응용통계연구 제 15권 2호, 2002년, pp. 233-250

잔차에 기초한 비례위험모형의 회귀진단법 고찰 * - PBC †자료를 통한 응용 연구

이성임[1] 박성현[2]

요 약

Cox의 비례위험모형(proportional hazards model)은 생존자료(survival data)에 대한 회귀모형으로 경제학 및 의·공학을 비롯한 여러 응용 분야에서 가장 널리 쓰이고 있는 모형 중 하나이다. 그러나, 이 모형은 일반선형모형에 비해 잔차 분석을 통한 회귀진단의 연구가 널리 알려져 있지 않아, 국내의 실제 자료 분석에서는 잔차 분석에 대한 활용이 거의 이루어지지 않고 있는 실정이다. 이에 본 논문에서는 그 동안 제안된 여러 잔차들을 비교 분석하고, S-plus 프로그램을 이용한 PBC(primary biliary cirrhosis) 자료 분석을 통해 각 잔차들의 의미를 고찰하고자 한다.

주요용어: 비례위험모형, 마팅게일 잔차, 회귀진단

1. 서론

생존시간(survival time)을 T라하고, 이에 영향을 주는 공변량을 $\mathbf{z}$($p \times 1$ 벡터)라고 할 때, Cox(1972, 1975)의 비례위험모형은 다음과 같이 정의된다:

$$\lambda(t;\mathbf{z}) = \lambda_0(t)\exp(\beta' z). \tag{1.1}$$

이때 $\lambda(t;\mathbf{z})$를 생존시간의 위험률 또는 위험함수(hazard rate or function)라 하고 λ_0는 $\mathbf{z} = \mathbf{z}_0$에서의 기저위험함수(baseline hazard function)를 가리킨다. 회귀계수 β에 관한 추론은 다음의 로그편우도함수(log partial likelihood function)에 기초한다.

$$\log L(\beta;\mathbf{z}) = \sum_{i=1}^{n}\int_0^{\infty}\left(Y_i(t)\beta'\mathbf{z}_i - \log\{\sum_{j=1}^{n}Y_j(t)\exp(\beta'\mathbf{z}_j)\}\right)\mathrm{d}N_i(t)\,. \tag{1.2}$$

위 함수는 Aalen(1975)이 생존자료에 다변량 계수확률과정(multivariate counting process) 이론을 적용시킨 것을 바탕으로 Anderson & Gill(1982)이 비례위험모형을 확장 적용하여 구한 결과이다. 이때 $N_i(t)$와 $Y_i(t)$의 정의는 다음과 같다.

* 본 연구는 서울대학교 복잡계통계연구센터를 통한 한국과학재단의 지원에 의하여 수행되었음.

† "http://lib.stat.cmu.edu/datasets/pbc"에서 자료와 변수의 설명을 볼 수 있음.

1) (151-742) 서울시 관악구 신림동 산56-1, 서울대학교 복잡계통계연구센터 연수연구원
E-mail: silee@stats.snu.ac.kr

2) (151-742) 서울시 관악구 신림동 산56-1, 서울대학교 통계학과, 교수
E-mail: E-mail: parksh@plaza.snu.ac.kr

$$N_i(t) = I\{X_i <= t, \delta_i = 1\}, \qquad X_i = T_i \wedge U_i (T_i : \text{생존시간}, U_i : \text{중도절단시간})$$
$$Y_i(t) = I\{X_i >= t\}, \qquad \delta_i = I(T_i <= U_i). \tag{1.3}$$

계수확률과정 $N_i(t)$는 i번째 환자가 시점 t에서 생존시간이 관측되면 1, 그렇지 않으면 0을 갖는 계수확률과정(counting process)이다. 즉, $(t, N_i(t))$의 그래프를 그려보면, i번째 환자의 생존시간이 관측되는 시점에서 크기 1의 점프가 생기는 확률과정이다. 또한 $Y_i(t)$는 i번째 환자가 시점 t에서 살아있으면 1, 아니면 0을 갖는 확률과정이다. 식 (1.3)을 이용한 모형 (1.1)의 확장은 회귀진단을 연구함에 있어 매우 유용한 정보를 제공해 주었다.

모형 (1.1)에서 회귀계수β는 함수 (1.2)를 최대로 하는 $\hat{\beta}$로 추정하고 이를 편우도추정량(maximum partial likelihood estimator, MPLE)이라 정의한다. 그리고, 기저위험함수의 추정은 로그편우도함수 (1.2)에서 알 수 있듯이 우도함수에 의존하지 않고 Breslow(1974)가 제안한 다음의 식이 일반적인 추정량이다.

$$\hat{\Lambda}_0(t) = \int_0^t \frac{1}{\sum_j^n Y_j(s) \exp(\hat{\beta}' \mathbf{z}_j)} \sum_i^n \mathrm{d}N_i(s)\,. \tag{1.4}$$

이러한 Cox의 비례위험모형은 경제학 및 의·공학 등 여러 분야에서 생존자료를 분석하는데 널리 사용되고 있다. 그러나 관측값들의 중도절단 비율이 높을 경우 위험률에 대한 비례성 가정은 위배되기 쉽다. 즉, 임의의 두 개인에 대한 위험률의 비가 상대적인 공변량의 크기에만 의존한다는 모형의 가정이 잘 맞지 않는다는 것이다. 그러므로, 분석 결과를 신뢰하고 분석의 질을 높이기 위해서는 일반선형회귀모형에서와 마찬가지로 비례위험의 가정을 검토할 수 있는 회귀진단 연구가 매우 중요하다. 특히, 일반선형회귀모형에서의 Belsley et al.(1980) 또는 Cook & Weisberg(1982)와 같은 잔차에 기초한 회귀진단 연구가 필요하다. 이를 위해 본 연구에서는 지금까지 소개된 비례위험모형의 잔차를 비교 검토하여, 이들이 회귀진단에 어떻게 사용될 수 있는지 알아보고, 또한 PBC 예제를 통하여 각 잔차의 역할을 비교해 봄으로써 비례위험모형을 통한 분석에 실질적인 도움을 주고자 하였다.

본 논문의 구성은 다음과 같다. 2절에서는 Cox & Snell(1968), Schoenfeld(1982), Cain & Lange(1984), Barlow & Prentice(1988) 그리고 Therneau, Grambsch & Flemming(1990) 등이 보여준 연구 결과를 바탕으로, 지금까지 정의된 각 잔차들을 소개하고, 3절에서는 이들 잔차와 모형의 평가 내용을 관련지어 고찰해보기로 한다. 모형의 진단은 첫째, 모형의 비례위험가정(proportional hazards assumption)을 검토하고, 둘째, 각 관측치에 대한 개별적인 평가로서 이상점(outliers)과 영향점(influential points)들을 검출하는 방법을 고찰하고, 마지막으로, 생존시간의 위험률에 미치는 공변량의 영향을 가장 잘 설명하는 공변량의 적절한 변수 형태가 무엇인지 검토하기로 한다. 4절에서는 3절에서 알아본 회귀진단 내용을 PBC 자료에 적용 평가해 보았다. 그리고 부록에 S-plus 프로그램을 함께 실어 실제의 자료분석에 도움이 되도록 하였다. 마지막으로 5절에서는 이들 잔차를 이용한 회귀진단의 제한점과 앞으로의 연구방향에 관하여 고찰해 보았다.

2. 비례위험모형의 잔차

비례위험모형은 일반선형모형에서처럼 "$T_i - \hat{T}_i$(생존시간의 관측값 - 생존시간의 추정값)" 형태의 잔차가 정의되지 않는다. 이것은 이 모형이 절대적인 시간에 관해 모형화된 것이 아니고 상대적인 시간에 모형화 되었기 때문이다. 따라서 이 절에서는 지금까지 제안된 여러 가지 새로운 형태의 잔차들을 소개하기로 한다.

2.1. Cox-Snell 잔차

Cox-Snell 잔차를 정의하기 전에 생존시간의 누적위험함수의 성질에 관하여 살펴보자. 생존시간 $T = t$에서 생존함수를 $S(t)$, 누적위험함수를 $\Lambda(t)$라 가정할 때 $S(t) = -\log \Lambda(t)$의 관계가 성립함을 알 수 있다. 이 때 생존시간의 누적위험함수를 $Y = \Lambda(T)$라 정의하자. 그렇다면 누적위험함수(Y)의 생존함수는

$$\begin{aligned} P(Y > y) &= P(\Lambda(T) > y) \\ &= P(T > \Lambda^{-1}(y)) = S(\Lambda^{-1}(y)) \\ &= \exp(-\Lambda(\Lambda^{-1}(y))) = \exp(-y) \end{aligned}$$

가 된다. 이것은 누적위험함수가 생존시간 T의 분포와 상관없이 평균이 1인 지수분포임을 알려준다. 따라서 (Y, δ)에 기초한 위험함수는 $\Lambda_Y(y) = y$의 관계가 있음을 알 수 있다. 여기서, β대신에 $\hat{\beta}$(MPLE)를 대입하고 Λ_0대신에 $\hat{\Lambda}_0$를 대입하여 구한 누적위험함수 $\hat{\Lambda}(T_i)$를 Cox-Snell 잔차로 정의한다:

$$r_i = \hat{\Lambda}_0(T_i) \exp(\hat{\beta}' \mathbf{z}_i). \tag{2.1}$$

2.2. 마팅게일 잔차(martingale residual)

마팅게일 잔차는 Barlow & Prentice(1988)가 처음 제안하고 Therneau, Grambsch & Flemming(1990)에 의해 그 성질이 자세히 연구되었다. 시점 t에서의 마팅게일 잔차는

$$M_i(t) = N_i(t) - \int_0^t Y_i(s) \exp(\beta' \mathbf{z}_i) \mathrm{d}\Lambda_0(s) \tag{2.2}$$

와 같이 정의한다. 위의 식에서 오른쪽에 있는 두 값의 차이가 마팅게일 확률과정(martingale process)을 따르므로(Fleming & Harrington, 1991; Andersen, Borgan, Gill & Keiding, 1993) 이를 마팅게일 잔차라고 부른다. 이 값은 "시간 [0, t]에서 관측된 사건의 수 − 모형으로부터 예측된 사건의 수"로 해석되고 이러한 값의 차이는 일반선형모형의 잔차 정의와 가장 흡사함을 알 수 있다. 물론 식 (1.1)의 비례위험모형에서는 관측된 사건의 수가 항상 0 또는 1이 된다. 식 (2.2)에서 회귀계수와 기저위험함수에 각각의 추정값을 대입하면 마팅게일 잔차의 추정값 $\hat{M}_i(t)$를 얻을 수 있다. 그리고 이들 잔차값은 $\sum_i \hat{M}_i(t) = 0$을 만족하고 근사적으로 $E(\hat{M}_i(t)) = \mathrm{cov}(\hat{M}_i(t), \hat{M}_j(t)) = 0$ (Gill, 1980)을 만족한다. 이로써 마팅게일 잔차는 그 의미 뿐 아니라 성질에 있어서도 일반선형모형에서의 잔차와 비슷하다는 것을 알 수

있다. 그리고, 식 (2.2)에서 $t = \infty$일때 추정한 마팅게일 잔차 $\hat{M}_i(\infty)$를 $\tilde{M}_i$로 나타내고 이를 마팅게일 잔차라고 부른다:

$$\begin{aligned} \tilde{M}_i &= \delta_i - \int_0^{T_i} \exp(\hat{\beta}'\mathbf{z}_i) \mathrm{d}\hat{\Lambda}_0(s) \\ &= \delta_i - \hat{\Lambda}(T_i). \end{aligned} \tag{2.3}$$

위에서 δ_i는 0 또는 1의 값을 갖고, $\hat{\Lambda}(T_i)$는 0부터 ∞까지 값을 취할 수 있으므로 마팅게일 잔차의 범위는 최대값이 1이고 최소값은 $-\infty$가 됨을 알 수 있다.

그러나 마팅게일 잔차가 일반선형모형에서의 잔차와 같은 역할을 하는 것은 아니다. 예를 들어 일반선형모형에서는 잔차의 제곱합(sum of squared residuals)이 모형의 적합도를 알려주거나 혹은 가장 작은 잔차의 제곱합을 갖는 모형을 최종모형으로 선택하기도 하지만 비례위험모형에서는 마팅게일 잔차의 제곱합이 작다고 해서 좋은 모형이라고는 말하지 않는다(Therneau & Grambsch, 2000). 그 대신 마팅게일 잔차와 공변량에 관해 그린 잔차그림을 통해 공변량이 위험률에 미치는 영향이 어떤 함수 형태인지를 평가할 수 있다(Therneau, Grambsch & Flemming 1990). 공변량 $\mathbf{z}$를 $j(j = 1, \cdots, p)$ 번째 공변량 z_j와 이를 제외한 나머지 $(p-1)$개의 공변량 $\mathbf{z}(j)$로 나누고 이들이 서로 독립이며 공변량 $\mathbf{z}(j)$들의 형태는 알려져 있다고 가정하면 위험률은 다음과 같이 나타낼 수 있다:

$$\lambda(t|\mathbf{z}(j), z_j) = \lambda_0(t) \exp(\beta^{*\prime}\mathbf{z}(j)) \exp(f(z_j)) .$$

위 식에서 j번째 공변량 z_j가 위험률에 미치는 영향을 나타내는 함수 f의 추정은 마팅게일 잔차를 통해 결정될 수 있다. 우선, 공변량 $\mathbf{z}(j)$들만 가지고 모형을 적합시킨 후 마팅게일 잔차 $\tilde{M}(j)$를 구하여 j번째 공변량에 관하여 산점도를 그린 후 그들간의 적당한 함수 관계를 찾으면, 이것이 f의 추정식임을 알 수 있다. 이러한 사실은 다음의 관계에 근거한다:

$$E(\tilde{M}(j)|\mathbf{z}(j)) \approx cf(z_j) . \tag{2.4}$$

이 때, 상수항 c는 공변량 z_j와는 독립으로 중도절단 비율에 관련이 있는 값으로 단지 y축의 스케일에 영향을 미치므로 전체적인 함수 형태에는 거의 영향을 주지 않는다.

2.3. 편차 잔차(deviance residual)

이 절에서는 식 (2.3)에서 정의된 마팅게일 잔차와는 달리 대칭적인 분포형태를 갖는 잔차를 소개하고자 한다. 마팅게일 잔차의 대부분의 값이 1에 몰려 있어 양의 값으로 잔차값이 큰 경우 상대적으로 값의 차이를 구별하기가 쉽지 않은 점이 있었다. 따라서 모형에 의해 예측된 것보다 더 많은 사건이 관측된 경우 즉, 생존시간이 예상외로 짧았던 관측치를 찾기에 어려운 단점이 있다. 또한 모형으로부터 추정된 것보다 사건의 수가 적은, 즉, 상대적으로 생존시간이 길었던 관측값이 인공적으로 나타날 수 있다. 이런 이유로 마팅게일 잔차가 좀 더 대칭적인 형태를 갖도록 변환할 필요가 있었고, McCullagh & Nelder(1989)의 편차(deviance)정의로부터 다음과 같은 편차잔차(deviance residual)가 제안되었다:

$$d_i = sign(\tilde{M}_i) \sqrt{2}\{ - \tilde{M}_i - \delta_i \log(\delta_i - \tilde{M}_i)\}^{1/2}. \tag{2.5}$$

중도절단된 자료의 비율이 약 25% 미만일 때, 위의 편차잔차는 0에 관하여 대칭이고 표준편차가 1인 정규분포의 형태와 매우 가깝다는 것이 알려져 있다(Therneau, Grambsch & Flemming. 1990).

2.4. 스코어 잔차(score residuals)

생존시간 t를 $0 = t_0 < t_1 < t_2 < \cdots (j \to \infty)$ 일때 $t_j \to \infty$의 구간으로 세분화하면 마팅게일 잔차 $\tilde{M}_i$를 다음과 같이 표현할 수 있다:

$$\begin{aligned} \tilde{M}_i &= N_i(\infty) - \int_0^{\infty} Y_i(s)\exp(\beta'\mathbf{z}_i)\mathrm{d}\Lambda_0(s) \\ &= \sum_{j=1}^{\infty}[\{N_i(t_j) - N_i(t_{j-1})\} - \int_{t_{j-1}}^{t_j} Y_i(s)\exp(\beta'\mathbf{z}_i)\mathrm{d}\Lambda_0(s)] \\ &\equiv \sum_{j=1}^{\infty}\Delta\hat{M}_{ij}. \end{aligned}$$

여기서 $\Delta\hat{M}_{ij}$는 구간 $(t_{j-1}, t_j]$에서 관측된 i번째 관측치의 마팅게일 잔차로 해석할 수 있고, $\Delta\hat{M}_{ij}$에 가중치를 주는 방법으로 새로운 형태의 일반화된 잔차가 제안될 수 있다(Barlow & Prentice, 1988). 예를 들어, 가중확률과정 $W_i = \{W_i(s) : s \geq 0\}$에 적당한 조건을 주게되면 $\int W_i(s)\mathrm{d}M_i(s)$의 형태는 다시 마팅게일 확률과정을 따르게 되는데, 이러한 형태를 갖는 통계량을 마팅게일 변환 잔차(martingale transform residual)라고 부른다. 이 절에서 소개할 스코어 잔차는 이러한 마팅게일 변환 잔차의 형태를 가진 잔차이다. 식 (1.2)를 j번째 회귀계수 β_j에 관하여 미분하면 다음과 같다.

$$\frac{\partial logL(\beta;\mathbf{z})}{\partial\beta_j}|_{\beta=\hat{\beta}} = \sum_i \int_0^{\infty}(z_{ij} - \bar{z}_j(\hat{\beta}, s))\mathrm{d}\hat{M}_i(s) = \sum_i U_{ij}(\hat{\beta}, \infty)$$

이때, $\bar{z}_j(\hat{\beta}, s) = \frac{\sum_i z_{ij}Y_i(s)\exp(\hat{\beta}'\mathbf{z}_i)}{\sum_i Y_i(s)\exp(\hat{\beta}\mathbf{z}_i)}$를 나타내고 이것은 s 시점에서 구한 공변량 z_j의 가중 기대값으로 해석할 수 있다. 그리고 $U_{ij}(\hat{\beta}, t)$는 i번째 관찰값의 j번째 공변량에 대한 스코어 과정(score process)이라고 부르고, 시점 $t = \infty$에서의 스코어 과정값, $U_{ij}(\hat{\beta}, \infty)$를 스코어 잔차(score residual)라 정의한다:

$$U_{ij}(\hat{\beta}, \infty) = \int_0^{\infty}\big(z_{ij} - \bar{z}_j(\hat{\beta}, s)\big)\mathrm{d}\hat{M}_i(s). \tag{2.6}$$

이 잔차는 앞에서 정의된 잔차들과는 달리 각 관측값 뿐 아니라 공변량별로도 서로 다른 잔차가 제안되고 있음을 알 수 있다.

Cain & Lange(1984)는 관측값마다 서로 다른 가중치를 준 편우도 함수를 생각함으로써 스코어 잔차로부터 다음의 결과를 유도하였다:

$$\hat{\beta} - \hat{\beta}(i) \approx \{-I(\hat{\beta}, \infty)^{-}1\}(U_{i1}(\hat{\beta}, \infty), U_{i2}(\hat{\beta}, \infty), \cdots, U_{ip}(\hat{\beta}, \infty)). \tag{2.7}$$

위에서 $\hat{\beta}(i)$는 i번째 관측치를 제외하고 구한 회귀계수를 의미하므로 식 (2.7)의 왼쪽은 i번째 관측치가 회귀계수의 추정에 미치는 변화량을 의미한다. 그리고 $I(\hat{\beta},\cdot) = \frac{-\partial^2 logL(\beta;\mathbf{z})}{\partial\beta^2}|_{\beta=\hat{\beta}}$이므로, 식 (2.7)의 오른쪽은 스코어 잔차를 회귀계수의 분산으로 스케일링한 값임을 의미한다. 다시 말해, 각 관측치가 회귀계수의 추정에 미치는 영향이 '스코어잔차 $\times\hat{\beta}$의 분산'으로 근사될 수 있음을 알려준다.

2.5. Schoenfeld 잔차

Schoenfeld 잔차(1982)는 식 (2.6)에 있는 스코어 잔차의 증가분(jump size)으로 정의된다:

$$r_{ij}(\hat{\beta}) = z_{ij} - \bar{z}_j(\hat{\beta}, t_i). \tag{2.8}$$

여기서 t_i는 i번째 관측치의 생존시간을 의미하고, 따라서 이 잔차는 생존시간이 관측된 관측값에 대해서만 잔차가 정의됨을 알 수 있다. 식 (2.6)에서의 스코어 잔차와 마찬가지로 관측값별 그리고 각 공변량별로 잔차가 정의되지만 이 잔차는 특별히 생존시간이 관측된 관측값에 대해서만 정의되는 잔차이다. 이 잔차의 의미는 "실제로 생존시간이 관측된 시점에서의 공변량값 - 공변량의 기대값"이다. 비례위험가정이 맞는다면 이 잔차의 기대값은 근사적으로 영이 되며, 시간과는 독립인 성질을 갖는다. 그러므로 시간에 대한 잔차의 산점도를 그리면 0에 관하여 임의로 흩어져 있는 상태가 될 것이다. Harrel(1986)은 생존시간의 순위값과 이 잔차의 상관계수를 통하여 비례위험성의 검정을 제안하였다.

3. 잔차에 기초한 회귀 진단

서론에서 언급했듯이 비례위험모형에서 잔차를 바탕으로 한 모형 진단의 연구가 비교적 활발하지 못했던 것은 비례위험모형이 절대적인 시간에 관하여 가정된 것이 아니고 상대적인 시간에 관해서만 모형화 되어 잔차의 정의가 어렵기 때문일 것이다. 비례위험모형의 통계적 추론이 관측된 생존시간의 순서에만 관계하는 사실을 생각할 때 그 이유를 짐작할 수 있다.

이와 같은 이유로 비례위험모형에서는 선형모형에서의 잔차와 같은 "$T_i - \hat{T}_i$ (=생존시간의 관측값 - 생존시간의 추정값)"이 아닌 새로운 형태의 잔차가 제안되었다. 이 절에서는 모형의 진단 내용에 따라 제2절에서 소개된 각 잔차들의 역할을 비교하고자 한다. 이를 위해 모형의 진단은 첫째, 모형의 비례위험가정(proportional hazards assumption)을 검토하고, 둘째, 이상점(outliers)과 영향점(influential points) 검출 등과 같은 각 관측치에 대한 개별적인 평가에 관해 잔차들의 역할을 비교해 보기로 한다. 그리고 마지막으로 생존시간의 위험률에 미치는 공변량의 영향을 가장 잘 설명할 수 있는 적절한 함수 형태가 무엇인지 조사한다.

3.1. 비례위험가정

모형의 비례위험가정에 대한 일반적인 검정법은 송혜향, 이선호(1994)와 장애방, 이재원(1997)을 참조하기 바란다.

제2절에서 소개한 잔차들 중 Cox-Snell 잔차, 스코어 잔차, Schoenfeld 잔차를 이용하여 위험률의 비례성 가정을 검토할 수 있다. 앞서 정의한 대로 이들 모두는 서로 다른 양을 측정하는 값이지만 동시에 같은 내용을 검정할 수 있는 통계량이다. Cox-Snell 잔차는 누적위험함수의 추정값으로 정의되고, Schoenfeld 잔차는 스코어 잔차의 증가분 즉, 공변량의 함수로 정의되고 있다. 이렇게 서로 다른 통계량에 바탕을 두고 정의된 잔차들이지만 이들 모두는 비례위험가정의 검정에 이용될 수 있다. 그러나 일반선형모형에서처럼 잔차 통계량의 근사 분포를 바탕으로 모형의 가정을 검토하는 것이 아니라, 각 잔차 그림을 그려봄으로써 시각적 판단에 의하여 검정을 하는 것이다.

Cox-Snell 잔차 그림은 y축에 Cox-Snell 잔차에 기초한 위험함수의 Nelson-Aalen 추정값을, x축에는 Cox-Snell 잔차를 나타내어 산점도를 그린 것이다. 생존시간의 분포와 상관없이 누적위험함수는 평균이 1인 지수분포를 따르므로, 이 잔차그림은 누적위험함수의 추정값이 정확할 때 즉, $\hat{\beta}$와 $\hat{\Lambda}(\cdot)$의 추정이 잘 맞을 때 기울기가 1인 직선관계를 나타낼 것이다. 그러므로 Cox-Snell 잔차는 비례위험모형의 가정이 적절하다면 기울기 1의 직선관계를 보여주는 잔차그림을 제공할 것이다.

Schoenfeld 잔차 그림은 y축을 잔차로 하고 x축을 생존시간으로 나타낸 것이다. 비례위험가정을 만족할 때 Schoenfeld 잔차가 시간과는 독립이라는 성질로부터, 잔차그림의 산점도가 랜덤하게 분포한다는 것은 가정된 모형이 적절했음을 알려준다. 따라서 산점도를 그려 보아 잔차값이 시간에 따라 증가 혹은 감소 등 특정한 경향을 보이는지 조사함으로써 비례위험가정을 검토할 수 있다. 특히, 추정된 회귀계수 $\hat{\beta}$의 분산으로 스케일링한 Schoenfeld 잔차, $I(\beta,\infty)^{-}1 \times (r_{i1}, r_{i2}, \cdots, r_{ip})'$는 시각적으로 좀 더 보기 좋은 형태의 그림을 제공한다.

위에서 설명한 것을 종합할 때 Cox-Snell 잔차는 모든 공변량들을 토대로 비례위험가정을 검토하는 그림을 제공하고, 스코어 잔차와 Schoenfeld 잔차는 각 공변량의 개별적인 비례위험가정을 조사하는 차이점이 있음을 알 수 있다.

3.2. 이상점과 영향점

이 절에서는 관측값의 이상점과 영향점을 판별하는 방법에 관하여 검토하고자 한다. 이상점과 영향점을 검출하는 것은 자료의 특성을 파악할 수 있을 뿐 아니라 그 결과를 이해하는 데에도 도움을 줄 수 있기 때문이다.

일반적으로 이상점은 관측값과 예측값의 차이가 큰 값으로 정의한다. 따라서 비례위험모형에서도 생존시간의 관측값과 모형으로부터 구한 예측값의 차이가 많이 나는 관측값을 이상점으로 검출할 수 있을 것이다. 그러나 앞에서 밝힌 대로 비례위험모형이 생존시간의 관측값 대신 관측 순서만을 고려한 모형이기 때문에 이러한 잔차는 정의되지 않는다. 따라서 2.2절에서 소개한 잔차들 중 일반적인 잔차의 의미와 가장 비슷하게 정의되어 있는 마팅게일 잔차로부터 이상점을 검출할 수 있을 것이다. 즉, 관측된 생존시간까지 일어난 사

건의 수와 모형으로부터 예측된 사건의 수의 차이가 큰 관측값을 이상점이라 할 수 있다. 이를 위해 y축을 마팅게일 잔차로 하고 x축을 위험점수(risk score: $\hat{\beta}'\mathbf{z}$)로 하는 잔차그림을 그릴 수 있다. 그런데 실제로는 마팅게일 잔차값의 분포가 매우 치우쳐 있기 때문에 마팅게일 잔차를 이용한 이상점 검토가 쉽지 않은 단점이 있다. 이 때문에 2.3절에서 소개된 편차 잔차를 이용한 잔차그림이 더 자주 쓰이고 있다. 편차 잔차는 마팅게일 잔차가 좀 더 대칭적인 모습으로 보이도록 변환된 형태이므로 대부분의 경우 마팅게일 잔차보다는 훨씬 대칭적인 모양을 나타내어 일반선형모형에서의 잔차그림과 가장 비슷한 모양을 나타낸다. 하지만 이 잔차는 자료의 중도절단 비율에 영향을 많이 받아 중도절단 비율이 높은 경우에는 이상점 검토를 위한 잔차로 추천할 만하지 않은 단점이 있다.

회귀계수의 추정에 영향력이 큰 관측치를 평가하기 위해서는 식 (2.7)의 관계를 근거로 스코어 잔차를 이용할 수 있다. 각 관측치에 대한 회귀계수의 변화량이 $\hat{\beta}$의 분산으로 표준화한 스코어 잔차값으로 근사되어, 표준화된 스코어 잔차가 개별적인 영향점들을 검출하는데 유용하게 쓰일 수 있음을 알려 준다. 따라서 각 관측번호에 따라 잔차의 크기를 나타내어 상대적인 변화량의 크기를 비교함으로써 회귀계수의 추정에 영향이 큰 관측값들을 검출 할 수 있을 것이다.

3.3. 공변량의 변수 형태

일반 선형모형에서의 잔차와 가장 흡사하게 정의된 마팅게일 잔차는 적합된 공변량의 형태를 평가하는 중요한 역할을 한다. 식 (2.4)의 결과로부터 마팅게일 잔차와 적합에 포함되지 않은 공변량의 산점도는 그 공변량의 근사적인 함수형태를 보여준다. 그런데 이러한 결과는 회귀계수 $\hat{\beta}$와 기저위험함수 Λ_0의 값이 참일 때 만족하는 것이므로 실제로 이들의 추정값을 대입하여 사용한 결과에는 추정으로 인한 불확실성이 항상 존재할 것이다. 그래프를 통한 시각적 판단으로 함수관계를 유추하기 때문에 주관적 판단의 여지 또한 있으므로 임의의 공변량이 위험률에 관계하는 임의의 함수형태를 찾는 것은 결코 쉽지 않은 일이다.

마팅게일 잔차그림을 통해 공변량의 변수형태를 검토할 수 있는 두 가지 방법을 알아보면 첫째, 관심있는 공변량의 함수형태에 관하여 변수변환 $z^2, \log(z), \frac{1}{z}$등을 먼저 취한 다음 잔차그림에서 추정된 함수가 직선과 가까운지를 판단한다. 추정된 함수가 직선과 비슷한 형태를 보인다면 이것은 곧 f가 항등 함수(identity function)가 되어 변환된 공변량의 형태가 적절한 함수관계임을 나타낸다. 둘째, 잔차그림으로부터 추정된 함수의 식에 역치(threshold)와 같은 값의 큰 변화가 존재하는지 검토한다. 이러한 값의 변화는 연속형 변수를 이산형 변수로 변환하는 것을 고려할 만 하다는 것을 의미한다. 이 두 가지 사실에 바탕을 두고 마팅게일 잔차그림을 분석한다면 새로운 공변량의 변수 형태를 좀 더 쉽게 짐작할 수 있다.

4. PBC 자료를 통한 회귀진단 연구

이 절에서는 앞에서 소개된 잔차들이 실제의 자료분석에서 어떻게 응용될 수 있는지S-

plus 2000 프로그램을 통한 결과를 보면서 제3절에서 논의된 회귀진단 방법들을 좀 더 구체적으로 검토하고자 한다. PBC 자료의 모형 적합 과정은 Flemming & Harrington(1991)과 Klein & Moeschberger(1997)를 비롯한 생존분석 관련 서적을 참고하여 최종모형을 찾고, 본 연구에서는 모형 적합 후의 회귀진단 부분에 한하여 이 자료를 분석해 보기로 한다.

PBC 자료는 간의 원발성 간경화(primary biliary cirrhosis)에 대하여 1974년과 1984년 사이에 있었던 Mayo Clinic 임상시험으로부터 얻어진 자료이다. PBC는 자가면역(autoimmune)이 원인인 진행성질환으로 그 이후의 염증과정이 간의 경화와 담도 파괴를 일으키고 결국은 환자를 죽음에 이르게 하는 병이다. 이러한 질병에 걸린 환자 312명의 자료로부터, 이들의 생존시간과 공변량들 간에 비례위험모형을 적합해 보았다. 앞으로 나오는 모든 결과표에 대해서는 부록을 참조하기 바란다.

이들 자료에 관한 최종모형으로, age, bilirubin, albumin, edema 그리고 protime을 적합시켜 부록에 있는 표 2의 결과를 얻었다. 분석시 PBC 이외의 질병으로 사망했거나 또는 간이식 수술을 받은 경우 또는 그 외 이유로 follow-up되지 않은 사람의 경우는 생존시간이 중도절단된 것으로 처리하였다. 이 절에서는 표 2의 적합 결과로부터 제2절에서 소개한 잔차들을 구해 보고 제3절의 내용에 맞추어 모형에 관한 적합결과를 검토하기로 한다.

4.1. 비례위험가정 검정

제3.1절에서 소개한 대로 Cox-Snell, 스코어, Schoenfeld 잔차 그림을 통한 위험률의 비례가정을 검토해 보기로 하자. 앞에서 소개한 대로 이들 잔차는 서로 다른 통계량으로 제안

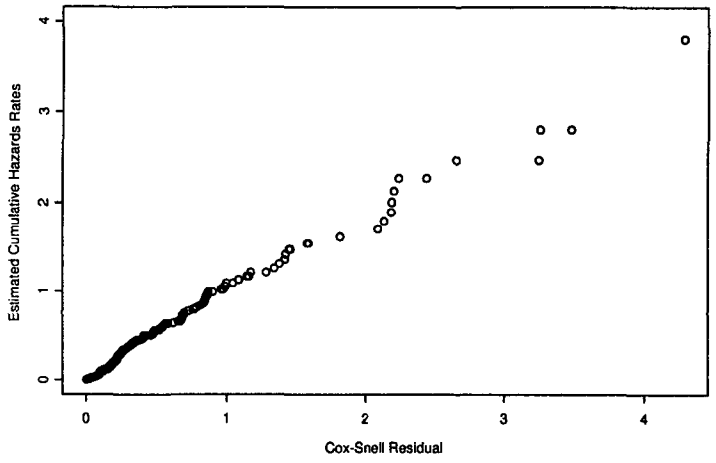

그림 4.1: Cox- Snell 잔차 그림

되었으며 각 통계량의 성질을 바탕으로 잔차그림이 서로 다른 모양을 나타낸다는 사실에 유의해야 할 것이다. 먼저 Cox-Snell 잔차는 그림 4.1과 같다. 그리고 이를 위한 S-plus 2000

프로그램은 부록의 표 3에 소개하였다. 제3.1절에서 살펴보았듯이 모형의 가정이 적절하다면 잔차그림이 기울기가 1인 직선에 가깝게 나타나야 할 것이다. 그림 4.1로부터 이들의 관계가 $y = x$의 직선 식과 많은 차이가 있다고 하기 어렵고 따라서 비례위험모형의 적합이 나쁘지 않음을 알 수 있다. 잔차가 2 근처인 부분에서 직선에서 많이 벗어나 보이긴 하지만, 잔차 2이상인 부분에서는 관측값의 개수가 많지 않았음을 고려해볼 때, 전체적인 적합도가 좋지 않다고 할 수 없을 것이다.

다음으로 Schoenfeld 잔차를 이용한 그림 4.2를 살펴보자. 식 (2.8)에서 정의한 Schoenfeld

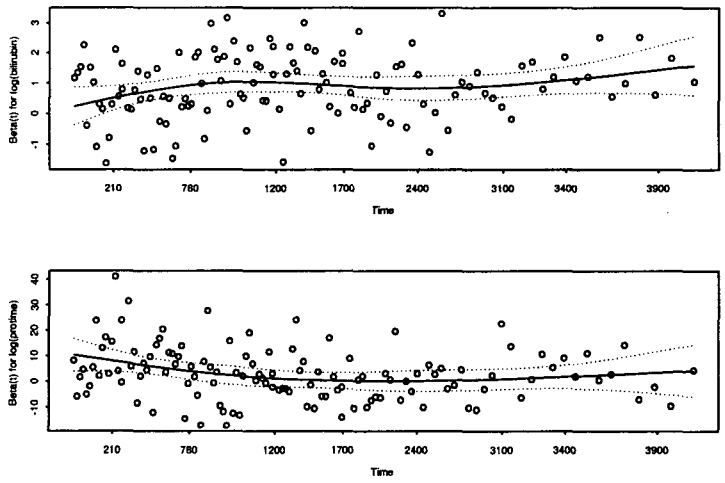

그림 4.2: Scaled Schoenfeld 잔차를 통한 비례위험가정 검토 (실선:비선형추정식, 점선: 비선형추정식의 신뢰 상한 혹은 하한)

잔차를 스케일링해서 y축에 나타내고, x축을 시간으로 해서 산점도를 그릴 수 있다. 앞에서 알아 본 것처럼 Schoenfeld 잔차는 비례위험 가정하에 시간과는 독립이기 때문에 이들 간에 산점도를 그려 시간에 따른 공변량의 효과 변화가 있었는지를 추정할 수 있다. 그래프에 의한 검토이기 때문에 판단에 도움을 주기 위해 잔차 그림에 비선형식을 적합하고 그것의 95% 신뢰구간을 그려보기도 하고 또한 생존시간의 순위값과 잔차와의 상관계수를 구하여 이 값이 0인지 검정해 보기로 한다. 프로그램 내용은 표 4를 참조하기 바란다. 표 4의 실행결과를 살펴보면 생존시간과 잔차 사이의 상관계수(ρ)를 추정하여 이것이 0인지 양측검정한 결과를 보여준다. log(bilirubin)과 log(protime)의 경우 유의확률 10%에서 시간이 흐름에 따라 위험률이 증가 또는 감소하는 경향을 보여주고 있는데 이들 두 변수에 관한 잔차그림을 살펴보면 그림 4.2와 같다. 점선은 $y = 0$ 의 직선을 그려보아 잔차들의 비선형 추정식과 구분해 보았는데, bilirubin의 경우 시간이 흐름에 따라 약간 증가하는 경향이 있음을 알 수 있고, 반대로 prothrombin time의 경우는 시간이 흐름에 따라 그 효과가 감소하는 경향이 있음을 알 수 있다. 즉 두 변수의 경우 시간에 따른 공변량의 효과 변화가 의심된다

고 할 것이다.

4.2. 이상점과 영향점 검출

이 절에서는 각 관측값의 개별적인 영향을 어떻게 검토할 수 있는지 알아보기로 한다. 먼저 이상점을 찾기위한 마팅게일 잔차 그림을 그려보면 그림 4.3과 같다. 이 그림에서 볼

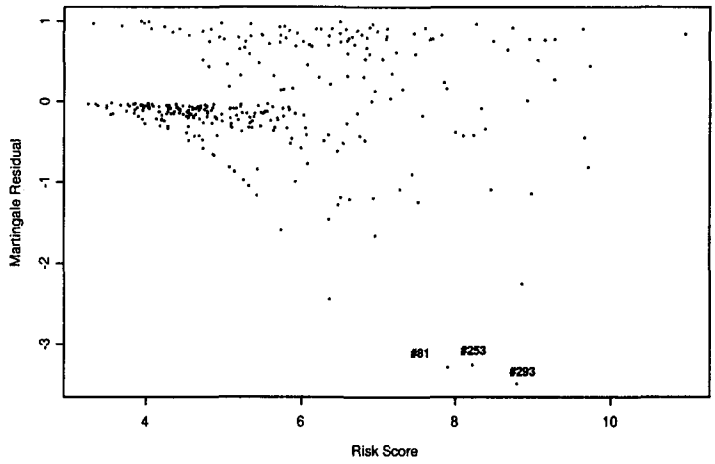

그림 4.3: 마팅게일 잔차

수 있듯이 마팅게일 잔차의 비대칭성 때문에 잔차가 큰 관측치를 검출하는 것이 쉽지 않다. 또한 값이 최대값 1 근처에 많이몰려있기 때문에 양의 방향으로 잔차가 큰 즉, 모형이 예측한 생존시간보다 빨리 사망한 관측값을 검출하기가 쉽지 않다. 단지, 81번, 253번 그리고 293번의 관측치가 다른 관측치에 비해 음의 방향으로 큰 차이를 보여 이상점으로 의심이 된다고 하겠다. 이상점의 검출을 위해서는 마팅게일 잔차 보다는 편차잔차를 이용한 잔차그림이 좀 더 유용하므로 이를 이용한 그림 4.4를 살펴보자. 그림 4.3에 비해서는 훨씬 대칭적 모습을 보여주고 있음을 알 수 있다. 이로 인해 87번 관측값의 경우 마팅게일 잔차를 관측했을 경우에는 알 수 없었지만 잔차의 크기가 크다는 것을 알 수 있었다. 또한 253, 293번째 관측치들은 여전히 잔차가 커서 이상점으로 의심이 되지만, 81번째 관측값의 경우는 그렇지 않다. 이처럼 이상점을 검토하기 위해서는 마팅게일 잔차보다는 편차 잔차를 이용한 잔차그림이 좀 더 유용함을 알 수 있다. 단 PBC 자료의 경우 관측값들의 중도절단 비율이 약 61%로 매우 높아 그림 4.4에서 보이는 것처럼 잔차가 0근처에 많이 몰려 있고 이 때문에 편차잔차 또한 이상점 검출에 적절한 잔차라 하기에 부족한 점이 있다.

이제 회귀분석의 결과에 크게 영향을 주는 관측값이 있는지 식별해 보기로 하자. 이를 위해 식 (2.7)의 스코어 잔차를 이용한 그림 4.5의 잔차그림을 살펴보자. 이를 위한 프로그램은 표 5를 참조하기 바란다. 이 절에서는 간략히 age, edema 두 변수에 대해 두 변수의

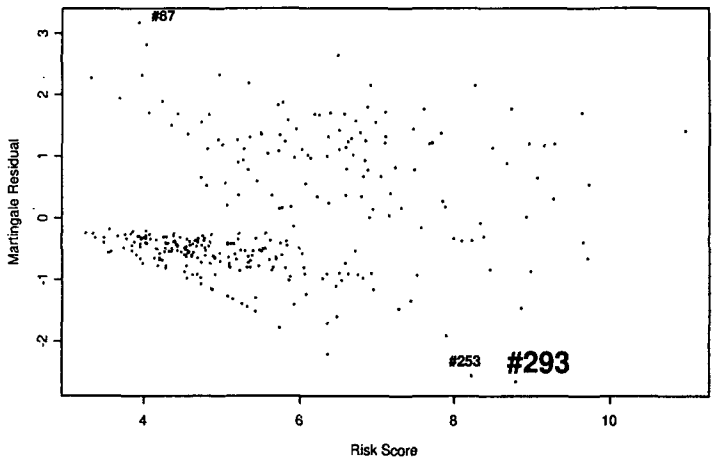

그림 4.4: 편차 잔차

추정에 영향을 가장 많이 미치는 관측치를 알아보기로 한다. 잔차그림은 y축에 스코어잔차를 두고 x 축을 관측번호로 하여 산점도를 그린후 그 값에 수직선을 그어 각 관측값별 잔차의 상대적인 크기를 식별할 수 있도록 하였다. 이때 스코어 잔차가 식 (2.7)에서 본 것처럼 그 관측치를 제외시키고 구한 회귀계수 추정값의 차이와 근사적인 관계가 있으므로, 잔

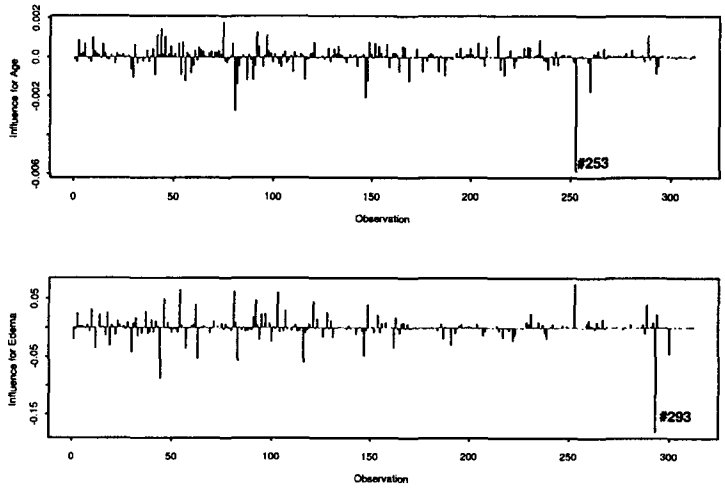

그림 4.5: 표준화된 스코어 잔차

차의 크기가 크다는 것은 그 관측값의 포함여부에 따라 회귀계수의 추정이 영향을 크게 받는다는 것을 의미한다. 따라서 age 변수 추정에 관하여 가장 영향을 많이 끼치는 관측치는 253번 관측치로 그 차이가 -0.63이다. 또 edema 변수의 경우에는 293번째 관측치가 다른 관측치들에 비해 회귀계수의 추정에 미치는 영향이 크다는 것을 알 수 있다.

4.3. 공변량의 변수형태

마팅게일 잔차는 공변량과 위험률 사이의 함수 관계를 결정하는 데 유용한 잔차이다. PBC 자료의 경우 표 2의 최종모형 적합 결과를 살펴보면, age와 edema를 제외한 모든 변수는 로그변환을 고려하였다. 이들 중 bilirubin 변수에 대하여 위험률에 미치는 적절한 함수관계로 로그변환이 적절한지 어떻게 판단할 수 있는지를 알아보기로 한다. 먼저 bilirubin 변수를 제외한 네 개의 변수 age, log(albumin), edema 그리고 log(protime)에 관하여 모형을 적합시킨 후 마팅게일 잔차를 구한다. 그리고 log(bilirubin)과 bilirubin에 대하여 잔차 그림을 그려보면 다음의 그림 4.6과 같다. 각각의 그래프에서 LOWESS 함수를 사용하여 식

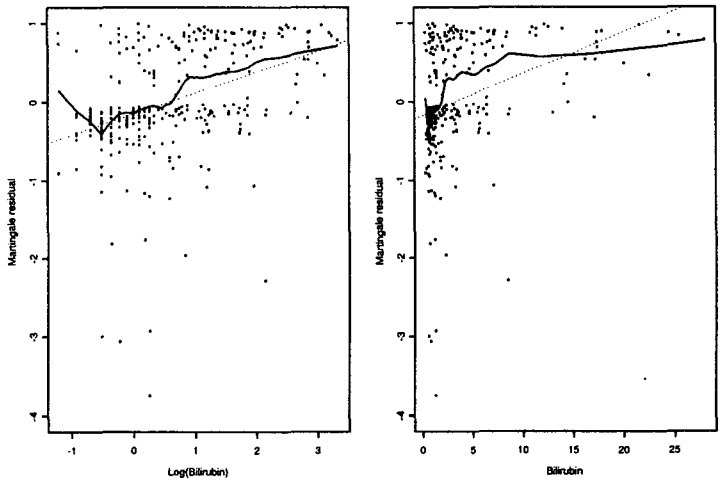

그림 4.6: 마팅게일 잔차(점선: 최소제곱식, 실선:비선형식)

(2.4)에서의 함수형태 f에 관한 비선형식을 추정해 보았다. 또한 점선으로 최소제곱추정식을 적합함으로써 선형 회귀식과 비선형 회귀식의 차이를 알아보았다. 이들 차이가 많이 나지 않는 잔차그림에서의공변량이 위험률에 미치는 영향을 가장 잘 표현한다고 할 수 있을 것이다. 그림 2의 오른쪽 그림에서 보면, bilirubin과 잔차 사이에는 추정된 회귀식이 최소제곱추정식과 많이 벗어나 있어, bilirubin으로 모형을 적합시키는 것이 옳지 않다는 것을 알 수 있다. 반면, log(bilirubin)의 경우 bilirubin과 비교하여, 최소제곱선과 거의 비슷한 선형관계를 보여주고 있으므로, 로그변환을 통한 모형적합이 적절하다 판단되어질 수 있다.

다른 변수들도 이와 같은 적합의 차이를 비교함으로써 로그 변환의 타당성을 검토해 볼 수 있을 것이다.

5. 고찰

비례성 가정 검토를 위한 잔차	이상점 검토를 위한 잔차	영향점 검토를 위한 잔차	공변량의 영향 평가를 위한 잔차
Cox-Snell 잔차	마팅게일 잔차	스코어 잔차*	마팅게일 잔차
Schoenfeld 잔차*	편차 잔차		

* : 잔차가 각 관측값마다 공변량별로 정의되므로 잔차의 형태가 행렬임.

표 1 :회귀진단 내용에 따른 잔차 분류

4절의 PBC 자료를 바탕으로 한 분석결과로부터 비례위험모형에서 제안된 잔차들이 모형의 진단에 어떻게 이용될 수 있는지를 살펴보았다. 표 1에 요약된 것처럼 모형의 평가내용에 따라 각각 서로 다른 잔차가 유용하게 쓰임을 알 수 있다. 또한 비례위험모형에서는 두 종류의 잔차가 정의되는데, 하나는 Cox-Snell, 마팅게일 그리고 편차잔차와 같이 관측값별로 정의되는 잔차가 있고, 또 다른 하나는 스코어 잔차와 Schoenfeld 잔차처럼 관측값마다 공변량별로 정의되는 잔차이다.

이들 잔차를 기초로 한 회귀진단 결과는 잔차그림을 통한 시각적 판단에 의한 것이고 잔차 통계량의 근사분포에 의한 것이 아니므로 분석가의 눈에 의한 이러한 검토는 분석결과가 매우 주관적일 수 있다는 단점이 있다. 잔차의 성질은 항상 회귀계수와 기저위험함수의 참값에서 만족되는 성질이므로 이들의 추정값을 사용한 잔차그림의 해석에는 융통성이 있다 하겠다. 또한 편차잔차의 경우를 보면 마팅게일 잔차의 대칭변환이라고는 하나 중도절단의 비율이 높은 경우 이 잔차 또한 성공적인 대칭변환이 될 수 없으며, 이상점을 검출하는 데 유용하지 않다고 한다. 따라서 개별적인 이상점 혹은 영향점을 검출할 수 있는 좀 더 객관적인 형태의 잔차가 연구, 제안될 필요가 있다.

참고문헌

[1] 송혜향, 이선호. (1994). Goodness of fit tests of Cox's proportional hazards model, *Journal of the Korean Statistical Society*, 24. 537-549쪽.

[2] 장애방, 이재원. (1997). 비례위험모형의 적합도 검정에 관한 연구, <응용통계연구>, 10. 85-103쪽.

[3] Aalen, O. O. (1975). *Statistical inference for a family of counting processes.* Ph.D. dissertation, University of California, Berkeley.

[4] Andersen, P.K., and Gill, R.D. (1982). Cox's regression model for counting processes: A large sample study. *The Annals of Statistics*, 10: 1100-20.

[5] Andersen, P.K., Borgan, O., Gill, R.D., and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

[6] Barlow, W.E., and Prentice, R. L. (1988). Residuals for relative risk regression. *Biometrika* 75: 65-74.

[7] Belsley, D.A., Kuh, E., and Welsch. R.E. (1980). *Regression Diagnostics*, New York: Wiley.

[8] Breslow, N.E. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89-99.

[9] Cain, K. C., and Lange, N.T. (1984). Approximate case influence for the proportional hazards regression model with censored data. *Biometrics* 40: 493-9.

[10] Cook, D.R., and Weisberg, S. (1982). *Residuals and Influence in Regression*, New York: Chapman and Hall.

[11] Cox, D.R. and Snell, E. J. (1968). A General Definition of Residuals(with Discussion). *Journal of the Royal Statistical Society B.* 30:248-275.

[12] Cox, D.R. (1972). Regression models and life tables(with discussion). *Journal of the Royal Statistical Society. B.* 34: 187-220.

[13] Cox, D.R. (1975). Partial likelihood. *Biometrika* 62: 269-76.

[14] Flemming, T.R., and Harrington, D.P. (1991). *Counting Processes and Survival Analysis.* New York: Wiley.

[15] Gill, R. D. (1980). *Censoring and Stochasic Integrals*, Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.

[16] Harrel, F. (1986). *The PHGLM procedure.* SAS Supplemental Library User's Guide, Version 5. SAS Institute, Inc., Cary, NC.

[17] Klein, J.P., and Moeschberger, M.L. (1997). *Survival analysis: Techniques for censored and truncated data*, New York: Springer-Verlag.

[18] McCullagh, P.,and Nelder, J.A. (1989). *Generalized linear Models*, 2nd Ed., London: Chapman and Hall.

[19] Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika* 69: 239-241.

[20] *S-plus4 Guide to statistics*, Data Analysis Products Division, MathSoft, Inc. Seattle, Washington (1997).

[21] Therneau, T. M., Grambsch, P.M., and Flemming, T.R. (1990). Martingale-based residuals for survival models. *Biometrika* 77:147-60

[22] Therneau, T. M. and Grambsch, P.M. (2000). *Modeling survival data: Extending the Cox Moddel.* New York: Springer-Verlag

[2002년 2월 접수, 2002년 5월 채택]

부록

표 2. 비례위험모형 적합 결과

```
> # Fit the Cox model
> pbcfit<-coxph(Surv(time,status==1) ~ age + log(albumin) +
                log(bilirubin) + edema+ log(protime), data=pbc)
> pbcfit
Call: coxph(formula = Surv(time, status == 1) ~ age + log(albumin)
+
        log(bilirubin) + edema + log(protime), data = pbc)


                coef exp(coef) se(coef)     z        p
          age  0.0332    1.0338  0.00866  3.84 0.000120
  log(albumin) -3.0600    0.0469  0.72404 -4.23 0.000024
log(bilirubin)  0.8801    2.4110  0.09874  8.91 0.000000
        edema  0.7859    2.1943  0.29897  2.63 0.008600
  log(protime)  3.0140   20.3687  1.02395  2.94 0.003200

Likelihood ratio test=199  on 5 df, p=0  n= 312
```

표 3. Cox-Snell 잔차 프로그램

```
>#  Martingale residual
>   mresid<-residuals.coxph(pbcfit,type="martingale")
>#  Cox-snell resdidual
>  csnell<-status-mresid
>fh.surv<-survfit(Surv(csnell,status)~1,type="fleming-harrington")$surv
>#  Nelson-Aalen estiamte of cumulative hazard function
>  cmhaz<- -log(fh.surv)
>  plot(sort(csnell),cmhaz,type="p",ylim=c(0,4),
   ylab="Estimated Cumulative Hazards Rates",xlab="Cox-Snell Residual")
```

표 4. Scaled Schoenfeld 잔차 프로그램

```
> # Check the proportionality using scaled Schoenfeld reisdual
> zph.pbc<-cox.zph(pbcfit)
> zph
                   rho  chisq      p
        age     -0.0369 0.1494 0.6991
log(albumin)    -0.0149 0.0288 0.8652 log(bilirubin)   0.1564
2.8505 0.0913
     edema  -0.1424 2.4557 0.1171
log(protime)    -0.1910 3.6932 0.0546
 GLOBAL         NA  9.5632 0.0886

> for (i in 1:5){
           plot(zph.pbc, var=i)
           abline(0,0,lty=3)
      }
```

표 5. 스코어 잔차 프로그램

```
> # sresid<-residuals.coxph(pbcfit, type="dfbeta")
> par(mfrow=c(2,1))
> plot(1:312, sresid[,1], type="h", ylab="Influence for Age", xlab="Observation")
> plot(1:312, sresid[,4], type="h", ylab="Influence for Edema", xlab="Observation
```

Review on proportional hazards regression diagnostics based on residuas

Sungim Lee [1] S.H. Park [2]

ABSTRACT

Cox's proportional hazard model is highly-used for the regression analysis of survival data in various fields. Regression diagnostics for the proportional hazards model, however, is not as well-known as the diagnostics for the classical linear models and so these diagnostic methods are not used widely in our practical data analyses. For this reason, we review the residuals proposed by several authors, and investigate how to use them in assessing the model. We also provide the results and interpretation with the analysis of PBC data using S-plus 2000 program.

Keywords: Proportional hazards model, martingale residual, regression diagnostics.

1) Post doctor, Statistical Research Center for Complex Systems, Seoul National University.
E-mail:silee@stats.snu.ac.kr
2) Professor, Department of Statistics, Seoul National University.
E-mail:parksh@plaza.snu.ac.kr