데이터마이닝
(Data Mining)

한국방송통신대학교
정보통계학과 장영재 교수

2 강 /

회귀모형

목차

2. 회귀모형

1) 선형모형

2) 로지스틱 회귀모형

3) 범주형 입력변수 처리

4) 모형 구축을 위한 변수 선택

1. 선형회귀모형

선형회귀모형

1) 모형의 정의

- 총 n의 객체(subject) 중에 i 번째 객체에 대한 연속형 목표변수 값을 Y_i, 입력변수들의 값을 $X_{1i}, X_{2i}, \cdots, X_{pi}$라고 할 때, 선형회귀 모형은 다음과 같이 정의

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi} + \epsilon_i, \quad i = 1, \ldots, n$$

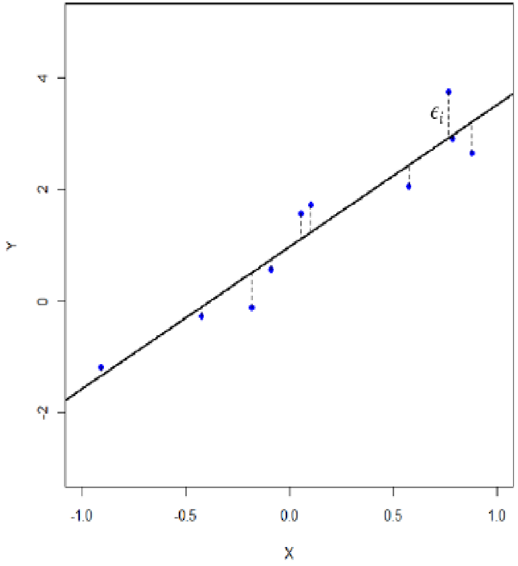
- $\beta_0, \beta_1, \beta_2, \cdots, \beta_p$ 를 회귀모수(regression parameters) 또는 회귀계수(regression coefficients)로서 알려지지 않은 상수(constant)

- ϵ_i 는 Y_i 의 근사에서 오차(error)

✓ 각 객체들의 오차는 서로 독립(independent)이고 평균이 0이고 일정한 분산(constant variance)을 가진 정규분포(normal distribution)를 따른다고 가정

선형회귀모형

2) 회귀모수의 추정



〈그림1〉 산점도와 회귀직선

선형회귀모형

2) 회귀모수의 추정

- 입력변수 X와 목표변수 Y의 산점도 〈그림 1〉에서 보듯이, 각 관측치로 부터 회귀 직선까지의 수직 거리 제곱의 합을 최소화하는 회귀모수를 찾는 최소제곱추정법 (least square estimation; LSE)을 주로 이용

- 오차 $\epsilon_i = Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \dots \beta_p X_{\pi}$ 의 제곱 합

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_{1i} - \beta_2 X_{2i} - \dots \beta_p X_{\pi})^2$$

을 최소화하는 추정값을 각각 라고 할 때, 이를 이용한 최소제곱 회귀직선(least square regression line)은

$$\widehat{Y_i} = \widehat{\beta_0} + \widehat{\beta_1} X_{1i} + \widehat{\beta_2} X_{2i} + \dots + \widehat{\beta_p} X_{pi}, \quad i = 1, \dots, n$$

선형회귀모형

3) 회귀계수의 해석

- 회귀계수 β_j 는 다른 입력변수들이 일정할 때 j 번째 입력변수가 한 단위 변동할 때 대응하는 Y의 변동 양으로 해석

- β_j는 다른 입력변수를 보정한 후에 Y에 대한 X의 기여도
  - 회귀계수 β_j가 양수(positive)라면 X_j가 증가할 때 Y도 증가하고, 반대로 β_j가 음수(negative)라면 X_j가 증가할 때 Y는 감소

선형회귀모형

4) 입력변수의 중요도

- 선형회귀모형에서 변수의 중요도는 t값으로 측정

- j 번째 입력변수 X_j에 대한 t값은 다음과 같이 정의

$$t_j = \frac{\hat{\beta}_j}{SE(\hat{\beta}_j)}$$

- 단, $SE(\hat{\beta}_j)$ 는 j번째 회귀계수의 추정치 $\hat{\beta}_j$ 의 표준오차
  - t값의 절대 값이 클수록 영향력이 크다고 할 수 있음

선형회귀모형

5) 모형의 적합도

- 모형의 상수항 β_0을 제외한 모든 회귀계수가 0인지 아닌지를 검정하는 척도가 F-값

- F-값은 회귀직선에 의해 평균적으로 설명할 수 있는 부분(mean squared regression; MSR)을 설명할 수 없는 부분(mean squared error; MSE)으로 나눈 값

$$F = \frac{MSR}{MSE} = \frac{SSR/p}{SSE/(n-p-1)} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2/p}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-p-1)}$$

  - F-값이 크면 개 입력변수들 중에 최소한 하나는 유의하다(회귀계수가 0이 아니다)라는 뜻이고, F-값이 작아서 P-값이 (보통 0.05보다) 크면 모든 입력변수가 유의하지 않아서 회귀직선이 유용하지 않음

선형회귀모형

5) 모형의 적합도

- 모형의 적합도(goodness-of-fit)를 결정계수 (coefficient of determination) R^2로 측정

- 결정계수 R^2 는 설명할 수 있는 부분의 총합을 변동의 총합으로 나눈 값으로 0과 1 사이의 값

$$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}, \quad 0 \leq R^2 \leq 1.$$

- 단, SST (sum of squared total variation)는 총 변동합

선형회귀모형

5) 모형의 적합도

- 모형에 포함된 변수의 수(p)가 증가하면 할수록 R^2는 증가하므로 변수의 수가 다른 모형을 비교할 때는 수정된(adjusted) R^2를 사용

$$R_a^2 = \text{adjusted}$$

$$R^2 = 1 - \frac{n-1}{n-p-1}(1-R^2), \quad 0 \le R_a^2 \le 1.$$

선형회귀모형

5) 모형의 적합도

- 입력변수의 수가 다른 모형들을 비교 평가하는 기준으로 AIC(Akaike Information Criterion)도 종종 사용

- 여러 후보 모형들 중에서 가장 작은 AIC를 가지는 모형을 선택한다.

$$AIC = n\log(SSE/n) + 2p$$

선형회귀모형

6) 모형을 이용한 예측

- 주어진 데이터에 기반하여 회귀식을 얻었을 때, 임의의 객체 i^*에 대해 관측한 입력변수의 값 $x_{1i^*}, \ x_{2i^*}, \ \cdots \ , \ x_{pi^*}$을 그 회귀식에 대입하여 목표변수의 예측 값 $\widehat{y_{i^*}}$을 얻을 수 있음

$$\widehat{y_{i^*}} = \widehat{\beta_0} + \widehat{\beta_1} x_{1i^*} + \widehat{\beta_2} x_{2i^*} + \ldots + \widehat{\beta_p} x_{pi^*}$$

선형회귀모형

7) 예측력

- 목표변수가 연속형인 경우에 모형의 예측력 측도로서 MSE(mean squared error)를 주로 사용

$$MSE = \sum_{i=1}^{n} (y_i - \hat{y_i})^2 / n$$

- 불편성(unbiasedness)을 가지게 하기 위해 n대신 (n-p-1) 로 나눈 값으로 사용하기도 함
- 시각적으로 관측값(y_i)과 예측값($\hat{y_i}$)의 차이를 확인하기 위해서는 이들을 가로 및 세로축에 놓고 그린 산점도가 45도 대각선을 중심으로 모여 있으면 예측력이 좋다고 평가

2. 로지스틱 회귀모형

2. 로지스틱 회귀모형

- 목표변수가 두 개의 범주를 가진 이항형인 경우, 선형회귀모형을 적용하면 0 또는 1과 다른 예측 값을 얻거나 범위를 넘어선 값을 얻게 될 가능성

  - ex) 목표변수의 두 범주 값 "신용이 좋다"는 1,
    "신용이 나쁘다"는 0인 경우

- 이를 방지하기 위해 목표변수의 값이 1인 확률의 로짓변환과 입력변수들의 선형 함수 관계로 나타내는 모형인 로지스틱 회귀모형을 이용

  - 목표변수가 두 값 중에 하나("실패" 0과 "성공" 1 중에 주로 "성공" 1)를 가지는 확률을 모형화

2. 로지스틱 회귀모형

1) 모형의 정의

- 이항형 목표변수 값을 y_i라고 하고 (범주 값은 1과 0) 목표변수가 “성공” 1을 가질 확률을 $\pi_i = \Pr(Y_i = 1)$할 때, 로지스틱 회귀모형은 다음과 같이 정의

$$\pi_i = \frac{\exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi})}{1 + \exp(\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi})}, \quad i = 1, \ldots, n$$

- 단, X_{1i}, X_{2i}, … , X_{pi} 는 입력변수의 값이고 이항형 목표변수는 이항분포 (binomial distribution)를 따른다고 가정
  - 로지스틱 회귀모형은 다음과 같이 변환하여 표시할 수 있음

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_p X_{pi}, \quad i = 1, \ldots, n$$

✓ β_0, β_1, β_2, … , β_p는 회귀모수(regression parameters) 또는 회귀계수 (regression coefficients)로서 추정의 대상

2. 로지스틱 회귀모형

1) 모형의 정의

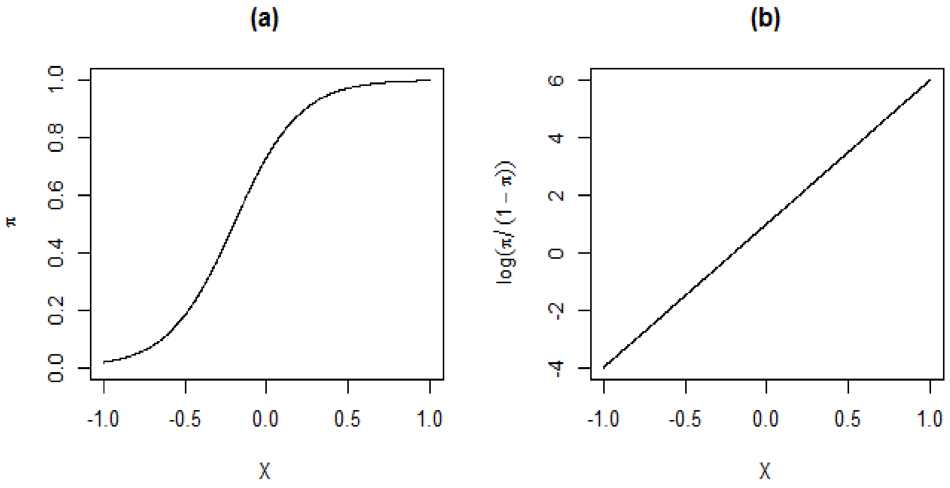
- 성공 확률 π과 입력변수와 관계는 종종 로지스틱 반응함수로 표현

- 대체로 S-형태의 곡선으로 입력변수가 증가함에 따라 초기에는 천천히 증가하다가 증가 속도가 점차 빨라지고 확률 1/2 이후에는 다시 증가 속도가 줄어드는 성장 곡선(growth curve) 형태
- 성공 확률과 실패 확률의 비를 오즈비(odds ratio)라고 하고 오즈비에 로그(log)를 취한 것을 로짓 변환(logit transformation)이라고 부름

✓ 입력변수와 로짓의 관계는 직선

2. 로지스틱 회귀모형

1) 모형의 정의



〈그림 2〉 로짓함수와 선형함수

2. 로지스틱 회귀모형

2) 회귀모수의 추정

- 로지스틱 회귀모형의 회귀모수는 최대우도추정법(maximum likelihood estimation method)에 의해 추정

- 데이터의 확률함수를 모수 β 의 함수로 취급한 것을 우도함수(likelihood function) $L(\beta)$라고 하고, 이 우도함수가 최대가 될 때 모수의 추정 값 $\widehat{\beta}_0,\ \widehat{\beta}_1,\ \ldots,\ \widehat{\beta}_p$ 이 최대우도추정치(maximum likelihood estimate, MLE)

– 우도함수를 최대화하는 모수의 추정 값은 뉴턴-랩슨 (Newton-Raphson) 또는 피셔 스코링 (Fisher scoring) 방법에 의해 반복적으로 계산

$$\widehat{\pi}_i = \frac{\exp(\widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \ldots + \widehat{\beta}_p X_{pi})}{1 + \exp(\widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \ldots + \widehat{\beta}_p X_{pi})}, \quad i = 1, \ldots, n$$

또는 $$\log\left(\frac{\widehat{\pi}_i}{1 - \widehat{\pi}_i}\right) = \widehat{\beta}_0 + \widehat{\beta}_1 X_{1i} + \widehat{\beta}_2 X_{2i} + \ldots + \widehat{\beta}_p X_{pi}, \quad i = 1, \ldots, n$$

2. 로지스틱 회귀모형

3) 회귀계수의 해석

- 회귀계수 β_j는 다른 입력변수들을 보정한 후 성공(Y=1)의 로그오즈(log odds = $\log(\pi/(1-\pi))$) 에 미치는 X_j의 효과

• 다른 입력변수가 일정할 때, $\exp(\beta_j)$ 는 j번째 입력변수 X_j가 한 단위 변동할 때 오즈에 미치는 기여도

– 회귀계수 β_j가 양수라면 X_j가 증가할 때 성공 확률 π와 로짓 $\log(\pi/(1-\pi))$ 은 증가하고, 반대로 β_j가 음수라면 X_j가 증가할 때 이들은 감소

2. 로지스틱 회귀모형

4) 변수의 중요도

- 선형회귀모형에서와 유사하게 로지스틱 회귀모형에서 변수의 중요도는 z값으로 측정할 수 있음

  - j 번째 입력변수 X_j 에 대한 z값은 다음과 같이 정의

$$z_j = \frac{\widehat{\beta}_j}{SE(\widehat{\beta}_j)}$$

  – 단, $SE(\widehat{\beta}_j)$ 는 번째 회귀계수의 추정치 $\widehat{\beta}_j$ 의 표준오차

2. 로지스틱 회귀모형

5) 모형의 적합도

- 모형의 적합도의 측도로서 이탈도(deviance)를 사용할 수 있음

• 이탈도란 어떤 모형 M의 최대로그우도(maximized log-likelihood) $\log(L_M)$ 에서 포화모형(saturated model) S의 최대로그우도 $\log(L_S)$ 를 뺀 것에 - 2를 곱한 값

$$\text{Deviance} = -2\left[\log(L_M) - \log(L_S)\right]$$

- 포화모형은 각 관측에 모수 하나씩 사용하여 완벽한 모형을 의미

- 이탈도가 클 경우에 포화모형에 비해 그 모형은 적합하지 않다고 평가

- 데이터를 모형에 적합하여 얻은 이탈도에 대응하는 P-값이 (보통, 〉0.05) 클 때 우리는 그 모형 M이 의미 있다고 해석

2. 로지스틱 회귀모형

5) 모형의 적합도

- 입력변수의 수가 다른 모형들을 비교 평가하는 기준으로 AIC(Akaike Information Criterion)도 종종 사용

$$AIC = -2\log(L_M) + 2p$$

- L_M 은 모형 M에 대한 우도함수(likelihood function)의 최대값
  p는 모수의 수

– AIC는 입력변수(또는 모수)의 수가 증가한다고 항상 작아지지는 않으므로, 여러 후보 모형들 중에서 가장 작은 AIC를 가지는 모형을 선택

2. 로지스틱 회귀모형

6) 모형을 이용한 예측

- 임의의 객체 i^*에 대해 관측한 입력변수의 값 $x_{1i^*},\ x_{2i^*},\ \ldots,\ x_{pi^*}$ 을 그 로지스틱회귀식에 대입하여 성공 확률 $\pi_{i^*} = \Pr(Y_{i^*} = 1)$의 예측 값 $\widehat{\pi_{i^*}}$ 을 산출

$$\widehat{\pi_{i^*}} = \frac{\exp(\widehat{\beta_0} + \widehat{\beta_1} x_{1i^*} + \widehat{\beta_2} X_{2i^*} + \ldots + \widehat{\beta_p} X_{pi^*})}{1 + \exp(\widehat{\beta_0} + \widehat{\beta_1} x_{1i^*} + \widehat{\beta_2} x_{2i^*} + \ldots + \widehat{\beta_p} x_{pi^*})}$$

2. 로지스틱 회귀모형

6) 모형을 이용한 예측

- 예측값 $\widehat{\pi_{i^*}}$ 이 크면 $\widehat{y_{i^*}} = 1$, $\widehat{\pi_{i^*}}$ 이 작으면 $\widehat{y_{i^*}} = 0$ 으로 분류

$$\widehat{y_{i^*}} = \begin{cases} 1, & \widehat{\pi_{i^*}} > \pi_0 \\ 0, & \widehat{\pi_{i^*}} \le \pi_0. \end{cases}$$

- 크고 작음을 분류하는 임계치(π_0)는 보통 0.5를 사용하지만, 적용 분야에 따라 달리 결정할 수 있음

7) 예측력

- 목표변수가 이항형인 경우에 모형에 의해 분류한 값에 기반하여 정오분류표를 만들어 예측력을 평가

- 실제 범주가 1일 때 1로 예측한 빈도 n_{11} 과 실제 범주가 0일 때 0으로 예측한 빈도 n_{00} 가 클수록 예측을 잘했다고 간주할 수 있음

2. 로지스틱 회귀모형

7) 예측력

〈표〉 정오분류표

		예측 범주		합계
		1	0	
실제 범주	1	n_{11}	n_{10}	n_{1+}
	0	n_{01}	n_{00}	n_{0+}
합계		n_{+1}	n_{+0}	n

2. 로지스틱 회귀모형

7) 예측력

- 예측력의 측도로 민감도(sensitivity)와 특이도(specificity) 등을 사용

- 민감도는 실제 양성(Y=1)일 때 양성으로 예측할 확률,
  특이도는 실제 음성(Y=0)일 때 음성으로 예측할 확률
- 예측 정확도(prediction accuracy)는 실제 양성인데 양성으로, 음성일 때 음성으로 제대로 예측할 확률로 민감도와 특이도의 가중평균
- 오분류율(misclassification rate)는 양성인데 음성으로, 음성일 때 양성으로 잘못 예측할 확률

2. 로지스틱 회귀모형

7) 예측력

- 민감도 $= \Pr(\hat{Y}=1 | Y=1) = n_{11}/n_{1+}$
- 특이도 $= \Pr(\hat{Y}=0 | Y=0) = n_{00}/n_{0+}$
- 예측 정확도 $= \Pr(\hat{Y}=1, Y=1) + \Pr(\hat{Y}=0, Y=0)$
  $= (n_{11}+n_{00})/n$
- 오분류율 $= \Pr(\hat{Y}\neq 1, Y=1) + \Pr(\hat{Y}\neq 0, Y=0)$
  $= (n_{10}+n_{01})/n$

2. 로지스틱 회귀모형

7) 예측력

- 민감도와 특이도는 임계치(π_0)에 따라 달라지고, 임계치는 상황에 따라 다르게 결정할 수 있음

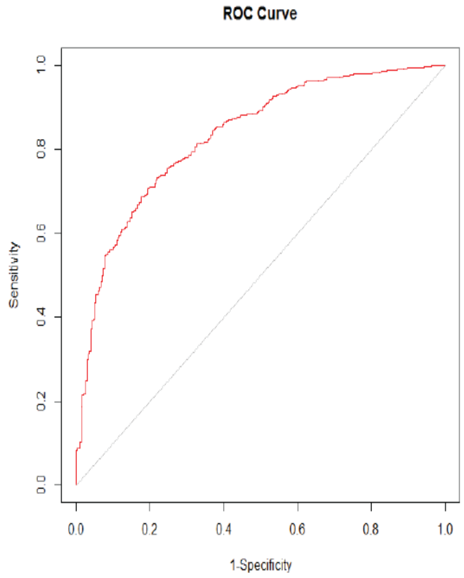
- 여러 가능한 임계치에 대해 (1-특이도)를 가로축에, 민감도를 세로축에 놓고 그린 그래프가 ROC(receiver operating characteristic) 곡선

- 민감도와 특이도가 높을수록 예측력이 좋다고 할 수 있기 때문에 ROC 곡선이 좌상단에 가까울수록 ROC 곡선 아래 면적(AUC; area under the ROC curve)이 커지고, AUC가 커질수록 예측력이 좋다고 평가

3. 범주형 입력변수 처리

3. 변수형 입력변수 처리

- 입력변수가 범주형일 경우에는 가변수(dummy variable)로 변환하여 처리



〈그림3〉 ROC 곡선

3. 변수형 입력변수 처리

- 어떤 입력변수 X가 3개의 범주(a, b, c)를 가진다고 할 때, 두 개의 가변수를 다음과 같이 새롭게 정의

$$X' = \begin{cases} 1, & X = a \\ 0, & X \neq a \end{cases} \qquad X'' = \begin{cases} 1, & X = b \\ 0, & X \neq b \end{cases}$$

- X가 범주 a를 가지는 경우 $X' = 1, \; X'' = 0$
- 범주 b를 가지는 경우 $X' = 0, \; X'' = 1$
- 범주 c를 가지는 경우 $X' = 0, \; X'' = 0$

– X가 L개 범주를 가지는 경우 L-1개의 가변수를 새롭게 생성한다.

4. 모형 구축을 위한 변수 선택

4. 모형 구축을 위한 변수 선택

1) 모형 구축을 위한 변수 선택

- 모형은 데이터를 잘 설명할 수 있을 만큼 충분히 복잡해야 하고, 과적합(overfitting)하지 않고 해석하기 좋게 단순해야 함

  - 입력변수가 너무 많으면 유지하기 비효율적이고, 예측오차(prediction error)가 큼
  - 중요한 변수를 제거하면 중요한 정보를 잃어버리고 편향(bias)이 발생

4. 모형 구축을 위한 변수 선택

1) 모형 구축을 위한 변수 선택

- 입력변수들의 모든 가능한 조합을 평가하여 유의한 입력변수만을 포함한 가장 적절한 모형을 선택하는 방법을 활용

① 후진소거법(backward elimination) : 모든 변수를 포함시킨 모형부터 시작하여 가장 유의하지 않은 변수를 하나씩 제거하여 유의한 변수만 남을 때까지 진행

② 전진선택법(forward selection) : 상수항만 가진 모형부터 시작하여 가장 유의한 변수를 하나 씩 포함시켜 포함되지 않고 남은 변수들이 모두 유의하지 않을 때까지 진행

③ 단계적선택법(stepwise selection) : 전진선택법처럼 상수항부터 시작하여 가장 유의한 변수를 하나씩 모형에 포함시키지만, 어떤 변수가 포함된 이후에 기존에 포함된 변수 중에 유의하지 않은 변수를 제거하는 과정이 포함

강의를 마쳤습니다.

다음시간에는...