Linear Regression

Yifei Sun

Contents

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).

```
library(RNHANES)
library(tidyverse)
library(summarytools)
library(stargazer)
library(leaps)
```

The data can be downloaded using functions in the package `RNHANES`.

```
dat <- nhanes_load_data("l13_B", "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)

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

dat$race <- as.factor(dat$race)

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

We first look at the 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 [numeric]	Min : 1 Mean : 1.5 Max : 2	1 : 3108 (48.3%) 2 : 3326 (51.7%)	IIIIIIII IIIIIIIII	6434 (100%)	0 (0%)
2	race [factor]	1. 1 2. 2 3. 3 4. 4 5. 5	1593 (24.8%) 262 (4.1%) 2910 (45.2%) 1448 (22.5%) 221 (3.4%)	IIII IIIIIIII IIII	6434 (100%)	0 (0%)

No	Variable	Stats / Values	Freqs (% of Valid)	Graph	Valid	Missing
3	age [numeric]	Mean (sd) : 35.3 (22.1) min < med < max: 5 < 29 < 85 IQR (CV) : 36 (0.6)	79 distinct values	: . : : : : : : : : : . . : : : : : : : : : :	6434 (100%)	0 (0%)
4	bmi [numeric]	Mean (sd) : 26 (6.5) min < med < max: 13.4 < 25.3 < 64.2 IQR (CV) : 8.2 (0.2)	2266 distinct values	. : : : : : : . : : : : : : : : .	6434 (100%)	0 (0%)
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%)

Simple linear regression

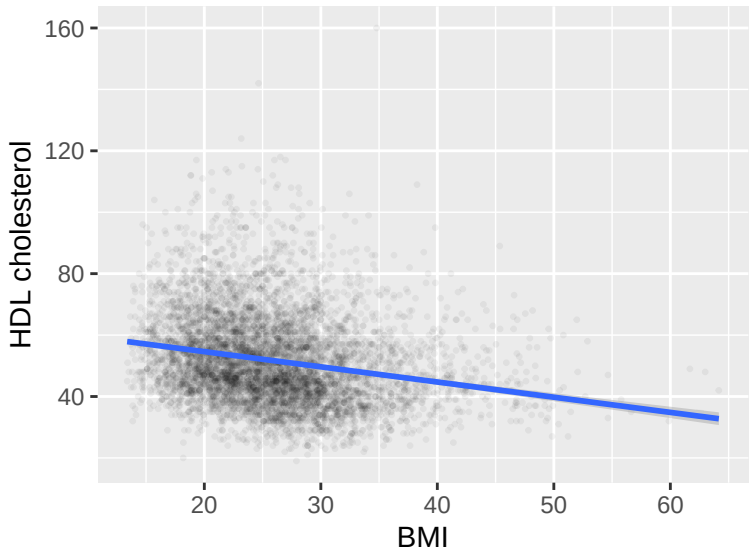
```r
fit0 <- lm(hdl ~ bmi,
           data = dat)

stargazer(fit0, header=FALSE, type='latex')
```

Table 2:

	Dependent variable:
	hdl
bmi	-0.495^{***} (0.027)
Constant	64.502^{***} (0.731)
Observations	6,434
R^2	0.049
Adjusted R^2	0.048
Residual Std. Error	14.153 (df = 6432)
F Statistic	328.634^{***} (df = 1; 6432)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
ggplot(dat, aes(bmi, hdl)) +
  geom_point(alpha = 0.05, size = 0.5, color = "black") +
  geom_smooth(method = "lm") +
  labs(x = "BMI", y = "HDL cholesterol")
```



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 = "predict")
```

```
   fit      lwr      upr
```

1 44.48379 18.50864 70.45895

```
predict(fit1, newdata, interval = "confidence")
```

```
   fit      lwr      upr
```

1 44.48379 43.83743 45.13016

Model selection

```
regsubsetsObj <- regsubsets(hdl ~ bmi + age + gender + race + sbp, data = dat,
                            method = "exhaustive")
plot(regsubsetsObj, scale = "adjr2")
```