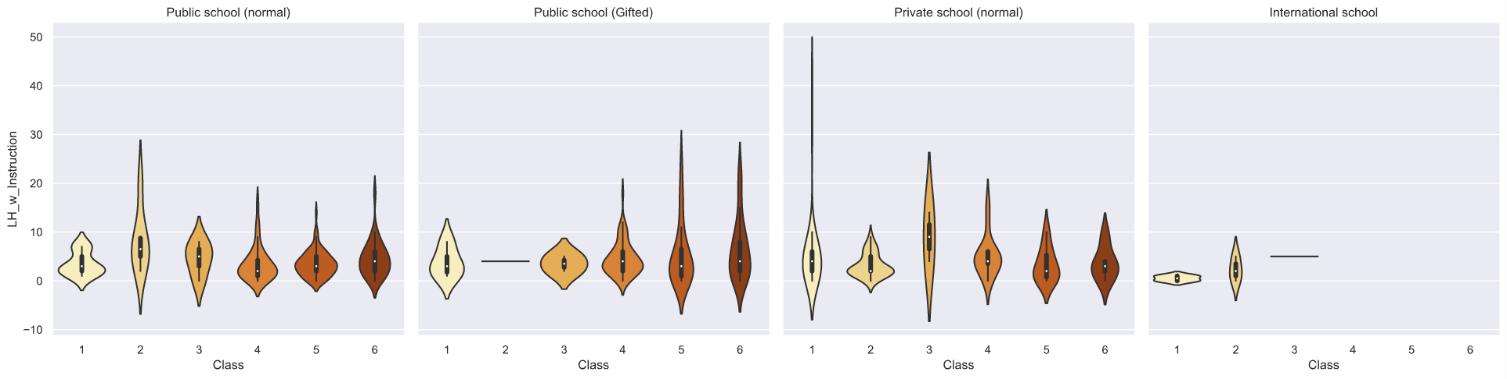
**Số giờ học dựa trên khối lớp và loại trường.**

Trực quan sự phân phối dữ liệu của số giờ học trung bình dựa trên khối lớp và phân loại theo các loại trường khác nhau.



Sử dụng **violin plot** giúp hiển thị toàn bộ phân phối dữ liệu. Đây là loại plot bao gồm box plot ở trọng tâm: chấm trắng thể hiện trung vị (median) của dữ liệu, thanh xám thể hiện tứ phân vị. Đối xứng 2 bên hiển thị phân phối mật độ của dữ liệu như Kernel density plot. Biểu đồ trên còn sử dụng màu sắc từ lợt đến đậm để hiển thị khối lớp tăng dần.

Từ biểu đồ vẽ được, ta thấy:

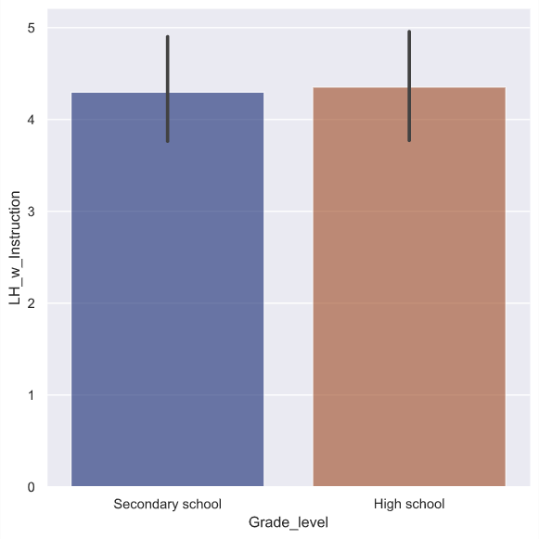
* Dữ liệu phân phối qua các khối lớp là đa dạng.
* *‘Lớp 1 trường công’* có phân phối 2 đỉnh rõ rệt, phần lớp quan sát tập trung ở đỉnh lớn. ‘*Trường quốc tế’* có phân bố dữ liệu hẹp, có thể do có ít quan sát.
* *‘Trường công’* có các trung vị tương đối gần nhau, khoảng **3 giờ**. Trong khi ‘*trường tư* *và trường quốc tế’* sự chênh lệch này là nhiều hơn, đặc biệt ở ‘*lớp 3 trường tư’* (gần **9 giờ**).
* *‘Lớp 1 trường tư’* có một vài giá trị ngoại lệ khiến phân phối bị kéo lệch lên trên. Một số điểm ngoại lệ có thể xuất hiện ở *‘lớp 2 trường công’*, *‘lớp 5, 6 trường chuyên’*.
* Một số loại không có hình như ‘*lớp 2 trường công chuyên’*, *‘lớp 3 trường quốc tế’* vì chỉ có 1 vài quan sát. Các ‘*lớp 4, 5, 6 trường quốc tế’* không có quan sát nào.

Thống kê trung bình số giờ tự học của học sinh *cấp 2* và *cấp 3*.

Table

Description automatically generated

Trực quan hóa dữ liệu trên:



Với **bar plot** trên, ta thấy:

* Nhìn chung, thời gian tự học của khối *cấp 2* và *cấp 3* là tương đối như nhau.
* Thanh error bars phía trên cho thấy đại diện cho sự thay đổi của dữ liệu cũng gần như tương đương nhau.

Để chắc chắn, ta sẽ kiểm định thống kê thời gian tự học trung bình 2 tổng thể trên. Sử dụng kiểm định 2 mẫu Z-test, ta có:

* Giả thuyết (null hypothesis): trung bình thời gian tự học của 2 mẫu là bằng nhau.
* Đối thuyết (alternative hypothesis): trung bình thời gian tự học của 2 mẫu là khác nhau.
* Mức ý nghĩa

Text

Description automatically generated

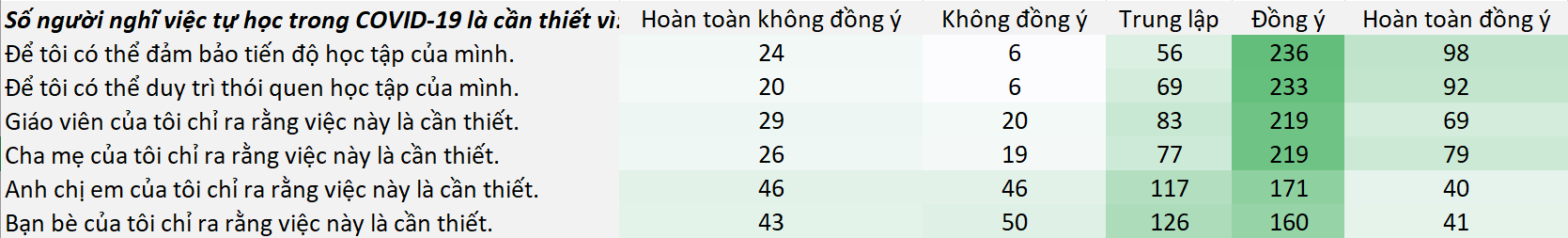
Kết luận: với mức ý nghĩa α = 0.05, ta không đủ cơ sở để bác bỏ giả thuyết . Nghĩa là trung bình thời gian tự học của cấp 2 bằng trung bình thời gian tự học cấp 3.

**Mối quan hệ giữa quan điểm cá nhân và lí do xúc tác ảnh hưởng đến thời gian tự học trong thời COVID-19.**

***Tại sao việc tự học thời Covid-19 lại cần thiết?***

Dữ liệu thu thập sử dụng thang đo Likert để cho phép cá nhân thể hiện mức độ họ đồng ý hoặc không đồng ý với một tuyên bố cụ thể *(Hoàn toàn không đồng ý, Không đồng ý, Trung lập, Đồng ý, Hoàn toàn đồng ý)*, với giả định rằng cường độ của một thái độ là tuyến tính.

Bảng thống kê mô tả bên dưới được kết hợp với **heat map** (số càng lớn thì màu càng đậm) cho dễ trực quan về định lượng.



Chart, bar chart

Description automatically generated

Với **area chart** bên trên được vẽ từ bảng thống kê, ta có thể thấy:

* Hầu hết người được khảo sát đều từ *‘đồng ý’* tới *‘hoàn toàn đồng ý’* với các ý kiến được đưa ra, chiếm khoảng **65%.**
* Số người có ý kiến *‘trung lập’* chiếm khoảng **21%**.
* Còn lại khoảng **14%** số người có phản hồi tiêu cực với ý kiến được đưa ra.
* Số người *‘hoàn toàn đồng ý’* chiếm tỉ lệ lớn hơn với số người *‘hoàn toàn không đồng ý’*, chỉ trừ ở 2 lí do đầu tiên.

Chart, bar chart

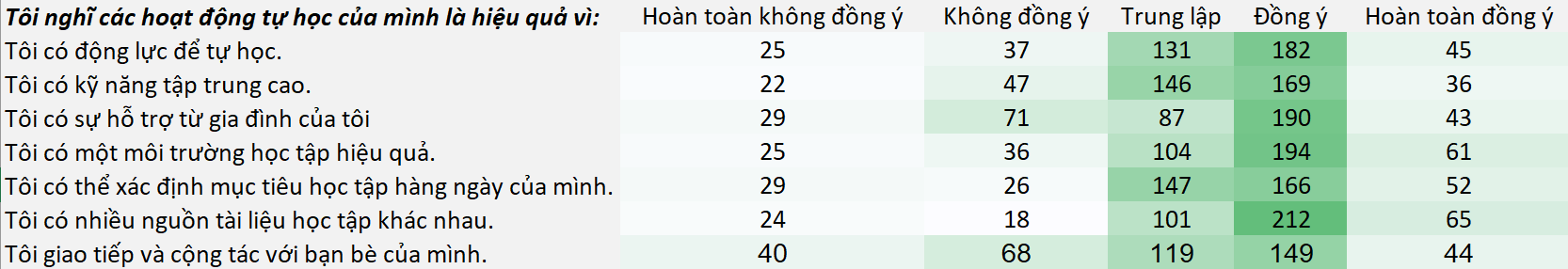
Description automatically generated

Để xem chi tiết số liệu hơn, ta quan sát **diverging stacked bar chart**. Có thể thấy:

* Sự chênh lệch trong nhóm *‘đồng ý’* và *‘trung lập’* có ít chênh lệch nhất giữa các ý kiến. Trong khi đó, 3 nhóm *‘hoàn toàn đồng ý’*, *‘không đồng ý’* và *‘hoàn toàn không đồng ý’* thì sự chênh lệch giữa các ý kiến là gần như gấp đôi.
* Lí do mà nhiều người cho rằng thúc đẩy họ trong việc tự học trong thời kỳ Covid-19 nhất là vì họ muốn đảm bảo tiến độ học tập của mình. Theo sau là việc họ muốn duy trì thói quen học tập từ trước khi diễn ra dịch bệnh. Đây cũng là 2 lí do nhận ít phản hồi tiêu cực nhất.
* 2 loại người xúc tác việc tự học trong thời Covid-19 không được đồng tình nhất là bạn bè và anh chị em. Với phản hồi tiêu cực nhiều **gấp đôi** so với cha mẹ và giáo viên.

***Điều gì giúp việc tự học thời Covid-19 lại hiệu quả?***

Bảng thống kê mô tả bên dưới được thực hiện tương tự nhưng với các ý kiến về mức độ hiệu quả của việc tự học của người được khảo sát.



Tương tự ta có thể vẽ 2 biểu đồ để xem từ tổng quát đến chi tiết bảng thống kê trên.

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

Một vài kết luận được rút ra:

* Giao tiếp và cộng tác với bạn bè là một nhân tố không được nhiều người đánh giá cao trong việc giúp việc tự học hiệu quả nhất.
* Sự chênh lệch về phản hồi tiêu cực giữa các ý kiến còn lại là không nhiều. Ý kiến về yếu tố hỗ trợ từ gia đình có số *‘không đồng ý’* khá nhiều hơn so với *‘hoàn toàn không đồng ý’*, chứng tỏ đây là ý kiến tiêu cực nhưng không bị đánh giá gay gắt.
* Lí do được nhiều người đồng tình nhất là việc có nhiều nguồn tài liệu học tập phong phú và môi trường học tập hiệu quả.

***Quan điểm có ảnh hưởng đến thời gian tự học của cá nhân?***

Bảng thống kê sau xem xét tới trung bình thời gian tự học của người được nhóm theo mức độ phản hồi của họ về các ý kiến về sự cần thiết của việc tự học.

****

Chart

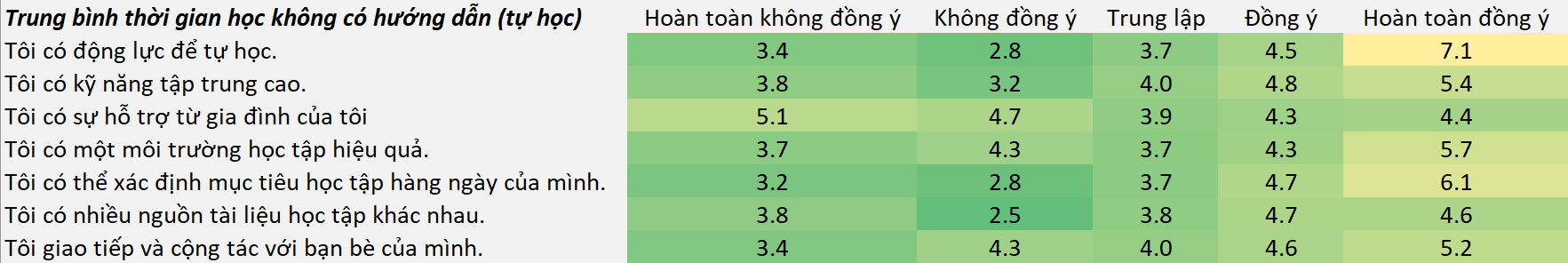
Description automatically generated

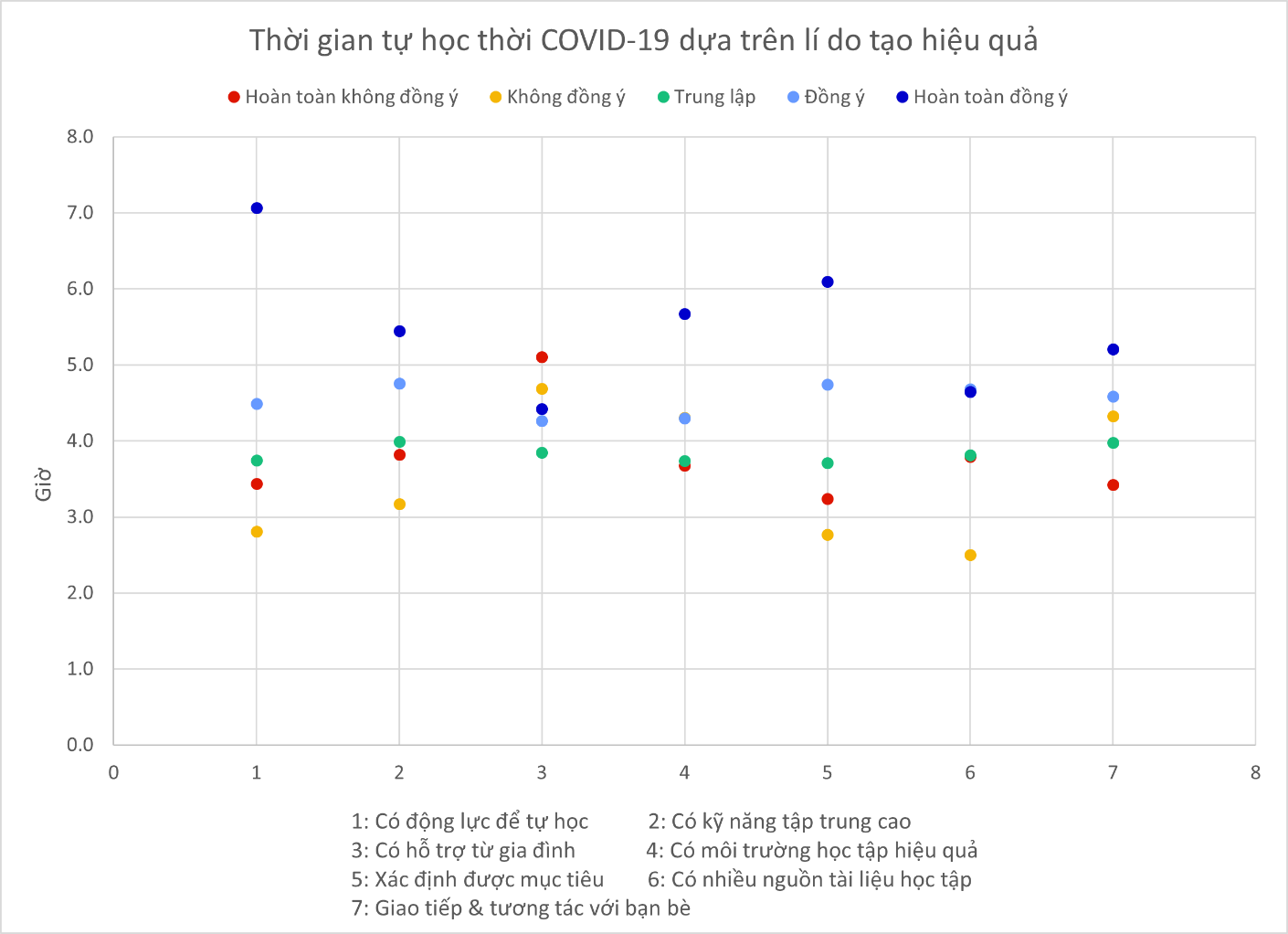
Sử dụng **box and whisker plot** để so sánh thời gian học online:

* Điểm thú vị là những người có quan điểm tích cực nhất *(Hoàn toàn đồng ý)* và những người có quan điểm tiêu cực nhất *(Hoàn toàn không đồng ý)* lại có thời gian tự học nhiều nhất, rơi vào khoảng từ **5** đến **6 giờ.**
* Trong khi đó nhóm *‘không đồng ý’* có thời gian tự học thấp nhất.
* 3 nhóm còn lại: *‘trung lập’* và *‘không đồng ý’* cũng có thời gian tự học trung bình rơi vào khoảng tương tự vào khoảng **4** **giờ**. Tuy nhiên nhóm *‘không đồng ý’* có biên độ dao động lớn hơn. Nhóm người *‘đồng ý’* có độ lệch nhỏ nhất.
* Đặc biệt, trung bình thời gian tự học của người *‘hoàn toàn đồng ý’* với ý kiến “Bạn bè của tôi chỉ ra việc này là cần thiết” là cao nhất (mặc cho đây là ý kiến nhận được nhiều phản hồi tiêu cực nhất). Điều này xảy ra ngược lại với nhóm *‘hoàn toàn không đồng ý’*, ý kiến “Tôi muốn đảm bảo tiến độ học tập của mình” được nhiều phản hồi tích cực nhất lại có thời gian tự học thấp nhất.

🡺 Cả 2 là những nhóm có quan điểm mạnh, họ có thời gian tự học cao hơn và không phụ thuộc vào xu hướng số đông.

Tương tự, ta xem xét thống kê trung bình thời gian tự học của người được nhóm theo mức độ phản hồi của họ về các ý kiến về sự hiệu quả của việc tự học.



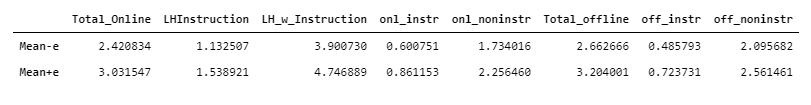


Với **scatter plot** bên trên, dựa trên quan điểm về độ hiệu quả, ta rút ra được:

* Trung bình thời gian tự học của nhóm người có ý kiến tích cực nhất cao hơn so với các nhóm còn lại.
* Tương tự phân tích trước, nhóm người *‘không đồng ý’* có thời gian tự học thấp nhất.
* Chỉ ở lí do “Việc tự học có hiệu quả nhờ có sự hỗ trợ từ gia đình”, ta thấy nhóm tiêu cực (Hoàn toàn không đồng ý, Không đồng ý) có thời gian tự học cao hơn nhóm tích cực (Hoàn toàn đồng ý, Đồng ý). Khoảng phân bố ở lí do này cũng nhỏ hơn các lí do khác, tức đây là lí do không quá ảnh hưởng đến sự giao động số giờ học.
* Nhóm *‘Trung lập’* là nhóm có trung bình thời gian học ổn định nhất, vào khoảng **4 giờ**.

**Thống kê mô tả số giờ tự học.**

Với độ tin cậy **95%**, ta có bảng ước lượng các giá trị trung bình các thuộc tính như sau:

****

Trước tiên, ta xem xét về thời gian tự học mà người được khảo sát bỏ ra trong thời kỳ Covid-19.

Chart

Description automatically generated Chart, pie chart

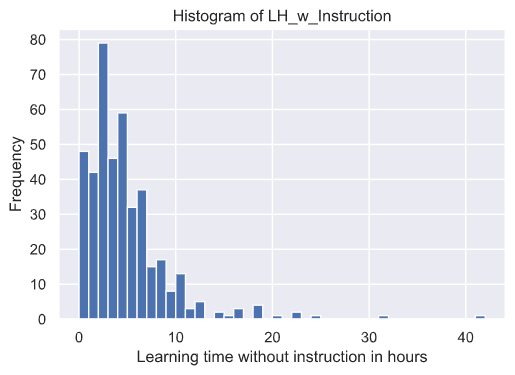
Description automatically generated

Từ **boxplot** trên, ta có thể rút ra một số dữ kiện về số giờ tự học:

* Thời gian tự học nhỏ nhất là **0 giờ**.
* Thời gian tự học lớn nhất là **42 giờ**. Lớn nhất không phải ngoại lệ là **12 giờ**.
* Trung bình (mean) là **4 giờ.** Trung vị (median) là **3 giờ.**
* Tứ phân vị Q1 (25th percentile) là **2 giờ**, tứ phân vị Q3 (75th percentile) là **6 giờ**.
* Nghĩa là có ít nhất **75%** người được khảo sát có thời gian tự học là **2 giờ** trở lên. Ít nhất **một nửa** người được khảo sát có thời gian tự học là từ **3 giờ** trở lên.
* Tập dữ liệu có **9** ngoại lệ (outlier) có khác biệt đáng kể so với quan sát khác. Một số có thể do trả lời khảo sát không hợp lệ (24, 31, 42).

Trong đó, thời gian học trực tuyến chiếm **46%**, thời gian học ngoại tuyến chiếm nhiều hơn **8%** (**54%**).

Ta sẽ xem xét số phân bố dữ liệu thời gian tự học.

****

Trực quan nhanh bằng **histogram chart**, ta thấy:

* Biểu đồ phân bố lệch về bên trái.
* Mức độ thường xuất hiện của thời gian tự học nằm trong khoảng từ **0** đến **6 giờ.**
* Giá trị thường xuất hiện nhất là **2 giờ**.
* Có 1 vài giá trị bất thường ở khoảng **20** đến **40 giờ**, có thể đây là dữ liệu không hợp lệ.

**Mối quan hệ giữa số giờ tự học và số giờ học trực tiếp, ngoại tuyến (offline).**

***Phân tích hồi quy tuyến tính.***

Ta sẽ ước tính hệ số cho mô hình quy luật này. Hệ số được ước tính bằng cách sử dụng phương pháp bình phương nhỏ nhất (OLS). Với hàm mô tả dạng tuyến tính:

Text

Description automatically generated Table

Description automatically generated

Như vậy có thể thấy:

* Hệ số chặn β0 (intercept) là 0.937791 với khoảng tin cậy là [0.524999, 1.350583]. Hệ số độ dốc β1 (slope) là 1.154324 với khoảng tin cậy là [1.052994, 1.255655].
* Nghĩa là khi số giờ học ngoại tuyến (Total\_offline) tăng 1 giờ thì số giờ tự học (LH\_w\_Instruction) tăng lên 1 giờ 9 phút.
* Phương trình hồi quy:

***Kiểm định thống kê mối quan hệ này.***

* Giả thuyết (null hypothesis): không có mối quan hệ giữa số giờ tự học và số giờ học ngoại tuyến.
* Đối thuyết (alternative hypothesis): có mối quan hệ giữa số giờ tự học và số giờ học ngoại tuyến.
* Mức ý nghĩa

Table

Description automatically generated

Xem rõ giá trị p-value:

Text

Description automatically generated with low confidence

statsmodels sử dụng độ tin cậy 95%, nghĩa là α=0.05. Ta thấy p-value của số giờ học ngoại tuyến (Total\_offline) có giá trị rất bé hơn so với α (1.494396e-73 << 0,05). Nghĩa là với mức ý nghĩa α:

* Bác bỏ giả thuyết H0.
* Nghĩa là có mối quan hệ giữa 2 biến *LH\_w\_Instruction* và *Total\_offline*.
* Biến *Total\_offline* có ý nghĩa đối với mô hình về mặt thống kê (p-value = 1.494396e-73).
* Mô hình phù hợp tốt với dữ liệu quan sát về mặt thống kê (p-value = 1.494396e-73).
* Biến *Total\_offline* có thể giải thích được **54.4%** sự thay đổi của biến *LH\_w\_Instruction.*

***Kiểm định hệ số tương quan.***

Như vậy 2 biến trên có mối quan hệ tuyến tính, nghĩa là chúng không còn độc lập lẫn nhau. Ta có thể sử dụng hệ số tương quan *Pearson* để đo lường mức độ quan hệ tuyến tính giữa hai biến.

A picture containing text, watch

Description automatically generated



***Phân tích hồi quy tuyến tính bội.***

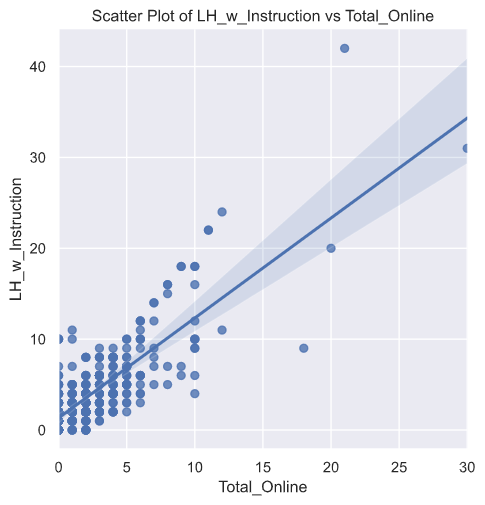
Vậy còn tương quan giữa số giờ tự học và các thuộc tính còn lại thì sao? Ta thử xem xét bằng phân tích hồi quy tuyến tính bội.

Table

Description automatically generated

Bảng trên tập hợp các hệ số tương quan để đo lường mức độ mạnh yếu của mối quan hệ giữa hai biến số.

Có thể thấy hệ số β1 tương ứng với thuộc tính *Total\_Online* là khá lớn so với các thuộc tính khác. Có nghĩa nó là thuộc tính tương quan tốt với *LH\_w\_Instruction*. Ta xem xét mối quan hệ giữa 2 thuộc tính này.



Dựa vào **scatter plot** trên, đúng như dự đoán:

* Biểu đồ có dạng tuyến tính. Bờ dốc dương, Y tăng khi X tăng.
* Các điểm phân tán không rộng mà tập trung xung quanh đường thẳng, mối quan hệ được coi là mạnh.
* Có thể thấy khi thời gian học trực tuyến tăng thì thời gian tự học cũng tăng theo.

Thực hiện huấn luyện bộ dữ liệu với `scikit-learn`*.*

Sau khi huấn luyện, ta có:

* Bộ train gồm 315 dòng. Sử dụng để huấn luyện và dự đoán giờ tự học.
* Bộ test gồm 105 dòng. Sử dụng để validate dự đoán.



Biến *Total\_Online* có thể giải thích được 76.35 % sự thay đổi của biến *LH\_w\_Instruction.*

A picture containing text

Description automatically generated

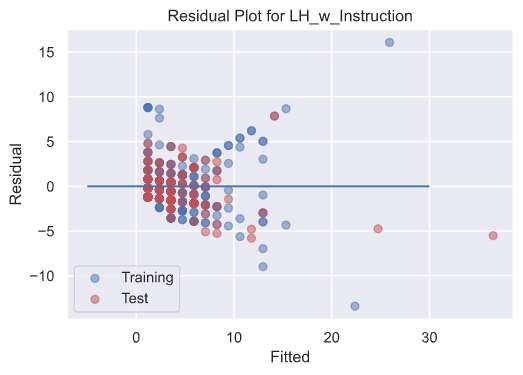
Ta nhận thấy *mean square error (MSE)* giữa bộ train và bộ test khá gần nhau.

***Biểu đồ phần dư.***

Trong phân tích hồi quy, hiệu giá trị quan sát của biến phụ thuộc và giá trị dự đoán được gọi là phần dư (residual). Mỗi điểm dữ liệu có một phần dư:

Residual = Observed value – Predicted value

Do đó để kiểm định mô hình đã huấn luyện, ta có thể trực quan hóa bằng **residual plots**:



Nhận xét:

* Residual plots là đồ thị biểu diễn phần dư trên trục tung và biến số độc lập trên trục hoành.
* Các điểm trong một ô dư được phân tán ngẫu nhiên quanh trục hoành. Ta có thể kết luận đây là một mô hình hồi quy tuyến tính là thích hợp cho dữ liệu.

Để dự đoán thời gian tự học của 1 người, ta chỉ cần sử dụng lại kết quả của hồi quy tuyến tính trên. Ví dụ:



***Phân tích hồi quy Logistic.***

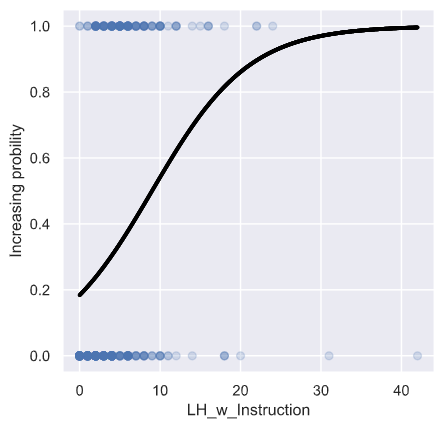
Ta thử tìm mối liên hệ thời gian tự học đến xác suất việc 1 học sinh chăm học hơn (số giờ học sau khi có Covid nhiều hơn trước khi có Covid).

Ta sẽ sử dụng hồi quy Logistic để xem mối liên hệ này. Với mọi giá trị t đầu vào, đầu ra luôn có giá trị từ 0 đến 1: . Truyền vào với tham số t = β0+β1x là dạng biểu diễn hồi quy tuyến tính đơn. Ta có:

Table

Description automatically generated

* Biến *LH\_w\_Instruction* (giờ tự học) có ý nghĩa đối với mô hình về mặt thống kê (p-value = 0.000)
* Nhưng mô hình chỉ giải thích được **7.2%** sự thay đổi của biến is\_increase (có chăm học hơn không). Ta có thể thử trực quan hàm xác suất:



Sau khi trực quan hàm xác suất thì đúng như vậy, ta có thể thấy:

* + Mặc dù dạng đường xác suất tăng. Tức là số giờ tự học tăng thì xác suất học sinh chăm học hơn cũng tăng.
  + Nhưng đây không hoàn toàn là đường cong cho hồi quy logit tốt, đường xác suất không chạm tới 0. Hơn nữa, nhìn vào 2 biểu đồ phân bố thì cũng thấy số giờ học phân bố rải rác từ 0 tới 10 là chủ yếu mặc cho xác suất tăng.

**Tài liệu tham khảo.**

[1] Bùi Tiến Lên. Data Analysis. Đại học Khoa học Tự nhiên, 2021.

[2] Nguyễn Tuấn. Biểu diễn dữ liệu với Box and Whisker Plot, 2021. URL: <https://chidokun.github.io/2021/02/box-and-whisker-plot/>

[3] Leila Gharani. Charting Survey Results in Excel, 2021. URL: <https://www.xelplus.com/charting-survey-results-excel/>

[4] Ritchie Ng. Evaluating a Linear Regression Model, 2021. URL: <https://www.ritchieng.com/machine-learning-evaluate-linear-regression-model/>

[5] Jose Marcial Portilla. Udemy's course note, 2021. URL: <https://github.com/jmportilla/Udemy---Machine-Learning/blob/master/Supervised%20Learning%20-%20%20Linear%20Regression.ipynb>

[6] Jason Brownlee. 17 Statistical Hypothesis Tests in Python (Cheat Sheet), 2021. URL: <https://machinelearningmastery.com/statistical-hypothesis-tests-in-python-cheat-sheet/>

[7] Ankit Das. Z-Tests vs T-Tests: How To Choose Among Two Important Hypothesis Tests, 2021. URL: <https://analyticsindiamag.com/z-test-t-test-statistcs/>