wk04-2-REG-df2015na Homework

회귀모형 과제

문제 설명

- 목적: sizekorea 자료를 이용하여 키(ht) 예측모형 개발
- Mlm: 모든 개인정보를 이용한 선형회귀모형
- Mstep: Mlm으로부터 변수 선택한 선형회귀모형
- 과제 요령
  - Rmd를 html로 만든 뒤 pdf로 출력한 것을 이클래스에 제출하면 됩니다.
  - 질문은 질의응답에 올리고 아는 사람은 대답해주세요. 나는 가급적 대답 안 할 생각입니다.
- 과제에 포함해야 되는 내용
  - 제일 앞에 아래표를 포함시킬 것 (Rmd에서 표만드는 방법은 검색해보거나 Help > Markdown Quick Reference 보세요)
  - 본인 포함하여 4인에 대해 예측키(yh), 잔차(res)를 제시할 것(자료 진위는 관심없음)
  - 나머지는 주어진 순서대로 R로 처리

모형	키예측식
Mlm	150+0.3wt+0.2wa+2.4alcY …
Mstep	155+0.2wt+0.1wa+2.4alcY …

모형	RSQ	MSE	MAE	AIC
Mlm				
Mstep				

모형	gnd	age	wt	wa	hdln	hdwd	ftln	ftwd	bld	lft	smk	alc	yhlm	reslm	yhstep	resstep
본인																
친구1																
친구2																
친구3																

자료설명

- Size Korea 2015년 인체계측자료 일부: n=300 (완전한 관측값 269개), d=13

변수	설명
gnd	성별{F, M}. 이진 판별분석시 타겟
age	나이
ht	키 (cm). 회귀분석시 타겟
wt	몸무게 (kg)
wa	허리둘레(cm)

변수	설명
hdln	손길이(cm)
hdwd	손너비(cm)
ftln	발길이(cm)
ftwd	발너비(cm)
bld	혈액형{A,AB,B,O}
lft	왼손잡이용 가변수 (1,0). 1이면 왼손잡이, 0이면 비왼손잡이임
smk	흡연 여부 (1,0). 1이면 흡연, 0이면 비흡연
alc	음주 여부 (1,0). 1이면 음주, 0이면 비음주

읽기

- df2015na.csv를 DF로 읽기
- df2015na-sc.csv(본인 포함 4인의 자료가 들어 있는 파일)를 만들고 SC로 읽기: df2015na.csv를 보고 만드세요

변수 조정

- 강의자료 참고(mutate예를 볼 것)
- 문자변수(gnd, bld)를 factor로 변환
- {0,1}로 코딩된 이산형 변수(lft, smk, alc)를 factor로 변환(수준을 {N, Y}로 변경할 것)
    - 예: smk = factor(smk, labels=c('N','Y'))

결측

- skimr::skim(data, …)로 결측탐색
    - 변수별 결측비율
    - 완전한 관측값 비율

간단 탐색

- caret::featurePlot(x:연속형, y:요인, plot='strip,scatter,box,density')으로 변수간 관계를 탐색할 것
- 연속 ~ 연속: 연속형 변수간 상관계수 분석
    - ht와 상관관계가 높은 순으로 변수를 정리
    - corrplot.mixed로 상관계수를 시각화하시오
- 이산 ~ 이산
    - 성별 흡연비율을 시각화 (3주차 강의자료 참고)
    - 성별 음주비율을 시각화
    - 성별 왼손잡이비율을 시각화

회귀분석

- SC: 본인포함 4인의 자료가 들어 있는 자료. csv로 만들세요

선형회귀분석

- ht를 모든 다른 변수를 이용하여 예측하는 회귀모형을 적합하시오(Mlm <- lm(ht~., data=DF))
- Mlm에 대해
    - ht의 예측식을 쓰시오(제일 앞부분 표완성)

    - Mlm에 대한 잔차를 시각화하시오(plot(Mlm))
    - DF에 대해 M0의 성능을 Rsq, MSE, MAE로 평가하시오(제일 앞부분 표완성)
    - Mlm으로 SC를 예측하시오(제일 앞부분 표완성)

변수선택 회귀분석

- stepAIC를 이용하여 M0로부터 변수선택한 모형 M1을 적합하시오(Mstep <- stepAIC(Mlm))
- Mstep에 대해
    - ht의 예측식을 쓰시오 (제일 앞부분 표완성)
    - Mstep에 대한 잔차를 시각화하시오(plot(Mstep))
    - DF에 대해 Mstep의 성능을 Rsq, MSE, MAE로 평가하시오(제일 앞부분 표완성. AIC확인할 것)
    - Mstep으로 SC를 예측하시오(제일 앞부분 표완성)