

Trường Đại học Khoa học tự nhiên  
Đại học Quốc gia Thành phố Hồ Chí Minh  
Khoa Công nghệ thông tin

**AMAS-IT**  
**Báo cáo lab project 3 – Linear Regression & Analysis**

Lớp: 22CLC04

Thực hiện:  
Ngũ Kiệt Hùng - 22127134

Giảng viên hướng dẫn:  
Thầy Trần Hà Sơn  
Thầy Nguyễn Ngọc Toàn

**Tháng 8, 2024**

# Mục lục

<b>1</b>	<b>Giới thiệu chủ đề .....</b>	<b>1</b>
<b>2</b>	<b>Ý tưởng tiếp cận vấn đề .....</b>	<b>1</b>
2.1	Cơ sở, phân tích toán .....	1
2.2	Phân tích, khám phá dữ liệu thông qua tương quan hồi quy tuyến tính đơn giữa các biến.....	2
2.3	Khám phá phân bố của các biến độc lập.....	6
2.4	Khám phá mối quan hệ giữa các biến độc lập .....	9
<b>3</b>	<b>Cấu trúc cài đặt chương trình, thư viện sử dụng .....</b>	<b>13</b>
<b>4</b>	<b>Cài đặt mô hình hồi quy tuyến tính đa biến bậc nhất.....</b>	<b>13</b>
4.1	Chi tiết lựa chọn và cài đặt mô hình .....	13
4.2	Thử nghiệm và kết quả của mô hình .....	14
<b>5</b>	<b>Cài đặt mô hình hồi quy tuyến tính đơn biến bậc nhất cho từng đặc trưng.....</b>	<b>15</b>
5.1	Chi tiết lựa chọn và cài đặt .....	15
5.2	Thử nghiệm và các kết quả của các mô hình.....	16
<b>6</b>	<b>Cài đặt mô hình hồi quy tuyến tính với các hàm cơ sở khác .....</b>	<b>17</b>
6.1	Chi tiết lựa chọn và cài đặt .....	17
6.2	Thử nghiệm và các kết quả của các mô hình.....	19
<b>7</b>	<b>Kết luận .....</b>	<b>22</b>
<b>8</b>	<b>Phụ lục .....</b>	<b>22</b>
<b>9</b>	<b>Tham khảo .....</b>	<b>23</b>

## 1 Giới thiệu chủ đề

Dự đoán trình độ học vấn của học sinh dựa vào chỉ số chất lượng thông qua các bài kiểm tra và các khảo sát xung quanh biến số ảnh hưởng đến việc học là một trong những công cụ quan trọng nhằm hỗ trợ các trường, trường đại học, học viện đánh giá xu hướng của thang đo trình độ cũng như đưa ra các kết luận về các yếu tố ảnh hưởng đến khả năng thể hiện trình độ học vấn/khả năng cải thiện trình độ của các sinh viên.

Việc khai thác dữ liệu số của các yếu tố cũng như phân tích các **đặc trưng** (feature engineering) đòi hỏi ta phải có các kiến thức cơ bản liên quan đến khoa học dữ liệu và các kỹ thuật học máy. Đối với các dạng dữ liệu như yếu tố sinh hoạt ảnh hưởng đến trình độ thể hiện học vấn của học sinh, tìm ra mối liên hệ phụ thuộc giữa các biến độc lập với nhau và với một biến phụ thuộc là trình độ thể hiện học vấn (performance index) đóng vai trò chủ đạo trong việc xây dựng nên các mô hình dự đoán có độ chính xác cao.

Trong bài thực hành này, ta sẽ tìm hiểu cơ bản các kỹ thuật phân tích dữ liệu nhiều biến, khám phá siêu mặt phẳng hình thành nên bởi hàm số đa biến. Các suy luận từ việc phân tích sẽ được sử dụng nhằm xác định các đặc trưng nổi trội, giúp ta xây dựng nên các mô hình tuyến tính từ đơn giản với các hàm cơ sở xác định bởi chính các biến độc lập; đến các mô hình phức tạp hơn, áp dụng mối liên hệ tự nhiên của các dữ liệu. Cuối cùng, ta sẽ xem xét so sánh giữa các mô hình đã xây dựng và đánh giá khả năng của chúng thông qua thang đo **trung bình tổng sai số tuyệt đối** (MAE).

## 2 Ý tưởng tiếp cận vấn đề

Các kỹ thuật phân tích dữ liệu đa số được tham khảo từ [1] [2] và sẽ không đi sâu vào các định lý hoặc chứng minh cho các định lý. Ngoài ra, bài thực hành cũng tham khảo các cách ký hiệu từ những bài blog của Tiến Sĩ Vũ Hữu Tiệp [3] vì tính dễ tiếp cận, giải thích nội dung của Tiến Sĩ. Xin cảm ơn.

### 2.1 Cơ sở, phân tích toán

Trước khi đi vào phân tích, khám phá dữ liệu, bài thực hành sẽ giới thiệu sơ lược về bài toán tối ưu trong Học máy có giám sát thông qua Hồi quy tuyến tính.

Lấy vấn đề được nêu trong chương 3 của [4], ta cần khảo sát mối liên hệ giữa các đại lượng trong dữ liệu, nhằm đưa ra các suy luận mang tính “thông tin”. Một trong các mảng thống kê suy diễn quan trọng được sinh ra nhằm giải quyết vấn đề này chính là **Phân tích Hồi quy** (regression analysis), nhằm đưa ra các mô hình thống kê giải thích sự tương quan giữa một (simple regression) hay nhiều biến độc lập (multiple regression) với một biến phụ thuộc.

Thông thường, có nhiều hướng để tiếp cận bài toán hồi quy tùy thuộc vào dữ liệu đầu vào và biến số đầu ra mong muốn [5] [6], Linear Regression là phương pháp **khớp dữ liệu** (data fitting) khi dữ liệu đầu vào ta có là các điểm dữ liệu là vector các **biến độc lập**  $x = (x_1, x_2, \dots, x_n) \in \mathbb{R}^n$  và  $y \in \mathbb{R}$  gọi là **biến phụ thuộc**. Ta mong muốn xây dựng một biểu thức (hay đúng hơn là mô hình) toán học sao cho

$$y \approx \hat{f}(x)$$

Với  $\hat{y} = \hat{f}(x)$  được gọi là **giá trị dự báo**.

Bài thực hành sẽ tập trung vào phương pháp Bình phương nhỏ nhất thông thường (Ordinary Least Square) được đề xuất lần đầu bởi Gauss.

Giả định mô hình ta có dạng

$$\hat{f}(x) = \theta_1 f_1(x) + \theta_2 f_2(x) + \cdots + \theta_p f_p, \quad (\text{eq0})$$

Các hàm  $f_i: \mathbb{R}^n \rightarrow \mathbb{R}$  được gọi là các **hàm cơ sở** (basis functions). Lưu ý rằng "tuyên tính" trong Hồi quy Tuyên tính chỉ sự tuyên tính trên tham số khi ta huấn luyện mô hình thông qua data fitting.

Gọi  $x^{(i)}, y^{(i)}$  là cặp vector các biến độc lập và biến phụ thuộc ứng với điểm dữ liệu  $(i)$ , giả sử ta đã chọn được các hàm cơ sở. Ta có:

$$r^{(i)} = y^{(i)} - \hat{y}^{(i)}, \quad \hat{y}^{(i)} = \hat{f}(x^{(i)})$$

Giá trị này được gọi là **phần dư** (residual, hoặc hiểu đơn giản là sai số). Giả sử với  $N$  điểm dữ liệu, thì ứng với mỗi điểm, ta sẽ có một giá trị phần dư sau khi dự báo, vector phần dư có dạng:

$$r = (r^{(1)}, \dots, r^{(N)}) \in \mathbb{R}^N$$

Phương pháp bình phương nhỏ nhất tìm  $\theta = (\theta_1, \theta_2, \dots, \theta_p) \in \mathbb{R}^p$  của mô hình được xác định bởi (eq0) sao cho vector phần dư có chuẩn nhỏ nhất.

Ta có thể viết lại bài toán cần giải dưới dạng:

$$\theta^* = \underset{\theta \in \mathbb{R}^p}{\operatorname{argmin}} (\|X\theta - y\|_2^2) = (X^\top X)^{-1} X^\top y, \quad (\text{eq1})$$

$X$  là ma trận tham số biến độc lập,  $y$  là vector biến phụ thuộc,

$$X = \begin{bmatrix} f_1(x^{(1)}) & \cdots & f_p(x^{(1)}) \\ \vdots & \ddots & \vdots \\ f_1(x^{(N)}) & \cdots & f_p(x^{(N)}) \end{bmatrix} \in \mathbb{R}^{N \times p}, \quad y = \begin{bmatrix} y^{(1)} \\ \vdots \\ y^{(N)} \end{bmatrix} \in \mathbb{R}^N, \quad \theta^* \in \mathbb{R}^p$$

Với dạng ma trận trên, giải bài toán tối ưu theo phương pháp bình phương nhỏ nhất trở nên tầm thường và có thể thực hiện rất nhanh. Chi tiết về các bước dẫn đến đẳng thức trên được giải thích kỹ trong chương 3 của [4].

## 2.2 Phân tích, khám phá dữ liệu thông qua tương quan hồi quy tuyến tính đơn giữa các biến

Một trong các bước quan trọng nhất của việc xây dựng mô hình chính là sự hiểu biết về mối tương quan giữa các biến độc lập và biến phụ thuộc, hoặc đặc biệt hơn là giữa các biến độc lập với nhau và với cả biến phụ thuộc. Mặt khác, vì thời lượng cũng như giới hạn của bài thực hành, ta sẽ không đi sâu vào các phân tích dữ liệu outlier, dữ liệu không tồn tại (NAs), các quá trình giải quyết các vấn đề trên... Thế nên, đầu tiên, bài thực hành sẽ trực quan hóa dữ liệu nhằm xác định mối tương quan tuyến tính giữa từng đặc tính của dữ liệu với biến phụ thuộc.

Bằng cách phát hoạ các giá trị của các điểm dữ liệu ứng với từng đặc tính với performance index, ta thu được các biểu đồ sau

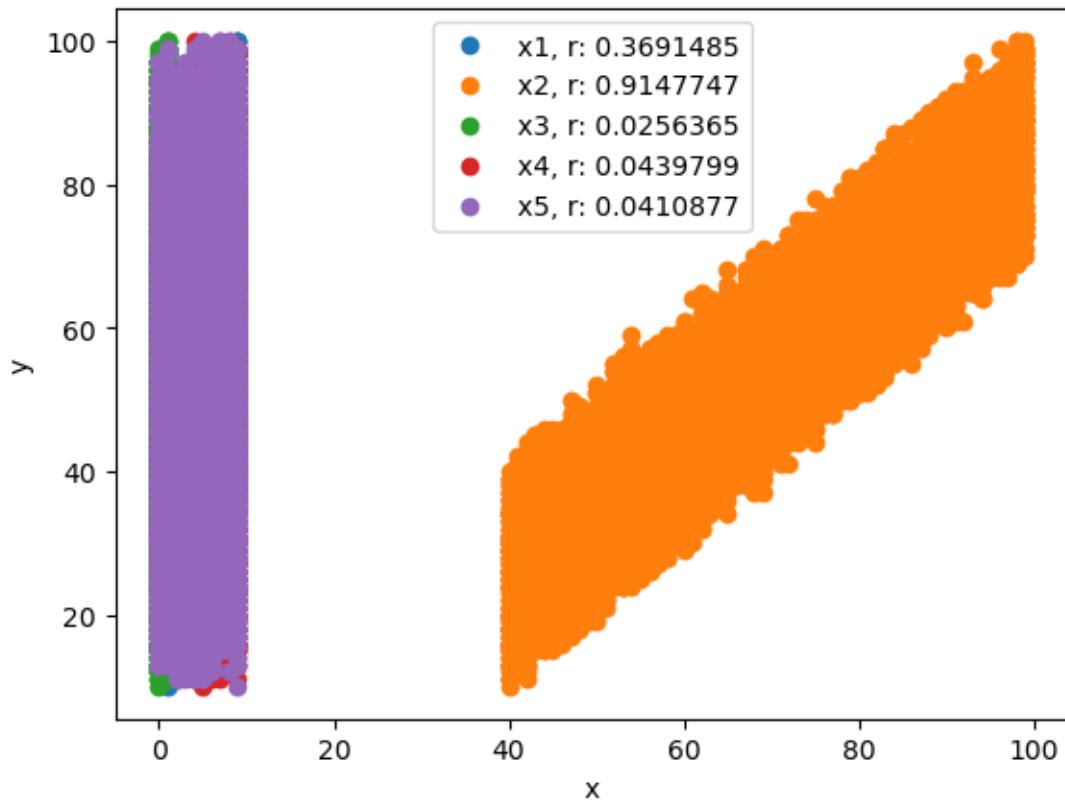


Figure 1 Tương quan giữa toàn bộ các đặc tính của dữ liệu

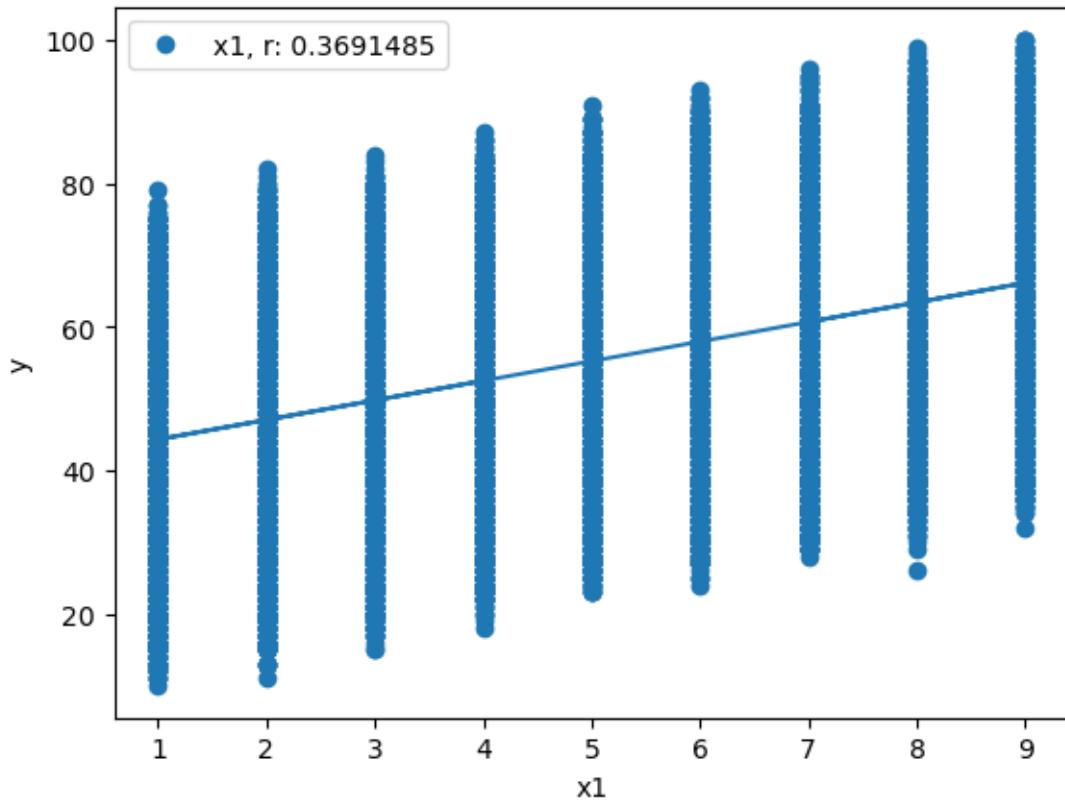


Figure 2 Tương quan giữa số giờ học mỗi ngày và điểm số hiện tại

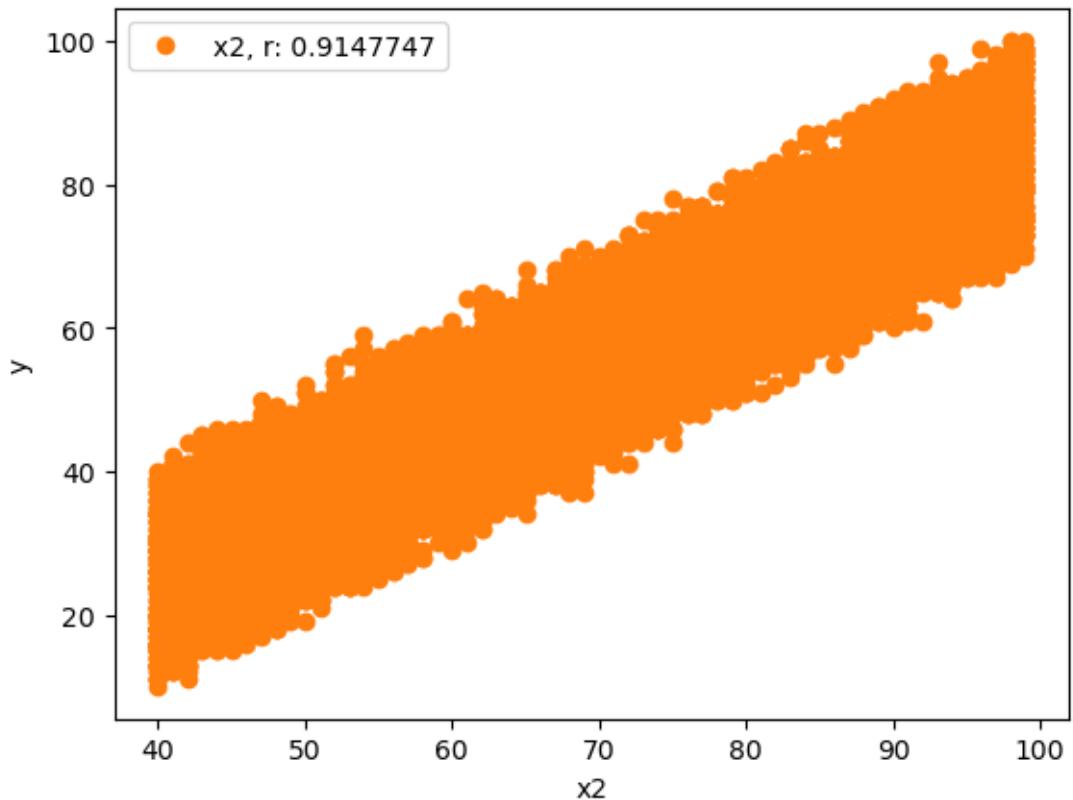


Figure 3 Tương quan giữa điểm các bài kiểm tra trước đó với điểm số hiện tại

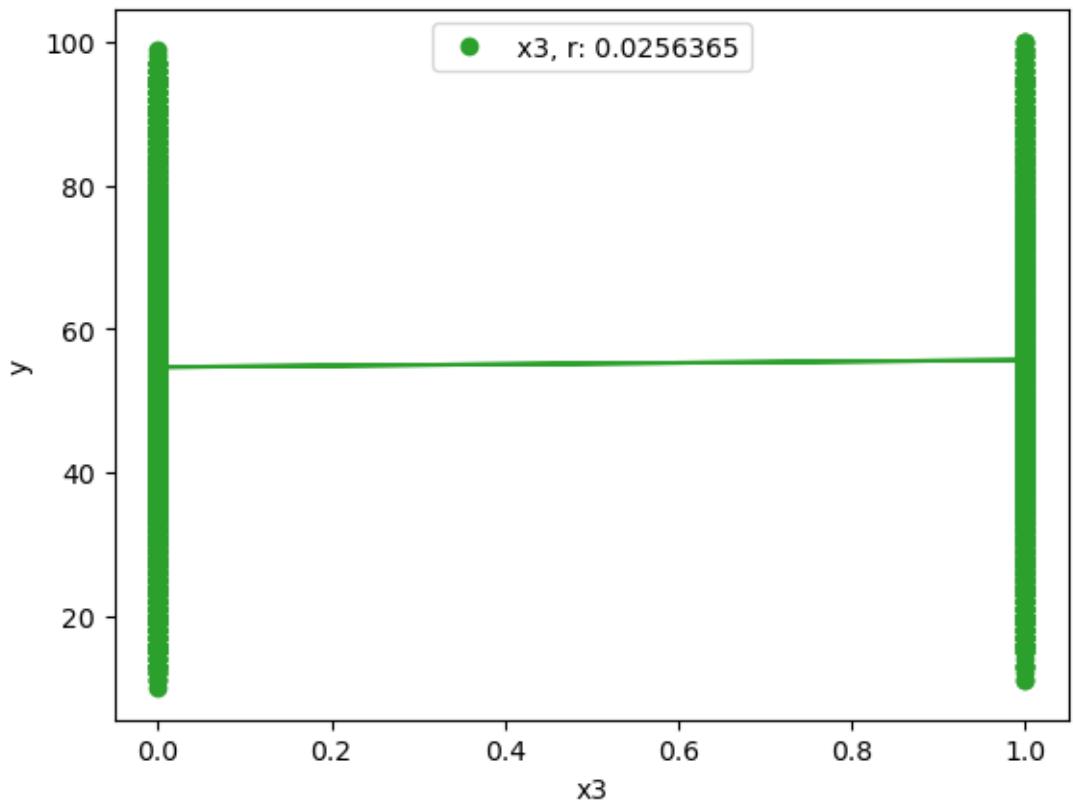


Figure 4 Tương quan giữa việc có hoạt động ngoại khoá với điểm số hiện tại

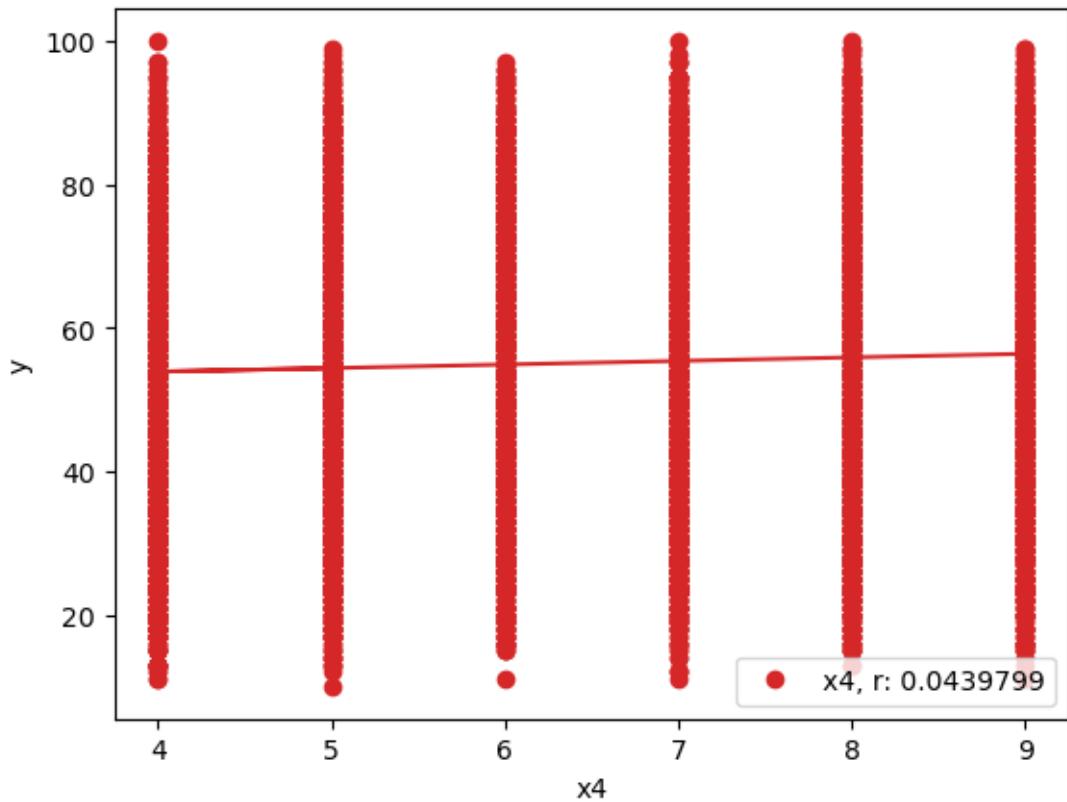


Figure 5 Tương quan giữa số giờ ngủ mỗi ngày với điểm số hiện tại

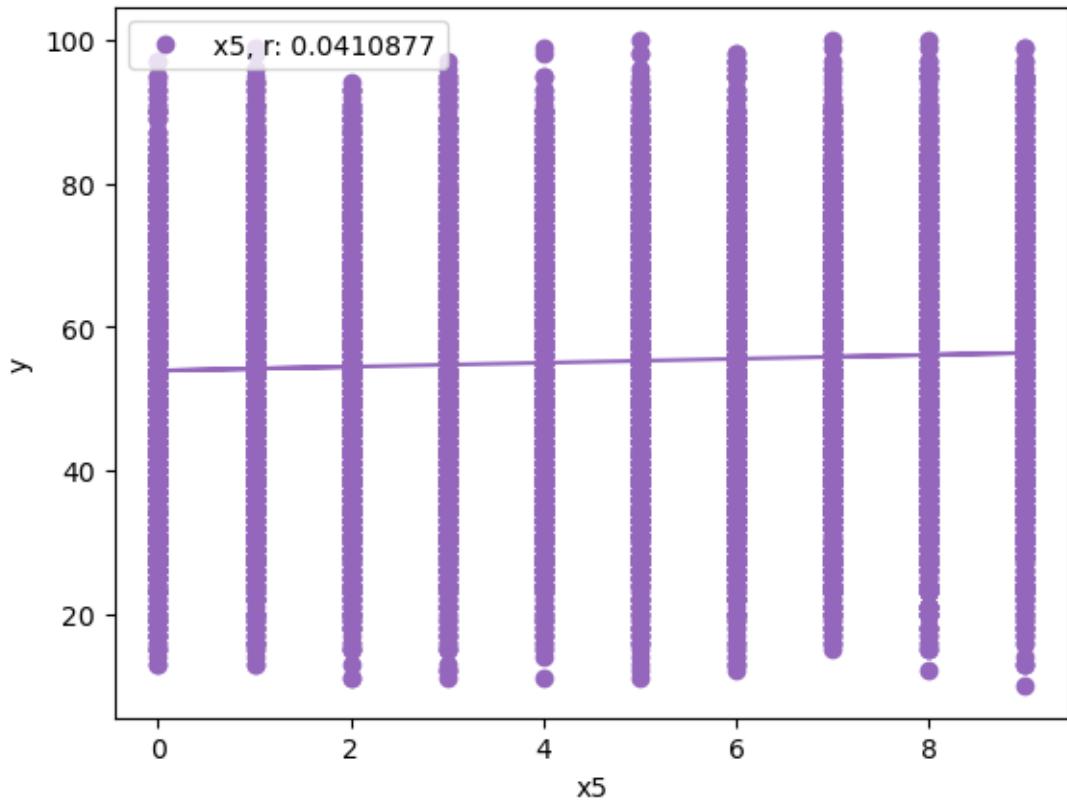


Figure 6 Tương quan giữa số bài kiểm tra mẫu đã luyện tập với điểm số hiện tại

Có thể nhận thấy, ngoài số giờ học mỗi ngày ( $x_1$ ) và số điểm của các bài kiểm tra trước ( $x_2$ ), hệ số tương quan của các biến độc lập  $x_3, x_4, x_5$  tương đối thấp, và với góc nhìn trực quan, ta cũng thấy đường thẳng hồi quy cho các biến độc lập đó tương đối ngang, thể hiện mối tương quan kém. Nói cách khác, ta rất khó có thể sử dụng mối quan hệ giữa  $x_3, x_4, x_5$  để giải thích cho sự thay đổi của  $y$ . Mặt khác, góc nhìn này cũng cho ta gợi ý rằng  $x_1, x_2$  đóng vai trò lớn trong việc liên hệ giữa các yếu tố và điểm số của học sinh;  $x_2$  có thể giải thích  $\sim 90\%$  sự thay đổi của  $y$ , trong khi  $x_1$  có thể giải thích  $\sim 30\%$ .

Ngoài ra, một lưu ý rằng giá trị của các biến ngoài  $x_2$  có giá trị tương đối gần nhau, trong khi đó  $x_2$  lại nằm rải rác về phía bên phải của đồ thị do miền giá trị của nó là điểm số. Khi sử dụng kết hợp nó với các biến khác, ta nên chuẩn hóa giá trị của các biến sử dụng nhằm tránh trường hợp biến số huấn luyện trở nên quá lớn, gây mất cân đối dữ liệu.

### 2.3 Khám phá phân bố của các biến độc lập

Ngoài ra, ta cũng nên thực hiện khảo sát phân phối của các biến độc lập. Một cách thông dụng và khá dễ thực hiện là thông qua biểu đồ histogram. Tại đây, ta khảo sát miền giá trị của các biến độc lập, và chia đều thành các khoảng bins. Với mỗi bins, ta đếm số lượng biến độc lập của đặc tính ứng với giá trị nằm trong khoảng của bins đó, và ta đã có biểu đồ histogram.

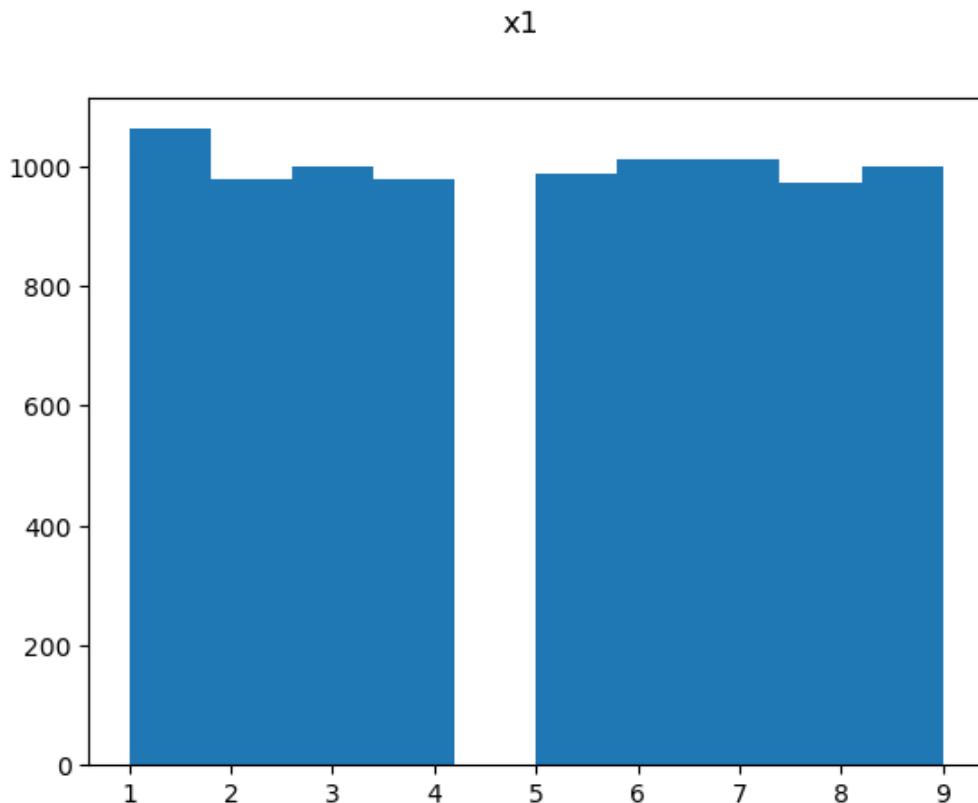


Figure 7 Phân phối của số giờ học mỗi ngày

x2

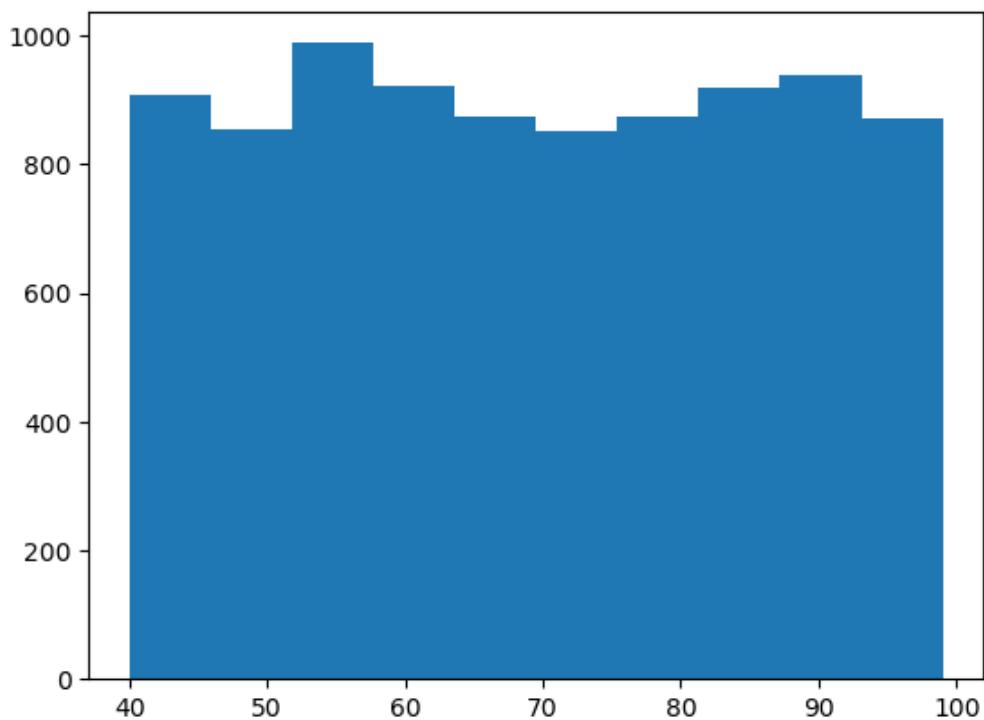


Figure 8 Phân phối của số điểm các bài kiểm tra trước

x3

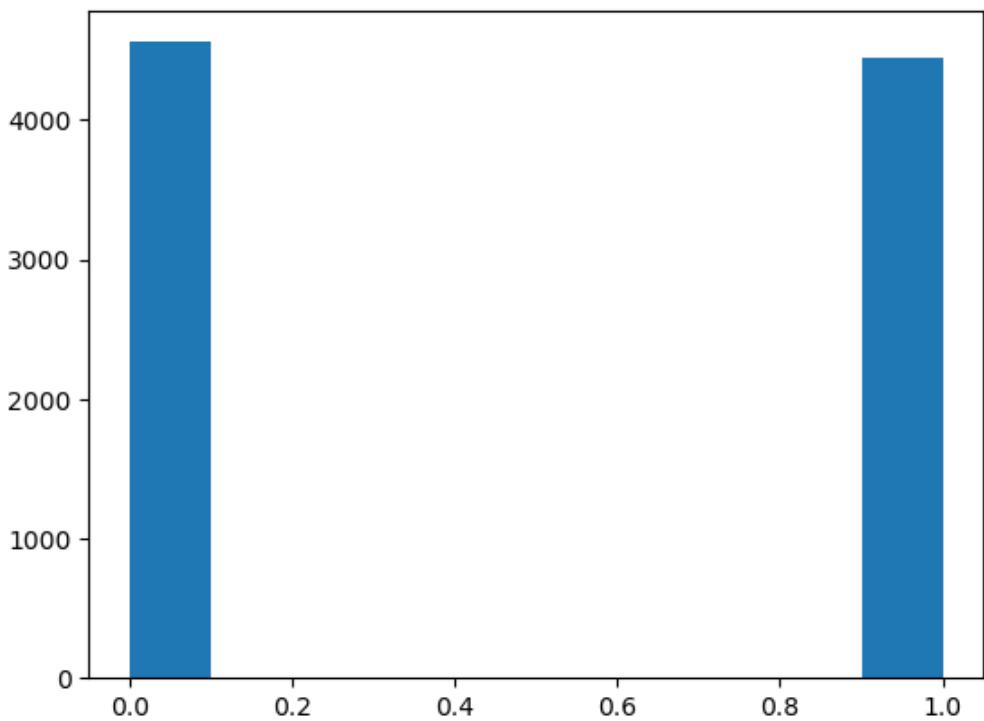


Figure 9 Phân phối của tham gia các hoạt động ngoại khóa

x4

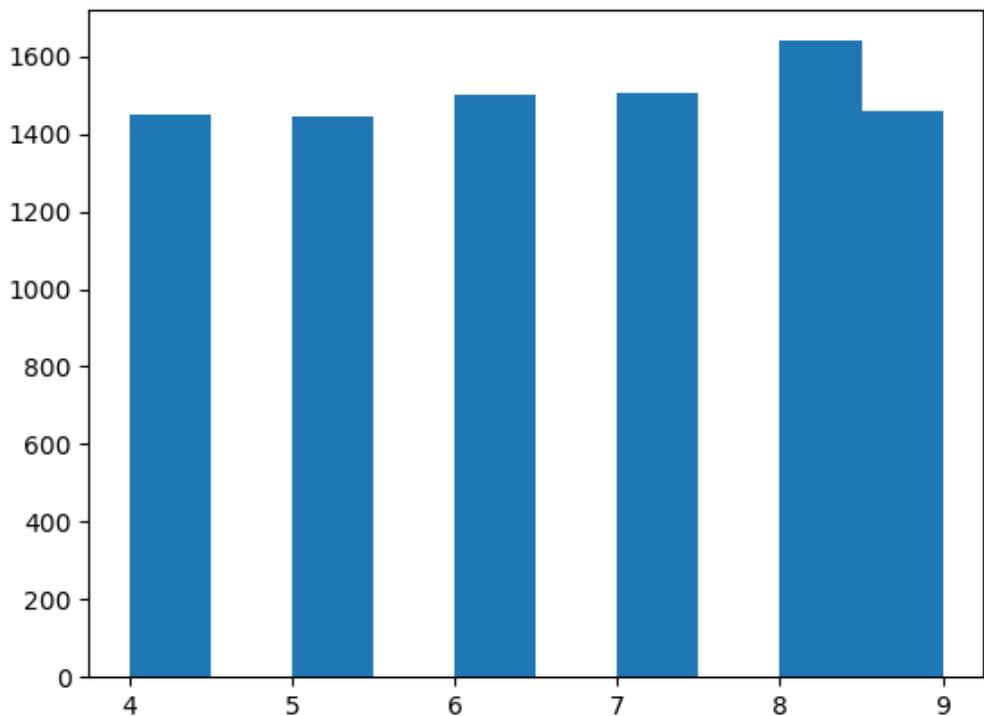


Figure 10 Phân phối của số giờ ngủ mỗi ngày

x5

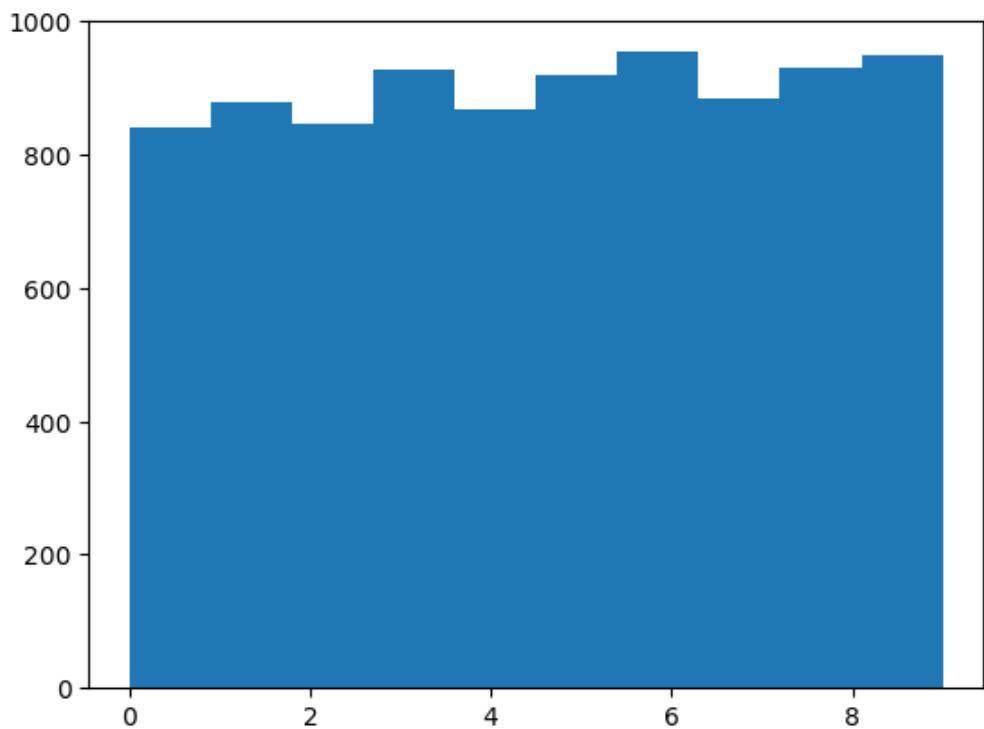


Figure 11 Phân phối của số đđ kiểm tra mẫu học sinh đã luyện tập

Thông qua các biểu đồ histogram, ta nhận thấy phân phối của các biến độc lập tương đối đều, ngoại trừ một số trường hợp:

- Với số giờ học mỗi ngày, ta không ghi nhận được bất kì quan sát nào mà giờ học có giá trị là 4. Ta có giả thuyết rằng các học sinh được chia làm hai nhóm, với nhóm đầu tiên là các bạn học thông thường, với số giờ học  $< 4$ . Nhóm thứ hai bao gồm các bạn học sinh chăm chỉ hơn, với số giờ học  $\geq 5$ . Sự phân bố rõ rệt này được chia ở giá trị 4.
- Với giá trị của tham gia hoạt động ngoại khoá chỉ là “có” hoặc “không”, ta cũng nhận thấy sự phân bố đồng đều giữa hai nhóm tham gia/không tham gia. Tuy nhiên, đối với các bài toán linear regression data fitting, dữ liệu có/không thông thường không đóng góp nhiều vào giá trị dự đoán.
- Tương tự, phân phối giá trị của số giờ ngủ và số bài kiểm tra mẫu đã luyện tập tương đối đều, có xu hướng lệch sang bên trái nhưng không nhiều. Tuy nhiên, ta chưa đủ cơ sở để kết luận hai biến này không có đóng góp trong mô hình.
- Các biểu đồ luôn tồn tại các giá trị mà tại đó có sự khác biệt nhiều đối với các cột xung quanh, ngầm lưu ý rằng dữ liệu sẽ có tồn tại outlier.

## 2.4 Khám phá mối quan hệ giữa các biến độc lập

Ngoài mối quan hệ giữa các biến độc lập với biến phụ thuộc, các mối quan hệ giữa các biến độc lập với nhau cũng đóng vai trò không kém; với các tập dữ liệu có mối liên kết suy diễn lớn giữa một hay nhiều cặp dữ liệu, mô hình huấn luyện có thể gặp phải tình trạng overfitting/underfitting mà các phần dưới sẽ đề cập thêm.

Để có thể chèn hết tất cả biểu đồ vào báo cáo sẽ phải khiến nội dung trở nên rời rạc, vì thế bài thực hành sẽ chỉ chèn cái biểu đồ giữa  $x_1$  với các biến độc lập còn lại, tất cả những nội dung còn lại sẽ được ghi chú thêm tại phần phụ lục.

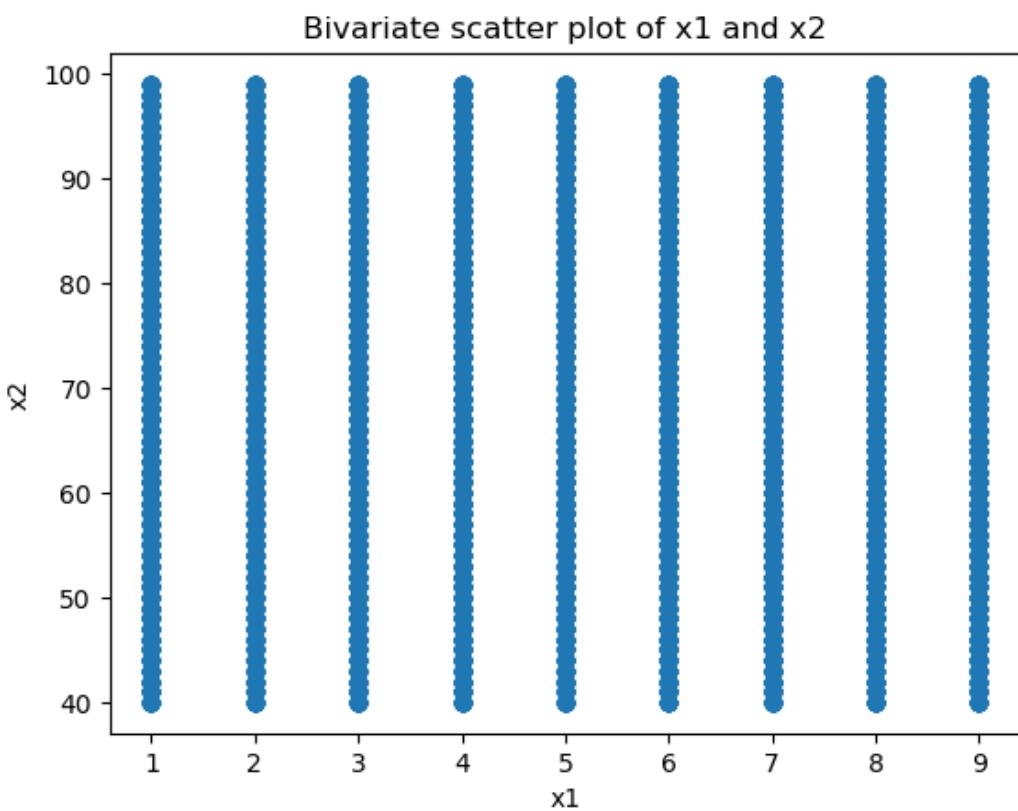


Figure 12 biểu đồ rời rạc giữa số giờ học và số điểm các bài kiểm tra trước

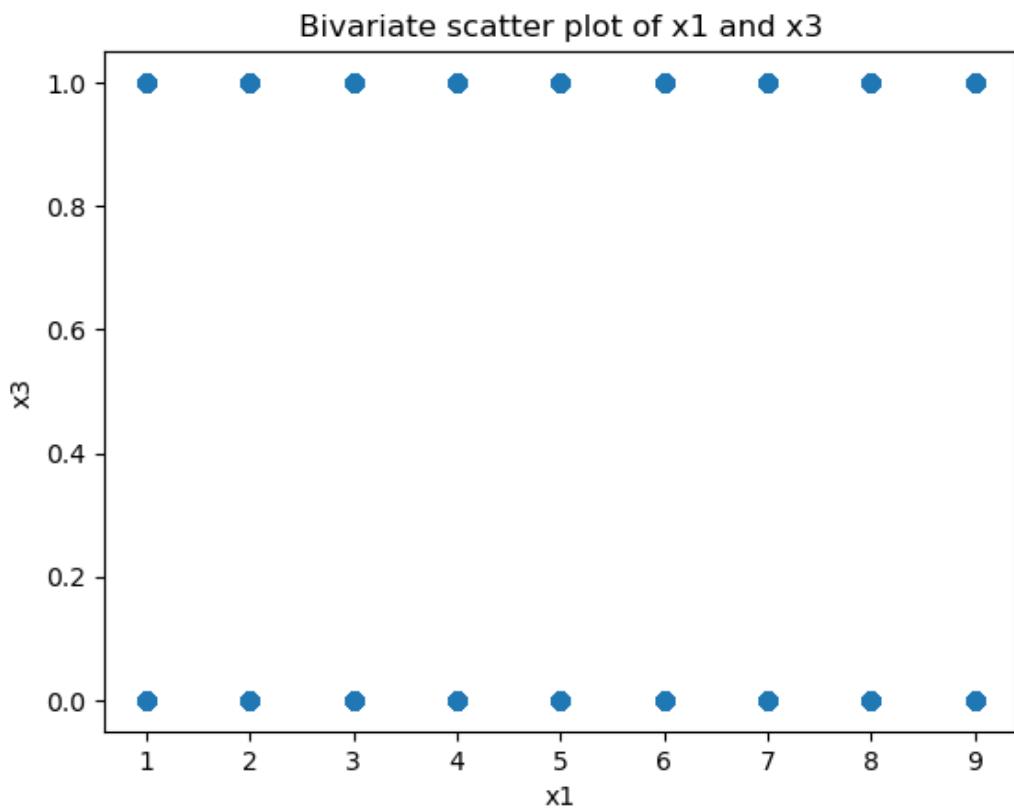


Figure 13 biểu đồ rời rạc giữa số giờ học và tham gia hoạt động ngoại khóa

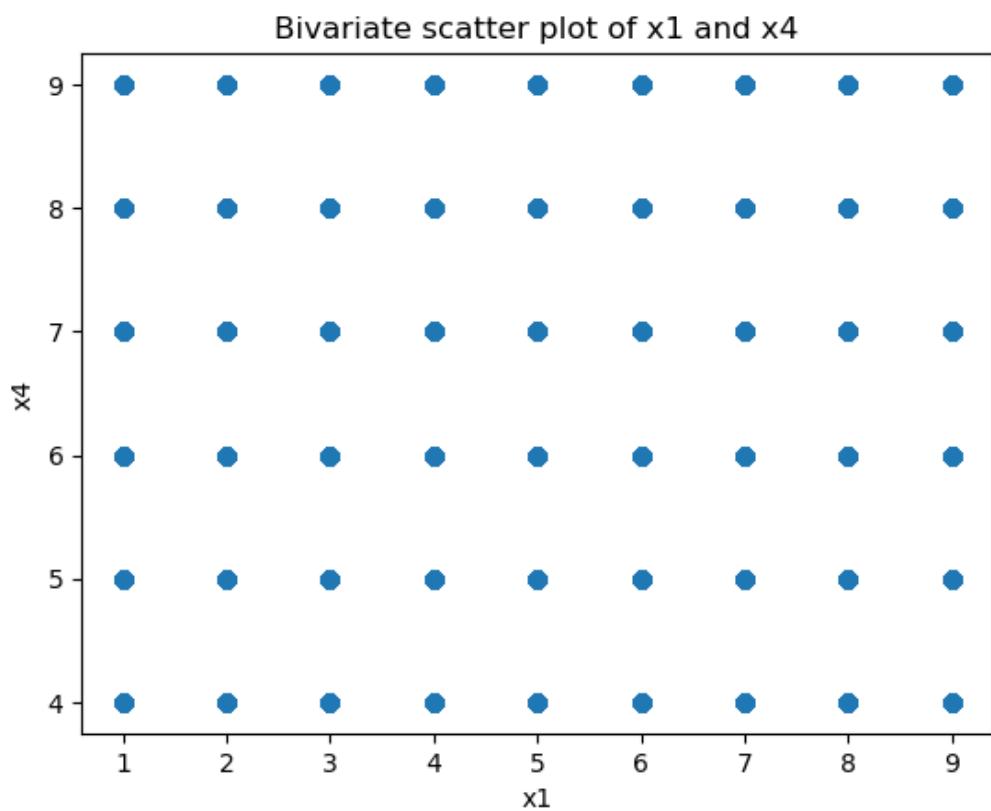


Figure 14 biểu đồ rời rạc số giờ học và số giờ ngủ

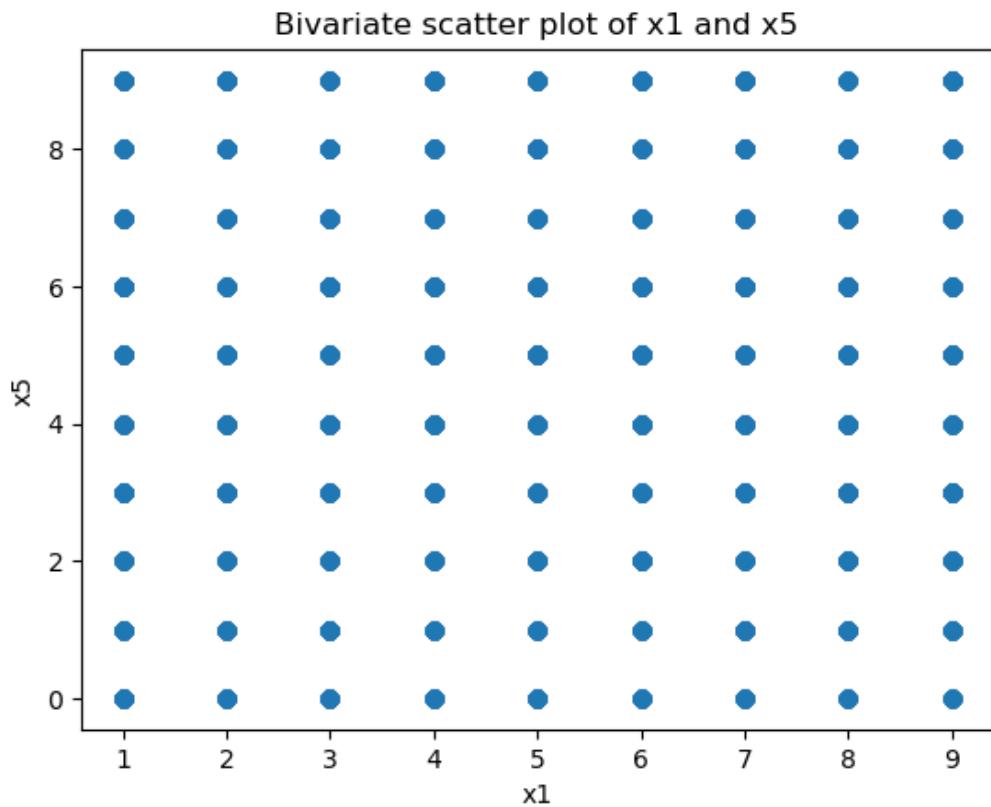


Figure 15 biểu đồ rời rạc số giờ học và số bài kiểm tra mẫu đã luyện tập

Có thể nhận thấy qua các biểu đồ trên, với  $x_1$ , hầu như không có mối quan hệ phụ thuộc tuyến tính giữa các cặp biến độc lập với  $x_1$ . Điều này cũng được thể hiện ở các cặp biến độc lập khác (tham khảo phụ lục). Với quan sát này, ta có thể, một cách gần đúng, kết luận rằng giữa các biến độc lập không có mối quan hệ tuyến tính, tức các cặp biến độc lập không thể suy diễn lẫn nhau. Điều này giúp ta có thể khảo sát mô hình hồi quy tuyến tính một cách tầm thường (trivially), và cũng là một gợi ý rằng các phép biến đổi tuyến tính mang lại mô hình tốt hơn.

Nhưng, có một lưu ý. Nếu ta khảo sát đồ thị 3 chiều giữa  $x_1, x_2, y$ , có một điểm thú vị

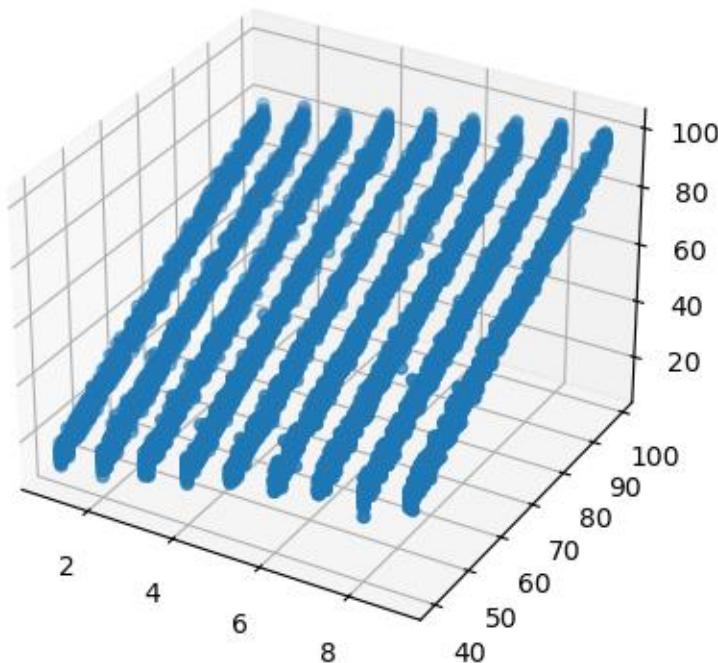


Figure 16 biểu đồ rời rạc số giờ học, số điểm các bài kiểm tra trước và điểm số hiện tại

Mặt phẳng hình thành bởi  $x_1, x_2, y$  trong có vẻ phẳng hơn các cặp quan hệ khác trong dữ liệu, gợi ý đến sự tương quan, không nhiều, nhưng tồn tại, giữa số giờ học và số điểm các bài kiểm tra trước. Mặt khác, các đồ thị giữa các cặp biến độc lập khác và biến phụ thuộc  $y$  hầu như có độ dày khá rõ ràng (noise), không thể dẫn đến kết luận gì.

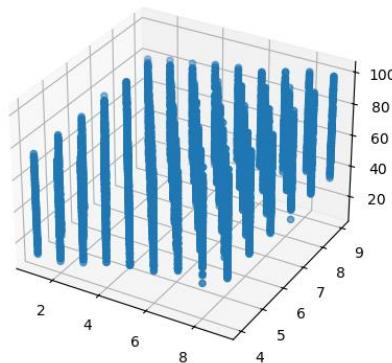


Figure 17 biểu đồ rời rạc số giờ học, số bài kiểm tra mẫu đã luyện tập và số điểm hiện tại

### 3 Cấu trúc cài đặt chương trình, thư viện sử dụng

Các mô hình được cài đặt trong ngôn ngữ Python, sử dụng NumPy quen thuộc làm thư viện xử lý đại số tuyến tính, các thư viện có sẵn của Python như statistic nhằm tính các thống kê của kết quả; tabulate nhằm tạo dựng bảng dữ liệu và hiển thị thông qua Jupyter Notebook IPython display module bằng các lớp Latex, Markdown, kết hợp với matplotlib.

Vì mô hình chính được quan tâm của bài thực hành là mô hình tuyến tính qua phương pháp bình phương nhỏ nhất thông thường, cấu trúc dữ liệu chính của các mô hình là lớp OLSLinearRegression (tham khảo từ các bài lab trước). Lớp chấp nhận dữ liệu thông qua hai hàm chính: fit(X, Y) và predict(X). Tại đây, ta có thể thiết kế mô hình với bất kì số lượng tham số và số lượng điểm dữ liệu, vì nghiệm của bài toán chỉ là phép tính số học. Trong hàm fit(X, Y), dữ liệu biến độc lập sẽ được gán thêm một cột vector  $\vec{1}$  để đơn giản hóa phép tính. Hàm predict(X) sẽ chấp nhận một vecto của một bộ điểm dữ liệu và trả về một số thực vô hướng là giá trị dự đoán của mô hình.

Ngoài ra, với thư viện statistic, bài thực hành sử dụng 2 module chính là NormalDist và variance. NormalDist giúp ta mô hình một phân phối chuẩn, và được dùng nhằm lấy giá trị phân phối tích luỹ của phân phối chuẩn. Module variance giúp đơn giản hóa việc tìm phương sai của một tập các số thực. Module tương đương với

$$Var(U) = \frac{\sum_{u \in U} (u - \bar{U})^2}{|U|}$$

$|U|$  là số phần tử trong  $U$

Các hàm chính cho cài đặt và phân tích cho các mô hình đơn/đa biến đơn giản bao gồm ols\_linregress(X, Y) và ols\_fitness\_stats(Model, X<sub>test</sub>, Y<sub>test</sub>). Như tên gọi của chúng, hàm đầu tiên đơn giản chỉ khai báo một object OLSLinearRegression và huấn luyện theo tập dữ liệu đã cung cấp, và trả về mô hình; hàm tiếp theo thực hiện tính các thống kê thường gặp cho dạng bài toán hồi quy như: sai số tuyệt đối trung bình (MAE) mẫu, sai số mẫu tối đa, sai số mẫu tối thiểu, kì vọng sai số mẫu, phương sai sai số mẫu.

Các hàm của chương trình thường nhận tập dữ liệu có dạng ma trận  $N$  dòng dữ liệu và  $n$  cột ứng với giá trị sau khi qua các hàm sơ cấp của các toán hạng trong đẳng thức tổng quát của mô hình. Vì thế, nếu  $X$  nhận vào có dạng (N,), chương trình thường sẽ broadcast  $X$  lên dạng ma trận (N, 1) hoặc (N, 2) khi huấn luyện/dự đoán.

### 4 Cài đặt mô hình hồi quy tuyến tính đa biến bậc nhất

#### 4.1 Chi tiết lựa chọn và cài đặt mô hình

Sau khi đã có cái nhìn tổng quan về dữ liệu mà ta sẽ làm việc, việc tìm và chọn các đặc trưng cũng như kiến trúc của mô hình là bước tiếp theo nhằm khẳng định các giả thuyết của bài toán.

Cũng như bao bài toán hồi quy tuyến tính khác, ta có ghi nhận sự độc lập giữa các biến đầu vào, cũng như có chứng cứ về tính tuyến tính của một số đặc tính với điểm số đầu ra hiện tại. Vì

thể phần này sẽ khảo sát mô hình tuyến tính đa biến bậc nhất cho tập dữ liệu Student Performance. Mô hình có dạng:

$$\hat{f}(x) = \widehat{\theta_1}x_1 + \widehat{\theta_2}x_2 + \widehat{\theta_3}x_3 + \widehat{\theta_4}x_4 + \widehat{\theta_5}x_5 + \epsilon$$

## 4.2 Thủ nghiệm và kết quả của mô hình

Sau huấn luyện bằng đăng thức (eq1), ta thu được mô hình:

$$\hat{y} = 2.852x_1 + 1.018x_2 + 0.604x_3 + 0.474x_4 + 0.192x_5 - 33.969$$

Sau khi đã có mô hình, ta mong muốn xác định độ chính xác của nó khi được đưa một tập dữ liệu chưa được gắp bao giờ. Nhằm giữ tính đơn giản, bài thực sẽ chỉ khảo sát giá trị *MAE* theo [7] [8] của các mô hình và đưa ra kết luận phù hợp.

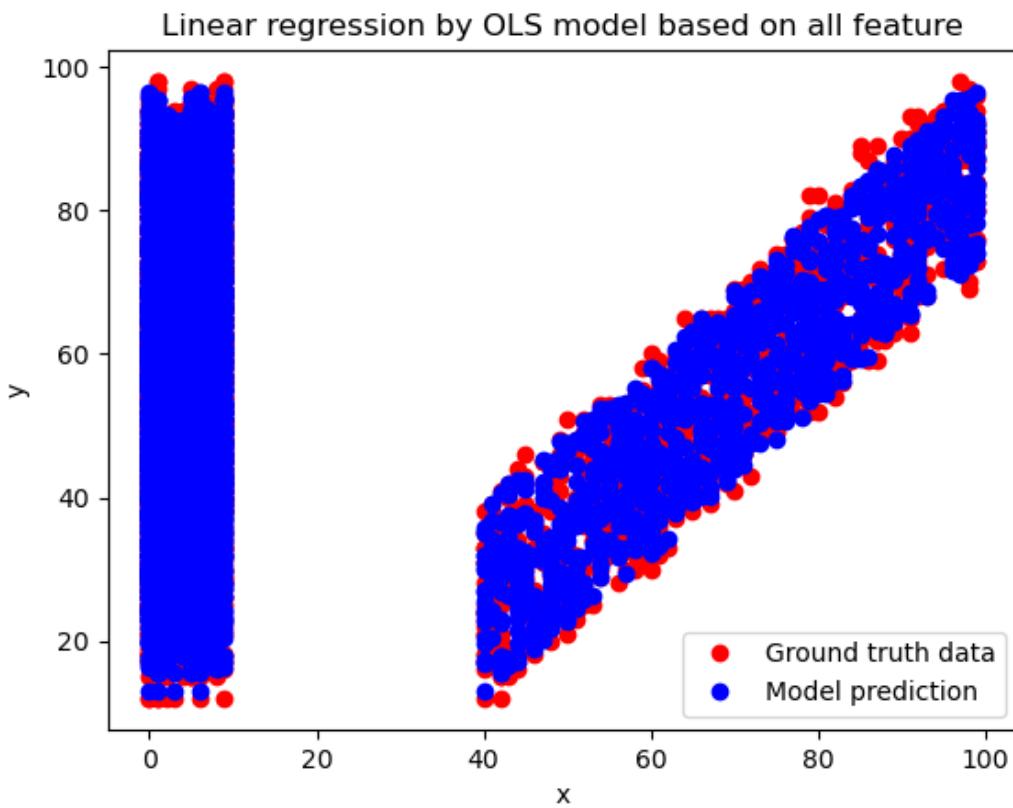


Figure 18 biểu đồ rời rạc các giá trị thực tế của tập dữ liệu kiểm tra và giá trị dự đoán tương ứng của mô hình.

Qua khảo sát trực quang, ta nhận thấy các điểm dự đoán có vẻ nằm khá gọn trong miền xác định của dữ liệu kiểm tra. Ngoài ra, qua hàm thống kê `ols_fitness_stats`, ta cũng nhận được giá trị  $MAE \sim 1.596$ , sai số tối đa  $\sim 7.084$  và kì vọng sai số rơi vào rất gần với 0, và phương sai của dữ liệu chỉ khoảng 4.093. Các thống kê cho thấy một cách tổng quan, rằng mô hình tuyến tính đa biến bậc nhất dù có vẻ đơn giản, nhưng mang lại tính tổng quát cao với giá trị dự đoán.

Mặt khác, như đã đề cập, dữ liệu có tồn tại outlier, khiến mô hình thể hiện không quá tốt do tính tuyến tính trên biến độc lập. tại vùng xác định của  $x_2$  phía bên phải, ta nhận thấy tại rìa trên

và dưới của vùng đó khó có thể được ước lượng thông qua mô hình. Tương tự, rìa trên và dưới của vùng bên trái cũng tồn tại, và là các nhân tố chính ảnh hưởng đến  $MAE$  của mô hình.

Ở cùng góc nhìn, ta cũng nhận thấy mặc dù đặc tính số điểm các bài kiểm tra trước có tính tuyến tính mạnh với số điểm hiện tại, mô hình vẫn chịu ảnh hưởng bởi nhiều cũng như là thiếu tính tổng quát khi xét về siêu mặt phẳng xác định của toàn bộ mô hình. Quá trình tìm hiểu giải pháp cho vấn đề này sẽ được đề cập đến trong mục 6.

## 5 Cài đặt mô hình hồi quy tuyến tính đơn biến bật nhất cho từng đặc trưng

### 5.1 Chi tiết lựa chọn và cài đặt

Để khảo sát đơn tính của từng đặc trưng và mối tương quan giữa chúng với biến phụ thuộc đang xét, ta thực hiện phương pháp khảo sát giá trị hồi quy tương quan giữa từng đặc trưng và số điểm hiện tại của học sinh.

Ta định nghĩa hàm  $k\_fold(X, Y, k)$  nhận vào tập dữ liệu huấn luyện mô hình và một số nguyên  $k$  và trả về một bộ gồm mô hình huấn luyện qua phương pháp  $OLS + k - fold cross validation$  và trung bình qua  $k$  bước huấn luyện.

Hiểu một cách đơn giản, khi huấn luyện mô hình, ta muốn xác định được khả năng dự đoán các dữ liệu chưa từng được thấy trước đó của mô hình. Vào một số trường hợp, nếu mô hình của ta quá phức tạp hoặc có mối quan hệ phụ thuộc nào đó giữa các biến được cho là độc lập, thì tình trạng *overfitting/underfitting* mà tại đó, một là mô hình quá khớp với dữ liệu huấn luyện, khiến việc dự đoán các giá trị ngoài tập huấn luyện trở nên mất chính xác, ta nói nó bị *overfitting*; hoặc khi mô hình quá đơn giản, không thể tổng quát hoá, ta nói nó bị *underfitting*. Tại phần này, ta chỉ quan tâm đến vấn đề *overfitting* khi mô hình chỉ có tính tuyến tính đáng chú ý với 2 đặc trưng.

Nhằm tránh việc “học tủ” [9] như vậy, trong trường hợp ta rất hạn chế về số lượng dữ liệu để xây dựng mô hình, phương pháp *cross validation* là một khởi đầu tốt để ta đánh giá hiệu năng và tránh bias của các mô hình. Trích một đoạn của Tiến Sĩ Tiệp:

“*Cross validation* là một cải tiến của *validation* với lượng dữ liệu trong tập validation là nhỏ nhưng chất lượng mô hình được đánh giá trên nhiều tập *validation* khác nhau. Một cách thường đường sử dụng là chia tập training ra  $k$  tập con không có phần tử chung, có kích thước gần bằng nhau. Tại mỗi lần kiểm thử, được gọi là *run*, một trong số  $k$  tập con được lấy ra làm *validate set*. Mô hình sẽ được xây dựng dựa vào hợp của  $k-1$  tập con còn lại. Mô hình cuối được xác định dựa trên trung bình của các *train error* và *validation error*. Cách làm này còn có tên gọi là ***k-fold cross validation***.” [9] Ts. Vũ Hữu Tiệp.

Trong cài đặt thực tế của hàm  $k\_fold(X, Y, k)$ , ta thực hiện các bước sau [10]:

1. Hoán vị bộ dữ liệu nhằm tránh trường hợp các điểm dữ liệu huấn luyện được chia không theo một tiến trình ngẫu nhiên.
2. Bộ dữ liệu huấn luyện được chia thành  $k$  tập con có số lượng điểm dữ liệu (gần) bằng nhau.

3. Một mô hình được huấn luyện trên  $k - 1$  tập con này và được kiểm thử trên tập con còn lại.
4. Lặp lại bước 3 k lần. Với mỗi lần lặp, ta chọn một tập con khác làm tập kiểm thử.
5. Lấy trung bình trên tham số giữa các mô hình trong k lần chạy để có mô hình đầu ra, ngoài ra cũng lấy trung bình sai số tuyệt đối trung bình ( $MAE$ ) của các mô hình. Sai số này được gọi là **sai số kiểm định chéo**.
6. Trả về mô hình trung bình và sai số kiểm định chéo.

Với giá trị sai số tuyệt đối trung bình, ta có thể suy diễn một cách đơn giản khả năng thể hiện trên tập dữ liệu kiểm thử của mỗi mô hình.

## 5.2 Thủ nghiệm và các kết quả của các mô hình

Sau khi đã huấn luyện theo  $k$ -fold với  $k = 10$ , ta có được các mô hình và  $MAE$  trung bình tương ứng:

Index	Feature Description	Model	MAE
1	Hours Studied	$\hat{y} = 2.730x_1 + 41.551$	15.449
2	Previous Scores	$\hat{y} = 1.011x_2 - 14.989$	6.618
3	Extracurricular Activities Participation	$\hat{y} = 0.984x_3 + 54.651$	16.194
4	Sleep Hours	$\hat{y} = 0.498x_4 + 51.884$	16.187
5	Sample Question Papers Practiced	$\hat{y} = 0.275x_5 + 53.873$	16.184

Figure 19 kết quả huấn luyện 5 mô hình tuyến tính đơn đặc trưng bậc nhất

Đúng theo khảo sát khi khám phá dữ liệu của ta, tính tuyến tính của số điểm các bài kiểm tra trước có mối liên hệ tương đối chặt chẽ với số điểm hiện tại. Các đặc trưng khác có  $MAE$  dao động từ  $\sim 15.5 - \sim 16.2$ , với đặc trưng số giờ họ thấp hơn đôi chút so với các đặc trưng khác. Qua kết quả này, ta có thể, một cách tương đối, kết luận rằng mô hình đơn biến bậc nhất cho bài toán hồi quy tuyến tính của tập dữ liệu này thể hiện rõ mối quan hệ tương quan nhất thông qua đặc trưng  $x_2$ , số điểm các kì kiểm tra trước.

Ngoài ra, ta cũng thực hiện kiểm định giả thuyết về độ dốc của tổng thể mô hình lý tưởng  $\beta_1$  thông qua độ dốc của mẫu  $\widehat{\theta}_1$  bằng hàm `hypothesis_test_feature_contribution(X, Y, Xtest, Ytest, alpha=0.05)` mà chi tiết cài đặt sẽ được đề cập phía dưới. Thông qua kiểm định, ta thu được p-giá trị tiệm cận 0, cung cấp cho giả thuyết rằng  $x_2$  có mối quan hệ tuyến tính mạnh với  $y$ .

Sau khi đã chọn được đặc trưng tốt nhất với mô hình tuyến tính đơn giản, ta huấn luyện lại nó thông qua hàm `ols_linregress(X, Y)` và chỉ lấy cộng giá trị tương ứng với  $x_2$  trong ma trận biến độc lập. Đáng chú ý rằng có vẻ phương pháp trung bình cộng các tham số trong quá trình  $k$ -fold có vẻ đưa ra mô hình có tham số tương đương với việc huấn luyện trên cả tập dữ liệu.

$$\hat{y} = 1.011x_2 - 14.989$$

Như quy trình phía trên, sau khi đã có mô hình, ta thực hiện kiểm tra nó với tập dữ liệu kiểm tra và nhận xét dựa trên các thống kê.

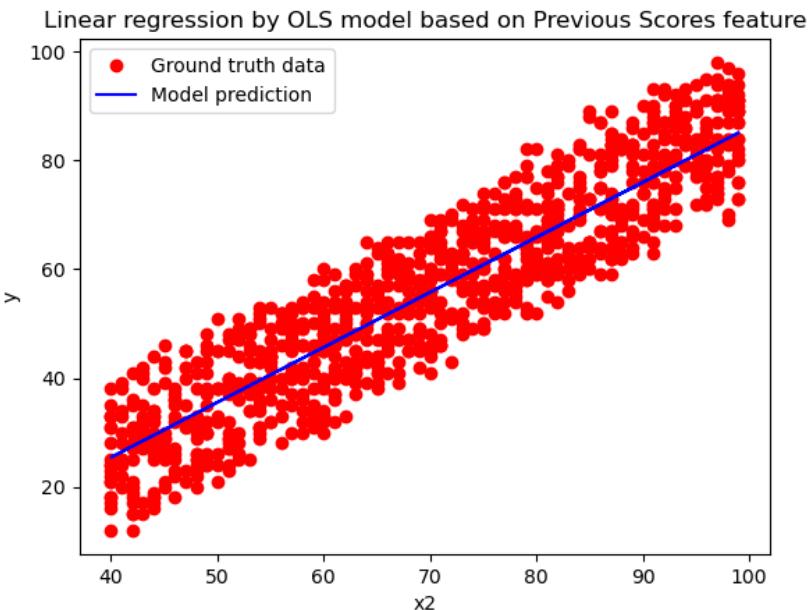


Figure 20 biểu đồ các giá trị thực tế của tập dữ liệu kiểm tra và đường thẳng dự đoán tương ứng của mô hình.

Khảo sát trực quang, ta thấy các điểm dữ liệu nằm gần đường thẳng dự đoán của mô hình, với vị trí của đường thẳng đi qua giữa các điểm dữ liệu. Vì là mô hình tuyến tính bậc nhất, mô hình chỉ có thể chọn một đường thẳng đi qua sao cho sai số là nhỏ nhất. Vì thế, kết hợp với các thông kê  $MAE = 6.544$ ,  $\max(r) = 18.096$ ,  $\min(r) = 0.009$ ,  $\mu_r = 0.384$ ,  $Var(r) = 58.8$ , mô hình có thể được kết luận là chỉ giải thích được xu hướng tương quan giữa hai biến dữ liệu, nhưng không thể dự đoán gần đúng điểm số trong bài kiểm tra của học sinh.

## 6 Cài đặt mô hình hồi quy tuyến tính với các hàm cơ sở khác

Theo chuẩn cấu trúc của bài thực hành, phần này sẽ đi vào một số mô hình được xây dựng trên cơ sở giả định liên hệ giữa các biến độc lập.

### 6.1 Chi tiết lựa chọn và cài đặt

Như đã có nhắc đến ở phía trên, nhằm khám phá một cách có logic các giả thuyết về liên hệ giữa các đặc trưng theo định nghĩa nghiệp vụ của lĩnh vực (ở đây là giáo dục), ta mong muốn có cơ sở để tập trung xây dựng các mô hình xung quanh một hay nhiều đặc trưng mang tính chủ đạo. Đầu tiên, ta có hàm `hypothesis_test_feature_contribution(X, Y, Xtest, Ytest, alpha)` nhận vào tập huấn luyện và tập kiểm định nào đó. Cùng với đó, hàm nhận vào một giá trị  $\alpha$  là giá trị của độ tin cậy khi ta kiểm định giả thuyết

$$\begin{cases} H_0: \theta_1 = 0 \\ H_1: \theta_1 \neq 0 \end{cases}$$

Để tiếp tục, ta giả định rằng với lượng mẫu lớn,  $\theta_1$  sẽ hội tụ về phân phối chuẩn theo định lý giới hạn trung tâm. Vì vậy, ta cần phẩm tìm độ lệch chuẩn của  $\theta_1$ , may mắn rằng nó đã được nghiên cứu:

$$\hat{\sigma}_{\beta_1} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum(x_i - \bar{x})^2}}$$

(Wikipedia)

Và từ đây, ta có thể thực hiện tìm thông kê kiểm định và p-giá trị.

Thông qua quy trình kiểm định tương tự, bài thực xác định được hai biến  $x_1, x_2$  tương ứng với số giờ học và số điểm các bài kiểm tra trước đóng vai trò chủ chốt trong việc dự đoán điểm số. Tuy nhiên, việc kiểm định không hoàn toàn bác bỏ liên hệ giữa các biến còn lại với biến phụ thuộc, vì thế, ta vẫn sẽ sử dụng chúng trong mô hình nhằm khai thác tối đa dữ liệu.

Sau khi đã có các thông tin cần thiết, ta tiến hành lựa chọn và thử nghiệm một số mô hình. Đầu tiên, để thuận tiện cho việc lập trình và mở rộng sau này, các hàm liên quan đến các mô hình sẽ được đặt tên theo dạng `model_xx(X, Y)` với `xx` là chỉ mục của mô hình; hàm nhận vào các tập dữ liệu huấn luyện và trả về một `OLSLinearRegression` object. Ngoài ra, còn có các hàm dạng `model_xx_expression(X)` nhận vào một tập đầy đủ các dữ liệu huấn luyện ngoại trừ biến phụ thuộc, và trả về ma trận đã biến đổi theo hàm số xác định mô hình.

Vì tính tuyến tính của cả  $x_1, x_2$  với  $y$ , sử dụng riêng hai biến này trên lý thuyết có thể hoạt động tốt hơn nhờ việc không bị nhiễu loạn bởi đầu vào của các biến độc lập khác. Mô hình đầu tiên của ta được xây dựng dựa trên suy luận này và có dạng

$$\hat{y} = \theta_1 x_1 + \theta_2 x_2 + \theta_3$$

Trước khi đến với mô hình thứ hai, ta quay trở lại với giả định rằng  $x_1$  và  $x_2$  có mối quan hệ tuyến tính với nhau và với  $y$ . Đối với các mô hình có mối quan hệ nhị phân giữa các biến được cho là độc lập, việc huấn luyện bị nhiễu bởi các ảnh hưởng của từng đặc trưng sẽ trở nên rõ ràng đối với lượng dữ liệu huấn luyện lớn, cuối cùng dẫn đến tình trạng overfitting hoặc underfitting. Vì thế, để tránh tình trạng này xảy ra, mô hình hai sẽ áp dụng một phương pháp ổn định theo chuẩn bình phương  $L_2$  (Ridge Regression).

Để không đi sâu vào lý thuyết, phương pháp  $L_2$  đơn giản chỉ thêm một số hạng theo dạng ma trận vào hàm mất mát (eq1) nhằm đánh giá độ phức tạp của mô hình [9]. Theo đó, hàm mất mát sẽ có dạng:

$$\mathcal{L}(\theta) = \|X\theta - y\|_2^2 + \lambda I$$

Trong đó,  $\lambda$  là ước số khuynh hướng và  $I$  là ma trận đơn vị  $\in \mathbb{R}^p$ . Từ đó, nghiệm của bài toán có thể được viết lại với dạng

$$\theta^* = \underset{\theta \in \mathbb{R}^p}{\mathbf{argmin}} (\|\mathcal{L}(\theta)\|) = (X^T X + \lambda I)^{-1} X^T y \quad (\text{eq2})$$

Trên thực tế,  $\lambda$  có thể áp dụng vào mô hình thông thường sẽ được tìm qua quá trình thử nghiệm chọn ngẫu nhiên (*ad hoc*) hoặc tối ưu theo [11]. Tuy nhiên, bài thực hành thông qua quá trình trial and error, đã lựa chọn  $\lambda = 0.01$  để mang lại hiệu quả huấn luyện tốt nhất.

Vậy, mô hình thứ 2 sẽ có dạng

$$\hat{y} = 2\theta_1 \sqrt{x_1 x_2 + \frac{3}{8}} + \theta_2 x_2 + \theta_3 x_3 + \theta_4 \frac{x_4}{x_1} + \theta_5 x_5 + \theta_6$$

Trong đây, toán hạng thứ nhất có dạng  $\sqrt{x_1 x_2 + \frac{3}{8}}$  là một biến đổi Anscombe; về cơ bản, nó có thể được dùng nhằm ổn định phương sai khi ta giả định một biến ngẫu nhiên trong đó có phân phối Poisson [12]; và vì thế, kì vọng  $\mu$  và phương sai  $\nu$  của biến có thể phụ thuộc lẫn nhau. Phép biến đổi nhằm đến việc chuyển đổi dữ liệu có phân phối Poisson sang sắp xỉ phân phối Gauss (chuẩn) và càng chính xác hơn với  $\mu$  càng lớn. Vì  $x_1$  là trung bình số giờ học mỗi ngày, có cơ sở để ta thử nghiệm một mô hình với giả định trên. Ngoài ra, ở toán hạng thứ 4 có dạng  $\frac{x_4}{x_1}$  vì ta cũng mong muốn khảo sát xem liệu tỉ lệ cân đối hoặc không cân đối giữa thời gian ngủ và thời gian học có mối quan hệ gì với điểm số không. Các biến còn lại được giữ dạng tuyến tính bậc nhất để ta dễ dàng xem xét những thay đổi trên.

Với mô hình thứ ba, ta sẽ không còn sử dụng phương pháp ổn định  $L_2$ . Mặt khác, ta muốn xem xét mức độ quan hệ giữa đại lượng của  $x_1$  và  $x_2$  sau khi chuẩn hoá, có thể giải quyết thiếu sót của mô hình hồi quy tuyến tính đơn biến bậc nhất qua đặc trưng  $x_2$  hay không. Để như vậy, ta phải chuẩn hoá giá trị của dữ liệu hai biến thông qua:

$$Norm(u)_U = \frac{u - \bar{U}}{\sigma_U}, \quad \sigma_U = \sqrt{\frac{\sum_{u \in U} (u - \bar{U})^2}{|U|}}$$

Với hy vọng việc chuẩn hoá sẽ giúp làm rõ hơn về tương quan giữa hai biến thông qua việc “chấp vá” các lỗ hỏng của biến độc lập mới  $x_1 \times x_2$  ở (Figure 24), mô hình 3 có dạng

$$\hat{y} = \theta_1 x_1 + \theta_2 x_2 + \theta_3 \frac{Norm(x_1)x_1}{\sigma_{x_1}} Norm(x_2)x_2 + \theta_4 x_3 + \theta_5 \frac{x_1}{x_4} + \theta_6 x_5 + \theta_7$$

Lưu ý rằng, ở toán hạng thứ 5, ta có tỉ số  $\frac{x_1}{x_4}$  khác với  $\frac{x_4}{x_1}$  của mô hình 2, do tại đây, ta muốn khám phá liệu tỉ lệ giữa thời gian học và thời gian ngủ có thể ảnh hưởng khác không. Với 3 mô hình như thế, ta tiến hành:

## 6.2 Thủ nghiệm và các kết quả của các mô hình

Tương tự như các mô hình trước, ta thực hiện huấn luyện mô hình thông qua lớp `OLSLinearRegression`. Tại mô hình 2, trước khi fit data, ta cần phải điều chỉnh cho mô hình sử dụng phương trình nghiệm theo (eq2). Bằng cách đặt cho biến `ridge_lambda` là thuộc tính của `OLSLinearRegression`, hàm fit data sẽ tự động chọn phương pháp regularization  $L_2$ .

Sau khi thực hiện huấn luyện theo các bước tương tự như *k-fold* tại **5.1**, ta thu được các mô hình:

Index	Model	MAE
1	$\hat{y} = 2.856x_1 + 0.804x_2 - 29.747$	1.816
2	$\hat{y} = 2 \times 0.820 \sqrt{x_1 x_2 + \frac{3}{8}} + 0.804x_3 + 0.570x_3 + 0.897 \frac{x_4}{x_1} + 0.191x_5 - 32.651$	1.98
3	$\begin{aligned} \hat{y} = & 3.291x_1 + 1.018x_2 - 0.017 \left( \frac{Norm(x_1)}{\sigma_{Norm(x_1)}} \right) Norm(x_2) + 0.595x_3 \\ & + 2.676 \left( \frac{x_1}{x_4} \right) + 0.194x_5 - 33.563 \end{aligned}$	1.649

Figure 21 kết quả huấn luyện 3 mô hình các hàm cơ sở khác nhau

Dựa vào bảng thống kê, ta có thể nhận thấy MAE của các mô hình tương đối thấp, dao động trong khoảng  $\sim 1.65 - \sim 1.98$ . Ta đồng thời thực hiện tính các thống kê liên quan khác, nhận thấy các mô hình có trung bình sai số khá gần với 0, và phương sai của sai số dao động từ  $\sim 4.211 - \sim 5.390$ .

Có thể dự đoán từ biểu đồ rời rạc giữa  $x_1, x_2, y$ , mô hình đầu tiên chỉ sử dụng hai biến độc lập nhưng đã giảm đáng kể MAE nhờ vào khả năng giải thích của cả hai biến tương đối đáng kể so với  $y$ . Điều này một phần có thể giải thích bởi biểu đồ

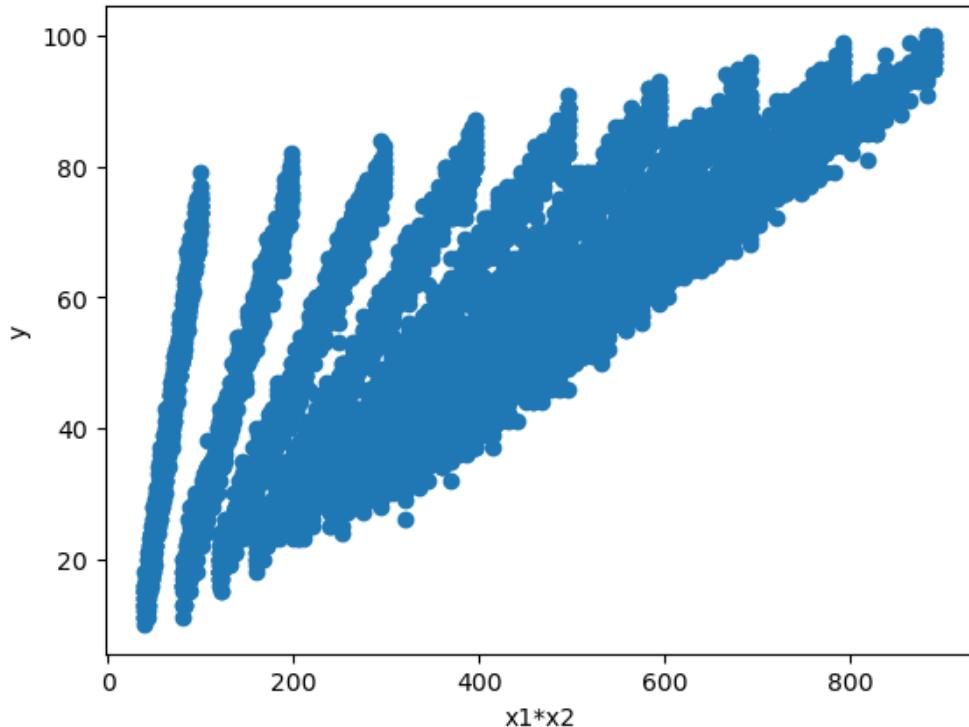


Figure 22 biểu đồ thể hiện mối quan hệ giữa  $x1 * x2$  và  $y$

Vì  $x_1 x_2$  có đơn vị đại lượng khác nhau, nên trục giá trị của đồ thị không thật sự rõ ràng và có nhiều khoảng trống giữa diện tích mặt phẳng đồ thị. Đó cũng là lý do cho phép biến đổi tại toán hạng đầu tiên của mô hình 2.

Ngoài ra, có điều thú vị rằng, dù mô hình 2 có vẻ phức tạp hơn nhiều và dựa vào đặc tính nghiệp vụ của các biến độc lập, mô hình 2 nên có khả năng dự đoán tốt hơn so với mô hình 1. Tuy nhiên, mô hình 2 lại thể hiện kém hơn khi được đưa vào dự đoán tập kiểm định. Một giả thuyết cho vấn đề này có thể kể đến như biến độc lập chỉ các bài kiểm tra mẫu đã luyện tập của học sinh không thật sự hiệu quả do tính chất, dạng đề của các bài kiểm tra không liên quan đến các bài mẫu đã ôn; ngoài ra, ta cũng không nói được gì nhiều về tỉ lệ giữa việc ngủ ít/ngủ nhiều, do có thể văn hóa học và sinh hoạt của mẫu được khảo sát không thật sự có khác biệt nhiều, tức có sự phân bố tập trung về tỉ lệ gần 1 của 2 biến độc lập.

Với mô hình 3, ta nhận được kết quả tốt nhất với cả  $MAE$  và phương sai của sai số thấp nhất trong cả 3 mô hình; ngoài ra, sai số tối đa và sai số tối thiểu cũng nằm trong khoảng tương đối thấp, gần với mô hình tuyến tính ở mục 4,  $Var(r_{M3}) \approx 4.211$ . Vì thế, ta có cơ sở để chọn mô hình này vì khả năng giải thích các dữ liệu chưa gặp tốt nhất của nó.

Sau khi đã chọn được mô hình phù hợp nhất, ta thực hiện huấn luyện lại với toàn bộ bộ dữ liệu huấn luyện, thu được:

$$\hat{y} = 3.291x_1 + 1.018x_2 - 0.017 \left( \frac{Norm(x_1)}{\sigma_{Norm(x_1)}} \right) Norm(x_2) + 0.595x_3 + 2.676 \left( \frac{x_1}{x_4} \right) + 0.194x_5 - 33.563$$

Ta khảo sát đồ thị và các thống kê của mô hình:

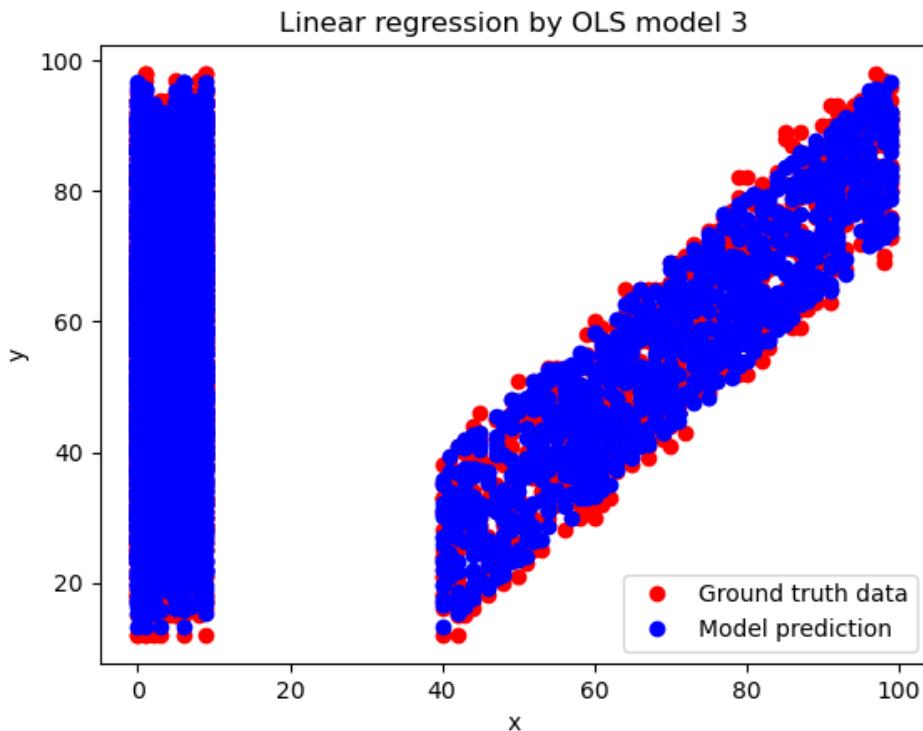


Figure 23 biểu đồ điểm dữ liệu thực tế và các điểm dự đoán của mô hình 3

Có thể nhận thấy, tương tự như mô hình tuyến tính ở mục 4, các dự đoán của mô hình 3 tương đối gần với các điểm thực tế. Ta cũng có thể đưa ra nhận xét rằng, một cách trực quang, đồ thị

của mô hình 3 so với mô hình tuyến tính đa biến bậc nhất ở mục 4 không có sự khác biệt rõ rệt. Tuy nhiên, khi khảo sát các thống kê của hai mô hình:

	Mô hình tuyến tính đa biến bậc nhất	Mô hình 3
MAE	1.596	1.621
Max Error	7.084	7.587
Min Error	0	0.002
Error Mean	-0.058	-0.068
Error Variance	4.093	4.211

Figure 24 Bảng thống kê mô hình tuyến tính đa biến bậc nhất và mô hình 3

Mô hình tuyến tính đa biến bậc nhất hầu như đều vượt trội hơn so với mô hình 3 ở mọi mặt. Điều này cũng chỉ rõ ra rằng, có vẻ dữ liệu mang tính tuyến tính nhiều hơn trong không gian sinh bởi các tham số tối ưu giả định của mô hình. Vì thế, từ đây ta có một bài học nhỏ: không nhất thiết mô hình ta phải được xây dựng trên các mối liên hệ mà con người thường nghĩ đến trong lĩnh vực của dữ liệu mà họ khảo sát; đôi khi, các mô hình đơn giản sẽ mang lại hiệu quả đú tốt.

## 7 Kết luận

Việc xác định, dự đoán các mối quan hệ cũng như số điểm của học sinh trong môi trường học vẫn đóng vai trò quan trọng đối với các cơ sở đào tạo nhằm thúc đẩy, cải thiện, theo dõi phương pháp học của học sinh và phương pháp giảng dạy của giảng viên. Trong bài thực hành này, ta đã tìm hiểu các bước để khảo sát dữ liệu, trích xuất dữ liệu trọng tâm và sau đó dùng nhiều mô hình khác nhau nhằm cố gắng giải thích, dự đoán điểm số của học sinh dựa trên các đặc trưng đầu vào. Bài thực hành cũng khảo sát một số các phương pháp thông dụng của Khoa học dữ liệu nhằm giảm thiểu độ lệch chuẩn của các sai số, đồng thời kiểm định mối tương quan giữa các biến độc lập với biến phụ thuộc, và giữa các cặp biến độc lập với nhau. Sau quá trình đó, bài thực hành kết luận rằng mô hình tuyến tính đa biến bậc nhất có dạng  $\sum_{i=1}^5 \theta_i x_i + \theta_6$  thể hiện tốt nhất trong việc dự đoán điểm số.

## Lời cảm ơn

Xin cảm ơn các tài liệu liên quan đến NumPy từ thầy Nguyễn Ngọc Toàn cũng như chi tiết các phương pháp thao tác, biến đổi ma trận từ thầy Trần Hà Sơn.

## 8 Phụ lục

Các tài liệu liên quan đến khảo sát dữ liệu tại đây: [Google Drive](#)

## 9 Tham khảo

- [1] J. S. Hair Jr, W. C. Black, B. J. Babin and R. E. Anderson, Multivariate Data Analysis, Pearson Prentice Hall, 2010.
- [2] O. El Aissaoui, Y. El Alami El Madani, L. Oughdir, D. A. and Y. El Alloui, "A Multiple Linear Regression-Based Approach to Predict Student Performance," *Advanced Intelligent Systems for Sustainable Development*, vol. 1102, pp. 9-23, 2020.
- [3] H. T. Vũ, "Machine Learning Cơ Bản, Linear Regression," 2016. [Online]. Available: <https://machinelearningcoban.com/2016/12/28/linearregression/>. [Accessed 2024].
- [4] T. B. Nguyễn, N. T. Đinh, Đ. T. Nguyễn and Đ. M. Nguyễn, Cơ sở toán cho khoa học dữ liệu, Ho Chi Minh City: Vietnam National University Ho Chi Minh City, 2022.
- [5] J. S. Cramer, The origins of logistic regression, 2020.
- [6] H. T. Vũ, "Machine Learning Cơ Bản, Regression Algorithms," 2016. [Online]. Available: <https://machinelearningcoban.com/2016/12/27/categories/#regression-algorithms>. [Accessed 2024].
- [7] A. Schneider, G. Hommel and M. Blettner, "Linear Regression Analysis," *Dtsch Arztebl Int*, pp. 776-782, 2010.
- [8] Duke University, "Linear Regression Analysis," [Online]. Available: <https://people.duke.edu/~rnau/compare.htm>.
- [9] H. T. Vũ, "Machine Learning Cơ Bản, Overfitting," 2017. [Online]. Available: <https://machinelearningcoban.com/2017/03/04/overfitting/>. [Accessed 2024].
- [10] Wikipedia, "Cross validation (statistics)," [Online]. Available: [https://en.wikipedia.org/wiki/Cross-validation\\_\(statistics\)](https://en.wikipedia.org/wiki/Cross-validation_(statistics)).
- [11] F. S. V. Bazán, "Simple and Efficient Determination of the Tikhonov Regularization Parameter Chosen by the Generalized Discrepancy Principle for Discrete Ill-Posed Problems," *Springer*, 2014.
- [12] Wikipedia, "Anscombe Transformation," [Online]. Available: [https://en.wikipedia.org/wiki/Anscombe\\_transform#Alternatives](https://en.wikipedia.org/wiki/Anscombe_transform#Alternatives).