# House Price Analysis and Prediction

MINH CHAU

```r
# load required packages
library(dplyr)
library(tidyr)
library(car)
library(ggplot2)
```

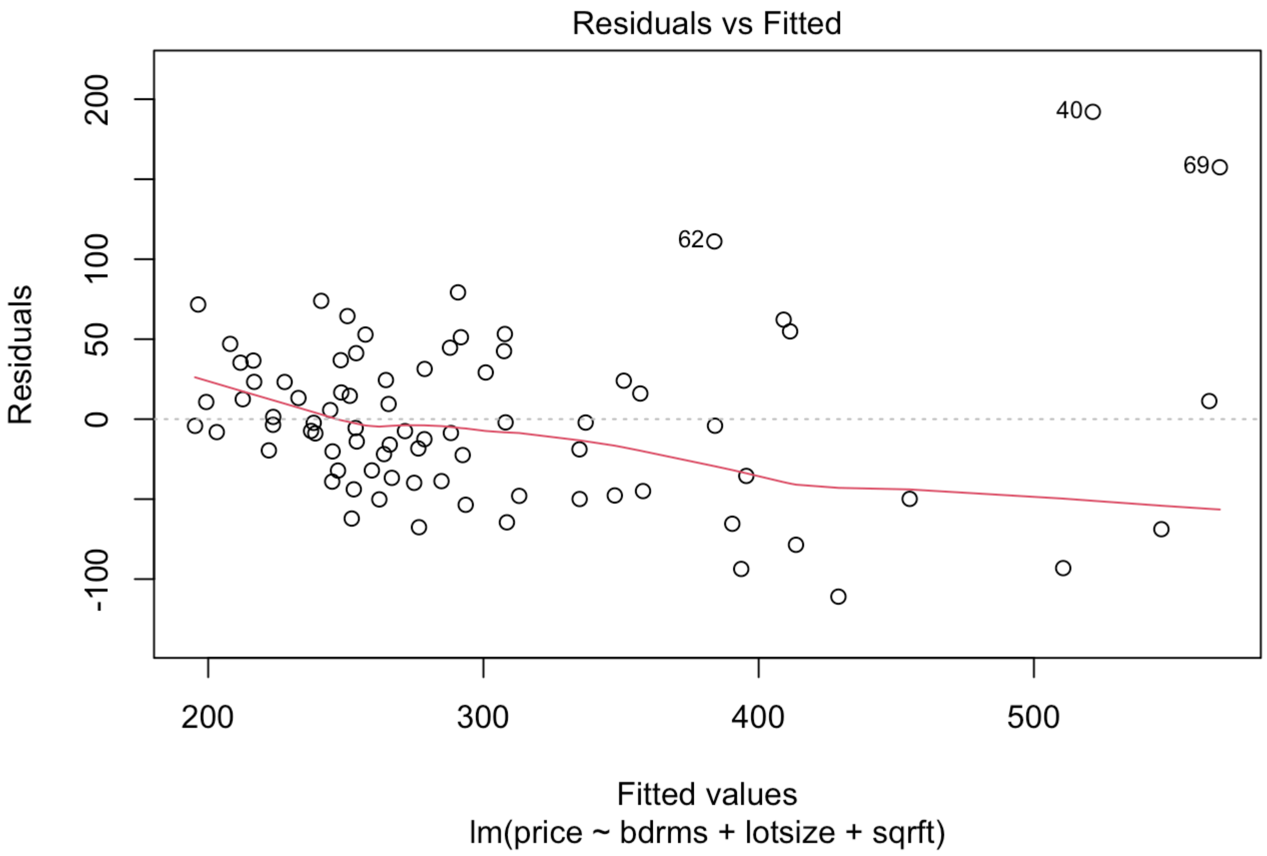
- Perform linear regression analysis on a dataset of homes that were sold in Boston, MA area.
  - **Dataset**: `housingprice.csv`
- There are 10 attributes in each case of the dataset. The description of the variables are :
  - **price:** house price, measured in thousands of dollars, $1000s
  - **assess**: assessed value of the house, $1000s
  - **bdrms**: number of bedrooms in the house
  - **lotsize**: size of the lot in square feet
  - **sqrft**: size of the house in square feet
  - **colonial**: =1 if the home is colonial style, 0 otherwise
  - **lprice**: log(price)
  - **lassess**: log(assess)
  - **llotsize**: log(lotsize)
  - **llotsize**: log(lotsize) **Note**: The term “log” refers to the natural logarithm. If the answer is greater than 1, round off to 2 decimal places. If the answer is less than 1, round off to 3 significant numbers. When rounding, also take note of the natural round points, for example, costs in dollars to round off to 2 decimal places.

**Purpose : Understand multiple linear regression and how to interpret its results.**

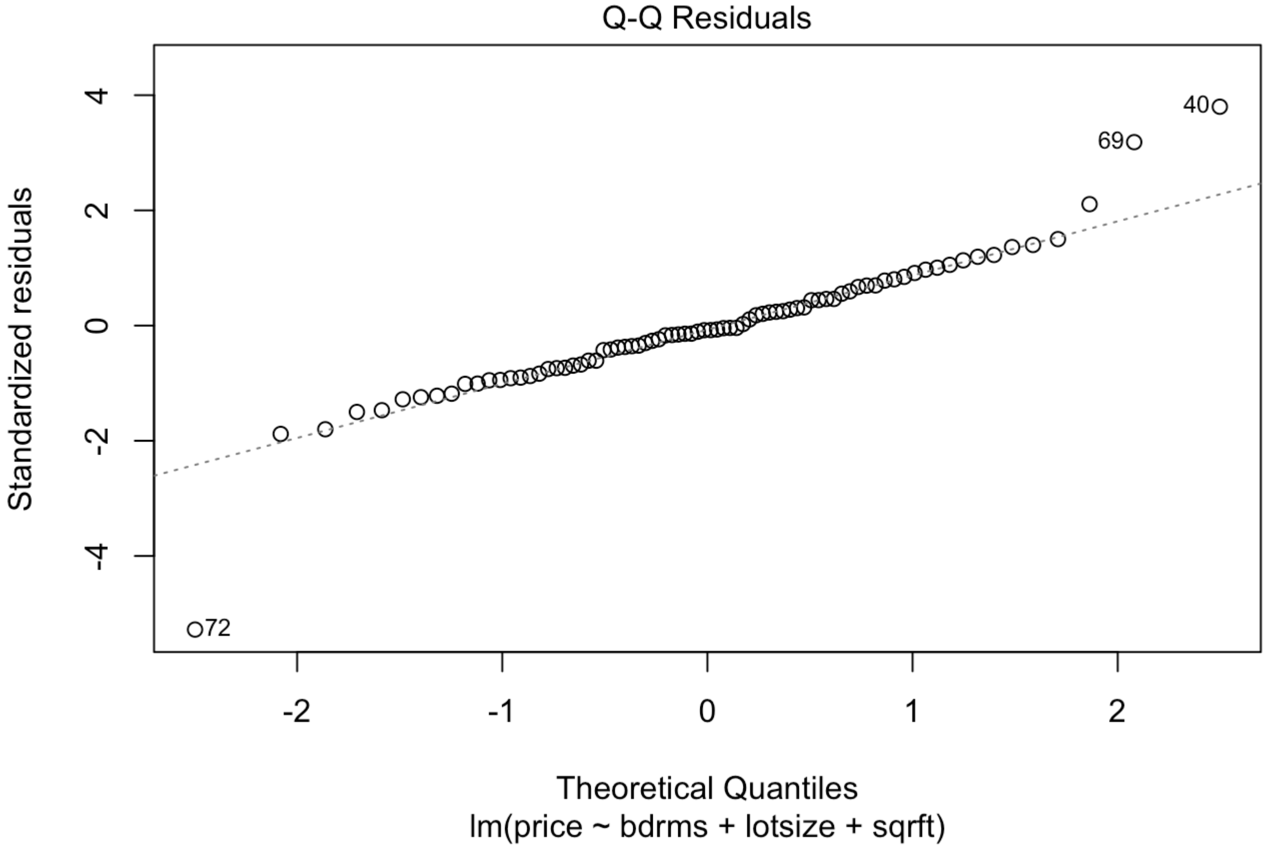
***Part 1.* Fitting a multiple linear regression to predict price using the predictors: bdrms, lotsize and sqrft.**

1. Fit a multiple linear regression model, *mod.multivariate*, where *price* as the response variable and *bdrms, lotsize*, and *sqrft* are the independent variables. Provide the fitted regression equation in mathematical form.

```r
HP <- read.csv("/Users/minhchau/Downloads/housingprice.csv")
mod.multivariate <- lm(price ~ bdrms + lotsize + sqrft, data = HP)
plot(mod.multivariate, 1) #For residual plot
```



```
plot(mod.multivariate, 2) #Plot the QQ plot
```



```
shapiro.test(residuals(mod.multivariate))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(mod.multivariate)
## W = 0.96051, p-value = 0.01445
```

```
summary(mod.multivariate)
```

```
##
## Call:
## lm(formula = price ~ bdrms + lotsize + sqrft, data = HP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.972  -37.232   -4.178   29.723  192.103
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.393e+01  2.842e+01  -1.546 0.126251
## bdrms        2.005e+01  8.404e+00   2.385 0.019550 *
## lotsize      1.991e-03  5.747e-04   3.464 0.000877 ***
## sqrft        1.228e-01  1.216e-02  10.094  1.1e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.39 on 76 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7238
## F-statistic: 70.01 on 3 and 76 DF,  p-value: < 2.2e-16
```

The fitted line model:

Y = -4.393e+01 + 2.005e+01 * bdrms + 1.991e-03 * lotsize + 1.228e-01 * sqrft where Y is the model’s predicted price.

We have RSE = 53.39 is not so good considering the price is around 200-400.

R-square value (Multiple): 0.7343, means we can explain 73.43% of data based on this model :)

The residuals is not randomly nor follow any pattern. This is a slight indication of HeteroSkedasticity, although I think it can also be classified residuals as randomly scattered.

The QQ plot have confirmed that the residuals is mostly normal,

The p-value of Shapiro-Wilk normality test also shown that p-value = 0.01445 < 0.05, suggesting the residuals is normally distributed.

2. Explain clearly what the regression coefficient estimate for ’*lotsize’* suggests. (2 marks)

The regression coefficient estimate for lotsize is 1.991e-03. It means when lotsize increase by 1 square feet, we expect to see an increase in the average price by approximately by 1.991 dollar.

3. State the null and alternative hypotheses for investigating the association between the predictor ‘*lotsize’* and the response’*price*’ clearly.

- Perform a hypothesis test at an appropriate level of significance to evaluate the association between ‘lotsize’ and ‘price’.
- Provide a clear conclusion based on your findings.

```
beta.estimate <- summary(mod.multivariate)$coefficients[3]
std.error <- summary(mod.multivariate)$coefficients[7]
t = beta.estimate / std.error
summary(mod.multivariate)
```

```
##
## Call:
## lm(formula = price ~ bdrms + lotsize + sqrft, data = HP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.972  -37.232   -4.178   29.723  192.103
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.393e+01  2.842e+01  -1.546 0.126251
## bdrms        2.005e+01  8.404e+00   2.385 0.019550 *
## lotsize      1.991e-03  5.747e-04   3.464 0.000877 ***
## sqrft        1.228e-01  1.216e-02  10.094  1.1e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.39 on 76 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7238
## F-statistic: 70.01 on 3 and 76 DF,  p-value: < 2.2e-16
```

Null hypothesis: H0: β2 = 0 Alternative hypothesis: H1: β2 != 0 Level of significant: 5% We have p-value = 0.000877 < 0.05 Conclusion: The p value associated with the t-statistic of the predictor “lotsize” is 0.000877 < 0.05 at 5% level of significant, we can reject H0 This indicate a real association of the predictor ‘lotsize’ and price at 5% significant level.

4. We can quantify the uncertainty in the **population slope parameter.** Find and report the 95% confidence interval for the population slope parameter of *‘lotsize’* using the multivariate regression model. Explain confidence interval

```
confint(mod.multivariate, level = 0.95)
```

```
##                     2.5 %      97.5 %
## (Intercept) -1.005341e+02 12.664168838
## bdrms        3.309320e+00 36.786462209
## lotsize      8.463815e-04  0.003135655
## sqrft        9.853050e-02  0.146971928
```

```
#or
t_value <- qt(1-0.05/2, nobs(mod.multivariate)-1-3)
lower <- 1.991e-03 - t_value * 5.747e-04
higher <- 1.991e-03 + t_value * 5.747e-04
print(cbind(lower, higher))
```

```
##               lower      higher
## [1,] 0.0008463858 0.003135614
```

The confidence interval for population slope parameter of lotsize is [8.463e-04, 0.003135] for 95% confidence level.

We achieve the same result with normal calculation: [0.0008464, 0.003136] A slight change is due to rounding.

***Part 2*** **Use multiple linear regression to make prediction, interpret the confidence interval for predictions**

1 A house in the sample has *bedrms* = 3, *lotsize = 5828 squarefeet* and *sqrft = 1715 square feet*. Find the *predicted selling price* for this house using the fitted OLS multivariate model.

The actual selling price of this in the sample is $236,000. Calculate the *residual* for this house.

Provide the the *95% confidence interval* for the price prediction and explain its significance. Buyer underpaid or overpaid for the house?

```
new.data <- data.frame(bdrms = c(3), lotsize = c(5828), sqrft = c(1715))
response_pre = predict(mod.multivariate, newdata = new.data, type = c('response'))
response_pre
```

```
##        1
## 238.3307
```

```
#Confidence interval calculation
#remember newdata = is important, not to mess up with data =
predict(mod.multivariate, newdata = new.data, interval = 'confidence')
```

```
##        fit      lwr      upr
## 1 238.3307 222.9378 253.7237
```

The predicted selling price for this house for given data is 238.3307 in thousands of dollars

The residuals in this cases is the actual price minus the predicted price, which is: 236000 - 238330.7 = -2330.7$

The confidence interval for the price prediction for 95% of confidence level is [222.9378, 253.7237] (in thousand of dollars).

The average price for the house that have bedrms = 3, lotsize = 5828 square feet and sqrft = 1715 square feet is 238.3307 thousand dollars with a 95% confidence interval of [222.9378, 253.7237] (in thousand of dollars).

The buyer buy the house within the expected range, so this suggest the buyer paid a fair price, neither overpaid nor underpaid significant

***Part 3***

**Examine the ANOVA table and calculate the components of variability. Derive the goodness-of-fit F-statistic from the ANOVA table.**

1. An ANOVA table is used to summarize the components of variability in a dataset. It helps in understanding how much of the total variation in the data can be attributed to different sources.

Using the anova() function in R, generate the ANOVA table for the fitted model mod.multivariate. - Identify and report the *Sum of Squares for Regression (SSR)* and *Sum of Squares for Error (SSE)* from the ANOVA table. - Compute the F-statistic using the values from the ANOVA table. Show all steps involved in the calculation.

```
anova(mod.multivariate)
```

```
## Analysis of Variance Table
##
## Response: price
##           Df Sum Sq Mean Sq F value    Pr(>F)
## bdrms      1 245404  245404  86.108 3.900e-14 ***
## lotsize    1  62793   62793  22.033 1.163e-05 ***
## sqrft      1 290370  290370 101.886 1.103e-15 ***
## Residuals 76 216597    2850
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#SSE = 216597
#SSR = 245404 + 62793 + 290370 = 598567
SSE = 216597
SSR = 598567
p = 3 #number of features or predictors
#Can take degree of freedom of residuals cause it the same as n - p - 1. n is the
#number of observation
f = (SSR/3) / (SSE/(76))
summary(mod.multivariate)
```

```
##
## Call:
## lm(formula = price ~ bdrms + lotsize + sqrft, data = HP)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -110.972  -37.232   -4.178   29.723  192.103
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.393e+01  2.842e+01  -1.546 0.126251
## bdrms        2.005e+01  8.404e+00   2.385 0.019550 *
## lotsize      1.991e-03  5.747e-04   3.464 0.000877 ***
## sqrft        1.228e-01  1.216e-02  10.094  1.1e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 53.39 on 76 degrees of freedom
## Multiple R-squared:  0.7343, Adjusted R-squared:  0.7238
## F-statistic: 70.01 on 3 and 76 DF,  p-value: < 2.2e-16
```

```
# f statistic is 70.00881 on 3 and 76 DF, p-value = 2.2e-16 < 0.05. Hence we reject H
0
```

Null hypothesis: β1 = β2 = β3 = 0 Alternative hypothesis: at least one βj is != 0 for j = 1,2,3

Since p-value < 0.05, we reject H0, providing statistical evidence that at least one independent variable is significantly associated with price of the house.

The overall model is statistically significant cat 5% significance interval.