다변량 통계분석 (Multivariate Statistical Analysis)

정여진

강의개요

- 강의시간: 토 9:00-10:30
- 교재: R을 활용한 응용 다변량분석 입문, Brian and Hothorn 지음, 김재희, 국광호 옮김, 교우사
- 강의노트는 가상대학 사이트에서 다운받아 출력
- 평가방법
  - ✓중간고사: 30%
  - ✓기말고사: 40%
  - ✓과제: 20%
  - ✓출석 및 기타: 10%

다변량 데이터와 다변량 분석

다변량 데이터란?

- 연구자가 관심을 갖는 많은 주체, 대상, 개체에 대해 여러 개의 확률변수 값(변수)을 기록할 때 발생
  - ✓심리학자들은 많은 피실험자에 대해 여러가지 인지변수들에 대한 측정값 기록
  - ✓교육 연구자들은 다양한 과목에 대해 학생들의 시험점수에 관심
  - ✓고고학자들은 관심있는 유물들에 대해 여러 가지 값을 측정
- 행(row): 특정 개체의 변수값
- 열(column): 특정 변수의 관측치

단위	변수1	…	변수q
1	x_{11}	$\cdots$	x_{1q}
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n	x_{n1}	$\cdots$	x_{nq}

- $n = 10, q = 7$
- NA: 결측치

```
> hypo
   individual    sex age  IQ depression    health weight
1           1   Male  21 120        Yes Very good    150
2           2   Male  43  NA         No Very good    160
3           3   Male  22 135         No   Average    135
4           4   Male  86 150         No Very poor    140
5           5   Male  60  92        Yes      Good    110
6           6 Female  16 130        Yes      Good    110
7           7 Female  NA 150        Yes Very good    120
8           8 Female  43  NA        Yes   Average    120
9           9 Female  22  84         No   Average    105
10         10 Female  80  70         No      Good    100
```

변수의 유형

- 명목(nominal)척도
  - ✓순서화 되지 않은 범주형 변수
  - ✓예: 성별, 머리색, 우울증의 존재 여부
- 순서(ordinal)척도
  - ✓순서는 존재하지만 척도의 서로 다른 값의 차이가 동일한 거리를 의미 하지 않음
  - ✓예: 사회적 계급, 건강의 자아 인식(5점 척도), 교육수준(초등이하, 중등, 고등 이상)
- 구간(interval)척도
  - ✓연속적인 점들 사이에 동일한 차이가 있음
  - ✓0의 위치는 임의적
  - ✓예: 섭씨, 화씨 온도
- 비(ratio)척도
  - ✓가장 높은 수준의 척도
  - ✓점수의 상대적 크기 조사 가능. 0의 위치 고정
  - ✓예: 나이, 무게, 길이

결측값 (missing value)

- 설문의 무응답, 경시적 데이터에서의 중도탈락, 설문 거절
- complete-case analysis
  - ✓결측값이 없는 unit만 포함하여 분석
  - ✓많은 양의 정보를 버림
  - ✓제외된 사례가 random한 subsample이 아닌 경우 추론의 bias 심각
- Available-case analysis
  - ✓관심있는 값 추정을 위해 이용가능한 모든 사례 이용
  - ✓예: 상관관계 추정을 위해 X_i와 X_j가 존재하는 모든 사례 이용
  - ✓표본이 상관관계를 추정할 때마다 변함
  - ✓추정된 상관행렬이 positive definite이라는 보장 없음

- Multiple imputation
  - ✓결측값을 대체
  - ✓Rubin(1987)에 의해 제안되어 가장 적절하다고 여겨지는 방법
  - ✓하나의 missing value에 대해 몬테카를로 방법으로 여러 값을 imputation
  - ✓결과를 적절한 방법으로 결합

Example: Measure Data

```
> measure
   chest waist hips gender
1     34    30   32   male
2     37    32   37   male
3     38    30   36   male
4     36    33   39   male
5     38    29   33   male
6     43    32   38   male
7     40    33   42   male
8     38    30   40   male
9     40    30   37   male
10    41    32   39   male
11    36    24   35 female
12    36    25   37 female
13    34    24   37 female
14    33    22   34 female
15    36    26   38 female
16    37    26   37 female
17    34    25   38 female
18    36    26   37 female
19    38    28   40 female
20    35    23   35 female
```

- 남자와 여자의 20명 표본에서 각 개인별 가슴, 허리, 엉덩이둘레에 대한 측정값을 기록
- Questions
  - ✓신체크기와 신체모양이 어떤 방식에 의해 세 개의 관측값이 결합되어 한 개의 숫자로 요약될 수 있을까?
    ➔ **주성분분석**
  - ✓남자들 사이에 그리고 여자들 사이에 각 집단 내에서는 유사한 신체모양을 가지고 그 집단 사이에서는 다른 모양을 갖는 신체모양의 하위 유형이 있을까?
    ➔ **군집분석**

Example: Pottery Data

```
> head(pottery,10)
   Al2O3 Fe2O3  MgO  CaO Na2O  K2O TiO2   MnO   BaO kiln
1   18.8  9.52 2.00 0.79 0.40 3.20 1.01 0.077 0.015    1
2   16.9  7.33 1.65 0.84 0.40 3.05 0.99 0.067 0.018    1
3   18.2  7.64 1.82 0.77 0.40 3.07 0.98 0.087 0.014    1
4   16.9  7.29 1.56 0.76 0.40 3.05 1.00 0.063 0.019    1
5   17.8  7.24 1.83 0.92 0.43 3.12 0.93 0.061 0.019    1
6   18.8  7.45 2.06 0.87 0.25 3.26 0.98 0.072 0.017    1
7   16.5  7.05 1.81 1.73 0.33 3.20 0.95 0.066 0.019    1
8   18.0  7.42 2.06 1.00 0.28 3.37 0.96 0.072 0.017    1
9   15.8  7.15 1.62 0.71 0.38 3.25 0.93 0.062 0.017    1
10  14.6  6.87 1.67 0.76 0.33 3.06 0.91 0.055 0.012    1
```

- 세 개의 다른 지역에서 만들어진 도자기에 대한 화학적 분석 결과
  - ✓지역 1: kiln(가마)=1
  - ✓지역 2: kiln=2,3
  - ✓지역 3: kiln=4,5
- Questions
  - ✓ 각 항아리의 화학적 프로필이 서로 다른 유형의 항아리를 암시하는가?

    ➔ **군집분석**

  - ✓그러한 유형이 가마 또는 지역과 관련이 있는가?

    ➔ **군집분석, 판별분석**

Example: Exam Data

```
> exam
  subject maths english history geography chemistry physics
1       1    60      70      75        58        53      42
2       2    80      65      66        75        70      76
3       3    53      60      50        48        45      43
4       4    85      79      71        77        68      79
5       5    45      80      80        84        44      46
```

- 심리학과 학생들의 6개 과목에서의 시험점수
- Question
  - ✓시험점수가 "일반적인 지능"과 같이 직접 측정할 수 없는 학생들의 내재하는 특성을 보여주는가?
    - ➔ **탐색적 인자분석**

Example: US Air Pollution Data

```
> head(USairpollution,10)
           SO2 temp manu popul wind precip predays
Albany      46 47.6   44   116  8.8  33.36     135
Albuquerque 11 56.8   46   244  8.9   7.77      58
Atlanta     24 61.5  368   497  9.1  48.34     115
Baltimore   47 55.0  625   905  9.6  41.31     111
Buffalo     11 47.1  391   463 12.4  36.11     166
Charleston  31 55.2   35    71  6.5  40.75     148
Chicago    110 50.6 3344  3369 10.4  34.44     122
Cincinnati  23 54.0  462   453  7.1  39.04     132
Cleveland   65 49.7 1007   751 10.9  34.99     155
Columbus    26 51.5  266   540  8.6  37.01     134
```

- 미국 도시들의 대기오염 연구 자료
- Questions
  - ✓ 이산화황(SO_2) 농도에 의해 측정된 오염 수준이 6개의 다른 변수와 어떤 관련이 있을까?

  ➔ **다중회귀분석**

(아주 약간의)

통계학 기초

공분산

- 두 확률변수의 선형 의존성에 대한 측도

$$s_{ij} = \frac{\sum_{k=1}^{n}(X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{n-1}$$

- X_i와 X_j가 서로 독립이면 공분산=0
- 공분산이 클 수록 두 변수 간의 선형 의존성의 정도가 크다
- q개의 변수

공분산행렬: $S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1q} \\ s_{21} & s_2^2 & \cdots & s_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ s_{q1} & s_{q2} & \cdots & s_q^2 \end{pmatrix}$

공분산 행렬: measure data

- 전체 데이터의 가슴, 허리, 엉덩이 둘레에 대한 공분산 행렬

```
> cov(measure[,1:3])
          chest     waist     hips
chest 6.631579  6.368421 3.000000
waist 6.368421 12.526316 3.578947
hips  3.000000  3.578947 5.944737
```

- 남자

```
> cov(measure[measure$gender=="male",1:3])
          chest     waist     hips
chest 6.7222222 0.9444444 3.944444
waist 0.9444444 2.1000000 3.077778
hips  3.9444444 3.0777778 9.344444
```

- 여자

```
> cov(measure[measure$gender=="female",1:3])
          chest    waist     hips
chest 2.277778 2.166667 1.555556
waist 2.166667 2.988889 2.755556
hips  1.555556 2.755556 3.066667
```

상관관계

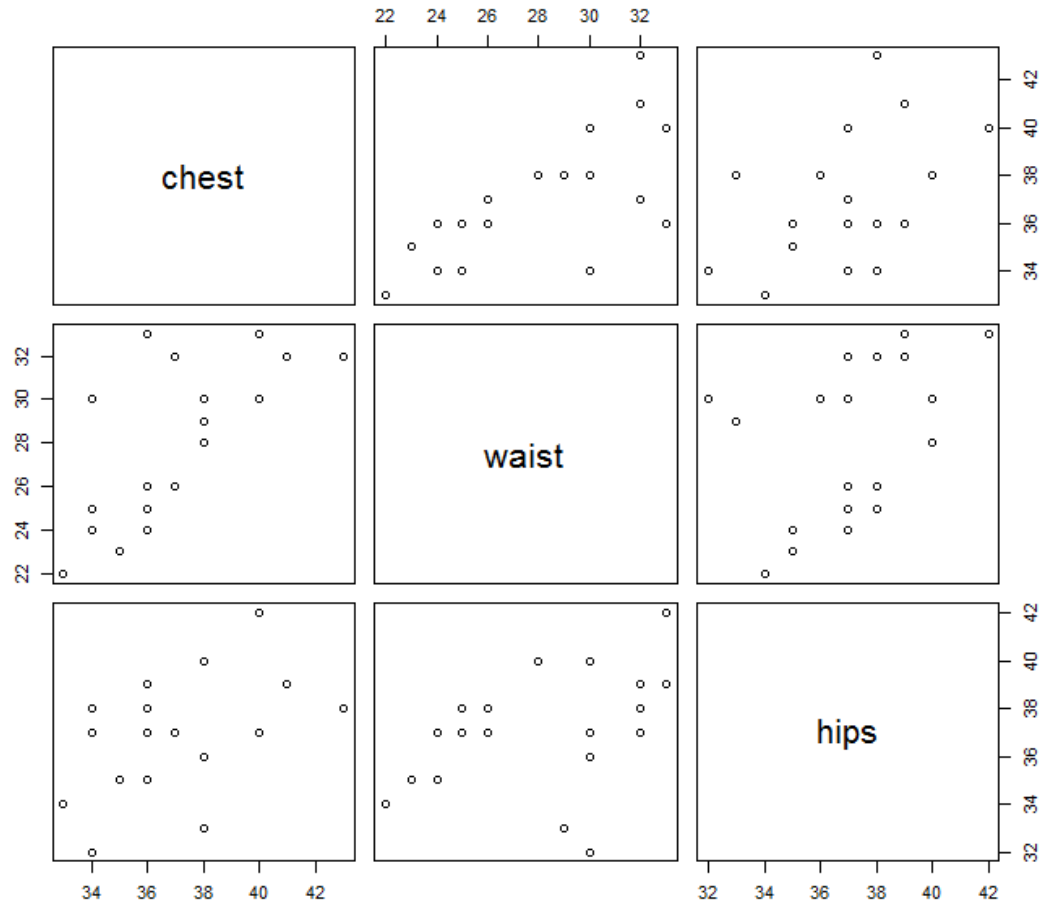
- 공분산은 두 변수를 측정하는 척도(m/cm, g/kg)에 의존하기 때문에 때로는 해석하기 어려움
- 상관계수

$$r_{ij} = \frac{s_{ij}}{s_i s_j}$$

✓1과 -1 사이의 값
✓척도에 독립적
✓두 변수의 관계가 선형이 아니면 오해를 불러일으킬 수 있음

상관계수 행렬: measure data

```
> cor(measure[,1:3])
          chest     waist      hips
chest 1.0000000 0.6987336 0.4778004
waist 0.6987336 1.0000000 0.4147413
hips  0.4778004 0.4147413 1.0000000
```



거리

- 데이터 내에서 개체들 사이의 거리를 측정
- 유클리드 거리 (Euclidean distance)

$$d_{ij} = \sqrt{\sum_{k=1}^{q} \left(x_{ik} - x_{jk}\right)^2}$$

- 데이터셋 내의 변수들이 서로 다른 척도를 갖는 경우 각 변수에 대해 표준화한 후 거리를 계산하는 것이 나음

```
> round(dist(scale(measure[,1:3])),2)
      1    2    3    4    5    6    7    8    9   10   11   12   13   14   15   16   17   18   19
2  2.43
3  2.26 0.80
4  3.09 0.95 1.68
5  1.63 1.89 1.26 2.82
6  4.31 2.37 2.18 2.76 2.95
7  4.79 2.38 2.72 1.98 3.94 2.03
8  3.63 1.41 1.64 1.22 2.88 2.18 1.41
9  3.10 1.29 0.88 1.95 1.84 1.36 2.22 1.46
10 3.99 1.76 1.79 1.96 2.85 0.88 1.32 1.36 1.07
11 2.23 2.44 1.91 3.03 1.81 3.74 4.14 2.77 2.44 3.40
12 2.61 2.02 1.66 2.40 2.14 3.39 3.42 2.03 2.10 2.89 0.87
13 2.66 2.54 2.34 2.78 2.66 4.18 4.01 2.61 2.88 3.63 1.13 0.83
14 2.44 3.45 3.09 3.90 2.80 5.07 5.27 3.86 3.74 4.67 1.36 1.89 1.41
15 2.82 1.79 1.60 2.02 2.35 3.20 3.00 1.60 1.96 2.61 1.35 0.50 1.04 2.31
16 2.62 1.70 1.26 2.18 1.89 2.91 3.08 1.72 1.62 2.44 1.07 0.48 1.29 2.28 0.56
17 2.84 2.33 2.25 2.43 2.81 4.02 3.64 2.25 2.76 3.39 1.48 0.88 0.50 1.89 0.83 1.27
18 2.47 1.74 1.43 2.14 2.00 3.23 3.24 1.84 1.92 2.70 1.00 0.28 0.96 2.04 0.41 0.39 0.92
19 3.67 1.72 1.74 1.66 2.88 2.39 1.81 0.57 1.56 1.67 2.47 1.68 2.28 3.56 1.26 1.41 1.95 1.56
20 2.36 2.78 2.33 3.29 2.21 4.20 4.47 3.08 2.89 3.82 0.48 1.07 0.95 0.92 1.54 1.41 1.41 1.24 2.75
```

다변량 정규분포

- q 개의 변수 벡터 $\boldsymbol{x^T} = (x_1, x_2, \dots, x_q)$에 대한 다변량 정규확률밀도함수

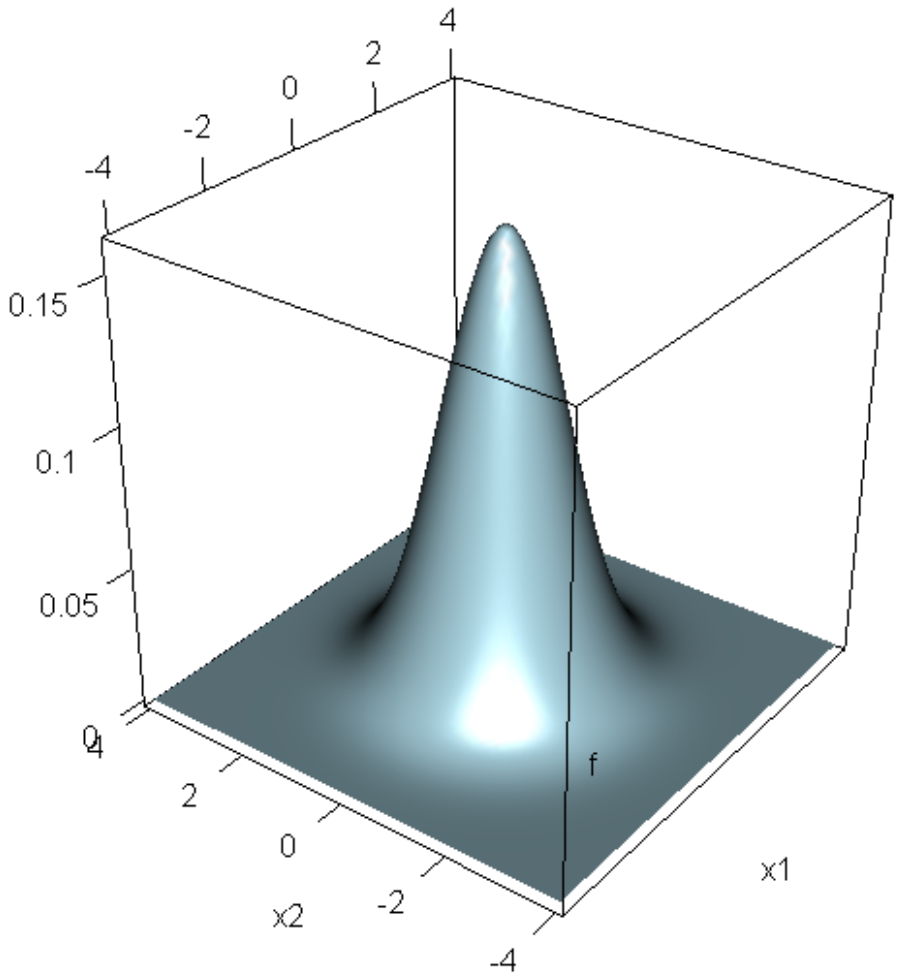
$$f(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = (2\pi)^{-q/2} \det(\Sigma)^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^{\boldsymbol{T}} \boldsymbol{\Sigma^{-1}} (\boldsymbol{x} - \boldsymbol{\mu})\right\}$$

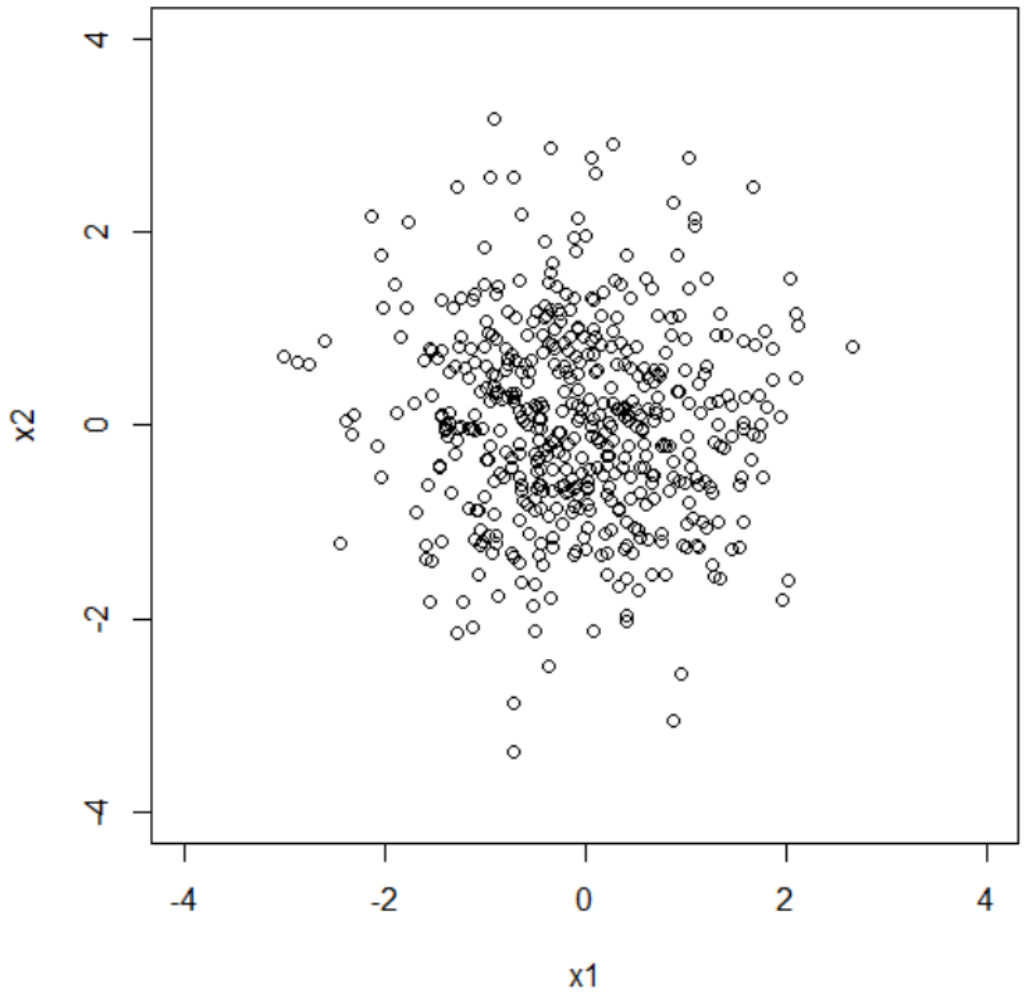
- q가 2인 경우

$$f((x_1, x_2))$$

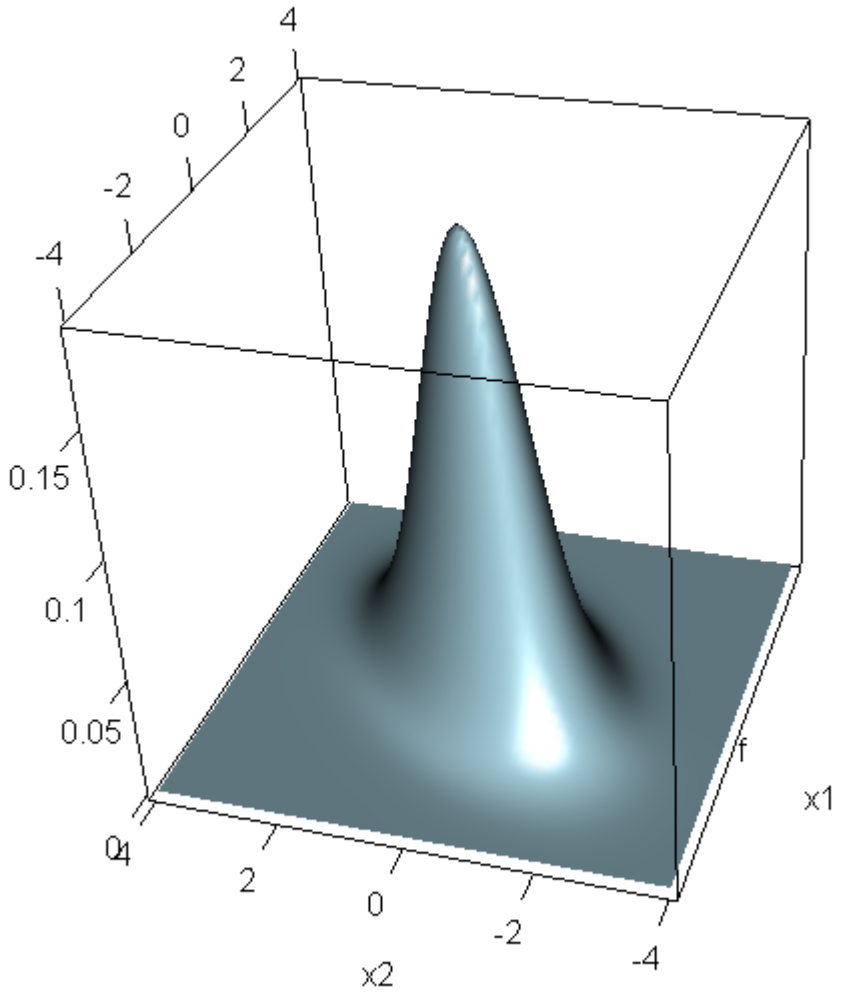
$$= \left(2\pi\sigma_1\sigma_2(1-\rho^2)\right)^{-1/2} \exp\left\{-\frac{1}{2(1-\rho^2)}\left(\left(\frac{x_1-\mu_1}{\sigma_1}\right)^2 - 2\rho\frac{x_1-\mu_1}{\sigma_1}\frac{x_2-\mu_2}{\sigma_2} + \left(\frac{x_2-\mu_2}{\sigma}\right)^2\right)\right\}$$

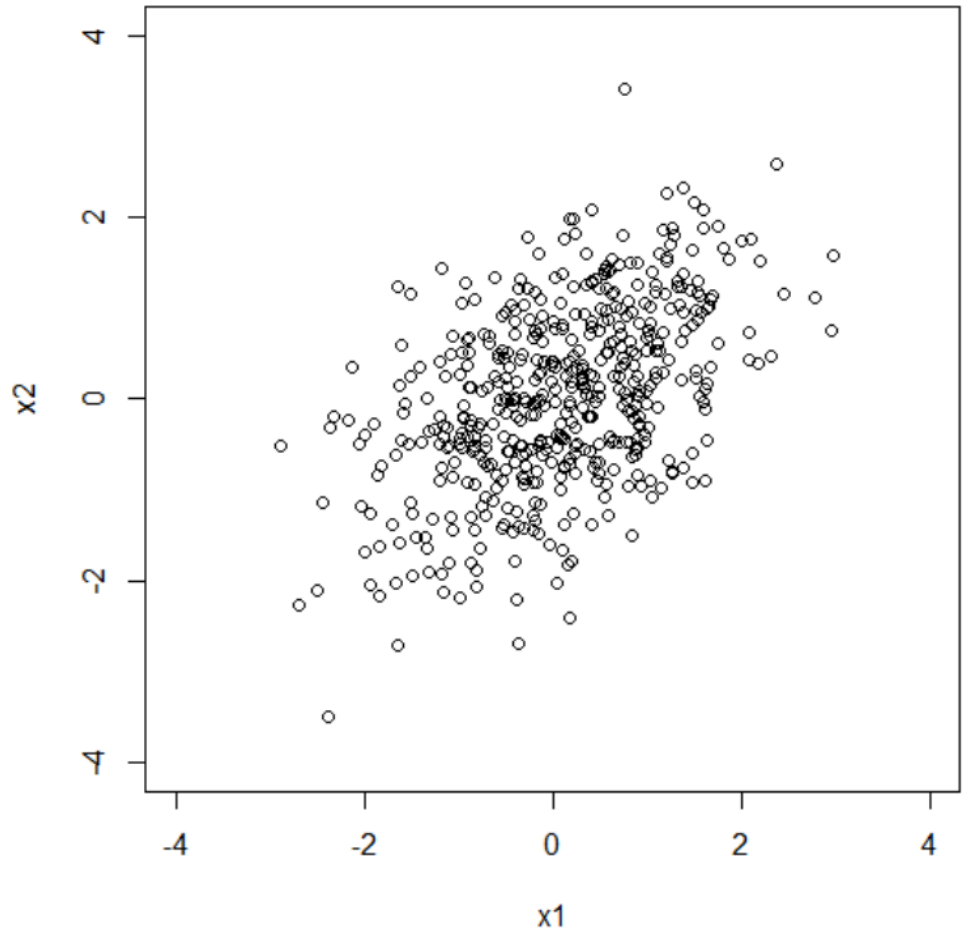
$$(\mu_1, \mu_2) = (0,0), \quad \sigma_1 = \sigma_2 = 1, \quad \sigma_{12} = \sigma_{21} = 0$$





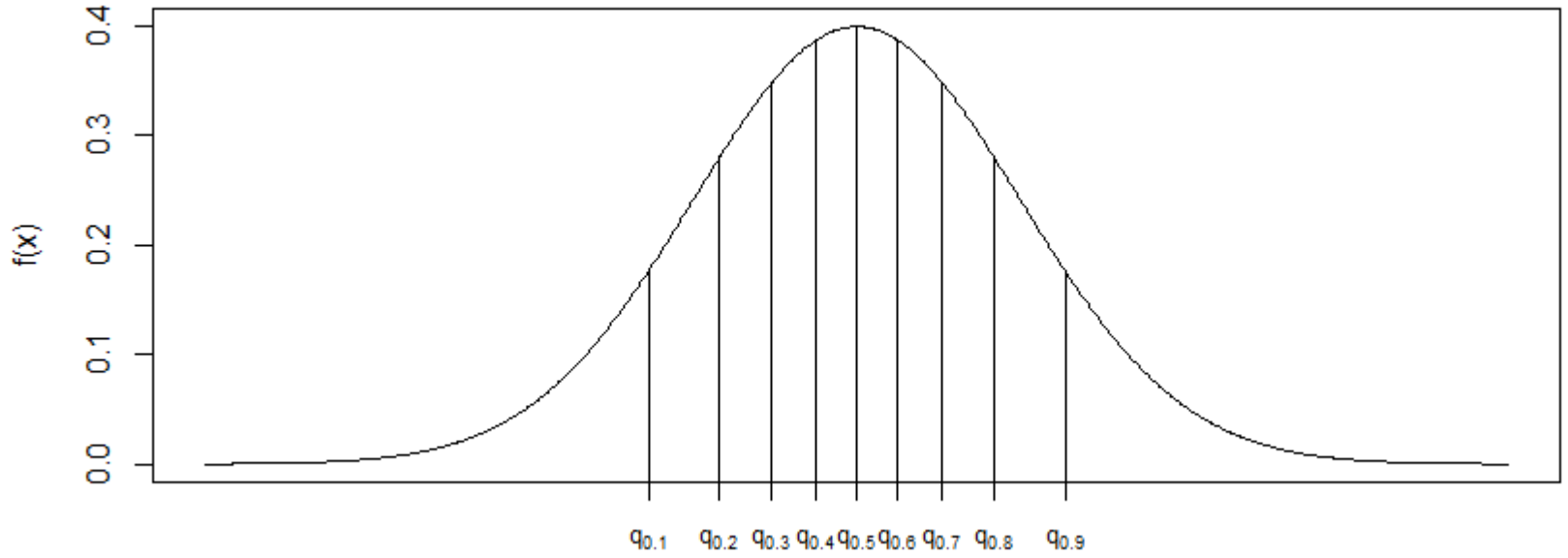
$$(\mu_1, \mu_2) = (0,0), \quad \sigma_1 = \sigma_2 = 1, \quad \sigma_{12} = \sigma_{21} = 0.5$$

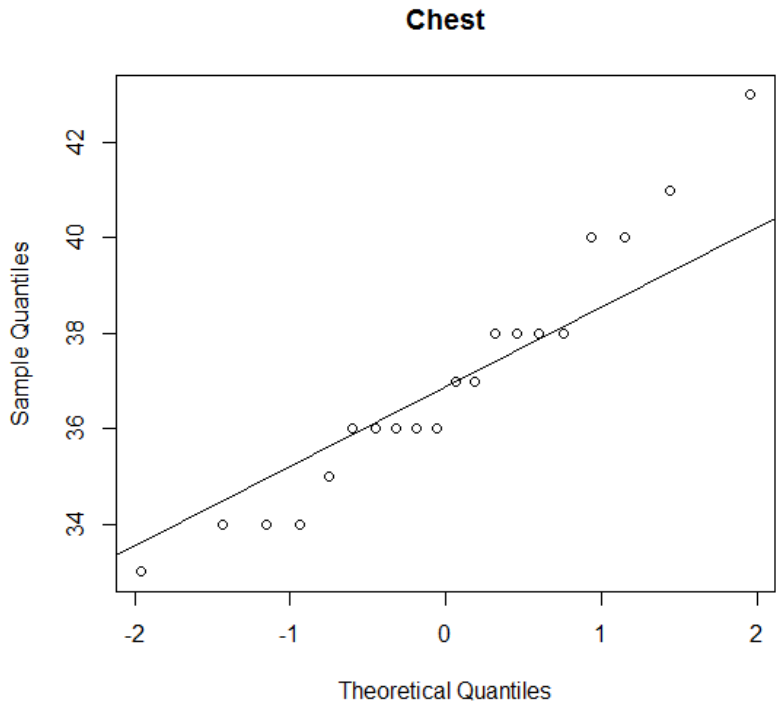


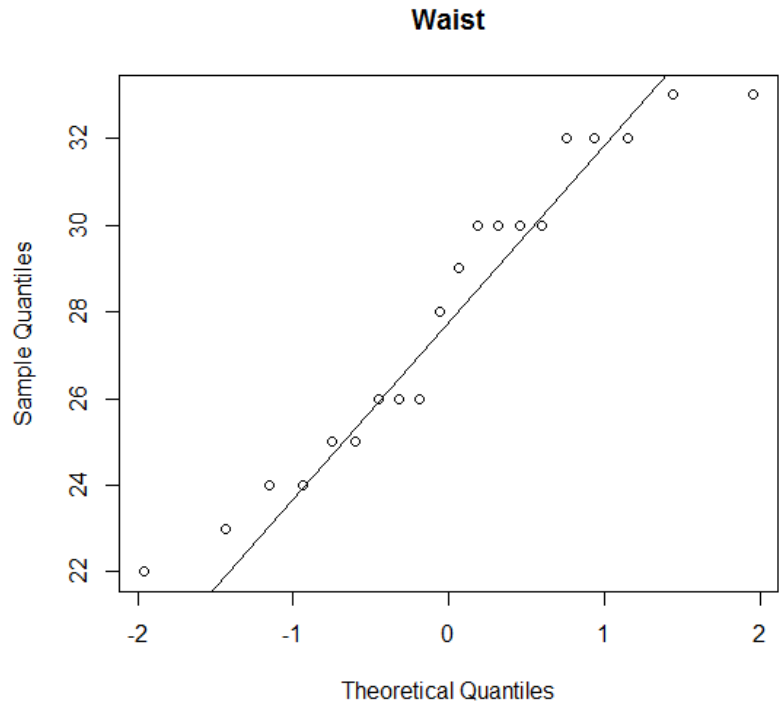


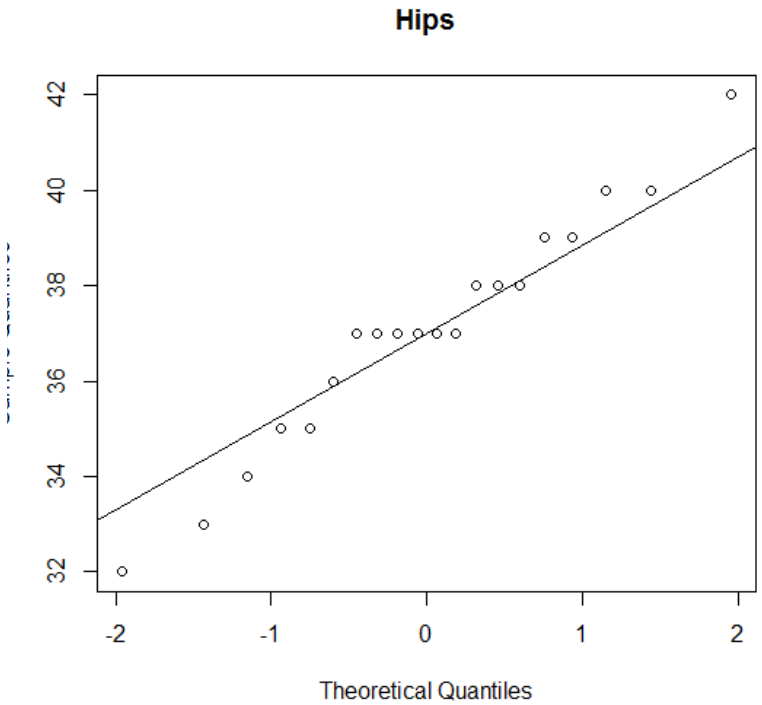
정규성 검정

- 각 변수의 정규성을 Q-Q plot을 통해 검정
  - ✓주의: 각 변수가 정규분포를 따른다고 해서 다변량 정규분포를 따르는 것은 아님 (역은 성립)
- q-분위수 (q-quantile)
  - ✓누적확률이 k/q가 되는 값들 (k=1,…,q)
- 정규 Q-Q plot
  - ✓정규분포의 q-분위수와 자료의 q-분위수를 비교한 산점도
  - ✓점들이 직선 상에 있으면 정규분포를 따른다고 판단










```
> qqnorm(measure$chest, main="Chest"); qqline(measure$chest)
> qqnorm(measure$waist, main="Waist"); qqline(measure$waist)
> qqnorm(measure$hips, main="Hips"); qqline(measure$hips)
```