2022년 2학기 경영데이터분석2 중간고사 문제&답안(통합)

[데이터와 변수 소개]

본 데이터는 은행의 고객에 대한 대출 관련 정보를 담고 있음

- id: 고객 식별용 번호
- credit: 대출금규모(단위: 만원)
- age: 고객 연령
- asset: 고객 자산규모(단위: 만원)
- income: 고객 월수입(단위: 만원)
- duration: 대출기간(단위: 월)
- purpose: 대출유형(house: 주택담보대출, credit: 신용대출, working: 운전자금대출)
- gender: 고객 성별(male과 female 두 가지로 구분)

※ credit.csv 파일을 불러와 데이터 프레임을 만들고, 다음 질문에 답하시오. 모든 경우 유의수준은 0.05로 하시오.

1. gender와 purpose 두 변수를 동시에 고려하여 전체 사례를 여섯 개 집단으로 구분했을 때, 변수 credit 평균이 가장 높은 집단의 gender는 (male)이고, purpose는 (house)이며, credit 평균값은 (5158)만원이다. 괄호 안을 채우시오(각 1점).

2. purpose 변수의 척도를 범주형으로 바꾸는 동시에 출력순서도 working, credit, house로 변경한 후, purpose 변수의 범주(측정값)별 빈도를 순서대로 쓰시오(각 1점).
working:269/credit:343/house:388

3. 다음 조건을 만족하는 새로운 변수 group을 만든 후, 네 가지 측정값(범주) 중에서 가장 빈도수가 낮은 범주와 해당 범주의 credit 변수 측정값 평균을 구하시오(각 1점).
B:160/B:1628

조건	group 측정값
asset ≤ 20,000	C
20,000 < asset ≤ 40,000	B
40,000 < asset ≤ 60,000	A
60,000 < asset	S

4. id 변수를 제외한 다섯 개의 계량형 척도로 측정된 변수들 간의 상관관계를 토대로 DV인 credit과 나머지 IV 간에 아래와 같은 대립가설을 수립하고자 한다. 아래 괄호 안에 들어갈 부호(+ 혹은 -)를 순서대로 쓰시오(각 0.5점).
H1: credit와 age 사이에는 (+)의 인과관계가 존재한다.
H2: credit와 asset 사이에는 (+)의 인과관계가 존재한다.

H3: credit와 income 사이에는 (+)의 인과관계가 존재한다.
H4: credit와 duration 사이에는 (+)의 인과관계가 존재한다.

※ 4번 문제에서 수립한 (대립)가설을 검정하기 위한 다중선형회귀식(lm1)을 수립하고, DV의 선형성과 등분산성 및 정규성 전제조건을 개선하기 위해 plot 함수를 통해 확인된 **다섯 개 사례(id = 379, 638, 888, 916, 918)를 모두 제거하여 credit 데이터프레임과 lm1을 모두 업데이트**한 후 이하의 질문에 답하시오.

5. DV인 credit 변수의 정규성 조건을 Shapiro test를 통해 확인한 결과 W는 (0.80797)로 나왔고, 결과적으로 정규성 조건을 만족하지 못한다(1점).

6. DW통계량과 2*p*-value를 근거로 오차의 자기상관 검토 결과를 간단히 설명하시오(2점).
DW통계량이 1.993598로 2에 가깝고 2p-value가 0.888로 0.05보다 크게 나왔기에 유의하지않다.(다만 rho!=0으로 나옴) 즉 상관관계가 없고 독립성이다.

7. 4번 문제에서 수립한 4개의 대립가설 중에서 채택된 대립가설은 무엇인가?(2점)
H1: credit와 age 사이에는 (+)의 인과관계가 존재한다. /H3: credit와 income 사이에는 (+)의 인과관계가 존재한다. /H4: credit와 duration 사이에는 (+)의 인과관계가 존재한다.

8. lm1의 모형적합도를 높이기 위해 세 가지 방식으로 추정한 결과 AIC값의 크기는 다음과 같다. 괄호 안에 적절한 등호(=)와 부등호(>, <)를 넣으시오(각 1점).
forward (>) backward (=) both

9. vif 함수를 통해 확인한 결과 네 개 독립변수의 VIF값은 age (3.428), asset (2.949), income (2.183), duration (3.100)으로 모두 5.3보다 작아서 다중공선성은 걱정할 필요가 없다. 괄호안에 VIF값을 **소수 넷째자리에서 반올림**하여 쓰시오(각 0.5점).

10. lm1에 새로운 IV인 purpose를 추가한 다중선형회귀식 lm2를 만든 후, purpose를 추가하는 것이 타당한지 확인하려고 한다. 결과적으로 adjusted R^2는 lm1에서는 (0.6234)이었는데 lm2에서 (0.6264)로 증가하였다. 또한 R^2변화량은 (0.003)인데, 이 변화량에 대한 통계적 유의성을 검토하니 *p*-value가 (0.0189)으로 나와 유의수준 보다 작으므로 새로운 변수인 purpose를 추가하는 것이 모형 설명력을 높일 수 있다. 괄호안에 수치를 **소수 다섯째자리에서 반올림**하여 쓰시오(각 0.5점).

11. lm2를 추정한 결과, purpose라는 새로운 IV가 DV에 미치는 영향에 대한 다음

설명 중 틀린 것은 무엇인가? (3) (1점)

① purpose 변수와 관련된 더미변수는 모두 두 개다.

② reference는 working(운전자금 대출)이다.

③ reference와 비교했을 때, purpose가 credit(신용자금 대출)인 경우 DV가 증가한다.

④ reference와 비교했을 때, purpose가 house(주택담보 대출)인 경우 DV가 증가한다.

12. lm2에서 DV에 대한 설명력이 가장 높은, 즉 가장 중요한 IV는 무엇인가?(2점). income(0.41261579)

13. lm2에 새로운 IV인 gender를 추가한 다중선형회귀식 lm3를 만든 후, gender를 추가하는 것이 타당한지 확인하려고 한다. 결과적으로 adjusted R^2는 lm2에서는 (0.6264)이었는데 lm3에서 (0.6284)로 증가하였다. 또한 R^2변화량은 (0.002)인데, 이 변화량에 대한 통계적 유의성을 검토하니 *p*-value가 (0.0209)로 나와 유의수준 보다 작으므로 새로운 변수인 gender를 추가하는 것이 모형설명력을 높일 수 있다. 괄호안에 수치를 **소수 다섯째자리에서 반올림**하여 쓰시오(각 0.5점).

14. lm3에서 gender 변수는 DV에 어떤 영향을 미치는지 간단하게 1줄로 서술하시오(2점). gender변수에서 reference는 female인데 female인 경우에 비해 male일 때 251.7 만큼 대출금규모가 증가한다. 또한 2p-value가 alpha인 0.05보다 작아 유의한데, 종속변수인 DV(credit)에 양의 인과관계를 미친다.

15. credit 데이터프레임을 남성(male)과 여성(female)으로 구분한 서브데이터프레임 credit_male과 credit_female을 만들고, 서브데이터프레임별로 credit을 DV로 하고 age, asset, income, duration, purpose를 IV로 하는 다중선형회귀식(lm_male과 lm_female)을 추정하시오. 다음 설명 중 틀린 것은 무엇인가(2점)? (4)

① 모형설명력은 lm_male(credit_male 기반) 보다 lm_female(credit_female 기반)이 더 좋다.

② 두 회귀식에서 공통적으로 age가 증가할수록 DV인 credit도 증가한다.

③ lm_female에서 purpose는 DV인 credit에 통계적으로 유의한 영향을 미치지 못한다.

④ 두 회귀식에서 공통적으로 asset은 DV인 credit에 통계적으로 유의한 영향을 미치지 못한다.

(끝)