Phân tích ảnh hưởng của các chỉ số sức khỏe đến tiến triển bệnh đái tháo đường *

Thái Minh Triết[1], Chu Hà Thảo Ngân[2], Võ Tuấn Anh[3], and Đỗ Trọng Hợp[4]

[1] Trường Đại học Công nghệ thông tin, Đại học quốc gia Thành phố Hồ Chí Minh
[2] Khoa Khoa học và Kỹ thuật Thông tin
{[1]19522397,[2]19521882,[3]19521226}@gm.uit.edu.vn
[4]hopdt@uit.edu.vn

Tóm tắt nội dung Bệnh đái tháo đường theo giới y khoa là một trong những mối đe doạ sức khoẻ toàn cầu đối với cộng đồng có tỉ lệ người mắc ngày càng tăng đều theo từng năm với nhiều nguyên nhân ảnh hưởng, trong đó nguyên nhân chính là rối loạn chức năng chuyển hóa insulin. Nhằm tìm hiểu được các yếu tố bên trong cơ thể có tác động như thế nào đến sự rối loạn này, nhóm chúng tôi đã tìm hiểu các chỉ số y tế mang tính thống kê có liên quan bằng bộ dữ liệu có sẵn và tiến hành phân tích, xử lí và đánh giá để có được những kết quả phù hợp cho mục đích của bài nghiên cứu. Trong quá trình phân tích có sử dụng phương pháp thống kê phân tích hồi quy để tìm ra sự ảnh hưởng và tương tác lẫn nhau của các yếu tố này. Sau khi thực hiện xây dựng các mô hình máy học phù hợp và đánh giá chúng, kết quả thu được là <KẾT QUẢ ĐÁNH GIÁ> cho mô hình phân tích tốt nhất <MÔ HÌNH TỐT NHẤT>. <Lược ý kết luận>

Keywords: Đái tháo đường · Bệnh tiểu đường · Phân tích hồi quy

1 Giới thiệu chung

Đái tháo đường được xem là một căn bệnh phổ biến trong cộng đồng, đặc biệt là Mỹ với xu hướng gia tăng ngày càng cao với con số ước tính lên đến 79 triệu người trưởng thành mắc bệnh và trong đó 50% người mắc đái tháo đường không biết mình có bệnh vì chưa có triệu chứng rõ ràng.

Đái tháo đường là một căn bệnh liên quan đến sự rối loạn chuyển hóa trong hormone tuyến tụy dẫn đến thiếu hụt lượng insuline trong máu. Các tiêu chuẩn để xác định, chuẩn đoán bệnh đái tháo đường hiện nay được Hiệp hội Đái tháo đường Mỹ công bố cần dựa vào nhiều chỉ số và quá trình phức tạp. Ví dụ khi kiểm tra chỉ số HbA1c của bệnh nhân cần được thực hiện ở phòng thí nghiệm được chuẩn hóa cao. Tuy nhiên, cũng nhờ đó mà việc chuẩn đoán đái tháo đường hay tiền đái tháo đường sẽ phần lớn phụ thuộc vào các chỉ số này từ cơ thể. Chính vì đó, ta cần phải có được một bộ dữ liệu với các chỉ số y tế liên quan để

* Hướng dẫn bởi TS. Đỗ Trọng Hợp

có thể đánh giá phần nào tình trạng hiện tại của bệnh, giúp tăng hiệu quả cũng như giảm đi chi phí cho quá trình chuẩn đoán.

Trong bài nghiên cứu này, bộ dữ liệu thu thập được từ các bệnh nhân đái tháo đường với các chỉ số y tế liên quan sẽ được đánh giá bằng các mô hình phân tích hồi quy. Từ đó không chỉ hỗ trợ tích cực cho việc chuẩn đoán mà còn có thể phân tích được sự ảnh hưởng và tương tác giữa các chỉ số trên với nhau.

Cấu trúc bài báo cáo được trình bày như sau: Ở mục 2, chúng tôi trình bày thông tin chi tiết của bộ dữ liệu được sử dụng trong bài nghiên cứu. Ở mục 3, chúng tôi tiếp cận bộ dữ liệu và thực hiện tiền xử lý dữ liệu. Sau khi bộ dữ liệu đã được xử lí, chúng sẽ được đưa vào phân tích bằng các mô hình hồi quy khác nhau, quá trình này sẽ được trình bày ở mục 4. Mục 5 sẽ trình bày kết quả đánh giá các mô hình, từ đó có được mô hình tốt nhất trên bộ dữ liệu. Và cuối cùng ở mục 6 là kết luận và hướng phát triển.

2 Tổng quan bộ dữ liệu

2.1 Giới thiệu chung về bộ dữ liệu

Bộ dữ liệu được xây dựng bao gồm thuộc tính cơ bản của 442 bệnh nhân bệnh tiểu đường và kết quả là một giá trị đáng giá quá trình tiến triển của bệnh nhân sau 1 năm ghi nhận. Dưới đây là ví dụ các điểm dữ liệu được trích ra từ bộ dữ liệu và codebook của bộ dữ liệu.

Bảng 1. Ví dụ về các điểm dữ liệu

AGE	**SEX**	**BMI**	**BF**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**Y**
59	2	32.1	101	157	93.2	38	4	4.8598	87	151
48	1	21.6	87	183	103.2	70	3	3.8918	69	75
72	2	30.5	93	156	93.6	41	4	4.6728	85	141
24	1	25.3	84	198	131.4	40	5	4.8903	89	206
50	1	23	101	192	125.4	52	4	4.2905	80	135

Bảng 2. Codebook của bộ dữ liệu

Thông tin	**Nội dung**
Tên bộ dữ liệu	Diabetes dataset
Nguồn thu nhập	`https://www4.stat.ncsu.edu/~boos/var.select/diabetes.html`
Số điểm dữ liệu	442
Số thuộc tính	11
Thông tin thuộc tính	**AGE:** độ tuổi của bệnh nhân **SEX:** giới tính của bệnh nhân - Giá trị 1 tương ứng với giới tính là Nam - Gái trị 2 tương ứng với giới tính là Nữ **BMI:** Body Mass Index (chỉ số khối cơ thể) **BP:** average blood pressure (giá trị huyết áp trung bình) **S1:** tổng lượng cholesterol trong huyết thanh (tc - total serum cholesterol) **S2:** giá trị lipoprotein tỷ trọng thấp (ldl - low-density lipoproteins) **S3:** giá trị lipoprotein tỷ trọng cao (hdl - high-density lipoproteins) **S4:** tỉ lệ giữa cholesterol toàn phần so với lượng HDL (tch - total cholesterol / HDL) **S5:** mức triglyceries có trong huyết thanh có thể ghi nhận (ltg - possibly log of serum triglycerides level) **S6:** chỉ số mức đường huyết (glu - blood sugar level) **Y:** giá trị định lượng về tiến triển bệnh của bệnh nhân sau 1 năm kể từ thời điểm ghi nhận
Tác giả	Bradley Efron, Trevor Hastie, Iain Johnstone và Robert Tibshirani, Đại học Stanford

2.2 Trực quan bộ dữ liệu

3 Tiền xử lý dữ liệu

3.1 Xử lý ngoại lệ

3.2 Chuyển đổi dữ liệu

3.3 Phân chia tập train và test

4 Phân tích hồi quy

4.1 Tổng quan về Phân tích Hồi quy (Regression Analysis)

Phân tích hồi quy là một phương pháp thống kê mạnh mẽ cho phép xem xét mối tương quan giữa các biến số trong việc đo lường. Có nhiều loại phân tích hồi quy khác nhau nhưng chúng tựu chung đều tập trung vào việc đánh giá sự ảnh hưởng của các biến số độc lập lên một biến phụ thuộc.

Đây cũng là một phương pháp đáng tin cậy để xác định xem liệu biến số nào có ảnh hưởng đến bài toán đo lường (topic of interest). Quá trình thực hiện hồi quy cho phép xác định được yếu tố nào đáng quan tâm nhất và những yếu tố nào có thể bỏ qua, hay sự ảnh hưởng, tương tác lẫn nhau của các yếu tố.

Để hiểu một cách đầy đủ về phương pháp phân tích hồi quy, trước hết cần phải nắm vững một vài thuật ngữ sau:

- **Biến phụ thuộc**: yếu tố chính cần được ước lượng, hiểu rõ và dự đoán
- **Biến độc lập**: những yếu tố được giả thiết rằng có ảnh hưởng đến biến phụ thuộc

4.2 Phân tích ảnh hưởng của yếu tố

4.3 Tương tác giữa các yếu tố

4.4 Các độ đo đánh giá mô hình hồi quy

Đánh giá mô hình là một bước quan trọng không thể thiếu trong quá trình xây dựng mô hình máy học. Đánh giá mô hình hỗ trợ trong việc đánh giá hiệu suất của mô hình trên dữ liệu huấn luyện và phát hiện ra những trường hợp gây ảnh hưởnng tới mô hình như over-fitting (hiện tượng mô hình thể hiện chính xác các điểm dữ liệu huấn luyện) hay under-fitting (hiện tượng mô hình không hoàn toàn có mối liên hệ nào với bộ dữ liệu huấn luyện), từ đó cải thiện chất lượng hiệu suất của mô hình và giúp mô hình đưa ra kết quả khả quan hơn khi xuất hiện bộ dữ liệu mới trong tương lai. Độ đo được sử đụng để đánh giá cho các mô hình máy học hồi quy trong bài báo cáo là Adjusted R-Squared.

R-Squared

$$R - Squared = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(Ytrue - Ypred)^2}{\sum(Ytrue - \overline{Y})^2} \tag{1}$$

Trong đó:

- **SSE** là Sum of Squared estimate of errors (tổng bình phương lỗi giữa biến phụ thuộc trong bộ dữ liệu với biến phụ thuộc được dự đoán từ mô hình).
- **SST** là sum of squares total (tổng bình phương lỗi giữa giá trị biến phụ thuộc trong bộ dữ liệu với tổng bình phương của nó).

Độ đo đánh giá R-Squared hay còn được còn có tên gọi khác là Coefficient of Determination, có khoảng giá trị từ 0 tới 1, giúp nhận xét mức độ thích hợp giữa mô hình và các điểm dữ liệu huấn luyện. Kết quả của độ đo R-Squared không đánh giá hiệu suất của mô hình một cách toàn vẹn, giá trị của độ đo thấp không có nghĩa là mô hình có hiệu suất thấp, ngược lại thì kết quả độ đo có cao thì không đồng nghĩa rằng mô hình hoạt động tốt. Ngoài ra, độ đo R-Squared còn có nhược điểm khi tăng số lượng thuộc tính độc lập thì giá trị của độ đo cũng tăng theo, dẫn tới tăng độ phức tạp cho mô hình mặc dù những thuộc tính mới không hoàn toàn mang ý nghĩa cho việc cải thiện dự đoán biến phụ thuộc.

Adjusted R-Squared

$$AdjustedR - Squared = 1 - (1 - R^2)\frac{n-1}{n-p-1} \tag{2}$$

Trong đó:

- $\boldsymbol{R^2}$: giá trị của R-Square trên bộ dữ liệu.
- **n:** số lượng điểm dữ liệu trong bộ dữ liệu.
- **p:** số lượng thuộc tính độc lập trong bộ dữ liệu.

Độ đo đánh giá Adjusted R-Squared đã khắc phục khuyết điểm vốn có của R-Squared, khi tăng số lượng thuộc tính độc lập nếu giá trị của độ đo Adjusted R-Squared giảm nghĩa là thuộc tính đó không có ý nghĩa trong việc nâng cao hiệu suất mô hình tránh trường hợp dư thừa hay tăng độ phức tạp cho mô hình một cách không cần thiết.

4.5 Xây dựng mô hình hồi quy

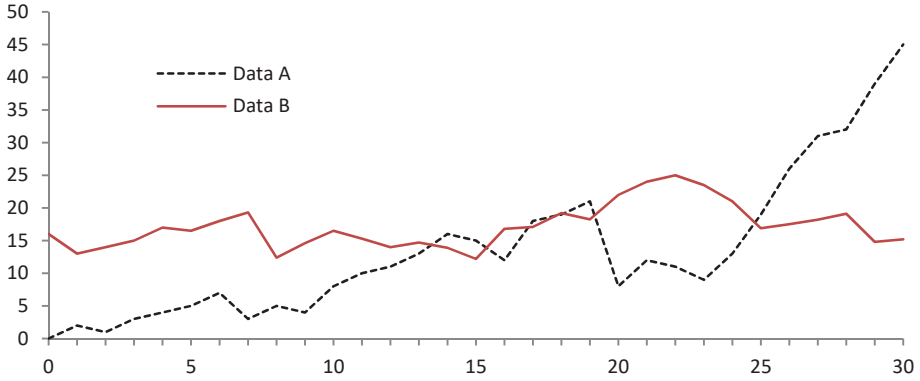
Simple Linear Regression là mô hình hồi quy tuyến tính đơn biến, nghĩa là mô hình chỉ có duy nhất một biến độc lập. Mô hình được biểu diễn qua phương trình: $y = \beta_0 + \beta_1 x$, trong đó y là biến phụ thuộc, x là biến độc lập và β_0, β_1 là các tham số mô hình. Đặc trưng của mô hình là một đường thẳng tuyến tính giúp đánh giá mối quan hệ giữa biến độc lập và biến phụ thuộc.

Vì bộ dữ liệu có 10 biến độc lập nên việc chọn thuộc tính dựa vào hệ số tương quan cao nhất giữa biến phụ thuộc Y và các thuộc tính còn lại.

Bảng 3. Hệ số tương quan giữa biến phụ thuộc Y và các biến độc lập

Biến độc lập	Hệ số tương quan
AGE	0.19
SEX	0.043
BMI	0.59
BP	0.44
S1	0.21
S2	0.17
S3	−0.39
S4	0.43
S5	0.57
S6	0.38

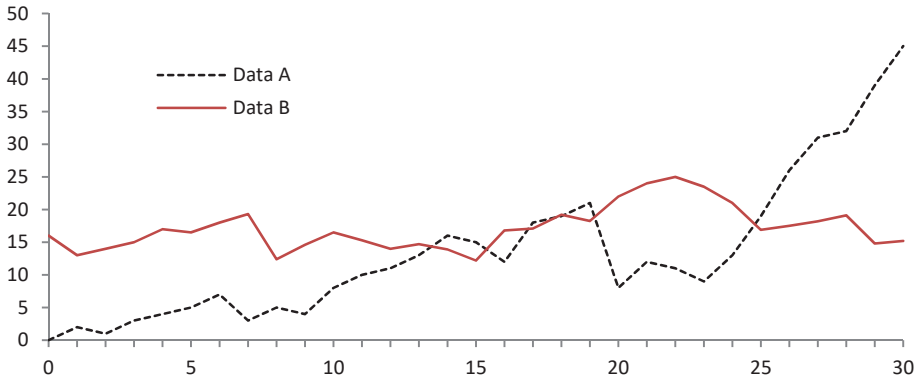
Thông qua bảng 4, biến độc lập được sử dụn trong mô hình là BMI với hệ số tương quan là 0.59. Quá trình xây dựng mô hình Simple Linear Regression được thực trên R - Studio thông qua hàm lm với các tham số: lm(formula = $Y \sim BMI$,data=train)



Hình 1. Phân tích hồi quy trên mô hình Simple Linear Regression trước khi tiền xử lí

Từ hình 1, mô hình Simple Linear Regression trước khi tiền xử lí có dạng là:

$$\mathbf{Y} = + * \mathbf{BMI}$$



Hình 2. Phân tích hồi quy trên mô hình Simple Linear Regression sau khi tiền xử lí

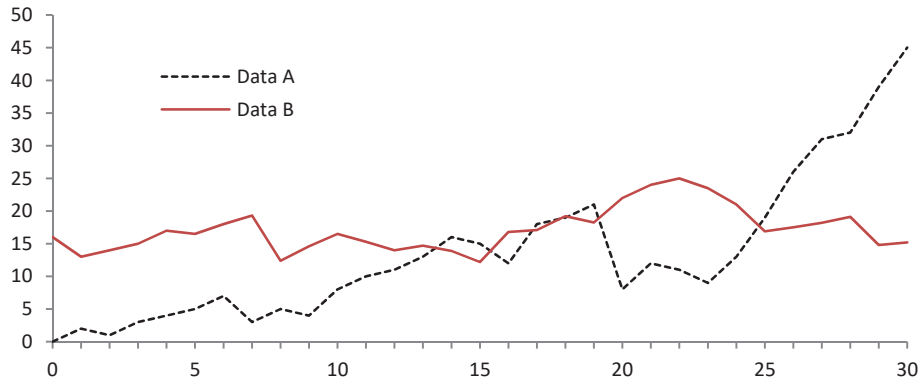
Từ hình 2, mô hình Simple Linear Regression sau khi tiền xử lí có dạng là:

$$\mathbf{Y} = + * \mathbf{BMI}$$

Multiple Linear Regression là trường hợp khi mô hình hồi quy tuyến tính tồn tại nhiều hơn một biến độc lập. Khi đó, phương trình biểu diễn cho mô hình

là: $y = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + ... + \beta_n * x_n$ với x,y,$\beta_0, \beta_1, ..., \beta_n$ có ý nghĩa tương tự như mô hình Simple Linear Regression.

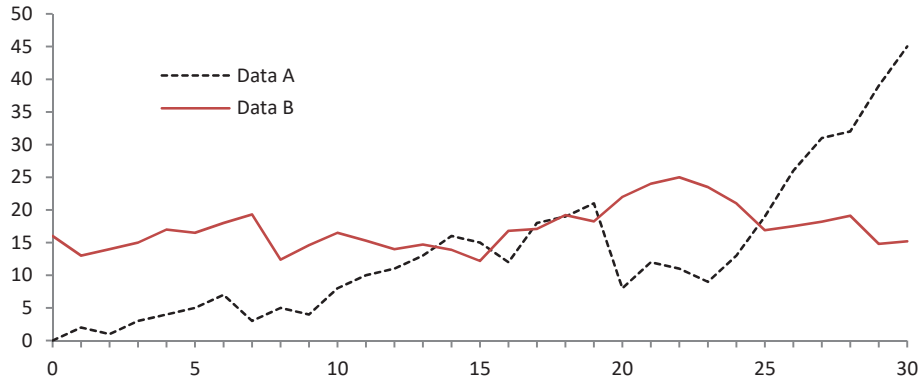
Dựa vào phân tích ảnh hưởng các yếu tố ở mục 4.2, các thuộc tính có ảnh hưởng sẽ được sử dụng để xây dựng mô hình Multiple Linear Regressionl là AGE, BMI, BP, S3, S5. Sau đó tiếp tục phân tích hồi quy để chọn ra thuộc tính có ảnh hưởng.Các tham số thuộc tính được chọn sau khi phân tích hồi quy trong hàm lm là lm(formula = $Y \sim BMI + BP + S3 + S5$,data=train).



Hình 3. Phân tích hồi quy trên mô hình Multiple Linear Regression trước khi tiền xử lí

Từ hình 3, mô hình Multiple Linear Regression trước khi tiền xử lí có dạng là:

$$\mathbf{Y} = + * \mathbf{BMI} + *\mathbf{BP} + *\mathbf{S3} + *\mathbf{S5}$$



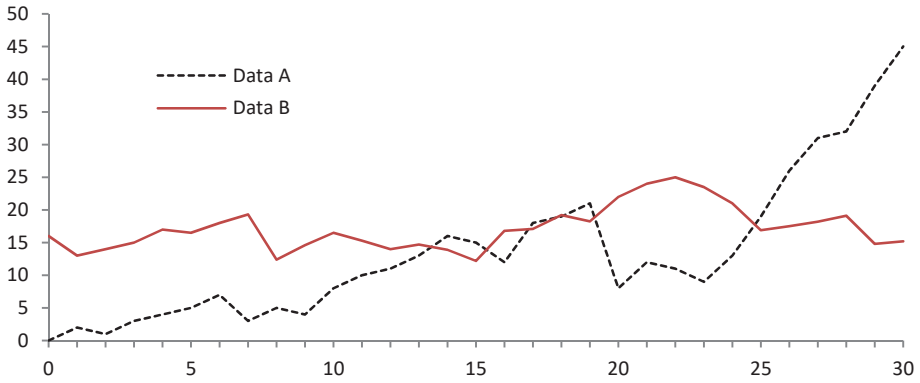
Hình 4. Phân tích hồi quy trên mô hình Multiple Linear Regression sau khi tiền xử lí

Từ hình 4, mô hình Multiple Linear Regression sau khi tiền xử lí có dạng là:

$$\mathbf{Y} = + * \mathbf{BMI} + *\mathbf{BP} + *\mathbf{S3} + *\mathbf{S5}$$

Polynomial Regression được sử dụng khi mối quan hệ giữa biến độc lập và biến phụ thuộc không thể trực quan trên một đường thẳng tuyến tính. Khi đó phương trình biểu diễn mô hình sẽ có dạng $y = \beta_0 + \beta_1 * x_1 + \beta_2 * x^2 + ... + \beta_n * x^n$ với x,y,$\beta_0, \beta_1, ..., \beta_n$ có ý nghĩa tương tự như mô hình Simple Linear Regression.

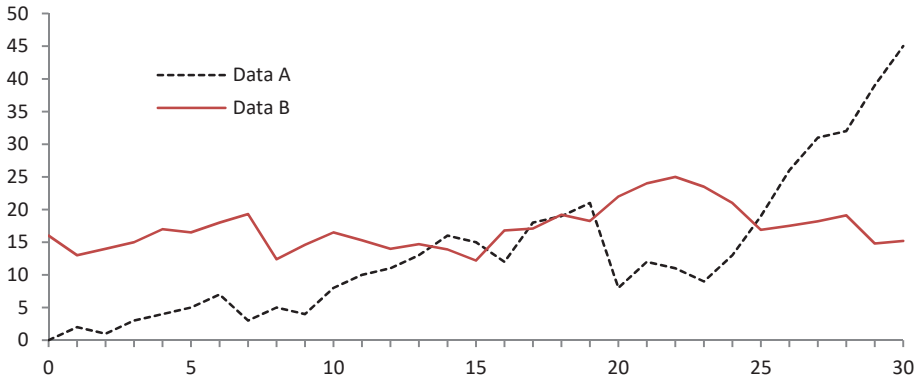
Thông qua phân tích ảnh hưởng các yếu tố ở mục 4.2 và tương tác giữa các yếu tố ở mục 4.3, các tham số được lựa chọn để xây dựng mô hình Polynomial Regression, sau đó thực hiện phân tích hồi quy để chọn ra tham số giúp mô hình đạt hiệu suất tốt nhất. Những tham số được chọn sau cùng trong hàm lm là lm(formula = $Y \sim$,data=train).



Hình 5. Phân tích hồi quy trên mô hình Polynomial Regression trước khi tiền xử lí

Từ hình 5, mô hình Polynomial Regression trước khi tiền xử lí có dạng là:

$$\mathbf{Y} = +$$



Hình 6. Phân tích hồi quy trên mô hình Polynomial Regression sau khi tiền xử lí

Từ hình 6, mô hình Multiple Linear Regression sau khi tiền xử lí có dạng là:

$$\mathbf{Y} =$$

Ridge Regression là phương pháp xác định hệ số của mô hình hồi quy đa biến trong điều kiện biến độc lập có sự tương quan chặt chẽ với nhau. Được

xem là dạng chính quy (regularization) của hồi quy tuyến tính. Dùng để giảm số lượng biến độc lập trong mô hình hồi quy, nhằm giảm các vấn đề Overfitting và Underfitting.

$$L_{ridge}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j=1}^{m}\hat{\beta}_j^2 = ||y - X\hat{\beta}||^2 + \lambda||\hat{\beta}||^2 \tag{3}$$

Lasso Regression là một phương phá dùng để giảm số lượng biến độc lập tring mô hình hồi quy nhưng sử dụng tổng các giá trị tuyệt đối của hệ số bên trong biểu thức. Vì về bản chất, Lasso Regression không chỉ giảm giá trị của hệ số mà thậm chí cố gắng loại bỏ nó. Sự khác biệt của Ridge và Lasso là Ridge sẽ không bao giờ cho giá trị của hệ số bằng 0. Hạn chế chính là không phù hợp cho một số loại dữ liệu.

$$L_{lasso}(\hat{\beta}) = \sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2 + \lambda\sum_{j=1}^{m}|\hat{\beta}_j| \tag{4}$$

Elastic Net Regression là sự kết hợp của cả hai mô hình Ridge Regression và Lasso Regression. Ưu điểm của mô hình này là không dễ dàng loại bỏ hệ số cộng tuyến.

$$L_{enet}(\hat{\beta}) = \frac{\sum_{i=1}^{n}(y_i - x_i'\hat{\beta})^2}{2n} + \lambda(\frac{1-\alpha}{2}\sum_{j=1}^{m}\hat{\beta}_j^2 + \alpha\sum_{j=1}^{m}|\hat{\beta}_j|) \tag{5}$$

5 Đánh giá mô hình hồi quy

5.1 Kết quả đánh giá mô hình

5.2 Mô hình tốt nhất

6 Kết luận

Tài liệu

7 First Section

7.1 A Subsection Sample

Please note that the first paragraph of a section or subsection is not indented. The first paragraph that follows a table, figure, equation etc. does not need an indent, either.

Subsequent paragraphs, however, are indented.

Sample Heading (Third Level) Only two levels of headings should be numbered. Lower level headings remain unnumbered; they are formatted as run-in headings.

Sample Heading (Fourth Level) The contribution should contain no more than four levels of headings. Table 4 gives a summary of all heading levels.

Bảng 4. Table captions should be placed above the tables.

Heading level	Example	Font size and style
Title (centered)	**Lecture Notes**	14 point, bold
1st-level heading	**1 Introduction**	12 point, bold
2nd-level heading	**2.1 Printing Area**	10 point, bold
3rd-level heading	**Run-in Heading in Bold.** Text follows	10 point, bold
4th-level heading	*Lowest Level Heading.* Text follows	10 point, italic

Displayed equations are centered and set on a separate line.

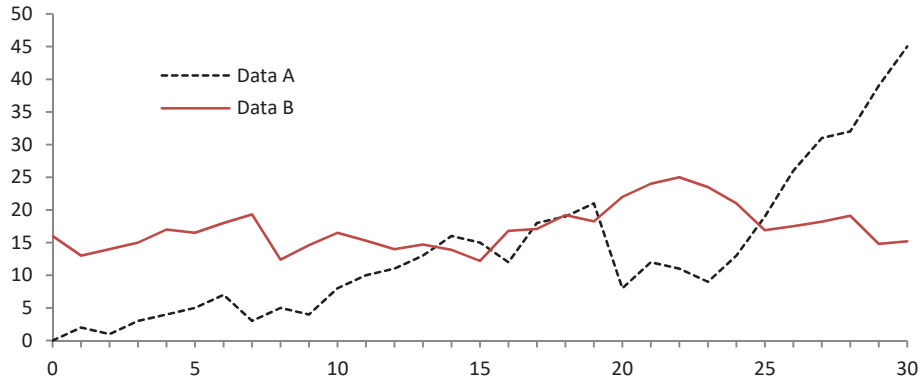
$$x + y = z \tag{6}$$

Please try to avoid rasterized images for line-art diagrams and schemas. Whenever possible, use vector graphics instead (see Fig. 7).

Theorem 1. *This is a sample theorem. The run-in heading is set in bold, while the following text appears in italics. Definitions, lemmas, propositions, and corollaries are styled the same way.*

Chứng minh. Proofs, examples, and remarks have the initial word in italics, while the following text appears in normal font.

For citations of references, we prefer the use of square brackets and consecutive numbers. Citations using labels or the author/year convention are also acceptable. The following bibliography provides a sample reference list with entries for journal articles [1], an LNCS chapter [2], a book [3], proceedings without editors [4], and a homepage [5]. Multiple citations are grouped [1–3], [1, 3–5].



Hình 7. A figure caption is always placed below the illustration. Please note that short captions are centered, while long ones are justified by the macro package automatically.

Tài liệu

1. Author, F.: Article title. Journal **2**(5), 99–110 (2016)
2. Author, F., Author, S.: Title of a proceedings paper. In: Editor, F., Editor, S. (eds.) CONFERENCE 2016, LNCS, vol. 9999, pp. 1–13. Springer, Heidelberg (2016). https://doi.org/10.10007/1234567890
3. Author, F., Author, S., Author, T.: Book title. 2nd edn. Publisher, Location (1999)
4. Author, A.-B.: Contribution title. In: 9th International Proceedings on Proceedings, pp. 1–2. Publisher, Location (2010)
5. LNCS Homepage, `http://www.springer.com/lncs`. Last accessed 4 Oct 2017