고급회귀분석론

Ch4. Model Adequacy Checking

양성준

선형회귀모형의 기본가정

- 선형회귀모형에서 취하였던 기본 가정은 다음과 같다.
  - (선형성) 예측변수와 반응변수의 관계는 선형이다.
  - (대칭성) 오차항의 평균은 0이다.
  - (등분산성) 오차항은 예측변수에 의존하지 않는 constant variance σ^2를 가진다.
  - (독립성) 오차항은 서로 독립이다.
  - (정규성) 오차항은 정규분포를 따른다.
- 최소제곱추정량의 불편성 등에는 위 조건이 모두 필요하지 않으나, 가설검정이나 신뢰구간 구성 등의 절차를 위해서는 위 가정들이 모두 만족되어야 한다.
- R^2와 같은 값은 위 가정의 충족여부에 대해서는 아무런 정보를 주지 못한다.

잔차 (residual)

- $e_i = y_i - \hat{y}_i$
- data와 fit 사이의 편차
- 반응변수 관측치 중 회귀모형에 의해 설명되지 않는 부분
- 오차항에 대한 관측치 or 실현값으로 간주할 수 있다
- 오차항에 대한 가정 검토에 있어서 핵심적인 역할을 한다

scaled residual

- 잔차를 적당하게 표준화시킨 것으로 이상치, 극단치의 탐색에 유용함

  - standardized residual

$$d_i = \frac{e_i}{\sqrt{MSR}},\ MSR = \hat{\sigma}^2$$

  - studentized residual

$$r_i = \frac{e_i}{\sqrt{MSR(1-h_{ii})}},\ var(e) = (I-H)\sigma^2,\ h_{ii} = (H)_{ii}$$

(큰 h_{ii}는 i번째 관측치가 leverage일 가능성을 내포)

  - press residual (deletion diagnostic)

$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1-h_{ii}}$$

($\hat{y}_{(i)}$: i번째 관측치를 제외하고 적합된 모형에 의한 예측치)

잔차도 (정규확률그림)

- 표준화잔차는 근사적으로 표준정규분포를 따른다. 따라서 $100\alpha\%$ 표본백분위수와 $N(0,1)$의 $100\alpha\%$ 모백분위수는 비슷한 값을 가져야 한다.
- $t_{[i]}$를 크기 순으로 i번째 표준화잔차라 할 때,

$$t_{[i]} \quad against \quad \Phi^{-1}\left(\frac{i-0.5}{n}\right)$$

  그림이 정규확률그림
- 점들이 직선 근처에 분포하면 정규성에서 벗어나지 않는 것으로 판단

잔차도 (잔차 against $\hat{y}_i$)

- (등분산성) : 잔차의 산포가 균일한가?
- (선형성) : 잔차에 남아있는 특정 패턴이 보이지 않는가?
- (정규성) : ± 2 (혹은 2.5)를 벗어나는 값들이 지나치게 많지 않은가?
- (독립성) : (시간순으로 잔차를 그렸을 때) 잔차의 부호가 동일한 것이 반복되는 경향이 있거나 부호가 계속 바뀌는 경향이 있지 않은가?
- 특별한 패턴 없이 균일하게 산포하고 ± 2를 넘는 것이 다수가 아니면 가정에서 크게 벗어나지 않는 것으로 판단

PRESS 통계량

- $PRESS = \sum_i (y_i - y_{(i)})^2 = \sum_i \left(\frac{e_i}{1-h_{ii}} \right)^2$
- $R^2_{pred} = 1 - \frac{PRESS}{SST}$
- 일종의 예측오차로, 모형의 비교의 목적으로 사용 가능함

이상치의 처리

- 삭제 : 관측오차, 기록오류, 전체 모델링에 도움되지 않는 경우
- 삭제X : 확률모형 하에서 발생이 가능한 경우, ex] 운영리스크, 홍수자료
- 삭제하지 않는 경우 robust 방법 혹은 대안적인 모형 등을 고려하여야 한다.

모형의 가정에 위배될 때

- 선형성 위배 : 비선형모형 고려, 자료변환
- 정규성 위배 : 자료변환, 정규성 외에 다른 분포 가정하에 추론
- 독립성 위배 : 시계열 모형 적합
- 등분산성 위배 : 가중회귀, 일반화회귀
- 가정으로부터 상대적으로 자유로운 방법이나 모형을 사용할 수 있다.
  - 비모수 모형 사용
  - 위 가정들은 대부분 모형이나 회귀계수에 대한 유의성 검정이나 신뢰구간 구성 등의 통계적 절차에 영향을 준다. 가정에 위배가 일어날 때는 그 가정들에 의존하는 전통적인 방법이 아닌 Bootstrap 방법과 같은 대안적인 방법을 이용하여 추론을 하는 것도 가능하다.

Leverage

- 예측변수의 공간에서 동떨어져 있는 관측치를 나타낸다.
- $\sum_i h_{ii} = p$이고 $\bar{h} = p/n$
- 보통 $2p/n$을 넘으면 leverage point로 간주한다.
- leverage point가 항상 influential point, 즉 모형에 심대한 영향을 주는 것은 아니다.

영향 측도

- Cook's distance
- DFFITS
- DFBETAS
- deletion diagnostic에 기반한 측도들

과제

- Delivery time data에 대하여
  - (1) residual, standardized residual, studentized residual, press residual을 각각 계산하여라.
  - (2) PRESS 통계량을 산출하여라. 예측변수를 n.prod, distance 각각 1개씩으로 했을 때와 두 예측변수를 모두 사용하였을 때 어떤 모형이 더 우수한지를 판단하여라.
  - (3) 잔차도를 통하여 (2)에서 우수한 것으로 판단된 선형모형 가정을 검증하여라.
  - (4) 가정이 잘 만족되는가? 만족되지 않는다면 어떤 식으로 해결할지 구체적으로 기술하여 보아라.