

# Linear Regression: A Revisit

Yifei Sun

## Contents

Data	2
Multiple linear regression: a small experiment	3
Prediction interval vs. confidence interval	4
Best subset model selection	4

```
library(RNHANES)
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`.

```
dat <- nhanes_load_data(file_name = "l13_B", year = "2001-2002")

dat = dat %>%
  left_join(nhanes_load_data("BMX_B", "2001-2002"), by="SEQN") %>%
  left_join(nhanes_load_data("BPX_B", "2001-2002"), by="SEQN") %>%
  left_join(nhanes_load_data("DEMO_B", "2001-2002"), by="SEQN")

dat = dat %>%
  select(SEQN, RIAGENDR, RIDRETH1, RIDAGEYR, BMXBMI, BPXSY1, LBDHDL) %>%
  mutate(RIAGENDR = as_factor(RIAGENDR), RIDRETH1 = as_factor(RIDRETH1))

colnames(dat) <- c("ID", "gender", "race", "age", "bmi", "sbp", "hdl")

dat <- na.omit(dat)
```

Summary statistics of the predictors and the response:

```
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 [factor]	1. 1	3108 (48.3%)	IIIIIIII	6434 (100%)	0 (0%)
		2. 2	3326 (51.7%)	IIIIIIII		
2	race [factor]	1. 1	1593 (24.8%)	III	6434 (100%)	0 (0%)
		2. 2	262 ( 4.1%)			
		3. 3	2910 (45.2%)	IIIIIIII		
		4. 4	1448 (22.5%)	III		
		5. 5	221 ( 3.4%)			



## Prediction interval vs. confidence interval

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

```
      fit      lwr      upr
```

```
1 44.48379 43.83743 45.13016
```

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

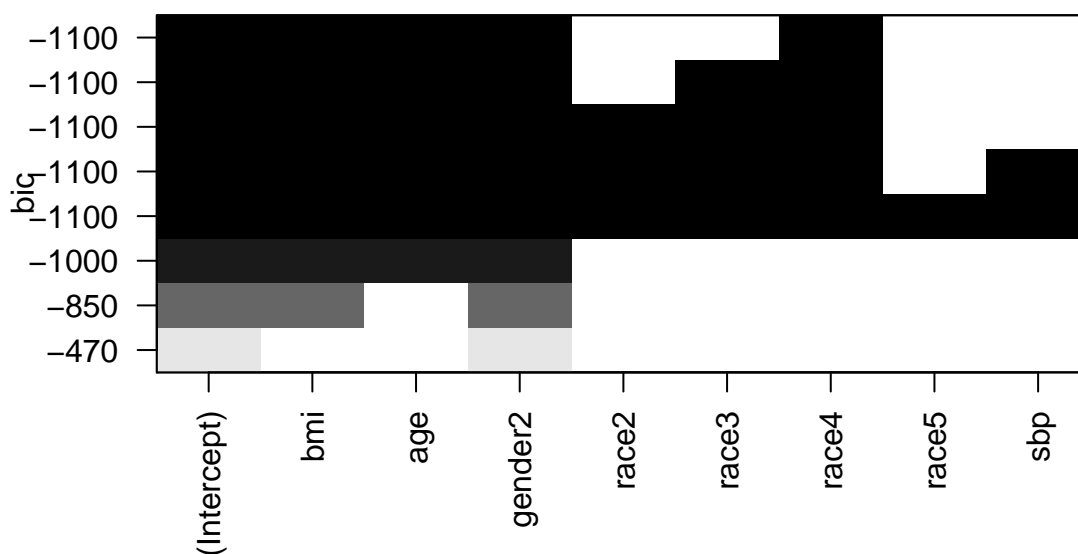
```
      fit      lwr      upr
```

```
1 44.48379 18.50864 70.45895
```

## Best subset model selection

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

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



```
# summary(regsubsetsObj)
```