제16장 단순회귀분석

회귀분석(regression analysis)

- 회귀분석: 변수와 변수 사이의 관계를 알아보기 위한 통계적 분석방법, 독립변수의 값에 의하여 종속변수의 값을 예측하위 위함
  - 독립변수(independent variable): 종속변수에 영향을 미치는 변수
  - 종속변수(dependent variable): 분석의 대상이 되는 변수
- 회귀분석의 분류
  - 단순회귀분석(simple regression analysis): 하나의 종속변수와 하나의 독립변수의 관계를 분석
  - 다중회귀분석(multiple regression analysis): 하나의 종속변수와 둘 이상의 독립변수간의 관계를 분석
  - 단순회귀분석이 간단하고 결과의 해석도 명확하지만 종속변수를 하나의 독립변수로 설명하기 어려운 경우가 많다.

두 변수간의 관계

- 두 변수간의 관계
  - 함수적 관계(functional relation): X값을 알면 Y값을 정확히 알 수 있는 관계
  - 통계적 관계(statistical relation): X값에 대한 Y의 값이 유일하게 결정되지 않는 관계
  - 산포도(scatter diagram): X와 Y의 통계적 관계를 그림으로 나타낸 것

단순선형회귀모형

- 선형모형(linear model): 두 변수간의 관계가 비례적인 선형관계일 때
- 비선형모형(nonlinear model): 두 변수간의 관계가 비선형관계로 나타날 때
- 비선형의 경우도 변수를 적절히 조작하여 선형으로 바꿀 수 있기 때문에 선형회귀모형이 중요한 의미를 가짐
- 산포도의 작성
  - 두 변수간의 관계가 선형인지 비선형인지 알아보기 위해 산포도를 그려본다.

선형회귀모형(1)

- X와 Y가 1차식으로 나타날 때 선형회귀모형이 되고 다음과 같은 전제가 필요
  - 독립변수 X의 각 값에 대한 Y의 확률분포가 존재한다.
  - Y의 확률분포의 평균은 X값의 변함에 따라 일정한 추세를 따라 움직인다.

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{회귀식}} + \underbrace{\varepsilon_i}_{\text{오차}}$$

회귀모형에서 ε_i에 대한 기본가정

1. ε_i는 정규분포의 형태를 이룬다.
2. ε_i의 기대치는 0이다. 즉, $E(\varepsilon_i) = 0$. 이 가정은 실제 관찰값이 회귀선상에 있는 점을 중심으로 분포되어 있다는 뜻이다.
3. ε_i의 분산은 모든 X값에서 동일하다.

   $\sigma^2(\varepsilon_i) = \sigma^2$
4. ε_i들은 서로 독립적이다.

선형회귀모형(2)

○ 단순회귀모형

$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad i = 1, 2, \cdots, n$

단, Y_i : i번째 종속변수의 값

X_i : i번째 독립변수의 값

β_0 : 선형회귀식의 절편

β_1 : 선형회귀식의 기울기

ε_i : 오차항으로 ε_i는 독립적이며 $N(0, \sigma^2)$의 분포를 이룬다.

Y의 확률분포

- 오차항 ε_i가 확률변수이므로 Y_i도 확률변수가 된다. 또한 $E(\varepsilon_i)=0$이므로 Y_i의 기대치는 다음과 같다.

$$E(Y_i) = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \beta_0 + \beta_1 X_i$$

위 식을 모든 X와 Y에 대하여 나타내면

$$E(Y) = \beta_0 + \beta_1 X_i$$

이것이 바로 회귀함수(*regression function*)가 된다.

Y_i의 확률분포

1. Y_i는 정규분포를 이룬다.
2. $E(Y_i) = \beta_0 + \beta_1 X_i$
3. $\sigma^2(Y_i) = \sigma^2$

선형회귀함수의 추정

- 선형회귀함수의 모수 β_0과 β_1은 표본으로 점추정한다.(구간추정은 다음장에서)
- 최소자승법(least square method)
  - 잔차의 제곱의 합을 최소화하는 회귀식

$E(Y) = \beta_0 + \beta_1 X$일 때 β_0과 β_1에 대한 점추정량은 b_0와 b_1으로 나타내고 추정회귀식은 $\hat{Y} = b_0 + b_1 X$과 같다.

여기서 $\hat{Y}$은 $E(Y)$에 대한 추정량이므로 X_i값에 대한 Y_i의 추정치는

$\hat{Y}_i = b_0 + b_1 X_i$

이때 관찰값 Y_i와 예측치 $\hat{Y}_i$가 일치하지 않는 것이 보통이고, 이 차이를 잔차(*residual*)라고 하고 e_i로 표시한다.

$e_i = Y_i - \hat{Y}_i$

추정회귀식의 산출

- 잔차제곱의 합을 Q로 놓으면

$$Q = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$$

Q를 최소로 하는 b_0와 b_1의 값을 귀하기 위하여 Q를 b_0와 b_1에 대하여 편미분하면

$$\sum Y_i = nb_0 + b_1 \sum X_i$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2$$

위식을 정규방정식(*nomal equation*)이라고 하며

이 방정식을 풀면 모집단회귀계수 β_0과 β_1의 추정량 b_0와 b_1을 구할 수 있다.

- 회귀계수의 추정
  - 모집단회귀계수 β_0과 β_1은에 대한 최소자승추정량은 다음과 같다.

$$b_1 = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2}$$

$$b_0 = \frac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \frac{\sum Y_i}{n} - \frac{b_1 \sum X_i}{n} = \bar{Y} - b_1 \bar{X}$$

잔차

- 오차(error term): ε_i는 관찰값 Y_i와 모집단회귀식과의 편차
- 잔차(residual): e_i는 관찰값 Y_i와 추정회귀식과의 편차

$$e_i = Y_i - \hat{Y}_i$$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \qquad Y_i = b_0 + b_1 X_i + e_i$$

$$\text{즉, } \varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i) \qquad e_i = Y_i - (b_0 + b_1 X_i)$$

- ε_i는 실제 알 수 없기 때문에 e_i에 의해 추정
- e_i는 통계적 추정에 있어 무작위변동의 크기를 측정, 회귀모형이 기본가정을 만족하는지를 평가(다음 장)
- 최소자승법의 속성상 잔차의 합은 항상 0이 된다.

$$\sum e_i = \sum (Y_i - \hat{Y}_i) = \sum (Y_i - b_0 - b_1 X_i) = \sum Y_i - n b_0 - b_1 \sum X_i$$

$$= \sum Y_i - n(\bar{Y} - b_1 \bar{X}) - b_1 \sum X_i = \sum Y_i - \sum Y_i + b_1 \sum X_i - b_1 \sum X_i = 0$$

단순선형회귀모형의 분산분석(1)

- 분산분석(analysis of variance)
  - 회귀모형의 설명력과 유용성을 분석하는 효과적인 방법을 보통 ANOVA라 부른다.
- 제곱의 합의 계산
  - SSTO (total sum of squares)
    - 독립변수 X를 이용하지 않고 종속변수 Y를 예측하는데 있어서 불확실성은 관찰값 Y의 변이성에 기인하며 다음의 편차에 의하여 측정한다.

      $Y_i - \bar{Y}$

      만약 모든 관찰값 Y_i가 동일하다면, 모든편차에 대하여 $Y_i - \bar{Y} = 0$ 즉, Y의 변이성이 클수록 $Y_i - \bar{Y}$는 커지게 되며 X를 이용하지 않고 Y를 예측하는데 불확실성이 커진다.

      $$SSTO = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

단순선형회귀모형의 분산분석(2)

- 제곱의 합의 계산
  - SSE (error sum of squares)
    - 종속변수 Y를 예측하는데 독립변수 X의 정보를 이용하게 되면, 예측의 불확실성은 추정회귀식 주위에 흩어진 Yi의 변이성에 해당되며 다음의 편차로 측정

      $Y_i - \hat{Y}_i$

      만약 모든 관찰값이 회쉬선상에 위치하면 $Y_i - \hat{Y}_i = 0$

      즉, $Y_i - \hat{Y}_i$의 값이 커질수록 X를이용한 Y의 예측에 불확실성이 커진다.

      $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$ 그런데 $Y_i - \hat{Y}_i = e_i$이므로 $SSE = \sum_{i=1}^{n} e_i^{\,2}$

단순선형회귀모형의 분산분석(3)

- 제곱의 합의 계산
  - SSR (regression sum of squares)
    - 종속변수 Y를 예측하는데 독립변수 X의 정보를 이용하여 총변동에서 줄어든 부분 즉, 회귀식에 의하여 변동이 줄어든 양

$$SSR = SSTO - SSE$$

$$SSR = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

만약 $SSR = 0$이라면 독립변수 X를 도입하여도 변이성이 전혀 줄어들지 않는 경우라고 할 수 있다.

- SSR: X를 도입함으로 인하여 Y_i의 총변동이 줄어든 부분
- SSE: X를 도입한 후에도 남아 있는 Y_i의 변동부분

단순선형회귀모형의 분산분석(4)

- 단순회귀모형에서 총변동은 다음과 같이 나누어 질 수 있다.

$$\underbrace{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}_{SSTO}=\underbrace{\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2}_{SSR}+\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}_{SSE}$$

$$SSTO=\sum Y_i^2-\frac{\left(\sum Y_i\right)^2}{n}$$

$$SSR=\frac{\left(\sum X_iY_i-\frac{\sum X_i\sum Y_i}{n}\right)^2}{\sum X_i^2-\frac{\left(\sum X_i\right)^2}{n}}$$

$$SSE=SSTO-SSR$$

분산분석표

- 자유도
  - 제곱의 합은 각각의 자유도(degree of freedom)를 가진다.
    - SSTO의 자유도 n-1
    - SSE의 자유도 n-2
    - SSR의 자유도 1
  - 자유도의 분해
    - SSTO의 자유도=SSR의 자유도 + SSE의 자유도
    - N-1=1+n-2
- 평균제곱(mean square)
  - 제곱의 합을 자유도로 나눈 값
  - MSR=SSR/1
  - MSE=SSE/n-2
- 분산분석표

결정계수와 상관계수(1)

- 회귀분석결과 독립변수가 종속변수를 얼마나 잘 설명하고 있는지 측정이 필요

- 결정계수
  - 회귀식의 적합도를 SSTO에 대한 SSR의 상대적인 크기에 의하여 측정

$$r^2 = \frac{SSR}{SSTO} = \frac{SSTO - SSE}{SSTO} = 1 - \frac{SSE}{SSTO}$$

$$0 \le r^2 \le 1$$

결정계수와 상관계수(1)

- 상관계수(coefficient of simple correlation)
  - 결정계수의 제곱근
    - 5차시에서는 공분산을 이용하여 계산

$$r = \pm\sqrt{r^2}$$

여기에서 부호는 b_1의 부호를 따른다.

$$-1 \le r \le 1$$

  - 상관계수는 실무에서 많이 이용되지만 회귀식에 의한 전체 변량의 비율 즉, 회귀식에 의하여 설명되는 부분을 직접적으로 나타내지는 않는다.
  - 그리고 0과 1사이의 값을 제곱근을 취하게 되면 그 값이 커지게 되므로 실제 보다 더 밀접한 관계가 있는 것 처럼 볼일 수 있다는 점에 주의하여 한다.