Introduction to Data Science (11372 & G 11516)

Semester 1 2021

UNIVERSITY OF CANBERRA

INTRODUCTION TO DATA SCIENCE

Lecture 12

Dr. Ibrahim Radwan

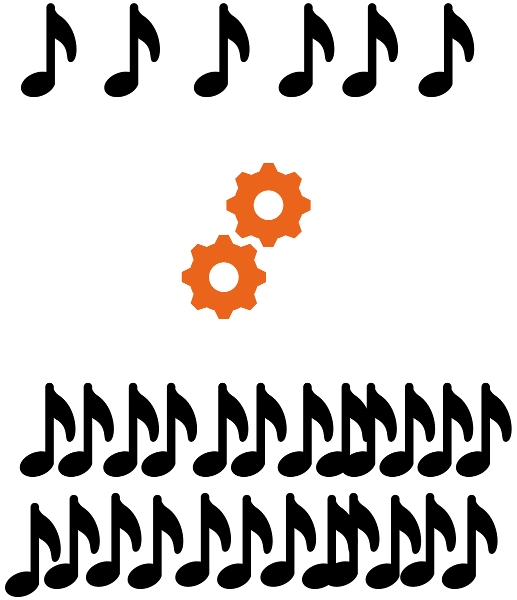
DISTINCTIVE BY DESIGN

OUTLINE

- Building a Linear Regression model
- Checking assumptions of the Linear Regression
- A Case study
- Next steps in data modelling

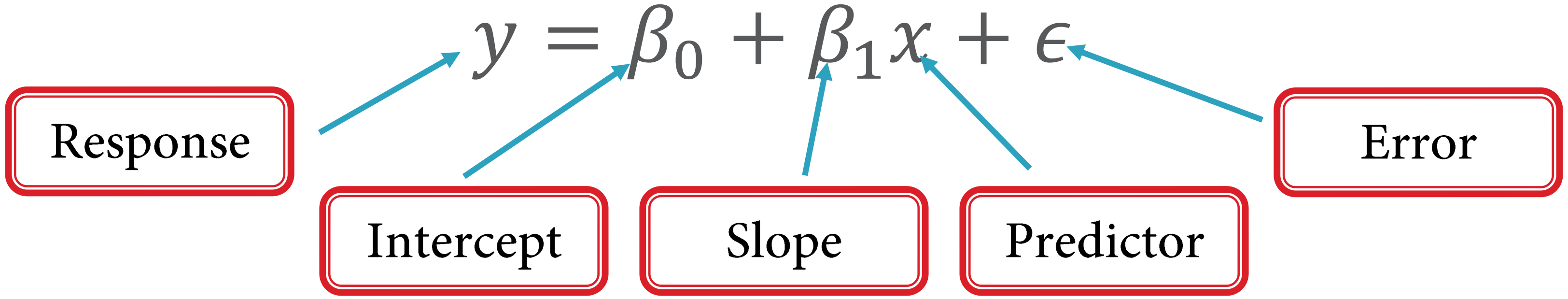
LINEAR REGRESSION

- Linear regression is a linear approximation of a causal relationship between two or more variables
- Process goes in the following steps:
  1. Get Sample Data
  2. Design a model that works for that sample
  3. Make predictions for the whole population
  4. Measure performance

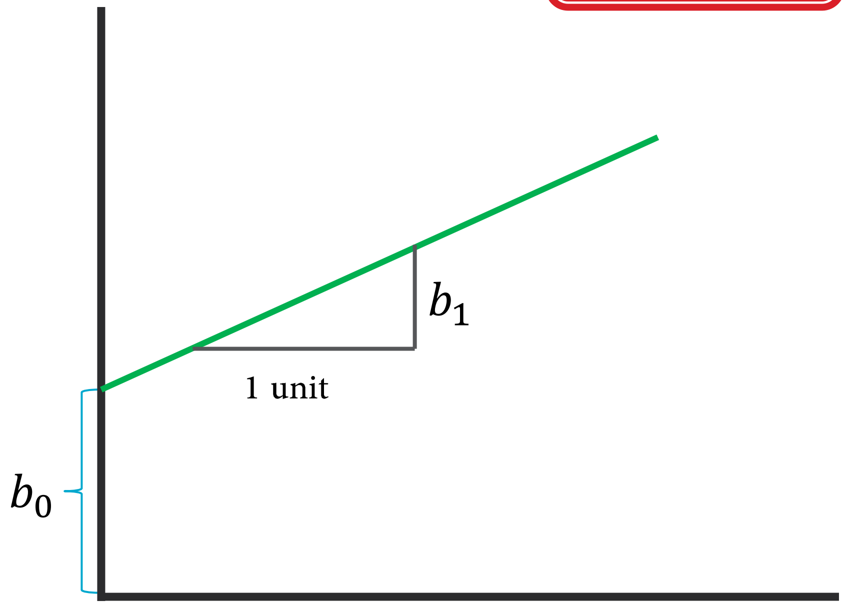


LINEAR REGRESSION (2)

- The linear regression can be estimated by following this formula:



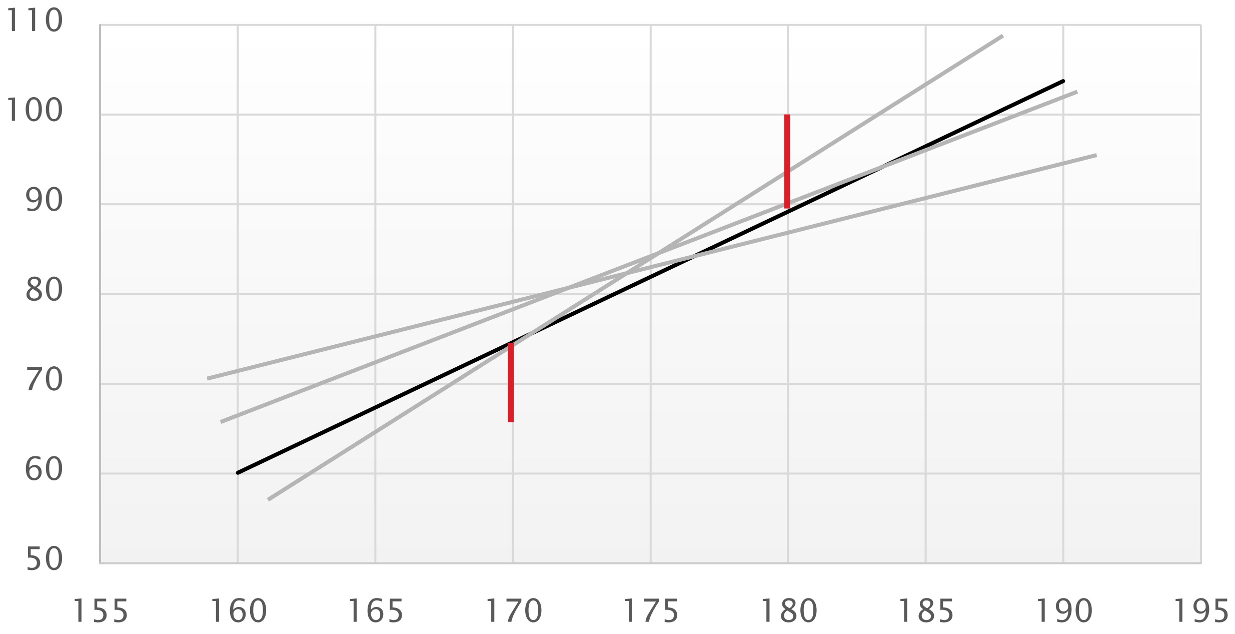
$$y = \beta_0 + \beta_1 x + \epsilon$$



…, so now given (x,y) pairs of data samples, can we estimate b_0 and b_1 ?

RESIDUALS

- To estimate the regression parameters (b_0 and b_1), we *minimise* the difference between the observed and the predicted values.
- This difference is called *residuals.*



BUILDING MODEL IN R

- In R, we use **lm** function to build the linear model
- When calling the lm function, the variable that we want to predict (i.e. response) is allocated to the left of the ~ symbol, and the variables that we use to predict (i.e. explanatory) is placed at the right of the ~ symbol.

```
> head(cars)
  speed dist
1     4    2
2     4   10
3     7    4
4     7   22
5     8   16
6     9   10
```

```
# fit regression line to predict cars’s stopping distances
from cars’s speed
model <- lm(dist~ speed, data = cars)
```

INTERPRETING MODEL OUTPUT

```
model
## Call:
## lm(formula = dist ~ speed, data=cars)
##
## Coefficients:
##(Intercept)          speed
##     -17.579         3.932
```

- The estimated regression can be written as follow:
  - dist = -17.579 + 3.932 x speed
- The intercept (b0) is -17.579 . It can be interpreted as the predicted dist for a zero speed is -17.579 ft.
- The slope (b1), is 3.932 . This means that, for every 1 mph increase in the speed, the stopping distance is getting increased by 3.932 feet.

INTERPRETING MODEL OUTPUT (2)

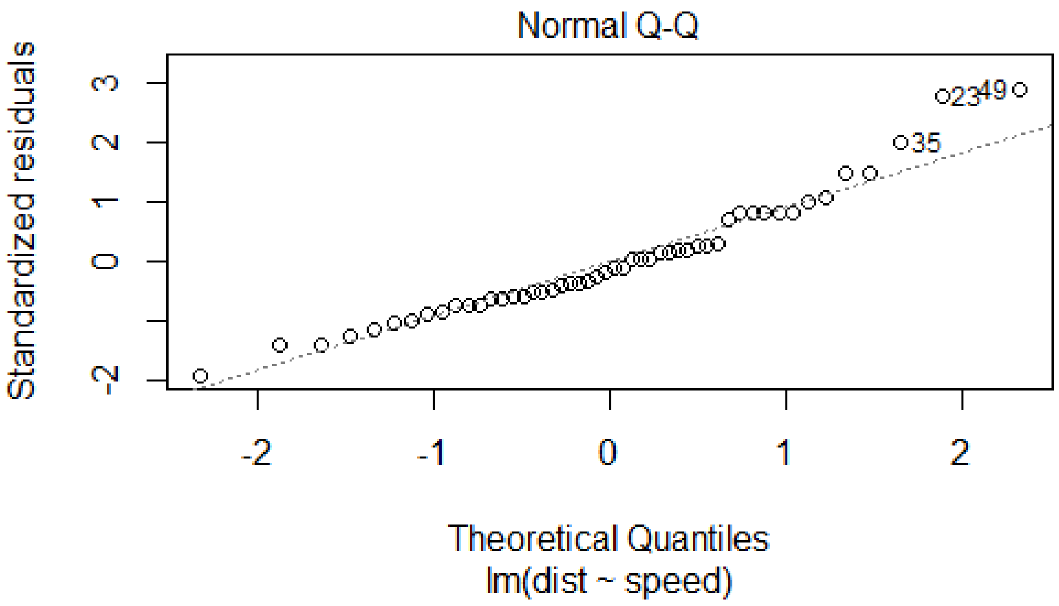
```
summary(model)
Call:
lm(formula = dist ~ speed, data = cars)

Residuals:
    Min      1Q  Median      3Q     Max
-29.069  -9.525  -2.272   9.215  43.201

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -17.5791     6.7584  -2.601   0.0123 *
speed         3.9324     0.4155   9.464 1.49e-12 ***
---
Signif. codes:
0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 15.38 on 48 degrees of freedom
Multiple R-squared:  0.6511,  Adjusted R-squared:  0.6438
F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

Overall p value on the basis of F-statistic, normally p value less than 0.05 indicates that overall model is significant

VISUALISING MODEL

- To visualize the fitted model, you can use Q-Q Plot:

```
> plot(model,2)
```



A straighter line of the residual quantiles means better model fitting

THE PREDICTION

- The predicted value is often denoted as $\hat{y}$
- The predict function in R can give us predictions directly.

```
# predict Y values
Y_hat <- predict(model, newdata=new_data,se.fit = TRUE)
names(Y_hat)
```

LINEAR REGRESSION – ASSUMPTIONS

UNIVERSITY OF CANBERRA

Pre-building assumptions

1- Linearity
2- Normality

Post-building assumptions

1- Error distribution
2- Patterns of Residual vs. fitted values

PRE-ASSUMPTIONS

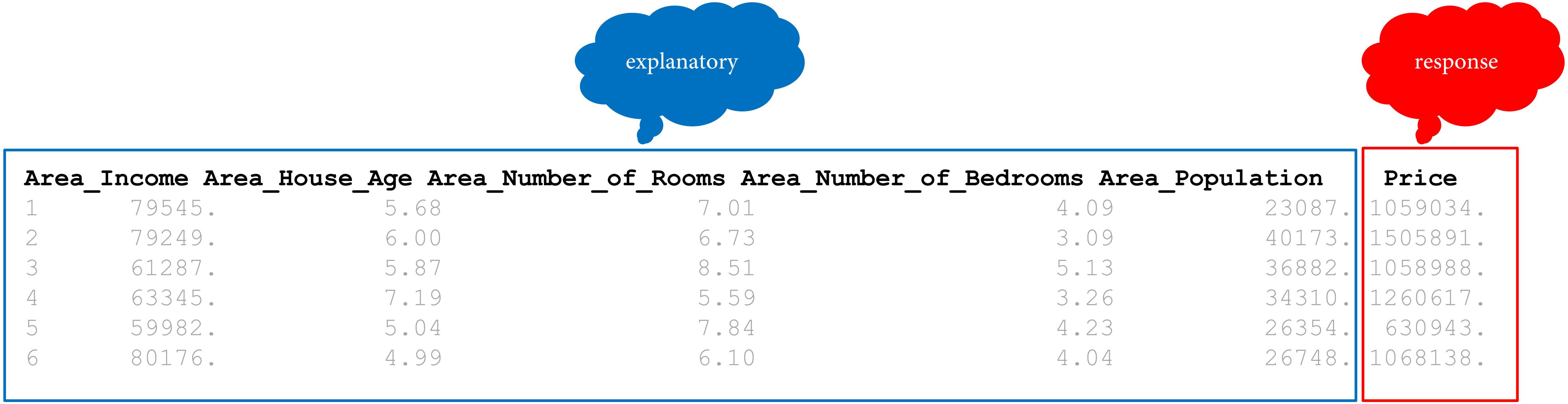
- Linearity and Co-linearity
  - The relationships between the response variable and the explanatory variables should be linear and strong enough to be encoded by a linear regression model.
    - To test this relationship, we either do that graphically (e.g. scatter plots) or numerically (e.g. Person correlation).
  - Also, there should <u>not</u> be co-linear relationship between the explanatory variables.
    - …, as these variables should be independent of each other.
- Normality
  - The distribution of the response variable should be normal
    - We can use the histogram plot to inspect the distribution of this variable.

POST-ASSUMPTIONS

- Residuals follow normal distribution
  - This can be checked by visualizing the histogram of the residuals (i.e. errors) or by generating Q-Q plot of the residuals.
- The variance of the residuals is constant
  - This can be checked by visualizing the scatter plot of residuals vs. fitted values.

Let us test these assumptions with the following case study.

LINEAR REGRESSION, EXAMPLE



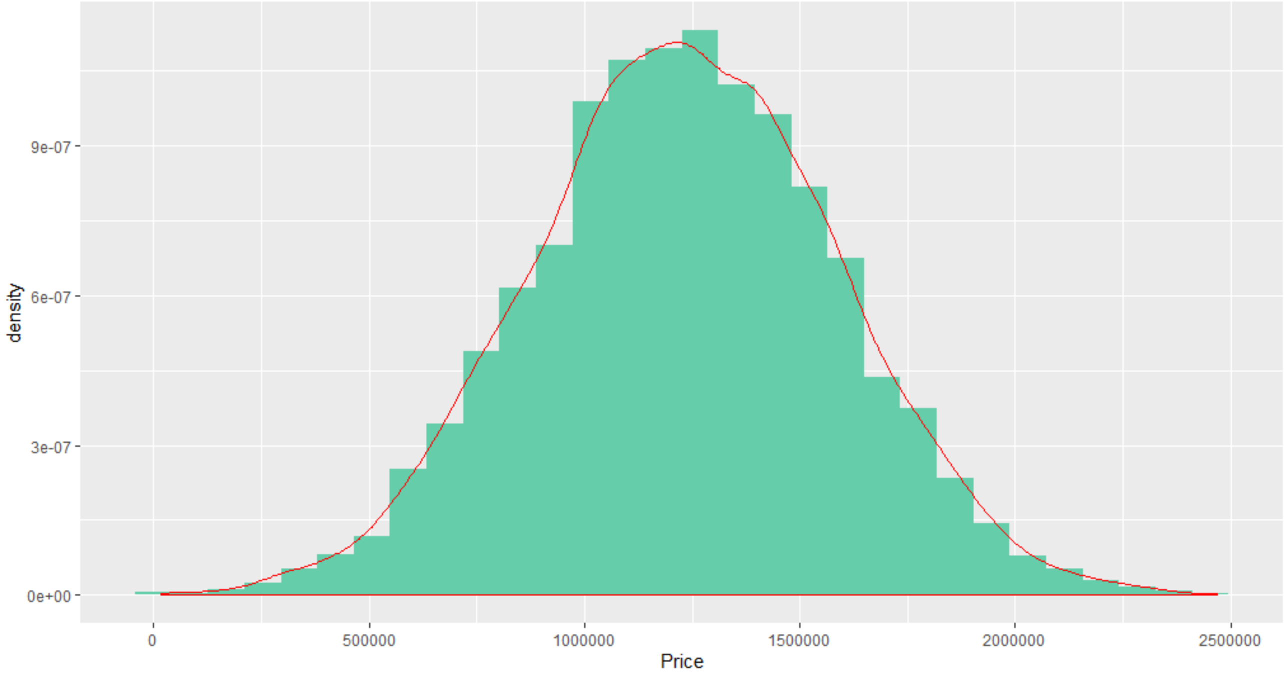
```
Area_Income Area_House_Age Area_Number_of_Rooms Area_Number_of_Bedrooms Area_Population  Price
1       79545.            5.68                 7.01                     4.09       23087. 1059034.
2       79249.            6.00                 6.73                     3.09       40173. 1505891.
3       61287.            5.87                 8.51                     5.13       36882. 1058988.
4       63345.            7.19                 5.59                     3.26       34310. 1260617.
5       59982.            5.04                 7.84                     4.23       26354.  630943.
6       80176.            4.99                 6.10                     4.04       26748. 1068138.
```

- This data is downloaded from Kaggle,
  - https://www.kaggle.com/aariyan101/usa-housingcsv
- We need to predict the price of the houses based on the other variables in the table.
- Can we use the linear regression to do so?

LINEAR REGRESSION, EXAMPLE (2)

- Inspect the distribution shape of the response variable by plotting its histogram and showing the density.

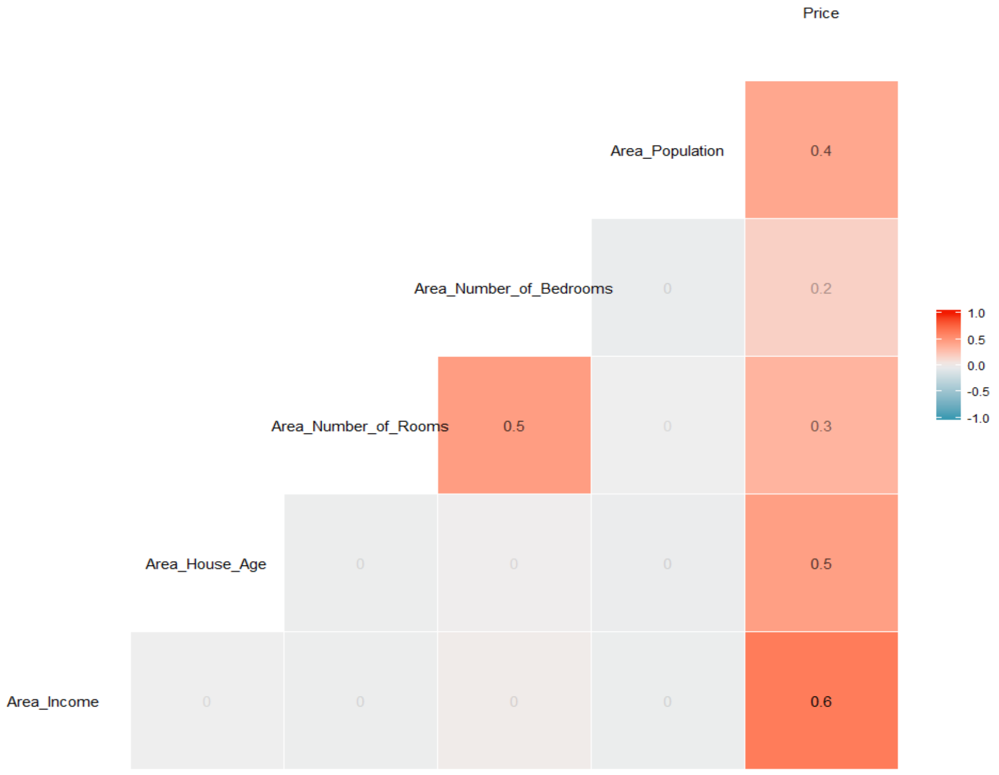
It is good that the response variable follows a normal distribution. This means the normality assumption holds.



LINEAR REGRESSION, EXAMPLE (3)

- Check the correlation matrix to check the linearity assumptions between the variables.

The correlation coefficients (strength of the relationship) between the explanatory variables and the response variables look moderate except for number_of_bedroom.



Co-Linearity

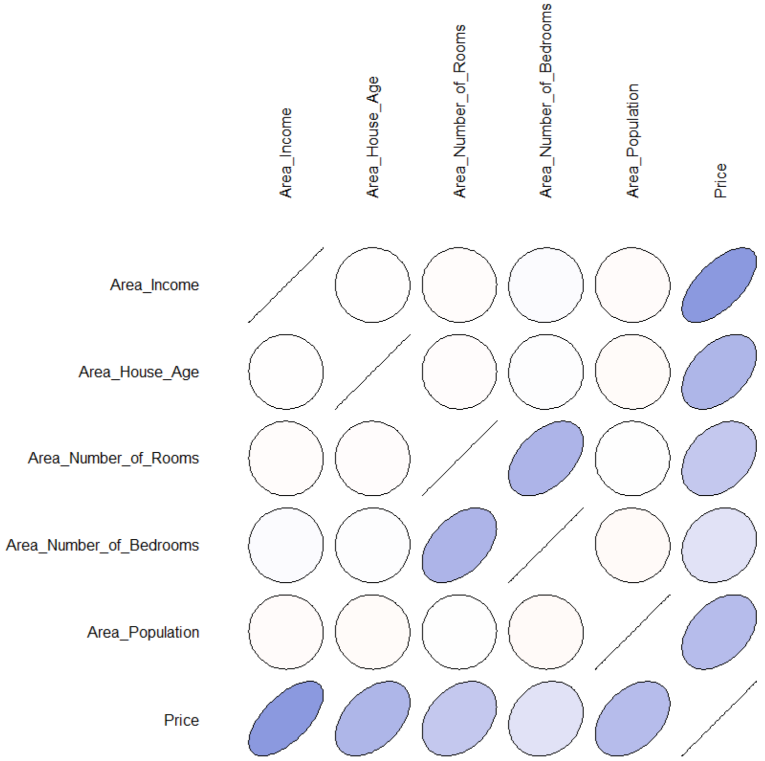
Linearity

LINEAR REGRESSION, EXAMPLE (4)

UNIVERSITY OF CANBERRA

- Check the correlation shapes to check the direction of the linear relationship between variables.

The correlation shape (direction of the relationship) between the explanatory variables and the response variables look mostly linear except for number_of_bedroom variable.



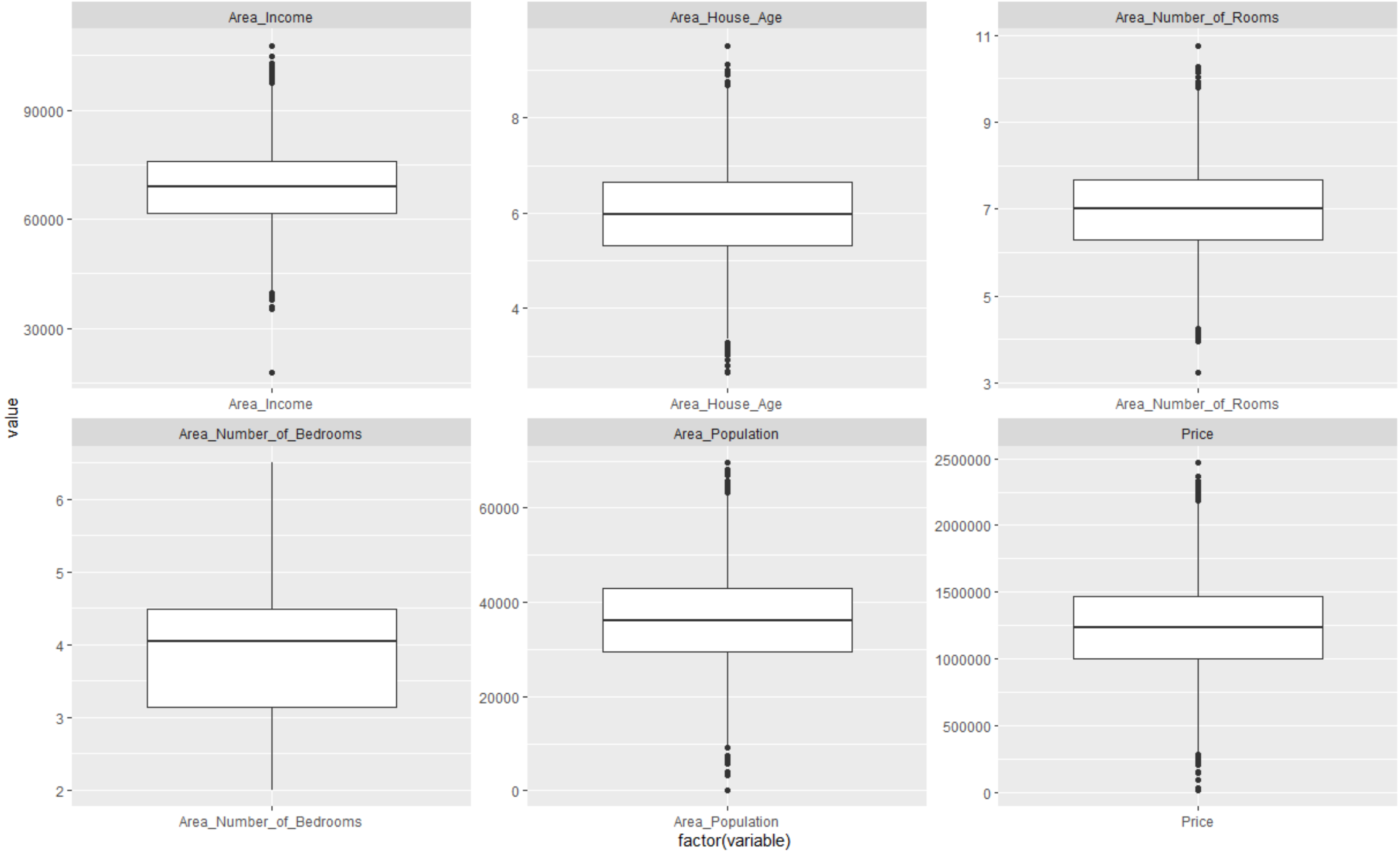
Co-Linearity

Linearity

LINEAR REGRESSION, EXAMPLE (5)

- Checking the outliers in the data

Most of the variables are obtaining outliers, except number_of_bedrooms variable.



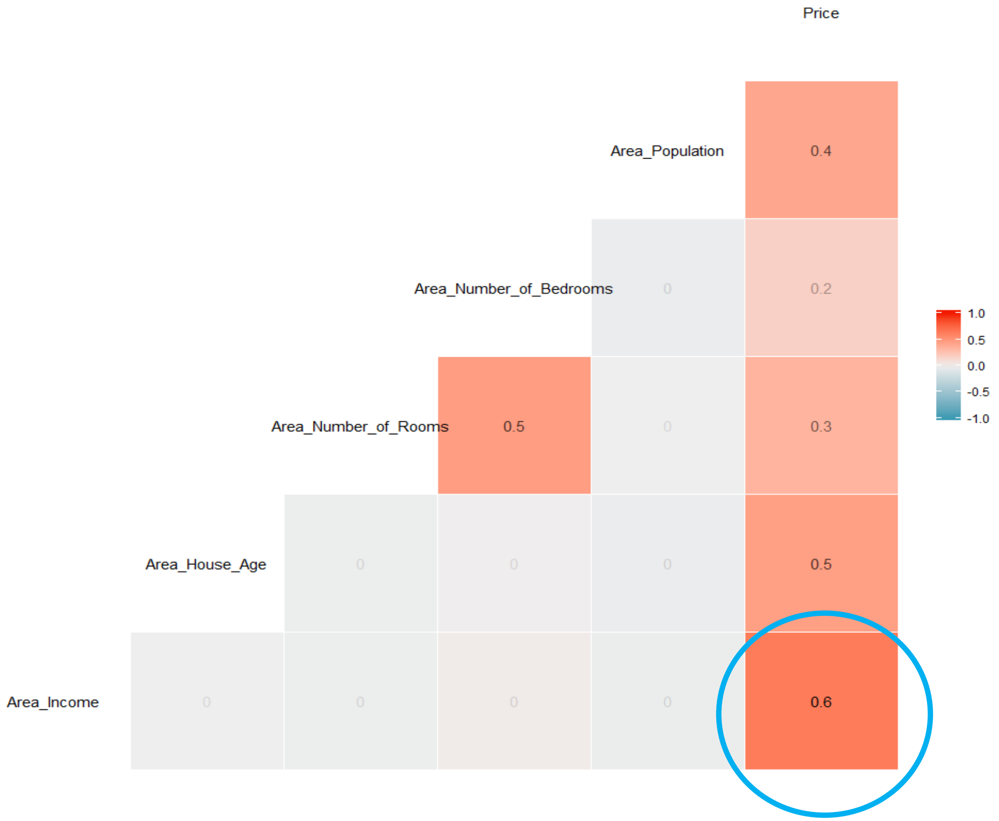
LINEAR REGRESSION, EXAMPLE (6)

- Building the model:
  - First, we need to split the data into two partitions, training and testing.
  - , then, training a model on the training set, while conducting the evaluation on the testing set.

```
library(caret)
index <- createDataPartition(data$Price, p = .70, list = FALSE)
train <- data[index, ]
test <- data[-index, ]
```

LINEAR REGRESSION, EXAMPLE (7)

- The First model will be created to predict the **Price** of the houses from the **Area_Income** variable.
- We have chosen this variable, because it shows the highest correlation coefficient with the predicted variable (i.e. Price)



LINEAR REGRESSION, EXAMPLE (8)

- Building the model:

```
lmModel1 <- lm(Price ~ Area_Income , data = train)
summary(lmModel1)
```

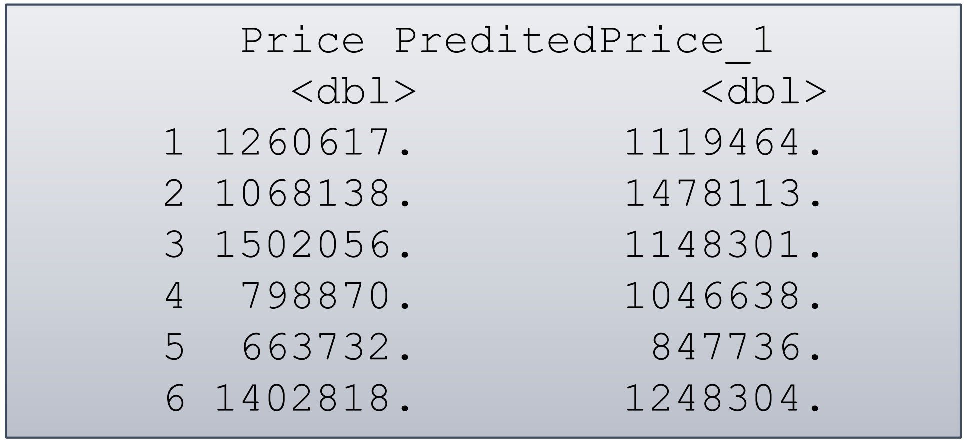
```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.304e+05  3.036e+04  -7.589  4.1e-14 ***
Area_Income  2.131e+01  4.375e-01  48.703  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 271900 on 3498 degrees of freedom
Multiple R-squared:  0.4041, Adjusted R-squared:  0.4039
F-statistic:  2372 on 1 and 3498 DF,  p-value: < 2.2e-16
```

Here, the R-Squared and adj R-Squared indicate that the explanatory variable can explain ~40% of the variation in the response variable.

LINEAR REGRESSION, EXAMPLE (9)

- Evaluating the model on the testing data:

```
test$PreditedPrice_1 <- predict(lmModel1, test)
head(test, c(“Price”, “PredictedPrice_1”)
```



```
       Price  PreditedPrice_1
       <dbl>            <dbl>
1 1260617.         1119464.
2 1068138.         1478113.
3 1502056.         1148301.
4  798870.         1046638.
5  663732.          847736.
6 1402818.         1248304.
```

R-squared for testing data = 0.4040877

Residual Mean Square Error (RMSE) = 271847.8

From these numbers, the built model looks **weak** in predicting the prices of the houses only from the average of the area income.

LINEAR REGRESSION, EXAMPLE (10)

- Building a multi-linear regression model by selecting all the variables (except Price) as the explanatory variables:

```
lmModel2 <- lm(Price ~ . , data = train)
summary(lmModel2)
```

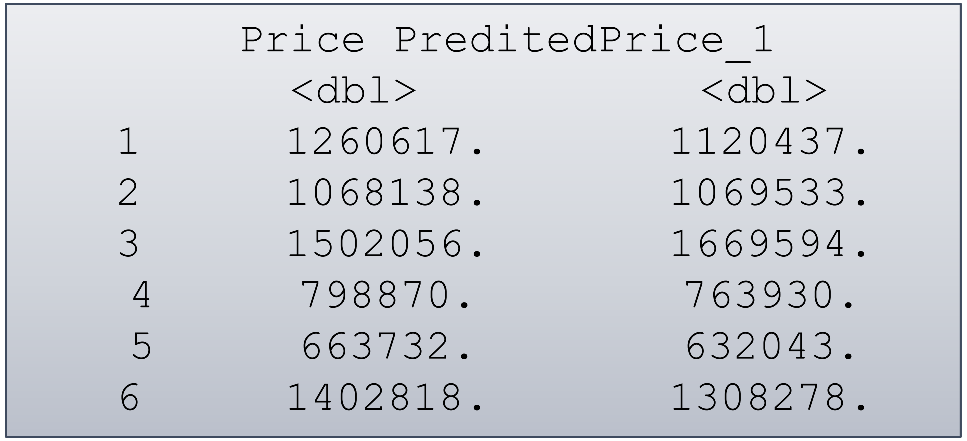
```
Coefficients:
                Estimate Std. Error  t value Pr(>|t|)
(Intercept) -2.641e+06  2.049e+04 -128.849   <2e-16 ***
Area_Income   2.155e+01  1.628e-01  132.347   <2e-16 ***
Area_House_Age 1.654e+05  1.733e+03   95.441   <2e-16 ***
Area_Number_of_Rooms 1.207e+05  1.929e+03   62.582   <2e-16 ***
Area_Number_of_Bedrooms  3.431e+03  1.566e+03    2.191   0.0285 *
Area_Population  1.519e+01  1.729e-01   87.870   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 101200 on 3494 degrees of freedom
Multiple R-squared:  0.9176,     Adjusted R-squared:  0.9175
F-statistic:  7782 on 5 and 3494 DF,  p-value: < 2.2e-16
```

Here, the R-Squared and adj R-Squared indicate that the explanatory variables can explain ~92% of the variation in the response variable.

LINEAR REGRESSION, EXAMPLE (11)

- Evaluating the model on the testing data:

```
test$PreditedPrice_2 <- predict(lmModel2, test)
head(test, c(“Price”, “PredictedPrice_2”)
```



```
         Price  PreditedPrice_1
         <dbl>            <dbl>
1     1260617.         1120437.
2     1068138.         1069533.
3     1502056.         1669594.
 4     798870.          763930.
 5     663732.          632043.
6     1402818.         1308278.
```

R-squared for testing data = 0.917601

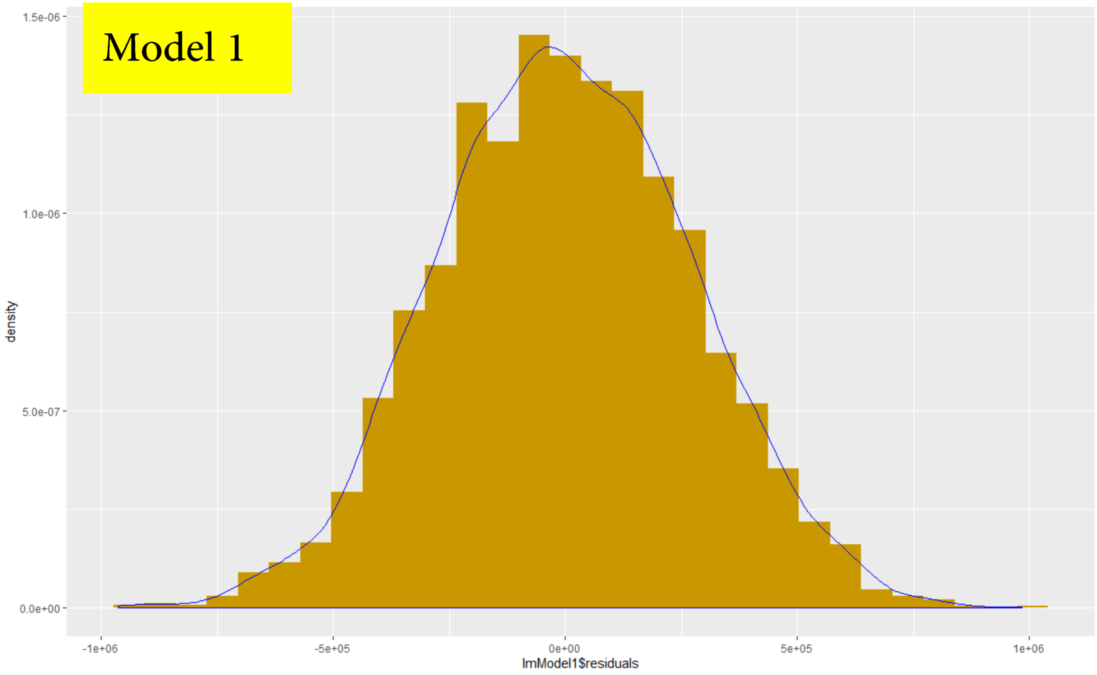
Residual Mean Square Error (RMSE) = 101087.1

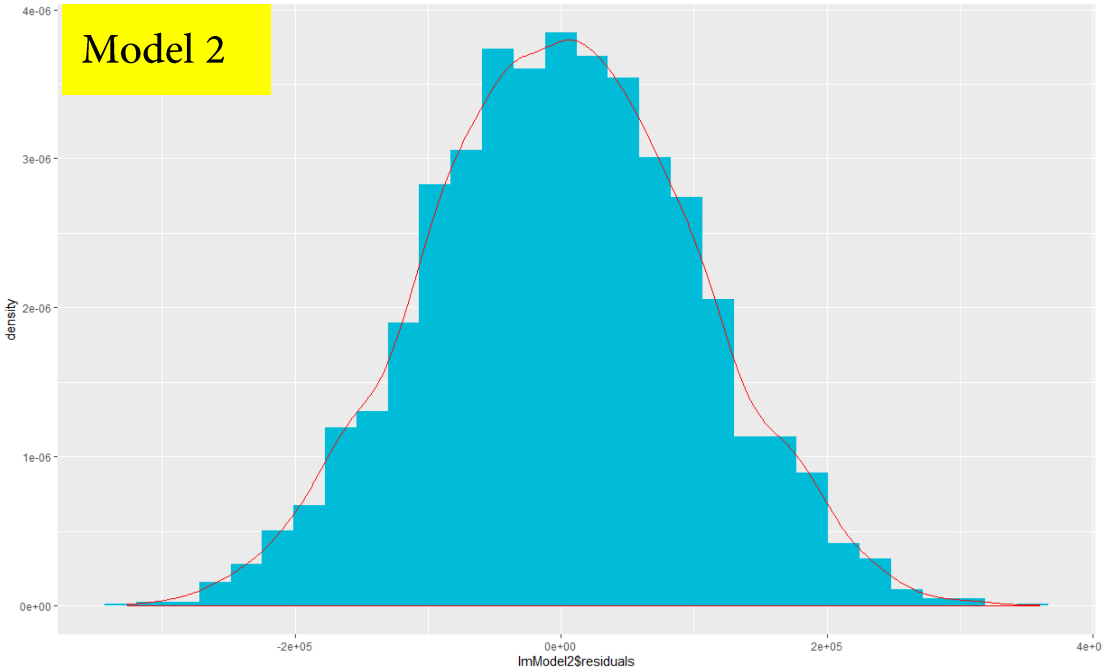
From these numbers, the built model looks **strong** in predicting the prices of the houses from all the explanatory variables.

LINEAR REGRESSION, EXAMPLE (12)

UNIVERSITY OF CANBERRA

- Checking the post-assumptions of the built models
  - Errors follow normal distributions:

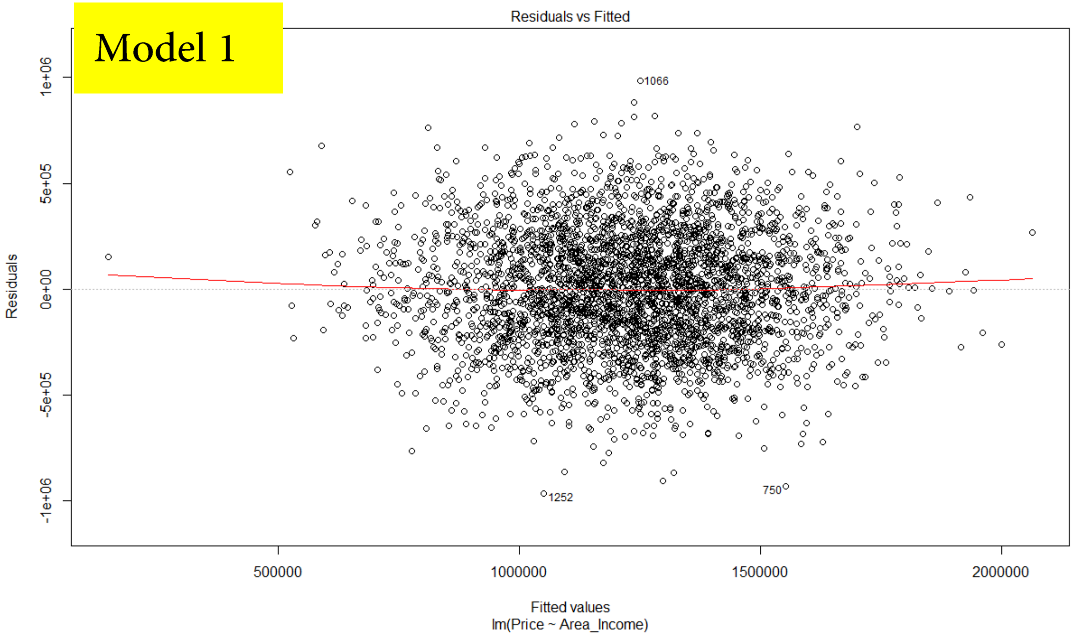


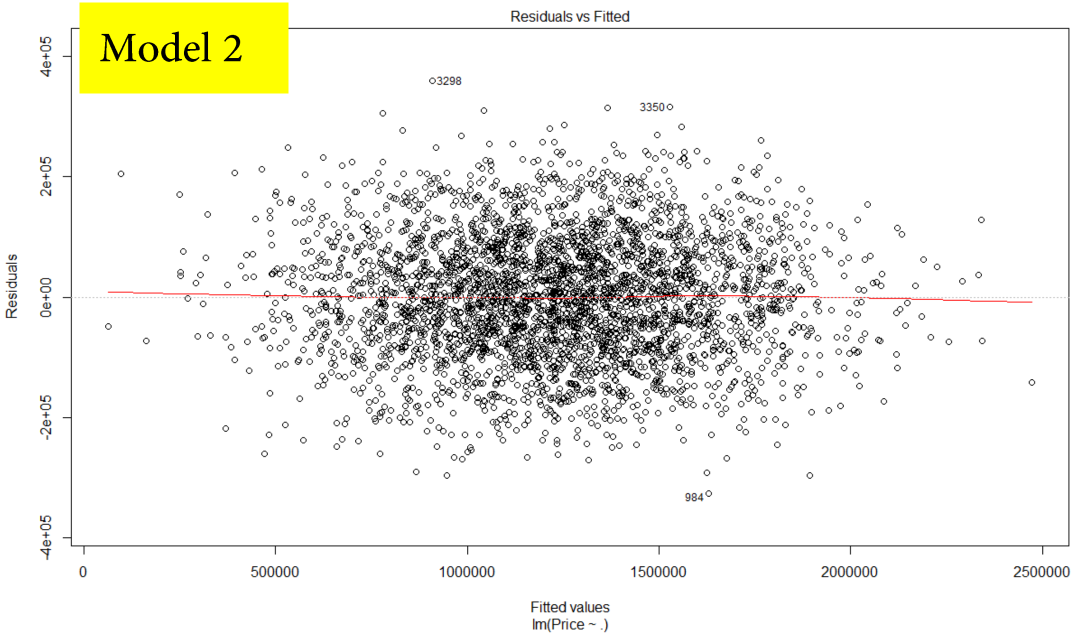


Both residuals of the two models are normally distributed.

LINEAR REGRESSION, EXAMPLE (13)

- Checking the post-assumptions of the built models
  - The variance of the errors is constant, i.e. no pattern on the residual vs. fitted scatter plot:

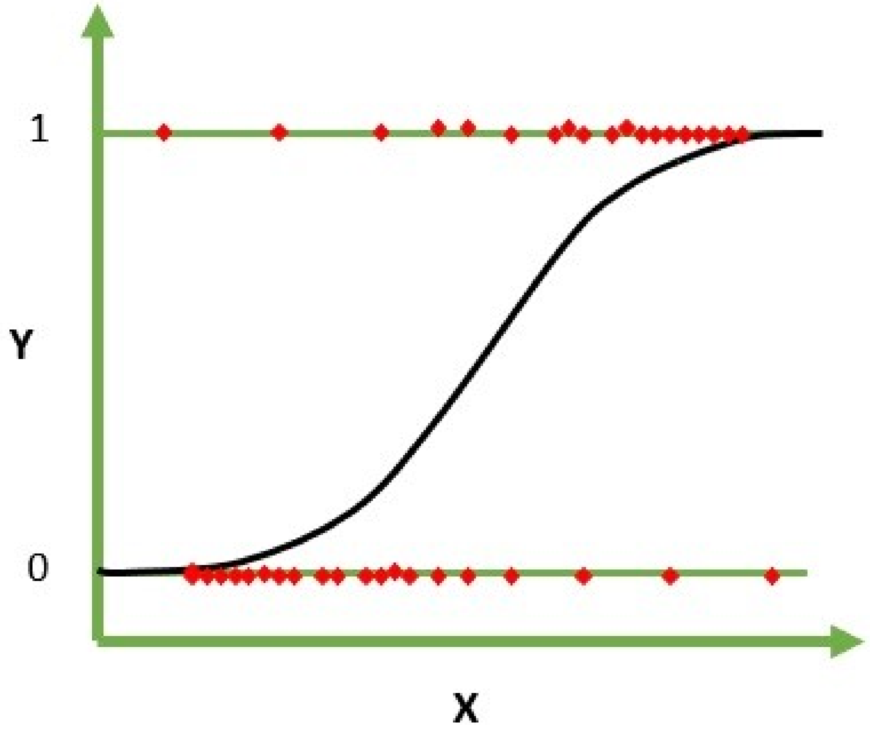




Model 2 looks more stable with respect to the residuals vs. the fitted values.

NEXT STEPS

- Where to go from here in data modelling?
  - What about the classification?
    - Logistic regression
    - Support vector machine
    - Neural Network
    - …



NEXT STEPS (2)

- Where to go from here in data modelling?
  - What about the unsupervised learning?
    - K-means clustering
    - Principal Component Analysis
    - Auto-Encoder
    - …

ANNOUNCEMENT

- Final assessment component is a take-home assignment
- Similar to the first assignment
- It is going to be available to you via Canvas from 7th of May 5:00 pm until 14th of May 5:00 pm
- Mandatory assessment component, you must score at least 50% to pass the unit