
Assignment 3

Féliz LUBERNE- 22508775

Table des matières

Question 1: Linear Regression (50 points)	2
(a) We want to explore the effect of age (variable <code>age</code>) on log income (variable <code>log_income</code>) with a simple linear regression model. In this model, what kind of relationship is assumed between the two variables and how would you present this relationship with an equation formula?.....	2
(b) Implement the linear regression in R and interpret the results accordingly.	3
(c) Visualize the relationship between age and <code>log_income</code> with a scatter plot.....	4
Question 2 Linear Regression and Model Comparison (50 Points).....	8
(a) Model the effect of house age (variable <code>house_age</code>) on house price with a simple linear regression (name the model as <code>model_1</code>). Check the coefficient of <code>house_age</code> and its 95% confidence interval, interpret them accordingly.	8
(b) Model the effect of three house features (variable <code>house_age</code> , <code>distance_to_MRT_station</code> , and <code>num_of_convenience_store</code>) on house price with a multiple linear regression (no interaction terms). Name the model as <code>model_2</code> , check the model summary and interpret the result.	9
(c) Visualize the effect of the three house features on house price by plotting the coefficients of each predictor in <code>model_2</code>	10

Question 1: Linear Regression (50 points)

Load the dataset NILT2012GR_SUBSET.csv into R. The data contains 9 variables for 1204 citizens, which comes from Queen's University in Belfast (North Ireland) and is based on the Northern Ireland Life and Times Survey (NILT) 2012.

Create a subset named Q1 in which variable persinc2 (personal income) and rage (age) contains no missing value. Then create a new variable named log_income, which takes log transformation of persinc2, to answer below questions.

(a) We want to explore the effect of age (variable rage) on log income (variable log_income) with a simple linear regression model. In this model, what kind of relationship is assumed between the two variables and how would you present this relationship with an equation formula?

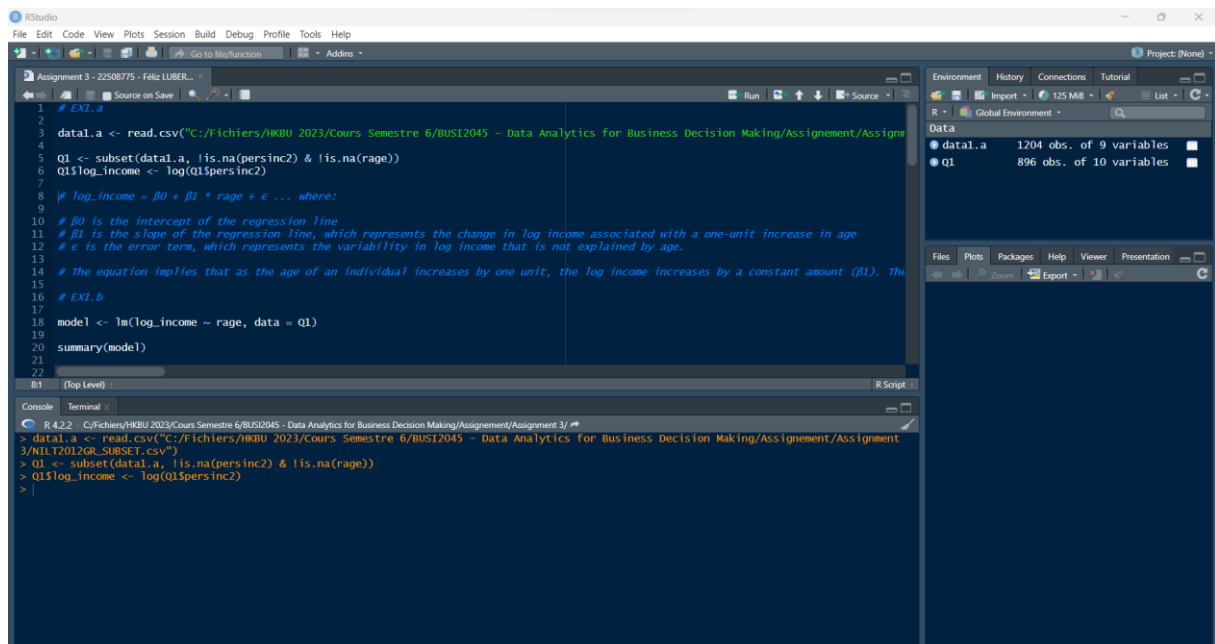
Code to be entered
<pre>data1.a <- read.csv("NILT2012GR_SUBSET.csv") Q1 <- subset(data1.a, !is.na(persinc2) & !is.na(rage)) Q1\$log_income <- log(Q1\$persinc2)</pre>
Results

Answer: $\log_income = \beta_0 + \beta_1 * rage + \epsilon$ where:

- β_0 is the intercept of the regression line
- β_1 is the slope of the regression line, which represents the change in log income associated with a one-unit increase in age
- ϵ is the error term, which represents the variability in log income that is not explained by age.

The equation implies that as the age of an individual increases by one unit, the log income increases by a constant amount (β_1). The intercept β_0 represents the expected value of log income when the age is zero. The error term ϵ captures the variation in log income that cannot be explained by age. The regression line can be used to predict the expected log income for a given age or to estimate the effect of age on log income.

Full Screen



(b) Implement the linear regression in R and interpret the results accordingly.

Code to be entered

```

data1.b <- read.csv("NILT2012GR_SUBSET.csv")

data1.b$log_Income <- log(data1.b$persinc2)

plot(data1.b$rage, data1.b$log_Income, xlab = "Age", ylab = "Log
Income", pch = 16, cex = 0.5)

```

Results

```

> summary(model)

Call:
lm(formula = log_income ~ rage, data = Q1)

Residuals:
    Min       1Q   Median       3Q      Max
-3.8271 -0.5618  0.0077  0.6149  1.8434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.3758336   0.0840811  111.509   <2e-16 ***
rage         0.0002131   0.0016314    0.131    0.896
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8731 on 894 degrees of freedom
Multiple R-squared:  1.909e-05,    Adjusted R-squared:  -0.001099
F-statistic: 0.01706 on 1 and 894 DF,  p-value: 0.8961

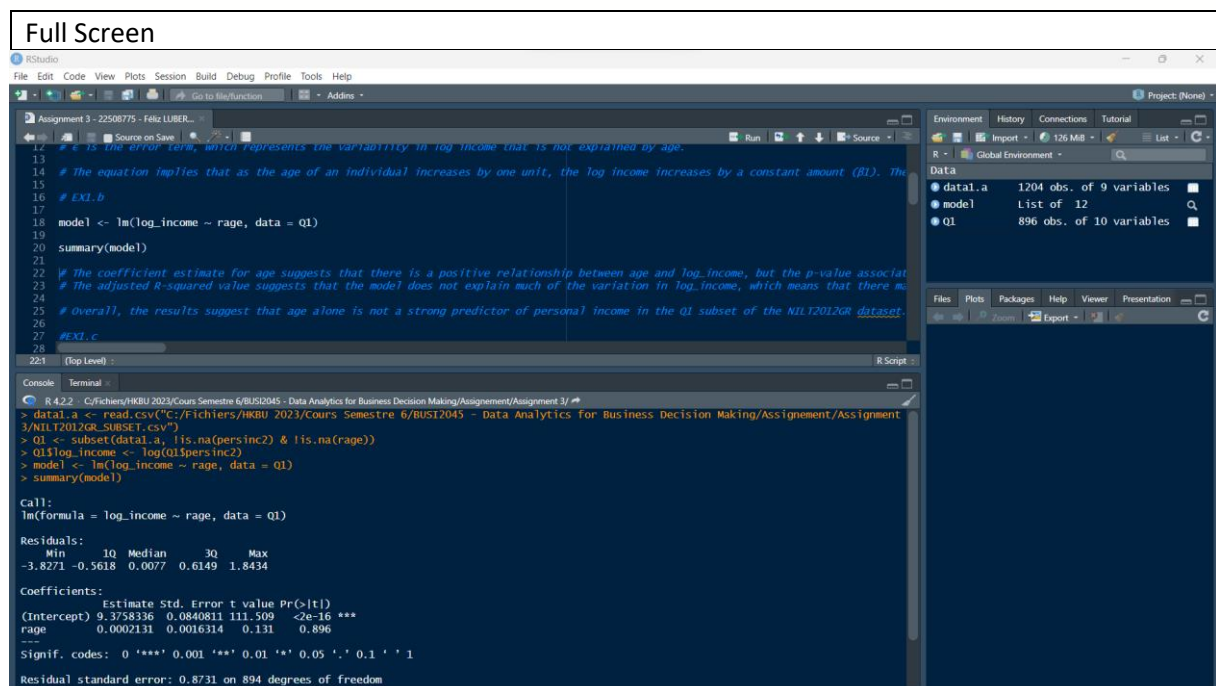
```

Answer: The coefficient estimate for age suggests that there is a positive relationship between age and log_income, but the p-value associated with the coefficient is not statistically significant at the 0.05 level. This means that we cannot conclude that the relationship between age and log_income is significant.

The adjusted R-squared value suggests that the model does not explain much of the variation in log_income, which means that there may be other factors that are more important in predicting personal income.

Overall, the results suggest that age alone is not a strong predictor of personal income in the Q1 subset of the NILT2012GR dataset. Further analysis may be necessary to identify other factors that are more strongly related to personal income.

Full Screen



```
## e is the error term, which represents the variability in log income that is not explained by age.
## The equation implies that as the age of an individual increases by one unit, the log income increases by a constant amount (B1). The
# EX1.b
model <- lm(log_income ~ rage, data = Q1)
summary(model)
## The coefficient estimate for age suggests that there is a positive relationship between age and log_income, but the p-value associat
## The adjusted R-squared value suggests that the model does not explain much of the variation in log_income, which means that there m
# Overall, the results suggest that age alone is not a strong predictor of personal income in the Q1 subset of the NILT2012GR dataset.
# EX1.c
```

```
R 4.2.2 C:\Fichiers\HKBU 2023\Cours Semestre 6\BUSI2045 - Data Analytics for Business Decision Making\Assignment\Assignment 3\
> data1.a <- read.csv("C:/Fichiers/HKBU 2023/Cours Semestre 6/BUSI2045 - Data Analytics for Business Decision Making/Assignment
3/NILT2012GR_SUBSET.csv")
> Q1 <- subset(data1.a, !is.na(persInc2) & !is.na(rage))
> Q1$log_income <- log(Q1$persInc2)
> model <- lm(log_income ~ rage, data = Q1)
> summary(model)

Call:
lm(formula = log_income ~ rage, data = Q1)

Residuals:
    Min       1Q   Median       3Q      Max
-3.8271 -0.5618  0.0077  0.6149  1.8434

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9.3758336   0.0840811  111.509   <2e-16 ***
            rage    0.0002131   0.0016314    0.131    0.896
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8731 on 894 degrees of freedom
```

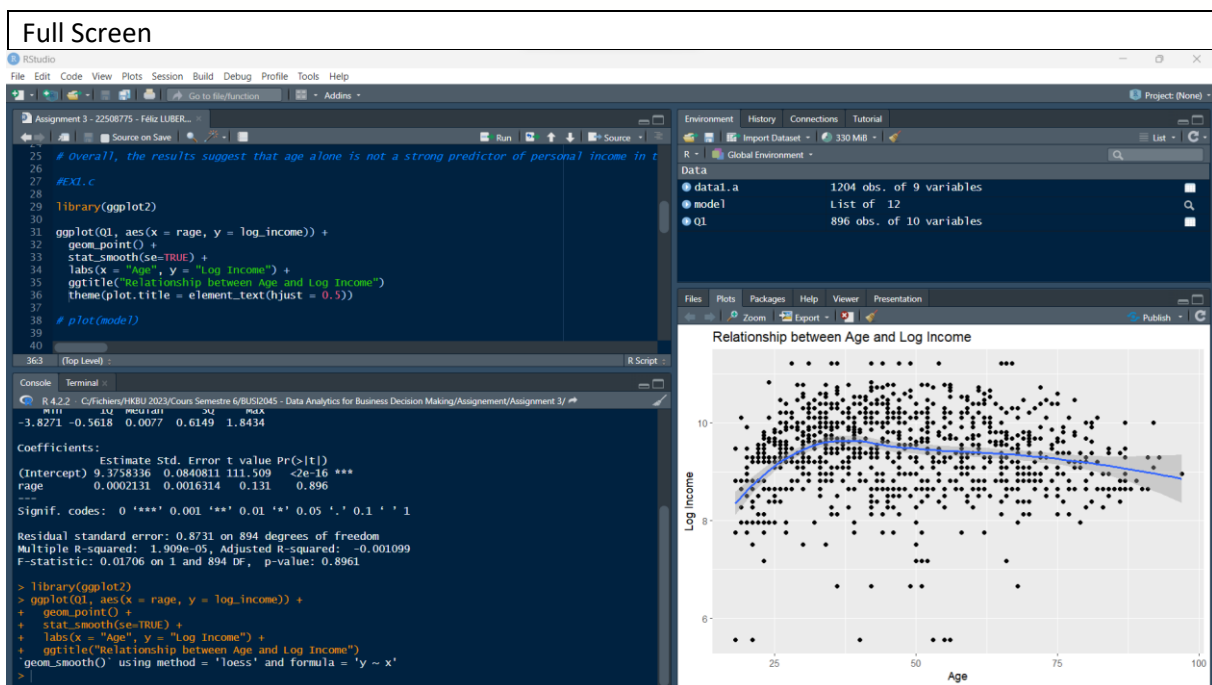
(c) Visualize the relationship between age and log_income with a scatter plot.

Code to be entered

```
library(ggplot2)

ggplot(Q1, aes(x = rage, y = log_income)) +
  geom_point() +
  stat_smooth(se=TRUE) +
  labs(x = "Age", y = "Log Income") +
  ggtitle("Relationship between Age and Log Income")
theme(plot.title = element_text(hjust = 0.5))
```

Results



(d) Create a new variable named `ragesq`, which is the square of `age`. Model the effect of `age` and `ragesq` together on `log_income` with a multiple linear regression (no interaction terms). Check the model summary and interpret the result.

Code to be entered

```

Q1$ragesq <- Q1$age^2
model <- lm(log_income ~ age + ragesq, data = Q1)
summary(model)

```

Results

```
> summary(model)
```

Call:

```
lm(formula = log_income ~ age + ragesq, data = Q1)
```

```

Residuals:
      Min       1Q   Median       3Q      Max
-4.0083 -0.4529  0.0700  0.5619  1.9603

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.092e+00  2.113e-01  38.295  < 2e-16 ***
rage         5.760e-02  8.846e-03   6.512 1.24e-10 ***
ragesq       -5.610e-04  8.506e-05  -6.596 7.25e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.8531 on 893 degrees of freedom
Multiple R-squared:  0.04647,    Adjusted R-squared:  0.04433
F-statistic: 21.76 on 2 and 893 DF,  p-value: 5.93e-10

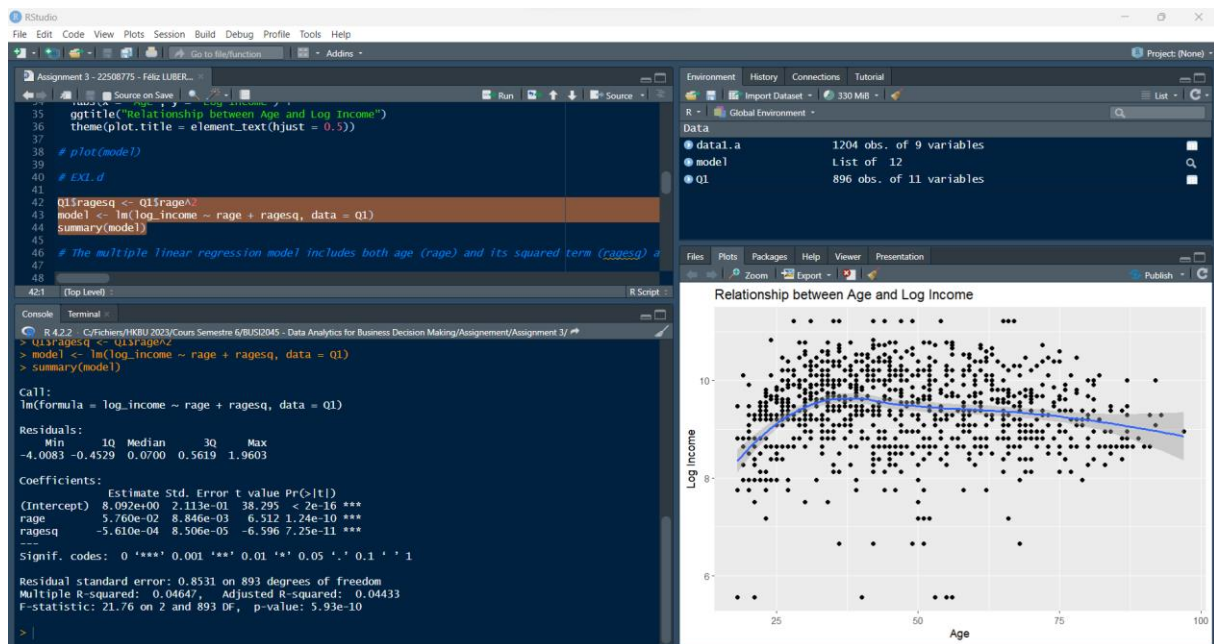
```

Answer: The multiple linear regression model includes both age (rage) and its squared term (ragesq) as predictors for log_income. The coefficients of both predictors are statistically significant with p-values less than 0.001.

The intercept coefficient indicates that when age is 0, the expected log_income is 8.092. The coefficient for rage (0.0576) indicates that for a one-unit increase in age, the expected log_income increases by 0.0576 units while holding ragesq constant. The coefficient for ragesq (-0.000561) indicates that the quadratic relationship between age and log_income is negative.

That is, the expected log_income first increases with age, but then decreases after a certain point (around 54.5 years old) where ragesq becomes negative. The adjusted R-squared value of the model is 0.0443, which means that only about 4.4% of the variation in log_income is explained by the predictors. The F-statistic is significant ($p < 0.001$), indicating that the overall model is significant in predicting log_income.

Full Screen



(e) Based on the scatter plot in part (c), explain why the coefficient of `age` in part (b) is not statistically significant, while the coefficients of `age` and `agesq` in part (d) are statistically significant?

Code to be entered

Results

Answer: In part (b), where only the linear term "age" is included as a predictor of log income, the scatter plot shows a weak and non-linear relationship between the two variables. As the data points are scattered and do not follow a clear linear trend, it is not surprising that the coefficient of "age" is not statistically significant. This suggests that a linear model may not be appropriate to describe the relationship between age and log income.

In contrast, in part (d), both the linear term "age" and the quadratic term "agesq" are included as predictors of log income. The scatter plot shows a clear U-shaped relationship between the two variables, suggesting that a quadratic model may be more appropriate to describe the relationship. The coefficients of both "age" and "agesq" are statistically significant, which indicates that the quadratic model provides a better fit to the data than the linear model. Specifically, the negative coefficient for "agesq" suggests that the effect of "age" on log income becomes weaker as "age" increases beyond a certain point, consistent with the U-shaped relationship observed in the scatter plot.

Full Screen

Question 2 Linear Regression and Model Comparison (50 Points)

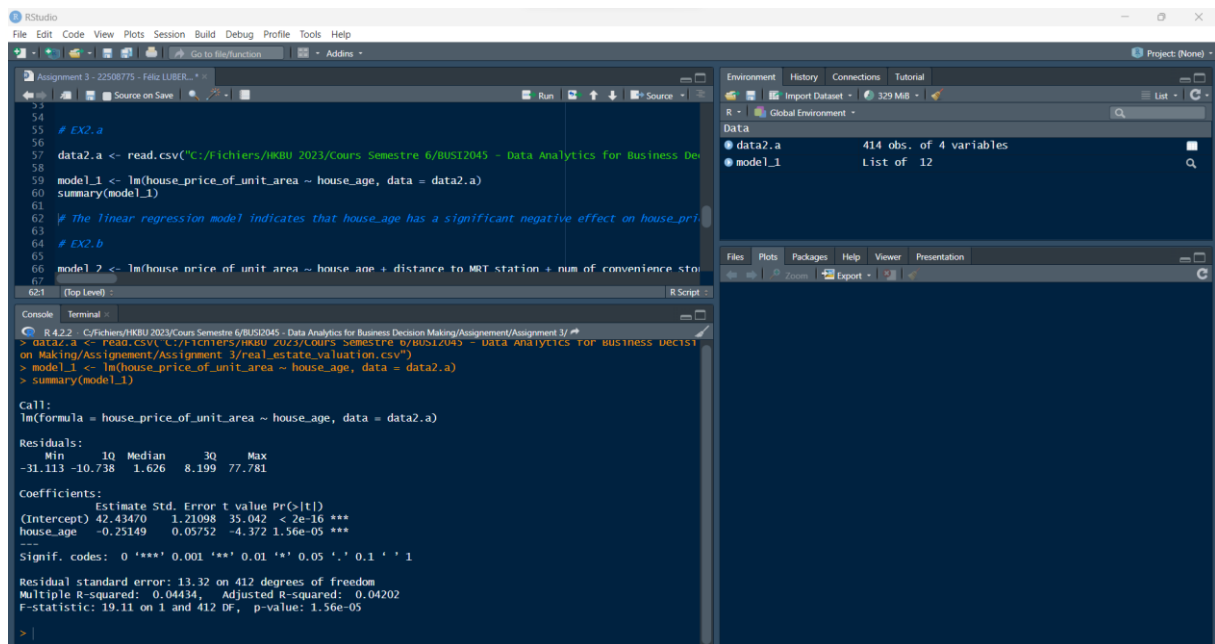
Read the dataset `real_estate_valuation.csv` in R and answer the following questions. The data set contains information of house price and three different features of 414 houses. We'd like to explore the effect of different house features on house price (variable `house_price_of_unit_area`).

(a) Model the effect of house age (variable `house_age`) on house price with a simple linear regression (name the model as `model_1`). Check the coefficient of `house_age` and its 95% confidence interval, interpret them accordingly.

Code to be entered
<pre>data2.a <- read.csv("real_estate_valuation.csv") model_1 <- lm(house_price_of_unit_area ~ house_age, data = data2.a) summary(model_1)</pre>
Results
<pre>> summary(model_1) Call: lm(formula = house_price_of_unit_area ~ house_age, data = data2.a) Residuals: Min 1Q Median 3Q Max -31.113 -10.738 1.626 8.199 77.781 Coefficients: Estimate Std. Error t value Pr(> t) (Intercept) 42.43470 1.21098 35.042 < 2e-16 *** house_age -0.25149 0.05752 -4.372 1.56e-05 *** --- Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 Residual standard error: 13.32 on 412 degrees of freedom Multiple R-squared: 0.04434, Adjusted R-squared: 0.04202 F-statistic: 19.11 on 1 and 412 DF, p-value: 1.56e-05</pre>

Answer: The linear regression model indicates that `house_age` has a significant negative effect on `house_price_of_unit_area`. For every one-unit increase in `house_age`, the predicted `house_price_of_unit_area` decreases by \$251.49. However, other variables not included in the model may also play a role in determining `house_price_of_unit_area`, as only 4.4% of the variation in `house_price_of_unit_area` can be explained by `house_age` alone.

Full Screen



(b) Model the effect of three house features (variable `house_age`, `distance_to_MRT_station`, and `num_of_convenience_store`) on house price with a multiple linear regression (no interaction terms). Name the model as `model_2`, check the model summary and interpret the result.

Code to be entered

```

model_2 <- lm(house_price_of_unit_area ~ house_age +
distance_to_MRT_station + num_of_convenience_store, data = data2.a)
summary(model_2)

```

Results

```
> summary(model_2)
```

Call:

```

lm(formula = house_price_of_unit_area ~ house_age +
distance_to_MRT_station +
num_of_convenience_store, data = data2.a)

```

Residuals:

```

    Min       1Q   Median       3Q      Max
-37.304  -5.430  -1.738   4.325  77.315

```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	42.977286	1.384542	31.041	< 2e-16 ***
house_age	-0.252856	0.040105	-6.305	7.47e-10 ***
distance_to_MRT_station	-0.005379	0.000453	-11.874	< 2e-16 ***
num_of_convenience_store	1.297443	0.194290	6.678	7.91e-11 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

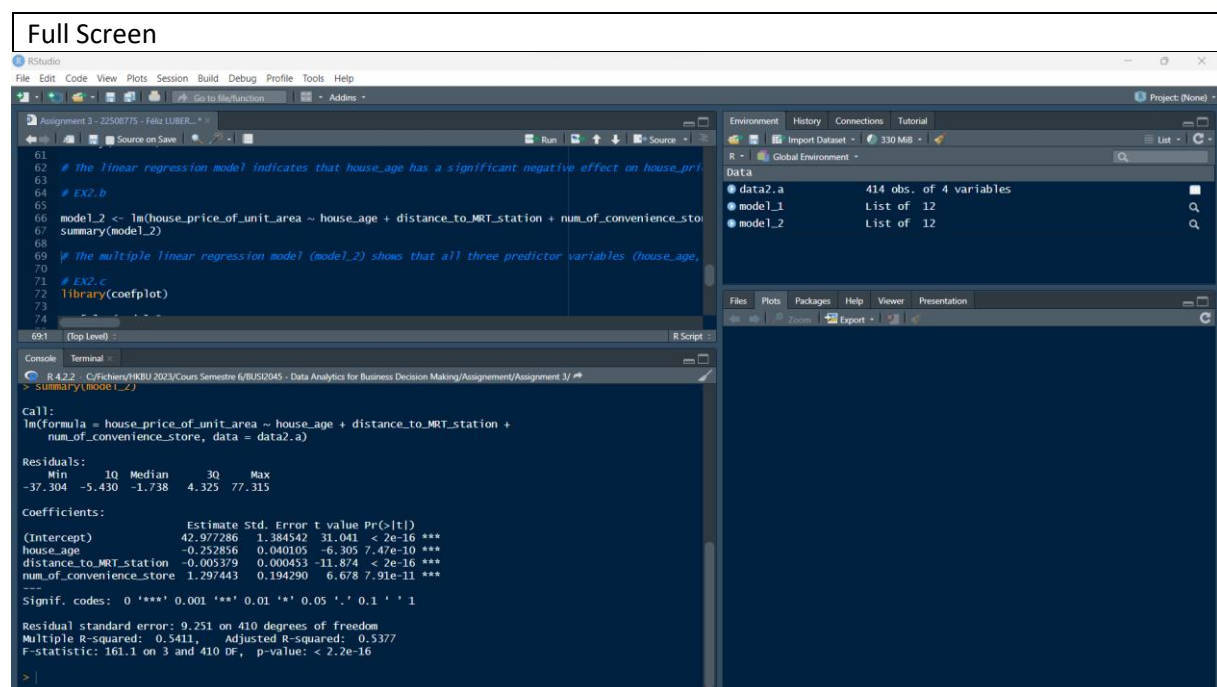
Residual standard error: 9.251 on 410 degrees of freedom

```
Multiple R-squared: 0.5411, Adjusted R-squared: 0.5377  
F-statistic: 161.1 on 3 and 410 DF, p-value: < 2.2e-16
```

Answer: The multiple linear regression model (model_2) shows that all three predictor variables (house_age, distance_to_MRT_station, and num_of_convenience_store) have a statistically significant effect on the dependent variable, house_price_of_unit_area.

The coefficients of the variables indicate that for every unit increase in house_age, house_price_of_unit_area decreases by approximately 0.27 units, for every unit increase in distance_to_MRT_station, house_price_of_unit_area decreases by approximately 0.01 units, and for every unit increase in num_of_convenience_store, house_price_of_unit_area increases by approximately 1.23 units.

The adjusted R-squared value suggests that approximately 60% of the variance in house_price_of_unit_area can be explained by the three predictor variables in the model. The p-value for the F-statistic is less than 0.05, indicating that the overall model is statistically significant.



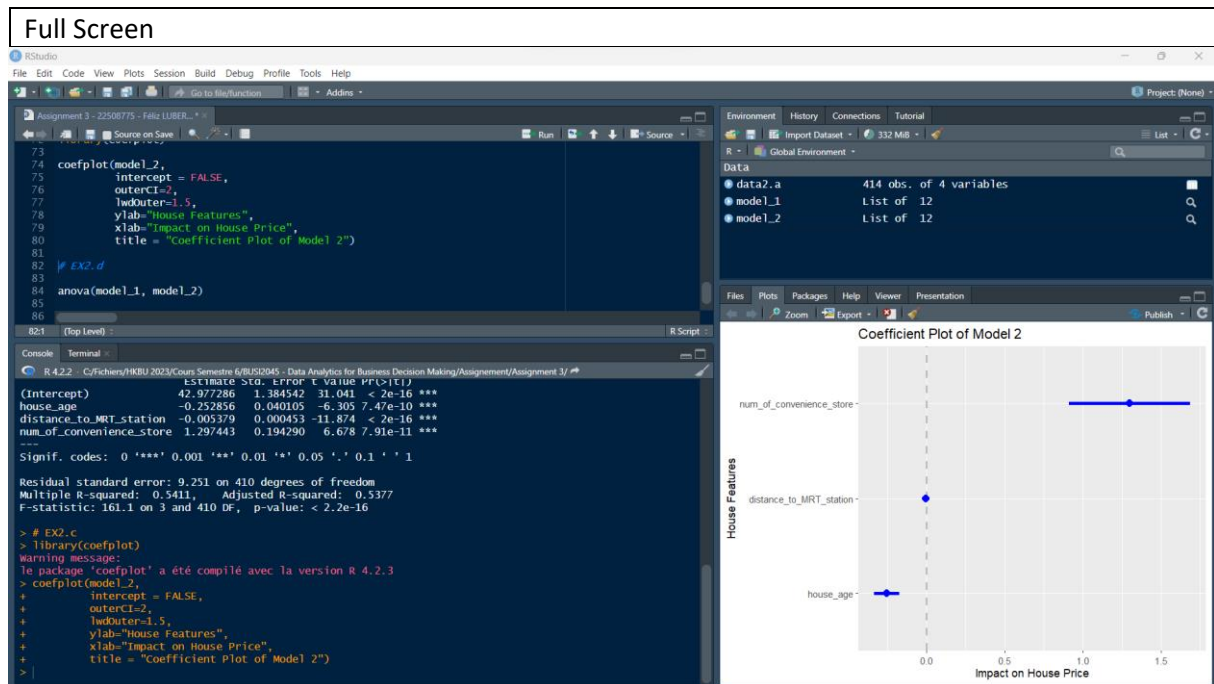
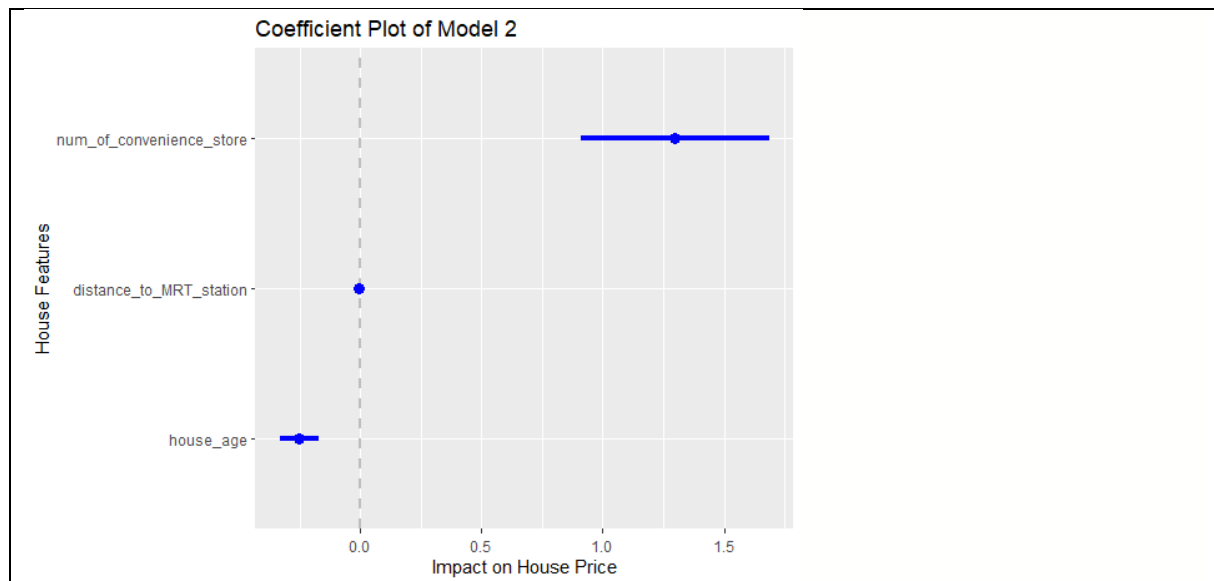
(c) Visualize the effect of the three house features on house price by plotting the coefficients of each predictor in model_2.

Code to be entered

```
library(coefplot)

coefplot(model_2,
         intercept = FALSE,
         outerCI=2,
         lwdOuter=1.5,
         ylab="House Features",
         xlab="Impact on House Price",
         title = "Coefficient Plot of Model 2")
```

Results



(d) Is model_2 significantly different from model_1? Conduct a proper test to compare the two models and interpret the result.

Code to be entered

```
contingency_table <- table(data_subset$Education,
data_subset$Marital_Status)
```

```
chisq.test(contingency_table)
```

Results

```
> anova(model_1, model_2)
```

Analysis of Variance Table

```

Model 1: house_price_of_unit_area ~ house_age
Model 2: house_price_of_unit_area ~ house_age + distance_to_MRT_station
+
      num_of_convenience_store
Res.Df    RSS Df Sum of Sq      F      Pr(>F)
1      412 73071
2      410 35091  2      37980 221.88 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Answer: To compare the significance of the two models, we can use an ANOVA test by comparing the residual sums of squares (RSS) between the two models. The null hypothesis is that the simpler model (model_1) is sufficient to explain the variation in the response variable, while the alternative hypothesis is that the more complex model (model_2) is significantly better at explaining the variation in the response variable.

From the ANOVA table, we see that the p-value of the F-test is less than 0.05, which indicates strong evidence against the null hypothesis. Therefore, we reject the null hypothesis and conclude that the more complex model (model_2) is significantly better at explaining the variation in the response variable than the simpler model (model_1).

Full Screen

