6.4 일반선형모델의 평가

잔차는 데이터와 모델의 괴리를 표현하는 중요한 지표. 모델의 손실을 파악하는 방법 배우겠음.

6.4.2 피어슨 잔차

이항분포에서**피어슨 잔차**는 다음과 같이 계산.

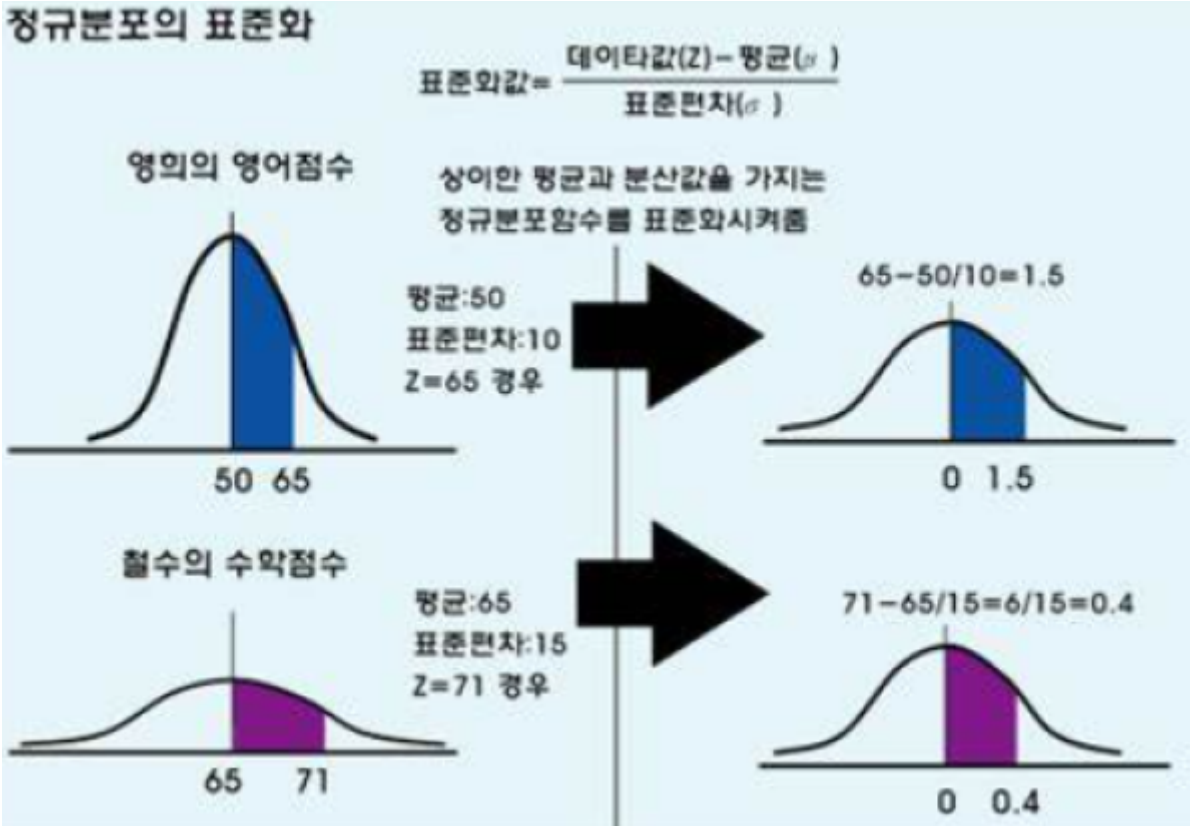
- y는 종속변수(이항확률 변수, 이 경우에는 합격여부 데이터)
- N은 시행횟수
- p̂은 추측한 성공확률(mod_glm.predict()로 계산한 예측치)
- 하나하나의 예측 결과에서 N이 당연히1

$$Pearson\ residuals = \frac{y - N\hat{p}}{\sqrt{N\hat{p}(1-\hat{p})}}$$

6.4.3 피어슨 잔차의 해석

피어슨 잔차의 분모의 Np̂(1-p̂)는 이항분포의 분산의 값과 일치. 그 값에 루트를 취한 것이므로 분모는 이항분포의 표준편차로 볼 수 있습니다.

정규선형모델에서는 종속변수와 predict() 함수로 구한 예측값의 차이를 잔차로 사용했다.y-p̂를 잔차로 사용하는 느낌. 피어슨은 일반 잔차를 분포의 표준편차로 나눈것



N을 고정했을 때 이항분포의 **분산 Np(1-p)가 최고로 클 때는 p=0.5.**

p=0.9와 같이 합격이 거의 확실히 예측되는 경우에는 **분산이 작아진다.** 이 때의 예측을 제외하면 큰 차이라고 간주할 수 있다. 이게 **피어슨 잔차**이다.

피어슨 잔차의 제곱합은 **피어슨 카이제곱통계량**이라고도 부른다.

$$e_i = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

6.4.5 deviance

6.4.6 deviance해석

deviance(디비언스)는 모델의 적합도를 평가하는 지표. deviance가 크면 모델이 맞지 않는다. deviance는 `잔차제곱합을 우도처럼 표현한것`. 즉, 최대우도법의 결과와 `deviance라는 손실을 최소화`하도록 파라미터를 추정한 결과는 일치한다.

$$deviance = 2[\log \mathcal{L}(\beta_{max}; y) - \log \mathcal{L}(\beta_{glm}; y)]$$

로지스틱 회귀의 로그우드함수를

$$\log \mathcal{L}(\beta_0, \beta_1; N, m)$$

라고 한다. 계수(B0,B1을 바꾸면 우도가 변한다.)

최대우도법을 추정한 로지스틱 회귀의 계수에 근거한 로그우도를

$$\log \mathcal{L}(\beta_{glm}; y)$$

라고 한다. 모든 합격 여부를 완전히 예측할 수 있을 때의 로그우도이다.

즉, 합격(1)이라면 성공확률 100%, 불합격이라면 성공확률0%로 예측할 때의 로그우도. 여기서 차이를 측정한 것이 deviance이다.

6.4.7 deviance와 우도비 검정

deviance할 때 2를 곱하는 이유는 우도비 검정을 할 때 편해서.

deviance= 잔차제곱합 (같은 의미를 갖는 지표)

때문에, 모델의 **deviance의 차이**를 통계량으로 한 검정은 **분산분석**처럼 해석이 가능하다. 이때 deviance를 앞서와 같이 정의해주면 **deviance의 차이가 카이제곱분포**로 불리는 분포에 점근적으로 따르게 된다.

deviance의 차이를 검정하는 것을 **우도비 검정**이라고 한다. R언어에서는 분산분석과 우도비 검정이 같은 anova 함수로 구현되어 있다.

6.4.8 deviance 잔차

이항분포에서 **deviance 잔차**는 *deviance잔차제곱합이 deviance가 된다*는 사실로 계산한다.

(주피터!!!)

6.4.9 교차 엔트로피 오차

머신러닝에서 로지스틱 회귀를 **교차 엔트로피 오차**의 최소화라는 관점으로 설명하는 일이 자주 있다.

$$\mathrm{Bin}(m \mid N, p) = {}_N C_m \cdot p^m \cdot (1-p)^{N-m}$$

데이터 하나하나니까 N=1

m=0 or 1

$$\mathrm{Bin}(m \mid N, p) = p^m \cdot (1-p)^{1-m}$$

y= 합(1)/ 불(0), 예측한 합격률을 p̂

$$\mathrm{Bin}(y \mid 1, \hat{p}) = \hat{p}^y \cdot (1-\hat{p})^{1-y}$$

아래는 우도함수, T는 샘플사이즈

$$\prod_{i=1}^{T} \hat{p}_i^{y_i} \cdot (1-\hat{p}_i)^{1-y_i}$$

우도함수에 -1를 곱하면 아래와 같다

$$-\sum_{i=1}^{T} [y_i \log \hat{p}_i + (1-y_i) \log(1-\hat{p}_i)]$$

이것을 **엔트로피 오차라고** 부른다.

모집단분포가 이항분포라고 가정하는 경우 deviance와 같은 의미를 갖는다.

교차 엔트로피 오차를 최소로 하는 것은 deviance를 최소로 하는 것과 동일

로지스틱 회귀의 로그우도를 최대로 하는 것과 같다.