Final Project

Daniel Zhang
PID: A16500214
Math 189
Spring 2023

Application Problems

1. a)

```
library(ISLR2)
head(Carseats)
```

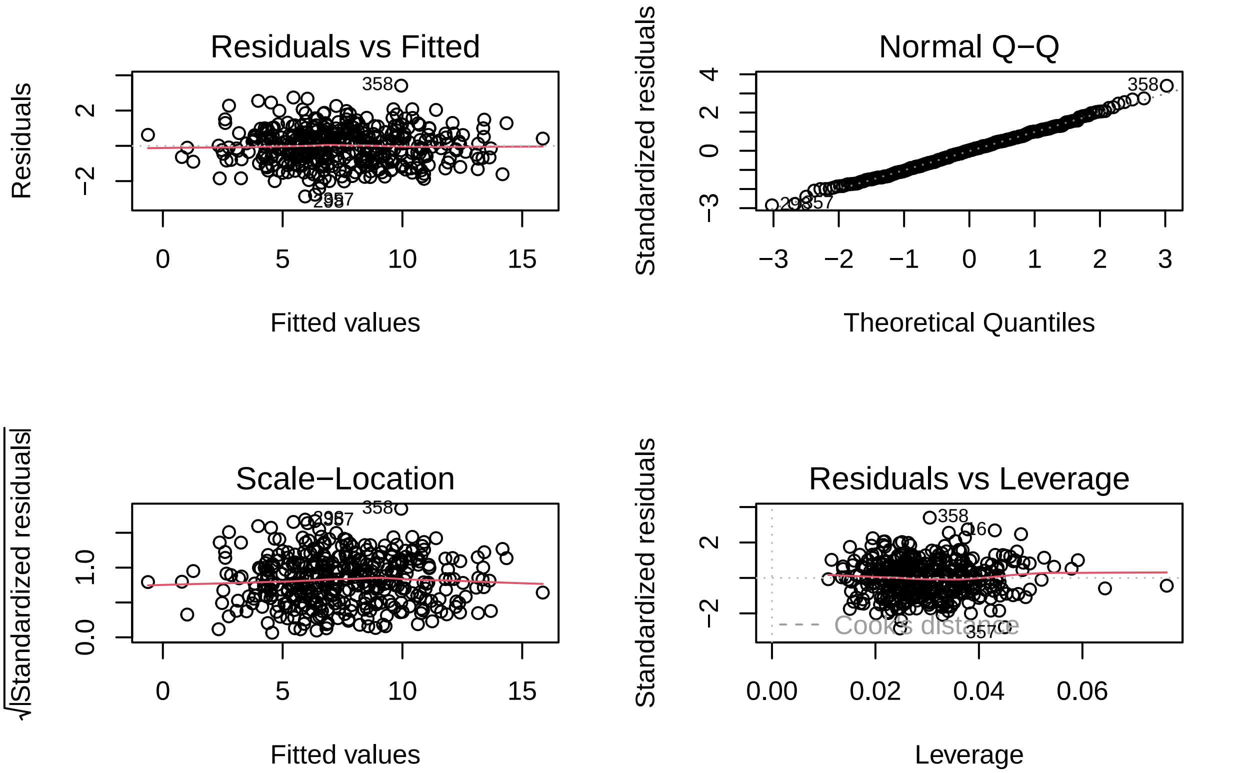
```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age Education
## 1  9.50       138     73          11        276   120       Bad  42        17
## 2 11.22       111     48          16        260    83      Good  65        10
## 3 10.06       113     35          10        269    80    Medium  59        12
## 4  7.40       117    100           4        466    97    Medium  55        14
## 5  4.15       141     64           3        340   128       Bad  38        13
## 6 10.81       124    113          13        501    72       Bad  78        16
##   Urban  US
## 1   Yes Yes
## 2   Yes Yes
## 3   Yes Yes
## 4   Yes Yes
## 5   Yes  No
## 6    No Yes
```

```
model <- lm(Sales ~., data = Carseats)
model
```

```
##
## Call:
## lm(formula = Sales ~ ., data = Carseats)
##
## Coefficients:
##     (Intercept)        CompPrice           Income      Advertising
##       5.6606231        0.0928153        0.0158028        0.1230951
##      Population            Price    ShelveLocGood  ShelveLocMedium
##       0.0002079       -0.0953579        4.8501827        1.9567148
##             Age        Education         UrbanYes            USYes
##      -0.0460452       -0.0211018        0.1228864       -0.1840928
```

1. b)

```
par(mfrow = c(2,2))
plot(model)
```



From the plots above, we see that the model is not violating any assumptions such as linearity or normality. The linear model should be appropriate.

1. c)

```
summary(model)
```

```
##
## Call:
## lm(formula = Sales ~ ., data = Carseats)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8692 -0.6908  0.0211  0.6636  3.4115
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      5.6606231  0.6034487   9.380  < 2e-16 ***
## CompPrice        0.0928153  0.0041477  22.378  < 2e-16 ***
## Income           0.0158028  0.0018451   8.565 2.58e-16 ***
## Advertising      0.1230951  0.0111237  11.066  < 2e-16 ***
## Population       0.0002079  0.0003705   0.561    0.575
## Price           -0.0953579  0.0026711 -35.700  < 2e-16 ***
## ShelveLocGood    4.8501827  0.1531100  31.678  < 2e-16 ***
## ShelveLocMedium  1.9567148  0.1261056  15.516  < 2e-16 ***
```

```
## Age              -0.0460452  0.0031817 -14.472  < 2e-16 ***
## Education        -0.0211018  0.0197205  -1.070    0.285
## UrbanYes          0.1228864  0.1129761   1.088    0.277
## USYes            -0.1840928  0.1498423  -1.229    0.220
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.019 on 388 degrees of freedom
## Multiple R-squared:  0.8734, Adjusted R-squared:  0.8698
## F-statistic: 243.4 on 11 and 388 DF,  p-value: < 2.2e-16
```

The null hypothesis is that the coefficients for CompPrice and Income are equal to zero. The alternative hypothesis is that the coefficients for CompPrice and Income are not equal to zero. The test statistic is the t-test statistic which has a normal distribution. An appropriate significance level is 0.05.

From the table above, we see that the p-value for CompPrice and Income are both below the significance level and we reject the null hypothesis.

2. a)

```
sample <- sample.int(n = nrow(Carseats), size = floor(0.8*nrow(Carseats)), replace = F)
train = Carseats[sample,]
nrow(train)
```

```
## [1] 320
```

```
test = Carseats[-sample,]
nrow(test)
```

```
## [1] 80
```

The proportions for the train/test split are 80/20.

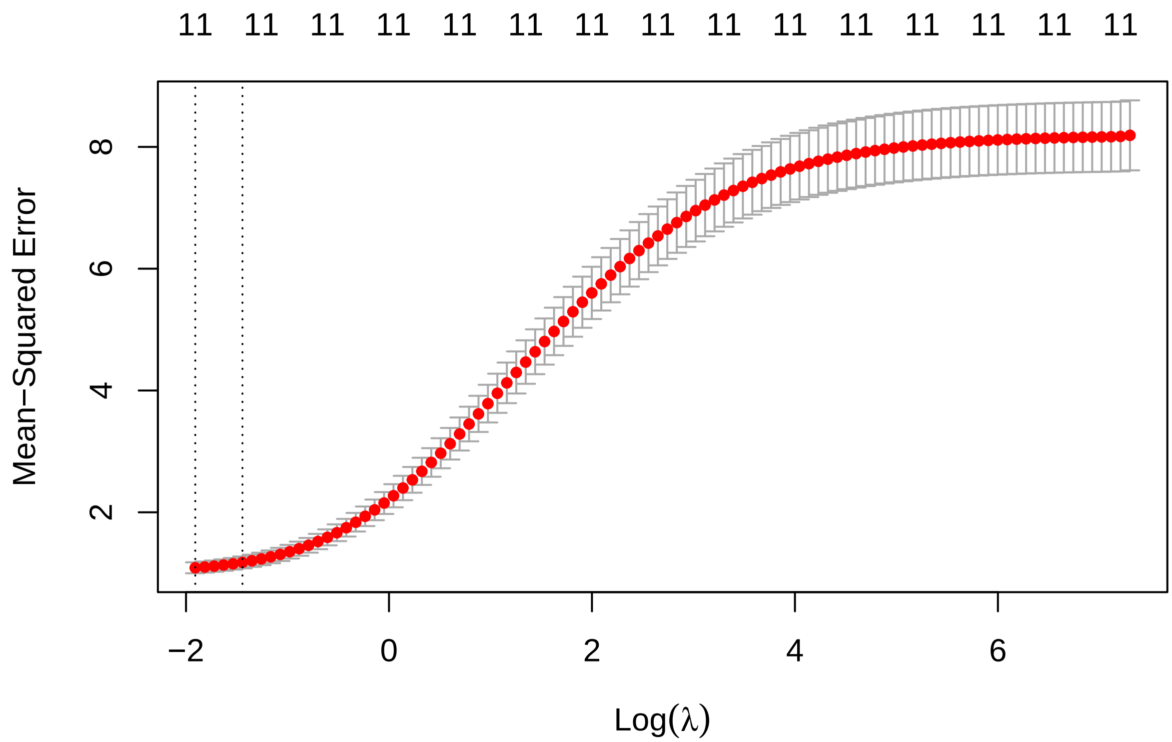
2. b)

```
library(glmnet)
```

```
## Loading required package: Matrix
```

```
## Loaded glmnet 4.1-7
```

```
x <- model.matrix(Sales ~ ., train)[, -1]
y <- train$Sales
set.seed(1)
cv.out <- cv.glmnet(x, y, alpha = 0)
plot(cv.out)
```




```
bestlam <- cv.out$lambda.min
bestlam
```

```
## [1] 0.1482363
```

```
coef(cv.out)
```

```
## 12 x 1 sparse Matrix of class "dgCMatrix"
##                           s1
## (Intercept)      6.8501486799
## CompPrice        0.0721633471
## Income           0.0149284924
## Advertising      0.1020216646
## Population       0.0005230278
## Price           -0.0810258502
## ShelveLocGood    4.1239547228
## ShelveLocMedium  1.2942413609
## Age             -0.0408033573
## Education       -0.0212549890
## UrbanYes         0.0604243760
## USYes            0.0379632895
```

2. c)

```
x <- model.matrix(Sales ~ ., test)[, -1]
data <- data.frame(pred = predict(cv.out, s = bestlam, newx = x), actual = test$Sales)
head(data)
```

```
##           s1 actual
## 2  12.140019  11.22
## 5   6.277755   4.15
## 6   9.818029  10.81
## 7   6.109526   6.63
## 20  7.498210   8.73
## 21  6.522399   6.41
```

```
sqrt(mean((data$actual - data$s1)^2))
```

```
## [1] 1.230363
```

2. d)

```
library(randomForest)
```

```
## randomForest 4.7-1.1
```

```
## Type rfNews() to see new features/changes/bug fixes.
```

```
library(Metrics)
set.seed(1)
rf <- randomForest(Sales ~ ., data = train, mtry = 10, ntree = 25, importance = TRUE)
rf
```

```
##
## Call:
##  randomForest(formula = Sales ~ ., data = train, mtry = 10, ntree = 25,      importance = TRUE)
##                Type of random forest: regression
##                      Number of trees: 25
## No. of variables tried at each split: 10
##
##           Mean of squared residuals: 2.841933
##                     % Var explained: 64.99
```

```
importance(rf)
```

```
##                 %IncMSE IncNodePurity
## CompPrice    9.0611605    241.371408
## Income       2.0756773    139.993412
## Advertising  5.6093425    211.799066
## Population   0.6957289     77.634353
## Price       13.6514033    815.261017
## ShelveLoc   16.3394765    740.199884
## Age          3.6688059    202.656870
## Education    3.6404284     71.252993
## Urban       -0.1668479      7.668564
## US           0.2569514     18.480809
```

```
rmse(test$Sales, predict(rf,test))
```

```
## [1] 1.858639
```

2. e) A marketing team may prefer the ridge regression model in (b) because it has a lower RMSE. Another marketing team may prefer the random forest model because it considers price as being important while the ridge regression model does not.

3. a)

```
set.seed(1)
X <- rt(200, 15)
summary(X)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -2.268942 -0.665461 -0.008478  0.065234  0.759181  3.230585
```

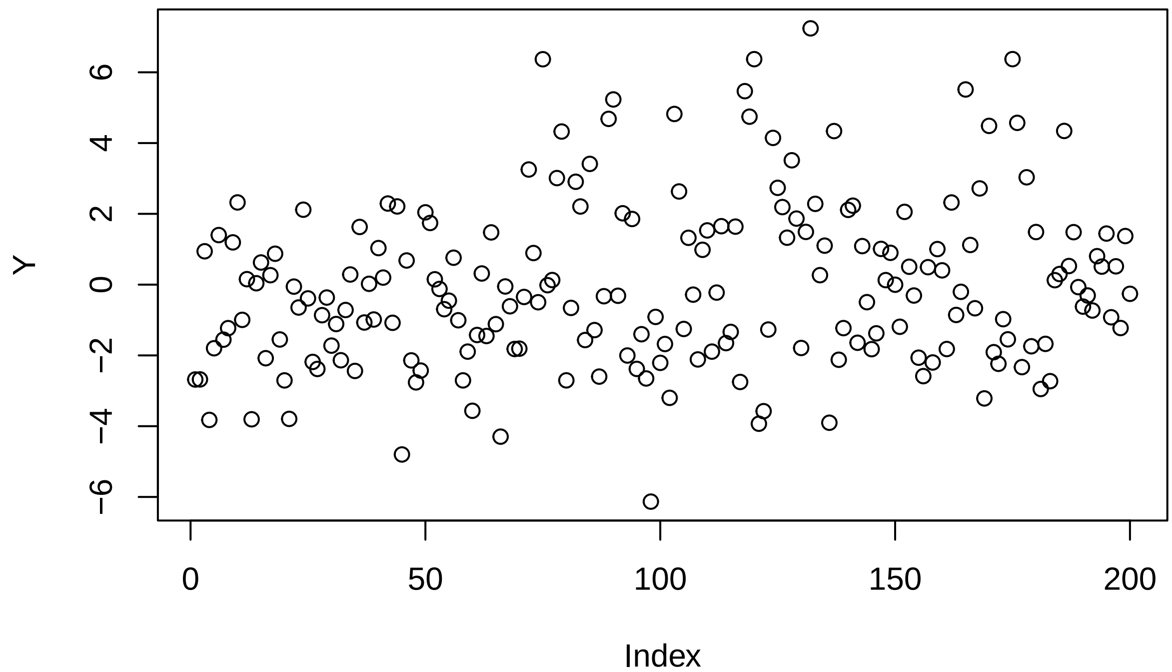
3. b)

```
noise <- rt(200, 5)
summary(noise)
```

```
##      Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
## -4.675369 -0.815115 -0.010529  0.007735  0.756192  4.353930
```

3. c)

```
Y = 5 + 2 * sin(X) - 7 * ( exp(2 * cos(X)) / (1 + exp(2 * cos(X)))) + noise
plot(Y)
```



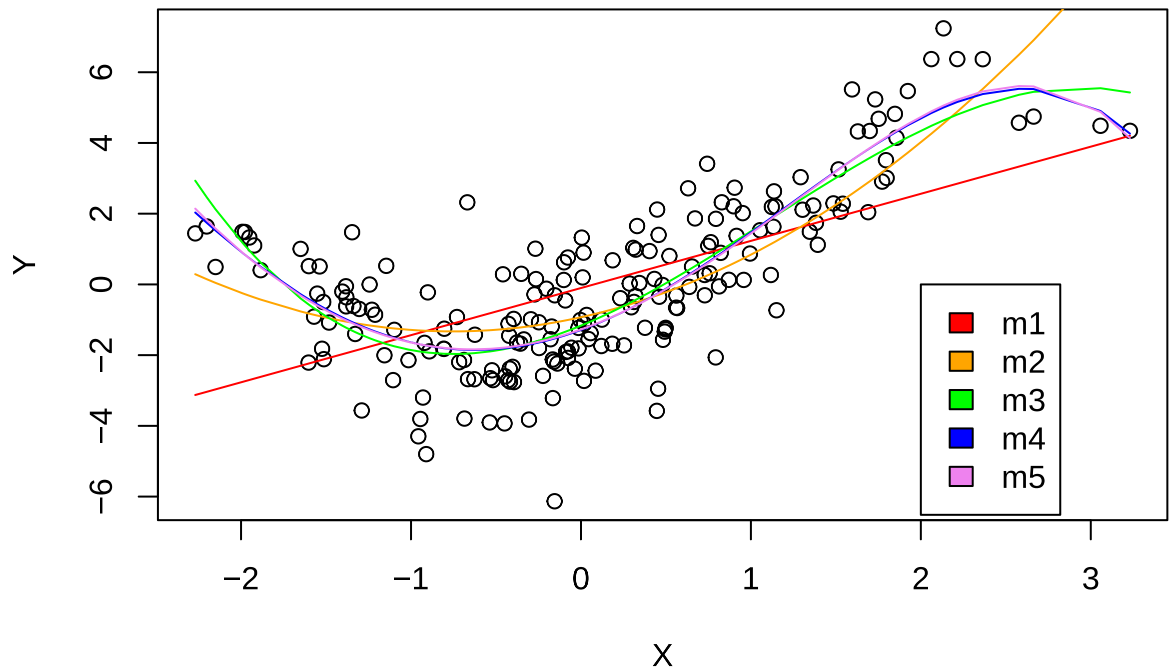

```
summary(Y)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -6.13140 -1.72791 -0.30768 -0.01605  1.41048  7.24075
```

3. d)

```
df <- data.frame(Y,X)

plot(X, Y)
color <- c("red","orange","green","blue","violet")
for (index in c(0:4))
{
  m <- lm(Y ~ poly(X, index + 1, raw = TRUE), data = df)
  c <- color[index + 1]
  x<-sort(X)
  y<-m$fitted.values[order(X)]
  lines(x, y, col=c)
}
legend(2, y= 0, paste0("m", 1:5), fill=color)
```



3. e) I prefer the model with X to the order of 2 or 3. They neither under-fitted like the linear m1 nor over-fitted like m4 and m5, which barely differ from each other.

   f)

```
m2 <- lm(Y ~ poly(X, 2, raw = TRUE), data = df)
predict(m2, newdata = data.frame(X=c(1)), interval = 'confidence')
```

```
##         fit       lwr      upr
## 1 0.8402292 0.5580371 1.122421
```

We are 90% confident that Y is between [0.5580371, 1.122421] when X = 1.

3. g)

```
library(boot)
fun <- function(data, idx)
{
  d <- data[idx, ]
  m2 <- lm(Y ~ poly(X, 2, raw = TRUE), data = d)
  predict(m2, newdata = data.frame(X=c(1)))
}
bootstrap <- boot(df, fun, R = 1000)
boot.ci(boot.out = bootstrap, type = c("norm"))
```

```
## BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
## Based on 1000 bootstrap replicates
##
## CALL :
## boot.ci(boot.out = bootstrap, type = c("norm"))
##
## Intervals :
## Level      Normal
## 95%   ( 0.5359,  1.1099 )
## Calculations and Intervals on Original Scale
```

We are 90% confident that Y is between [0.5359, 1.1099] when X = 1.

4. a)

```
data(College)
head(College)
```

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

```
set.seed(1)
sample <- sample.int(n = nrow(College), size = floor(0.8*nrow(College)), replace = F)
train = College[sample,]
nrow(train)
```

```
## [1] 621
```

```r
test = College[-sample,]
nrow(test)
```

```
## [1] 156
```

4. b)

```r
logreg <- glm(Private ~ ., train,family="binomial")
logreg
```

```
##
## Call:  glm(formula = Private ~ ., family = "binomial", data = train)
##
## Coefficients:
## (Intercept)         Apps       Accept       Enroll    Top10perc    Top25perc
##   7.607e-02   -4.834e-04    7.605e-04    6.249e-04    1.576e-03    4.751e-03
## F.Undergrad  P.Undergrad     Outstate   Room.Board        Books     Personal
##  -7.738e-04    1.876e-04    6.896e-04    2.002e-05    1.519e-03   -7.255e-05
##         PhD     Terminal    S.F.Ratio  perc.alumni       Expend    Grad.Rate
##  -5.717e-02   -3.445e-02   -5.199e-02    4.593e-02    2.056e-04    1.652e-02
##
## Degrees of Freedom: 620 Total (i.e. Null);  603 Residual
## Null Deviance:       712.9
## Residual Deviance: 186.5     AIC: 222.5
```

The statistic of Top10perc is the percentage of new students being from the top 10% of high school classes. The coefficient for Top10perc can be understood as how important this statistic is as a factor of a college being public or private. Currently, it seems the percentage of new students from the top 10% of high school classes is not an important factor in whether a college is public or private.

4. c)

```r
prob <- predict(logreg,newdata = test, type = "response")
predicted <- ifelse(prob > 0.5, "Yes", "No")
1 - mean(predicted == test$Private)
```

```
## [1] 0.05769231
```

4. d)

```r
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked from 'package:ISLR2':
##
##     Boston
```

```
lda_model = lda(Private ~ ., train)
predicted <- predict(lda_model,newdata = test)$class
1 - mean(predicted == test$Private)
```

```
## [1] 0.03846154
```

4. e)

```
qda_model = qda(Private ~ ., train)
predicted <- predict(qda_model,newdata = test)$class
1 - mean(predicted == test$Private)
```

```
## [1] 0.07692308
```

4. f)

```
library(e1071)
svm_model <- svm(Private ~ ., train)
predicted <- predict(svm_model,newdata = test)
1 - mean(predicted == test$Private)
```

```
## [1] 0.05128205
```

4. g) I picked the LDA model because it has the lowest test error.

5. a)

```
library(MultBiplotR)
```

```
##
## Attaching package: 'MultBiplotR'

## The following object is masked from 'package:MASS':
##
##     ginv

## The following object is masked from 'package:boot':
##
##     logit
```

```
data(Protein)
head(Protein)
```

```
##                 Comunist Region Red_Meat White_Meat Eggs Milk Fish Cereal Starch
## Albania              Yes  South     10.1        1.4  0.5  8.9  0.2   42.3    0.6
## Austria               No Center      8.9       14.0  4.3 19.9  2.1   28.0    3.6
## Belgium               No Center     13.5        9.3  4.1 17.5  4.5   26.6    5.7
## Bulgaria             Yes  South      7.8        6.0  1.6  8.3  1.2   56.7    1.1
## Czechoslovakia       Yes Center      9.7       11.4  2.8 12.5  2.0   34.3    5.0
```

```
## Denmark              No  North     10.6       10.8  3.7 25.0  9.9   21.9    4.8
##                Nuts Fruits_Vegetables
## Albania         5.5               1.7
## Austria         1.3               4.3
## Belgium         2.1               4.0
## Bulgaria        3.7               4.2
## Czechoslovakia  1.1               4.0
## Denmark         0.7               2.4
```

```
p <- subset(Protein, select = -c(Comunist,Region))
pca = prcomp(p, scale. = TRUE, rank. =5)
summary(pca)
```

```
## Importance of first k=5 (out of 9) components:
##                           PC1    PC2    PC3    PC4     PC5
## Standard deviation     2.0016 1.2787 1.0620 0.9771 0.68106
## Proportion of Variance 0.4452 0.1817 0.1253 0.1061 0.05154
## Cumulative Proportion  0.4452 0.6268 0.7521 0.8582 0.90976
```

5. b)

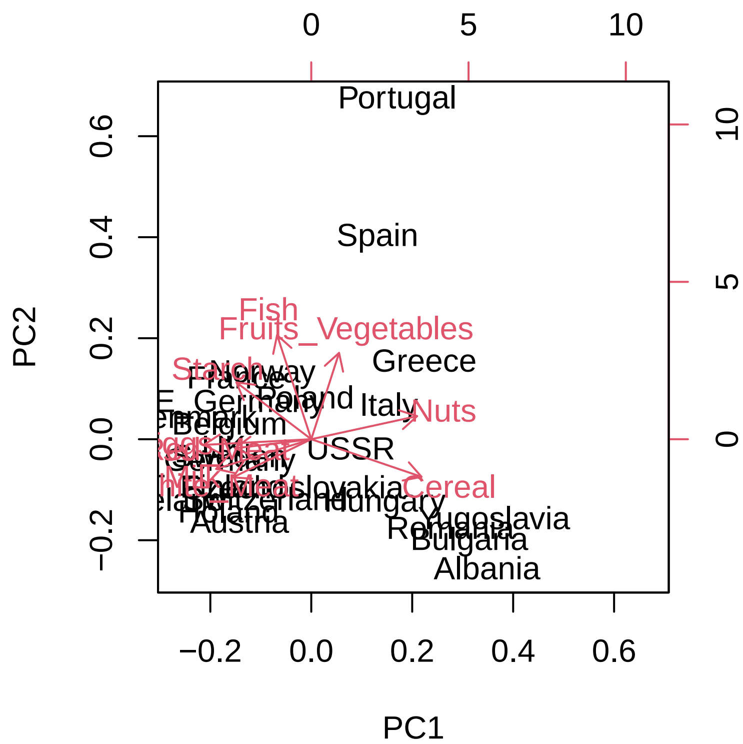
```
pca
```

```
## Standard deviations (1, .., p=9):
## [1] 2.0016087 1.2786710 1.0620355 0.9770691 0.6810568 0.5702026 0.5211586
## [8] 0.3410160 0.3148204
##
## Rotation (n x k) = (9 x 5):
##                          PC1         PC2         PC3          PC4         PC5
## Red_Meat          -0.3026094 -0.05625165 -0.29757957 -0.646476536  0.32216008
## White_Meat        -0.3105562 -0.23685334  0.62389724  0.036992271 -0.30016494
## Eggs              -0.4266785 -0.03533576  0.18152828 -0.313163873  0.07911048
## Milk              -0.3777273 -0.18458877 -0.38565773  0.003318279 -0.20041361
## Fish              -0.1356499  0.64681970 -0.32127431  0.215955001 -0.29003065
## Cereal             0.4377434 -0.23348508  0.09591750  0.006204117  0.23816783
## Starch            -0.2972477  0.35282564  0.24297503  0.336684733  0.73597332
## Nuts               0.4203344  0.14331056 -0.05438778 -0.330287545  0.15053689
## Fruits_Vegetables  0.1104199  0.53619004  0.40755612 -0.462055746 -0.23351666
```

The first principle component has negative associations with non-fish meat and starch, while also having large positive associations with cereal and nuts. The second principle component has large positive associations with fish as well as fruits and vegetables. These two components measure different dietary habits.

5. c)

```
biplot(pca)
```



Based on the plot above, milk is most positively correlated with white meat, most negatively correlated with nuts, and uncorrelated with fish and fruits.

5. d)

```r
reg <- Protein[Protein$Region == 'North' | Protein$Region == "Center",]
subset(reg, select = Region)
```

```
##                Region
## Austria        Center
## Belgium        Center
## Czechoslovakia Center
## Denmark         North
## E_Germany      Center
## Finland         North
## France         Center
## Hungary        Center
## Ireland        Center
## Holand         Center
## Norway          North
## Poland         Center
## Sweden          North
## Switzerland    Center
## UK             Center
## USSR           Center
## W_Germany      Center
```

```
summary(pca)
```

```
## Importance of first k=5 (out of 9) components:
##                           PC1    PC2    PC3    PC4     PC5
## Standard deviation     2.0016 1.2787 1.0620 0.9771 0.68106
## Proportion of Variance 0.4452 0.1817 0.1253 0.1061 0.05154
## Cumulative Proportion  0.4452 0.6268 0.7521 0.8582 0.90976
```

Countries in the north and central regions are grouped close together in the biplot. However, some countries in the north region such as Denmark and Norway are located higher on PC2 compared to countries in the center region. This suggests that countries in the north region have higher consumption of fish and fruits/vegetables.

Conceptual Problems

6. For linear regression, bootstrapping can help validate the model and its confidence intervals. For random forest, bagging uses bootstrapping to reduce variability, which is more helpful.
7. FEWR and FDR are about the rates of type I errors, which are false positives. Correcting for FEWR and FDR may decrease type I errors, but it will also increase type II errors, which are false negatives. This should not be done if the cost of false negatives is higher than the cost of false positives, such as covid test results.
8. Assumptions such as linearity and normality need to be checked because they affect the accuracy of the model. For example, if there is a pattern in the residual plot for a linear model, then it might not have a linear relationship and the model should not be used for inference or prediction.