ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA KHOA HỌC ỨNG DỤNG

Xác suất và Thống kê

MT2013

Báo cáo bài tập lớn

GVHD: TS. Nguyễn Bá Thi

Nhóm 9_L13:

Lê Văn Tuấn Kiệt	MSSV: 2110300
Nguyễn Thành Nhân	MSSV: 2012522
Diệp Bảo Phong	MSSV: 2014114
Nguyễn Xuân Triều	MSSV: 2110610
Phan Trần Minh Đạt	MSSV: 2111025
Trần Cao Nguyên	MSSV: 2111882
Nguyễn Sinh Thành	MSSV: 2112302

Hồ Chí Minh, 27/11/2022

MỤC LỤC

I. CƠ SỞ LÝ THUYẾT

1. Thống kê mô tả (Descriptive statistics, summary)

Nói đến thống kê mô tả là mô tả dữ liệu bằng các phép tính và chỉ số thống kê như số trung bình (mean), số trung vị (median), phương sai (variance) độ lệch chuẩn (standard deviation) … cho các biến số liên tục, và tỉ số (proportion) cho các biến số không liên tục.

- Số trung bình (mean): $\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Hàm R: mean(x).

- Phương sai (var): $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2$.Hàm R: var(x).

- Độ lệch chuẩn (sd): $s = \sqrt{s^2}$. Hàm R: sd(x).

- Sai số chuẩn (standard error): $SE = \frac{s}{\sqrt{n}}$.

- Giá trị nhỏ nhất min. Hàm R: min(x).

- Giá trị lớn nhất max. Hàm R: max(x).

Mục tiêu chính của phân tích thống kê mô tả là tìm những ước số của mẫu. Có hai loại đo lường: liên tục (continuous measurement) và không liên tục hay rời rạc (discrete measurement).

2. Biểu đồ

2.1. Biểu đồ Histogram

Biểu đồ phân bố tần số (còn được gọi là biểu đồ phân bố mật độ, biểu đồ cột) dùng để đo tần số xuất hiện của một vấn đề nào đó, cho ta thấy rõ hình ảnh sự thay đổi, biến động của một tập dữ liệu. Ba đặc trưng quan trọng của biểu đồ phân bố tần số là tâm điểm, độ rộng, độ dốc.

2.2. Biểu đồ hộp (box-plot):

Biểu đồ hộp (Box-plot)hay còn gọi là biểu đồ hộp-và-râu (box-and-whisker plot) là biểu đồ diễn tả 5 vị trí phân bố của dữ liệu, đó là : giá trị nhỏ nhất (min), tứ phân vị thứ nhất (Q1), trung vị (median), tứ phân vị thứ 3 (Q3) và giá trị lớn nhất (max).

2.3. Biểu đồ pairs:

Là biểu đồ thể hiện mối liên hệ giữa các biến.

2.4. Biểu đồ phân bố (scatter plot):

Để tìm hiểu mối liên hệ giữa hai biến, chúng ta dùng biểu đồ phân bố. Để vẽ biểu đồ phân bố, chúng ta sử dụng hàm plot.

3. Chọn mẫu ngẫu nhiên (random sampling):

Trong xác suất và thống kê, lấy mẫu ngẫu nhiên rất quan trọng, vì nó đảm bảo tính hợp lí của các phương pháp phân tích và suy luận thống kê. Với R, chúng ta có thể lấy mẫu một mẫu ngẫu nhiên bằng cách sử dụng hàm sample.

4. Kiểm định t (t.test):

Kiểm định t dựa vào giả thiết phân phối chuẩn. Có hai loại kiểm định t: kiểm định t cho một mẫu (one-sample t-test), và kiểm định t cho hai mẫu (two-sample t-test). Kiểm định t một mẫu nằm trả lời câu hỏi dữ liệu từ một mẫu có phải thật sự bằng một thông số nào đó hay không. Còn kiểm định t hai mẫu thì nhằm trả lời câu hỏi hai mẫu có cùng một luật phân phối, hay cụ thể hơn là hai mẫu có thật sự có cùng trị số trung bình hay không.

5. Phân tích hồi qui tuyến tính:

5.1. Hệ số tương quan:

Hệ số tương quan (r) là một chỉ số thống kê đo lường mối liên hệ tương quan giữa hai biến số, như giữa độ tuổi (x) và cholesterol (y). Hệ số tương quan có giá trị từ -1 đến 1. Hệ số tương quan bằng 0 (hay gần 0) có nghĩa là hai biến số không có liên hệ gì với nhau; ngược lại nếu hệ số bằng -1 hay 1 có nghĩa là hai biến số có một mối liên hệ tuyệt đối. Nếu giá trị của hệ số tương quan là âm ($r < 0$) có nghĩa là khi x tăng cao thì y cũng tăng, và khi x tăng cao thì y cũng giảm theo.

5.2. Mô hình của hồi qui tuyến tính đơn biến:

Mô hình hồi qui tuyến tính đơn biến dùng để xem xét mối quan hệ tuyến tính giữa biến phụ thuộc y (biến kết cục) và biến độc lập x (biến dự đoán). Phương trình tuyến tính (đường thẳng) đơn biến có dạng: $y = \alpha + \beta . x_i + \epsilon_i$.

Trong đó α là điểm cắt trên trục tung, β là độ dốc (trong thống kê gọi là hệ số hồi qui) và ϵ là phần dư.

5.2.1 Phương trình tổng quát

$$\hat{Y}_{|X} = B_0 + BX$$

$$B_0 = \bar{Y} - B\bar{X}$$

$$B = \frac{\sum X_i Y_i - \sum X_i Y_i / N}{X_i - \bar{X}}$$

Y : biến số phụ thuộc (dependent/reponse variable)

X : biến số độc lập (independent/predictor variable)

B_0, B : các hệ số hồi quy (regression coefficients)

5.2.2. Giá trị thống kê

Giá trị R bình phương (R square)

$$R = \frac{SSR}{SST}$$

Với $100R^2$ là tỉ lệ phần trăm của biến đổi trên Y được giải thích bởi X.

Độ lệch chuẩn (Standard Error)

$$S = \sqrt{\frac{1}{N-2}\sum (Y_i - Y_j)^2}$$

Sự phân tán của dữ liệu càng ít thì giá rị của S càng gần zero.

5.2.2. Trắc nghiệm thống kê

Đối với một phương trình hồi quy, $\hat{Y}_{|X} = B_0 + BX$, ý nghĩa thống kê của các hệ số B_i (B_0 hay B) được đánh giá bằng trắc nghiệm t (phân phối Student) trong khi tính chất thích hợp của phương trình $\hat{Y}_X = f(X)$ được đánh giá bằng trắc nghiệm F (phân phối Fisher).

Trắc nghiệm t

- Giả thiết:

H_0: $\beta_i = 0$, "Hệ số hồi quy không có ý nghĩa"

H_0: $\beta_i \neq 0$, "Hệ số hồi quy có ý nghĩa"

- Giá trị thống kê:

$$t = \frac{|B_i - \beta_i|}{\sqrt{S_n^2}};\ S_n^2 = \frac{S^2}{\sum (X_i - \bar{X})^2} = \frac{B}{\sqrt{S_n^2}}$$

Phân phối Student $\gamma = N - 2$

- Biện luận:

Nếu $t < t_\alpha (N-2) \Rightarrow$ Chấp nhận giả thiết H_0.

Trắc nghiệm F

- Giả thiết:

H_0: $\beta_i = 0$, "Phương trình hồi quy không thích hợp"

H_0: $\beta_i \neq 0$, "Phương trình hồi quy thích hợp"

- Giá trị thống kê:

$$F = \frac{MSR}{MSE}$$

- Kết luận:

Nếu $F < F_\alpha (1, N-2) \Rightarrow$ Chấp nhận giả thiết H_0.

5.3. Mô hình hồi qui tuyến tính đa biến:

Mô hình được diễn đạt qua phương trình có một yếu tố duy nhất (đó là x), và vì thế thường được gọi là mô hình hồi qui tuyến tính đơn giản : $y = \alpha + \beta . x_i + \epsilon_i$ (simple linear regression model). Trong thực tế, chúng ta có thể phát triển mô hình này thành nhiều biến, chứ không chỉ giới hạn một biến như trên, chẳng hạn như: $Y = \Theta_0 + \sum_{i=1}^{n} \Theta_i . X_i + \epsilon$. Trong đó X_i và Y là các biến độc lập ngẫu nhiên cho trước, thì sai số trung bình bình phương $E(|\epsilon|)$ là nhỏ nhất.

Chú ý trong phương trình trên, chúng ta có nhiều biến x (x_1, x_2, … đến x_k), và mỗi biến có một thông số β_j (j = 1, 2, …, k) cần phải ước tính. Vì thế mô hình này còn được gọi là mô hình hồi qui tuyến tính đa biến.

5.3.1. Phương trình tổng quát

$$\hat{Y}_{|X_0, X_1 \dots X_k} = B_0 + B_1 X_1 + B_2 X_2 + \cdots + B_k X_k$$

Phương trình hồi quy đa tham số có thể được trình bày dưới dạng ma trận.

5.3.2. Giá trị thống kê

Giá trị R^2 được hiệu chỉnh (Adjusted R Square)

$$R^2 = \frac{SSR}{SST} = \frac{kF}{(N-k-1) + kF}$$

Giá trị R^2 được hiệu chỉnh (Adjusted R Square)

$$R_u^2 = \frac{(N-1)R^2 - k}{N-k-1} = R^2 - \frac{k(1-R^2)}{N-k-1}$$

(R_u^2 sẽ trở nên âm hay không xác định nếu R^2 hay N nhỏ)

Độ lệch chuẩn (Standard Error)

$$S = \sqrt{\frac{SSE}{N-k-1}}$$

($S \leq 0{,}30$ là khá tốt)

5.3.3 Trắc nghiệm thống kê

Tương tự hồi quy đơn giản, song bạn cần chú ý:

Trắc nghiệm t

H_0: $\beta_i = 0$, "Các hệ số hồi quy không có ý nghĩa"

H_0: $\beta_i \neq 0$, "Có ít nhất vài hệ số hồi quy có ý nghĩa"

Bậc tự do của giá trị t: $\gamma = N - k - 1$.

$$t = \frac{|B_i - \beta_i|}{\sqrt{S_n^2}};\ S_n^2 = \frac{S^2}{\sum(X_i - \bar{X})^2}$$

Trắc nghiệm F

H_0: $\beta_i = 0$, "Phương trình hồi quy không thích hợp"

H_0: $\beta_i \neq 0$, "Phương trình hồi quy thích hợp" với ít nhất vài B_i.

Bậc tự do của giá trị F: $v_1 = 1$; $v_2 = N - k - 1$.

5.4. Sai số hồi quy (Regression residual):

Là khoảng cách theo chiều dọc giữa điểm dữ liệu và đường hồi quy. Mỗi dữ liệu sẽ có một khoảng cách. Những điểm dữ liệu nằm trên đường hồi quy sẽ có sai số dương. Những điểm dữ liệu nằm dưới đường hồi quy sẽ có sai số âm. Những điểm dữ liệu thuộc đường hồi quy sẽ có sai số bằng 0.

II. HOẠT ĐỘNG CHUNG

1. Đề tài 1:

Tập tin **gia_nha.csv** chứa thông tin về giá bán ra thị trường (đơn vị: Đô la) của 21613 ngôi nhà ở quận King, nước Mỹ trong khoảng thời gian từ tháng 5/2014 đến tháng 5/2015. Bên cạnh giá nhà, dữ liệu còn bao gồm các thuộc tính mô tả chất lượng ngôi nhà. Dữ liệu gốc được cung cấp tại https://www.kaggle.com/datasets/harlfoxem/housesalesprediction.

Các biến chính của bộ dữ liệu được dùng trong bài:

- price: Giá nhà được bán ra.
- sqft_living15: Diện tích trung bình của 15 ngôi nhà gần nhất trong khu dân cư.
- floors: Số tầng của ngôi nhà được phân loại từ 1-3.5.
- condition: Điều kiện kiến trúc của ngôi nhà từ 1 - 5 (1: rất tệ - 5: rất tốt).

- sqft_above: Diện tích ngôi nhà.
- sqft_living: diện tích khuôn viên nhà.

1.1. Đọc dữ liệu (Import data):

Đầu tiên, ta import file dữ liệu kc_house_data.csv vào Rstudio. Khởi tạo một data frame raw_table và lưu nội dung file vào data frame đó.

```
   #Get current working directory
   wd <- getwd()
   if (!is.null(wd)) setwd(wd)

   #Read CSV file
   raw = read.csv("gia_nha.csv")
```

Có thể thấy, dữ liệu ban đầu có thông tin của 21613 ngôi nhà (observation) về 21 yếu tố (variable). Mỗi hàng của bảng dữ liệu là một nhóm 21 thông tin khác nhau về một ngôi nhà, bao gồm các thông tin về giá nhà, điều kiện, chất lượng của ngôi nhà, …

1.2. Làm sạch dữ liệu (Data cleaning)

Trước tiên, ta sẽ tiến hành trích xuất 6 trường dữ liệu chính cần sử dụng như đã trình bày ở phần trên. Sau đó, các giá trị bị khuyết (NA) có trong các trường dữ liệu sẽ được lược bỏ.

```
   new_DF                                                                    =
raw[,c("price","sqft_living15","floors","condition","sqft_above","sqft_living")]
   new_DF = na.omit(new_DF)
```

1.3. Làm rõ dữ liệu (Data visualization)

Để dễ dàng tính toán với bộ dữ liệu, ta sẽ chuyển đổi một số biến có giá trị tương đối lớn thành các giá trị logarit cơ số *e* của chính nó. **Từ đây ta hiểu rằng mọi sự tính toán từ các biến trên đều đã được đổi sang dạng log**.

```
   #Convert initial data to logarithm
   new_DF       [,c("price","sqft_living15","sqft_above","sqft_living")]<-log(new_DF
[,c("price","sqft_living15","sqft_above","sqft_living")])
```

Tính toán các giá trị thống kê mô tả của các biến liên tục. Lưu các giá trị đó vào data frame c_table.

```
   c_mean                                                                    <-
c(mean(new_DF[,c("price")]),mean(new_DF[,c("sqft_living15")]),mean(new_DF[,c("sqft_ab
ove")]),mean(new_DF[,c("sqft_living")]))
   c_median                                                                  <-
c(median(new_DF[,c("price")]),median(new_DF[,c("sqft_living15")]),median(new_DF[,c("s
qft_above")]),median(new_DF[,c("sqft_living")]))
```

```
   c_sd                                                                    <-
c(sd(new_DF[,c("price")]),sd(new_DF[,c("sqft_living15")]),sd(new_DF[,c("sqft_above")]
),sd(new_DF[,c("sqft_living")]))
   c_min                                                                   <-
c(min(new_DF[,c("price")]),min(new_DF[,c("sqft_living15")]),min(new_DF[,c("sqft_above
")]),min(new_DF[,c("sqft_living")]))
   c_max                                                                   <-
c(max(new_DF[,c("price")]),max(new_DF[,c("sqft_living15")]),max(new_DF[,c("sqft_above
")]),max(new_DF[,c("sqft_living")]))

   c_table <- data.frame(rbind(c_mean,c_median,c_sd,c_min,c_max))

   colnames(c_table) <- c("price","sqft_living15","sqft_above","sqft_living")
   rownames(c_table) <- c("Mean","Median","Standard Deviation","Min","Max")
   head(c_table,10)
```

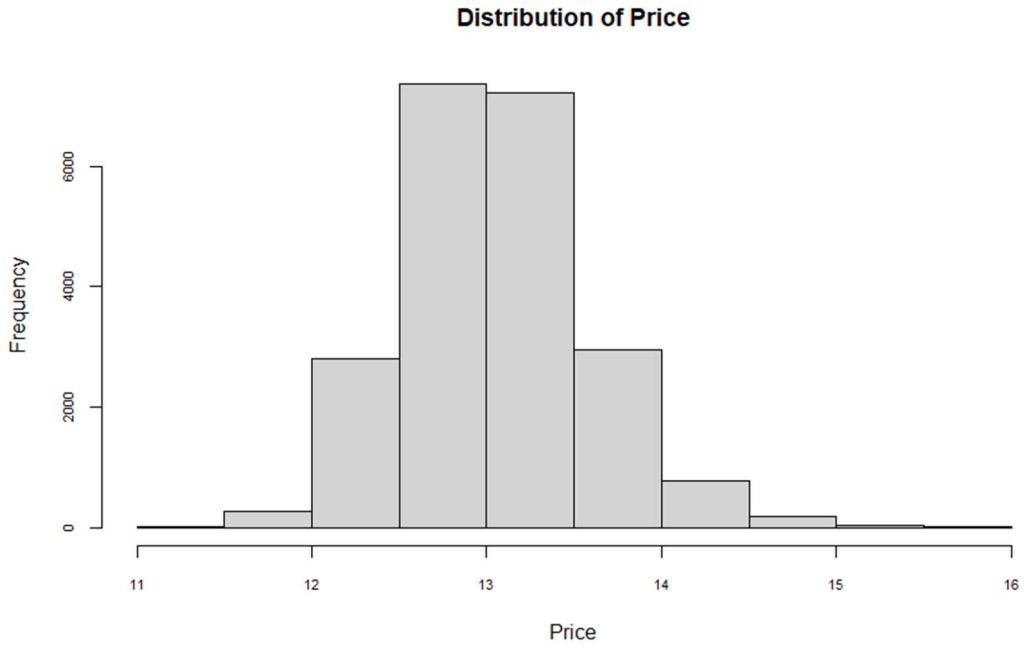
	price	sqft_living15	sqft_above	sqft_living
Mean	13.047841	7.5394471	7.3948826	7.5503286
Median	13.017003	7.5175209	7.3524411	7.5548585
Standard Deviation	0.526574	0.3274562	0.4276433	0.4247722
Min	11.225243	5.9889614	5.6698809	5.6698809
Max	15.856731	8.7339162	9.1495282	9.5134035

Lập bảng thống kê tần số đối với các biến rời rạc. Lưu các bảng đó vào data frame d_table_floors và d_table_condition.

```
   #Create a frequency table for discrete values
   table_floors <- data.frame(table(new_DF$floors))
   table_condition <- data.frame(table(new_DF$condition))

   colnames(table_floors) <- c("floors","freq")
   colnames(table_condition) <- c("condition","freq")
```

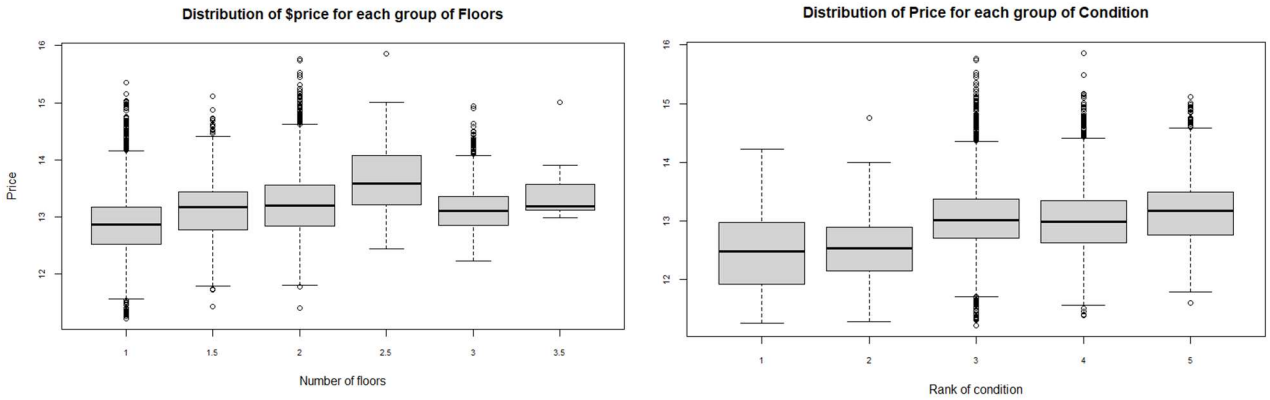
Để minh họa cho bộ dữ liệu, ta sẽ tiến hành vẽ các loại đồ thị liên quan, ta sử dụng các lệnh hist, boxplot và pairs để vẽ các đồ thị.

```
   #Plotting histogram
   hist(
     new_DF$price,
     main = "Distribution of Price",
     xlab = "Price",
     ylab = "Frequency",
   )
```



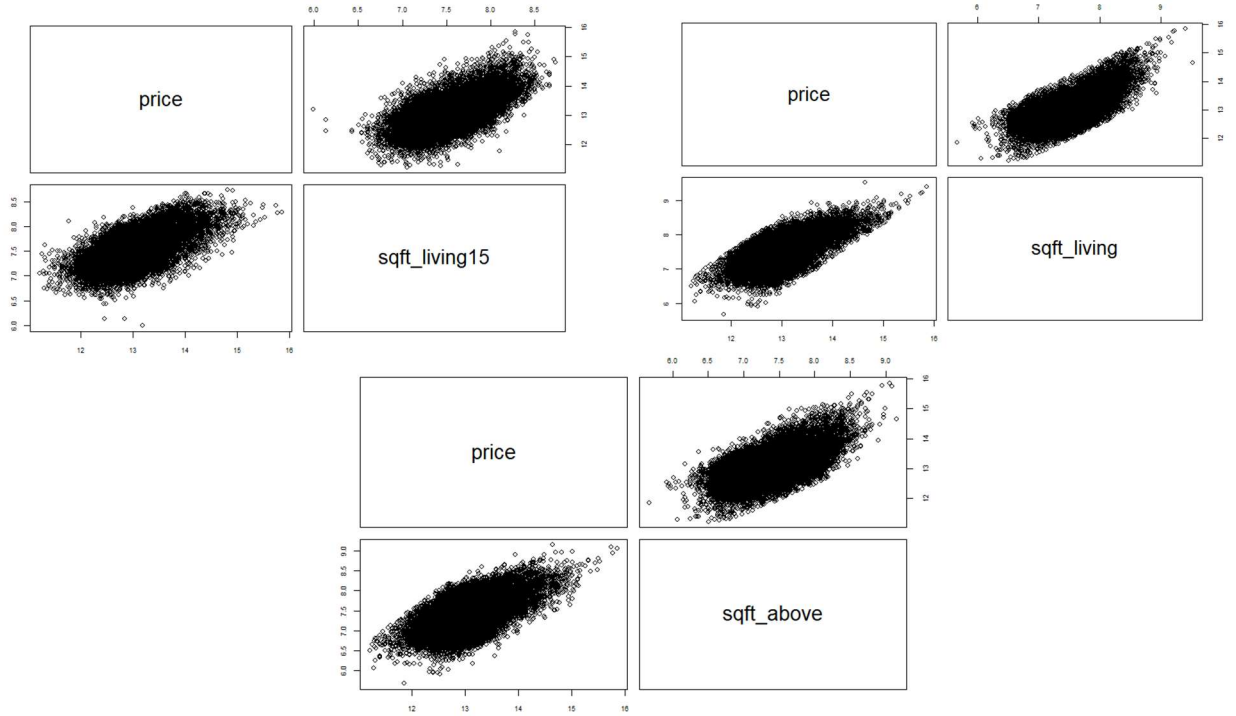
Ta thấy rằng phần lớn số nhà bán ra thị trường có mức giá tầm trung.

```
#Plotting box plots
boxplot(
  new_DF$price~ new_DF $floors,
  main = "Distribution of Price for each group of Floors",
  xlab = "Number of floors",
  ylab = "Price")
boxplot(
  new_DF $price~ new_DF $condition,
  main = "Distribution of Price for each group of Condition",
  xlab = "Rank of condition",
  ylab = "Price")
```




```
#Plotting pairs
pairs(price~sqft_living15, data = new_DF)
pairs(price~sqft_above, data = new_DF)
```

```
   pairs(price~sqft_living, data = new_DF)
```



Đây là biểu đồ phân tán giữa price và các biến định lượng. Ta có thể thấy rằng dường như price có mối quan hệ tuyến tính với các biến định lượng sqft_living15, sqft_above, sqft_living, nhưng để chắc chắn ta sẽ đi kiểm định lại xem các biến này có thực sự ảnh hưởng đến price hay không.

1.4 Xây dựng các mô hình hồi quy tuyến tính

Xét mô hình tuyến tính bao gồm biến **price** là một biến phụ thuộc, và tất cả các biến còn lại đều là biến độc lập. Ta dùng lệnh lm() để thực hiện mô hình hồi quy tuyến tính dưới dạng bội

```
   model = lm(price ~ sqft_living15 + floors + condition + sqft_above + sqft_living,
data = new_DF)
   summary(model)
```

```
Call:
   lm(formula = price ~ sqft_living15 + floors + condition + sqft_above +
       sqft_living, data = new_DF)

Residuals:
     Min       1Q   Median       3Q      Max
-1.25277 -0.27502  0.00764  0.24359  1.50543

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.442270   0.062423   87.18   <2e-16 ***
```

```
sqft_living15  0.430556   0.011977   35.95   <2e-16 ***
floors         0.137069   0.005952   23.03   <2e-16 ***
condition      0.085465   0.004076   20.97   <2e-16 ***
sqft_above    -0.178957   0.014021  -12.76   <2e-16 ***
sqft_living    0.686935   0.013186   52.09   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3727 on 21587 degrees of freedom
Multiple R-squared:  0.499, Adjusted R-squared:  0.4989
F-statistic:  4301 on 5 and 21587 DF,  p-value: < 2.2e-16
```

Mô hình này có thể được viết lại như sau:

```
   price = 5.44 + 0.43*sqft_living15 + 0.137*floors + 0.085*condition - 0.18*sqft_above
+ 0.69*sqft_living
```

Trong đó Residuals chính là phần dư (hay còn gọi là nhiễu), Intercept là hệ số chặn. Ở đây ta quan tâm đến giá trị ước lượng điểm (`Estimate`) của các hệ số B_i của các biến độc lập. Pr(>|t|) chính là hệ số P-value của kiểm định. Ta thấy rằng mọi `Pr(>|t|) <2e-16 < 0.05` nên tất cả các hệ số hồi quy đều có ý nghĩa, tức là sự thay đổi của biến độc lập tương ứng sẽ gây ra sự thay đổi về mặt thống kê đối với biến **`price`**. Ta cũng thấy `P-value` của `F-static` nhỏ hơn `2.2e-16,` có nghĩa rằng ta có “Mô hình hồi quy này phù hợp với ít nhất vài B_i”và R^2 khác 0.

Hệ số R^2 là `0.499`, tức là sự biến động của **`price`** thì có `49.9%` là do mô hình hồi quy tuyến tính này gây ra.

Với mức tin cậy 5%, vì `P-value` của mọi hệ số góc đều nhỏ hơn 5% nên ta bác bỏ giả thiết “H_0: $B_i = 0$”, vậy nên ta sẽ không loại bỏ biến nào khỏi mô hình này.

Ta xét mô hình M_2 là mô hình khi đã loại bỏ biến `condition` từ mô hình ban đầu, ta sử dụng hàm `anova()` để tìm mô hình hợp lý hơn.

```
   model2 = lm(price ~ sqft_living15 + floors  + sqft_above + sqft_living, data =
new_DF)
   anova(model, model2)
```

Kết quả:

```
   Analysis of Variance Table


   Model 1: price ~ sqft_living15 + floors + condition + sqft_above +
sqft_living
```

```
Model 2: price ~ sqft_living15 + floors + sqft_above + sqft_living
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1  21587 2999.3
2  21588 3060.4 -1   -61.072 439.55 < 2.2e-16 ***

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

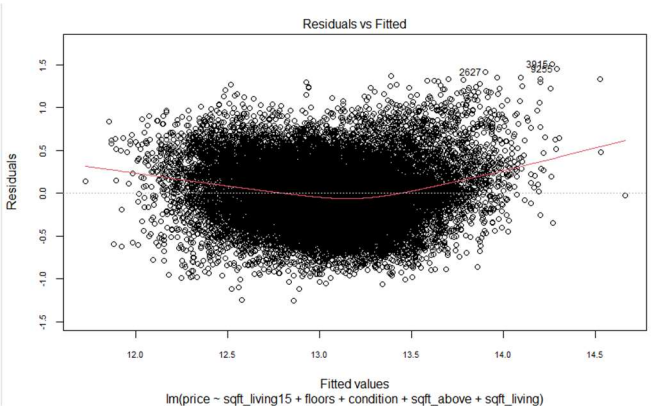
Ta thấy rằng `P-value < 2.2e-16,` với mức ý nghĩa là 5% thì việc loại bỏ biến `condition` có ý nghĩa về mặt thống kê, nên ta sẽ giữ lại biến condition và mô hình ban đầu hợp lý hơn.

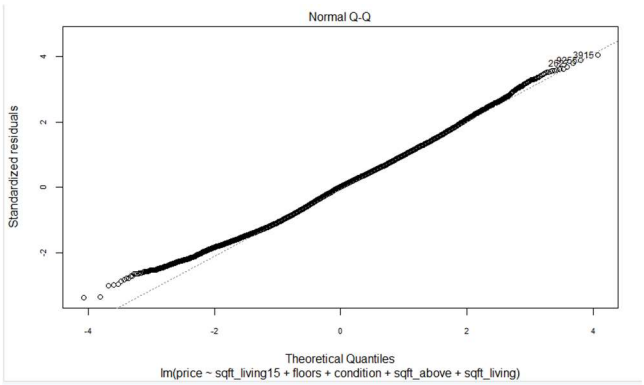
Ta kết luận được, mọi biến đều có quan hệ tuyến tính với giá nhà.

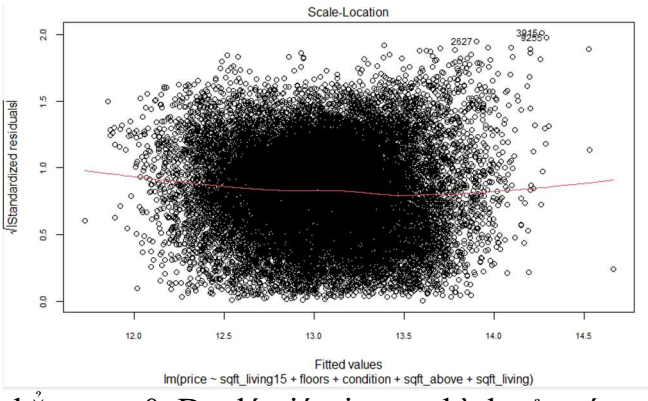
Ta dùng lệnh `plot` để biểu thị sai số hồi quy và giá trị dự báo

```
plot(model)
```

Quan sát đồ thị Residual vs Fitted, ta thấy các sai số ước lượng tập trung quanh đường







màu đỏ xấp xỉ đường thẳng y = 0. Do đó giá trị trung bình của các sai số ước lượng xấp xỉ 0. Nhìn vào biểu đồ QQ, các điểm tương ứng với phân vị của sai số ước lượng chuẩn hóa

ứng với phân vị của phân phối chuẩn nằm theo đường thẳng chuẩn nên khẳng định phân phối của sai số ước lượng là phân phối chuẩn.

1.5 Dự báo

Sau khi xây dựng mô hình hồi quy tuyến tính cho giá nhà, ta tiến hành dự đoán cho giá nhà. Trong báo cáo nhóm tiến hành dự đoán giá nhà ở điều kiện diện tích trung bình, số tầng là 2.

```
x1 = data.frame(sqft_living15 =     c_table["Mean", "sqft_living15"],
                        sqft_above = c_table["Mean", "sqft_above"],
                        sqft_living = c_table["Mean", "sqft_living"],
                        floors = 2,
                        condition = 3)
predict(model, newdata= x1, interval = "confidence")
```

Kết quả:

```
  fit       lwr      upr
1 13.08217 13.07425 13.0901
```

Điều này chỉ ra rằng, để mua được một căn nhà với điều kiện trung bình ở quận King, số tiền ta phải bỏ ra khoảng từ $e^{13.074}$ đến $e^{13.090}$. Tiếp theo, ta tiến hành dự đoán giá nhà tại điều kiện kiện diện tích lớn nhất, số tầng là 2 và điều kiện kiến trúc là 3.

```
x2 = data.frame(sqft_living15 = c_table["Max", "sqft_living15"],
                sqft_above =    c_table["Max", "sqft_above"],
                sqft_living =   c_table["Max", "sqft_living"],
                floors = 2,
                condition = 3)
predict(model, newdata= x2, interval = "confidence")
```

Kết quả:

```
  fit       lwr       upr
  1  14.63096 14.60666 14.65525
```

2. Đề tài 2:

Chăn nuôi gà là một ngành công nghiệp trị giá nhiều tỷ đô la ở Mỹ. Bất kỳ phương pháp nào có thể làm tăng tốc độ tăng trưởng của gà con đều giúp giảm chi phí tổng chăn nuôi và làm tăng lợi nhuận của công ty, có thể giá trị đến hàng triệu đô la. Một thí nghiệm đã được thực hiện để đo lường và so sánh hiệu quả của các loại thức ăn khác nhau đối với tốc độ tăng trưởng của gà con. Thí nghiệm được thực hiện như sau: người ta chia ngẫu nhiên những gà con mới nở vào sáu nhóm và mỗi nhóm được cung cấp một loại thức ăn khác nhua. Sáu loại thức ăn được thử nghiệm là casein, đậu răng ngựa (horsebean), hạt

lanh (linseed), thịt xay (meatmeal), đậu tương (soybean) và hoa hướng dương (sunflower). Kết quả của thí nghiệm được cung cấp trong tập tin **chicken_feed.csv**, gồm hai biến weight là trọng lượng của gà con sau thời gian dài được ăn loại thức ăn thử nghiệm, feed là biến nhân tố với các giá trị là tên 6 loại thức ăn được thử nghiệm.

2.1. Đọc file dữ liệu, thực hiện thống kê mô tả và kiểm định.

Ta đọc dữ liệu vào R và đưa biến feed thành biến nhân tố. Biến weight có chứa một số giá trị khuyến (NA), trong báo cáo, ta sử dụng phương pháp bỏ qua những giá trị khuyết này.

```
   bigTable <- read.csv("chicken_feed.csv", header=TRUE)
   bigTable$feed <- as.factor(bigTable$feed)
   bigTable <- na.omit(bigTable)
   foods = unique(bigTable $feed)
```

Tiếp theo, ta thực hiện tín toán những giá trị cho thống kê mô tả cho biến weight theo từng loại thức ăn tương ứng.

```
   #tạo những bảng rỗng để sử dụng
   set = list()
   c_max=c_sd = c_range = c_mean = c_max = c_mean = c()
   c_summary=data.frame(range(6)) #rỗng

   for (i in c( 1:length(foods)))
   {
     data = subset(bigTable, feed == foods[i]) $weight #tách được dữ liệu
     set <- list.append(set, data)
     print(foods[i])
     c_max    = c(c_max,    max(data))
     c_median = c(c_median, median(data))
     c_mean   = c(c_mean,   mean (data))
     c_sd     = c(c_sd,     sd(data))
     c_summary = data.frame(c_summary, data.frame (as.array(summary(data)))[2] ) #giá
trị lớn nhất, nhỏ nhất, tứ phân vị, độ lệch chuẩn, outlier
   }
   c_summary <- c_summary[ , -1] #để xóa cột đầu tiên
   summary_table = rbind(c_summary, c_sd)
   rownames(summary_table) <- c("Min", "1st Qu.", "Median", "Mean", "3rd Qu.", "Max",
"Standard deviation")
   colnames(summary_table) <- foods
   summary_table
```

Kết quả thu được:

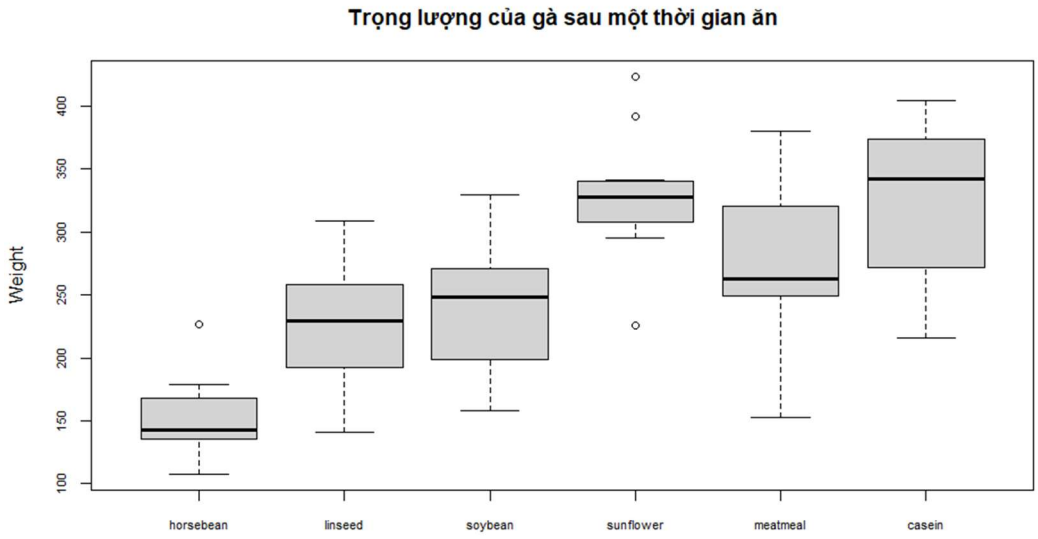
```
                horsebean   linseed   soybean sunflower  meatmeal    casein
Min              108.0000 141.00000 158.00000 226.00000 153.00000 216.00000
1st Qu.          136.0000 192.00000 206.75000 312.75000 249.50000 277.25000
Median           143.0000 229.00000 248.00000 328.00000 263.00000 342.00000
Mean             153.8889 225.18182 246.42857 328.91667 276.90909 323.58333
3rd Qu.          168.0000 258.50000 270.00000 340.25000 320.00000 370.75000
```

```
Max                 227.0000 309.00000 329.00000 423.00000 380.00000 404.00000
Standard deviation   35.0765  49.55163  54.12907  48.83638  64.90062  64.43384
```

Để dữ liệu được trực quan hơn, ta vẽ biểu đồ boxplot cho trọng lượng của gà con theo từng loại thức ăn tương ứng.

```
boxplot(set, names = foods,par(cex.axis=0.7),
        main = "Trọng lượng của gà sau một thời gian ăn", ylab ="Weight" )
```

Kết quả:



Nhận xét: Biểu đồ boxplot biểu hiện sự dao động của cân nặng theo từng loại thức ăn. Ta thấy rằng mỗi loại thức ăn ảnh hưởng đến trọng lượng của gà con theo một cách khác nhau. Casein có lẽ cho cân nặng cao nhất nhưng nhiều biến động. Khoảng biến động của sự ảnh hưởng của Sunflower đến cân nặng của gà con nhỏ hơn các loại còn lại. Horsebean cho trọng lượng của gà con rất nhỏ.

2.2. Phân tích phương sai một nhân tố (one way ANOVA)

Từ biểu đồ boxplot trên, có lẽ ta cũng nhận thấy rằng sự khác biệt của ảnh hưởng của các loại thức ăn đến trọng lượng của gà con. Ta xem xét trung bình của các nhóm thì ta thấy có sự khác biệt rõ rệt. Tuy nhiên ta cần xem xét thêm sự biến động trong từng nhóm. Ta thấy rằng sự biến động của sự ảnh hưởng thức ăn đến tăng trưởng của gà con trong từng nhóm là khá lớn, điều này do ta lấy mẫu ngẫu nhiên. Vậy sự khác biệt của trọng lượng của gà con là do tác động của các loại thức ăn hay thực sự chỉ do ngẫu nhiên. Để đi trả lời câu hỏi trên, ta đi phân tích phương sai (ANOVA). Biến phụ thuộc của ta sẽ là `weight`, biến độc lập của ta là `feed`. Cặp giả thuyết của ta là:

H_0: "Tất cả trung bình của các nhóm gà con ăn các loại thức ăn khác nhau bằng nhau"

$$u_1 = u_2 = \cdots = u_6$$

H_1: "Có ít nhất hai nhóm có trung bình khác nhau."

$$\exists i, j \; u_i \neq u_j$$

Để áp dụng phương pháp phân tích phương sai, ta cần mô hình thỏa mãn: Các quần thể có phân phối chuẩn với cùng phương sai σ^2.

Đầu tiên ta kiểm tra giả định về phân phối chuẩn, ta sử dụng kiểm định Shapiro-Wilk với giả thiết H_0: Tổng thể có phân phối chuẩn.

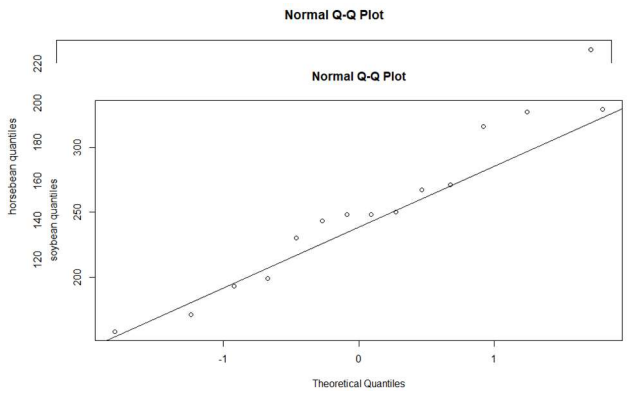
```
sharipowilk = c(1,2,3,4)
for (i in c( 1:length(foods)))
{
  sharipowilk <- rbind (sharipowilk, shapiro.test(set[[i]]))
  qqnorm(set[[i]], ylab = paste(foods[i], "quantiles", sep = ' ') )
  qqline(set[[i]])
}
sharipowilk = sharipowilk[-1, ]
rownames(sharipowilk) = foods
sharipowilk = sharipowilk[ , c(1,2,3)]
```

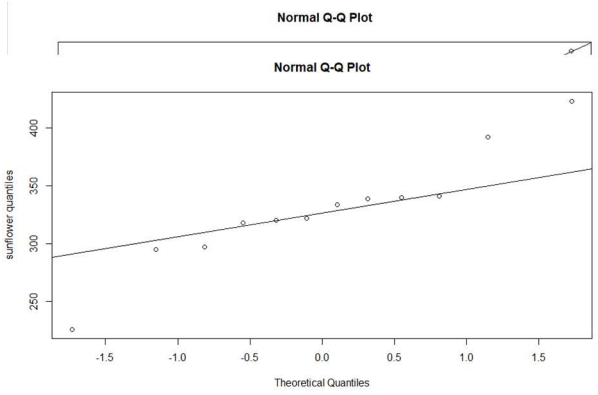
Kết quả:

```
           statistic p.value   method
horsebean 0.93885   0.5698468 "Shapiro-Wilk normality test"
linseed   0.986724  0.9921462 "Shapiro-Wilk normality test"
soybean   0.9464029 0.5063768 "Shapiro-Wilk normality test"
sunflower 0.9280884 0.3602904 "Shapiro-Wilk normality test"
meatmeal  0.9791381 0.9611795 "Shapiro-Wilk normality test"
casein    0.9166257 0.2591841 "Shapiro-Wilk normality test"
```

Với mức ý nghĩa 5%, P-value của tất cả các nhóm đều lớn hơn 0,05, ta chấp nhận giả thiết các tổng thể cân nặng của gà con theo từng loại thức ăn được cho ăn có phân phối chuẩn.

Để trực quan dữ liệu, ta vẽ biểu đồ QQ:



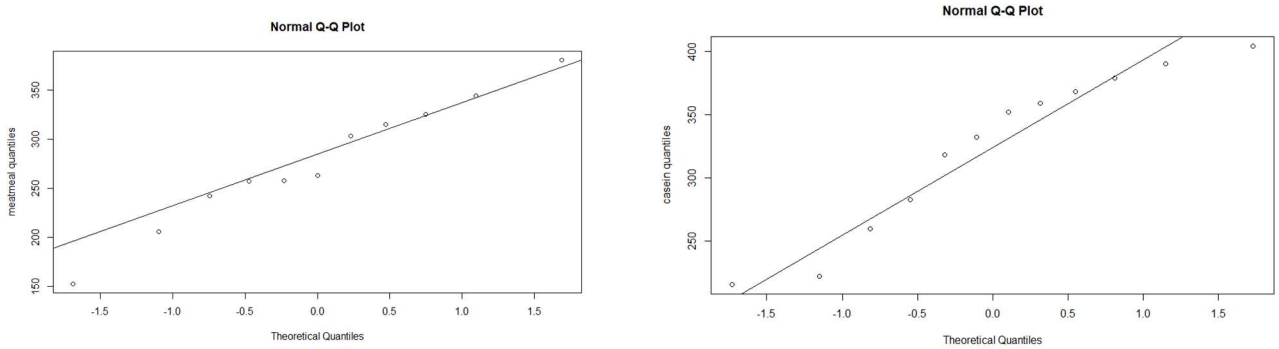


QQ-plots cho phép ta so sánh phân bố của tổng thể của một mẫu với phân phối chuẩn. Bởi vì tất cả các điểm đều ở gần đường thẳng (không nhất thiết phải là đường chéo $y = x$) nên ta thấy rằng các tổng thể có phân phối chuẩn.

Tiếp theo ta kiểm định giả thiết về phương sai, ta sử dụng kiểm định Leneve, với giả thiết $H_0: \sigma_1 = \sigma_2 = \cdots = \sigma_6$ và giả thiết $H_1: \exists i, j\ \sigma_i \neq \sigma_j$.

```
leveneTest(weight~feed,bigTable)
```

Kết quả:




```
Levene's Test for Homogeneity of Variance (center = median)
      Df F value Pr(>F)
group  5  0.9016 0.4858
      63
```

Vì với mức ý nghĩa 5%, ta chấp nhận giả thiết phương sai của các tổng thể cân nặng của gà theo các loại thức ăn được cho ăn có phương sai bằng nhau.

Vậy, mô hình này có thể áp dụng phương pháp phân tích phương sai để kiểm định trung bình của các nhóm gà con ăn các loại thức ăn khác nhau.

```
anova(lm(weight~feed, data = bigTable))
```

Kết quả:

```
Response: weight
          Df Sum Sq Mean Sq F value    Pr(>F)
feed       5 225012   45002  15.201 8.833e-10 ***
Residuals 63 186511    2960
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Với mức ý nghĩa 5%, ta có $P_{value} = 8.833e - 10 < 0.05$ nên ta bác bỏ giả thiết H_0: Tất cả trung bình của các nhóm gà con ăn các loại thức ăn khác nhau bằng nhau, $u_1 = u_2 = \cdots = u_6$

Vậy nên có ít nhất hai tổng thể gà ăn hai loại thức ăn khác nhau có trung bình cân nặng khác nhau, $\exists i, j\ u_i \neq u_j$.

Phương pháp ANOVA chỉ cho ta biết rằng có ít nhất 2 nhóm có trung bình khác nhau nhưng không cho ta biết các nhóm khác nhau từng đôi một như thế nào. Để so sánh sự khác biệt đôi một giữa trung bình của các nhóm, ta sử dụng phương pháp kiểm định Tukey HSD, với giả thiết H_0: sự khác biệt giữa u_i và và u_j **không** có ý nghĩa về mặt thống kê và giả thiết H_1: sự khác biệt giữa u_i và u_j có ý nghĩa về mặt thống kê. Phương pháp này đi so sánh $|u_i - u_j|$ với SEq_α với q_α từ phân phối *Student*.

```
model = aov(weight~feed, data=bigTable)
TukeyHSD(model)
```

Kết quả:

```
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weight ~ feed, data = bigTable)

$feed
                          diff          lwr       upr     p adj
horsebean-casein    -169.694444 -240.2146894 -99.17420 0.0000000
linseed-casein       -98.401515 -165.1579285 -31.64510 0.0007400
meatmeal-casein      -46.674242 -113.4306557  20.08217 0.3240085
soybean-casein       -77.154762 -140.0688745 -14.24065 0.0077887
sunflower-casein       5.333333  -59.9557269  70.62239 0.9998856
linseed-horsebean     71.292929   -0.5879602 143.17382 0.0531327
meatmeal-horsebean   123.020202   51.1393125 194.90109 0.0000620
soybean-horsebean     92.539683   24.2123152 160.86705 0.0023904
sunflower-horsebean  175.027778  104.5075328 245.54802 0.0000000
meatmeal-linseed      51.727273  -16.4649266 119.91947 0.2391269
soybean-linseed       21.246753  -43.1888185  85.68232 0.9259562
sunflower-linseed    103.734848   36.9784352 170.49126 0.0003279
soybean-meatmeal     -30.480519  -94.9160912  33.95505 0.7325632
sunflower-meatmeal    52.007576  -14.7488376 118.76399 0.2135819
sunflower-soybean     82.488095   19.5739826 145.40221 0.0035948
```

Có hai cách để ta có thể rút ra kết luận từ bảng này, hoặc sử dụng P-value hoặc sử dụng khoảng tin cậy của hiệu hai trung bình mẫu. Với khoảng tin cậy với độ tin cậy 95%, hay là với mức ý nghĩa 5%, ta thấy rằng các cặp `horsebean-casein`, `linseed-casein`, `meatmeal-horsebean`, `soybean-horsebean`, `sunflower-horsebean`, `sunflower-linseed`, `sunflower-soybean` có sự khác biệt có ý nghĩa về mặt thống kê.

Loại thức ăn tốt nhất cho sự tăng trưởng của gà con ta rút ra được là **Sunflower.** Kết luận rút ra này hợp lý với biểu đồ Boxplot mà ta vẽ ở trên.

III. PHẦN RIÊNG

Bộ dữ liệu **Computer.csv** chứa thông tin về các thông số của phần cứng máy tính và giá tiền được bán ra ngoài thị trường của 1 nhà phân phối máy tại Mĩ và tháng 1 năm 2015.

price: Giá tiền máy tính

speed: tốc độ RAM, đơn vị MHz, nhận giá trị 25, 33, 50, 66

hd: Dung lượng đĩa cứng, đơn vị GB

ram: dung lượng RAM, đơn vị GB, nhận giá trị 2, 4, 6, 8

screen: Độ dài đường chéo màn hình, tính bằng inch, nhận các giá trị 14, 15, 17

cd: Sự xuất hiện của ổ đĩa CD trong máy tính (yes/no)

trend: Điểm dự đoán mức độ thịnh hành của máy tính trong tương lai (thang điểm từ 1-50, điểm càng thấp thì máy tính càng thịnh hành)

Ta coi biến price là biến phụ thuộc, các biến còn lại là biến độc lập, ta sẽ nghiên cứu sự ảnh hưởng của các biến độc này lên **price.** Biến trend, hd và price là biến định lượng. Các biến còn lại là ra, speed, cd là các biến định tính do các biến nay là các tiêu chí để phân loại máy tính, tập giá trị của chúng chỉ gồm một số ít các giá trị rời rạc. Ngoài ra, ta có thể khảo sát xu hướng chuộng máy tính như thế nào.

1. Đọc và làm sạch dữ liệu

```
raw <- read.csv("Computers.csv")
attach(raw)

table <- data.frame(price, speed, hd, ram, screen, cd, trend)
detach(raw)
attach(table)

table[table[,] == ""] = NA
table <- na.omit(table)
```

2. Làm rõ dữ liệu

Ta chuyển kiểu dữ liệu của các biến định tính sang kiểu factor

```
speed <- as.factor(speed)
ram <- as.factor(ram)
screen <- as.factor(screen)
cd <- as.factor(cd)
```

Ta viết hàm SUMMARY() để in bảng thống kê của các biến định tính.

```
SUMMARY <- function(x)
{
  df <- data.frame(table(x))
```

```
    colnames(df) <- c("Type", "Quantity")
    df$Percent <- df$Quantity / sum(df$Quantity) * 100
    print(df)
  }

  SUMMARY(speed)
  SUMMARY(ram)
  SUMMARY(cd)
  SUMMARY(screen)
```

Kết quả:

```
> SUMMARY(speed)
  Type Quantity   Percent
1   25      566  9.042978
2   33     2033 32.481227
3   50      994 15.881131
4   66     2028 32.401342
5   75      122  1.949193
6  100      516  8.244128
> SUMMARY(ram)
  Type Quantity    Percent
1    2      394  6.2949353
2    4     2236 35.7245566
3    8     2320 37.0666241
4   16      996 15.9130852
5   24      297  4.7451670
6   32       16  0.2556319
> SUMMARY(cd)
  Type Quantity Percent
1   no     3351 53.5389
2  yes     2908 46.4611
> SUMMARY(screen)
  Type Quantity   Percent
1   14     3661 58.491772
2   15     1992 31.826170
3   17      606  9.682058
```

Ta thống kế các biến định lượng:

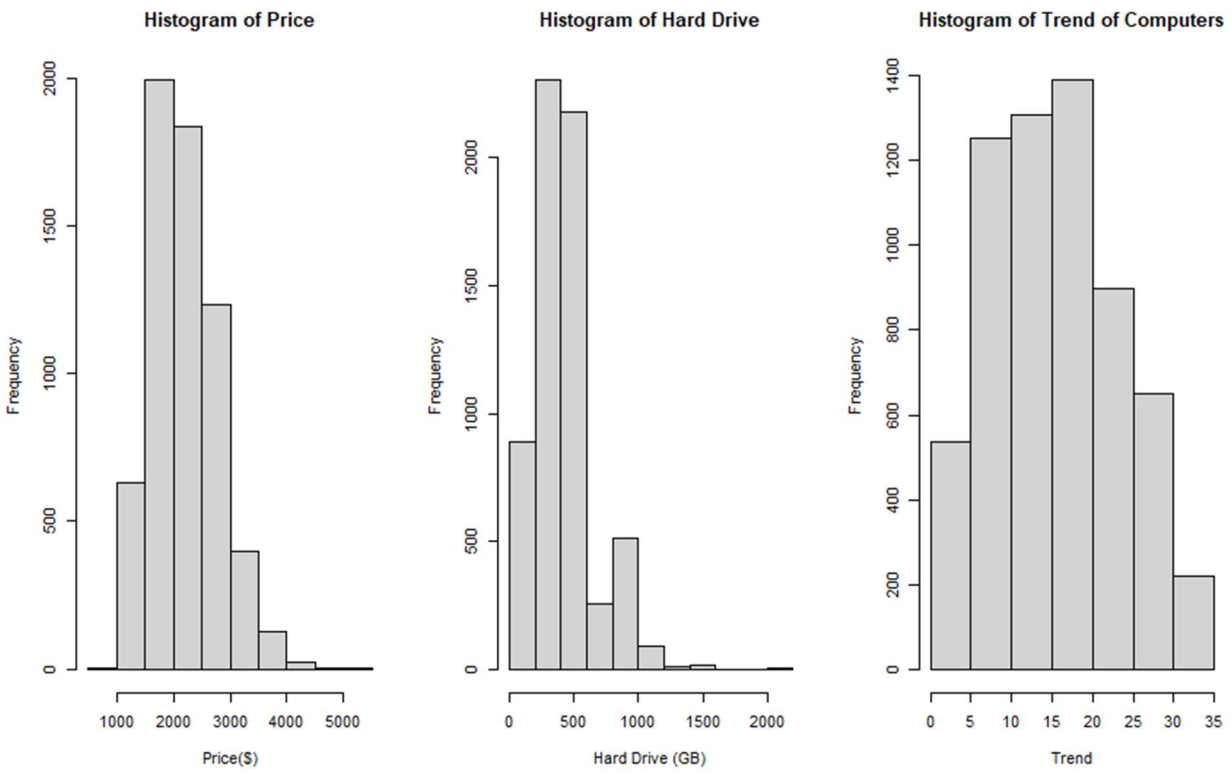
```
  analyze <- function(x){
    parameter = c("Min:", "Qua1:", "Median:", "Qua3:", "Max:", "Min:", "Sd:", "Var:")
    value = c(min(x), summary(x)[2], median(x), summary(x)[5], max(x), min(x), sd(x),
var(x))
    print(data.frame(parameter, value))
  }
  analyze(price)
  analyze(hd)
  analyze(trend)
```

Kết quả:

```
> analyze(price)
  parameter      value
1      Min:    949.000
2     Qua1:   1794.000
3   Median:   2144.000
4     Qua3:   2595.000
5      Max:   5399.000
6      Min:    949.000
7       Sd:    580.804
8      Var: 337333.235
> analyze(hd)
  parameter      value
1      Min:    80.0000
2     Qua1:   214.0000
3   Median:   340.0000
4     Qua3:   528.0000
5      Max:  2100.0000
6      Min:    80.0000
7       Sd:   258.5484
8      Var: 66847.2985
> analyze(trend)
  parameter     value
1      Min:  1.000000
2     Qua1: 10.000000
3   Median: 16.000000
4     Qua3: 21.500000
5      Max: 35.000000
6      Min:  1.000000
7       Sd:  7.873984
8      Var: 61.999622
```

Ta vẽ đồ thị phân phối của 3 biến định lượng này

```
hist(price, breaks = 10,
     xlab = "Price($)", main = "Histogram of Price",
     )
hist(hd, breaks = 10,
     xlab = "Hard Drive (GB)", main = "Histogram of Hard Drive",
     )
hist(trend, breaks = 10,
     xlab = "Trend", main = "Histogram of Trend of Computers",
     )
```

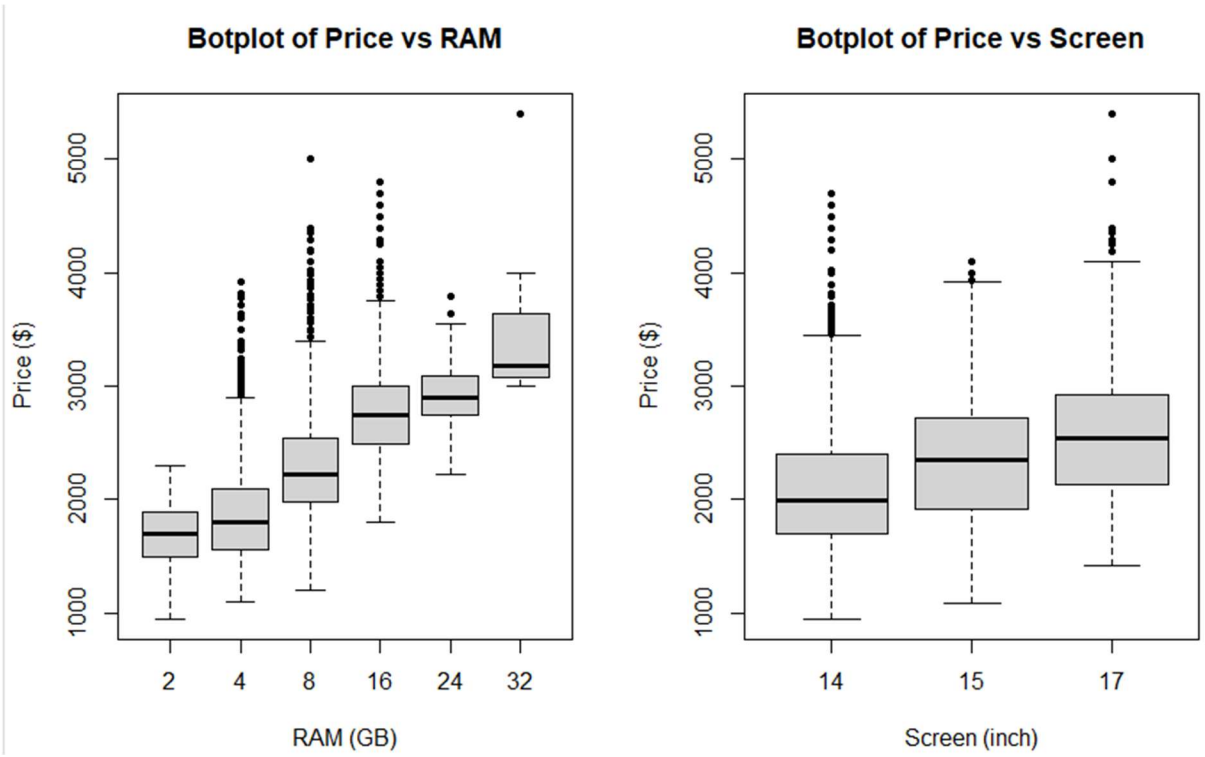


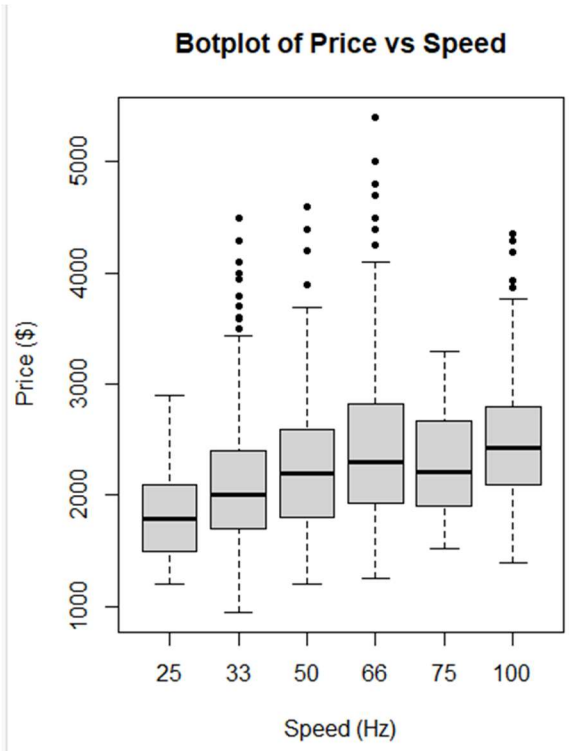
Ta thấy rằng khách hàng có xu hướng mua những máy tính ở phân khúc tầm trung.

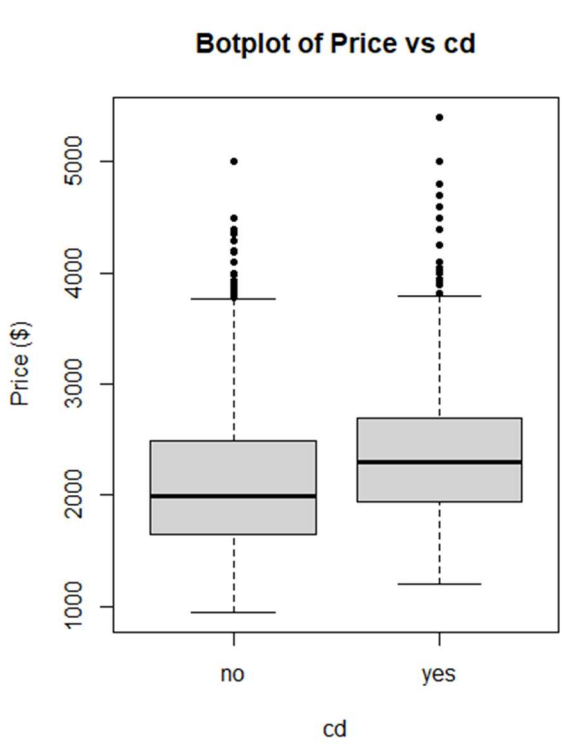
3. Khảo sát sự ảnh hưởng của các biến độc lập lên biến price.

Ta vẽ biểu đồ Boxplot để khảo sát sự ảnh hưởng của từng biến định tính lên biến price

```
boxplot(price ~ ram,
        xlab = "RAM (GB)",
        ylab = "Price ($)",
        main = "Botplot of Price vs RAM",
        )
boxplot(price ~ screen,
        xlab = "Screen (inch)",
        ylab = "Price ($)",
        main = "Botplot of Price vs Screen")
boxplot(price ~ speed,
        xlab = "Speed (Hz)",
        ylab = "Price ($)",
        main = "Botplot of Price vs Speed")
boxplot(price ~ cd,
        xlab = "cd",
        ylab = "Price ($)",
        main = "Botplot of Price vs cd")
```







Ta thấy dung lượng RAM, màn hinh, tốc độ bus càng lớn thì máy tính càng đắt tiền. Có thêm ổ đĩa CD cũng khiến máy tính tăng giá.

Với mức ý nghĩa là 5%, ta kiểm định xem sự thay đổi ở những biến định tính này có thực sự gây ra sự thay đổi giá thành máy tính hay không, ta sẽ sử dụng hàm `anova()`

```
anova(lm(price ~ ram))
anova(lm(price ~ screen))
anova(lm(price ~ speed))
anova(lm(price ~ cd))
```

Hàm anova chính là kiểm định xem thử sự thay đổi ở biến định tính có thực sự gây ra sự thay đổi ở biến độc lập hay không.

```
> anova(lm(price ~ ram))
Analysis of Variance Table

Response: price
            Df     Sum Sq   Mean Sq F value    Pr(>F)
ram          5  887135680 177427136  906.49 < 2.2e-16 ***
Residuals 6253 1223895704    195729
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(lm(price ~ screen))
Analysis of Variance Table

Response: price
            Df     Sum Sq  Mean Sq F value    Pr(>F)
screen       2  195114561 97557281  318.55 < 2.2e-16 ***
Residuals 6256 1915916823   306253
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(lm(price ~ speed))
Analysis of Variance Table

Response: price
            Df     Sum Sq  Mean Sq F value    Pr(>F)
speed        5  235645782 47129156  157.14 < 2.2e-16 ***
Residuals 6253 1875385602   299918
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(lm(price ~ cd))
Analysis of Variance Table

Response: price
            Df     Sum Sq  Mean Sq F value    Pr(>F)
cd           1   82212838 82212838  253.55 < 2.2e-16 ***
Residuals 6257 2028818546   324248
---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Tất cả các biến định tính đều cho P-value nhỏ hơn 2.2e-16, giá trị này rất nhỏ hơn 0.05, do đó các biến định tính này có ý nghĩa về mặt thống kê đối với biến price. Chúng có sự ảnh hưởng đến giá thành của máy tính trên thị trường.

4. Xây dựng mô hình hồi quy tuyến tính

Ta lượng hóa cá biến định tính

```
cd <- as.character(cd)
cd <- replace(cd, cd=="yes", 1)
cd <- replace(cd, cd=="no", 0)
cd<-as.numeric((cd))
speed=as.numeric(speed)
hd = as.numeric(hd)
trend = as.numeric(trend)
screen = as.numeric(screen)
ram = as.numeric(ram)
```

Ta xây dựng mô hình tuyến tính:

```
model = lm(price~cd + speed + hd + trend + screen)
summary(model)
```

Kết quả:

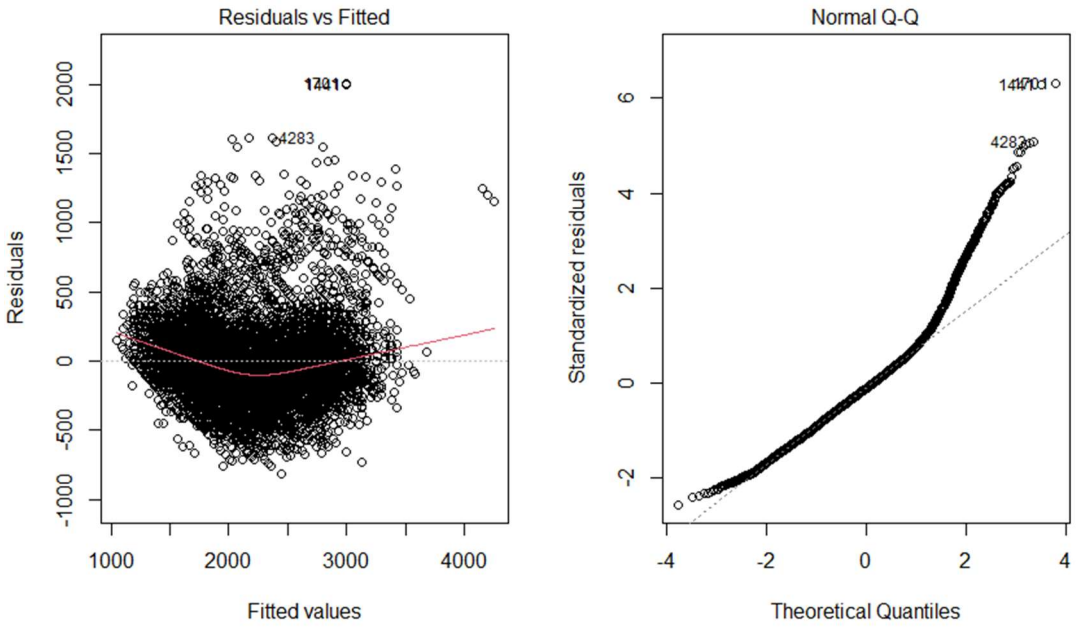
```
Residuals:
    Min      1Q  Median      3Q     Max
-812.54 -200.41  -31.75  145.58 2006.04

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1314.3100    18.0216  72.930  < 2e-16 ***
cd1           40.5592     9.8291   4.126 3.73e-05 ***
speed        131.0916     3.3652  38.955  < 2e-16 ***
ram          283.7783     7.2854  38.951  < 2e-16 ***
hd             0.5966     0.0311  19.185  < 2e-16 ***
trend        -48.8642     0.7139 -68.447  < 2e-16 ***
screen       145.9372     6.4064  22.780  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 318.8 on 6252 degrees of freedom
Multiple R-squared:  0.699,   Adjusted R-squared:  0.6987
F-statistic:  2420 on 6 and 6252 DF,  p-value: < 2.2e-16
```

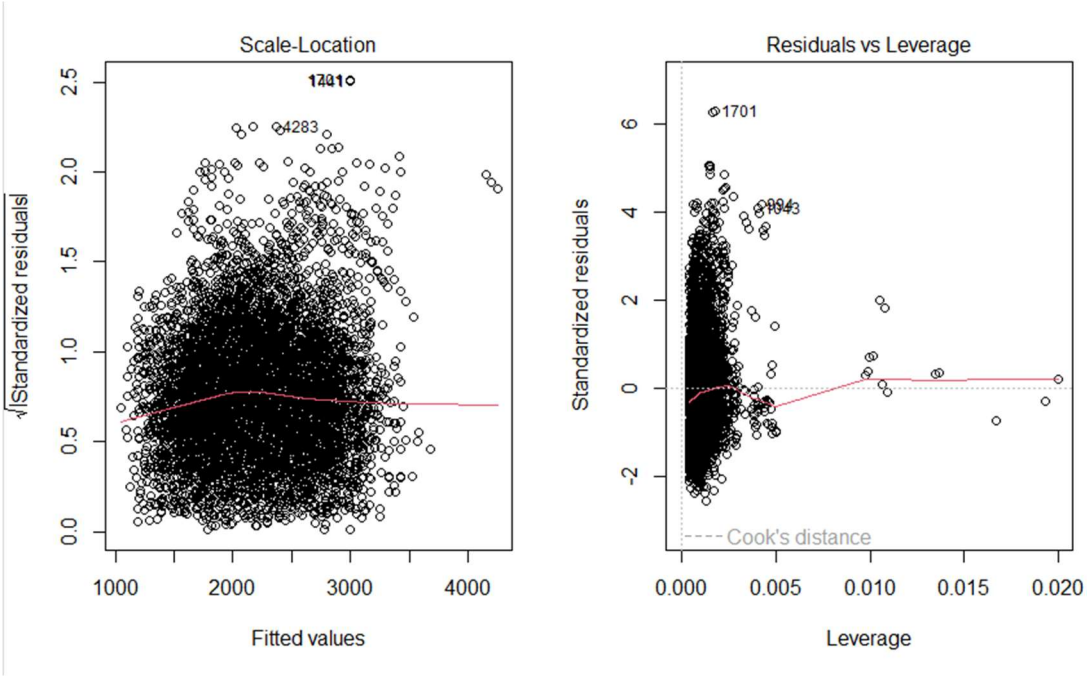
Mô hình: `price = 1314.31 + 40.56*cd1 + 131.1*speed + 283.78*ram +  0.6*hd - 48.86*trend + 145.94*screen`

Với mức ý nghĩa 5%, bởi vì P-value của tất cả các biến độc lập nhỏ hơn 0.05, điều đó chỉ ra rằng sự thay đổi của các biến độc lập này gây ra sự thay đổi thực sự ở biến price. Ngoài ra, P-value của F-statistic nhỏ hơn 2.2e-16, có nghĩa rằng mô hình hồi quy này phù hợp *với ít nhất một vài* B_i, và R^2 khác 0. Hệ số R-square bằng 0.699 chỉ ra rằng 69.9% sự thay đổi của biến price có thể được giải thích bằng mô hình này, ngoài ra còn có nhiều tác động bên ngoài lên giá thành của máy tính.

Ta vẽ đồ thị để thấy được sự hợp lý của mô hình này:

```
plot(model)
```





Đồ thị Residuals vs fitted cho thấy số dư tập trung xung quanh trục 0, đồ thị QQ cho thấy rằng số dư chuẩn hóa mang phân phối chuẩn. Vậy mô hình tuyến tính vừa tìm được phù hợp và 6 biến độc lập trên ảnh hưởng tới giá tiền một chiếc máy tính.

TÀI LIỆU THAM KHẢO

[1] *Giáo trình Xác suất và Thống kê*, Nguyễn Đình Huy, Đậu Thế Cấp, Lê Xuân Đại

[2] D. C. Montgomery, G. C. Runger, Applied Statistics and Probability for Engineers, 7th Edtion.