

다루니

제출기한: 6월 19일

Problem 1

첨부된 geyser.txt는 미국 옐로우스톤 국립공원의 간헐천인 'Old Faithful' 간헐천의 분출시간 (단위 분)과 분출 간격 (단위 분)을 기록한 자료이다. 자료는 2개의 변수에 대한 299개의 관측치로 이루어져있다. (R code: Faraway-PRA pp. 14-15, pp. 62-63 참고)

참고문헌

Azzalini, A. and Bowman, A. W. (1990) A look at some data on the Old Faithful geyser. Applied Statistics 39, 357-365.

X Y ?

- 자료의 산점도와 분출간격을 분출시간에 선형회귀모형을 최소제곱법으로 적합한 회귀선을 그리시오. 산점도와 회귀 결과의 특이점을 설명하시오.
- 구해진 잔차의 제곱과 분출시간의 산점도를 그리고 발견한 현상을 기술하시오.
- B에서 관찰한 내용에 따라 A의 적합의 잠재적 문제가 무엇인지 설명하고 이를 해결하기위한 방법으로 가중최소제곱법을 적용해보시오.
- A와 C의 결과를 비교하여 설명하시오.

Problem 2

첨부된 data2.txt는 첫 열이 반응변수이고 나머지 10개의 열이 설명변수로 구성된 200개의 관측치를 포함하는 자료이다. (R code: Faraway-PRA pp. 124-133, pp. 32-33 참고)

- 부분 F검정법에 기반한 후진제거 방법을 이용하여 적합한 모형을 구하고 결과를 설명하시오.
- 수정결정계수에 기반한 전진선택 방법을 이용하여 적합한 모형을 구하고 결과를 설명하시오.
- AIC에 기반한 단계적 회귀적합 방법을 이용하여 적합한 모형을 구하고 결과를 설명하시오.

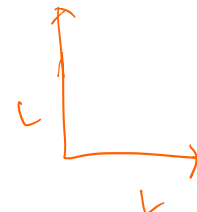
Problem 3

첨부된 calif.txt는 1990년 조사된 미국 캘리포니아주의 20,640 구역의 주택가격의 중앙값, 소득의 중앙값, 주택연령의 중앙값, 총 방의 수, 총 침실의 수, 인구, 가구수, 위도, 경도를 기록한 자료이다.

- 각 변수들의 특징을 요약하시오. 이중 특이한 관측치가 있으면 그 관측치를 보고하고 이유를 설명하시오.
- 아래에 제공되는 R code를 이용하여 주택가격의 중위수를 캘리포니아 지도 위에 표현하시오. 생성된 그림을 간단히 설명하시오.

```
plot(calif$longitude, calif$latitude, pch=21,
     col=heat.colors(11)[11-floor(calif$value/50000)],
     bg=heat.colors(11)[11-floor(calif$value/50000)],
```

value .



인데...
value.
어떻게
필요?

```
cex=sqrt(calif$population/median(calif$population)),  
xlab="Longitude",ylab="Latitude",main="Median House Prices")  
legend(x="topright",legend=(50*(11:1)),fill=heat.colors(11))
```

어디에!

- C. 위도, 경도를 제외한 모든 변수를 이용하여 주택가격의 분위수를 설명하는 선형회귀모형을 적합 하고, 그 결과를 설명하시오.
- D. 위에서 적합된 결과에서 구한 잔차를 지도위에 표시하는 그림을 그리고 그 결과를 설명하시오.
- E. 반응변수를 로그 변환한 자료로 C와 D를 반복하고 그 결과를 비교하시오.

이런 거!