DACA

NGO 데이터 분석 : 2주차

2020.10.28 ~ 2020.11.4

빅데이터 응용학과 석사 1기 양선욱

목차

1. 신뢰성분석 – 1. 내적일관성분석

사용 데이터

PAY_RATE_NOPAY : 미납율

PAY_NUM : 납입후원횟수

PLED_FIRST_LONGEVITY : 첫 플릿지 기준 고객 기간

PAY_SUM_PAYMENTAMOUNT : 납입총후원금액

분석 목적

- 우대 후원자를 선별하기 위해 책임감, 지속성, 후원력, 납부력을 측정하는 측정항목을 개발하여 고객을 조사하였다.
- 4가지 항목으로 구성된 우대 후원자 평가가 신뢰성 있는 조사인지 확인한다.
- 우대 후원자는 전체의 10% 정도로 선정한다.

분석 과정

1) 결측값 제거

2) 측정항목 생성

3) 내적일관성분석 수행

측정항목

측정항목	분류기준	측정단위
책임감	미납율	1 ~ 7
지속성	월후원율	1 ~ 7
후원력	총후원횟수	1 ~ 7
납부력	총후원금액	1 ~ 7

핵심코드

```
X = df_1[['책임감','지속성','후원력','납부력']]
X.to_csv('12.df_1.csv', encoding='utf-8-sig')

#전체의 크론바흐 알파 계수 출력
df_X = pd.read_csv('12.df_1.csv', sep=',', encoding='utf-8-sig')
X1 = df_X[['지속성','후원력','납부력']]
X2 = df_X[['책임감','후원력','납부력']]
X3 = df_X[['책임감','지속성','납부력']]
X4 = df_X[['책임감','지속성','후원력']]
#print(X)

CA_all = pg.cronbach_alpha(data=df_X)
CA_X1 = pg.cronbach_alpha(data=X1)
CA_X2 = pg.cronbach_alpha(data=X2)
CA_X3 = pg.cronbach_alpha(data=X3)
CA_X4 = pg.cronbach_alpha(data=X4)
```

1. 신뢰성분석 – 1. 내적일관성분석

결과

				알파 계수
책임감	지속성	후원력	납부력	-0.005
	지속성	후원력	납부력	0.141
책임감		후원력	납부력	0.793
책임감	지속성		납부력	0.198
책임감	지속성	후원력		0.212

책임감 + 후원력 + 납부력 = 총합

총합의 평균 : 15.5

총합의 중간값 : 16.0

총합의 상위 10% : 19.0

결과 해석

- 측정 항목 중에서 책임감, 후원력, 납부력으로 구성된 측정 항목이 크론바흐 알파 계수가 0.793으로 유의미하게 나왔다.
- 책임감, 후원력, 납부력의 총합의 평균은 15.5이며, 총합의 상위 10% 값은 19.0이다.

실무적 인사이트

- 우대 고객 판별을 위한 측정 항목으로 지속성을 제외하고 책임감, 후원력, 납부력을 채택한다.
- 고객 중 10%만 우대고객으로 지정하기위해 총합의 기준을 19로 잡는다.
- 총합의 중간값 16과 상위 10%의 값 19의 차이가 미비하여 변별력이 없다고 판단된다. 측정 척도를 수정할 필요가 있다.

1. 신뢰성분석 – 1. 내적일관성분석

총 인원 : 977

측정 척도

미납율	0.9 ~ 1	0.8 ~ 0.9	0.6 ~ 0.8	0.4 ~ 0.6	0.2 ~ 0.4	0.1 ~ 0.2	0 ~ 0.1
책임감	1	2	3	4	5	6	7
인원수	4	3	29	23	24	37	851
비율	0.004	0.003	0.029	0.023	0.024	0.037	0.871

월후원율	0 ~ 0.1	0.1 ~ 0.2	0.2 ~ 0.4	0.4 ~ 0.6	0.6 ~ 0.8	0.8 ~ 0.9	0.9 ~ 1
지속성	1	2	3	4	5	6	7
인원수	228	283	219	80	38	24	105
비율	0.233	0.289	0.224	0.081	0.038	0.024	0.107

납입횟수	0~2	3~5	6~8	9~11	12~14	15~17	18~
후원력	1	2	3	4	5	6	7
인원수	36	64	50	102	624	8	93
비율	0.036	0.065	0.051	0.104	0.638	0.008	0.095

총후원금액 *0.01	0~500	500~1000	1000~2000	2000~3000	3000~4000	4000~5000	5000~
후원력	1	2	3	4	5	6	7
인원수	35	59	192	192	392	23	84
비율	0.035	0.060	0.196	0.196	0.401	0.023	0.085

2. 분산분석 – 1. 일원분산분석

사용 데이터

내적일관성분석에서 나온 책임감, 후원력, 납입력, 총합

LONGEVITY_D : 가입일수

PAY_SUM_PAYMENTAMOUNT : 총납입금액

분석 목적

- 총합을 기준으로 고객 등급을 '우대','Gold', 'Silver', 'Bronze', '새싹'으로 나누었다.
- 나누어진 고객 등급간 가입일수의 평균이 차이가 있는지 알아본다.
- 나누어진 고객 등급간 총납입금액의 평균이 차이가 있는지 알아본다.

분석 과정

1) 결측값 제거
2) 파생변수 생성
3) 등분산 검정
4) 일원분산분석 수행

측정항목

고객등급	총합점수 기준	고객 비율 기준
우대	19점 이상	10%
골드	17점 이상	20%
실버	15점 이상	30%
브론즈	11점 이상	30%
새싹	0점 이상	10%

결과

등분산검정

독립변수	종속변수	F값	p-value
고객등급	납입총금액	39.7	7.78e-31
	가입일수	61.4	2.56e-46

일원분산분석

독립변수	종속변수	F값	p-value
고객등급	납입총금액	456	0.000
	가입일수	28.5	0.000

2. 분산분석 – 1. 일원분산분석

결과

고객등급x총납입금액

	우대	골드	실버	브론즈	새싹
우대	1.0				
골드	0.00	1.0			
실버	0.00	0.00	1.0		
브론즈	0.00	0.00	0.573	1.0	
새싹	0.00	0.00	0.00	0.000	1.0

고객등급x가입일수

	우대	골드	실버	브론즈	새싹
우대	1.0				
골드	0.15	1.0			
실버	0.71	0.66	1.0		
브론즈	0.000	0.058	0.002	1.0	
새싹	0.00	0.14	0.012	0.14	1.0

결과 해석

- 등분산 검정으로 부터 고객등급x총납입금액, 고객등급x가입일수 모두 고객등급 집단의 분산이 동일하지 않다.
- 일원산분산분석 결과로 부터 1개의 고객등급은 다른 고객등급과 총납입금액, 가입일의 차이가 있다.

실무적 인사이트

- 브론즈와 실버 사이를 제외하고 우대-골드, 골드-실버, 브론즈-새싹 구간에 총납입금액 평균의 차이가 있다.
- 총납입금액에 따른 고객등급의 분할이 유의미 하다.
- 브론즈와 실버 사이를 제외하고 우대-골드, 골드-실버, 브론즈-새싹 구간에 가입일수 평균의 차이가 없다.
- 가입일수에 상관없이 고객의 활동 수치에 따라 고객등급이 매겨진다.
- 실버-브론즈 구간에 대한 조사가 필요하다.

2. 분산분석 – 2. 이원분산분석

사용 데이터

일원분산분석에서 나온 고객등급

MOTI_CHANNEL : 주요채널

PAY_SUM_PAYMENTAMOUNT : 총납입금액

분석 목적

- 고객등급과 주요채널에 따른 총납입금액 평균의 차이가 있는지 알아보고 싶다.
- 고객등급이 주요채널와도 연관이 있을 것이라 추론되어 이원분산분석을 통해 알아본다.

분석 과정

1) 결측값 제거
2) 파생변수 생성
3) 등분산 검정
4) 이원분산분석 수행
5) 사후분석

파생변수

- 주요채널의 결측값과 UNKNOWN은 제거한다.

ex BRODCASE : 1, DIGITAL : 2

결과

등분산검정

변수	변수	F값	p-value
고객등급	총납입금액	39.4	1.34e-30
주요채널		2.95	0.007

이원분산분석

변수	변수	F값	p-value
고객등급	총납입금액	81.4	0.000
주요채널		70.8	0.000
고객등급 x 주요채널		33.8	0.000

2. 분산분석 – 2. 이원분산분석

결과

고객등급x총납입금액

	우대	골드	실버	브론즈	새싹
우대	1				
골드	0.00	1			
실버	0.00	0.00	1		
브론즈	0.00	0.00	0.598	1	
새싹	0.00	0.00	0.00	0.00	1

주요채널x총납입금액

	BROADCAST	DIGITAL	DM	EVENT	GENERAL AD	RELATIONSH	TM
BROADCAST	1						
DIGITAL	0.192	1					
DM	0.999	0.999	1				
EVENT	0.743	0.061	0.997	1			
GENERAL AD	0.549	0.999	0.999	0.134	1		
RELATIONSH	0.035	0.999	0.999	0.023	1.000	1	
TM	0.976	0.643	0.997	1.000	0.740	0.542	1

결과 해석

- 등분산 검정으로 부터 고객등급x총납입금액, 주요채널x가입일수 모두 고객등급 집단의 분산이 동일하지 않다.
- 이원분산분석 결과로 부터 고객등급(주요채널)은 적어도 1개의 항목이 다른 고객등급(주요채널)과 총납입금액 차이가 있다.
- 고객등급과 주요채널은 상호작용 효과가 있다.

실무적 인사이트

- 총납입금액은 고객등급에 따라, 주요채널에 따라 달라진다.
- 주요채널에선 BROADCAST와 RELATIONSH, RELATIONSH와 EVENT의 그룹간 총납입금액의 차이가 유의하게 나타난다.
- 해당 그룹간 차이를 자세히 보기 위해 고객등급, 주요채널, 총납입금액의 피벗 테이블이 필요하다.

2. 분산분석 – 2. 이원분산분석

주요채널	BROADCAST	DIGITAL	DM	EVENT	GENERAL AD	RELATIONSH	TM
고객등급							
우대	660000	717647	NAN	360000	751375	861343	440000
골드	359489	362705	360000	366428	351724	375804	345000
실버	222173	199814	120000	144761	221428	178823	175000
브론즈	183469	169736	NAN	140000	210833	173285	60000
새싹	70909	85000	320000	65000	85000	71538	18333

실무적 인사이트

- BROADCAST와 RELATIONS,
- RELATIONSH와 EVENT
- 두 가지 주요채널 관계 모두 우대 고객등급에서 가장 큰 차이를 보이고 있다.

- 우대 고객에게는 DM, EVENT, TM의 채널방식이 높은 총납입금액을 유도하지 못한다.
- 새싹 단계의 회원 중에 DM을 주요채널로 사용하는 후원자는 총납입금액이 높다. 그럼에도 새싹단계에 머물러 있는 원인파악이 필요하다.

2. 분산분석 – 3. 다변량분산분석

사용 데이터

-고객등급 -MOTI_CHANNEL : 주요채널

- PAY_SUM_PAYMENTAMOUNT : 총납입금액

- PAY_NUM : 납입후원횟수

분석 목적

-이원분산분석을 통해 고객등급과 주요채널에 따른 총납입금액의 차이가 있었다.

- 사전에 총납입금액과 납입후원횟수 사이에 유의한 상관관계가 존재한다고 밝혀냈다.

- 납입후원횟수를 추가하여 다변량분산분석을 수행한다.

분석 과정

1) 결측값 제거
2) 등분산 검정
3) 다변량분산분석
4) 사후분석
5) 산점도

결과

등분산검정

변수	변수	F값	p-value
고객등급	납입금액	39.1	2.24e-30
고객등급	납입횟수	56.5	7.92e-43
주요채널	납입금액	2.95	0.007
주요채널	납입횟수	2.95	0.007

다변량분산분석

	F값	p-value
전체모형	846	0.000
고객등급	384	0.000
주요채널	0.77	0.461

2. 분산분석 – 3. 다변량분산분석

사후분석

1) 고객등급x납입횟수

	우대	골드	실버	브론즈	새싹
우대	1				
골드	0.00	1			
실버	0.00	0.827	1		
브론즈	0.00	0.00	0.00	1	
새싹	0.00	0.00	0.00	0.00	1

3) 고객등급x주요채널의 평균납입횟수

주요채널	BROADCAST	DIGITAL	DM	EVENT	GENERAL AD	RELATIONSH	TM
고객등급							
우대	23.8	23.9	nan	24.0	33.1	28.6	20.0
골드	12.4	12.8	12.0	12.0	12.4	12.4	17.0
실버	11.9	12.1	12.0	12.0	11.9	12.0	12.0
브론즈	8.3	8.5	nan	9.6	10.3	8.6	12.0
새싹	3.4	3.4	8.0	3.6	3.5	3.0	3.5

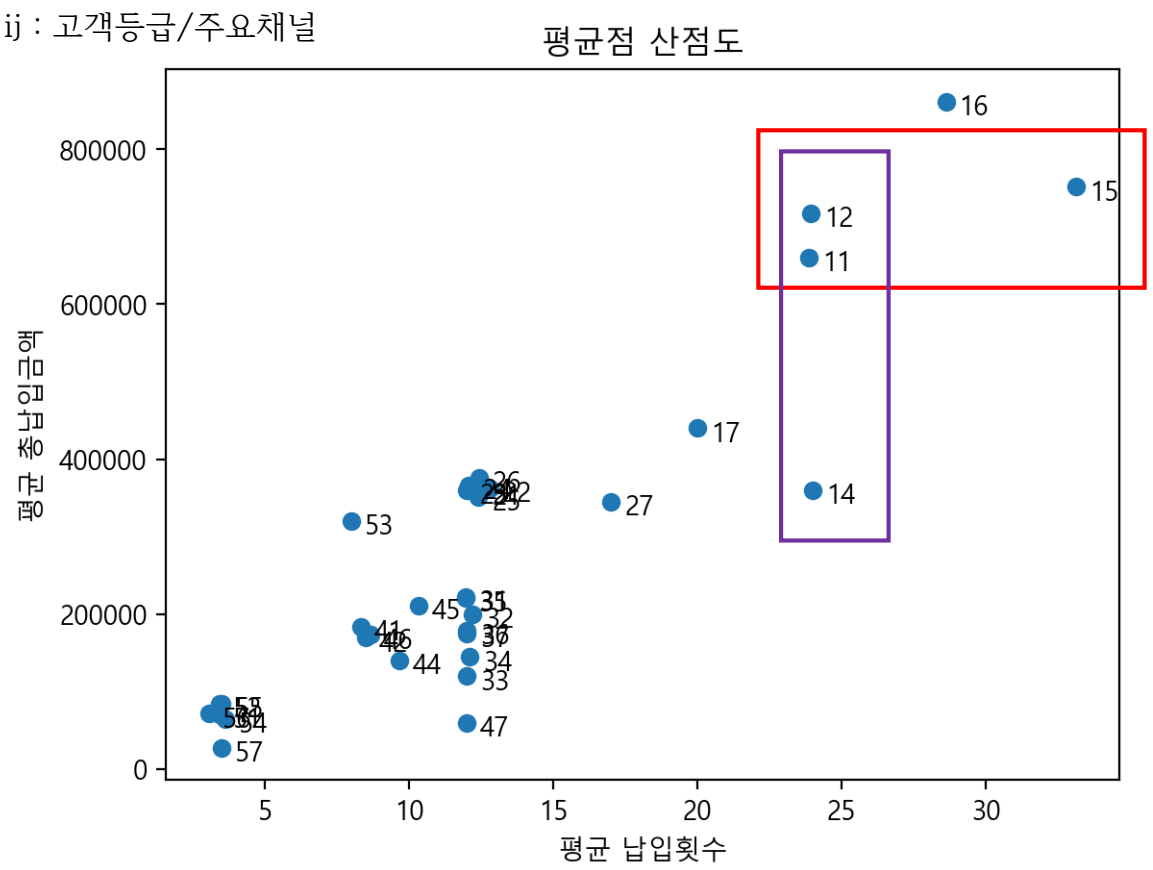
2) 주요채널x납입횟수

	BROADCAST	DIGITAL	DM	EVENT	GENERAL AD	RELATIONSH	TM
BROADCAST	1						
DIGITAL	0.259	1					
DM	1	0.99	1				
EVENT	0.99	0.649	1	1			
GENERAL AD	0.144	0.953	0.99	0.320	1		
RELATIONSH	0.173	1	0.99	0.617	0.95	1	
TM	0.986	1	0.99	0.971	0.99	1	1

2. 분산분석 – 3. 다변량분산분석

사후분석

ij : 고객등급/주요채널



결과 해석

- 다변량분산분석 모형과 고객등급별 총납입금액과 납입횟수의 차이가 유의하다.
- 주요채널별 총납입금액과 납입횟수에 유의한 차이가 없다.
- 이전의 결과와 연동하여 주요채널별 총납입금액은 차이가 있지만 납입횟수는 차이가 없다.

실무적 인사이트

- 우대등급의 값이 커서 이를 배제한 분석이 필요하다.
- GENERAL AD는 우대 등급에서만 더 높은 납입횟수를 보여준다. 우대 등급 후원자들에게 GENERAL AD를 주요채널로 지정하도록 홍보를 해야한다.
- EVENT는 우대 등급에서 낮은 평균 납부액을 보여준다. 우대 등급 후원자 중에 EVENT를 주요채널로 지정한 후원자가 다른 채널을 지정하도록 유도해야 한다.

2. 분산분석 – 4. 공분산분석

사용 데이터

-고객등급 - PAY_NUM : 납입후원횟수

- PAY_SUM_PAYMENTAMOUNT : 총납입금액

분석 목적

- 다변량 분석으로 부터 고객등급별 총납입금액과 후원횟수의 차이가 있다는 것을 알았다.
- 총납입금액과 후원횟수는 상관관계가 있다.
- 고객등급별 총납입금액의 순수한 차이를 알기위해 후원횟수를 통제한 일원공분산분석을 수행한다.

결과

일원공분산분석 - 고객등급x총납입금액

	F값	p-value
후원횟수 통제	85.7	0.000
후원횟수 미통제	294.3	0.000

실무적 인사이트

- 후원횟수는 총납입금액에 일정한 영향을 주고 있었다.
- 고객등급별 총납입금액을 비교할 때 과대평가된 요소가 있다.
- 고객등급 이외의 요소에서 총납입금액을 비교할 때 후원횟수의 통제가 필요하다.

3. 회귀분석 – 1. 단순회귀분석

사용 데이터

- PAY_NUM : 납입후원횟수
- PAY_SUM_PAYMENTAMOUNT : 총납입금액

분석 목적

- 앞선 분석들로 부터 총납입금액과 후원횟수 사이에 유의미한 관계가 있다.
- 후원횟수가 얼마나 총납입금액에 영향을 주는지 알고 싶다.

분석 결과

회귀분석 결과

측정항목	통계값
R-squared	0.666
F-statistic	1943
Prob (F-statistic)	2.22e-234

분석결과

회귀계수 테이블

변수명	coef	p-value
Intercept	-2.831e+04	0.001
PAY_NUM	2.666e+04	0.000

잔차의 평가

측정항목	통계값
Omnibus	331.824
Durbin-Watson	2.038

실무적 인사이트

- 총납입금액 = 26660*납입횟수 -28310
- 총납입금액을 늘리기 위해 납입횟수를 늘려야한다.
- 1회 납입 시 납입금액을 늘려야한다.

3. 회귀분석 – 2. 다중회귀분석

사용 데이터

- 고객등급
- PAY_NUM_NOPAY: 미납횟수
- PAY_SUM_PAYMENTAMOUNT : 총납입금액
- PAY_NUM : 납입후원횟수

분석 목적

- 총납입금액에 영향을 주는 변수들을 더 알아본다.

실무적 인사이트

- 총납입금액 = -74280*고객등급 + 15760*미납횟수 + 20710*납부횟수 + 238000
- 총납입금액에 가장 큰 영향을 주는 것은 고객등급이다.
- 고객등급을 높이기 위한 마케팅 전략이 필요
- 미납횟수의 회귀계수가 '+'인 것은 납부 횟수가 많아질 수록 미납횟수가 늘어나기 때문이라 추정. 이에 대한 확인 필요

회귀분석 결과

측정항목	통계값
R-squared	0.727
F-statistic	845.6
Prob (F-statistic)	5.21e-268

다중공선성 검사

변수	VIF
고객등급	1.978
미납횟수	1.25
납부횟수	1.66

회귀계수 테이블

변수명	coef	p-value
Intercept	2.38e+05	0.000
고객등급	-7.428e+04	0.000
미납횟수	1.576e+04	0.000
납부횟수	2.071e+04	0.000

잔차의 평가

측정항목	통계값
Omnibus	845.188
Durbin-Watson	2.006

3. 회귀분석 – 3. 더미변수를 이용한 회귀분석

사용 데이터

- 고객등급
- PAY_SUM_PAYMENTAMOUNT : 총납입금액

분석 목적

- 이전의 분석으로 총납입금액에 고객등급이 가장 큰 영향을 준다는 것을 알았다.
- 고객등급별 총납입금액에 얼마나 영향을 주는지 알아보자.

분석 과정

1) 결측값 제거
2) 더미변수 생성
3) 다중회귀분석 수행
4) 비교, 검정

더미변수 범례

고객등급	우대	골드	실버	브론즈	새싹
더미변수	first_drop	2	3	4	5

핵심코드

```
#더미변수 생성
df2 = pd.get_dummies(df['고객등급'], prefix='고객등급',drop_first=True)
df3 = pd.concat([df1, df2], axis = 1)

Model1 = smf.ols(formula = 'PAY_SUM_PAYMENTAMOUNT ~  고객등급_2+고객등급_3+ 고객등급_4 + 고객등급_5', data=df3).fit()
print(Model1.summary())
```

3. 회귀분석 – 3. 더미변수를 이용한 회귀분석

결과

회귀분석 결과

측정항목	통계값
R-squared	0.552
F-statistic	294
Prob (F-statistic)	9.83e-165

잔차의 평가

측정항목	통계값
Omnibus	1652
Durbin-Watson	2.044

회귀계수 테이블

변수명	coef	p-value	기대값	상승비율
Intercept=우대	7.516e+05	0.00	751,600	105%
골드	-3.861e+05	0.00	365,000	81.8%
실버	-5.509e+05	0.00	200,700	15.6%
브론즈	-5.78e+05	0.00	173,600	131%
새싹	-6.766e+05	0.00	75,000	

결과 해석

- F통계량이 294이고 유의확률이 0.01 이하이므로 본 회귀모형은 유효하다.
- 설명력이 55.2%이고 절편을 포함한 모든 변수가 유효하다.
- 최종 도출된 회귀식은 총납입금액 = 751600 – 386100*D1 – 550900*D2 – 578000*D3 – 676600*D4 이다.

실무적 인사이트

- 총납입금액의 기대값은 우대 등급이 가장 높다.
: 고객등급을 우대로 올리거나 우대를 유지시키는 전략이 필요하다.
- 새싹 -> 브론즈의 총납입금액 상승 비율이 가장 높다.
: 신규 후원자의 적응을 높고 이탈을 막을 전략이 필요하다.
- 브론즈 -> 실버의 총납입금액 상승 비율이 가장 낮다.
: 해당 등급의 승급에 총납입금액 이외의 요소가 중요하게 작용한다.