Quiz 2

Modern Data Mining

March 2, 2021

Instruction: This is an open book, 10-15 minute quiz. Answer all 9 questions and choose the correct answer.

The first portion of the quiz uses a subset of 200 subjects that are randomly chosen from `IQ.Full.csv`. From this dataset we extracted their 4 AFQT tests: `Arith`, `Word`, `Parag` and `Math`. The dataset is named `afqt`.

```
data.full <- read.csv("IQ.Full.csv")
data1 <- data.full %>% select(Arith, Word, Parag, Math)
set.seed(1)
n <- dim(data1)[1]
afqt <- data1[sample(n, 200, replace = FALSE), ]  # take 200 people
names(afqt)
```

```
## [1] "Arith" "Word"  "Parag" "Math"
```

```
afqt.stat <- summary(afqt)
afqt.mean <- colMeans(afqt)
afqt.sd <- apply(afqt, 2, sd)
afqt.mean
```

```
## Arith  Word Parag  Math
##  18.4  26.4  11.1  14.2
```

```
afqt.sd
```

```
## Arith  Word Parag  Math
##  7.07  7.37  3.26  6.42
```

1. We first perform PCA to summarize the set of four tests. The four tests are first centered and scaled.

```
afqt.pca <- prcomp(afqt, center = TRUE, scale. = TRUE)
afqt.pca$rotation
```

```
##         PC1    PC2      PC3     PC4
## Arith 0.502 -0.518  0.00136 -0.6928
## Word  0.503  0.394 -0.76584  0.0682
## Parag 0.489  0.603  0.62320 -0.0956
## Math  0.506 -0.461  0.15846  0.7115
```

PC1 scores are *approximately* equal to:

(A) .5 (Arith + Word + Parag + Math)

(B) .5 [(Arith - 18.41)/7.068) + (Word - 26.39)/7.374) + (Parag - 11.095)/3.256) + (Math - 14.21)/6.416)]

Answer (B): Since we ran PCA for centered and scaled scores, that means PC1 is obtained by (B)

2. The `PC1` score of `afqt.pca` in question 1 has the largest variance among all 4 PC scores.

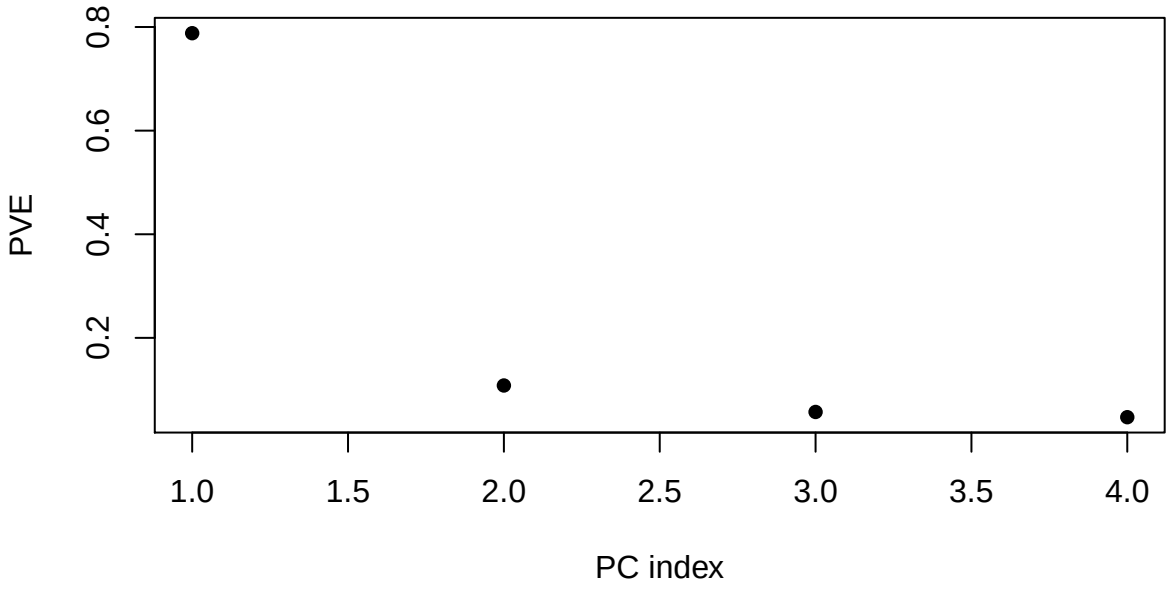
(A) True

(B) False

Answer (A): The goal of doing PCA is to find a new set of uncorrelated scores such that `PC1` has largest variance, then `PC2`...

3. Based on the following PVE plot we see that

```
plot(summary(afqt.pca)$importance[2, ], pch = 16, xlab = "PC index", ylab = "PVE")
```



(A) PC1 accounts for approximately 80% of the total variance among the 4 PCs

(B) PC1 accounts for approximately 20% of the total variance among the 4 PCs

Answer (A): By definition of PVE we know that var(PC1) is 80% of the total variances among 4 PC's.

We next run a kmeans clustering analysis specifying 2 clusters.

```
afqt.kmeans <- kmeans(afqt, centers = 2)
afqt.kmeans$size
```

```
## [1] 116  84
```

4 Choose the correct answer:

(A) There are 100 subjects in cluster 1 and another 100 in cluster 2

(B) There are 116 in cluster 1 and 84 in cluster 2.

Answer (B)

The remaining quiz questions are about regression. We will use a subset from the `Cars_04` data that has been used in class. We will use `MPG_Hwy` as the response variable.

Let us first take a subset of the data and name it car.data.

```
set.seed(10)
car.temp <- read.csv("Cars_04.csv")
s.index <- sample(nrow(car.temp), 200)
car.data <- car.temp[s.index, ]
summary(car.data)
```

```
##                     Make.Model  Continent    MPG_City      Horsepower
##  Acura_MDX               :  1   Am:61     Min.   :10.0   Min.   : 65
##  Acura_NSX               :  1   As:81     1st Qu.:16.0   1st Qu.:160
##  Acura_RL                :  1   E :58     Median :19.0   Median :203
##  Acura_RSX               :  1             Mean   :19.4   Mean   :226
##  Acura_TSX               :  1             3rd Qu.:22.0   3rd Qu.:275
```

```
##  Aston_Martin_V12_Vanquish:  1             Max.   :60.0   Max.   :605
##  (Other)                   :194
##      Weight        Seating         Length       MPG_Hwy         Origin
##  Min.   :1.98   Min.   :2.00   Min.   :143   Min.   :14.0   Min.   :1.00
##  1st Qu.:3.11   1st Qu.:5.00   1st Qu.:177   1st Qu.:22.0   1st Qu.:1.00
##  Median :3.54   Median :5.00   Median :187   Median :26.0   Median :2.00
##  Mean   :3.67   Mean   :4.93   Mean   :186   Mean   :25.9   Mean   :2.06
##  3rd Qu.:4.06   3rd Qu.:5.00   3rd Qu.:192   3rd Qu.:29.0   3rd Qu.:3.00
##  Max.   :5.82   Max.   :8.00   Max.   :224   Max.   :56.0   Max.   :3.00
##
##         Transmission      EPA_Class      Width       Displacement
##  automatic    :184   suv2wd    :38   Min.   :65.4   Min.   :1.00
##  cont_variable:  6   compact   :35   1st Qu.:69.5   1st Qu.:2.40
##  manual       : 10   midsize   :30   Median :71.7   Median :3.20
##                      two_seater:22   Mean   :72.1   Mean   :3.31
##                      suv4wd    :19   3rd Qu.:74.7   3rd Qu.:4.20
##                      large     :18   Max.   :80.5   Max.   :8.30
##                      (Other)   :38
##    Cylinders            Make                Model        Turndiam
##  Min.   : 2.00   Chevrolet : 12   3              :  1   Min.   :30.2
##  1st Qu.: 4.00   Toyota    : 12   300M           :  1   1st Qu.:35.4
##  Median : 6.00   Volkswagen:  9   360_Modena     :  1   Median :37.1
##  Mean   : 5.88   Honda     :  8   4RunnerSR5     :  1   Mean   :37.2
##  3rd Qu.: 6.00   Mitsubishi:  8   525i           :  1   3rd Qu.:38.7
##  Max.   :12.00   Cadillac  :  7   575M_Maranello:  1   Max.   :43.5
##                  (Other)   :144   (Other)       :194   NA's   :51
```

We then fit a linear model `fit1: MPG_Hwy vs. Horsepower`

```
fit1 <- lm(MPG_Hwy ~ Horsepower, car.data)
fit1.s <- summary(fit1)
fit1.s
```

```
##
## Call:
## lm(formula = MPG_Hwy ~ Horsepower, data = car.data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -10.324  -2.785   0.042   2.322  23.024
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 35.82117    0.81791    43.8   <2e-16 ***
## Horsepower  -0.04377    0.00334   -13.1   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.44 on 198 degrees of freedom
## Multiple R-squared:  0.465,  Adjusted R-squared:  0.462
## F-statistic:  172 on 1 and 198 DF,  p-value: <2e-16
```

5. Based on summary of `fit1`, choose correct answer(s).

(A) On average `MPG_Hwy` decreases 0.044 when `Horsepower` increases by 1.

(B) Take two cars, car1 with `Horsepower=220` and car2 with `Horsepower=221`; `fit1` tells us `MPG_Hwy` is

guaranteed to be higher in car1 than car2.

Answer (A): Though the mean of `MPG_Hwy` for cars with `Horsepower=220` is higher than that of cars with `Horsepower=221`, the `MPG_Hwy`s can be larger or small comparing two individual cars from each group.

Next, we add one variable `Weight` to `fit1` and store the result in `fit2`.

```
fit2 <- lm(MPG_Hwy ~ Horsepower + Weight, car.data)
fit2.s <- summary(fit2)
fit2.s
```

```
##
## Call:
## lm(formula = MPG_Hwy ~ Horsepower + Weight, data = car.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.871 -1.815 -0.316  1.738 18.753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.98899    1.20757   38.91   <2e-16 ***
## Horsepower  -0.02803    0.00301   -9.33   <2e-16 ***
## Weight      -4.01044    0.36626  -10.95   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.51 on 197 degrees of freedom
## Multiple R-squared:  0.667,  Adjusted R-squared:  0.664
## F-statistic:  197 on 2 and 197 DF,  p-value: <2e-16
```

6. From `fit2`, we see that 1 unit increase in horsepower always results in a decrease in `MPG_Hwy` on average by 0.028.

(A) True

(B) False

Answer (B): Only if they have the same `Weight`s.

7. Based on `fit2`, we would like to estimate the mean of `MPG_Hwy` for all cars with the following measurements: `Horsepower = 240`, `Weight = 3.5`, with 4 seats and 180" long.

(A) We can not do it since `Seats` and `Length` are not included in the `fit2`

(B) It is

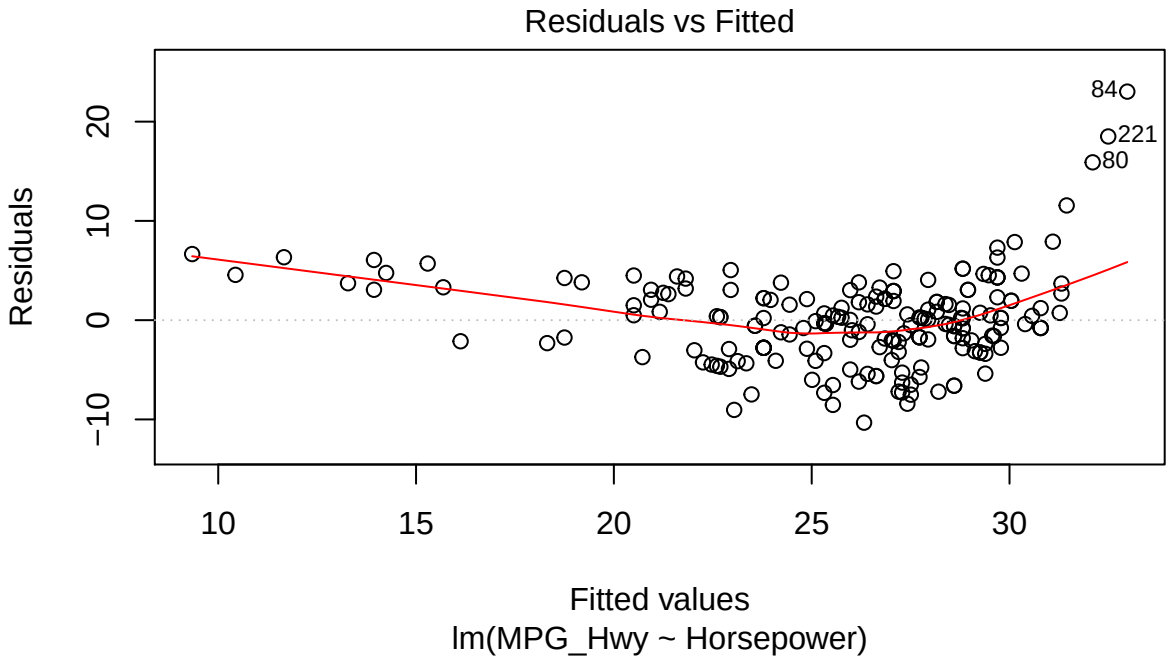
$$46.989 - 0.028 \times 240 - 4.01 \times 3.5$$

Answer (B): the predicition equation can be used as long as the predictors are give. On the other hand we can't use `fit2` to estimate the mean of cars with `Horsepower = 240`.

We didn't grade this question due to a typo.

Model diagnoses for `fit2`. Choose the correct answers.

```
plot(fit1, 1)
```



8. Choose one answer.

(A) The linearity might be a problem since cars with smaller `MPG_Hwy` seem to be underestimated.

(B) The linearity might be a problem since cars with smaller `MPG_Hwy` seem to be overestimated.

Answer: (A) The residuales ($y - \hat{y}$ are all > 0 for smaller `MPG_Hwy`.

9. `fit2` can be used to reject $H_0 : \beta_1 = \beta_2 = 0$ at a significance level of 0.001 for the following reason:

(A) Because of a large R^2.

(B) Because the F test in the summary report has a p-value much smaller than .001.

Answer: (B) Once again the larger R^2 is the more useful a model is. But we need to use F to see precisely how large R^2 to be to reject the null hypothesis at an $\alpha = .001$. Check the precise equation between R^2 and F in our lecture please.