Linear Regression: A Revisit

Yifei Sun

Contents

Data **2**

Multiple linear regression: a small experiment **3**

Prediction interval vs. confidence interval **3**

Best subset model selection **4**

```r
library(tidyverse)
library(summarytools)
library(leaps)
```

Data

In this example, we assess the association between high density lipoprotein (HDL) cholesterol and body mass index, blood pressure, and other demographic factors (age, gender, race) using the NHANES data (https://wwwn.cdc.gov/nchs/nhanes/ContinuousNhanes/Default.aspx?BeginYear=2001). The data can be downloaded using functions in the package RNHANES.

```r
load("L4_data.RData")
```

Summary statistics of the predictors and the response:

```r
st_options(plain.ascii = FALSE,
           style = "rmarkdown",
           dfSummary.silent = TRUE,
           footnote = NA,
           subtitle.emphasis = FALSE)

dfSummary(dat[,-1])
```

Data Frame Summary

dat
Dimensions: 6434 x 6
Duplicates: 0

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
1	gender	1. 1	3108 (48.3%)	IIIIIIIII	6434	0
	[factor]	2. 2	3326 (51.7%)	IIIIIIIIII	(100%)	(0%)
2	race	1. 1	1593 (24.8%)	IIII	6434	0
	[factor]	2. 2	262 (4.1%)		(100%)	(0%)
		3. 3	2910 (45.2%)	IIIIIIIII		
		4. 4	1448 (22.5%)	IIII		
		5. 5	221 (3.4%)			
3	age	Mean (sd) : 35.3 (22.1)	79 distinct values		6434	0
	[numeric]	min < med < max:		:	(100%)	(0%)
		5 < 29 < 85		. :		
		IQR (CV) : 36 (0.6)		: :		
				: : : : : : . .		
				: : : : : : : : : :		
4	bmi	Mean (sd) : 26 (6.5)	2266 distinct		6434	0
	[numeric]	min < med < max:	values	. :	(100%)	(0%)
		13.4 < 25.3 < 64.2		: :		
		IQR (CV) : 8.2 (0.2)		: : :		
				. : : :		
				: : : : : .		

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
5	sbp [numeric]	Mean (sd) : 119.5 (20.1) min < med < max: 74 < 116 < 228 IQR (CV) : 22 (0.2)	73 distinct values	: : : : . : : : : : : : .	6434 (100%)	0 (0%)
6	hdl [numeric]	Mean (sd) : 51.6 (14.5) min < med < max: 19 < 49 < 160 IQR (CV) : 17 (0.3)	102 distinct values	: . : : : : : : : : : : : .	6434 (100%)	0 (0%)

Multiple linear regression: a small experiment

```
fit1 <- lm(hdl ~ bmi + age + gender + race + sbp,
           data = dat)

fit2 <- lm(bmi ~ age + gender + race + sbp,
           data = dat)

r2 <- fit2$residuals

fit3 <- lm(hdl ~ r2,
           data = dat)

coef(fit1)["bmi"]
```

```
   bmi
```

-0.6649902

```
coef(fit3)["r2"]
```

```
    r2
```

-0.6649902

Prediction interval vs. confidence interval

```
newdata <- dat[1,]
predict(fit1, newdata, interval = "confidence")
```

```
   fit      lwr      upr
```

```
1 44.48379 43.83743 45.13016
```

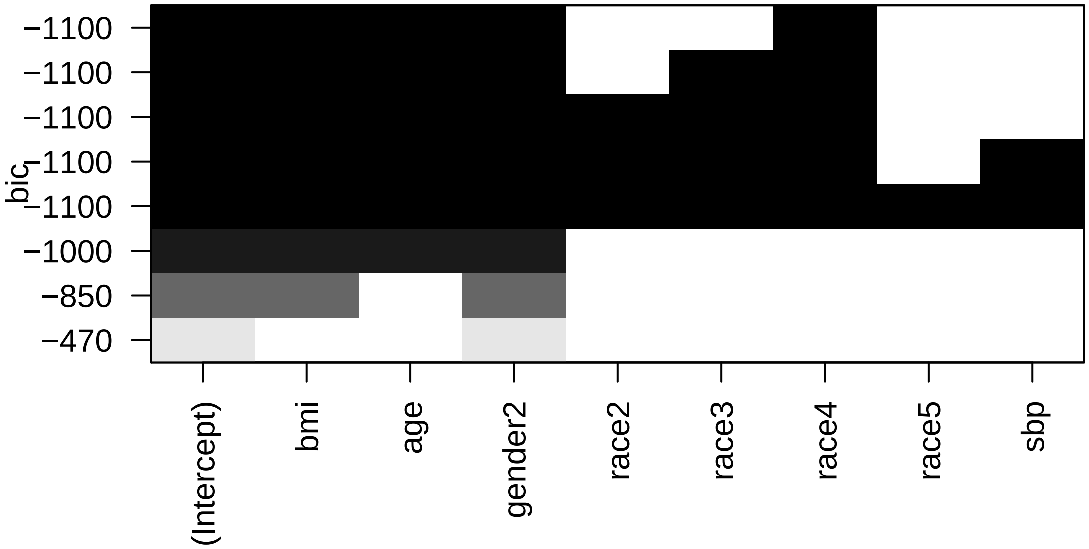
```r
predict(fit1, newdata, interval = "predict") # much wider!
```

```
   fit       lwr       upr
1 44.48379 18.50864 70.45895
```

Best subset model selection

```r
regsubsetsObj <- regsubsets(hdl ~ bmi + age + gender + race + sbp, data = dat,
                            method = "exhaustive", nbest = 1)

plot(regsubsetsObj, scale = "bic")
```



```r
# summary(regsubsetsObj)
```