통계학실험(005)
과제 8 (9장)
2022-18608 문현호

1

예제1. 어느 시장 조사기관은 여러 가지 대중매체가 주는 정보의 양을 비교하기 위해 다음과 같은 실험을 계획하였다. 40명의 성인을 랜덤하게 추출하여 철저한 면접을 통해 TV, 신문, 라디오, 잡지 중 어느 매체를 많이 접하는지에 따라 분류하였다. 다음 표는 최근에 일어난 사건들에 대한 조사 대상자들의 인지도를 측정한 실험에서 얻어진 값들을 나타내고, 값이 클수록 인지도가 높은 것을 의미한다. 이 자료를 이용하여 사람들의 인지도가 대중매체에 따라 다르다고 할 수 있는지 유의수준 5%에서 검정해보자.

조사대상 대중매체			
TV	신문	라디오	잡지
16	13	18	11
19	14	18	15
25	15	15	11
22	16	14	17
21	15	14	17
15	13	10	13
16	19	18	14
22	16	15	16
21	20	15	13
18	14		11
	11		

Python 코드

```
# make a dataframe
arr = {'col': [16, 19, 25, 22, 21, 15, 16, 22, 21, 18] + [13, 14, 15, 16,
15, 13, 19, 16, 20, 14, 11]
+ [18, 18, 15, 14, 14, 10, 18, 15, 15] + [11, 15, 11, 17, 17, 13, 14, 16,
13, 11],
'num': [1 for   in range(11)] + [2 for   in range(10)] + [3 for   in
range(9)] + [4 for   in range(10)]
}
print(arr)
media = pd.DataFrame(arr)

# try anova
model = ols("num ~ C(col)", data=media).fit()
print(model.summary())
table = sm.stats.anova_lm(model, typ=2)
print(table)
```

OLS Regression Results

Dep. Variable:	num	R-squared:	0.499
Model:	OLS	Adj. R-squared:	0.277
Method:	Least Squares	F-statistic:	2.245
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	0.0396
Time:	04:26:56	Log-Likelihood:	-47.369
No. Observations:	40	AIC:	120.7
Df Residuals:	27	BIC:	142.7
Df Model:	12		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.0000	0.963	3.117	0.004	1.025	4.975
C(col)[T.11]	0.5000	1.076	0.465	0.646	-1.708	2.708
C(col)[T.13]	8.438e-15	1.076	7.84e-15	1.000	-2.208	2.208
C(col)[T.14]	-0.2000	1.054	-0.190	0.851	-2.363	1.963
C(col)[T.15]	-0.4286	1.029	-0.417	0.680	-2.540	1.683
C(col)[T.16]	-1.0000	1.054	-0.948	0.351	-3.163	1.163
C(col)[T.17]	1.0000	1.179	0.848	0.404	-1.419	3.419
C(col)[T.18]	-0.5000	1.076	-0.465	0.646	-2.708	1.708
C(col)[T.19]	-1.5000	1.179	-1.272	0.214	-3.919	0.919
C(col)[T.20]	-1.0000	1.361	-0.735	0.469	-3.793	1.793
C(col)[T.21]	-2.0000	1.179	-1.697	0.101	-4.419	0.419
C(col)[T.22]	-2.0000	1.179	-1.697	0.101	-4.419	0.419
C(col)[T.25]	-2.0000	1.361	-1.469	0.153	-4.793	0.793

Omnibus:	0.146	Durbin-Watson:	0.824
Prob(Omnibus):	0.930	Jarque-Bera (JB):	0.007
Skew:	-0.019	Prob(JB):	0.996
Kurtosis:	2.947	Cond. No.	24.4

	sum_sq	df	F	PR(>F)
C(col)	24.960714	12.0	2.245181	0.039576
Residual	25.014286	27.0	NaN	NaN

- 자료의 종류에 따른 인지도의 차이가 존재하는지를 확인했을 때, 분산분석 및 F 검정 결과 p value는 0.0396으로, 0.05보다 작다. 즉 자료의 종류에 따라 인지도의 차이가 존재하지 않는다는 귀무가설을 기각한다.
- 따라서 유의수준 5%에서 인지도는 자료의 종류에 따라서 달라졌다.

add 1

추가과제1. 강의노트 9.4 예제에서 예제1에 주어진 표에서 신문, 라디오, 잡지에 해당하는 열만 추출한 data를 사용한다. 이 자료를 이용하여 사람들의 인지도가 신문, 라디오, 잡지에 따라 다르다고 할 수 있는지 유의수준 5%에서 검정해보자.

Python 코드

```
# make a dataframe
arr = {'col': [16, 19, 25, 22, 21, 15, 16, 22, 21, 18] + [13, 14, 15, 16,
15, 13, 19, 16, 20, 14, 11]
+ [18, 18, 15, 14, 14, 10, 18, 15, 15],
'num': [1 for _ in range(11)] + [2 for _ in range(10)] + [3 for _ in
range(9)]
}
print(arr)
media = pd.DataFrame(arr)

# try anova
model = ols("num ~ C(col)", data=media).fit()
print(model.summary())
table = sm.stats.anova_lm(model, typ=2)
print(table)
```

OLS Regression Results

Dep. Variable:	num	R-squared:	0.270
Model:	OLS	Adj. R-squared:	-0.058
Method:	Least Squares	F-statistic:	0.8230
Date:	Mon, 12 Dec 2022	Prob (F-statistic):	0.603
Time:	04:26:56	Log-Likelihood:	-32.468
No. Observations:	30	AIC:	84.94
Df Residuals:	20	BIC:	98.95
Df Model:	9		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	3.0000	0.875	3.430	0.003	1.176	4.824
C(col)[T.11]	0.5000	0.978	0.511	0.615	-1.540	2.540
C(col)[T.13]	2.887e-15	0.978	2.95e-15	1.000	-2.040	2.040
C(col)[T.14]	-0.2000	0.958	-0.209	0.837	-2.199	1.799
C(col)[T.15]	-0.1667	0.945	-0.176	0.862	-2.137	1.804
C(col)[T.16]	-0.3333	1.010	-0.330	0.745	-2.440	1.773
C(col)[T.17]	1.0000	1.071	0.934	0.362	-1.235	3.235
C(col)[T.18]	2.922e-15	1.010	2.89e-15	1.000	-2.107	2.107
C(col)[T.19]	-1.0000	1.237	-0.808	0.428	-3.580	1.580
C(col)[T.20]	-1.0000	1.237	-0.808	0.428	-3.580	1.580

Omnibus:	0.321	Durbin-Watson:	0.533
Prob(Omnibus):	0.852	Jarque-Bera (JB):	0.491
Skew:	-0.005	Prob(JB):	0.782
Kurtosis:	2.374	Cond. No.	18.8

	sum_sq	df	F	PR(>F)
C(col)	5.666667	9.0	0.823045	0.602577
Residual	15.300000	20.0	NaN	NaN

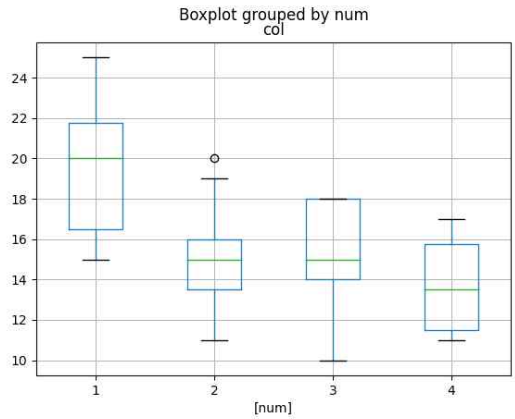
- 마찬가지로 인지도에 위 3개의 자료의 종류에 따라 차이가 있는지를 확인했을 때, 분산분석과 F 검정 결과 p value는 0.603으로, 0.05보다 더 크다. 즉 자료의 종류에 따라 인지도의 차이가 존재하지 않는다는 귀무가설을 기각할 수 없다.
- 따라서 인지도는 유의수준 5%에서, 세 자료의 종류에 따라서 별로 달라지지 않았다.

- 이때 이 이유는 다음과 같은 내용으로 확인할 수 있다.

Python 코드

```
# make a dataframe
arr = {'col': [16, 19, 25, 22, 21, 15, 16, 22, 21, 18] + [13, 14, 15, 16,
15, 13, 19, 16, 20, 14, 11]
+ [18, 18, 15, 14, 14, 10, 18, 15, 15] + [11, 15, 11, 17, 17, 13, 14, 16,
13, 11],
'num': [1 for _ in range(10)] + [2 for _ in range(11)] + [3 for _ in
range(9)] + [4 for _ in range(10)]
}
print(arr)
media = pd.DataFrame(arr)

# box plot of all data
media.boxplot('col', by=['num'])
plt.show()
```



- 박스플롯을 보면 알 수 있듯이, 1번, TV라는 매체의 인지도가 다른 것들보다 상당히 높고, 다른 세 매체끼리는 인지도의 분포가 유사한 반면 TV의 분포가 큰 차이를 나타낸다. 따라서 1번부터 4번을 모두 포함한 경우는 인지도의 차이가 존재한다는 결과가 나온 반면, 1번을 제외하면 인지도의 차이가 거의 없다는 결과가 나타나게 되었다.

2

예제2. 어느 회사의 마케팅 부서에서는 하나의 상품에 대해 세 가지 다른 디자인의 포장을 적용한 후 이 상품들을 서로 다른 5군데의 상점에서 한 달 동안 판매하였다. 그리고 그 판매 결과는 아래와 같다. 제품의 매출은 판매되는 상점과 제품의 포장 디자인에 따라 다르다고 할 수 있는가? 적절한 가설을 쓰고 유의수준 5%에서 이를 검정하시오.

	상점1	상점2	상점3	상점4	상점5
상자1	210	230	190	180	190
상자2	195	170	200	190	193
상자3	295	275	290	275	265

Python 코드

```
# make a dataframe
arr = {'y': [210, 230, 190, 180, 190, 195, 170, 200, 190, 193, 295, 275,
290, 275, 265],
'a': ["Box " + str(i) for i in range(1, 4) for _ in range(5)],
'b': ["Shop" + str(i) for i in range(1, 6)]*3
}
print(arr)
sell = pd.DataFrame(arr)

# try anova
model = ols("y ~ C(a) + C(b)", data=sell).fit()
print(model.summary())
table = sm.stats.anova_lm(model, typ=2)
print(table)
```

OLS Regression Results

Dep. Variable:	y	R-squared:	0.926
Model:	OLS	Adj. R-squared:	0.870
Method:	Least Squares	F-statistic:	16.60
Date:	Fri, 09 Dec 2022	Prob (F-statistic):	0.000405
Time:	14:29:37	Log-Likelihood:	-58.063
No. Observations:	15	AIC:	130.1
Df Residuals:	8	BIC:	135.1
Df Model:	6		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	210.1333	10.861	19.348	0.000	185.088	235.178
C(a)[T.Box 2]	-10.4000	10.055	-1.034	0.331	-33.587	12.787
C(a)[T.Box 3]	80.0000	10.055	7.956	0.000	56.813	103.187
C(b)[T.Shop2]	-8.3333	12.981	-0.642	0.539	-38.268	21.601
C(b)[T.Shop3]	-6.6667	12.981	-0.514	0.621	-36.601	23.268
C(b)[T.Shop4]	-18.3333	12.981	-1.412	0.196	-48.268	11.601
C(b)[T.Shop5]	-17.3333	12.981	-1.335	0.219	-47.268	12.601

Omnibus:	1.945	Durbin-Watson:	2.073
Prob(Omnibus):	0.378	Jarque-Bera (JB):	0.567
Skew:	0.435	Prob(JB):	0.753
Kurtosis:	3.388	Cond. No.	6.45

	sum_sq	df	F	PR(>F)
C(a)	24467.200000	2.0	48.398787	0.000034
C(b)	711.066667	4.0	0.703284	0.611391
Residual	2022.133333	8.0	NaN	NaN

- 디자인(상자, a)에 따라, 그리고 상점(b)에 따라 판매량의 차이가 존재하는지를 확인했을 때, 분산분석 및 F 검정 결과 디자인에 따른 p value는 0.000034로 귀무가설(판매량은 디자인에 따라 변하지 않는다)의 유의수준인 0.05보다 작았고, 상점에 따른 p value는 0.61로 귀무가설(판매량은 상점에 따라 변하지 않는다)의 유의수준인 0.05보다 컸다.
- 따라서 유의수준 5%에서 판매량은 디자인에 따라 달라졌고, 판매 상점에 따라서는 달라지지 않았다.

add 2

추가과제2. 강의노트 9.4 예제에서 예제 2와 동일한 표를 이용하여 문제를 풀이한다.

(1) 각각의 포장 디자인 별 제품 매출의 평균과 표준편차를 구하여라.

(2) 각각의 포장 디자인 별 제품 매출을 boxplot을 이용하여 시각화하여라.

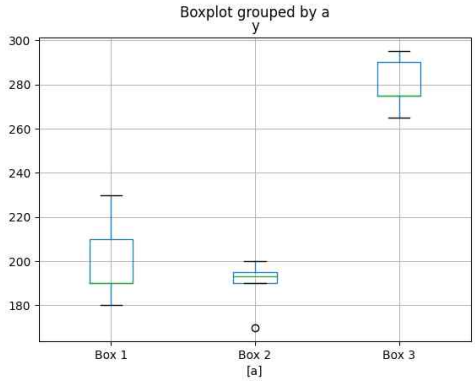
(3) 소문제 1,2 에서 구한 요약통계량과 boxplot을 예제 2에서 실시한 검정 결과와 종합하여 서술하여라.

Python 코드

```
# mean and std for each column
sell  = pd.DataFrame({"Box 1": [210, 230, 190, 180, 190],
"Box 2": [195, 170, 200, 190, 193],
"Box 3": [295, 275, 290, 275, 265]})
print(sell .describe())

# box plot of all data
sell.boxplot('y', by=['a'])
plt.show()
```

	Box 1	Box 2	Box 3
count	5.0	5.000000	5.000000
mean	200.0	189.600000	280.000000
std	20.0	11.545562	12.247449
min	180.0	170.000000	265.000000
25%	190.0	190.000000	275.000000
50%	190.0	193.000000	275.000000
75%	210.0	195.000000	290.000000
max	230.0	200.000000	295.000000



- 1) 평균과 표준편차는 다음과 같다. 1번(200.0, 20.0), 2번(189.6, 11.6), 3번(280.0, 12.3)
- 2) 제품 매출을 Boxplot으로 나타낸 결과는 위 그림과 같다.
- 3) 위에서 나타낸 결과대로, 판매량은 박스 디자인에 따라 충분히 크게 달라진다는 점을 확인할 수 있다. 평균 사이의 차이의 범위(20~100)가 그 값의 범위(200~300)보다 매우 컸다.

3

예제3. 남녀의 성별과 고단백질로 구성된 아침 식사의 섭취 여부가 성인의 신체적 활동 능력에 영향을 미치는지를 알아보기 위하여 랜덤하게 선택된 남녀 10명에 대해 각각 5명씩 고단백질 아침식사와 저단백질 아침식사를 섭취하게 한 후, 신체적 능력을 테스트를 통해 측정하였다. 측정된 점수가 높을수록 신체 활동 능력이 더 우수하다는 것을 의미한다. 실험 결과가 아래와 같을 때, 주어진 자료에 대해 이원배치법을 적용한 후 그 결과를 해석하여라.

	고단백질 식사	저단백질 식사
남성	10 7 9 6 8	5 4 7 4 5
여성	5 4 6 3 2	3 4 5 1 2

Python 코드

```
# make a dataframe
arr = {'y': [10, 7, 9, 6, 8, 5, 4, 7, 4, 5, 5, 4, 6, 3, 2, 3, 4, 5, 1, 2],
'a': ["Male" for _ in range(10)] + ["Female" for _ in range(10)],
'b': (["High-protein" for _ in range(5)] + ["Low-Protein" for _ in
range(5)])*2
}
print(arr)
ability = pd.DataFrame(arr)

# try anova without considering interaction
model = ols("y ~ C(a) + C(b)", data=ability).fit()
print(model.summary())
table = sm.stats.anova_lm(model, typ=2)
print(table)
```

OLS Regression Results

Dep. Variable:	y	R-squared:	0.613
Model:	OLS	Adj. R-squared:	0.568
Method:	Least Squares	F-statistic:	13.48
Date:	Fri, 09 Dec 2022	Prob (F-statistic):	0.000312
Time:	15:03:27	Log-Likelihood:	-35.557
No. Observations:	20	AIC:	77.11
Df Residuals:	17	BIC:	80.10
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.5000	0.601	7.482	0.000	3.231	5.769
C(a)[T.Male]	3.0000	0.695	4.320	0.000	1.535	4.465
C(b)[T.Low-Protein]	-2.0000	0.695	-2.880	0.010	-3.465	-0.535

Omnibus:	1.825	Durbin-Watson:	2.220
Prob(Omnibus):	0.402	Jarque-Bera (JB):	1.072
Skew:	0.204	Prob(JB):	0.585
Kurtosis:	1.942	Cond. No.	3.19

	sum_sq	df	F	PR(>F)
C(a)	45.0	1.0	18.658537	0.000465
C(b)	20.0	1.0	8.292683	0.010400
Residual	41.0	17.0	NaN	NaN

- 성별(a), 식단(b)에 따라 신체적 능력이 변화하는가에 대해 분산분석 및 F 검정을 진행한 결과는 위와 같다.
- 먼저 성별에 따라 신체적 능력이 변화하는지에 대한 F 검정 결과의 p value는 0.000465로 유의수준 0.05보다 훨씬 낮았으며, 따라서 성별에 따라 신체적 능력이 변화하지 않는다는 귀무가설을 기각해야 한다.
- 식단에 따라 신체적 활동 능력이 변화하는지에 대한 F 검정 결과의 p value는 0.010400으로 마찬가지로 유의수준 0.05보다 훨씬 낮고, 따라서 식단에 따라 신체적 능력이 변화하지 않는다는 귀무가설을 기각한다.
- 따라서 유의수준 0.05에서, 신체적 능력은 성별에 따라, 그리고 아침 식사가 고단백인지 저단백인지에 따라 달라졌음을 확인할 수 있다.

add 3

추가과제3. 강의노트 9.4 예제에서 예제 3과 동일한 표를 이용하여 문제를 풀이한다.

(1) 성별과 고단백질로 구성된 아침 식사의 섭취 여부에 대한 상호작용의 유무를 평균 그림(interaction plot)을 통해 확인하고 그래프를 해석하여라.

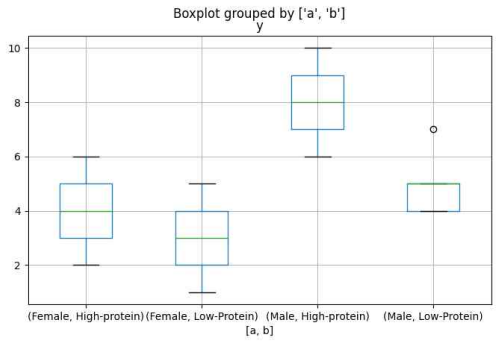
(2) 상호작용을 포함한 이원배치 분산분석 검정을 통해 유의수준 5%에서 그 결과를 해석하여라.

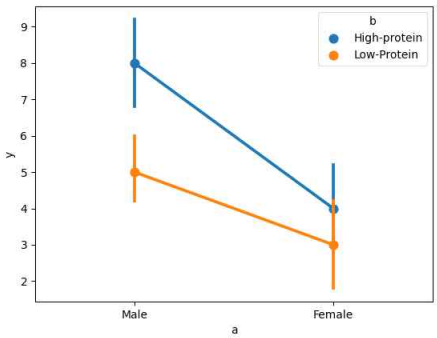
Python 코드

```
# box plot of all data
ability.boxplot('y', by=['a', 'b'])
plt.show()

# interaction plot of all data
sns.pointplot(x='a', y='y', hue='b', data=ability)
plt.show()

# try anova with considering interaction
model = ols("y ~ C(a) * C(b)", data=ability).fit()
print(model.summary())
table = sm.stats.anova_lm(model, typ=2)
print(table)
```





- 1) 두 그래프, 특히 interaction plot을 보면, 전체적으로 남성이 여성보다 신체적 능력이 높고, 고단백 아침 식사를 섭취한 그룹이 저단백 그룹보다 신체적 능력이 높은 것을 알 수 있다. 다만 이 전체적인 변화는 서로 간의 조합에서 크게 변하지 않았다.

OLS Regression Results

Dep. Variable:	y	R-squared:	0.660
Model:	OLS	Adj. R-squared:	0.597
Method:	Least Squares	F-statistic:	10.37
Date:	Fri, 09 Dec 2022	Prob (F-statistic):	0.000494
Time:	15:03:27	Log-Likelihood:	-34.257
No. Observations:	20	AIC:	76.51
Df Residuals:	16	BIC:	80.50
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	4.0000	0.671	5.963	0.000	2.578	5.422
C(a)[T.Male]	4.0000	0.949	4.216	0.001	1.989	6.011
C(b)[T.Low-Protein]	-1.0000	0.949	-1.054	0.307	-3.011	1.011
C(a)[T.Male]:C(b)[T.Low-Protein]	-2.0000	1.342	-1.491	0.155	-4.844	0.844

Omnibus:	2.411	Durbin-Watson:	2.361
Prob(Omnibus):	0.300	Jarque-Bera (JB):	1.150
Skew:	0.124	Prob(JB):	0.563
Kurtosis:	1.852	Cond. No.	6.85

	sum_sq	df	F	PR(>F)
C(a)	45.0	1.0	20.000000	0.000385
C(b)	20.0	1.0	8.888889	0.008814
C(a):C(b)	5.0	1.0	2.222222	0.155487
Residual	36.0	16.0	NaN	NaN

- 2) 성별(a), 식단(b)에 따라, 추가로 그 둘 사이 interaction에 의해 신체적 능력이 변화하는가에 대해 분산분석 및 F 검정을 진행한 결과는 위와 같다.

- 먼저 성별에 따라서의 F 검정 결과의 p value는 0.000385, 식단에 따라서의 F 검정 결과의 p value는 0.008814로 0.05보다 훨씬 낮고, 따라서 두 조건 각각에 따라 신체적 능력이 변화하지 않는다는 귀무가설을 위와 마찬가지로 기각한다.
- 또한, 이들과 별개로 성별과 식단의 interaction에 따라 신체적 활동 능력이 변화하는지에 대한 F 검정 결과 p value는 0.155로, 유의수준 0.05보다 크다. 따라서 성별과 식단의 interaction에 따라 신체적 능력이 변화하지 않는다는 귀무가설을 채택한다.
- 이에 따라, 결과적으로 신체적 능력은 유의수준 0.05에서, 성별에 따라, 그리고 아침 식사의 단백질 함량에 따라 달라졌으며, 그 둘 사이의 interaction은 존재하지 않았다.

감사합니다.