COMP2025 Introduction to Data Science

Spring 2023

Assignment 1

By including this statement, I the author of this work, verify that:

• I hereby certify that no part of this assignment/product has been copied from any other student's work or from any other source except where due acknowledgement is made in the assignment.

• No part of this assignment/product has been written/produced for us by another person except where such collaboration has been authorised by the subject lecturer/tutor concerned.

• I am aware that this work may be reproduced and submitted to plagiarism detection software programs for the purpose of detecting possible plagiarism (which may retain a copy on its database for future plagiarism checking).

• I hereby certify that we have read and understand what the School of Computer, Data and Mathematical Sciences defines as minor and substantial breaches of misconduct as outlined in the learning guide for this unit.

Regression

The dataset "Realestate.csv" is a historical dataset of real estate valuation collected from a city in Taiwan. It includes the following variables.

age: the age of the house (in years)

year: transaction year

distanceMRT: the distance to the nearest metro rail transit (in meters)

convstore: the number of convenience stores in the living circle on foot

price: the house price of unit area (USD per square meter)

Question 1

Import the dataset and explore (Identify the types of variables, number of observations and view first few rows).

- We can used read.csv() function to read CSV files:

```
realstate <- read.csv("Realestate.csv") #since our file iris.csv is in same
#folder as the project we are in so we do not need assign path.
```

- Using head() function to see the first six rows of the data:

```
head(realstate)
```

```
##    age year distanceMRT convstore   price
## 1 32.0 2012       84.88        10 3675.15
## 2 19.5 2012      306.59         9 4092.12
## 3 13.3 2013      561.98         5 4586.67
## 4 13.3 2013      561.98         5 5313.94
## 5  5.0 2012      390.57         5 4179.39
## 6  7.1 2012     2175.03         3 3112.73
```

- attach() function attaches the dataset to the R search path, so variables in the dataset can be accessed simply by giving their names.

```
attach(realstate)
```

- dim() function gives the dimension or say number of rows and columns in our dataset.

```
dim(realstate)
```

```
## [1] 413   5
```

- names() function gives the name of variables/columns in our dataset

```
names(realstate)
```

```
## [1] "age"         "year"        "distanceMRT" "convstore"   "price"
```

- sapply() function gives the types of each variable in our dataset.

```
sapply(realstate, class)
```

```
##         age        year distanceMRT   convstore       price
##   "numeric"   "integer"   "numeric"   "integer"   "numeric"
```

- str() function gives information about the data frame's structure including other details as well.

```
str(realstate)
```

```
## 'data.frame':    413 obs. of  5 variables:
##  $ age        : num  32 19.5 13.3 13.3 5 7.1 34.5 20.3 31.7 17.9 ...
##  $ year       : int  2012 2012 2013 2013 2012 2012 2012 2013 2013 2013 ...
##  $ distanceMRT: num  84.9 306.6 562 562 390.6 ...
##  $ convstore  : int  10 9 5 5 5 3 7 6 1 3 ...
##  $ price      : num  3675 4092 4587 5314 4179 ...
```

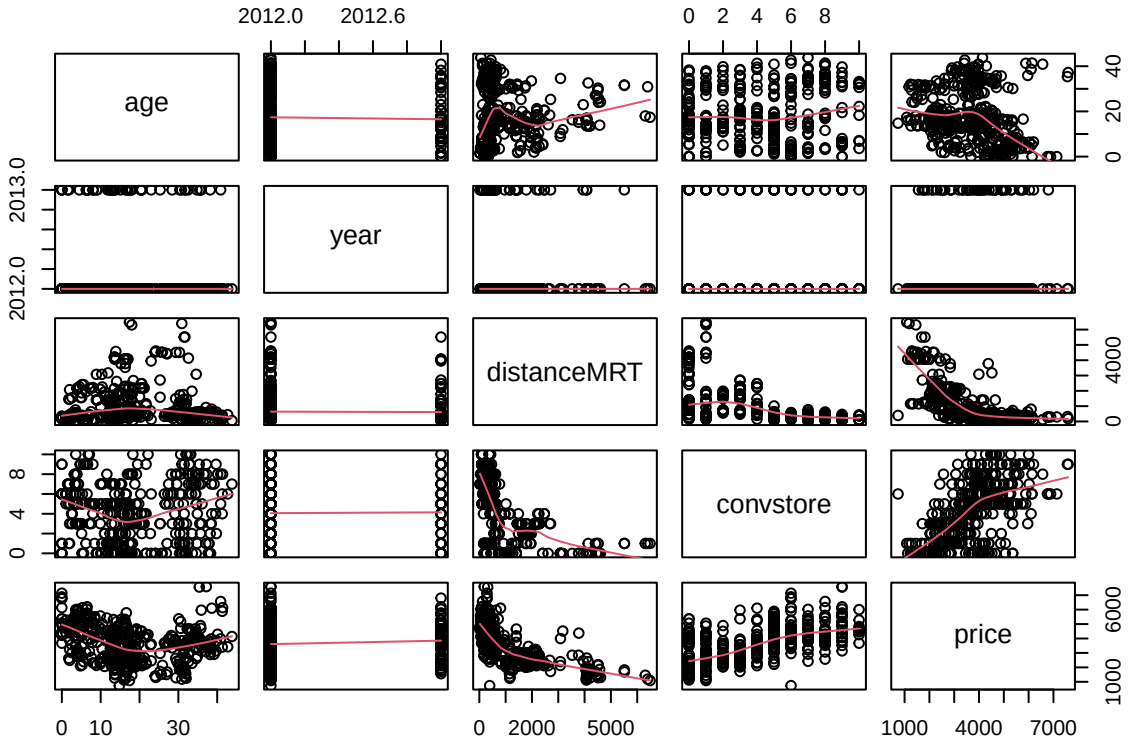
Explanation: There are 413 observations with 5 variables in the above realestate data set. The 5 variables are age(numeric), year(integer), distanceMRT(numeric), convstore(integer), and price(numeric).

Question 2

Construct the matrix plot and correlation matrix. Comment on the relationship among variables.

- matrix plot:

```
pairs(realstate,panel=panel.smooth)
```



From above plot, it can be depicted that there is a moderate linear relationship between 'convstore' and 'price' variables. There is no significant relationship of 'year' variable with any other variables. 'distanceMRT' and 'price' , 'distanceMRT' and 'price', 'distanceMRT' and convostore' seems to have negative correlation. It seems that except 'convstore' and 'price' variables, most of them have either very weak relationship or negative relationship or no significant relationship at all. Let's look into the more precise figure using cor() to accurately access their correlation.

- correlation matrix:

```
cor(realstate,method="pearson")
```

```
##                    age        year distanceMRT   convstore       price
## age          1.00000000 -0.02481960  0.02467605  0.04813145 -0.21098481
## year        -0.02481960  1.00000000 -0.01807399  0.01169100  0.05934157
## distanceMRT  0.02467605 -0.01807399  1.00000000 -0.60532789 -0.69400899
## convstore    0.04813145  0.01169100 -0.60532789  1.00000000  0.61263531
## price       -0.21098481  0.05934157 -0.69400899  0.61263531  1.00000000
```

```
# by default, R computes Pearson Correlation coefficient even if we don't specify.
```

Comment on the relationship among variables:

age and price: There is a negative correlation of approximately -0.21 between 'age' and 'price'. This suggests that older house tend to have slightly lower prices.

age and convostore: 'age' and 'convstore' also show a negligible correlation(0.04813145), indicating a lack of significant linear relationship between the house's transaction year and its proximity to convenience stores.

age and distanceMRT: There is a very minimal correlation between age and distanceMRT,around 0.024, which suggest that there's no significant linear relationship between the age of the house and its distance to the nearest metro rail transit.

age and year: There is a very weak negative correlation of approximately -0.024 between age and year.This indicates that there's no significant linear relationship between the age of the house and its transaction year.

year and price: 'year' and 'price' exhibit a positive correlation of around 0.06. This implies that there might be a slight increase in house price as the transaction year increases, but still the correlation is weak.

year and convstore: Similarly, 'year' and 'convstore' show a very weak positive correlation of about 0.012, indicating a very weak relationship between the transaction year and the number of convenience stores in the living circle on foot.

year and distanceMRT: There is a very weak negative correlation of approximately -0.018 between 'Year' and 'distanceMRT'. This suggests there is no significant relationship between the transaction year and the the distance to the nearest metro rail transit.

distanceMRT and price: There is a strong negative correlation of about -0.69 between 'distanceMRT' and 'price'. This indicates that house located closer to metro rail transit tend to have higher prices.

distanceMRT and convstore: There is a negative correlation of roughly -0.61 between 'distanceMRT' and 'convstore'. This suggests that properties located closer to metro rail transit tend to be have less number of convenience stores in the living circle on foot.

convstore and price: 'convstore' and 'price' has a positive correlation of around 0.61. This suggests that the house with higher number of convenience stores around its living circle tend to have higher prices.

Question 3

Simple Linear Regression

i. Fit a model to predict price in terms of distanceMRT

```
model1 <- lm(price ~ distanceMRT)
model1
```

```
##
## Call:
## lm(formula = price ~ distanceMRT)
##
## Coefficients:
## (Intercept)  distanceMRT
##   4419.1338      -0.6952
```

ii. Test the significance of the slope parameter (Write down the relevant hypothesis).

–> For to test the significance of the slope of the linear model we perform the hypothesis test. We'll perform this hypothesis test at 5% significance level.

H0: There is no linear relationship between price and distanceMRT.(OR, $H_0 : \beta = 0$)

HA: There is some linear relationship between price and distanceMRT.(OR, $(H_A : \beta \neq 0)$)

We calculate the p-value using summary(model) function and based on that p-value we conclude our test of significance of the slope of our model.

```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ distanceMRT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3408.8  -578.9   -98.0   491.2  3365.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4419.13375   59.20875   74.64   <2e-16 ***
## distanceMRT   -0.69516    0.03557  -19.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 911.9 on 411 degrees of freedom
## Multiple R-squared:  0.4816, Adjusted R-squared:  0.4804
## F-statistic: 381.9 on 1 and 411 DF,  p-value: < 2.2e-16
```

Here we can see, the p-value is 2.2e-16 which is less than 0.05. So we have strong evidence to reject the null hypothesis at 5% level of significance and support the alternative hypothesis ,i.e, HA : $(\beta \neq 0)$

Therefore, strong evidence to support that there is a significant linear relationship between price and distanceMRT.

iii. Interpret the slope and the intercept.

From the summary of our model above, we can deduce the following equation:

$$\hat{\text{price}} = 4419.13375 - 0.69516 \times \text{distanceMRT}$$

where, $\hat{\text{price}}$ is predicted price

From above equation we can interpret our slope and intercept as:

Slope: For every unit increase in 'distanceMRT', the price of the house is predicted to decrease by approximately $0.69516.

Intercept: It is the predicted value of 'price' when 'distanceMRT' is zero. However, in this context, it doesn't have a meaningful interpretation because having a 'distanceMRT' of zero might not be realistic.

Note: In the dataset, the distanceMRT are given in meters and price are given in USD per square meter.

iv.Discuss the accuracy of the parameter estimates (standard errors/confidence intervals).

By looking at the Model(model1) Summary Output above,

Standard errors of the estimates:

- On average the estimated value for intercept, can differ from the true value by 4419.13375 units.

- On average the estimated value for slope parameter, can differ from the true value by -0.69516 units.

95% Confidence Intervals (CI):

```
confint(model1)
```

```
##                     2.5 %       97.5 %
## (Intercept) 4302.7439928 4535.5235074
## distanceMRT   -0.7650898   -0.6252367
```

$$95\%\ \text{CI for } \hat{\text{price}}\text{: } [\ 4302.74,\ 4535.52\ ]$$

When the distance to MRT is zero, on average, the predicated price will range from 4302.7439928 to 4535.5235074 USD per square meter with 95% chance.

$$95\%\ \text{CI for distanceMRT: } [\ -0.7650898,\ -0.6252367\ ]$$

The coefficient of "distanceMRT" represents the change in the predicted price ($\hat{\text{price}}$) for a one-unit increase in "distanceMRT." In this case, the negative coefficient suggests that as "distanceMRT" increases by one unit, the predicted price tends to decrease by an amount between 0.6252367 and 0.7650898 units on average with 95% chance.

v. Discuss the overall accuracy of the model (R2, residual standard error etc)

We can assess the overall accuracy of the model by looking at the Model Summary Output or by using Anova table:

Model Summary:

```
summary(model1)
```

```
##
## Call:
## lm(formula = price ~ distanceMRT)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3408.8  -578.9   -98.0   491.2  3365.9
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4419.13375   59.20875   74.64   <2e-16 ***
## distanceMRT   -0.69516    0.03557  -19.54   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 911.9 on 411 degrees of freedom
## Multiple R-squared:  0.4816, Adjusted R-squared:  0.4804
## F-statistic: 381.9 on 1 and 411 DF,  p-value: < 2.2e-16
```

ANOVA table:

```
anova(model1)
```

```
## Analysis of Variance Table
##
## Response: price
##              Df    Sum Sq   Mean Sq F value    Pr(>F)
## distanceMRT   1 317585575 317585575   381.9 < 2.2e-16 ***
## Residuals   411 341786536    831597
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Therefore,

1. p-value = 2.2e-16 is extremely small. Supports strong linear relationship.

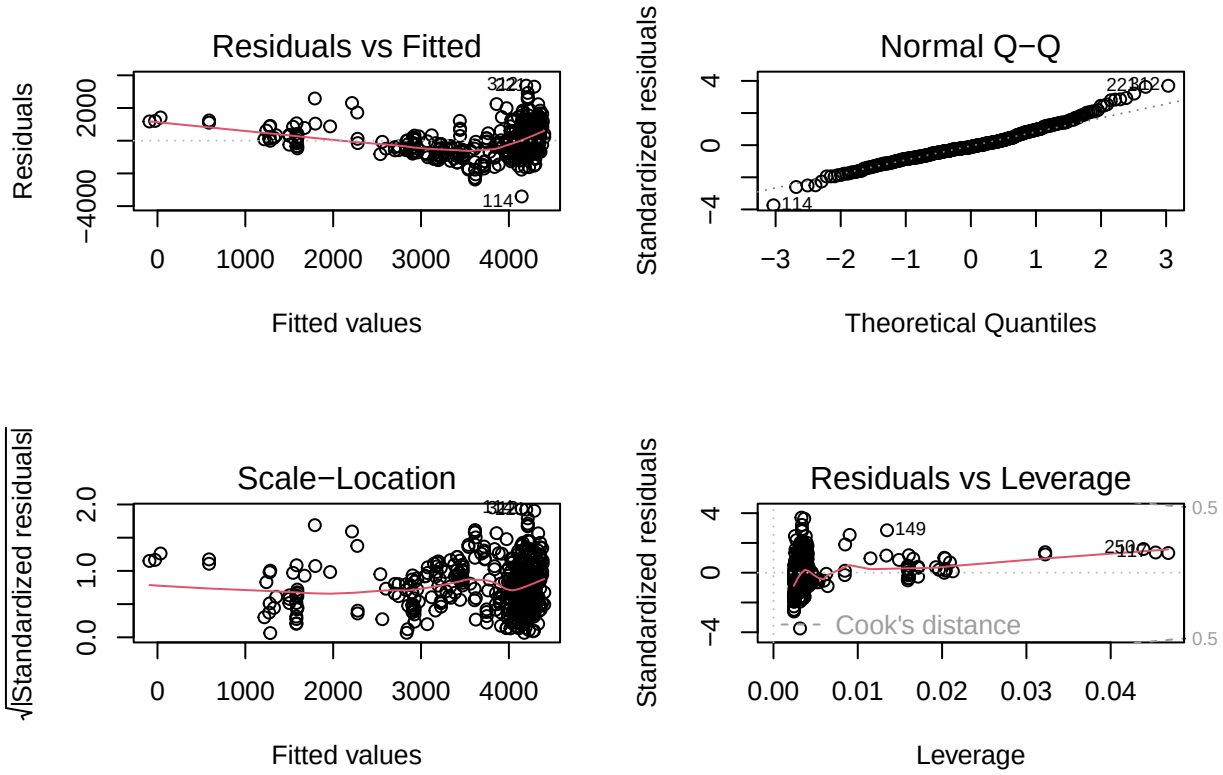
2. Residual standard error = 911.9, which means, on average, the predicted values deviate from the true regression line by 911.9.

3. R2 = 0.4816, that means 48.16% variation in price is explained by the regression model.

vi. Check for the model assumptions.

We can check whether the model assumptions were satisfied or not by using model diagnostic plots:

```
par(mfrow=c(2,2)) #this code divides the output page in 2-2
plot(model1)
```



plot1 (Residuals vs Fitted):

- we use this plot to check the linear relationship(or Linearity),
- the points must be scattered in a way that there is no pattern,
- for the case of this model, Graph 1 shows a non-linear pattern in the residual implying that the pattern in the data set is not completely captured by the model. Additionally, we can see that the variation is comparatively higher towards the end in compare to middle and start. This suggest that the residuals variance is not constant.

plot2 (Normal Q-Q):

- we use this plot to check the normality assumptions,
- if its normal all the data points should lie in the above diagonal (black) line which we can see in the plot.
- for the case of this model, normality assumptions is violated as observations are deviating from the straight line.

plot3 (Scale - Location):

- we use this plot to check the constant variation assumptions (Homoscedasticity),
- for this plot the fitted line should be flat, which means constant,
- In other words, for the value of X the variance of Y should be constant,
- for the case of this model, constant variation assumptions is violated as the variances are not constant.

plot4 (Residuals vs leverages):

- we use this plot to check influential observations and outliers
- for the case of this model, there are no influential observations.

vii. Write down the model equation.

The model equation for our linear regression model, where we are predicting "price" based on the "distanceMRT" , can be written as:

$$\text{price} = \beta_0 + \beta_1 \times \text{distanceMRT} + \varepsilon$$

Here:

β_0 is the intercept coefficient (the value of "price" when "distanceMRT" is 0).

β_1 is the coefficient for the "distanceMRT" (how much "price" changes for a one-unit increase in "distanceMRT").

ε represents the error term, accounting for the variability in the response variable that is not explained by the model.

So, in words, the equation represents how the "price" variable is predicted based on the linear relationship with "distanceMRT," while accounting for the intercept and error.

Substituting the value in our equation:

$$\text{price} = 4419.13375 - 0.69516 \times \text{distanceMRT} + \varepsilon$$

OR, we can also write the equation as:

$$\hat{\text{price}} = 4419.13375 - 0.69516 \times \text{distanceMRT}$$

viii. Predict the unit price of a house which is 500 meters away from MRT using the model in part vii.

To predict E[Y] value for a given X value (distanceMRT = 500meters):

$$\hat{\text{price}} = 4419.13375 - 0.69516 \times 500$$

$$\hat{\text{price}} = 4071.552$$

OR,

```
predict(model1,list(distanceMRT = 500.0))
```

```
##        1
## 4071.552
```

Thus, according to our model in part vii, the price of a house which is 500 meters away from MRT is predicted to be approximately 4071.552 USD per square meter.

Question4

Multiple Linear Regression

i. Fit a model to predict price in terms of all the other variables in the dataset.

```
model2_test <- lm(price ~ ., data = realstate)
model2_test
```

```
##
## Call:
## lm(formula = price ~ ., data = realstate)
##
## Coefficients:
## (Intercept)          age         year  distanceMRT    convstore
##  -3.130e+05   -2.369e+01    1.575e+02   -4.920e-01    1.396e+02
```

ii. Remove insignificant variables (if there is any) and fit a model including the rest of the variables.

–> To check for insignificant variables we check the model summary:

```
summary(model2_test)
```

```
##
## Call:
## lm(formula = price ~ ., data = realstate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3582.7  -520.6  -136.5   424.4  3283.0
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.130e+05  2.438e+05  -1.284    0.200
## age         -2.369e+01  3.541e+00  -6.689 7.41e-11 ***
## year         1.575e+02  1.212e+02   1.300    0.194
## distanceMRT -4.920e-01  4.011e-02 -12.268  < 2e-16 ***
## convstore    1.396e+02  1.721e+01   8.113 5.87e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 816.4 on 408 degrees of freedom
## Multiple R-squared:  0.5876, Adjusted R-squared:  0.5835
## F-statistic: 145.3 on 4 and 408 DF,  p-value: < 2.2e-16
```

From our 'model2_test' summary above, we can clearly see that the 'year' variable is not significant. Thus,it could be removed from our model and new model could be fit with the rest of variables:

```
model2 <- lm(price ~ age + distanceMRT + convstore)
summary(model2)
```

```
##
## Call:
## lm(formula = price ~ age + distanceMRT + convstore)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3603.4  -493.8  -143.3   446.4  3264.5
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 4048.20346  122.96886  32.921  < 2e-16 ***
## age          -23.80111    3.54310  -6.718 6.21e-11 ***
## distanceMRT   -0.49266    0.04014 -12.275  < 2e-16 ***
## convstore    139.70095   17.22623   8.110 5.96e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 817.1 on 409 degrees of freedom
## Multiple R-squared:  0.5858, Adjusted R-squared:  0.5828
## F-statistic: 192.9 on 3 and 409 DF,  p-value: < 2.2e-16
```

Now we can see that all the terms in the model above are significant. R2 = 58.58% and Residual Standard error = 817.1. This shows an improvement from the previous model.

iii. Add the interaction term distanceMRT * convstore to the model above (part ii).

```
model2_i <- lm(price ~ age + distanceMRT + convstore + distanceMRT * convstore)
summary(model2_i)
```

```
##
## Call:
## lm(formula = price ~ age + distanceMRT + convstore + distanceMRT *
##     convstore)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3594.4  -487.7   -83.3   384.3  3215.7
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           4052.44028  118.07977  34.320  < 2e-16 ***
## age                    -23.91084    3.40222  -7.028 8.85e-12 ***
## distanceMRT             -0.39236    0.04205  -9.331  < 2e-16 ***
## convstore              181.60659   17.97089  10.106  < 2e-16 ***
## distanceMRT:convstore   -0.12822    0.02149  -5.965 5.30e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 784.6 on 408 degrees of freedom
## Multiple R-squared:  0.6191, Adjusted R-squared:  0.6153
## F-statistic: 165.8 on 4 and 408 DF,  p-value: < 2.2e-16
```

Using hypothesis testing as mentioned above and from the output, it can be seen clearly that the interaction term is highly significant. Additionally, adding this interaction term has significantly improved our model, with the increased R2 = 61.91% and the decreased Residual Standard error = 784.6. This shows an improvement from the previous models.

iv. Comment on the significance of the parameters of the model above (part iii).

age (-23.91084): The coefficient for "age" is -23.91084. This means that for every one-unit increase in "age," the predicted "price" is expected to decrease by approximately $23.91. The p-value (8.85e-12) is very small, indicating that the "age" variable is statistically significant in predicting "price."

distanceMRT (-0.39236): The coefficient for "distanceMRT" is -0.39236. This suggests that for every one-unit increase in "distanceMRT," the predicted "price" is expected to decrease by approximately $0.39. The p-value (< 2e-16) indicates that "distanceMRT" is statistically significant.

convstore (181.60659): The coefficient for "convstore" is 181.60659. This means that the presence of a "convstore" is associated with an increase of approximately $181.61 in the predicted "price." The p-value (< 2e-16) suggests that "convstore" is statistically significant.

distanceMRT * convstore (-0.12822): The interaction term coefficient indicates how the relationship between "distanceMRT" and "price" changes depending on the presence of a "convstore." In this case, the coefficient is -0.12822. This suggests that the interaction term has a small but significant effect on the "price" prediction.

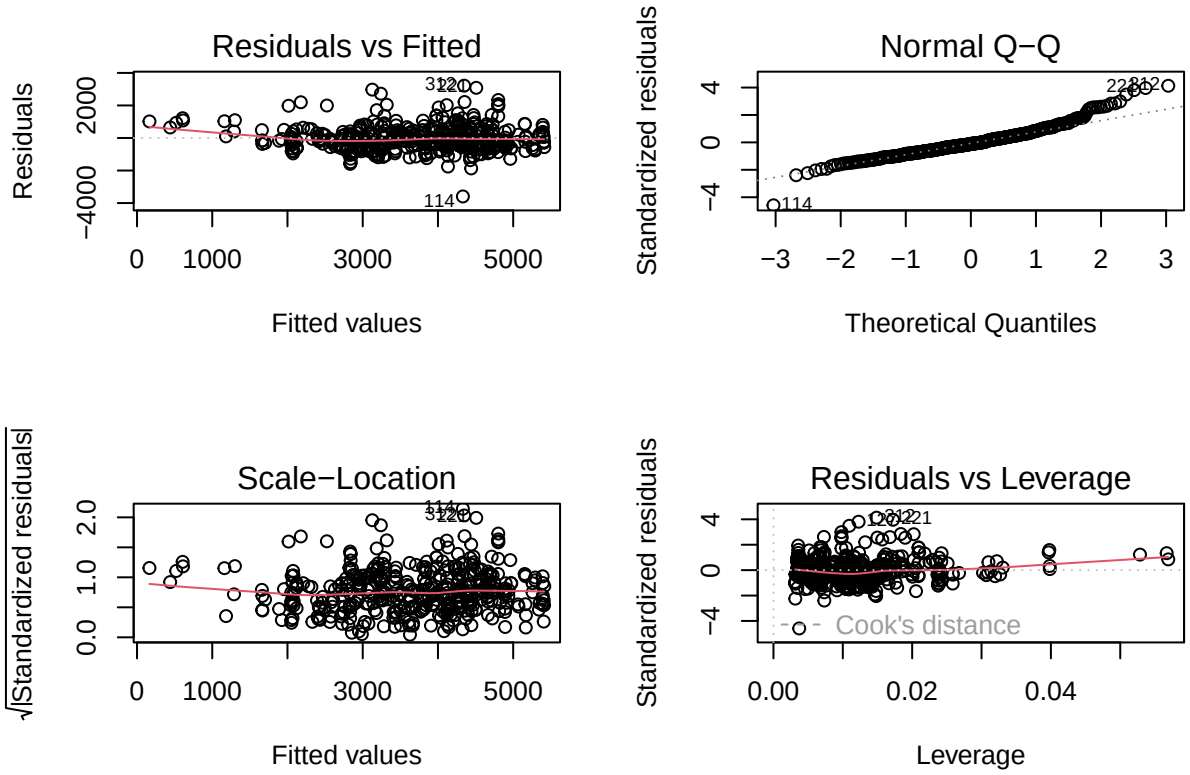
Overall, the p-values for all coefficients are very small (< 0.001), indicating that all independent variables and the interaction term are statistically significant in the model. The multiple R-squared value (0.6191) indicates that approximately 61.91% of the variability in "price" can be explained by the independent variables in the model.

Note: In the dataset, the age is given in years, distanceMRT are given in meters and price are given in USD per square meter.

v. Check for the model assumptions (model in part iii).

We can check for model assumptions by using model diagnostic plots:

```
par(mfrow=c(2,2)) #this code divides the output page in 2-2
plot(model2_i)
```



plot1 (Residuals vs Fitted):

- we use this plot to check the linear relationship(or Linearity),
- the points must be scattered in a way that there is no pattern,
- for the case of this model, Graph 1 shows a non-linear pattern in the residual implying that the pattern in the data set is not completely captured by the model.
- Additionally, we can see that the variation is comparatively higher towards the end and the middle in compare to start. This suggest that the residuals variance is not constant.

plot2 (Normal Q-Q):

- we use this plot to check the normality assumptions,
- if its normal all the data points should lie in the above diagonal (black) line which we can see in the plot.
- for the case of this model, normality assumptions is violated as observations are deviating from the straight line.

plot3 (Scale - Location):

- we use this plot to check the constant variation assumptions (Homoscedasticity),
- for this plot the fitted line should be flat, which means constant,
- In other words, for the value of X, the variance of Y should be constant,
- for the case of this model, constant variation assumptions is violated as the variances are not constant.

plot4 (Residuals vs leverages):

- we use this plot to check influential observations,
- for the case of this model, we can see that there are some outliers in the data points.

vi. Compare and comment on the accuracy of the models in part ii and part iii.

Metric	Model in Part ii	Model in Part iii	Comment
Residual Standard Error	817.1	784.6	A lower value indicates that the predictions are closer to the observed values, favoring the model in Part iii for better accuracy.
R-squared	0.5858	0.6191	A higher R-squared suggests better explanation of variance. Part iii has a higher R-squared, implying it fits the data more effectively.
Adjusted R-squared	0.5828	0.6153	Adjusted R-squared accounts for model complexity. The higher value in Part iii implies it's a better-fitting model considering predictor variables.
F-statistic	192.9	165.8	The F-statistic tests overall significance. While Part ii has a higher F-statistic, it should be interpreted alongside other metrics for a comprehensive assessment.

Overall, the model in Part iii generally demonstrates better accuracy and fit to the data based on lower residual standard error, higher R-squared and adjusted R-squared values. The F-statistic difference should be evaluated along with other metrics to provide a complete understanding of model performance.

Question 5

Polynomial Regression and Transformations

i. Fit a polynomial regression model to predict price using distanceMRT of order 3 and test the model significance. Give the resulting model.

```
model3 <- lm(price ~ distanceMRT +I(distanceMRT*distanceMRT)+
               I(distanceMRT*distanceMRT*distanceMRT),
             data = realstate)
summary(model3)
```

```
##
## Call:
## lm(formula = price ~ distanceMRT + I(distanceMRT * distanceMRT) +
##     I(distanceMRT * distanceMRT * distanceMRT), data = realstate)
##
## Residuals:
```

```
##     Min      1Q  Median      3Q     Max
## -3493.1  -475.4   -62.9   415.3  3188.1
##
## Coefficients:
##                                           Estimate Std. Error t value
## (Intercept)                               5.070e+03  8.511e+01  59.572
## distanceMRT                              -2.379e+00  1.905e-01 -12.488
## I(distanceMRT * distanceMRT)              6.387e-04  8.752e-05   7.298
## I(distanceMRT * distanceMRT * distanceMRT) -5.787e-08  1.057e-08  -5.473
##                                          Pr(>|t|)
## (Intercept)                               < 2e-16 ***
## distanceMRT                               < 2e-16 ***
## I(distanceMRT * distanceMRT)             1.53e-12 ***
## I(distanceMRT * distanceMRT * distanceMRT) 7.73e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 815.3 on 409 degrees of freedom
## Multiple R-squared:  0.5877, Adjusted R-squared:  0.5846
## F-statistic: 194.3 on 3 and 409 DF,  p-value: < 2.2e-16
```

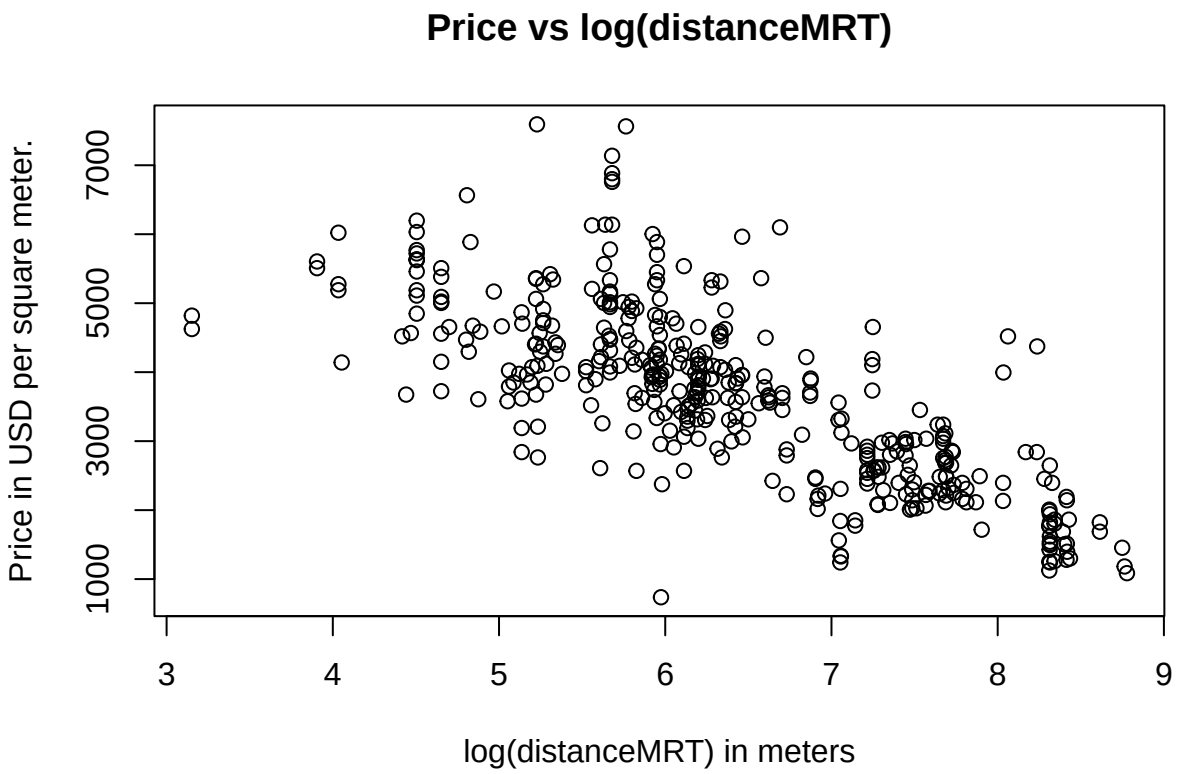
Using hypothesis testing as mentioned above and from the above output, it can be seen clearly that 'distanceMRT' variable with order 2 and order 3 are highly significant. Hence, this polynomial regression model is adequate.

The resulting model is:

$$\hat{\text{price}} = 5070 + (-2.379) \times \text{distanceMRT} + 0.0006387 \times \text{distanceMRT}^2 + (-0.00000005787) \times \text{distanceMRT}^3$$

ii. Construct a scatter plot to visualize the relationship between price and log transformed values of distanceMRT (price vs log(distanceMRT)).

```
plot(log(distanceMRT), price,
     main="Price vs log(distanceMRT)",
     xlab="log(distanceMRT) in meters", ylab="Price in USD per square meter.",
     col="black")
```



It seems there is a negative relationship between price and log-transformed distance to MRT. But let's fit a model and look at the summary for more confirmation.

iii. Fit a model to predict price in terms of log(distanceMRT).

```
model4 <- lm(price ~ log(distanceMRT), data = realstate)
summary(model4)
```

```
##
## Call:
## lm(formula = price ~ log(distanceMRT), data = realstate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3285.3  -463.2   -77.0   355.3  3361.5
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9122.61     236.68   38.54   <2e-16 ***
## log(distanceMRT)  -853.70      36.46  -23.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 829.1 on 411 degrees of freedom
## Multiple R-squared:  0.5715, Adjusted R-squared:  0.5705
```

```
## F-statistic: 548.2 on 1 and 411 DF,  p-value: < 2.2e-16
```

So, our resulting model will be:

$$\hat{\text{price}} = 9122.61 - 853.70 \times \log(\text{distanceMRT})$$

Where,

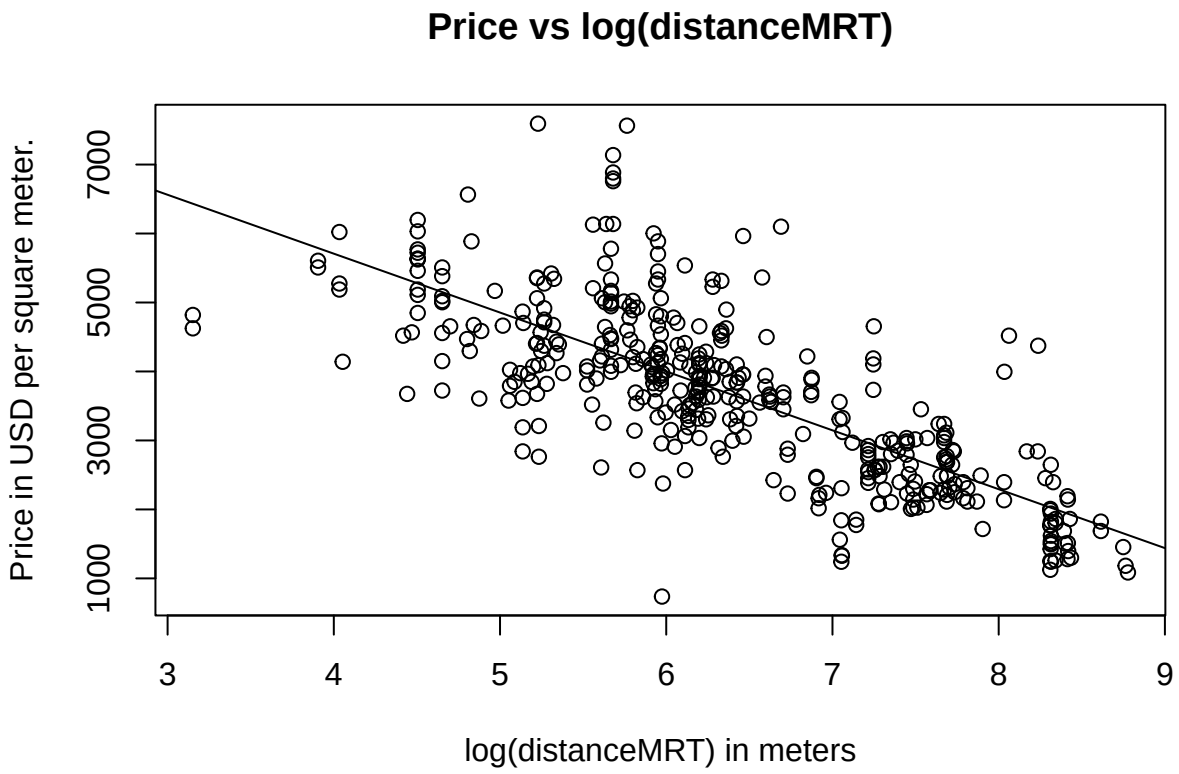
$\hat{\text{price}}$ is the predicted price.

log(distanceMRT) is the natural logarithm of the distance to MRT.

This equation suggests that there is a negative linear relationship between the log-transformed distance to MRT and the price on our linear regression model. Let's plot our model with fitted line and see what outcome we will get.

iv. Plot the straight line (regression line) corresponding to part iii within the scatter plot in part ii (i.e., draw both the scatter plot and the fitted line on one plot).

```
plot(log(distanceMRT), price,
     main="Price vs log(distanceMRT)",
     xlab="log(distanceMRT) in meters", ylab="Price in USD per square meter.",
     col="black")
abline(a=9122.61, b=-853.70)
```



```
#abline(model4)
```

As expected, the visual representation of our model, shows a decreasing trend which confirms the negative relationship. i.e, As the log-transformed distance to MRT increases, the predicted price tends to decrease along the regression line.