ĐẠI HỌC QUỐC GIA TP. HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC CÔNG NGHỆ THÔNG TIN

BÁO CÁO ĐỒ ÁN

MÔN HỌC:

PHÂN TÍCH DỮ LIỆU KINH DOANH

Năm học: 2020 – 2021

UIT-HCM

Lớp: IS403.L22.HTCL

Giảng viên hướng dẫn: PGS.TS.Nguyễn Đình Thuân

Nhóm thực hiện: Nhóm 7

18521320 - Đoàn Thục Quyên
18521554 - Nguyễn Thành Trung
18520454 – Nguyễn Đức Anh

TP. Hồ Chí Minh, tháng 06 năm 2021

LỜI CẢM ƠN

Đầu tiên, nhóm chúng em xin gửi lời cảm ơn chân thành đến quý Thầy cô giảng viên Trường Đại học Công nghệ thông tin – Đại học Quốc gia TP.HCM và quý thầy cô khoa Hệ thống Thông tin đã giúp cho nhóm chúng em có những kiến thức cơ bản làm nền tảng để thực hiện đề tài này.

Đặc biệt, nhóm chúng em xin gửi lời cảm ơn và lòng biết ơn sâu sắc nhất tới PGS. TS. Nguyễn Đình Thuân, người đã hướng dẫn cho em trong suốt thời gian làm đề tài. Cô đã trực tiếp hướng dẫn tận tình, sửa chữa và đóng góp nhiều ý kiến quý báu giúp nhóm chúng em có thể hoàn thành tốt Báo cáo Đồ án môn học.

Trong thời gian một học kỳ thực hiện đề tài, nhóm chúng em đã vận dụng những kiến thức nền tảng đã tích lũy đồng thời kết hợp với việc học hỏi và nghiên cứu những kiến thức mới từ thầy cô, bạn bè cũng như nhiều nguồn tài liệu tham khảo. Từ đó, nhóm chúng em vận dụng tối đa những gì đã thu thập được để hoàn thành một báo cáo đồ án tốt nhất.

Tuy nhiên, vì kiến thức chuyên môn còn hạn chế và bản thân còn thiếu nhiều kinh nghiệm thực tiễn nên nội dung của báo cáo không tránh khỏi những thiếu sót, nhóm em rất mong nhận được sự góp ý, chỉ bảo thêm của Cô nhằm hoàn thiện những kiến thức của mình để nhóm chúng em có thể dùng làm hành trang thực hiện tiếp các đề tài khác trong tương lai cũng như là trong việc học tập và làm việc sau này.

Một lần nữa xin gửi đến Cô và các bạn lời cảm ơn chân thành và tốt đẹp nhất!

Thành phố Hồ Chí Minh, tháng 06 năm 2021

Nhóm sinh viên thực hiện

NHẬN XÉT CỦA GIẢNG VIÊN

…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
…………………………………………………………………………………………
………………………………………………………………………

MỤC LỤC

Loại 1: Phân tích phương sai (Levene, ANOVA, Tukey)

Mỗi nhóm chọn ba trong bốn loại sau, và mỗi loại giải 1 bài tập với dữ liệu thực tế (kinh tế, xã hội) tùy chọn của Việt Nam:

1. Dùng MS Excel và ngôn ngữ R thực hiện.

2. Mỗi bài tập cần thực hiện trong báo cáo:

- Phát biểu bài toán, nêu ý nghĩa của bài toán cần giải quyết.

- Tính lại và giải thích các giá trị trong bảng kết quả.

Loại 1: Phân tích phương sai (ANOVA, Levene, Tukey)

- Dữ liệu đây là danh sách các loại đồ dùng bán chạy nhất trên web Amazon trong năm 2021.
- Nguồn dữ liệu: https://www.kaggle.com/hussainaliarif/amazon-best-seller-june-2021-products
- **Phát biểu bài toán: Kiểm định xem giá tiền của 7 loại đồ dùng có giống nhau hay không. Với mức ý nghĩa α = 5%.**

a) LENEVE TEST : Kiểm định phương sai có bằng nhau hay không giữa các nhóm

Xác định giả thuyết:

- Giả thuyết H0 : Không có sự khác biệt về phương sai của 7 loại đồ dùng
- Giả thuyết H1 : Có sự khác biệt về phương sai của 7 loại đồ dùng.

1. Nhập dữ liệu:

```
> data = read.csv("C:/Users/Admin/Downloads/DoAn(ANOVA).csv",header=TRUE)
```

2. Kiểm tra dữ liệu:

	ï..Category	Price
1	Electronics	39.99
2	Electronics	34.99
3	Electronics	44.99
4	Electronics	28.48
5	Electronics	49.99
6	Electronics	89.99
7	Electronics	34.99
8	Electronics	149.99
9	Electronics	39.00
10	Electronics	24.99
11	Electronics	139.99
12	Electronics	249.99
13	Electronics	49.99
14	Electronics	44.99
15	Electronics	59.99
16	Electronics	129.99
17	Electronics	49.99
18	Electronics	13.22
19	Electronics	34.99
20	Electronics	149.99
21	Electronics	89.99
22	Electronics	179.99
23	Electronics	99.99
24	Electronics	7.99
25	Electronics	99.99
26	Electronics	88.34
27	Electronics	164.94
28	Electronics	37.98
29	Electronics	74.99
30	Electronics	16.99
31	Electronics	12.75

3. Kiểm định LEVENE:

```
> library(car)
> leveneTest(Price ~ ï..Category, data)
Levene's Test for Homogeneity of Variance (center = median)
       Df F value    Pr(>F)
group   6  32.921 < 2.2e-16 ***
      700
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Vì p-value > α nên chúng ta chấp nhận giả thuyết H0. Vậy không có sự khác nhau về phương sai của 7 loại đồ dùng.

➔ Đủ điều kiện để phân tích ANOVA

b) ANOVA TEST: Kiểm định ANOVA

Xác định giả thuyết:

- Giả thuyết H0 được đặt ra là giá tiền trung bình của 7 loại đồ dùng là như nhau.
- Giả thuyết H1 được đặt ra là có ít nhất một giá tiền trung bình trong 7 loại đồ dùng khác nhau.

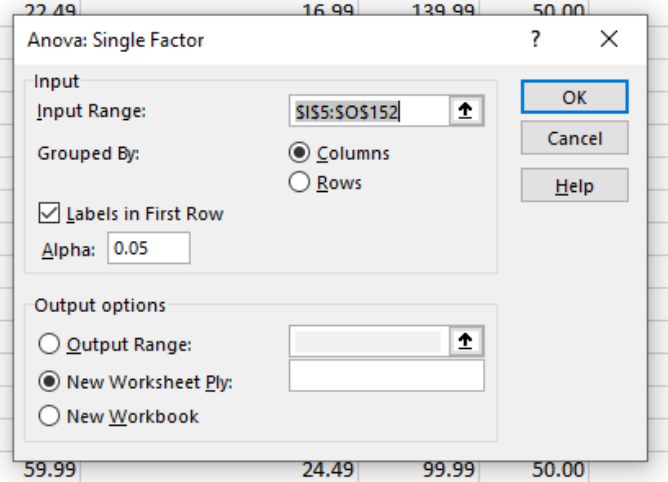
- Excel:

1. Nhập dữ liệu theo cột

Books	Camera & Photo	Clothing, Shoes & Jewelry	Electronics	Gift Cards	Toys & Games	Video Games
4.00	34.99	19.99	39.99	50.00	14.93	10.00
16.30	99.99	29.90	34.99	50.00	23.99	10.00
11.40	37.98	8.68	44.99	50.00	24.00	5.00
16.80	9.98	12.01	28.48	50.00	4.99	10.00
13.29	99.99	15.97	49.99	25.00	15.00	59.88
11.98	35.99	13.00	89.99	10.00	37.99	25.00
5.77	49.99	20.00	34.99	50.00	12.60	9.99
12.20	36.99	14.63	149.99	50.00	29.99	20.00
9.42	99.99	5.75	39.00	50.00	14.99	69.00
5.06	49.99	23.13	24.99	50.00	11.59	29.99
4.44	22.49	16.99	139.99	50.00	24.88	37.99
14.99	33.99	18.99	249.99	25.00	9.99	19.99
5.51	24.87	22.86	49.99	25.00	7.99	19.99
9.98	69.99	13.50	44.99	20.00	10.99	14.99
18.35	29.99	23.99	59.99	50.00	7.20	299.00
10.00	123.56	18.00	129.99	50.00	0.88	49.00
10.00	99.99	12.99	49.99	50.00	4.99	394.99
4.22	68.86	15.99	13.22	25.00	8.99	44.99
18.00	51.99	23.35	34.99	25.00	7.99	29.99
10.35	11.98	19.95	149.99	50.00	19.70	25.00
8.89	34.99	9.99	89.99	54.95	6.89	5.99
14.49	46.00	20.70	179.99	50.00	25.00	808.29
15.05	23.99	49.95	99.99	50.00	8.99	12.00
7.52	31.99	29.99	7.99	50.00	23.99	59.00
9.63	59.99	24.49	99.99	50.00	11.48	19.99
7.36	39.99	13.68	88.34	50.00	6.99	399.00
14.49	34.99	14.98	164.94	10.00	18.99	26.95
13.80	29.99	28.97	37.98	105.95	10.95	57.00
12.85	21.99	16.70	74.99	25.00	13.98	74.99
13.98	84.99	9.00	16.99	50.00	6.88	10.00
7.59	24.99	24.99	12.75	25.00	69.99	59.99
8.37	25.99	6.90	49.99	206.95	19.99	39.99
8.45	39.99	6.99	99.99	10.00	9.99	59.99
8.99	39.99	15.99	159.99	50.00	12.49	25.00
9.59	47.99	15.98	334.98	15.00	6.19	29.99
17.40	28.99	10.98	57.99	50.00	25.00	299.99
16.79	24.99	7.69	139.99	50.00	8.99	29.99
5.00	39.99	19.99	99.99	50.00	20.99	59.00
12.01	29.99	19.60	89.99	50.00	7.40	26.13
7.78	26.99	15.95	37.09	25.00	8.49	12.96

2. Chọn tab data -> Chọn Data Analysis -> Chọn Anova Single Factor -> input range chọn dữ liệu từ file excel đã nhập ở trên

Books	Camera & Photo	Clothing, Shoes & Jewelry	Electronics	Gift Cards	Toys & Games	Video Games
4.00	34.99	19.99	39.99	50.00	14.93	10.00
16.30	99.99	29.90	34.99	50.00	23.99	10.00
11.40	37.98	8.68	44.99	50.00	24.00	5.00
16.80	9.98	12.01	28.48	50.00	4.99	10.00
13.29	99.99	15.97	49.99	25.00	15.00	59.88
11.98	35.99	13.00	89.99	10.00	37.99	25.00
5.77	49.99	20.00	34.99	50.00	12.60	9.99
12.20	36.99	14.63	149.99	50.00	29.99	20.00
9.42	99.99	5.75	39.00	50.00	14.99	69.00
5.06	49.99	23.13	24.99	50.00	11.59	29.99
4.44	22.49	16.99	139.99	50.00	24.88	37.99
14.99					9.99	19.99
5.51					7.99	19.99
9.98					10.99	14.99
18.35					7.20	299.00
10.00					0.88	49.00
10.00					4.99	394.99
4.22					8.99	44.99
18.00					7.99	29.99
10.35					19.70	25.00
8.89					6.89	5.99
14.49					25.00	808.29
15.05					8.99	12.00
7.52					23.99	59.00
9.63	59.99	24.49	99.99	50.00	11.48	19.99
7.36	39.99	13.68	88.34	50.00	6.99	399.00
14.49	34.99	14.98	164.94	10.00	18.99	26.95
13.80	29.99	28.97	37.98	105.95	10.95	57.00
12.85	21.99	16.70	74.99	25.00	13.98	74.99
13.98	84.99	9.00	16.99	50.00	6.88	10.00
7.59	24.99	24.99	12.75	25.00	69.99	59.99
8.37	25.99	6.90	49.99	206.95	19.99	39.99
8.45	39.99	6.99	99.99	10.00	9.99	59.99
8.99	39.99	15.99	159.99	50.00	12.49	25.00
9.59	47.99	15.98	334.98	15.00	6.19	29.99
17.40	28.99	10.98	57.99	50.00	25.00	299.99
16.79	24.99	7.69	139.99	50.00	8.99	29.99
5.00	39.99	19.99	99.99	50.00	20.99	59.00
12.01	29.99	19.60	89.99	50.00	7.40	26.13



3. Em được bảng ANOVA như sau

Anova: Single Factor

SUMMARY

Groups	Count	Sum	Average	Variance
Books	70	804.67	11.49529	23.6408
Camera & Photo	100	5063.08	50.6308	1008.055
Clothing, Shoes & Jewelry	100	1760.87	17.6087	74.3502
Electronics	147	19907.48	135.425	19867.32
Gift Cards	100	4400.74	44.0074	790.6348
Toys & Games	95	1639.33	17.25611	245.639
Video Games	95	5794.08	60.99032	11206.07

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1375526	6	229254.4	38.538	1.6E-40	2.111514
Within Groups	4164152	700	5948.788			
Total	5539678	706				

Nhìn bảng em thấy F > F crit (38.538 > 2.111514) nên ta bác bỏ giả thuyết H0, chấp nhận giả thuyết H1 là có ít nhất một giá tiền trung bình trong 7 loại đồ dùng khác nhau.

Giải thích một số ký tự trong bảng kết quả ANOVA:

Bảng ANOVA

Nguồn sai số	Tổng bình phương SS	Bậc tự do df	Bình phương trung bình MS	Giá trị thống kê F
Yếu t ố (Between Group)	SSA	k-1	$MSA = \frac{SSA}{k-1}$	$F = \frac{MSA}{MSE}$
Sai số (Within Group)	SSE = SST - SSA	n-k	$MSE = \frac{SSE}{n-k}$	
Tổng cộng	SST	n-1		

Trong đó:

- SSA = SSG (between-groups sum of squares) là tổng bình phương độ lệch giữa các nhóm
- SSE = SSW (Within-groups sum of squares) là tổng bình phương độ lệch trong nhóm
- SST (Total- sum of squares) là tổng bình phương các độ lệch giữa từng quan sát với trung bình của tất cả các quan sát.

Cách tính một số giá trị :

- ❖ SSW

$$SSW = SS_1 + SS_2 + \ldots + SS_k = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

Em tính các giá trị SSi bằng công thức:

$$SS_i = \sum_{j=1}^{n_i}(x_{ij} - \bar{x}_i)^2$$

- ○ Các giá trị (Xij – Xi)^2

=(C6-V5)^2

I	J	K	L	M	N	O	P	Q
eo Games		SS(books)	SS(Camera & Photo)	SS(Clothing, Shoes & Jewelry)	SS(Electronics)	SS(Gift Cards)	SS(Toys & Games)	SS(Video Games)
10.00		56.18	244.63	5.67	9107.85	35.91	5.41	2600.01
10.00		23.09	2436.33	151.08	10087.20	35.91	45.35	2600.01
5.00		0.01	160.04	79.72	8178.50	35.91	45.48	3134.92
10.00		28.14	1652.49	31.35	11437.24	35.91	150.46	2600.01
59.88		3.22	2436.33	2.69	7299.15	361.28	5.09	1.23
25.00		0.23	214.35	21.24	2064.34	1156.50	429.89	1295.30
9.99		32.78	0.41	5.72	10087.20	35.91	21.68	2601.03
20.00		0.50	186.07	8.87	212.14	35.91	162.15	1680.21
69.00		4.31	2436.33	140.63	9297.79	35.91	5.14	64.16
29.99		41.41	0.41	30.48	12195.90	35.91	32.10	961.02
37.99		49.78	791.90	0.38	20.84	35.91	58.12	529.01
19.99		12.21	276.92	1.91	13125.13	361.28	52.80	1681.03
19.99		35.82	663.62	27.58	7299.15	361.28	85.86	1681.03
14.99		2.30	374.78	16.88	8178.50	576.36	39.26	2116.03
299.00		46.99	426.04	40.72	5690.44	35.91	101.13	56648.61
49.00		2.24	5318.67	0.15	29.54	35.91	268.18	143.77
394.99		2.24	2436.33	21.33	7299.15	35.91	150.46	111555.79
44.99		52.93	332.30	2.62	14934.07	361.28	68.33	256.01

- ○ Cộng theo dòng thì sẽ ra các giá trị SS của mỗi nhóm

=SUM(K6:K75)

	Q	R	S	T	U
es)	SS(Video Games)				
.41	2600.01		Books	1631.21	
.35	2600.01		Camera & Photo	99797.49	
.48	3134.92		Clothing, Shoes &	7360.67	
.46	2600.01		Electronics	2900629.03	
.09	1.23		Gift Cards	78272.85	
.89	1295.30		Toys & Games	23090.06	
.68	2601.03		Video Games	1053370.25	
.15	1680.21				
.14	64.16				

- Cộng các giá trị SSi lại sẽ ra được SSW

=SUM(T6:T12)

	Q	R	S	T	U	V	W
es)	SS(Video Games)						
41	2600.01		Books	1631.21		SSW	4164151.57
35	2600.01		Camera & Photo	99797.49			
48	3134.92		Clothing, Shoes &	7360.67			
46	2600.01		Electronics	2900629.03			
09	1.23		Gift Cards	78272.85			
89	1295.30		Toys & Games	23090.06			
68	2601.03		Video Games	1053370.25			
15	1680.21						
14	64.16						
10	961.02						

❖ SSG

$$SSG = \sum_{i=1}^{n_i} n_i (\bar{x}_i - \bar{x})^2$$

- Tính các giá trị trung bình Xi theo từng nhóm

=AVERAGE(C6:C75)

U	V	W	X	Y	Z
				Giá trị TB của nhóm Books	11.50
	SSW	4164151.57		Giá trị TB của nhóm Camera & Photo	50.63
				Giá trị TB của nhóm Clothing, Shoes &	17.61
				Giá trị TB của nhóm Electronics	135.43
				Giá trị TB của nhóm Gift Cards	44.01
				Giá trị TB của nhóm Toys & Games	17.26
				Giá trị TB của nhóm Video Games	60.99

- Giá trị trung bình của toàn bộ X

=AVERAGE(C6:I152)

U	V	W	X	Y	Z	AA	AB	AC
				Giá trị TB của nhóm Books	11.50		Tổng giá trị TB	55.69
	SSW	4164151.57		Giá trị TB của nhóm Camera & Photo	50.63			
				Giá trị TB của nhóm Clothing, Shoes &	17.61			
				Giá trị TB của nhóm Electronics	135.43			
				Giá trị TB của nhóm Gift Cards	44.01			

- Tính các giá trị (Xi – X)^2

=(Z5-AC5)^2

U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF
				Giá trị TB của nhóm Books	11.50		Tổng giá trị TB	55.69		SS	1952.850231
	SSW	4164151.57		Giá trị TB của nhóm Camera & Photo	50.63						25.558593668
				Giá trị TB của nhóm Clothing, Shoes &	17.61						1449.907488766
				Giá trị TB của nhóm Electronics	135.43						6358.257604159
				Giá trị TB của nhóm Gift Cards	44.01						136.397891273
				Giá trị TB của nhóm Toys & Games	17.26						1476.883770326
				Giá trị TB của nhóm Video Games	60.99						28.132044844

- Đếm các mẫu dữ liệu theo từng nhóm

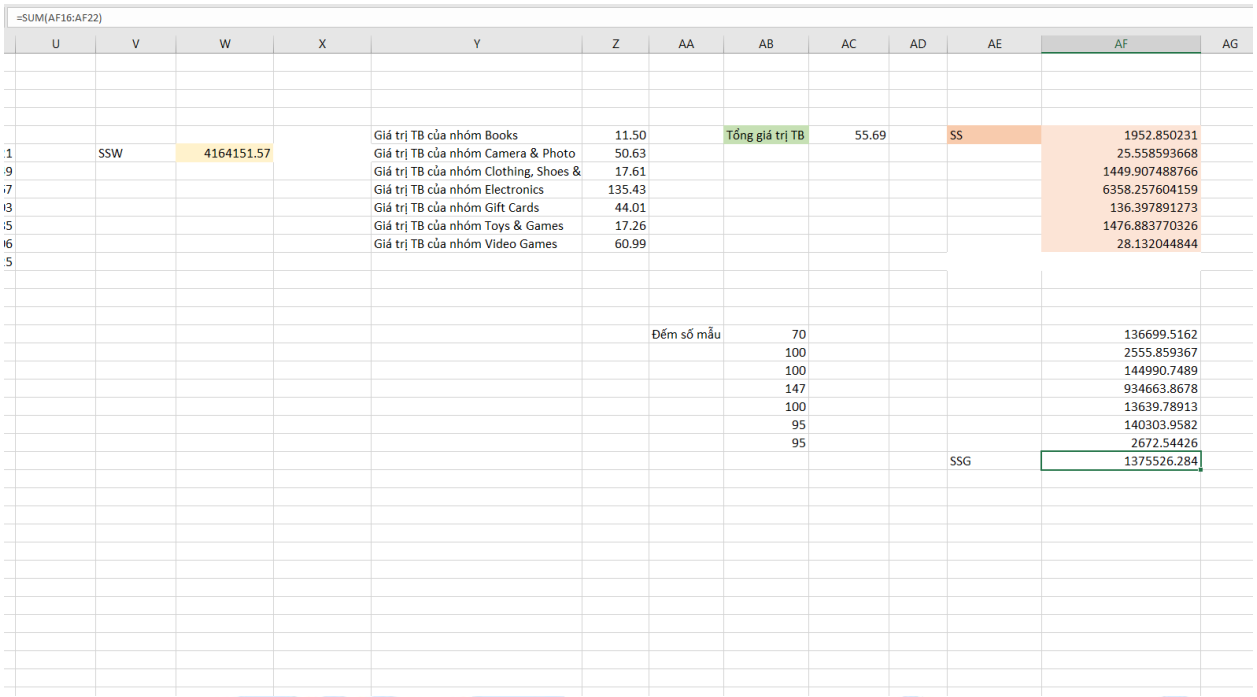
=COUNT(C6:C75)

U	V	W	X	Y	Z	AA	AB
				Giá trị TB của nhóm Books	11.50		Tổng giá trị TB
	SSW	4164151.57		Giá trị TB của nhóm Camera & Photo	50.63		
				Giá trị TB của nhóm Clothing, Shoes &	17.61		
				Giá trị TB của nhóm Electronics	135.43		
				Giá trị TB của nhóm Gift Cards	44.01		
				Giá trị TB của nhóm Toys & Games	17.26		
				Giá trị TB của nhóm Video Games	60.99		
						Đếm số mẫu	70
							100
							100
							147
							100
							95
							95

- Tính SSG

=AF5*AB16

U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
				Giá trị TB của nhóm Books	11.50		Tổng giá trị TB	55.69		SS	1952.850231	
	SSW	4164151.57		Giá trị TB của nhóm Camera & Photo	50.63						25.558593668	
				Giá trị TB của nhóm Clothing, Shoes &	17.61						1449.907488766	
				Giá trị TB của nhóm Electronics	135.43						6358.257604159	
				Giá trị TB của nhóm Gift Cards	44.01						136.397891273	
				Giá trị TB của nhóm Toys & Games	17.26						1476.883770326	
				Giá trị TB của nhóm Video Games	60.99						28.132044844	
						Đếm số mẫu	70				136699.5162	
							100				2555.859367	
							100				144990.7489	
							147				934663.8678	
							100				13639.78913	
							95				140303.9582	
							95				2672.54426	
										SSG	1375526.284	



	U	V	W	X	Y	Z	AA	AB	AC	AD	AE	AF	AG
					Giá trị TB của nhóm Books	11.50		Tổng giá trị TB	55.69		SS	1952.850231	
1		SSW	4164151.57		Giá trị TB của nhóm Camera & Photo	50.63						25.558593668	
9					Giá trị TB của nhóm Clothing, Shoes &	17.61						1449.907488766	
7					Giá trị TB của nhóm Electronics	135.43						6358.257604159	
3					Giá trị TB của nhóm Gift Cards	44.01						136.397891273	
5					Giá trị TB của nhóm Toys & Games	17.26						1476.883770326	
6					Giá trị TB của nhóm Video Games	60.99						28.132044844	
5													
							Đếm số mẫu	70				136699.5162	
								100				2555.859367	
								100				144990.7489	
								147				934663.8678	
								100				13639.78913	
								95				140303.9582	
								95				2672.54426	
											SSG	1375526.284	

❖ MSW = SSW/(n – k) = 4164152/(707 – 7) = 5948.788

❖ MSG = SSG/(k – 1) = 1375526.284/(7 – 1) = 229254.3807

❖ F = MSG/MSW = 229254.3807 / 5948.788 = 38.537

R

1. Nhập file

- data <- read.csv("C:/Users/Admin/Downloads/DoAn(ANOVA).csv")

```
> data
                ï..Category  Price
1                Electronics  39.99
2                Electronics  34.99
3                Electronics  44.99
4                Electronics  28.48
5                Electronics  49.99
6                Electronics  89.99
7                Electronics  34.99
8                Electronics 149.99
9                Electronics  39.00
10               Electronics  24.99
11               Electronics 139.99
12               Electronics 249.99
13               Electronics  49.99
14               Electronics  44.99
15               Electronics  59.99
16               Electronics 129.99
17               Electronics  49.99
18               Electronics  13.22
```

2. Tiến hành kiểm định phương sai

```
> oneway <- aov(Price ~ ï..Category, data = data)
> summary(oneway)
              Df  Sum Sq Mean Sq F value Pr(>F)
ï..Category    6 1375526  229254   38.54 <2e-16 ***
Residuals    700 4164152    5949
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

c) Tukey Test: Kiểm định Tukey Test

- Vì đã bác bỏ H0 nên bây giờ em sẽ thực hiện kiểm định chuyên sâu ANOVA nhằm xác định giá trị trung bình của nhóm nào khác với nhóm nào, lớn hơn hay nhỏ hơn.

Cách tính :

B1 : Tính Q- statistic

Ta có k = 7, df = 700, α = 0.05 -> Tra bảng phân phối Tukey, ta được Q-statistic = 4.170

B2: Tính tiêu chuẩn so sánh:

$$T = q_{(k,n-k),\alpha}\sqrt{\frac{MSW}{n_{\min}}}$$

Trong đó:

Q: q-statistic

K: Số nhóm

n - k: k là số nhóm, n là số lượng các giá trị quan sát

α = 0.05

MSW: bình phương trung bình trong nội bộ của các nhóm

N_{min}: Số lượng các giá trị quan sát nhỏ nhất

C24 =C23 * (SQRT(D17/B5))

	A	B	C	D	E	F	G	H
1	Anova: Single Factor							
2								
3	SUMMARY							
4	*Groups*	*Count*	*Sum*	*Average*	*Variance*			
5	Books	70	804.67	11.49529	23.6408			
6	Camera & Photo	100	5063.08	50.6308	1008.055			
7	Clothing, Shoes & Jewelry	100	1760.87	17.6087	74.3502			
8	Electronics	147	19907.48	135.425	19867.32			
9	Gift Cards	100	4400.74	44.0074	790.6348			
10	Toys & Games	95	1639.33	17.25611	245.639			
11	Video Games	95	5794.08	60.99032	11206.07			
12								
13								
14	ANOVA							
15	*Source of Variation*	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*	
16	Between Groups	1375526.284	6	229254.4	38.538	1.6E-40	2.111514	
17	Within Groups	4164151.569	700	5948.788				
18								
19	Total	5539677.853	706					
20								
21								
22								
23		q- statistic	4.17					
24		T	38.44158					
25								

B3: Tính sự khác biệt giữa 2 nhóm

$$D_{ij} = \left| \overline{x_i} - \overline{x_j} \right|$$

Trong đó:

x_i: Giá trị trung bình của nhóm i

x_j: Giá trị trung bình của nhóm j

- Books || Camera & Photo = | 11.4952857142857 - 50.6307999999999| = 38.5 > 38.44 (significant)
- Books || Clothing, Shoes & Jewelry = |11.4952857142857 - 17.6087| = 6.1 < 38.44 (not significant)
- **Books || Electronics = |11.4952857142857 - 135.425034013605| = 123.9 > 38.44 (significant)**
- Books || Gift Cards = |11.4952857142857 - 44.0074 | = 32.5 < 38.44 (not significant)

- Books || Toys & Games = |11.4952857142857 - 17.2561052631579| = 5.75 < 38.44
- Books || Video Games = |11.4952857142857 - 60.9903157894736| = 49.5 > 38.44 (significant)
- Camera & Photo || Clothing, Shoes & Jewelry = |50.6307999999999 - 17.6087| = 33 < 38.44 (not significant)
- **Camera & Photo || Electronics = |50.6307999999999 - 135.425034013605| = 84.8 > 38.44 (significant)**
- Camera & Photo || Gift Cards = |50.6307999999999 - 44.0074| = 5.4 < 38.44 (not significant)
- Camera & Photo || Toys & Games = |50.6307999999999 - 17.2561052631579| = 33.35 < 38.44 (not significant)
- Camera & Photo || Video Games = |50.6307999999999 - 60.9903157894736| = 10 < 38.44 (not significant)
- **Clothing, Shoes & Jewelry || Electronics = |17.6087 - 135.425034013605| = 117.75 > 38.44 (significant)**
- Clothing, Shoes & Jewelry || Gift Cards = |17.6087 - 44.0074| = 26.4 < 38.44 (not significant)
- Clothing, Shoes & Jewelry || Toys & Games = |17.6087 - 17.2561052631579| = 0.35 < 38.44 (not significant)
- Clothing, Shoes & Jewelry || Video Games = |17.6087 - 60.9903157894736| = 43.4 > 38.44 (significant)
- Electronics || Gift Cards = | 135.425034013605 - 44.0074| = 91.4 > 38.4 (significant)
- **Electronics || Toys & Games= | 135.425034013605 - 17.2561052631579| = 118.25 4 > 38.4 (significant)**
- **Electronics || Video Games= | 135.425034013605 - 60.9903157894736| = 74.4 > 38.4 (significant)**
- Gift Cards || Toys & Games= | 44.0074 - 17.2561052631579| = 26.75 < 38.4 (not significant)
- Gift Cards || Video Games= | 44.0074 - 60.9903157894736| = 19 < 38.4 (not significant)
- Toys & Games || Video Games= | 17.2561052631579 - 60.9903157894736| = 43.75> 38.4 (significant)

⇨ Kết luận : Nhìn chung thì có sự khác biệt rõ ràng đều liên quan đến nhóm đồ dùng Electronics

Tính bằng R :

```
> summary(onwway)
             Df  Sum Sq Mean Sq F value Pr(>F)
ï..Category   6 1375526  229254   38.54 <2e-16 ***
Residuals   700 4164152    5949
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> TukeyHSD(onwway)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Price ~ ï..Category, data = data)

$ï..Category
                                                diff         lwr         upr     p adj
Camera & Photo-Books                       39.1355143    3.594347  74.6766818 0.0201947
Clothing, Shoes & Jewelry-Books             6.1134143  -29.427753  41.6545818 0.9987382
Electronics-Books                         123.9297483   90.810656 157.0488409 0.0000000
Gift Cards-Books                           32.5121143   -3.029053  68.0532818 0.0984912
Toys & Games-Books                          5.7608195  -30.163405  41.6850438 0.9991535
Video Games-Books                          49.4950301   13.570806  85.4192543 0.0010087
Clothing, Shoes & Jewelry-Camera & Photo  -33.0221000  -65.275184  -0.7690160 0.0407737
Electronics-Camera & Photo                 84.7942340   55.231400 114.3570676 0.0000000
Gift Cards-Camera & Photo                  -6.6234000  -38.876484  25.6296840 0.9965925
Toys & Games-Camera & Photo               -33.3746947  -66.049406  -0.6999839 0.0416384
Video Games-Camera & Photo                 10.3595158  -22.315195  43.0342266 0.9663726
Electronics-Clothing, Shoes & Jewelry     117.8163340   88.253500 147.3791676 0.0000000
Gift Cards-Clothing, Shoes & Jewelry       26.3987000   -5.854384  58.6517840 0.1915820
Toys & Games-Clothing, Shoes & Jewelry     -0.3525947  -33.027306  32.3221161 1.0000000
Video Games-Clothing, Shoes & Jewelry      43.3816158   10.706905  76.0563266 0.0018295
Gift Cards-Electronics                    -91.4176340 -120.980468 -61.8548004 0.0000000
Toys & Games-Electronics                 -118.1689288 -148.191194 -88.1466632 0.0000000
Video Games-Electronics                   -74.4347182 -104.456984 -44.4124527 0.0000000
Toys & Games-Gift Cards                   -26.7512947  -59.426006   5.9234161 0.1913036
Video Games-Gift Cards                     16.9829158  -15.691795  49.6576266 0.7223927
Video Games-Toys & Games                   43.7342105   10.643245  76.8251765 0.0019628
```

Trong đó:

P adj : chỉ số p sau khi đã điều chỉnh

Dff : Hiệu số hay khác biệt giữa 2 nhóm

Bảng tin cậy 95% theo lwr và upr, nếu đều < 0 hay > 0 thì có sự khác biệt có ý nghĩa thống kê :

+ Do đó ta thấy Camera & Photo-Books có lwr và upr đều > 0 nên nó có sự khác biệt mang ý nghĩa thống kê. Tương tự các trường hợp còn lại.

Nguồn tài liệu tham khảo:

https://www.youtube.com/watch?v=zQr190cacC0&t=279s

https://www.geeksforgeeks.org/levenes-test-in-r-programming/

https://www.youtube.com/watch?v=f6h6Y8PEOt8

https://www.scribbr.com/statistics/anova-in-r/

Loại 2: Chọn bài tập với dữ liệu thực tế (kinh tế, xã hội) tùy chọn của Việt Nam để thực hiện bài toán Hồi qui tuyến tính đa biến hoặc Hồi qui phi tuyến đa biến

1. Dùng MS Excel và ngôn ngữ R thực hiện.

2. Mỗi bài tập cần thực hiện trong báo cáo:

- Phát biểu bài toán, nêu ý nghĩa của bài toán cần giải quyết.

- Tính lại và giải thích các giá trị trong bảng kết quả.

Thông tin tập dữ liệu:

Tập dữ liệu này chứa dữ liệu được thu thập trong một cuộc khảo sát về nhà ở được thực hiện vào năm 2016, là một phần của dự án Luận án Tiến sĩ Phan Anh Nguyên. Dự án nghiên cứu này được thực hiện nhờ khoản tài trợ từ Học bổng 911 của Chính phủ Việt Nam. Ngày đăng trực tuyến đầu tiên trên https://figshare.com vào ngày 27.10.2020. Kết quả nghiêm cứu phục vụ cho việc nâng cao hiệu quả sử dụng năng lượng trong nhà ống Việt Nam.

Dữ liệu được lấy từ một cuộc khảo sát với 153 người được hỏi ở ba vùng khí hậu chính của Việt Nam. Cuộc khảo sát tập trung vào các đặc điểm của tòa nhà, hiệu suất môi trường, hiệu suất năng lượng và các hoạt động tân trang. Dữ liệu thu thập từ cuộc khảo sát được phân tích thống kê để cung cấp cái nhìn sâu sắc về hiệu suất của các loại nhà ở hiện tại và tiềm năng tiết kiệm năng lượng của nó.

Nguồn dữ liệu:

https://figshare.com/articles/dataset/Survey_of_housing_energy_consumption_and_refurbishment_in_Vietnam_/13109924?file=25143299

	A	B	C	D	E	F	G	H	I	J	K	L
1	Timestamp	Age	Climatic region	Ownership	Number of occupants	House age	House typology	Function other than residential	Number of exposed façade	Main orientation	Number of floor	Total floor area
2	2016-06-10 17:15:19	25	North	shared rent	4	5	Attached row house	No	3	West	3	110
3	2016-06-11 01:18:33	39	Center	privately own	6	5	Detached house	No	2	South	5	180
4	2016-06-13 13:11:39	25	North	privately rent	4	5	Attached row house	No	3	West	3	30
5	2016-06-14 12:35:00	25	North	privately rent	1	5	Attached row house	No	1	North East	5	30
6	2016-06-13 13:33:25	25	North	privately own	7	5	Attached row house	office	1	South West	5	240
7	2016-06-13 16:12:35	39	North	privately own	7	15	Attached row house	No	3	South East	4	240
8	2016-06-13 16:12:40	39	North	privately own	6	15	Attached row house	No	3	South East	4	240
9	2016-06-13 15:18:57	25	North	privately own	4	15	Attached row house	No	2	South	2	110
10	2016-06-13 18:08:07	39	North	privately own	5	5	Attached row house	No	1	South West	4	30
11	2016-06-13 15:36:59	25	North	privately own	5	5	Attached row house	Commercial	2	North West	3	70
12	2016-06-13 15:38:48	39	North	privately rent	2		Attached row house	No	2	South West	4	
13	2016-06-13 15:43:39	39	North	privately own	4	15	Attached row house	No	2	North East	5	240
14	2016-06-13 15:44:11	39	South	privately rent	2	5	Apartment	Commercial	4		5	70
15	2016-06-14 16:14:55	39	North	privately own	4	5	Attached row house	No	2	North	4	110
16	2016-06-14 17:17:47	25	North	shared rent	1		Attached row house	No	1	North	5	

Phát biểu bài toán:

Xây dựng mô hình hồi quy đa tuyến tính hoặc phi tuyến đặc trưng cho mối quan hệ giữa mức năng lượng tiêu thụ điện kWh trên đầu người và nhiều biến độc lập độc lập khác(tất cả đều là số).

Thực hiện: bài báo cáo sẽ gồm các phần sau:

- **Xác định biến phụ thuộc:** *energy kWh per person*
- **Xác định biến độc lập**
- **Tính thủ công các giá trị trong bảng kết quả bằng excel**
- **Giải thích các giá trị trong bảng kết quả bằng excel và bằng R**

I. Tiền xử lí dữ liệu:

- Làm sạch dữ liệu, xóa các cột thuộc tính không cần thiết:
    - Các cột dữ liệu kiểu chuỗi (ngoại trừ các cột dữ liệu mang tính phân loại thì xem xét sau).
    - Chuyển dữ liệu phân loại đã xét ở trên thành số tương ứng: Yes = 1 và No = 0
    - Em tính mực độ tương quan giữa biến phụ thuộc và các biến còn lại (dùng trên excel bằng hàm CORREL() hoặc trên R bằng hàm cor( )). Từ kết quả thu được, em chỉ giữa lại các thuộc tính có giá trị tương quan cao (giá trị dương).

Age	Number of occupan	House age	Number of flo	Total floor are	summer dalyl	summer therma	summe	Winter	Winter	Winter	Gas col	maximum ele	minimum ele	number	Frequel	Frequel	Frequel	Frequ
39	2	5	5	70	3	4	4	3	4	4	100000	240000	100000	2	1	1	0	
25	2	15	2	110	5	1	3	4	3	3	100000	240000	100000	1	2	1	1	
60	5	15	3	240	3	3	3	4	3	3		1330000	675000	3	1	2	0	
25	4	15	1	70	5	1	3	5	1	3	225000	930000	440000	1	2	2	0	
39	3	25	4	180	4	2	3	4	2	3		930000	675000	2	2	2	1	
60	5	5	5	240	4	3	4	4	3	4	100000	930000	440000	2	1	1	0	
80	3	25	4	240	3	3	3	3	4	3		1330000	930000	4	1	1	1	
60	6	15	3	240	5	4	4	5	4	4		1330000	440000	4	1	2	0	
60	4	25	3	180	3	3	2	2	3	3	225000	1800000	930000	2	1	2	0	
39	2	5	3	110	3	3	3	3	3	3		930000	675000	1	2	2	2	
39	4	15	2	70	4	4	4	4	4	4	100000	675000	100000	1	2	2	0	
39	3	5	4	180	4	2	3	4	3	3		1330000	930000	1	2	1	1	
60	2	5	1	110	2	2	2	1	1	1	100000	675000	240000	3	2	2	2	
39	2	5	2	180	3	3	3	3	3	2	100000	240000	100000	1	0	1	0	
39	4	15	2	240	3	3	3	3	3	3	375000	440000	675000	3	1	1	1	
CORREL																		
0.24287	0.129424465	0.1544574	0.243571583	0.239282842	-0.076271748	-0.034664341	-0.03841	-0.01209	-0.06337	-0.03282	0.139307	0.439350235	0.203909539	0.167767	0.033694	0.071747	-0.02123	0.03

- Làm mượt dữ liệu:
    - Đối với cột mang giá trị phân loại: tính MODE() của cột và điền vào các ô bị khuyết.

- Đối với cột mang giá trị liên tục: tính AVERAGE() của cột và điền vào các ô bị khuyết.
- Đối với những hàng có quá nhiều dữ liệu 0 hoặc trống, em xóa.

- Tất cả các biến độc lập trong mô hình hồi quy tuyến tính không phải lúc nào cũng đáng kể. Chúng em sẽ học cách xây dựng các mô hình hồi quy tốt bao gồm bộ biến "tốt nhất".

II. Xác định các biến phụ thuộc

Bước 1: Thêm dữ liệu

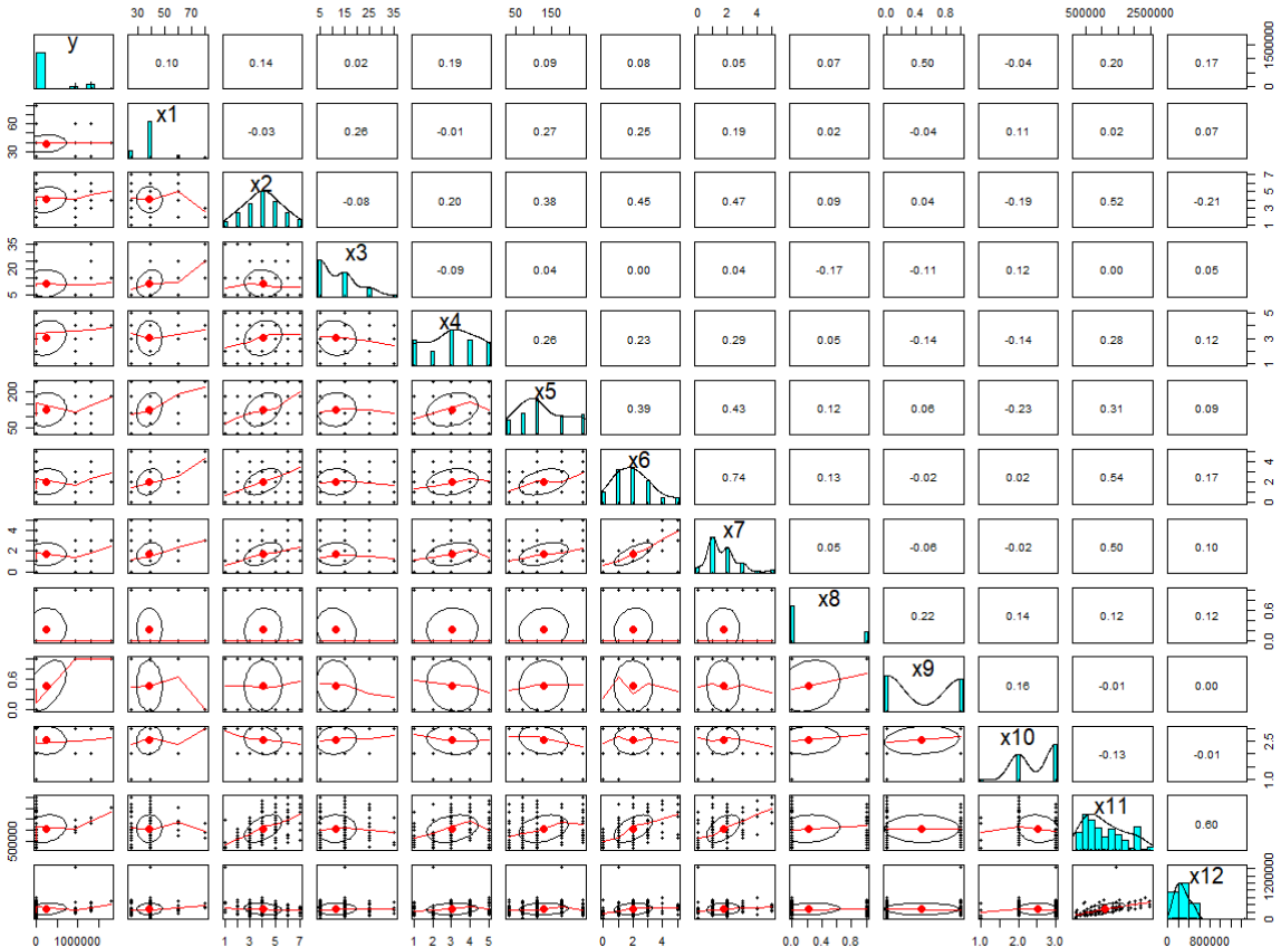
```
> survey_data <-
read_csv("F:/_Nam03/HKII/5.PhanTichDuLieuKinhDoanh/R/DoAn/survey_data.csv")
> View(survey_data)
> x1 = survey_data $Age
> x2 = survey_data$`Number of occupants`
> x3 = survey_data $`House age`
> x4 = survey_data $`Number of floor`
> x5 = survey_data $`Toeml floor area`
> x6 = survey_data $`number of airconditioning`
> x7 = survey_data $`electricity water heater`
> x8 = survey_data $`Solar hotwater`
> x9 = survey_data $`Energy efficiency equipment`
> x10= survey_data $`saving energy attitude`
> x11= survey_data $`toeml energy consumption`
> x12= survey_data $`energy oeroerson
> y  =  survey_data $`energy kWh per person`
```

Bước 2: Kiểm tra dữ liệu

```
> library(psych)
> vars = cbind(y,x1,x2,x3,x4,x5,x6,x7,x8,x9,x10,x11,x12)
```

```
> pairs.panels(vars)
```

Em thu được bảng sau:



- Em thấy hệ số tương quan giữa x6 và x7 bằng 0.74 > 0.7 □ có khả năng tồn tại multicollinearity (Đa đối chiếu)
- Em càng chắc chắn hơn khi dùng hàm alias(). Kết quả cho thấy x6 và x7 là cộng tuyến hoàn hảo. Khắc phục điều này, em tiến hành loại bỏ x6 và x7.

```
> alias(model)
```

```
Complete :
    (Intercept) x1 x2 x4 x5 x6 x7 x8 x9 x10 x11
x12 0           0  0  0  0  1  1  0  0  0   0
```

- Để chắc chắn hơn, em tính Variance Inflation Factors(VIF) với câu lệnh như bên dưới.

```
> library(tidyverse)
```

```
> library(caret)
> model = lm(y ~ x1 + x2 + x4 + x5 + x6 + x7 + x8 + x9 + x10 + x11 + x12)
> car::vif(model)
```

```
      x1       x2       x4       x5       x6       x7       x8       x9      x10      x11      x12
1.218877 3.745165 1.185458 1.576672 2.620732 2.574164 1.128642 1.120204 1.212951 5.752730 4.073085
```

- Em phát hiện giá trị VIF của x11 là rất cao (VIF = 5.75), lớn hơn 5, em cũng tiến hành loại bỏ x11.

Bước 3: Import lại dữ liệu, sau khi xóa 3 cột thuộc tính

```
> x1 = survey_data $Age
> x2 = survey_data$`Number of occupants`
> x3 = survey_data $`House age`
> x4 = survey_data $`Number of floor`
> x5 = survey_data $`Toeml floor area`
> x8 = survey_data $`electricity water heater`
> x9= survey_data $`electricity bill`
> x10= survey_data $`energy oeroerson`
> y  =  survey_data $`energy kWh per person`
> model = lm(y ~ x1 + x2 + x4 + x5 + x6 + x7 + x8 + x9 + x10)
```

- Xem kết quả mô hình

```
> summary(model)
```

```
Call:
lm(formula = y ~ x1 + x2 + x4 + x5 + x6 + x7 + x8 + x9 + x10)

Residuals:
    Min      1Q  Median      3Q     Max
-990966 -438905 -107632  496770 1290990

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.135e+06  3.273e+05  -3.468 0.000695 ***
x1           1.435e+04  5.164e+03   2.779 0.006192 **
x2           1.003e+05  4.042e+04   2.480 0.014298 *
x4           7.932e+04  3.403e+04   2.331 0.021190 *
x5           8.433e+02  7.678e+02   1.098 0.273925
x6          -1.546e-01  1.876e-01  -0.824 0.411215
x7          -3.052e+03  5.464e+04  -0.056 0.955540
x8          -6.384e+04  6.365e+04  -1.003 0.317517
x9           1.811e+04  7.046e+04   0.257 0.797538
x10          1.458e+00  3.627e-01   4.019 9.48e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 536700 on 141 degrees of freedom
Multiple R-squared:  0.2534,    Adjusted R-squared:  0.2057
F-statistic: 5.317 on 9 and 141 DF,  p-value: 3.033e-06
```

- Vì x7 và x9 có giá trị p-value lớn nhất nên t lần lượt bỏ x7 và x9 và xây dựng lại mô hình hồi quy.

```
Call:
lm(formula = y ~ x1 + x2 + x4 + x5 + x6 + x8 + x10)

Residuals:
     Min       1Q   Median       3Q      Max
-1000464  -438653  -104813   498794  1286488

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.085e+06  2.620e+05  -4.140 5.91e-05 ***
x1           1.424e+04  5.003e+03   2.847  0.00507 **
x2           1.005e+05  3.958e+04   2.540  0.01217 *
x4           7.919e+04  3.380e+04   2.343  0.02052 *
x5           8.454e+02  7.621e+02   1.109  0.26919
x6          -1.573e-01  1.834e-01  -0.858  0.39245
x8          -6.513e+04  5.268e+04  -1.236  0.21836
x10          1.481e+00  3.461e-01   4.279 3.42e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 533100 on 143 degrees of freedom
Multiple R-squared:  0.253,     Adjusted R-squared:  0.2164
F-statistic: 6.919 on 7 and 143 DF,  p-value: 4.311e-07
```

- Mặc dù Adjusted R-squared tăng lên nhưng các biến x5,x6,x8 vẫn không có ý nghĩa thống kê vì giá trị p-value lớn hơn 0.05.
- Vì giá trị Adjusted R-squared nhỏ nên em thử với mô hình sau:

```
> model = lm(y ~ x1 + x2 + x4 + x5 + x6 + poly(x8 + x10,3) )
> summary(model)

Call:
lm(formula = y ~ x1 + x2 + x4 + x5 + x6 + poly(x8 + x10, 3))

Residuals:
     Min       1Q   Median       3Q      Max
-1045890  -421686   -45351   425183  1192998

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -4.724e+05  2.283e+05  -2.069  0.04035 *
x1                  1.042e+04  4.707e+03   2.213  0.02851 *
x2                  1.192e+05  3.777e+04   3.155  0.00196 **
x4                  5.506e+04  3.230e+04   1.704  0.09048 .
x5                  9.907e+02  7.345e+02   1.349  0.17957
x6                 -5.075e-01  1.760e-01  -2.883  0.00455 **
poly(x8 + x10, 3)1  3.617e+06  6.674e+05   5.420  2.5e-07 ***
poly(x8 + x10, 3)2 -1.851e+06  5.615e+05  -3.296  0.00124 **
poly(x8 + x10, 3)3  1.349e+06  5.173e+05   2.609  0.01007 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 508500 on 142 degrees of freedom
Multiple R-squared:  0.3251,    Adjusted R-squared:  0.2871
F-statistic:  8.55 on 8 and 142 DF,  p-value: 1.829e-09
```

- Giá trị Adjusted R-squared tăng lên nhiều nhưng x5 và x4 không có ý nghĩa thống kê (giá trị p-value lớn hơn 0.05), em thử bỏ biến x5, xây dựng lại mô hình.

```
> model = lm(y ~ x1 + x2 + x4 + x6 + poly(x8 + x10,3) )
> summary(model)

Call:
lm(formula = y ~ x1 + x2 + x4 + x6 + poly(x8 + x10, 3))

Residuals:
     Min       1Q   Median       3Q      Max
-1039485  -405001   -30671   398888  1235203

Coefficients:
                     Estimate Std. Error t value Pr(>|t|)
(Intercept)        -5.158e+05  2.267e+05  -2.275  0.02437 *
x1                  1.246e+04  4.468e+03   2.790  0.00600 **
x2                  1.269e+05  3.743e+04   3.391  0.00090 ***
x4                  6.303e+04  3.185e+04   1.979  0.04975 *
x6                 -4.396e-01  1.692e-01  -2.599  0.01034 *
poly(x8 + x10, 3)1  3.568e+06  6.683e+05   5.339 3.59e-07 ***
poly(x8 + x10, 3)2 -1.730e+06  5.558e+05  -3.112  0.00224 **
poly(x8 + x10, 3)3  1.398e+06  5.175e+05   2.702  0.00772 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 510000 on 143 degrees of freedom
Multiple R-squared:  0.3164,    Adjusted R-squared:  0.283
F-statistic: 9.457 on 7 and 143 DF,  p-value: 1.288e-09
```

- Các biến độc lập x đều có ý nghĩa thống kê.
- Tiếp theo em thực hiện tính thủ công các giá trị có trong bảng phân tích bằng Excel.

III. Tính thủ công các giá trị trong bảng kết quả bằng Excel

1 Tính các giá trị trong bảng SUMMARY:

Thực hiện thêm các cột dữ liệu như phần phân tích trên vào Excel

x1	x2	x4	x6	(x8+x10)	(x8+x10)^2	(x8+x10)^3	y	y hồi quy
25	4	3	100000	135001	18225270001	2.46043E+15	250	37216.402
39	6	5	1800000	362505	1.3141E+11	4.76367E+16	800	554973.06
25	4	3	240000	85001	7225170001	6.14147E+14	150	-320443.2
25	1	5	100000	200000	40000000000	8E+15	75	84729.689
25	7	5	1330000	257147.8571	66125020411	1.70039E+16	800	501494.57
39	7	4	1800000	353576.4286	1.25016E+11	4.42028E+16	800	607548.06
39	6	4	1800000	412505	1.7016E+11	7.0192E+16	800	535350.38
25	4	2	675000	225001	50625450001	1.13908E+16	350	116996.25
39	5	4	930000	360003	1.29602E+11	4.66572E+16	800	744349.45
25	5	3	930000	495001	2.45026E+11	1.21288E+17	800	559635.3
39	2	4	240000	270001	72900540001	1.96832E+16	250	493101
39	4	5	440000	168752	28477237504	4.80559E+15	350	356425.96
39	2	5	675000	465001	2.16226E+11	1.00545E+17	930000	597583.79
39	4	4	930000	450003	2.02503E+11	9.11268E+16	800	675953.38
25	1	5	100000	240000	57600000000	1.3824E+16	150	228198.21
39	4	3	240000	110002	12100440004	1.33107E+15	250	9375.7782
39	4	3	1800000	450003	2.02503E+11	9.11268E+16	800	230514.63

Các bước thực hiện:

- Vào mục Data > Chọn Data Analysis > Chọn Regression
- Chọn Y Range là cột y và X Range là 7 cột còn lại.

Phân tích hồi qui bằng Excel, em có kết quả:

Regression Sta	Cột1
Multiple R	0.562533616
R Square	0.316444069
Adjusted R Squa	0.282983289
Standard Error	509952.1107
Observations	151

ANOVA

	df	SS	MS	F	Significance F
Regression	7	1.72154E+13	2.45935E+12	9.457163586	1.28834E-09
Residual	143	3.71873E+13	2.60051E+11		
Total	150	5.44027E+13			

Cột1	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%
Intercept	-1867486.595	329914.0199	-5.660525114	7.97412E-08	-2519625.064	-1215348.13	-2519625.064	-1215348.126
x1	12464.47632	4468.20565	2.789593251	0.005997202	3632.209078	21296.74357	3632.209078	21296.74357
x2	126942.142	37434.09662	3.391083356	0.000900318	52946.45526	200937.8287	52946.45526	200937.8287
x4	63028.81525	31850.54026	1.978893128	0.049748852	70.10081937	125987.5297	70.10081937	125987.5297
x6	-0.439551648	0.169154377	-2.598523636	0.010343186	-0.773917786	-0.10518551	-0.773917786	-0.10518551
(x8+x10)	8.687769954	2.119267889	4.09942037	6.91501E-05	4.498629652	12.87691026	4.498629652	12.87691026
(x8+x10)^2	-1.35743E-05	4.48778E-06	-3.024734671	0.002951085	-2.24453E-05	-4.7034E-06	-2.24453E-05	-4.70338E-06
(x8+x10)^3	5.98664E-12	2.21544E-12	2.702236155	0.00772129	1.6074E-12	1.03659E-11	1.6074E-12	1.03659E-11

Em thực hiện tính lại bằng công thức như sau:

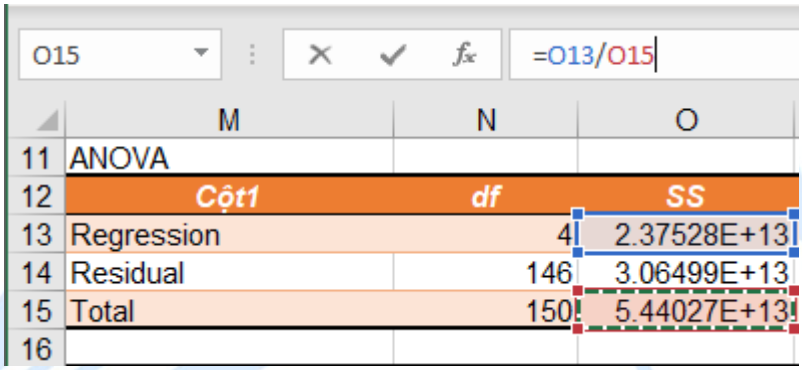
1.1 Bảng 1

Regression Statistics	
Multiple R	0.562533616

R Square	0.316444069
Adjusted R Square	0.282983289
Standard Error	509952.1107
Observations	151

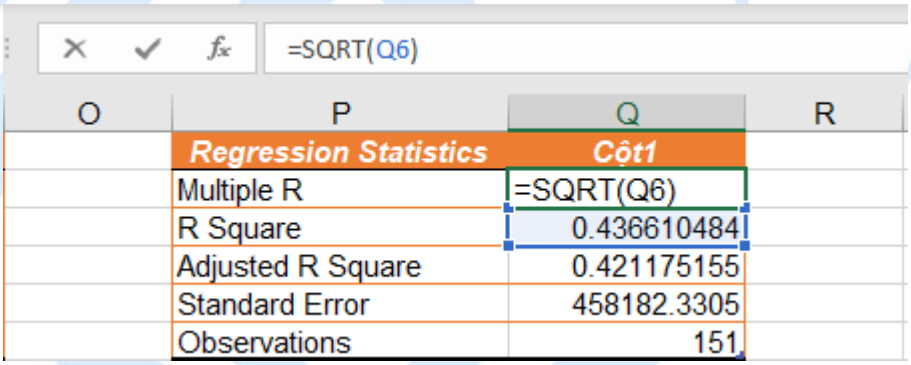
- Giá trị R Square:

  R Square = SSR/SST = 0.316444069



- Giá trị Multiple R

  R = SQRT(R Square) = 0.562533616



- Giá trị Adjusted R Square:

- *Áp dụng công thức:*

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-(k+1))}{SST/(n-1)}$$

- Từ công thức trên, biến đổi và suy ra:
  - Adjusted R Square = 1 – (MSE/MST) = 0.282983289

- Giá trị Standard Error

$$\sqrt{\frac{SSE}{n-k}}$$

Công thức tương đương với:

Standard Error = SQRT(MSE) = 509952.1107

- Giá trị Observations

Có 151 dòng dữ liệu đang xét

1.2 Bảng 2

Công thức: (Theo slide Tiếng Việt của thầy Thuân)

Nguồn	*Df*	*SS*	*MS*
Hồi quy	4	$SSR = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$	*MSR = SSR / 4*
Số dôi	*n - 5*	$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	*MSE = SSE / (n-5)*
Tổng	*n - 1*	$SSTO = \sum_{i=1}^{n} (y_i - \bar{y})^2$	

Dựa vào công thức,, em lần lượt tính từng cột:

- Cột 1:

- Số dòng dữ liệu: n = 151
- Số cột dữ liệu: m = 5
- df của Regression: (m - 1) = 6 – 1 = 5
- df của Residual: (n – 5) = 151 – 6 = 145
- df của Toeml: (n – 1) = 151 – 1 = 150

ANOVA	
Cột1	*df*
Regression	5
Residual	145
Total	150

- Cột 2:

- Dựa vào công thức ở trên
- Gọi tên cột y là y.
- y trung bình = y TB = AVERAGE(y)
- y hồi quy = - 1406919.871 + 9140.650058*x1 - 0.35024374*x2
  + 1.108070403*x3 - 1.41264E-07*x4
- Từ 3 cột trên, em tính lần lượt 3 cột còn lại theo công thức.
- Sau cùng, tính tổng 3 cột cuối. Em tính được SSR, SSE và SST.

y	y TB	y hồi quy	(y hồi quy - y TB)^2	(y - y hồi quy) ^2	(y - y TB)^2
250	445736.09	37216.40197	1.66888E+11	1366514874	1.98458E+11
150		-105848.654	3.04246E+11	11235714645	1.98547E+11
1330000		870645.5918	1.80548E+11	2.11006E+11	7.81923E+11
930000		342461.9496	10665548645	3.45201E+11	2.34512E+11
930000		521938.4606	5806800874	1.66514E+11	2.34512E+11
930000		902819.4413	2.08925E+11	738782770.7	2.34512E+11
1330000		1059090.821	3.76204E+11	73391783107	7.81923E+11
1330000		961733.2241	2.66253E+11	1.3562E+11	7.81923E+11
1800000		864425.6451	1.75301E+11	8.75299E+11	1.83403E+12
930000		471526.2083	665130059.5	2.10198E+11	2.34512E+11
350		425736.4633	399985176.5	1.80954E+11	1.98369E+11
1330000		548046.9285	10467507121	6.11451E+11	7.81923E+11
350		768378.5192	1.04098E+11	5.89868E+11	1.98369E+11
150		68654.01456	1.42191E+11	4692800011	1.98547E+11
250		213013.3927	54159855089	45268261285	1.98458E+11
			SSR	SSE	SST
			1.72154E+13	3.71873E+13	5.44027E+13

- Giá trị tính được trùng khớp với bảng kết quả.

ANOVA

	df	*SS*
Regression	7	1.72154E+13
Residual	143	3.71873E+13
Toeml	150	5.44027E+13

- Cột 3:

  MSR = SSR/(df của Regression) = (2.37528E+13)/4 = 5.9382E+12

  MSE = SSE/(df của Residual) = (3.06499E+13)/146 = 2.09931E+11

ANOVA

	df	*SS*	*MS*
Regression	7	1.72154E+13	2.45935E+12
Residual	143	3.71873E+13	2.60051E+11
Toeml	150	5.44027E+13	

- Cột 4:

  F = MSR/MSE = 5.9382E+12/2.09931E+11 = 28.28644

ANOVA

	df	SS	MS	F
Regression	7	1.72154E+13	2.45935E+12	9.457163586
Residual	143	3.71873E+13	2.60051E+11	
Toeml	150	5.44027E+13		

- Cột 5:

  Significance F = 1 - F.DIST(28.28644 , 4, 146, TRUE) = 2.11774E-17

ANOVA

	df	SS	MS	F	Significance F
Regression	7	1.72154E+13	2.45935E+12	9.457163586	1.28834E-09
Residual	143	3.71873E+13	2.60051E+11		
Toeml	150	5.44027E+13			

1.3 Bảng 3:

- Cột 1:

  Dựa và công thức: (slide Tiếng Việt của giảng viên Nguyễn Đình Thuân)

  - Với các ma trận:

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{k1} \\ 1 & x_{12} & \dots & x_{k2} \\ \dots & \dots & \dots & \dots \\ 1 & x_{1n} & \dots & x_{kn} \end{pmatrix} \quad B = \begin{pmatrix} b_0 \\ b_1 \\ \dots \\ b_k \end{pmatrix} \quad Y = \begin{pmatrix} y_1 \\ y_2 \\ \dots \\ y_n \end{pmatrix}$$

  - Ta có hệ phương trình trước được viết lại:

$$X^T.X.B = X^T.Y$$

$$B=(X^T.X)^{-1} X^T.Y$$

Cột1	Coefficients
Intercept	-1867486.595
x1	12464.47632
x2	126942.142
x4	63028.81525
x6	-0.439551648
(x8+x10)	8.687769954
(x8+x10)^2	-1.35743E-05
(x8+x10)^3	5.98664E-12

- **Bước 1:** Tính ma trận chuyển vị X.T. Dùng PASTE (TRANSPOSE) của excel.
- **Bước 2:** Nhân ma trân X.T*X bằng hàm MMULT() của excel.
- **Bước 3:** Tính ma trận khả nghịch của ma trận ở trên: (X.T*X)^(-1) bằng hàm MMULT() của excel.
- **Bước 4:** Nhân ma trận tính được ở bước 3 với ma trận chuyển vị X.T: ((X.T*X)^(-1))*X.T bằng hàm MINVERSE () của excel
- **Bước 5:** Tinh ma trận ở bước 4 với y: ((X.T*X).T)*X.T * Y bằng hàm MMULT() của excel.
- **Cụ thể:** Kết quả giống với bảng kết luận của excel.

X.T

1	1	1	1	1	1	1	1	1	1	1	1	1
25	39	25	25	25	39	39	25	39	25	39	39	39
4	6	4	1	7	7	6	4	5	5	2	4	2
3	5	3	5	5	4	4	2	4	3	4	5	5
100000	1800000	240000	100000	1330000	1800000	1800000	675000	930000	930000	240000	440000	675000
135001	362505	85001	200000	257147.9	353576.4	412505	225001	360003	495001	270001	168752	465001
18225270001	1.31E+11	7.23E+09	4E+10	6.61E+10	1.25E+11	1.7E+11	5.06E+10	1.3E+11	2.45E+11	7.29E+10	2.85E+10	2.16E+11
2.46043E+15	4.76E+16	6.14E+14	8E+15	1.7E+16	4.42E+16	7.02E+16	1.14E+16	4.67E+16	1.21E+17	1.97E+16	4.81E+15	1.01E+17

X.T*X

151	5844	611	460	77815000	42452355	1.58E+13	8.31E+18
5844	239436	23576	17789	3.02E+09	1.66E+09	6.19E+14	3.25E+20
611	23576	2813	1924	3.55E+08	1.64E+08	5.5E+13	2.28E+19
460	17789	1924	1686	2.57E+08	1.33E+08	4.9E+13	2.36E+19
77815000	3.02E+09	3.55E+08	2.57E+08	6.12E+13	2.55E+13	1.04E+19	5.66E+24
42452355.24	1.66E+09	1.64E+08	1.33E+08	2.55E+13	1.58E+13	8.31E+18	6.67E+24
1.5798E+13	6.19E+14	5.5E+13	4.9E+13	1.04E+19	8.31E+18	6.67E+24	7.49E+30
8.30758E+18	3.25E+20	2.28E+19	2.36E+19	5.66E+24	6.67E+24	7.49E+30	9.94E+36

(X.T*X).T

0.418545576	-0.00251	-0.02343	-0.00582	8.19E-08	-1.9E-06	3.61E-12	-1.7E-18
-0.002509552	7.68E-05	-6.6E-06	1.33E-05	1.17E-10	-3.8E-09	6.66E-15	-3E-21
-0.023433772	-6.6E-06	0.005389	-0.00067	-1.5E-08	5.8E-08	-5.3E-14	1.88E-20
-0.005818215	1.33E-05	-0.00067	0.003901	-9.9E-10	-1.8E-08	1.91E-14	-5E-21
8.19471E-08	1.17E-10	-1.5E-08	-9.9E-10	1.1E-13	-4.4E-13	5.38E-19	-2.1E-25
-1.87985E-06	-3.8E-09	5.8E-08	-1.8E-08	-4.4E-13	1.73E-11	-3.6E-17	1.71E-23
3.61399E-12	6.66E-15	-5.3E-14	1.91E-14	5.38E-19	-3.6E-17	7.74E-23	-3.8E-29
-1.71009E-18	-3E-21	1.88E-20	-5E-21	-2.1E-25	1.71E-23	-3.8E-29	1.89E-35

((X.T*X).T)*X.T

0.060688471	0.01048	0.129557	0.066386	-0.00183	-0.00759	0.023778	0.046262	-0.03172	0.044975	-0.00756	0.011374	0.035394
-0.000964449	7.7E-05	-0.00083	-0.00104	-0.001	5.87E-05	6.47E-05	-0.00106	-3.1E-05	-0.00109	-4.7E-05	0.00011	-1E-05
0.001336173	-0.00726	-0.00316	-0.01345	0.002178	-0.00145	-0.00532	-0.00311	0.001253	0.005128	-0.00708	-0.00383	-0.00917
0.001337382	0.004119	0.001901	0.010359	0.004543	-0.00039	-6E-05	-0.00419	0.001767	-0.00293	0.00514	0.008561	0.007413
-1.80576E-08	8.99E-08	1.37E-08	7.94E-09	3.87E-08	7.68E-08	8.51E-08	2.24E-08	1.08E-08	-2.5E-09	-4.7E-09	9.26E-09	1.7E-08
-1.15732E-07	-1.6E-07	-6.8E-07	1.16E-07	1.38E-07	-6.9E-08	-2.7E-07	2.05E-07	1.86E-07	-2.4E-07	3.16E-07	-9.6E-08	-2.4E-07
1.9274E-13	8.54E-14	1.27E-12	-4.5E-13	-4.6E-13	-3.4E-14	4.31E-13	-5.5E-13	-3.6E-13	8.26E-13	-7.4E-13	1.07E-14	6.05E-13
-8.2575E-20	-8.2E-21	-5.8E-19	2.4E-19	2.5E-19	4.09E-20	-1.9E-19	2.79E-19	1.67E-19	-4.5E-19	3.61E-19	2.68E-20	-3.2E-19

((X.T*X).T)*X.T * Y

-1867486.595
12464.47632
126942.142
63028.81525
-0.439551648
8.687769954
-1.35743E-05
5.98664E-12

- Cột 2:

Cột1	Coefficients	Standard Error
Intercept	-1867486.595	329914.0199
x1	12464.47632	4468.20565
x2	126942.142	37434.09662
x4	63028.81525	31850.54026
x6	-0.439551648	0.169154377
(x8+x10)	8.687769954	2.119267889
(x8+x10)^2	-1.35743E-05	4.48778E-06
(x8+x10)^3	5.98664E-12	2.21544E-12

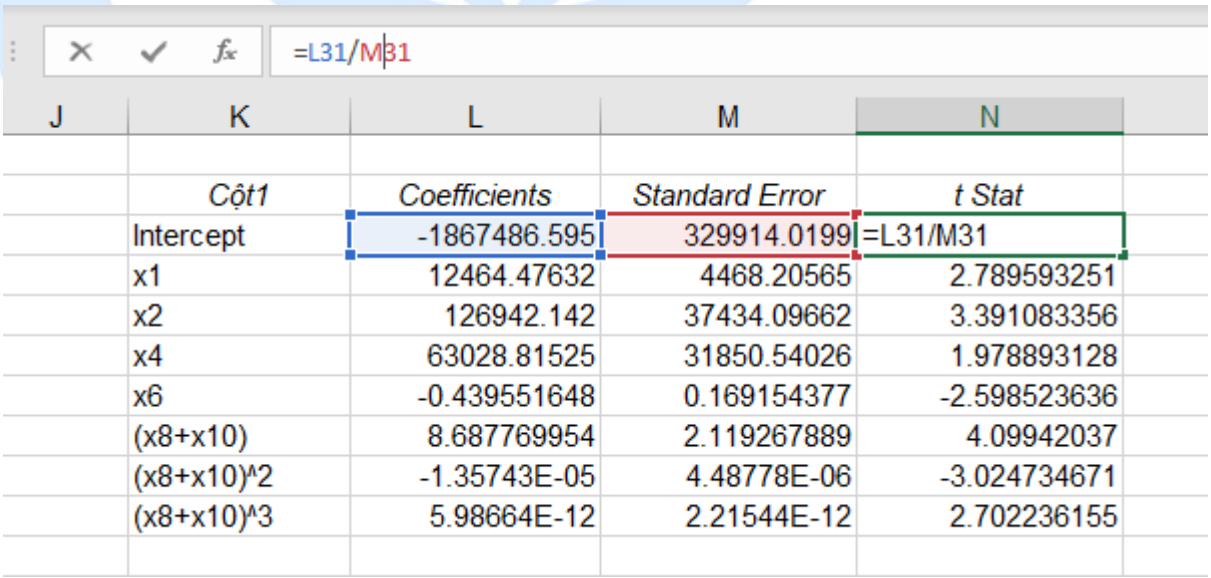
Standard Error of Coefficient (sai số chuẩn của hệ số hồi quy) được tính bằng cách nhân ma trận gồm các giá trị đường chéo của ma trận $(X^TX)^{-1}$ với MSE

- Cột 3:

- *Dựa vào công thức: t Stat = Coefficients/ Standard Error*

Cột1	Coefficients	Standard Error	t Stat
Intercept	-1867486.595	329914.0199	-5.660525114
x1	12464.47632	4468.20565	2.789593251
x2	126942.142	37434.09662	3.391083356
x4	63028.81525	31850.54026	1.978893128
x6	-0.439551648	0.169154377	-2.598523636
(x8+x10)	8.687769954	2.119267889	4.09942037
(x8+x10)^2	-1.35743E-05	4.48778E-06	-3.024734671
(x8+x10)^3	5.98664E-12	2.21544E-12	2.702236155

- Sau khi tính, em cũng ra được kết quả như bảng trên



- Cột 4

- *Em dùng hàm trong Excel: T.DIST.2T()*

Cột1	*Coefficients*	*Standard Error*	*t Stat*	*P-value*
Intercept	-1867486.595	329914.0199	-5.660525114	7.97412E-08
x1	12464.47632	4468.20565	2.789593251	0.005997202
x2	126942.142	37434.09662	3.391083356	0.000900318
x4	63028.81525	31850.54026	1.978893128	0.049748852
x6	-0.439551648	0.169154377	-2.598523636	0.010343186
(x8+x10)	8.687769954	2.119267889	4.09942037	6.91501E-05
(x8+x10)^2	-1.35743E-05	4.48778E-06	-3.024734671	0.002951085
(x8+x10)^3	5.98664E-12	2.21544E-12	2.702236155	0.00772129

- Sau khi tính, em cũng ra được kết quả như bảng trên

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	1.72154E+13	2.45935E+12	9.457163586	1.28834E-09
Residual	143	3.71873E+13	2.60051E+11		
Total	150	5.44027E+13			

Cột1	*Coefficients*	*Standard Error*	*t Stat*	P-value
Intercept	-1867486.595	329914.0199	-5.660525114	=T.DIST.2T(N31,L15)
x1	12464.47632	4468.20565	2.789593251	0.005997202
x2	126942.142	37434.09662	3.391083356	0.000900318
x4	63028.81525	31850.54026	1.978893128	0.049748852
x6	-0.439551648	0.169154377	-2.598523636	0.010343186
(x8+x10)	8.687769954	2.119267889	4.09942037	6.91501E-05
(x8+x10)^2	-1.35743E-05	4.48778E-06	-3.024734671	0.002951085
(x8+x10)^3	5.98664E-12	2.21544E-12	2.702236155	0.00772129

- Cột 5, 6

- *Dựa vào và công thức tổng quát của khoảng tin cậy:[1]*

estimator ± (reliability coefficient) x (standard error)

- *Ở đây, em có tỷ lệ 95% nên em áp dụng vào bài này theo công thức:*

$$\left[\hat{\beta}_j - 1.96 \times SE(\hat{\beta}_j), \hat{\beta}_j + 1.96 \times SE(\hat{\beta}_j)\right]$$

- *Áp dụng công thức, em có:*
  - ***Lower 95%:*** **Coefficient – 1.96 x Standard Error**
  - ***Upper 95%:*** **Coefficient + 1.96 x Standard Error**

- Kết quả tính toán cũng ra kết quả giống với khi tính bằng Excel.

Cột1	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%
Intercept	-1867486.595	329914.0199	-5.660525114	7.97412E-08	-2519625.064	-1215348.13
x1	12464.47632	4468.20565	2.789593251	0.005997202	3632.209078	21296.74357
x2	126942.142	37434.09662	3.391083356	0.000900318	52946.45526	200937.8287
x4	63028.81525	31850.54026	1.978893128	0.049748852	70.10081937	125987.5297
x6	-0.439551648	0.169154377	-2.598523636	0.010343186	-0.773917786	-0.10518551
(x8+x10)	8.687769954	2.119267889	4.09942037	6.91501E-05	4.498629652	12.87691026
(x8+x10)^2	-1.35743E-05	4.48778E-06	-3.024734671	0.002951085	-2.24453E-05	-4.7034E-06
(x8+x10)^3	5.98664E-12	2.21544E-12	2.702236155	0.00772129	1.6074E-12	1.03659E-11

IV. Giải thích các giá trị trong bảng kết quả

1. Bảng kết quả khi sử dụng ngôn ngữ R:

```
Call:
lm(formula = y ~ x1 + x2 + x4 + x6 + poly(x8 + x10, 3))

Residuals:
    Min      1Q  Median      3Q     Max
-1039485  -405001   -30671   398888  1235203

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)       -5.158e+05  2.267e+05  -2.275  0.02437 *
x1                 1.246e+04  4.468e+03   2.790  0.00600 **
x2                 1.269e+05  3.743e+04   3.391  0.00090 ***
x4                 6.303e+04  3.185e+04   1.979  0.04975 *
x6                -4.396e-01  1.692e-01  -2.599  0.01034 *
poly(x8 + x10, 3)1 3.568e+06  6.683e+05   5.339 3.59e-07 ***
poly(x8 + x10, 3)2 -1.730e+06  5.558e+05  -3.112  0.00224 **
poly(x8 + x10, 3)3 1.398e+06  5.175e+05   2.702  0.00772 **
---
```

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

Residual Standard error: 510000 on 143 degrees of freedom

Multiple R-squared: 0.3164, Adjusted R-squared: 0.283

F-Statistic: 9.457 on 7 and 143 DF, p-value: 1.288e-09

2. Bảng kết quả khi sử dụng phần mềm Excel:

SUMMARY OUTPUT

Regression Statistics	Cột1
Multiple R	0.562533616
R Square	0.316444069
Adjusted R Square	0.282983289
Standard Error	509952.1107
Observations	151

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	7	1.7215E+13	2.4593E+12	9.45716359	1.28834E-09
Residual	143	3.7187E+13	2.6005E+11		
Toeml	150	5.4403E+13			

Cột1	*Coefficients*	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-2E+06	3E+05	-5.661	8E-08	-3E+06	-1215348
x1	12464	4468	2.79	0.006	3632.2	21296.74
x2	126942	37434	3.391	0.0009	52946	200937.8
x4	63029	31851	1.979	0.0497	70.101	125987.5
x6	-0.4396	0.169	-2.599	0.0103	-0.7739	-0.10519
(x8+x10)	8.6878	2.119	4.099	7E-05	4.4986	12.87691
(x8+x10)^2	-1E-05	4E-06	-3.025	0.003	-2E-05	-4.7E-06
(x8+x10)^3	6E-12	2E-12	2.702	0.0077	2E-12	1.04E-11

3. **Giải thích:**

Em có mô hình hồi qui phi tuyến như sau:

energy kWh per person = **- 1867486.595**

+ 12464.47632 * `Age`

+ 126942.142 * `Number of occupants"

+ 63028.81525 * `Number of floor`

- 0.439551648 * `minimum electricity consumption`

+ 8.687769954 * (`electricity water heater` + `energy oeroerson`)

- 1.35743E-05 * (`electricity water heater` + `energy oeroerson`)^2

+ 5.98664E-12 * (`electricity water heater` + `energy oeroerson`)^3

- *Observations bằng 151:* nghĩa là có 151 quan sát hay dòng dữ liệu.
- *Hệ số tương quan Multiple R bằng 0.56:*

  o Giá trị dương -> Biểu đồ hồi quy biến thiên hướng lên.

o Giá trị R nằm trong khoảng [-1; 1]. Trong trường hợp này nằm trong khoảng [0.5; 0.6] -> Mối quan hệ giữa biến phụ thuộc và biến độc lập là mối tương quan trung bình [3].

- *Hệ số xác định R Square bằng 0.3164:*

  o Điều đó có nghĩa là có 31.64% giá trị của biến "*energy kWh per person"* có thể giải thích bằng giá trị của các biến độc lập.

  o Giá trị R square nằm trong khoảng [0;1]. Trường hợp này, nằm trong khoảng gần 0 nên t có thể nói khả năng giải thích giá trị biến phụ thuộc (biến "*energy kWh per person"*) của các biến độc lập (các biến X) không tốt.

- *Giá trị Adjusted R Square 0.28298: [4]*

  o Cả hai R square và adjusted R square cho biết mức độ phù hợp của các mô hình với một đường cong hoặc đường thẳng, nhưng adjusted R square không bị ảnh hưởng bởi số lượng các biến trong một mô hình. Nghĩa là thêm càng nhiều biến vô ích vào một mô hình, adjusted R square sẽ giảm, ngược lại thêm nhiều biến hữu ích, adjusted R square sẽ tăng lên. Adjusted R square sẽ luôn nhỏ hơn hoặc bằng R square .

  o Ở đây, có một sự khác biệt chính giữa R square và adjusted R square: R square giả định rằng mỗi biến duy nhất giải thích sự thay đổi trong các biến phụ thuộc . adjusted R square cho biết tỷ lệ bao nhiêu phần trăm biến phụ thuộc biến thiên mà chỉ được giải thích bởi các biến độc lập thực sự ảnh hưởng đến biến phụ thuộc .

  o Giá trị Adjusted R Square nằm trong khoảng [0,1]. Vì vậy, em có mức độ tốt của mô hình hồi quy là 0.28298

- *Giá trị sai số chuẩn Standard Error bằng 509952.1107: phản ánh mức độ dao động của các quan sát trong một tổng thể.* Từ đây, em có thể kết luận khoảng tin cậy.

o 95% số trung bình tính từ mẫu có giá trị từ x ─ 1.96×SE đến x +1.96×SE, với x tương ứng là các Coefficients tương ứng của nó.

95% số trung bình của Intercept sẽ dao động trong khoảng [*-3E+06;-1215348]*

o 95% số trung bình của Age (x1) sẽ dao động trong khoảng *[3632.2;21296.74]*

o 95% số trung bình của Number of occupants (x2) sẽ dao động trong khoảng *[52946;200937.8]*

o 95% số trung bình của x4 sẽ dao động trong khoảng *[70.101;125987.5]*

o 95% số trung bình của x6 sẽ dao động trong khoảng *[-0.7739;-0.10519]*

o 95% số trung bình của (`electricity water heater` + `energy oeroerson`) sẽ dao động trong khoảng *[4.4986;12.87691]*

o 95% số trung bình của (`electricity water heater` + `energy oeroerson`)^2 sẽ dao động trong khoảng *[-2E-05;-4.7E-06]*

o 95% số trung bình của (`electricity water heater` + `energy oeroerson`)^3 sẽ dao động trong khoảng *[2E-12;1.04E-11]*

- *Giá trị Significance F bằng 1.28834E-09* nhỏ hơn alpha (alpha = 0.05). Chúng em có thể kết luận mô hình này có ý nghĩa thống kê. Nhưng không phải tất cả biến trong mô hình này đều có ý nghĩa thống kê. Đó là lí do, em phải kiểm tra thêm giá trị p-value của của từng biến. Và em có thể thấy tất cả giá trị p-value đều nhỏ hơn 0.05, nghĩa là các biến đều có giá trị thống kê.

- Các biến này ảnh hưởng đến biến phụ thuộc như thế nào, em kiểm tra *giá trị Coefficients.* Giá trị Coefficients biến phụ thuộc dự kiến sẽ tăng bao nhiêu khi một biến độc lập tăng một và giữ tất cả các biến độc lập khác không đổi. Cụ thể:

  o Khi thêm 1 lầu(number of floor) thì năng lượng điện tiêu thụ kWh trên đầu người tăng 63029 kWh.

  o Khi giảm 1 người ở cùng (Number of occupants) thì năng lượng điện tiêu thụ kWh trên đầu người giảm 126942.142 kWh.

- *Giá trị SSR là 1.72154E+13 và giá trị SSE là 3.71873E+13. [5]*

    - *Giá trị SSE lớn trong khi SSR lại nhỏ hơn so với SSE. Điều này có thể hiểu giá trị chúng em dự đoán qua mô hình nằm xa so với giá trị thực tế.*

➔ **Kết luận: Mắc dù mô hình có ý nghĩa thống kê nhưng đây không phải một mô hình thực sự tốt và chỉ 31.64% giá trị của biến "*energy kWh per person*" có thể giải thích bằng giá trị của các biến độc lập.**

TÀI LIỆU THAM KHẢO

[1]https://www.fmu.ac.jp/home/public_h/ebm/report/images/10%2095%20percent%20CI%20and%20P_VN.pdf

[2] Slile bài giảng của giảng viên Nguyễn Đình Thuân

[3] http://bomonnoiydhue.edu.vn/upload/file/lstk12_phantichtuongquan.pdf

[4]https://www.Statisticshowto.com/probability-and-Statistics/Statistics-definitions/adjusted-r2/

[5]https://www.youtube.com/watch?v=Hc3Z1OjYAQA

LOẠI 3 : HỒI QUY LOGISTIC

Chọn bài tập với dữ liệu thực tế (kinh tế, xã hội) tùy chọn của Việt Nam để thực hiện bài toán Hồi qui logistic

1. Dùng MS Excel và ngôn ngữ R thực hiện.

2. Mỗi bài tập cần thực hiện trong báo cáo:

- Phát biểu bài toán, nêu ý nghĩa của bài toán cần giải quyết.

- Tính lại và giải thích các giá trị trong bảng kết quả.

I. Thông tin tập dữ liệu:

SUV là dòng xe thể thao đa dụng viết tắt của cụm từ Sport Utility Vehicle với thiết kế vuông vắn, mạnh mẽ, cơ bắp cùng kết cấu thân trên khung như xe tải và khoảng sáng gầm cao, động cơ mạnh cho khả năng vượt nhiều địa hình, có nội thất rộng rãi cho 5-7 người bao gồm cả hành lý

BẢNG DOANH SỐ CÁC MẪU XE SUV/CROSSOVER THÁNG 2/2021

STT	Mẫu xe	Phân khúc	Chỗ ngồi	Giá tham khảo (triệu đồng)	Sản lượng (chiếc)
1.	Kia Seltos	B	5	599 - 719	1.012
2.	Toyota Corolla Cross	B	5	720 – 910	726
3.	Mazda CX-5	C	5	839 – 1.059	551
4.	Hyundai Santa Fe	D	7	995 – 1.245	412
5.	Kia Sorento	D	7	1.079 – 1.349	306
Nguồn: TC Motor, VAMA, VinFast					

Theo số liệu từ Tổng cục thống kê, Tính đến tháng 7/2018 tổng số ô tô đang lưu hành tại Việt Nam đạt hơn 3 triệu xe. Ô tô được tiêu thụ nhiều nhất tại Hà Nội và TP.HCM. Hai thành phố này chiếm khoảng 45% tổng lượng xe được đăng ký tại Việt Nam hàng năm.

Cụ thể, tính đến năm 2016, các loại xe du lịch đã đăng ký tại TP.HCM chiếm 211.000 xe và Hà Nội là 291.000 xe. Khoảng 600.000 xe còn lại được tiêu thụ rải rác tại các tỉnh thành khách.

Tỷ lệ sở hữu ô tô của người Việt ở mức thấp

Dù tỉ lệ tăng trưởng số lượng ô tô ở Việt Nam ở mức cao nhưng trung bình người Việt vẫn sở hữu ô tô vẫn ít, ở mức 16 xe/1.000 dân. Một con số khá thấp so với các nước trong khu vực như: Malaysia (341 xe/1.000 dân), Thái Lan (196 xe/1.000 dân) và Indonesia (55 xe/1.000 dân).

Trong suốt nhiều tháng của năm 2017, các chương trình ưu đãi, giảm giá liên tục được các hãng ôtô tung ra để thu hút khách hàng. Tuy nhiên, sức mua trên thị trường vẫn rơi vào tình cảnh ảm đạm và thiếu ổn định, thị trường ôtô Việt chỉ tiêu thụ được gần 273.000 xe.

Nguồn dữ liệu:

Logistic-Regression-on-SUV/SUV_Data.csv at master · ShukurShaik/Logistic-Regression-on-SUV · Git

401 lines (401 sloc) | 10.7 KB | Raw | Blame

Search this file...

	User_ID	Gender	Age	EstimatedSalary	Purchased
2	15624510	Male	19	19000	0
3	15810944	Male	35	20000	0
4	15668575	Female	26	43000	0
5	15603246	Female	27	57000	0
6	15804002	Male	19	76000	0
7	15728773	Male	27	58000	0
8	15598044	Female	27	84000	0
9	15694829	Female	32	150000	1
10	15600575	Male	25	33000	0
11	15727311	Female	35	65000	0
12	15570769	Female	26	80000	0
13	15606274	Female	26	52000	0

II. Phát biểu bài toán:

Dữ liệu SUV (Sport Utility Vehicle) này chứa thông tin của khách hàng Giới tính, Độ tuổi, Mức lương ước tính và Chi tiết mua hàng. Bài toán đặt ra tìm ra loại người thích mua SUV dựa vào thông tin khách hàng.

Ý nghĩa: đưa ra đề xuất cải thiện kinh doanh, tăng doanh thu cho cửa hàng và phân tích trên từng thuộc tính và xây dựng mô hình Hồi quy Logistic để dự đoán giá trị.

III. **Thực hiện:** bài báo cáo sẽ gồm các phần sau:

- **Xác định biến độc lập:** *GENDER, AGE , ESTIMATESALARY*
- **Xác định biến phụ thuộc: PURCHASED**
- **Tính thủ công các giá trị trong bảng kết quả bằng excel**
- **Giải thích các giá trị trong bảng kết quả bằng excel và bằng R**

IV. **Tiền xử lí dữ liệu:**

1. **Dữ liệu cột GENDER dạng chữ được chuyển đổi thành dạng số bằng cách sử dụng Các cột trích xuất thống kê thực từ công cụ phân tích dữ liệu Phạm vi dữ liệu.**

Gender#1	Age	Estimated	Purchased
1	19	19,000	0
1	35	20,000	0
0	26	43,000	0
0	27	57,000	0
1	19	76,000	0
1	27	58,000	0
0	27	84,000	0
0	32	150,000	1
1	25	33,000	0
0	35	65,000	0
0	26	80,000	0
0	26	52,000	0
1	20	86,000	0
1	32	18,000	0
1	18	82,000	0
1	29	80,000	0
1	47	25,000	1
1	45	26,000	1
1	46	28,000	1
0	48	29,000	1
1	45	22,000	1
0	47	49,000	1
1	48	41,000	1
0	45	22,000	1
1	46	23,000	1
1	47	20,000	1

2. EXCEL

Bảng Logistic Regression:

Logistic Regression								LL statistics		Covariance matrix				Converge	Classification Table			
		# Iter	20		Alpha	0.05												
	coeff	s.e.	Wald	p-value	exp(b)	lower	upper	LL	-137.922	1.847551	-0.1134	-0.03424	-5.32E-06	1.67E-14		Obs Suc	Obs Fail	Total
intercept	-12.7836	1.359247	88.45292	0	2.81E-06			LL0	-260.786	-0.1134	0.093164	0.001266	1.96E-07	7.93E-15	Pred Suc	104	20	124
Gender#1	0.333843	0.305227	1.196299	0.274063	1.396324	0.767674	2.53978	Chi-sq	245.7297	-0.03424	0.001266	0.000696	7.60E-08	9.41E-13	Pred Fail	39	237	276
Age	0.236969	0.026377	80.70987	0	1.267402	1.203545	1.334648	df	3	-5.32E-06	1.96E-07	7.60E-08	3.00E-11	7.57E-10	Total	143	257	400
Estimated!	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036	1.000026	1.000047	p-value	0						Accuracy	0.727273	0.922179	0.8525
								R-sq (L)	0.471132							0.727273	0.077821	0.8525
								R-sq (CS)	0.458994						Cutoff	0.5		
								R-sq (N)	0.63002									
								AIC	283.8432						AUC	0.927403		
								BIC	299.8091									

Cột p-value cho thấy các biến đều có ý nghĩa thống kê (<0.05).

- Cột coeff chứa các hệ số để lập thành phương trình hồi quy. Cách tính cột coeff:

+ Bước đầu tiên, ta thiết lập hệ số hồi quy cho các biến = -0.0001

M	N	O	P
-12.7836	0.333843	0.236969	3.59E-05
-0.0001	-0.0001	-0.0001	-0.0001

+ Tiếp theo, ta tính hệ số Logit theo công thức

$$Logit = b_0 + b_1 * X_1 + b_2 * X_2 + ... + b_n * X_n$$

D	E	F	G	H
Gender#1	Age	Estimated!	Purchased	logit
1	19	19000	0	-7.26523
1	35	20000	0	-3.77177
0	26	43000	0	-5.07867
0	27	57000	0	-4.33909
1	19	76000	0	-5.55273

+ Bước tiếp theo ta tính hệ số e^Logit

D	E	F	G	H	I	J
Gender#1	Age	Estimated	Purchased	logit	e^logit	like
1	19	19000	0	-7.26523	0.000699	0.999301
1	35	20000	0	-3.77177	0.022494	0.977506
0	26	43000	0	-5.07867	0.00619	0.99381
0	27	57000	0	-4.33909	0.01288	0.98712
1	19	76000	0	-5.55273	0.003862	0.996138
1	27	58000	0	#VALUE!	#VALUE!	#VALUE!
0	27	84000	0	-3.36977	0.033254	0.966746

+ Ta tính tiếp hệ số Probability theo công thức sau:

$$\text{Probability} = \frac{e^{l}}{1+e^{l}}$$

D	E	F	G	H	I	J	K
Gender#1	Age	Estimated	Purchased	logit	e^logit	like	LL
1	19	19000	0	-7.26523	0.000699	0.999301	-0.0007
1	35	20000	0	-3.77177	0.022494	0.977506	-0.02275
0	26	43000	0	-5.07867	0.00619	0.99381	-0.00621
0	27	57000	0	-4.33909	0.01288	0.98712	-0.01296
1	19	76000	0	-5.55273	0.003862	0.996138	-0.00387

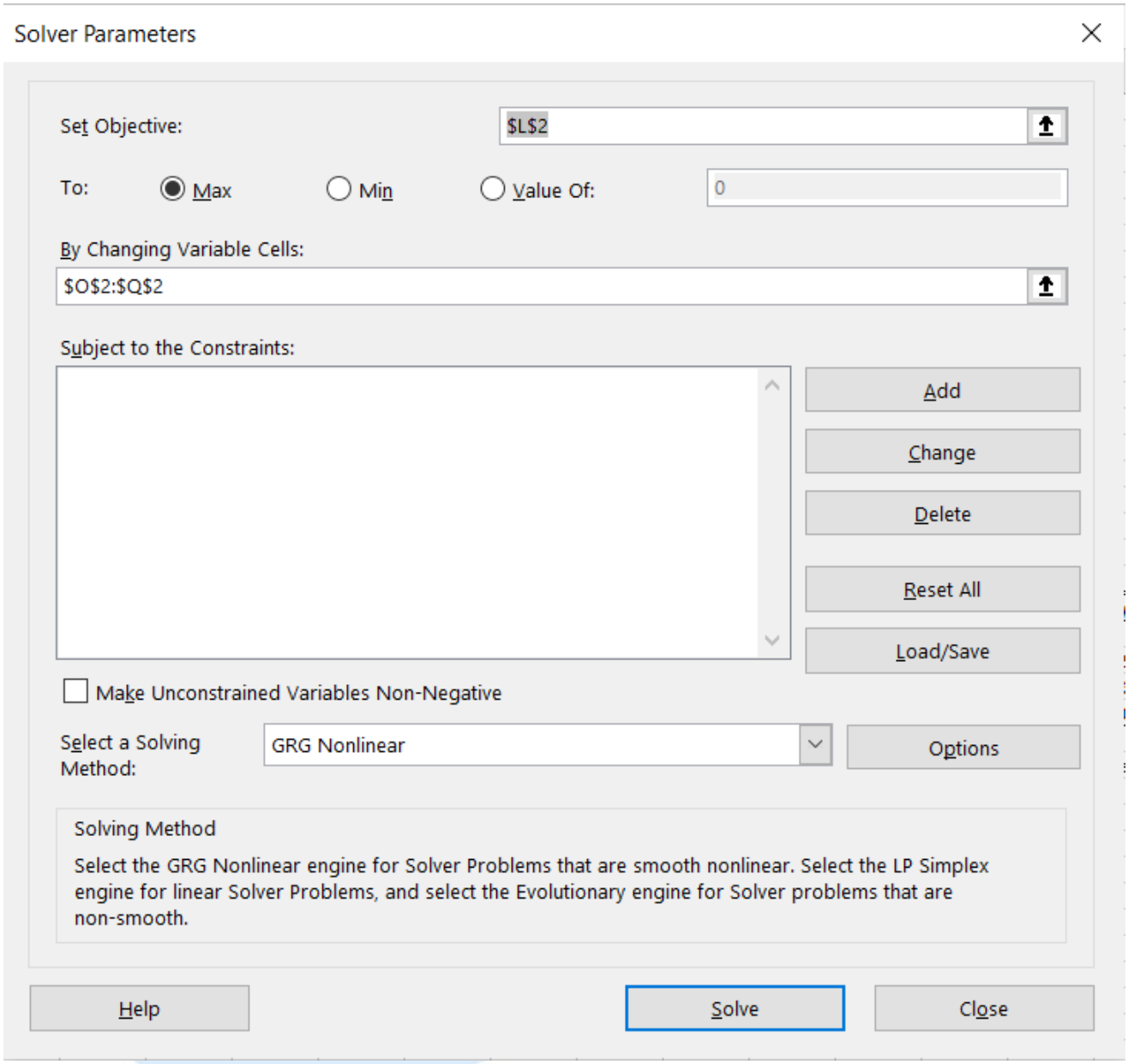
+ Tính Log-Likelihood với công thức:

$$LogLikelihood = y^{*}\ln(pro) + (1-y)^{*}(1-pro)$$

D	E	F	G	H	I	J	K
Gender#1	Age	Estimated	Purchased	logit	e^logit	like	LL
1	19	19000	0	-7.26523	0.000699	0.999301	-0.0007
1	35	20000	0	-3.77177	0.022494	0.977506	-0.02275
0	26	43000	0	-5.07867	0.00619	0.99381	-0.00621
0	27	57000	0	-4.33909	0.01288	0.98712	-0.01296
1	19	76000	0	-5.55273	0.003862	0.996138	-0.00387

+ Sau khi tính Log-Likelihood, ta cộng tổng cột này lại:

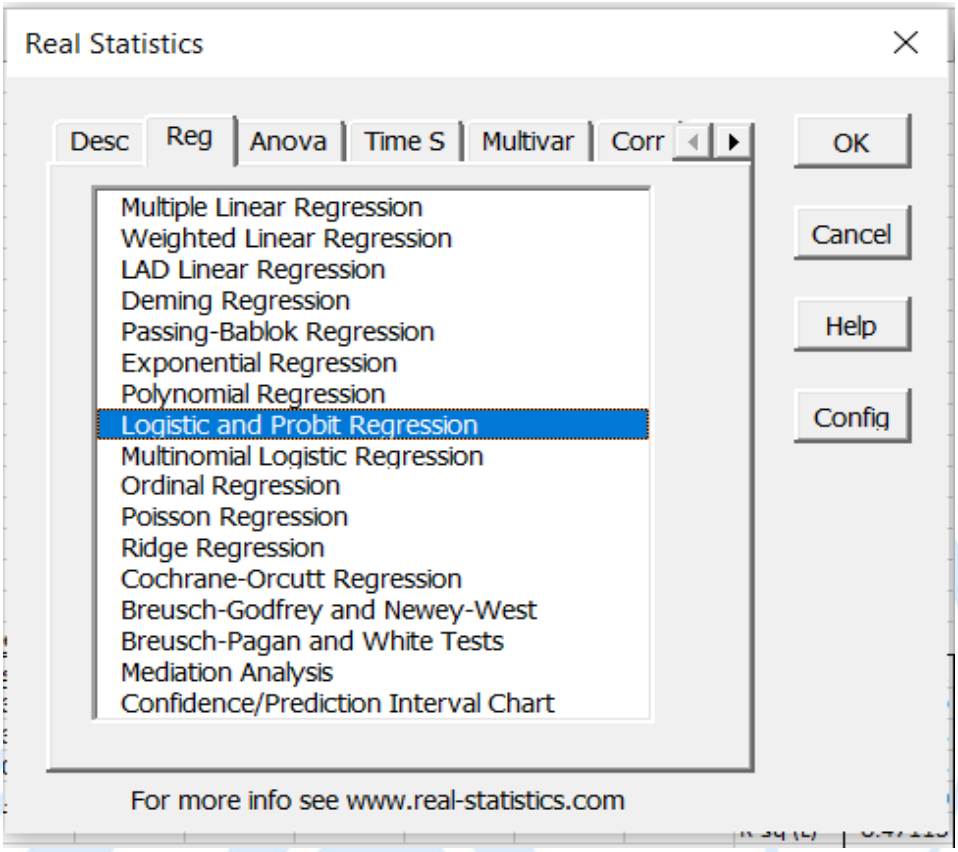
+ Cuối cùng, ta sử dụng tính năng Solver trong excel để giúp ta ước lượng được các hệ số hồi quy: Data/Solver

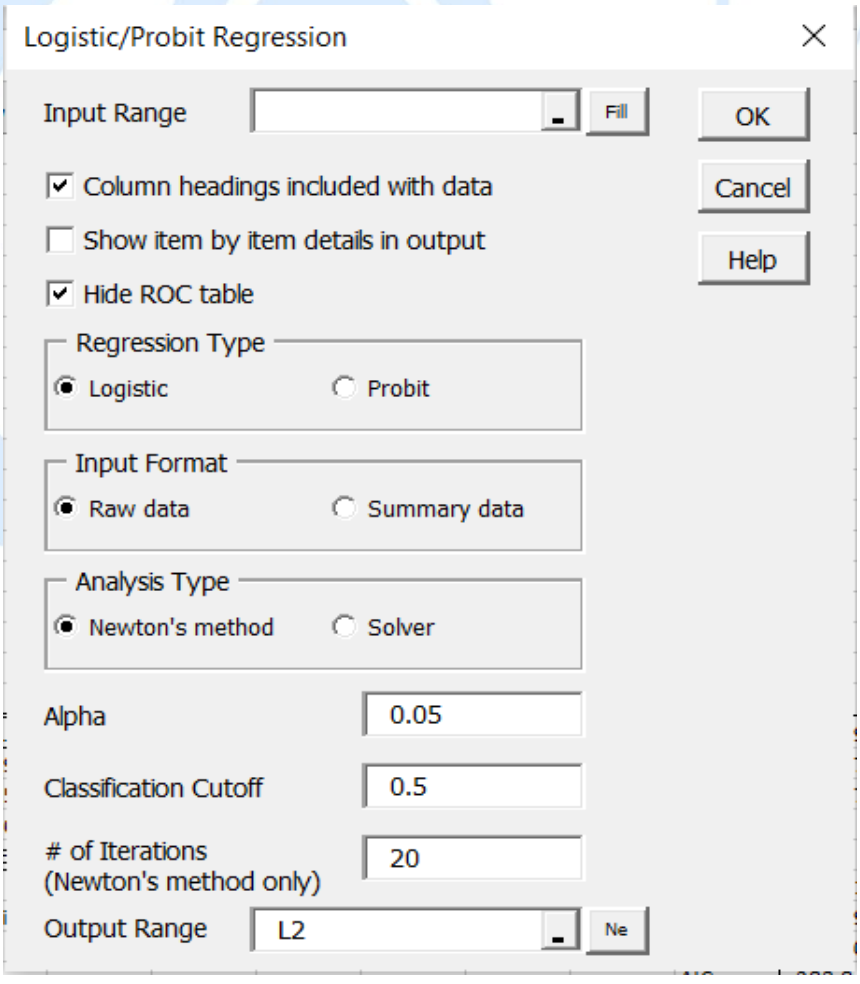


Kết quả :

M	N	O	P
-12.7836	0.333843	0.236969	3.59E-05

2.2 Dùng logistic and probit regression





Logistic Regression								LL statistics		Covariance matrix				Converge	Classification Table			
		# Iter	20		Alpha	0.05												
	coeff	*s.e.*	*Wald*	*p-value*	*exp(b)*	*lower*	*upper*	LL	-137.922	1.84755	-1.13E-01	-0.03424	-5.32E-06	1.7E-14		Obs Suc	Obs Fail	Total
intercept	-12.7836	1.35925	88.4529	0	2.8E-06			LL0	-260.786	-1.13E-01	9.32E-02	0.00127	1.96E-07	7.9E-15	Pred Suc	104	20	124
Gender#1	3.34E-01	0.30523	1.20E+00	0.27406	1.39632	0.76767	2.53978	Chi-sq	245.73	-0.03424	0.00127	0.0007	7.6E-08	9.4E-13	Pred Fail	39	237	276
Age	0.23697	0.02638	80.7099	0	1.2674	1.20354	1.33465	df	3	-5.3E-06	2E-07	7.6E-08	3E-11	7.6E-10	Total	143	257	400
Estimatec	3.6E-05	5.5E-06	44.3357	2.8E-11	1.00004	1.00003	1.00005	p-value	0						Accuracy	0.72727	0.92218	0.8525
								R-sq (L)	0.47113									
ta có mô hình		y=logit=ln(odds)=12.8+3.3E*gender#1+0.24*age+3.6E*estimatedsalary						R-sq (CS)	0.45899						Cutoff	0.5		
								R-sq (N)	0.63002									
								AIC	283.843						AUC	0.9274		
								BIC	299.809									

- **Cột s.e là sai số chuẩn :**

=SQRT(Z8)

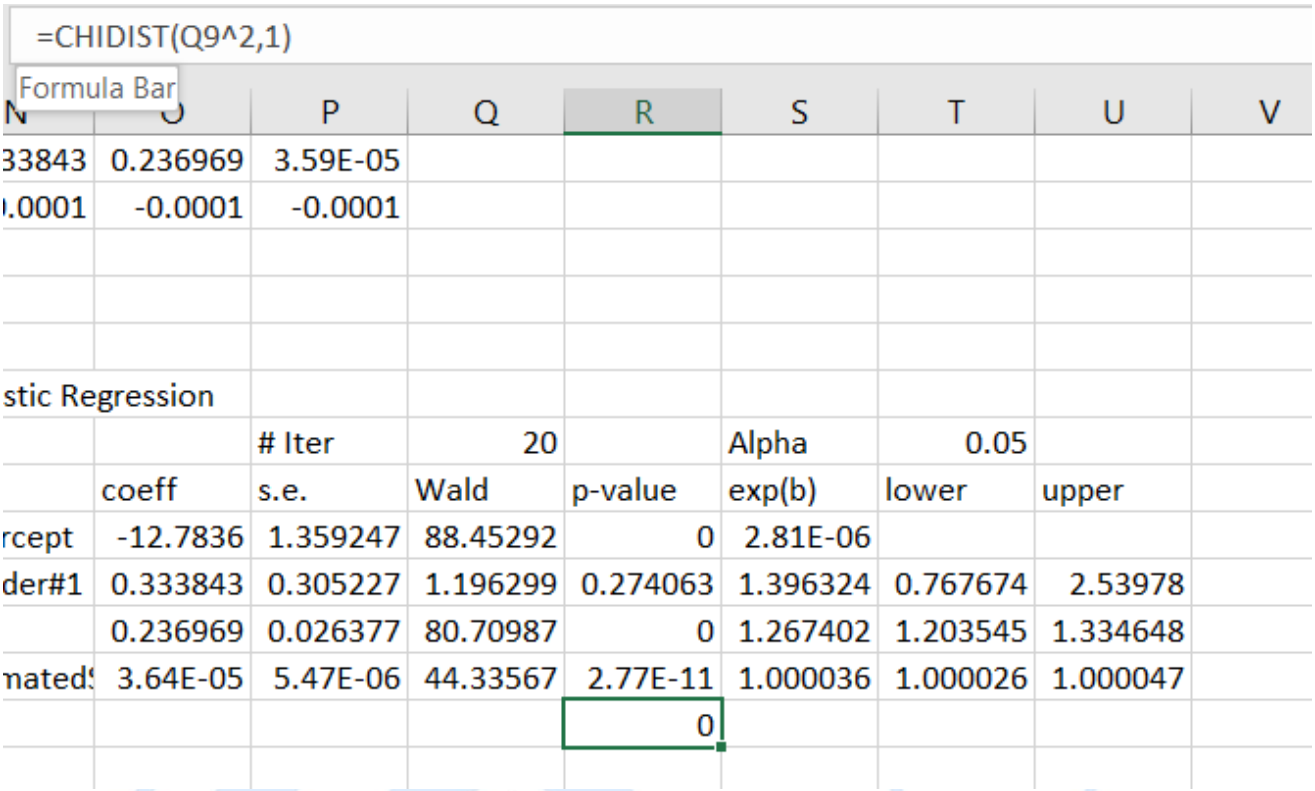
N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC
33843	0.236969	3.59E-05													
.0001	-0.0001	-0.0001													
stic Regression									LL statistics			Covariance matrix			
		# Iter	20		Alpha	0.05									
	coeff	s.e.	Wald	p-value	exp(b)	lower	upper								
rcept	-12.7836	1.359247	88.45292	0	2.81E-06				LL	-137.922		1.847551	-0.1134	-0.03424	-5.32E-06
der#1	0.333843	0.305227	1.196299	0.274063	1.396324	0.767674	2.53978		LL0	-260.786		-0.1134	0.093164	0.001266	1.96E-07
	0.236969	0.026377	80.70987	0	1.267402	1.203545	1.334648		Chi-sq	245.7297		-0.03424	0.001266	0.000696	7.60E-08
nated	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036	1.000026	1.000047		df	3		-5.32E-06	1.96E-07	7.60E-08	3.00E-11
		1.359247							p-value	0					
									R-sq (L)	0.471132					
									R-sq (CS)	0.458994					
									R-sq (N)	0.63002					
									AIC	283.8432					
									BIC	299.8091					

- **Cột wald hệ số kiểm định:**

=(O9/P9)^2

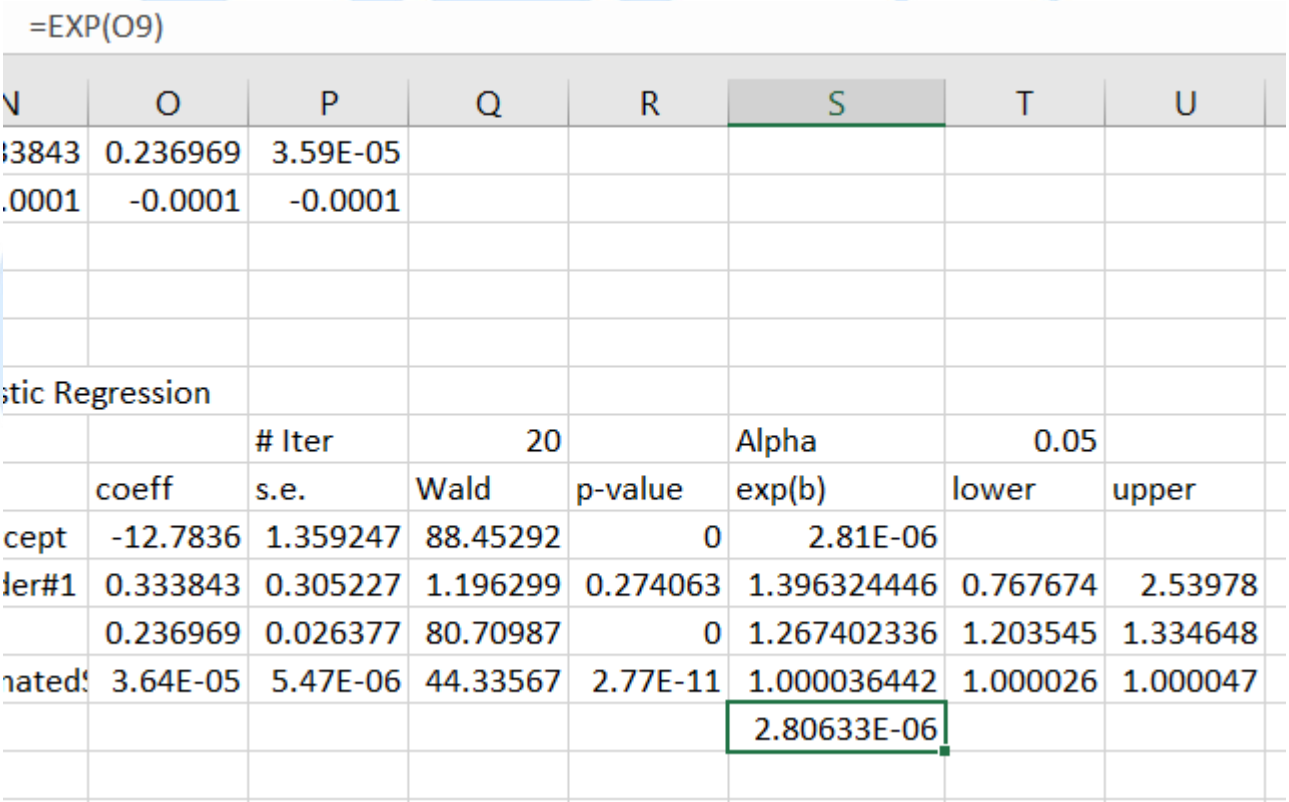
	N	O	P	Q	R	S	T	U
'836	0.333843	0.236969	3.59E-05					
)001	-0.0001	-0.0001	-0.0001					
	Logistic Regression							
			# Iter	20		Alpha	0.05	
		coeff	s.e.	Wald	p-value	exp(b)	lower	upper
	intercept	-12.7836	1.359247	88.45292	0	2.81E-06		
	Gender#1	0.333843	0.305227	1.196299	0.274063	1.396324	0.767674	2.53978
	Age	0.236969	0.026377	80.70987	0	1.267402	1.203545	1.334648
	Estimated	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036	1.000026	1.000047
				88.45292				

- **Cột p-value cho thấy các biến đều có ý nghĩa thống kê :**



33843	0.236969	3.59E-05						
.0001	-0.0001	-0.0001						
stic Regression								
		# Iter	20		Alpha	0.05		
	coeff	s.e.	Wald	p-value	exp(b)	lower	upper	
rcept	-12.7836	1.359247	88.45292	0	2.81E-06			
der#1	0.333843	0.305227	1.196299	0.274063	1.396324	0.767674	2.53978	
	0.236969	0.026377	80.70987	0	1.267402	1.203545	1.334648	
nated	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036	1.000026	1.000047	
				0				

- **Cột exp(b):**



3843	0.236969	3.59E-05					
.0001	-0.0001	-0.0001					
stic Regression							
		# Iter	20		Alpha	0.05	
	coeff	s.e.	Wald	p-value	exp(b)	lower	upper
cept	-12.7836	1.359247	88.45292	0	2.81E-06		
der#1	0.333843	0.305227	1.196299	0.274063	1.396324446	0.767674	2.53978
	0.236969	0.026377	80.70987	0	1.267402336	1.203545	1.334648
nated	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036442	1.000026	1.000047
					2.80633E-06		

- **Cột lower: tỉ lệ ước lượng hệ số hồi quy**

=EXP(O10-P10*NORMSINV(1-0.05/2))

N	O	P	Q	R	S	T	U	V
3843	0.236969	3.59E-05						
.0001	-0.0001	-0.0001						
stic Regression								
		# Iter	20		Alpha	0.05		
	coeff	s.e.	Wald	p-value	exp(b)	lower	upper	
cept	-12.7836	1.359247	88.45292	0	2.81E-06			
der#1	0.333843	0.305227	1.196299	0.274063	1.396324446	0.767674	2.53978	
	0.236969	0.026377	80.70987	0	1.267402336	1.203545	1.334648	
nated	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036442	1.000026	1.000047	
						0.767674		

- **Cột upper tỉ lệ ước lượng hệ số hồi quy**

=EXP(O10+P10*NORMSINV(1-0.05/2))

N	O	P	Q	R	S	T	U	V
3843	0.236969	3.59E-05						
0001	-0.0001	-0.0001						
tic Regression								
		# Iter	20		Alpha	0.05		
	coeff	s.e.	Wald	p-value	exp(b)	lower	upper	
cept	-12.7836	1.359247	88.45292	0	2.81E-06			
der#1	0.333843	0.305227	1.196299	0.274063	1.396324446	0.767674	2.53978	
	0.236969	0.026377	80.70987	0	1.267402336	1.203545	1.334648	
nated	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036442	1.000026	1.000047	
							2.53978	

Bảng LL statistic

- **Giá trị LL(log-likelihood): (đã thực hiện tính phía trên (*)) là một trong số các chỉ tiêu để đánh giá sự phù hợp của mô hình hồi quy. LL càng lớn thì mô hình**

càng phù hợp.

=sum(k2:k401)

D	E	F	G	H	I	J	K
der#1	Age	Estimated	Purchased	logit	e^logit	like	LL
1	19	19000	0	-7.26523	0.000699	0.999301	-0.0007
1	35	20000	0	-3.77177	0.022494	0.977506	-0.02275
0	26	43000	0	-5.07867	0.00619	0.99381	-0.00621
0	27	57000	0	-4.33909	0.01288	0.98712	-0.01296
1	19	76000	0	-5.55273	0.003862	0.996138	-0.00387
1	27	58000	0	#VALUE!	#VALUE!	#VALUE!	#VALUE!
0	27	84000	0	-3.36977	0.033254	0.966746	-0.03382
0	32	150000	1	#VALUE!	#VALUE!	#VALUE!	#VALUE!
1	25	33000	0	#VALUE!	#VALUE!	#VALUE!	#VALUE!
0	35	65000	0	#VALUE!	#VALUE!	#VALUE!	#VALUE!
0	26	80000	0	#VALUE!	#VALUE!	#VALUE!	#VALUE!
0	26	52000	0	#VALUE!	#VALUE!	#VALUE!	#VALUE!
1	20	86000	0	-4.95675	0.006987	0.993013	-0.00701
1	32	18000	0	-4.55438	0.010412	0.989588	-0.01047
1	18	82000	0	-5.57429	0.00378	0.99622	-0.00379
1	29	80000	0	-3.03944	0.045676	0.954324	-0.04675
1	47	25000	1	-0.74854	0.32114	0.32114	-1.13588
1	45	26000	1	-1.18657	0.233872	0.233872	-1.45298
1	46	28000	1	-0.8778	0.293633	0.293633	-1.22542
0	48	29000	1	-0.36796	0.409033	0.409033	-0.89396
1	45	22000	1	-1.33018	0.20913	0.20913	-1.5648
0	47	49000	1	0.113082	0.52824	0.52824	-0.6382
1	48	41000	1	0.062845	0.515706	0.515706	-0.66222
0	45	22000	1	-1.33018	0.20913	0.20913	-1.5648
1	46	23000	1	-1.05731	0.257824	0.257824	-1.35548
1	47	20000	1	-0.92804	0.283322	0.283322	-1.26117

- **Cột Chi-sq:**

=-2*X9-(-2*X8)

L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
(LL)	-12.7836	0.333843	0.236969	3.59E-05									
LUE!	-0.0001	-0.0001	-0.0001	-0.0001									
		Logistic Regression									LL statistics		
				# Iter	20		Alpha	0.05					
			coeff	s.e.	Wald	p-value	exp(b)	lower	upper				
		intercept	-12.7836	1.359247	88.45292	0	2.81E-06				LL	-137.922	
		Gender#1	0.333843	0.305227	1.196299	0.274063	1.396324446	0.767674	2.53978		LL0	-260.786	
		Age	0.236969	0.026377	80.70987	0	1.267402336	1.203545	1.334648		Chi-sq	245.7297	245.72974
		Estimated	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036442	1.000026	1.000047		df	3	
									2.53978		p-value	0	
											R-sq (L)	0.471132	
											R-sq (CS)	0.458994	
											R-sq (N)	0.63002	
											AIC	283.8432	
											BIC	299.8091	

- **Cột df: biến độc lập**
- **Cột R-sq(L):**

=1-X8/X9

L	M	N	O	P	Q	R	S	T	U	V	W	X	Y
(LL)	-12.7836	0.333843	0.236969	3.59E-05									
LUE!	-0.0001	-0.0001	-0.0001	-0.0001									
		Logistic Regression									LL statistics		
				# Iter	20		Alpha	0.05					
			coeff	s.e.	Wald	p-value	exp(b)	lower	upper				
		intercept	-12.7836	1.359247	88.45292	0	2.81E-06				LL	-137.922	
		Gender#1	0.333843	0.305227	1.196299	0.274063	1.396324446	0.767674	2.53978		LL0	-260.786	
		Age	0.236969	0.026377	80.70987	0	1.267402336	1.203545	1.334648		Chi-sq	245.7297	
		Estimated	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036442	1.000026	1.000047		df	3	
									2.53978		p-value	0	
											R-sq (L)	0.471132	0.47113205
											R-sq (CS)	0.458994	
											R-sq (N)	0.63002	
											AIC	283.8432	
											BIC	299.8091	

- **Cột AIC:**

=-2*X8+2*((2-1)+X11)

	M	N	O	P	Q	R	S	T	U	V	W	X	Y
LL)	-12.7836	0.333843	0.236969	3.59E-05									
LUE!	-0.0001	-0.0001	-0.0001	-0.0001									
		Logistic Regression									LL statistics		
				# Iter	20		Alpha	0.05					
			coeff	s.e.	Wald	p-value	exp(b)	lower	upper		LL	-137.922	
		intercept	-12.7836	1.359247	88.45292	0	2.81E-06				LL0	-260.786	
		Gender#1	0.333843	0.305227	1.196299	0.274063	1.396324446	0.767674	2.53978		Chi-sq	245.7297	
		Age	0.236969	0.026377	80.70987	0	1.267402336	1.203545	1.334648		df	3	
		Estimated	3.64E-05	5.47E-06	44.33567	2.77E-11	1.000036442	1.000026	1.000047		p-value	0	
									2.53978		R-sq (L)	0.471132	
											R-sq (CS)	0.458994	
											R-sq (N)	0.63002	
											AIC	283.8432	283.843225
											BIC	299.8091	

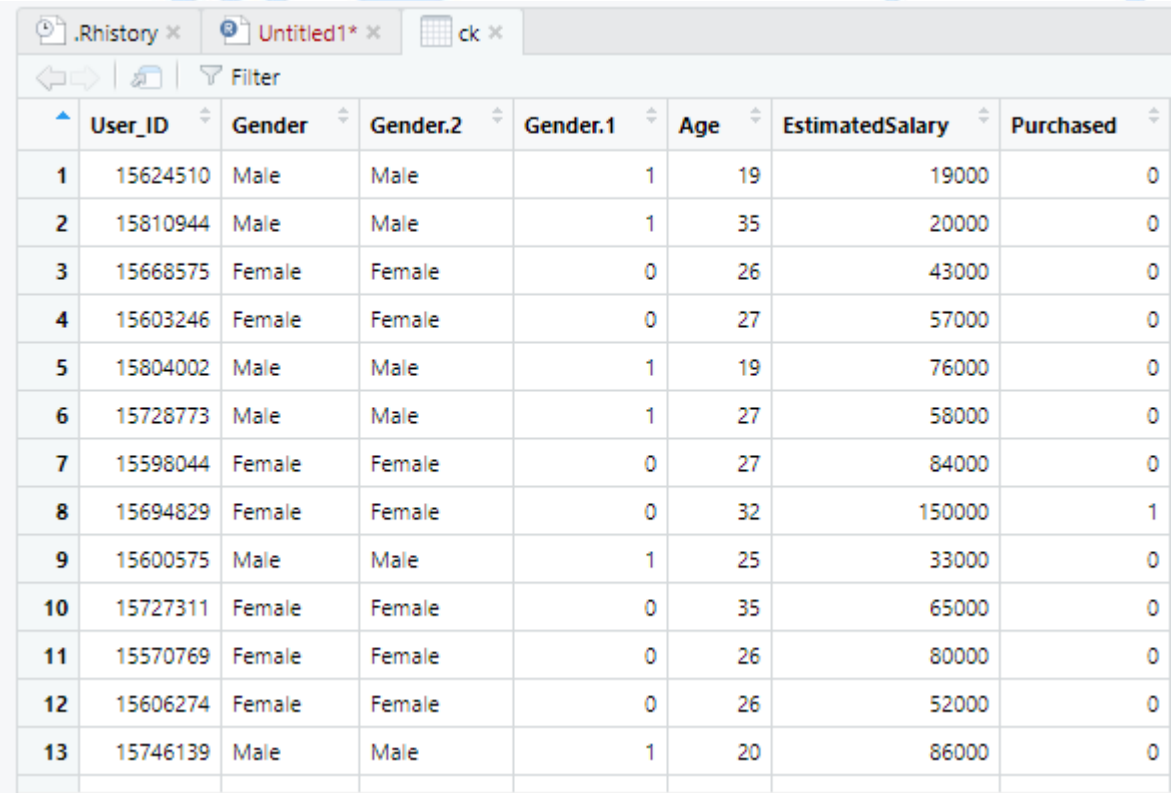
Bảng Classification Table:

- **Với mô hình này , ta dự đoán được 104 trường hợp loại người thích xe SUV, 39 người không thích xe SUV,20 người không mua xe, 237 người mua xe**
- **Cột accuracy :**
  - **0.727273 : xác suất người thích loại xe trong tổng số người thích và k thích**
  - **0.922179 : xác xuất người mua xe trong tổng số người không mua và mua**
  - **0.8525 : xác xuất người thích và người mua trong tổng số dự liệu 400 người**
- **cutoff tham số để tìm các thuộc tính xác xuất**

Classification Table			
	Obs Suc	Obs Fail	Total
Pred Suc	104	20	124
Pred Fail	39	237	276
Total	143	257	400
Accuracy	0.727273	0.922179	0.8525
	0.727273	0.077821	0.8525
Cutoff	0.5		
AUC	0.927403		

3. Dùng R

3.1 import dữ liệu



.Rhistory | Untitled1* | ck

Filter

	User_ID	Gender	Gender.2	Gender.1	Age	EstimatedSalary	Purchased
1	15624510	Male	Male	1	19	19000	0
2	15810944	Male	Male	1	35	20000	0
3	15668575	Female	Female	0	26	43000	0
4	15603246	Female	Female	0	27	57000	0
5	15804002	Male	Male	1	19	76000	0
6	15728773	Male	Male	1	27	58000	0
7	15598044	Female	Female	0	27	84000	0
8	15694829	Female	Female	0	32	150000	1
9	15600575	Male	Male	1	25	33000	0
10	15727311	Female	Female	0	35	65000	0
11	15570769	Female	Female	0	26	80000	0
12	15606274	Female	Female	0	26	52000	0
13	15746139	Male	Male	1	20	86000	0

3.2 Dùng hàm GLM để hồi quy logistic

```
Logistic <- glm(ck$Purchased~ck$Gender.1+ck$Age+ck$EstimatedSalary,data=ck, family = binomial)
summary(logistic)
```

```
Call:
glm(formula = ck$Purchased ~ ck$Gender.1 + ck$Age + ck$EstimatedSalary,
    family = binomial, data = ck)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.9109  -0.5218  -0.1406   0.3662   2.4254

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.278e+01  1.359e+00  -9.405  < 2e-16 ***
ck$Gender.1         3.338e-01  3.052e-01   1.094    0.274
ck$Age              2.370e-01  2.638e-02   8.984  < 2e-16 ***
ck$EstimatedSalary  3.644e-05  5.473e-06   6.659 2.77e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 521.57  on 399  degrees of freedom
Residual deviance: 275.84  on 396  degrees of freedom
AIC: 283.84

Number of Fisher Scoring iterations: 6
```

PHÂN CÔNG CÔNG VIỆC

MSSV	Họ tên	Công việc	Mức độ hoàn thành (%)
18520445	Nguyễn Đức Anh	- Làm loại 1(ANOVA) trên Excel - Viết báo cáo phần giới thiệu, tiền xử lý dữ liệu và phân tích ANOVA, LEVENE, TURKEY - Làm các trang thuyết trình có liên quan - Chạy ANOVA, Turkey trên R - Quay video phần ANOVA trên Excel và R và Turkey trên R	100% 100% 100% 100% 100%
18521320	Đoàn Thục Quyên	- Tìm dữ liệu loại 3 và làm phần tiền xử lý dữ liệu - Viết báo cáo hồi quy Logistic - Làm các trang thuyết trình có liên quan - Chạy hồi quy đa biến trên R - Chạy hồi quy đa biến trên Excel. Tính giá trị và nêu ý nghĩa các đại lượng trong Excel. - Quay video giải thích phần hồi quy đa biến trên powerpoint.	90% 100% 100% 100% 100% 100%
18521554	Nguyễn Thành Trung	- Tìm dữ liệu loại 3 (hồi quy Logistic) và làm phần tiền xử lý dữ liệu - Viết báo cáo hồi quy Logistic - Làm các trang thuyết trình có liên quan - Chạy hồi quy Logistic trên R	80% 80% 80% 80%

		- Chạy hồi quy Logistic trên Excel - Quay video phần hồi quy Logistic trên Excel và R	80% 50%