# Exercise-1 R Markdown

Santanu Mukherjee

06/12/2022

# R Markdown

## Exercise 1

## Q1

Consider a well-known dataset on per capita income and per capita spending in public schools by state in the United States in 1979. (Available on blackboard). This dataset has been widely analyzed in various statistical. As in those previous analyses, we take per capita spending (Expenditure) as the dependent variable and per capita income as the predictor variable.

```
df.inc.exp <- read.table("https://github.com/santum4/dataexam/raw/main/STA-6543-Summer-2022/Exercise%201/Income.txt", header
= T)


str(df.inc.exp)
```

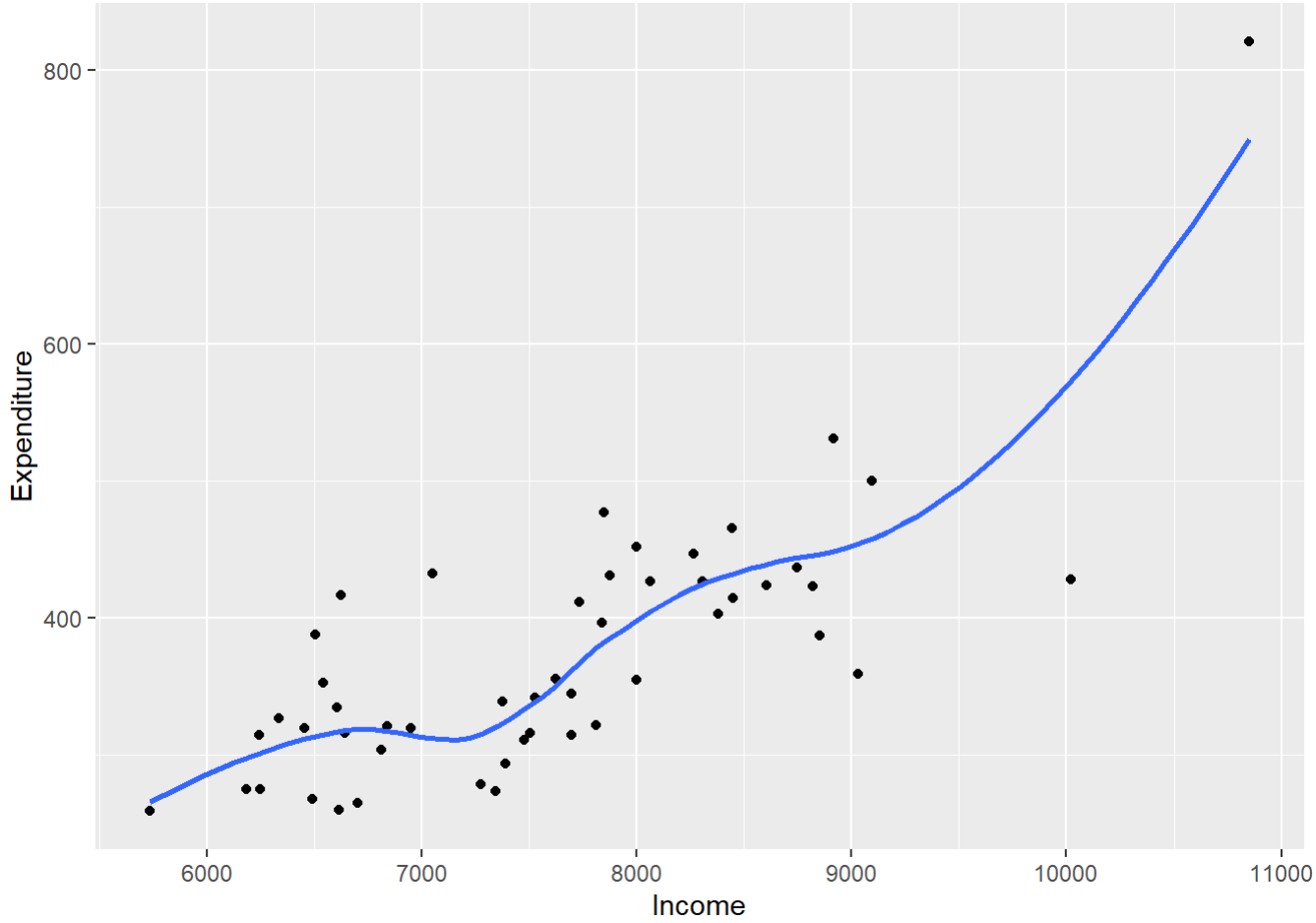
```
## 'data.frame':    50 obs. of  3 variables:
##  $ State      : chr  "AL" "AR" "CT" "FL" ...
##  $ Expenditure: int  275 275 531 316 304 431 316 427 259 294 ...
##  $ Income     : int  6247 6183 8914 7505 6813 7873 6640 8063 5736 7391 ...
```

### Q 1 a

a. Draw a scatter-plot to check the relationship between Income and Expenditure and interpret the relationship between Income and Expenditure.

## Answer 1 (a)

```
# Scatter Plot

df.work = df.inc.exp[,-1]
df.work = df.work[,c(2,1)]

ggplot(df.work, aes(x=Income, y=Expenditure))+geom_point()+geom_smooth(se=FALSE)
```




From the above scatterplot, we can say that the relationship between Income and Expenditure is non-linear. Also it can be said that we see an upward rising trend in expenditure with rising income.

# Q 1 b

b. Find and interpret the slope for the least squares regression line

# Answer 1 (b)

```
# Least square regression

set.seed(1)
lm.fit = lm(Expenditure~Income, data=df.work)
summary(lm.fit)
```

```
##
## Call:
## lm(formula = Expenditure ~ Income, data = df.work)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -112.390  -42.146   -6.162   30.630  224.210
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -151.26509   64.12183  -2.359   0.0224 *
## Income         0.06894    0.00835   8.256 9.05e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 61.41 on 48 degrees of freedom
## Multiple R-squared:  0.5868, Adjusted R-squared:  0.5782
## F-statistic: 68.16 on 1 and 48 DF,  p-value: 9.055e-11
```

The slope of the least square is **positive** $(0.06894)$ which means that on an average, for every unit increase in Income, the least square regression model predicts an increase of 0.06894 in Expenditure.

## Q 1 c

c. Find and interpret y-intercept for the least squares regression line

# Answer 1 (c)

The y-intercept is **negative** $(-151.26509)$. This means that when x(Income) is zero, then y(Expenditure) will be -151.26509. In reality, this cannot happen and so this tells me that the linear model is not the best fit for this dataset.

## Q 1 d

d. Find the least square regression equation and circle the results from your outputs.

# Answer 1 (d)

$Expenditure = -151.26509 + (0.06894)Income$

## Q 1 e

e. Find proportion of the variation that can be explained by the least squares regression line (i.e., $R^2$).

# Answer 1 (e)

So, we know that Total Sum of Squares (TSS) = Model Sum of Squares(MSS) + Residuals (RSS)

$R^2$ = $\frac{(TSS-RSS)}{TSS}$ = $1 - \frac{RSS}{TSS}$ = $\frac{MSS}{TSS}$ = $\frac{ExplainedVariationbyModel}{TotalVariation}$ = Coefficient of Variation

Range of $R^2$ is between $0$ and $1$

Here the value of $R^2$ is $0.5868$. This means that 58% of the variability observed in the target variable is explained by the regression model.

## Q 1 f

f. Find the estimator of $\sigma^2$ (i.e., $s^2$) and interpret the value of this estimator.

# Answer 1 (f)

We use $\hat{\sigma}^2$ = $s^2$ = $MSE$ = $\frac{RSS}{n-2}$ as an estimate of $\sigma^2$ . Here $n$ = $50$

Also if we take the square of **Residual Standard Error** (which in this case is $61.41$) , we would get $\hat{\sigma}^2$.

So the estimator of $\sigma^2$ is $61.41^2$ = $3771.1881$

Along with $R^2$, the **Residual standard Error** metric is often used to measure the goodness of fit , meaning it measures how well a regression model fits a dataset. The smaller the value of the residual standard error, the better the regression model fits the dataset. Here the **Residual standard Error** is $significantly\ LARGE$, which means that the Linear model is $NOT$ a good fit.

## Q 1 g

g. Check if the data contain any outlier or influential points?

# Answer 1 (g)

```
#Studentized deleted residuals to detect outliers, Measures detecting outlying Y

p.rstud = 1 - pt(abs(rstudent(lm.fit)), length(df.work$Expenditure) - 3)
p.rstud
```

```
##            1            2            3            4            5            6
## 4.711220e-01 4.998919e-01 1.308291e-01 2.076857e-01 4.071561e-01 2.605867e-01
##            7            8            9           10           11           12
## 4.382537e-01 3.579934e-01 4.018406e-01 1.475348e-01 2.900632e-01 3.071557e-01
##           13           14           15           16           17           18
## 4.487150e-01 3.392969e-01 3.225365e-01 1.903494e-01 3.336049e-01 3.325775e-06
##           19           20           21           22           23           24
## 1.171221e-01 3.849738e-01 2.276037e-01 4.053697e-01 2.305935e-01 2.460082e-01
##           25           26           27           28           29           30
## 2.822809e-01 9.294777e-02 2.928843e-02 6.605810e-02 1.932627e-01 4.493779e-01
##           31           32           33           34           35           36
## 2.763561e-01 1.472639e-01 3.832987e-01 3.846943e-01 2.001048e-01 2.526521e-02
##           37           38           39           40           41           42
## 3.512321e-01 2.890659e-01 2.315503e-01 4.633266e-01 7.637975e-02 5.296621e-02
##           43           44           45           46           47           48
## 1.218584e-01 3.222584e-01 1.436914e-01 3.124083e-01 4.957477e-01 3.141842e-02
##           49           50
## 3.954018e-01 3.443695e-01
```

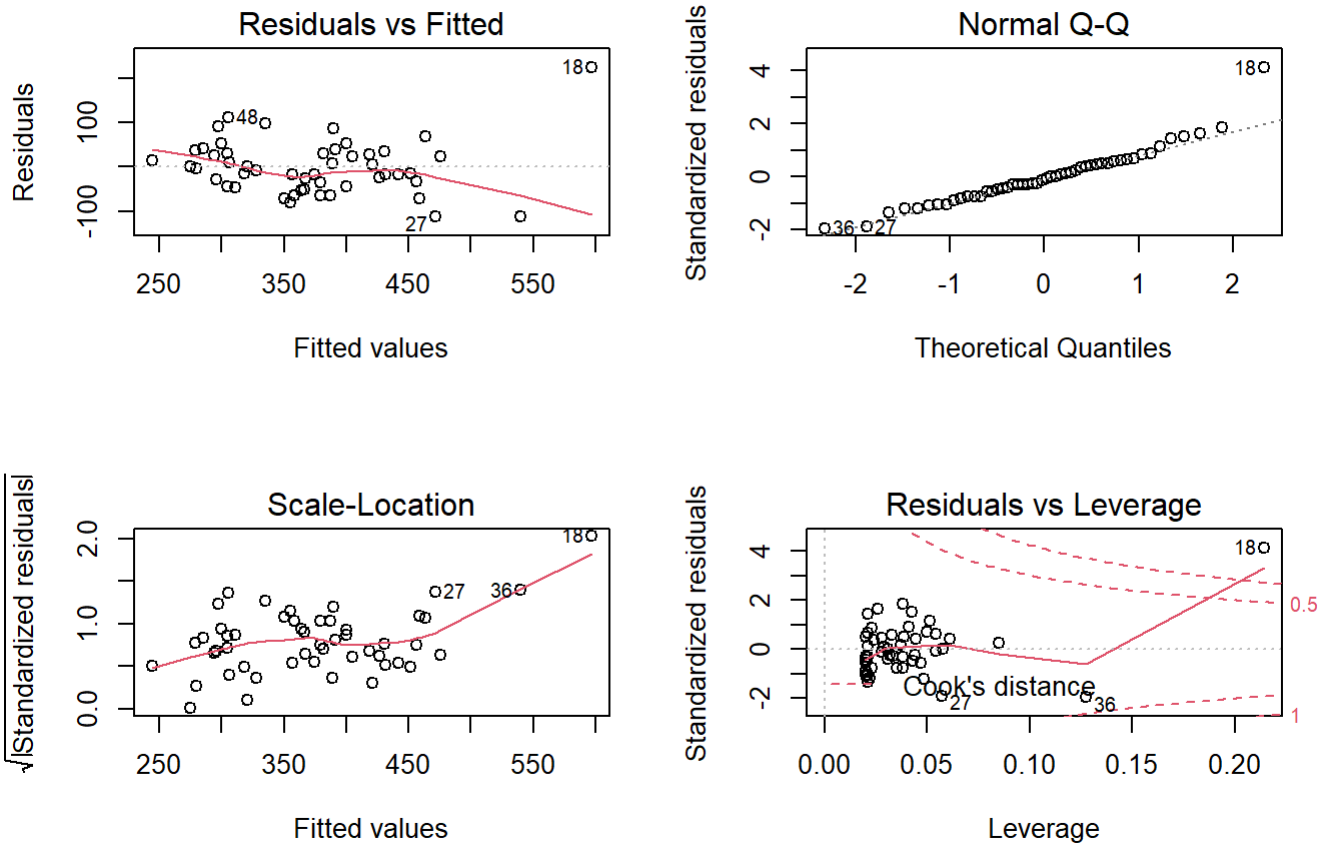
```
p.rstud[p.rstud<0.05]
```

```
##           18           27           36           48
## 3.325775e-06 2.928843e-02 2.526521e-02 3.141842e-02
```

```
p.rstud[p.rstud<0.025]
```

```
##           18
## 3.325775e-06
```

```
par(mfrow = c(2,2))
plot(lm.fit)
```




```
#Identifying outlying X observations - HAT matrix leverage values

hat.lm = hatvalues(lm.fit); hat.lm[hat.lm>2*2/length(df.work$Expenditure)]
```

```
##         9        18        36
## 0.08482819 0.21437317 0.12768753
```

The outliers are identified through using Studentized deleted residuals (y values) and HAT matrix leverage values (x values).

```
# Cooks Distance to identify influential points

cooksD <- cooks.distance(lm.fit)
influential <- cooksD[(cooksD > (3 * mean(cooksD, na.rm = TRUE)))]
influential
```

```
##        18        36
## 2.3149630 0.2772922
```

The influential points are identified by utilizing the **cooks.distance** function on the model and then filtering out any values greater than $3x$ the mean.

## Q 1 h

h. Fit a single linear model and conduct 10-fold CV to estimate the error. In addition, draw the scatter plot with the fitted line and the scatter plot between the observed and fitted values below.

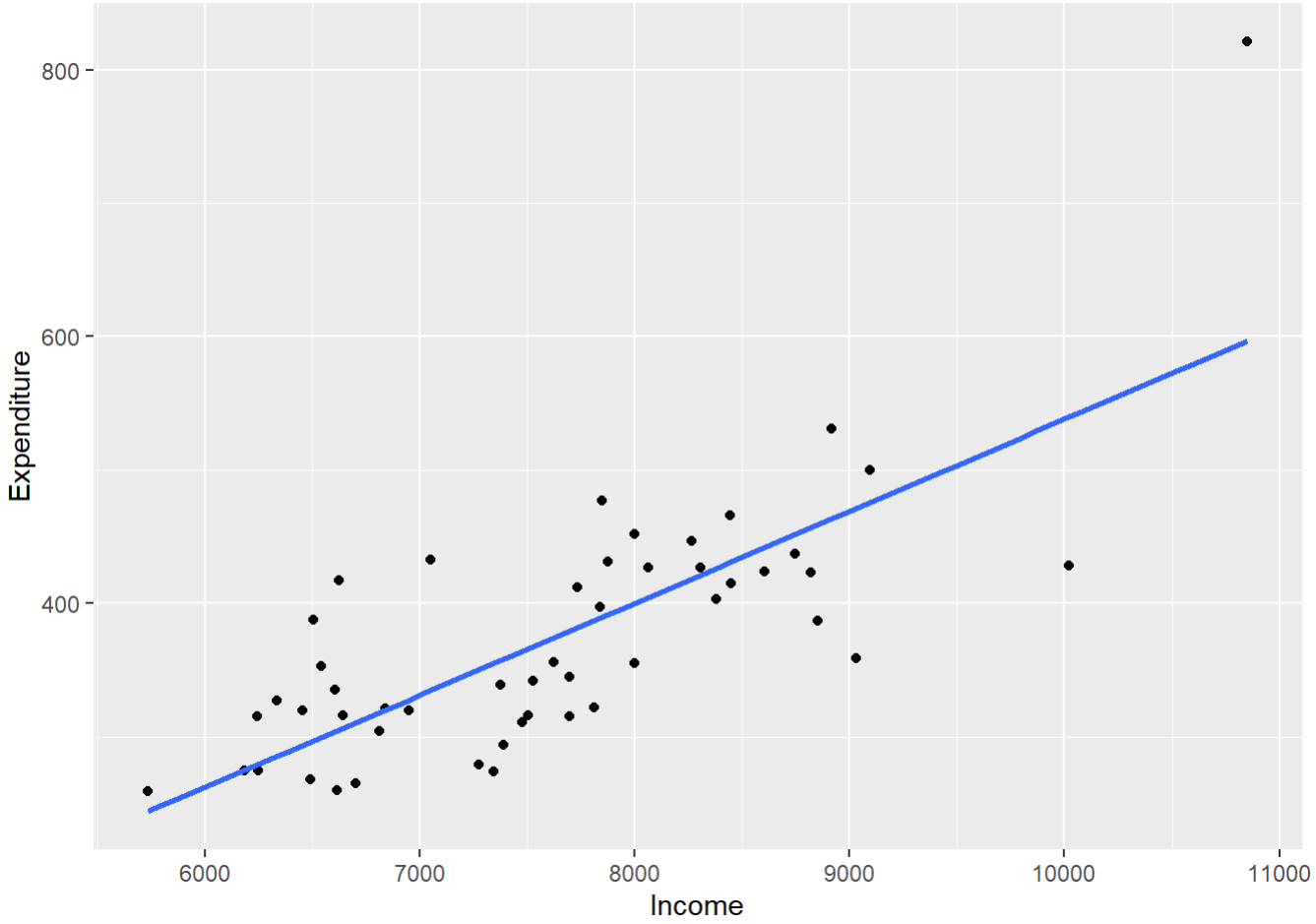
# Answer 1 (h)

```
# Single Linear regression with 10 fold CV

# define the number of folds , here CV = 10
train_control = trainControl(method = "CV", number = 10)

# Run linear regression with Cross Validation 10 fold (CV = 10)
set.seed(1)
lm.cv.fit1 = train(Expenditure~Income, data = df.inc.exp, method = 'lm', trControl = train_control )
print(lm.cv.fit1)
```

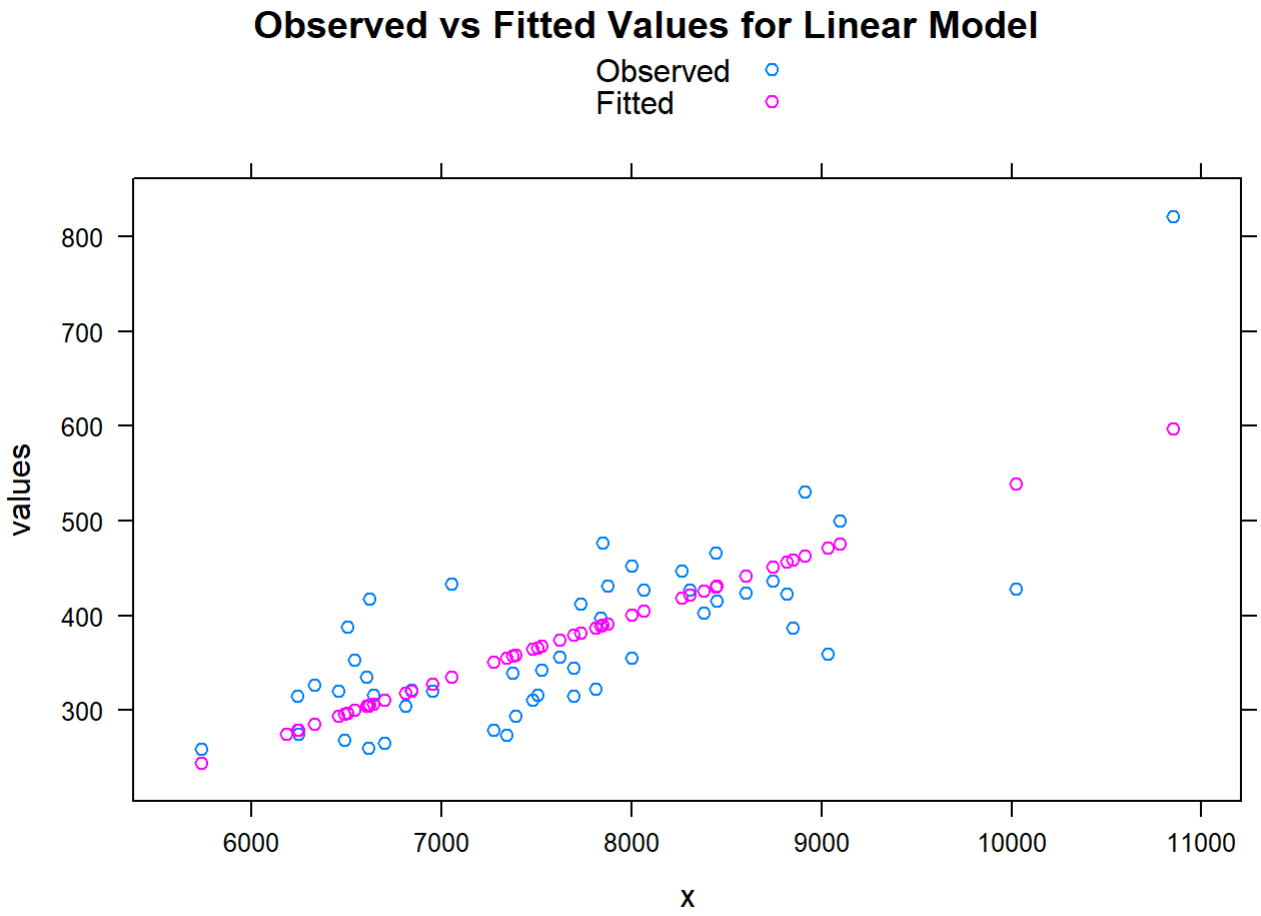
```
## Linear Regression
##
## 50 samples
##  1 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 45, 45, 45, 43, 46, 46, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   58.40613  0.5929131  47.86736
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Based on the results, for **Linear Regression**, the error ($RMSE$) is $58.40613$ and $R^2$ is $0.5929131$.

```
# Scatter plot with the fitted line and the residual vs fitted values graph

ggplot(df.inc.exp, aes(x=Income, y=Expenditure)) + geom_point() + geom_smooth(method=lm, se=FALSE)
```




```
res <- stack(data.frame(Observed = df.inc.exp$Expenditure, Fitted = fitted(lm.fit)))
res <- cbind(res, x = rep(df.inc.exp$Income, 2))

require("lattice")

xyplot(values ~ x, data = res, group = ind, auto.key = TRUE, main = "Observed vs Fitted Values for Linear Model")
```




## Q 1 i

i. Fit a quadratic model and conduct 10-fold CV to estimate the error and draw the scatter plot with the fitted line and the scatter plot between the observed and fitted values.

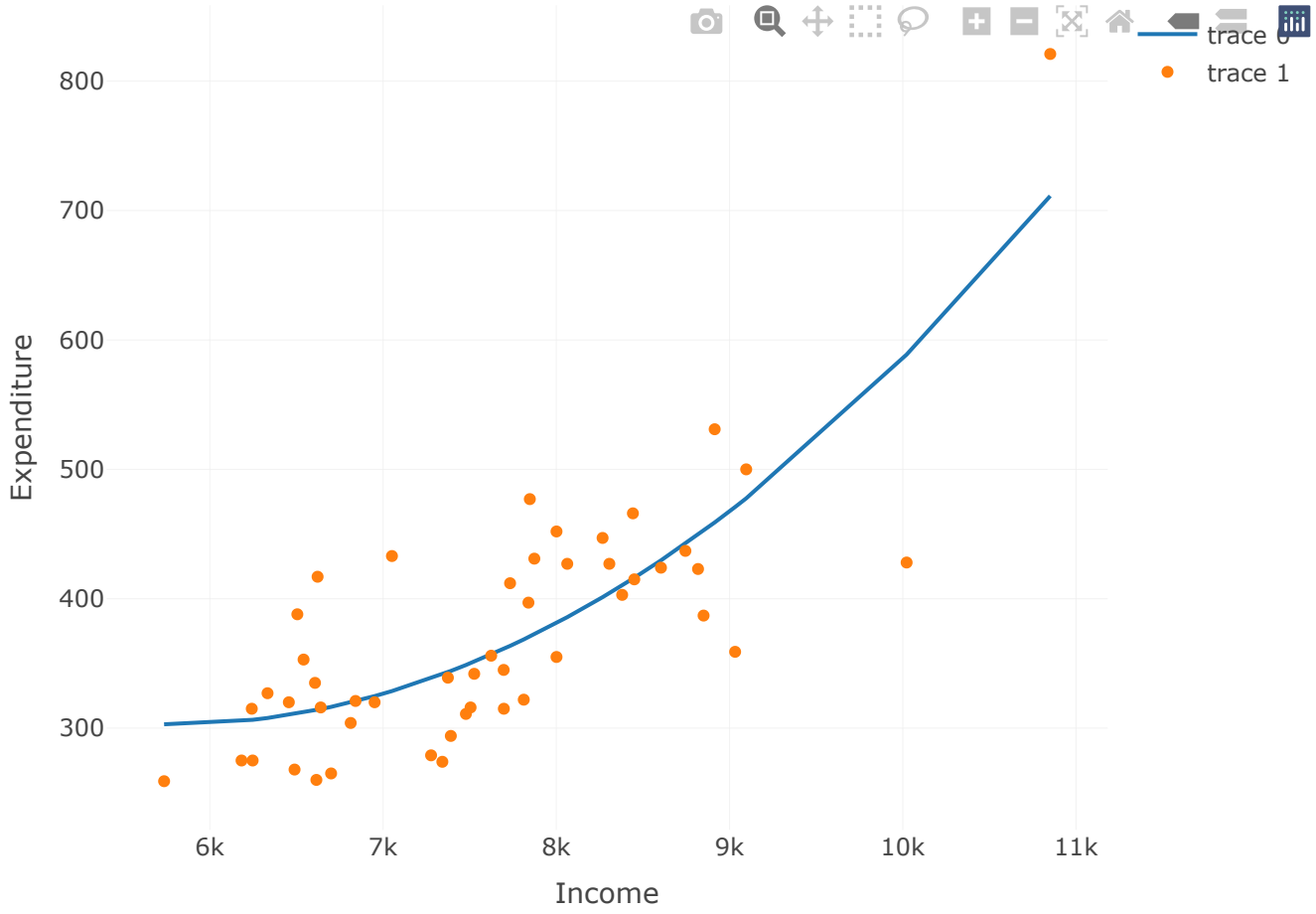
# Answer 1 (i)

```
# Quadratic regression with 10 fold CV

# define the number of folds , here CV = 10
train_control = trainControl(method = "CV", number = 10)

# Run linear regression with Cross Validation 10 fold (CV = 10)
set.seed(1)
df.inc.exp$Income2 = df.inc.exp$Income^2
lm.qd.cv.fit2 = train(Expenditure~Income + Income2, data = df.inc.exp, method = 'lm', trControl = train_control )
print(lm.qd.cv.fit2)
```

```
## Linear Regression
##
## 50 samples
##  2 predictor
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 45, 45, 45, 43, 46, 46, ...
## Resampling results:
##
##   RMSE      Rsquared   MAE
##   63.52598  0.5506899  49.27421
##
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

Based on the results, for **Quadratic Model**, the error ($RMSE$) is $63.52598$ and $R^2$ is $0.5506899$.

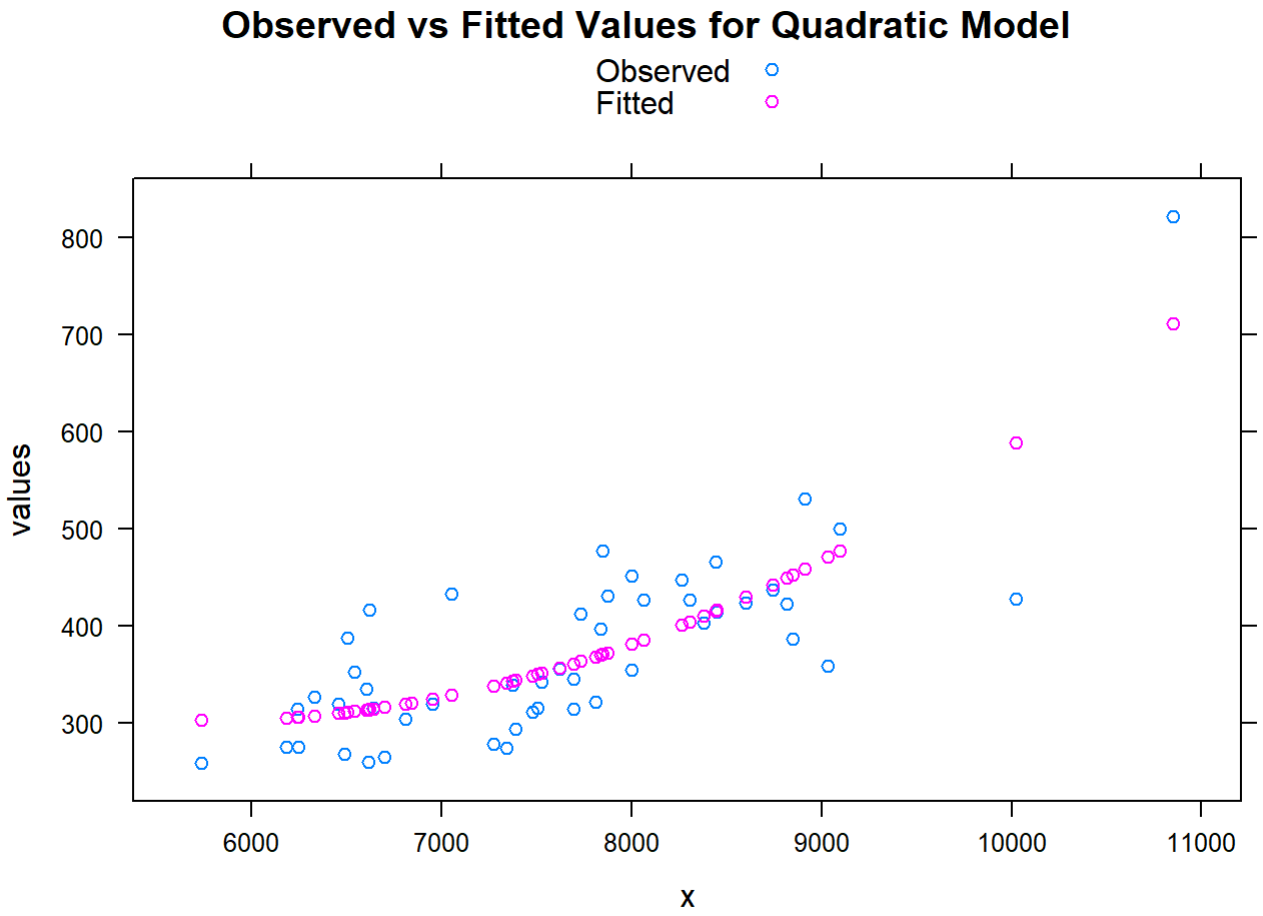
```
# Scatter plot with the fitted line and the residual vs fitted values graph

lm.qd.fit2 = lm(Expenditure~ poly(Income,2) + Income2, data = df.inc.exp)

df.inc.exp %>%
plot_ly() %>%
add_lines(x = ~Income, y = fitted(lm.qd.fit2)) %>%
add_trace(x=~Income, y=~Expenditure)
```




```
res <- stack(data.frame(Observed = df.inc.exp$Expenditure, Fitted = fitted(lm.qd.fit2)))
res <- cbind(res, x = rep(df.inc.exp$Income, 2))
head(res)
```

```
##   values      ind    x
## 1    275 Observed 6247
## 2    275 Observed 6183
## 3    531 Observed 8914
## 4    316 Observed 7505
## 5    304 Observed 6813
## 6    431 Observed 7873
```

```
require("lattice")

xyplot(values ~ x, data = res, group = ind, auto.key = TRUE, main = "Observed vs Fitted Values for Quadratic Model")
```




## Q 1 j

j. Fit a mars model with optimal tuning parameters that you choose and conduct 10-fold CV to estimate the error and draw the scatter plot with the fitted line and the scatter plot between the observed and fitted values.

# Answer 1 (j)

```
# Mars Model fitment

library(earth)
library(caret)

#create a tuning grid
hyper_grid <- expand.grid(degree = 1:3,
                          nprune = seq(2, 50, length.out = 10) %>%
                          floor())

set.seed(1)

#fit MARS model using 10-fold cross-validation
model.cv_mars <- train(
  x = subset(df.work, select = -Expenditure),
  y = df.work$Expenditure,
  method = "earth",
  metric = "RMSE",
  trControl = trainControl(method = "cv", number = 10),
  tuneGrid = hyper_grid)

#display model with lowest test RMSE
print(model.cv_mars$bestTune)
```

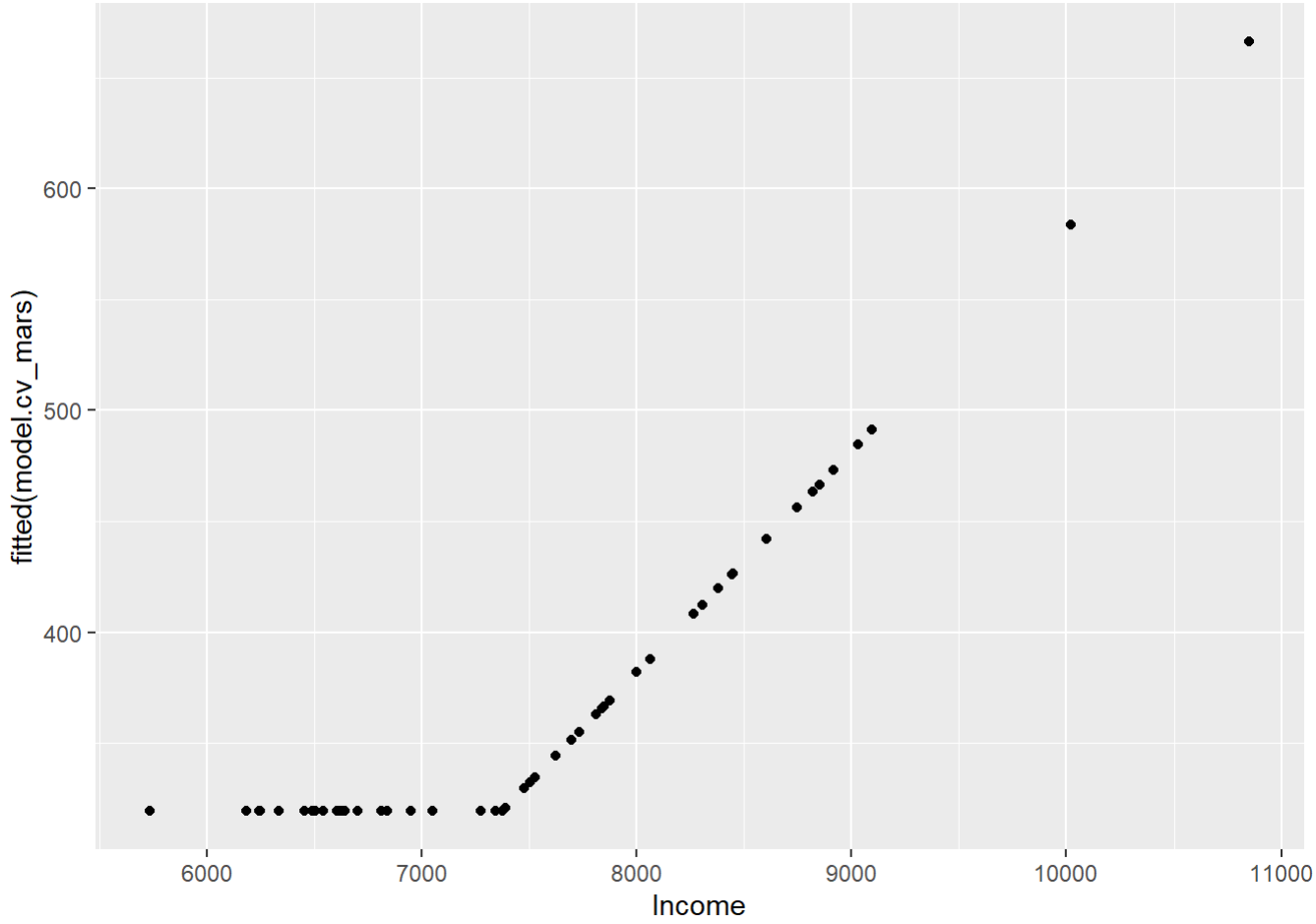
```
##   nprune degree
## 1      2      1
```

```
model.cv_mars$results %>%
  filter(nprune==model.cv_mars$bestTune$nprune, degree == model.cv_mars$bestTune$degree)
```

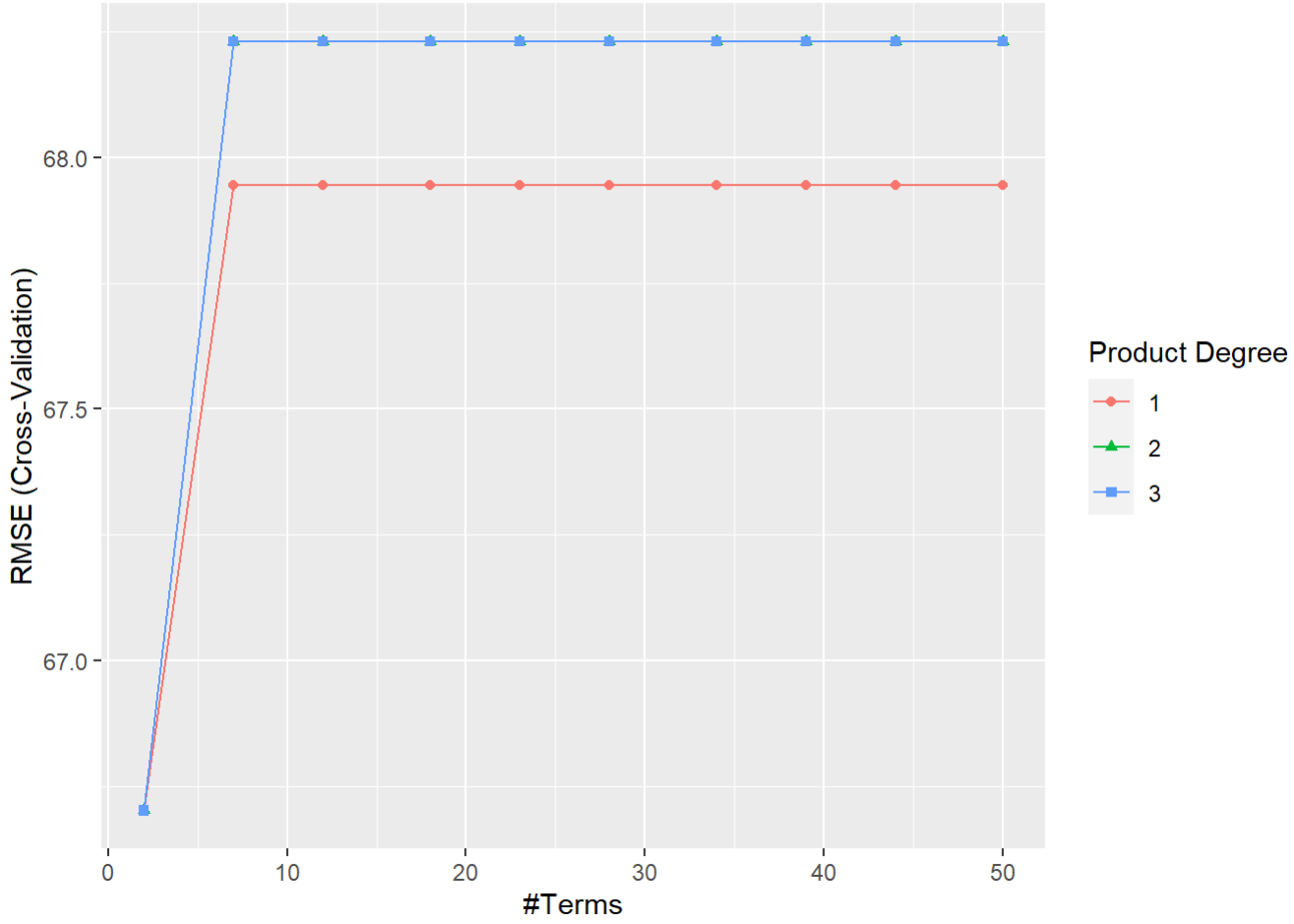
```
##   degree nprune     RMSE  Rsquared     MAE   RMSESD RsquaredSD    MAESD
## 1      1      2 66.70138 0.5563583 50.0171 43.89027  0.2998567 24.94631
```

Based on the **MARS** model results, the lowest error ($RMSE$) is $66.70138$ and $R^2$ is $0.5563583$.

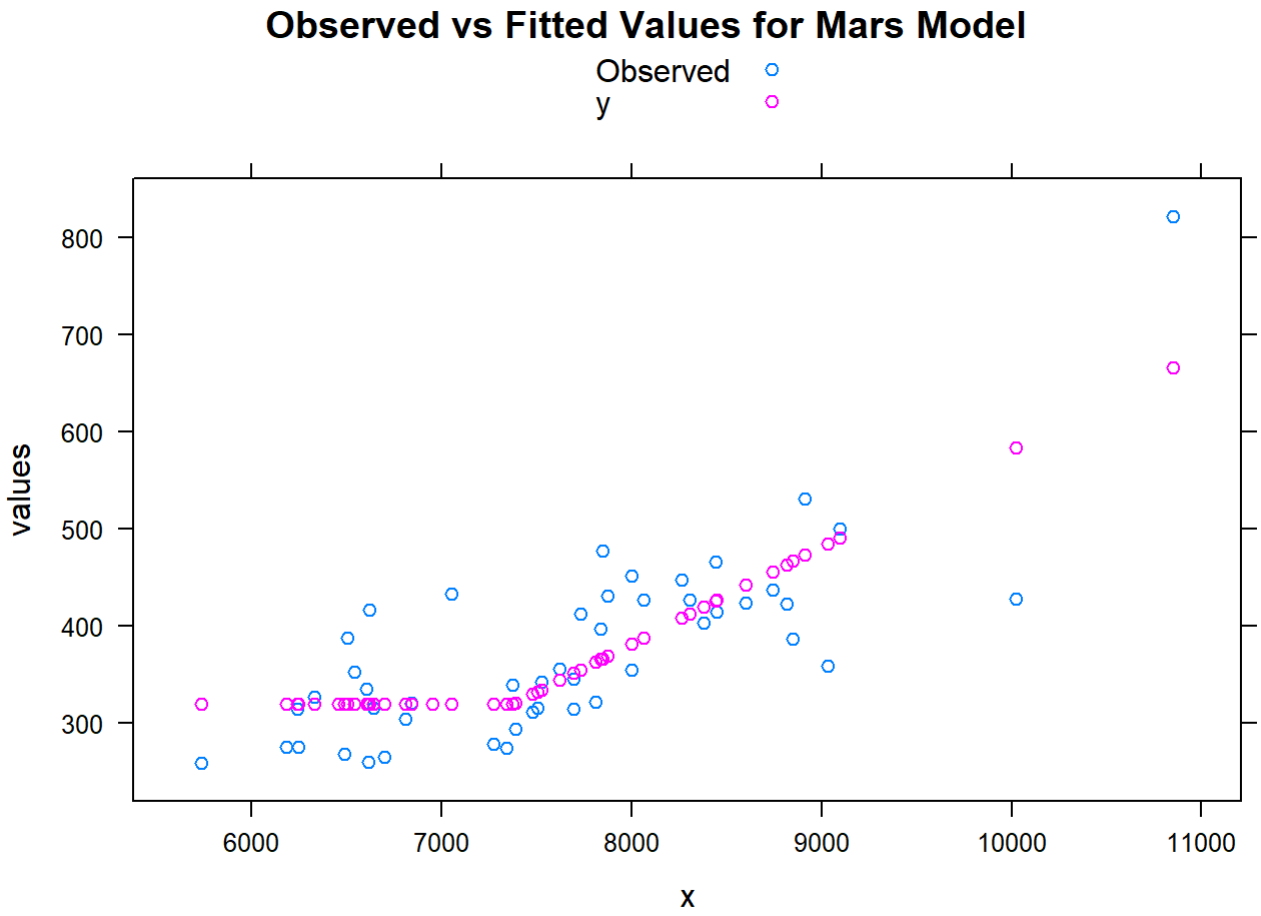
```
# Mars Scatter plot with the fitted line


ggplot(df.work, aes(x=Income, y=fitted(model.cv_mars))) + geom_point() + geom_smooth(method=earth, se=FALSE)
```




```
ggplot(model.cv_mars)
```




```
# Observed values vs fitted values graph

res <- stack(data.frame(Observed = df.inc.exp$Expenditure, Fitted = fitted(model.cv_mars)))
res <- cbind(res, x = rep(df.inc.exp$Income, 2))

require("lattice")

xyplot(values ~ x, data = res, group = ind, auto.key = TRUE, main = "Observed vs Fitted Values for Mars Model")
```




## Q 1 k

k. Compare the three fitted models in terms of $RMSE$ and $R^2$, and then make a recommendation based on your criteria.

# Answer 1 (k)

Based on the data for the 3 fitted models in terms of $RMSE$ and $R^2$, the best model is the **Linear Regression** Model, because it has the **lowest** $RMSE$ and **highest** $R^2$.