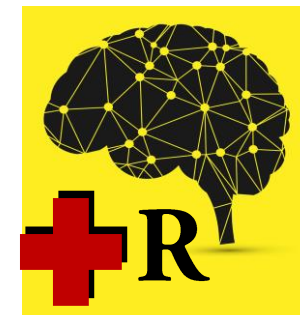




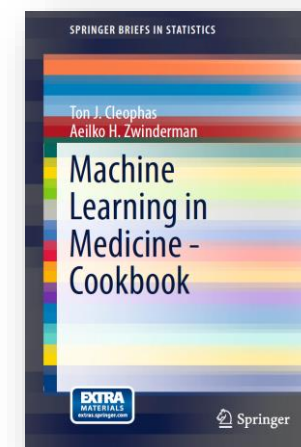
Series: Machine learning ứng dụng trong y học

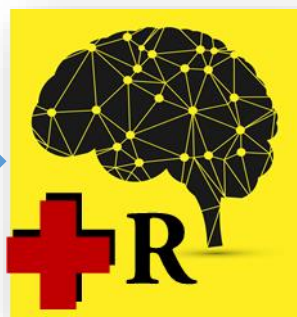
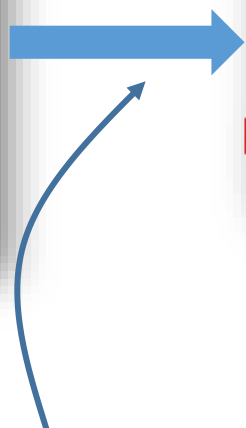
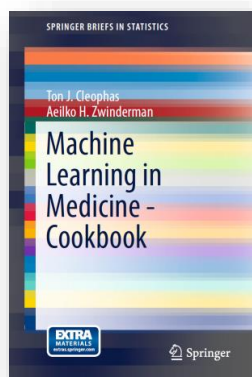
Case study số 3



Hồi quy tuyến tính + bootstrap sử dụng caret

Thực hiện: Lê Đông Nhật Nam





Nhóm gồm 12 thành viên
sử dụng R như công cụ chính

Giới thiệu

Machine learning in Medicine – Cookbook là một trong số ít tài liệu trình bày về ứng dụng của Machine learning trong Y học.

Các tác giả đã sử dụng cách tiếp cận theo case study và algorithm với 65 dataset có giá trị thực hành cao.

Tuy nhiên, rất đáng tiếc vì toàn bộ tài liệu được đặt trên nền tảng IBM-SPSS, là một công cụ thương mại, đắt tiền, kém linh hoạt và không cho phép người học hiểu được căn cứ về bản chất của quy trình và cách kiểm soát nó.

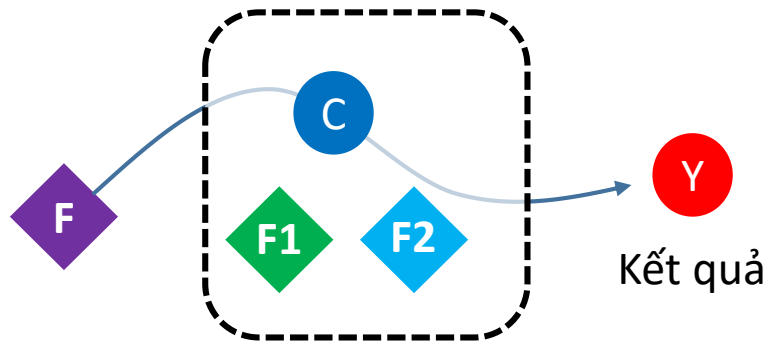
Do đó, nhóm tác giả dự án « Machine learning trong y học sử dụng R » sẽ lần lượt tái hiện lại nội dung của những case study trong sách nhưng hoàn toàn sử dụng R.

Chúng tôi hy vọng dự án này sẽ giúp thực hiện được 3 mục tiêu sau:

- Phổ biến và khuyến khích ứng dụng Machine learning trong nghiên cứu Y học tại Việt Nam
- Tạo cầu nối giữa Machine learning và Thống kê cổ điển
- Khuyến khích việc sử dụng R, một công cụ thống kê miễn phí và hiệu quả cao.

1

1.1 Giới thiệu case study và dataset



Mẫu = 20 bệnh nhân

F: phân nhóm điều trị (yếu tố chính, bắt buộc phải được xét)

Các biến phụ :

F1: giới tính (nam/nữ)

F2: bệnh lý kèm theo (có/không)

C: Tuổi

Y: thời gian ngủ mỗi ngày = biến kết quả

Mục tiêu

Câu hỏi nghiên cứu trong case study số 3 này có thể được diễn đạt theo nhiều cách :

1. Khảo sát hiệu quả trị liệu của một loại thuốc ngủ X
2. Dự báo hiệu quả điều trị của thuốc ngủ X dựa vào tuổi và một số yếu tố khác
3. So sánh hiệu quả điều trị giữa 2 phân nhóm sử dụng giả dược và thuốc ngủ X

...

1

1.2 Tải data SPSS vào R

```
library(foreign)
```

```
spssdata<- read.spss("linoutcomeprediction.sav",to.data.frame=TRUE)
```

```
data=spssdata
```

```
data$treatment<-factor(data$treatment,levels=c(0,1),labels=c("Placebo", "Treatment"))
```

```
data$gender<-factor(data$gender,levels=c(0,1),labels=c("Female", "Male"))
```

```
data$comorbidity<-factor(data$comorbidity,levels=c(0,1),labels=c("No", "Yes"))
```

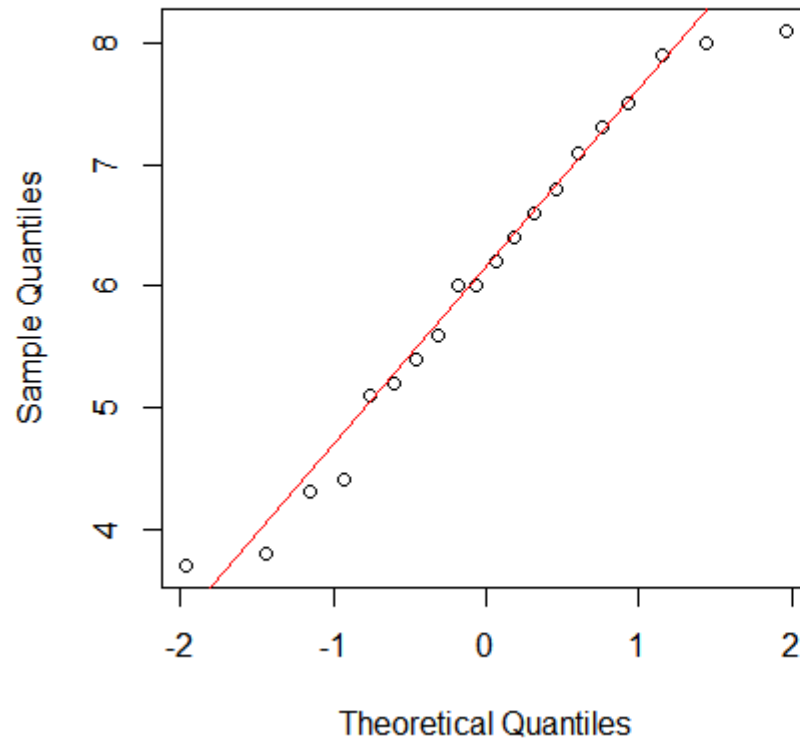
```
head(data)
```

	treatment	hoursofsleep	age	gender	comorbidity
1	Placebo	6.0	65	Female	Yes
2	Placebo	7.1	75	Female	Yes
3	Placebo	8.1	86	Female	No
4	Placebo	7.5	74	Female	No
5	Placebo	6.4	64	Female	Yes
6	Placebo	7.9	75	Male	Yes

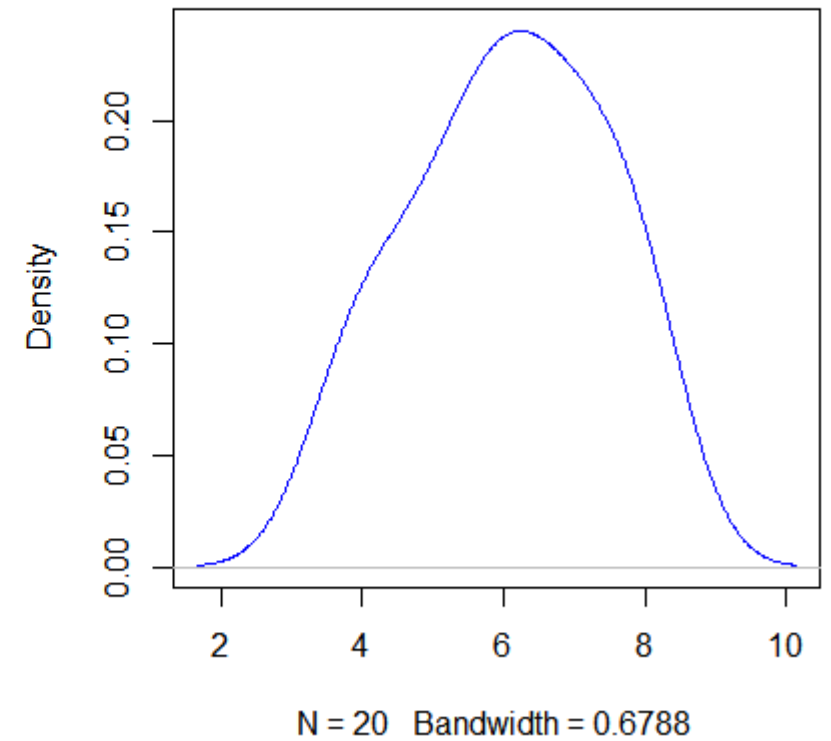
2

2.1 Thăm dò trực quan

Normal Q-Q Plot



density.default(x = data\$hoursofsleep)



```
plot(density(data$age))
```

```
par(mfrow=c(1,2))
```

```
qqnorm(data$hoursofsleep)
```

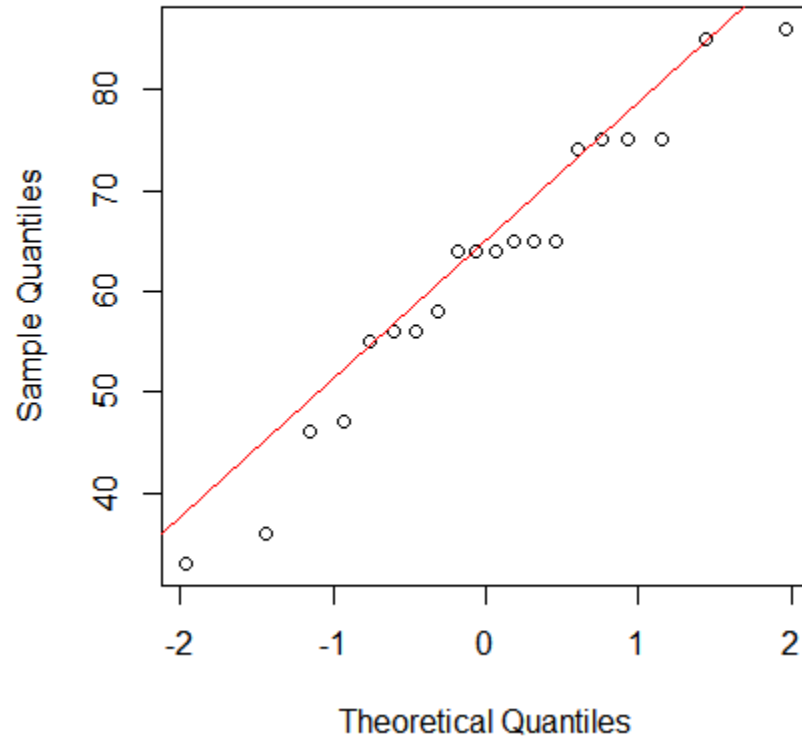
```
qqline(data$hoursofsleep,col="red")
```

```
plot(density(data$hoursofsleep),col="blue")
```

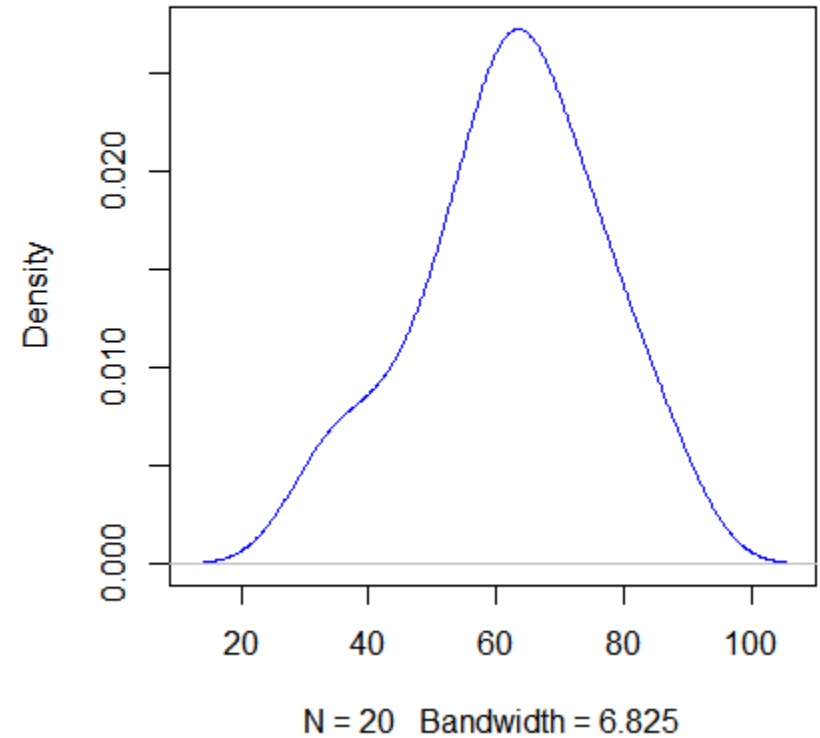
2

2.1 Thăm dò trực quan

Normal Q-Q Plot



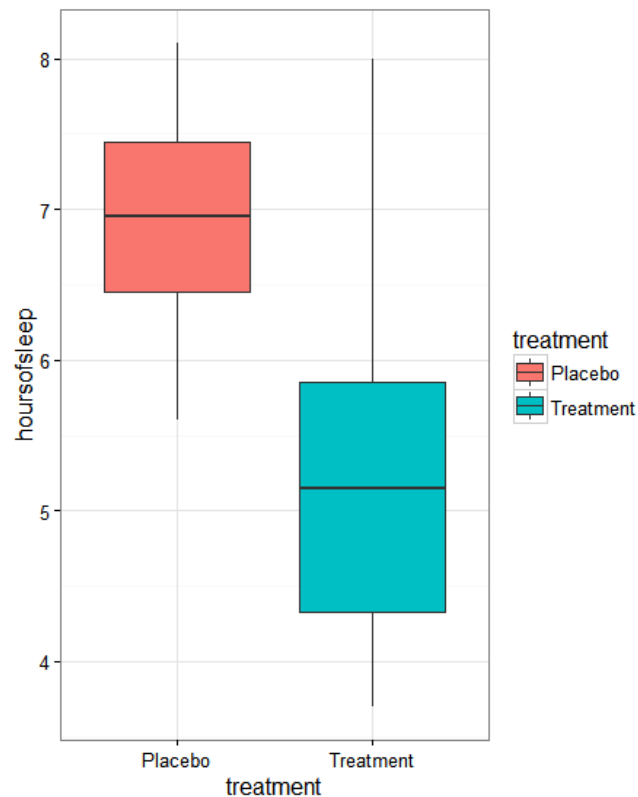
density.default(x = data\$age)



```
par(mfrow=c(1,2))  
qqnorm(data$age)  
qqline(data$age,col="red")  
plot(density(data$age),col="blue")
```

2

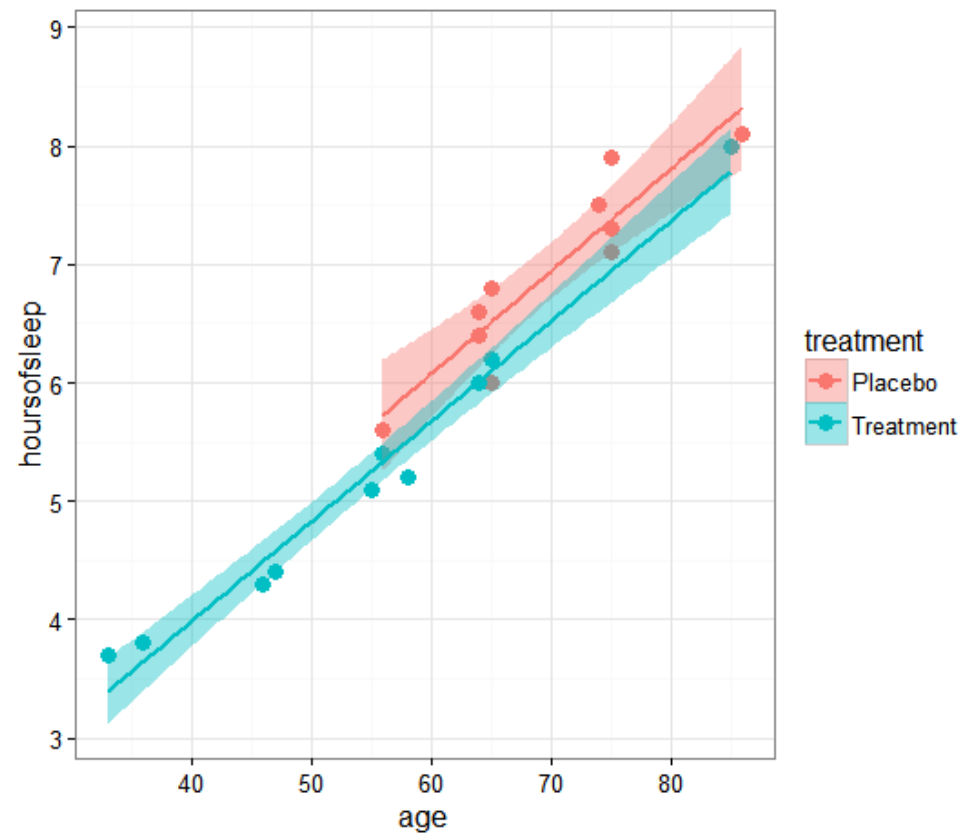
2.1 Thăm dò trực quan



```
library(ggplot2)
```

```
ggplot(data,aes(treatment,hoursofsleep,fill=treatment))+geom_boxplot()+theme_bw()
```

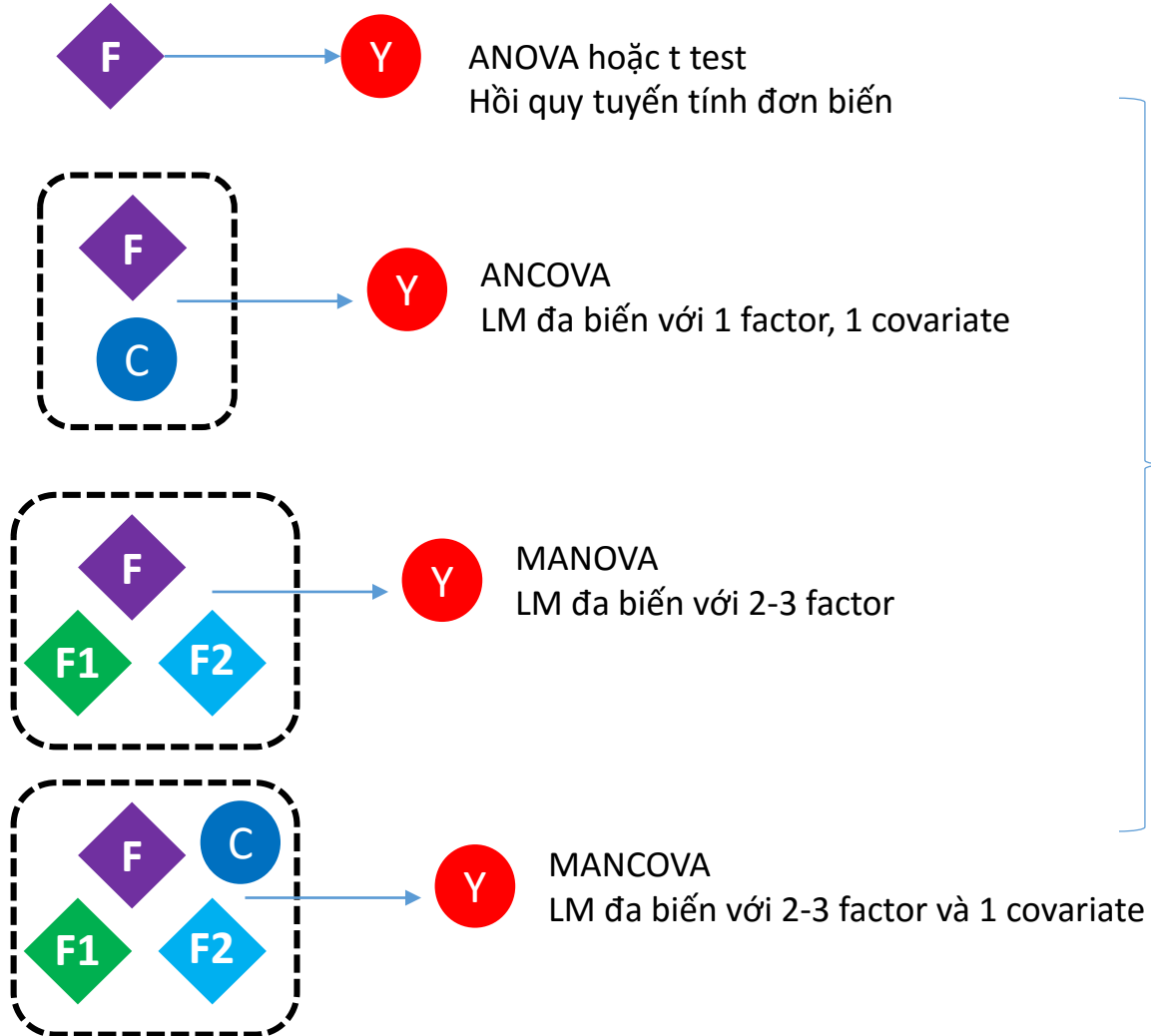
```
ggplot(data,aes(age,hoursofsleep,col=treatment,fill=treatment),alpha=0.5)+geom_point(size=3)
+geom_smooth(method="lm")+theme_bw()
```



3

3.1 Biện luận về algorithm

Thí dụ này đặt ra 1 tình huống đơn giản và thường gặp trong phân tích thống kê, chúng ta có thể nhận ra các giải pháp quen thuộc như sau:



Tất cả đều dựa vào 1 mô hình hồi quy tuyến tính

Một algorithm trong machine learning

Tài liệu gốc sử dụng case study 3 như 1 thí dụ minh họa cho algorithm Hồi quy tuyến tính

Do đó chúng tôi cũng sử dụng chính phương pháp HQTТ này, ở đây chỉ bổ sung thêm 1 số biện luận về ưu điểm của phương pháp này như sau:

Những ưu điểm của algorithm hồi quy tuyến tính

1. Tính phổ quát và đơn giản: Đây là một phương pháp rất đơn giản và thông dụng
2. Tính đa dụng : có thể dùng cho cả 2 mục đích dự báo và suy diễn thống kê, vì mô hình tuyến tính là cầu nối để thực hiện các phân tích phương sai (ANOVA, MANOVA, ANCOVA, MANCOVA).
3. Tính linh hoạt: mô hình hồi quy tổng quát có thể được biến đổi để thích ứng dễ dàng với nhiều họ phân phối và quan hệ phức tạp hơn (phi tuyến tính, count data, logistic...)
4. Tính dễ hiểu : hồi quy tuyến tính là một trong các algorithm dễ hiểu nhất và cho phép diễn giải được kết quả.

Chúng ta sẽ sử dụng giao thức CARET

Lợi ích của package caret (ngay cả khi áp dụng cho một phân tích hồi quy tuyến tính thông thường)

1. Caret cung cấp một giao thức tổng quát, có thể tái sử dụng cho nhiều algorithm khác trong tương lai
2. Đơn giản hóa quy trình Bootstrap (chọn mẫu ngẫu nhiên lặp lại) và/hoặc Cross-validation (kiểm chứng chéo).
3. Cho phép kiểm tra được độ chính xác dự báo của mô hình bằng RMSE và R2 trung bình trước khi suy diễn thống kê

Tài liệu hướng dẫn về caret có thể tìm thấy trên diễn đàn

<http://machinelearningvn.freeforums.net/>

4

4.1 Bước 1: Thăm dò bằng stepwise AIC

```
library(caret)
```

```
# Huấn luyện 1 mô hình hồi quy tuyến tính đa biến với bootstrap và lựa chọn mô hình tối ưu dựa vào AIC
```

```
Control<- trainControl(method= "boot",number=1000,summaryFunction=defaultSummary)  
set.seed(123)
```

```
glm1<-train(hoursofsleep~.,data=data,method ="glmStepAIC",trace=FALSE,trControl=Control,metric="RMSE")
```

4

4.1 Bước 1: Thăm dò bằng stepwise AIC

glm1\$results

```
parameter  RMSE  Rsquared    RMSESD  RsquaredSD
1 none 0.3432448 0.9457544 0.09191613 0.04909263
```

summary(glm1)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-0.43381	-0.18504	0.00233	0.18152	0.42197

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.763992	0.392258	1.948	0.06922	.
treatmentTreatment	-0.411078	0.139194	-2.953	0.00935	**
age	0.087228	0.005223	16.700	1.51e-11	***
genderMale	0.171940	0.124868	1.377	0.18748	

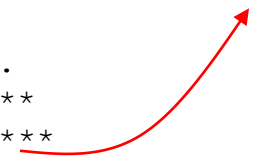
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 0.06763859)

Null deviance: 35.8220 on 19 degrees of freedom
 Residual deviance: 1.0822 on 16 degrees of freedom
 AIC: 8.4231

Number of Fisher Scoring iterations: 2

Ta đã loại được yếu tố Comorbidity
và cũng có thể loại luôn yếu tố gender
khỏi mô hình



4

4.2 Bước 2: Thăm dò hiệu ứng tương tác giữa Age và Treatment bằng stepwise AIC

```
set.seed(123)
```

```
glm2<-train(hoursofsleep~treatment*age,data=data,method="glmStepAIC",trace=FALSE,trControl=Control,metric="RMSE")
```

```
glm2$results
```

	parameter	RMSE	Rsquared	RMSESD	RsquaredSD
1	none	0.3292994	0.9529643	0.1087258	0.04232158

```
summary(glm2)
```

```
Deviance Residuals:
```

	Min	1Q	Median	3Q	Max
	-0.51352	-0.17628	-0.02299	0.17801	0.53652

```
Coefficients:
```

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.988729	0.365971	2.702	0.0151	*
treatmentTreatment	-0.411051	0.142815	-2.878	0.0104	*
age	0.084997	0.005095	16.684	5.66e-12	***

```
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for gaussian family taken to be 0.07120382)
```

```
Null deviance: 35.8220  on 19  degrees of freedom
Residual deviance:  1.2105  on 17  degrees of freedom
AIC: 8.663
```

```
Number of Fisher Scoring iterations: 2
```

Không có sự tương tác ý nghĩa giữa Tuổi và phân nhóm điều trị

Đây là mô hình cuối cùng (dùng cho mục đích dự báo)
Chỉ gồm 2 biến số: Tuổi và Phân nhóm điều trị

4

4.3 Bước 3: Dựng mô hình tối ưu bằng algorithm « lm »

Dựng mô hình hồi quy tuyến tính Hoursofsleep ~ Age + Treatment bằng phương pháp « lm » và bootstrap 1000 lần

```
set.seed(123)
```

```
lm=train(hoursofsleep~treatment+age,data=data,method ="lm",trControl=Control,metric="RMSE")
```

```
mod=lm$finalModel
```

Ghi chú: Lý do sử dụng « lm » thay vì « glm » là để chuẩn bị cho công đoạn suy diễn thống kê bằng ANOVA, cũng như kiểm tra các giả định của hồi quy tuyến tính

5

5.1 Xuất kết quả mô hình

lm\$results

	parameter	RMSE	Rsquared	RMSESD	RsquaredSD
1	none	0.2995411	0.9595418	0.07062634	0.03683456

summary(lm)

Call:

lm(formula = .outcome ~ ., data = dat)

Residuals:

	Min	1Q	Median	3Q	Max
	-0.51352	-0.17628	-0.02299	0.17801	0.53652

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	0.988729	0.365971	2.702	0.0151	*
treatmentTreatment	-0.411051	0.142815	-2.878	0.0104	*
age	0.084997	0.005095	16.684	5.66e-12	***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2668 on 17 degrees of freedom

Multiple R-squared: 0.9662, Adjusted R-squared: 0.9622

F-statistic: 243 on 2 and 17 DF, p-value: 3.125e-13

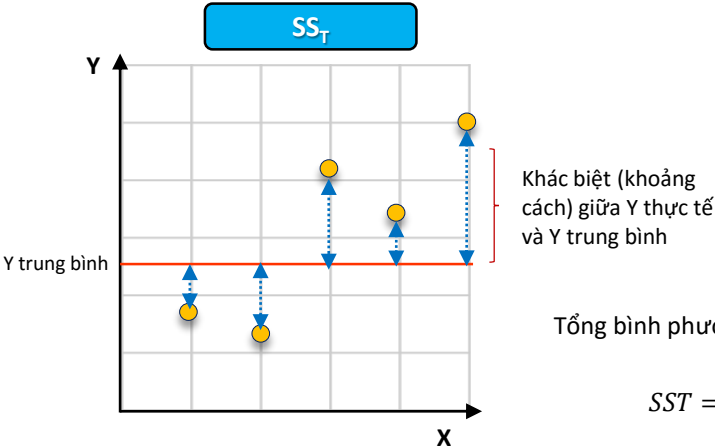
AIC(mod)

[1] 8.662987

BIC(mod)

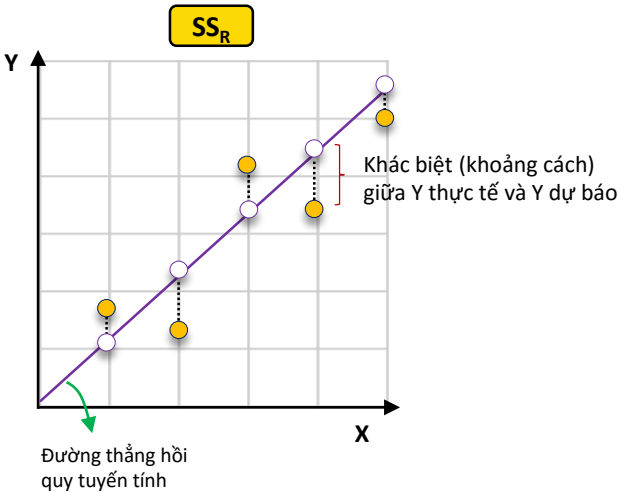
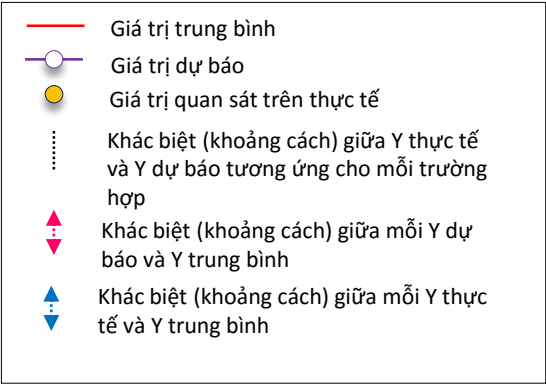
[1] 12.64592

Trên thực tế, không có mô hình nào hoàn hảo, giá trị dự báo do mô hình tính ra luôn sai biệt ít nhiều với giá trị thực tế quan sát được.



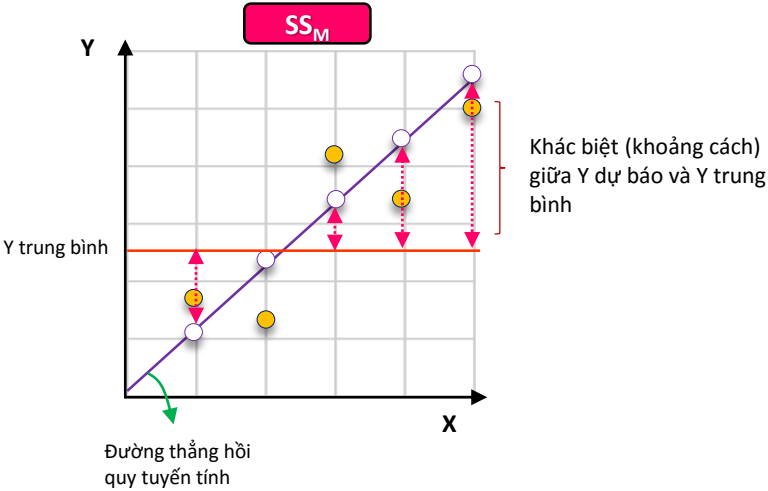
Tổng bình phương sai số: Total sum of square (TSS)

$$SST = \sum (thực - trung\ bình\ của\ dự\ báo)^2$$



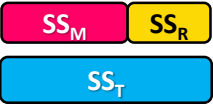
Phần sai số mà mô hình không giải thích được, gọi là sai số thặng dư hay ngẫu nhiên, bất định; được đánh giá bằng tổng bình phương sai số thặng dư: Residual sum of square

$$SSR = \sum (thực - dự\ báo)^2$$



$$SSM = \sum (trung\ bình\ của\ dự\ báo - giá\ trị\ dự\ báo)^2$$

Hệ số R2 cho biết tỉ lệ sai số do mô hình gây ra trên tổng sai số



$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST}$$

Phần sai số mà mô hình có thể giải thích được (do bản thân hiệu ứng chính của mô hình gây ra): ước tính bằng Model sum of square (MSS) hay còn gọi là regression sum of square

RMSE = Root Mean Squared Error

$$RMSE = \sqrt{\frac{1}{n} \sum (\hat{y}_i - y_i)^2}$$

AIC

Tiêu chí thông tin Akaike (1974)

$$\begin{cases} AIC = 2k + n * \ln\left(\frac{RSS}{n}\right) \\ AIC \text{ hiệu chỉnh} = 2k + n * \ln\left(\frac{RSS}{n}\right) + \frac{2k(k+1)}{(n-k-1)} \end{cases}$$

BIC

Tiêu chuẩn thông tin Bayesian (Schwarz - 1978)

$$BIC = n * \ln\left(\frac{RSS}{n}\right) + k * \ln(n)$$

Wherry **W** $R^2_{cor} = 1 - (1 - R^2) * \frac{(n-1)}{(n-k)}$

McNemar **M** $R^2_{cor} = 1 - (1 - R^2) * \frac{(n-1)}{(n-k-1)}$

Lord **L** $R^2_{cor} = 1 - (1 - R^2) * \frac{(n+k-1)}{(n-k-1)}$

Stein **S** $R^2_{cor} = 1 - \left[\left(\frac{n-1}{n-k-1} \right) * \left(\frac{n-2}{n-k-2} \right) * \left(\frac{n+1}{n} \right) * (1 - R^2) \right]$

Hệ số R^2 được định nghĩa như phần biến thiên của Y mà mô hình có khả năng giải thích được, hay nói cách khác, $R^2 = 1$ trừ cho phần bất định (ngẫu nhiên, không giải thích được, hay SSR/SST).

Chúng ta nên dùng giá trị R^2 hiệu chỉnh thay vì giá trị gốc. Có 4 công thức khác nhau cho R^2 hiệu chỉnh. Package stats trong R sử dụng công thức của Wherry.

Các phương pháp còn lại gồm McNemar, Lord và Stein.
Một số tác giả đề nghị dùng R^2 hiệu chỉnh theo Stein (2002).

Ghi chú: RSS = tổng bình phương sai số (được cung cấp trong bảng ANOVA)
k= số lượng tham số trong mô hình, bao gồm hằng số bo và những biến độc lập
n= cỡ mẫu

5

5.2 Phân tích phương sai

anova(mod)

Analysis of Variance Table

Response: .outcome

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
treatmentTreatment	1	14.7920	14.7920	207.74	5.822e-11 ***
age	1	19.8195	19.8195	278.35	5.660e-12 ***
Residuals	17	1.2105	0.0712		

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Phân tích phương sai

library(lmSupport)

modelEffectSizes(mod)

lm(formula = .outcome ~ ., data = dat)

Coefficients

	SSR	df	pEta-sqr	dR-sqr
(Intercept)	0.5197	1	0.3004	NA
treatmentTreatment	0.5899	1	0.3276	0.0165
age	19.8195	1	0.9424	0.5533

Sum of squared errors (SSE): 1.2

Sum of squared total (SST): 35.8

Tính Effect-size cho mỗi yếu tố = partial Etasquared

5.2 Phân tích phương sai

Giải thích: Bảng ANOVA trình bày kết quả phân tích phương sai (test Fisher) cho mô hình, mục tiêu là chứng minh mô hình với k tham số cho phép giải thích phần lớn biến thiên của Y so với giá trị trung bình quan sát được trong mẫu khảo sát n trường hợp.

Trị số F được tính bằng tỉ lệ giữa trung bình bình phương của hiệu ứng chính (do mô hình gây ra: MSM) và trung bình bình phương do sai số ngẫu nhiên (MSR): $F = \text{MSM} / \text{MSR}$.

F có phân phối χ^2 , với độ tự do của MSM và MSR, ta có thể tính được xác suất để có giá trị $F = F$ quan sát được.

Nếu $p < 0,05$ tức là F tìm được thực sự lớn, hay nói cách khác hiệu ứng chính của mô hình cao hơn nhiều so với sai số ngẫu nhiên → Mô hình có ý nghĩa thống kê.

Ngoài ra, bảng ANOVA còn cung cấp thông tin khác rất quan trọng về tổng bình phương (Sum of square): có 3 loại SS:

MSS hay Regression SS = phần biến thiên mà mô hình có thể giải thích được.

RSS = Residual SS hay phần biến thiên mô hình không giải thích được = phần bất định, ngẫu nhiên;

và TSS hay Total SS = Tổng biến thiên quan sát được trong quần thể n.

Giá trị RSS cho phép tính được (thủ công) 2 trị số : AIC (tiêu chuẩn thông tin Akaike) và BIC (Tiêu chuẩn thông tin Bayesian hay Schwarz), là 2 tiêu chí để so sánh phẩm chất giữa các mô hình với nhau, nhằm chọn ra mô hình tối ưu.

5.3 Khảo sát quy trình bootstrap

```
resamp=as.data.frame(lm$resample)
```

```
library(ggfortify)
```

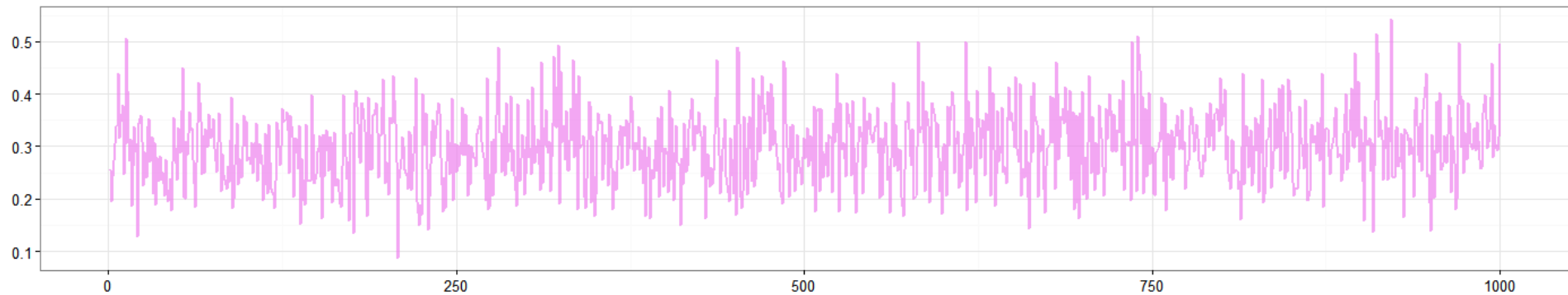
```
RMSE=ts(resamp$RMSE)
```

```
Rsquared=ts(resamp$Rsquared)
```

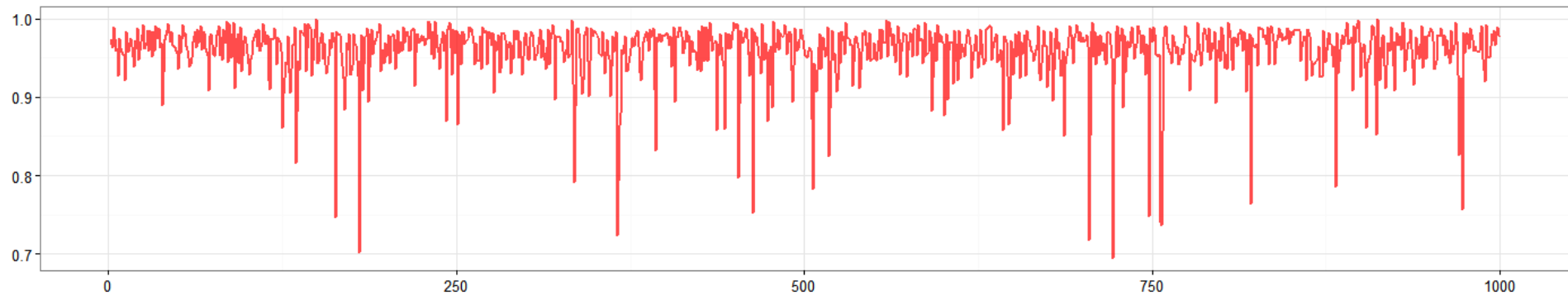
```
autoplot(RMSE,ts.colour="violet",size=0.8,alpha=0.7)+ggplot2::theme_bw()
```

```
autoplot(Rsquared,ts.colour="red",size=0.8,alpha=0.7)+ggplot2::theme_bw()
```

MRSE



R²

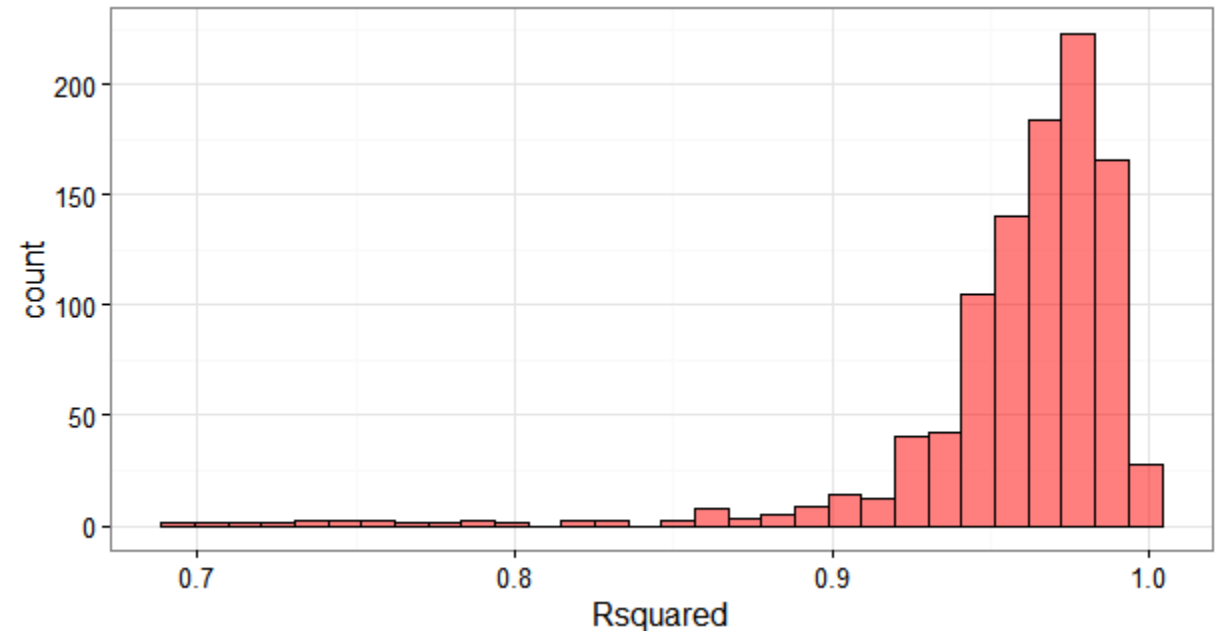
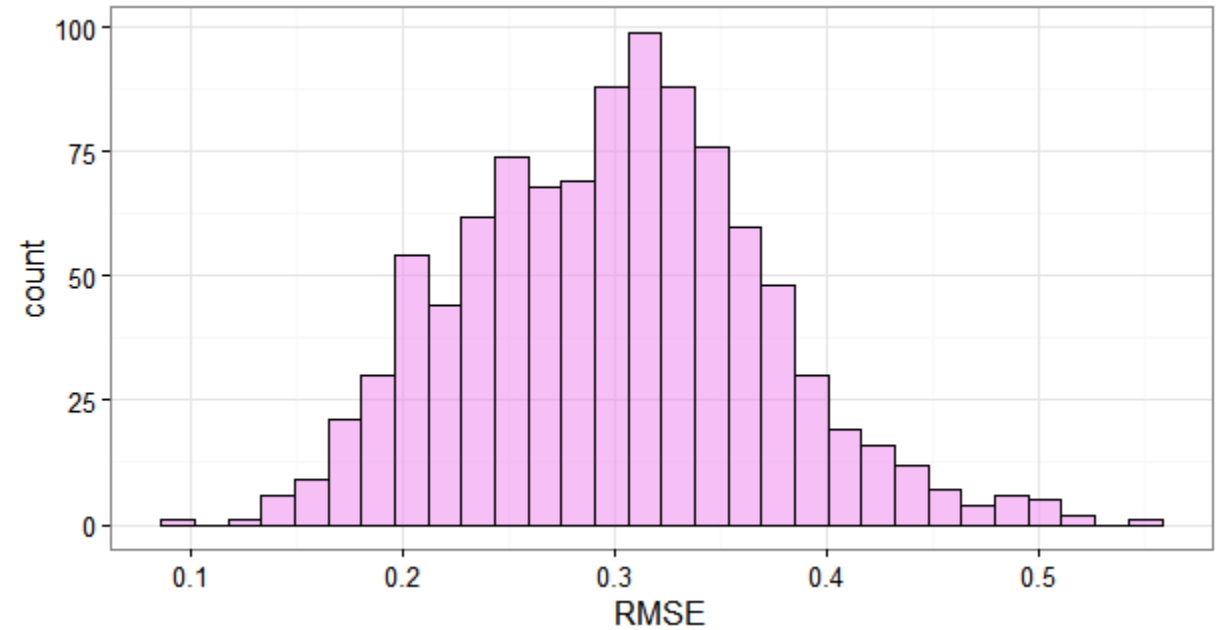


5

5.3 Khảo sát quy trình bootstrap

```
ggplot(data=resamp,aes(x=RMSE))+geom_histogram(colour="black",fill="violet",alpha=0.5)+theme_bw()
```

```
ggplot(data=resamp,aes(x=Rsquared))+geom_histogram(colour="black",fill="red",alpha=0.5)+theme_bw()
```



5

5.4 Vai trò của mỗi yếu tố

```
library(relaimpo)
```

```
boot <- boot.relimp(mod, b = 1000, type = c("lmg", "last", "first", "pratt"), rank = TRUE, diff = TRUE, rela = TRUE)
```

```
booteval.relimp(boot)
```

```
plot(booteval.relimp(boot, sort=FALSE))
```

```
Response variable: .outcome
Total response variance: 1.885368
Analysis based on 20 observations
```

```
2 Regressors:
treatmentTreatment age
Proportion of variance explained by model: 96.62%
Metrics are normalized to sum to 100% (rela=TRUE).
```

```
Relative importance metrics:
```

	lmg	last	first	pratt
treatmentTreatment	0.2222071	0.02890117	0.3030298	0.1021346
age	0.7777929	0.97109883	0.6969702	0.8978654

```
Average coefficients for different model sizes:
```

	1X	2Xs
treatmentTreatment	-1.72000000	-0.41105052
age	0.09305202	0.08499672

5.4 Vai trò của mỗi yếu tố

Confidence interval information (1000 bootstrap replicates, bty= perc):
Relative Contributions with confidence intervals:

			Lower	Upper
	percentage	0.95	0.95	0.95
treatmentTreatment.lmg	0.2222	_B	0.0614	0.4076
age.lmg	0.7778	A_	0.5924	0.9386
treatmentTreatment.last	0.0289	_B	0.0058	0.1191
age.last	0.9711	A_	0.8809	0.9942
treatmentTreatment.first	0.3030	_B	0.1031	0.4480
age.first	0.6970	A_	0.5520	0.8969
treatmentTreatment.pratt	0.1021	_B	0.0334	0.2456
age.pratt	0.8979	A_	0.7544	0.9666

Letters indicate the ranks covered by bootstrap CIs.
(Rank bootstrap confidence intervals always obtained by percentile method)
CAUTION: Bootstrap confidence intervals can be somewhat liberal.

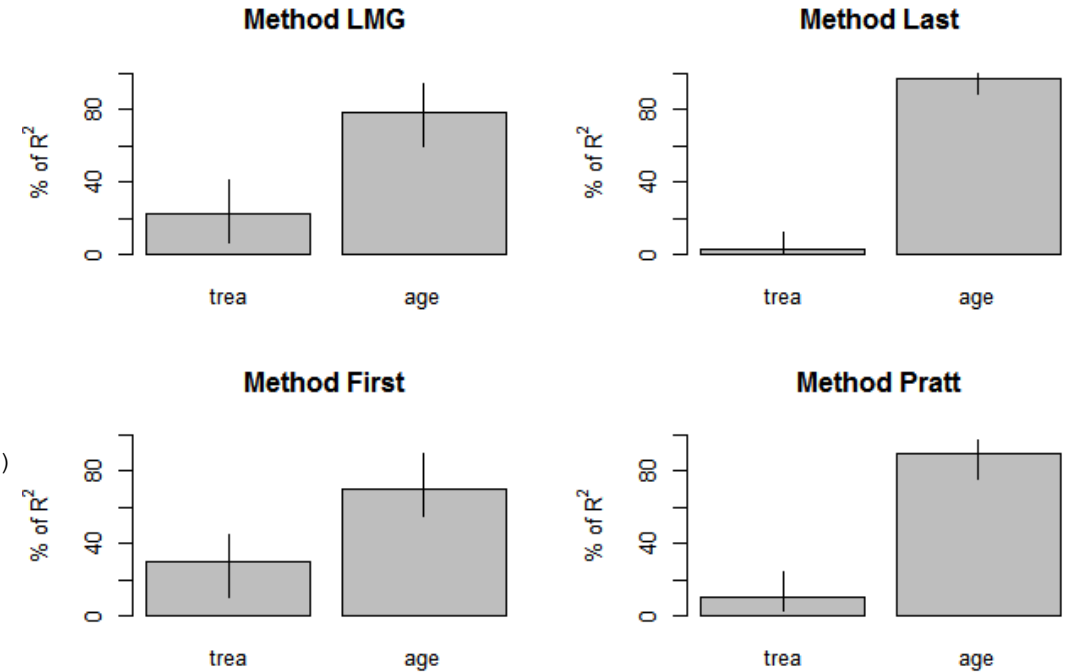
Differences between Relative Contributions:

			Lower	Upper
	difference	0.95	0.95	0.95
treatmentTreatment-age.lmg	-0.5556	*	-0.8771	-0.1848
treatmentTreatment-age.last	-0.9422	*	-0.9885	-0.7619
treatmentTreatment-age.first	-0.3939	*	-0.7937	-0.1039
treatmentTreatment-age.pratt	-0.7957	*	-0.9333	-0.5088

* indicates that CI for difference does not include 0.

CAUTION: Bootstrap confidence intervals can be somewhat liberal.

**Relative importances for .outcome
with 95% bootstrap confidence intervals**



$R^2 = 96.62\%$, metrics are normalized to sum 100%.

5

5.5 Marginal effect

```
confint(mod, level=0.95)
```

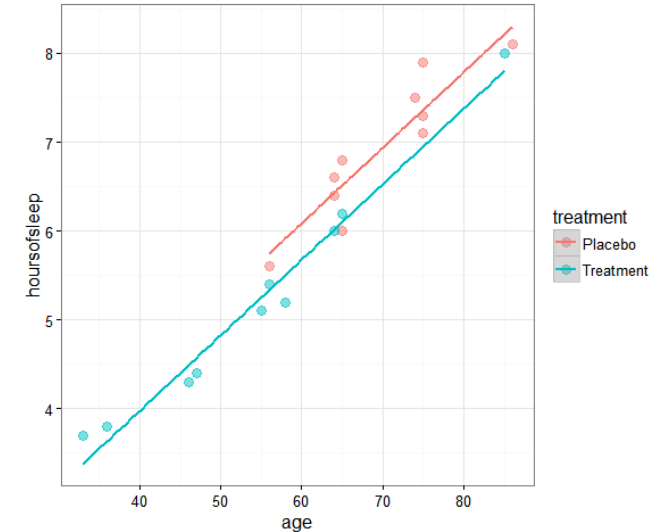
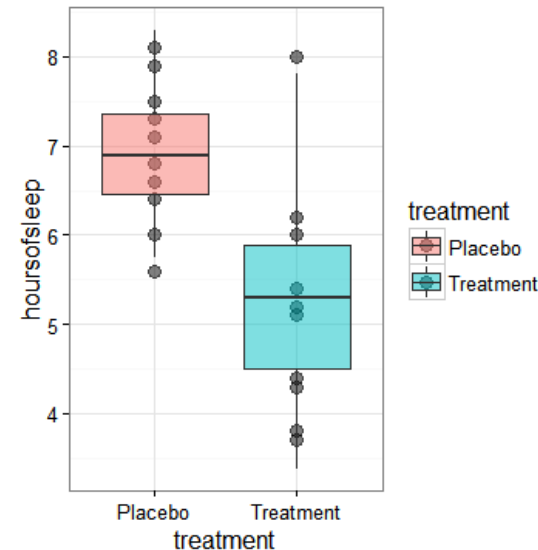
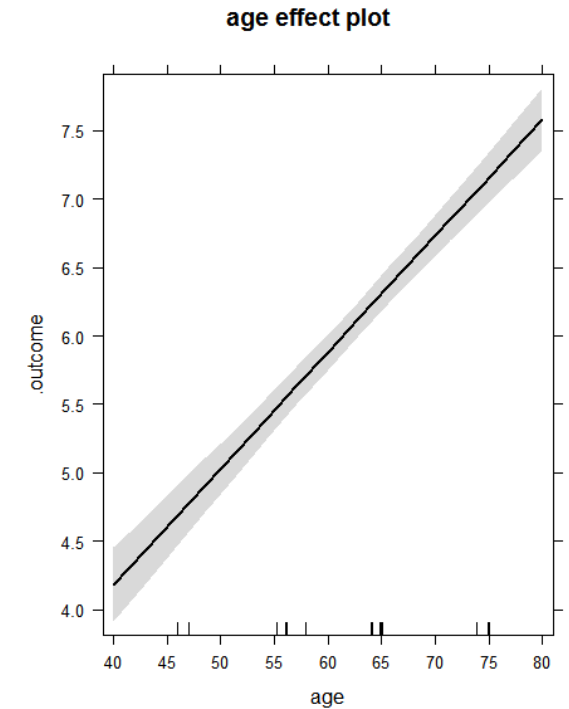
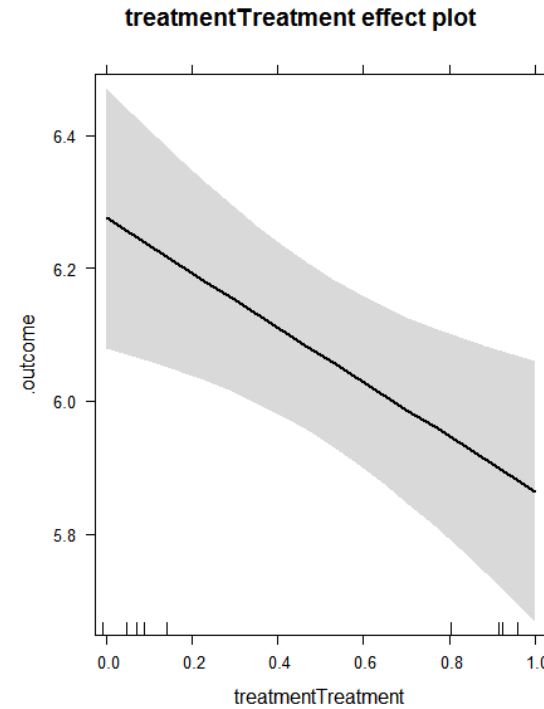
	2.5 %	97.5 %
(Intercept)	0.21659801	1.76086062
treatmentTreatment	-0.71236406	-0.10973698
age	0.07424813	0.09574531

```
library(effects)
plot(allEffects(mod))
```

```
data$pred=predict(lm,data)
```

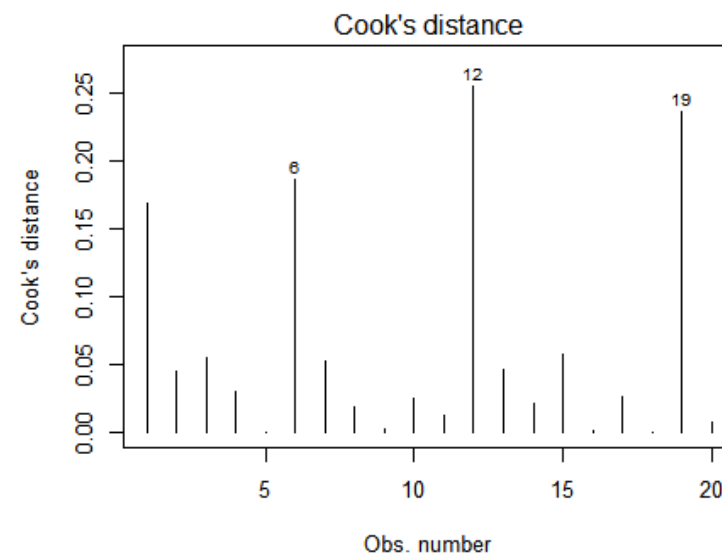
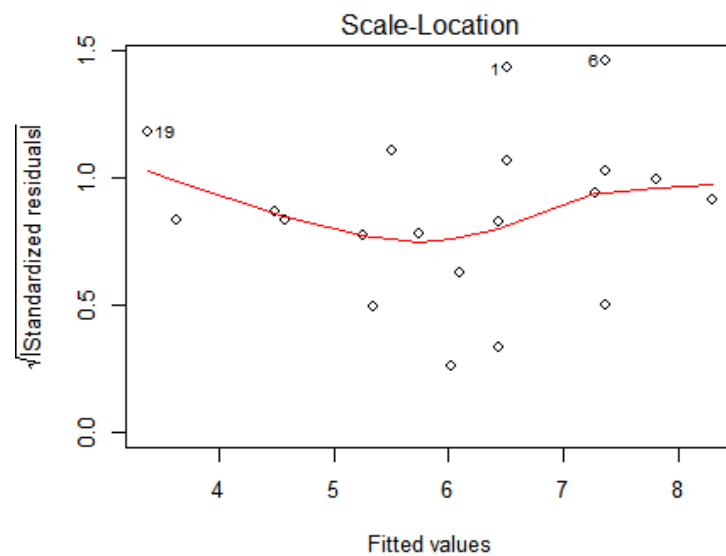
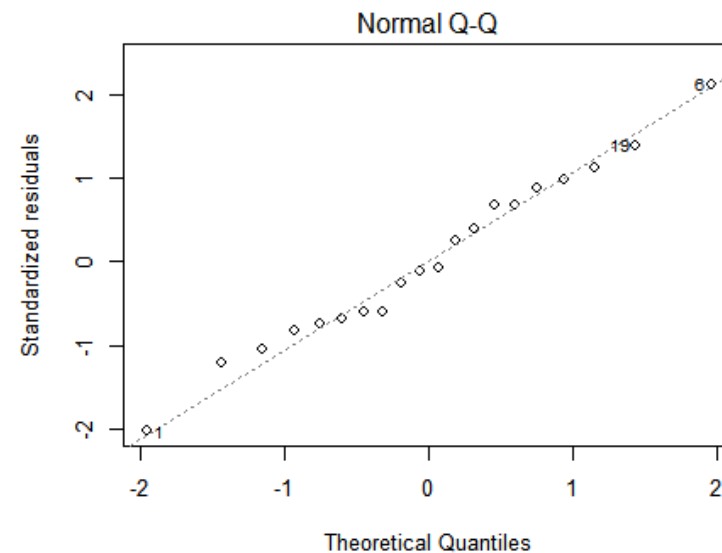
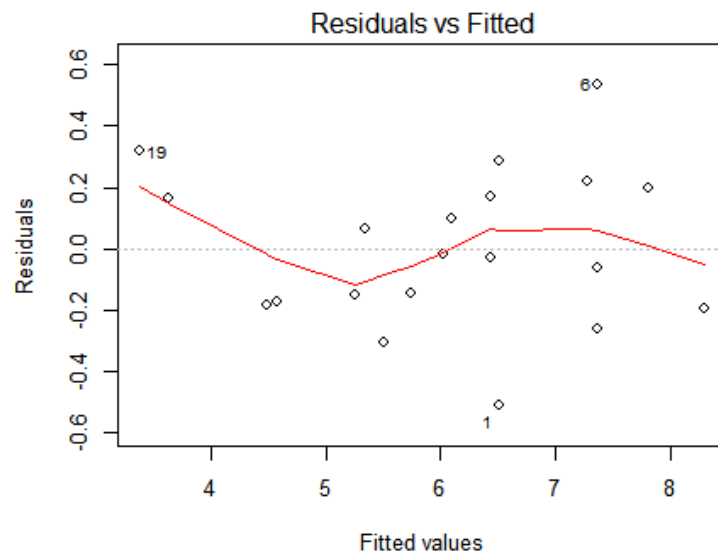
```
ggplot(data,aes(color=treatment))+geom_point(aes(x=age,y=hoursofsleep),size=3,alpha=0.5)+geom_smooth(aes(x=age,y=pred),method="lm")+theme_bw()
```

```
ggplot(data,aes(fill=treatment))+geom_point(aes(x=treatment,y=hoursofsleep),size=3,alpha=0.5)+geom_boxplot(aes(x=treatment,y=pred),alpha=0.5)+theme_bw()
```



6.1 Kiểm tra mô hình

```
par(mfrow=c(2,2))
plot(mod, which=1:4)
```



6

6.1 Kiểm tra mô hình

Chẩn đoán về Outliers và những trường hợp ảnh hưởng bất thường

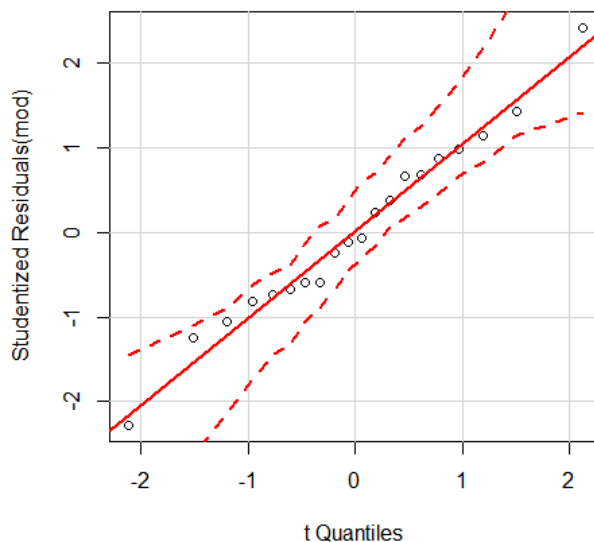
```
library(car)
```

```
outlierTest(mod)
```

```
qqPlot(mod)
```

```
avPlots(mod)
```

```
leveragePlots(mod)
```

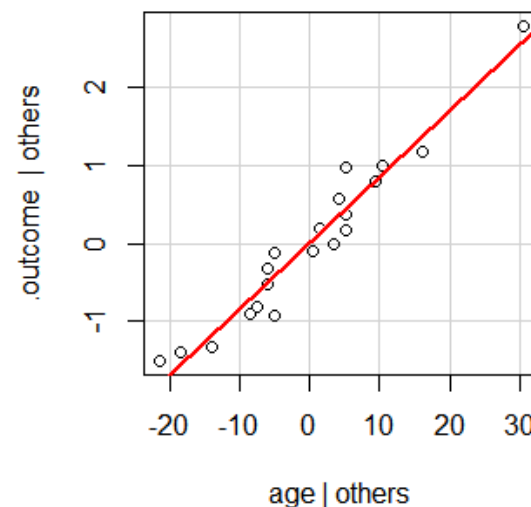
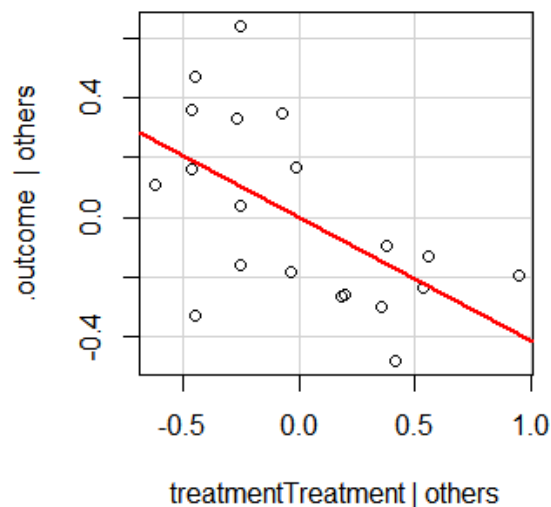


```
No Studentized residuals with Bonferonni  $p < 0.05$ 
```

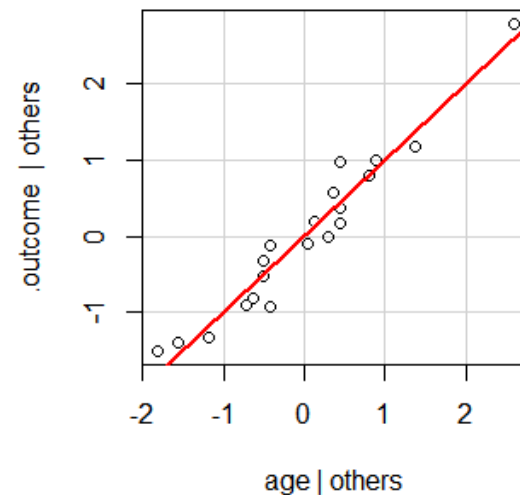
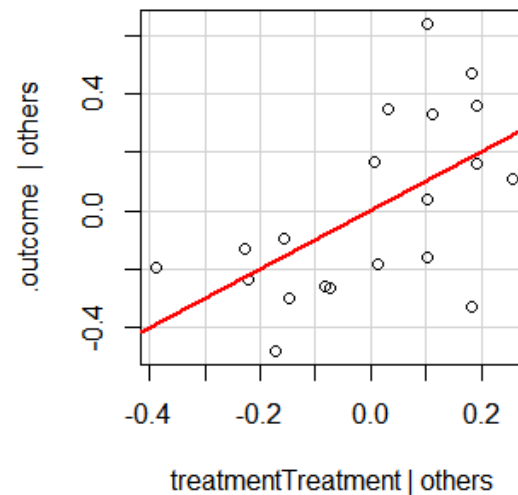
```
Largest |rstudent|:
```

	rstudent	unadjusted p-value	Bonferonni p
6	2.414373	0.028103	0.56206

Added-Variable Plots



Leverage Plots



influence.measures(mod)

```
Influence measures of
lm(formula = .outcome ~ ., data = dat) :

   dfb.1_ dfb.trtT dfb.age dffit cov.r  cook.d  hat inf
1 -0.39509  0.57413  0.22545 -0.7947 0.581 0.169014 0.109
2  0.02428  0.14830 -0.10829 -0.3680 1.103 0.044868 0.109
3  0.20681  0.01644 -0.28105 -0.4032 1.316 0.055259 0.194
4 -0.00300 -0.13264  0.07218  0.3004 1.167 0.030511 0.106
5 -0.02134  0.02908  0.01317 -0.0392 1.349 0.000545 0.113
6 -0.05586 -0.34118  0.24912  0.8465 0.530 0.186035 0.109
7  0.19933 -0.28965 -0.11374  0.4009 1.061 0.052604 0.109
8  0.13004 -0.17720 -0.08025  0.2392 1.244 0.019703 0.113
9  0.00567  0.03463 -0.02528 -0.0859 1.332 0.002605 0.109
10 -0.21799  0.21897  0.17470 -0.2718 1.353 0.025581 0.170
11  0.00578 -0.11963 -0.00595 -0.1970 1.249 0.013456 0.100
12 -0.74753  0.66853  0.76823  0.8742 1.790 0.255099 0.439 *
13  0.26558 -0.00556 -0.27293  0.3663 1.420 0.046194 0.225
14 -0.10079 -0.07825  0.10358 -0.2511 1.253 0.021708 0.121
15  0.08496 -0.29210 -0.08731 -0.4223 1.019 0.057655 0.104
16 -0.00706  0.05129  0.00725  0.0804 1.320 0.002281 0.101
17 -0.12533 -0.07754  0.12880 -0.2821 1.240 0.027249 0.126
18  0.01293 -0.02099 -0.01329 -0.0267 1.382 0.000253 0.133
19  0.66939 -0.06479 -0.68792  0.8684 1.141 0.236686 0.268
20 -0.08082  0.12304  0.08306  0.1551 1.357 0.008445 0.140
```

summary(influence.measures(mod))

```
Potentially influential observations of
lm(formula = .outcome ~ ., data = dat) :

   dfb.1_ dfb.trtT dfb.age dffit cov.r  cook.d hat
12 -0.75  0.67  0.77  0.87  1.79_*  0.26  0.44
```

1) Trị số DfBeta chính là khác biệt giữa 2 hệ số hồi quy beta cho 1 yếu tố dự báo X :
(1) Hệ số Beta ước tính từ toàn bộ các trường hợp trong mẫu khảo sát n và (2) Hệ số Beta ước tính sau khi loại bỏ trường hợp i mà ta đang xét. Ngưỡng nguy cơ là giá trị tuyệt đối của dfBeta = 1, bất cứ trường hợp nào có dfBeta > 1 được xem là có ảnh hưởng lớn đến hệ số hồi quy beta và có khả năng làm sai lệch dự báo

2) Leverage value (còn gọi là Hat value) : Giá trị Leverage đo lường ảnh hưởng của giá trị thực tế quan sát được lên giá trị dự báo của Y bởi mô hình. Leverage càng cao > giá trị trung bình càng cho thấy ảnh hưởng lên giá trị dự báo (1 trường hợp cá biệt)

Những trường hợp nào có Leverage > 2-3 lần giá trị trung bình cần được nghi ngờ vì mức độ ảnh hưởng cá biệt làm thay đổi giá trị dự báo.

3) tỉ số hiệp phương sai (Covariance ratio): cov.r

Tỉ số này đánh giá khả năng 1 trường hợp cá biệt có thể ảnh hưởng đến phương sai của các tham số hồi quy (yếu tố dự báo).

Giá trị CVR gần bằng 1 chứng tỏ ảnh hưởng ít (tốt)
Ta diễn giải CVR dựa vào giá trị ngưỡng trên và ngưỡng dưới:

Ngưỡng trên UL = $1 + 3(\text{Leverage trung bình})$ hay $= 1 + (3(k+1)/n)$
Ngưỡng dưới LL = $1 - 3(\text{Leverage trung bình})$ hay $= 1 - (3(k+1)/n)$

Nếu 1 trường hợp i có CVRi > UL: Loại bỏ trường hợp i này sẽ làm tổn hại đến độ chính xác của mô hình (ý nghĩa cảnh báo về tính cần thiết của trường hợp i)

Nếu 1 trường hợp có CVRi < LL: Loại bỏ trường hợp i này sẽ làm tăng tính chính xác của mô hình (ý nghĩa cảnh báo về nguy cơ làm sai lệch dự báo của trường hợp i).

4) Cook's distance cho phép đánh giá ảnh hưởng của một trường hợp cá thể lên mô hình (thiết lập năm 1982 bởi Cook và Weisberg). Xem slide tiếp theo

6

6.1 Kiểm tra mô hình

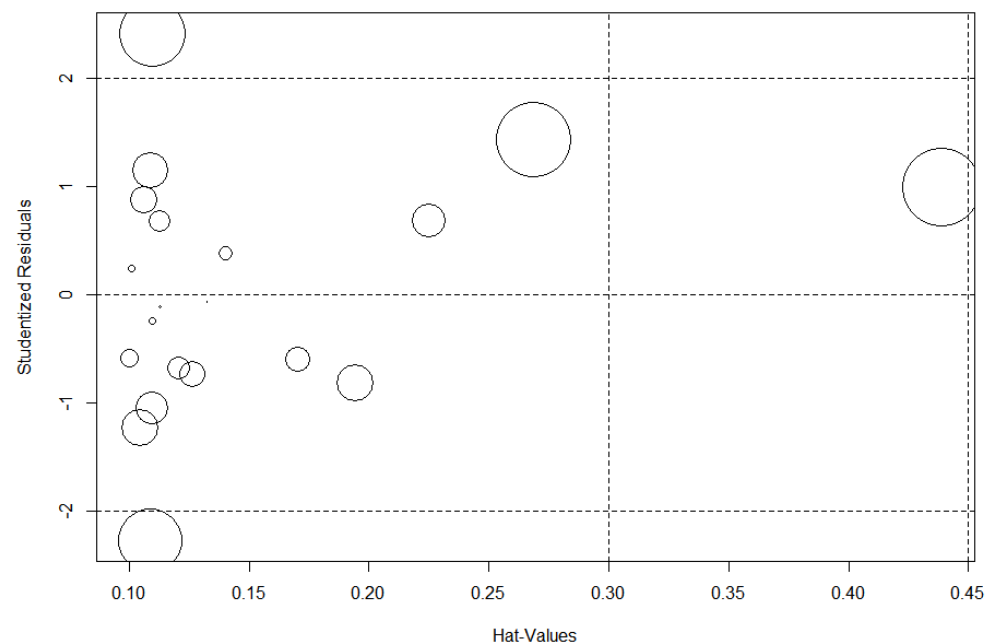
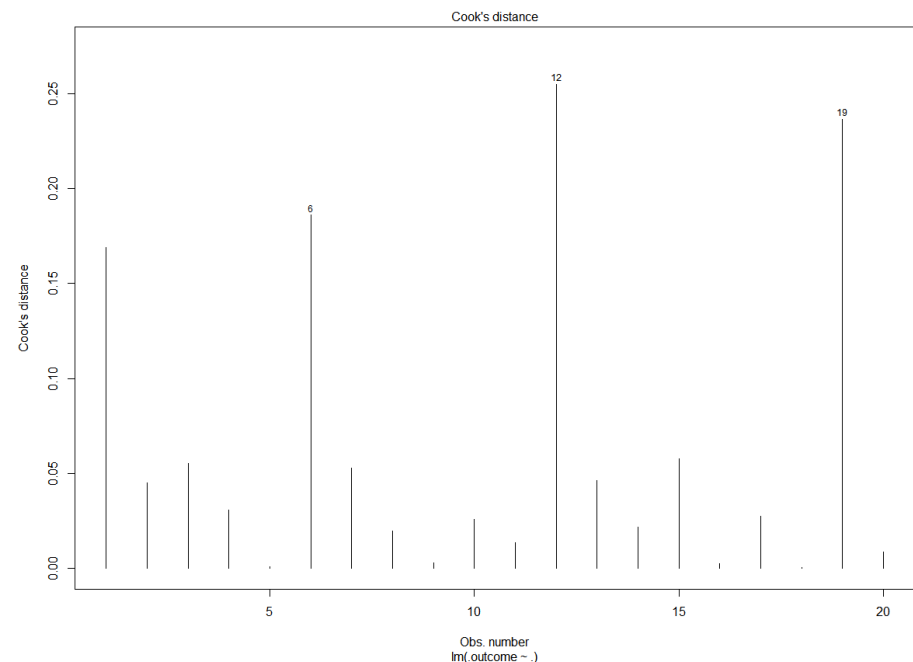
Chẩn đoán Cook's distance

```
cutoff <- 4/((nrow(data)-length(mod$coefficients)-2))
cutoff
summary(cooks.distance(mod)>=cutoff)
plot(mod, which=4, cook.levels=cutoff)
```

```
cutoff
[1] 0.2666667
```

```
Mode      FALSE      NA's
logical    20        0
```

```
influencePlot(mod, id.method="identify")
```

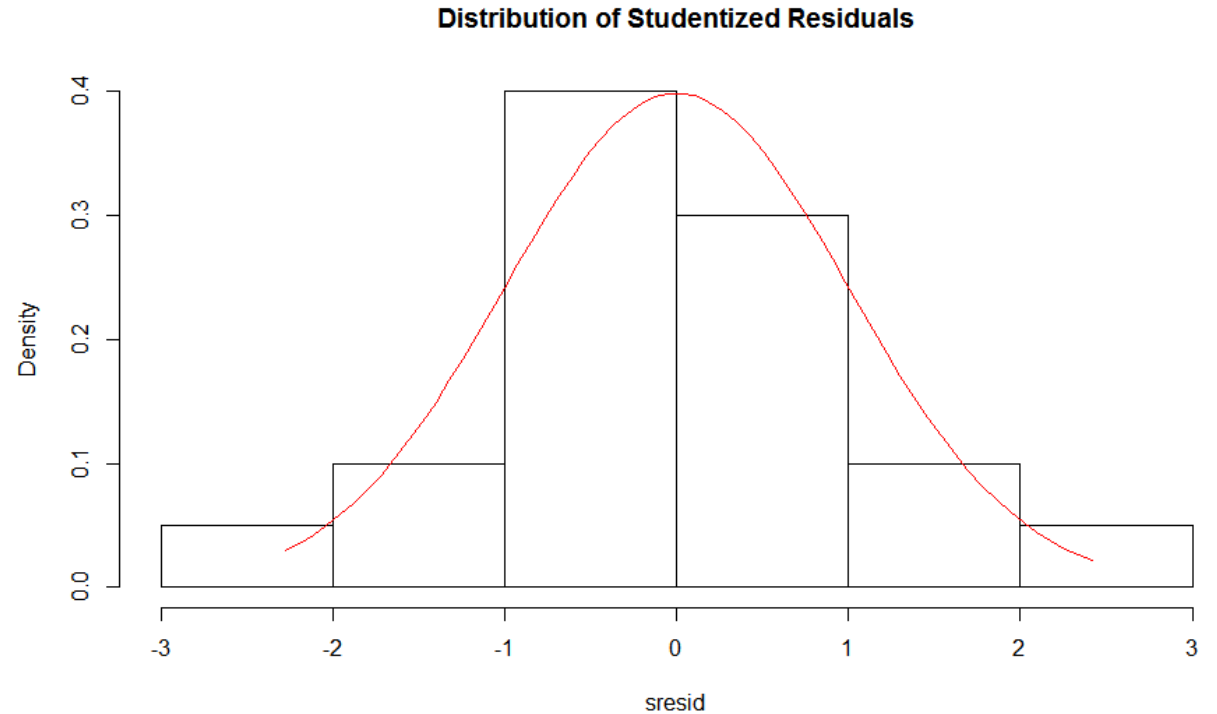


6

6.1 Kiểm tra mô hình

Chẩn đoán về phân phối của sai số chuẩn hóa

```
library(MASS)
sresid <- studres(mod)
hist(sresid, freq=FALSE, main="Distribution of Studentized Residuals")
xfit<-seq(min(sresid),max(sresid),length=60)
yfit<-dnorm(xfit)
lines(xfit, yfit,col="red")
```



Vẽ histogram của sai số chuẩn hóa: Nếu phân phối bình thường là tốt, vì sai số được phân phối chuẩn.

Dấu hiệu phân phối bình thường là đường cong hình chuông úp có đỉnh trùng với giá trị ZRE = 0,

Lưu ý: phân phối này nhạy với cỡ mẫu, n quá nhỏ dễ dẫn tới phân phối không bình thường.

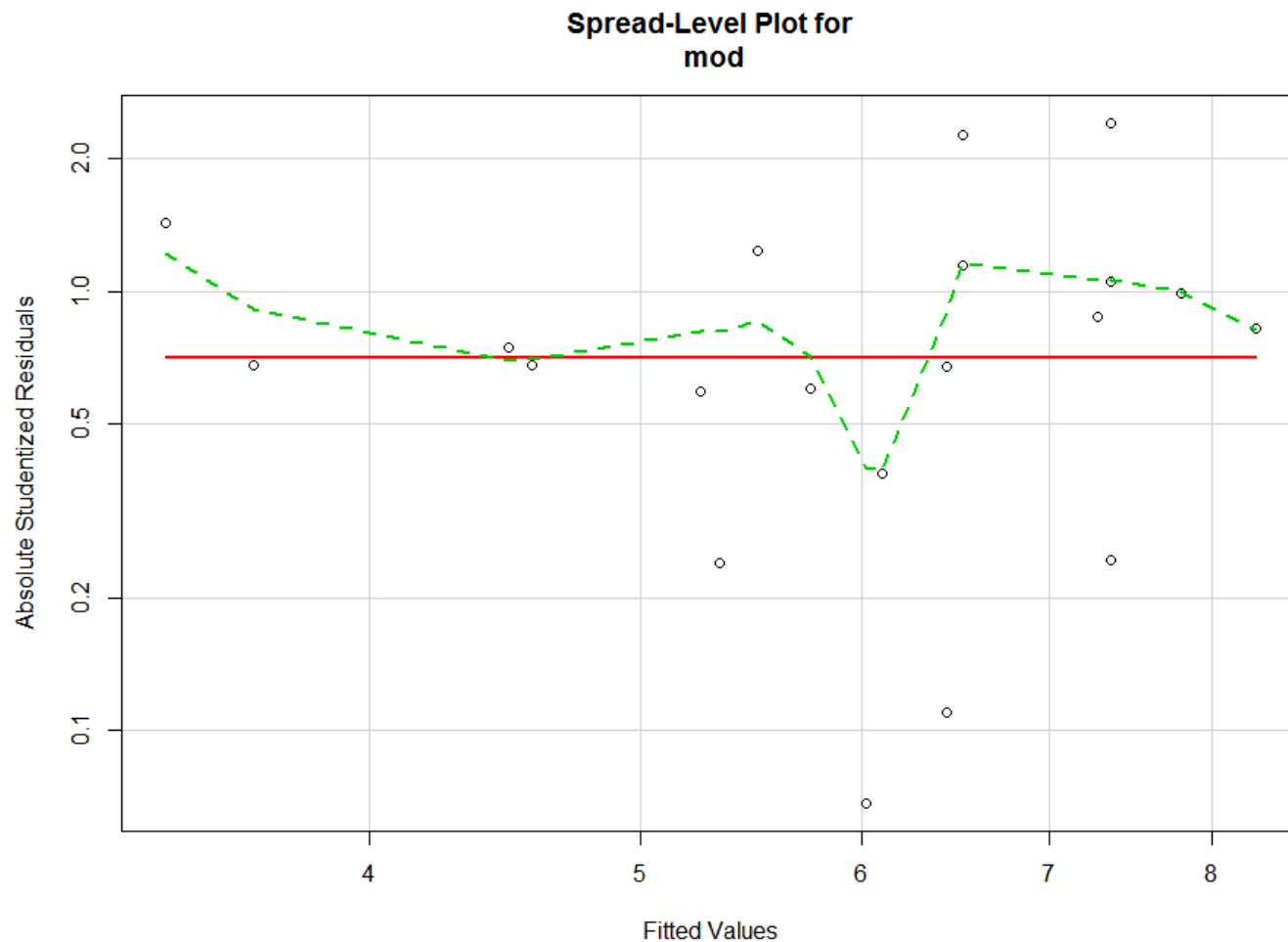
Khảo sát Non-constant Error Variance

`ncvTest(mod)`

Non-constant Variance Score Test
Variance formula: $\sim \text{fitted.values}$
Chisquare = 0.4273279 Df = 1
p = 0.513303

`spreadLevelPlot(mod)`

Suggested power transformation: 1.008191



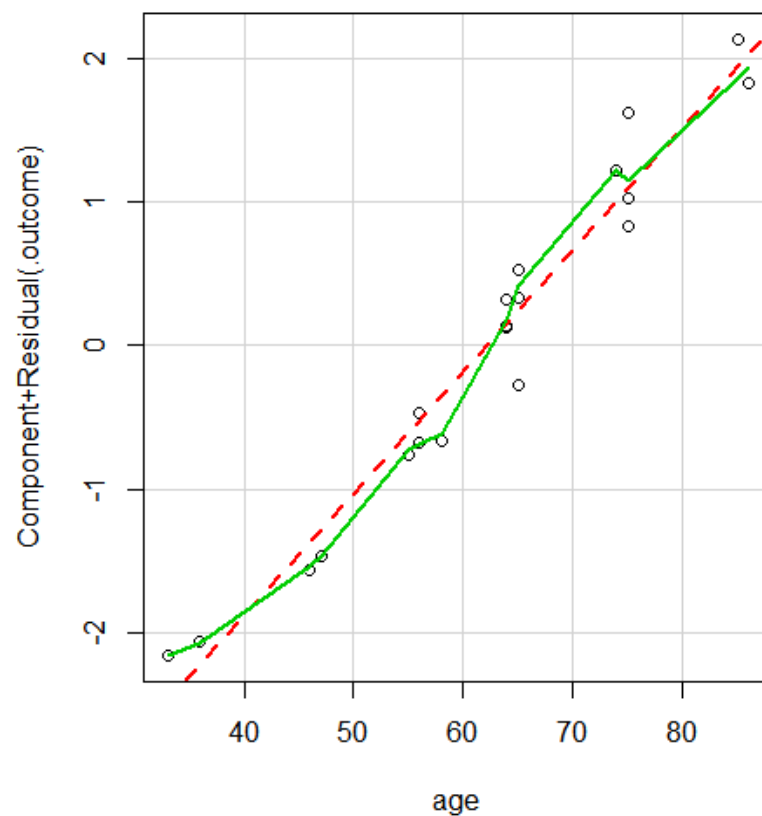
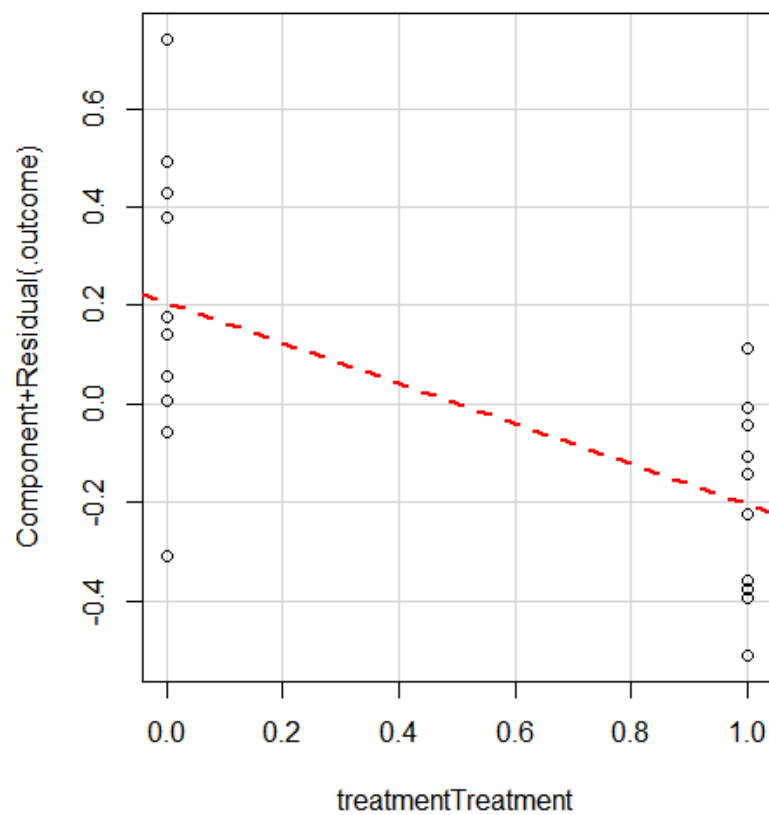
6

6.1 Kiểm tra mô hình

Chẩn đoán quan hệ phi tuyến tính

`crPlots(mod)`

Component + Residual Plots



6

6.1 Kiểm tra mô hình

Chẩn đoán Tự tương quan / Đa cộng tuyến

durbinWatsonTest(mod)

```
lag Autocorrelation D-W Statistic p-value
1      0.3063112      1.161669    0.034
Alternative hypothesis: rho != 0
```

vif(mod)

```
treatmentTreatment      age
1.432237                1.432237
```

sqrt(vif(mod)) > 2

```
treatmentTreatment      age
FALSE                FALSE
```

Kiểm tra các giả định về phân phối

```
library(gvlma)
gvmodel <- gvlma(mod)
summary(gvmodel)
```

```
ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance = 0.05
```

```
Call:
gvlma(x = mod)
```

	Value	p-value	Decision
Global Stat	4.03904	0.4007	Assumptions acceptable.
Skewness	0.03291	0.8561	Assumptions acceptable.
Kurtosis	0.10841	0.7420	Assumptions acceptable.
Link Function	1.35134	0.2450	Assumptions acceptable.
Heteroscedasticity	2.54639	0.1105	Assumptions acceptable.

Những kỹ năng :

1. Sử dụng caret như một giao thức đơn giản để huấn luyện mô hình hồi quy tuyến tính bằng 2 phương pháp: stepwise và lm
2. Trích xuất nội dung mô hình từ caret để thực hiện kiểm tra phẩm chất mô hình và phân tích phương sai (ANOVA)

Thông điệp của bài thực hành:

Nội dung Machine learning trong thí dụ này rất ít , bao gồm:

- + Hồi quy tuyến tính (vay mượn từ thống kê truyền thống) như 1 algorithm để dự báo kết quả định lượng từ dataset nhiều biến số
- + Bootstrap là một phương pháp lấy mẫu ngẫu nhiên
- + Huấn luyện mô hình bằng caret
- + Sử dụng stepwise regression để thăm dò mô hình tối ưu và loại bỏ những biến số không có đóng góp quan trọng.

Nghiên cứu này đã được đơn giản hóa với cỡ mẫu nhỏ, ít biến số (và có phân phối chuẩn), vì mục tiêu của tác giả là dùng cách tiếp cận đơn giản, quen thuộc để

- + giới thiệu về hồi quy tuyến tính như 1 algorithm trong Machine learning

- + cho thấy cách mà Machine learning giao thoa với Thống kê học thông qua những công cụ thống kê cổ điển

Sau khi thực hành xong, chúng ta có thể nhận ra Machine learning có thể trợ giúp cho chúng ta như thế nào một khi tình huống trở nên phức tạp hơn, thí dụ

Một dataset với rất nhiều biến số: thí dụ không chỉ có 5 mà tới hàng chục biến

Một mẫu có kích thước lớn hơn rất nhiều: không chỉ 20 mà là hàng trăm, hàng ngàn trường hợp quan sát, cho phép sử dụng nhiều cách huấn luyện và kiểm định, thí dụ k-fold cross-validation

R làm được tất cả những gì SPSS có thể làm, và hơn thế nữa

	IBM-SPSS	R
Stepwise regression	Có	Có (MASS)
Bootstrap	Có	Có (+ tái lập được kết quả)
Bảng ANOVA và Etasquared	Có	Có
RMSE	Không	Có (caret)
R2 và Adjusted R2	Có	Có
AIC	Không	Có
Chẩn đoán mô hình	Không	Có (car và những packages khác)
Đồ thị	Giới hạn	Rất phong phú và đẹp
Resample	Không	Có (caret)
Marginal effect	Có	Có