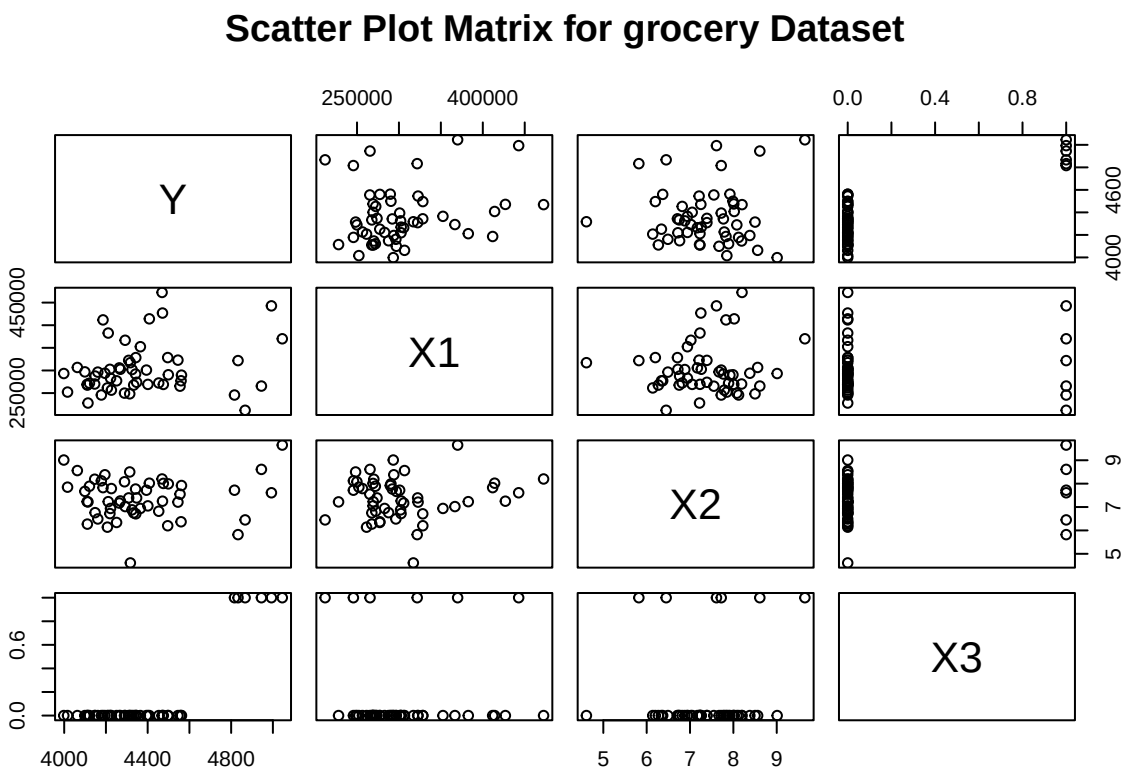
hw4 Q3

Lisong he

2024-03-27

a)

```
grocery_data <- read.csv("/Users/lisonghe/Library/CloudStorage/OneDrive-JohnsHopkins/Semester 2/613 Appl
pairs(grocery_data, main = "Scatter Plot Matrix for grocery Dataset")
```




```
correlation_matrix <- cor(grocery_data)
print(correlation_matrix)
```

```
##            Y         X1         X2         X3
## Y  1.0000000 0.20766494 0.06002960 0.81057940
## X1 0.2076649 1.00000000 0.08489639 0.04565698
## X2 0.0600296 0.08489639 1.00000000 0.11337076
## X3 0.8105794 0.04565698 0.11337076 1.00000000
```

Based on the scatter plot matrix, there seems to be a weakly positive linear relationship between Y and X1 and a even weaker positive linear relationship between Y and X2. The plot of Y and X3 indicates that X3 is a categorical variable. Also, X1 and X2 and X3 are weakly linearly related as all proved in the correlation matrix.

b)

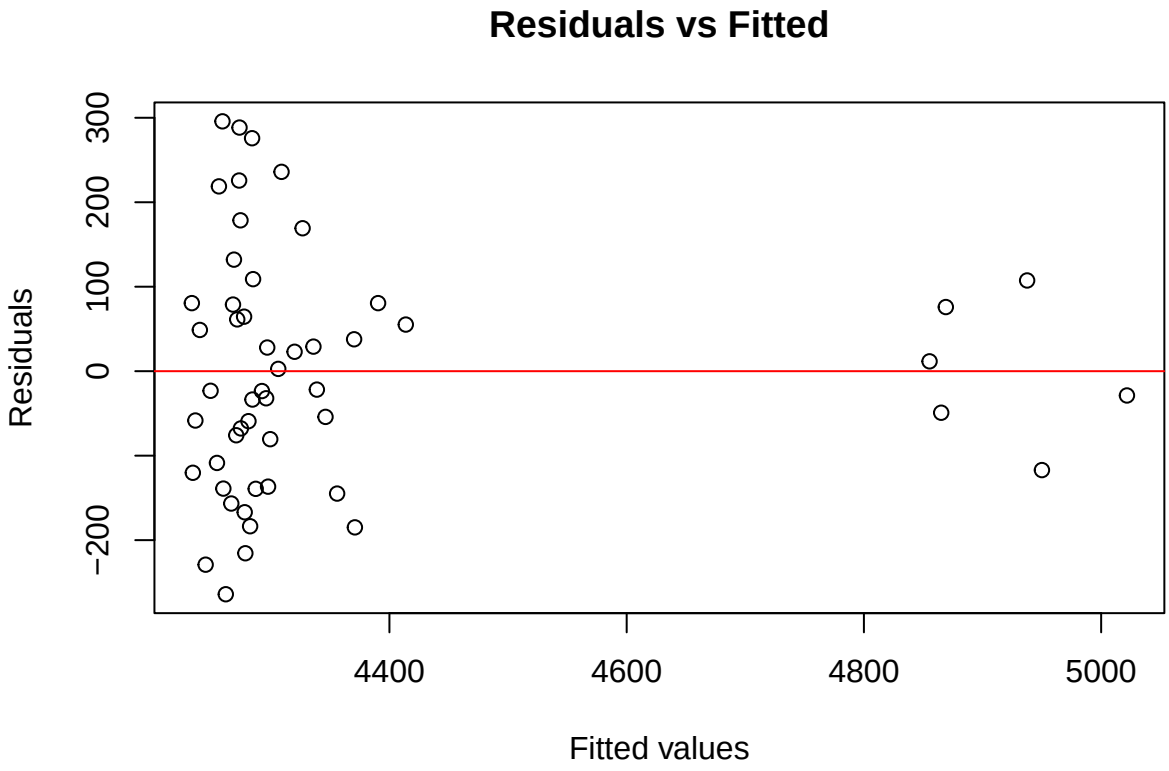
```
model <- lm(Y ~ X1 + X2 + X3, data = grocery_data)
summary(model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + X3, data = grocery_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -264.05 -110.73  -22.52   79.29  295.75
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.150e+03  1.956e+02  21.220  < 2e-16 ***
## X1           7.871e-04  3.646e-04   2.159   0.0359 *
## X2          -1.317e+01  2.309e+01  -0.570   0.5712
## X3           6.236e+02  6.264e+01   9.954 2.94e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 143.3 on 48 degrees of freedom
## Multiple R-squared:  0.6883, Adjusted R-squared:  0.6689
## F-statistic: 35.34 on 3 and 48 DF,  p-value: 3.316e-12
```

Based on the regression model, the estimated regression function is Y = 0.0007871 * X1 -13.17e * X2 + 623.6 * X3 + 4150 Here, b1 and b3 are statistically significant whereas b2 is not.

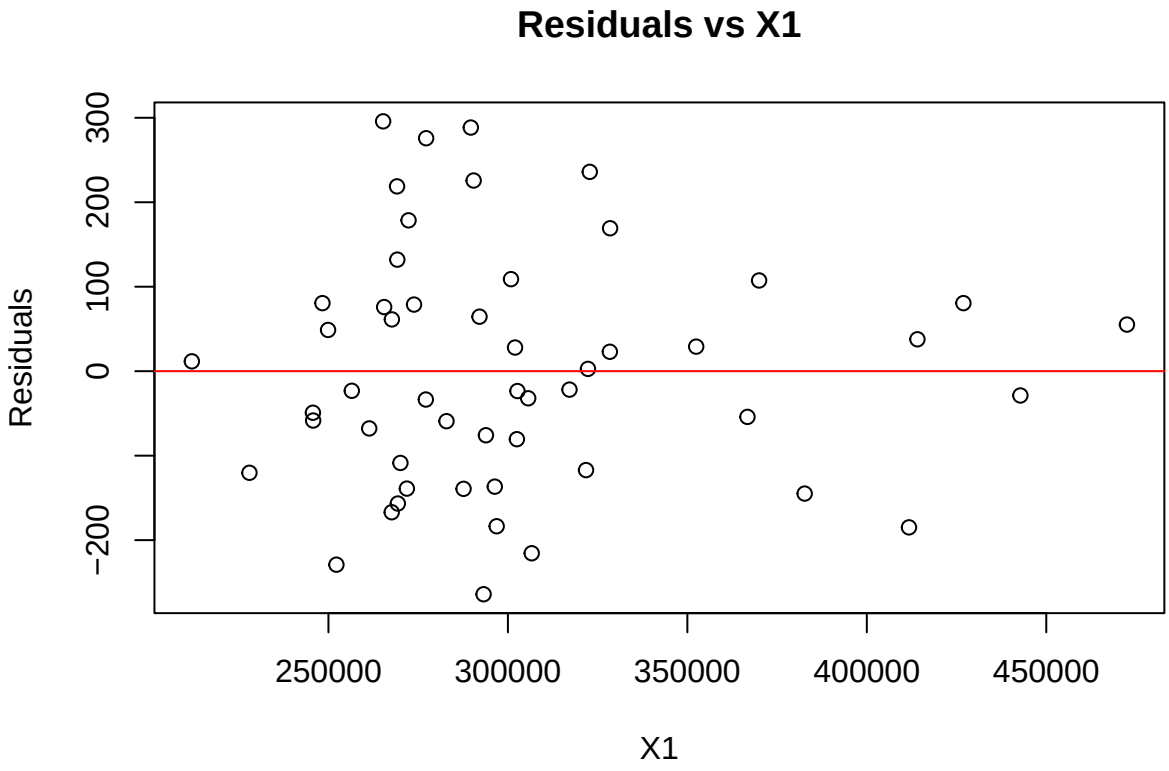
c)

```
# Residuals vs Fitted
plot(model$fitted.values, resid(model), xlab="Fitted values", ylab="Residuals", main="Residuals vs Fitt
abline(h=0, col="red")
```

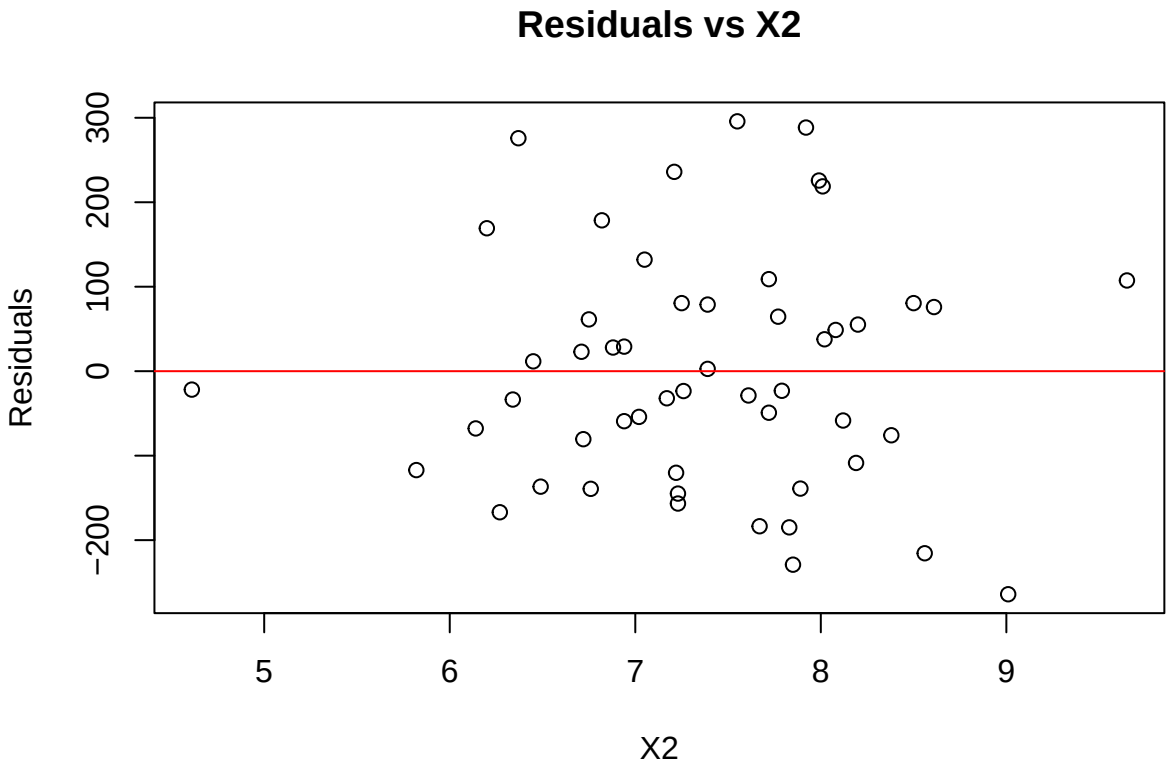


This effectively checks homoscedasticity. Here, there is no pattern detected which proves constant variance.

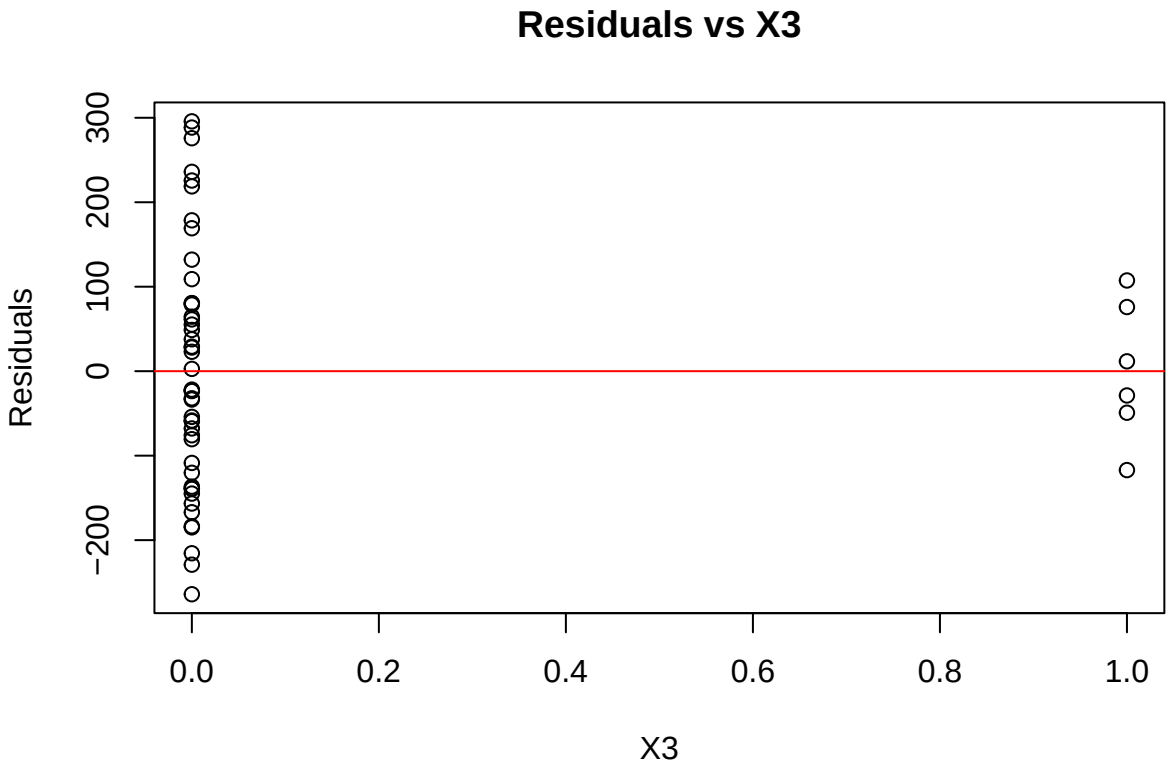
```
# Residuals vs X1
plot(grocery_data$X1, resid(model), xlab="X1", ylab="Residuals", main="Residuals vs X1")
abline(h=0, col="red")
```




```r
# Residuals vs X2
plot(grocery_data$X2, resid(model), xlab="X2", ylab="Residuals", main="Residuals vs X2")
abline(h=0, col="red")
```

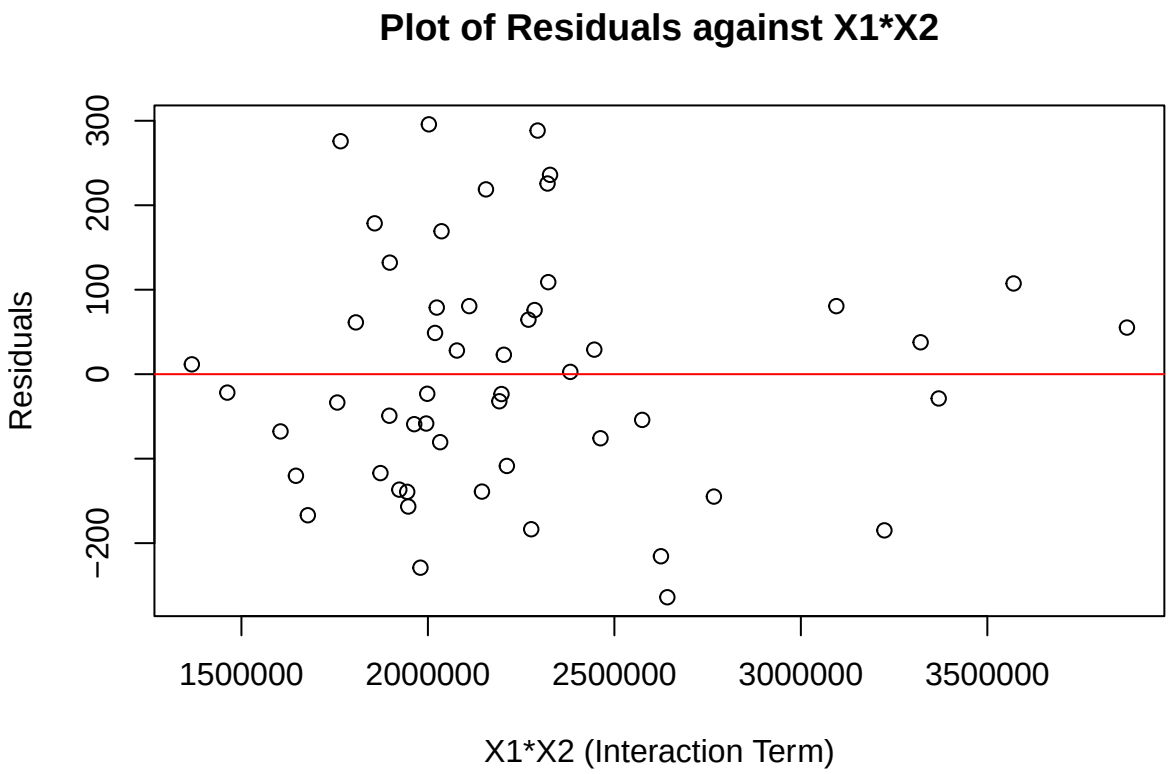



```
# Residuals vs X3
plot(grocery_data$X3, resid(model), xlab="X3", ylab="Residuals", main="Residuals vs X3")
abline(h=0, col="red")
```



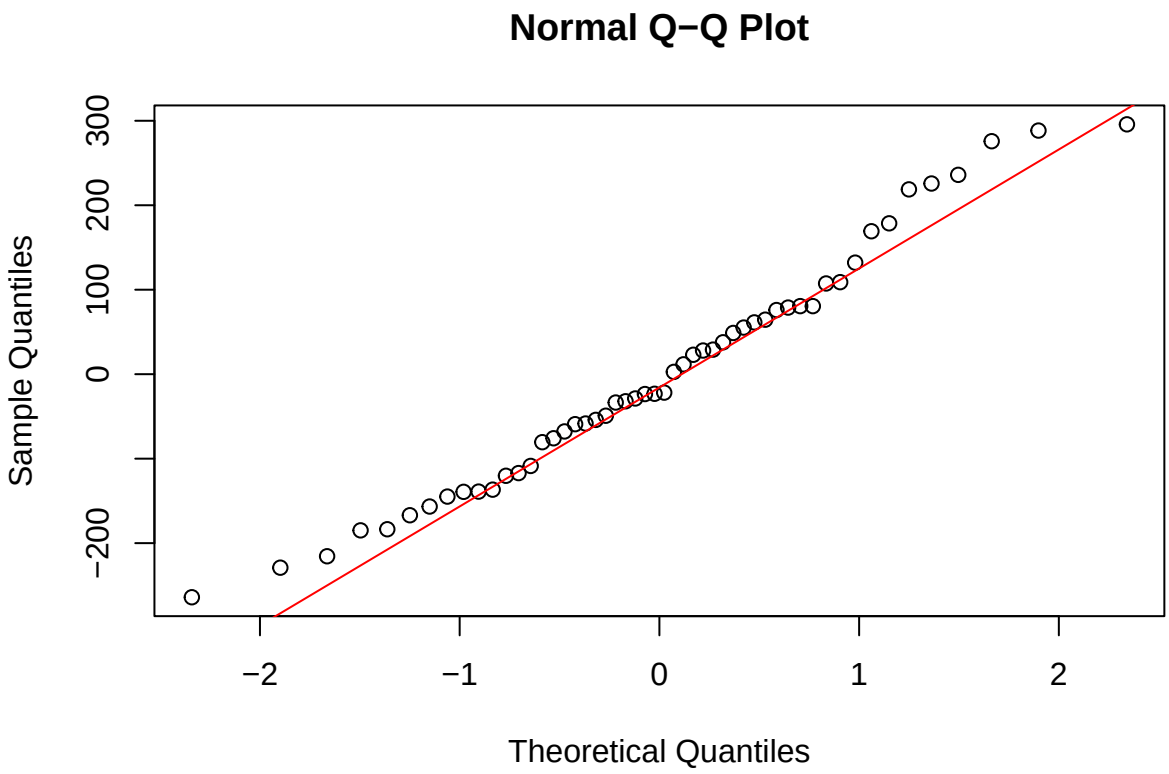
Here, the three plots of residuals against X1, X2, and X3 check the necessity of transformation of predictors. Here, all three plots show no patterns which means no transformation is needed.

```
grocery_data$X1X2 <- grocery_data$X1 * grocery_data$X2
residuals <- resid(model)
if(length(residuals) == nrow(grocery_data)) {
  plot(grocery_data$X1X2, residuals,
       xlab = "X1*X2 (Interaction Term)",
       ylab = "Residuals",
       main = "Plot of Residuals against X1*X2")
  abline(h = 0, col = "red")
} else {
  stop("The lengths of 'residuals' and 'grocery_data$X1X2' do not match.")
}
```



Plotting residuals against interaction terms checks the presence of the interaction term

```
# Normal Q-Q plot
qqnorm(resid(model), main="Normal Q-Q Plot")
qqline(resid(model), col="red")
```



This testifies the normality of residuals. At two ends of the distribution, the dots deviate from the normal line which stands for longer tails than a normal curve. The residual is roughly normal.

d)

```
model_2 <- lm(Y ~ X1 + X2, data=grocery_data)
summary(model_2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2, data = grocery_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -376.21 -173.77  -49.36  123.73  601.11
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.995e+03  3.378e+02  11.829 5.72e-16 ***
## X1          9.192e-04  6.312e-04   1.456    0.152
## X2          1.212e+01  3.977e+01   0.305    0.762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 248.3 on 49 degrees of freedom
## Multiple R-squared:  0.04494,    Adjusted R-squared:  0.005953
```

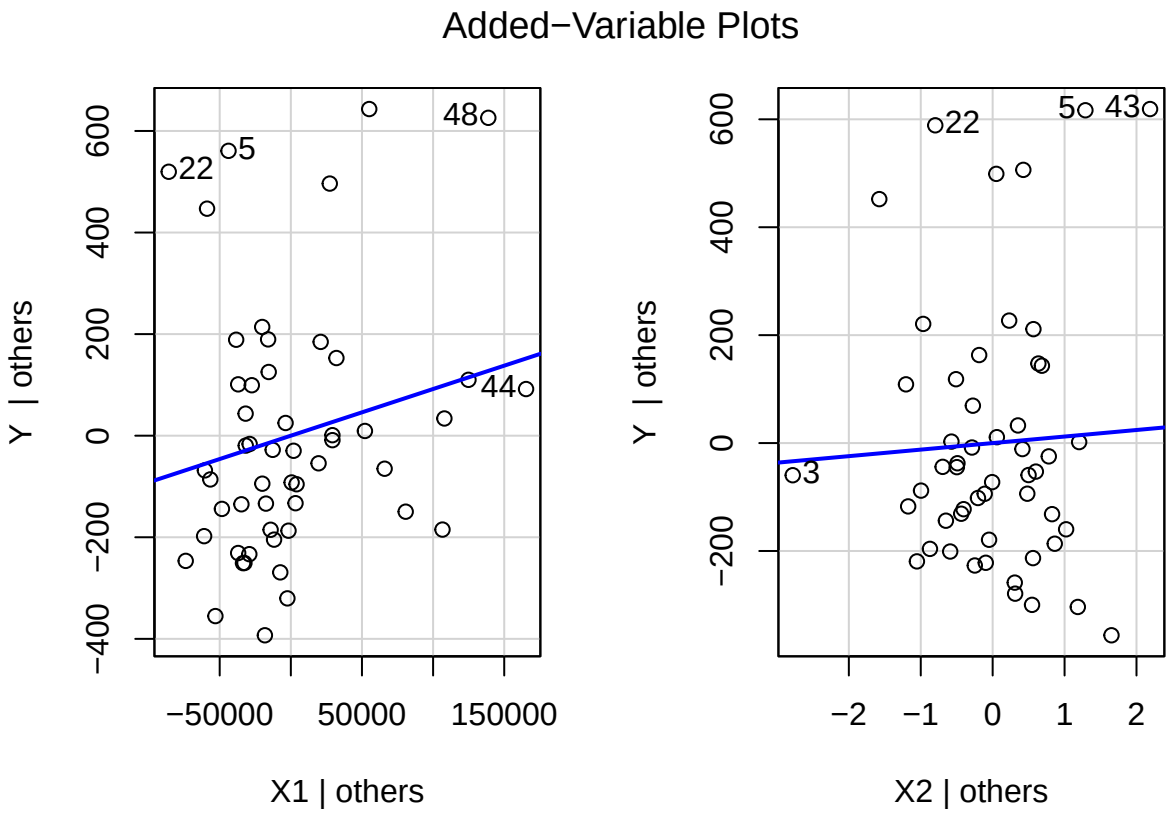
```
## F-statistic: 1.153 on 2 and 49 DF,  p-value: 0.3242
```

e)

```
library(car)
```

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

```
avPlots(model_2)
```



Added-Variable plots plot residuals of Y after regressing on all predictors except Xi against that of Xi. It is essentially the plot of part of Y unexplained by other predictors against part of Xi unexplained by other predictors. Both plots show non-zero slopes, indicating that X1 explains Y.

f)

```
model_Y_on_X1 <- lm(Y ~ X1, data=grocery_data)
model_X2_on_X1 <- lm(X2 ~ X1, data=grocery_data)
model_resids <- lm(resid(model_Y_on_X1) ~ resid(model_X2_on_X1))
summary(model_resids)
```

```
##
## Call:
## lm(formula = resid(model_Y_on_X1) ~ resid(model_X2_on_X1))
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -376.21 -173.77  -49.36  123.73  601.11
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)           6.525e-15  3.408e+01   0.000    1.000
## resid(model_X2_on_X1) 1.212e+01  3.937e+01   0.308    0.759
##
## Residual standard error: 245.8 on 50 degrees of freedom
## Multiple R-squared:  0.001892,   Adjusted R-squared:  -0.01807
## F-statistic: 0.0948 on 1 and 50 DF,  p-value: 0.7594
```