회귀모형 이론과 계산

이용희

2020년 9월 10일

차 례

서론

이 책은 가장 기초적인 선형회귀모형부터 일반화 선형모형, 혼합모형과 같은 복잡한 회귀모형 대한 이론을 가능도 함수 위주로 설명합니다. 또한 모형을 적합하는 계산법과 연관된 행렬이론에 대하여 다루고자 합니다.

이 책에서 사용된 기호와 표기법은 다음과 같습니다.

- 스칼라(scalar)와 확률변수는 모두 소문자 또는 대문자인 보통 글씨체로 표기한다.
- 다변량 확률벡터와 행렬은 굵은 글자체로 표기한다.

```
library(DT)
library(ggplot2)
library(xfun)
library(JuliaCall)
# 아래 3 문장은 한글을 포함한 ggplot 그림이 포함된 HTML, PDF로 만드는 경우 사용
library(showtext)
font_add_google("Nanum Pen Script", "gl")
showtext_auto()
```

이 책에서 사용된 기호와 표기법은 다음과 같습니다.

- 스칼라(scalar)와 확률변수는 모두 소문자 또는 대문자인 보통 글씨체로 표기한다.
- 다변량 확률벡터와 행렬은 굵은 글자체로 표기한다.

참고문헌은 Xie (2015) 이렇게

제 1 장

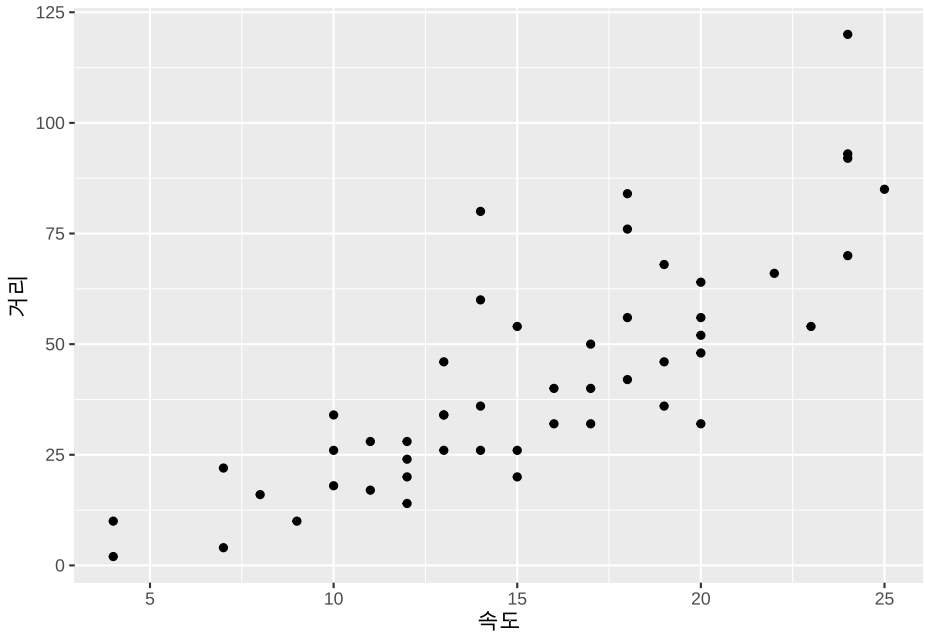
회귀모형에서의 기초 추정법

1.1 단순선형회귀분석

1.1.1 예제: 자동차의 속도와 제동거리

자동차가 달리는 속도(speed,mph)와 제동거리(dist,ft)의 관계를 알아보기 위하여 50대의 자동차로 실험한 결과 자료와 산포도가 아래와 같다.

```
ggplot(cars, aes(x=speed, y=dist)) +
  geom_point() +
  labs(x = "속도", y = "거리")
```



위와 같은 자료를 이용하여 자동차의 속도가 주어졌을 경우 제동거리의 평균에 대한 예측을 하려고 한다면 어떤 방법을 사용해야 할까? 회귀분석(regression analysis)는 여러 가지 변수들의 관계를 분석하는 통계적 방법이다. 일반적으로 한 개 또는 여러 가지의 설명변수들(explanatory variables)이 관심있는 종속변수(response variable)에 어떤 형태로 영향을 미치는지에 파악하고 설명변수와 종속변수의 관계를 통계적으로 추론하는 것이 회귀분석의 목적이다.

자동차의 속도를 x 라고 하고 제동거리를 y 라고 하면 다음과 같은 선형식으로 자동차의 속도와 제동거리의 관계를 나타내는 것을 단순선형회귀식(simple linear regression equation)이라고 한다.

$$E(y|x) = \beta_0 + \beta_1 x \tag{1.1}$$

식 (1.1)은 y의 평균이 x 의 선형식으로 나타나는 관계를 가정한 것이며 절편 β_0 와 기울기 β_1은 모르는 모수로서 자료를 통하여 추정해야 한다.

위에서 본 cars 예제와 같이 n 개의 자료를 독립적으로 추출하였다면 자료의 생성 과정을 다음과 같은 선형회귀모형(linear regression model)로 나타낸다. 종속변수 y는 설명변수 x의 선형식으로 나타내어지는 결정적인 요인과 확률 변수로 나타내어지는 임의의 오차항

e의 합으로 나타내어진다.

$$y_i = E(y_i|x_i) + e_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \dots, n \tag{1.2}$$

위에서 오차항 e_i 평균이 0이고 분산이 σ^2 인 임의의 확률분포를 따르며 서로 독립이다.

$$E(e_i) = 0, \quad V(e_i) = \sigma^2 \quad i = 1, 2, \dots, n$$

1.1.2 최소제곱법

단순회귀식 (1.2)에서 모수 β_0와 β_1를 회귀계수(regression coefficient) 라고 하며 자료(observation; data)를 수집하여 추정해야 한다. n개의 자료를 이용하여 회귀계수 β_0와 β_1를 추정하려고 할 때 가장 쉽고 오래됬으며 또한 가장 유용한 방법인 최소제곱법(least square method)을 사용할 수 있다. 일단 위의 식 (1.2)에서 종속변수의 관측값 y_i을 대응하는 설명변수 $x = x_i$를 이용하여 예측한 값은 $\beta_0 + \beta_1 x_i$이다. 여기서 실제 관측하여 얻어진 값 y_i와 예측값 $\beta_0 + \beta_1 x_i$사이에는 차이가 발생한다. 그 차이를 잔차(residual)라고 하며 표현하면 다음과 같다.

$$r_i = y_i - E(y_i|x_i) = y_i - (\beta_0 + \beta_1 x_i)$$

잔차는 위에 식에서 알 수 있듯이 관측값과 회귀식을 통한 예측값의 차이를 나타낸 것이다. 그러면 자료를 가장 잘 설명할 수 있는 회귀직선을 얻기 위해서는 잔차 r_i를 가장 작게하는 회귀모형을 세워야 한다. 잔차들을 최소로 하는 방법들 중 하나인 최소제곱법은 잔차들의 제곱합을 최소로 하는 회귀계수를 추정하는 방법이다. 잔차들의 제곱합은 다음과 같이 표현된다.

$$S(\beta_0, \beta_1) = \sum_{i=1}^{n} r_i^2 = \sum_{i=1}^{n} [y_i - (\beta_0 + \beta_1 x_i)]^2 \tag{1.3}$$

위의 잔차 제곱합 $S(\beta_0, \beta_1)$ 을 최소화하는 β_0와 β_1의 값을 구하는 방법은 잔차 제곱합이 β_0와 β_1의 미분 가능한 2차 함수이고 아래로 볼록한 함수(convex function)임을 이용한다. 각각의 회귀계수에 대해서 편미분을 하고 0으로 놓으면 아래와 같이 정리된다.

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_0} = \sum_{i=1}^{n}(-2)[y_i - (\beta_0 + \beta_1 x_i)] = 0 \tag{1.4}$$

$$\frac{\partial S(\beta_0,\beta_1)}{\partial \beta_1} = \sum_{i=1}^{n}(-2x_i)[y_i - (\beta_0 + \beta_1 x_i)] = 0 \tag{1.5}$$

위의 연립방정식을 행렬식으로 표시하면 다음과 같이 나타낼 수 있다.

$$\begin{bmatrix} n & \sum_i x_i \\ \sum_i x_i & \sum_i x_i^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} \sum_i y_i \\ \sum_i x_i y_i \end{bmatrix}$$

위의 방정식을 풀어서 구한 회귀계수의 추정치를 $\hat{\beta}_0$, $\hat{\beta}_1$ 이라고 하면 다음과 같이 주어진다.

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2}$$

1.2 회귀식의 행렬형식

일반적으로 회귀모형에서 종속변수의 수는 하나인 경우가 많지만 설명변수의 수는 여러 개인 경우가 많다. 이런 경우 중선형회귀식(multiple linear regression)은 다음과 같이 표현할 수 있고, p개의 설명변수가 있다고 가정하고 $(x_1, x_2, \cdots, x_p)$ 표본의 크기 n인 자료가 얻어지면 선형회귀식을 행렬로 다음과 같이 표현할 수 있다.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + e_i \tag{1.6}$$

$$= \mathbf{x}_i^t \boldsymbol{\beta} + e_i \tag{1.7}$$

위의 식을 다시 표현하면 다음과 같이 쓸 수 있다.

$$y_i = \begin{bmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + e_i$$

n개의 관측치가 있을때 n개의 회귀식을 행렬식으로 표현하면 다음과 같다.

$$\begin{bmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \vdots \\ \mathrm{y}_\mathrm{n} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

위의 식을 벡터를 이용하여 표시하면 다음과 같다.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{1.8}$$

위의 행렬식에서 각 벡터와 행렬의 차원은 다음과 같다.

- $\mathbf{y}$ $n \times 1$
- $\mathbf{X}$ $n \times (p+1)$
- $\boldsymbol{\beta}$ $(p+1) \times 1$
- $\mathbf{e}$ $n \times 1$

여기서 회귀분석의 오차항은 서로 독립이고 동일한 분산을 갖는다. 즉, 오차항은 다음의 분포를 따른다. 즉, $E(\mathbf{e}) = \mathbf{0}$ 이므로 관측값 벡터 $\mathbf{y}$의 평균을 보면

$$E(\mathbf{y}|\mathbf{X}) = E(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}) = \mathbf{X}\boldsymbol{\beta} + E(\mathbf{e}) = \mathbf{X}\boldsymbol{\beta} \tag{1.9}$$

1.3 최소제곱추정

최소제곱추정법(least square estimation)은 자료의 관계을 잘 반영하는 회귀식을 구한 다음 실제 관측값 y_i과 예측값 $\mathbf{x}_i^t\boldsymbol{\beta}$ 간에 차이인 잔차를 가장 작게 만드는 것이 목적이다.

모든 잔차항의 제곱의 합을 최소화하는 방법을 최소제곱법이라고 하며 이를 이용하여 회귀계수의 추정량을 찾는다.

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2 = \min_{\boldsymbol{\beta}} (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \tag{1.10}$$

1.3.1 방법 1

$\hat{\boldsymbol{\beta}}$는 잔차의 제곱합 (1.10) 을 최소로 하는 최소제곱 추정량이다. 잔차의 제곱합을 $S(\boldsymbol{\beta})$ 이라고 하면

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^t\mathbf{y} - \mathbf{y}^t\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^t\mathbf{y} - 2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} \end{aligned} \tag{1.11}$$

여기서 $S(\boldsymbol{\beta})$를 최소로 하는 회귀계수벡터의 값을 구하기 위하여 $S(\boldsymbol{\beta})$를 회귀계수벡터 $\boldsymbol{\beta}$로 미분한후 $\mathbf{0}$ 으로 놓고 선형 방정식을 풀어야 한다.

앞 절에 나오는 벡터미분을 이용하면

$$\begin{aligned} \frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}} (\mathbf{y}^t\mathbf{y} - 2\boldsymbol{\beta}^t\mathbf{X}^t\mathbf{y} + \boldsymbol{\beta}^t\mathbf{X}^t\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{0} - 2\mathbf{X}^t\mathbf{y} + 2\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{0} \end{aligned}$$

최소제곱 추정량을 구하기 위한 정규방정식은 다음과 같이 쓸 수 있다.

$$\mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{X}^t\mathbf{y} \tag{1.12}$$

방정식 (1.12)를 정규방정식(normal equation)이라고 한다. 만약 $\mathbf{X}^t\mathbf{X}$가 정칙행렬일 경우 최소제곱법에 의한 회귀계수 추정량 $\hat{\boldsymbol{\beta}}$ 다음과 같다.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y} \tag{1.13}$$

예측값 벡터 $\hat{\mathbf{y}}$ 는 $E(\mathbf{y}|\mathbf{X})$의 추정치로서 다음과 같다.

$$\hat{E}(\mathbf{y}|\mathbf{X}) = \hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t y$$

만약 $\mathbf{X}^t\mathbf{X}$가 정칙행렬이 아닐 경우 최소제곱법에 의한 회귀계수 추정량 $\hat{\boldsymbol{\beta}}$은 $\mathbf{X}^t\mathbf{X}$의 일반화 역행렬 $(\mathbf{X}^t\mathbf{X})^-$를 이용하여 다음과 같이 구한다. 이 경우 일반화 역행렬이 유일하지 않기 때문에 회귀계수 추정량도 유일하지 않다.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^-\mathbf{X}^t\mathbf{y}$$

1.3.2 방법 2

식 (1.10)에서 나오는 오차벡터를 정의하고 $\mathbf{e} = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$ 오차벡터를 모수벡터 $\boldsymbol{\beta}$로 미분하면 다음과 같은 결과를 얻는다.

$$\frac{\partial \mathbf{e}}{\partial \boldsymbol{\beta}} = \frac{\partial(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{\partial \mathbf{X}\boldsymbol{\beta}}{\partial \boldsymbol{\beta}} \equiv -\frac{\partial \boldsymbol{\beta}^t\mathbf{X}^t}{\partial \boldsymbol{\beta}} = -\mathbf{X}^t$$

이제 오차제곱합 $S(\boldsymbol{\beta}) = \mathbf{e}^t\mathbf{e}$ 를 모수벡터로 미분하면 이차형식의 미분공식과 합성함수 미분공식을 차례로 적용하면 된다.

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \frac{\partial \mathbf{e}^t\mathbf{e}}{\partial \boldsymbol{\beta}} = \frac{\partial \mathbf{e}}{\partial \boldsymbol{\beta}}\frac{\partial \mathbf{e}^t\mathbf{e}}{\partial \mathbf{e}} = -\mathbf{X}^t\,(2\mathbf{e}) = -2\mathbf{X}^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

위의 방정식을 $\mathbf{0}$으로 놓으면 최소제곱 추정량 (열) 벡터를 구한다.

$$\mathbf{X}^t\mathbf{y} - \mathbf{X}^t\mathbf{X}\boldsymbol{\beta} = \mathbf{0} \quad \rightarrow \quad \hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

제 2 장

가능도함수

2.1 가능도함수 개요

2.2 가능도함수와 그 성질

확률변수 y 가 확률밀도함수 $f(y;\boldsymbol{\theta})$를 따른다고 하자. 모수벡터 $\boldsymbol{\theta}$ 에 대한 가능도함수 $L(\boldsymbol{\theta};y)$와 로그가능도함수 $\ell(\boldsymbol{\theta};y)$는 다음과 같이 정의한다.

$$L(\boldsymbol{\theta}) = L(\boldsymbol{\theta};y) \equiv f(y;\boldsymbol{\theta}), \quad \ell(\boldsymbol{\theta}) = \ell(\boldsymbol{\theta};y) \equiv \log L(\boldsymbol{\theta};y)$$

로그가능도함수를 모수 $\boldsymbol{\theta}$로 한 번 미분한 도함수(greadient)를 스코어함수(score function) $\mathbf{s}(\boldsymbol{\theta};y)$ 로 아래와 같이 정의한다. 또한 두 번 미분한 헤시안(hessian)의 음수를 관측피셔정보(observed Fisher information) $J(\boldsymbol{\theta};y)$ 라고 정의한다.

$$\mathbf{s}(\boldsymbol{\theta}) = \mathbf{s}(\boldsymbol{\theta};y) \equiv \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta};y), \quad \mathbf{J}(\boldsymbol{\theta}) = \mathbf{J}(\boldsymbol{\theta};y) \equiv -\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\ell(\boldsymbol{\theta};y)$$

위의 식에서 만약 모수벡터 $\boldsymbol{\theta}$의 차원이 p라면 $\mathbf{s}(\boldsymbol{\theta})$는 $p\times 1$ 벡터이고 $\mathbf{J}(\boldsymbol{\theta};y)$는 $p\times p$ 행렬이다.

로그가능도함수는 다음의 두 가지 중요한 방정식을 만족한다.

$$E\left\{\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta};y)\right\}=0 \tag{2.1}$$

$$E\left\{\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta};y)\right]\left[\frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta};y)\right]^t\right\}+E\left\{\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\ell(\boldsymbol{\theta};y)\right\}=0 \tag{2.2}$$

식 (2.1)와 식 (2.2) 으로부터 다음과 같은 식이 유도되며

$$E[\mathbf{s}(\boldsymbol{\theta};y)]=0,\quad E[\mathbf{s}(\boldsymbol{\theta};y)s^t(\boldsymbol{\theta};y)]=E[\mathbf{J}(\boldsymbol{\theta};y)]$$

다음과 같은 공식이 주어진다.

$$\begin{aligned}Var[\mathbf{s}(\boldsymbol{\theta};y)]&=E[\mathbf{s}(\boldsymbol{\theta};y)\mathbf{s}^t(\boldsymbol{\theta};y)]-\{E[\mathbf{s}(\boldsymbol{\theta};y)]E[\mathbf{s}^t(\boldsymbol{\theta};y)]\}\\&=E[\mathbf{s}(\boldsymbol{\theta};y)\mathbf{s}^t(\boldsymbol{\theta};y)]-\mathbf{0}\\&=-E\left[\frac{\partial^2}{\partial\boldsymbol{\theta}^2}\log f(y;\boldsymbol{\theta})\right]\\&=E[\mathbf{J}(\boldsymbol{\theta};y)]\\&\equiv\mathbf{I}(\boldsymbol{\theta})\end{aligned}$$

위의 식에서 스코어함수의 분산을 피셔정보(Fisher information)이라고 부르며 $\mathbf{I}(\boldsymbol{\theta})$로 표기한다.

관측피셔정보(observed Fisher information) $J(\boldsymbol{\theta};y)$ 는 모수와 확률변수로 정의되는 확률 이며 피셔정보(Fisher information) $\mathbf{I}(\boldsymbol{\theta})$는 관측피셔정보의 기대값으로 모수만의 함수로서 더이상 확률변수가 아니다.

첫 번째 방정식 (2.1)는 다음과 같이 적분과 미분의 교환에 의해 증명할 수 있다.

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\theta}} \int f(y;\boldsymbol{\theta})\, dy \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} f(y;\boldsymbol{\theta})\, dy \\
&= \int \frac{\frac{\partial}{\partial \boldsymbol{\theta}} f(y;\boldsymbol{\theta})}{f(y;\boldsymbol{\theta})} f(y;\boldsymbol{\theta})\, dy \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) f(y;\boldsymbol{\theta})\, dy \\
&= E\left\{ \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right\}
\end{aligned}
$$

두 번째 방정식 (2.2)는 아래와 같이 증명할 수 있다.

$$
\begin{aligned}
0 &= \frac{\partial}{\partial \boldsymbol{\theta}} \int \frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) f(y;\boldsymbol{\theta})\, dy \\
&= \int \frac{\partial}{\partial \boldsymbol{\theta}} \left\{ f(y;\boldsymbol{\theta}) \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right]^t \right\} dy \\
&= \int \left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} f(y;\boldsymbol{\theta}) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right]^t + f(y;\boldsymbol{\theta}) \frac{\partial}{\partial \boldsymbol{\theta}} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right]^t \right\} dy \\
&= \int \left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right]^t f(y;\boldsymbol{\theta}) + \left[\frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta};y) \right] f(y;\boldsymbol{\theta}) \right\} dy \\
&= E\left\{ \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right] \left[\frac{\partial}{\partial \boldsymbol{\theta}} \ell(\boldsymbol{\theta};y) \right]^t \right\} + E\left\{ \frac{\partial^2}{\partial \boldsymbol{\theta}^2} \ell(\boldsymbol{\theta};y) \right\}
\end{aligned}
$$

2.3 독립표본

표본 $y_1, y_2, \ldots, y_n$ 가 분포 $f(y_i;\boldsymbol{\theta})$에서 독립적으로 얻어졌다면 표본에 대한 가능도함수 $L_n(\boldsymbol{\theta})$ 은 다음과 같다.

$$
L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(y_i;\boldsymbol{\theta}) \tag{2.3}
$$

또한 표본에 대한 로그가능도함수 $\ell_n(\boldsymbol{\theta})$ 은 다음과 같다.

$$\ell_n(\boldsymbol{\theta}) = \ell_n(\boldsymbol{\theta}; \mathbf{y}) = \log L_n(\boldsymbol{\theta}) = \log \prod_{i=1}^{n} f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta}; y_i) \tag{2.4}$$

표본에 의한 로그 가능도함수 $\ell_n(\boldsymbol{\theta})$를 미분한 값, 즉 표본에 의한 스코어 함수 $s_n(\boldsymbol{\theta})$는 다음과 같이 정의한다.

$$s_n(\boldsymbol{\theta}) = \frac{\partial}{\partial \boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}; \mathbf{y})$$

n개의 표본에 대한 관측피셔정보 $\mathbf{J}_n(\boldsymbol{\theta})$와 피셔정보 $\mathbf{I}_n(\boldsymbol{\theta})$도 한 개의 확률 변수 경우와 유사하게 다음과 같이 정의된다.

$$\mathbf{I}_n(\boldsymbol{\theta}) = E\left[\mathbf{J}_n(\boldsymbol{\theta}; \mathbf{y})\right] = E\left[-\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^t} \ell_n(\boldsymbol{\theta}; \mathbf{y})\right]$$

2.4 지수군 분포

확률변수 y가 다음과 같은형태의 분포를 따른다면 y의 분포는 지수군(exponential family)에 속한다고 한다.

$$f(y; \boldsymbol{\theta}, \phi) = \exp\left\{\frac{\mathbf{t}(y)^t \boldsymbol{\xi}(\boldsymbol{\theta}) - b(\boldsymbol{\theta})}{a(\phi)} + c(y, \phi)\right\} \tag{2.5}$$

다시 쓰면

$$\log f(y; \boldsymbol{\theta}, \phi) = \frac{\mathbf{t}(y)^t \boldsymbol{\xi}(\boldsymbol{\theta}) - b(\boldsymbol{\theta})}{a(\phi)} + c(y, \phi)$$

지수군 분포 (2.5) 에서 $\mathbf{t}(y)^t = [t_1(y), \ldots, t_k(y)]$ 는 k 개의 충분통계량으로 구성된 벡터이고 k-차원 벡터 $\boldsymbol{\xi}(\boldsymbol{\theta})^t = [\xi_1(\boldsymbol{\theta}), \ldots, \xi_k(\boldsymbol{\theta})]$를 **기본형 모수**(canonical parameter)라고 부른다. 또한 $a(\phi)$를 **스케일모수**(scale parameter)라고 부르며 많은 경우 $a(\phi) = a \times w$의 형태로 나타나며 여기서 w는 보통 가중치와 같은 역활을 한다. 또한 기본형 모수 $\boldsymbol{\xi}(\boldsymbol{\theta})$는 y 의 평균 $\mu = E(y)$와 특별한 함수관계를 가진다.

지수군 분포는 일반적으로 식 (2.5) 과 같은 나타낼 수 있지만 많은 경우 모수 $\boldsymbol{\theta}$와 기본형

모수 $\boldsymbol{\xi}(\boldsymbol{\theta})$는 일대일 대응관계를 가진다. 일대일 대응 관계가 아닌 경우 이를 곡선형 지수군 (curved exponential family)라고 부른다. 따라서 지수군 분포의 성질을 간결하게 유도하고 설명하기 위해서 다음과 같은 단순화된 형태의 식을 사용하는 것이 편리하다. 이후 모든 성질의 유도는 아래 식 (2.6) 의 형태를 사용할 것이다.

$$\log f(y;\boldsymbol{\theta},\phi) = \frac{\mathbf{t}^t\boldsymbol{\theta} - b(\boldsymbol{\theta})}{a(\phi)} + c(y,\phi) \tag{2.6}$$

위의 식 (2.6) 은 충분통계량 벡터를 $\mathbf{y}$으로 나타내고 대응하는 기본형 모수를 $\boldsymbol{\theta}$로 표시한 것이다. 또한 여기서 충분통계량 벡터 $\mathbf{t}$의 평균을 $\boldsymbol{\mu}$라고 하자.

$$E(\mathbf{t}) = \boldsymbol{\mu}$$

2.4.1 평균, 분산과 기본형 모수의 관계

이제 (2.6) 에 나오는 $b(\boldsymbol{\theta})$의 미분을 다음과 같이 표시하자.

$$b'(\boldsymbol{\theta}) = \frac{\partial b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \quad b''(\boldsymbol{\theta}) = \frac{\partial^2 b(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}^t \partial \boldsymbol{\theta}}$$

방정식 (2.1)으로부터 다음과 같은 식이 유도된다..

$$\begin{aligned} 0 &= E\left(\frac{\partial \ell}{\partial \boldsymbol{\theta}}\right) \\ &= E\left(\mathbf{t} - b'(\boldsymbol{\theta})\right)/a(\phi)) \\ &= [\boldsymbol{\mu} - b'(\boldsymbol{\theta})]/a(\phi) \end{aligned}$$

따라서 평균 $\boldsymbol{\mu}$와 함수 $b(\boldsymbol{\theta})$는 다음과 같은 관계가 성립한다.

$$E(\mathbf{t}) = \boldsymbol{\mu} = b'(\boldsymbol{\theta}) \tag{2.7}$$

또한 방정식 (2.2) 으로부터 다음이 성립하고

$$E\left\{a^{-2}(\phi)[\mathbf{t}-b'(\boldsymbol{\theta})][\mathbf{t}-b'(\boldsymbol{\theta})]^t\right\}+E\left\{-a^{-1}(\phi)b''(\boldsymbol{\theta})\right\}=0$$

따라서 $\mathbf{t}$의 분산과 함수 $b(\boldsymbol{\theta})$는 다음과 같은 관계가 성립한다.

$$Var(\mathbf{t})=a(\phi)b''(\boldsymbol{\theta})\equiv a(\phi)v(\boldsymbol{\mu}) \tag{2.8}$$

위의 식에서 $v(\boldsymbol{\mu})=b''(\boldsymbol{\theta})$ 으로 정의하고 분산함수(variance function)라고 부른다.

이제 분산함수의 정의로부터 $\boldsymbol{\mu},\boldsymbol{\theta}$에 대하여 다음과 같은 관계가 얻어지고

$$\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}}=\frac{\partial b'(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}=b''(\boldsymbol{\theta})=v(\boldsymbol{\mu}) \tag{2.9}$$

도함수의 역관계가 다음과 같이 주어진다.

$$\frac{\partial\boldsymbol{\theta}}{\partial\boldsymbol{\mu}}=\left[\frac{\partial\boldsymbol{\mu}}{\partial\boldsymbol{\theta}}\right]^{-1}=[b''(\boldsymbol{\theta})]^{-1}=v^{-1}(\boldsymbol{\mu}) \tag{2.10}$$

2.4.2 지수군 분포의 예제

예제 2.1 (이항분포). 확률변수 S 가 이항분포 $B(n,\mu)$를 따른다면(여기서 $\mu=p$ 성공확률) 표본 비율 $y=S/n$의 로그확률밀도함수는 다음과 같다.

$$\begin{aligned}\log f(y;\theta,\phi)&=\log\left\{\binom{n}{ny}\mu^{ny}(1-\mu)^{n-ny}\right\}\\&=\frac{y\log\frac{\mu}{1-\mu}+\log(1-\mu)}{n^{-1}}+\log\binom{n}{ny}\\&=\frac{y\theta-b(\theta)}{a(\phi)}+c(y,\phi)\end{aligned}$$

충분통계량 t는 표본비율 y이고 기본형 모수 θ와 평균 $E(y)=\mu=p$, 스케일 모수 $a(\phi)$은 다음과 같은 관계가 있다.

$$\theta=\log\frac{\mu}{1-\mu}=\log\frac{p}{1-p},\quad b(\theta)=-\log(1-\mu),\quad a(\phi)=\frac{1}{n} \tag{2.11}$$

평균 μ를 기본형모수 θ의 함수로 역변환하면 θ의 로지스틱함수로 표현된다.

$$\mu = \frac{\exp(\theta)}{1+\exp(\theta)} = \frac{1}{1+\exp(-\theta)}, \quad 1-\mu = \frac{1}{1+\exp(\theta)}$$

또한 함수 $b(\theta)$를 기본형 모수로 나타내면

$$b(\theta) = -\log(1-\mu) = \log[1+\exp(\theta)]$$

특별히 $n=1$인 경우는 베르누이 분포이다.

이항분포에서 평균 μ는 함수 b와 다음과 같은 관계가 있다.

$$b'(\theta) = \frac{\exp(\theta)}{1+\exp(\theta)} = \mu$$

또한

$$b''(\theta) = \frac{\exp(\theta)[1+\exp(\theta)] - [\exp(\theta)]^2}{[1+\exp(\theta)]^2} = \mu(1-\mu) = v(\mu)$$

따라서

$$Var(y) = a(\phi)v(\mu) = \frac{\mu(1-\mu)}{n}$$

예제 2.2 (포아송분포). 확률변수 y 가 포아송 분포 $poi(\mu)$를 따른다고 하자. 여기서 $E(y)=\mu$ 이다. 확률변수 y의 로그확률밀도함수는 다음과 같다.

$$\log f(y;\theta,\phi) = y\log\mu - \mu - log(y!)$$

충분통계량 t는 반응값 y 자체이고 기본형 모수 θ와 평균 $E(y)=\mu$, 스케일 모수 $a(\phi)$은 다음과 같은 관계가 있다. 기본형 모수 θ와 평균 μ, 스케일 모수 $a(\phi)$은 다음과 같은 관계가 있다.

$$\theta = \log\mu, \quad b(\theta) = \mu, \quad a(\phi) = 1 \tag{2.12}$$

평균 μ를 기본형모수 θ의 함수로 역변환하면 θ의로그함수로 표현된다. 또한 함수 $b(\theta)$를

기본형 모수로 나타내면 다음과 같다.

$$\mu = \exp(\theta), \quad b(\theta) = \mu = \exp(\theta)$$

따라서 포아송분포에서는 평균 μ는 함수 b와 다음과 같은 관계가 있다.

$$b'(\theta) = \exp(\theta) = \mu, \quad b''(\theta) = \exp(\theta) = \mu = v(\mu)$$

따라서

$$Var(y) = a(\phi)v(\mu) = \mu$$

예제 2.3 (정규분포). 확률변수 y 가 정규분포 $N(\mu, \sigma^2)$를 따른다고 하자. 확률변수 y의 로그확률밀도함수는 다음과 같다. 이때 모수벡터를 $\boldsymbol{\theta}^t = (\mu, \sigma^2)$이다.

$$\begin{aligned}
\log f(y; \boldsymbol{\theta}, \phi) &= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{(y-\mu)^2}{2\sigma^2} \\
&= -\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\sigma^2 - \frac{y^2 - 2y\mu + \mu^2}{2\sigma^2} \\
&= \left[y\frac{\mu}{\sigma^2} - y^2\frac{1}{2\sigma^2}\right] - \left[\frac{1}{2}\log\sigma^2 + \frac{\mu^2}{2\sigma^2}\right] - \frac{n}{2}\log(2\pi)
\end{aligned}$$

충분 통계량과 기본형 모수는 다음과 같다.

$$\mathbf{t}(y)^t = (y, y^2), \quad \boldsymbol{\xi}(\boldsymbol{\theta})^t = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right), \quad a(\phi) = 1$$

또한

$$b(\boldsymbol{\theta}) = \frac{1}{2}\log\sigma^2 + \frac{\mu^2}{2\sigma^2}$$

예제 2.4 (다항분포). 다항분포(multinomial distribution)인 경우도 고려해 보자. 확률벡터 $y^t = (y_1, \ldots, y_k)$를 성공확률이 $\mu^t = (\mu_1, \ldots, \mu_k)$ 인 다항분포를 따른다고 하자 ($\sum_j \mu_j = 1$, $\sum y_j = m$). 확률벡터 y의 확률함수는 다음과 같다.

$$\log f(y, \mu, m) = \log\frac{m!}{y_1! \ldots y_k!}\mu_1^{y_1}\mu_2^{y_2}\ldots\mu_k^{y_k} = \sum_{j=1}^{k-1} y_j \log\frac{\mu_j}{1-\sum_1^{k-1}\mu_j} + m\log(1-\sum_1^{k-1}\mu_j) + c(m, y)$$

2.4.3 최대가능도추정법

모수 $\boldsymbol{\theta}$ 에 대한 최대가능도 추정량(Maximum Likelihood Estimator;MLE) $\hat{\boldsymbol{\theta}}$는 가능도 함수를 최대로 하는 값으로 정의된다.

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} L_n(\boldsymbol{\theta})$$

많은 경우 가능도 함수를 최대화하는 값을 구하기 어려우므로 가능도 함수의 로그 함수, 즉 로그가능도함수를 최대로 하는 값으로 최대가능도 추정량을 구한다.

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\max_{\boldsymbol{\theta}} \ell_n(\boldsymbol{\theta})$$

만약 로그가능도 함수가 모수 $\boldsymbol{\theta}$에 대하며 미분가능한 함수이면 최대가능도 추정량은 다음과 같은 방정식에 의하여 구할 수 있다.

$$\frac{\partial}{\partial \boldsymbol{\theta}} \ell_n(\boldsymbol{\theta}; \mathbf{y}) = s_n(\boldsymbol{\theta}) = \mathbf{0}$$

최대가능도 추정량은 적당한 조건하에서 다음과 같은 점근적 성질(Asymptotical properties)을 가진다.

- $\hat{\boldsymbol{\theta}}_{MLE}$는 모수의 참값 $\boldsymbol{\theta}_0$로 확률적 수렴한다.

$$\hat{\boldsymbol{\theta}}_{MLE} \to_p \boldsymbol{\theta}_0 \quad \text{as } n \to \infty$$

- 최대가능도추정량 $\hat{\boldsymbol{\theta}}_{MLE}$는 점근적으로 정규분포를 따른다.

$$\hat{\boldsymbol{\theta}}_{MLE} \sim_d N(\boldsymbol{\theta}_0, \mathbf{I}_n^{-1}(\boldsymbol{\theta}_0))$$

2.5 선형모형

반응변수가 y이고 $p-1$개의 독립변수 $(x_1, x_2, \cdots, x_{p-1})$가 있다고 가정하고 표본의 크기 n인 자료가 얻어지면 선형회귀식을 행렬로 다음과 같이 표현할 수 있다.

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{i,p-1} + e_i \\ &= \mathbf{x}_i^t \boldsymbol{\beta} + e_i, \quad i = 1, 2, \cdots, n \end{aligned}$$

위의 식을 다시 표현하면 다음과 같이 쓸 수 있다. 이와 같은 회귀모형을 선형중회귀모형이라 부르며. 각 개체에 대한 모형의 방정식을 하나의 식으로 표현하면 다음과 같다.

$$\begin{bmatrix} \mathrm{y}_1 \\ \mathrm{y}_2 \\ \vdots \\ \mathrm{y}_\mathrm{n} \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1,p-1} \\ 1 & x_{21} & \cdots & x_{2,p-1} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

위의 선형모형(linear model)을 벡터와 행렬을 이용하여 표시하면 다음과 같다.

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e} \tag{2.13}$$

여기서 $\mathbf{y}$ 는 $n \times 1$ 반응변수 벡터, $\mathbf{X}$ 는 p 개의 독립변수로 이루어진 $n \times p$ 계획행렬(design matrix)이다. 모수 벡터 $\boldsymbol{\beta}$ 는 $p \times 1$ 벡터로 각 독립변수에 대한 회귀계수 벡터이다. $\mathbf{e}$는 $n \times 1$ 오차벡터이다.

여기서 회귀분석의 오차항의 가정을 살펴보면 오차항이 서로 독립이고 동일한 분산을 갖는다. 즉, 오차항은 다음의 분포를 따른다. 즉, $\mathbf{e} \sim (0, \sigma^2\mathbf{I})$. 관측값 벡터 $\mathbf{y}$의 평균과 분산을 보면

$$E(\mathbf{y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}, \quad Var(\mathbf{y}|\mathbf{X}) = \sigma^2\mathbf{I}$$

여기서 오차항이 정규분포를 따른다면 ($\mathbf{e} \sim N(0, \sigma^2\mathbf{I})$ 관측값 벡터 $\mathbf{y}$ 또한 정규분포를 따른다

$$\mathbf{y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I})$$

또한 X 가 완전계수(full rank) 행렬이라고 가정하자.

$p+1$개의 모수를 모아놓은 모수벡터는 $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \sigma^2)^t$이다. 여기서 편의상 오차항의

분산을 $\tau = \sigma^2$ 로 표시하하고자 한다, 즉 $\boldsymbol{\theta} = (\boldsymbol{\beta}^t, \tau)^t$

2.5.1 최소제곱 추정

위의 선형 모형 가정하에서, 최소제곱 추정량 $\hat{\boldsymbol{\beta}}$ (least square estimator)는 다음과 같이 오차제곱합(Error Sum of Sqaures) SSE 를 최소로 하는 추정량이다.

$$\begin{aligned} SSE(\boldsymbol{\beta}) &= \sum_{i=1}^{n} (y_i - \mathbf{x}_i^t \boldsymbol{\beta})^2 \\ &= (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t (\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) \\ \hat{\boldsymbol{\beta}} &= \arg\min_{\boldsymbol{\beta}} SSE(\boldsymbol{\beta}) \end{aligned}$$

따라서 $\hat{\boldsymbol{\beta}}$는 오차제곱합을 최소로 하는 계수 벡터이며 최소제곱 추정량은 다음과 같이 주어진다.

$$\hat{\boldsymbol{\beta}}_{LS} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{y}$$

오차제곱합에 최소제곱 추정량을 사용하면 이를 잔차제곱합(Residual Sum of Squares)라고 하며 이를 $SSE(\hat{\boldsymbol{\beta}})$로 표시한다.

$$SSE(\hat{\boldsymbol{\beta}}) = (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})$$

오차항의 분산에 대한 추정은 정규분포 가정을 오차항에 대한 정규분포를 가정하고 다음과 같은 잔차제곱합의 분포에 대한 결과를 이용하면

$$SSE(\hat{\boldsymbol{\beta}}) \sim \sigma^2 \chi^2(n-p)$$

오차항의 분산 σ^2 의 불편추정량 S^2 을 구할 수 있다.

$$S^2 = \frac{SSE(\hat{\boldsymbol{\beta}})}{(n-p)} = \frac{\sum_{i=1}^{n} (y_i - \mathbf{x}_i^t \hat{\boldsymbol{\beta}})^2}{(n-p)}$$

즉,

$$E(S^2) = \sigma^2$$

여기서 자유도 $n-p$은 자료의 개수 n에서 절편을 포함한 회귀계수의 개수 p를 뺀 수이다.

2.5.2 가능도함수

선형모형 (3.1)에 대한 가능도 함수는 다음과 같이 주어진다.

$$\begin{aligned} L_n(\boldsymbol{\theta}; \mathbf{y}) &= L(\boldsymbol{\beta}, \sigma^2 | \mathbf{y}) \\ &= \prod_{i=1}^{n} f(y_i) \\ &= \prod_{i=1}^{n} (2\pi\sigma^2)^{-\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}(y_i - \mathbf{x}_i^t\boldsymbol{\beta})^2\right] \\ &= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right] \end{aligned}$$

또한 분산에 대한 모수를 $\tau = \sigma^2$ 과 같이 쓰면 로그 가능도함수는 다음과 같다.

$$\begin{aligned} \ell_n(\boldsymbol{\theta}; \mathbf{y}) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\sigma^2} \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\tau - \frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^t(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}{2\tau} \end{aligned}$$

이제 로그가능도함수로부터 구할 수 있는 스코어함수 $s(\boldsymbol{\theta}; y)$ 와 그에 대한 관측 피셔정보 $\mathbf{J}_n(\boldsymbol{\theta}; \mathbf{y})$ 은 다음과 같이 주어진다.

$$
\begin{aligned}
s(\boldsymbol{\theta};\mathbf{y}) &= \frac{\partial}{\partial\boldsymbol{\theta}}\ell_n(\boldsymbol{\theta};\mathbf{y}) \\
&= \begin{bmatrix} \frac{\partial}{\partial\boldsymbol{\beta}}\ell_n(\boldsymbol{\theta};\mathbf{y}) \\ \frac{\partial}{\partial\tau}\ell_n(\boldsymbol{\theta};\mathbf{y}) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{X}^t(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\tau \\ -\frac{n}{2\tau} + \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2\tau^2} \end{bmatrix} \\
\mathbf{J}_n(\boldsymbol{\theta};\mathbf{y}) &= -\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^t}\ell_n(\boldsymbol{\theta};y) \\
&= -\begin{bmatrix} \frac{\partial^2}{\partial\beta\partial\boldsymbol{\beta}^t}\ell_n(\boldsymbol{\theta};y) & \frac{\partial^2}{\partial\boldsymbol{\beta}\partial\tau^t}\ell_n(\boldsymbol{\theta};y) \\ \frac{\partial^2}{\partial\tau\partial\boldsymbol{\beta}^t}\ell_n(\boldsymbol{\theta};y) & \frac{\partial^2}{\partial\tau\partial\tau^2}\ell_n(\boldsymbol{\theta};y) \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{X}^t\mathbf{X}/\tau & -\mathbf{X}^t(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\tau^2 \\ -\mathbf{X}^t(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})/\tau^2 & -\frac{n}{2\tau^2} + \frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^t(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{\tau^3} \end{bmatrix}
\end{aligned}
$$

2.5.3 최대가능도 추정량

이제 회귀계수 $\boldsymbol{\beta}$에 대한 최대가능도 추정량은 스코어함수로 부터 얻어진 방정식 $s(\boldsymbol{\theta};\mathbf{y}) = \mathbf{0}$ 으로부터 얻어지며 다음과 같은 형태를 가진다.

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{y}$$

$$\hat{\sigma}^2 = \hat{\tau} = (\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^t(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/n = \frac{SSE(\hat{\boldsymbol{\beta}})}{n}$$

여기서 유의할 점은 회귀계수 $\boldsymbol{\beta}$ 의 최대가능도 추정량은 최소제곱법으로 구한 추정량과 동일하다. 따라서 $\hat{\boldsymbol{\beta}}$은 최소분산 불편 추정량이다. 하지만 오차항의 분산 σ^2 에 대한 최대가능도 추정량은 불편추정량이 아니다.

$$E(\hat{\sigma}^2) = E\left[(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})^t(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/n\right] = E\left[\frac{SSE}{n}\right] \neq \sigma^2$$

참고로 오차항의 분산 σ^2에 대한 불편추정량은 $SSE/(n-p)$이다.

최대가능도 추정량의 점근적 분포를 이용하면 다음과 같이 말할 수 있다. 오차항이 정규분

포인 선형모형인 경우 아래의 분포는 점근분포가 아닌 정확한 분포이다.

$$\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \sim N(\mathbf{0}, \mathbf{I}_n^{-1}(\boldsymbol{\theta}_0))$$

여기서

$$\mathbf{I}_n(\boldsymbol{\theta}) = E[\mathbf{J}(\boldsymbol{\theta}; \mathbf{y})] = \begin{bmatrix} \mathbf{X}^t\mathbf{X}/\tau & 0 \\ 0 & \frac{n}{2\tau^2} \end{bmatrix}$$

그리고

$$\mathbf{I}_n^{-1}(\boldsymbol{\theta}) = \begin{bmatrix} \tau(\mathbf{X}^t\mathbf{X})^{-1} & 0 \\ 0 & \frac{2\tau^2}{n} \end{bmatrix} = \begin{bmatrix} \sigma^2(\mathbf{X}^t\mathbf{X})^{-1} & 0 \\ 0 & \frac{2\sigma^4}{n} \end{bmatrix}$$

따라서 회귀계수 $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0$의 분포는 평균이 $\mathbf{0}$ 이고 공분산이 $\sigma^2(\mathbf{X}^t\mathbf{X})^{-1}$ 인 정규분포를 따른다.

여기거 주목할 점은 가능도함수에 최대가능도추정량을 대입하면 그 값이 $SSE(\hat{\boldsymbol{\beta}})$의 함수로 나타난다.

$$\begin{aligned} L_n(\hat{\boldsymbol{\theta}}) &= L_n(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) \\ &= (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\hat{\sigma}^2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^t(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})\right] \\ &= (2\pi\hat{\sigma}^2)^{-\frac{n}{2}} \exp\left[-\frac{n}{2}\right] \\ &= \left(2\pi\frac{SSE(\hat{\boldsymbol{\beta}})}{n}\right)^{-\frac{n}{2}} \exp\left[-\frac{n}{2}\right] \\ l_n(\hat{\boldsymbol{\theta}}) &= l_n(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2) \\ &= \text{constant} - \frac{n}{2}\log SSE(\hat{\boldsymbol{\beta}}) \end{aligned}$$

따라서 잔차제곱함 $SSE(\hat{\boldsymbol{\beta}})$ 작아지면 가능도함수는 커진다.

2.6 최대가능도 추정량의 점근적 성질 (update 필요)

로그가능도 함수의 테일러 전개를 고려해 보자.

$$\ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}) + s(\hat{\boldsymbol{\theta}})^t(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t \nabla(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o_p(1)$$

따라서

$$2(\ell(\boldsymbol{\theta}) = \ell(\hat{\boldsymbol{\theta}}) + 0 - \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t I(\boldsymbol{\theta})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^t [\nabla(\hat{\boldsymbol{\theta}}) + I(\boldsymbol{\theta})](\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) + o_p(1)$$

제 3 장

일반화 선형모형

3.1 일반화선형모형

1 장에서 살펴본 회귀모형을 보통 선형모형이라고 부른다. 선형모형의 의미는 모형에서 고려하는 설명변수가 변할 때 반응값의 평균이 변하는 관계가 선형이라는 것이다. 즉, 반응값의 평균을 설명하는 회귀식이 회귀계수에 대하여 선형이라는 의미이다. 참고로 아래 식 (3.1)의 오른쪽에 나타나는 식을 선형예측식(linear predictor, η)라고 부른다.

$$E(y|x_1, x_2, \dots, x_p) = \beta_1 x_1 + \cdots + \beta_p x_p \equiv \eta \tag{3.1}$$

이러한 선형성의 가정이 적절하지 않은 경우가 있다. 예를 들어 공학에서나 생물학에서 사용되는 비선형 회귀모형(nonlinear regression model)처럼 반응변수의 변화가 설명변수들의 복잡한 비선형 관계(예를 들어 미분방정식의 관계)로 나타나는 경우로 흔히 나타난다. 또한 반응변수가 가질 수 있는 평균값에 제한이 있을 수 있다. 예를 들어 베르누이 분포의 경우 평균이 성공확률이기 때문에 0과 1사이에 있으며 포아송 분포의 경우 반응값은 음의 값을 가질 수 없다. 따라서 식 (3.1)의 선형예측식 η와 반응값의 평균 $E(y|\mathbf{x})$의 관계를 선형모형 (3.1)처럼 정의할 수 없다.

이렇게 반응변수의 평균과 선형예측식의 범위가 일치하지 않는 경우 임의의 단조증가 함수 g를 사용하여 그 범위를 일치하게 만들어 줄 수 있다. 예를 들어 베르누이 분포의 경우 표준정규분포의 누적분포함수 Φ를 사용하여 확률의 범위와 선형예측식의 범위를 맞추어 줄

수 있다. 이러한 회귀모형을 프로빗(probit)모형이라고 부른다.

$$\Phi^{-1}[p(y|\mathbf{x})] = \beta_1 x_1 + \cdots + \beta_p x_p = \eta \tag{3.2}$$

이제부터 정규분포하에서 정의되는 선형모형을 다른 분포들로 확장한 모형인 일반화 선형모형(Generalized Linear Model; GLM)을 살펴보기로 하자.

3.1.1 지수군 분포와 일반화 선형모형

일변량 확률변수 y가 식 (2.6)와 같이 다음과 같은 지수군 분포를 따른다고 가정하자.

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

충분통계량이 관측값 y이고 1차원의 기본형 모수θ로 정의된 분포임을 유의하자.

확률변수 y의 평균을 $\mu = E(y)$ 이라고 하고 독립변수 벡터 $\mathbf{x}$와 회귀계수 벡터 $\boldsymbol{\beta}$로 구성된 선형예측식을 η라고 하자.

$$\eta = \mathbf{x}^t\boldsymbol{\beta} \tag{3.3}$$

일반화 선형모형은 분포의 특성에 따라 주어진 단조증가함수 g 를 이용하여 y 의 평균과 선형예측식의 관계를 설정하는 모형이다. 이러한 함수 g 를 연결함수(link function)라고 부른다.

$$g(\mu) = g(E[y|\mathbf{x}]) = \mathbf{x}^t\boldsymbol{\beta} \tag{3.4}$$

반응값의 분포가 주어진 경우 연결함수는 평균의 범위와 선형예측식의 범위를 연속적으로 1-1 대응하게 해주는 함수이며 사용할 수 있는 가능한 함수는 무한히 많다. 예를 들어 베르부이 분포의 경우 위에서 정의된 프로빗 모형 (3.2)에서 Φ^{-1} 같이 (0,1) 에서 실수전체 집합으로 단조증가하는 함수는 모두 연결함수로 고려할 수 있다.

지수군 분포에서 모수 θ를 기본형모수(canonical parameter)라고 부르며 일반적으로 θ는 평균 μ의 비선형 함수로 나타난다. 만약 다음과 같은 관계를 나타내는 연결함수 g 가 있다면 그 함수를 기본형 연결함수(canonical link function)이라고 부른다.

$$\theta = \eta \tag{3.5}$$

예를 들어 예제 2.1을 보면 이항분포의 확률밀도함수를 지수군 분포의 형태로 표현했을 때 식 (2.11)의 형태를 보면 다음과 같은 관계를 알 수 있다.

$$\theta = \log \frac{\mu}{1-\mu} = \log \frac{p}{1-p}$$

따라서 식 (3.5) 을 만족하는 연결함수는 다음과 같고

$$\log \frac{p}{1-p} = \eta = \mathbf{x}^t \boldsymbol{\beta} \tag{3.6}$$

이를 로짓 연결함수(logit link function)이라고 부르며 이는 이항분포의 기본형 연결함수이다.

만약 y의 분포가 정규분포이며 연결함수 g가 $g(\mu) = \mu$이면 선형회귀모형이 된다.

$$E[y|\mathbf{x}] = \mathbf{x}^t \boldsymbol{\beta}$$

3.2 일반화 선형모형의 가능도함수

하나의 표본 y에 대하여 기본형 모수 θ 하나인 로그가능도함수(log likelihood function) ℓ은 다음과 같이 정의된다.

$$\ell = \log f(y) = \frac{y\theta - b(\theta)}{a(\phi)} + \log c(y, \phi)$$

표본 $y_1, y_2, \ldots, y_n$ 가 각각 설명변수 벡터$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$에서 독립적으로 얻어졌다면 로그가능도함수 ℓ_n 은 다음과 같다.

$$\ell_n = \log \prod_{i=1}^{n} f(y_i) = \sum_{i=1}^{n} \frac{y_i\theta_i - b(\theta_i)}{a(\phi_i)} + \sum_{i=1}^{n} \log c(y_i, \phi) \tag{3.7}$$

여기서 θ_i는 i 번째 관측값에 대한 모수로서 첨자 i 를 붙이는 이유는 관측치의 기대값 $\mu_i = E(y_i|\mathbf{x}_i)$가 독립변수의 값 $\mathbf{x}_i$에 따라 다를 수 있고 θ_i는 평균 μ_i의 함수이기 때문이다.

이제 식 (3.4)과 같이 설명변수와 반응변수 평균과의 관계가 연결함수 g로 정의되었다고 하자.

$$g(\mu_i) = g(E[y_i|\mathbf{x}_i]) = \mathbf{x}_i^t\boldsymbol{\beta} \equiv \eta_i, \quad i = 1, 2, \dots, n \tag{3.8}$$

Nelder and Wedderburn (1972) 에서 연결 함수(link function)의 개념이 제시할 때 다음과 같은 작업 변량(working variate) z_i를 이용하여 선형모형을 일반화하고자 하였다. 즉, 작업 변량 z_i

$$z_i = g(\mu_i) + g_\mu(\mu)(y_i - \mu_i) \tag{3.9}$$
$$= \mathbf{x}_i^t\boldsymbol{\beta} + g_\mu(\mu)(y_i - \mu_i) \tag{3.10}$$
$$\simeq \mathbf{x}_i^t\boldsymbol{\beta} + r_i \tag{3.11}$$

위의 식에서 오른쪽 식의 두번째 항의 기대값이 0이므로 이를 오차항과 같이 생각하면 위의 모형을 오차항의 분산이 다른 선형모형으로 생각할 수 있다. 지수군 분포의 성질 (2.8)를 이용하면 작업 변량 z_i의 분산은 다음과 같다.

$$Var(z_i) = Var(r_i) = [g_\mu(\mu_i)]^2 Var(y_i) = [g_\mu(\mu_i)]^2[a(\phi_i)v(\mu_i)] \tag{3.12}$$

이러한 가정하에서 작업 변량 z_i를 반응변수로 놓고 분산의 역수를 가중치로하는 가중 선형모형(wighted linear regression) 을 최소제곱법으로 적합하고 계수의 값이 수렴할 때까지 반복적으로 수행하는하는 계산법을 제공하였다. 이러한 방법을 반복가중최소최곱법(iterative weighted least square method; IWLS)라고 부른다.

(Searle and McCulloch, 2001, 의 136 쪽 참조)

3.3 최대가능도추정

이제 회귀계수 $\boldsymbol{\beta}$를 최대가능도추정법(Maximum Likelihood Estimation)으로 구하기 위하여 로그가능도함수 (3.7) 를 회귀계수 벡터 $\boldsymbol{\beta}$로 미분한 가능도함수 방정식을 고려하자.

$$\frac{\partial \ell_n}{\partial \boldsymbol{\beta}} = \mathbf{0} \tag{3.13}$$

여기서 일반화 선형모형에서 나타나는 모수들 $\boldsymbol{\beta}$, μ_i, θ_i의 관계를 살펴보자.

1. 회귀계수 벡터 $\boldsymbol{\beta}$는 설명변수 벡터 $\mathbf{x}_i$와 내적 형태로 연결되어 있으며 이를 선형 예측식 이라고 한다.
$$\eta_i = \mathbf{x}_i^t \boldsymbol{\beta}$$
2. 선형 예측식 η_i는 관측값의 평균 μ_i와 연결함수 g로 연결되어 있다.
$$g(\mu_i) = \eta_i$$
3. 관측값의 평균 μ_i는 기본형 모수 θ_i와 함수 b로 연결되어 있다.
$$b'(\theta_i) = \mu_i$$

일반화 선형모형에서 최종적으로 추정해야 하는 모수는 회귀계수 벡터 $\boldsymbol{\beta}$ 이며 모수 $\boldsymbol{\beta}$, μ_i, θ_i 다음과 연결되어 있음을 알 수 있다.

$$\boldsymbol{\beta} \underset{g}{\longrightarrow} \mu_i \underset{b}{\longrightarrow} \theta_i \tag{3.14}$$

최대가능도 추정량을 구하는 방정식을 유도할 때 다음과 같은 일반화 선형모형에 대한 tp 가지 지수군 분포에서 나타나는 미분공식이 적용된다.

1. **선형 예측식의 미분**:
$$\frac{\partial \eta}{\partial \boldsymbol{\beta}} = \frac{\partial \mathbf{x}^t \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = \mathbf{x}$$
2. **연결함수의 역함수에 대한 미분**: $g(\mu) = \eta$ 의 관계를 이용하면
$$\frac{\partial \mu}{\partial \eta} = \frac{\partial \mu}{\partial g(\mu)} = \left[\frac{\partial g(\mu)}{\partial \mu}\right]^{-1} == [g'(\mu)]^{-1} = g_\mu^{-1}(\mu)$$

여기서 $g_\mu(\mu) = g'(\mu)$로서 연결함수의 미분을 나타내는 기호이다.

3. **평균과 기본형모수의 미분, 분산함수**: 식 (2.9)에서 관계식을 이용하면

$$\frac{\partial \theta}{\partial \mu} = \left[\frac{\partial \mu}{\partial \theta}\right]^{-1} = [b''(\theta)]^{-1} = \frac{1}{v(\mu)} \tag{3.15}$$

3.3.1 가능도 방정식의 유도: 첫 번째 방법

이제 가능도 함수 (3.7) 의 형태를 이용하여 방정식 (3.13) 를 을 유도해 보자.

$$\begin{aligned}
\mathbf{0} &= \frac{\partial \ell_n}{\partial \boldsymbol{\beta}} && (3.16)\\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)}\right] \left[y_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} - \frac{\partial b(\theta_i)}{\partial \boldsymbol{\beta}}\right] && (3.17)\\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)}\right] \left[y_i \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} - \frac{\partial \theta_i}{\partial \boldsymbol{\beta}} \frac{\partial b(\theta_i)}{\partial \theta_i}\right] && (3.18)\\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)}\right] \left[\frac{\partial \theta_i}{\partial \boldsymbol{\beta}} (y_i - \mu_i)\right] && (3.19)\\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)}\right] \left[\frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \frac{\partial \mu_i}{\partial \eta_i} \frac{\partial \theta_i}{\partial \mu_i} (y_i - \mu_i)\right] && (3.20)\\
&= \sum_{i=1}^{n} \left[\frac{1}{a(\psi_i)}\right] \left[\mathbf{x}_i \frac{(y_i - \mu_i)}{v(\mu_i) g_\mu(\mu_i)}\right] && (3.21)\\
&= \sum_{i=1}^{n} \left[\mathbf{x}_i w_i g_\mu(\mu_i)(y_i - \mu_i)\right] && (3.22)\\
&= \begin{bmatrix} \sum_{i=1}^{n} x_{i1} w_i g_\mu(\mu_i)(y_i - \mu_i) \\ \sum_{i=1}^{n} x_{i2} w_i g_\mu(\mu_i)(y_i - \mu_i) \\ \vdots \\ \sum_{i=1}^{n} x_{ip} w_i g_\mu(\mu_i)(y_i - \mu_i) \end{bmatrix} && (3.23)
\end{aligned}$$

여기서 가중치 w_i는 다음과 정의한다. 가중치 w_i는 앞에서 설명한 작업 변량의 분산의 역수와 동일하다. 식 (3.12) 을 참조하자.

$$w_i \equiv \frac{1}{g_\mu^2(\mu_i) a(\phi_i) v(\mu_i)}$$

.

3.3.2 가능도 방정식의 유도: 두 번째 방법

위의 방정식 (3.13) 에 미분의 연쇄법칙(chain rule)을 적용하면 다음과 같은 방정식을 얻는다.

$$\frac{\partial \ell_n}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} \frac{\partial \ell_n}{\partial \boldsymbol{\theta}} = 0 \tag{3.24}$$

위의 식에서 $\boldsymbol{\eta}$, $\boldsymbol{\mu}$, $\boldsymbol{\theta}$는 다음과 같이 n개의 대응되는 원소로 이루어진 벡터이다.

$$\boldsymbol{\eta} = \begin{bmatrix} \eta_1 \\ \eta_2 \\ \vdots \\ \eta_n \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix}, \quad \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}$$

이제 식 (3.24)에서 나타난 도함수들을 각각 구해보자. 먼저

$$\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} = \begin{bmatrix} \frac{\partial \eta_1}{\partial \boldsymbol{\beta}} & \frac{\partial \eta_2}{\partial \boldsymbol{\beta}} & \cdots & \frac{\partial \eta_n}{\partial \boldsymbol{\beta}} \end{bmatrix} \tag{3.25}$$

$$= \begin{bmatrix} \frac{\partial \mathbf{x}_1^t \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} & \frac{\partial \mathbf{x}_2^t \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} & \cdots & \frac{\partial \mathbf{x}_n^t \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} \end{bmatrix} \tag{3.26}$$

$$= [\mathbf{x}_1 \ \ \mathbf{x}_2 \ \ \ldots \ \ \mathbf{x}_n] \tag{3.27}$$

$$= \mathbf{X}^t \tag{3.28}$$

위의 식에서 행렬 $\mathbf{X}$는 $n \times p$ 계획행렬이다.

또한

$$\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} = \begin{bmatrix} \frac{\partial \mu_1}{\partial \boldsymbol{\eta}} & \frac{\partial \mu_2}{\partial \boldsymbol{\eta}} & \cdots & \frac{\partial \mu_n}{\partial \boldsymbol{\eta}} \end{bmatrix} \tag{3.29}$$

$$= \begin{bmatrix} \frac{\partial g^{-1}(\eta_1)}{\partial \boldsymbol{\eta}} & \frac{\partial g^{-1}(\eta_2)}{\partial \boldsymbol{\eta}} & \cdots & \frac{\partial g^{-1}(\eta_2)}{\partial \boldsymbol{\eta}} \end{bmatrix} \tag{3.30}$$

$$= \begin{bmatrix} \frac{1}{g'(\mu_1)} & 0 & \cdots & 0 \\ 0 & \frac{1}{g'(\mu_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{g'(\mu_n)} \end{bmatrix} \tag{3.31}$$

$$= \begin{bmatrix} g_\mu^{-1}(\mu_1) & & & \\ & g_\mu^{-1}(\mu_2) & & \\ & & \ddots & \\ & & & g_\mu^{-1}(\mu_n) \end{bmatrix} \tag{3.32}$$

위에서 $\partial \boldsymbol{\mu}/\partial \boldsymbol{\eta}$는 n-차원 대각행렬이며 $g_\mu(\mu) = g'(\mu)$로서 연결함수 g를 1차 미분한 함수이다.

또한 다음과 같은 결과를 얻는다.

$$\frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} = \begin{bmatrix} \frac{\partial \theta_1}{\partial \boldsymbol{\mu}} & \frac{\partial \theta_2}{\partial \boldsymbol{\mu}} & \cdots & \frac{\partial \theta_n}{\partial \boldsymbol{\mu}} \end{bmatrix} \tag{3.33}$$

$$= \begin{bmatrix} \frac{1}{b''(\theta_1)} & 0 & \cdots & 0 \\ 0 & \frac{1}{b''(\theta_2)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \frac{1}{b''(\theta_n)} \end{bmatrix} \tag{3.34}$$

$$= \begin{bmatrix} v^{-1}(\mu_1) & & & \\ & v^{-1}(\mu_2) & & \\ & & \ddots & \\ & & & v^{-1}(\mu_n) \end{bmatrix} \tag{3.35}$$

마지막으로 식 (2.7) 과 (3.7) 를 이용하면 로그가능도함수 ℓ_n 을 $\boldsymbol{\theta}$로 미분한 n-차원 벡터는 다음과 같이 얻어진다.

$$\frac{\partial \ell_n}{\partial \boldsymbol{\theta}} = \begin{bmatrix} \frac{y_1 - b'(\theta_1)}{a(\phi_1)} \\ \frac{y_2 - b'(\theta_2)}{a(\phi_2)} \\ \vdots \\ \frac{y_n - b'(\theta_n)}{a(\phi_n)} \end{bmatrix} = \begin{bmatrix} \frac{y_1 - \mu_1}{a(\phi_1)} \\ \frac{y_2 - \mu_2}{a(\phi_2)} \\ \vdots \\ \frac{y_n - \mu_n}{a(\phi_n)} \end{bmatrix} \tag{3.36}$$

이제 가능도추정을 위한 방정식 (3.24) 을 위에서 유도한 도함수 벡터와 행렬을 이용하여 다시 쓰면 다음과 같다.

$$\mathbf{0} = \frac{\partial \ell_n}{\partial \boldsymbol{\beta}} = \frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \frac{\partial \boldsymbol{\theta}}{\partial \boldsymbol{\mu}} \frac{\partial \ell}{\partial \boldsymbol{\theta}} \tag{3.37}$$

$$= \mathbf{X}^t \begin{bmatrix} \frac{1}{g_\mu(\mu_1)} & & & \\ & \frac{1}{g_\mu(\mu_1)} & & \\ & & \ddots & \\ & & & \frac{1}{g_\mu(\mu_n)} \end{bmatrix} \begin{bmatrix} \frac{1}{v(\mu_1)} & & & \\ & \frac{1}{v(\mu_2)} & & \\ & & \ddots & \\ & & & \frac{1}{v(\mu_n)} \end{bmatrix} \begin{bmatrix} \frac{y_1 - \mu_1}{a(\phi_1)} \\ \frac{y_2 - \mu_2}{a(\phi_2)} \\ \vdots \\ \frac{y_n - \mu_n}{a(\phi_n)} \end{bmatrix} \tag{3.38}$$

$$= \mathbf{X}^t \begin{bmatrix} \frac{1}{g_\mu^2(\mu_1) a(\phi_1) v(\mu_1)} & & & \\ & \frac{1}{g_\mu^2(\mu_1) a(\phi_2) v(\mu_2)} & & \\ & & \ddots & \\ & & & \frac{1}{g_\mu^2 a(\phi_n)(\mu_n) v(\mu_n)} \end{bmatrix} \tag{3.39}$$

$$\times \begin{bmatrix} g_\mu(\mu_1) & & & \\ & g_\mu(\mu_2) & & \\ & & \ddots & \\ & & & g_\mu(\mu_n) \end{bmatrix} \begin{bmatrix} y_1 - \mu_1 \\ y_2 - \mu_2 \\ \vdots \\ y_n - \mu_n \end{bmatrix} \tag{3.40}$$

$$= \mathbf{X}^t \mathbf{W} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}) \tag{3.41}$$

위의 식에서 가중값 대각행렬 $\mathbf{W}$, 연결함수 미분값 대각행렬 $\boldsymbol{\Delta}$, 관측값 벡터 $\mathbf{y}$, 평균 벡터 $\boldsymbol{\mu}$는 다음과 같이 정의된다.

$$\mathbf{W} = \begin{bmatrix} w_1 & & & \\ & w_2 & & \\ & & \ddots & \\ & & & w_n \end{bmatrix} = \begin{bmatrix} \frac{1}{g_\mu^2(\mu_1)a(\phi_1)v(\mu_1)} & & & \\ & \frac{1}{g_\mu^2(\mu_1)a(\phi_2)v(\mu_2)} & & \\ & & \ddots & \\ & & & \frac{1}{g_\mu^2 a(\phi_n)(\mu_n)v(\mu_n)} \end{bmatrix} \tag{3.42}$$

$$\mathbf{\Delta} = \begin{bmatrix} \delta_1 & & & \\ & \delta_2 & & \\ & & \ddots & \\ & & & \delta_n \end{bmatrix} = \begin{bmatrix} g_\mu(\mu_1) & & & \\ & g_\mu(\mu_2) & & \\ & & \ddots & \\ & & & g_\mu(\mu_n) \end{bmatrix} \tag{3.43}$$

$$\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_n \end{bmatrix} \tag{3.44}$$

3.4 최대가능도추정량의 계산

이제 회귀계수 β를 최대가능도추정법(Maximum Likelihood Estimation)으로 가능도 방정식을 식 (3.41)를 이용하면 다음과 같은 행렬 방정식으로 표시된다.

$$\mathbf{X}^t\mathbf{W}\mathbf{\Delta}\mathbf{y} = \mathbf{X}^t\mathbf{W}\mathbf{\Delta}\boldsymbol{\mu} \tag{3.45}$$

위의 방정식은 일반적으로 회귀계수 벡터 $\boldsymbol{\beta}$에 대하여 선형방정식이 아니므로 최소제곱법과 같이 최대가능도 추정량을 직접 구할 수 없다.

1. 정규분포 가정 하에서 선형회귀 모형에서는 식 (**??**)이 최소제곱법의 방정식 $\mathbf{X}^t\mathbf{y} = \mathbf{X}^t\mathbf{X}\boldsymbol{\beta}$로 유도되고 직접적으로 구할 수 있다.
2. 많은 경우 스케일 모수 $a(\phi_i)$는 관측값 y_i에 따라 변하지 않고 상수인 경우가 흔하다. 즉 $a(\phi_i) \equiv a(phi)$. 이러한 경우 가능도 방정식 (3.23) 또는 (3.41)에서 스케일 모수 $a(\phi_i)$를 1로 놓고 방정식을 푼다.

최대가능도추정량을 실제 계산하기 위하여 로그 가능도 함수의 2차 도함수(헤시안) 행렬을

구해보자. 식 (3.23) 에서 얻은 1차 도함수를 한번 더 미분하면 다음과 같은 결과를 얻는다.

$$
\begin{aligned}
\frac{\partial^2 \ell_n}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} &= \sum \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{x}_i w_i g_\mu(\mu_i)(y_i - \mu_i) \right] \\
&= \sum \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{x}_i c_i (y_i - \mu_i) \right] \quad [c_i \equiv w_i g_\mu(\mu_i)] \\
&= \sum \left[\frac{\partial c_i (y_i - \mu_i)}{\partial \boldsymbol{\beta}} \mathbf{x}_i^t \right] \\
&= \sum \left[\frac{\partial c_i}{\partial \boldsymbol{\beta}} (y_i - \mu_i) + \frac{\partial (y_i - \mu_i)}{\partial \boldsymbol{\beta}} c_i \right] \mathbf{x}_i^t \\
&= \sum \left[\frac{\partial c_i}{\partial \boldsymbol{\beta}} (y_i - \mu_i) - \frac{\partial \eta_i}{\partial \boldsymbol{\beta}} \frac{\partial \mu_i}{\partial \eta_i} c_i \right] \mathbf{x}_i^t \\
&= \sum \left[\frac{\partial c_i}{\partial \boldsymbol{\beta}} (y_i - \mu_i) - \mathbf{x}_i [g_\mu(\mu)]^{-1} c_i \right] \mathbf{x}_i^t \quad [c_i [g_\mu(\mu)]^{-1} = w_i] \\
&= \sum \left[\frac{\partial c_i}{\partial \beta} (y_i - \mu_i) \right] \mathbf{x}_i^t - \sum \mathbf{x}_i w_i \mathbf{x}_i^t \\
&= \sum \left[\frac{\partial c_i}{\partial \beta} (y_i - \mu_i) \right] \mathbf{x}_i^t - \mathbf{X}^t \mathbf{W} \mathbf{X}
\end{aligned}
$$

그러므로 피셔정보 $\mathbf{I}(\boldsymbol{\beta})$는 다음과 같이 얻어진다.

$$
\begin{aligned}
\mathbf{I}(\boldsymbol{\beta}) &= -E \left[\frac{\partial^2 \ell_n}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} \right] && (3.46) \\
&= E \left[-\sum \left[\frac{\partial c_i}{\partial \beta} (y_i - \mu_i) \right] \mathbf{x}_i^t + \mathbf{X}^t \mathbf{W} \mathbf{X} \right] && (3.47) \\
&= \mathbf{0} + \mathbf{X}^t \mathbf{W} \mathbf{X} && (3.48)
\end{aligned}
$$

또는 식 (3.41) 에서 얻은 1차 도함수 방정식을 를 한번 더 미분하여 기대값을 취하면 식 (3.48)와 동일한 결과를 얻는다.

$$\frac{\partial^2 \ell_n}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t} = \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{X}^t \mathbf{W} \boldsymbol{\Delta} (\mathbf{y} - \boldsymbol{\mu}) \right] \tag{3.49}$$

$$= \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{X}^t \mathbf{W} \boldsymbol{\Delta} \right] \right\} (\mathbf{y} - \boldsymbol{\mu}) + \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \left[(\mathbf{y} - \boldsymbol{\mu})^t \right] \right\} \boldsymbol{\Delta} \mathbf{W} \mathbf{X} \tag{3.50}$$

$$= \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{X}^t \mathbf{W} \boldsymbol{\Delta} \right] \right\} (\mathbf{y} - \boldsymbol{\mu}) - \left[\frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\beta}} \right] \boldsymbol{\Delta} \mathbf{W} \mathbf{X} \tag{3.51}$$

$$= \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{X}^t \mathbf{W} \boldsymbol{\Delta} \right] \right\} (\mathbf{y} - \boldsymbol{\mu}) - \left[\frac{\partial \boldsymbol{\eta}}{\partial \boldsymbol{\beta}} \frac{\partial \boldsymbol{\mu}}{\partial \boldsymbol{\eta}} \right] \boldsymbol{\Delta} \mathbf{W} \mathbf{X} \tag{3.52}$$

$$= \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{X}^t \mathbf{W} \boldsymbol{\Delta} \right] \right\} (\mathbf{y} - \boldsymbol{\mu}) - \left[\mathbf{X}^t \boldsymbol{\Delta}^{-1} \right] \boldsymbol{\Delta} \mathbf{W} \mathbf{X} \tag{3.53}$$

$$= \left\{ \frac{\partial}{\partial \boldsymbol{\beta}} \left[\mathbf{X}^t \mathbf{W} \boldsymbol{\Delta} \right] \right\} (\mathbf{y} - \boldsymbol{\mu}) - \mathbf{X}^t \mathbf{W} \mathbf{X} \tag{3.54}$$

$$\tag{3.55}$$

최대가능도추정량 $\hat{\boldsymbol{\beta}}$는 가능도 방정식 (3.45) 을 직접 풀어서 계산할 수 있지만 많은 경우 직접해(explicit solution)를 구하는 것이 불가능 하다. 따라서 보통의 경우 선형화된 작업변량에 반복가중최소제곱법(iterative weighted least square; IWLS)을 적용하여 최대가능도 추정량을 구하며 IWLS로 구하는 해는 가능도 방정식 (3.45)의 해와 동일하다.

주어진 분포에서 기본 연결함수를 g 라고 하고 관측값 y를 변환한 작업변량 $z = g(y)$의 테일러 전개를 다음과 같이 고려해 보자.

$$z \equiv g(y) \cong g(\mu) + g_\mu(\mu)(y - \mu)$$

작업 변량 z 의 분산은 식 (3.12) 와 같이 다음으로 주어진다.

$$var(z) = v(\mu) g_\mu^2(\mu) \equiv w^{-1}$$

회귀계수 벡터의 초기값을 $\boldsymbol{\beta}_0$라고 하자. 그러면 작업 변량의 초기값 z_0는 $\boldsymbol{\beta}_0$ 로 계산된 μ_0를 이용하여 다음과 같이 구할 수 있다.

$$z_0 = g(\mu_0) + g_\mu(\mu_0)(y - \mu_0) = \eta_0 + g_\mu(\mu_0)(y - \mu_0)$$

IWLS 추정량 $\hat{\boldsymbol{\beta}}$는 z_0 를 설명변수 벡터 $\mathbf{x}$로 선형회귀분석을 적합할 때 가중치를 w_0 로 이용하는 가중최소제곱법으로 반복적으로 적용하여 개선할 수 있다.

실제로 IWLS 추정량은 피셔정보를 이용한 스코어 방법(Fisher scoring method)로 구한 최대가능도 추정량과 동일함을 보일 수 있다. 일단 회귀계수 벡터의 초기값 $\hat{\boldsymbol{\beta}}^0$ 으로 계산된 피셔정보 행렬을 $\mathbf{A}$로 아래와 같이 정의하자.

$$\mathbf{A} = -E\left[\frac{\partial^2 \ell_n}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^t}\right]_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^0}$$

새로운 추정량 $\hat{\boldsymbol{\beta}}^1$ 가 이전의 추정량 $\hat{\boldsymbol{\beta}}^0$에서 다음과 같은 축차식으로 계산되는 방법이 피셔 스코링 방법이다.

$$0 = \frac{\partial \ell_n}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^0} - \mathbf{A}(\hat{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^0) \quad \Leftrightarrow \quad \mathbf{A}(\hat{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^0) = \frac{\partial \ell_n}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^0}$$

위의 피셔의 스코링 방법을 더 정리하면 다음과 같이 유도할 수 있다.

$$\begin{aligned}
\mathbf{A}(\hat{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^0) &= \frac{\partial \ell_n}{\partial \boldsymbol{\beta}}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}^0} \\
\Leftrightarrow \mathbf{X}^t\mathbf{W}_0\mathbf{X}(\hat{\boldsymbol{\beta}}^1 - \hat{\boldsymbol{\beta}}^0) &= \mathbf{X}^t\mathbf{W}_0\boldsymbol{\Delta}_0(\mathbf{y} - \boldsymbol{\mu}_0) \\
\Leftrightarrow \mathbf{X}^t\mathbf{W}_0\mathbf{X}\hat{\boldsymbol{\beta}}^1 &= \mathbf{X}^t\mathbf{W}_0[\mathbf{X}\hat{\boldsymbol{\beta}}^0 + \boldsymbol{\Delta}_0(\mathbf{y} - \boldsymbol{\mu}_0)] \\
\Leftrightarrow \mathbf{X}^t\mathbf{W}_0\mathbf{X}\hat{\boldsymbol{\beta}}^1 &= \mathbf{X}^t\mathbf{W}_0[\boldsymbol{\eta}_0 + \boldsymbol{\Delta}_0(\mathbf{y} - \boldsymbol{\mu}_0)] \\
\Leftrightarrow \mathbf{X}^t\mathbf{W}_0\mathbf{X}\hat{\boldsymbol{\beta}}^1 &= \mathbf{X}^t\mathbf{W}_0\mathbf{z}_0
\end{aligned}$$

위의 방정식은 $z_0 = \eta_0 + g_\mu(\mu_0)(y - \mu_0)$를 가중치 w_0를 사용하여 얻은 가중최소제곱법에서 나온 방정식임을 알 수 있다. 따라서 최대가능도 추정량을 구하는 피셔의 스코링 방법은 앞에서 알아본 반복가중최소제곱법과 동일하다.

예제 3.1 (이항분포).

3.5 Maximum Quasi-Likelihood

Each observation y is estimated by its the estimator$\hat{\mu}$ and the goodness of fit should be assessed. We shall be concerned with the formed form the logarithm of a ratio of likelihoods, to be called *deviance*.

We consider two extreme models: the null model and the full (saturated) model.

The null model consider only one parameter for all y's so that it is a simplest model. The full model assumes there are n parameters for n observations so that each observation contribute to one parameter only (perfect fit).

Let $\ell(y, \theta, \psi) = \ell(\theta, \psi; y)$ be the log likelihood function and we can describe it as a function of μ such as $\ell(\theta, \psi; y) = \ell(\mu, \psi; y)$. The deviance is defined as

$$D(y, \mu, \psi) = 2\{\ell(\hat{\mu}_F, \psi; y) - \ell(\hat{\mu}, \psi; y)\}$$

where $\hat{\mu}_F$ is estimator under the full model and $\hat{\mu}$ is the maximum likelihood estimator under the model considered. The scaled deviance is defined as

$$D(y, \mu) = D(y, \mu, \psi)a(\psi)$$

부록 A

다변량 확률변수의 성질

A.1 일변량분포

일변량 확률변수 X가 확률밀도함수 $f(x)$를 가지는 분포를 따를때 기대값과 분산은 다음과 같이 정의된다.

$$E(X) = \int xf(x)dx = \mu, \quad V(X) = E[X - E(X)]^2 = \int (x-\mu)^2 f(x)dx = \sigma^2$$

새로운 확률변수 Y가 확률변수 X의 선형변환으로 표시된다면 (a와 b는 실수)

$$Y = aX + b$$

그 기대값(평균)과 분산은 다음과 같이 계산된다.

$$\begin{aligned} E(Y) &= E(aX+b) \\ &= \int (ax+b)f(x)dx \\ &= a\int xf(x)dx + b \\ &= aE(X)+b \\ &= a\mu + b \\ V(Y) &= Var(aX+b) \\ &= E[aX+b-E(aX+b)]^2 \\ &= E[a(X-\mu)]^2 \\ &= a^2E(X-\mu)^2 \\ &= a^2\sigma^2 \end{aligned}$$

A.2 확률벡터와 분포

확률벡터 $\mathbf{X}$가 p 차원의 다변량분포를 따른다고 하고 결합확률밀도함수 $f(\mathbf{x}) = f(x_1, x_2, \dots, x_p)$를 를 가진다고 하자.

$$\mathbf{X} = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ .. \\ X_p \end{pmatrix}$$

다변량 확률벡터의 기대값(평균벡터)과 공분산(행렬)은 다음과 같이 계산된다.

$$\mathbf{E}(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ E(X_3) \\ .. \\ E(X_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ .. \\ \mu_p \end{pmatrix} = \boldsymbol{\mu}$$

$$V(\mathbf{X}) = Cov(\mathbf{X}) = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^t = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{12} & \sigma_{22} & \cdots & \sigma_{2p} \\ & \cdots & \cdots & \\ \sigma_{1p} & \sigma_{2p} & \cdots & \sigma_{pp} \end{pmatrix} = \boldsymbol{\Sigma}$$

여기서 $\sigma_{ii} = V(X_i)$, $\sigma_{ij} = Cov(X_i, X_j) = Cov(X_j, X_i)$이다. 따라서 공분산 행렬 $\boldsymbol{\Sigma}$는 대칭행렬(symmetric matrix)이다. 다음 공식은 유용한 공식이다.

$$\boldsymbol{\Sigma} = E(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^t = E(\mathbf{X}\mathbf{X}^t) - \boldsymbol{\mu}\boldsymbol{\mu}^t$$

두 확률변수의 상관계수 ρ_{ij}는 다음과 같이 정의된다.

$$\rho_{ij} = \frac{Cov(X_i, X_j)}{\sqrt{V(X_i)V(X_j)}} = \frac{\sigma_{ij}}{\sqrt{\sigma_{ii}\sigma_{jj}}}$$

새로운 확률벡터 $\mathbf{Y}$가 확률벡터 $\mathbf{X}$ 의 선형변환라고 하자.

$$\mathbf{Y} = \mathbf{A}\mathbf{X} + \mathbf{b}$$

단 여기서 $\mathbf{A} = \{a_{ij}\}$는 $p \times p$ 실수 행렬이고 $\mathbf{b} = (b_1 b_2 \ldots b_p)^t$는 $p \times 1$ 실수 벡터이다.

확률벡터 $\mathbf{Y}$의 기대값(평균벡터)과 공분산은 다음과 같이 계산된다.

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{A}\mathbf{X}+\mathbf{b}) \\ &= \mathbf{A}E(\mathbf{X}) + \mathbf{b} \\ &= \mathbf{A}\boldsymbol{\mu} + \mathbf{b} \\ V(\mathbf{Y}) &= Var(\mathbf{A}\mathbf{X}+\mathbf{b}) \\ &= E[\mathbf{A}\mathbf{X}+\mathbf{b}-E(\mathbf{A}\mathbf{X}+\mathbf{b})][\mathbf{A}\mathbf{X}+\mathbf{b}-E(\mathbf{A}\mathbf{X}+\mathbf{b})]^t \\ &= E[\mathbf{A}\mathbf{X}-\mathbf{A}\boldsymbol{\mu}][\mathbf{A}\mathbf{X}-\mathbf{A}\boldsymbol{\mu}]^t \\ &= E[\mathbf{A}(\mathbf{X}-\boldsymbol{\mu})][\mathbf{A}(\mathbf{X}-\boldsymbol{\mu})]^t \\ &= \mathbf{A}E[(\mathbf{X}-\boldsymbol{\mu})(\mathbf{X}-\boldsymbol{\mu})^t]\mathbf{A}^t \\ &= \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^t \end{aligned}$$

만약 표본 $\mathbf{X}_i, \mathbf{X}_2, \ldots, \mathbf{X}_n$ 이 독립적으로 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$ 인 분포에서 추출되었다면 표본의 평균벡터 $\bar{\mathbf{X}}$ 는 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\frac{1}{n}\boldsymbol{\Sigma}$ 인 분포를 따른다.

$$\bar{\mathbf{X}} = \begin{pmatrix} \sum_{i=1}^{n} X_{i1}/n \\ \sum_{i=1}^{n} X_{i2}/n \\ \sum_{i=1}^{n} X_{i3}/n \\ .. \\ \sum_{i=1}^{n} X_{ip}/n \end{pmatrix}$$

여기서 X_{ij} 는 i번째 표본벡터 $\mathbf{X}_i = (X_{i1} X_{i2} \ldots X_{ip})^t$ 의 j번째 확률변수이다.

A.3 다변량 정규분포

일변량 확률변수 X가 평균이 μ 이고 분산이 σ^2인 정규분포를 따른다면 다음과 같이 나타내고

$$X \sim N(\mu, \sigma^2)$$

확률밀도함수 $f(x)$ 는 다음과 같이 주어진다.

$$f(x) = (2\pi\sigma^2)^{-1/2} \exp\left(-\frac{(x-\mu)^2}{2}\right)$$

p-차원 확률벡터 $\mathbf{X}$가 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$인 다변량 정규분포를 따른다면 다음과 같이 나타내고

$$\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

확률밀도함수 $f(\mathbf{x})$ 는 다음과 같이 주어진다.

$$f(\mathbf{x}) = (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{(\mathbf{x}-\boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})^t}{2}\right)$$

다변량 정규분포 $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$를 따르는 확률벡터 $\mathbf{X}$를 다음과 같이 두 부분으로 나누면

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \quad \mathbf{X}_1 = \begin{bmatrix} \mathbf{X}_{11} \\ \mathbf{X}_{12} \\ \vdots \\ \mathbf{X}_{1p} \end{bmatrix}, \quad \mathbf{X}_2 = \begin{bmatrix} \mathbf{X}_{21} \\ \mathbf{X}_{22} \\ \vdots \\ \mathbf{X}_{2q} \end{bmatrix}$$

각각 다변량 정규분포를 따르고 다음과 같이 나타낼 수 있다.

$$\begin{bmatrix} E(\mathbf{X}_1) \\ E(\mathbf{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \quad \begin{bmatrix} V(\mathbf{X}_1) & Cov(\mathbf{X}_1, X_2) \\ Cov(\mathbf{X}_2 X_1) & V(\mathbf{X}_2) \end{bmatrix} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^t & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix} \sim N_{p+q} \left(\begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \Sigma_{12} \\ \boldsymbol{\Sigma}_{12}^t & \Sigma_{22} \end{bmatrix} \right)$$

확률벡터 $\mathbf{X}_2 = \mathbf{x}_2$가 주어진 경우 $\mathbf{X}_1$의 조건부 분포는 p-차원 다변량 정규분포를 따르고 평균과 공분산은 다음과 같다.

$$E(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) = \boldsymbol{\mu}_1 + \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\boldsymbol{\mu}_2 - \mathbf{x}_2), \quad V(\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2) = \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^t$$

예를 들어 2-차원 확률벡터 $\mathbf{X} = (X_1, X_2)^t$가 평균이 $\boldsymbol{\mu} = (\mu_1, \mu_2)^t$ 이고 공분산 $\boldsymbol{\Sigma}$가 다음과 같이 주어진

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{pmatrix}$$

이변량 정규분포를 따른다면 확률밀도함수 $f(\mathbf{x})$에서 exp함수의 인자는 다음과 같이 주어진다.

$$(\mathbf{x} - \boldsymbol{\mu})\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})^t =$$
$$-\frac{1}{2(1-\rho^2)} \left[\left(\frac{(x_1-\mu_1)^2}{\sigma_{11}} \right) + \left(\frac{(x_2-\mu_2)^2}{\sigma_{22}} \right) - 2\rho \left(\frac{(x_1-\mu_1)}{\sqrt{\sigma_{11}}} \right) \left(\frac{(x_2-\mu_2)}{\sqrt{\sigma_{22}}} \right) \right]$$

그리고 $p = 2$인 경우 확률밀도함수의 상수부분은 다음과 같이 주어진다.

$$(2\pi)^{-p/2}|\boldsymbol{\Sigma}|^{-1/2} = \frac{1}{2\pi\sqrt{\sigma_{11}\sigma_{22}(1-\rho^2)}}$$

여기서 $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$

만약 $X_2 = x_2$가 주어졌을 때 X_1의 조건부 분포는 정규분포이고 평균과 분산은 다음과 같이 주어진다.

$$E(X_1|X_2 = x_2) = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(\mu_2 - x_2) = \mu_1 + \rho\frac{\sqrt{\sigma_{11}}}{\sqrt{\sigma_{22}}}(\mu_2 - x_2)$$

$$V(X_1|X_2 = x_2) = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} = \sigma_{11}(1 - \rho^2)$$

다변량 정규분포에서 공분산이 0인 두 확률 변수는 독립이다.

$$\sigma_{ij} = 0 \leftrightarrow X_i \text{ and } X_j \text{ are independent}$$

A.4 표준정규분포로의 변환

일변량 확률변수 X가 평균이 μ 이고 분산이 σ^2인 경우 다음과 같은 선형변환을 고려하면.

$$Z = \frac{X - \mu}{\sigma} = (\sigma^2)^{-1/2}(X - \mu)$$

확률변수 Z 는 평균이 0 이고 분산이 1인 분포를 따른다.

p차원 확률벡터 $\mathbf{X}$ 가 평균이 $\boldsymbol{\mu}$ 이고 공분산이 $\boldsymbol{\Sigma}$인 분포를 가진다고 가정하자. 공분산 행렬 $\boldsymbol{\Sigma}$는 양정치 행렬(positive definite matrix)이며 다음과 같은 행렬의 분해가 가능하다.

$$\Sigma = \mathbf{C}\mathbf{C}^t$$

여기서 $\mathbf{C}$ 는 정칙행렬이며 역행렬 $\mathbf{C}^{-1}$가 존재한다. 위와 같은 행렬의 분해는 스펙트럴 분해(spectral decomposition)을 이용하여 구할 수 있다. 공분산 행렬 $\boldsymbol{\Sigma}$는 양정치 행렬이므로 고유치(eigen value) $(\lambda_1, \lambda_2, \dots, \lambda_p)$가 모두 양수이고 정규직교 고유벡터(orthonormal eigen vector)의 행렬 $\mathbf{P}$을 이용하여 다음과 같은 분해가 가능하다.

$$\Sigma = \mathbf{P}\boldsymbol{\Lambda}\mathbf{P}^t = \mathbf{P}\boldsymbol{\Lambda}^{1/2}\Lambda^{1/2}\mathbf{P}^t$$

여기서 $\boldsymbol{\Lambda}$는 고유치 $(\lambda_1, \lambda_2, \dots, \lambda_p)$를 대각원소로 가지는 대각행렬이며 $\boldsymbol{\Lambda}^{1/2}$는 고유치의

제곱근을 대각원소로 가지는 대각행렬이다. 따라서 $\mathbf{C} = \mathbf{P}\mathbf{\Lambda}^{1/2}$로 하면 위와 같은 행렬의 분해가 가능하다. 정규직교 고유벡터(orthonormal eigen vector)의 행렬 $\mathbf{P}$는 직교행렬이므로

$$\mathbf{C}^{-1} = (\mathbf{P}\mathbf{\Lambda}^{1/2})^{-1} = \mathbf{\Lambda}^{-1/2}\mathbf{P}^t$$

p차원 확률벡터 $\mathbf{X}$의 다음과 같은 선형변환을 고려하면.

$$\mathbf{Z} = \mathbf{C}^{-1}(\mathbf{X} - \boldsymbol{\mu}) = \mathbf{\Lambda}^{-1/2}\mathbf{P}^t(\mathbf{X} - \boldsymbol{\mu})$$

확률벡터 $\mathbf{Z}$는 평균이 $\mathbf{0}$이고 공분산이 $\mathbf{I}$인 분포를 따른다 (why?).

확률벡터 $\mathbf{X}$가 정규분포를 따른다면 선형변환한 확률벡터 $\mathbf{Z}$도 정규분포를 따른다.

A.5 예제

예를 들어 이변량확률벡터 $\mathbf{X}$가 다음과 같은 평균벡터와 공분산을 가진 정규분포를 따른다고 하자

$$\boldsymbol{\mu} = \begin{pmatrix} 1 \\ 2 \end{pmatrix} \quad \mathbf{\Sigma} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}$$

공분산행렬 $\boldsymbol{\sigma}$의 고유치는 $|\boldsymbol{\sigma} - \lambda\mathbf{I}| = 0$의 방정식을 풀어 구할 수 있다.

$$|\boldsymbol{\sigma} - \lambda\mathbf{I}| = \begin{pmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{pmatrix} = \lambda^2 - 4\lambda + 3 = 0$$

방정식을 풀면 고유치는 $(\lambda_1, \lambda_2) = (3, 1)$이다. 각 고유치에 대한 고유벡터 $\mathbf{p} = (p_1, p_2)^t$는 $\mathbf{\Sigma p} = \lambda\mathbf{p}$ 으로 구할 수 있다. 각 고유치에 대하여 방정식을 구하면 다음 두 개의 방정식을 얻을 수 있다.

$$p_1 - p_2 = 1 \text{ and } p_1 + p_2 = 0$$

정규직교 벡터의 조건을 만족 시키기 위해서 $p_1^2 + p_2^2 = 1$의 조건을 적용하면 다음과 같은 정규직교 고유행렬을 얻을 수 있다.

$$\mathbf{P} = \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

또한

$$\mathbf{\Lambda} = \begin{pmatrix} 3 & 0 \\ 0 & 1 \end{pmatrix} \quad \mathbf{\Lambda}^{1/2} = \begin{pmatrix} \sqrt{3} & 0 \\ 0 & 1 \end{pmatrix}$$

따라서 $C^{-1} = \Lambda^{-1/2}\mathbf{P}^t$ 이며

$$\mathbf{C}^{-1} = \mathbf{\Lambda}^{-1/2}\mathbf{P}^t = \begin{pmatrix} \frac{1}{\sqrt{3}} & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} \\ -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix}$$

위의 계산을 R 프로그램으로 다음과 같이 구현할 수 있다.

```
mu <- c(1,2)
S <- matrix(c(2,1,1,2),2,2)
res<- eigen(S)
res
```

```
## eigen() decomposition
## $values
## [1] 3 1
##
## $vectors
##        [,1]    [,2]
## [1,] 0.7071 -0.7071
## [2,] 0.7071  0.7071
```

```
L <- res$values
P <- res$vectors
Lsqrt <- diag(sqrt(L))
C <- P %*% Lsqrt
C
```

```
##       [,1]    [,2]
## [1,] 1.225 -0.7071
## [2,] 1.225  0.7071
```

```
Cinv <- solve(C)
Cinv
```

```
##         [,1]   [,2]
## [1,]  0.4082 0.4082
## [2,] -0.7071 0.7071
```

```
Cinv %*% S %*% t(Cinv)
```

```
##      [,1] [,2]
## [1,]    1    0
## [2,]    0    1
```

부록 B

벡터미분

B.1 스칼라미분

벡터미분(Vector diffrential) 또는 행렬미분(Matrix differential)은 벡터와 행렬의 미분식에 대한 표기법을 정의하는 방법이다. 보통 스칼라(scalar)에 대한 미분은 일분수 함수 $f : \Re^1 \rightarrow \Re^1$ 또는 다변수 함수(function of several variables) $f : \Re^p \rightarrow \Re^1$ 에서 쉽게 정의된다. 만약 $y = f(x)$ 또는 $y = f(\mathbf{x})$라고 하면 다음과 같이 미분이 주어진다.

$$\frac{\partial y}{\partial x} = \frac{\partial f(x)}{\partial x} = f'(x)$$

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \left(\frac{\partial f(\mathbf{x})}{\partial x_1}, \ \frac{\partial f(\mathbf{x})}{\partial x_2}, \ \cdots, \frac{\partial f(\mathbf{x})}{\partial x_p} \right) = \nabla f(x)$$

함수가 다변수함수일 경우 함수의 값을 각 축의 변수로 미분한 것(partial derivative)을 벡터로 표시하는 것을 gradient 라고 한다.

B.2 벡터미분의 표기 방법

이제 다변량함수(multivariate function), $f : \Re^p \rightarrow \Re^q$에 대한 미분을 생각해보자. 앞 절에서 본것과 같이 스칼라 함수를 여러 변수로 미분하여 partial derivative를 구한 뒤 gradient를 만드는 경우 열벡터와 행벡터 중 하나를 선택해야 한다. 이러한 선택은

절대적인 것이 아니며 각 분야의 특성과 편의에 따라 다르게 선택 될 수 있다.

이제 간단한 예제를 고려해 보자. 두 열벡터 $\mathbf{x} = (x_1, x_2)^t \in \Re_2$, $\mathbf{y} = (y_1, y_2, y_3)^t \in \Re^3$ 를 고려하고 다음과 같은 함수로 두 벡터의 관계가 정의된다고 하자.

$$y_1 = x_1^2 + x_2, \quad y_2 = \exp(x_1) + 3x_2, \quad y_3 = \sin(x_1) + x_2^3$$

일단 각각의 partial derivative $\partial y_i / \partial x_j$를 구해야 하며 이는 scalar 미분으로 쉽게 구해진다.

$$\begin{array}{lll} \dfrac{\partial y_1}{\partial x_1} = 2x_1, & \dfrac{\partial y_2}{\partial x_1} = \exp(x_1), & \dfrac{\partial y_3}{\partial x_1} = \cos(x_1) \\ \dfrac{\partial y_1}{\partial x_2} = 1, & \dfrac{\partial y_2}{\partial x_1} = 3, & \dfrac{\partial y_3}{\partial x_1} = 3x_2^2 \end{array}$$

통계학에서는 벡터 $\mathbf{y}$를 벡터 $\mathbf{x}$로 미분하려면 다음과 같이 분모 표기법 (Denominator layout)을 사용하여 표기한다.

$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \equiv \frac{\partial \mathbf{y}^t}{\partial \mathbf{x}} \underset{def}{\equiv} \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \frac{\partial y_3}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \frac{\partial y_3}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 & \exp(x_1) & \cos(x_1) \\ 1 & 3 & 3x_2^2 \end{bmatrix}$$

즉 분모표기법은 분모를 열벡터로, 분자를 행벡터로 보고 각각 위치에 있는 변수들에 대하여 미분을 표기하는 방법이다.

B.3 핵심공식

디음은 분모표기법을 이용한 가장 기본적이고 핵심적인 미분 공식들이다. 공식을 유도하는 경우 분모표기법에서는 $\partial \mathbf{y}/\partial \mathbf{x} \equiv \partial \mathbf{y}^t/\partial \mathbf{x}$ 임을 이용한다. 변환이 있거나 여러가지 곱이 있는 경우 미분할 대상 벡터를 가장 왼쪽에 전치형태(즉, 행벡터의 형태로)로 놓는 것이 필요하다. 예를 들어

$$\frac{\partial \mathbf{a}^t \mathbf{V} \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{f}(\mathbf{x})^t \mathbf{V}^t \mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{f}(\mathbf{x})^t}{\partial \mathbf{x}} \mathbf{V}^t \mathbf{a} = \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} \mathbf{V}^t \mathbf{a}$$

또한 행렬은 교환법칙이 성립하지 않기 때문에 연산의 순서를 유지해야 하는 것을 유념하자.

B.3.1 기본행렬 미분

벡터 $\mathbf{c}$를 상수벡터하고 하자.

$$\frac{\partial \mathbf{c}}{\partial \mathbf{x}} = \mathbf{0}, \quad \frac{\partial \mathbf{x}}{\partial \mathbf{x}} = \mathbf{I}$$

B.3.2 벡터-스칼라 미분

이 경우는 $\mathbf{x} \in \Re^1$, $\mathbf{y} \in \Re^q$ 인 경우이며 결과는 다음과 같이 행벡터로 결과가 주어진다.

$$\frac{\partial \mathbf{y}}{\partial x} \underset{def}{\equiv} \frac{\partial \mathbf{y}^t}{\partial x} = \left[\frac{\partial y_1}{\partial x}, \ \frac{\partial y_2}{\partial x}, \ \cdots, \frac{\partial y_q}{\partial x} \right]$$

B.3.3 스칼라-벡터 미분

이 경우는 $\mathbf{x} \in \Re^p$, $\mathbf{y} \in \Re^1$ 인 경우이며 결과는 다음과 같이 열벡터로 결과가 주어진다.

$$\frac{\partial y}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y}{\partial x_1} \\ \frac{\partial y}{\partial x_2} \\ \vdots \\ \frac{\partial y}{\partial x_p} \end{bmatrix}$$

B.3.4 상수벡터와 내적에 대한 미분

열벡터 $\mathbf{a}$를 $p \times 1$ 상수벡터이라고 하고 $y = \mathbf{a}^t\mathbf{x} = \mathbf{x}^t\mathbf{a}$라 하자.

$$\frac{\partial y}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^t\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^t\mathbf{a}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial \mathbf{a}^t\mathbf{x}}{\partial x_1} \\ \frac{\partial \mathbf{a}^t\mathbf{x}}{\partial x_2} \\ \vdots \\ \frac{\partial \mathbf{a}^t\mathbf{x}}{\partial x_p} \end{bmatrix} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \mathbf{a}$$

B.3.5 선형변환에 대한 미분

행렬 $\mathbf{A}$를 $q \times p$ 행렬이라고 하고 $\mathbf{y} = \mathbf{A}\mathbf{x}$라 하자. 여기서 행렬 $\mathbf{A}$를 다음과 같이 나타내자.

$$\mathbf{A} = \begin{bmatrix} \mathbf{a}_1^t \\ \mathbf{a}_2^t \\ \vdots \\ \mathbf{a}_q^t \end{bmatrix} \text{ or } \mathbf{A}^t = [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_q]$$

위의 내적에 대한 미분 결과를 이용하면 다음은 결과를 얻는다.

$$\begin{aligned} \frac{\partial \mathbf{y}}{\partial \mathbf{x}} &= \frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} \\ &\equiv \frac{\partial \mathbf{x}^t \mathbf{A}^t}{\partial \mathbf{x}} \\ &= \frac{\partial}{\partial \mathbf{x}}[\mathbf{x}^t\mathbf{a}_1\ \mathbf{x}^t\mathbf{a}_2\ \cdots\ \mathbf{x}^t\mathbf{a}_q] \\ &= [\frac{\partial \mathbf{x}^t\mathbf{a}_1}{\partial \mathbf{x}}\ \frac{\partial \mathbf{x}^t\mathbf{a}_2}{\partial \mathbf{x}}\ \cdots\ \frac{\partial \mathbf{x}^t\mathbf{a}_q}{\partial \mathbf{x}}] \\ &= [\mathbf{a}_1\ \mathbf{a}_2\ \cdots\ \mathbf{a}_q] \\ &= \mathbf{A}^t \end{aligned}$$

위의 결과를 응용하면 다음의 결과를 얻는다.

$$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \mathbf{A}^t \quad \text{and} \quad \frac{\partial \mathbf{x}^t\mathbf{A}}{\partial \mathbf{x}} = \mathbf{A}$$

B.3.5.1 이차형식

$$\frac{\partial \mathbf{x}^t\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = \frac{\partial \mathbf{x}^t}{\partial \mathbf{x}}\mathbf{A}\mathbf{x} + \frac{\partial \mathbf{x}^t\mathbf{A}^t}{\partial \mathbf{x}}\mathbf{x} = \mathbf{A}\mathbf{x} + \mathbf{A}^t\mathbf{x}$$

만약 행렬 $\mathbf{A}$가 대칭이면

$$\frac{\partial \mathbf{x}^t\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = 2\mathbf{A}\mathbf{x}$$

B.4 합성함수에 대한 미분공식

두 개의 다변량 함수를 고려하고

$$g : \Re^p \to \Re^q, \quad f : \Re^q \to \Re^r$$

$$
\begin{aligned}
\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{x}))}{\partial \mathbf{x}} &= \frac{\partial \mathbf{f}^t(\mathbf{g}(\mathbf{x}))}{\partial \mathbf{x}} \\
&= \left[\frac{\partial f_1(\mathbf{g}(\mathbf{x}))}{\partial \mathbf{x}}, \frac{\partial f_2(\mathbf{g}(\mathbf{x}))}{\partial \mathbf{x}}, \cdots, \frac{\partial f_r(\mathbf{g}(\mathbf{x}))}{\partial \mathbf{x}} \right] \\
&= \begin{bmatrix} \frac{\partial f_1(\mathbf{g}(\mathbf{x}))}{\partial x_1} & \frac{\partial f_2(\mathbf{g}(\mathbf{x}))}{\partial x_1} & \cdots & \frac{\partial f_r(\mathbf{g}(\mathbf{x}))}{\partial x_1} \\ \frac{\partial f_1(\mathbf{g}(\mathbf{x}))}{\partial x_2} & \frac{\partial f_2(\mathbf{g}(\mathbf{x}))}{\partial x_2} & \cdots & \frac{\partial f_r(\mathbf{g}(\mathbf{x}))}{\partial x_2} \\ & & \vdots & \\ \frac{\partial f_1(\mathbf{g}(\mathbf{x}))}{\partial x_p} & \frac{\partial f_2(\mathbf{g}(\mathbf{x}))}{\partial x_p} & \cdots & \frac{\partial f_r(\mathbf{g}(\mathbf{x}))}{\partial x_p} \end{bmatrix} \\
&= \begin{bmatrix} \sum_{k=1}^q \frac{\partial f_1}{\partial g_k}\frac{\partial g_k}{\partial x_1} & \sum_{k=1}^q \frac{\partial f_2}{\partial g_k}\frac{\partial g_k}{\partial x_1} & \cdots & \sum_{k=1}^q \frac{\partial f_r}{\partial g_k}\frac{\partial g_k}{\partial x_1} \\ \sum_{k=1}^q \frac{\partial f_1}{\partial g_k}\frac{\partial g_k}{\partial x_2} & \sum_{k=1}^q \frac{\partial f_2}{\partial g_k}\frac{\partial g_k}{\partial x_2} & \cdots & \sum_{k=1}^q \frac{\partial f_r}{\partial g_k}\frac{\partial g_k}{\partial x_2} \\ & & \vdots & \\ \sum_{k=1}^q \frac{\partial f_1}{\partial g_k}\frac{\partial g_k}{\partial x_p} & \sum_{k=1}^q \frac{\partial f_2}{\partial g_k}\frac{\partial g_k}{\partial x_p} & \cdots & \sum_{k=1}^q \frac{\partial f_r}{\partial g_k}\frac{\partial g_k}{\partial x_p} \end{bmatrix} \\
&= \sum_{k=1}^q \begin{bmatrix} \frac{\partial f_1}{\partial g_k}\frac{\partial g_k}{\partial x_1} & \frac{\partial f_2}{\partial g_k}\frac{\partial g_k}{\partial x_1} & \cdots & \frac{\partial f_r}{\partial g_k}\frac{\partial g_k}{\partial x_1} \\ \frac{\partial f_1}{\partial g_k}\frac{\partial g_k}{\partial x_2} & \frac{\partial f_2}{\partial g_k}\frac{\partial g_k}{\partial x_2} & \cdots & \frac{\partial f_r}{\partial g_k}\frac{\partial g_k}{\partial x_2} \\ & & \vdots & \\ \frac{\partial f_1}{\partial g_k}\frac{\partial g_k}{\partial x_p} & \frac{\partial f_2}{\partial g_k}\frac{\partial g_k}{\partial x_p} & \cdots & \frac{\partial f_r}{\partial g_k}\frac{\partial g_k}{\partial x_p} \end{bmatrix} \\
&= \sum_{k=1}^q \begin{bmatrix} \frac{\partial g_k}{\partial x_1} \\ \frac{\partial g_k}{\partial x_2} \\ \vdots \\ \frac{\partial g_k}{\partial x_p} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial g_k} & \frac{\partial f_2}{\partial g_k} & \cdots & \frac{\partial f_r}{\partial g_k} \end{bmatrix} \\
&= \sum_{k=1}^q \frac{\partial g_k}{\partial \mathbf{x}} \frac{\partial \mathbf{f}}{\partial g_k} \\
&= \begin{bmatrix} \frac{\partial g_1}{\partial \mathbf{x}} \frac{\partial g_2}{\partial \mathbf{x}} \cdots \frac{\partial g_q}{\partial \mathbf{x}} \end{bmatrix} \begin{bmatrix} \partial \mathbf{f}/\partial g_1 \\ \partial \mathbf{f}/\partial g_2 \\ \vdots \\ \partial \mathbf{f}/\partial g_q \end{bmatrix} \\
&= \frac{\partial \mathbf{g}}{\partial \mathbf{x}} \frac{\partial \mathbf{f}}{\partial \mathbf{g}} \quad (p \times q)(q \times r)
\end{aligned}
$$

특별히 f가 일변량인 경우($r = 1$),

$$\frac{\partial f(\mathbf{g}(\mathbf{x}))}{\partial \mathbf{x}} = \frac{\partial \mathbf{g}}{\partial \mathbf{x}} \frac{\partial f}{\partial \mathbf{g}} \quad (p \times q)(q \times 1)$$

더 나아가 다음도 보일 수 있다.

$$\frac{\partial \mathbf{f}(\mathbf{g}(\mathbf{u}(\mathbf{x})))}{\partial \mathbf{x}} = \frac{\partial \mathbf{u}}{\partial \mathbf{x}} \frac{\partial \mathbf{g}}{\partial \mathbf{u}} \frac{\partial \mathbf{f}}{\partial \mathbf{g}}$$

참고 문헌

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384.

Searle, S. R. and McCulloch, C. E. (2001). *Generalized, linear and mixed models*. Wiley.

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.