ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC BÁCH KHOA
KHOA CƠ KHÍ

BÁO CÁO BÀI TẬP LỚN

MÔN XÁC SUẤT THỐNG KÊ

ĐỀ TÀI 4

"XÁC ĐỊNH MỨC ĐỘ ẢNH HƯỞNG CỦA CÁC THÔNG SỐ ĐIỀU CHỈNH TRONG MÁY IN 3D ĐẾN CHẤT LƯỢNG IN, ĐỘ CHÍNH XÁC VÀ DỘ BỀN"

Sinh viên:	**MSSV**
Huỳnh Thị Ngọc Bích	2112894
Võ Minh Đăng	2210744
Phạm Khoa Nguyên	2114240

Giảng viên hướng dẫn : Phan Thị Hường

Lớp: L10 – Nhóm 1

TP.HCM, tháng 11 năm 2023

Bảng đánh giá quá trình công việc của L10 - NHÓM 1

Tên thành viên	Nhiệm vụ	% Nhiệm vụ	Điểm cộng/ trừ
Huỳnh Thị Ngọc Bích - 2112894	- Thống kê tả - Thống kê suy diễn - Code R	100%	0
Võ Đăng Khoa – 2210744	- Thống kê tả - Code R	100%	0
Phạm Khoa Nguyên – 2114240	- Tổng quan dữ liệu - Kiến thức nền - Mở rộng - Word	100%	0

MỤC LỤC

1. TỔNG QUAN DỮ LIỆU

1.1 Mục đích nghiên cứu

Xác định mức độ ảnh hưởng của các thông số điều chỉnh trong máy in 3d đến chất lượng in, độ chính xác và độ bền. Trong đó có chín thông số cài đặt và ba thông số đầu ra được đo.

1.2 Nguồn dữ liệu

Dữ liệu được cung cấp tại: 3D Printer Dataset for Mechanical Engineers (kaggle.com)

1.3 Miêu tả các biến

Gồm 12 biến:

Layer_height	*Độ dày của từng lớp trong quá trình in 3D. Nó đo bằng đơn vị millimet (mm) và có thể ảnh hưởng đến chất lượng*
Wall_thickness	*Độ dày của thành sản phẩm hoặc vật thể in 3D. Nó đo bằng đơn vị millimet (mm) và quyết định độ mạnh của vật thể*
Infill_density	Tỷ lệ khoảng trống trong sản phẩm in 3D. Nó có thể ảnh hưởng đến trọng lượng của sản phẩm và sử dụng nguyên liệu in
Infill_pattern	Đây là mẫu hoặc cấu trúc sử dụng để điền vào khoảng trống bên trong sản phẩm in 3D. Ví dụ: "grid" có thể ám chỉ việc điền vào bằng lưới, "honeycomb" có thể ám chỉ việc điền vào bằng cấu trúc tổ ong.
Nozzle_temperature	Đây là nhiệt độ của đầu phun trong quá trình in 3D. Nhiệt độ này ảnh hưởng đến việc nóng chảy nguyên liệu in và quá trình in tổng thể.
Bed_temperature	Đây là nhiệt độ của mặt giường hoặc nền sản phẩm in 3D. Nó có thể ảnh hưởng đến việc bám dính và nhiệt độ tổng thể của sản phẩm in.
Print_speed	Đây là tốc độ in 3D, đo bằng đơn vị millimet trên giây (mm/s). Tốc độ in ảnh hưởng đến thời gian in và chất lượng sản phẩm.
Material	Đây là loại nguyên liệu sử dụng để in sản phẩm 3D. Nó có thể là "abs" hoặc "pla," hai loại nguyên liệu phổ biến trong in 3D.

Fan_speed	Đây có thể là tốc độ của quạt được sử dụng trong quá trình làm nguội sản phẩm in 3D.
Roughness	Đây là độ xù lớp bề mặt của sản phẩm in 3D, thường được đo bằng đơn vị millimet (mm).
Tension_strenght	Đây có thể là sức căng của sản phẩm in 3D, đo bằng đơn vị MegaPascals (MPa). Nó đo lường khả năng chịu tải và kéo của sản phẩm.
Elongation	Đây là độ gia dãn của sản phẩm in 3D, thường được biểu thị dưới dạng phần trăm (%). Nó cho biết khả năng của sản phẩm để kéo dài mà không bị gãy.

- **Các loại biến:**

- Biến số liên tục: Các biến như layer_height, wall_thickness, nozzle-temperature, print-speed, roughness, tension_strenght và elongation và fan_speed

- Biến số rời rạc: Các biến như infill_density, infill_pattern, material

- **Các bước thực hiện:**

- *Bước 1.* Đọc dữ liệu (Import data):

- *Bước 2.* Làm sạch dữ liệu (Data cleaning)

- *Bước 3.* Làm rõ dữ liệu: (Data visualization

(a) Chuyển đổi biến (nếu cần thiết).

(b) Thống kê mô tả: dùng thống kê mẫu và dùng đồ thị.

- *Bước 4.* Dữ liệu này có thể được sử dụng để phân tích mối quan hệ giữa các biến để hiểu cách chúng ảnh hưởng đến chất lượng và tính chất của sản phẩm in 3D. Các thông số như layer_height, wall_thickness, infill_density, infill_pattern, nozzle_temperature, bed_temperature,….và các thông số về nguyên liệu và quá trình in khác có thể đã được đo lường trong quá trình thực hiện vì vậy các mô hình hồi quy sẽ là phương pháp nhóm chúng em sẽ tiếp cận sử dung tệp dữ liệu trong suốt quá trình làm bài tập.

2. KIẾN THỨC NỀN

2.1 Hồi quy

Hồi quy chính là một phương pháp thống kê để thiết lập mối quan hệ giữa một biến phụ thuộc và một nhóm tập hợp các biến độc lập. Mô hình với một biến phụ thuộc với hai hoặc nhiều biến độc lập được gọi là hồi quy bội (hay còn gọi là hồi quy đa biến). Ví dụ: Chi tiêu của hộ gia đình về thực phẩm phụ thuộc vào quy mô hộ gia đình, thu nhập, vị trí địa lý,…; Tỷ lệ tử vong trẻ em của một quốc gia phụ thuộc vào thu nhập bình quân đầu người, trình độ giáo dục,…; Lương của một người phụ thuộc vào chức vụ, kinh nghiệm, độ tuổi,…

a) Hồi quy tuyến tính bội

Mô hình hồi quy tuyến tính bội có dạng tổng quát như sau:

$$Y = \beta 1 + \beta 2X2 + \beta 3X3 + \ldots + \beta kXk + \varepsilon$$

Trong đó:

Y: biến phụ thuộc

Xi: biến độc lập

β1: hệ số tự do (hệ số chặn)

β1: hệ số hồi quy riêng. β1 đo lường tác động riêng phần của biến Xi lên Y với điều kiện các biến số khác trong mô hình không đổi. Cụ thể hơn, nếu các biến khác trong mô hình không đổi, giá trị kỳ vọng của Y sẽ tăng βi đơn vị nếu Xi tăng 1 đơn vị, ε: sai số ngẫu nhiên. Hồi quy đa biến là một phương pháp thống kê trong phân tích hồi quy, được sử dụng khi có nhiều hơn một biến độc lập (hoặc biến đầu vào) được sử dụng để dự đoán một biến phụ thuộc (hoặc biến đầu ra) không phải lúc nào cũng có biến phản hồi hoặc biến giải thích trong phân tích thống kê dữ liệu mối quan hệ tuyến tính là những mối quan hệ xuất phát từ 2 điểm tương đối với nhau dựa trên các yếu tố dự đoán của nó sau khi áp dụng và quy đa biến.

- **Đặc điểm chính của mô hình hồi quy đa biến**

- Nhiều biến độc lập: Trong hồi quy đa biến, có ít nhất hai hoặc nhiều hơn các biến độc lập, được ký hiệu bằng $x1, x2, x3, ..., xn$. Mỗi biến độc lập đại diện cho một yếu tố hoặc thuộc tính có thể ảnh hưởng đến biến phụ thuộc.

- Tính chuẩn tắc phần dư tuân theo lệnh phân phối chuẩn (chênh lệch giữa giá trị quan sát và giá trị dự toán).

- Tối ưu hóa: Trong hồi quy đa biến, quá trình tối ưu hóa được sử dụng để điều chỉnh giá trị của các hệ số mô hình sao cho sai số giữa dự đoán và giá trị thực tế là nhỏ nhất. Phương

pháp thông thường là tối thiểu hóa tổng sai số bình phương (RSS – Residual Sum of Squares) hoặc hàm mục tiêu tương tự.

- Sự thụ động này có thể thể hiện qua các hệ số hoặc trọng số (coefficients) trong mô hình hồi quy. Mỗi biến độc lập có một hệ số tương ứng, thể hiện mức độ ảnh hưởng của biến này đối với biến phụ thuộc. Hệ số dương cho biết mối quan hệ tích cực (tăng một biến dẫn đến tăng biến phụ thuộc), trong khi hệ số âm cho biết mối quan hệ tiêu cực (tăng một biến dẫn đến giảm biến phụ thuộc.

b) Kiểm định ý nghĩa của mô hình

- Trong mô hình hồi quy đa biến, giả thuyết "không" cho rằng mô hình không có ý nghĩa được hiểu là tất cả các hệ số hồi quy riêng đều bằng 0.

- Ứng dụng kiểm định Wald (thường được gọi là kiểm định F) được tiến hành cụ thể như sau:

Bước 1: Giả thuyết "không" là H0: $\beta 2 = \beta 3 = \ldots = \beta k = 0$.

Bước 2: Trước tiên hồi quy Y theo một số hạng không đổi và X2, X3, …, Xk, sau đó tính tổng bình phương sai số RSSU, RSSR. Phân phối F là tỷ số của hai biến ngẫu nhiên phân phối khi bình phương độc lập.

Bước 3: Tra số liệu trong bảng F tương ứng với bậc tự do (k – 1) cho tử số và (n – k) cho mẫu số, và với mức ý nghĩa α cho trước.

Bước 4: Bác bỏ giả thuyết H0 ở mức ý nghĩa α nếu Fc > F(α, k-1, n-k). Đối với phương pháp giá trị p, tính giá trị p = P (F>Fc|H0) và bác bỏ giả thuyết H0 nếu p < α.

c) Kiểm tra các giả định của mô hình

Nhắc lại các giả định của mô hình hồi quy:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_i X_i + \epsilon_i \text{ với } i = 1,\ldots, n.$$

- *Giả thuyết1:* Tính tuyến tính của dữ liệu: mối quan hệ giữa biến dự báo X và biến phụ thuộc Y được giả sử là tuyến tính.
- *Giả thuyết 2:* Sai số có phân phối chuẩn
- *Giả thuyết 3:* Phương sai của các sai số là hằng số.
- *Giả thuyết 4:* Các sai số ε có kỳ vọng = 0.
- *Giả thuyết 5:* Các sai số $\varepsilon_1,\ldots, \varepsilon_n$ thì độc lập với nhau.

2.2 Anova

- Mục tiêu của phân tích phương sai (Analysis of Variance ANOVA) là so sánh trung bình của nhiều nhóm (tổng thể) dựa trên các trị trung bình của các mẫu quan sát từ các nhóm này,

và thông qua kiểm định giả thuyết để kết luận về sự bằng nhau của các trung bình tổng thể này. Trong nghiên cứu, phân tích phương sai được dùng như một công cụ để xem xét ảnh hưởng của một yếu tố nguyên nhân (định tính) đến một yếu tố kết quả (định lượng).

• **Một số giả định khi phân tích ANOVA.**

- Các nhóm so sánh phải độc lập và được chọn một cách ngẫu nhiên.

- Các nhóm so sánh phải có phân phối chuẩn hoặc cỡ mẫu phải đủ lớn để được xem như tiệm cận phân phối chuẩn. Phương sai của các nhóm so sánh phải đồng nhất.

- *Lưu ý:* nếu giả định tổng thể có phân phối chuẩn với phương sai bằng nhau không đáp ứng được thì bạn có thể dùng kiểm định phi tham số Kruskal-Wallis sẽ để thay thế cho ANOVA. Sự biến thiên trong dữ liệu phân tích là mấu chốt để kiểm tra sự khác biệt về kì vọng giữa các nhóm

- Phân chia sự biến thiên: Sự biến thiên toàn phần trong dữ liệu, hay tổng bình phương toàn phần, bằng tổng các tổng bình phương nghiệm thức và tổng bình phương sai số

$$SST = SSW + SSB$$

$$\sum_{i=1}^{K}\sum_{j=1}^{n}(y_{ij}-y)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(y_{ij}-y_i)^2 + n\sum_{i=1}^{k}(y_i-y)^2$$

• ***Trong đó***

SST = Tổng bình phương toàn phần (Total Sum of Squares).

SSW = Tổng bình phương bên trong các các nhóm (Sum of Squares Within Groups).

SSB = Tổng bình phương giữa các nhóm (Sum of Squares Between Groups

- Tổng các bình phương của ANOVA với cỡ mẫu bằng nhau trong mỗi phương thức xử lý thường được tính bởi các công thức rút gọn sau:

$STT=\sum_{i=1}^{k}\sum_{j=1}^{n}y_{ij}^2 - \frac{y^2....}{N}$

$SSB=\sum_{i=1}^{k}\frac{y_i^2}{n} - \frac{y^2}{N}$

SSW=SST-SSB

• **Trung bình bình phương toàn phần:**

$$MST = \frac{SST}{kn-1}$$

• **Trung bình bình phương trong từng nhóm:** Tình phương sai trong nội bộ nhóm (MSW) bằng cách lấy tổng các chênh lệch bình phương trong nội bộ các nhóm (SSW) chia cho bậc tự do tương ứng là n-k (n là số quan sát, k là số nhóm so sánh). MSW là ước lượng phần biến thiên của yếu tố kết quả do các yếu tố khác gây ra (hay giải thích).

$$MSW = \frac{SSW}{k(n-1)}$$

• **Tính phương sai giữa các nhóm (MSG):** bằng cách lấy tổng các chênh lệch bình phương giữa các nhóm chia cho bậc tự do tương ứng là k – 1. MSG là ước lượng phần biến thiên của yếu tố kết quả do yếu tố nguyên nhân đang nghiên cứu gây ra (hay giải thích được).

$$MSG = \frac{SSG}{K-1}$$

- Ta chỉ so sánh bội hậu phân tích phương sai khi ANOVA khi giả thuyết H_0 bị bác bỏ, tức F>F_0 Phép phân tích phương sai ANOVA chỉ cho ra sự khác biệt giữa các nhóm, nhưng không chỉ ra được cốt lõi sự khác biệt (nhóm hoặc các nhóm gây ra sự khác biệt). Để chỉ ra được sự khác biệt do kỳ vọng nhóm nào gây ra, ta dùng so sánh bội. Ở phép ANOVA một nhân tố, ta sẽ sử dụng so sánh bội đơn giản: phương pháp ý nghĩa độ lệch nhỏ nhất LSD.

3. TIỀN XỬ LÝ SỐ LIỆU

3.1 Đọc dữ liệu

Đọc tệp tin "data.csv" và gán với tên data

Code:

```
data<-read.csv("C:/Users/DELL/Downloads/archive/data.csv")
View(data)
head(data,10)
```

Giải thích:

- Đọc dữ liệu và lưu với tên data

- Trích 10 dòng đầu của dữ liệu data

Kết quả

```
   layer_height wall_thickness infill_density infill_pattern nozzle_temperature
1          0.02              8             90          grid                220
2          0.02              7             90     honeycomb                225
3          0.02              1             80          grid                230
4          0.02              4             70     honeycomb                240
5          0.02              6             90          grid                250
6          0.02             10             40     honeycomb                200
7          0.02              5             10          grid                205
8          0.02             10             10     honeycomb                210
9          0.02              9             70          grid                215
10         0.02              8             40     honeycomb                220
   bed_temperature print_speed material fan_speed roughness tension_strenght elongation
1               60          40      abs         0        25               18        1.2
2               65          40      abs        25        32               16        1.4
3               70          40      abs        50        40                8        0.8
4               75          40      abs        75        68               10        0.5
5               80          40      abs       100        92                5        0.7
6               60          40      pla         0        60               24        1.1
7               65          40      pla        25        55               12        1.3
8               70          40      pla        50        21               14        1.5
9               75          40      pla        75        24               27        1.4
10              80          40      pla       100        30               25        1.7
>
```

Hình 3.1 Kết quả khi xem 10 dòng đầu tiên của tệp tin "data.csv"

3.2 Xử lý dữ liệu khuyết

Kiểm tra dữ liệu khuyết trong data

```
> apply(is.na(data),2,sum)
      layer_height     wall_thickness     infill_density     infill_pattern
                 0                  0                  0                  0
nozzle_temperature    bed_temperature        print_speed           material
                 0                  0                  0                  0
         fan_speed          roughness   tension_strenght         elongation
                 0                  0                  0                  0
```

Hình 3.2 Kết quả đếm dữ liệu khuyết trong tệp tin data

➔ Không tìm thấy dữ liệu khuyết trong tệp tin data

3.3 Chuyển đổi biến về dạng factor

2 biến : " infill_ pattern", " matarial" về dạng factor

```
> data$infill_pattern=as.factor(data$infill_pattern)
> summary(data$infill_pattern)
    grid honeycomb
      25        25
> data$material=as.factor(data$material)
> summary(data$material)
abs pla
 25  25
```

Hình 3.3 Kết quả khi dùng factor chuyển đổi 2 biến " infill_ pattern", " matarial"

4. THỐNG KÊ TẢ

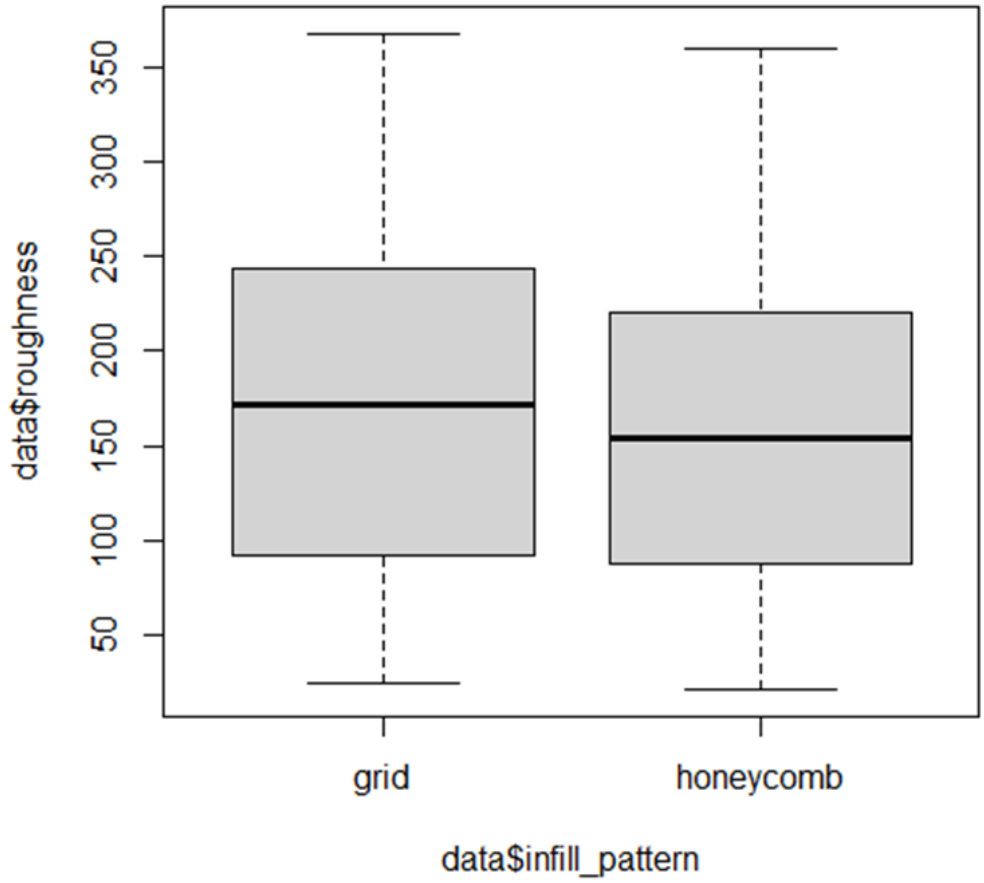
4.1 Thống kê tả dạng bảng

```
> summary(data)
  layer_height   wall_thickness  infill_density   infill_pattern nozzle_temperature
 Min.   :0.020   Min.   : 1.00   Min.   :10.0   grid     :25     Min.   :200.0
 1st Qu.:0.060   1st Qu.: 3.00   1st Qu.:40.0   honeycomb:25     1st Qu.:210.0
 Median :0.100   Median : 5.00   Median :50.0                    Median :220.0
 Mean   :0.106   Mean   : 5.22   Mean   :53.4                    Mean   :221.5
 3rd Qu.:0.150   3rd Qu.: 7.00   3rd Qu.:80.0                    3rd Qu.:230.0
 Max.   :0.200   Max.   :10.00   Max.   :90.0                    Max.   :250.0
 bed_temperature  print_speed  material   fan_speed     roughness
 Min.   :60      Min.   : 40   abs:25   Min.   :  0   Min.   : 21.0
 1st Qu.:65      1st Qu.: 40   pla:25   1st Qu.: 25   1st Qu.: 92.0
 Median :70      Median : 60            Median : 50   Median :165.5
 Mean   :70      Mean   : 64            Mean   : 50   Mean   :170.6
 3rd Qu.:75      3rd Qu.: 60            3rd Qu.: 75   3rd Qu.:239.2
 Max.   :80      Max.   :120            Max.   :100   Max.   :368.0
 tension_strenght   elongation
 Min.   : 4.00    Min.   :0.400
 1st Qu.:12.00    1st Qu.:1.100
 Median :19.00    Median :1.550
 Mean   :20.08    Mean   :1.672
 3rd Qu.:27.00    3rd Qu.:2.175
 Max.   :37.00    Max.   :3.300
```

Hình 4.1 Kết quả khi xem 5 dòng đầu tiên của tệp tin "data.csv"

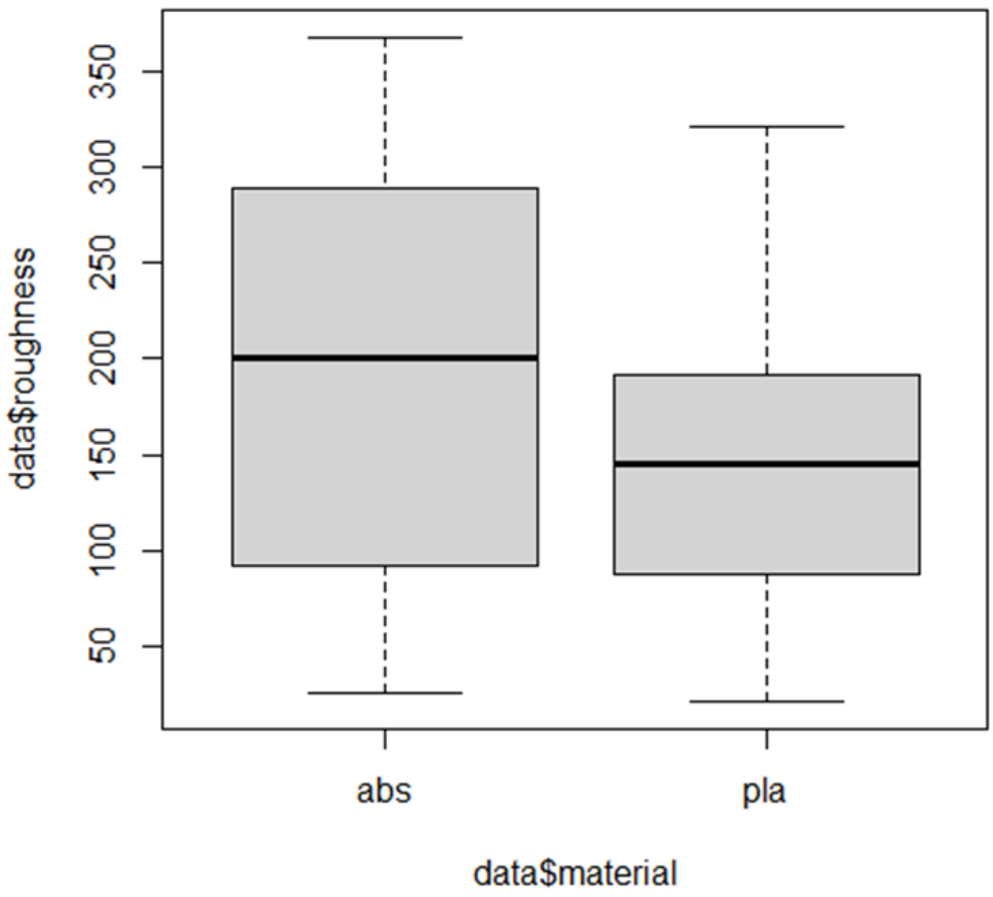
4.2 Một số đồ thị

a) Đồ thị Boxplot



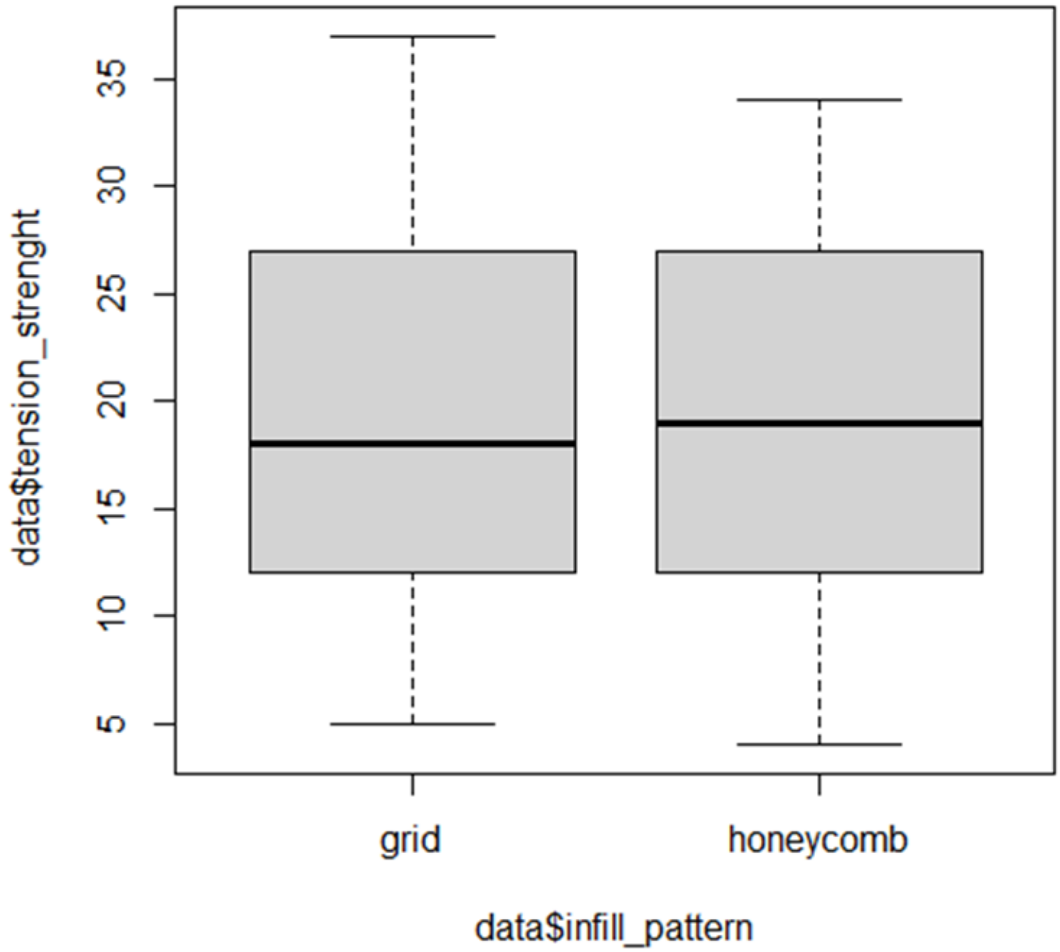
Hình 4.2.1. Đồ thị Boxplot của infill_pattern và roughness

Ở hình 1 ta có thể thấy mức trung vị của grid cao hơn so với mức trung vị của honeycomb. Tuy nhiên nhìn chung hai đồ thị là tương đương nhau và không có điểm cụ thể nào phân biệt rỏ mức độ ảnh hưởng của infill_pattern tác động đến roughness.



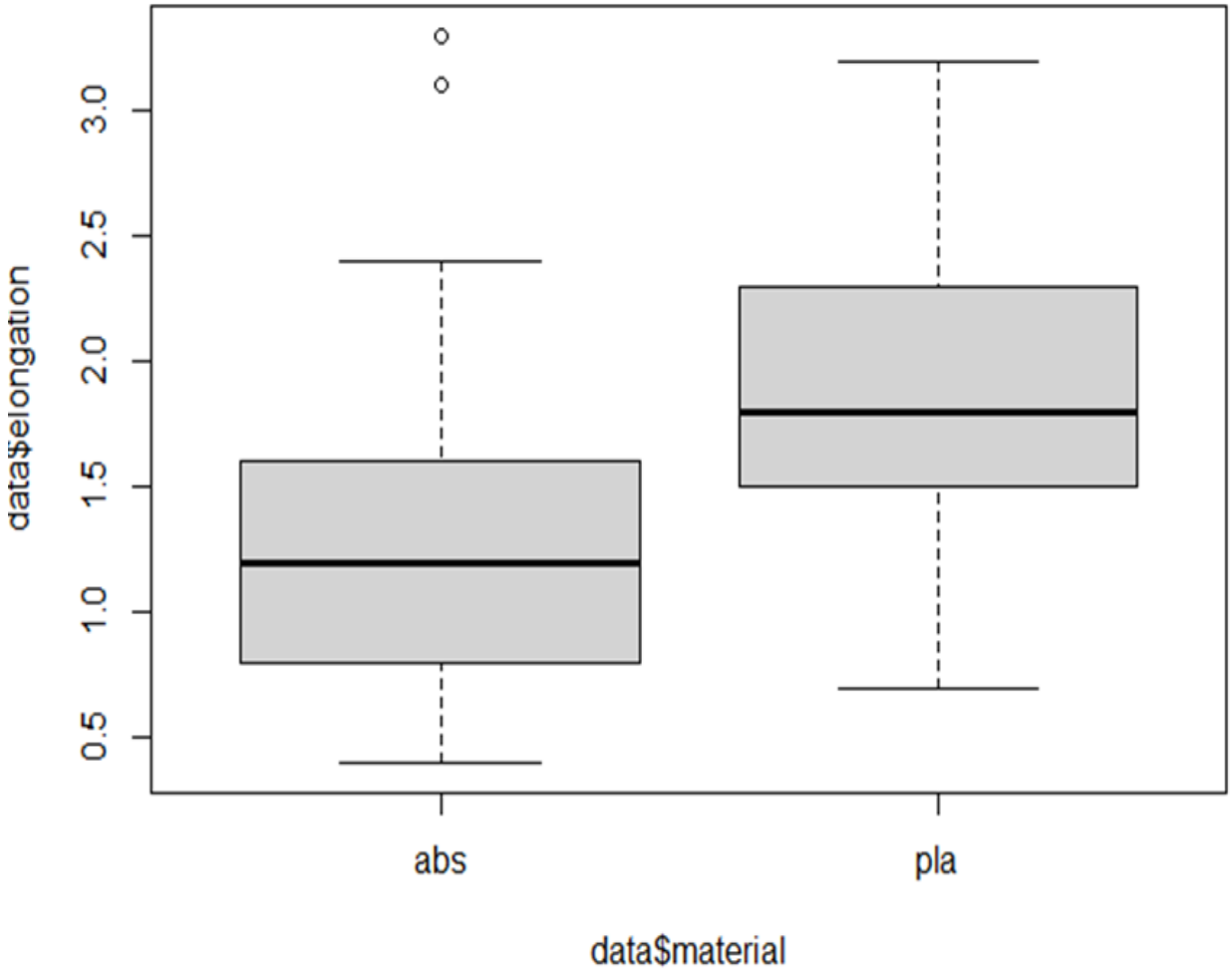
Hình 4.2.2. Đồ thị Boxplot của material và roughness

Về boxplot của material và roughness cả hai đều phân bố lệch so với mức trung vị. Ở "abs" các giá trị roughness phân bố từ khoảng 90 đến 280. Trong khi "pla" chỉ phân bố đến bé hơn mức trung vị của "abs".



Hình 4.2.3. Đồ thị của infill_patter và tension_strenght

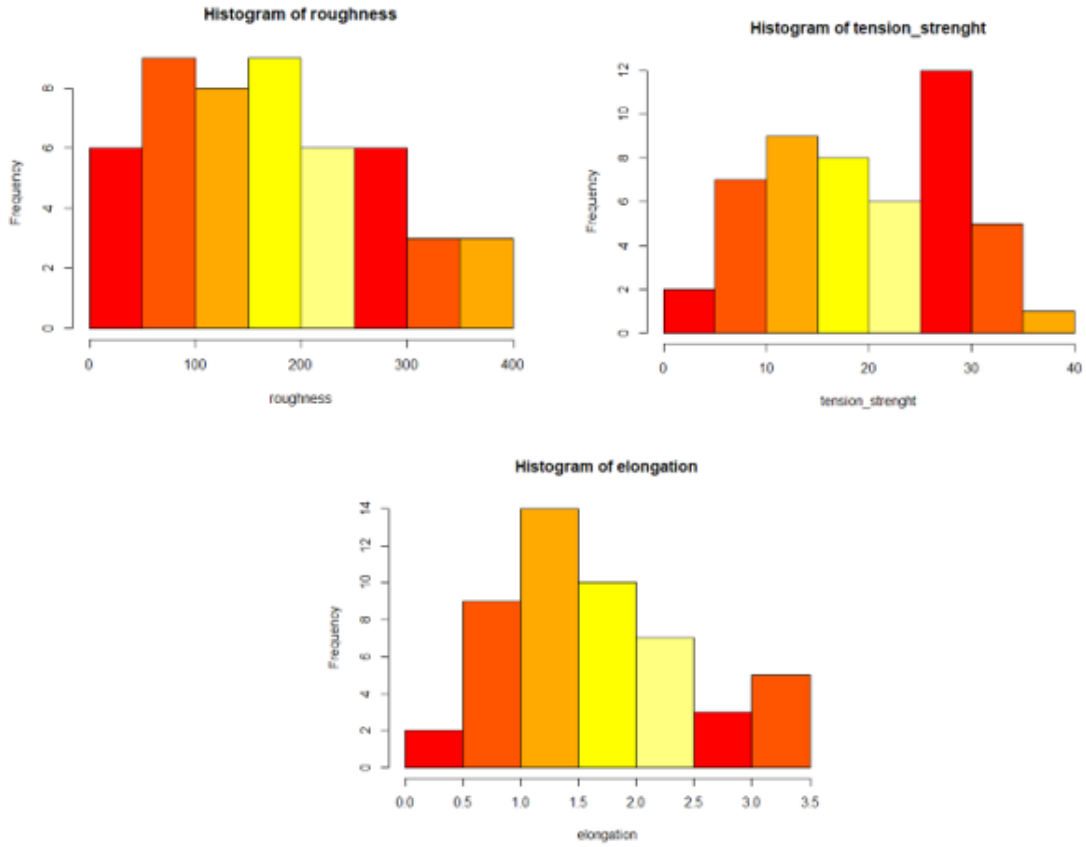
Ở hình 3 ta có thể thấy được 2 đồ thi gần như giống nhau chỉ có sự chênh lệch nhỏ về mức trung vị của 2 đồ thị cụ thể là mức trung vị của "honeycomb" cao hơn so với mức trung vị của "grid".



Hình 4.4.4. Đồ thị Boxplot của material và elongation

Về boxplot của dữ liệu material. Cả 2 boxplot đều phân bố lệch so với trung vị. Ở "abs" có điểm ngoại lai và hầu như các giá trị của "pla" đều lớn hơn các giá trị của "abs".

b) Đồ thị Histogram

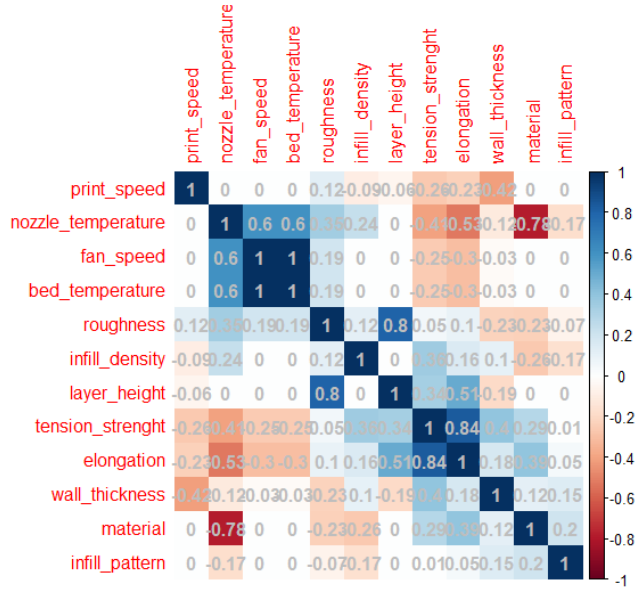


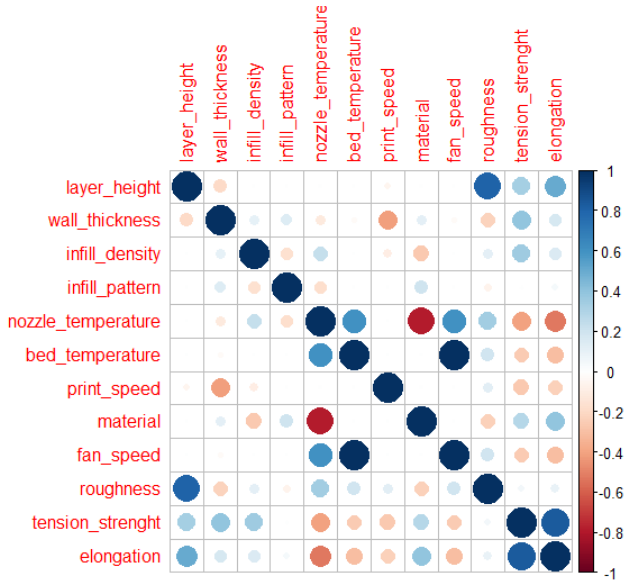
Hình 4.4.5. Đồ thị histogram của roughness, tension_strenght và elongation

Từ hình trên ta có thể thấy các đồ thị phân bố không đều. Như ở đồ thị histogram của roughness các giá trị lớn thường tập trung trong khoảng từ 50 đến 200. Trong khi đó ở đồ thị histogram của tension_strenght thì lại tập trung về phía bên trái từ 25 đến 30. Còn đồ thị histogram của elongation các giá trị lớn lại tập trung về giữa từ khỏag 1 đến 2.

c) Hệ số tương quan của các biến.

Để thấy mối quan hệ tuyến tính giữa từng biến, ta sẽ vẽ hệ số tương quan của tất cả các biến:





5. THỐNG KÊ SUY DIỄN *"Xây dựng mô hình hồi quy tuyến tính bội và Anova so sánh mô hình"*

Các biến " roughness", " tension_strenght", "elongation" là các biến phụ thuộc, còn lại là các biến độc lập.

5.1 " roughness"

a) Xây dựng mô hình và anova tìm mô hình lý tưởng nhất

Dùng lệnh lm () để xây dựng mô hình hồi quy tuyến tính bội và dùng lệnh summary để tóm tắt kết quả thu được:

- **Code và kết quả model_1**

```
> model_1<-lm(roughness~layer_height+wall_thickness+infill_density+nozzle_temperature+b
ed_temperature+infill_pattern+print_speed+material,data)
> summary(model_1)

Call:
lm(formula = roughness ~ layer_height + wall_thickness + infill_density +
    nozzle_temperature + bed_temperature + infill_pattern + print_speed +
    material, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-72.746 -24.332  -1.641  20.304  96.552

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)             -2.371e+03  3.716e+02  -6.379 1.25e-07 ***
layer_height             1.269e+03  8.765e+01  14.483  < 2e-16 ***
wall_thickness           2.334e+00  2.189e+00   1.066  0.29259
infill_density          -4.231e-02  2.341e-01  -0.181  0.85742
nozzle_temperature       1.506e+01  2.529e+00   5.953 5.05e-07 ***
bed_temperature         -1.613e+01  3.251e+00  -4.962 1.27e-05 ***
infill_patternhoneycomb -1.255e-01  1.128e+01  -0.011  0.99117
print_speed              6.496e-01  2.060e-01   3.153  0.00302 **
materialpla              2.985e+02  5.836e+01   5.114 7.78e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 38.24 on 41 degrees of freedom
Multiple R-squared:  0.8752,    Adjusted R-squared:  0.8509
F-statistic: 35.95 on 8 and 41 DF,  p-value: 3.834e-16
```

Hình 5.1.1.1 Kết quả mô hình hồi quy tuyến tính model_1

Kiểm định hệ số hồi quy *(Dùng p-value :mức ý nghĩa quan sát, xác suất quan sát)*

- Nếu p-value < α ⇒ bác bỏ H0, chấp nhận H1

- Nếu p-value ≥ α ⇒ chưa bác bỏ H0

- Giả thuyết H0: Hệ số hồi quy không có ý nghĩa thống kê ($\beta i = 0$)

- Giả thuyết H1: Hệ số hồi quy có ý nghĩa thống kê ($\beta i \neq 0$)

- Đối với mức tin vậy 5%:

+ Pr (>|t|) của các hệ số ứng với biến layer_height, nozzle_temperature, bed_temperature, print_speed, materia, bé hơn mức ý nghĩa α = 0,05 nên ta bác bỏ H0 và chấp nhận H1. Do đó các hệ số ứng với biến này có ý nghĩa thống kê đối với mô hình hồi quy mà ta xây dựng.

+ Pr (>|t|) của các hệ số ứng với biến infill_pattern, wall_thickness, infill_density, lớn hơn mức ý nghĩa α = 0,05 nên ta chưa thể bác bỏ H0. Do đó các hệ số này ứng với các biến này không có ý nghĩa thống kê với mô hình hồi quy mà ta xây dựng, có thể cân nhắc để loại bỏ các biến wall_thickness, infill_density, infill_pattern.

- **Code và kết quả model_2** (bỏ các biến wall_thickness, infill_density,infill_pattern.)

```
> model_2<-lm(roughness~layer_height+nozzle_temperature+bed_temperature+print_speed+mat
erial,data)
> summary(model_2)

Call:
lm(formula = roughness ~ layer_height + nozzle_temperature +
    bed_temperature + print_speed + material, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-74.084 -26.500  -1.662  22.585  92.356

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        -2310.7356   353.2009  -6.542 5.38e-08 ***
layer_height        1246.5353    83.1780  14.986  < 2e-16 ***
nozzle_temperature    14.7774     2.3979   6.163 1.95e-07 ***
bed_temperature      -15.8078     3.0895  -5.117 6.55e-06 ***
print_speed            0.5538     0.1804   3.070  0.00366 **
materialpla          294.1610    56.1586   5.238 4.38e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 37.44 on 44 degrees of freedom
Multiple R-squared:  0.8717,    Adjusted R-squared:  0.8571
F-statistic: 59.78 on 5 and 44 DF,  p-value: < 2.2e-16
```

Hình 5.1.1.2 Kết quả mô hình hồi quy tuyến tính model_2

- **So sánh Model_1 và Model_2**

```
> anova(model_1,model_2)
Analysis of Variance Table

Model 1: roughness ~ layer_height + wall_thickness + infill_density +
    nozzle_temperature + bed_temperature + infill_pattern + print_speed +
    material
Model 2: roughness ~ layer_height + nozzle_temperature + bed_temperature +
    print_speed + material
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1     41 59968
2     44 61667 -3   -1699.4 0.3873 0.7627
```

Hình 5.1.1.3 Kết quả so sánh model_1 và model_2

Nhận xét:

Giả thuyết H0: model_2 hiệu quả hơn

Giả thuyết H1: model_1 hiệu quả hơn

→ Ta nhận thấy giá trị Pr (>F) bằng o.7627 lớn hơn mức ý nghĩa α = 0,05 nên chưa bác bỏ được giả thuyết H0 , nên model_2 hiệu quả hơn.

b) Kiểm tra các giả định của mô hình model_2

Nhắc lại các giả định của mô hình hồi quy:

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_i X_i + \epsilon_i$ *với* i = 1,..., n.

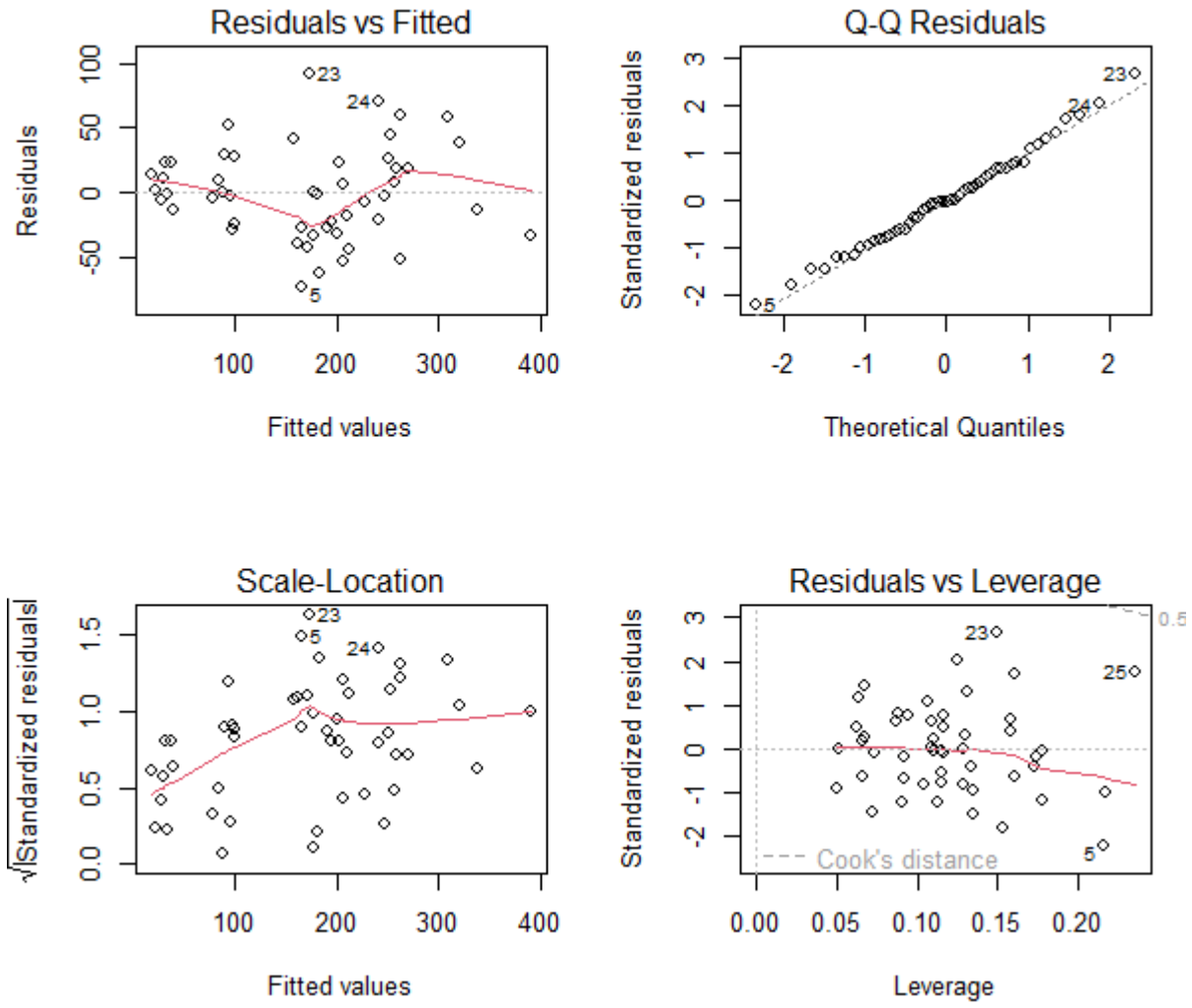
- *Giả thuyết1:* Tính tuyến tính của dữ liệu: mối quan hệ giữa biến dự báo X và biến phụ thuộc Y được giả sử là tuyến tính.
- *Giả thuyết 2:* Sai số có phân phối chuẩn
- *Giả thuyết 3:* Phương sai của các sai số là hằng số.
- *Giả thuyết 4:* Các sai số ε có kỳ vọng = 0.
- *Giả thuyết 5:* Các sai số ε_1,..., ε_n thì độc lập với nhau.

<u>Cách 1:</u> Thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình

Code:

```
> par(mfrow=c(2,2))
> plot(model_2)
```

Kết quả:



Hình 5.1.1.4 Kết quả khi vẽ các đồ thị phân tích thặng dư

Nhận xét:

- Đồ thị Residuals vs Fitted là đường cong có độ dóc chưa thỏa được ***giả thuyết 1***
- Đồ thị Normal Q-Q: kiểm tra giả định về phân phối chuẩn của các sai số. Nếu các điểm sai số nằm trên cùng một đường thẳng thì điều kiện về ***giả thuyết 2*** phân phối chuẩn được thoả.
- Đồ thị Scale-Location: vẽ căn bậc hai của các sai số được chuẩn hoá bởi các giá trị dự báo, được dùng để kiểm tra ***giả thuyết 3*** (phương sai của các sai số là hằng số), các điểm đường màu đỏ có độ dốc và các điểm thặng dư phân tán không đều xung quanh đường thẳng này nên **giả thiết 3** bị vi phạm
- Đồ thị Residuals vs Leverage: Các điểm thứ 5, 23 và 25 là những điểm ảnh hưởng cao trong bộ dữ liệu. Tuy nhiên những điểm ảnh hưởng cao này chưa vượt qua đường thẳng khoảng cách Cook (Cook's distance) nên chúng không phải là các điểm outliers và ta không cần loại bỏ chúng khỏi bộ dữ liệu.

<u>Cách 2:</u> Kiểm tra các gải thiết dựa vào các kiểm định:

Giả thiết 2: Sai số có phân phối chuẩn.

H_0: Các sai số hồi quy có phân phối chuẩn.

H_1: Dữ liệu không có phân phối chuẩn.

```
> re<-residuals(model_2)
> shapiro.test(re)

        Shapiro-Wilk normality test

data:  re
W = 0.99094, p-value = 0.9656
```

Hình 5.1.1.5 Kết quả kiểm tra giả thiết sai số có phân phối chuẩn

Nhận xét:

- Từ kết quả trên ta có p - $value$ = 1 > α = 5%, nên không bác bỏ H0. Vậy giả thiết 2: Sai số có phân phối chuẩn thỏa mãn.

Giả thiết 4: Các sai số ε có kỳ vọng = 0.

H0: Các sai số có kỳ vọng μ = 0

H1: Các sai số có kỳ vọng μ ≠ 0\

```
> re<-residuals(model_2)
> t.test(re,mu=0)

        One Sample t-test

data:  re
t = -9.9042e-17, df = 49, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -10.08202  10.08202
sample estimates:
    mean of x
-4.968942e-16
```

Hình 5.1.1.6 Kết quả kiểm tra giả thiết sai số hồi quy

Giả thuyết:

H_0: Các sai số có kỳ vọng μ = 0

H_1: Các sai số có kỳ vọng μ ≠ 0

Miền bác bỏ

$$RR = \left(-\infty; -t_{\frac{\alpha}{2}}^{n-1}\right) \cup \left(t_{\frac{\alpha}{2}}^{n-1}; +\infty\right) = (-\infty; -t_{0.025}^{50-1}) \cup (t_{0.025}^{50-1}; +\infty)$$

$$= (-\infty;\ -10)\ \cup\ (10;\ +\infty)$$

Vì n − 1 = 49 ≥ 30 nên $t_{\frac{\alpha}{2}}^{n-1} \approx z_{\alpha/2}$

Tiêu chuẩn kiểm định: $z_0 = \frac{\bar{X}-\mu_0}{S/\sqrt{n}}$ = -9.9043e-17

Cách 2.1: Kiểm định theo tiêu chuẩn kiểm định

Từ kết quả R cho ta thấy z_0 không thuộc miền bác bỏ, chưa bác bỏ được giả thuyết H_0 nên giả định về các sai số có kỳ vọng bằng 0 được thoả mãn.

Cách 2.2: Kiểm định theo p-value

Ta nhận thấy p-value =1 chưa bác bỏ được giả thuyết H_0, nên giả định về các sai số có kỳ vọng bằng 0 được thoả mãn.

5.2 “tension_strenght”

a) Xây dựng mô hình và anova tìm mô hình lý tưởng nhất

- **Code và kết quả model_3**

```
> model_3<-lm(tension_strenght~layer_height+wall_thickness+infill_density+nozzle_temper
ature+bed_temperature+infill_pattern+print_speed+material,data)
> summary(model_3)

Call:
lm(formula = tension_strenght ~ layer_height + wall_thickness +
    infill_density + nozzle_temperature + bed_temperature + infill_pattern +
    print_speed + material, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4101 -3.8491  0.0338  3.9073 13.5086

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           171.62809   54.21779   3.166  0.00292 **
layer_height           55.59721   12.78771   4.348 8.87e-05 ***
wall_thickness          1.06871    0.31942   3.346  0.00176 **
infill_density          0.16286    0.03415   4.769 2.35e-05 ***
nozzle_temperature     -1.04681    0.36901  -2.837  0.00705 **
bed_temperature         1.00534    0.47426   2.120  0.04012 *
infill_patternhoneycomb -1.14271    1.64581  -0.694  0.49140
print_speed            -0.01559    0.03006  -0.519  0.60679
materialpla           -17.30508    8.51530  -2.032  0.04864 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.58 on 41 degrees of freedom
Multiple R-squared:  0.673,     Adjusted R-squared:  0.6092
F-statistic: 10.55 on 8 and 41 DF,  p-value: 6.91e-08
```

Hình 5.1.2.1 Kết quả mô hình hồi quy tuyến tính model_3

Kiểm định hệ số hồi quy (Dùng p-value :mức ý nghĩa quan sát, xác suất quan sát)

Nếu p-value < α ⇒ bác bỏ H0, chấp nhận H1.

Nếu p-value ≥ α ⇒ chưa bác bỏ H0.

- Giả thuyết H0: Hệ số hồi quy không có ý nghĩa thống kê ($\beta i = 0$).

- Giả thuyết H1: Hệ số hồi quy có ý nghĩa thống kê ($\beta i \neq 0$).

- Đối với mức tin vậy 5%.

+ Pr (>|t|) của các hệ số ứng với biến: layer_height, materia, wall_thickness, infill_density,nozzle_temperature, bed_temperature , bé hơn mức ý nghĩa $\alpha = 0{,}05$ nên ta bác bỏ H0 và chấp nhận H1. Do đó các hệ số ứng với biến này có ý nghĩa thống kê đối với mô hình hồi quy mà ta xây dựng.

+ Pr (>|t|) của các hệ số ứng với biến: infill_pattern, print_speed, lớn hơn mức ý nghĩa $\alpha =$ 0,05 nên ta chưa thể bác bỏ H0. Do đó các hệ số này ứng với các biến này không có ý nghĩa

thống kê với mô hình hồi quy mà ta xây dựng, có thể cân nhắc để loại bỏ các biến :infill_pattern, print_speed.

- **Code và kết quả model_4** (bỏ các biến infill_pattern, print_speed)

```
> model_4<-lm(tension_strenght~layer_height+wall_thickness+infill_density+nozzle_temper
ature+bed_temperature+material,data)
> summary(model_4)

Call:
lm(formula = tension_strenght ~ layer_height + wall_thickness +
    infill_density + nozzle_temperature + bed_temperature + material,
    data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-9.4717 -4.1695  0.0914  3.8586 13.9581

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       167.02463   53.19880   3.140 0.003055 **
layer_height       56.36695   12.44845   4.528 4.67e-05 ***
wall_thickness      1.11169    0.28024   3.967 0.000271 ***
infill_density      0.16643    0.03339   4.985 1.06e-05 ***
nozzle_temperature -1.02890    0.36321  -2.833 0.006996 **
bed_temperature     0.98346    0.46685   2.107 0.041026 *
materialpla       -17.10365    8.39202  -2.038 0.047722 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.501 on 43 degrees of freedom
Multiple R-squared:  0.6667,    Adjusted R-squared:  0.6201
F-statistic: 14.33 on 6 and 43 DF,  p-value: 6.783e-09
```

Hình 5.1.2.2 Kết quả mô hình hồi quy tuyến tính model_4

- **So sánh model_3, model_4**

```
> anova(model_3,model_4)
Analysis of Variance Table

Model 1: tension_strenght ~ layer_height + wall_thickness + infill_density +
    nozzle_temperature + bed_temperature + infill_pattern + print_speed +
    material
Model 2: tension_strenght ~ layer_height + wall_thickness + infill_density +
    nozzle_temperature + bed_temperature + material
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     41 1276.5
2     43 1301.3 -2   -24.756 0.3976 0.6745
```

Hình 5.1.2.3 Kết quả so sánh model_3 và model_4

Nhận xét:

Giả thuyết H0: model_4 hiệu quả hơn

Giả thuyết H1: model_3 hiệu quả hơn

➔ Ta nhận thấy giá trị Pr (>F) bằng 0.6745 lớn hơn mức ý nghĩa α = 0,05 nên chưa bác bỏ được giả thuyết H0 , nên model_4 hiệu quả hơn

b) Kiểm tra các giả định của mô hình model_4

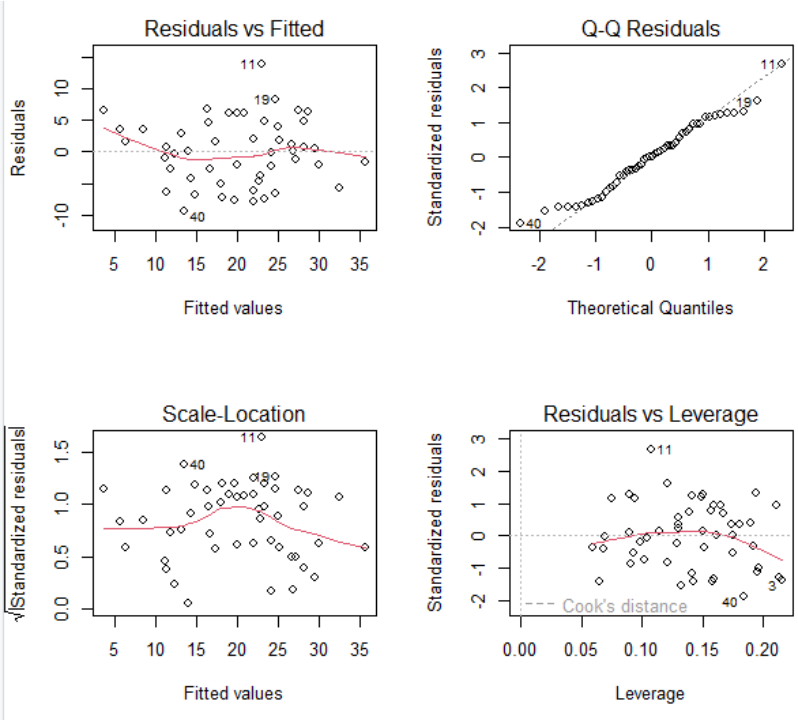
Nhắc lại các giả định của mô hình hồi quy:

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_i X_i + \epsilon_i \text{ với } i = 1,\ldots, n.$$

- *Giả thuyết 1:* Tính tuyến tính của dữ liệu: mối quan hệ giữa biến dự báo X và biến phụ thuộc Y được giả sử là tuyến tính.
- *Giả thuyết 2:* Sai số có phân phối chuẩn
- *Giả thuyết 3:* Phương sai của các sai số là hằng số.
- *Giả thuyết 4:* Các sai số ε có kỳ vọng = 0.
- *Giả thuyết 5:* Các sai số $\varepsilon_1,\ldots, \varepsilon_n$ thì độc lập với nhau.

Cách 1: Thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình

Kết quả



Hình 5.1.2.4 Kết quả khi vẽ các đồ thị phân tích thặng dư

Nhận xét:

- Đồ thị Residuals vs Fitted: Ta nhận thấy đường màu đỏ gần như là đường thẳng nên giả định tuyến tính của dữ liệu thỏa mãn.
- Đồ thị Normal Q-Q: kiểm tra giả định về phân phối chuẩn của các sai số, các điểm sai số không nằm trên cùng một đường thẳng thì điều kiện về ***giả thuyết 2*** phân phối chuẩn không được thoả.

- Đồ thị Scale-Location: vẽ căn bậc hai của các sai số được chuẩn hoá bởi các giá trị dự báo, được dùng để kiểm tra ***giả thuyết 3*** (phương sai của các sai số là hằng số), các điểm đường màu đỏ có độ dốc và các điểm thặng dư phân tán không đều xung quanh đường thẳng này nên **giả thiết 3** bị vi phạm
- Đồ thị Residuals vs Leverage: Các điểm thứ 3, 11 và 40 là những điểm ảnh hưởng cao trong bộ dữ liệu. Tuy nhiên những điểm ảnh hưởng cao này chưa vượt qua đường thẳng khoảng cách Cook (Cook's distance) nên chúng không phải là các điểm outliers và ta không cần loại bỏ chúng khỏi bộ dữ liệu

<u>Cách 2:</u> ***Kiểm tra các gải thiết dựa vào các kiểm định:***

Giả thiết 2: Sai số có phân phối chuẩn

H_0: Các sai số hồi quy có phân phối chuẩn.

H_1: Dữ liệu không có phân phối chuẩn.

```
> re<-residuals(model_4)
> shapiro.test(re)

        Shapiro-Wilk normality test

data:  re
W = 0.97458, p-value = 0.3517
```

Hình 5.1.2.5 Kết quả kiểm tra giả thiết sai số có phân phối chuẩn

Nhận xét:

➢ Từ kết quả trên ta có p - $value$ = 0.3517> α = 5%, nên không bác bỏ H_0,. Vậy giả thiết 2: Sai số có phân phối chuẩn thỏa mãn.

Giả thuyết 4 :

H_0: Các sai số có kỳ vọng $\mu = 0$

H_1: Các sai số có kỳ vọng $\mu \neq 0$

```
> re<-residuals(model_4)
> t.test(re,mu=0)

        One Sample t-test

data:  re
t = -7.5559e-16, df = 49, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.464546  1.464546
sample estimates:
    mean of x
-5.506598e-16
```

Hình 5.1.2.6 Kết quả kiểm tra giả thiết sai số hồi quy

Miền bác bỏ

$$RR = \left(-\infty; -t_{\frac{\alpha}{2}}^{n-1}\right) \cup \left(t_{\frac{\alpha}{2}}^{n-1}; +\infty\right) = (-\infty; -t_{0.025}^{50-1}) \cup (t_{0.025}^{50-1}; +\infty)$$

$$= (-\infty;\ -1{,}46)\ \cup\ (1{,}46;\ +\infty)$$

Vì $n - 1 = 49 \geq 30$ nên $t_{\frac{\alpha}{2}}^{n-1} \approx z_{\alpha/2}$

Tiêu chuẩn kiểm định: $z_0 = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} =$ -7.5559e-16

Cách 2.1: Kiểm định theo tiêu chuẩn kiểm định

Từ kết quả R cho ta thấy z_0 không thuộc miền bác bỏ, chưa bác bỏ được giả thuyết H_0 nên giả định về các sai số có kỳ vọng bằng 0 được thoả mãn.

Cách 2.2: Kiểm định theo p-value

Ta nhận thấy p-value =1 chưa bác bỏ được giả thuyết H_0, nên giả định về các sai số có kỳ vọng bằng 0 được thoả mãn.

5.3 “elongation”

a) Xây dựng mô hình và anova tìm mô hình lý tưởng nhất

- **Code và kết quả model_5**

```
> model_5<-lm(elongation~layer_height+wall_thickness+infill_density+nozzle_temperature+
bed_temperature+infill_pattern+print_speed+material,data)
> summary(model_5)

Call:
lm(formula = elongation ~ layer_height + wall_thickness + infill_density +
    nozzle_temperature + bed_temperature + infill_pattern + print_speed +
    material, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99066 -0.37634  0.03137  0.26680  0.86971

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           18.435640   4.466801   4.127 0.000175 ***
layer_height           6.382037   1.053532   6.058 3.58e-07 ***
wall_thickness         0.030827   0.026316   1.171 0.248194
infill_density         0.009890   0.002813   3.516 0.001086 **
nozzle_temperature    -0.109642   0.030401  -3.607 0.000834 ***
bed_temperature        0.104223   0.039073   2.667 0.010894 *
infill_patternhoneycomb -0.062282   0.135592  -0.459 0.648422
print_speed           -0.003375   0.002477  -1.363 0.180405
materialpla           -1.783730   0.701544  -2.543 0.014876 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4597 on 41 degrees of freedom
Multiple R-squared:  0.7154,    Adjusted R-squared:  0.6598
F-statistic: 12.88 on 8 and 41 DF,  p-value: 4.765e-09
```

Hình 5.1.3.1 Kết quả mô hình hồi quy tuyến tính model_5

Kiểm định hệ số hồi quy (Dùng p-value :mức ý nghĩa quan sát, xác suất quan sát)

Nếu p-value < α ⇒ bác bỏ H_0, chấp nhận H_1.

Nếu p-value ≥ α ⇒ chưa bác bỏ H_0.

- Giả thuyết H0: Hệ số hồi quy không có ý nghĩa thống kê ($\beta i = 0$).
- Giả thuyết H1: Hệ số hồi quy có ý nghĩa thống kê ($\beta i \neq 0$).
- Đối với mức tin vậy 5%.

+ Pr (>|t|) của các hệ số ứng với biến: layer_height ,nozzle_temperature, materia, infill_density, bed_temperature, bé hơn mức ý nghĩa α = 0,05 nên ta bác bỏ H0 và chấp nhận H1. Do đó các hệ số ứng với biến này có ý nghĩa thống kê đối với mô hình hồi quy mà ta xây dựng.

+ Pr (>|t|) của các hệ số ứng với biến: infill_pattern, print_speed, lớn hơn mức ý nghĩa α = 0,05 nên ta chưa thể bác bỏ H0. Do đó các hệ số này ứng với các biến này không có

ý nghĩa thống kê với mô hình hồi quy mà ta xây dựng, có thể cân nhắc để loại bỏ các biến :infill_pattern, print_speed.

- **Code và kết quả model_6**

```
> model_6<-lm(elongation~layer_height+infill_density+nozzle_temperature+bed_temperature
+material,data)
> summary(model_6)

Call:
lm(formula = elongation ~ layer_height + infill_density + nozzle_temperature +
    bed_temperature + material, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.09844 -0.31878 -0.01446  0.37965  0.87425

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        18.88608    4.54962   4.151 0.000149 ***
layer_height        6.19739    1.05168   5.893 4.85e-07 ***
infill_density      0.01103    0.00284   3.885 0.000340 ***
nozzle_temperature -0.11352    0.03116  -3.643 0.000707 ***
bed_temperature     0.10870    0.04009   2.712 0.009515 **
materialpla        -1.84929    0.72176  -2.562 0.013901 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4741 on 44 degrees of freedom
Multiple R-squared:  0.6751,    Adjusted R-squared:  0.6382
F-statistic: 18.29 on 5 and 44 DF,  p-value: 8.722e-10
```

Hình 5.1.3.2 Kết quả mô hình hồi quy tuyến tính model_6

- **So sánh model_5, model_6**

```
> anova(model_5,model_6)
Analysis of Variance Table

Model 1: elongation ~ layer_height + wall_thickness + infill_density +
    nozzle_temperature + bed_temperature + infill_pattern + print_speed +
    material
Model 2: elongation ~ layer_height + infill_density + nozzle_temperature +
    bed_temperature + material
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     41 8.6643
2     44 9.8887 -3   -1.2245 1.9314 0.1396
```

Hình 5.1.3.3 Kết quả so sánh model_5 và model_6

- **Nhận xét:**
  - Giả thuyết H0: model_6 hiệu quả hơn.
  - Giả thuyết H1: model_5 hiệu quả hơn.

→ Ta nhận thấy giá trị Pr (>F) bằng 0.1396 lớn hơn mức ý nghĩa α = 0,05 nên chưa bác bỏ được giả thuyết H0 , nên model_6 hiệu quả hơn.

b) Kiểm định giả định của mô hình model_6.

Các giả định của mô hình hồi quy:

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots \beta_i X_i + \epsilon_i$ *với* i = 1,..., n.

+ Giả thiết 1: Tính tuyến tính của dữ liệu: mối quan hệ giữa biến dự báo X và biến phụ thuộc Y được giả sử là tuyến tính.

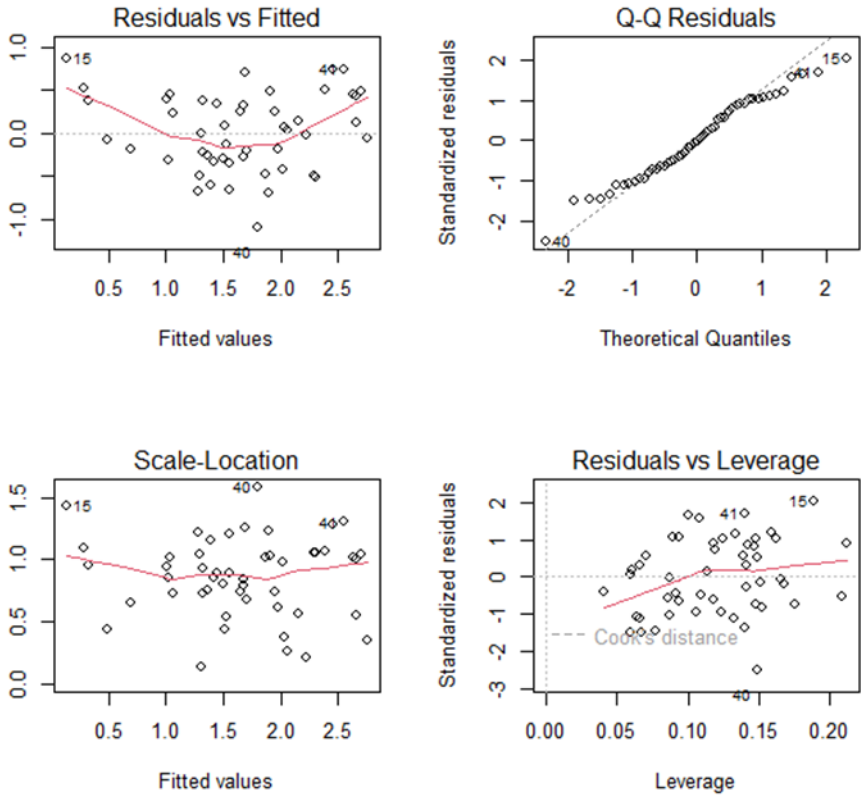
+ Giả thiết 2: Sai số có phân phối chuẩn.

+ Giả thiết 3: Phương sai của các sai số là hằng số.

+ Giả thiết 4: Các sai số ε có kỳ vọng = 0.

+ Giả thiết 5: Các sai số ε_1,..., ε_n thì độc lập với nhau.

Cách 1: Thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình



Hình 5.1.3.4 Kết quả khi vẽ các đồ thị phân tích thặng dư

- **Nhận xét**

➢ Đồ thị Residuals vs Fitted là đường cong có độ dóc chưa thỏa được ***giả thuyết 1.***

➢ Đồ thị Normal Q-Q: kiểm tra giả định về phân phối chuẩn của các sai số, các điểm sai số không nằm trên cùng một đường thẳng thì điều kiện về ***giả thuyết 2*** phân phối chuẩn không được thoả.

- Đồ thị Scale-Location: vẽ căn bậc hai của các sai số được chuẩn hoá bởi các giá trị dự báo, được dùng để kiểm tra ***giả thuyết 3*** (phương sai của các sai số là hằng số), các điểm đường màu đỏ có độ dốc và các điểm thặng dư phân tán không đều xung quanh đường thẳng này nên **giả thiết 3** bị vi phạm.
- Đồ thị Residuals vs Leverage: Các điểm thứ 15,41 là những điểm ảnh hưởng cao trong bộ dữ liệu. Tuy nhiên những điểm ảnh hưởng cao này chưa vượt qua đường thẳng khoảng cách Cook (Cook's distance) nên chúng không phải là các điểm outliers và ta không cần loại bỏ chúng khỏi bộ dữ liệu.

Cách 2: Kiểm tra các gải thiết dựa vào các kiểm định:

Giả thiết 2: Sai số có phân phối chuẩn.

H_0: Các sai số hồi quy có phân phối chuẩn.

H_1: Dữ liệu không có phân phối chuẩn.

```
> re<-residuals(model_6)
> shapiro.test(re)

        Shapiro-Wilk normality test

data:  re
W = 0.97867, p-value = 0.4972
```

Hình 5.1.3.5 Kết quả kiểm tra giả thiết sai số có phân phối chuẩn

- ***Nhận xét:***
  - Từ kết quả trên ta có p - $value$ = 0.4972 >α = 5%, nên không bác bỏ H0. Vậy giả thiết 2: Sai số có phân phối chuẩn thỏa mãn.

Giả thuyết 4 :

H_0: Các sai số có kỳ vọng $\mu = 0$

H_1: Các sai số có kỳ vọng $\mu \neq 0$

```
> re<-residuals(model_6)
> t.test(re,mu=0)

        One Sample t-test

data:  re
t = 3.3301e-16, df = 49, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -0.1276708  0.1276708
sample estimates:
   mean of x
2.115658e-17
```

Hình 5.1.3.6 Kết quả kiểm tra giả thiết sai số hồi quy

Miền bác bỏ

$$RR = \left(-\infty; -t_{\frac{\alpha}{2}}^{n-1}\right) \cup \left(t_{\frac{\alpha}{2}}^{n-1}; +\infty\right) = (-\infty; -t_{0.025}^{50-1}) \cup (t_{0.025}^{50-1}; +\infty)$$

$$= (-\infty;\ -1{,}13)\ \cup\ (1{,}13;\ +\infty)$$

Vì n − 1 = 49 ≥ 30 nên $t_{\frac{\alpha}{2}}^{n-1} \approx z_{\alpha/2}$

Tiêu chuẩn kiểm định: $z_0 = \frac{\bar{X}-\mu_0}{S/\sqrt{n}} = 3.3301e\text{-}16$

Cách 2.1: Kiểm định theo tiêu chuẩn kiểm định.

Từ kết quả R cho ta thấy z_0 không thuộc miền bác bỏ, chưa bác bỏ được giả thuyết H_0 nên giả định về các sai số có kỳ vọng bằng 0 được thoả mãn.

Cách 2.2: Kiểm định theo p-value

Ta nhận thấy p-value =1 chưa bác bỏ được giả thuyết H_0, nên giả định về các sai số có kỳ vọng bằng 0 được thoả mãn.

6. MỞ RỘNG

a) Phân Tích Phương Sai (ANOVA)

- Ưu điểm: Kiểm tra sự khác biết của nhiều nhóm: ANOVA rất hiệu quả khi muốn biết dữ liệu có sự khác biệt ý nghĩa nào đó giữa ba hoặc nhiều nhóm về các yếu tố như layer height, wall thickness, infill density, vv. Điều này có thể hữu ích nếu bạn muốn so sánh hiệu suất của máy in 3D trong các điều kiện khác nhau.

+ Phân Tích Phương Sai: ANOVA cho phép bạn phân tích mức độ biến động giữa các nhóm và mức độ biến động bên trong các nhóm. Có thể cung cấp thông tin về độ đồng nhất hoặc độ chệch lệch giữa chúng.

+ Phân loại các yếu tố ảnh hưởng: ANOVA cho phép xác định xem yếu tố nào (ví dụ: nhiệt

độ nozzle, tốc độ in) có ảnh hưởng đáng kể đến các biến đo lường.

- Nhược Điểm: Giới hạn về tuyến tính: ANOVA giả định về tuyến tính giữa biến độc lập và biến phụ thuộc, và nếu mối quan hệ không tuyến tính, phương pháp này có thể không hiệu quả. Phụ thuộc vào giả định: ANOVA đòi hỏi các giả định như phân phối chuẩn và đồng nhất của phương sai giữa các nhóm. ANOVA có thể mở rộng để xử lý nhiều biến độc lập nếu cần thiết.

- Hạn chế: Phương pháp ANOVA giả định rằng các nhóm có phân phối chuẩn. Nếu dữ liệu không tuân theo phân phối chuẩn, kết quả có thể không chính xác. ANOVA chỉ phản ánh mối quan hệ thống kê giữa các biến mà không thể xác định được mối quan hệ nguyên nhân - hiệu quả giữa chúng

b) Hồi Quy Tuyến Tính

- Ưu điểm: Hồi quy tuyến tính giúp mô hình hóa mối quan hệ tuyến tính giữa các biến, giúp bạn hiểu rõ hơn về cách các yếu tố ảnh hưởng đến kết quả các biến như nhiệt độ, tốc độ in và chất lượng in

+Dự Đoán Giá Trị: Hồi quy tuyến tính sử dụng để dự đoán giá trị của biến phụ thuộc dựa trên giá trị của các biến độc lập. Dự đoán chất lượng in dựa trên các giá trị cụ thể của các yếu tố đầu vào.

+ Xác Định Mức Độ Ảnh Hưởng Của Từng Biến: Hồi quy tuyến tính cung cấp thông tin về mức độ ảnh hưởng của từng yếu tố lên chất lượng in, giúp xác định yếu tố nào quan trọng nhất.

- Hạn chế: Hồi quy tuyến tính cho rằng mối quan hệ giữa biến độc lập và biến phụ thuộc là tuyến tính. Nếu mối quan hệ này không tuyến tính, mô hình có thể không phản ánh đúng mối quan hệ thực tế. Mô hình hồi quy tuyến tính yêu cầu dữ liệu độc lập và đồng đều, tức là các quan sát không ảnh hưởng lẫn nhau và có cùng phương sai. Nếu không thỏa mãn, kết quả có thể không chính xác. Khi số lượng biến tăng lên, mô hình có thể trở nên không ổn định và dễ làm giảm hiệu xuất, đặc biệt nếu kích thước mẫu nhỏ.

7. CODE VÀ DỮ LIỆU

Link:https://drive.google.com/drive/folders/1lPu5Mrc4HQXsOKdE5hHSCoGrPTykeiKo?usp=drive_link

8. TÀI LIỆU THAM KHẢO

1. Nguyễn Tiến Dũng (chủ biên) & Nguyễn Đình Huy, 2019, Xác suất - Thống kê & Phân tích số liệu.

2. Pham Thi Hong Anh, 2020, Xử lý missing data trong Data analysis, truy cập từ https://viblo.asia/p/xu-ly-missing-data-trong-data-analysis-maGK7qaAlj2

3. Cách nhận xét biểu đồ hộp, truy cập từ https://toploigiai.vn/cach-nhan-xet-bieu-do-hop.

4. Hướng dẫn sử dụng phần mềm Rstudio, otworzumysl.com, 21/04/2021, truy cập từ https://otworzumysl.com/huong-dan-su-dung-phan-mem-r-studio/.

5. Nguyễn Văn Tuấn, Phân tích số liệu và biểu đồ bằng R, truy cập từ https://cran.r-project.org/doc/contrib/Intro_to_R_Vietnamese.pdf6

6. Đạt Vũ, 22/07/2019, Biểu đồ Boxplots (box and whiskers), truy cập từ https://www.diendat.net/bieu-do-box-and-whiskers/

7. Minh Đức, 19/05/2021, Những điều cần biết về biểu đồ Histogram, truy cập từ https://vietquality.vn/nhung-dieu-can-biet-ve-histogram-diagram-bieu-do-phan-bo-tan-suat/