.dioxide   -0.137     0.031  -0.067
## total.sulfur.dioxide  -0.215     0.014  -0.183
## density               -0.338    -0.100  -0.474
## pH                     1.000    -0.088   0.297
## sulphates             -0.088     1.000   0.323
## alcohol                0.297     0.323   1.000
```

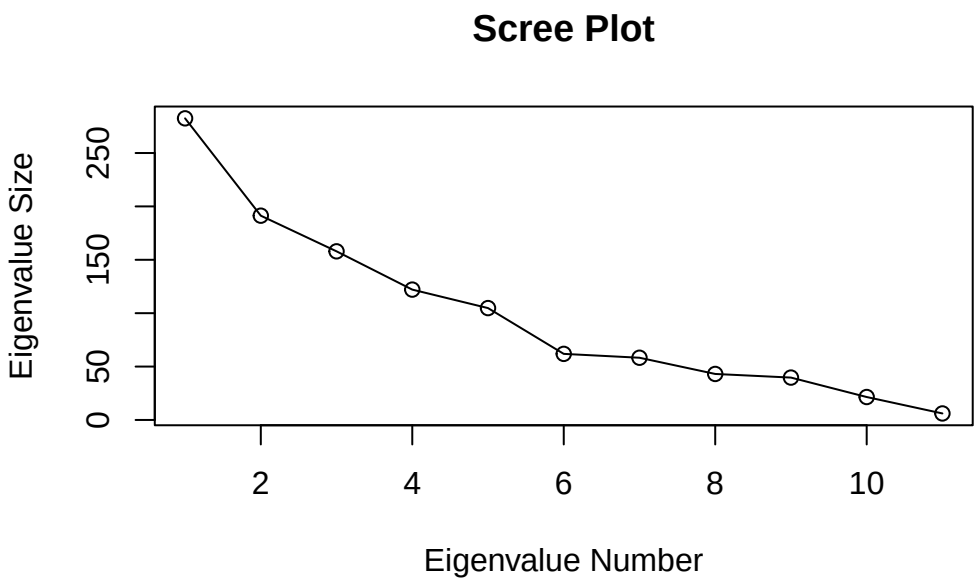
The correlation matrix indicate that some of the predictors are moderately correlated ($|r| \approx 0.6$).

```
e <- eigen(t(X)%*%X)
e
```

```
## eigen() decomposition
## $values
##  [1] 282.462251 191.304334 157.923033 122.062516 104.767989  61.878626
##  [7]  58.262598  43.076181  39.685033  21.487305   6.090135
##
## $vectors
##             [,1]        [,2]        [,3]        [,4]        [,5]        [,6]
##  [1,]  0.4941696  0.06474701  0.16640027  0.01061186  0.35258084 -0.35404666
##  [2,] -0.3217839 -0.36819062  0.17058929  0.18942121 -0.05988056 -0.62930697
##  [3,]  0.4826579  0.17290252 -0.18236160 -0.17695205 -0.08164836  0.03527191
##  [4,]  0.1002701 -0.14342594 -0.23917753 -0.63375374 -0.45754141 -0.22113361
##  [5,]  0.1487404  0.08163769  0.14334671  0.39245575 -0.74986701 -0.17336707
##  [6,]  0.0514107 -0.36329523 -0.54370364  0.16953824  0.08867628 -0.00077800
##  [7,]  0.1070639 -0.43280481 -0.48967231  0.13868080  0.01769085  0.02035660
##  [8,]  0.3969300 -0.25454961  0.28543067 -0.21824311 -0.07651978  0.00628748
##  [9,] -0.4323245  0.15040567 -0.08105673 -0.33431759 -0.15511924  0.18185787
## [10,]  0.1340290  0.38052369 -0.29941954  0.39562714 -0.17948839  0.12336070
## [11,] -0.1028963  0.50159088 -0.34513009 -0.11460777  0.15323238 -0.59133637
##              [,7]        [,8]        [,9]       [,10]        [,11]
##  [1,] -0.06228374  0.12033076 -0.01492625 -0.16511326  0.654443526
```

```
##  [2,] -0.30237107 -0.05328439  0.09860193  0.43813430  0.018580760
##  [3,]  0.08460628  0.17578312  0.37869737  0.68607640 -0.111791312
##  [4,]  0.03181937 -0.35481942 -0.29648291 -0.02254452  0.192253313
##  [5,]  0.25808561  0.28411259  0.09082152 -0.20981402  0.068458022
##  [6,]  0.05803944  0.52611233 -0.49355681  0.09670135  0.006350623
##  [7,] -0.05530374 -0.21892907  0.61638273 -0.32843400  0.051900022
##  [8,] -0.54546811  0.24907648 -0.03004912 -0.26035282 -0.463027457
##  [9,] -0.36368486  0.50226534  0.27949811 -0.04427071  0.387794667
## [10,] -0.62689739 -0.29818317 -0.20554800  0.08920785  0.107282522
## [11,]  0.02615413  0.13121701  0.07546105 -0.26575058 -0.371740070
```

```
V <- e$vectors
lam <- e$values
plot(lam, xlab='Eigenvalue Number', ylab='Eigenvalue Size', main='Scree Plot')
lines(lam)
```



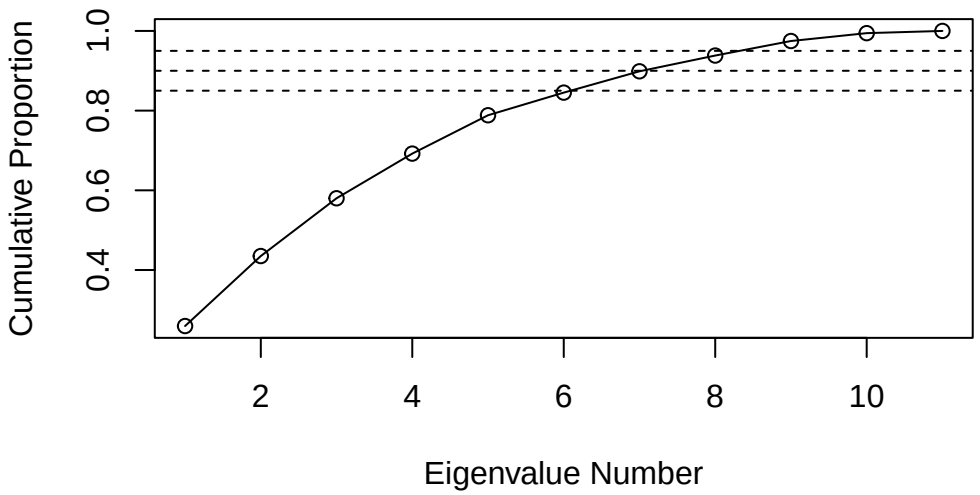

```
pi <- lam/sum(lam)
pi
```

```
##  [1] 0.25937764 0.17566973 0.14501656 0.11208679 0.09620568 0.05682151
##  [7] 0.05350101 0.03955572 0.03644172 0.01973123 0.00559241
```

```
pi.cumul <- c()
for (i in 1:length(lam)) {
    pi.cumul[i] <- sum(pi[1:i])
}
pi.cumul
```

```
##  [1] 0.2593776 0.4350474 0.5800639 0.6921507 0.7883564 0.8451779 0.8986789
##  [8] 0.9382346 0.9746764 0.9944076 1.0000000
```

```
plot(pi.cumul, xlab='Eigenvalue Number', ylab='Cumulative Proportion', main='')
lines(pi.cumul)
abline(0.85, 0, lty=2)
abline(0.90, 0, lty=2)
abline(0.95, 0, lty=2)
```




```
Z <- X%*%V
diag(var(Z))
```

```
##  [1] 2.85315405 1.93236701 1.59518215 1.23295471 1.05826252 0.62503663
##  [7] 0.58851109 0.43511294 0.40085892 0.21704348 0.06151651
```

```
pcr <- lm(quality ~ Z[,1:6])
summary(pcr)
```

```
##
## Call:
## lm(formula = quality ~ Z[, 1:6])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.88467 -0.26027 -0.02567  0.36085  1.61053
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.530000   0.061849  89.412  < 2e-16 ***
## Z[, 1:6]1   -0.009637   0.036800  -0.262 0.794006
## Z[, 1:6]2    0.260612   0.044717   5.828 8.02e-08 ***
## Z[, 1:6]3   -0.183066   0.049216  -3.720 0.000341 ***
## Z[, 1:6]4    0.015021   0.055981   0.268 0.789050
## Z[, 1:6]5    0.050420   0.060425   0.834 0.406185
## Z[, 1:6]6   -0.276454   0.078625  -3.516 0.000679 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6185 on 93 degrees of freedom
```

```
## Multiple R-squared:  0.3961, Adjusted R-squared:  0.3572
## F-statistic: 10.17 on 6 and 93 DF,  p-value: 1.257e-08
```

I decided to use the first 6 principle components, as they cover about 85% of the variance in wine quality. Comparing this to the original full model, we see that this model has the residual SE (0.62) and R^2_{adj} (0.36) really close to the original model's values. And adding more principle components probably will not help too much.

c.

```
pcr2 <- lm(quality ~ Z)
b_hat_pca <- as.matrix(pcr2$coefficients[2:12], nr=11)
b_hat <- V%*%b_hat_pca
b_hat
```

```
##               [,1]
##  [1,] -0.106631192
##  [2,] -0.007932931
##  [3,]  0.021256628
##  [4,]  0.035488595
##  [5,] -0.005475145
##  [6,]  0.062746500
##  [7,] -0.119487511
##  [8,] -0.041640818
##  [9,] -0.216977283
## [10,]  0.061484941
## [11,]  0.470853872
```

Compared to the original model, these estimates for beta are now all really small, instead of a mix of small and big values. There are also some coefficients, like the last one (alcohol) is is kept about the same, probably because it was significant.