



Day36; 20221027

📅 날짜	@2022년 10월 27일
👤 유형	@2022년 10월 27일
☰ 태그	

GitHub - u8yes/Python

You can't perform that action at this time. You signed in with another tab or window. You signed out in another tab or window. Reload to refresh your session. Reload to refresh your session.

<https://github.com/u8yes/Python>

u8yes/Python



1 Contributor 0 Issues 1 Star 0 Forks

ADsP) 통계적 추정(Estimation) - 신뢰 구간

- 99% 신뢰수준에 대한 신뢰구간이 95% 신뢰수준에 대한 신뢰구간보다 길다
- 표본의 크기가 커지면 신뢰구간의 길이는 줄어든다

▣ 2. 다음 중 신뢰구간에 대한 설명 중 가장 적절하지 않은 것은?

- 1 99%의 신뢰구간이 95%의 신뢰구간보다 길다
- 2 관측치의 크기가 커지면 신뢰구간의 길이는 줄어든다
- 3 95% 신뢰구간은 미지의 수가 포함되지 않을 확률이 95%를 의미한다
- 4 점추정의 정확성을 보완하는 방법이다

- 신뢰수준 95% 의미는 실제 모수값이 신뢰구간에 존재할 확률이 95%라 할 수 있다

가설 검정

기각역
critical region

- 검정통계량(t-value)의 분포에서 유의수준의 크기에 해당하는 영역
- 계산한 검정통계량의 유의성(귀무가설의 기각)을 판정하는 기준

제 1종 오류

α error, 귀무가설이 참인데 기각하게 되는 오류

제 2종 오류

β error, 귀무가설이 거짓인데 채택하는 오류

- 두 가지 오류가 작을 수록 바람직함

- 두 가지를 동시에 줄일 수 없기 때문에 1종오류를 범할 확률의 최대 허용치를 미리 어떤 특정값(유의수준)으로 지정해 놓고 제 2종 오류의 확률을 가장 작게 해주는 검정 방법을 사용함

유의수준(α)

▪ Significance level, 제 1종 오류의 최대 허용 한계

▪ 유의수준 0.05(5%) : 100번 실험에서 1종 오류 범하는 최대 허용 한계가 5번

유의확률 = p-value

- Probability Value, $0 \leq p\text{-value} \leq 1$, 1종 오류를 범할 확률, 귀무가설을 지지하는 정도
- 귀무가설이 사실일 때 기각하는 1종 오류 시 우리가 내린 판정이 잘못되었을 확률
- 검정 통계량들은 거의 대부분이 귀무가설을 가정하고 얻게 되는 값
- 검정 통계량에 관한 확률로, 극단적인 표본 값이 나올 확률
- p-value가 작을 수록 그 정도가 약하다고 보면, $p\text{-value} < \alpha$ 귀무가설을 기각, 대립가설을 채택함
- p-value가 0.05(5%) : 귀무가설을 기각했을 때 기각 결정이 잘못될 확률이 5%임

모수적 검정

- 검정하고자 하는 모집단의 분포에 대해 가정을 하고, 그 가정하에서 검정 통계량과 검정 통계량의 분포를 유도해 검정을 실시함
- 1) 가정된 분포의 모수에 대해 가설 설정
- 2) 관측된 자료를 이용해 구한 표본 평균, 표본 분산 등을 이용해 검정 실시

모수적 통계의 전제조건

- 표본의 모집단이 정규분포를 이루어야 하며, 집단 내의 분산은 같아야 함
- 변인(=변수)은 등간척도나 비율척도로 측정되어야 함 (아니면 비모수 통계 사용)

모수 검정방법

- T test, Paired T test, Two sample T test, ANOVA test, z분포, t분포, F분포

모수 검정방법 사용 예

- 모평균과 표본평균과의 차이, 표본평균 간의 차이 : z 분포, t 분포
- 모분산과 표본분산과의 차이, 표본분산 간의 차이 : F 분포

3-61. 모수적 추론 : T-test

T test

- 평균값이 올바른지, 두 집단의 평균 차이가 있는지를 검증하는 방법으로 t값을 사용함
- t값이 커질수록 p-value는 작아지며, 집단간 유의한 차이를 보일 가능성이 높아짐

t-검정 방법	예시
One Sample t-test	<ul style="list-style-type: none">▪ 단일 표본의 평균 검정을 위한 방법▪ S사 USB의 평균 수명은 20000 시간이다
Paired t-test 대응표본 t-검정	<ul style="list-style-type: none">▪ 동일 개체에 어떤 처리를 하기 전, 후의 자료를 얻을 때 차이 값에 대한 평균 검정을 위한 방법▪ 예) 매일 1시간 한달 걸으면 2Kg이 빠진다 (걷기 수행 전/수행 후)▪ 가능한 동일한 특성을 갖는 두 개체에 서로 다른 처리를 하여 그 처리의 효과를 비교하는 방법(matching)▪ 예) X질병 환자들을 두 집단으로 나누어 A, B 약을 투약해 약의 효과 비교
Two sample t-test 독립표본 t-검정	<ul style="list-style-type: none">▪ 서로 다른 두 그룹의 평균을 비교하여 두 표본의 차이가 있는지 검정하는 방법▪ 귀무가설 - 두 집단의 평균 차이 값이 0이다▪ 2학년과 3학년의 결석률은 같다

3-61. Paired t-test (대응표본 t-검정)

- ☞ 수면유도제 데이터를 통한 ‘두 집단의 평균이 같다’는 가설에 대한 Paired t-test

```
> t.test(extra~group, data=sleep, paired=TRUE)
```

Paired t-test

```
data: extra by group
t = -4.0621, df = 9, p-value = 0.002833
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
-2.4598858 -0.7001142
sample estimates:
mean of the differences
-1.58
```

- paired=TRUE: Paired t-test, 짹을 이루는 데이터인 경우 분석 전 등분산성 검정 필요 없음
- df = 9 : 그룹별 데이터의 수 10개 → 분석 전 정규성 검정 실시
- p-value 가 0.002833 으로 두 집단의 평균이 같다는 귀무가설을 기각할 수 있다
- 신뢰구간에 0이 포함되지 않음

3-61. Two Sample t-test(=독립표본 t-test)

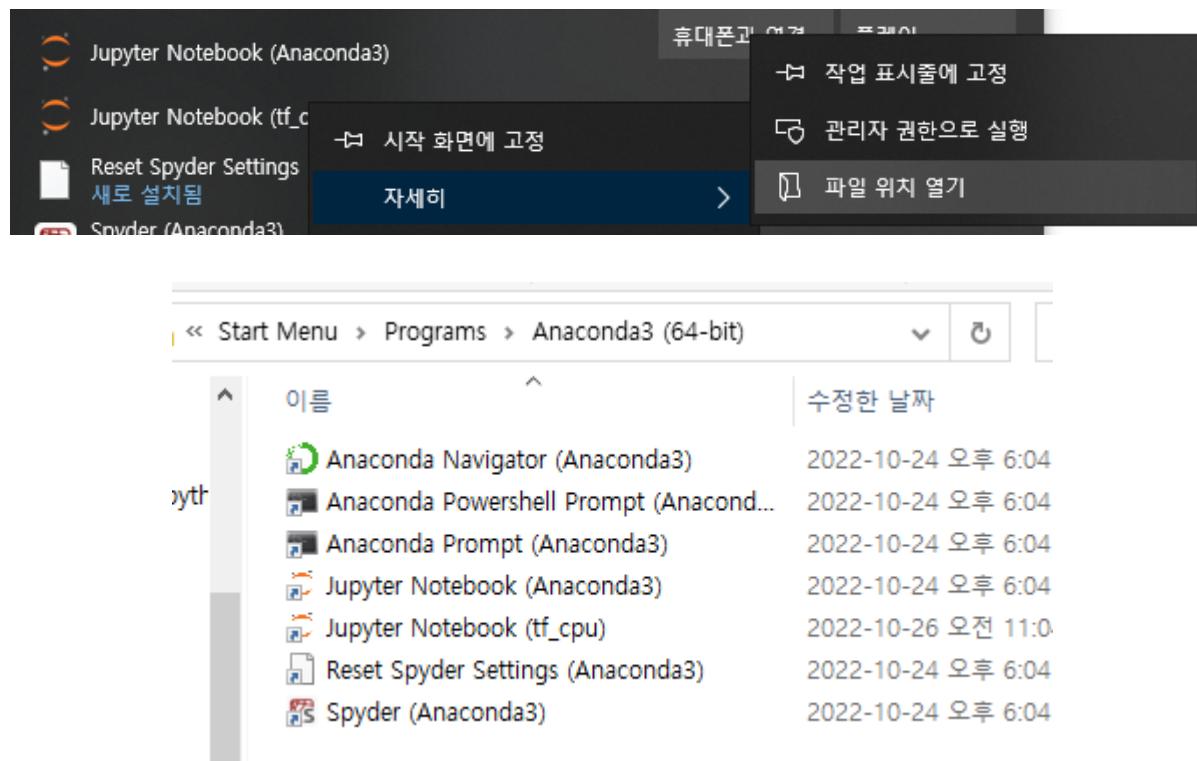
▣ 수면유도제 데이터를 통한 '집단 간 평균이 같다'는 가설에 대한 t-test

```
> t.test(extra~group, data=sleep, var.equal=TRUE)
```

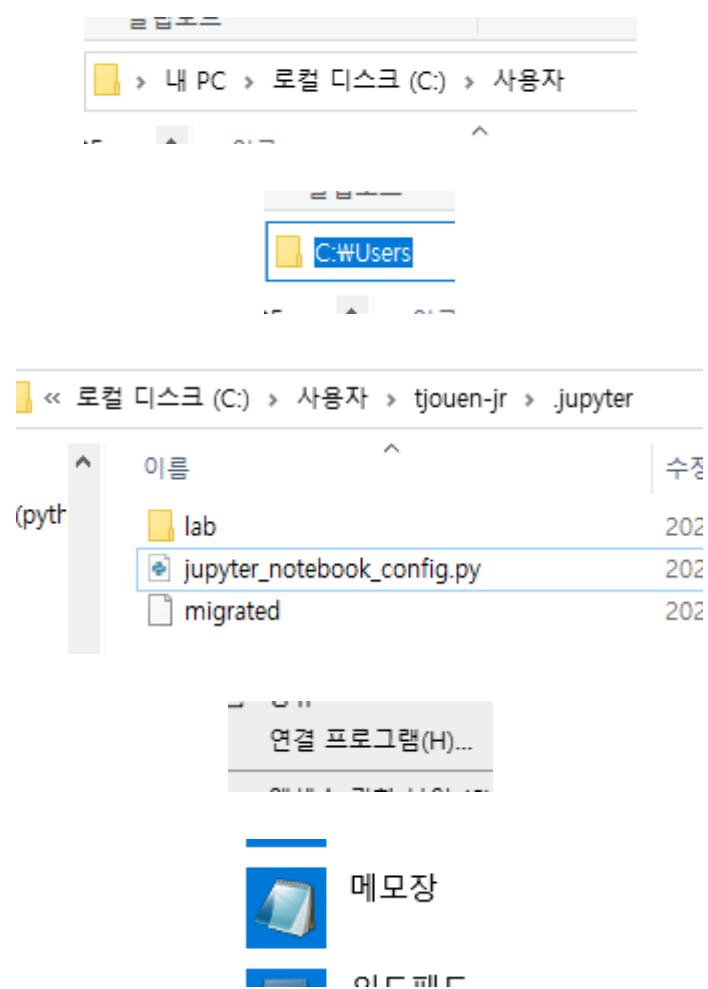
Two Sample t-test

```
data: extra by group  
t = -1.8608, df = 18, p-value = 0.07919  
alternative hypothesis: true difference in means is not equal to 0  
95 percent confidence interval:  
-3.363874 0.203874  
sample estimates:  
mean in group 1 mean in group 2  
0.75 2.33
```

- var.equal=TRUE: 두 집단의 모분산이 같다는 등분산성 만족 → 분석 전 등분산성 검정 실시
- df = 18 : 그룹이 2개이므로 데이터의 수 20개 → 분석 전 정규성 검정 실시
- p-value 가 0.07919 로 두 집단의 평균이 같다는 귀무가설을 기각할 수 없다
- 신뢰구간에 0이 포함되므로 두 집단간 평균에 차이가 없다고 해석할 수 있음



```
(tf_cpu) D:\heaven_dev\workspaces\Python\src\python>jupyter notebook --generate-config
```



jupyter_notebook_config.py - Windows 메모장

파일(F) 편집(E) 서식(O) 보기(V) 도움말(H)

```

## Sets the maximum allowed size of the client request body, specified in the
# Content-Length request header field. If the size in a request exceeds the
# configured value, a malformed HTTP message is returned to the client.
#
# Note: max_body_size is applied even in streaming mode.
# Default: 536870912
# c.NotebookApp.max_body_size = 536870912

## Gets or sets the maximum amount of memory, in bytes, that is allocated for use
# by the buffer manager.
# Default: 536870912
# c.NotebookApp.max_buffer_size = 536870912

## Gets or sets a lower bound on the open file handles process resource limit.
# This may need to be increased if you run into an OSError: [Errno 24] Too many
# open files. This is not applicable when running on Windows.
# Default: 0
# c.NotebookApp.min_open_files_limit = 0

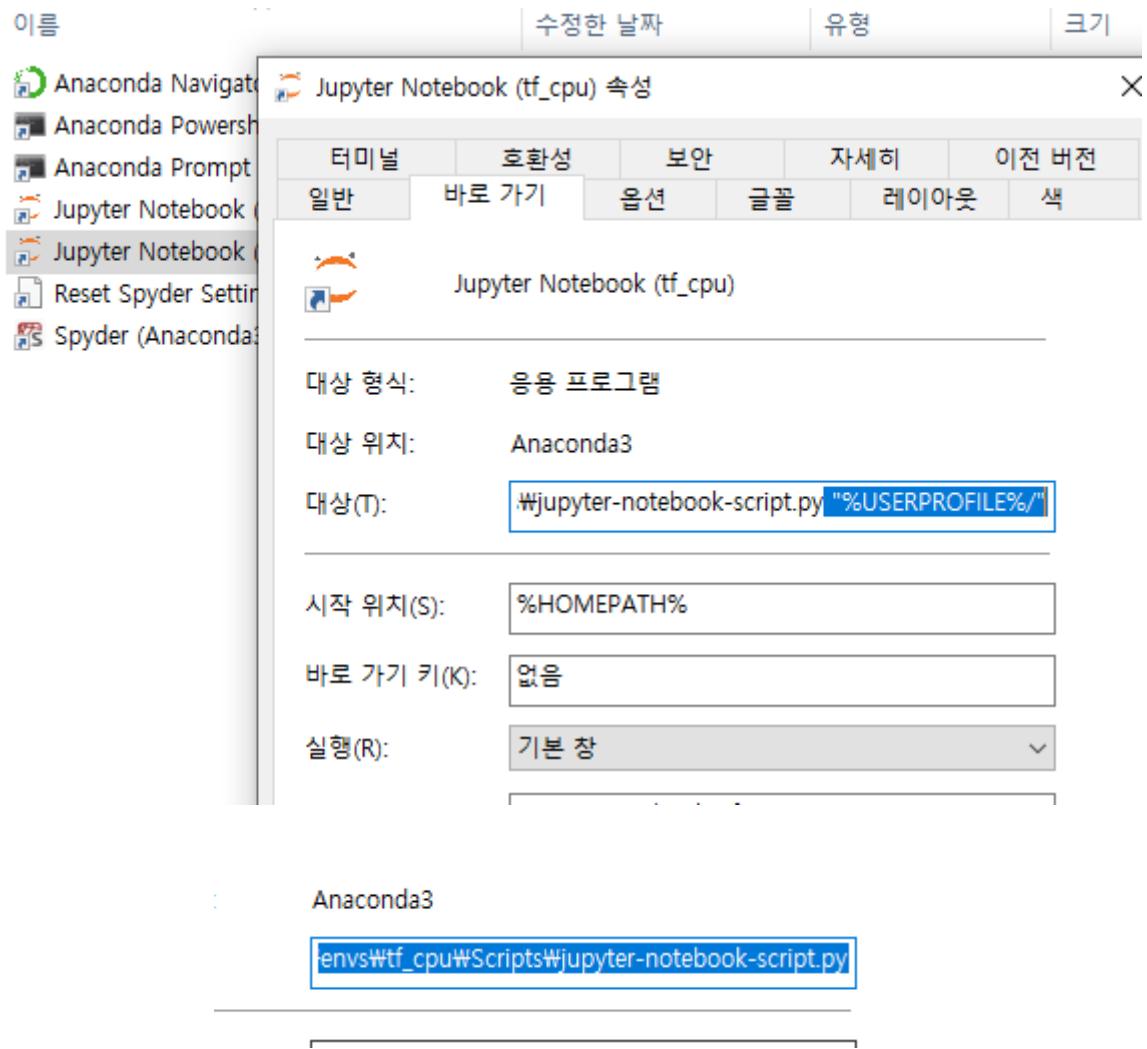
## Dict of Python modules to load as notebook server extensions. Entry values can
# be used to enable and disable the loading of the extensions. The extensions
# will be loaded in alphabetical order.
# Default: {}
# c.NotebookApp.nbserver_extensions = {}

## The directory to use for notebooks.
# Default: ""
# c.NotebookApp.notebook_dir = ""
```

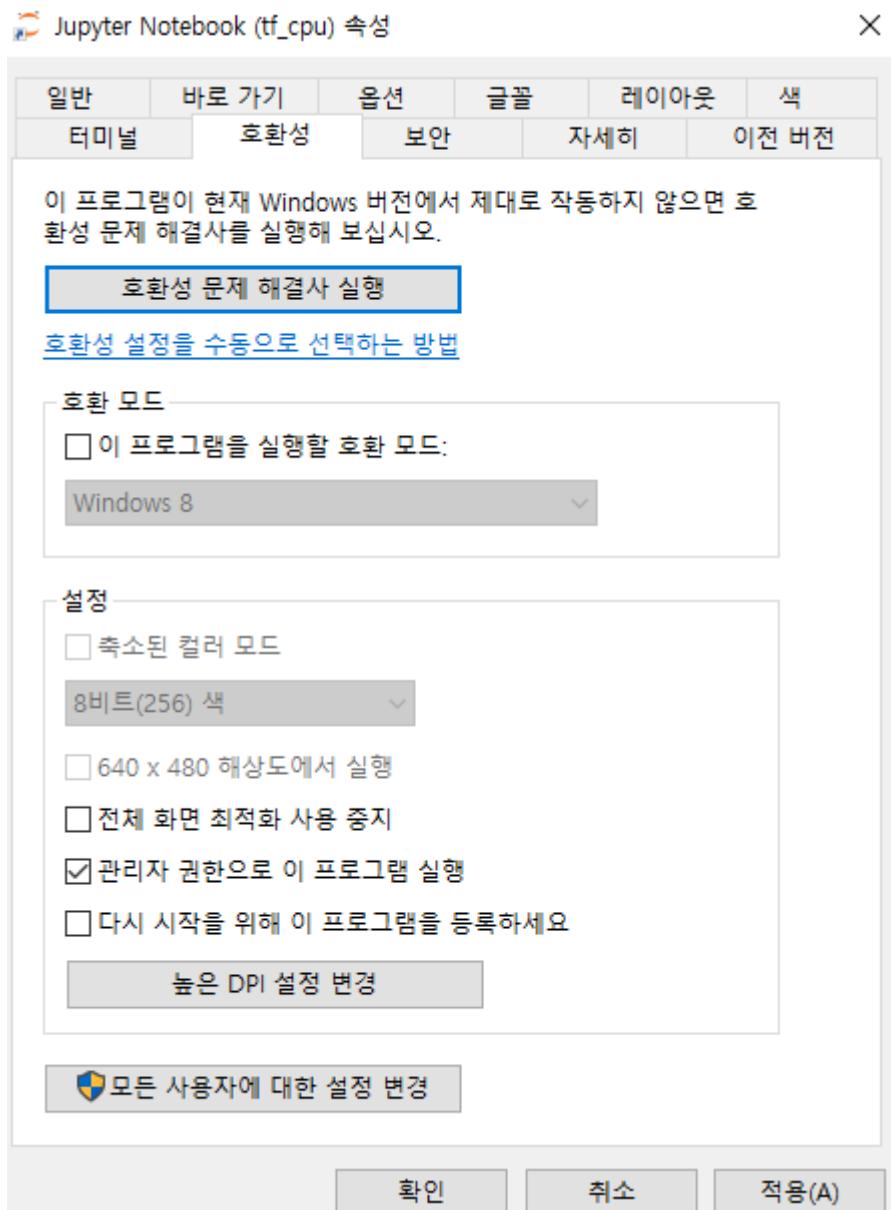
#을 지우고 공백 지워서 아래처럼 만들어줌.

```
# Default: ""
c.NotebookApp.notebook_dir = 'D:\heaven_dev\workspaces\Python\src\python'
```

userprofile 삭제 해줌 (C:\ProgramData\Microsoft\Windows\Start
Menu\Programs\Anaconda3 (64-bit))



'관리자 권한으로 이 프로그램 실행'을 클릭했기에 앞으로는 따로 관리자 권한 실행을 따로 누를 필요가 없음.



C ⓘ localhost:8890/tree

jupyter

Files Running Clusters Conda

Select items to perform actions on them.

0 /

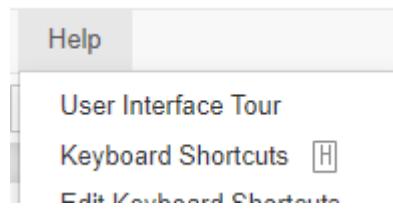
chap01_출력,연산기능,변수.ipynb

chap02_변수명규칙.ipynb

부호비트 - 가장 왼쪽에 양수 0, 음수 1

음수 -1 비트 표현

2의 보수법을 통해 음수가 양수되고, 양수가 음수되게 할 수 있음.



ADsP)비모수(명목), 비모수(서열), 모수

3-61~63. 추론(inference) - 문제 4

4. 다음의 통계 검정 중 표본특성이 2개 표본 이상일 때의 비모수 검정이 아닌 것은?

- 1 부호 검정
- 2 크루스칼-왈리스 검정
- 3 맨-위트니 검정
- 4 카이스퀘어 검정

비교대상 집단수	관계	비모수-명목척도	비모수-서열척도	모수
1		카이스퀘어 검정	Kolmogorov-Smirnov test	One sample T test
2	독립	Crosstab	Mann-Whitney	Two sample T test
	대응 자료	McNemar test	Wilcoxon signed -rank test Sign test	Paired T test
k(다변량)	독립		Kruskal-Wallis test	ANOVA test
	대응 자료	Cochran test	Friedman test	

3-64. 회귀 모형의 가정

16

▶ 회귀 모형의 가정

- 선형성 : 독립변수의 변화에 따라 종속변수도 변화하는 **선형(linear) 모형**이다
- 독립성 : 잔차와 독립변수의 값이 관련되어 있지 않다 (Durbin-Watson 통계량 확인)
- 정규성 : 잔차항이 **정규분포**를 이뤄야 한다
- 등분산성 : 잔차항들의 분포는 **동일한 분산**을 갖는다
- 비상관성 : 잔차들끼리 **상관이 없어야** 한다

Normal Q-Q plot

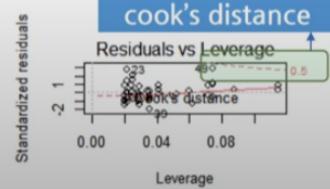
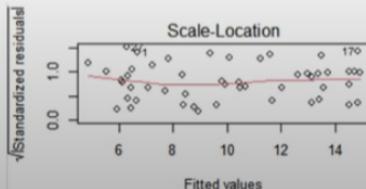
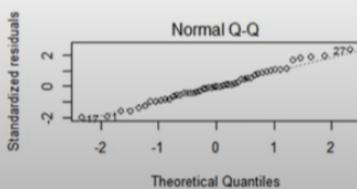
- **정규성**(정상성), 잔차가 정규분포를 잘 따르고 있는지를 확인하는 그래프
- 잔차들이 그래프 선상에 있어야 이상적임

Scale-Location

- **등분산성**, y축이 표준화 잔차를 나타내며, 기울기 0인 직선이 이상적임

Cook's Distance

- 일반적으로 1값이 넘어가면 관측치를 영향점(influence points)로 판별



▶ 다중공선성(Multicollinearity)

- 모형의 일부 설명변수(=예측변수)가 다른 설명변수와 상관되어 있을 때 발생하는 조건
- 중대한 다중공선성은 **회귀계수의 분산을 증가**시켜 불안정하고 해석하기 어렵게 만들기 때문에 문제가 됨
- R의 vif 함수를 사용해 구할 수 있으며, **VIF** 값이 10이 넘으면 다중공선성이 존재한다고 봄

▶ 해결방법

variance inflation factor

- 높은 상관 관계가 있는 설명변수를 모형에서 제거하는 것으로 해결함
- 설명변수를 제거하면 대부분 R-square가 감소함
- 단계적 회귀분석을 이용하여 제거함

▶ 설명변수의 선택 원칙

- y에 영향을 미칠 수 있는 모든 설명변수 x들은 y의 값을 예측하는 데 참여시킴
- 설명변수 x들의 수가 많아지면 관리에 많은 노력이 요구되므로 가능한 범위 내에서 적은 수의 설명변수를 포함시켜야 함
- 두 원칙이 이율배반적이므로 적절한 설명변수 선택이 필요함

3-67. 실병 변수 선택 방법

▶ step 함수를 사용한 후진제거법

```
> step(lm(y~x1+x2+x3+x4, df), direction='backward')
Start: AIC=26.94
y ~ x1 + x2 + x3 + x4
```

- 후진제거법 : direction = 'backward'
- 전진선택법 : direction = 'forward'
- 단계선택법 : direction = 'both'

```
Df Sum of Sq    RSS    AIC
- x3     1   0.1091 47.973 24.974
- x4     1   0.2470 48.111 25.011
- x2     1   2.9725 50.836 25.728
<none>          47.864 26.944
- x1     1  25.9509 73.815 30.576
```

```
Step: AIC=24.97
y ~ x1 + x2 + x4
```

```
Df Sum of Sq    RSS    AIC
<none>          47.97 24.974
- x4     1     9.93 57.90 25.420
- x2     1    26.79 74.76 28.742
- x1     1  820.91 868.88 60.629
```

최종 선택 설명 변수 : x1, x2, x4

Call:
lm(formula = y ~ x1 + x2 + x4, data = df)

Coefficients:

(Intercept)	x1	x2	x4
71.6483	1.4519	0.4161	-0.2365

```
In [24]: # '+' 연산자를 통한 튜플 간의 결합 가능
m = (1,2,3)
n = (4,5,6)

m + n
```

Out [24]: (1, 2, 3, 4, 5, 6)

```
In [27]: print(m[1])
m[1] = 888 # TypeError # tuple은 한번 입력하면 더 이상 배정이 안 된다. 고정된다.

2
```

```
TypeError                                         Traceback (most recent call last)
~\AppData\Local\Temp\ipykernel_10444\2415895351.py in <module>
      1 print(m[1])
----> 2 m[1] = 888
```

TypeError: 'tuple' object does not support item assignment

ADsP)

3. 다음 주성분 분석에 대한 설명 중 적절하지 않은 것은?

```
> data3 <- princomp(data1, cor=TRUE) # ISLR 패키지 data (Hitters)
> data3
Call :
princomp(x = data1, cor = TRUE)

Standard deviations:
          Comp.1        Comp.2        Comp.3        Comp.4        Comp.5        Comp.6        Comp.7
2.77339679 2.03026013 1.31485574 0.95454099 0.84109683 0.7237422 0.69841796

생략
17 variables and 263 observations.
> summary(data3)
Importance of components:

          Comp.1        Comp.2        Comp.3        Comp.4        Comp.5
Standard deviation 2.7733967 2.0302601 1.3148557 0.9575410 0.8410968
Proportion of Variance 0.4524547 0.2424680 0.1016968 0.0539344 0.04161435
Cumulative Proportion 0.4524547 0.6949227 0.7966195 0.8505539 0.89216822
```

- 1 80%이상 설명하려면 주성분 4개 이상 선택하면 된다.
- 2 제1성분의 설명력은 45%이다
- 3 공분산행렬을 활용한 결과이다.
- 4 차원을 2차원으로 줄이면 데이터 손실율은 약 30.51%이다
 - cor=TRUE 이므로 상관계수 행렬을 사용한 것이다

4. 다음 중 주성분분석의 해석으로 올바르지 않은 것은?

```
> data_1 <- prcomp(data, scale=TRUE)
> data_1
Standard deviations (1, ..., p=4):
[1] 1.4154072 1.3086525 0.4377899 0.3039594

Rotation (n * k) = (4 * 4)
          PC1         PC2         PC3         PC4
x1  0.2398128 -0.6895993  0.5325178  0.4287728
x2  0.4604720 -0.5393126 -0.5603653 -0.4278997
x3  0.6038420  0.3514805 -0.3277028  0.6359616
x4  0.6052345  0.3317472  0.5431634 -0.4781303

> summary(data_1)
Importance of components:
          PC1    PC2    PC3    PC4
Standard deviation 1.4154 1.3087 0.43779 0.3040
Proportion of Variance 0.5008 0.4281 0.04791 0.0231
Cumulative Proportion 0.5008 0.9290 0.97690 1.0000
```

- 1 제2변수 구하는 주성분함수식은 $-0.69*x1 + -0.54*x2 + 0.35*x3 + 0.33 * x4$ 이다
- 2 주성분 2개의 누적 기여율은 92.9%이다.
- 3 변수들의 scale이 많이 다른 경우 특정 변수가 전체적인 경향을 좌우하기 때문에 상관계수 행렬을 사용하여 분석하는 것이 좋다.
- 4 princomp(data, cor=TRUE)와 결과가 다르다
 - scale = TRUE 이므로 상관행렬을 사용한 것이다
 - princomp(data, cor=TRUE)와 결과가 같다

• 연산자

산술 연산자	1순위
비교(관계) 연산자	2순위
논리 연산자	3순위
대입 연산자(=)	4순위

파이썬에서는 0이 아니면 전부 True로 설정을 해놨다. True = 1

자격증 ADsP)

정상 시계열의 조건

- 평균은 모든 시점(시간 t)에 대해 일정하다
- 분산은 모든 시점(시간 t)에 대해 일정하다
- 공분산은 시점(시간 t)에 의존하지 않고, 단지 시차에만 의존한다

Eduatoz 교재 449페이지에서

4. 다음 시계열 자료의 정상성(stationary)에 대한 설명 중 가장 부적절한 것은?

- 1) 모든 시점에 대해 일정한 평균을 가진다.
- 2) 모든 시점에 대해 일정한 분산을 가진다.
- 3) 공분산은 단지 시차에만 의존하고 시점 자체에는 의존하지 않는다.
- 4) 모든 분산이 시점에 의존하지 않는다.

답) 4번에서 '모든 분산'이라고 했는데

그냥 '모든' 때문에 가장 부적절한 것으로 본 것으로 판단됩니다.

☞ 1. 다음 중 정상시계열에 대한 설명 중 가장 적절하지 않은 것은?

- 1 추세요인 : 자료의 그림을 그렸을 때 그 형태가 오르거나 내리는 등 자료가 어떤 특정한 형태를 취할 때
- 2 계절요인 : 고정된 주기에 따라 자료가 변화하는 경우
- 3 순환요인 : 물가상승률, 급격한 인구 증가 등의 이유로 주기를 가지고 변화하는 자료
- 4 불규칙요인 : 위 세 가지 요인으로 설명할 수 없는 요인에 의해 발생

순환요인 : 물가상승률, 급격한 인구 증가 등의 이유로 알려지지 않은 주기를 가지고 자료가 변화하는 경우

