

Sử dụng R Markdown

Tuan V. Nguyen

Garvan Institute of Medical Research

University of New South Wales (UNSW Sydney), Australia

University of Technology, Sydney (UTS), Australia

Ton Duc Thang University, Vietnam



RMarkown, RStudio, R

- **R** là nền tảng
- **R Studio** được xây dựng trên R
 - có 'menu' và giao diện biểu đồ
- **R Markdown** là một phần của RStudio
 - giúp 'document' mã R
 - tái lập (reproducibility)



R Studio

- Download R Studio
<https://www.rstudio.com/products/RStudio/#Desktop>
- R Markdown có trong R Studio

RStudio

File Edit Code View Plots Session Build Debug Profile Tools Window Help

RStudio

Addins

Project: (None)

Untitled1

Source on Save | Import Dataset | Global Environment

Run | Source | Data

1

R markdown

Data objects

Mā R

Output

Vận hành của R Studio

RStudio
Untitled1 x Go to file/function Addins Project: (None)

1:1 (Top Level) R Script

Console ~/ ↻

```
> # Ví dụ vận hành
> set.seed(123)
> height = rnorm(n=1000, mean=156, sd=5.5)
> error = rnorm(1000)
> p.weight = -30 + 0.55*height + error
> # fit linear regression model
> m = lm(p.weight ~ height)
> summary(m)

Call:
lm(formula = p.weight ~ height)

Residuals:
    Min      1Q  Median      3Q     Max 
-3.0279 -0.6914  0.0043  0.7087  3.2911 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -32.456296   0.911762 -35.60 <2e-16 ***
height       0.566009   0.005838  96.96 <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.006 on 998 degrees of freedom
Multiple R-squared:  0.904,    Adjusted R-squared:  0.9039 
F-statistic: 9401 on 1 and 998 DF,  p-value: < 2.2e-16

> hist(height, col="blue", border="white")
> |
```

Có thể lưu mã R và output bằng cách dùng File/Save

Environment History Connections

Import Dataset Global Environment

Data

m	List of 12
t	3649 obs. of 6 variables

Values

error	num [1:1000] -0.996 -1.04 -0.018 -0.132 ...
height	num [1:1000] 153 155 165 156 157 ...
p.weight	num [1:1000] 53.1 54.1 60.5 55.9 53.6 ...

Files Plots Packages Help Viewer

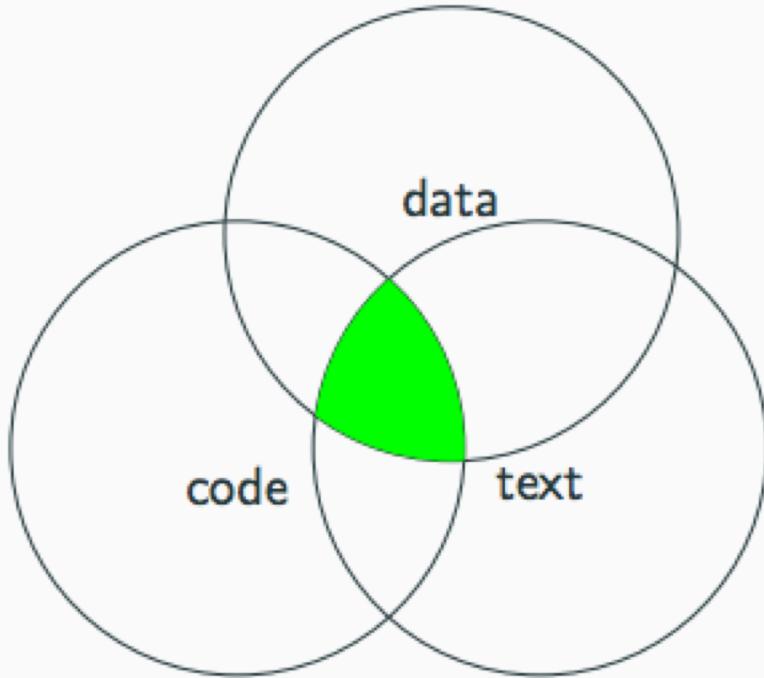
Zoom Export

Histogram of height

The histogram displays the frequency distribution of height. The x-axis is labeled 'height' and ranges from 140 to 175. The y-axis is labeled 'Frequency' and ranges from 0 to 300. The distribution is roughly bell-shaped, centered around 156.

Bin Range (height)	Frequency
140 - 145	~10
145 - 150	~140
150 - 155	~280
155 - 160	~320
160 - 165	~180
165 - 170	~50
170 - 175	~10

R Markdown và mục tiêu tái lập

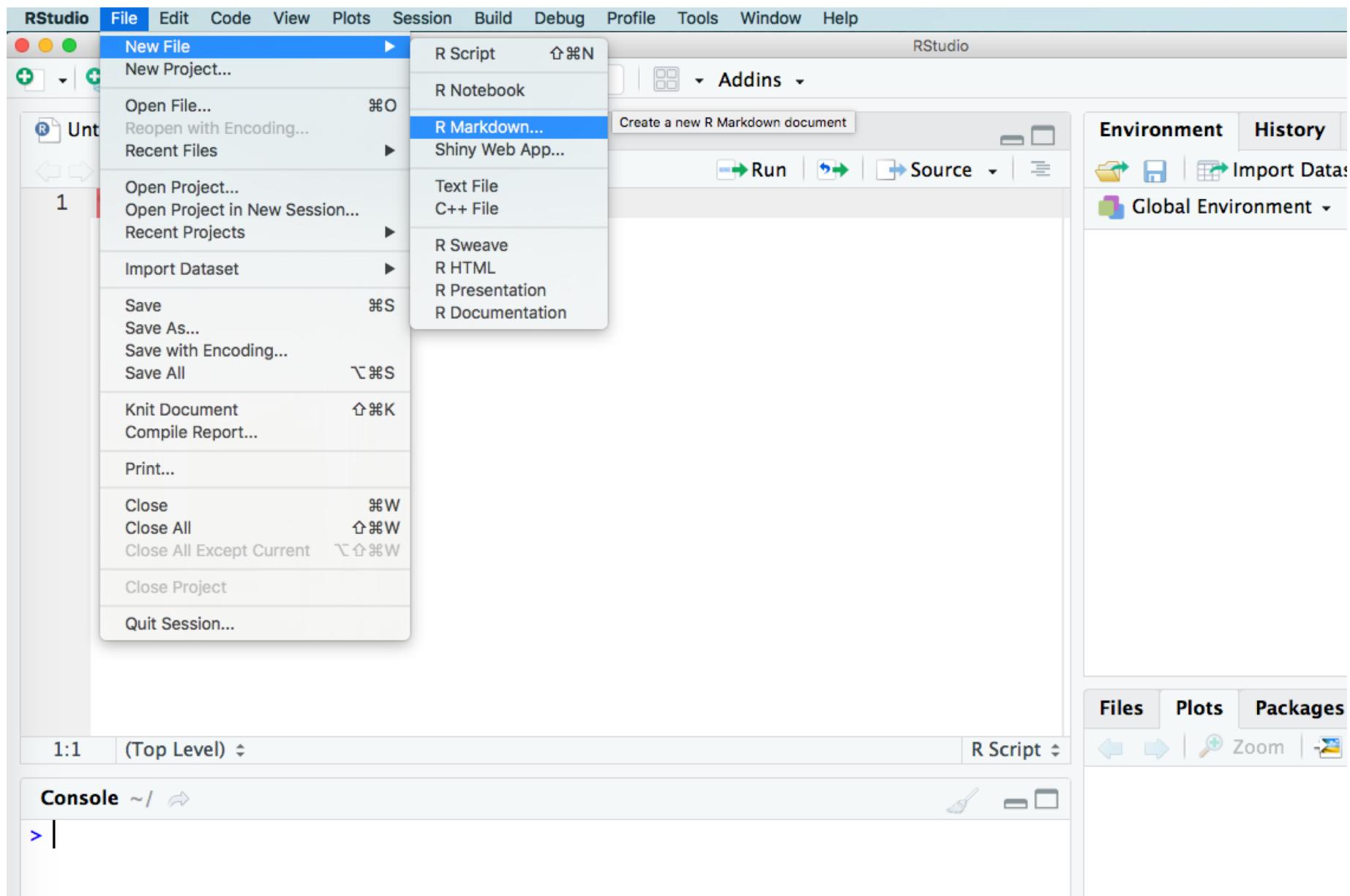


Đảm bảo tính reproducibility: **dữ liệu + mã R + văn bản**

R Markdown

- R + Markdown = RMarkdown
- Có thể xem là một "markup language"
- Lưu trữ mã R
- Export output sang **Word, pdf, html**

Một session với R Markdown



New R Markdown

Document

Title: An example with RMarkdown

Author: Tuan Nguyen

Default Output Format:

HTML
Recommended format for authoring (you can switch to PDF or Word output anytime).

PDF
PDF output requires TeX (MiKTeX on Windows, MacTeX 2013+ on OS X, TeX Live 2013+ on Linux).

Word
Previewing Word documents requires an installation of MS Word (or Libre/Open Office on Linux).

OK Cancel

Untitled1 x Untitled2 x

ABC Knit Insert Run

```
1 ---  
2 title: "An.example.with.RMarkdown"  
3 author: "Tuan.Nguyen"  
4 date: "4/22/2019"  
5 output: html_document  
6 ---  
7  
8 ```{r setup, include=FALSE}  
9 knitr::opts_chunk$set(echo = TRUE)  
10 ````  
11  
12 ## R Markdown  
13  
14 This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.  
15  
16 When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:  
17  
18 ```{r cars}  
19 summary(cars)  
20 ````  
21  
22 ## Including Plots  
23  
24 You can also embed plots, for example:  
25  
2:1 # An example with RMarkdown
```

R Markdown

Console ~/ ↵

Một document của R Markdown

- Header
- R codes
- Output

The image shows two side-by-side windows. On the left is the RStudio IDE with the file 'example.Rmd' open. The code in the Rmd file is:

```
1 # Header 1
2
3 This is an R Markdown document. Markdown is a
4 simple formatting syntax for authoring webpages.
5
6 Use an asterisk mark to provide emphasis, such
7 as *italics* or **bold**.
8
9 - Item 1
10 - Item 2
11 - Item 3
12
13 ``
14 Use back ticks to
15 create a block of code
16 ``
17
18 Embed LaTeX or MathML equations,
19 $\frac{1}{n} \sum_{i=1}^n x_i$"
20
21 Or even footnotes, citations, and a
22 bibliography. [^1]
23
24 [^1]: Markdown is great.
```

The right window shows the generated HTML file 'example.html'. The content is:

Header 1

This is an R Markdown document. Markdown is a simple formatting syntax for authoring web pages.

Use an asterisk mark to provide emphasis, such as *italics* or **bold**.

Create lists with a dash:

- Item 1
- Item 2
- Item 3

Use back ticks to
create a block of code

Embed LaTeX or MathML equations, $\frac{1}{n} \sum_{i=1}^n x_i$

Or even footnotes, citations, and a bibliography.^[^1]

1. Markdown is great. ↪

Một document của R Markdown

- Header
- R codes
- Output

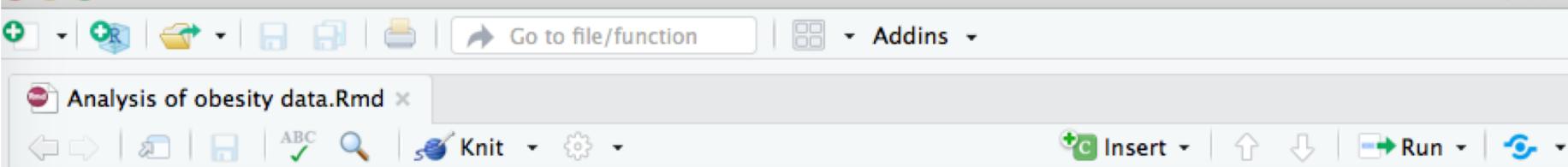
Giữa `---` và `---`
là R codes

```
# Phân tích tương quan giữa BMI và pcfat  
## Dữ liệu obesity data, Việt Nam  
### PI: Tuan Nguyen
```

```
# Đọc dữ liệu vào R  
---  
ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data  
Analytics 2019/obesity data.csv")  
head(ob)  
summary(lm(pcfat ~ bmi, data=ob))
```

```
# Vẽ mối tương quan giữa BMI và pcfat  
library(ggplot2)  
p = ggplot(data=ob, aes(x=bmi, y=pcfat, ol=gender))  
p + geom_point() + geom_smooth(method="lm")  
---
```

`<text giữa hai dấu này không có trong phần output>`



```
1 ----  
2 title: "Analysis of obesity"  
3 author: "Tuan Nguyen"  
4 date: "4/22/2019"  
5 output: html_document  
6 ----  
7 ~  
8 # Analysis of obesity dataset  
9 ## Data are from a study in Vietnam  
10 ### PI: Tuan Nguyen  
11 ~  
12 # Đọc dữ liệu vào R ...  
13 ```{r}  
14 ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")  
15 head(ob)  
16 ```  
17 # Phân tích mô hình hồi qui tuyến tính  
18 ```{r}  
19 summary(lm(pcfat ~ bmi, data=ob))  
20 ```  
21 ### Summary  
22 The mean of percent body fat is `r mean(ob$pcfat)`.  
23 ~  
24 ~  
25 # Vẽ mối tương quan giữa BMI và pcfat ...  
26 ```{r}  
27 library(ggplot2)  
28 p = ggplot(data=ob, aes(x=bmi, y=pcfat, col=gender))  
29 p + geom_point() + geom_smooth(method="lm")  
30 ```  
31 ~
```

Analysis of obesity

Tuan Nguyen

4/22/2019

Analysis of obesity dataset

Data are from a study in Vietnam

PI: Tuan Nguyen

Đọc dữ liệu vào R

Output

```
ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")
head(ob)
```

```
##   id gender height weight bmi age WBBMC wbbmd   fat   lean pcfat
## 1  1       F     150     49 21.8  53 1312  0.88 17802 28600  37.3
## 2  2       M     165     52 19.1  65 1309  0.84  8381 40229  16.8
## 3  3       F     157     57 23.1  64 1230  0.84 19221 36057  34.0
## 4  4       F     156     53 21.8  56 1171  0.80 17472 33094  33.8
## 5  5       M     160     51 19.9  54 1681  0.98  7336 40621  14.8
## 6  6       F     153     47 20.1  52 1358  0.91 14904 30068  32.2
```

Phân tích mô hình hồi qui tuyến tính

```
summary(lm(pcfat ~ bmi, data=ob))

##
## Call:
## lm(formula = pcfat ~ bmi, data = ob)
##
## Residuals:
##    Min      1Q  Median      3Q     Max 
## -19.612  -4.181   1.392   4.690  18.241 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 8.39889   1.36777  6.141 1.11e-09 ***
## bmi         1.03619   0.06051 17.123 < 2e-16 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 6.45 on 1215 degrees of freedom
## Multiple R-squared:  0.1944, Adjusted R-squared:  0.1937 
## F-statistic: 293.2 on 1 and 1215 DF,  p-value: < 2.2e-16
```

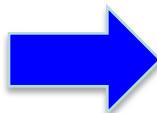
Summary

The mean of percent body fat is 31.6047859.

R Markdown codes and output

```
title: "Analysis of obesity"  
author: "Tuan Nguyen"  
date: "4/22/2019"  
output: html_document
```

```
# Analysis of obesity dataset  
## Data are from a study in Vietnam  
### PI: Tuan Nguyen
```



Analysis of obesity

Tuan Nguyen

4/22/2019

Analysis of obesity dataset

Data are from a study in Vietnam

PI: Tuan Nguyen

R Markdown codes and output

```
# Đọc dữ liệu vào R
```{r}
ob = read.csv("~/Dropbox/_Lectures
and Talks/UTS Data Analytics
2019/obesity data.csv")
head(ob)
```

```



Đọc dữ liệu vào R

```
ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")
head(ob)
```

```
##   id gender height weight  bmi age WBBMC wbbmd   fat   lean pcfat
## 1  1      F    150     49 21.8  53  1312  0.88 17802 28600  37.3
## 2  2      M    165     52 19.1  65  1309  0.84  8381 40229  16.8
## 3  3      F    157     57 23.1  64  1230  0.84 19221 36057  34.0
## 4  4      F    156     53 21.8  56  1171  0.80 17472 33094  33.8
## 5  5      M    160     51 19.9  54  1681  0.98  7336 40621  14.8
## 6  6      F    153     47 20.1  52  1358  0.91 14904 30068  32.2
```

```
# Phân tích mô hình hồi qui tuyến tính
```{r}
summary(lm(pcfat ~ bmi, data=ob))
```

#### Summary
The mean of percent body fat is `r mean(ob$pcfat)`.
```



Phân tích mô hình hồi qui tuyến tính

```
summary(lm(pcfat ~ bmi, data=ob))

##
## Call:
## lm(formula = pcfat ~ bmi, data = ob)
##
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -19.612  -4.181   1.392   4.690  18.241 
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 8.39889   1.36777   6.141 1.11e-09 ***
## bmi         1.03619   0.06051  17.123 < 2e-16 ***
## ---        
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
##
## Residual standard error: 6.45 on 1215 degrees of freedom
## Multiple R-squared:  0.1944, Adjusted R-squared:  0.1937 
## F-statistic: 293.2 on 1 and 1215 DF,  p-value: < 2.2e-16
```

Summary

The mean of percent body fat is 31.6047859.

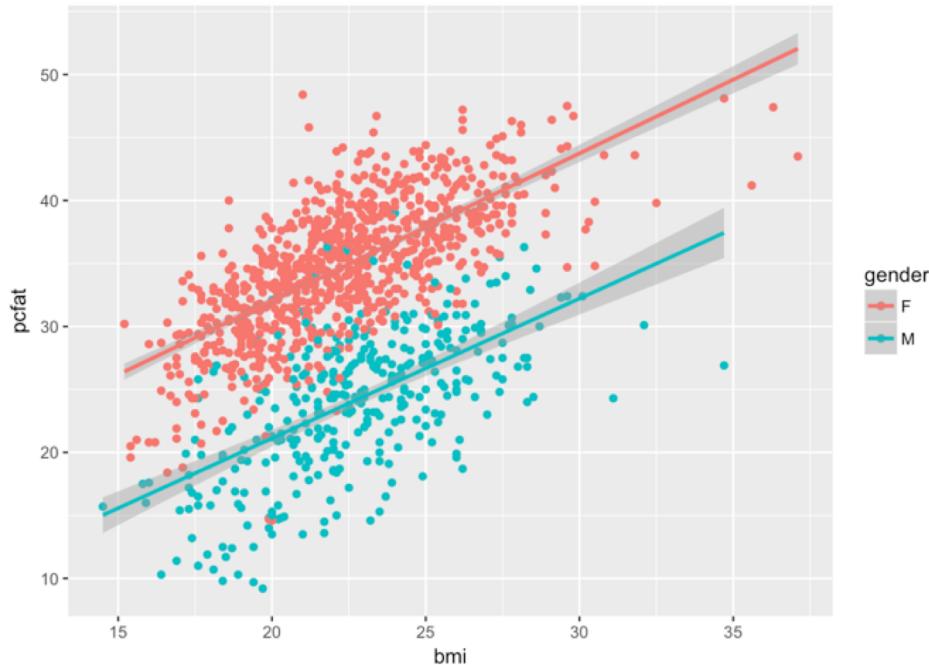
```
# Vẽ mối tương quan giữa BMI và pcfat
```
{r}
library(ggplot2)
p = ggplot(data=ob, aes(x=bmi, y=pcfat, col=gender))
p + geom_point() + geom_smooth(method="lm")
```

```



Vẽ mối tương quan giữa BMI và pcfat

```
library(ggplot2)
p = ggplot(data=ob, aes(x=bmi, y=pcfat, col=gender))
p + geom_point() + geom_smooth(method="lm")
```



```
1 ----  
2 title: "Analysis of obesity"  
3 author: "Tuan Nguyen"  
4 date: "4/22/2019"  
5 output: html_document  
6 ----  
7  
8 # Analysis of obesity dataset  
9 ## Data are from a study in Vietnam  
10 ### PI: Tuan Nguyen  
11  
12 # Đọc dữ liệu vào R  
13 ```{r}  
14 ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")  
15 head(ob)  
16```  
17 # Phân tích mô hình hồi qui tuyến tính  
18 ```{r}  
19 summary(lm(pcfat ~ bmi, data=ob))  
20```  
21 ### Summary  
22 The mean of percent body fat is `r mean(ob$pcfat)`.  
23  
24  
25 # Vẽ mối tương quan giữa BMI và pcfat  
26 ```{r}  
27 library(ggplot2)  
28 p = ggplot(data=ob, aes(x=bmi, y=pcfat, col=gender))  
29 p + geom_point() + geom_smooth(method="lm")  
30```  
31
```

Sau khi viết xong, có thể "Knit" để cho ra html output trên rpubs.com

Analysis of obesity

Tuan Nguyen
4/22/2019

Analysis of obesity dataset

Data are from a study in Vietnam

PI: Tuan Nguyen

Đọc dữ liệu vào R

```
ob = read.csv("~/Dropbox/_Lectures and Talks/UTS Data Analytics 2019/obesity data.csv")
head(ob)
```

| | id | gender | height | weight | bmi | age | WBBMC | wbmd | fat | lean | pcfat |
|------|----|--------|--------|--------|------|-----|-------|------|-------|-------|-------|
| ## 1 | 1 | F | 150 | 49 | 21.8 | 53 | 1312 | 0.88 | 17802 | 28600 | 37.3 |
| ## 2 | 2 | M | 165 | 52 | 19.1 | 65 | 1309 | 0.84 | 8381 | 40229 | 16.8 |
| ## 3 | 3 | F | 157 | 57 | 23.1 | 64 | 1230 | 0.84 | 19221 | 36057 | 34.0 |
| ## 4 | 4 | F | 156 | 53 | 21.8 | 56 | 1171 | 0.80 | 17472 | 33094 | 33.8 |
| ## 5 | 5 | M | 160 | 51 | 19.9 | 54 | 1681 | 0.98 | 7336 | 40621 | 14.8 |
| ## 6 | 6 | F | 153 | 47 | 20.1 | 52 | 1358 | 0.91 | 14904 | 30068 | 32.2 |

Phân tích mô hình hồi qui tuyến tính

```
summary(lm(pcfat ~ bmi, data=ob))
```

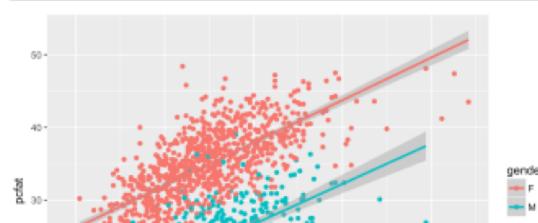
```
##
## Call:
## lm(formula = pcfat ~ bmi, data = ob)
##
## Residuals:
##   Min     1Q   Median     3Q    Max
## -19.612 -4.181  1.392  4.690 18.241
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.39889   1.36777  6.141 1.11e-09 ***
## bmi         1.03619   0.06051 17.123 < 2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.45 on 1215 degrees of freedom
## Multiple R-squared: 0.1944, Adjusted R-squared: 0.1937
## F-statistic: 293.2 on 1 and 1215 DF, p-value: < 2.2e-16
```

Summary

The mean of percent body fat is 31.6047859.

Vẽ mối tương quan giữa BMI và pcfat

```
library(ggplot2)
p = ggplot(data=ob, aes(x=bmi, y=pcfat, col=gender))
p + geom_point() + geom_smooth(method="lm")
```



RStudio và RMarkdown

- Hai 'additions' rất quan trọng cho phân tích dữ liệu với R
- **RStudio** cung cấp giao diện 'thân thiện' hơn R
- **RMarkdown** là một 'ngôn ngữ bị chú' cho một công trình phân tích dữ liệu
 - một phương tiện rất có ích cho lưu trữ mã R
 - đảm bảo tính tái lập trong phân tích dữ liệu
 - có thể xuất sang Word, html và pdf

