TẠP CHÍ KHOA HỌC VÀ CÔNG NGHỆ NĂNG LƯỢNG - TRƯỜNG ĐẠI HỌC ĐIỆN LỰC
(ISSN: 1859 - 4557)

SỬ DỤNG R TRONG PHÂN TÍCH HỒI QUY ÁP DỤNG CHO DỰ ÁN ĐIỆN MẶT TRỜI ÁP MÁI

USE R IN REGRESSION ANALYSIS APPLIED TO ROOFTOP SOLAR POWER PROJECT

Chu Văn Tuấn, Nguyễn Thúy Ninh

Đại học Điện lực

Ngày nhận bài: 02/06/2022, Ngày chấp nhận đăng: 12/08/2022, Phản biện: TS. Đỗ Anh Tuấn

Tóm tắt:

Hiện nay có nhiều phương pháp, phần mềm dùng để phân tích hồi quy, trong bài báo này tác giả sử dụng R. R là một ngôn ngữ thống kê học, nhưng cũng có thể xem là một phần mềm có thể sử dụng cho các phân tích thống kê và đồ thị. R có thể sử dụng cho nhiều mục tiêu khác nhau, từ tính toán đơn giản, toán học giải trí, tính toán ma trận đến các phân tích thống kê phức tạp.

Sử dụng R trong phân tích các yếu tố ảnh hưởng đến sản lượng điện năng của nhà máy điện mặt trời có công suất 1195kWp bằng phương pháp hồi quy tuyến tính. Từ đó chỉ ra ý nghĩa của các tham số trong mô hình, cách đánh giá tầm quan trọng của các biến tiên lượng, quy trình xây dựng và kiểm định mô hình dự báo xem xét đến cả các vấn đề đa cộng tuyến và hoán chuyển dữ liệu. Trong một tương lai không xa, khi thị trường điện phát triển, kết quả của việc nghiên cứu mô hình dự báo hay chào giá sản lượng điện năng do các dự án mặt trời tạo ra có ý nghĩa vô cùng quan trọng.

Từ khóa:

Phân tích, thống kê, đồ thị, R, hồi quy, điện mặt trời.

Abstract:

There are many methods and software used for regression analysis, in this paper the author used R. R is not only a statistical language but also a software that can be used for statistical analysis and graphs. Additionaly R can be used for a variety of purposes, from simple calculations, recreational math, matrix calculations to complex statistical analyses.

Using R in analyzing factors affecting power output of a solar power plant with a capacity of 1195kWp by linear regression. It shows the meaning of the parameters in the model, how to evaluate the importance of prognostic variables, and the process of building and testing the predictive model considering both multicollinearity and transformation problems data. In the not-so-distant future, when the electricity market develops, the results of studying the forecasting model or the price of electricity generated by solar projects are extremely important.

Keywords:

Analysis, statistics, graph, R, regression, solar power.

I/ ĐẶT VẤN ĐỀ

Phân tích hồi quy là một tập hợp các phương pháp thống kê được sử dụng để ước tính các mối quan hệ giữa một biến phụ thuộc và một hoặc nhiều biến độc lập. Nó có thể được sử dụng để đánh giá mối quan hệ giữa các biến và mô hình hóa mối quan hệ trong tương lai giữa chúng.

Trong các dự án, phân tích hồi quy được sử dụng để xác định biến nào trong số những biến đó thực sự có tác động. Nó trả lời các câu hỏi: Yếu tố nào quan trọng nhất? Yếu tố nào có thể bỏ qua? Các yếu tố đó tương tác với nhau như thế nào? Ứng dụng mô hình hồi quy đòi hỏi kỹ năng về mô hình hóa, kiến thức, không ứng dụng sai mô hình, không kiểm tra các giả định, và phải xem xét các hiện tượng đa cộng tuyến, hoán vị dữ liệu … Xây dựng mô hình tiên lượng phải có độ chính xác cao đồng thời phải đơn giản, thực tế và dễ áp dụng.

Hiện nay có nhiều phương pháp, phần mềm dùng để phân tích hồi quy, trong bài báo này tác giả sử

dụng R. R là một ngôn ngữ thống kê học, nhưng cũng có thể xem là một phần mềm có thể sử dụng cho các phân tích thống kê và đồ thị. Phần cơ bản của R bao gồm một số lệnh/hàm phổ biến có thể sử dụng cho phân tích đơn giản. Các hàm rnorm, mean, sd, hist, lm, glm… có sẵn trong Base R. Tuy nhiên, trong thực tế, chúng ta phân tích chuyên biệt như mô hình hồi quy phi tuyến tính thì Base R không làm được. Trong trường hợp phân tích chuyên biệt, chúng ta cần dùng đến các package chuyên biệt. Trong R có rất nhiều package (hơn 10.000 packages), và mỗi package chỉ tập trung làm một số phân tích chuyên sâu. Các package có trên CRAN. Mỗi package có lệnh/hàm riêng mà nhà thiết kế đã cài sẵn. Do đó, để sử dụng package, chúng ta cài đặt trực tiếp bằng install.packages. Trước khi dùng R cho phân tích dữ liệu, dữ liệu phải được đọc vào R. R có thể đọc hầu hết các loại dữ liệu dạng Excel, Stata, SPSS… Đối với các dữ liệu đơn giản có thể nhập trực tiếp vào R mà không cần dùng chương trình (package) nào bằng cách dùng hàm c() sau đó đưa vào một dataset (R gọi dataser là data.frame) để phân tích [1], [2].

Sử dụng R trong phân tích các yếu tố ảnh hưởng đến sản lượng điện năng của nhà máy điện mặt trời có công suất 1195kWp. Có rất nhiều phương pháp phân tích hồi quy như hồi quy logistics, hồi quy Cox, hồi quy Poisson … tuy nhiên tùy thuộc vào từng đối tượng phân tích, bộ dữ liệu thu thập được, tác giả lựa chọn phương pháp hồi quy tuyến tính để phân tích các yếu tố ảnh hưởng đến sản lượng điện năng do tấm pin mặt trời sản xuất ra (Quantity.PV). Qua đây tác giả chỉ ra ý nghĩa của các tham số trong mô hình, cách đánh giá tầm quan trọng của các biến tiên lượng, quy trình xây dựng và kiểm định mô hình dự báo xem xét đến cả các vấn đề đa cộng tuyến và hoán chuyển dữ liệu [3]. Trong một tương lai không xa, khi thị trường điện phát triển, kết quả của việc nghiên cứu mô hình dự báo, đưa ra chiến lược chào giá dựa trên sản lượng điện năng do các dự án mặt trời tạo ra có ý nghĩa vô cùng quan trọng.

II/ CƠ SỞ LÝ THUYẾT

Để xây dựng mô hình để định lượng hóa và dự báo, một trong những mô hình phổ biến nhất là mô hình hồi quy tuyến tính (line regression model). Gọi (x_i, y_i) là cặp giá trị x và y của đối tượng i (i=1,2,3…n). Mô hình hồi quy tuyến tính: $y_i = \alpha+\beta x_i$

Tuy nhiên chúng ta kỳ vọng rằng mô hình đường này không thể nối kết tất cả các giá trị (x_i,y_i) được. Sẽ có một số giá trị lệch khỏi mô hình. Do đó, chúng ta thêm một yếu tố khác của mô hình là ε_i.

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad (1)$$

Đó là mô hình cho tổng thể. Chúng ta không biết giá trị của 2 tham số α và β, nhưng chúng ta có mẫu quan sát để ước tính cho các tham số. Mô hình cho mẫu nghiên cứu là:

$$y_i = a + bx_i + e_i \quad (2)$$

a là ước số của α và b là ước số của β. Biến e là phần dư tức là phần còn lại của y mà mô hình a + bx không giải thích được. Nói cách khác, mô hình hồi quy tuyến tính: Giá trị quan sát của y = giá trị tiên lượng + phần dư hay y = $\hat{y}$ + e

Phần dư = giá trị quan sát – giá trị tiên lượng

$$e = y-\hat{y} = y - (a + bx) \quad (3)$$

Phương pháp bình phương cực tiêu có mục tiêu là cực tiểu hóa tổng phần dư.

$\min\sum (y - \alpha - \beta x)^2$ hay mục tiêu là cần tìm a và b sao cho tổng bình phương phần dư là nhỏ nhất.

$$b = \frac{\sum (x_i-\underline{x})(y_i-\underline{y})}{\sum (x_i-\underline{x})^2} \text{ và } a=\underline{y}-bx \quad (4)$$

Sau khi đã có các giá trị ước lượng a và b, ta có thể ước lượng các giá trị y cho từng giá trị x:

$$\hat{y}_i = a + bx_i \quad (5)$$

Hai chỉ số chính để đánh giá sự hữu dụng của một mô hình hồi quy tuyến tính là hệ số R^2 và MSE (mean square error).

Chỉ số đơn giản để thể hiện độ biến thiên là tổng bình phương (sum of squares hay SS). Nhưng SS cần một điểm tham chiếu. Chúng ta có thể thấy rằng điểm tham chiếu của biến y là giá trị trung bình và chúng ta có thể tính SS cho y (ký hiệu là TSS) nhu sau:

$$TSS = \sum_{i=1}^{n} (y_i - \underline{y})^2 \quad (6)$$

Tổng bình phương từ giá trị tiên lượng và giá trị trung bình là:

$$RSS = \sum_{i=1}^{n} (\hat{y}_i - \underline{y}_i)^2 \quad (7)$$

Tổng bình phương của phần dư:

$$ESS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad (8)$$

Hệ số xác định (R^2) của mô hình hồi quy là tỷ số của RSS và TSS:

$$R^2 = \frac{RSS}{TSS} \quad (9)$$

R^2 nằm trong khoảng 0 và 1. Hệ số xác định R^2 là phần trăm phương sai của y có thể giải thích bởi mô hình hồi quy tuyến tính.

Một chỉ số quan trọng khác là MSE (mean squared error là phương sai của y sau khi hiệu chỉnh cho x. Trong thực tế, MSE được ước tính từ phần dư, bởi vì phần dư phản ánh phần phương sai mà mô hình không giải thích được.

$$MSE = \frac{\sum (y_i - \hat{y}_i)^2}{n-2} = \frac{\sum_{i=1}^{n} (e_i)^2}{n-2} \quad (10)$$

Độ lệch chuẩn của y sau khi đã hiệu chỉnh cho x:

$$\text{RMSE} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}} = \sqrt{\frac{\sum_{i=1}^{n} (e_i)^2}{n-2}} \quad (11)$$

Để đánh giá một mô hình hồi quy tuyến tính có đại diện cho dữ liệu, chúng ta sử dụng hệ số xác định R^2 và MSE. Mô hình có R^2 càng cao có nghĩa là mô hình giải thích nhiều phương sai và giảm độ bất định nên MSE sẽ thấp. Mô hình có R^2 thấp thì tính bất định của tiên lượng sẽ cao và điều này cũng phản ánh qua giá trị MSE tăng [4], [5].

III/ KẾT QUẢ NGHIÊN CỨU

1. Dự án điện mặt trời: Dự án điện mặt trời áp mái có công suất lắp đặt 1195kWp bằng phương pháp hồi quy tuyến tính. Chủ đầu tư: Công ty TNHH NTPM (Việt Nam), đơn vị tổng thầu: Công ty TNHH Năng lượng bền vững Việt Nga.

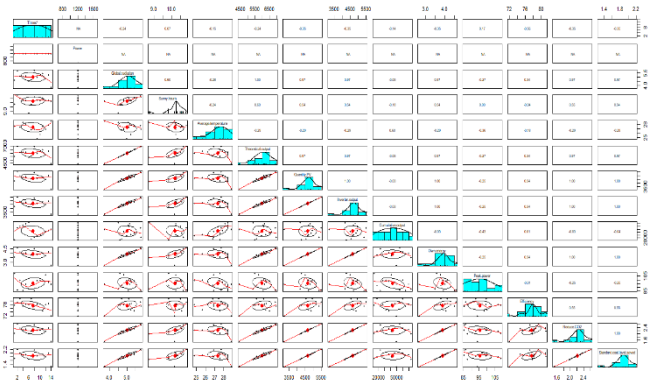
2. Nhập dữ liệu và phân tích

Như chúng ta biết, lượng điện năng do tấm pin mặt trời được sinh ra chính nhờ ánh nắng mặt trời. Vì thế chỉ số về ánh nắng vô cùng quan trọng, quyết định sản lượng điện năng được sinh ra nhiều hay không. Trong giai đoạn dự án hiện nay, các số liệu được thu thập, khảo sát và ghi lại theo thời gian vào file dữ liệu Excel. Có rất nhiều đối tượng được quan sát trong file thu thập. Sử dụng R vào thống kê mô tả các đối tượng trong dữ liệu nghiên cứu, chúng ta sử dụng hàm describe trong package psych[9], [10].

```
>library(psych)
>describe(m)
```

Trước khi đi sâu vào phân tích và lựa chọn mô hình hồi quy tuyến tính phù hợp, tác giả muốn chỉ ra độ tương quan giữa các biến độc lập trong bộ số liệu thu thập cùng một lúc bằng cách gọi library(psych) trong package: ggplot2.

```
>pairs.panels(m)
```



Hình 1. Biểu đồ tương quan giữa các biến

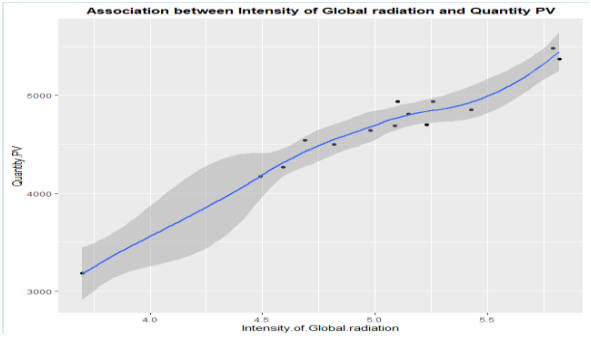
Biểu đồ trên là một ma trận biểu đồ cung cấp cho chúng ta biểu đồ tương quan từng biến và đường biểu diễn một cách trực quan. Phần phía trên của ma trận là hệ số tương quan. Các ô trong đường chéo vẽ phân bố của từng biến. Trong bài báo này, tác giả tập trung phân tích các biến ảnh hưởng đến sản lượng điện năng do tấm pin mặt trời sản xuất ra (Quantity.PV). Kết quả từ hàm pairs.panels(m) cho thấy biến Quantity.PV có mối liên quan mật thiết với các biến: Cường độ bức xạ (Intensity.of.Global.radiation) và thời gian có nắng (Sunny.hours) do có hệ tương quan cao là 0.97 và 0.54.

Mô hình 1: Quantity.PV~ Intensity.of Intensity.of.Global.radiation

Chúng ta tập trung phân tích sự ảnh hưởng của biến "Intensity.of.Global.radiation" đến biến "Quantity.PV".

Để phân tích biểu đồ tương quan giữa 2 biến, chúng ta gọi hàm ggplot2[5], [6].

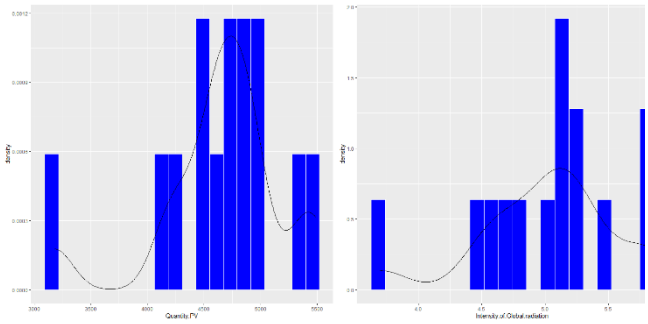
```
>library(ggplot2)
>p=ggplot(data=m,aes(x=Intensity.of.Global.radiation,y=Quantity.PV))
>p=p+geom_point()+geom_smooth()+ggtitle("Association between Intensity of Global radiation and Quantity PV")+theme(plot.title=element_text(lineheight=0.8,face="bold",hjust=0.5))
> p=p+theme(legend.position="centre")
>p
```



Hình 2. Biểu đồ tương quan giữa hai biến

Nhìn vào biểu đồ tương quan của hai biến trên, ta thấy Intensity.of.Global.radiation và Quantity.PV tương quan gần như một đường thẳng. Phân tích từng biến sâu hơn, tác giả sử dụng biểu đồ phân bố. Đây là một phương tiện rất có ích để thể hiện sự phân bố của một biến số liên tục. Để thể hiện phân bố của biến Intensity.of.Global.radiation, Quantity.PV ta dùng hàm geom_histogram như sau:

```
>g=ggplot(data=m,aes(Intensity.of.Global.radiation))
>g=g+geom_histogram(bins=20,aes(y=..density..),col="white",fill="blue",lwd=0.5 )
>g=g+geom_density()
>n=ggplot(data=m,aes(Quantity.PV ))
>n=n+geom_histogram(bins=20,aes(y=..density..),col="white",fill="blue",lwd=0.5)
>n=n+geom_density()
```



Hình 3. Biểu đồ phân bố

Biểu đồ thanh cũng có thể dùng để thể hiện một biến liên tục, trình bày theo dạng ngang để nhấn mạnh hai đối tượng đang phân tích. Qua biểu đồ phân tích mối tương quan ở trên x=Intensity.of.Global.radiation, y=Quantity.PV, chúng ta thấy có mối tương quan thuận. Khi cường độ bức xạ tại điểm đo giảm thì sản lượng điện năng tấm pin mặt trời sản xuất ra cũng giảm và ngược lại. Vấn đề đặt ra là làm thế nào để định lượng hóa mối tương quan này.

Hệ số tương quan Pearson và Spearman cho chúng ta thấy được mối quan hệ đó. Theo Pearson để đo lường mối tương quan, cần xác định một chỉ số đó là covariance (hiệp phương sai).

Trong R, theo phương pháp Pearson ta dùng cor.test(x,y). Kết quả phân tích:

```
>cor.test(x=m$Intensity.of.Global.radiation,y=m$Quantity.PV)
```

Pearson's product-moment correlation

data: m$Intensity.of.Global.radiation and m$Quantity.PV

t = 14.954, df = 12, p-value = 4.026e-09

alternative hypothesis: true correlation is not equal to 0

95 percent confidence interval:

0.9182656 0.9920164

sample estimates:

cor

0.9742015

r=0.97>0 gần bằng 1, mối tương quan giữa Intensity.of.Global.radiation, Quantity.PV là khá chặt chẽ và tương quan thuận với nhau. Trong trường hợp(x,y) không tuân theo quy luật phân bố chuẩn, để đánh giá mối tương quan, thay vì dùng hệ số Pearson, ta dùng hệ số Spearman (ρ).

```
>cor.test(x=m$Intensity.of.Global.radiation,y=m$Quantity.PV,method="spearman")
```

Kết quả:

Spearman's rank correlation rho

data: m$Intensity.of.Global.radiation and m$Quantity.PV

S = 22, p-value < 2.2e-16

alternative hypothesis: true rho is not equal to 0

sample estimates:

rho

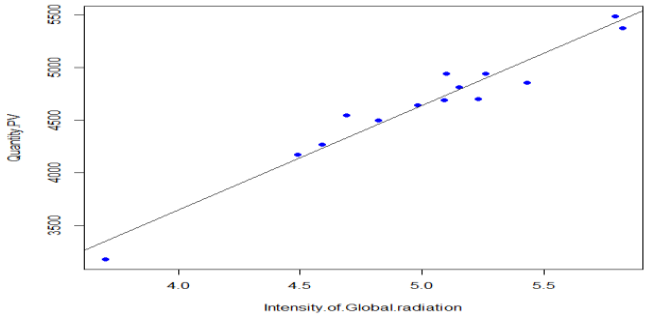
0.9516484

Hệ số tương quan Spearman là 0.95, tuy thấp hơn hệ số tương quan Pearson, nhưng vẫn có ý nghĩa thống kê (P<0.001)

Chúng ta đã thấy hệ số tương quan để định lượng hóa một mối liên quan giữa biến x = Quantity.PV và y= Intensity.of.Global.radiation. Tuy nhiên chúng ta muốn xây dựng mô hình để định lượng hóa và dự báo. Một trong những mô hình phổ biến nhất là mô hình hồi quy tuyến tính (line regression model)[4].

>plot(Quantity.PV~Intensity.of.Global.radiation,data=m,pch=16,col="blue")

>abline(lm(Quantity.PV~Intensity.of.Global.radiation,data=m))



Hình 4. Mô hình hồi quy tuyến tính giữa sản lượng điện năng và cường độ bức xạ mặt trời

Phân tích hồi quy tuyến tính bằng R và kết quả như sau:

>M1=lm(Quantity.PV~Intensity.of.Global.radiation,data=m)

>summary(M1)

Call:

lm(formula = Quantity.PV ~ Intensity.of.Global.radiation, data = m)

Residuals:

Min	1Q	Median	3Q	Max
-213.56	-69.63	27.30	40.50	211.28

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-314.68	333.81	-0.943	0.364
Intensity.of.Global.radiation	990.89	66.26	14.954	4.03e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 130.7 on 12 degrees of freedom

Multiple R-squared: 0.9491, Adjusted R-squared: 0.9448

F-statistic: 223.6 on 1 and 12 DF, p-value: 4.026e-09

Trong trường hợp này $R^2 = 0.94$ có nghĩa là biến độc lập Intensity.of.Global.radiation "giải thích" khoảng 94% sự biến thiên của biến phụ thuộc Quantity.PV. Phần còn lại 6% được giải thích bởi các biến ngoài mô hình và sai số ngẫu nhiên. Cột Estimate cho ta kết quả ước tính hai tham số của mô hình hồi quy tuyến tính. Theo đó, a= -314.68 và b=990.89. Do đó, mô hình bây giờ là:

M1: Quantity.PV = -314.68 + 990.89 Intensity.of.Global.radiation

Trong mô hình này, ý nghĩa của b=990.89 là khi cường độ bức xạ tăng lên 1kWh/m^2 thì sản lượng điện năng do tấm pin mặt trời sinh ra tăng lên 990.89 kWh.

Hằng số a=-314.68 có nghĩa là khi cường độ bức xạ = 0 thì sản lượng điện năng là -314.68 kWh. Điều này hơi vô lý vì thực thế sản lượng tạo ra không thể là số âm. Tuy nhiên chúng ta có thể hoán đổi biến cường độ bức xạ Intensity.of.Global.radiation sang đơn vị z:

zIntensity.of.Global.radiation

$$= \frac{Intensity.of.Global.radiation - trung\ bình\ (Intensity.of.Global.radiation)}{Độ\ lệch\ chuẩn\ của\ Intensity.of.Global.radiation}$$

Giá trị trung bình của Intensity.of.Global.radiation là 5.01 kWh/m^2 và độ lệch chuẩn là 0.55. Điều này có nghĩa là khi cường độ bức xạ có giá trị bằng giá trị trung bình thì zIntensity.of.Global.radiation = 0.

Chúng ta có thể hoán đổi bằng cách dùng hàm scale như sau:

m$zIntensity.of.Global.radiation=scale(m$Intensity.of.Global.radiation). Đưa biến số zIntensity.of.Global.radiation vào bộ dữ liệu ban đầu. Biến số này có giá trị trung bình là 0 và độ lệch chuẩn là 1[2].

Phân tích mô hình với biến zIntensity.of.Global.radiation.

>zM1=lm(Quantity.PV~zIntensity.of.Global.radiation,data=m)

>summary(zM1)

Kết quả như sau:

Call:

lm(formula = Quantity.PV ~ zIntensity.of.Global.radiation, data = m)

Residuals:

Min	1Q	Median	3Q	Max
-213.56	-69.63	27.30	40.50	211.28

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)

(Intercept) 4649.66 34.92 133.15 < 2e-16 ***

zIntensity.of.Global.radiation 541.91 36.24 14.95 4.03e-09 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 130.7 on 12 degrees of freedom

Multiple R-squared: 0.9491, Adjusted R-squared: 0.9448

F-statistic: 223.6 on 1 and 12 DF, p-value: 4.026e-09

Các chỉ số R-squared = 0.9491 không thay đổi so với mô hình có biến là Intensity.of.Global.radiation. Tuy nhiên ý nghĩa của tham số a và b thì khác so với mô hình trước.

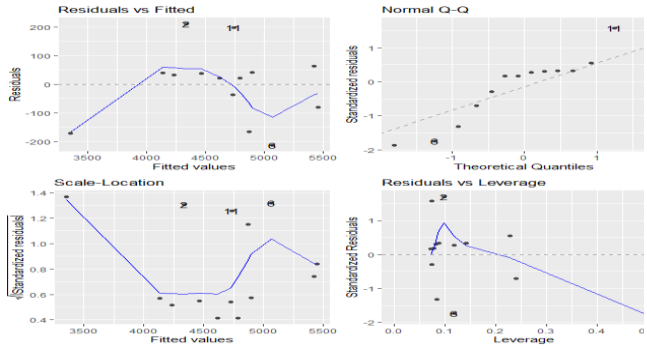
- Tham số a = 4649.66 có nghĩa khi zIntensity.of.Global.radiation =0 (tức khi Intensity.of.Global.radiation = 5.01 kWh/m^2 = giá trị trung bình)
- Tham số b = 541.91, có nghĩa là khi Intensity.of.Global.radiation tăng 1 độ lệch chuẩn (0.55) sản lượng điện năng do tấm pin mặt trời sản xuất ra tăng 541.91 kWh.

Trong trường hợp nghiên cứu mô hình hồi quy tuyến tính này, tác giả hoán vị sang đơn vị z vì ý nghĩa thực tế của tham số.

Tiến hành kiểm tra các giả định bằng hàm autoplot() trong gói ggfortify.

>library(ggfortify)

>autoplot(M1)



Hình 5. Biểu đồ phân tích giả định

Biểu đồ phần trên và bên trái trình bày mối liên quan giữa giá trị dự báo với phần dư, cho thấy các phần dư xoay quanh giá trị 0, tức là đúng với giả định rằng giá trị trung bình của phần dư bằng 0.

Biểu đồ phía trên và bên phải trình bày mối tương quan giữa giá trị lý thuyết và thực tế của phần dư. Nếu phần dư tuân theo luật phân bố bình thường thì các giá trị nằm trên đường lý thuyết, và trong trường hợp phân tích này các phần dư đều xấp xỉ xoay quanh đường lý thuyết. Điều này có nghĩa là giả định về phân bố bình thường của mô hình là có thể chấp nhận được.

Biểu đồ phần dưới bên trái chỉ ra mối tương quan giữa giá trị dự báo và căn bậc hai của phần dư. Biểu đồ này cho chúng ta biết phương sai của phần dư có hay không có liên quan với giá trị của biến x. Biểu đồ cho thấy không có mối liên quan.

Biểu đồ bên dưới và bên phải trình bày giá trị "leverage" và phần dư chuẩn hóa. Biểu đồ này cho chúng ta biết có những giá trị có ảnh hưởng cao hay không. Tất cả đều có giá trị phần dư nằm trong khoảng -2 đến +2, chúng ta chấp nhận không có giá trị ngoại vi ảnh hưởng đến mô hình.

Như vậy, phần phân tích trên cho chúng ta mô hình hồi quy tuyến tính giản đơn. Hai chỉ số chính để đánh giá sự hữu dụng của mô hình hồi quy tuyến tính là hệ số R^2 và phương sai. Mô hình có R^2 cao có nghĩa là mô hình giải thích nhiều phương sai giảm độ bất định nên MSE sẽ thấp. Mô hình có R^2 thấp thì tính bất định của tiên lượng sẽ cao và điều này cũng phản ánh giá trị MSE tăng [2], [4].

Tiếp tục dùng lệnh để phân tích phương sai:

> anova(M1)

Analysis of Variance Table

Response: Quantity.PV

Df Sum Sq Mean Sq F value Pr(>F)

Intensity.of.Global.radiation 1 3817645 3817645 223.61 4.026e-09 ***

Residuals 12 204872 17073

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

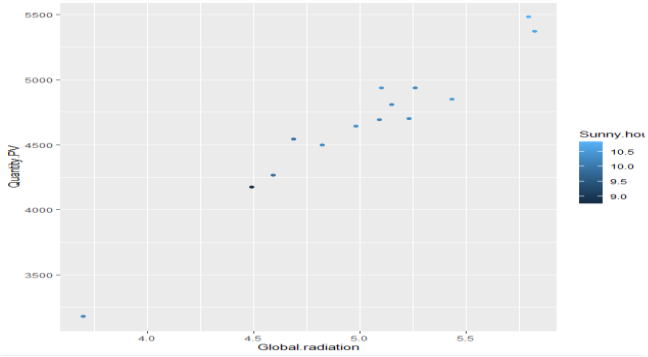
Phương sai của mô hình mean squared error (MSE) có thể hiểu là phương sai của y sau khi hiệu chỉnh cho x. Trong thực tế, MSE được ước tính từ phần dư bởi vì phần dư phản ánh phương sai mà mô hình không giải thích được. Trong phân tích phương sai ở bảng trên thì MSE = 17073

Vậy mô hình 1 đáp ứng các giả định và có hệ số R^2 rất cao. Một yếu tố có thể ảnh hưởng

đến sản lượng đó là thời gian có nắng. Tác giả tiếp tục đưa ra mô hình thứ hai, phân tích biến thời gian có nắng **Sunny.hours** đến sản lượng điện năng do tấm pin mặt trời sản xuất ra.

Mô hình 2: Quantity.PV~ Sunny.hours

>k=ggplot(data=m,aes(x=Intensity.of.Global.radiation,y=Quantity.PV,col=Sunny.hours))+geom_point()



Hình 6. Biểu đồ mối tương quan giữa sản lượng và thời gian có nắng

Biểu đồ cho thấy điểm có màu xanh nhạt là thời gian có nắng nhiều, sản lượng sản xuất ra cũng có xu hướng tăng.

>
M2=lm(Quantity.PV~Sunny.hours,data=m)

> summary(M2)

Call:

lm(formula = Quantity.PV ~ Sunny.hours, data = m)

Residuals:

Min	1Q	Median	3Q	Max
-1453.34	-74.06	62.09	223.81	454.06

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1491.9	2750.9	-0.542	0.5975
Sunny.hours	602.4	269.5	2.235	0.0452 *

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

Residual standard error: 486.5 on 12 degrees of freedom

Multiple R-squared: 0.2939, Adjusted R-squared: 0.2351

F-statistic: 4.996 on 1 and 12 DF, p-value: 0.04519

Tương tự, tác giả kiểm tra mối tương quan của biến số giờ có nắng Sunny.hours và Quantity.PV theo phương pháp Pearson qua hàm:

>cor.test(x=m$Sunny.hours,y=m$Quantity.PV).

Kết quả là r = 0.54>0, mối tương quan giữa Sunny.hours và Quantity.PV là quan thuận với nhau. R^2 = 0.29 có nghĩa là biến độc lập Sunny.hours “giải thích” khoảng 29% sự biến thiên của biến phụ thuộc Quantity.PV. Phần còn lại 71% được giải thích bởi các biến ngoài mô hình và sai số ngẫu nhiên. Cột Estimate cho ta kết quả ước tính hai tham số của mô hình hồi quy tuyến tính. Theo đó, a= -1491.9 và b=602.4. Do đó, mô hình bây giờ là:

M2: Quantity.PV = -1491.9 + 602.4 Sunny.hours

Phân tích phương sai:

>anova(M2)

Analysis of Variance Table

Response: Quantity.PV

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Sunny.hours	1	1182365	1182365	4.9956	0.04519 *
Residuals	12	2840152	236679		

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

MSE = 236679

Vậy so với mô hình 1, R^2 thấp hơn rất nhiều, tính bất định của biến dự báo cao và điều này cũng phản ánh giá trị MSE tăng.

Mô hình 3:

Quantity.PV~Intensity.of.Global.radiation+Sunny.hours

Tác giả thử xem xét đưa biến sunny.hours vào mô hình.

>M3=lm(Quantity.PV~Intensity.of.Global.radiation+Sunny.hours,data=m)

>summary(M3)

Call:

lm(formula = Quantity.PV ~ Intensity.of.Global.radiation + Sunny.hours, data = m)

Residuals:

Min 1Q Median 3Q Max

-201.47 -71.88 21.08 46.35 218.90

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 274.77 760.34 0.361 0.725

Intensity.of.Global.radiation 1034.69 83.96 12.323 8.85e-08 ***

Sunny.hours -79.34 91.72 -0.865 0.405

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 132.1 on 11 degrees of freedom

Multiple R-squared: 0.9523, Adjusted R-squared: 0.9436

F-statistic: 109.8 on 2 and 11 DF, p-value: 5.385e-08

Phân tích phương sai của mô hình trên:

> anova(M3)

Analysis of Variance Table

Response: Quantity.PV

Df Sum Sq Mean Sq F value Pr(>F)

Intensity.of.Global.radiation 1 3817645 3817645 218.9217 1.316e-08 ***

Sunny.hours 1 13050 13050 0.7483 0.4055

Residuals 11 191822 17438

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Kết quả phân tích cho ra mô hình:

M3: Quantity.PV = 274.77 + 1034.69 Intensity.of.Global.radiation -79.34 Sunny.hours

Xem xét hiện tượng đa cộng tuyến:

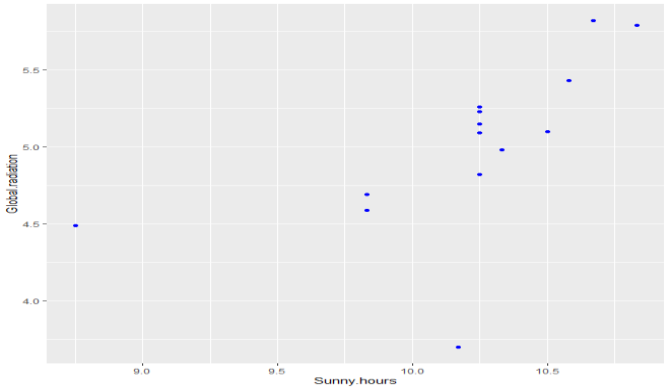
Kiểm tra hiện tượng đa cộng tuyến khi ước lượng hệ số hồi quy cho biến **Sunny.hours** là -79.34 tức là sản lượng điện năng giảm khi số giờ có nắng tăng.

Xem xét mối liên quan giữa cường độ bức xạ và số giờ có nắng. Hệ số tương quan ở mức 0,603.

>with(data=m,cor(Intensity.of.Global.radiation,Sunny.hours))

[1] 0.6031506

>ggplot(data=m,aes(x=Sunny.hours,y=Intensity.of.Global.radiation))+geom_point(col='blue')



Hình 7. Biểu diễn tương quan giữa cường độ bức xạ và số giờ có nắng

Phương pháp phát hiện và định lượng đa cộng tuyến. Trong R chúng ra có thể tính VIF qua hàm vif trong chương trình car.

>f=lm(Quantity.PV~Intensity.of.Global.radiation+Sunny.hours,data=m)

> library(car)

> vif(f)

Intensity.of.Global.radiation Sunny.hours

1.57181 1.57181

VIF=1.57<5 nên giữa Intensity.of.Global.radiation và Sunny.hours không xảy ra hiện tượng đa cộng tuyến.

3. Lựa chọn mô hình

1	Quantity.PV = -314.68 + 990.89 Intensity.of.Global.radiation	R^2 = 0.94
		MSE = 17073
		r=0.97>0
2	Quantity.PV = -1491.9 + 602.4 Sunny.hours	R^2 = 0.29
		MSE = 236679
		r = 0.54>0
3	Quantity.PV = 274.77 + 1034.69 Intensity.of.Global.radiation - 79.34 Sunny.hours	R^2 =0.9523

		MSE = 17438
		VIF=1.57<5

Mô hình so với mô hình 1 và 2, mô hình 3 có một số điểm chú ý:

- Khi số giờ có nắng tăng lên 1h thì sản lượng điện năng giảm xuống 79.34kWh điều này trái với lý thuyết kinh tế.

R^2 =0.9523 và phương sai của phần dư MSE = 17438 tăng hơn so với mô hình 1. Như vậy mô hình dự báo không tốt so với mô hình 1 đồng thời ảnh hưởng của yếu tố thời gian có nắng Sunny.hours bây giờ không có ý nghĩa thống kê (P=0.405)

Vậy giữa 3 mô hình đưa ra và phân tích trên ngôn ngữ R, tác giả lựa chọn mô hình 1. Trong các biến biến thu thập, có biến cường độ bức xạ Intensity.of.Global.radiation có ảnh hưởng nhiều đến sản lượng điện năng tấm pin mặt trời sản xuất ra, còn biến số giờ có nắng Sunny.hours có mức độ ảnh hưởng không đáng kể.

Dùng mô hình 1 để dự báo sản lượng điện năng tạo ra từ tấm pin mặt trời khi bức xạ thay đổi.

Chúng ta có thể dùng hàm predict với đối số interval = "prediction" để ước tính giá trị dự báo và khoảng tin cậy 95% của giá trị dự báo.

```
>j=data.frame(Intensity.of.Global.radiation =c(4.5,5,5.5))
>predict(d,j,interval="prediction")
      fit     lwr     upr
1 4144.306 3840.564 4448.047
2 4639.748 4345.064 4934.433
3 5135.191 4832.137 5438.245
```

Nếu cường độ bức xạ tại điểm đo là 4.5 kWh/m^2 thì sản lượng điện năng tấm pin mặt trời sản xuất ra là 4144.306 kWh, dao động trong khoảng 3840.564 đến 4448.047kWh.

KẾT LUẬN

Bài báo sử dụng R trong phân tích các yếu tố ảnh hưởng đến sản lượng điện năng của nhà máy điện mặt trời có công suất 1195kWp. Với bộ dữ liệu khảo sát, thu thập được, tác giả phân tích các yêu tố ảnh hưởng đến sản lượng điện năng do tấm pin mặt trời sản xuất ra (Quantity.PV) và lựa chọn mô hình hồi quy tuyến tính phù hợp. Hàm pairs.panels cung cấp biểu đồ tương quan từng biến và đường biểu diễn một cách trực quan, tốc độ xử lý nhanh hơn rất nhiều so với Excel, SPSS… Kết quả cho thấy biến Quantity.PV có mối liên quan mật thiết với các biến: Intensity.of.Global.radiation và Sunny.hours do có hệ tương quan cao là 0.97 và 0.54.

Bằng cách sử dụng plot và abline(lm), tác giả chỉ ra ý nghĩa của các tham số trong mô hình, cách đánh giá tầm quan trọng của các biến tiên lượng, quy trình xây dựng và kiểm định mô hình dự báo xem xét đến cả các vấn đề đa cộng tuyến và hoán chuyển dữ liệu qua hàm scale. Mô hình được lựa chọn là:

Quantity.PV = -314.68 + 990.89 Intensity.of.Global.radiation, có R^2 = 0.94, MSE = 17073.

Hàm predict cho kết quả nếu cường độ bức xạ tại điểm đo là 4.5 kWh/m^2 thì sản lượng điện năng tấm pin mặt trời sản xuất ra là 4144.306 kWh, dao động trong khoảng 3840.564 đến 4448.047kWh.

Khi thị trường điện phát triển, các dự án điện mặt trời nối lưới, vấn đề dự báo hay chào giá sản lượng điện năng do các dự án mặt trời tạo ra có ý nghĩa rất quan trọng. Mô hình hồi quy tuyến tính mà tác giả lựa chọn chỉ ra biến cường độ bức xạ mặt trời ảnh hưởng chủ yếu đến sản lượng điện năng mà tấm pin mặt trời sản xuất ra. Đồng thời, thay vì việc đo bức xạ mặt trời theo thiết bị đo cầm tay, việc cập nhật bức xạ mặt trời nên được gắn với hệ thống thiết bị đo quan trắc và được tích hợp cùng với hệ quản lý năng lượng từ xa bao gồm các thông tin theo chuỗi thời gian về cường độ bức xạ mặt trời, sản lượng điện năng để liên tục cập nhật số liệu, phục vụ cho công tác thu thập số liệu, phân tích số liệu để dự báo sản lượng điện năng.

TÀI LIỆU THAM KHẢO

[1]. Nguyễn Văn Tuấn, "Mô hình hồi quy và khám phá khoa học", 323, NXB tổng hợp thành phố Hồ Chí Minh, 2020.

[2]. Nguyễn Văn Tuấn, "Phân tích dữ liệu với R", 520, NXB Thành phố Hồ Chí Minh, 2020

[3]. Cole nussbaumer knaflic, dịch giả: Hồ Vũ Thanh Phong, "Storytelling with data let practice, 419, Wiley, 2020

[4]. Robert I.Kabacoff, "R in action data analysis and graphics with R", 608, Manning publications, 2015.

[5]. Nina Zumel, John Mount, "Practical data science with R", 519, Manning publications, 2020.

[6]. Peter Bruce, Andrew Bruce, and Peter Gedeck, "Practical Statistics for Data Scientists 50 + Essential Concepts using R and python, 342, O'Reilly, 2020

[7]. Joseph F.Hair JR, William C.Black, Barry J.Babin, Rolph E. Anderson, "Multivariate data analysis, 760, Pearson Prentice Hall, 210.

[8]. Peter Dalgaard, "Introductory statistics with R" 200, Springer, 2004.

[9]. Julian Faraway, "Linear Models with R", 213, Chapman & Hall/CRC, 2004

[10]. Paul Murrell, "R Graphics (Computer Science and Data Analysis)", 250, Chapman & Hall/CRC, 2005.

Giới thiệu tác giả:

Tác giả Chu Văn Tuấn, tốt nghiệp trường Đại học Điện Lực năm 2012, nhận bằng thạc sĩ ngành Hệ thống điện năm 2014 tại trường Đại học Điện Lực.

Lĩnh vực nghiên cứu: bù trơn công suất phản kháng, lưới điện thông minh, năng lượng tái tạo, tinh gọn chuỗi giá trị, khởi nghiệp đổi mới sáng tạo.

Tác giả Nguyễn Thúy Ninh, tốt nghiệp trường Đại học Điện Lực năm 2012, nhận bằng thạc sĩ ngành Quản lý Năng lượng năm 2014 tại trường Đại học Điện Lực.

Lĩnh vực nghiên cứu: dự báo nhu cầu phụ tải, thị trường điện, năng lượng tái tạo, nhiên liệu than và lò hơi.