Báo cáo bài tập 4

1612174 - Phùng Tiến Hào - tienhaophung@gmail.com

07/04/2019

Contents

1 Mô tả tổng quan dataset (nguồn gốc, mục đích nghiên cứu, tổng thể nghiên cứu, cách thu thập dữ liệu, cỡ mẫu, số lượng biến).

- Nguồn gốc: Gelman and Hill phân tích dữ liệu sử dụng hồi quy và mô hình đa cấp /phân cấp của đại học Cambridge tại New York năm 2007
- Mục đích nghiên cứu: Khảo sát và phân tích các đánh giá của các học sinh về bạn khác giới của mình trong cuộc gặp gỡ 4 phút của trường Columbia về việc tham dự sự kiện ”SpeedDating”.
- Tổng thể nghiên cứu: Những người tham dự sự kiện ”SpeedDating” là học sinh của trường Columbia được chọn lọc bởi các trợ lý nghiên cứu.
- Cách thu thập dữ liệu:
  - Lấy dữ liệu ngày đầu tiên của cuộc hội giữa người tham dự và bạn tình của họ
  - Các cuộc gặp gỡ được chọn ngẫu nhiên và thời lượng 4 phút
  - Sau đó, người tham dự đánh giá các thuộc tính trên thang điểm 1-10.
- Cỡ mẫu: 276 quan sát
- Số lượng biến: 22 biến

2 Chọn ra 5 biến quan tâm, trong đó có ít nhất 2 biến định tính và 2 biến định lượng. Mô tả sơ lược ý nghĩa 5 biến này và nêu lí do chọn.

a) Các biến định tính:

- DecisionMale (Yes/No): Quyết dịnh nam có muốn 1 ngày hẹn nào khác không?
- RaceF (Asian, Black,...): Chủng tộc của bạn nữ

→ Lý do: vì cái kết quan trọng nhất là bạn nam tham dự có muốn tiến đến cuộc hẹn hò thật sự vào một ngày khác không.

b) Các biến định lượng:

- AttractiveM (num): Nam đánh giá về sức quyến rũ của bạn nữ.
- LikeM (num): Mức độ thích của người nam đối với nữ.
- SincereM (num): Mam đánh giá về độ chân thành của nữ.

→ Lý do: Đây các yếu tố mang tính cảm tính để quyết định người nam có ấn tượng ban đầu tốt đối với người phụ nữ và ảnh hưởng đến DecisionMale.

3 Phân tích thăm dò riêng từng biến đã chọn:

1 2

a) Biến định tính:

- DecisionMale:

```
tab1 = table(DecisionMale) # Count so luong nam yes va no
# Them total
addmargins(tab1)
> DecisionMale
No Yes Sum
130 146 276

prop.table(tab1) # Proportions
> DecisionMale
No       Yes
0.4710145 0.5289855

barplot(tab1) # Ve barchart

```

[1]Nên dùng attach(SpeedDating) để khỏi phải gõ $ trước tên biến mỗi khi truy cập
[2]Nếu attach rồi thì phải detach(SpeedDating) khi đã dùng xong

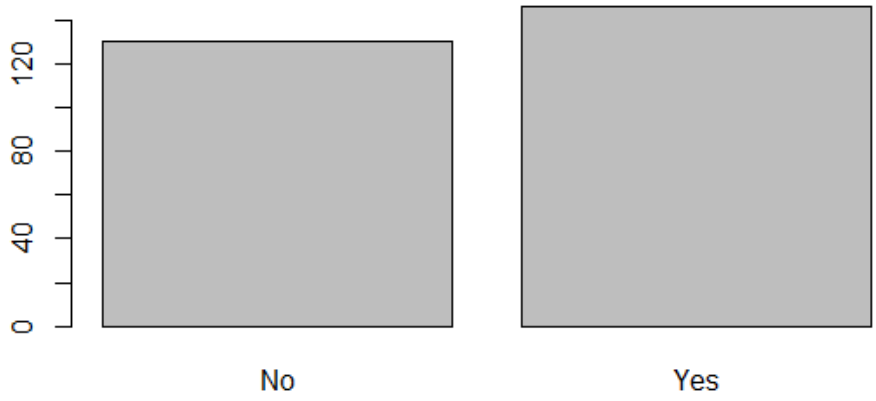

Figure 1: DecisionMale barchar

NX: Tỉ lệ nam đồng ý nhiều hơn là không đồng ý. $(0.529 - 0.471 = 0.058)$

- RaceF:

```
tab2 = table(RaceF) #Count so luong nu cho tung chung toc
# Them total
addmargins(tab2)
> RaceF
   Asian    Black Caucasian    Latino    Other      Sum
4      70      15      148      23       16         276

prop.table(tab2) # Proportions
> RaceF
       Asian      Black  Caucasian     Latino      Other
0.01449275 0.25362319 0.05434783 0.53623188 0.08333333 0.05797101

barplot(tab2)

```

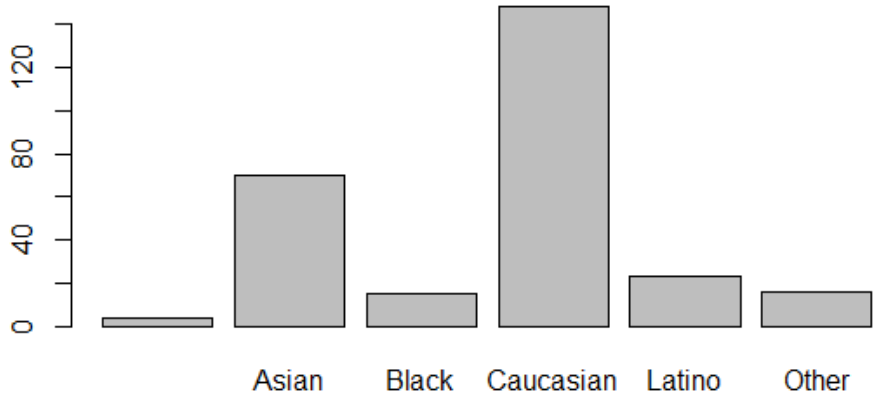

Figure 2: RaceF barchar

NX chung: Tỷ lệ nữ trắng nhiều nhất, tiếp đến là nữ châu Á.

b) Biến định lượng:

- AttractiveM:

```
# AttractiveM
summary(AttractiveM) # 5−number summary
>
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
1.000   5.000   7.000   6.687   8.000  10.000       3

# Histogram
hist(AttractiveM)

```

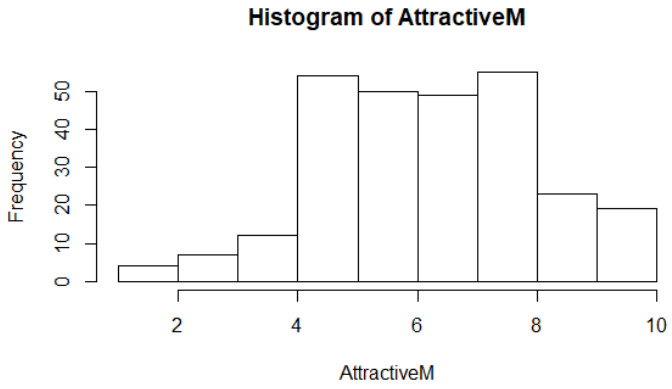

Figure 3: Histogram of AttractiveM

  - Đồ thị có dạng bell-shape
  - 25% Nam đánh giá sức quyến rũ của bạn nữ ít nhất 1 - 5 điểm
  - 50% Nam đánh giá sức quyến rũ của bạn nữ trong khoảng 5 - 8 điểm
  - 25% Nam đánh giá sức quyến rũ của bạn nữ nhiều nhất từ 8 - 10 điểm
  - Kết luận:
    * 50% Nam đánh giá sức quyến rũ của bạn nữ ít hơn 7 điểm.
    * 50% Nam đánh giá sức quyến rũ của bạn nữ trên 7 điểm.

- LikeM:

```
# LikeM
summary(LikeM) # 5−number summary
>
Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
1.000   6.000   7.000   6.682   8.000  10.000       2

# Histogram
hist(LikeM)

```

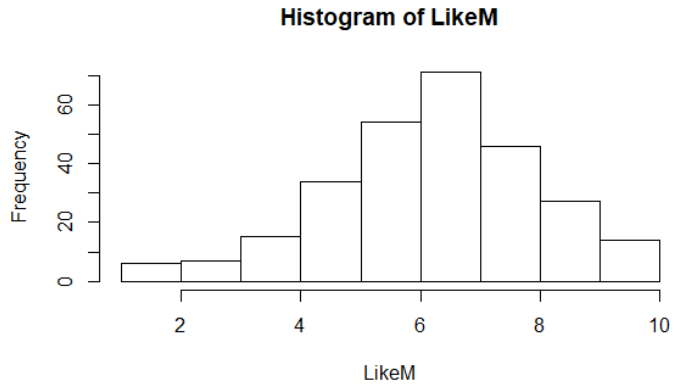

Figure 4: Histogram of LikeM

- Đồ thị có dạng bell-shape
- 25% mức độ thích thú của nam ít nhất 1 - 6 điểm
- 50% mức độ thích thú của nam trong khoảng 6 - 8 điểm
- 25% mức độ thích thú của nam nhiều nhất từ 8 - 10 điểm
- Kết luận:
  - 50% mức độ thích thú của nam ít hơn 7 điểm.
  - 50% mức độ thích thú của nam trên 7 điểm.

- SincereM:

```
# SincereM
summary(SincereM) # 5−number summary
>
 Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
1.000   7.000   8.000   7.856   9.000  10.000       5

# Histogram
hist(SincereM)

```

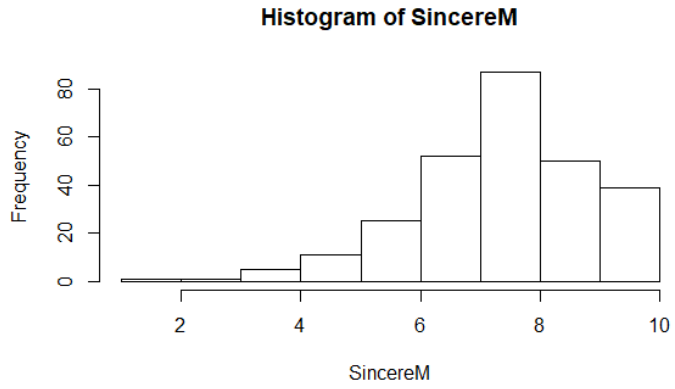

Figure 5: Histogram of SincereM

- Đồ thị hơi nghiêng về bên trái.
- 25% nam đánh giá độ chân thành của nữ ít nhất 1 - 7 điểm
- 50% nam đánh giá độ chân thành của nữ trong khoảng 7 - 9 điểm
- 25% nam đánh giá độ chân thành của nữ nhiều nhất từ 9 - 10 điểm
- Kết luận:
  * 50% nam đánh giá độ chân thành của nữ ít hơn 8 điểm.
  * 50% nam đánh giá độ chân thành của nữ trên 8 điểm.

NX: Độ chân thành và sức quyến rũ của bạn nữ dễ ảnh hưởng đến DecisionM và LikeM của nam.

4 Chọn ra 2 biến định tính (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.

Chọn 2 biến định tính: DecisionMale (Yes/No), RaceF (Asian, Black, Caucasian, Latino, or Other)

```
# 2 bien dinh tinh
tab1 = table(DecisionMale, RaceF)
# Them margin
addmargins(tab1)

>
          RaceF
DecisionMale   Asian Black Caucasian Latino Other Sum
   No    2    32     7       72     7    10 130
```

```
   Yes  2   38    8      76    16    6 146
   Sum  4   70   15     148    23   16 276

# 2-way table
# Ti le nam (yes/no) dieu kien chung toc nu (Asian, Black, ...)
prop.table(tab1, margin = 1)

>
         RaceF
DecisionMale             Asian      Black  Caucasian     Latino      Other
   No  0.01538462 0.24615385 0.05384615 0.55384615 0.05384615 0.07692308
   Yes 0.01369863 0.26027397 0.05479452 0.52054795 0.10958904 0.04109589

# Segmented barchart
barplot(tab1, legend = TRUE)
```

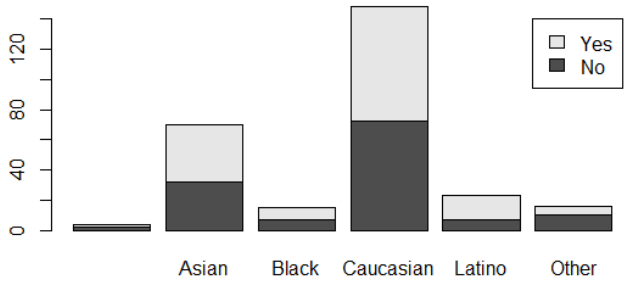

Figure 6: Segmented barchart of 2 categorial variables

- Ta thấy rằng tỉ lệ phản hồi (no/yes) của nam đối với chủng tộc nữ da trắng (Caucasian) là cao nhất (0.553, 0.52). Tiếp đến là nữ châu Á (Asian) (0.246, 0.26).
- Tỉ lệ phản hồi ”yes” và ”no” đối với nữ da trắng:

$$p_{yes} \approx 0.52$$
$$p_{no} \approx 0.553$$

- Tỉ lệ khác biệt (difference proportion) giữa tỉ lệ phản hồi ”yes” và phản hồi

”no” đối với nữ da trắng:

$$p_{yes} - p_{no} = 0.52 - 0.553 = -0.033$$

→ Tỷ lệ người nam phản hồi ’no’ cao hơn phản hồi ’yes’ đối với nữ da trắng 0.033.

- Tỷ lệ nam phản hồi ”yes” đối với người da trắng cao hơn với người châu Á: $0.553 - 0.246 = 0.307$
- Từ bảng 2-way table, ta thấy rằng có 4 người nữ không có chủng tộc: 2 nhận phản hồi ”yes” và 2 nhận phản hồi no ”no”.

  NX: người da trắng (Caucasian) và người da màu (Asian) nhận được sự phản hồi cao hơn các tộc còn lại.

5 Chọn ra 1 biến định tính, 1 biến định lượng (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.

Chọn 1 biến định tính và 1 biến định lượng: DecisionMale (yes/no), AttractiveM (1-10)

```
# 1 quantitative and 1 categorical varibles
# statistics for the quantitative variable within each category
by(AttractiveM, DecisionMale, mean, na.rm=TRUE)
>
DecisionMale: No
[1] 5.641732
------------------------------------
DecisionMale: Yes
[1] 7.59589

# side-by-side boxplots
boxplot(AttractiveM ~ DecisionMale, xlab = 'DecisionMale', ylab = 'AttractiveM')

```

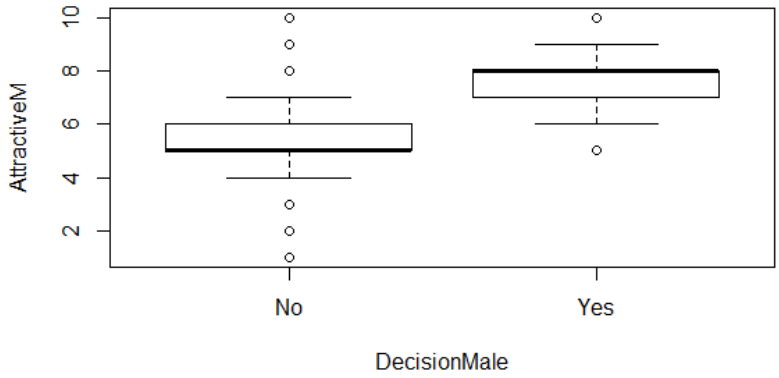

Figure 7: Side-by-side boxplots

- Ở trường hợp phản hồi ”no”, Ta thấy median trùng với 1st quartile bằng 5.
- Ở trường hợp phản hồi ”yes”, ta thấy median trùng với 3rd quartile bằng 8.
- Các outlier của phản hồi ”no” xuất hiện nhiều hơn của phản hồi ”yes”. Điều này làm tần suất của dữ liệu trải dài đều từ 1 - 10.
- Ta thấy rằng có TH outlier bên phản hồi ”no” có attractiveM = 10. Điều này khá bất thường. Còn bên phản hồi ”yes” thì các outlier không đáng kể lắm, phân bố cũng gần range của dữ liệu.
- Thêm vào đó, ta thấy rằng mean của DecisionMale: No (5.641) < mean của DecisionMale: Yes (7.59). Do đó ta có thể nói mức điểm quyến rũ của phản hồi ”yes” cao hơn của phản hồi ”no”.
- Đối với phản hồi ”no”, ta thấy rằng median < mean do có nhiều outlier trải dải từ 1-10. Ngược lại, phản hồi ”yes” có median > mean do ít outlier hơn và phân bố của outlier cũng không quá xa range của dữ liệu.
- Range của phản hồi ”yes” và ”no” cũng tương đối giống nhau.
- Chúng ta thấy có 1 liến kết khi attractiveM càng cao thì khả năng phản hồi yes cũng cao tuy rằng liên kết này không quá mạnh.

6 Chọn ra 2 biến định lượng (từ 5 biến quan tâm) và phân tích thăm dò quan hệ giữa chúng.

Chọn 2 biến định lượng: AttractiveM (1-10) và LikeM (1-10)

```
# 2 quantitative varibles
# Summary statistics: correlation, regression line
> cor(AttractiveM, LikeM, use = "complete.obs") # avoid missing value NA
[1] 0.7240187

> lm(LikeM~AttractiveM) # Linear regression for 2 varibles

Call:
lm(formula = LikeM ~ AttractiveM)

Coefficients:
(Intercept)  AttractiveM
1.9110       0.7139


# Graphical display: scatterplot
plot(AttractiveM, LikeM, main = "Scatter plot example", pch=19)
# Add fit lines
abline(lm(LikeM~AttractiveM), col="red") # regression line (y~x)

```

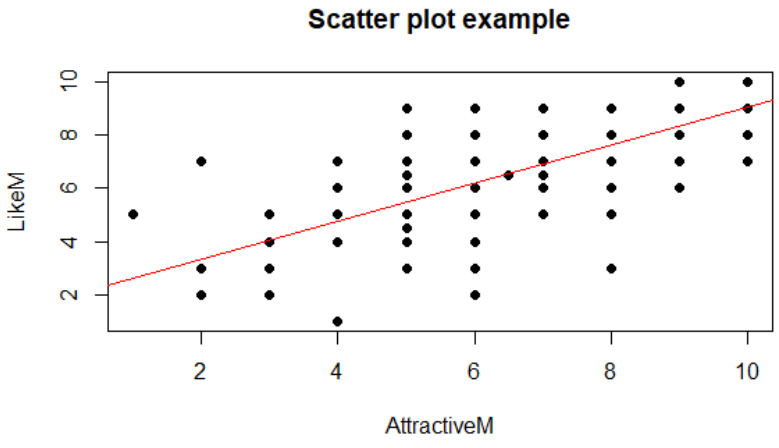

Figure 8: Scatterplot of 2 quantitative variables

- Ta thấy rằng giữa 2 biến định lượng này có liên kết dương khá mạnh (positive

association) $(r = 0.724)$ do đó nhìn chung khi AttractiveM tăng thì LikeM cũng tăng.

- Tuy rằng, có liên kết mạnh nhưng điều này không thể khẳng định giữa LikeM và AttractiveM có quan hệ nhân quả (causation)
- Chúng ta có thể fit 1 đường thẳng (best-fit line) $y = 1.911 + 0.7139x$ để chia tập dữ liệu thì ta thấy rằng y sẽ tăng 71.39% nếu x tăng lên 1 đơn vị.
- Cụ thể ở đây khi AttractiveM tăng 1 đơn vị thì LikeM sẽ tăng 71.39%. The intercept 1.911 chỉ rằng $LikeM = 1.911$ nếu $AttractiveM = 0$ nhưng hầu như rất hiếm $AttractiveM = 0$.
- Ta thấy dữ liệu phân bố không gần best-fit line.

→ Việc dùng đường thẳng này để dự đoán ở đây là khả thi nhưng hiệu quả không cao do dữ liệu không phân bố không gần regression line.

7 (Cộng điểm) Phân tích thăm dò quan hệ giữa nhiều hơn 2 biến (từ 5 biến quan tâm).

Multiple regression: LikeM ∼ AttractiveM + SincereM
[3]

```
# Multiple regression
> fit <- lm(LikeM~AttractiveM + SincereM)
> summary(fit) # show the results
Call:
lm(formula = LikeM ~ AttractiveM + SincereM)

Residuals:
Min      1Q  Median      3Q     Max
−3.9329 −0.5840  0.0905  0.7111  3.3394

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.06763    0.39650   0.171    0.865
AttractiveM  0.62837    0.04165  15.085  < 2e−16 ***
SincereM     0.30639    0.05023   6.100  3.7e−09 ***
−−−
Signif. codes:  0 ‘’*** 0.001 ‘’** 0.01 ‘’* 0.05 ‘’. 0.1 ‘ ’ 1

```

[3] Sử dụng hàm ggpairs của GGally package để hiển thị hệ số tương quan cũng như scatterplot cho từng cặp biến và density plot cho từng biến

```
Residual standard error: 1.149 on 267 degrees of freedom
(6 observations deleted due to missingness)
Multiple R−squared:  0.5903,  Adjusted R−squared:  0.5872
F−statistic: 192.3 on 2 and 267 DF,  p−value: < 2.2e−16

#shows the correlation coefficient of multiple variables
#in conjunction with a scatterplot
#(including a line of best fit with a confidence interval) and a density plot.
ggpairs(SpeedDating,
columns = c("AttractiveM", "SincereM", "LikeM"),
upper = list(continuous = wrap("cor",
size = 10)),
lower = list(continuous = "smooth"))
```

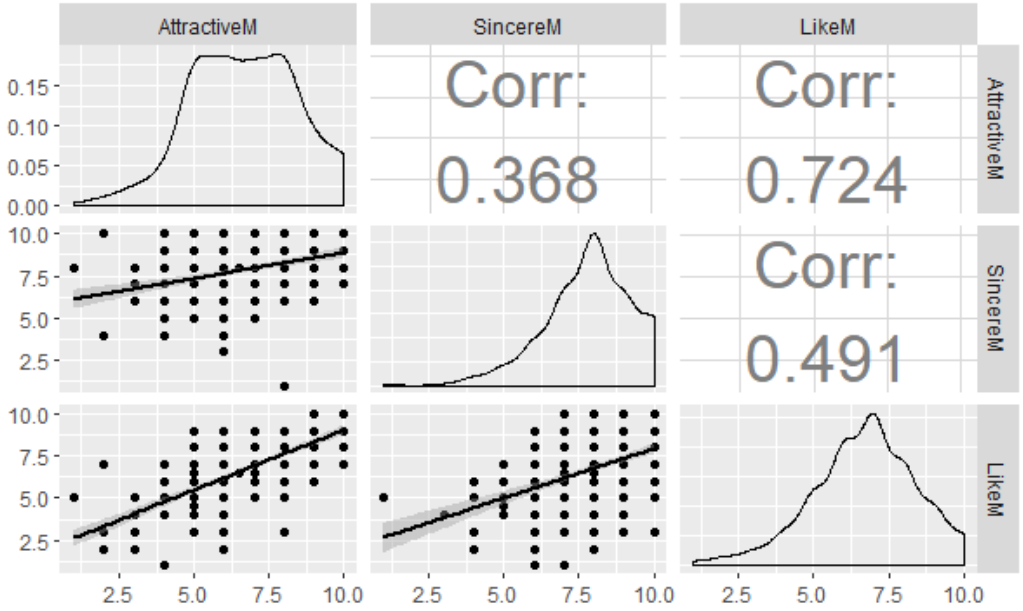

Figure 9: Multivariate plot

- Ta thấy rằng p-value của $F - statistic < 2.2e - 16$ điều này có nghĩa là ít nhất 1 biến x thay đổi thì sẽ ảnh hưởng đến output y.
- R-squared: 59.03% điều này cho thấy mô hình này có residual khá cao do đó khả năng đem đi dự đoán mang lai chinh xác không cao.
- Đường thẳng phân tách dữ liệu của ta là: $y = 0.067 + 0.628x_1 + 0.306x_2$

- Ta thấy cả attractiveM và likeM đều ảnh hưởng đáng kể đến output y. Cụ thể, ta thấy AttractiveM ảnh hưởng nhiều đến LikeM hơn là SincereM (0.628 > 0.306).
- Ví dụ: khi attractiveM tăng 1 thì đầu ra y sẽ tăng 0.067 lần so với khi SincereM tăng 1 thì y chỉ tăng 0.306 lần.
- Như đã phân tích ở câu 6, ta thấy rằng giữa AttractiveM và LikeM có liên kết dương mạnh ($r_{AL} = 0.724$). Và giữa SincereM và LikeM cũng có liên kết dương trung tính ($r_{SL} = 0.491 \approx 0.5$). Điều này lại một lần nữa khẳng định sự ảnh hưởng của AttractiveM, SincereM lên LikeM là đáng kể.
- Thêm vào đó, hệ số tương quan của AttractiveM và SincereM $r_{AS} = 0.368$ không quá mạnh. Điều này cho thấy sự thay đổi của cả 2 không ảnh hưởng đến nhau nhiều.
- Ta thấy rằng các scatter plot giữa (AttractiveM, LikeM), (SincereM, LikeM) và (AttractiveM, SincereM) có phần smooth curve màu xám. Phần xám này khá to ở phần đầu khi số lượng point nhỏ và dần về sau phần xám nhỏ đi và biến mất khi số point nhiều hơn.
- Các smooth curves này giúp chúng ta thấy được số lượng các liên kết không chắc chắn (uncertain association) với regression line. Sự không chắc chắn này tăng khi dữ liệu quan sát nhỏ và giảm khi dữ liệu quan sát lớn.

8 Tham khảo

[1] Lock5withR pdf-file, Lock5withR.
[2] Quick-R by DataCamp, Quick-R
[3] Official blog - R Correlation tutorial, DataCamp-Blog
[4] Multiple Linear Regression in R, STHDA - Articles - Regression Analysis
[5] Book: Unlocking the power of data, Chapter 2 – Describing Data, Robin H. Lock, Patti Frazer Lock, Dennis F. Lock, Kari Lock Morgan, Eric F. Lock - Statistics: Unlocking the Power of Data (2012, Wiley)