q3

Lisong He

2024-04-10

a)

```
data <- read.csv('grocery-retailer.csv')
model <- lm(Y ~ X1 + X2 + X3, data=data)
hat_values <- hatvalues(model)
print(hat_values)
```

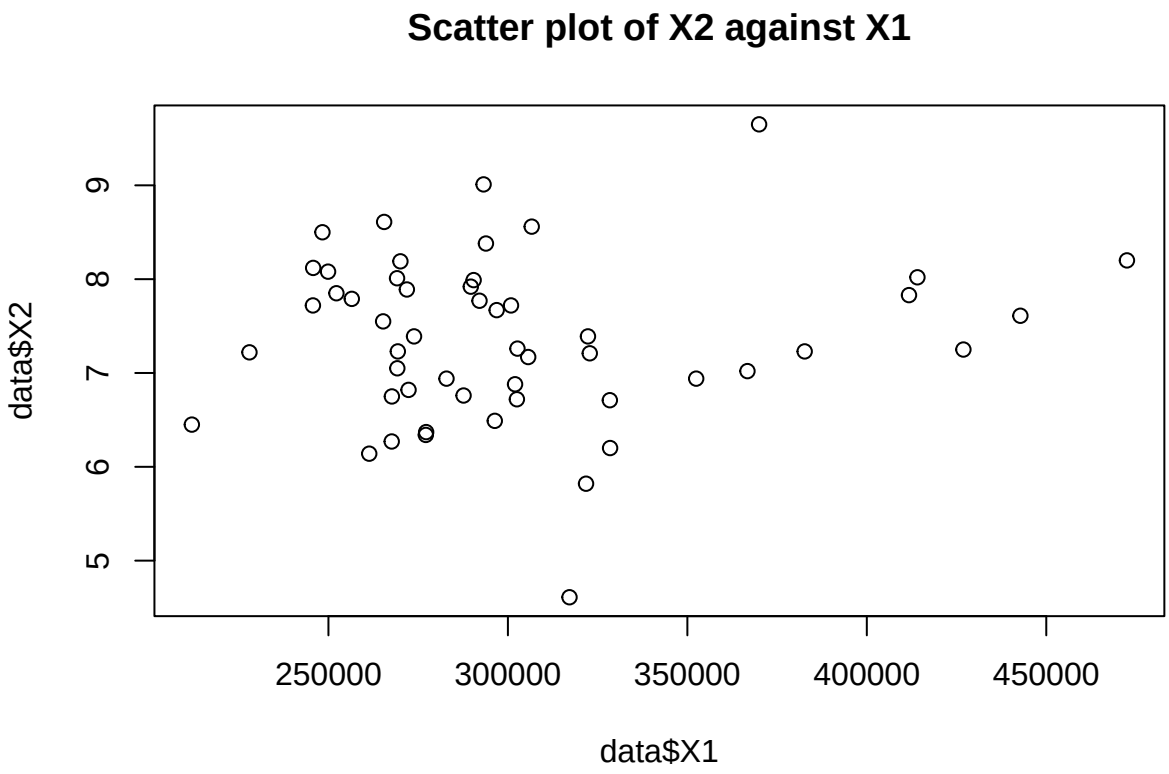
```
##          1          2          3          4          5          6          7
## 0.02258497 0.06179963 0.21887726 0.05297322 0.20632818 0.02712212 0.02861964
##          8          9         10         11         12         13         14
## 0.05635264 0.04017169 0.04826901 0.03011634 0.04977033 0.02761134 0.06047246
##         15         16         17         18         19         20         21
## 0.03756448 0.25542493 0.03324965 0.05104935 0.02561758 0.02491881 0.19360472
##         22         23         24         25         26         27         28
## 0.25771995 0.05677233 0.07959049 0.05613301 0.02189441 0.02697280 0.06097409
##         29         30         31         32         33         34         35
## 0.03684681 0.04174658 0.03663401 0.09602318 0.04193292 0.02517837 0.04621057
##         36         37         38         39         40         41         42
## 0.06622225 0.03108517 0.03204566 0.04903249 0.03210502 0.04373193 0.12395571
##         43         44         45         46         47         48         49
## 0.28685861 0.22002363 0.11050577 0.03159426 0.06494377 0.28177664 0.02446692
##         50         51         52
## 0.03420197 0.10278142 0.02754093
```

```
outlying_X <- which(hat_values > 2*4/52)
outlying_X
```

```
##  3  5 16 21 22 43 44 48
##  3  5 16 21 22 43 44 48
```

b)

```
plot(data$X1, data$X2, main="Scatter plot of X2 against X1")
```




```
X_new <- cbind(1, 300, 7.2, 0) # Assuming your model has an intercept
predict(model, newdata=data.frame(X1=300, X2=7.2, X3=0))
```

```
##        1
## 4055.328
```

Visually at X1 = 300000 and x2 = 7.2, the point is in the center of the data so no extrapolation.

```
X <- model.matrix(model, data)
XTX <- t(X) %*% X
XTX_inv <- solve(XTX)
X_new <- matrix(c(1, 300, 7.2, 0), nrow = 1)
lev <- X_new %*% XTX_inv %*% t(X_new)
print(paste('Hnew,new is ',lev, 'for this measurement.'))
```

```
## [1] "Hnew,new is  0.60789073186237 for this measurement."
```

```
print(paste('Range of hat values is [', min(hat_values), ',', max(hat_values), ']'))
```

```
## [1] "Range of hat values is [ 0.0218944051050263 , 0.2868586074036 ]"
```

Apparently, Hnew,new exceeds the bounds so it is a hidden extrapolation, which contradicts my previous observation.

c) Looking at DFFITS:

```
cases <- c(16, 22, 43, 48, 10, 32, 38, 40)
for (case in cases) {
  cat("Case", case, "has DFFITS of ", dffits(model)[case], "\n")
  if (abs(dffits(model)[case]) > 2 * sqrt(4/52)) {
    cat("Case", case, "is an influential observation on Y in terms of DFFITS", "\n")
  } else {
    cat("Case", case, "is not a influential observatio on Y in terms of DFFITS", "\n")
  }
  cat("\n")
}
```

```
## Case 16 has DFFITS of  -0.5539903
## Case 16 is not a influential observatio on Y in terms of DFFITS
##
## Case 22 has DFFITS of  0.05508583
## Case 22 is not a influential observatio on Y in terms of DFFITS
##
## Case 43 has DFFITS of  0.5616519
## Case 43 is an influential observation on Y in terms of DFFITS
##
## Case 48 has DFFITS of  -0.1468415
## Case 48 is not a influential observatio on Y in terms of DFFITS
##
## Case 10 has DFFITS of  0.458633
## Case 10 is not a influential observatio on Y in terms of DFFITS
##
## Case 32 has DFFITS of  -0.6510771
## Case 32 is an influential observation on Y in terms of DFFITS
##
## Case 38 has DFFITS of  0.3855177
## Case 38 is not a influential observatio on Y in terms of DFFITS
##
## Case 40 has DFFITS of  0.3967203
## Case 40 is not a influential observatio on Y in terms of DFFITS
```

Looking at cook's distance:

```
cases <- c(16, 22, 43, 48, 10, 32, 38, 40)
for (case in cases) {
  cook <- cooks.distance(model)[case]
  cat("Case", case, "has Cook's distance of", cook, "\n")
  if (cook <= 0.2) {
    cat("Case", case, "has no influence in terms of Cook's distance\n")
  } else if (cook > 0.2 && cook <= 0.5) {
    cat("Case", case, "has moderate influence in terms of Cook's distance\n")
  } else if (cook > 0.5) {
    cat("Case", case, "has major influence in terms of Cook's distance\n")
  }
  cat("\n")
}
```

```
## Case 16 has Cook's distance of 0.07689508
## Case 16 has no influence in terms of Cook's distance
##
## Case 22 has Cook's distance of 0.0007746088
## Case 22 has no influence in terms of Cook's distance
##
## Case 43 has Cook's distance of 0.07921931
## Case 43 has no influence in terms of Cook's distance
##
## Case 48 has Cook's distance of 0.005498867
## Case 48 has no influence in terms of Cook's distance
##
## Case 10 has Cook's distance of 0.04935012
## Case 10 has no influence in terms of Cook's distance
##
## Case 32 has Cook's distance of 0.09975974
## Case 32 has no influence in terms of Cook's distance
##
## Case 38 has Cook's distance of 0.03463803
## Case 38 has no influence in terms of Cook's distance
##
## Case 40 has Cook's distance of 0.03649915
## Case 40 has no influence in terms of Cook's distance
```

Looking at DFBETAS:

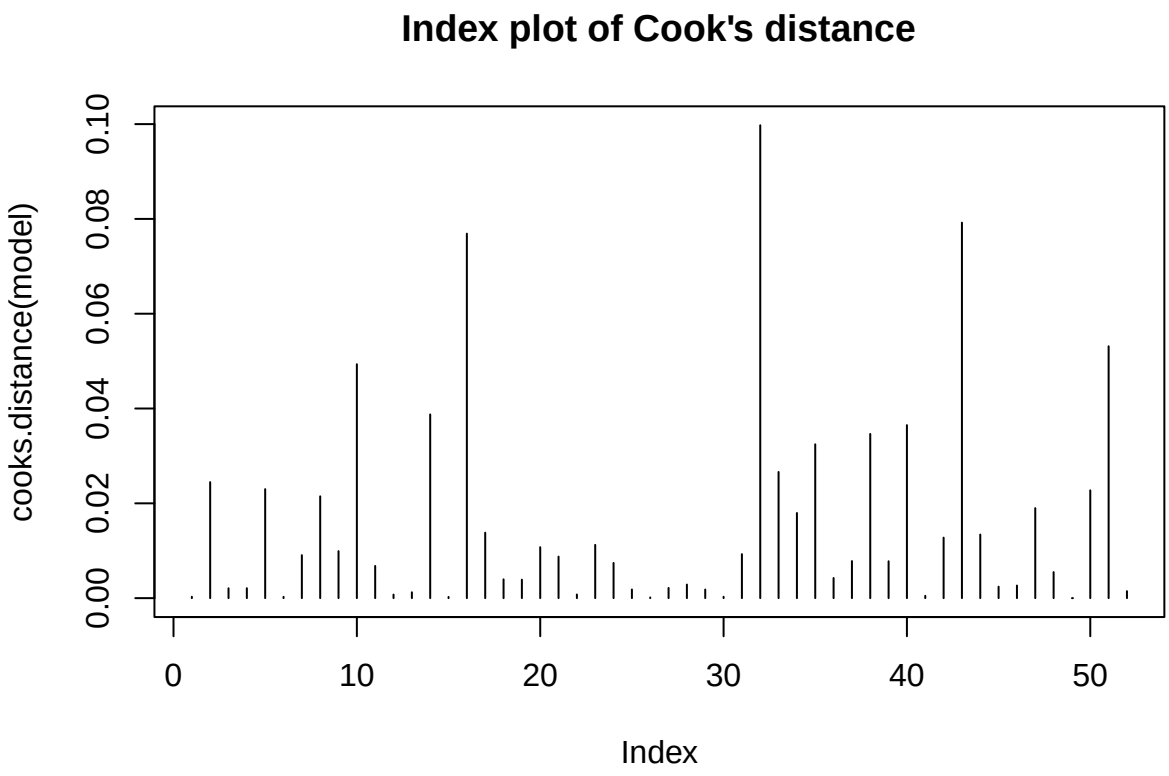
```r
cases <- c(16, 22, 43, 48, 10, 32, 38, 40)
for (case in cases) {
  cat("Case", case, "has DFBETAS of ", dfbetas(model)[case], "\n")
  if (abs(dfbetas(model)[case]) > 2/sqrt(52)) {
    cat("Case", case, "is an influential observation on Y in terms of DFBETAS", "\n")
  } else {
    cat("Case", case, "is not a influential observatio on Y in terms of DFBETAS", "\n")
  }
  cat("\n")
}
```

```
## Case 16 has DFBETAS of  -0.2476887
## Case 16 is not a influential observatio on Y in terms of DFBETAS
##
## Case 22 has DFBETAS of  0.03042319
## Case 22 is not a influential observatio on Y in terms of DFBETAS
##
## Case 43 has DFBETAS of  -0.3577973
## Case 43 is an influential observation on Y in terms of DFBETAS
##
## Case 48 has DFBETAS of  0.0449858
## Case 48 is not a influential observatio on Y in terms of DFBETAS
##
## Case 10 has DFBETAS of  0.3640749
## Case 10 is an influential observation on Y in terms of DFBETAS
##
## Case 32 has DFBETAS of  0.4095415
```

```
## Case 32 is an influential observation on Y in terms of DFBETAS
##
## Case 38 has DFBETAS of  -0.09961479
## Case 38 is not a influential observatio on Y in terms of DFBETAS
##
## Case 40 has DFBETAS of  0.07379876
## Case 40 is not a influential observatio on Y in terms of DFBETAS
```

d)

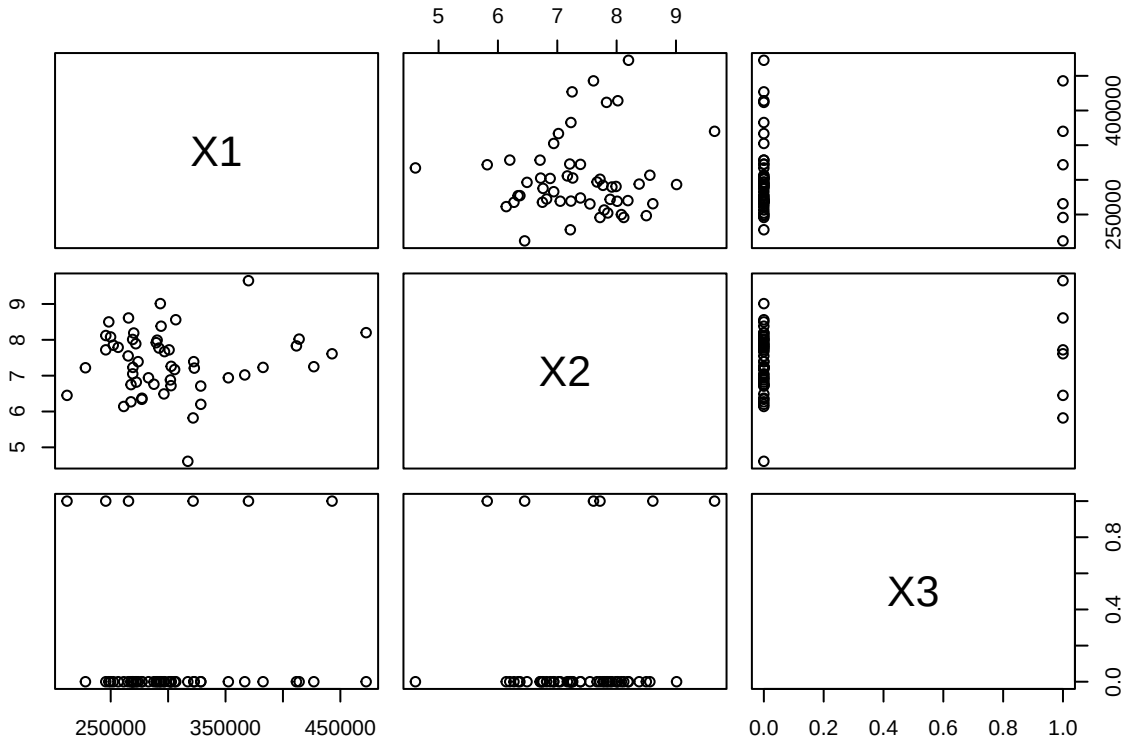
```
plot(cooks.distance(model), type="h", main="Index plot of Cook's distance")
```



According to the plot, none of the cases has cook's distance larger than 0.1 so they are all considered non-influential.

e)

```
# Scatter plot matrix
pairs(data[, c("X1", "X2", "X3")])
```




```
# Correlation matrix
cor(data[, c("X1", "X2", "X3")])
```

```
##            X1         X2         X3
## X1 1.00000000 0.08489639 0.04565698
## X2 0.08489639 1.00000000 0.11337076
## X3 0.04565698 0.11337076 1.00000000
```

The scatter plots and correlation coefficients indicate that there are weak linear relationships between the variables. The binary nature of X3 results in distinct groupings rather than a traditional scatter pattern, which could influence the low correlation coefficients with X1 and X2.

f)

```
library(car)
```

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

```
vif(model)
```

```
##       X1       X2       X3
## 1.008596 1.019598 1.014364
```