PART 3. 데이터 분석 - 4장. 통계분석

데이터 준전문가

ADSP, Advanced Data Analytics semi-Professional

류영표 강사
ryp1662@gmail.com

강사 소개

류영표

Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료

現 Upstage AI X 네이버 부스트 캠프 AI tech 1~4기 멘토

前 Innovation on Quantum & CT(IQCT) 이사

前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)

前 ㈜셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 공공데이터 청년 인턴 / SW공개개발자대회 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의
- 이젠 종로 아카데미(파이썬, ADSP 강사) / 강남 : ADSP
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융사 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문위원
- 보건 · 바이오 AI활용 S/W개발 및 응용전문가 양성과정 강사
- Upstage AI X KT 융합기술원 기업교육 모델최적화 담당 조교

주요 프로젝트 및 기타사항

- 개인 맞춤형 당뇨병 예방·관리 인공지능 시스템 개발 및 고도화(안정화)
- 폐플라스틱 이미지 객체 검출 경진대회 3위
- 인공지능(AI)기반 데이터 사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

Test

수업시간에 다뤘던 문제들에 대한 출제이다.

객관식 8문제, 주관식 2문제

연습문제

1. 확률이란 “특정사건이 일어날 가능성의 척도”라고 정의할 수 있다. 통계적 실험을 실시할 때 나타날 수 있는 모든 결과들의 집합이라고 하고, 사건이란 표본공간의 부분집합을 말한다. 다음 중 확률 및 확률 분포에 대한 설명으로 가장 부적절한 것은?

① 모든 사건의 확률값은 0과 1사이에 있다.

② 서로 배반인 사건들의 합집합의 확률은 각 사건들의 확률의 합이다.

③ 두 사건 A, B가 독립이라면, 사건 B의 확률은 A가 일어난다는 가정하에서의 B의 조건부확률과 동일하다.

④ 확률변수 X가 구간 또는 구간들의 모임인 숫자 값을 갖는 확률분포함수를 이산형확률밀도함수라 한다.

연습문제

2. 자료의 정보를 이용해 집단에 관한 추측, 결론을 이끌어내는 과정인 통계적 추론에 대한 설명으로 가장 부적절한 것은?

① 전수조사가 불가능하면, 모집단에서 표본을 추출하고 표본을 근거로 확률론을 활용하여 모집단의 모수들에 대해 추론하는 것을 추정이라 한다.

② 점 추정은 표본의 정보로부터 모집단의 모수를 하나의 값으로 추정하는 것이다.

③ 통계적 추론은 제한된 표본을 바탕으로 모집단에 대한 일반적인 결론을 유도하려는 시도이므로 본질적으로 불확실성을 수반한다.

④ 구간추정은 모수의 참값이 포함되어 있다고 추정되는 구간을 결정하는 것이며, 실제 모집단의 모수는 신뢰구간에 포함되어야 한다.

연습문제

3. 모집단내에서 모집단의 특성을 잘 나타낼 수 있는 일부를 추출하여 이들로부터 자료를 수집하고 수집된 자료를 토대로 모집단의 특성을 추정하게 된다. 이 때 조사하는 모집단의 일부분을 표본(sample)이라 한다. 다음 중 표본조사에 대한 설명으로 가장 부적절한 것은?

① 표본 오차(Sampling error)는 모집단을 대표할 수 있는 표본 단위들이 조사대상으로 추출되지 못함으로서 발생하는 오차를 말한다.

② 표본편의(sampling bias)는 모수를 작게 또는 크게 할 때 추정하는 것과 같이 표본추출법에서 기인하는 오차를 의미한다.

③ 표본편의는 확률화(randomization)에 의해 최소화하거나 없앨 수 있다. 확률화란 모집단으로부터 편의되지 않은 표본을 추출하는 절차를 의미하며, 확률화 절차에 의해 추출된 표본을 확률표본(random sample)이라 한다.

④ 비표본오차(non-sampling error)는 표본오차를 제외한 모든 오차로 조사 과정에서 발생하는 모든 부주의나 실수, 알 수 없는 원인 등 모든 오차를 의미하며 조사대상이 증가한다고 해서 오차가 커지지는 않는다.

연습문제

4. 표본공간은 어떤 실험이나 시도의 결과로 나올 수 있는 모든 가능한 결과의 집합이다. 사건이랑 표본공간의 부분집합을 말한다. 다음 중 확률 및 확률분포에 관한 설명으로 부적절한 것은?

① (사건 A가 일어나는 경우의 수) / (일어날 수 있는 모든 경우의 수)를 P(A)라 할 때, 이를 A의 수학적 확률이라 한다.

② 한 사건 A가 일어날 확률을 P(A)라 할 때, n번의 반복시행에서 사건 A가 일어난 횟수를 r라하면, 상대도수는 r/n는 n이 커짐에 따라 확률 P(A)에 가까워짐을 알 수 있다. P(A)를 사건 A의 통계적 확률이라 한다.

③ 두 사건 A,B가 독립일 때, 사건 B의 확률은 A가 일어났다는 가정 하에서의 B의 조건부확률과는 다르다.

④ 표본공간 임의의 사건 A가 일어날 확률 P(A)는 항상 0과 1사이에 있다.

연습문제

5. 귀무가설이 사실인데도 불구하고 사실이 아니라고 판정할 때 (귀무가설을 기각하는 오류) 이를 제1종오류라고 한다. 이때 우리가 내린 판정이 잘못되었을 실제 확률은 무엇으로 나타낼 수 있느가?

① α(알파)

② p-value

③ 검정통계량

④ 1- α

연습문제

6. 다음 중 모분산의 추론에 대한 설명으로 가장 부적절한 것은?

① 모집단의 변동성 또는 퍼짐의 정도에 관심이 있는 경우, 모분산이 추론의 대상이 된다.

② 정규모집단으로부터 n개를 단순임의 추출한 표본의 분산은 자유도가 n-1인 t분포를 따른다.

③ 모집단이 정규분포를 따르지 않더라도 중심극한정리를 통해 정규모집단으로부터 모분산에 대한 검정을 유사하게 시행할 수 있다.

④ 이 표본에 의한 분산비 검정은 두 표본의 분산이 동일한지를 비교하는 검정으로 검정통계량은 F분포를 따른다.

연습문제

7. 통계적 추론이란 자료의 정보를 이용하여 모집단에 관한 추측이나 결론을 이끌어 내는 과정이다.

이 과정은 추정과 가설검정을 통하여 이루어진다. 다음 중 추정과 가설검정에 대한 설명으로 가장 부적절한 것은?

① 가장 참값이라고 여겨지는 하나의 모수 값을 택하는 것을 점추정이라고 한다. 즉, 점추정은 모수가 특정한 값일 것이라고 추정하는 것이다.

② 구간추정이란 일정한 크기의 신뢰구간으로 모수가 특정한 구간에 있을 것이라고 선언 하는 것으로 구해진 구간을 신뢰구간이라 한다.

③ 귀무가설이 사실일 때, 관측된 검정통계량의 값보다 귀무가설을 지지하는 방향으로 검정통계량이 나올 확률을 p값이라고 한다.

④ 검정력이란 대립가설이 맞을 때 그것을 받아들이는 확률을 의미한다.

연습문제

8. 표본조사나 실험을 실시하는 과정에서 추출된 원소들이나 실험 단위로부터 주어진 목적에 적합하도록 관측해 자료를 얻는 것을 측정(Measurement)라 한다. 다음 중 자료의 종류에 대한 설명으로 가장 부적절한 것은?

① 명목척도 - 측정 대상이 어느 집단에 속하는지 분류할 때 사용

② 순서척도 - 측정 대상의 특성이 가지는 서열관계를 관측하는 척도

③ 구간척도 - 측정 대상이 갖는 속성의 양을 측정하는 것으로 구간이나 구간사이의 간격이 의미가 있는 자료

④ 비율척도 - 절대적 기준인 원점이 존재하지 않으며, 모든 사칙연산이 가능한 척도

연습문제

9. 아래는 chickwts 데이터프레임을 분석한 것이다. 다음 중 결과에 대한 해석이 잘못된 것은?

```
> t.test(chickwts$weight)

        One Sample t-test

data:  chickwts$weight
t = 28.202, df = 70, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 242.8301 279.7896
sample estimates:
mean of x
 261.3099
```

① 전체 관측치 수는 70개이다.

② 99% 신뢰구간을 구하기 위해서는 “conf.level=0.99”라는 옵션을 사용할 수 있다.

③ 닭 무게의 점 추정량은 261.3이며, 95% 신뢰구간은 242.8에서 279.8이다.

④ 닭 무게에 대한 p-value는 p-value<2.2e-16이므로 귀무가설이 기각된다.

연습문제

10. Wage 데이터에서 wage에 대한 t-test를 실시하였다 다음 설명중 부적절한 것은?

```
> t.test(Wage$wage,mu=100)

        One Sample t-test

data:  Wage$wage
t = 15.362, df = 2999, p-value < 2.2e-16
alternative hypothesis: true mean is not equal to 100
95 percent confidence interval:
 110.2098 113.1974
sample estimates:
mean of x
 111.7036
```

① 한 집단의 평균에 대한 t-test(one sample t-test) 결과이다.

② 양측검정 결과를 보여주고 있다.

③ t-test의 자유도는 2999이다.

④ 평균에 대한 95% 신뢰구간은 귀무가설에서 설정한 평균의 참값을 포함한다.

연습문제

11. Carseats 데이터프레임은 400개 상점에서 판매 중인 유아용 카시트에 대한 자료이다. 이 데이터의 일부 변수들의 상관분석 결과로 가장 부적절한 것은?

```
> rcorr(as.matrix(Carseats[,c(1:6,8)]),type='pearson')
            Sales CompPrice Income Advertising Population Price   Age
Sales        1.00      0.06   0.15        0.27       0.05 -0.44 -0.23
CompPrice    0.06      1.00  -0.08       -0.02      -0.09  0.58 -0.10
Income       0.15     -0.08   1.00        0.06      -0.01 -0.06  0.00
Advertising  0.27     -0.02   0.06        1.00       0.27  0.04  0.00
Population   0.05     -0.09  -0.01        0.27       1.00 -0.01 -0.04
Price       -0.44      0.58  -0.06        0.04      -0.01  1.00 -0.10
Age         -0.23     -0.10   0.00        0.00      -0.04 -0.10  1.00

n= 400


P
            Sales  CompPrice Income Advertising Population Price  Age
Sales               0.2009    0.0023 0.0000      0.3140     0.0000 0.0000
CompPrice   0.2009            0.1073 0.6294      0.0584     0.0000 0.0451
Income      0.0023 0.1073           0.2391      0.8752     0.2579 0.9258
Advertising 0.0000 0.6294    0.2391              0.0000     0.3743 0.9276
Population  0.3140 0.0584    0.8752 0.0000                  0.8087 0.3948
Price       0.0000 0.0000    0.2579 0.3743      0.8087            0.0411
Age         0.0000 0.0451    0.9258 0.9276      0.3948     0.0411
```

① Sales와 CompPrice 간의 상관계수는 유의하지 않다.

② Sales와 가장 강한 상관관계를 보이는 변수는 Price이다.

③ Price가 올라갈수록 Sales는 낮아지는 경향이 있다.

④ Sales와 Price는 양의 선형관계를 갖는다.

연습문제

12. 다중 회귀분석에서 가장 적합한 회귀모형을 찾기 위한 과정의 설명으로 가장 부적절한 것은?

① 독립변수의 수가 많아지면 모델의 설명력이 증가하지만 모형이 복잡해지고, 독립변수들 간에 서로 영향을 미치는 다중공선성의 문제가 발생하므로 상대적인 조정이 필요하다.

② 회귀식에 대한 검정은 독립변수의 기울기(회귀계수)가 0이 아니라는 가정을 귀무가설, 기울기가 0인 것을 대립가설로 놓는다.

③ 잔차의 독립성, 등분산성 그리고 정규성을 만족하는지 확인해야 한다.

④ 회귀분석의 가설검정에서 p값이 0.05보다 작은 값이 나와야 통계적으로 유의한 결과로 받아들일 수 있다.

연습문제

13. Default 데이터셋은 10,000명의 신용카드 고객에 대한 연체여부(default:1-default, 0-not default), 카드대금 납입 후 남은 평균 카드잔고(balance), 연봉(income)을 토함하고 있다. 아래는 연체 가능성을 95% 신뢰수준으로 모형화한 결과이다. 다음 설명이 부적절한 것은 무엇인가?

```
> summary(glm(default~.,family=binomial(),data=Default))

Call:
glm(formula = default ~ ., family = binomial(), data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

① 로지스틱 회귀모형이 적합 결과이다.

② Balance는 default를 설명하는 데 통계적으로 유의하다.

③ Balance가 높을수록 default 가능성이 높다.

④ Income이 높을수록 default 가능성이 낮다.

연습문제

14. 회귀분석에서 결정계수(R^2)에 한 설명으로 부적절한 것은?

① 총 변동 중에서 설명이 되지 않는 오차에 의한 변동이 차지하는 비율이다.

② 회귀모형에서 입력 변수가 증가하면 결정계수도 증가한다.

③ 다중 회귀분석에서는 최적 모형의 선정기준으로 결정계수 값보다는 수정된 결정계수 값을 사용하는 것이 적절하다.

④ 수정된 결정계수는 유의하지 않는 독립변수들이 회귀식에 포함되었을 때 그 값이 감소한다.

연습문제

15. 다음 중 데이터의 정규성을 확인하기 위한 방법으로 부적절한 것은?

① 히스토그램

② Q-Q plot

③ Shapiro-Wilks test

④ Durbin Watson test

연습문제

16. 최적방정식을 선택하기 위한 방법 중 모든 독립변수 후보를 포함한 모형에서 시작하여 가장 적은 영향을 주는 변수를 하나씩 제거하면서 더 이상 유의하지 않는 변수가 없을 때까지 설명변수를 제거하는 방법은 무엇인가?

연습문제

17. 아래는 스위스의 47개 프랑스어 사용지역의 출산율(Fertility)과 교육수준(Education)과의 관계를 회귀 모형으로 추정한 것이다. 아래의 결과를 사용하여 결정계수(R^2)을 계산하시오.

```
> out=lm(Fertility~Education,data=swiss)
> anova(out)
Analysis of Variance Table

Response: Fertility
           Df Sum Sq Mean Sq F value   Pr(>F)
Education   1 3162.7  3162.7  35.446 3.659e-07 ***
Residuals  45 4015.2    89.2
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

연습문제

18. 다음 시계열분석에서 정상성의 특징이 아닌 것은?

① 평균이 일정하다. 즉, 모든 시점에 대해 일정한 평균을 가진다.

② 분산도 시점에 의존하지 않는다.

③ 자기회귀식에는 백색잡음이 없다.

④ 공분산은 단지 시차에만 의존하고 실제 어느 시점 t,s에는 의존하지 않는다.

연습문제

19. 시계열을 구성하는 4가지 요소에 해당되지 않은 것은?

① 계절요인

② 교호요인

③ 순환요인

④ 추세요인

연습문제

20. Data는 메이저리그에서 활약하는 263명의 선수에 대한 타자 기록으로 연봉(Salary)을 비롯한 17개 변수를 포함하고 있다. 아래는 17개의 변수들을 사용하여 주성분분석을 시행한 결과이다. 다음 설명 중 잘못된 것은?

```
> pca=princomp(data,cor=TRUE)
> summary(pca)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5     Comp.6     Comp.7     Comp.8     Comp.9
Standard deviation     2.7733967 2.0302601 1.3148557 0.9575410 0.84109683 0.72374220 0.69841796 0.50090065 0.42525940
Proportion of Variance 0.4524547 0.2424680 0.1016968 0.0539344 0.04161435 0.03081193 0.02869339 0.01475891 0.01063797
Cumulative Proportion  0.4524547 0.6949227 0.7966195 0.8505539 0.89216822 0.92298014 0.95167354 0.96643244 0.97707042
                          Comp.10     Comp.11     Comp.12     Comp.13     Comp.14      Comp.15      Comp.16
Standard deviation     0.363901982 0.312011679 0.243641510 0.232044829 0.163510472 0.1186398422 0.0693395039
Proportion of Variance 0.007789685 0.005726546 0.003491834 0.003167341 0.001572687 0.0008279654 0.0002828216
Cumulative Proportion  0.984860104 0.990586651 0.994078485 0.997245826 0.998818513 0.9996464785 0.9999293001
                          Comp.17
Standard deviation     3.466841e-02
Proportion of Variance 7.069994e-05
Cumulative Proportion  1.000000e+00
```

① 최소 80% 이상의 분산설명력을 갖기 위해서는 4개 이상의 주성분을 사용해야 한다.

② 가장 큰 분산설명력을 가지는 주성분은 전체 분산의 45.25%를 설명한다.

③ 공분산행렬을 사용하여 주성분분석을 시행한 것이다.

④ 17차원을 2차원으로 축소한다면 잃게 되는 정보량은 약 30.5%이다.

연습문제

21. 교차분석은 2개 이상의 변수를 결합하여 자료의 빈도를 살펴보는 기법이다. 다음 중 교차분석에 대한 설명으로 부적절한 것은 무엇인가?

① 범주의 관찰도수에 비교될 수 있는 기대도수를 기대한다.

② 교차분석은 두 문항 모두 범주형 변수가 아니어도 사용할 수 있으며, 두 변수 간 관계를 보기 위해 실시한다.

③ 교차분석은 교차표를 작성하여 교차빈도를 집계할 뿐 아니라 두 변수들 간의 독립성 검정을 할 수 있다.

④ 기대빈도가 5미만인 셀의 비율이 20%를 넘으면 카이제곱분포에 근사하지 않으며, 이런 경우 표본의 크기를 늘리거나 변수의 수준을 합쳐 셀의 수를 줄이는 방법 등을 사용한다.

연습문제

22. 시계열의 요소분해법은 시계열 자료가 몇 가지 변동들의 결합으로 이루어져 있다고 보고 변동요소별로 분해하여 쉽게 분석하기 위한 것이다. 다음 중 분해 요소에 대한 설명이 부적절한 것은?

① 추세 분석은 장기적으로 변해가는 큰 흐름을 나타내는 것으로 자료가 장기적으로 커지거나 작아지는 변화를 나타내는 요소이다.

② 계절변동은 일정한 주기를 가지고 반복적으로 같은 패턴을 보이는 변화를 나타내는 요소이다.

③ 순환변동은 경제 전반이나 특정 산업의 부침을 나타내 주는 것을 말한다.

④ 불규칙변동은 불규칙하게 변동하는 급격한 환경변화, 천재기변 같은 것으로 발생하는 변동을 말한다.

연습문제

23. 다음 중 아래 주성분 분석을 시행한 결과에 대한 설명으로 가장 부적합한 것은?

```
> college_s<-scale(college)
> summary(college_s)
    Outstate          Room.Board          Books             Personal          Grad.Rate       
 Min.   :-2.0136   Min.   :-2.3503   Min.   :-2.7460   Min.   :-1.6108   Min.   :-3.22880  
 1st Qu.:-0.7757   1st Qu.:-0.6935   1st Qu.:-0.4808   1st Qu.:-0.7247   1st Qu.:-0.72555  
 Median :-0.1120   Median :-0.1436   Median :-0.2991   Median :-0.2077   Median :-0.02697  
 Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.0000   Mean   : 0.00000  
 3rd Qu.: 0.6175   3rd Qu.: 0.6314   3rd Qu.: 0.3066   3rd Qu.: 0.5308   3rd Qu.: 0.72982  
 Max.   : 2.7987   Max.   : 3.4344   Max.   :10.8453   Max.   : 8.0632   Max.   : 3.05842  
> fit<-princomp(college_s)
> fit$loadings

Loadings:
           Comp.1 Comp.2 Comp.3 Comp.4 Comp.5
Outstate    0.587         0.155  0.142  0.779
Room.Board  0.531  0.230  0.155  0.574 -0.557
Books              0.812 -0.561 -0.153       
Personal   -0.329  0.532  0.776              
Grad.Rate   0.514         0.187 -0.789 -0.279
```

① 두 번째 주성분은 0.230*Room.Board+0.812*Books+0.532*Personal로 계산된다.

② 두 번째 주성분에 가장 큰 영향을 미치는 원변수는 Books이다.

③ Personal 값이 클수록 첫 번째 주성분은 작아진다.

④ fit<-princomp(college, cor=T)의 결과는 위의 결과와 다르다.

연습문제

24. 아래 주성분분석의 결과에서 두 개의 주성분을 사용할 때 설명 가능한 전체 분산의 비율은?

```
> model<-princomp(Car)
> summary(model)
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
Standard deviation        1.503     1.075     0.840     0.752     0.555
Proportion of Variance    0.453     0.231     0.141     0.113     0.061
Cumulative Proportion     0.453     0.684     0.825     0.938     1.000
```

Test

복원문제에 다뤘던 문제들에 대한 출제이다.

객관식 8문제, 주관식 2문제

연습문제

1. 측정대상이 갖고 있는 속성의 양을 측정하는 것으로 측정결과가 숫자로 표현되나 해당 속성이 전혀 없는 상태인 절대적인 영점이 없어 두 관측 값 사이의 비율은 별 의미가 없게 된다. 온도, 지수, 등이 해당되는 이 척도는 무엇인가?

① 명목척도

② 순서척도

③ 구간척도

④ 비율척도

연습문제

2. 다음 중 아래의 R코드를 수행한 결과에 대한 설명으로 옳은 것은?

```
> c(2,4,6,8)+c(1,3,5,7,8)
```

① 경고 메시지와 함께 결과가 출력된다.

② 4개의 숫자로 이루어진 벡터가 출력된다.

③ 9개의 숫자로 이루어진 벡터가 출력된다.

④ 에러 메시지가 출력되고, 명령 수행이 중단된다.

연습문제

3. 다음 다중회귀분석을 위해 사용되는 변수선택방법에 대한 설명 중 변수선택방법에 대한 설명 중 변수선택방법과 설명이 잘못 연결되어 있는 것은??

① 전진선택법(forward selection)은 상수항만 포함된 모형에서 출발하여 설명력이 좋은 변수를 하나씩 추가하는 방법이다.

② 단계적 방법(stepwise method)은 설명력이 나쁜 변수를 제거하거나 모형에서 제외된 변수 중 모형의 설명력을 가장 잘 개선하는 변수를 추가하는 방법이다.

③ 후진제거법(backward elimination)은 모든 변수가 포함된 모형에서 설명력이 나쁜 변수를 하나씩 제거하는 방법이다.

④ 최적선택법(optimum selection)은 전진선택법과 후진제거법을 결합한 방법으로 회귀식이 최적의 변수를 선택하는 방법이다.

연습문제

4. 이상치를 찾는 것은 데이터 분석에서 데이터 전처리를 어떻게 할지 검정할 때 사용할 수 있다.

다음 중 상자그림을 이용하여 이상치를 판정하는 방법에 대한 설명으로 가장 부적절한 것은?

① IQR = Q3-Q1이라고 할 때, Q1-1.5*IQR < x < Q3+1.5*IQR을 벗어나는 x를 이상치라고 규정한다.

② 평균으로부터 3*표준편차 범위를 벗어나는 것들을 비정상이라 규정하고 제거한다.

③ 이상치는 변수의 분포에서 벗어난 값으로 상자 그림을 통해 확인할 수 있다.

④ 이상치는 분포를 왜곡할 수 있으나 실제 오류인자인지에 대해서는 통계적으로 판단하지 못하기 때문에 제거여부는 실무자를 통해 결정하는 것이 바람직하다.

연습문제

5. 통계분석에서 자료를 수집하고 그 수집된 자료로부터 어떤 정보를 얻고자 하는 경우에는 항상 수립된 자료가 특정한 확률분포를 따른다고 가정한다. 다음 중 연속형 확률분포가 아닌 것은?

① 이항분포

② 정규분포

③ T분포

④ F분포

연습문제

6. 다음 표본 추출 방법에 관한 설명 중 잘못된 것은 무엇인가?

① 표본의 크기를 결정할 때 가장 중요한 부분은 표본이 모집단을 얼마나 설명하는지에 대한 대표성의 확보이다.

② 단순랜덤추출법은 모집단에서 샘플을 뽑을 때, 각각의 샘플이 모두 동등한 확률을 가지고 무작위로 추출되는 방법이다.

③ 계통추출법은 모집단으로 군집으로 구분하고 선정된 군집의 원소를 모두 샘플로 추출하는 다단계 추출 방법이다.

④ 층화추출법은 모집단을 몇 개의 집단으로 구분하고, 각 집단의 크기와 분산을 고려하여 각 집단마다 샘플을 추출하는 방법이다.

연습문제

7. 다음 중 비모수검정이 아닌 것을 고르시오.

① 윌콕슨의 순위합검증

② 만-위트니의 U검정

③ 스피어만의 순위상관계수

④ 자기상관검증

연습문제

8. 두 변량 X,Y의 상관분석에 관한 내용이다. 설명이 옳지 않은 것은?

① 등간척도로 측정된 두 변수간의 상관관계는 피어슨 상관계수(Pearson correlation)를 통해 확인할 수 있다.

② 상관계수가 0이면 두 변량 X,Y 사이에 선형관계가 없다.

③ 서열척도로 측정된 두 변수간의 상관관계는 스피어만 상관계수(Spearman Correlation)를 통해 확인할 수 있다.

④ R에서 상관계수를 구하기 위해서는 rcor()함수를 사용하면 되고 type인자를 통해 피어슨과 스피어만 상관계수를 선택할 수 있다.

연습문제

9. 다음 중 회귀분석에서 나온 결정계수(R^2)에 대한 설명으로 옳지 않은 것은?

① 총 제곱의 합 중 설명된 제곱의 합의 비율을 뜻한다.

② 종속변수에 미치는 영향이 적은 독립변수가 추가된다면 결정계수는 변하지 않는다.

③ R^2의 값이 클수록 회귀선으로 실제 관측치를 예측하는데 정확성이 높아진다.

④ 독립변수와 종속변수 간의 표본상관계수 r의 제곱값과 같다.

연습문제

10. 다음 시계열 분석의 기초가 되는 개념인 정상성(Stationarity)의 특징에 관한 설명이다. 설명이 옳지 않은 것은?

① 평균이 일정하다. 즉 모든 시점에 대한 일정한 평균을 가진다.

② 시계열 분석에서 비정상 시계열 자료는 시계열 분석을 할 수 없다.

③ 분산도 시점에 의존하지 않는다.

④ 공분산은 단지 시차에만 의존하고 실제 어느 시점 t,s에는 의존하지 않는다.

연습문제

11. 시계열에 관한 설명 중 틀린 것은?

① 대부분의 시계열은 비정상 자료이다. 그러므로 비정상 자료를 정상성 조건에 만족시켜 정상시계열로 만든 후 시계열 분석을 한다.

② 시계열이 정상 시계열인지 비정상 시계열인지 판단하기 위해 폭발적인 추세를 보이거나 시간에 따라 분산이 변화하는지 관찰해야 한다.

③ 비정상 시계열은 정상 시계열로 변경하고자 할 때, 변환과 차분의 방법을 사용한다.

④ 일반적으로 평균이 일정하지 않은 비정상 시계열은 변환을 통해, 분산이 일정하지 않은 비정상 시계열은 차분을 통해 정상 시계열로 바꾼다.

연습문제

12. 다음 중 중심극한정리(Central Limit Theorem)에 대한 설명으로 가장 부적절한 것은?

① 여러 통계적 방법론에는 정규데이터가 필요하지만 중심극한 정리를 사용하면 비정규적인 모집단에도 이와 유사한 절차를 적용할 수 있다.

② 표본평균의 분포는 표본의 크기가 커짐에 따라 정규분포로 근사한다.

③ 모집단의 분포가 정규분포에 가까워져야 표본평균의 분포가 정규분포로 근사하게 된다.

④ 모집단의 분포가 대칭이면 표본의 크기가 작아도 되지만 모집단의 분포가 비대칭이면 표본의 크기가 30이상이 되어야 한다.

연습문제

13. 다음은 데이터의 척도에 관한 설명이다. 설명이 틀린 것은?

① 명목척도는 측정 대상이 어느 집단에 속하는지 분류할 때 사용되며 성별, 출생지 정보가 해당된다.

② 순서척도는 측정 대상이 순서를 갖는 자료를 의미하며, 만족도, 선호도, 학년, 신용등급 정보가 해당된다.

③ 구간척도는 측정 대상의 순서와 순서 사이의 간격이 의미가 있는 자료를 의미하며, 온도, 물가지수, 주가지수 정보가 해당된다.

④ 비율척도는 측정대상의 값이 비율로 정의되는 자료를 의미하며, 물가성장율, 흡연감소율의 정보가 해당된다.

연습문제

14. 분해시계열에 대한 설명 중 잘못된 것은?

① 분해시계열이란 시계열에 영향을 주는 일반적인 요인을 시계열에서 분리해 분석하는 방법을 말한다.

② 분해 시계열의 분해 요소는 추세요인, 계절요인, 순환요인, 회귀요인으로 크게 4가지로 이루어진다.

③ 추세요인은 자료의 형태가 오르거나 내리는 추세를 따르는 경우로 선형적 형태, 지수형태 등이 있다.

④ 순환요인은 경제적이나 자연적인 이유가 없이 알려지지 않은 주기를 가지고 변화하는 자료형태이다.

연습문제

15. 두 개 이상의 독립변수를 사용해 하나의 종속변수의 변화를 설명하는 다중회귀분석을 실시할 것이다. 다음 중 모형을 적합 시킨 후, 모형이 적절한지 확인하기 위해 체크해야 할 사항으로 부적절한 것은?

① 상관계수를 통해 모형의 설명력을 확인한다.

② F-value를 통해 모형이 통계적으로 유의한지 확인한다.

③ 모형이 데이터에 잘 적합되어 있는지를 확인한다.

④ t-value, p-value를 통해 유의한지 확인한다.

연습문제

16. 주성분분석은 차원의 단순화를 통해 서로 상관되어 있는 변수들 간의 복잡한 구조를 분석하는 것이 목적이다. 다음 중 주성분분석에 대한 설명으로 적절하지 않은 것은 무엇인가?

① 다변량 자료를 저차원의 그래프로 표시하여 이상치(Outlier) 탐색에 사용한다.

② 변수들끼리 상관성이 있는 경우, 해석상의 복잡한 구조적 문제가 발생하는데 이를 해결하기 위해 사용한다.

③ 회귀분석에서 다중공선성(Multicollinearity)의 문제를 해결하기 위해 활용한다.

④ P개의 변수들을 중요한 m(p)개의 주성분으로 표현하여 전체 변동을 설명하는 것으로 m개의 주성분은 원래 변수와의 관계없이 생성된 변수들이다.

연습문제

17. 아래는 데이터프레임 mtcars를 이용해 회귀분석을 수행한 R 명령의 결과이다. 다음 중 이 결과에 대한 설명으로 가장 부적절한 것은?

```
> summary(lm(mpg~., data=mtcars))

Call:
lm(formula = mpg ~ ., data = mtcars)

Residuals:
    Min      1Q  Median      3Q     Max
-3.4506 -1.6044 -0.1196  1.2193  4.6271

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12.30337   18.71788   0.657   0.5181
cyl         -0.11144    1.04502  -0.107   0.9161
disp         0.01334    0.01786   0.747   0.4635
hp          -0.02148    0.02177  -0.987   0.3350
drat         0.78711    1.63537   0.481   0.6353
wt          -3.71530    1.89441  -1.961   0.0633
qsec         0.82104    0.73084   1.123   0.2739
vs           0.31776    2.10451   0.151   0.8814
am           2.52023    2.05665   1.225   0.2340
gear         0.65541    1.49326   0.439   0.6652
carb        -0.19942    0.82875  -0.241   0.8122
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.65 on 21 degrees of freedom
Multiple R-squared:  0.869,     Adjusted R-squared:  0.8066
F-statistic: 13.93 on 10 and 21 DF,  p-value: 3.793e-07
```

① 오차의 표준편차 추정치는 2.65이다.

② 모든 독립변수가 유의수준 0.1에서 유의하지 않다.

③ 후진제거법을 적용할 때 가장 먼저 제거될 독립변수는 cyl이다.

④ 유의수준 0.01하에서 이 회귀모형은 유의하다.

연습문제

18. 데이터마이닝을 위한 데이터 분할에 대한 설명으로 틀린 것은 어느 것인가?

① 데이터를 구축용(training), 검증용(Validation), 시험용(test)으로 분리한다.

② 일반적으로 데이터 구축용, 검정용, 시험용 데이터는 50%, 30%, 20%로 정한다.

③ 데이터가 충분하지 않을 때는 구축용과 시험용 데이터만 구분하여 활용한다.

④ 통계학에 적용되는 교차확인(Cross-validation)은 데이터마이닝에서 활용할 수 없다.

연습문제

19. Default 데이터는 10,000명의 신용카드 고객에 대한 체납 여부(default)와 학생여부(student), 카드 잔고(balance), 연봉(income)을 포함하고 있다. 고객의 체납 확률을 예측하기 위한 아래 결과에 대한 설명으로 가장 부적절한 것은?

```
> summary(glm(default~.,data=Default,family="binomial"))

Call:
glm(formula = default ~ ., family = "binomial", data = Default)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4691  -0.1418  -0.0557  -0.0203   3.7383

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
income       3.033e-06  8.203e-06   0.370  0.71152
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2920.6  on 9999  degrees of freedom
Residual deviance: 1571.5  on 9996  degrees of freedom
AIC: 1579.5

Number of Fisher Scoring iterations: 8
```

① 로지스틱 회귀모형을 사용한 결과이다.

② 카드 잔고와 연봉이 동일한 수준일 때, 학생(studentYes)이 학생이 아닌 고객보다 체납확률이 낮다.

③ 세 설명변수 모두 체납확률을 예측하는데 유익한 영향이 있다.

④ 동일한 신분과 연봉 수준일 때 카드 잔고가 높을수록 체날 확률이 높다.

연습문제

20. 다음 중 비모수적 방법에 대한 설명으로 가장 부적절한 것은?

① 관측된 자료가 주어진 분포를 따른다는 가정을 받아들일 수 없을 때 이용하는 검정법이다.

② 자료가 추출된 모집단의 분포에 대해 제약을 가하지 않고, 검정을 실시하는 방법이다.

③ 관측된 자료가 구한 표본평균과 표본분산 등을 이용해 검정을 실시한다.

④ 관측된 자료가 특정 분포를 따른다고 가정할 수 없을 때 이용한다.

연습문제

21. 다음 중 연속형 확률 변수의 분포 중 정규분포로부터 유도되었으며, 정규 분포의 평균을 측정할 때 주로 사용되는 분포로 두 집단의 평균 차이 검증을 등에 활용되는 분포는?

① 균일 분포(uniform distribution)

② 지수분포(exponential distribution)

③ t-분포(t-distribution)

④ F-분포(F-distribution)

연습문제

22. 다음 중 비모수 검정 방법 중 하나로 표본들이 서로 관련되어 있는 경우 짝지어진 두 개의 관찰치들의 크고 작음을 표시하여 그 개수를 가지고 두 분포의 차이가 있는지에 대한 가설을 검증하는 방법은?

① 런 검정(run test)

② 만-위트니의 U검정

③ 부호 검정(sign test)

④ 스피어만 순위상관계수

연습문제

23. 다음 중 소득 수준과 같이 정규 분포를 따르지 않고 오른쪽 꼬리가 긴(right-skewed)분포를 나타내는 자료의 평균과 중앙값의 관계로 옳은 것은 무엇인가?

① 자료의 크기(scale)에 따라 달라진다.

② 평균이 중앙값보다 작은 경향을 보인다.

③ 평균이 중앙값이 일치하는 경향을 보인다.

④ 평균이 중앙값보다 큰 경향을 보인다.

연습문제

24. 다음 중 변수를 단조 증가 함수로 변환하여 다른 변수를 나타낼 수 있는 정도를 나타내며 두 변수의 선형 관계의 크기 뿐만 아니라 비선형적인 관계도 나타낼 수 있는 상관계수는 무엇인가?

① 코사인 유사도

② 피어슨 상관계수

③ 스피어만 상관계수

④ 자카드 인덱스

연습문제

25. 여러 대상 간의 거리가 주어졌을 때, 대상들을 동일한 상대적 거리를 가진 실수 공간의 점들로 배치시키는 방법을 무엇이라 하는가?

연습문제

26. 회귀 모형의 가정 중 잔차항이 정규분포를 이루어야 하는 가정을 의미하는 용어는 무엇인가?

연습문제

27. 데이터의 한 부분으로 특정 사용자가 관심을 갖고 있는 데이터를 담은 비교적 작은 규모의 데이터 웨어하우스는 무엇이라고 하는가?

① 데이터베이스

② 데이터 마트

③ 데이터 마이닝

④ 데이터 프레임

연습문제

28. 측정 대상이 어느 집단에 속하는지 분류할 때 사용되는 척도로, 성별(남, 여) 구분, 출생지(서울특별시, 부산광역시, 경기도 등) 구분 등을 할 때 사용되는 척도는?

① 명목척도

② 순서척도

③ 구간척도

④ 비율척도

연습문제

29. 다음 중 회귀 모형에 사용되는 독립 변수 간의 상관관계가 존재하여 회귀 계수 추정치가 불안하고 해석하기 어려워지는 현상을 나타내는 것은?

① 다중공선성

② 등분산성

③ 정상성

④ 독립성

연습문제

30. 데이터 분석 시 원 데이터는 불완전한 내용을 담고 있는 경우가 많다. 데이터 전처리는 이를 제거하거나 보정하여 데이터의 품질을 높이는 작업이라 할 수 있다. 데이터 전처리 작업 중 이상치(Outlier) 검색은 분석에서 전처리를 어떻게 할지 결정할 때 사용한다. 다음 이상치 판정 방법 중 가장 부적절한 것은?

① 3-Sigma 방법은 “평균으로부터 표준편차의 3배가 넘는 범위의 데이터“ 를 비정상이라 규정한다.

② 회귀분석 적합 후 잔차분석을 실시하여 이상치를 판정하는 방법이 있다.

③ Q2(중위수)+1.5*IQR 보다 크거나 작은 데이터를 이상치로 규정한다.

④ 통계 모형에 기반한 방법으로는 Grubb’s Test, Hotelling’s T2 test 등이 있다.

연습문제

31. 다음 중 lasso 회귀모형에 대한 설명으로 부적절한 것은?

① 모형에 포함된 회귀계수들의 절대값의 크기가 클수록 Penalty를 부여하는 방법이다.

② 자동적으로 변수선택을 하는 효과가 있다.

③ Lambda 값으로 penalty의 정도를 조정한다.

④ L2 penalty 를 사용한다.

연습문제

32. 다음은 데이터의 회귀분석 결과이다. 다음 설명 중 옳지 않은 것을 고르시오.

아래

```
> summary(lm(wage~.,wage))

Call:
lm(formula = wage ~ ., data = wage)

Residuals:
    Min      1Q  Median      3Q     Max
-112.31  -19.94   -3.09   15.33  222.56

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                84.104      2.231  37.695  < 2e-16 ***
education2. HS Grad        11.679      2.520   4.634 3.74e-06 ***
education3. Some College   23.651      2.652   8.920  < 2e-16 ***
education4. College Grad   40.323      2.632  15.322  < 2e-16 ***
education5. Advanced Degree 66.813     2.848  23.462  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 36.53 on 2995 degrees of freedom
Multiple R-squared:  0.2348,    Adjusted R-squared:  0.2338
F-statistic: 229.8 on 4 and 2995 DF,  p-value: < 2.2e-16
```

① 잔차의 IQR은 35.27이다.

② education의 더미변수가 4개이다.

③ education5 그룹의 임금이 가장 높다.

④ 모든 변수가 wage를 설명하는데 유의하게 나타나게 있다.

연습문제

33. 분해 시계열의 요인에 해당하지 않는 것은?

① 정상 요인

② 추세요인

③ 계절요인

④ 불규칙요인

연습문제

34. 다음 중 가설검정과 관련된 내용 중 가장 부적절한 것은?

① 귀무가설을 기각시키는 검정통계량들의 범위를 기각역이라 한다.

② 가설을 검정하기 위한 기준으로 사용하는 값을 검정통계량이라 한다.

③ 현재까지 주장되어 온 것이나 변화나 차이가 없음을 설명하는 가설을 귀무가설이라 한다.

④ P-value 값이 미리 정해놓은 유의수준(alpha)값보다 클 경우, 귀무가설을 기각하므로 대립가설의 가정이 옳다고 할 수 있다.

연습문제

35. 통계적 추론에서 모집단의 모수를 검증하기 위해 사용하는 모수적 방법과 비교하여 비모수적 방법의 특징으로 가장 부적절한 것은?

① 비모수적 검정은 모집단의 분포에 대해 아무런 제약을 가하지 않는다.

② 관측된 자료가 특정 분포를 따른다고 가정할 수 없는 경우에 이용된다.

③ 분포의 모수에 대한 가설을 설정하지 않고 분포의 형태에 대해 가설을 설정한다.

④ 비모수 검정에서는 관측값의 절대적 크기에 의존하여 평균, 분산 등을 이용해 검정을 실시한다.

연습문제

36. 다음 정상시계열에 대한 설명 중 적절하지 않은 것은?

① 대부분의 시계열은 비정상 자료이다. 그러므로 비정상 자료를 정상성 조건에 만족시켜 정상시계열로 만든 후 시계열 분석을 한다.

② 시계열이 정상 시계열인지 비정상 시계열인지 판단하기 위해 폭발전인 추세를 보이거나 시간에 따라 분산이 변화하는지 관찰해야 한다.

③ 비정상 시계열은 정상 시계열로 변경하고자 할 때 변환과 차분의 방법을 사용한다.

④ 일반적으로 평균이 일정하지 않은 비정상 시계열은 변환을 통해, 분산이 일정하지 않은 비정상 시계열은 차분을 통해 정상 시계열로 바꾼다.


Thank you.
ADSP / 류영표 강사
ryp1662@gmail.com