A. Chọn 2 đề tài tự do để tìm hiểu, vận dụng kiến thức đã học từ mô này

1.

2. Hồi quy cho các thành phần chính

- Tên đề tài: Phân tích xây dựng mô hình dự đoán Cường độ nén bê tông (MPa - **M**ega **Pa**scal) dựa trên bộ dữ liệu

- Nguồn gốc bộ dữ liệu: **Concrete Compressive Strength Data Set**

Được công bố trên trang web khá nổi tiếng về dữ liệu mở: [UCI Machine Learning Repository: Concrete Compressive Strength Data Set](https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength)

Thông tin về bộ dữ liệu:

* Số quan trắc: 1030
* Số biến (thuộc tính): 9. Trong đó có 9 biến đầu vào và 1 biến đầu ra
* Không có missing value trong bộ dữ liệu này.

Mô tả các biến sử dụng: toàn bộ đều là biến định lượng, là cơ sở đầu tiên cho thấy có thể áp dụng hồi quy thành phần chính để xây dựng mô hình.

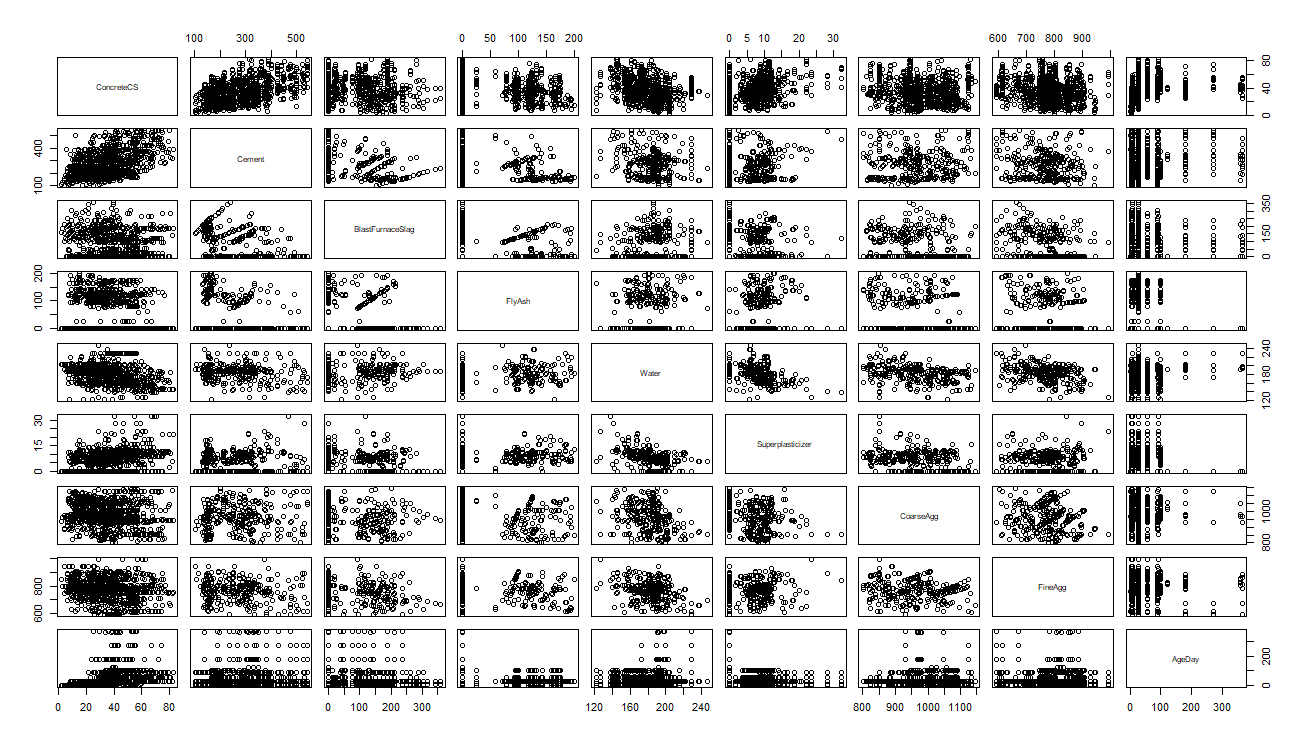
|  |  |  |
| --- | --- | --- |
| Tên | Đơn vị | Mô tả |
| Cement  (component 1) | (kg in a m^3 mixture) | Khổi lượng **Xi măng** trên 1m3 hỗn hợp bê tông |
| Blast Furnace Slag  (component 2) | (kg in a m^3 mixture) | Khổi lượng **Xỉ lò cao** (xỉ sắt nóng chảy lò cao) trên 1m3 hỗn hợp bê tông. Đây là phế thải dạng hạt của ngành công nghiệp luyện  gang  thép. |
| Fly Ash  (component 3) | (kg in a m^3 mixture) | Khổi lượng **Tro bay** trên 1m3 hỗn hợp bê tông. Tro bay là bụi khí thải dưới dạng hạt mịn thu được từ quá trình đốt cháy nhiên liệu than đá trong các nhà máy nhiệt điện... |
| Water  (component 4) | (kg in a m^3 mixture) | Khổi lượng **Nước** trên 1m3 hỗn hợp bê tông |
| Superplasticizer  (component 5) | (kg in a m^3 mixture) | Khổi lượng **Phụ gia siêu dẻo** trên 1m3 hỗn hợp bê tông |
| Coarse Aggregate  (component 6) | (kg in a m^3 mixture) | Khổi lượng **Cốt liệu thô** trên một trên 1m3 hỗn hợp bê tông. Ví dụ như đá, sỏi. |
| Fine Aggregate  (component 7) | (kg in a m^3 mixture) | Khổi lượng **Cốt liệu mịn** (tinh) trên một trên 1m3 hỗn hợp bê tông. Ví dụ như cát, đá mạt. |
| Age | (day) | **Tuổi** bê tông, tính bằng ngày từ 1-365 |
| Concrete compressive strength | (MPa, megapascals) | **Cường độ nén bê tông**.  Đây chính là biến phụ thuộc cần nghiên cứu xây dựng mô hình. |

Thực hiện nhập dự liệu vào R, điều chỉnh tên biến cho thuận tiện sử dụng:

|  |
| --- |
| Concrete <- readxl::read\_xls('Concrete\_Data.xls')  colnames(Concrete)[1] <- 'Cement'  colnames(Concrete)[2] <- 'BlastFurnaceSlag'  colnames(Concrete)[3] <- 'FlyAsh'  colnames(Concrete)[4] <- 'Water'  colnames(Concrete)[5] <- 'Superplasticizer'  colnames(Concrete)[6] <- 'CoarseAgg'  colnames(Concrete)[7] <- 'FineAgg'  colnames(Concrete)[8] <- 'AgeDay'  colnames(Concrete)[9] <- 'ConcreteCS' |

Quan sát biểu đồ phân tán:

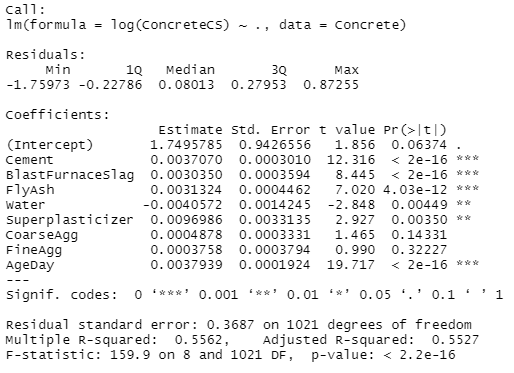
|  |
| --- |
| pairs(ConcreteCS~., data = Concrete) |



Chỉ có khối lượng xi măng là có mối quan hệ tuyến tính rõ ràng với cường độ nén bê tông. Biểu đồ phân tán này cho thấy các biến đổi đơn giản như log, mũ sẽ không thể tạo ra quan hệ tuyến tính đối với các biến.

Thực tế thì khối lượng các thành phần ảnh hưởng đến chất lượng bê tông là điều rất rõ ràng. Do đó, ta thử xem xét mô hình với đầy đủ biến là khối lượng các thành phần này.

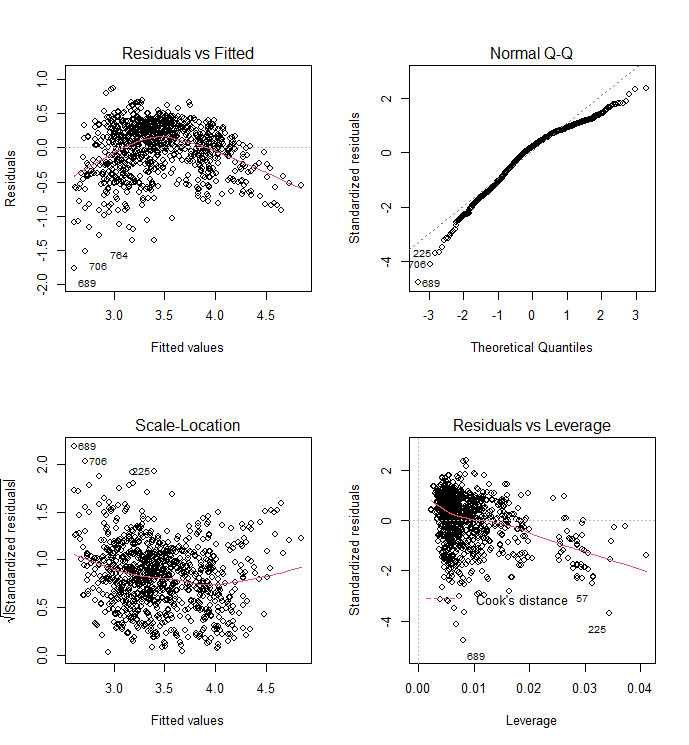
|  |
| --- |
| fullModel <- lm(log(ConcreteCS) ~ ., data = Concrete)  summary(fullModel) |



Đa số biến đều có ý nghĩa thông kế kê ngoại trừ CoarseAgg và FineAgg. Việc loại bỏ 2 biến này có hợp lý? Mô hình dự đoán chất lượng bê tông nhưng không quan tâm đến thành phần là đá và cát sẽ không là sự lựa chọn đúng của nhà thống kê. Điều này rõ ràng không thuyết phục. Ngoài ra mô hình chỉ giải thích được 55.6% giá trị cường độ nén bê tông.

Xem xét thêm biểu đồ để có thêm chứng minh rằng đây không là 1 mô hình đúng

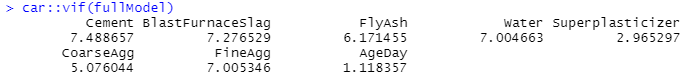
|  |
| --- |
| par(mfrow=c(2,2))  plot(fullModel) |



Phần dư không tuân theo phân phối chuẩn, phương sai thay đổi mạnh, trung bình cũng dao động chứ không bằng 0.

Tiếp theo, thử xem xét hệ số lạm phát phương sai để kiểm tra hiện tượng đa cộng tuyến

|  |
| --- |
| car::vif(fullModel) |

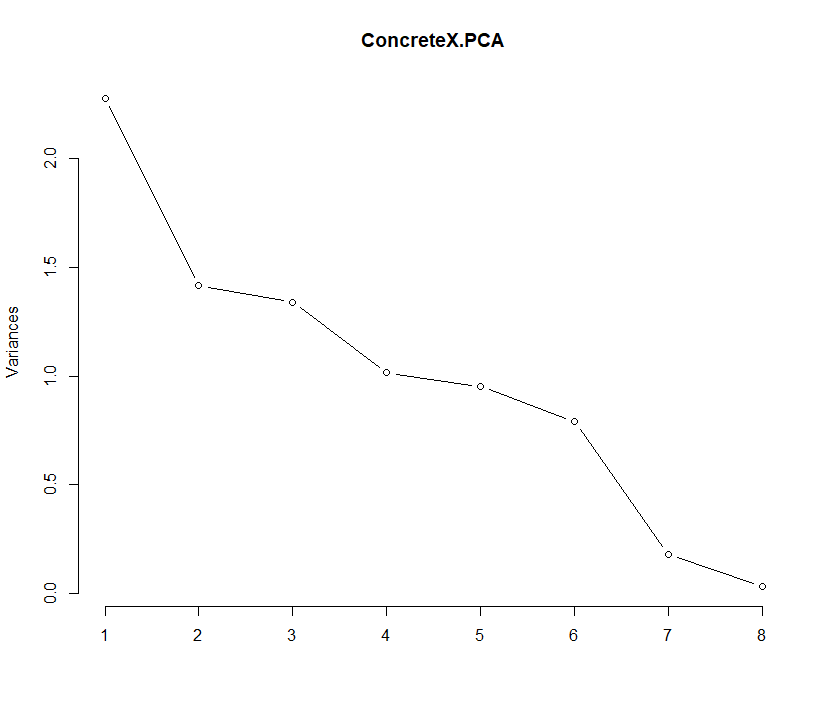


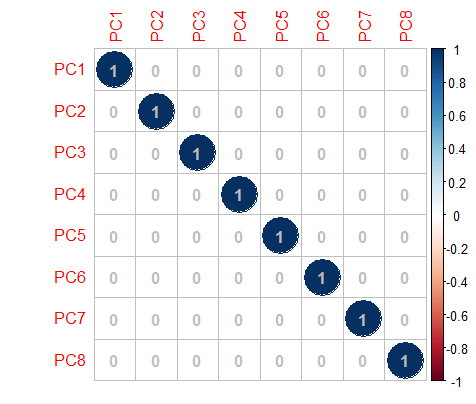
Có thể thấy được lý do có của việc mô hình này không tồn tại chính là hiện tượng đa cộng tuyến. Thông thường, ta có thể bỏ đi biến có VIF cao nhất. Tuy vậy, ở đây khối lượng xi măng chắc chắn không nên là biến có thể bỏ ra khỏi mô hình dự đoán.

Như vậy có thể kết luận là mô hình hồi quy tuyến tính đa biến thông thường không thể sử dụng để giải thích chất lượng bê tông được. Để vượt qua vấn đề này, ta có thể sử dụng mô hình hồi quy bậc cao hơn. Tuy nhiên, ta vừa kết luận rằng đa cộng tuyến là nguyên nhân, một cách tiếp cận rất tốt để loại bỏ nó đó là PCA, phân tích thành phần chính. Một lý do khác để sử dụng PCA cho bộ dữ liệu này đó là toàn bộ biến đều là định lượng, rất phù hợp để phân tích thành phần chính mà không phải loại bỏ bất kì biến nào.

Trên bộ dữ liệu đang có, thực hiện loại ra biến phụ thuộc và tiến hành phân tích thành phần chính cho toàn bộ biến độc lập đầu vào.

|  |
| --- |
| ConcreteX <- Concrete[,-9]  ConcreteX.PCA <- prcomp(ConcreteX, center = TRUE,scale. = TRUE)  plot(ConcreteX.PCA, type = c("lines"))  summary(ConcreteX.PCA)  corrplot::corrplot(cor(ConcreteX.PCA$x), addCoef.col = "gray") |

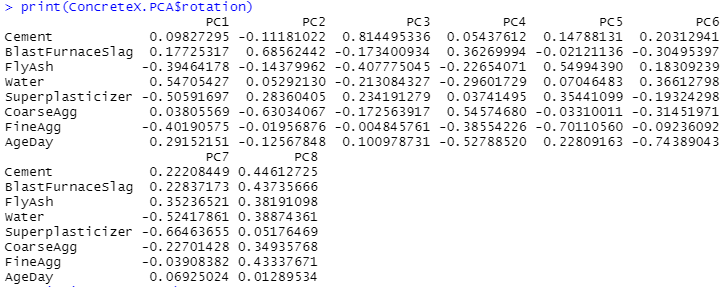




Các thành phần chính không hề có tương quan lẫn nhau.

Một số thông tin khác về kết quả phân tích thành phần chính:

|  |
| --- |
| print(ConcreteX.PCA$rotation)  eigen(cor(ConcreteX))$value  diag(cov(ConcreteX.PCA$x[,])) |

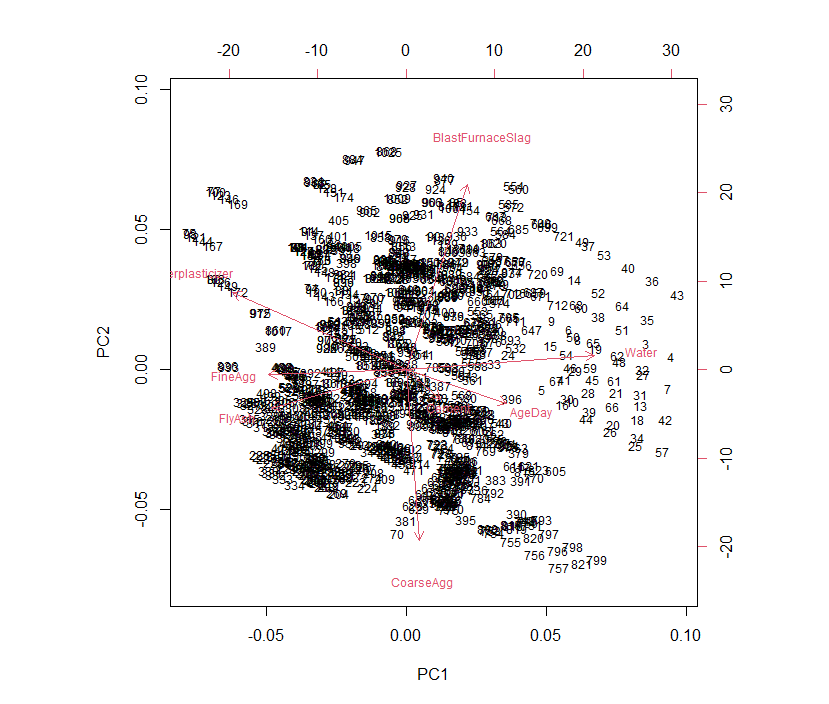


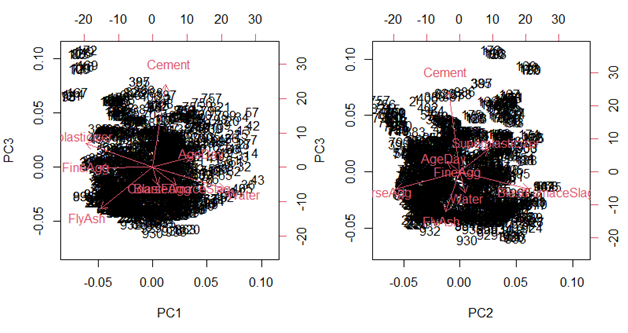
Chuyển dữ liệu từ hệ không gian gốc sang hệ không gian PCA bằng các hệ số rotation

Ví dụ: PC1 = 0.55\*Water -0.51\*Superplasticizer – 0.4\*FineAgg – 0.4\*FlyAsh + 0.3\*AgeDay, các thành phần còn lại ảnh hưởng rất nhỏ bởi hệ số nhỏ của nó.

Đồ thị biểu diễn các mẫu dữ liệu trên 3 thành phần chính đầu tiên sẽ cho cái nhìn trực quan hơn về điều này:

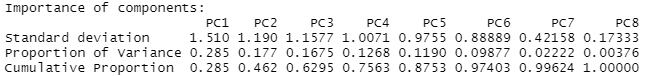
|  |
| --- |
| biplot(ConcreteX.PCA, cex = 0.75)  biplot(ConcreteX.PCA, choices = c(1, 3))  biplot(ConcreteX.PCA, choices = c(2, 3)) |





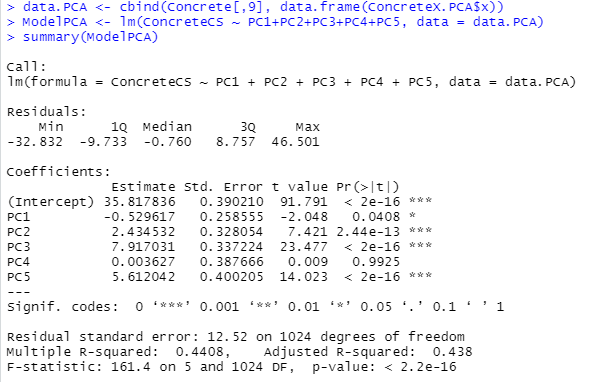
Dựa trên biểu đồ này, có thể dễ nhận thấy rằng PC1 đại diện bởi tỉ lệ nước (tương quan thuận) , phụ gia dẻo, bụi bay và cốt liệu liệu mịn (tương quan nghịch). Đây là cá thành phần rắn có kích thước nhỏ và lỏng. Ngược lại PC được đại diện bởi các thành phần rắn có kích thước lớn hơn như cốt liệu thô hay xỉ lò cao.

PC3 chịu ảnh hưởng lớn nhất bởi khối lượng xi măng và một phần nào đó là bụi bay, là các thành phần kích thước cực nhỏ, dạng bột mịn.



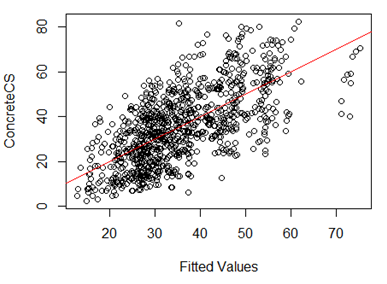
Ta thấy 5 thành phần chính đầu tiên tích lũy được hơn 87% tỉ trọng phương sai. Phù hợp để xây dựng mô hình hồi quy tuyến tính.

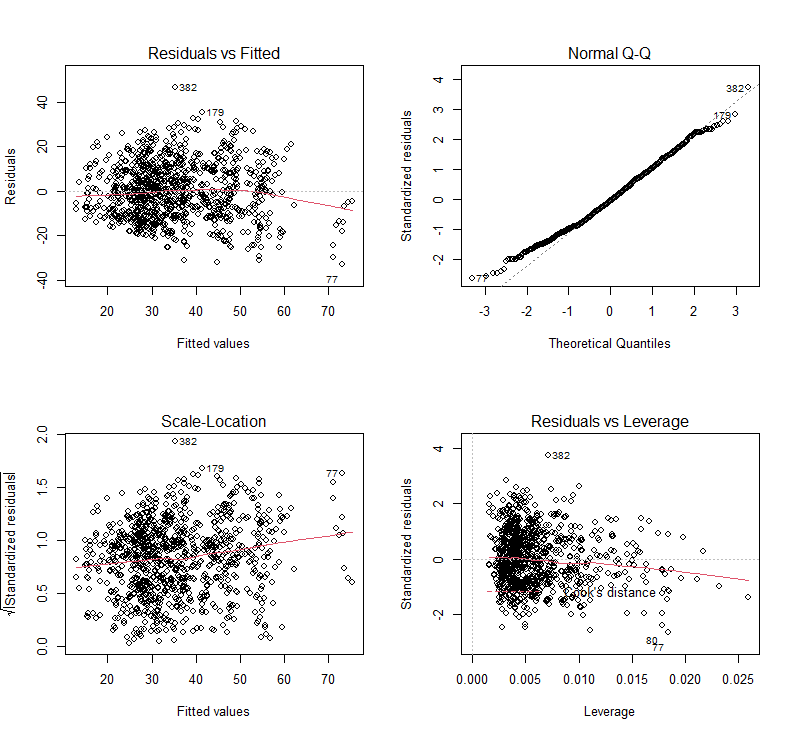
|  |
| --- |
| data.PCA <- cbind(Concrete[,9], data.frame(ConcreteX.PCA$x))  ModelPCA <- lm(ConcreteCS ~ PC1+PC2+PC3+PC4+PC5, data = data.PCA)  summary(ModelPCA) |



Mô hình có hệ số của PC4 là không có ý nghĩa thống kê, chỉ giải thích được hơn 43% của biến dự đoán. Tuy vậy ta xem xét thêm biểu đồ diagnostic để xem mô hình có cải thiện được so với mô hình các biến gốc

|  |
| --- |
| par(mfrow=c(1,1))  plot(ModelPCA$fitted.values,ConcreteCS,xlab="Fitted Values")  abline(lsfit(ModelPCA$fitted.values,ConcreteCS), col = "red")  par(mfrow=c(2,2))  plot(ModelPCA) |



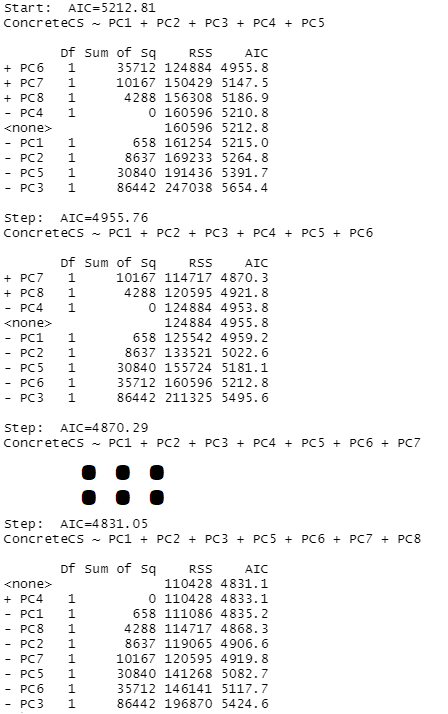


Ta thấy những dấu hiệu rõ ràng về một mô hình đạt tiêu chuẩn tồn tại đó là Phần dư tuân theo phân phối chuẩn, phương sai cố định dao động khoảng 0.8-1.0, trung bình cũng bằng 0. Và rõ ràng các thành phần chính sẽ loại bỏ hoàn toàn hiện tượng đa cộng tuyến.

Dựa trên những thông tin này ta tiếp tục chọn mô hình dựa trên các thành phần chính này.

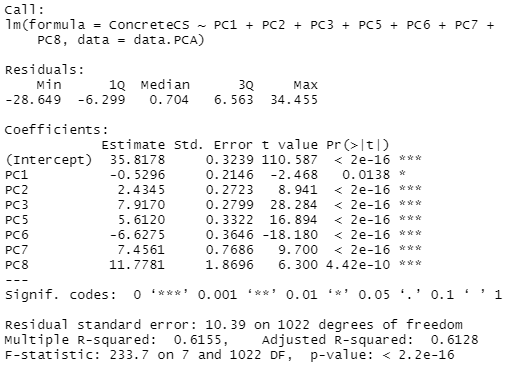
Lựa chọn mô hình trung bình chính là mô hình với với 5 thành phần chính với tích lũy phương sai 87% đã chọn.

|  |
| --- |
| ModelPCABase <- lm(ConcreteCS ~ 1 , data = data.PCA)  ModelPCAFull <- lm(ConcreteCS ~ . , data = data.PCA)  ModelPCA.AIC <- MASS::stepAIC(ModelPCA, direction = "both",  scope = list(lower = ModelPCABase, upper = ModelPCAFull)) |



|  |
| --- |
| summary(ModelPCA.AIC) |

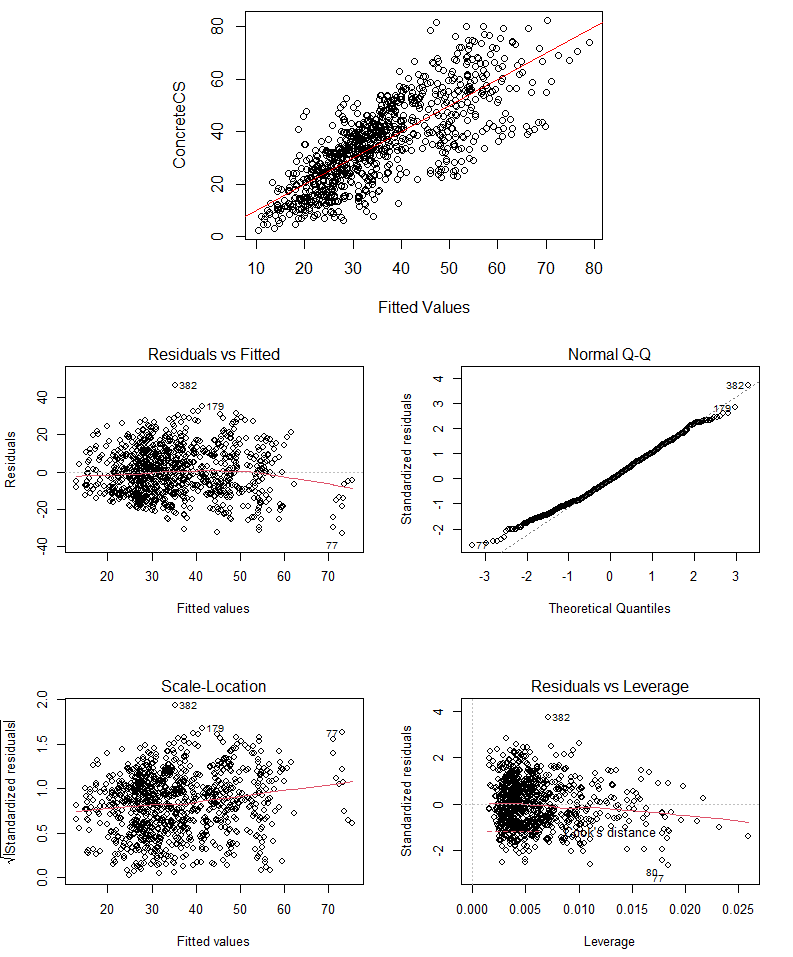
Mô hình chọn được các biến đều có ý nghĩa thống kê cao, là một mô hình có tồn tại.



Các biểu đồ bề giá trị fited và giá trị thực, biểu đồ diagnostic và hệ số VIF = 1 giúp ta khẳng định mô hình được chọn là mô hình valid.

|  |
| --- |
| par(mfrow=c(1,1))  plot(ModelAIC$fitted.values,ConcreteCS,xlab="Fitted Values")  abline(lsfit(ModelAIC$fitted.values,ConcreteCS), col = "red")  par(mfrow=c(2,2))  plot(ModelAIC)  par(mfrow=c(1,1))  hist(ModelAIC$residuals)  car::vif(ModelAIC) |

|  |  |
| --- | --- |
|  |  |



Kết luận: Bộ dữ liệu về đánh giá chất lượng bê tông với số biến là các thành phần cấu tạo nên bê tông, toàn bộ biến này đều quan trọng về mặt thực tế. Tuy nhiên hồi quy tuyến tính lại không thể giải thích và có xu hướng muốn bỏ bớt biến để chọn mô hình - điều là không hợp lý. Với PCA ta có thể vượt qua vấn đề này khi không những tạo ra được mô hình hồi quy valid mà mô hình này còn giải thích được biến phụ thuộc với tỉ lệ cao hơn là 61.3% so với mô hình hồi quy thông thường chỉ giải thích được 55,3%.