Faraway

Harry Woo

2020-5-12

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>.

## Chapter 2 Estimation

## Problem 1

*The dataset teengamb concerns a study of teenage gambling in Britain. Fit a regression model with the expenditure on gambling as the response and the sex, status, income and verbal score as predictors. Present the output.*

data(teengamb)  
head(teengamb)

## sex status income verbal gamble  
## 1 1 51 2.00 8 0.0  
## 2 1 28 2.50 8 0.0  
## 3 1 37 2.00 6 0.0  
## 4 1 28 7.00 4 7.3  
## 5 1 65 2.00 8 19.6  
## 6 1 61 3.47 6 0.1

tg\_lm <- lm(gamble ~ sex + status + income + verbal, data = teengamb)  
tg\_lms <- summary(tg\_lm)  
print(tg\_lms)

##   
## Call:  
## lm(formula = gamble ~ sex + status + income + verbal, data = teengamb)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -51.082 -11.320 -1.451 9.452 94.252   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 22.55565 17.19680 1.312 0.1968   
## sex -22.11833 8.21111 -2.694 0.0101 \*   
## status 0.05223 0.28111 0.186 0.8535   
## income 4.96198 1.02539 4.839 1.79e-05 \*\*\*  
## verbal -2.95949 2.17215 -1.362 0.1803   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22.69 on 42 degrees of freedom  
## Multiple R-squared: 0.5267, Adjusted R-squared: 0.4816   
## F-statistic: 11.69 on 4 and 42 DF, p-value: 1.815e-06

### (a) What percentage of variation in the response is explained by these predictors?

전체 제곱합(SST)에서 회귀 제곱합(SSR)이 설명하는 비중, 즉 모형의 설명력은 결정 계수 R2 이다. 위 Summary 에서와 같이 동 모형의 결정계수 **Multiple R-squared = 0.5267** 이다.

var\_ex <- data.frame(Var\_explained = tg\_lms$r.squared)  
var\_ex %>% gt() %>%   
 fmt\_percent(columns = vars(Var\_explained),  
 decimals = 2)

Var\_explained

52.67%

### (b) Which observation has the largest (positive) residual? Give the case number.

회귀모형의 residuals 를 데이터프레임으로 변환하여 잔차값 기준으로 내림차순 정렬을 시행해 largest residual의 case number를 추출한 결과, **해당 case number는 24** 이다.

res <- data.frame(case\_no = c(1:47), residual = tg\_lm$residuals)  
res %>%   
 arrange(desc(residual)) %>%   
 slice(1) %>%   
 gt()

case\_no

residual

24

94.25222

### (c) Compute the mean and median of the residuals.

.

res %>%   
 summarise(mean = mean(residual), median = median(residual)) %>%   
 gt()

mean

median

-3.065293e-17

-1.451392

### (d) Compute the correlation of the residuals with the fitted values.

.

data.frame(correlation = cor(tg\_lm$residuals, tg\_lm$fitted.values)) %>%   
 gt()

correlation

-1.070659e-16

### (e) Compute the correlation of the residuals with the income.

.

data.frame(correlation = cor(teengamb$income, tg\_lm$fitted.values)) %>%   
 gt()

correlation

0.857142

### (f) For all other predictors held constant, what would be the difference in predicted expenditure on gambling for a male compared to a female?

.

data.frame(Gender\_coef = tg\_lm$coefficients["sex"]) %>%   
 gt()

Gender\_coef

-22.11833

## Chapter 3 Interference

## Problem 1

*For the prostate data, fit a model with lpsa as the response and the other variables as predictors.*

data(prostate)  
head(prostate)

## lcavol lweight age lbph svi lcp gleason pgg45 lpsa  
## 1 -0.5798185 2.7695 50 -1.386294 0 -1.38629 6 0 -0.43078  
## 2 -0.9942523 3.3196 58 -1.386294 0 -1.38629 6 0 -0.16252  
## 3 -0.5108256 2.6912 74 -1.386294 0 -1.38629 7 20 -0.16252  
## 4 -1.2039728 3.2828 58 -1.386294 0 -1.38629 6 0 -0.16252  
## 5 0.7514161 3.4324 62 -1.386294 0 -1.38629 6 0 0.37156  
## 6 -1.0498221 3.2288 50 -1.386294 0 -1.38629 6 0 0.76547

ps\_lm <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostate)  
ps\_lms <- summary(ps\_lm)  
ps\_lms

##   
## Call:  
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +   
## gleason + pgg45, data = prostate)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -1.7331 -0.3713 -0.0170 0.4141 1.6381   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.669337 1.296387 0.516 0.60693   
## lcavol 0.587022 0.087920 6.677 2.11e-09 \*\*\*  
## lweight 0.454467 0.170012 2.673 0.00896 \*\*   
## age -0.019637 0.011173 -1.758 0.08229 .   
## lbph 0.107054 0.058449 1.832 0.07040 .   
## svi 0.766157 0.244309 3.136 0.00233 \*\*   
## lcp -0.105474 0.091013 -1.159 0.24964   
## gleason 0.045142 0.157465 0.287 0.77503   
## pgg45 0.004525 0.004421 1.024 0.30886   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.7084 on 88 degrees of freedom  
## Multiple R-squared: 0.6548, Adjusted R-squared: 0.6234   
## F-statistic: 20.86 on 8 and 88 DF, p-value: < 2.2e-16

### (a) Compute 90 and 95% CIs for the parameter associated with age. Using just these intervals, what could we have deduced about the p-value for age in the regression summary?

/

confint(ps\_lm, parm = "age", level = 0.90)

## 5 % 95 %  
## age -0.0382102 -0.001064151

confint(ps\_lm, parm = "age", level = 0.95)

## 2.5 % 97.5 %  
## age -0.04184062 0.002566267

names(ps\_lms)

## [1] "call" "terms" "residuals" "coefficients"   
## [5] "aliased" "sigma" "df" "r.squared"   
## [9] "adj.r.squared" "fstatistic" "cov.unscaled"

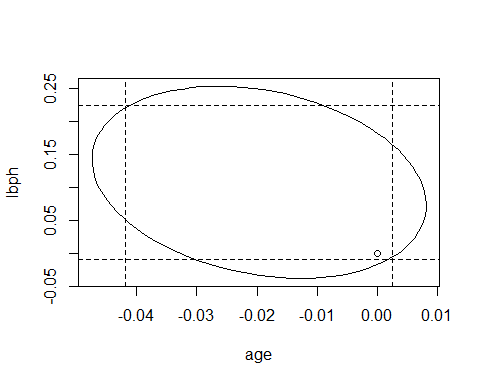
ps\_lms$coefficients["age", "Pr(>|t|)"]

## [1] 0.08229321

### (b) Compute and display a 95% joint confidence region for the parameters associated with age and lbph. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome.

/

library(ellipse)   
plot(ellipse(ps\_lm, c("age", "lbph")), type = "l")   
points(0, 0, pch = 1)   
abline(v = confint(ps\_lm)['age', ], lty = 2)   
abline(h = confint(ps\_lm)['lbph', ], lty = 2)



### (c) Suppose a new patient with the following values arrives:

`data.framelcavol lweight age Ibph svi lcp 1.44692 3.62301 65.00000 0.30010 0.00000 -0.79851 gleason pgg45 7.00000 15.00000

### Predict the lpsa for this patient along with an appropriate 95% CI.

/

new\_patient <- data.frame(  
 "lcavol" = 1.44692,  
 "lweight" = 3.62301,  
 "age" = 65.00000,  
 "lbph" = 0.30010,  
 "svi" = 0.00000,  
 "lcp" = -0.79851,  
 "gleason" = 7.00000,  
 "pgg45" = 15.00000  
)  
  
new\_patient

## lcavol lweight age lbph svi lcp gleason pgg45  
## 1 1.44692 3.62301 65 0.3001 0 -0.79851 7 15

predict(ps\_lm, newdata = new\_patient, interval = "prediction")

## fit lwr upr  
## 1 2.389053 0.9646584 3.813447

### (d) Repeat the last question for a patient with the same values except that he or she is age 20. Explain why the CI is wider.

/

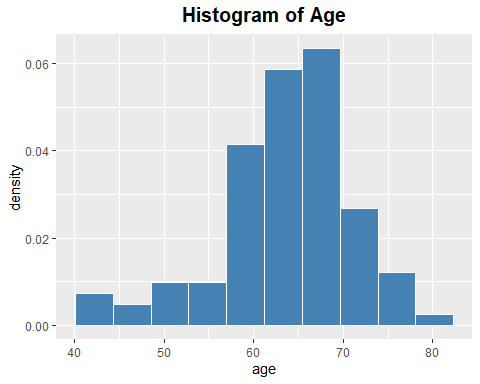
new\_patient2 <- new\_patient  
new\_patient2[3] = 20  
rbind(new\_patient, new\_patient2)

## lcavol lweight age lbph svi lcp gleason pgg45  
## 1 1.44692 3.62301 65 0.3001 0 -0.79851 7 15  
## 2 1.44692 3.62301 20 0.3001 0 -0.79851 7 15

predict(ps\_lm, newdata = new\_patient2, interval = "prediction")

## fit lwr upr  
## 1 3.272726 1.538744 5.006707

ggplot(data = prostate, aes(x = age, y = ..density..)) +  
 geom\_histogram(bins = 10, fill = "steelblue", colour = "white") +  
 ggtitle(label = "Histogram of Age") +  
 theme(plot.title = element\_text(size = 15, hjust = 0.5, vjust = 1.5, face = "bold"))



### (e) In the text, we made a permutation test corresponding to the F-test for the significance of all the predictors. Execute the permutation test corresponding to the t-test for age in this model. (Hint: {summary (g) $coef [4,3] gets you the t-statistic you need if the model is called g.)

/

t\_value <- summary(ps\_lm) %>%   
 coef() %>%   
 .["age", "t value"]   
  
permute\_tmod <- function(nsims) {  
 map\_dbl(1:nsims,  
 ~ lm(sample(lpsa) ~ ., data = prostate) %>%  
 summary() %>%  
 coef() %>%  
 .["age", "t value"])   
}   
   
mean(abs(permute\_tmod(100)) > abs(t\_value))

## [1] 0.07

mean(abs(permute\_tmod(1000)) > abs(t\_value))

## [1] 0.083

mean(abs(permute\_tmod(10000)) > abs(t\_value))

## [1] 0.0823