Journal of the Korean
Data & Information Science Society
2006, Vol. 17, No. 4, pp. 1053~1065

A Study on One Factorial Longitudinal Data Analysis with Informative Drop-out

Ki Hoon Lee[1)]

Abstract

This paper proposes a method in one-way layouts for longitudinal data with informative drop-out. When dropouts are informative, that is, correlated with unobserved data and/or the previous observed data, the simple imputation methods such as 'last observation carried forward' (LOCF) methods would arise the bias of the testing models. The maximum likelihood procedure combined with a logit model for the drop-out process is proposed to test treatment effects for one factorial designs and compared with LOCF method in two examples.

Keywords : Drop-out, Longitudinal data, Maximum likelihood

1. 서론

대부분 의학적인 임상실험(clinical trial)에서는 자료를 경시적(다시점, longitudinal) 방법으로 수집하게 된다. 이는 연구목적이 경시적 자료분석을 위한 것이 아니고 일반적인 단변량 분산분석인 경우에도 마찬가지이다. 본 논문에서 다루고자하는 처리효과에 대한 일원배치 분산분석을 하고자 할 때도 자료를 다시점으로 수집하고 최종적인 분석에서 마지막 시점에서의 측정자료만을 이용하여 효과분석을 하는 것이 일반적인 임상실험의 분석행태이다. 그런데 일정기간동안 한 개체에서 반복적으로 자료를 수집하게 되면 거의 반드시 일부 실험개체에서 중도탈락(drop-out)이 발생하게 되는데, 이 중도탈락 개체를 어떤 형태로 분석에 포함시킬 것인지 또는 제외할 것인지가 중요한 논점이 된다. 분석하는 대상에 따라 실험설계는 중도탈락을 제외하고 실험조건에 만족하는 개체만을 분석하는 PP(per protocol)와 실험에 참여한 모든 개체를 분석하는 ITT(intention to treat) 등으로 분류할 수 있는데, PP의 경우 중도탈락을 제외하고 분석하지만 ITT의 경우에는 중도탈락하여 비관측된 값을 탈락직전의 값으로 대체하는

1) 전라북도 전주시 완산구 효자동 3가 전주대학교 경영학부 교수
E-mail : khlee@jj.ac.kr

LOCF(last observation carried forward) 방법이 주로 사용되어 왔다.

그러나 이러한 방법들은 중도탈락의 형태에 따라 그 유효성이 판단될 수 있을 것이다. Rubin(1976)과 Little과 Rubin(1987)은 어떤 시점 t에서 결측값(missing)이 발생하였을 때 결측값을 다음과 같이 세 가지 종류로 분류하였다.

1) 완전무작위결측(MACR : missing completely at random): t시점에서 결측은 전에 관측된 값들과 비관측된 t시점의 값과 무관하다.

2) 무작위결측 (MAR : missing at random) : t시점에서의 결측은 전에 관측된 값들과는 관계가 있으나 비관측된 t시점의 값과는 무관하다.

3) 무시할 수 없는 비 응답(MNI : non-ignorable non-response) : t시점에서의 결측은 전의 관측값들과 관계가 있고 t시점에서의 비관측값과 관련될 가능성이 있다.

Diggle과 Kenward(1994)는 이를 중도탈락에 접목하여 중도탈락을 각각 완전무작위 중도탈락(CRD : completely random drop-out), 비관측값과 독립이지만 기존의 값들에는 종속된 무작위 중도탈락(RD : random drop-out), 기존의 값들과 비관측값 모두에 종속된 정보형 중도탈락(ID : informative drop-out) 등의 세 가지로 분류하였다.

임상실험에서 중도탈락(drop-out)은 여러 가지 이유로 발생하는데, 대부분 사망, 질병, 치료중단, 치료효과미비, 추적실패, 연구종료 등이 원인이다. 생존이 주요 끝점(endpoint)인 경우, 환자들은 대체로 병환중이기 때문에 질병이나 사망이 중도탈락의 주요 원인이 될 수 있고 중도탈락이 비관측된 값(질병의 발생, 악화, 사망)과 연관이 있으면 이를 정보형 중도탈락이라 규정할 수 있다. 치료효과가 주요 끝점인 경우에도 치료효과미비로 인한 실망으로 인해 중도포기, 또는 기존 치료로 복귀하는 경우 이도 역시 완전 무작위 중도탈락이라 볼 수 없다. 그런데 이러한 결측 패턴을 모형에 포함하지 않고 단순한 대체(imputation)에 의한 분석을 할 경우 추론에 편의(bias)가 존재하고 검정력에 손실이 있음이 Gould(1980), Pledger와 Hall(1982), Laird(1988), Zwinderman(1992), Myers(2000), Hogan, Roy와 Korkontzelou(2004), Roy와 Lin(2005) 등에 의해서 언급되었다. 특히 Shao와 Zhong(2003)는 정보형 중도탈락이 존재할 때 LOCF 방법을 사용하면 검정력이 떨어짐을 보였다.

결측값이 완전 무작위일 때는 일반적인 통계방법을 사용하여도 무방하고, 무작위(MAR)일 때도 Murray와 Findlay(1988)에 의하면 우도함수(likelihood function)에 기초하지 않은 방법은 편의가 존재하지만 우도함수에 기초한 분석법을 사용한다면 결합우도함수가 완전관측자료와 결측과정 등의 두 부분의 곱으로 인수분해할 수 있기 때문에 결측과정이 무시될 수 있다고 한다. 본 논문에서는 완전관측자료와 결측과정이 독립적이지 않은 ID인 경우 우도함수를 이용한 분석방법을 고찰하고자 한다. 2장에서는 Diggle과 Kenward(1994), Verbyla와 Cullis(1990)의 방법을 응용하여 주어진 자료모형에서의 우도함수를 유도하고 이를 구체적으로 계산하는 방법을 설명하였고, 3장에서는 두 가지 예제에 이를 적용하여 일반적인 LOCF 일원배치와 비교하고 그 특성을 설명하였다.

2. 제안한 통계량

2.1 자료 모형과 가설

본 논문에서 고려하고자 하는 일원배치 자료는 처리(treatment) 수가 r이고, 각 처리에 n개의 개체가 있다고 가정하고, 자료를 경시적으로 수집하였으므로 각 개체는 l회 시점에서 반복 측정한다. 이때, i번째 처리의 k번째 개체의 t시점의 관측값은 다음과 같이 모형화 할 수 있다.

$$\begin{aligned} y_k &= \mu + e_k \\ &= \mu + \alpha_{i1} + \alpha_{i2} + \cdots + \alpha + e_k \,;\, i = 1,...,r;\, t = 1,...,l;\, k = 1,...,n. \end{aligned} \tag{2.1}$$

여기서 α는 i번 째 처리의 관측시점 t에서 발생하는 부가적인 처리효과로서 가산적으로 가정하고, 오차항은 정규분포 $N(0,\, \sigma^2)$를 가정한다.

일원배치분석에서 l시점의 최종 자료 $y_{1lk}, y_{2lk}, \cdots, y_{rlk} (k = 1,...,n)$의 효과에만 관심이 있으므로, 부가적인 처리효과의 합을

$$\alpha_{i1} + \alpha_{i2} + \cdots + \alpha_{il} = \sum_{j=1}^{l} \alpha_{ij} = \tau_i$$

이라 하면 우리가 검정하고자 하는 가설은 다음과 같다.

$$H_0 : \mu_{1l} = \mu_{2l} = \cdots = \mu_{rl} = \mu \ (\tau_1 = \tau_2 = \cdots = \tau_r = 0). \tag{2.2}$$

여기서 $\alpha_{ij}, j = 1,...,l$는 서로 독립이 아니지만 $\tau_i, i = 1,...,r$는 서로 독립이다.

중도탈락이 발생하여 (2.1)의 자료들을 모두 얻을 수 있지 않으므로 관측 가능하여 우리가 얻은 (2.1)에 대응하는 관측값을 다음과 같이 표시한다.

$$x_k = \begin{cases} y_k, & t = 1, \cdots, d-1 \\ 0, & t = d, \cdots, l\,. \end{cases} \tag{2.3}$$

여기서 $d (d = 2, \cdots, l+1)$는 중도탈락이 발생하는 시점이다. 단, 중도탈락이 발생하지 않으면 $d = l+1$. 예를 들어 어떤 개체가 d시점에 중도탈락이 발생하면 그 개체의 원자료와 관측값의 관계를 다음과 표현할 수 있다.

$$\boldsymbol{x_{ik}} = (x_{i1k}, x_{i2k}, \cdots, x_{ilk})' = (y_{i1k}, \cdots, y_{i,d-1,k}, 0, \cdots, 0)'.$$

2.2 우도함수의 유도

관측값들의 우도함수(likelihood function)를 구하면 처리간, 개체간에는 서로 독립이

므로 다음과 같이 표시할 수 있다.

$$f(\boldsymbol{x_{11}}, \cdots, \boldsymbol{x_{1n}}, \boldsymbol{x_{21}}, \cdots, \boldsymbol{x_{2n}}, \cdots, \boldsymbol{x_{r1}}, \cdots, \boldsymbol{x_{rn}}) = \prod_{i=1}^{r}\prod_{k=1}^{n} f(\boldsymbol{x_{ik}}).$$

여기서 한 개체의 관측값만을 분리하여 첨자의 복잡성을 피하여 한 개체 내에서 경시적 자료벡터를 $\boldsymbol{x_{ik}} = \boldsymbol{x}$라 표시하고 원자료의 형태를 이용하여 우도함수를 다시 적으면 다음과 같다.

첫째 중도탈락 없이 모든 값이 관측되었을 때, 즉 $\boldsymbol{x} = (x_1, x_2, \cdots, x_l)' = (y_1, y_2, \cdots, y_l)'$일 때

$$\begin{aligned} f(\boldsymbol{x}) &= f_1(x_1) f_2(x_2|x_1) f_3(x_3|x_1, x_2) \cdots f_l(x_l|x_1, \cdots, x_{l-1}) \\ &= f_1(y_1) f_2(y_2|y_1)(1-p_2) f_3(y_3|y_1, y_2)(1-p_3) \cdots f_l(y_l|y_1, \cdots, y_{l-1})(1-p_l) \\ &= f(y_1, y_2, \cdots, y_l) \prod_{t=1}^{l} (1-p_t). \end{aligned}$$

여기서 p_t는 t시점에서 중도탈락이 처음 발생할 확률로서 변수 D를 중도탈락 시점이라 하면 $p_t = \Pr(D = t | x_{t-1} \neq 0) = \Pr(x_t = 0 | x_{t-1} \neq 0)$와 같이 표현할 수 있다. 그리고 $p_1 = 0$이라 가정하고, 중도탈락이 없으면($p_t = 0, \forall t$), $f(\boldsymbol{x}) = f(\boldsymbol{y})$이 된다.

둘째로 만약 d시점에 중도탈락이 발생해 $\boldsymbol{x} = (y_1, y_2, \cdots, y_{d-1}, 0, \cdots, 0)'$이 되면 이때 우도함수는

$$\begin{aligned} f(\boldsymbol{x}) &= f(x_1, x_2, \cdots, x_l) \\ &= f_1(y_1) f_2(y_2|y_1) \cdots f_{d-1}(y_{d-1}|y_1, \cdots, y_{d-2}) \prod_{t=1}^{d-1} (1-p_t) \cdot \Pr(x_d = 0 | x_{d-1} \neq 0) \\ &= f(y_1, y_2, \cdots, y_{d-1}) \prod_{t=1}^{d-1} (1-p_t) \cdot \Pr(x_d = 0 | x_{d-1} \neq 0). \end{aligned}$$

그러므로 $\boldsymbol{x_{ik}}$개체에서 중도탈락 시점이 d_{ik}라 가정하면 전체 모형에 대한 로그우도함수(log-likelihood function)는 일반적으로 다음과 같이 표시할 수 있다.

$$\begin{aligned} \log f(\boldsymbol{x_{11}}, \cdots, \boldsymbol{x_{1n}}, \boldsymbol{x_{21}}, \cdots, \boldsymbol{x_{2n}}, \cdots, \boldsymbol{x_{r1}}, \cdots, \boldsymbol{x_{rn}}) &= \sum_{i=1}^{r}\sum_{k=1}^{n} \log f(\boldsymbol{x_{ik}}) \\ &= \sum_{i=1}^{r}\sum_{k=1}^{n} \log f(y_1, y_2, \cdots, y_{d_{ik}-1}) + \sum_{i=1}^{r}\sum_{k=1}^{n}\sum_{t=1}^{d_{ik}-1} \log(1-p_t) \\ &\quad + \sum_{i=1}^{r}\sum_{k=1}^{n} \log \Pr(x_{d_{ik}} = 0 | x_{d_{ik}-1} \neq 0) \\ &= (1) + (2) + (3). \end{aligned}$$

(1)에서 로그함수부분을 정리하면 다음과 같다.

$$\log f(y_1, y_2, \cdots, y_{d_{ik}-1}) = [-(d_{ik}-1)\log 2\pi - \log|\Sigma_{ik}| - (\boldsymbol{y_{ik}} - \boldsymbol{\mu_{ik}})' \Sigma_{ik}^{-1} (\boldsymbol{y_{ik}} - \boldsymbol{\mu_{ik}})]/2.$$

여기서 $\boldsymbol{y_{ik}} = (y_{i1k}, \cdots, y_{i,d_{ik}-1,k})'$는 관측 가능한 자료의 $(d_{ik}-1)\times 1$ 벡터이고 $\boldsymbol{\mu_{ik}}$와 Σ_{ik}는 각각 이에 대응하는 기대값 벡터와 공분산 행렬을 표시한다.

(2)식을 정리하기 위하여 중도탈락에 관한 로짓확률(logit)을 다음과 같이 표현한다.

$$\log\left(\frac{p_t}{1-p_t}\right) = \beta_{t0} + \beta_1 y_t + \beta_2 y_{t-1} + \cdots + \beta_t y_1$$
$$= \beta_{t0} + \sum_{m=1}^{t} \beta_m y_{t+1-m}.$$

그러므로 (2)에서 $p_t\,(t<d)$는 중도탈락 전까지의 범위만 포함하므로 다음과 같이 비관측된 $y_{d_{ik}}$값에 무관하게 구할 수 있다.

$$(2) = -\sum_{i=1}^{r}\sum_{k=1}^{n}\sum_{t=2}^{d_{ik}} \log[1+\exp(\beta_{t0} + \sum_{m=2}^{t} \beta_{im}\, y_{t+1-m})].$$

여기서 $\beta,\ (i=1,\cdots,r;\ t=1,\cdots,l)$는 i번째 처리그룹에서 logit에 관한 회귀계수이다.

그러나 (3)은 비관측된 $y_{d_{ik}}$값에 의존하는데 이를 정리하기 위해 중도탈락시점 d에서의 logit을 $\widehat{y_d}$을 사용하여 다음과 같이 표현한다.

$$\log\left(\frac{p_d}{1-p_d}\right) = \beta_{d0} + \beta_1 y_d + \beta_2 y_{d-1} + \cdots + \beta_d y_1$$
$$\simeq \beta_{d0} + \beta_1 \widehat{y_d} + \sum_{m=2}^{d} \beta_m y_{d+1-m}.$$

만약 중도탈락의 형태가 RD나 CRD이면 비관측된 y_d와 중도탈락확률이 무관하므로 (3)의 확률을 다음과 같이 표시할 수 있다.

$$\log \Pr(x_d = 0 \mid x_{d-1} \neq 0) = \beta_{d0} + \sum_{m=2}^{d} \beta_m y_{d+1-m} - \log[1+\exp(\beta_{d0} + \sum_{m=2}^{d} \beta_m y_{d+1-m})].$$

여기서 CRD인 경우는 β_{d0}를 제외한 나머지 회귀계수가 0이 된다.

그러므로 RD나 CRD인 경우 (3)식은 다음과 같이 정리된다.

$$(3) = \sum_{i=1}^{r}\sum_{k=1}^{n} \log \Pr(x_{d_{ik}} = 0 \mid x_{d_{ik}-1} \neq 0)$$
$$= \sum_{i=1}^{r}\sum_{k=1}^{n} (\beta_{d_{ik}0} + \sum_{m=2}^{d_{ik}} \beta_{im} y_{i,d_{ik}+1-m} - \log[1+\exp(\beta_{d_{ik}0} + \sum_{m=2}^{d_{ik}} \beta_{im} y_{i,d_{ik}+1-m})]).$$

처리간의 효과를 검정하기 위해서는 로그우도함수를 최대화하여 추정하는 모수의

수에 따라 자유도를 정하고 카이제곱 검정을 적용할 수 있다. 최대화하는 방법으로는 Nelder와 Mead(1965), Broyden-Fletcher-Goldfarb-Shanno(BFGS) 등이 있다. 이에 관한 자세한 적용은 3장의 예제에서 직접 다루도록 한다.

2.3 정보형 중도탈락의 모형화

만약 중도탈락의 형태가 ID 이면 $\hat{y_d}$를 구하는 문제가 남게 된다. 이를 위하여 각 그룹마다 다음을 가정한다.

$$\boldsymbol{y}_{ik} = (y_{i1k}, y_{i2k}, \cdots, y_{ilk})' \sim N(\boldsymbol{\mu_i}, \Sigma_i), \ k = 1, \cdots, n .$$

여기서 $\boldsymbol{\mu_i} = (\mu_{i1}, \mu_{i2}, \cdots, \mu_{il})'$, Σ_i는 $(l \times l)$ 공분산 행렬이다.

정리 $\boldsymbol{x^{(d)}} = (x_1, x_2, \cdots, x_d)'$가 평균$\boldsymbol{\mu^{(d)}} = (\mu_1, \mu_2, \cdots, \mu_d)'$와 공분산 $\Sigma^{(d)} = ((\sigma_{ij}))_{d \times d}$을 갖는 d변량 정규분포를 따를 때 $f(x_d | (y_1, y_2, \cdots, y_{d-1}))$는 다음과 같은 평균과 분산을 갖는 정규분포 확률밀도함수이다.

$$m_d = \mu_d + (\sigma_{1d}, \sigma_{2d}, \cdots, \sigma_{d-1,d})(\Sigma^{(d-1)})^{-1}(\boldsymbol{x^{(d-1)}} - \boldsymbol{\mu^{(d-1)}}).$$
$$v_d = \sigma_{dd} - (\sigma_{1d}, \sigma_{2d}, \cdots, \sigma_{d-1,d})(\Sigma^{(d-1)})^{-1}(\sigma_{1d}, \sigma_{2d}, \cdots, \sigma_{d-1,d})'.$$

중도탈락이 ID인 경우 (3)의 식에 y_d를 포함해야 하는데 위의 정리를 이용하여 이를 m_d로 대체하면 (3)식을 일반적으로 다음과 같이 표현할 수 있다.

$$\begin{aligned}(3) &= \sum_{i=1}^{r}\sum_{k=1}^{n} \log \Pr(x_{d_{ik}} = 0 \mid x_{d_{ik}-1} \neq 0) \\ &= \sum_{i}^{r}\sum_{k=1}^{n} (\beta_{d_{ik}0} + \beta_{i1} m_d + \sum_{m=2}^{d_{ik}} \beta_{im} y_{i,d_{ik}+1-m} \\ &\quad - \log[1 + \exp(\beta_{d_{ik}0} + \beta_{i1} m_d + \sum_{m=2}^{d_{ik}} \beta_{im} y_{i,d_{ik}+1-m})]).\end{aligned}$$

그런데 우리가 개체내의 상관관계를 가정하였기 때문에 실제적인 계산에서는 복잡함을 피하기 위해 중도탈락시점의 값과 직전 값만을 사용하여 이를 다음과 같이 간략하게 변형할 수 있다.

$$(3) \simeq \sum_{i}^{r}\sum_{k=1}^{n} (\beta_{i0} + \beta_{i1} y_{i,d_{ik}} + \beta_{i2} y_{i,d_{ik}-1} - \log[1 + \exp(\beta_{i0} + \beta_{i1} y_{i,d_{ik}} + \beta_{i2} y_{i,d_{ik}-1})]).$$

위에 식에 근거하여 로그우도함수의 구체적인 형태를 유도하는 간단한 예를 들어보자. 처리간에 구별 없이 다음 표와 같이 3개의 개체가 각 5번 반복측정된 것을 가정하고 이들의 중도탈락은 각각 '시점 4', '시점 5', '발생 안함' 등이라 하자. 관측값은 y, 비관측된 값은 0으로 표시하였다.

관측시점		1	2	3	4	5
개체	1	y_{11}	y_{12}	y_{13}	0	0
	2	y_{21}	y_{22}	y_{23}	y_{24}	0
	3	y_{31}	y_{32}	y_{33}	y_{34}	y_{35}

첫 번째 개체의 로그우도함수는 다음과 같다.

$$\begin{aligned}\log f(y_1, y_2, y_3, 0, 0) &= \log f(y_1, y_2, y_3) + \log p_4 \\ &= -\frac{3}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_3| - \frac{1}{2}((y_1, y_2, y_3)' - \boldsymbol{\mu_3})'\Sigma_3^{-1}((y_1, y_2, y_3)' - \boldsymbol{\mu_3}) \\ &\quad + \beta_{40} + \beta_1\widehat{y_{14}} + \beta_2 y_{13} - \log(1 + \exp(\beta_{40} + \beta_1\widehat{y_{14}} + \beta_2 y_{13})).\end{aligned}$$

두 번째 개체의 로그우도함수는 다음과 같다.

$$\begin{aligned}\log f(y_1, y_2, y_3, y_4, 0) &= \log f(y_1, y_2, y_3, y_4) + \log(1 - p_4) + \log p_5 \\ &= -\frac{4}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_4| - \frac{1}{2}((y_1, \cdots, y_4)' - \boldsymbol{\mu_4})'\Sigma_4^{-1}((y_1, \cdots, y_4)' - \boldsymbol{\mu_4}) \\ &\quad - \log(1 + \exp(\beta_{40} + \beta_1 y_{24} + \beta_2 y_{23})) \\ &\quad + \beta_{50} + \beta_1\widehat{y_{25}} + \beta_2 y_{24} - \log(1 + \exp(\beta_{50} + \beta_1\widehat{y_{25}} + \beta_2 y_{24})).\end{aligned}$$

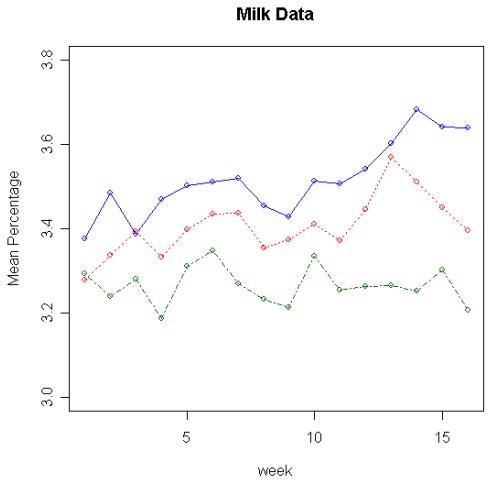
마지막으로 세 번째 개체의 로그우도함수는 다음과 같이 표시할 수 있다.

$$\begin{aligned}\log f(y_1, y_2, y_3, y_4, y_5) &= \log f(y_1, y_2, y_3, y_4, y_5) + \log(1 - p_4) + \log(1 - p_5) \\ &= -\frac{5}{2}\log 2\pi - \frac{1}{2}\log|\Sigma_5| - \frac{1}{2}((y_1, \cdots, y_5)' - \boldsymbol{\mu_5})'\Sigma_5^{-1}((y_1, \cdots, y_5)' - \boldsymbol{\mu_5}) \\ &\quad - \log(1 + \exp(\beta_{40} + \beta_1 y_{34} + \beta_2 y_{33})) - \log(1 + \exp(\beta_{50} + \beta_1 y_{35} + \beta_2 y_{34})).\end{aligned}$$

3. 예제

3.1 우유의 단백질 비율 자료

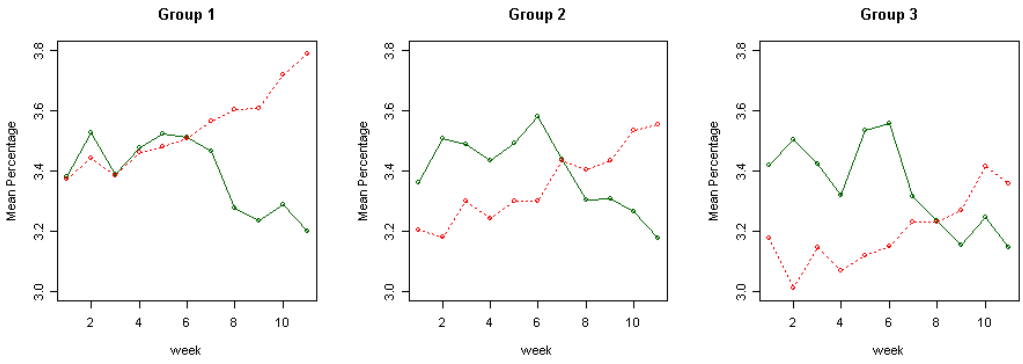
이절에서는 Verbyla와 Cullis(1990)가 인용한 후 Diggle과 Kenward(1994), Thijs, Molenberghs와 Verbeke(2000) 등에 의해 분석되어졌던 우유 단백질 자료를 분석하고자 한다. 이는 79마리의 소에게 3종류의 사료를 먹인 후 이들이 생산하는 우유의 단백질 자료를 조사한 것으로 79마리를 각각 25, 27, 27마리의 세 군으로 나누어 각각 보리(사료 1), 보리-루핀혼합(사료 2), 루핀(사료 3) 사료를 19주간 투여하고 이들이 생산하는 우유의 단백질 비율을 측정하였다. 그리고 각 사료간에 단백질 비율의 차이가 있는지를 검정하였다. 최초 3주를 제외한 16주간의 자료를 이용하였는데 원자료의 각 사료군 별 16주간의 평균변화 추이도는 다음 그림과 같다.



<그림 1> 사료에 따른 우유 단백질 비율 평균 변화 추이
(── : 사료1; ------ : 사료 2; -·-·- : 사료 3)

19주간의 실험이 종료되기 전에 질병 또는 노화, 사망 등으로 더 이상 우유를 생산 못하는 경우가 38마리로 전체 실험대상의 48%를 차지하고 있다. 이들 자료에서 중도탈락을 빼고 실험이 완료된 자료만을 사용한 PP(per protocol)의 경우 분산분석 결과 $F(2,38) = 6.5255$로 유의확률이 0.003663이었고, 중도탈락한 결측값을 LOCF로 대체하여 분석하는 ITT(intention to treat)의 경우 $F(2, 76) = 3.9398$ 로 유의확률이 0.02355이다. 즉 PP인 경우 사료의 효과가 1%의 유의수준에서 유의하지만 ITT인 경우는 5% 유의수준에서 유의한 약간 상이한 결과를 보여준다.

중도탈락의 형태(pattern)를 조사하기 위해 중도탈락이 한건도 발생하지 않은 11주차까지 중도탈락한 개체와 실험을 완료한 개체들을 나누어 평균변화 추이를 그려보았는데 다음 <그림 2>와 같다. 즉 중도탈락한 개체들의 단백질 비율은 변동이 심하고 6주 이후에는 비율이 꾸준히 줄어들고 있음을 알 수 있다.



<그림 2> 사료군 별 탈락개체와 실험완료개체 평균 변화 추이
(── : 중도탈락개체; ------ : 실험완료개체)

Diggle과 Kenward(1994)에서와 같이 (2.1)의 모형에서 처리효과는 실험기간 동안 일정하다고 가정하고 다음과 같이 정의한다.

$$\alpha_{ij} = \begin{cases} \alpha_i, & j = 1 \\ 0, & j \geq 2, \end{cases}$$

또는

$$\alpha_{i1} = \alpha_i, \quad \alpha_{i2} = \alpha_{i3} = \cdots = \alpha_{il} = 0\,, \quad i = 1, 2, \cdots, r\,.$$

또한 검정하고자 하는 귀무가설은 다음과 같다.

$$H_0 : \mu_{1l} = \mu_{2l} = \cdots = \mu_{rl} = \mu \ (\alpha_1 = \alpha_2 = \cdots = \alpha_r = 0).$$

이때 검정통계량으로 사용되는 2배의 로그우도함수는 다음과 같다.

$$\begin{aligned} 2L(\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}) = & -\sum_{k \in k^*} k \log 2\pi - \Big(\sum_{k \in k^*} \log |V^{(k)}| + \sum_{k \in k^*} (\boldsymbol{y^{(k)}} - \boldsymbol{\mu^{(k)}})' [V^{(k)}]^{-1} (\boldsymbol{y^{(k)}} - \boldsymbol{\mu^{(k)}}) \\ & - 2 \sum_{11 \leq k < k^*} \log(1 + \exp(\beta_{k0} + \beta_1 y_{k+1} + \beta_2 y_k)) \\ & + 2 \sum_{k \in k^*} [\beta_{k0} + \beta_1 y_{k+1} + \beta_2 y_k - \log(1 + \exp(\beta_{k0} + \beta_1 y_{k+1} + \beta_2 y_k))], \end{aligned}$$

여기서 $k^* = d - 1$ (d는 탈락시점)으로 본 자료에서는 11, 13, 14, 15 등의 시점에서 탈락이 발생하였다. 그리고 $2L(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})|_{H_1} - 2L(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n})|_{H_0} \sim \chi^2(2)$가 된다.

<표 1> 우유 단백질 자료의 추정량과 우도함수값

	μ_1	μ_2	μ_3	$\beta_{0,11}$	$\beta_{0,13}$	$\beta_{0,14}$	$\beta_{0,15}$	β_1	β_2	$2L(\boldsymbol{x})$
H_1	3.37	3.30	3.12	16.21	15.00	14.13	15.64	5.03	-10.50	390.95
H_0	μ= 3.263119			16.21	15.00	14.13	15.64	5.03	-10.50	344.63
p-값	$\Pr(\chi^2(2) \leq 390.95 - 344.63) = 8.95211e-24$									

<표 1>에 의하면 평균의 동일성 검정은 LOCF와는 달리 매우 작은 유의수준으로 사료간 처리효과가 존재함을 지지하고 있다. 각 가설하에서 탈락에 관한 로짓확률의 회귀계수 추정값은 거의 같아 소수 5자리 이하에서 다르게 나올 뿐이었다. 일반적인 검정과 현격한 차이가 나는 이유는 사료 1군에서는 탈락개체와 완료개체의 초기 단백질 비율이 거의 동일한데 반해 사료 2, 3 군에서는 초반에 단백질비율이 높은 소는 점점 낮아지며 탈락하고 초반에 낮은 소가 점점 증가하며 완료하는 사실을 모형에 반영함으로써 뚜렷하게 처리효과를 도출해 낼 수 있기 때문이다.

이와 같은 모형에서 중도탈락형태에 따른 가설도 검정할 수 있다. 중도탈락 형태가 CRD이면 '$\beta_1 = \beta_2 = 0$'을 의미하는데 이때 $2L(\boldsymbol{x})$가 301.2233이 되어 두 회귀계수를 포함한 ID 경우인 390.95와 큰 차이가 나서 CRD가 적합하지 않음을 알 수 있다. 또

한 RD, 즉 '$\beta_1 = 0$'인 경우, $2L(\boldsymbol{x})$=367.9093, 비관측값에만 의존하는 '$\beta_2 = 0$'인 경우 $2L(\boldsymbol{x})$=304.1236 등으로 두 회귀계수가 매우 유의하게 0이 아니므로 중도탈락의 모형이 ID가 적합함을 알 수 있다. 이는 <그림 2>에서 중도탈락 자료의 추이가 실험완료 자료의 추이와 완연히 다르다는 점에서 예상할 수 있는 결과이다.

직관적으로 β_1, β_2의 추정값이 유사하게 나와야 함에도 불구하고 서로 다른 부호의 값이 나온 것은 자료들의 종속성 때문이다. Diggle과 Kenward(1994)가 제안한 바와 같이 다음과 같이 고치면 해석이 용이하다.

$$\theta_1 = (\beta_1 + \beta_2)/2\,,\ \theta_2 = (\beta_1 - \beta_2)/2\,.$$

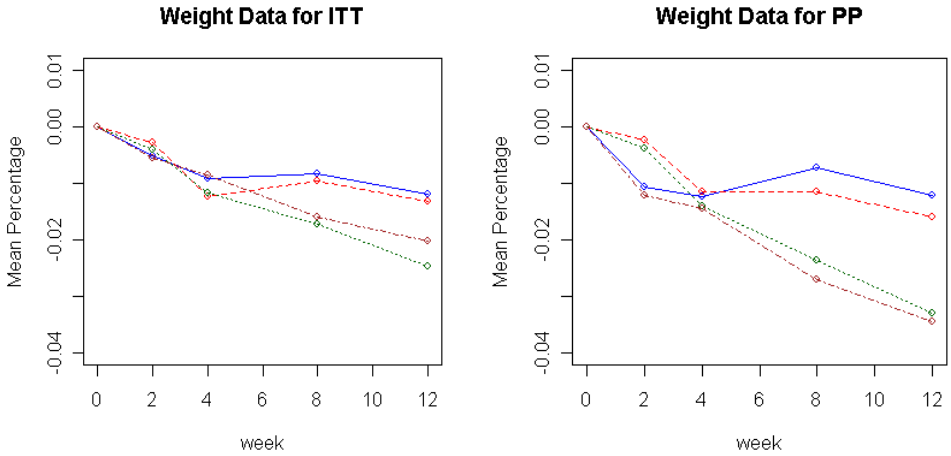
여기서 θ_1는 전체적인 값이 중도탈락에 미치는 영향을 의미하고 θ_2는 마지막 관측된 값과 비교하여 비관측값이 중도탈락에 주는 영향을 표현하고 있다.

이러한 변환을 자료에 적용하여 분석하면 이들 추정값이 $\widehat{\theta_1}$=−2.74, $\widehat{\theta_2}$= 7.9이므로 중도탈락확률은 단백질비율이 적어지거나 직전에 비해 단백질 비율이 증가할 때 높아짐을 추측할 수 있다.

3.2 비만 치료제 자료

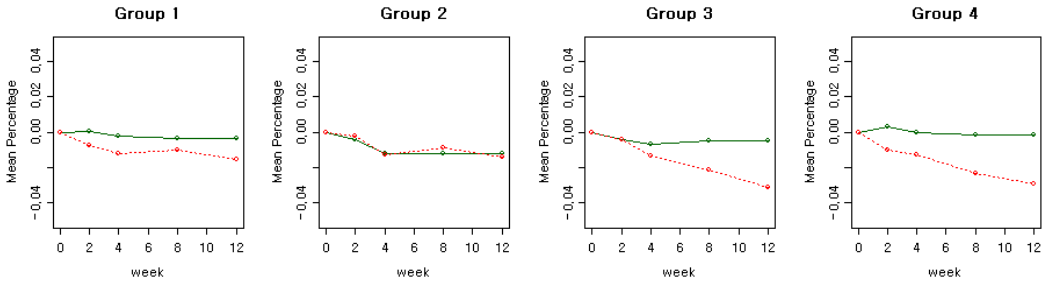
이 절에서 분석하고자 하는 자료는 어떤 비만치료제의 체중감소 효과를 측정하고자 119명의 자원자를 4개의 투약군(위약군, 600㎎, 1200㎎, 1800㎎)으로 나눠 12주간 약을 투여하고 체중변화를 조사한 것이다. 환자가 치료차 병원방문시 이루어지는 강제적인 조사가 아니고 자원자의 자발적인 참여에 의존하기 때문에 다른 임상실험에 비해 많은 중도탈락이 발생하여 최종적으로 67명만이 실험을 마치었다. 이들 중도탈락의 원인은 여러 가지로 찾을 수 있으나 가장 설득력 있는 원인은 효과미비에 의한 실망으로 인한 이탈, 체중감소 효과는 있으나 약물 투여에 대한 부담, 기타 질병에 의한 이탈 등을 들 수 있다. 이들 자료를 각 투약군 별로 체중감소추이를 보면 다음 <그림 3>과 같다.

중도탈락한 개체의 결측값을 최종측정값으로 대체한 LOCF의 경우보다 실험을 완료한 개체만을 분석한 PP의 경우에 도표상에서 처리군들간의 차이가 뚜렷이 보이고 있다. 그러나 이러한 도표상의 결과에 비해 처리간 효과를 일원배치 분산분석한 결과는 ITT인 경우 F(3, 115)=1.2272, p-값=0.3031, PP일 때 F(3,63)=1.8519, p-값=0.1469으로 둘 다 유의수준 10%에서 유의한 차이를 보이지 못하고 있다



<그림 3> 실험에 참여한 모든 개체(ITT)와 실험을 완료한 개체(PP)의 투약군별 체중감소율 평균 변화 추이
(── : 위약군; ------ : 600mg군; -·-·- : 1200mg군; ·····: 1800mg군)

중도탈락의 형태를 조사하기 위해 결측값을 직전 값으로 대체한 후 12주차까지 중도탈락한 개체와 실험을 완료한 개체들을 나누어 평균변화 추이를 다음 <그림 4>에 도시하였다. 그림에서 보면 대부분의 중도탈락이 치료효과의 미비로 추측할 수 있어 중도탈락이 RD 또는 ID로 예상된다. 그러나 앞의 자료와는 달리 각 투약군 별로 중도탈락의 패턴이 거의 동일하고 1200㎎, 1600㎎ 투약군에서는 오히려 효과가 적은 환자가 이탈함으로써 PP로 분석시 처리효과의 존재를 과장시킬 여지가 있다.



<그림 4> 투약군 별 탈락개체와 실험완료개체 평균 변화 추이
(── : 중도탈락개체; ------ : 실험완료개체)

검정하고자 하는 귀무가설은 (2.2)와 같고, 중도탈락 시점은 $d=2,3,4$(각 4, 8, 12주)이므로 4개의 β_0계수와 각 투약군에 동일한 β_1, β_2를 모형에 포함시켰다. 각 시점에서의 처리효과보다는 최종적인 처리효과의 합에 관심 있으므로 귀무가설과 대립가

설의 모수의 개수차이는 3으로 $2L(\boldsymbol{x_1},\cdots,\boldsymbol{x_n})|_{H_1} - 2L(\boldsymbol{x_1},\cdots,\boldsymbol{x_n})|_{H_0} \sim \chi^2(3)$인 사실을 이용해 평균차이를 검정한다. 추정된 모수와 로그우도함수의 값은 다음 <표 2>에 주어져있다.

<표 2> 체중변화 자료의 추정량과 우도함수값

	α_1	α_2	α_3	$\alpha_1+\alpha_2+\alpha_3+\alpha_4$	
H_1	0.0049	0.0057	-0.0022	μ_1	0.0131
	0.0027	0.0101	-0.0035	μ_2	0.0143
	0.0041	0.0082	0.0057	μ_3	0.0279
	0.0052	0.0043	0.0084	μ_4	0.0238
H_0	0.0063	0.0073	-0.0001	μ	0.0201
	0.0039	0.0114	-0.0016		
	0.0024	0.0064	0.0034		
	0.0045	0.0035	0.0072		

	β_{20}	β_{30}	β_{40}	β_1	β_2	$2L(\boldsymbol{x})$
H_1	-2.2495	-1.7541	-2.2171	0.9241	-15.7529	2119.8
H_0	-2.2495	-1.7541	-2.2171	0.9269	-15.7577	2115.7
p-값	$\Pr(\chi^2(3) \le 2119.8-2115.7) = 0.42520$					

<표 2>에 따르면 분산분석의 결과와 마찬가지로 평균차이 검정의 유의확률이 0.42로 처리효과가 존재하지 않는 것으로 결론을 내리고 있다. <그림 3>에서 보면 어느 정도 투약효과가 있어 보이지만 <그림 4>에서 중도탈락의 패턴을 보면 치료효과가 없는 환자가 실험에서 이탈하므로 실험이 완료된 환자만을 대상으로 분석하는 PP의 경우 처리효과의 존재가 과장될 가능성이 있었다. 중도탈락을 모형에 포함하여 처리효과를 검출한다고 해서 무조건 유의확률을 낮추는 결과를 가져오지는 않는다는 사실을 보여주는 예라 할 수 있다.

4. 결론

본 논문에서는 처리효과를 검정하는 일원배치 자료분석에서 경시적 방법으로 자료를 수집하고 이들 중 중도탈락이 발생할 경우, 중도탈락의 정보를 우도함수에 포함하여 처리효과를 분석하는 통계방법을 제안하였다. 예제를 통해 구체적으로 우도함수를 어떻게 모형화하고 이를 최적화하여 검정통계량을 유도하는지를 실증하였고, 또한 이러한 분석방법을 결측값을 제거한 방법, 최종자료로 대체한 LOCF방법과 비교하여 그 특성을 살펴보았다. 제안한 방법이 중도탈락 자료의 형태를 추론모형에 포함시키기 때문에 기존의 방법보다는 직관적으로 우수한 검정 결과를 나타냄을 알 수 있었다.

중도탈락의 형태를 ID로 정의하면 비관측된 값이 모형에 포함되어야 하기 때문에 이를 추정하여 우도함수 모형에 포함시켰는데, 이렇게 추정값을 사용한다고 해도 중도탈락의 로짓확률이 관측값과 추정값을 동시에 이용하여 유도되므로 완전히 종속되

지는 않는다. 그러나 이러한 모형으로 중도탈락이 ID인가를 검정할 때 종속성이 존재하여 유의확률을 낮게 추정할 가능성이 있다. 그러므로 본 논문에서 제안한 분석방법을 중도탈락의 형태에 관한 검정으로 사용하는 데는 제한점이 있다고 할 수 있다.

참고문헌

1. Diggle, P. and Kenward M. G. (1994). Informative Drop-out in Longitudinal Data Analysis (with discussion). *Applied Statistics*, Vol. 43, 49-94.
2. Hogan, J. W., Roy, J. and Korkontzelou, C. (2004). Tutorial in Biostatistics : Handling Drop-out in Longitudinal Studies. *Statistics in Medicine*, Vol. 23, 1455-1497.
3. Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.
4. Murray, G. D., Findlay, J. G.(1988). Correcting for the Bias caused by Drop-outs in Hypertension Trials. *Statistics in Medicin*e. Vol. 7, 941-946.
5. Myers, W. R. (2000). Handling Missing Data in Clinical Trials: An Overview. *Drug Information Journal*, Vol. 34, 525-533.
6. Rubin, D. B. (1976). Inference and Missing Data, *Biometrika*, Vol. 63, 581-592.
7. Robins, J., Rotnisky, A. and Zhao, L. P. (1995). Analysis of Semiparametric Regression Models for Repeated Outcomes in the Presence of Missing Data. *Journal of American Statistical Association*, Vol. 90, 106-121.
8. Roy, J. and Lin, X. (2005). Missing Covariates in Longitudinal Data with Informative Dropouts: Bias Analysis and Inference. *Biometrics*, Vol. 61, 837-846.
9. Shao, J. and Zhong, B. (2003). Last Observation Carry-forward and Last Observation Analysis. *Statistics in Medicine*, Vol. 22, 2429-2441.
10. Thijs, H., Molenberghs, G. and Verbeke, G. (2000). The Milk Protein Trial : Influence Analysis of the Dropout Process, *Biometrical Journal*, Vol. 42, No. 5, 617-646.
11. Wu, M. C. and Bailey, K. R. (1988). Analysing Changes in the Presence of Informative Right Censoring Caused by Death and Withdrawal. *Statistics in Medicine,* Vol. 7, 337-346.
12. Wu, M. C. and Bailey, K. R. (1989). Estimation and Comparison of Changes in the Presence of Informative Right Censoring: Conditional Linear Model. *Biometrics*, Vol. 45, 939-955.
13. Zwinderman, A. H.(1992). Statistical Analysis of Longitudinal Quality of Life Data with Missing Measurements. *Qual Life Res*, Vol. 1, 219-224.

[2006년 7월 접수, 2006년 9월 채택]