

ĐẠI HỌC QUỐC GIA TP HCM
TRƯỜNG ĐẠI HỌC BÁCH KHOA



BÁO CÁO BÀI TẬP LỚN

Môn : Xác suất thống kê

Khoa : Điện – Điện tử

GVHD: Phan Thị Hường

Lớp : L10

Nhóm : 13

STT	Mã số SV	Họ	Tên
1	2213803	Trần Anh	Tuấn
2	2213821	Nguyễn Duy	Tuyên
3	2213962	Huỳnh Thanh	Vinh
4	2213799	Phạm Đức	Tuấn
5	2211517	Nguyễn	Khánh

Thành Phố Hồ Chí Minh, Tháng 11-2023

Nhóm trưởng : Trần Anh Tuấn - tuank22@hcmut.edu.vn

Thành viên	Điểm chia +2	Đóng góp
Trần Anh Tuấn	0	20%
Nguyễn Duy Tuyên	0	20%
Huỳnh Thanh Vinh	0	20%
Phạm Đức Tuấn	0	20%
Nguyễn Khánh	0	20%

Điểm của GVHD

Thành viên	Điểm	Nhận xét
Trần Anh Tuấn		
Nguyễn Duy Tuyên		
Huỳnh Thanh Vinh		
Phạm Đức Tuấn		
Nguyễn Khánh		

CONTENTS

1. Tổng quan dữ liệu.....	4
1.1. Ngữ cảnh của dữ liệu	4
1.2. Các biến.....	4
2. Kiến thức nền.....	5
2.1. Phân tích phương sai – ANOVA một nhân tố	5
2.2. Mô hình hồi quy tuyến tính bội.....	6
2.3. Đánh giá sự phù hợp của mô hình	9
2.4. Phương pháp bình phương nhỏ nhất	12
2.5. Các giả định của mô hình hồi quy	12
2.6. Lý thuyết về hồi quy binary logistic	13
2.7. Phương trình hồi quy binary logistic	14
3. Tiền xử lý số liệu:.....	16
3.1 Đọc dữ liệu:	16
3.2. Làm sạch dữ liệu (Data cleaning)	16
3.3. Xử lý dữ liệu khuyết:	17
3.4. Làm rõ dữ liệu:	17
4. Thống kê tả:	18
5) Thống kê suy diễn :	29
6. Thảo luận và mở rộng:.....	36
6.1 Mục đích xây dựng mô hình hồi quy logistic:	36
6.2. Lợi ích đạt được:	36
6.3. Ứng dụng thực tế:.....	36
7. Nguồn dữ liệu và nguồn code.....	37

1. Tổng quan dữ liệu

1.1. Ngữ cảnh của dữ liệu

Tập dữ liệu chứa thông tin về một cửa hàng điện tử trực tuyến. Cửa hàng có ba kho để giao hàng cho khách hàng. Dựa vào dữ liệu trên tìm được mối quan hệ giữa các biến để từ đó đưa ra được phỏng đoán, xây dựng được mô hình dự báo mức độ hài lòng của khách hàng

- Tiêu đề : Transactional Retail Dataset of Electronics Store
- Thông tin nguồn :
 - Tác giả : SHAHRAYAR
 - Ngày : 2 năm trước
- Giá trị quan trắc : 500
- Số lượng biến : 16

1.2. Các biến

Dữ liệu bao gồm các biến:

- **order_id**: Một id duy nhất cho mỗi đơn hàng
- **customer_id**: Một id duy nhất cho mỗi khách hàng
- **date**: Ngày đặt hàng, được đưa ra ở định dạng YYYY-MM-DD
- **nearest_warehouse**: Một chuỗi biểu thị tên kho gần khách hàng nhất
- **shopping_cart**: Danh sách các bộ dữ liệu đại diện cho các mục đơn hàng: phần tử đầu tiên của bộ dữ liệu là mục được sắp xếp và phần tử thứ hai là số lượng đặt hàng cho mặt hàng đó
- **order_price**: Một số float biểu thị giá đặt hàng bằng USD. Giá đặt hàng là giá của mặt hàng trước khi có bất kỳ khoản giảm giá và/hoặc phí giao hàng nào được áp dụng.
- **delivery_charges**: Một hình nổi thể hiện phí giao hàng của đơn hàng
- **customer_lat**: Vĩ độ vị trí của khách hàng
- **customer_long**: Kinh độ vị trí của khách hàng
- **coupon_discount**: Một số nguyên biểu thị phần trăm chiết khấu được áp dụng cho order_price.
- **order_total**: một số float biểu thị tổng đơn đặt hàng bằng USD sau khi giảm giá và/hoặc phí giao hàng được áp dụng
- **season**: Một chuỗi biểu thị mùa mà đơn hàng được đặt.
- **is_expedited_delivery**: Một boolean biểu thị liệu khách hàng có yêu cầu giao hàng nhanh hay không

- **distance_to_nearest_warehouse:** Một số float biểu thị khoảng cách vòng cung, tính bằng km, giữa khách hàng và nhà kho gần nhất với họ.
- **latest_customer_review:** Một chuỗi thể hiện đánh giá mới nhất của khách hàng về đơn hàng gần đây nhất của họ
- **is_happy_customer:** Một boolean biểu thị liệu khách hàng có hài lòng hay không hoặc có vấn đề với đơn hàng gần đây nhất của họ.

2. Kiến thức nền

2.1. Phân tích phương sai – ANOVA một nhân tố

- Phân tích phương sai là một mô hình dùng để xem xét sự biến động của một biến ngẫu nhiên định lượng X chịu tác động trực tiếp của một hay nhiều yếu tố nguyên nhân.
- Trong mô hình phân tích phương sai một yếu tố, ta kiểm định so sánh trung bình của biến ngẫu nhiên X ở những tổng thể khác nhau dựa vào các mẫu quan sát lấy từ những tổng thể này. Các tổng thể được phân biệt bởi các mức độ khác nhau của yếu tố đang xem xét.
 - Giả thiết của bài toán ANOVA
 - Các tổng thể có phân phối chuẩn $N(\mu_i, \sigma_i^2)$ với $i = 1, 2, \dots, k$ với k là số tổng thể ($k \geq 3$)
 - Phương sai các tổng thể bằng nhau $\sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2$
 - Các mẫu quan sát (từ các tổng thể) được lấy độc lập
 - Đặt giả thuyết kiểm định
 - Giả thuyết không H_0 (null hypothesis): $\mu_1 = \mu_2 = \dots = \mu_i$
 - Giả thuyết đối H_1 (alternative hypothesis): $\exists \mu_i \neq \mu_j$ với $i \neq j$
 - Tính giá trị kiểm định thống kê

Source of variation	Tổng bình phương chênh lệch	Bậc tự do	Phương sai (Trung bình BPCL)	Tiêu chuẩn kiểm định
Between Groups	$SSB = \sum_{j=1}^k n_j * (\bar{x}_j - \bar{x})^2$	k-1	$MSB = \frac{SSB}{k-1}$	$F = \frac{MSB}{MSW}$
Within Groups	$SSW = \sum_{j=1}^k \sum_{i=1}^{n_i} (x_{ij} - \bar{x}_j)^2$	N-k	$MSW = \frac{SSW}{N-k}$	
Total	$SST = \sum_{j=1}^k \sum_{i=1}^{n_i} (x_{ij} - \bar{x})^2$ SST=SSB+SSW	N-1		

2.2. Mô hình hồi quy tuyến tính bội

- Hồi quy tuyến tính bội là một phần mở rộng của hồi quy tuyến tính đơn. Nó được sử dụng khi ta muốn dự đoán giá trị của một biến phản hồi dựa trên giá trị của hai hoặc nhiều biến giải thích khác. Biến mà chúng ta muốn dự đoán được gọi là biến phản hồi (biến phụ thuộc). Các biến mà ta đang sử dụng để dự đoán giá trị của biến phản hồi được gọi là các biến giải thích (biến dự báo, biến phụ thuộc). Ví dụ: sử dụng hồi quy bội số để dự đoán kết quả kỳ thi XSTK dựa trên thời gian ôn tập, niên khóa, giới tính của sinh viên.
- Hồi quy bội cũng cho phép chúng ta xác định sự phù hợp tổng thể của mô hình và đóng góp tương đối của từng yếu tố dự báo và tổng phương sai được giải thích.
- Vì một biến dự báo đơn lẻ không cung cấp sự mô tả đầy đủ vì một số lượng các biến dự báo chìa khóa tác động đến biến đáp ứng theo các cách đặc biệt và quan trọng. Các dự báo của biến đáp ứng dựa vào mô hình chỉ có một biến dự báo riêng lẻ là không chính xác. Vì thế mô hình hồi quy tuyến tính bội được đưa ra:

- Xét trường hợp có $p - 1$ biến dự báo X_1, \dots, X_{p-1} . Mô hình hồi quy:
 - $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{ip-1} + \varepsilon_i$ được gọi là mô hình bậc nhất với $p - 1$ biến dự báo.
 - Hay : $Y_i = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{ik} + \varepsilon_i$
 - Trong đó:
 - $\beta_1, \beta_2, \dots, \beta_{p-1}$: là các tham số
 - $X_{i1}, X_{i2}, \dots, X_{ip-1}$: là các hằng số đã biết
 - $\varepsilon_i \sim N(0; \sigma^2)$
 - $i = 1, 2, \dots, n$
 - Khi $p - 1 = 1$ mô hình hồi quy là: $Y_i = \beta_0 + \beta_1 X_{i1} + \varepsilon_i$ là mô hình hồi quy tuyến tính đơn.
 - Giả sử: $E\{\varepsilon_i\} = 0$, hàm đáp ứng với mô hình:
 - $E\{Y\} = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{ip-1}$
 - Tham số β_k chỉ sự thay đổi của trung bình đáp ứng $E(Y)$ với 1 đơn vị tăng trong biến dự báo X_k khi tất cả các biến dự báo còn lại được coi là hằng số.
 - Ảnh hưởng của biến dự báo bất kỳ trong trung bình đáp ứng là như nhau khi các biến dự báo khác được cố định. Do đó, mô hình hồi quy bậc nhất được thiết kế cho các biến dự báo mà ảnh hưởng của nó trên trung bình đáp ứng là cộng tính hay không có tương tác.

- Vậy mô hình hồi quy tuyến tính tổng quát với điều kiện sai số chuẩn có các quan sát Y_i là các biến chuẩn độc lập, với trung bình $E\{Y\}$ và phương sai không đổi σ^2 .

➤ Các biến dự báo định tính

Mô hình hồi quy tuyến tính tổng quát: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \dots + \beta_{p-1} X_{ip-1} + \varepsilon_i$ bao gồm các biến dự báo định lượng và các biến dự báo định tính. Nên ta sử dụng các biến số nhận giá trị 0 và 1 để định nghĩa các lớp giá trị của biến định tính.

➤ Ước lượng các hệ số hồi quy

Tiêu chuẩn bình phương cực tiểu được tổng quát hóa cho mô hình hồi quy tuyến tính tổng quát như sau:

$$Q = \sum_{i=1}^n \varepsilon_i^2 = \sum_{i=1}^n (Y_i - \beta_0 - \beta_1 X_{i1} - \dots - \beta_{p-1} X_{ip-1})^2$$

Các ước lượng bình phương cực tiểu là các giá trị của $\beta_0, \beta_1, \dots, \beta_{p-1}$ làm cực tiểu hóa Q. Ta biểu diễn vector ước lượng các hệ số hồi quy $b_0, b_1, \dots, \beta_{p-1}$ là b:

$$b = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix}$$

Các phương trình chuẩn bình phương cực tiểu cho mô hình hồi quy tuyến tính tổng quát là: $(X'X)b = X'Y$ và các ước lượng bình phương cực tiểu là: $b = (X'X)^{-1}X'Y$

➤ Bảng ANOVA cho mô hình hồi quy tuyến tính tổng quát

Nguồn biến đổi	SS	df	MS
Hồi quy	$SSR = b'X'Y - \left(\frac{1}{n}\right)Y'JY$	p-1	$MSR = \frac{SSR}{p-1}$
Sai số	$SSR = Y'Y - b'X'Y$	n-p	$MSE = \frac{SSE}{n-p}$
Tổng số	$SSR = Y'Y - \left(\frac{1}{n}\right)Y'JY$	n-1	

➤ Kiểm định F cho quan hệ hồi quy

Để kiểm định liệu có hay không quan hệ hồi quy giữa biến đáp ứng và các biến X: X_1, \dots, X_{p-1} tức là lựa chọn giữa các giả thuyết:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{p-1} = 0$$

H_a : không phải tất cả $\beta_k, k = 1, \dots, p - 1$ đều = 0. Ta dùng một thống kê kiểm định: $F^* = \frac{MSR}{MSE}$

Nếu $F^* \leq F(1 - \alpha; p - 1; n - p)$ chấp nhận H_0

Nếu $F^* > F(1 - \alpha; p - 1; n - p)$ chấp nhận H_a

➤ Hệ số xác định bội R^2

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

Hệ số xác định bội cho biết các biến dự báo trong mô hình giải thích được bao nhiêu phần trăm sự thay đổi của biến đáp ứng. Vì thế ta có: $0 \leq R^2 \leq 1$. $R^2 = 0$ khi tất cả các giá trị $b_k = 0$ ($k = 1, \dots, p - 1$). $R^2 = 1$ khi tất cả các quan sát nằm trên mặt phẳng đáp ứng, tức $Y_i = \hat{Y}_i$ với mọi i

Thêm nhiều hơn các biến dự báo X vào mô hình có thể chỉ làm tăng thêm R^2 .

Hệ số xác định bội hiệu chỉnh, ký hiệu là R_a^2 , điều chỉnh R^2 bằng cách chia mỗi tổng bình phương cho bậc tự do của nó:

$$R_a^2 = 1 - \frac{\frac{SSE}{n - p}}{\frac{SSTO}{n - 1}} = 1 - \left(\frac{n - 1}{n - p} \right) \frac{SSE}{SSTO}$$

Hệ số xác định bội hiệu chỉnh thực sự có thể nhỏ hơn khi biến X khác được đưa vào trong mô hình.

➤ Hệ số tương quan bội

Hệ số tương quan bội $R = \sqrt{R^2}$. Khi có một biến X trong mô hình hồi quy (khi $p - 1 = 1$), hệ số tương quan bội R bằng trị tuyệt đối của hệ số tương quan r trong tương quan đơn.

Phương trình hồi quy tổng thể với k biến độc lập có dạng như sau:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} \dots + \beta_k X_{ki} + \varepsilon_i \quad \text{Trong đó:}$$

– β_0 : là hệ số tung độ góc

- β_1 : là hệ dốc của Y theo biến X_1 và giữa các biến X_2, X_3, \dots, X_k không đổi.
- β_2 : là hệ dốc của Y theo biến X_2 và giữa các biến X_1, X_3, \dots, X_k không đổi.
- β_3 : là hệ dốc của Y theo biến X_3 và giữa các biến X_1, X_2, \dots, X_k không đổi.
- ...
- β_k : là hệ dốc của Y theo biến X_k và giữa các biến X_1, X_2, \dots, X_k không đổi.
- ϵ_i : là thành phần ngẫu nhiên (yếu tố nhiễu), có kì vọng bằng 0 và phương sai không đổi σ^2 .

Giả sử có một mẫu quan sát với giá trị thực tế là $(Y_i, X_{2i}, \dots, X_{ki})$ với $(i=1,2,3,\dots,k)$. Ta sẽ sử dụng thông tin từ mẫu để xây dựng các ước lượng cho các hệ số β_j (với $j=1,2,3,\dots,k$). Từ các giá trị ước lượng này có thể viết thành hàm hồi quy mẫu như sau:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} \dots + \hat{\beta}_k X_{ki}$$

Trong đó \hat{Y}_i là giá trị ước lượng cho Y_i và sai lệch giữa hai giá trị này được gọi là phần dư

2.3. Đánh giá sự phù hợp của mô hình

Có một số phương pháp thống kê để tiến hành đánh giá sự phù hợp của mô hình là: tính toán hệ số xác định, dùng thống kê F để đánh giá mức ý nghĩa toàn diện của mô hình, tính toán sai số chuẩn của ước lượng và đánh giá ý nghĩa của từng biến độc lập.

* Tính toán hệ số xác định bội

Khi có nhiều biến độc lập trong mô hình thì R^2 vẫn được sử dụng để xác định phần biến thiên trong biến phụ thuộc và tất cả các biến độc lập trong mô hình, tuy nhiên lúc này R^2 được gọi là hệ số xác định bội, công thức tính toán hệ số xác định bội như sau:

$$R^2 = \frac{SSR}{SST}$$

Cụ thể trong trường hợp khi $R^2 = 0,82$ thì ta có thể kết luận rằng 82% biến thiên trong giá trị của biến phụ thuộc có thể được giải thích bởi mối liên hệ tuyến tính giữa biến phụ thuộc với các biến độc lập trong mô hình, tuy nhiên chú ý rằng không phải tất cả các biến độc lập này đều có tầm quan trọng ngang nhau đối với khả năng giải thích cho biến thiên trong biến phụ thuộc của mô hình.

Sự gia tăng trong R^2 có thể không bù đắp được thiệt hại do mất thêm bậc tự do khi thêm biến, thế nhưng R^2_{adj} có tính đến chi phí này và điều chỉnh giá trị R^2_{adj} theo nó một cách phù hợp.

Khi một biến độc lập được thêm vào không có đóng góp xứng đáng vào khả năng giải thích cho biến phụ thuộc thì R_{adj}^2 sẽ luôn luôn giảm đi mặc dù R^2 thì tăng.

Điều đó cho thấy với mô hình hồi quy đa biến, nhất là khi số biến độc lập khá lớn trong tương quan với cỡ mẫu thì ta nên dùng R_{adj}^2 để đánh giá khả năng giải thích của mô hình. Vì vậy thông thường khi đánh giá độ phù hợp của mô hình hồi quy bội, bên cạnh thông tin về R_{adj}^2 người ta cũng dùng thêm thông tin về R_{adj}^2 để tham khảo.

Đánh giá ý nghĩa toàn diện của mô hình

Mô hình hồi quy mà chúng ta xây dựng là dựa trên dữ liệu của một mẫu lấy từ tổng thể vì vậy nó có thể bị ảnh hưởng của sai số lấy mẫu, vì thế chúng ta cần kiểm định ý nghĩa thống kê của toàn bộ mô hình.

Chúng ta có thể dựng một giả thuyết như sau:

- $H_0: \beta_1 = \beta_2 = \dots = \beta_k$ (hay $R^2 = 0$)
- $H_1: \nexists \beta_j \neq 0$ (hay $R^2 \neq 0$)

Nếu giả thuyết H_0 trên đúng nghĩa là tất cả các hệ số độ dốc đều đồng thời bằng 0 thì mô hình hồi quy đa biến xây dựng không hề có tác dụng trong việc dự đoán hay mô tả về biến phụ thuộc.

Đại lượng F thống kê (trong bảng ANOVA) chính là con số thống kê được sử dụng để kiểm định giả thuyết về ý nghĩa toàn diện của mô hình hồi quy, công thức của đại lượng F được hình thành như sau:

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n - k - 1}}$$

Trong đó SSR là tổng bình phương hồi quy (Regression Sum of Squares), SSE là tổng bình phương sai số (Error Sum of Squares), n và k lần lượt là cỡ mẫu và biến độc lập.

Chú ý là để quyết định ta phải tra bảng thống kê F tìm giá trị tới hạn tương ứng với mức ý nghĩa ta chọn trước. Mà muốn tra bảng F ta phải có thêm thông tin về bậc tự do ở tử số và mẫu số, ta qui ước bậc tự do của tử số k và bậc tự do của mẫu số là $(n - k - 1)$.

Từ đây, ta có quy trình đánh giá ý nghĩa toàn diện của mô hình như sau:

Bước 1 : Đặt giả thuyết:

- $H_0: \beta_1 = \beta_2 = \dots = \beta_k$

$$- H_1: \nexists \beta_j \neq 0$$

Bước 2 : Chọn độ tin cậy cho kiểm định từ đó có mức ý nghĩa α .

Bước 3 : Với bậc tự do xác định như trên, tra bảng phân phối F ta được giá trị F tới hạn.

Bước 4 : So sánh giá trị F kiểm định tính được theo công thức trên và giá trị F tới hạn.

Bước 5 : Kết luận.

Nếu F kiểm định > F tới hạn, ta có thể kết luận rằng mô hình hồi quy bội với các biến độc lập ta đưa vào có thể giải thích một cách có ý nghĩa cho biến thiên giá trị của biến phụ thuộc.

Tính toán sai số chuẩn ước lượng :

Mục tiêu của việc xây dựng mô hình hồi quy là để có thể xác định được giá trị của biến phụ thuộc khi biết trước các giá trị cụ thể của biến độc lập. Một số thống kê cho thấy mô hình hồi quy thực hiện mục tiêu này tốt đến đâu là lệch chuẩn của mô hình hồi quy (còn gọi tên là Sai số chuẩn ước lượng). Giá trị ước lượng từ thông tin mẫu của độ lệch chuẩn của mô hình hồi quy (sai số chuẩn ước lượng) được tính toán như sau đây:

$$s_{Y/X} = \sqrt{\frac{SSE}{n - k - 1}}$$

Trong đó n là cỡ mẫu, k là biến độc lập trong mô hình. Sai số chuẩn ước lượng đo lường sự phân tán của các giá trị thực tế đo lường được của biến phụ thuộc quanh những giá trị của biến phụ thuộc được dự đoán bằng đường hồi quy.

Đánh giá ý nghĩa của của từng biến độc lập riêng biệt :

Ở kiểm định F, giả sử H_1 được chấp nhận ta kết luận rằng mô hình toàn diện có ý nghĩa. Điều này có ý nghĩa là có ít nhất một biến độc lập trong mô hình có thể giải thích được một cách có ý nghĩa cho biến thiên phụ thuộc. Tuy nhiên điều này không có ý nghĩa là tất cả các biến độc lập đưa vào mô hình đều có ý nghĩa, để xác định biến độc lập nào có ý nghĩa chúng ta kiểm định giả thuyết sau:

$$- H_0: \beta = 0$$

$$- H_1: \beta_j = 0$$

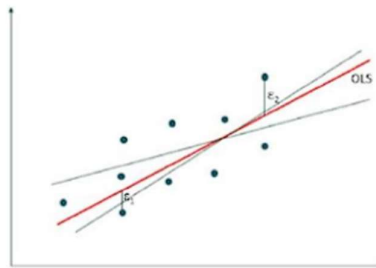
Chúng ta có thể dùng kiểm định t để kiểm định nghĩa của mỗi hệ số hồi quy với độ tin cậy được chọn trước, t được xác định bằng công thức:

$$t = \frac{b_j - 0}{s_{b_j}}$$

Trong đó β_j là hệ số dốc trong mô hình hồi quy mẫu cho biến độc lập thứ j, s_{b_j} là sai số chuẩn ước lượng lượng của hệ số độ dốc của biến độc lập thứ j. Giá trị t tính toán được sẽ được so sánh với giá trị t tới hạn tra từ bảng phân phối student với $(n - k - 1)$ bậc tự do và mức ý nghĩa $\frac{\alpha}{2}$.

2.4. Phương pháp bình phương nhỏ nhất

Phương pháp bình phương nhỏ nhất được đưa ra bởi nhà Toán học Đức Carl Friedrich Gauss - đây là một trong các phương pháp ước lượng hồi quy tuyến tính phổ biến nhất. Với tổng thể, sai số (error) kí hiệu là e, còn trong mẫu nghiên cứu lúc này được gọi là phần dư và được kí hiệu là ε . Biến thiên phần dư được tính bằng tổng bình phương tất cả các phần dư cộng lại. Nguyên tắc của phương pháp hồi quy OLS là làm cho biến thiên phần dư này trong phép hồi quy là nhỏ nhất. Khi biểu diễn trên mặt phẳng Oxy, đường hồi quy là đường thẳng đi qua đám đông các điểm dữ liệu mà ở đó, khoảng cách từ các điểm dữ liệu (tuyệt đối của ε đến đường hồi quy là ngắn nhất).



Từ đồ thị scatter biểu diễn mối quan hệ giữa các biến độc lập và biến phụ thuộc, các điểm dữ liệu sẽ nằm phân tán nhưng có xu hướng chung tạo thành một đường thẳng. Chúng ta có thể có rất nhiều đường thẳng hồi quy đi qua đám đông các điểm dữ liệu này chứ không phải chỉ một đường duy nhất, vấn đề là ta phải chọn ra đường thẳng nào mô tả sát nhất xu hướng dữ liệu. Bình phương nhỏ nhất OLS sẽ tìm ra đường thẳng đó dựa trên nguyên tắc cực tiểu hóa khoảng cách từ các điểm dữ liệu đến đường thẳng. Trong hình ở trên đường màu đỏ là đường hồi quy OLS.

2.5. Các giả định của mô hình hồi quy

a. Hàm hồi quy là tuyến tính theo các tham số.

Điều này có nghĩa là quá trình thực hành hồi quy trên thực tế được miêu tả bởi mối quan hệ dưới dạng:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \dots + \beta_k x_k + \varepsilon$$

hoặc mối quan hệ thực tế có thể được viết lại ví dụ như dưới dạng lấy loga cả hai vế

b. $E(\varepsilon_i) = 0$: Kỳ vọng của các yếu tố ngẫu nhiên bằng 0.

Trung bình tổng thể sai số là bằng 0. Điều này có nghĩa là có một số giá trị sai số mang dấu dương và một số mang dấu âm. Do hàm xem như là đường trung bình nên có thể giả định rằng các sai số ngẫu nhiên trên sẽ bị loại trừ nhau, ở mức trung bình trong tổng thể.

c. $\text{Cov}(\varepsilon_i, x_i) = 0$: Không có sự tương quan giữa các ε_i .

Không có sự tương quan giữa các quan sát của yếu tố sai số. Nếu ta xem xét các chuỗi số liệu thời gian (dữ liệu được thu nhập từ một nguồn trong nhiều khoảng thời gian khác nhau), yếu tố sai số ε_i trong khoảng thời gian này không có bất kỳ một tương quan nào với yếu tố sai số trong khoảng thời gian trước đó.

d. $\text{Cov}(\varepsilon_i, x_i) = 0$: ε và X không có sự tương quan với nhau.

Khi bất kỳ biến giải thích nào lớn hơn hay nhỏ đi thì yếu tố sai số sẽ không thay đổi theo nó.

e. $\text{Var}(\varepsilon_i) = \sigma^2$: Phương sai bằng nhau và thuần nhất với mọi ε_i .

Tất cả giá trị ε_i được phân phối giống nhau với cùng σ^2 , sao cho: $\text{Var}(\varepsilon_i) = E(\varepsilon^2) = \sigma^2$.

f. ε_i phân phối chuẩn.

Điều này rất quan trọng khi phát sinh khoảng tin cậy và thực hiện kiểm định giả thuyết trong những phạm vi mẫu là nhỏ. Nhưng phạm vi mẫu lớn hơn, điều này trở nên không mấy quan trọng.

g. Giữa các x^2, x^3, \dots, x^k không có quan hệ tuyến tính.

Nếu x^2, x^3, \dots, x^k có quan hệ tuyến tính thì người ta nói rằng có hiện tượng đa cộng tuyến.

Hay không tồn tại $\lambda_i \equiv 0$: $\lambda_1 x_{1i} + \lambda_2 x_{2i} + \lambda_3 x_{3i} + \dots + \lambda_k x_{ki} + v_i = 0$

2.6. Lý thuyết về hồi quy binary logistic

Hồi quy Binary Logistic là mô hình phổ biến trong nghiên cứu dùng để ước

lượng xác suất một sự kiện sẽ xảy ra. Đặc trưng của hồi quy nhị phân là biến phụ

thuộc chỉ có hai giá trị: 0 và 1. Trên thực tế, có rất nhiều hiện tượng tự nhiên, hiện tượng kinh tế, xã hội, ... mà chúng ta cần dự đoán khả năng xảy ra của nó như chiến dịch quảng cáo có được chấp nhận hay không, người vay có trả được nợ hay không, công ty có phá sản hay không, khách

hàng có mua hay không,... Những biến nghiên cứu có hai biểu hiện như vậy được mã hóa thành hai giá trị 0 và 1, được gọi là biến nhị phân.

Khi biến phụ thuộc ở dạng nhị phân, chúng ta không thể phân tích với dạng hồi quy tuyến tính thông thường vì mô hình sẽ vi phạm các giả định hồi quy ...

Các giả định quan trọng này bị vi phạm sẽ làm mất hiệu lực thống kê của các kiểm định trong hồi quy, dẫn đến kết quả ước lượng không còn chính xác. Trong khi đó, hồi quy Binary Logistic lại không cần thiết phải thỏa mãn các giả định này

2.7. Phương trình hồi quy binary logistic

Thay vì chúng ta ước lượng giá trị của biến phụ thuộc Y theo biến độc lập X như ở hồi quy đa biến, thì trong hồi quy Binary Logistic, chúng ta sẽ ước lượng xác suất xảy ra sự kiện Y (probability) khi biết giá trị X. Biến phụ thuộc Y có hai giá trị 0 và 1, với 0 là không xảy ra sự kiện và 1 là xảy ra sự kiện. Từ đặc điểm này, chúng ta có thể đánh giá được khả năng xảy ra sự kiện ($Y = 1$) nếu xác suất dự đoán lớn hơn 0.5, ngược lại, khả năng không xảy ra sự kiện ($Y = 0$) nếu xác suất dự đoán nhỏ hơn 0.5. Ta có hàm xác suất như sau:

$$P_i = P(Y = 1) = E(Y = 1 / X) = \frac{e^{(B_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n)}}{1 + e^{(B_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n)}}$$

Trong đó $P_i = P(Y = 1) = E(Y = 1 / X)$ là xác suất xảy ra sự kiện. Thực hiện các phép chuyển đổi toán học, chúng ta thu được phương trình hồi quy Binary Logistic như sau:

$$\log_e \left[\frac{P_i}{1 - P_i} \right] = B_0 + B_1X_1 + B_2X_2 + \dots + B_nX_n$$

Trong đó:

- P_i : xác suất xảy ra sự kiện ($Y = 1$)
- $1 - P_i$: xác suất không xảy ra sự kiện ($Y = 0$)
- B_0 : hằng số hồi quy
- B_1, B_2, \dots, B_n : hệ số hồi quy

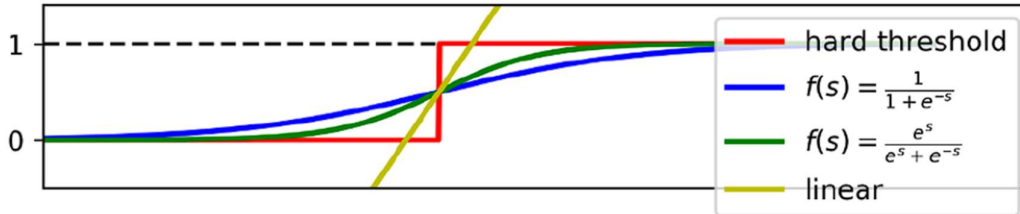
2.8. Mô hình Logistic Regression

Đầu ra dự đoán của:

- Linear Regression: $f(x) = wTx$
- PLA: $f(x) = \text{sgn}(wTx)$

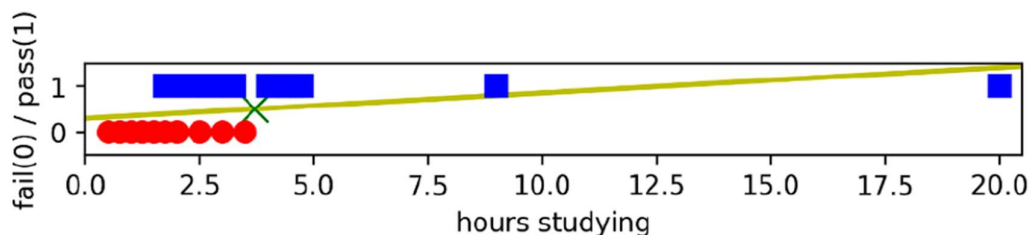
Đầu ra dự đoán của logistic regression thường được viết chung dưới dạng: $f(x) = \theta(wTx)$

Trong đó θ được gọi là logistic function. Một số activation cho mô hình tuyến tính được cho trong hình dưới đây:



Các activation function khác nhau.

Đường màu vàng biểu diễn linear regression. Đường này không bị chặn nên không phù hợp cho bài toán này. Có một trick nhỏ để đưa nó về dạng bị chặn: cắt phần nhỏ hơn 0 bằng cách cho chúng bằng 0, cắt các phần lớn hơn 1 bằng cách cho chúng bằng 1. Sau đó lấy điểm trên đường thẳng này có tung độ bằng 0.5 làm điểm phân chia hai class, đây cũng không phải là một lựa chọn tốt. Giả sử có thêm vài bạn sinh viên tiêu biểu ôn tập đến 20 giờ và, tất nhiên, thi đỗ. Khi áp dụng mô hình linear regression như hình dưới đây và lấy mốc 0.5 để phân lớp, toàn bộ sinh viên thi trượt vẫn được dự đoán là trượt, nhưng rất nhiều sinh viên thi đỗ cũng được dự đoán là trượt (nếu ta coi điểm x màu xanh lục là ngưỡng cứng để đưa ra kết luận). Rõ ràng đây là một mô hình không tốt. Anh chàng sinh viên tiêu biểu này đã kéo theo rất nhiều bạn khác bị trượt.



Tại sao Linear Regression không phù hợp?

- Đường màu đỏ (chỉ khác với activation function của PLA ở chỗ hai class là 0 và 1 thay vì -1 và 1) cũng thuộc dạng ngưỡng cứng (hard threshold). PLA không hoạt động trong bài toán này vì dữ liệu đã cho không linearly separable.

- Các đường màu xanh lam và xanh lục phù hợp với bài toán của chúng ta hơn. Chúng có một vài tính chất quan trọng sau:
 - Là hàm số liên tục nhận giá trị thực, bị chặn trong khoảng (0,1)(0,1).
 - Nếu coi điểm có tung độ là 1/2 làm điểm phân chia thì các điểm càng xa điểm này về phía bên trái có giá trị càng gần 0. Ngược lại, các điểm càng xa điểm này về phía phải có giá trị càng gần 1. Điều này khớp với nhận xét rằng học càng nhiều thì xác suất đỗ càng cao và ngược lại.
 - Mượt (smooth) nên có đạo hàm mọi nơi, có thể được lợi trong việc tối ưu.

3. Tiền xử lý số liệu:

3.1 Đọc dữ liệu:

```
setwd("C:/")
```

```
df <- read.csv("LopL10_Nhom13.csv")
```

```
df
```

	order_id	customer_id	date	nearest_warehouse	shopping_cart	order_price	delivery_charges	customer_lat	customer_long
1	ORD182494	ID6197211592	6/22/2019	Thompson	[(('Lucent 330S', 1), ('Thunder line', 2), ('iStream', 2), ('pearTV', ...	12200	79.89	-37.81511	144.9328
2	ORD395518	ID0282825849	12/29/2019	Thompson	[(('Thunder line', 1), ('Universe Note', 2)]	9080	62.71	-37.80274	144.9511
3	ORD494479	ID0579391891	3/2/2019	Nickolson	[(('Thunder line', 1), ('pearTV', 2)]	10670	65.87	-37.82130	144.9576
4	ORD019224	ID4544561904	1/12/2019	Nickolson	[(('Universe Note', 1), ('Alcon 10', 2), ('Olivia x460', 1), ('iAssist ...	24800	57.61	-37.81142	144.9731
5	ORD104032	ID6231506320	11/28/2019	Nickolson	[(('Universe Note', 1), ('Olivia x460', 1), ('iStream', 1), ('Toshika ...	9145	75.54	37.82386	144.9699
6	ORD146760	ID0311654900	9/16/2019	Bakers	[(('Thunder line', 2), ('Universe Note', 1)]	7810	71.22	37.82025	145.0149
7	ORD337984	ID3394768956	9/14/2019	Thompson	[(('Candle Inferno', 1), ('Alcon 10', 1), ('Toshika 750', 1)]	13700	74.84	-37.80774	144.9516
8	ORD072312	ID0774517121	5/23/2019	Thompson	[(('Universe Note', 1), ('Thunder line', 2), ('iStream', 1)]	7960	52.28	-37.80634	144.9595
9	ORD377837	ID4769265355	10/9/2019	Bakers	[(('Alcon 10', 2), ('Thunder line', 1), ('Candle Inferno', 2), ('iAssi...	25390	107.58	-37.81081	145.0141
10	ORD462194	ID5301568579	3/21/2019	Thompson	[(('Universe Note', 1), ('Lucent 330S', 1), ('Toshika 750', 2)]	13320	62.26	-37.80868	144.9423
11	ORD034800	ID4283908179	8/3/2019	Bakers	[(('Alcon 10', 2), ('pearTV', 2), ('iStream', 1), ('Olivia x460', 1)]	31895	78.25	-37.81133	145.0087
12	ORD361636	ID0589500304	12/5/2019	Nickolson	[(('Lucent 330S', 1), ('pearTV', 2)]	13850	77.29	-37.82023	144.9574
13	ORD124395	ID0702352304	2/11/2019	Thompson	[(('Alcon 10', 1), ('Universe Note', 1), ('pearTV', 1), ('iStream', 2)]	19010	94.75	-37.80543	144.9413
14	ORD255642	ID3085953531	12/24/2019	Nickolson	[(('iAssist Line', 2), ('Alcon 10', 1), ('pearTV', 1)]	19710	75.64	-37.81617	144.9753
15	ORD496722	ID0589449820	4/9/2019	Nickolson	[(('pearTV', 2), ('iStream', 1), ('Lucent 330S', 1), ('Alcon 10', 2)]	31900	79.78	-37.80946	144.9724
16	ORD449130	ID0356449717	5/17/2019	Bakers	[(('Toshika 750', 2), ('Alcon 10', 1), ('Thunder line', 1), ('Candle ...	20200	46.35	-37.80813	144.9868

	coupon_discount	order_total	season	is_expedited_delivery	distance_to_nearest_warehouse	latest_customer_review	is_happy_customer
	10	11059.89	Winter	1	1.2800	perfect phone and trusted seller. phone itself is amazing. i g...	1
	0	9142.71	Summer	0	1.1621	it keeps dropping calls the wifi don't work this is a waste of ...	0
	10	9668.87	Autumn	0	1.0949	five stars this is a great cheap phone.	1
	15	21137.61	Summer	0	0.8571	charger did not fit the charger didn't fit.	0
	25	6934.29	Spring	0	0.5867	four stars good	1
	10	7100.22	Spring	0	2.0752	stolen phone sold us a stolen phone so we couldn't activate...	0
	5	13089.84	Spring	0	0.6767	love our inferno stick.easy to set up and have loads of show...	1
	5	10789.79	Autumn	0	1.3043	it sucks mine came with dead pixels	0
	10	22958.58	Spring	1	1.6595	this is how top phone should look like! super fast phone. ov...	1
	15	11384.26	winter	1	0.6093	does not live up to its reputation. customer service at olivia l...	0
	0	31973.25	Winter	1	1.1919	i love this phone it is so user friendly and the battery life is a...	1
	25	10464.79	Summer	0	1.0829	great phone great deall phone in fantastic condition 3 mont...	1
	0	926057.25	Summer	1	0.9509	the cult's alive i love this little dummy-phone. the standby ti...	1
	0	19785.64	Summer	1	0.5716	phone had a problem phone seemed great but constantly h...	0
	0	31979.78	Autumn	1	66.6483	five stars good speed. like stanley	1
	25	15196.35	Autumn	0	0.7706	junk that what receive in the mail .	0

3.2. Làm sạch dữ liệu (Data cleaning)

Từ dữ liệu trong df , trích ra một dữ liệu con bao gồm các biến chính của

đề bài và đặt tên là "new_df"

```
new_df <- df[,c("is_expedited_delivery", "order_price", "delivery_charges", "coupon_discount",  
"order_total", "distance_to_nearest_warehouse", "is_happy_customer")]
```

	is_expedited_delivery	order_price	delivery_charges	coupon_discount	order_total	distance_to_nearest_warehouse	is_happy_customer
1	1	12200	79.89	10	11059.89	1.2800	1
2	0	9080	62.71	0	9142.71	1.1621	0
3	0	10670	65.87	10	9668.87	1.0949	1
4	0	24800	57.61	15	21137.61	0.8571	0
5	0	9145	75.54	25	6934.29	0.5867	1
6	0	7810	71.22	10	7100.22	2.0752	0
7	0	13700	74.84	5	13089.84	0.6767	1
8	0	7960	52.28	5	10789.79	1.3043	0
9	1	25390	107.58	10	22958.58	1.6595	1
10	1	13320	62.26	15	11384.26	0.6093	0
11	1	31895	78.25	0	31973.25	1.1919	1
12	0	13850	77.29	25	10464.79	1.0829	1
13	1	19010	94.75	0	926057.25	0.9509	1
14	1	19710	75.64	0	19785.64	0.5716	0
15	1	31900	79.78	0	31979.78	66.6483	1
16	0	20200	46.35	25	15196.35	0.7706	0
17	1	14810	80.69	25	11188.19	1.6104	1

3.3. Xử lý dữ liệu khuyết:

```
apply(is.na(new_df),2,which)
```

```
> apply(is.na(new_df),2,which)  
integer(0)
```

=> Ta thấy trong dữ liệu "new_df" không dữ liệu nào bị khuyết

3.4. Làm rõ dữ liệu:

Tạo 1 data mới có tên "new_df2" (gồm các biến như trong "new_df" đã làm sạch dữ liệu) và chuyển đổi các biến is_expedited_delivery, order_price, delivery_charges, coupon_discount, order_total, distance_to_nearest_warehouse, lần lượt thành log(is_expedited_delivery +1), log(order_price+1), log(delivery_charges+1), log(coupon_discount+1), log(order_total+1), log(distance_to_nearest_warehouse+1)

```
new_df2 <- new_df
```

```
new_df2[,c("is_expedited_delivery", "order_price", "delivery_charges", "coupon_discount", "order_total", "distance_to_nearest_warehouse")] <- log(new_df2[,c("is_expedited_delivery", "order_price", "delivery_charges", "coupon_discount", "order_total", "distance_to_nearest_warehouse")]+1)
```

	is_expedited_delivery	order_price	delivery_charges	coupon_discount	order_total	distance_to_nearest_warehouse	is_happy_customer
1	0.6931472	9.409273	4.393090	2.397895	9.311171	0.8241754	1
2	0.0000000	9.113940	4.154342	0.000000	9.120821	0.7710800	0
3	0.0000000	9.275285	4.202750	2.397895	9.176770	0.7395058	1
4	0.0000000	10.118639	4.070905	2.772589	9.958857	0.6190161	0
5	0.0000000	9.121072	4.337813	3.258097	8.844378	0.4616564	1
6	0.0000000	8.963288	4.279717	2.397895	8.868022	1.1233699	0
7	0.0000000	9.525224	4.328626	1.791759	9.479668	0.5168276	1
8	0.0000000	8.982310	3.975561	1.791759	9.286448	0.8347769	0
9	0.6931472	10.142150	4.687487	2.397895	10.041491	0.9781381	1
10	0.6931472	9.497097	4.147253	2.772589	9.340075	0.4757993	0
11	0.6931472	10.370236	4.372607	0.000000	10.372686	0.7847687	1
12	0.0000000	9.536113	4.360420	3.258097	9.255867	0.7337612	1
13	0.6931472	9.852773	4.561741	0.000000	13.738692	0.6682908	1
14	0.6931472	9.888932	4.339119	0.000000	9.892762	0.4520942	0
15	0.6931472	10.370393	4.391729	0.000000	10.372890	4.2143222	1
16	0.0000000	9.913487	3.857567	3.258097	9.628876	0.5713185	0
17	0.6931472	9.603125	4.102932	3.258097	9.322703	0.9595035	1

Giải thích lý do chuyển sang dạng $\log(x+1)$: + Cải thiện sự phù hợp của mô hình : giả định khi ta xây dựng mô hình hồi quy thì các sai số hồi quy (phần dư) phải có phân phối chuẩn, do đó trong trường hợp sai số hồi quy (phần dư) không có phân phối chuẩn thì việc lấy log của một số biến giúp thay đổi tỉ lệ và làm cho biến đó có phân phối chuẩn. Ngoài ra, trong trường hợp phần dư (phương sai thay đổi) do các biến độc lập gây ra, ta cũng có thể chuyển đổi các biến đó sang dạng log. + Diễn giải: đây là lý do giúp ta có thể diễn giải mối quan hệ giữa 2 biến thuận tiện hơn. Nếu ta lấy log của biến phụ thuộc Y và biến độc lập X, khi đó hệ số hồi quy β sẽ là hệ số co giãn và diễn giải sẽ như sau: X tăng 1% sẽ dẫn đến tăng việc kỳ vọng Y tăng lên β % (về mặt trung bình của Y). + Ước lượng mô hình phi tuyến: việc lấy log cho phép ta ước lượng các mô hình này bằng hồi quy tuyến tính. + Ngoài ra, việc chuyển sang dạng $\log(x+1)$ thay vì $\log(x)$ bởi do biến sqft_basement có nhiều giá trị bằng 0 (do một số ngôi nhà không có tầng hầm). Nếu chuyển sang dạng log thì sẽ nhận được các giá trị infinity. Do đó ta sẽ chuyển sang $\log(x+1)$ thay vì $\log(x)$.

4. Thống kê mô tả:

Tính các giá trị thống kê mô tả (trung bình, độ lệch chuẩn, min, max, trung vị) cho các biến, xuất dưới dạng bảng:

```
mean<-apply(new_df,2,mean)
```

```
sd<-apply(new_df,2,sd)
```

```
median<-apply(new_df,2,median)
```

```
Q1<-apply(new_df,2,quantile,probs=0.25)
```

```
Q3<-apply(new_df,2,quantile,probs=0.75)
```

```
max<-apply(new_df,2,max)
min<-apply(new_df,2,min)
otput<-cbind(mean,median,sd,min,max,Q1,Q3)
```

	mean	median	sd	min	max	Q1	Q3
is_expedited_delivery	0.498000	0.0000	5.004967e-01	0.0000	1.00000e+00	0.000000	1.000000
order_price	25522.216000	12807.5000	8.633373e+04	585.0000	9.47691e+05	7050.000000	20360.000000
delivery_charges	76.658200	76.3100	1.448146e+01	46.3500	1.14040e+02	65.982500	82.555000
coupon_discount	10.890000	10.0000	8.649134e+00	0.0000	2.50000e+01	5.000000	15.000000
order_total	39209.672180	11293.9600	2.741940e+05	639.2900	5.68827e+06	6454.735000	18119.187500
distance_to_nearest_warehouse	2.204224	1.0301	8.812416e+00	0.1078	9.49734e+01	0.751425	1.408625
is_happy_customer	0.718000	1.0000	4.504240e-01	0.0000	1.00000e+00	0.000000	1.000000

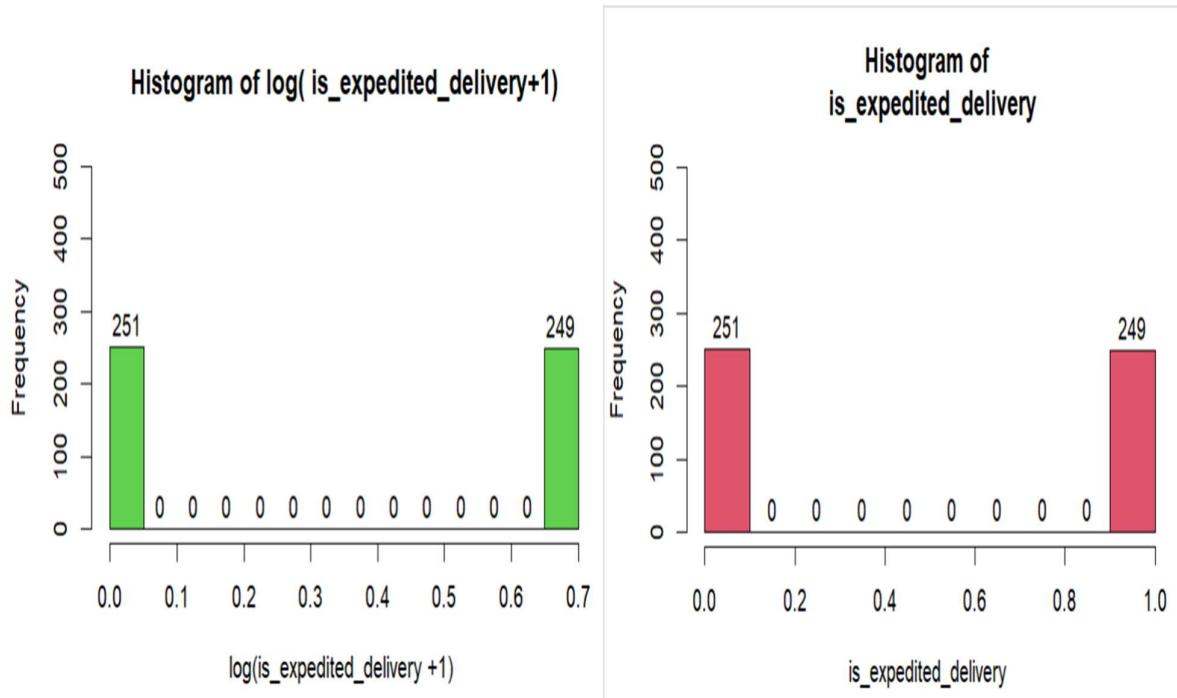
Tính các giá trị thống kê mô tả (trung bình, độ lệch chuẩn, min, max, trung vị) cho các biến sau khi đã chuyển qua dạng $\log(x+1)$

```
mean<-apply(new_df2,2,mean)
sd<-apply(new_df2,2,sd)
median<-apply(new_df2,2,median)
Q1<-apply(new_df2,2,quantile,probs=0.25)
Q3<-apply(new_df2,2,quantile,probs=0.75)
max<-apply(new_df2,2,max)
min<-apply(new_df2,2,min)
otput<-cbind(mean,median,sd,min,max,Q1,Q3)
```

	mean	median	sd	min	max	Q1	Q3
is_expedited_delivery	0.3451873	0.000000	0.3469179	0.0000000	0.6931472	0.0000000	0.6931472
order_price	9.3882814	9.457864	0.9293305	6.3733198	13.7617848	8.8609247	9.9213766
delivery_charges	4.3352537	4.347823	0.1847459	3.8575668	4.7452799	4.2044311	4.4255039
coupon_discount	2.0227016	2.397895	1.1432004	0.0000000	3.2580965	1.7917595	2.7725887
order_total	9.3173597	9.332112	1.0041486	6.4619212	15.5539168	8.7727192	9.8047767
distance_to_nearest_warehouse	0.7607225	0.708085	0.5127867	0.1023761	4.5640711	0.5604297	0.8790560
is_happy_customer	0.7180000	1.000000	0.4504240	0.0000000	1.0000000	0.0000000	1.0000000

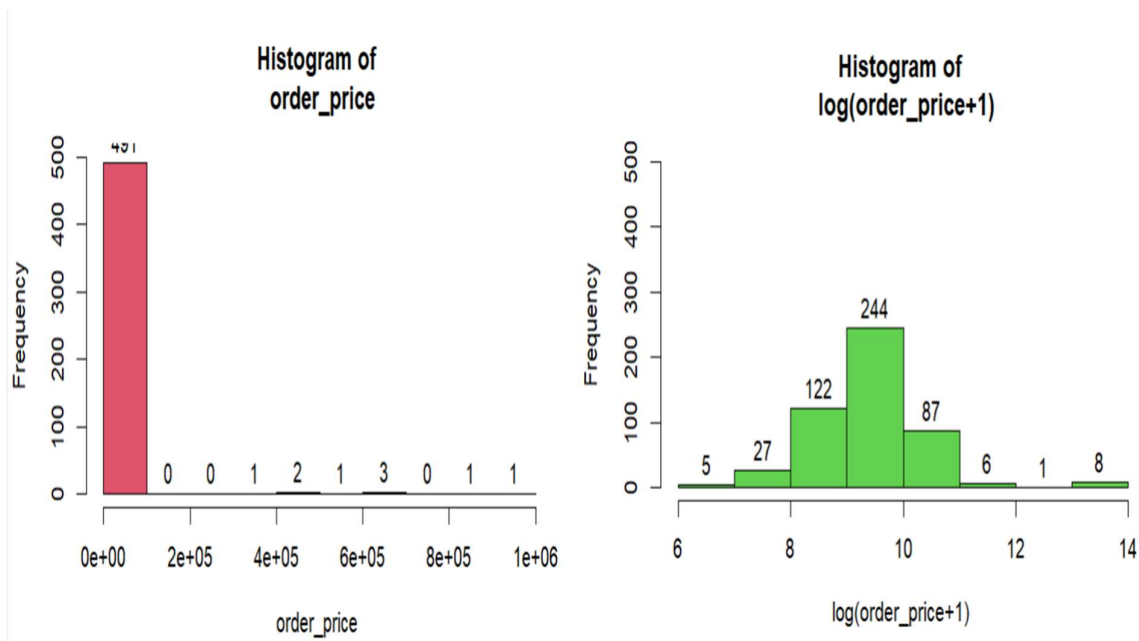
Vẽ biểu đồ Histogram thể hiện phân phối của biến `is_expedited_delivery` trước và sau khi chuyển sang dạng $\log(x+1)$

```
hist(new_df[,"is_expedited_delivery"],xlab="is_expedited_delivery ",main="Histogram of
is_expedited_delivery",ylim=c(0,500),col=2,labels=T)
hist(new_df2[,"is_expedited_delivery"],xlab="log(is_expedited_delivery +1) ",main="Histogram of log( is_expedited_delivery+1)",ylim=c(0,500),col=3,labels=T)
```



Về biểu đồ Histogram thể hiện phân phối của biến order_price trước và sau khi chuyển sang dạng $\log(x+1)$

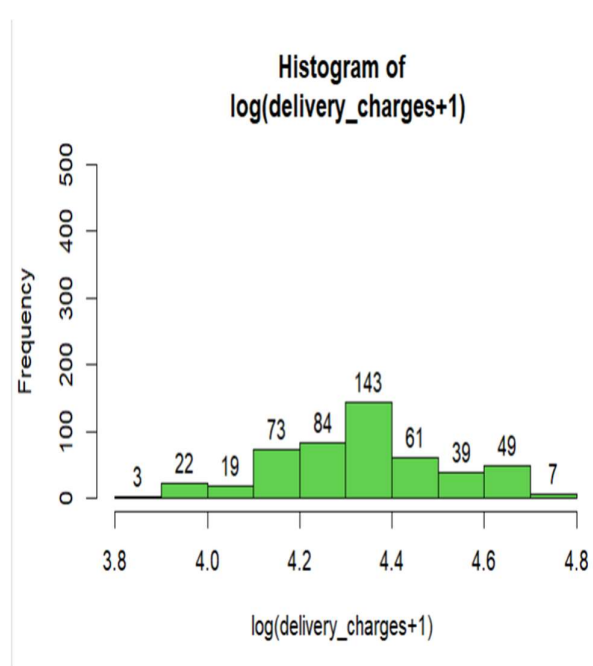
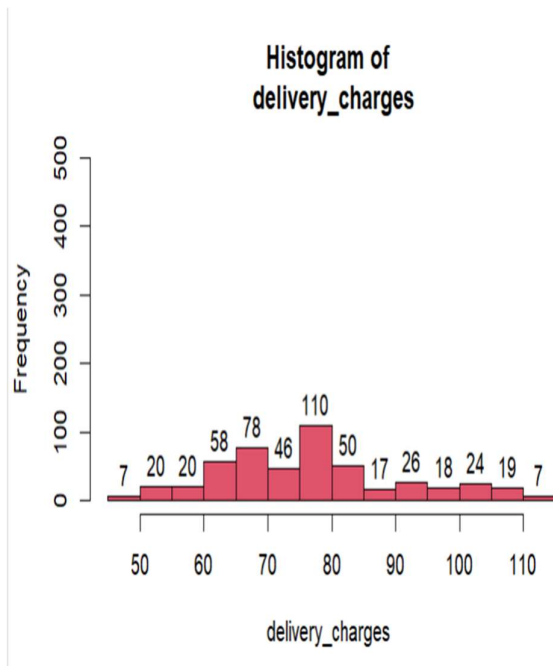
```
hist(new_df[,"order_price"],xlab="order_price ",main="Histogram of
order_price",ylim=c(0,500),col=2,labels=T)
hist(new_df2[,"order_price"],xlab="log(order_price+1) ",main="Histogram of
log(order_price+1)",ylim=c(0,500),col=3,labels=T)
```



Về biểu đồ Histogram thể hiện phân phối của delivery_charges order_price trước và sau khi chuyển sang dạng $\log(x+1)$

```
hist(new_df["delivery_charges"],xlab="delivery_charges ",main="Histogram of
delivery_charges",ylim=c(0,500),col=2,labels=T)
```

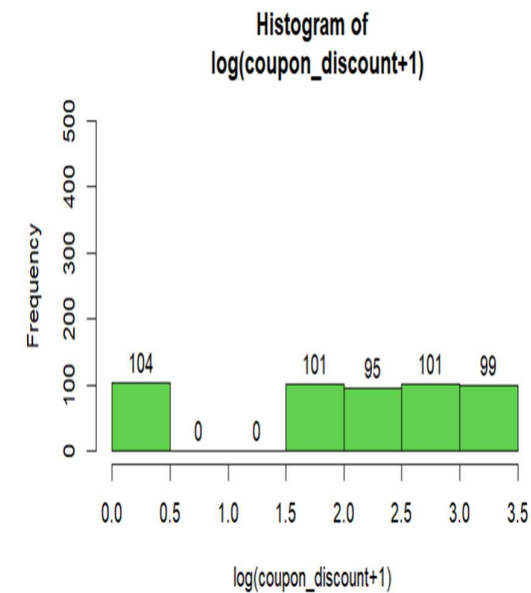
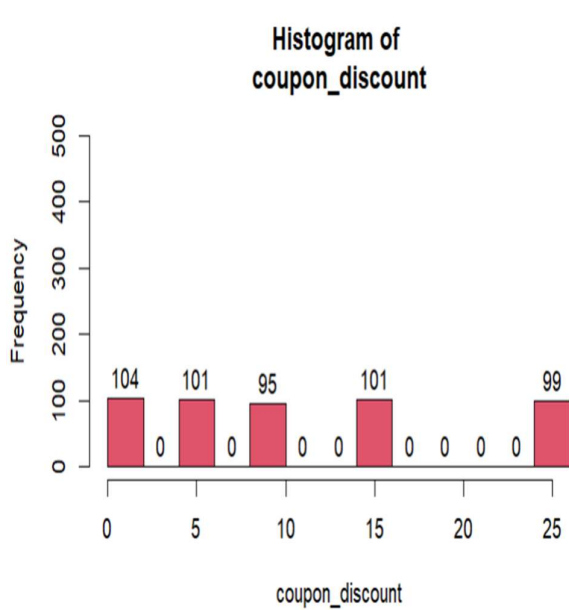
```
hist(new_df2["delivery_charges"],xlab="log(delivery_charges+1) ",main="Histogram of
log(delivery_charges+1)",ylim=c(0,500),col=3,labels=T)
```



Vẽ biểu đồ Histogram thể hiện phân phối của delivery_charges trước và sau khi chuyển sang dạng $\log(x+1)$

```
hist(new_df[,"coupon_discount"],xlab=" coupon_discount ",main="Histogram of coupon_discount",ylim=c(0,500),col=2,labels=T)
```

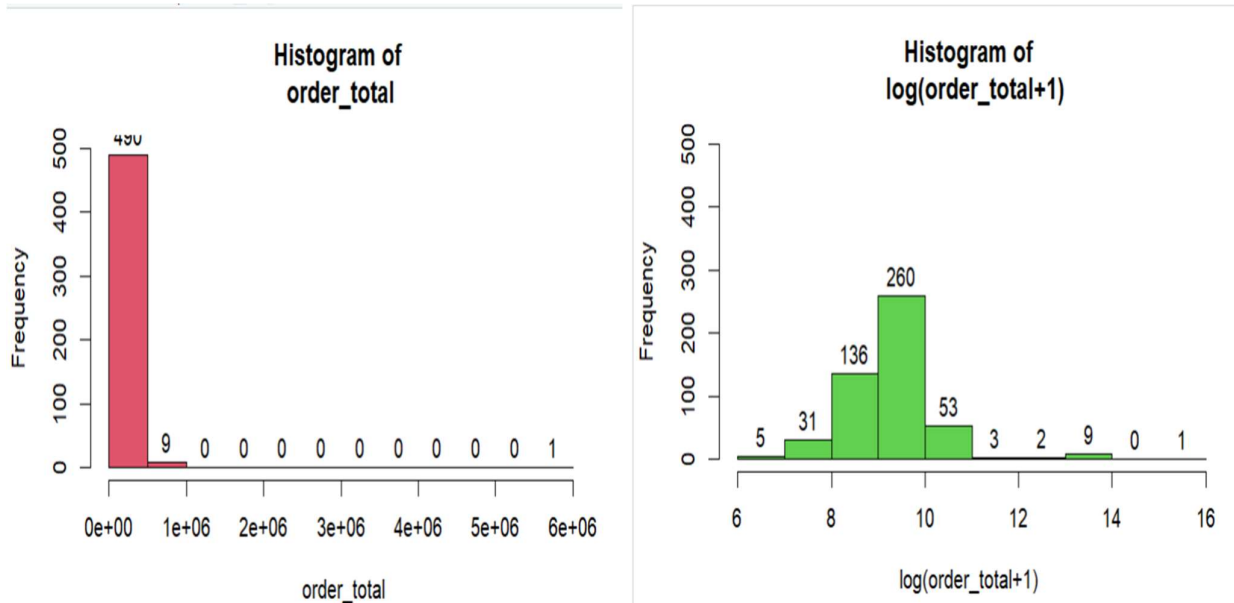
```
hist(new_df2[,"coupon_discount"],xlab="log(coupon_discount+1) ",main="Histogram of log(coupon_discount+1)",ylim=c(0,500),col=3,labels=T)
```



Vẽ biểu đồ Histogram thể hiện phân phối của `order_total` trước và sau khi chuyển sang dạng $\log(x+1)$

```
hist(new_df[,"order_total"],xlab=" order_total",main="Histogram of
order_total",ylim=c(0,500),col=2,labels=T)
```

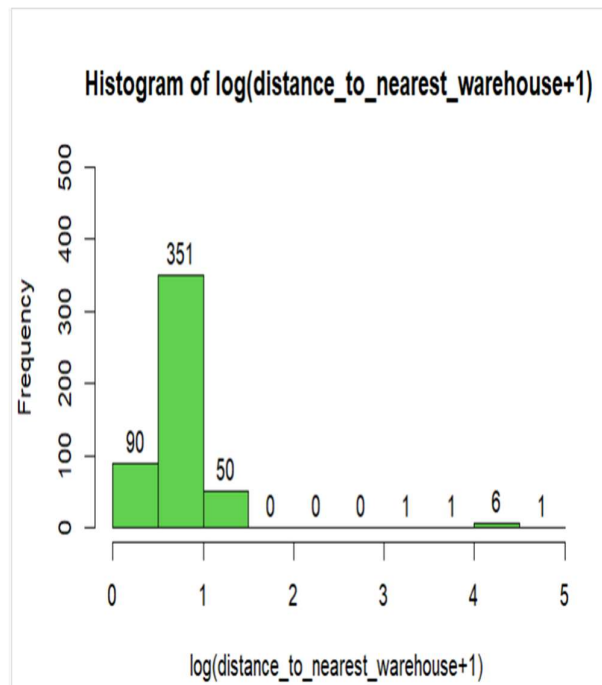
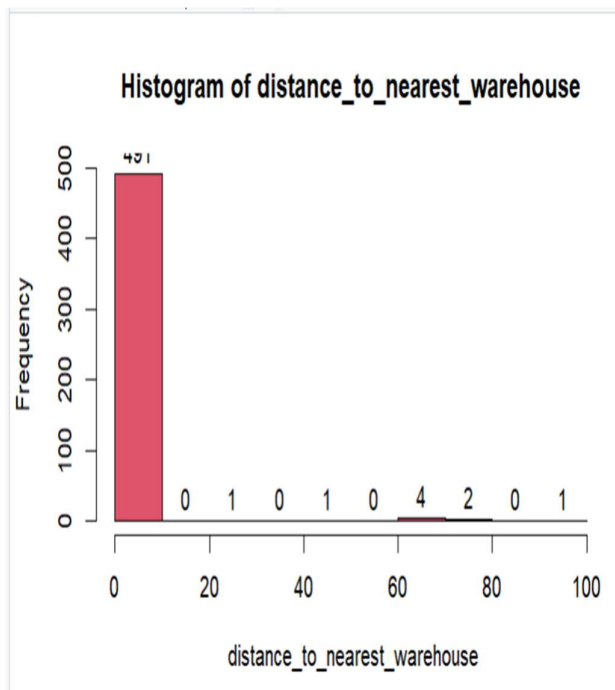
```
hist(new_df2[,"order_total"],xlab="log(order_total+1) ",main="Histogram of
log(order_total+1)",ylim=c(0,500),col=3,labels=T)
```



Vẽ biểu đồ Histogram thể hiện phân phối của `distance_to_nearest_warehouse` trước và sau khi chuyển sang dạng $\log(x+1)$

```
hist(new_df[,"distance_to_nearest_warehouse"],xlab=" distance_to_nearest_warehouse",main =
"Histogram of distance_to_nearest_warehouse",ylim=c(0,500),col=2,labels=T)
```

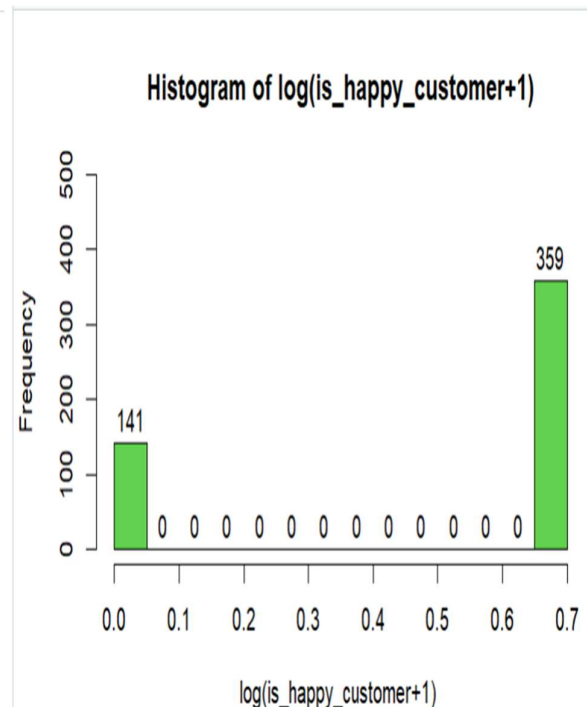
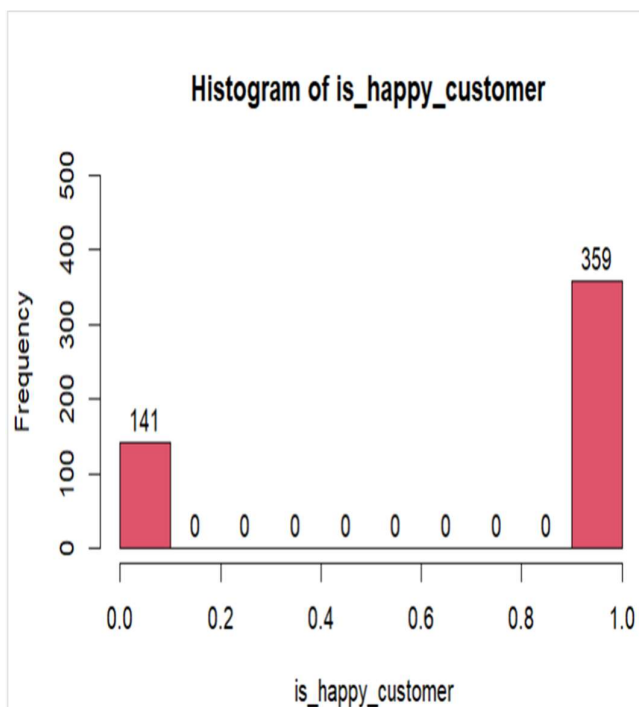
```
hist(new_df2[,"distance_to_nearest_warehouse"],xlab="log(distance_to_nearest_warehouse+1
) ",main="Histogram of log(distance_to_nearest_warehouse+1)",ylim=c(0,500),col=3,labels=T
)
```

Vẽ biểu đồ Histogram thể hiện phân phối của `is_happy_customer` trước và sau khi chuyển sang dạng $\log(x+1)$

```
hist(new_df[,"is_happy_customer"],xlab=" is_happy_customer",main="Histogram of is_happy_customer",ylim=c(0,500),col=2,labels=T)
```

```
hist(new_df2[,"is_happy_customer"],xlab="log(is_happy_customer+1) ",main="Histogram of log(is_happy_customer+1)",ylim=c(0,500),col=3,labels=T)
```



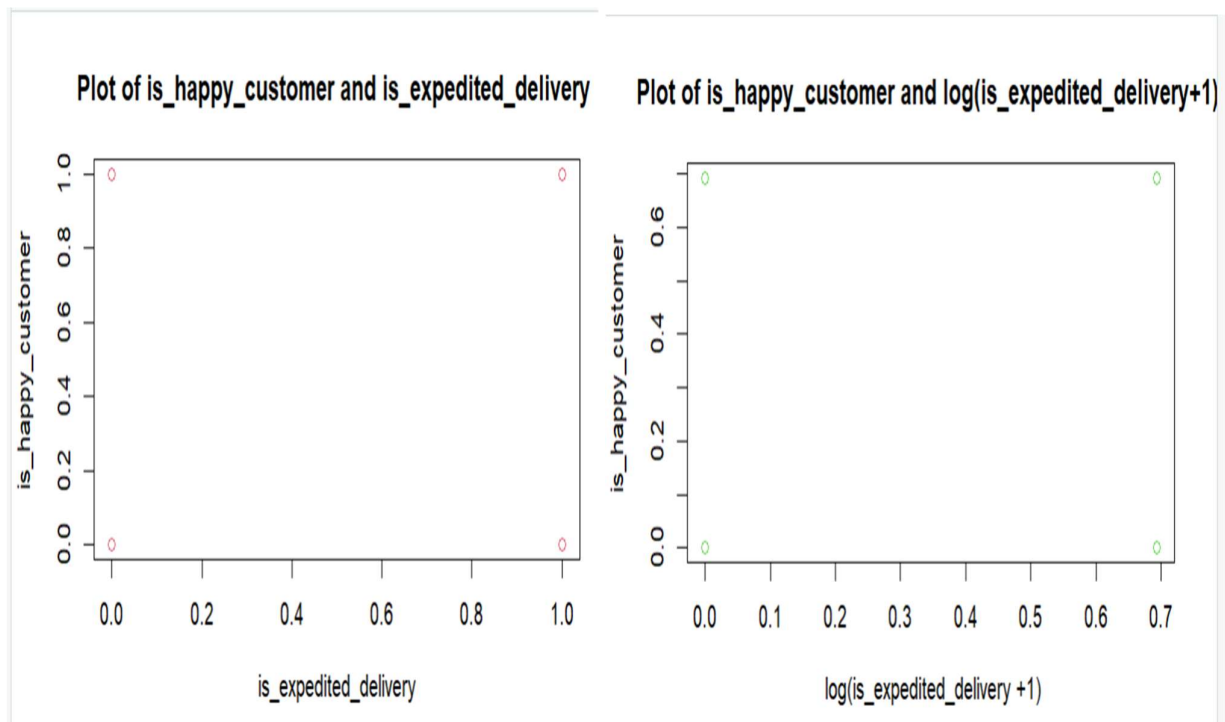
Nhận xét:

Nhìn vào biểu đồ histogram, ta thấy đa số các biến không có phân phối chuẩn do đồ thị bị lệch về một bên hoặc không có dạng hình chuông.

Vẽ biểu đồ phân tán thể hiện sự phân phối của biến `is_happy_customer` theo biến `is_expedited_delivery` trước và sau khi chuyển sang dạng $\log(x+1)$.

```
plot(is_happy_customer~ is_expedited_delivery,data=new_df,xlab=" is_expedited_delivery ",  
ylab="is_happy_customer",main="Plot of is_happy_customer and is_expedited_delivery ",col  
=2)
```

```
plot(is_happy_customer~ is_expedited_delivery,data=new_df2,xlab=" log(is_expedited_delive  
ry +1)",ylab="is_happy_customer",main="Plot of is_happy_customer and log(is_expedited_de  
livery+1) ",col=3)
```



Vẽ biểu đồ phân tán thể hiện sự phân phối của biến `is_happy_customer` theo biến `order_price` trước và sau khi chuyển sang dạng $\log(x+1)$.

```
plot(is_happy_customer~ order_price,data=new_df,xlab=" order_price",ylab="is_happy_custo  
mer",main="Plot of is_happy_customer and order_price",col=2)
```

```
plot(is_happy_customer~ order_price,data=new_df2,xlab=" log(order_price+1)",ylab="is_happy_customer",main="Plot of is_happy_customer and log(order_price+1) ",col=3)
```



Về biểu đồ phân tán thể hiện sự phân phối của biến `is_happy_customer` theo biến `delivery_charges` trước và sau khi chuyển sang dạng $\log(x+1)$.

```
plot(is_happy_customer~ delivery_charges,data=new_df,xlab=" delivery_charges",ylab="is_happy_customer",main="Plot of is_happy_customer and delivery_charges",col=2)
```

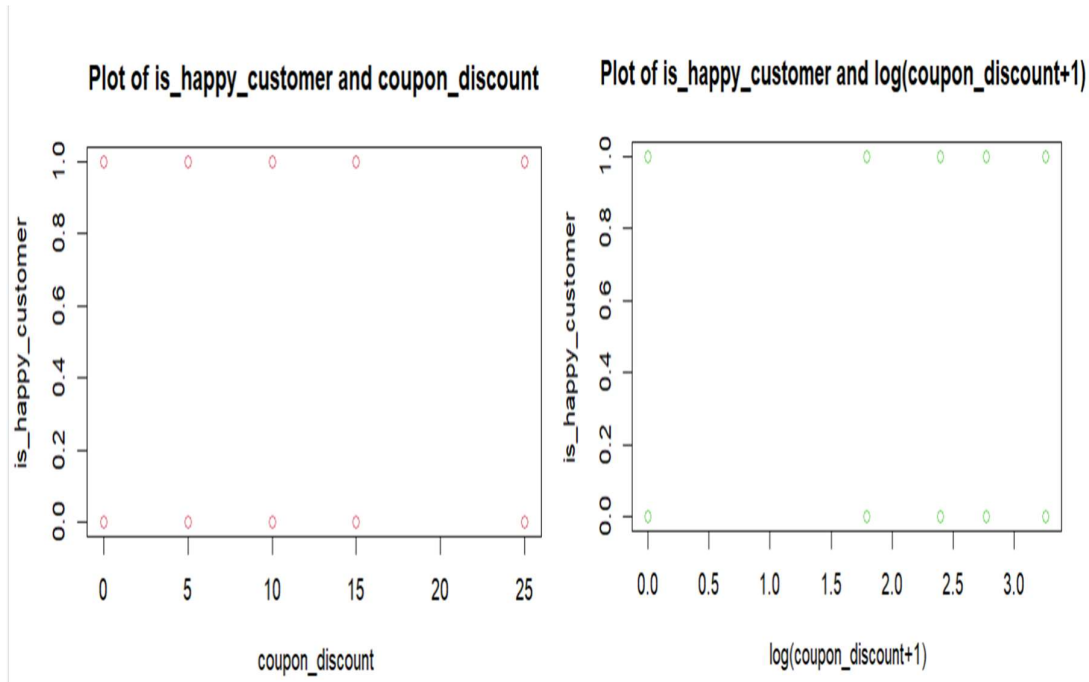
```
plot(is_happy_customer~ delivery_charges,data=new_df2,xlab=" log(delivery_charges+1)",ylab="is_happy_customer",main="Plot of is_happy_customer and log(delivery_charges+1) ",col=3)
```



Vẽ biểu đồ phân tán thể hiện sự phân phối của biến `is_happy_customer` theo biến `coupon_discount` trước và sau khi chuyển sang dạng $\log(x+1)$.

```
plot(is_happy_customer~ coupon_discount,data=new_df,xlab=" coupon_discount",ylab="is_happy_customer",main="Plot of is_happy_customer and coupon_discount",col=2)
```

```
plot(is_happy_customer~ coupon_discount,data=new_df2,xlab=" log(coupon_discount+1)",ylab="is_happy_customer",main="Plot of is_happy_customer and log(coupon_discount+1) ",col=3)
```



Vẽ biểu đồ phân tán thể hiện sự phân phối của biến `is_happy_customer` theo biến `order_total` trước và sau khi chuyển sang dạng $\log(x+1)$.

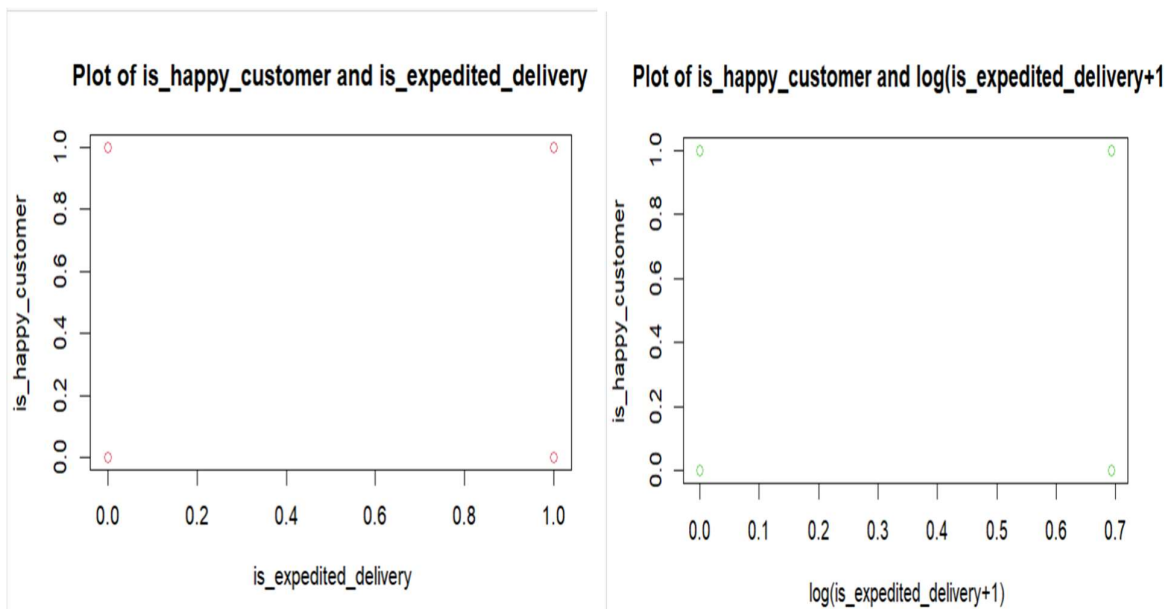
```
plot(is_happy_customer~ order_total,data=new_df,xlab=" order_total",ylab="is_happy_customer",main="Plot of is_happy_customer and order_total",col=2)
```

```
plot(is_happy_customer~ order_total,data=new_df2,xlab=" log(order_total+1)",ylab="is_happy_customer",main="Plot of is_happy_customer and log(order_total+1) ",col=3)
```



Về biểu đồ phân tán thể hiện sự phân phối của biến `is_happy_customer` theo biến `is_expedited_delivery` trước và sau khi chuyển sang dạng $\log(x+1)$.

```
plot(is_happy_customer~ is_expedited_delivery,data=new_df,xlab=" is_expedited_delivery",y
lab="is_happy_customer",main="Plot of is_happy_customer and is_expedited_delivery",col=2
)
plot(is_happy_customer~ is_expedited_delivery,data=new_df2,xlab=" log(is_expedited_delive
ry+1)",ylab="is_happy_customer",main="Plot of is_happy_customer and log(is_expedited_del
ivery+1) ",col=3)
```



- Nhận xét

Dựa vào biểu đồ phân tán, ta thấy có rất nhiều điểm không tập trung thành một vệt thẳng. Điều này thể hiện mối tương quan tuyến tính tương đối không rõ giữa các biến được vẽ.

5) Thống kê suy diễn :

- Ta có biến phụ thuộc là `is_happy_customer` và biến độc lập là tất cả các biến còn lại:

```
install.packages('caTools')
```

```
library(caTools)
```

#Chia bộ dữ liệu thành mẫu kiểm định và mẫu xây dựng:

```
set.seed(100)
```



```
split= sample.split(new_df2$is_happy_customer,SplitRatio = 0.65)
```

<code>split</code>	<code>logi [1:500] TRUE TRUE TRUE TRUE TRUE TRUE ...</code>
--------------------	---

#Sử dụng lệnh `subset` để chia thành mẫu train và test

```
mauxaydung = subset(new_df2,split==TRUE)
```

```
maukiemdinh = subset(new_df2,split==FALSE)
```

 maukiemdinh	175 obs. of 7 variables
 mauxaydung	325 obs. of 7 variables

#Xây dựng mô hình logistic:

```
mohinh = glm(is_happy_customer~.,data=mauxaydung,family = binomial)
```

```
summary(mohinh)
```

```
Call:
glm(formula = is_happy_customer ~ ., family = binomial, data = mauxaydung)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.8708  -0.4797   0.3220   0.7600   2.4418

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)    -49.2833     6.6432  -7.419 1.18e-13 ***
is_expedited_delivery -4.0207     0.6476  -6.208 5.35e-10 ***
order_price      0.2424     0.2184   1.110  0.2671
delivery_charges 12.3243     1.5568   7.916 2.44e-15 ***
coupon_discount  -0.2322     0.1362  -1.705  0.0881 .
order_total      -0.3163     0.1748  -1.809  0.0705 .
distance_to_nearest_warehouse -0.3134     0.2382  -1.316  0.1882
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 387.29  on 324  degrees of freedom
Residual deviance: 268.91  on 318  degrees of freedom
AIC: 282.91

Number of Fisher Scoring iterations: 5
```

Dự báo trên bộ mẫu xây dựng:

```
dubaoxaydung = predict(mohinh,type="response",newdata= mauxaydung)
summary(dubaoxaydung)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.02261	0.62582	0.75762	0.71692	0.94213	0.99569

Dự báo trên bộ mẫu kiểm định:

```
dubaokiemdinh = predict(mohinh,type="response",newdata= maukiemdinh)
summary(dubaokiemdinh)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.03489	0.52880	0.77852	0.69912	0.95139	0.99298

Xây dựng ma trận nhầm lẫn với mức ngưỡng = 0,5

```
table(maukiemdinh$sis_happy_customer,dubaokiemdinh>0.5)
```

	FALSE	TRUE
0	28	21
1	11	115

Độ chính xác của mô hình:

`(21+115)/nrow(maukiemdinh)`

Mô hình có độ chính xác cao: 77,71 %

- Xây dựng mô hình hồi quy logistic “mohinh2” khi đã loại bỏ các biến không có ý nghĩa thống kê ($p\text{-value} > 0,05$)

```
mohinh2 = glm(is_happy_customer ~ is_expedited_delivery + delivery_charges, data=mauxaydung, family = binomial)
summary(mohinh2)
```

Call:

```
glm(formula = is_happy_customer ~ is_expedited_delivery + delivery_charges,
    family = binomial, data = mauxaydung)
```

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.8843	-0.4998	0.3574	0.7635	2.4264

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-48.9321	6.3214	-7.741	9.89e-15	***
is_expedited_delivery	-3.8007	0.6216	-6.114	9.70e-10	***
delivery_charges	11.9007	1.5117	7.873	3.47e-15	***

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

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 387.29 on 324 degrees of freedom
 Residual deviance: 275.93 on 322 degrees of freedom
 AIC: 281.93

Number of Fisher Scoring iterations: 5

Dự báo trên bộ mẫu kiểm định:

```
dubaokiemdinh = predict(mohinh2,type="response",newdata= maukiemdinh)
summary(dubaokiemdinh)
```


Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.04499	0.55705	0.75676	0.70245	0.94714	0.98921

Xây dựng ma trận nhằm lần với mức ngưỡng = 0,5

```
table(maukiemdinh$is_happy_customer,dubaokiemdinh>0.5)
```

	FALSE	TRUE
0	28	21
1	9	117

Độ chính xác của mô hình:

```
(21+117)/nrow(maukiemdinh)
```

Mô hình có độ chính xác cao: 78,86 %

- So sánh giữa 2 mô hình mohinh, mohinh2:

```
anova(mohinh, mohinh2, test = "Chisq")
```

với giả thuyết:

H0: 2 mô hình có hiệu quả như nhau

H1: 2 mô hình có hiệu quả khác nhau

Analysis of Deviance Table

```
Model 1: is_happy_customer ~ is_expedited_delivery + order_price + delivery_charges +
  coupon_discount + order_total + distance_to_nearest_warehouse
Model 2: is_happy_customer ~ is_expedited_delivery + delivery_charges
Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      318      268.91
2      322      275.93 -4    -7.022    0.1347
```

Ta thấy p-value > 0,05. Chưa đủ cơ sở để bác bỏ H0

➔ 2 mô hình có hiệu quả như nhau

Ta chọn mô hình 2 vì nó có ít biến hơn

➔ Ta tìm được mô hình hồi quy logistic có dạng:

$$\text{is_happy_customer} = \frac{1}{1 + e^{-(11,9\text{delivery_charg} - 3,8\text{is_expedite_delivery} - 48,9)}}$$

- Kiểm tra giả định của mô hình hồi quy

Các giả định của mô hình hồi quy: $Y_i = \beta_0 + X_1B_1 + \dots + X_iB_i$, $i = 1, \dots, n$

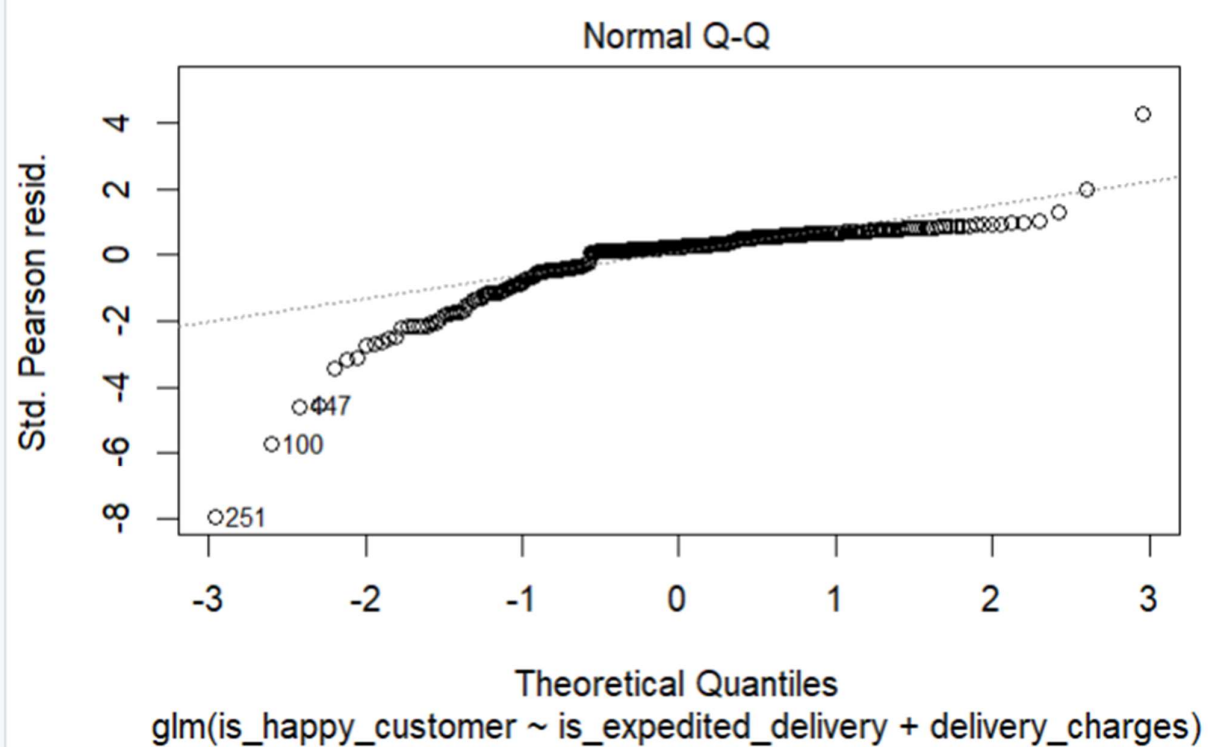
Tính tuyến tính của dữ liệu: mối quan hệ giữa biến dự báo X và biến phụ thuộc Y được giả sử là tuyến tính. Các sai số có kì vọng bằng 0. Phương sai của các sai số là hằng số. - Sai số có phân phối chuẩn. - Các sai số $\varepsilon_1, \dots, \varepsilon_n$ thì độc lập với nhau. Ta thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình. Các giả định bao gồm:

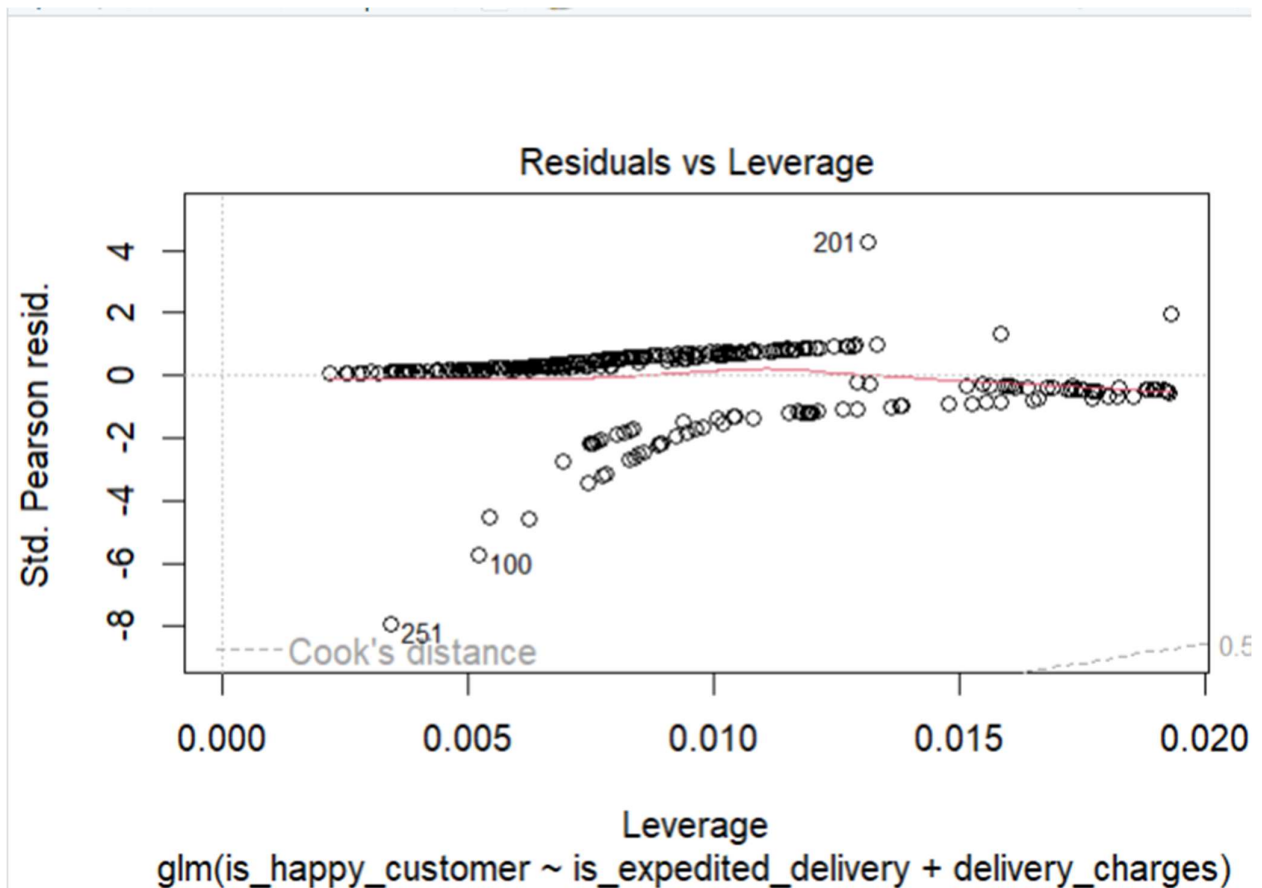
- Giả định 1: Tính tuyến tính của dữ liệu: mối quan hệ giữa biến độc lập và biến phụ thuộc được giả sử là tuyến tính.
- Giả định 2: Sai số có phân phối chuẩn.
- Giả định 3: Phương sai của các sai số là hằng số và có kì vọng bằng 0.
- Giả định 4: Các sai số độc lập với nhau.

Ta thực hiện phân tích thặng dư để kiểm tra các giả định của mô hình:

`plot(mohinh2)`







- Đồ thị thứ 1 (Residuals vs Fitted) vẽ các giá trị sai số hồi quy tương ứng với các giá trị dự báo, dùng để kiểm tra 3 giả định: tính tuyến tính của dữ liệu (giả định 1), phương sai các sai số là hằng số và có kỳ vọng bằng 0 (giả định 3). Nhìn đồ thị ta thấy đường màu đỏ là đường nằm ngang nên (giả định 1) Y có quan hệ tuyến tính với các biến độc lập thỏa mãn. Đường màu đỏ nằm sát đường $Y=0$ và các điểm sai số phân tán đều đường $Y=0$ nên (giả định 3) phương sai của sai số là hằng số và có kỳ vọng $=0$ không được thỏa mãn.

- Đồ thị thứ 2 (Normal Q-Q) cho phép kiểm tra giả định về phân phối chuẩn của các sai số. Ta thấy có nhiều điểm sai số không nằm trên đường kỳ vọng phân phối chuẩn. Nên (giả định 2) sai số có phân phối chuẩn là chưa được thỏa mãn.

- Đồ thị thứ 3 (Scale - Location) vẽ căn bậc hai của các giá trị tuyệt đối được chuẩn hóa với các giá trị dự báo, được dùng để kiểm tra giả định thứ 3 (phương sai của các sai số là hằng số). Nếu như đường màu đỏ trên đồ thị là đường thẳng nằm ngang và các điểm tuyệt đối phân tán đều xung quanh đường thẳng này thì giả định thứ 3 được thỏa. Nếu như đường màu đỏ có độ dốc (hoặc cong) hoặc các điểm tuyệt đối phân tán không đều xung quanh đường thẳng này, thì giả

định thứ 3 bị vi phạm. Ta thấy rằng các giá trị sai số trong đồ thị phân tán đều xung quanh và đường màu đỏ nằm ngang nên giả định về phương sai của các sai số là hằng số được thỏa mãn.

- Đồ thị thứ 4 (Residuals vs Leverage) cho phép xác định những điểm có ảnh hưởng cao trong bộ dữ liệu, ta thấy không có điểm nào vượt ra khỏi đường Cook's distance nên không cần phải loại bỏ điểm nào hết.

6. Thảo luận và mở rộng:

6.1 Mục đích xây dựng mô hình hồi quy logistic:

Mục đích chính của việc xây dựng mô hình hồi quy logistic là phân loại cảm xúc của khách hàng khi mua hàng thành các nhóm hay lớp. Điều này giúp ta hiểu được xác suất của việc khách hàng có cảm xúc tích cực hay tiêu cực dựa trên các biến độc lập như `customer_id`, `nearest_warehouse`, `order_price`, và `coupon_discount`. Mô hình hồi quy logistic thích hợp cho các bài toán phân loại.

6.2. Lợi ích đạt được:

Dự đoán cảm xúc của khách hàng: Mô hình có khả năng dự đoán xác suất một khách hàng cảm thấy hạnh phúc hoặc không hài lòng khi mua hàng, dựa trên các biến độc lập.

Tối ưu hóa chiến lược kinh doanh: Hiểu rõ về cảm xúc của khách hàng giúp doanh nghiệp tối ưu hóa chiến lược giảm giá, quản lý kho, và cải thiện dịch vụ để tăng cường hài lòng khách hàng.

Phát hiện tình huống đặc biệt: Mô hình có thể giúp phát hiện những tình huống đặc biệt khi cảm xúc của khách hàng không như dự kiến.

6.3. Ứng dụng thực tế:

Quản lý chất lượng dịch vụ: Mô hình có thể được sử dụng để phân loại khách hàng thành nhóm cảm xúc để doanh nghiệp có thể tập trung nâng cao chất lượng dịch vụ cho từng nhóm khách hàng.

Chiến lược tiếp thị tùy chỉnh: Doanh nghiệp có thể sử dụng mô hình để tùy chỉnh chiến lược tiếp thị và quảng cáo để thu hút và giữ chân khách hàng.

Phản hồi nhanh chóng: Khi mô hình được triển khai trong thời gian thực, doanh nghiệp có thể phản ứng nhanh chóng đối với các thay đổi trong tâm trạng của khách hàng và thực hiện các biện pháp cần thiết.

7. Nguồn dữ liệu và nguồn code.

Nguồn dữ liệu : https://www.kaggle.com/datasets/muhammadshahrayar/transactional-retail-dataset-of-electronics-store?select=dirty_data.csv

Nguồn code : <https://drive.google.com/drive/u/1/folders/1B1bZoI-Dt-3KNxQZV1Wa-WHLLNZf23nF>