Multiple Linear Regression Analysis

Eric Wu

2023-07-16

Background

The data set “housing” is from Kaggle.com, an online community with enriched data sets from various contributors for data analysis and machine learning projects. The primary owner of the “housing” data set is Ashish.

Figure 1: (Cited from https://opendatascience.com/10-tips-to-get-started-with-kaggle/)

Skill Set Objectives

- Multiple Regression Analysis
- Backward Elimination
- Perform Cross Validations on the model

Dataset

The data set contains the data of 545 houses with 13 variables including the price, area(square feet), number of bedrooms/bathrooms, stories(floors without basement), whether it's close to the main road/preferred neighborhood, has a guest room/basement/hot water heating/air conditioning, number of parking spaces, and furnish status. The data set also has 0 null values.

Table 1: Basic statistics of quantitative variables

	vars	n	mean	sd	min	max	range	se
price	1	545	4766729.25	1870439.62	1750000	13300000	11550000	80120.83
area	2	545	5150.54	2170.14	1650	16200	14550	92.96
bedrooms	3	545	2.97	0.74	1	6	5	0.03
bathrooms	4	545	1.29	0.50	1	4	3	0.02
stories	5	545	1.81	0.87	1	4	3	0.04
parking	6	545	0.69	0.86	0	3	3	0.04

Table 1 demonstrated some simple statistics of the quantitative variables within the data set. Based on the table, there's a huge difference in scale between the price and other variables. Which showed a need for scaling the data.

Research Question

Here are some research questions to begin with:

1. Which model is selected to be the "best" model using backward-elimination?
2. Which predictor is the most influential?
3. What's the estimated price for a house with 7000 square feet, 3 bedrooms, 3 bathrooms, 1 stories, next to main road, doesn't have a guestroom, have a basement, hot water heating, air conditioning, 1 parking space, at a preferred neighborhood, and is furnished?

Implementing Multiple Linear Regression

Before implementing the multiple linear regression, it's important to know that the categorical variable can also an effect on the prediction of the pricing of the rent. Therefore, it's crucial to create indicator variables to evaluate the effect of all the categorical variables in the dataset.

Table 2: Scaled statistics of quantitative variables

	vars	n	mean	sd	min	max	range	se
price	1	545	15.31	0.37	14.38	16.40	2.03	0.02
area	2	545	8.47	0.40	7.41	9.69	2.28	0.02
bedrooms	3	545	2.97	0.74	1.00	6.00	5.00	0.03
bathrooms	4	545	1.29	0.50	1.00	4.00	3.00	0.02
stories	5	545	1.81	0.87	1.00	4.00	3.00	0.04
parking	6	545	0.69	0.86	0.00	3.00	3.00	0.04

Table 2 demonstrated the statistics of scaled data, primarily on the price and area variable to prevent the difference in scaling.

The full model is a great model to start with since the effect of each variables is not pre-known, and the coefficient of the model are listed in the following

```
##          (Intercept)                 area             bedrooms 
##               12.004                0.302                0.034 
##            bathrooms              stories              parking 
##                0.165                0.091                0.047 
##         mainroad_var        guestroom_var         basement_var 
##                0.106                0.056                0.103 
##  hotwaterheating_var  airconditioning_var          prefarea_var 
##                0.179                0.162                0.131 
## furnishingstatus_var 
##                0.039
```

The “var” variables are the indicator of the original categorical variables, and the data is scaled to ensure the possibility in comparing different predictor variables. In addition to the coefficients, the Rˆ2 value of the full model are 69%. The Rˆ2 value of 69% indicates that the model explains 69% of the total variation in the price variable.

```
## Linear Regression 
## 
## 545 samples
##  12 predictor
## 
## No pre-processing
## Resampling: Cross-Validated (10 fold) 
## Summary of sample sizes: 490, 492, 490, 490, 489, 492, ... 
## Resampling results:
## 
##   RMSE       Rsquared   MAE      
##   0.2108718  0.6876142  0.1634109
## 
## Tuning parameter 'intercept' was held constant at a value of TRUE
```

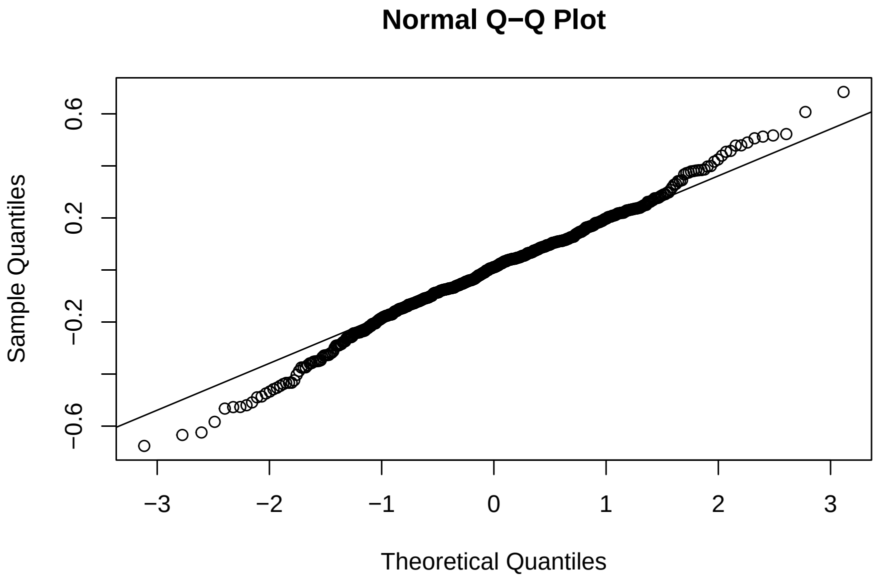
Here are the results of the 10-fold cross validation, with a RMSE of 0.2108718.

Testing Linear Assumptions

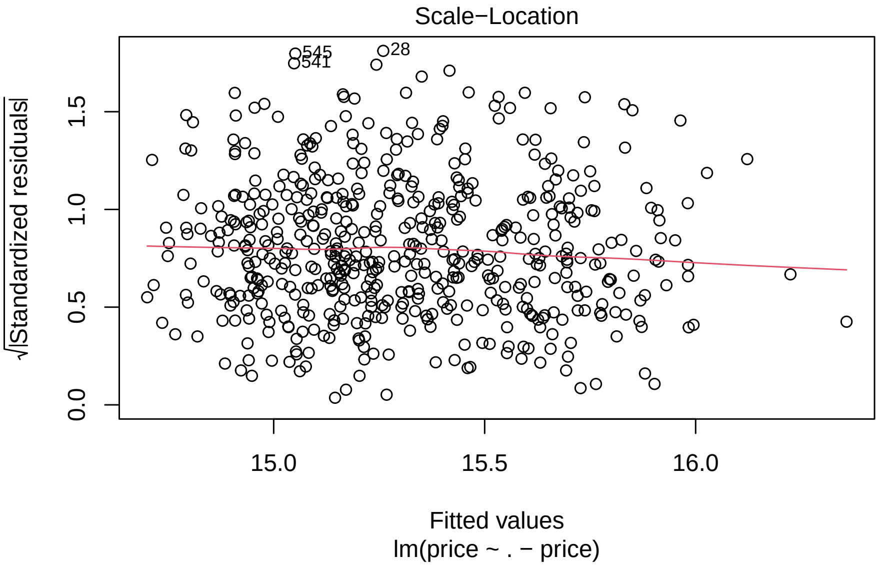
Before performing model selection, it’s important to check for the linear assumptions

```
## 
##  Shapiro-Wilk normality test
## 
## data:  full_model$residuals
## W = 0.99249, p-value = 0.007704
```

First we can check the normality of the residuals. Based on the result from the Shapiro-wilk normality test, we reject the null hypothesis and concluded that the residual does not come from a normal distribution.



The above QQ Plot suggested a similar conclusion and it demonstrates that there's extremes or outliers in the data set.



Next to check is the constant variance of residuals. Based on the graph, the variance of the residuals are not constant either.

Backward-Elimination

Next step is to perform backward-elimination using the step function from the stats library to determine the best model by the Alkaika Information Criterion (AIC) that penalized the likelihood by the number of terms in the model.

```
## Start:  AIC=-1688.21
## price ~ (area + bedrooms + bathrooms + stories + parking + mainroad_var +
##     guestroom_var + basement_var + hotwaterheating_var + airconditioning_var +
##     prefarea_var + furnishingstatus_var) - price
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               23.463 -1688.2
## - furnishingstatus_var  1    0.1466 23.610 -1686.8
## - guestroom_var         1    0.2066 23.670 -1685.4
## - bedrooms              1    0.2540 23.717 -1684.3
## - mainroad_var          1    0.6218 24.085 -1676.0
## - parking               1    0.7341 24.198 -1673.4
## - hotwaterheating_var   1    0.7355 24.199 -1673.4
## - basement_var          1    1.0069 24.470 -1667.3
## - prefarea_var          1    1.4673 24.931 -1657.2
## - stories               1    2.2783 25.742 -1639.7
## - airconditioning_var   1    2.5085 25.972 -1634.9
## - bathrooms             1    2.9066 26.370 -1626.6
## - area                  1    5.6600 29.123 -1572.4
```

```
##
## Call:
## lm(formula = price ~ (area + bedrooms + bathrooms + stories +
##     parking + mainroad_var + guestroom_var + basement_var + hotwaterheating_var +
##     airconditioning_var + prefarea_var + furnishingstatus_var) -
##     price, data = housing_prep)
##
## Coefficients:
##          (Intercept)                  area              bedrooms
##             12.00384               0.30191               0.03420
##            bathrooms               stories               parking
##              0.16497               0.09062               0.04688
##         mainroad_var         guestroom_var          basement_var
##              0.10609               0.05635               0.10349
##  hotwaterheating_var   airconditioning_var          prefarea_var
##              0.17892               0.16166               0.13067
## furnishingstatus_var
##              0.03869
```

Surprisingly, the best model selected by the backward-elimination is the full model, which suggested that it's the most accurate model that uses minimum possible number of predictors in AIC standard. Based on the result, the most influential predictor suggested by the model is the "area" variable because it has the largest coefficient and the data set is scaled to make sure the comparison are valid. Moreover, the predicted price of the third research question according to the model are $\$1.0133951 \times 10^{7}$ dollars.

Conclusion

This rmd file discussed and analyzed the possible usage of a multiple linear regression model on a housing price prediction data, including model selections using backward-elimination method as will as tested some of the assumptions of multiple linear regression. It's after testes that the data set doesn't fit many of the linear model assumptions, but given that the full model has a residual of 69%, and that means we can still use the multiple regression model for the estimation of housing prices and the average rmse of the full model are 0.2108718 which suggest an average difference of 21.0871776%.

Code Appendix

```
knitr::opts_chunk$set(
  echo = FALSE,
  warning = FALSE,
  fig.pos = 'H')

rm(list = ls()) # Clean environment
# Load libraries
library(psych)
library(tidyverse)
library(kableExtra)
library(modelr)
library(stats)
library(esquisse)
library(float)
library(caret)

housing <- read.csv("housing.csv")

quanti_vars <- c("price", "area", "bedrooms", "bathrooms", "stories", "parking")
summary_stat <- psych::describe(housing[quanti_vars], skew = FALSE) %>%
  round(2)

summary_stat %>%
  kable(
    caption = 'Basic statistics of quantitative variables',
    booktabs = TRUE,
    align = c('l', rep('c', 8))
  ) %>%
  kableExtra::kable_styling(
    bootstrap_options = c('striped', 'condensed'),
    font_size = 10,
    latex_options = "hold_position"
  )

# Scale the data
testing <- housing %>%
  mutate(
    price = log(price),
    area = log(area)
  )

summary_stat2 <- psych::describe(testing[quanti_vars], skew = FALSE) %>%
  round(2)

summary_stat2 %>%
  kable(
    caption = 'Scaled statistics of quantitative variables',
    booktabs = TRUE,
    align = c('l', rep('c', 8))
  ) %>%
  kableExtra::kable_styling(
```

```r
    bootstrap_options = c('striped', 'condensed'),
    font_size = 10,
    latex_options = "HOLD_position"
  )

# Convert categorical to indicator variables
testing <- testing %>%
  mutate(
    mainroad_var = ifelse(mainroad == "yes", 1, 0),
    guestroom_var = ifelse(guestroom == "yes", 1, 0),
    basement_var = ifelse(basement == "yes", 1, 0),
    hotwaterheating_var = ifelse(hotwaterheating == "yes", 1, 0),
    airconditioning_var = ifelse(airconditioning == "yes", 1, 0),
    prefarea_var = ifelse(prefarea == "yes", 1, 0),
    furnishingstatus_var = ifelse(furnishingstatus == "furnished", 1, 0) # Set other level of furnished
  )

# Create a data set to store all variables except categorical variable
housing_prep <- testing %>%
  select(-mainroad, -basement, -guestroom, -hotwaterheating, -airconditioning, -prefarea, -furnishingsta

# 10-fold Cross Validation
train_control <- trainControl(method = "cv",
                              number = 10)

cv_full <- train(price ~ .-price, data = housing_prep,
               method = "lm",
               trControl = train_control)

full_model <- lm(price ~ .-price, data = housing_prep) # Create a full model
round(full_model$coefficients, 3)

print(cv_full)
shapiro.test(full_model$residuals)

qqnorm(full_model$residuals)
qqline(full_model$residuals)

plot(full_model, 3)

int_only_model <- lm(price ~ 1, data = housing_prep)
stats::step(object = full_model,
            scope = list(lower = int_only_model, upper = full_model),
            data = housing_prep,
            direction = "backward")

# Created a scaled data frame
housing_prep[546, ] <- list(0, log(7000), 3, 3, 1, 1, 1, 0, 1, 1, 1, 1, 1)
prediction <- housing_prep[546, ]

# Calculated the prediction
price <- predict(full_model, newdata = prediction)
```