Báo cáo bài tập 9

1612174 - Phùng Tiến Hào - tienhaophung@gmail.com

09/06/2019

Contents

1 Categorical variables with 2 levels **1**

2 Categorical variables with more than 2 levels **4**

Dữ liệu khảo sát: SpeedDating trong package Lock5withR

Load package và thêm các thư viện cần thiết trước khi đi vào xử lý:

```
require(Lock5withR)
library(Lock5withR)
library(mosaic)

# View data
# View(SpeedDating)
# Avoid using $
attach(SpeedDating)

```

	DecisionM	DecisionF	LikeM	LikeF	PartnerYesM	PartnerYesF	AgeM	AgeF	RaceM	RaceF	AttractiveM	AttractiveF
1	0	0	6.0	7.0	5	5	27	21	Caucasian	Caucasian	6.0	5
2	1	0	8.0	7.0	4	3	22	22	Caucasian	Asian	7.0	6
3	1	0	10.0	6.0	10	2	22	23	Asian	Asian	10.0	4
4	1	1	9.0	7.0	7	8	23	24	Caucasian	Caucasian	9.0	7
5	1	1	7.0	5.0	8	5	24	26	Latino	Other	7.0	5
6	0	0	6.0	6.0	6	1	25	26	Caucasian	Caucasian	NA	5
7	0	1	2.0	6.0	1	5	30	21	Caucasian	Asian	3.0	7
8	0	0	7.0	6.0	7	6	27	23	Caucasian	Caucasian	6.0	5
9	1	1	8.0	7.0	8	10	28	25	Caucasian	Caucasian	8.0	8
10	0	0	5.0	8.0	5	7	24	25	Caucasian	Caucasian	5.0	8
11	0	1	3.0	7.0	1	7	25	25	Asian	Caucasian	5.0	6
12	0	0	7.0	8.0	NA	5	30	23	Asian	Caucasian	10.0	8
13	0	0	8.0	8.0	5	1	23	23	Asian	Black	6.0	8

Figure 1: View data with the first 13 rows.

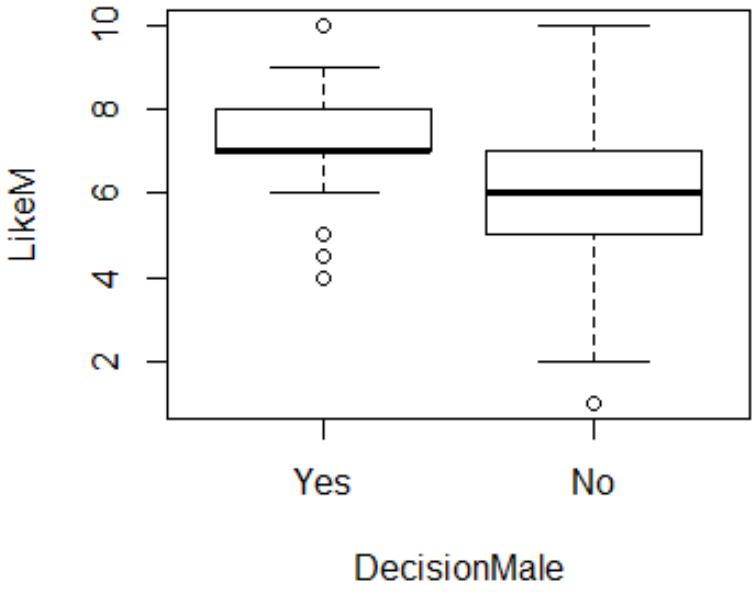
1 Categorical variables with 2 levels

Biến khảo sát: LikeM (Num: 1-10), Decision (Yes/No)
Ta có phương trình tuyến tính như sau: $y = b_0 + b_1 * x$. Với b_0 là hệ số chặn intercept hay còn gọi là bias và b_1 là hệ số slope.

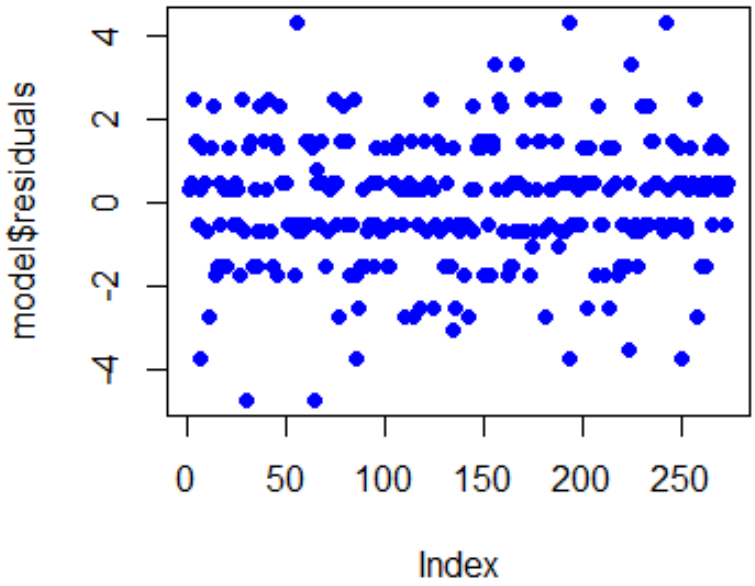
Trước tiên, ta sẽ plot dữ liệu để khảo sát:

```
# Plot data
plot(LikeM~DecisionMale)
# Plot residual to check if data contains pattern or not
plot(model$residuals, pch = 16, col = "blue")

```



(a) Boxplot of LikeM and DecisionMale



(b) Residual plot

Figure 2: Visualize data

Ta nhìn vào hình 2b, ta thấy dữ liệu được phân bố ngẫu nhiên, không chứa pattern ẩn nào.

Do đó, ta có thể yên tâm sử dụng linear regression.

Chúng ta cần kiểm tra xem sự khác biệt về LikeM - mức độ thích của nam đối với nữ giữa phản hồi Yes và No.

Ở đây, ta sẽ tạo ra dummy variables như sau: $x = \begin{cases} 0 : \text{No} \\ 1 : \text{Yes} \end{cases}$

Khi dự đoán LikeM - mức độ thích bằng phương trình hồi qui, ta sẽ có:

- b_0 nếu phản hồi No.
- $b_0 + b_1$ nếu phản hồi Yes

Có thể giải thích hệ số trên như sau:

- b_0 là mức độ thích trung bình của phản hồi No
- $b_0 + b_1$ là mức độ thích trung bình của phản hồi Yes
- b_1 là sự khác biệt về trung bình mức độ thích giữa phản hồi Yes và No.

Kiểm tra sự khác biệt về mức độ thích giữa phản hồi No và Yes bằng việc tính model hồi qui tuyến tính:

```r
# Compute Linear regression model
model <- lm(LikeM~DecisionMale)
summary(model)

```

```
Call:
lm(formula = LikeM ~ DecisionMale)

Residuals:
   Min     1Q Median     3Q    Max
-4.707 -0.707  0.293  1.293  4.293

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       5.7070     0.1351  42.232   <2e-16 ***
DecisionMaleYes   1.8306     0.1851   9.889   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.529 on 272 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.2644,    Adjusted R-squared:  0.2617
F-statistic: 97.78 on 1 and 272 DF,  p-value: < 2.2e-16
```

Figure 3: Summary model

Ta thấy rằng DeisionMaleYes có liên kết mạnh với mức độ thích LikeM khi mà $p-value < 2e-16$. Với mức ý nghĩa 0.05, thì ta hoàn toàn tin tưởng rằng có sự khác biệt về mức độ thích giữa phản hồi Yes và No.

Cụ thể hơn, ta thấy rằng mức độ thích trung bình của phản hồi No ước tính là 5.7070 và của phản hồi Yes là $5.7070 + 1.8306 = 7.537671$.

Để kiểm tra xem model có tốt không thì ta sẽ nhìn vào hệ số R-squared = 0.2644. Tức là model giải thích được 26.44 độ biến thiên của dữ liệu. Có thể thấy rằng model này vẫn chưa tốt, trong thức tế nếu hệ số R-squared này lớn 0.5 là có thể xem là tốt.

Để kiểm tra dummy variables của DecisionMale được phát sinh tự động bởi R:

```
> # Use contrasts() function to return codes that R have used to create dummy var
> contrasts(DecisionMale)
Yes
No   0
Yes  1

```

Ta có thể qui định phản hồi Yes là baseline (tức là bằng 0) và No là bằng 1.

```
# We can specify the baseline to Yes by function relevel()
SpeedDating2 <- SpeedDating %>% mutate(DecisionMale = relevel(DecisionMale, ref = "Yes"))

```

Tiếp đến, ta sẽ tính lại các hệ số của hồi qui:

```
model2 <- lm(LikeM~DecisionMale, data = SpeedDating2)
summary(model2)

```

```
Call:
lm(formula = LikeM ~ DecisionMale)

Residuals:
   Min     1Q Median     3Q    Max
-4.707 -0.707  0.293  1.293  4.293

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.5377     0.1265  59.571   <2e-16 ***
DecisionMaleNo -1.8306     0.1851  -9.889   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.529 on 272 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.2644,    Adjusted R-squared:  0.2617
F-statistic: 97.78 on 1 and 272 DF,  p-value: < 2.2e-16
```

Figure 4: Summary model

Ở đây, ta thấy hệ số của DecisionMaleNo là $-1.8306 < 0$ do đó nếu là phản hồi No thì mức độ thích sẽ giảm.

Khi đó, ước tính cho DecisionMaleYes là $b_0 = 7.5377$ và cho DecisionMaleNo là $b_0 + b_1 = 5.707031$. Ta thấy kết quả ước tính vẫn không thay đổi.

2 Categorical variables with more than 2 levels

Ở đây, ta sẽ thực hiện kiểm tra đối với biến định tính có từ 2 levels trở lên. Cụ thể, ta sẽ thực hiện khảo sát mức độ thích LikeM bởi các biến giải thích: AttractiveM - mức độ quyến rũ (Num: 0-10), DecisionMale - Quyết định làm quen của nam (Yes, No), RaceF - chủng tộc của

nữ (Black, Asian, ...) + SincereM - mức độ chân thành của nữ (Num: 0-10).

Trước tiên, việc phải làm là plot dữ liệu

```
plot(LikeM~AttractiveM + DecisionMale + RaceF + SincereM)

```

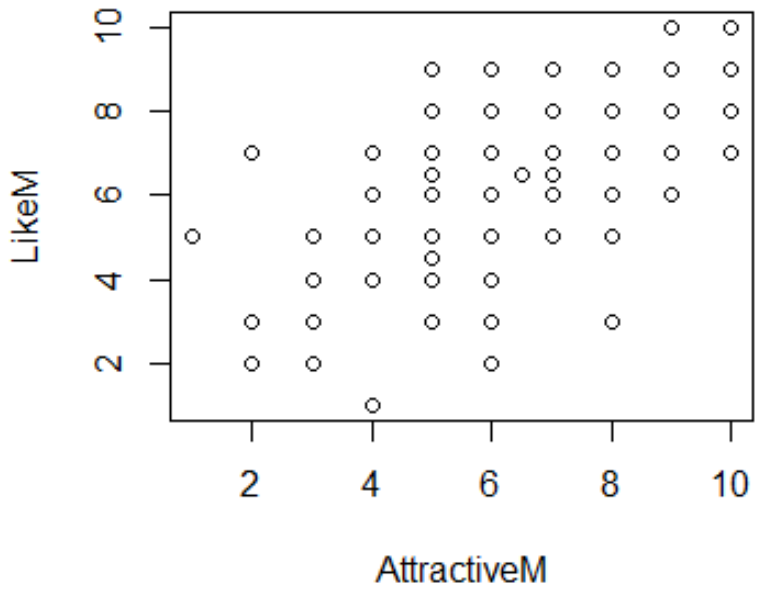


Figure 5: Scatter plot của LikeM và AttractiveM

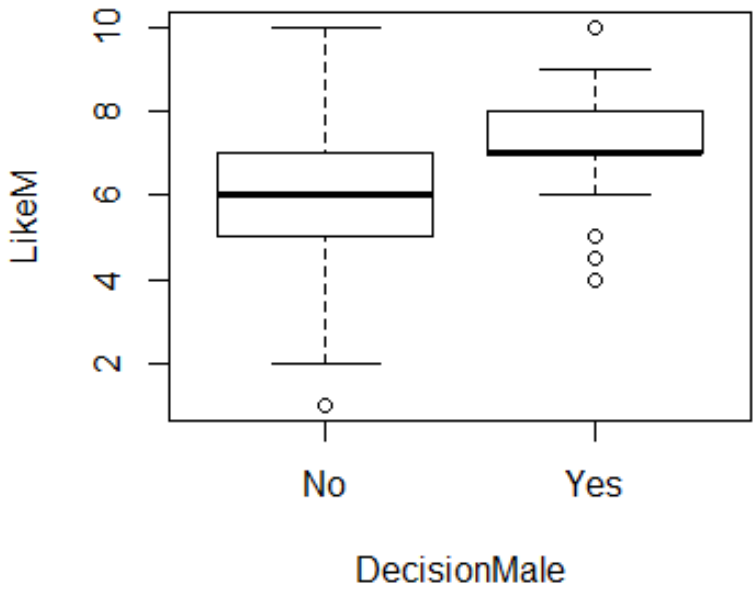


Figure 6: Boxplot của LikeM và DecisionMale

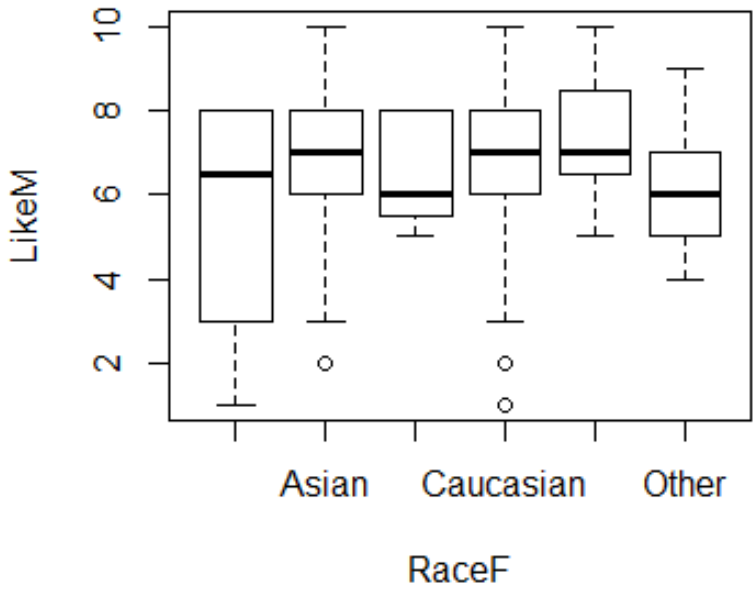


Figure 7: Boxplot của Like và RaceF

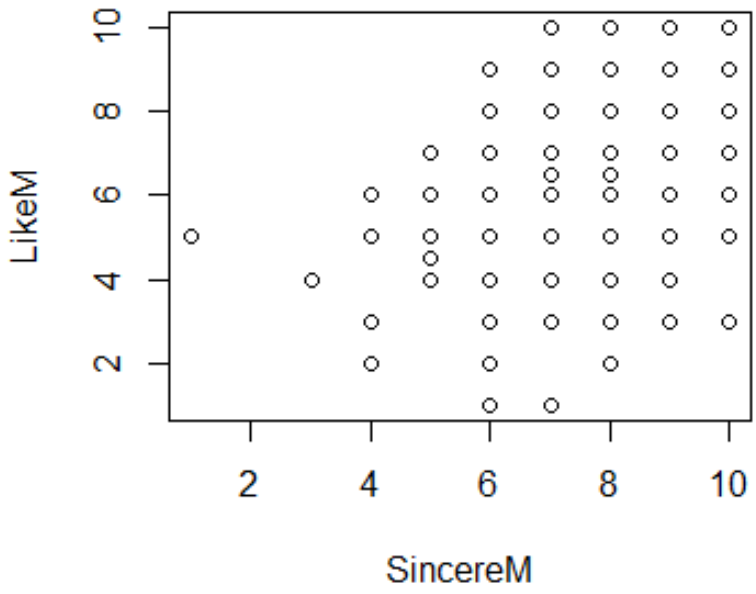


Figure 8: Scatterplot của LikeM và SincereM

Ở đây, nếu một biến định tính có 6 level như RaceF thì ta sẽ tạo dummy code như sau:

- 000000: None (Trường bị bỏ trống trong dữ liệu)
- 000001: Asian
- 000010: Black
- 000100: Caucasian
- 001000: Latino
- 010000: Other

Để phát sinh các dummy code tự động trong R ta có thể dùng hàm model.matrix():

```
# Check dummy code of RaceF
> dummy_code <- model.matrix(~RaceF)
> head(dummy_code[, -1], 6)

```

```
  RaceFAsian RaceFBlack RaceFCaucasian RaceFLatino RaceFOther
1          0          0              1          0          0
2          1          0              0          0          0
3          1          0              0          0          0
4          0          0              1          0          0
5          0          0              0          0          1
6          0          0              1          0          0
```

Figure 9: Dummy coding của biến RaceF

Tiếp đến, ta sẽ đi phân tích phương sai của dữ liệu cũng như kiểm định mối liên kết giữa các biến bằng Anova:

```
model <- lm(LikeM~AttractiveM + DecisionMale + RaceF + SincereM)
anova(model)

```

```
> anova(model)
Analysis of Variance Table

Response: LikeM
             Df Sum Sq Mean Sq  F value    Pr(>F)
AttractiveM   1 458.67  458.67 380.6784 < 2.2e-16 ***
DecisionMale  1  19.02   19.02  15.7820 9.200e-05 ***
RaceF         5   7.52    1.50   1.2479    0.2871
SincereM      1  60.64   60.64  50.3275 1.223e-11 ***
Residuals   261 314.47    1.20
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 10: Analyst of variance for linear regression model

Ta thấy rằng, RaceF liên kết không đáng kể với sự biến thiên dữ liệu của LikeM khi mà $p - value = 0.2871 > alpha = 0.05$. Nghĩa là ta có 28.71% mà việc dữ đoán của model sẽ không có ý nghĩa. Trong đây, ta chỉ thấy rằng 2 biến DecisionMale, SincereM và AttractiveM có liên kết đáng kể nhất. Cụ thể hơn AttractiveM có $p - value = 2.2e - 16$ rất bé cho thấy nó là thành phần cực tốt cần đưa vào model.

Ta sẽ xem qua các hệ số của model:

```
summary(model)

```

```
Call:
lm(formula = LikeM ~ AttractiveM + DecisionMale + RaceF + SincereM)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3811 -0.6044  0.1147  0.6934  3.4344

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.55999    0.65499  -0.855   0.3934
AttractiveM      0.48494    0.04815  10.072  < 2e-16 ***
DecisionMaleYes  0.81122    0.16158   5.021 9.55e-07 ***
RaceFAsian       0.64347    0.56498   1.139   0.2558
RaceFBlack       0.77667    0.61853   1.256   0.2104
RaceFCaucasian   0.90251    0.55680   1.621   0.1063
RaceFLatino      1.06533    0.59612   1.787   0.0751 .
RaceFOther       0.59441    0.61573   0.965   0.3352
SincereM         0.34980    0.04931   7.094 1.22e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.098 on 261 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.6345,    Adjusted R-squared:  0.6233
F-statistic: 56.63 on 8 and 261 DF,  p-value: < 2.2e-16
```

Figure 11: Summary of model

Ta thấy rằng: DecisionMaleYes có liên kết dương đáng kể khi mà nó sẽ tăng 0.81122 mức điểm trung bình của LikeM so với của DecisionMaleNo. Ta thấy rằng biến RaceF có hệ số cũng

đáng kể nhưng nó lại chẳng có ý nghĩa cho việc dự đoán của model vì mức độ khó xảy ra của nó khá cao. Đồng nghĩa với việc nó chẳng đóng góp gì đáng kể cho model. Nên ta có thể loại bỏ nó ra khỏi model và kiểm định lại.

```
# Because RaceF is not significantly associated with LikeM, we remove it
model2 <- lm(LikeM~AttractiveM + DecisionMale + SincereM)
summary(model2)

```

```
Call:
lm(formula = LikeM ~ AttractiveM + DecisionMale + SincereM)

Residuals:
    Min      1Q  Median      3Q     Max
-3.6159 -0.5995  0.0661  0.7017  3.5429

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      0.22103    0.38112   0.580    0.562
AttractiveM      0.49789    0.04772  10.434  < 2e-16 ***
DecisionMaleYes  0.80333    0.16107   4.988 1.10e-06 ***
SincereM         0.34332    0.04869   7.052 1.52e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.101 on 266 degrees of freedom
  (6 observations deleted due to missingness)
Multiple R-squared:  0.6253,    Adjusted R-squared:  0.6211
F-statistic:   148 on 3 and 266 DF,  p-value: < 2.2e-16
```

Figure 12: Summary of model

Nhận xét:

- Do dữ liệu có tính phân tán nên residuals của nó khá lớn. Chứng tỏ có sự chênh lệch khá đáng kể giữa giá trị dự đoán và giá trị quan sát.
- Thêm nữa, hệ số R-squared = 62.53%, nghĩa là model chiếm 62.53 độ biến thiên dữ liệu của tổng độ biến thiên của dữ liệu. Tuy rằng, không cao nhưng nếu so thực tế thì lớn 0.5 đã được cân nhắc là khá tốt.

References

[1] Randall Pruim and Lana Park. Lock5WithR. Chapter 9: Inference for regression and Chapter 10: Multiple regression. PDF.

[2] R Users Guide. Chapter 9: Inference for regression and Chapter 10: Multiple regression. PDF.

[3] Regression with Categorical Variables: Dummy Coding Essentials in R. Retrieved from http://www.sthda.com/english/articles/40-regression-analysis/163-regression-with-categorical-variables-dummy-coding-essentials-in-r/

[4] "Linear Regression R." DataCamp Community. Retrieved from https://www.datacamp.com/community/ tutorials/linear-regression-R