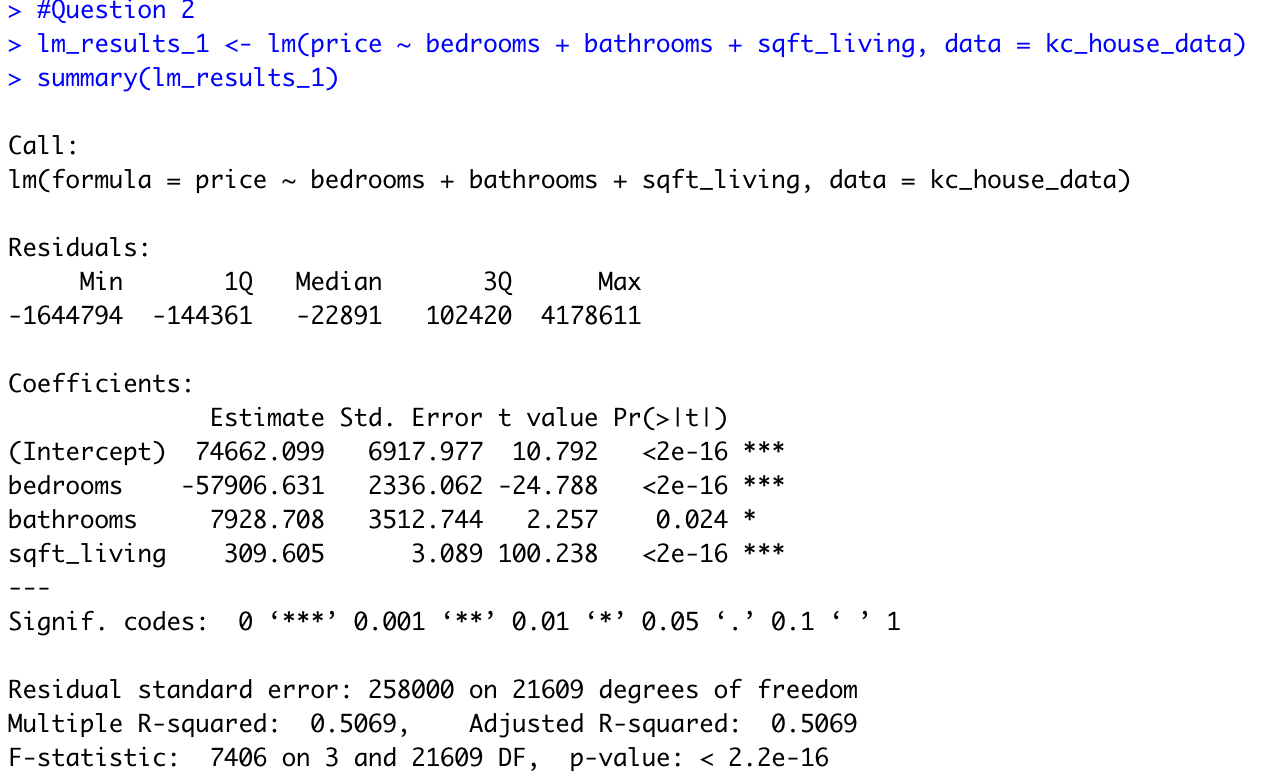
Symphony Hopkins

DSCI 502: R Programming

Professor Xiaobo Li

November 23, 2022

Project 6

1. Load the dataset in kc\_house\_data.csv into R. Call the loaded data kc\_house\_data. Make sure that you have the directory set to the correct location for the data.
   1. See R script.
2. Build a linear model to forecast the **price** using **bedrooms**, **bathrooms** and **sqft\_living**.
   1. Then write down the corresponding math formula.
      1. price = 74662.099 - 57906.631 \* bedrooms + 7928.708 \* bathrooms + 309.605 \* sqft\_living
   2. Is it a good model based on R square or adjusted R square?
      1. No, this is not a good model.
3. Build a linear model to forecast the **price** using **bedrooms**, **bathrooms** and **sqft\_living** and all the **cross effects** between them.Text

   Description automatically generated
   1. Then write down the corresponding math formula.
      1. price = 4.339e+05 - 4.670e+04 \* bedrooms - 1.203e+05 \* bathrooms - 2.490e+01 \* sqft\_living - 4.825e+03 \* bedrooms:bathrooms + 3.227e+01 \* bedrooms:sqft\_living + 1.121e+02 \* bathrooms:sqft\_living - 1.015e+01 \* bedrooms:bathrooms:sqft\_living
   2. Is it a better model than the model in Question 2 based on adjusted R square?
      1. Because the adjusted R square increased, we can say that this is a better model than the previous model.
4. Build a linear model to forecast the **price** using **bedrooms**, **bathrooms** and **sqft\_living,** **waterfront** and **grade**.Text

   Description automatically generated with medium confidence
   1. Then write down the corresponding math formula.
      1. price = - 4.883e+05 - 3.198e+04 \* bedrooms - 2.554e+04 \* bathrooms + 2.134e+02 \* sqft\_living + 7.991e+05 \* waterfront + 9.669e+04 \* grade
   2. Is it a better model than the model in Question 3 based on adjusted R square?
      1. Because the adjusted R square increased, we can say that this is a better model than the previous model.
5. Build a linear model to forecast the **price** using all other columns except **id**, **date**, **zipcode**, **lat**, and **long without** a y-intercept. If we only consider the models defined in Q2, Q3, Q4 and Q5, which model do you recommend based on the adjusted R squared value?Text

   Description automatically generated
   1. Based on the R square values, this (Q5) is currently the best model compared to the previous models (Q2, Q3, and Q4).

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **bedrooms** | **bathrooms** | **sqft\_living** | **sqft\_lot** | **floors** | **waterfront** | **view** | **condition** | **grade** |
| 4 | 2 | 2560 | 7650 | 1.5 | 1 | 3 | 5 | 10 |

1. You are asked to build a linear model to forecast **price** using **bedrooms, bathrooms, sqft\_living, sqft\_lot, floors, waterfront, view, condition,** and **grade**. Then you are given the flowing new house info:

Text

Description automatically generated with medium confidenceText

Description automatically generated

* 1. Predict the average sales price for this house.
     1. The predicted average sales price for this house is $1,689,762.
  2. Predict the 95% predicted interval for this house.
     1. The 95% predicted interval for this house is between $1,235,440 and $2,144,084.