통계학 개론

제8장 통계모형: 상관분석과 회귀분석

8.1 상관분석

상관분석(correlation analysis): 두 변수 간 상호의존 관계가 있을 경우 이 관계가 어느 정도 밀접한가를 측정하는 분석방법

공분산(covariance): 두 변수 사이의 상호관계를 나타내는 측도

$$C_{XY} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$
$$= \frac{1}{n-1}\left(\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}\right)$$

여기서, $\bar{x}$, $\bar{y}$는 X, Y의 표본평균

공분산은 X와 Y의 단위에 의존하기 때문에 비교할 때 불편하는 단점
표본상관계수(sample correlation coefficient): 변수의 종류나 특정 단위에 관계없는 측도를 구하기 위해 표본공분산 S_{XY}를 X와 Y의 표본표준편차인 SX와 SY의 곱으로 나누어 표준화시킨 측도

$$r = \frac{S_{XY}}{S_X S_Y} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$
$$= \frac{\sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}\sqrt{\sum_{i=1}^{n} y_i^2 - n\bar{y}^2}}$$

❖ 표본상관계수 r의 성질

① r은 -1과 +1 사이의 값을 가지며, r의 값이 +1에 가까울수록 강한 양의 선형관계를, -1에 가까울수록 강한 음의 상관관계를 나타내며, r의 값이 0에 가까울수록 선형관계는 약해진다.

② X와 Y의 대응되는 모든 값이 한 직선상에 위치하면 r의 값은 -1(직선의 기술기가 음인 경우)이나 +1(직선의 기울기가 양인 경우)의 값을 가진다.

③ 표본상관계수 r은 단지 두 변수의 선형관계만을 나타내는 측도이다. 그러므로 두 변수의 선형상관관계는 없지만 다른 관계는 가질 때 r은 0에 가까울 수 있다.

8.2 단순회귀분석

회귀분석(regression analysis): 변수 간의 함수적 관련성을 규명하기 위하여 어떤 수학적 모형을 가정하고, 이 모형을 측정된 변수의 데이터로부터 추정하는 통계적 분석방법

회귀식(regression equation): 변수 간의 관계를 나타내는 수학적 모형

종속변수(dependent variable): 서로 관계를 가지고 있는 변수 중에서 다른 변수에 의해 영향을 받는 변수. 반응변수(response variable)라고도 함

독립변수(independent variable): 종속변수에 영향을 주는 변수. 설명변수(explanatory variable)라고도 함

1) 단순선형회귀모형

단순선형회귀분석(simple linear regression analysis): 한 개의 독립변수로 이루어진 회귀식

$$Y = \alpha + \beta X$$

α, β: 회귀계수(regression coefficient)

모집단 회귀모형

$$y_i = \alpha + \beta x_i + \epsilon_{i,} \quad i = 1, 2, \cdots, n$$

ϵ_i는 평균이 0, 분산이 σ^2인 서로 독립인 오차를 나타내는 확률변수

적합된 회귀식(fitted regression equation), 적합 회귀 방정식

$$\hat{y}_i = a + bx_i$$

$\hat{y}$는 적합 회귀 방정식에 의해 $X = x_i$에서 예측된 Y의 값

a, b는 표본을 이용하여 추정된 회귀계수

잔차(residual): 예측된 Y값과 실제 관측된 값의 차이

$$e_i = y_i - \hat{y}_{i,} \quad i = 1, 2, \cdots, n$$

2) 회귀계수의 추정

❖ 최소제곱법(method of least squares)

적합된 회귀식에서 계산된 예측값 $\hat{y}_i$와 관찰값 y_i의 차이인 잔차들의 제곱의 합이 최소가 되도록 회귀계수를 추정하는 방법으로 다음의 식을 최소화하는 a, b의 값을

구한다.

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \bar{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

정규방정식: 잔차의 제곱합을 a와 b에 대해 각각 편미분하여 0으로 놓고 a와 b에 대해 푼 것

$$a \cdot n + b\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i$$

❖ 최소제곱추정량(least squares estimator)

$$b = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$a = \bar{y} - b\bar{x}$$

b의 분모와 분자를 n-1로 나누면

$$b = \frac{S_{XY}}{S_{XX}} = r \cdot \frac{\sqrt{S_{XX}}\sqrt{S_{YY}}}{S_{XX}} = r \cdot \frac{\sqrt{S_{YY}}}{\sqrt{S_{XX}}}$$

3) 회귀직선의 적합도

회귀식의 타당성 조사에는 추정의 표준호차(standard error of estimate)와 결정계수(coefficient of determination)가 사용된다.

추정의 표준오차 s: 관측값들이 추정회귀직선의 주위에 흩어져 있는 정도를 나타내는 측도

$$s^2 = \frac{1}{n-2}\sum_{i=1}^{n} (y_i - \hat{y})^2$$

❖ 제곱합과 자유도의 분할

제곱합: $\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y})^2$

SST = SSR + SSE

자유도: n-1 = 1 + n-2

결정계수(R^2; coefficient of determination): 총변동 SST에서 설명된 변동 SSR이 차지하는 비

$$R^2 = \frac{SSR}{SST}$$

4) 회귀의 분산분석

회귀의 분산분석: 제곱합의 분할을 이용하여 회귀분석과 관련된 문제를 다루는 것

❖ 단순선형회귀의 분산분석표

요인	제곱합	자유도	평균제곱	F비
회귀 오차	SSR SSE	1 n-2	$MSR = \frac{SSR}{1}$ $MSE = \frac{SSE}{n-2}$	$F_0 = \frac{MSR}{MSE}$
전체	SST	n-1		

❖ 단순선형회귀분석에서의 검정

가설: $H_0 : \beta = 0, \quad H_1 : \beta \neq 0$

검정: $F_0 = \frac{MSR}{MSE} > F_{1,n-2,\alpha}$ 이면 H_0를 기각

5) 회귀분석에서의 추론

(1) β에 관한 추론

❖ β에 대한 가설검정

H_0: $\beta = \beta_0$

검정통계량: $t = \frac{b-\beta_0}{SE(b)}, \quad SE(b) = \frac{s}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

H_0 기각역 → H_1: $\beta < \beta_0$ 이면 $t < -t_{n-2,\alpha}$

H_1: $\beta > \beta_0$ 이면 $t > t_{n-2,\alpha}$

H_1: $\beta \neq \beta_0$ 이면 $|t| > t_{n-2,\alpha/2}$

(2) 모수 α에 관한 추론

❖ α에 대한 가설검정

H_0: $\alpha = \alpha_0$

검정통계량: $t = \frac{\alpha - \alpha_0}{SE(\alpha)}$, $SE(\alpha) = S \cdot \sqrt{\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$

H_0 기각역 → H_1: $\alpha < \alpha_0$ 이면 $t < -t_{n-2,\alpha}$

H_1: $\alpha > \alpha_0$ 이면 $t > t_{n-2,\alpha}$

H_1: $\alpha \neq \alpha_0$ 이면 $|t| > t_{n-2,\alpha/2}$

(3) Y의 평균값에 관한 추론

X의 임의의 점 $X = x_0$ 에서 종속변수 Y는 평균값 $\mu_{Y|X} = \alpha + \beta x_0$ 를 가진다.

이 값을 추정한다는 것은 Y의 평균값을 예측하는 것과 같은 의미이므로 $\mu_{Y|X}$ 역시 중요한 모수로 취급된다. $\mu_{Y|X}$ 의 점추정량은 다음과 같다.

$$\hat{y}_0 = a + bx_0$$

$\mu_{Y|X}$ 의 100(1-α)% 신뢰구간은

$$\left[\hat{y}_0 - t_{n-2,\alpha/2} \cdot SE(y_0),\ \hat{y}_0 + t_{n-2,\alpha/2} \cdot SE(y_0)\right]$$

$$SE(y_0) = s \cdot \sqrt{\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

6) 잔차분석

각 모수에 대한 추론은 모두 모집단 회귀모형에 포함된 오차항 ϵ_i에 대한 몇가지 가정을 바탕으로 하고 있다. 그러므로 추론이 타당하려면 이 가정들이 성립되어야 한다.

회귀분석에서의 가정

A1: 가정된 모형 $y_i = \alpha + \beta x_i + \epsilon_i$는 옳다.

A2: 오차 ε_i의 평균값은 0이다.

A3: (등분산성) 모든 ε_i의 분산은 σ^2으로 동일하다.

A4: (독립성) 오차 ε_i들은 서로 독립이다.

A5: (정규성) 오차 ε_i들은 정규분포를 따른다.

이 가정의 타당성은 일반적으로 잔차의 산점도를 이용해 조사되는데, 각각의 가정을 위해 주로 사용되는 산점도는 다음과 같다.

① 잔차 대 예측값(즉, e_i 대 $\hat{y}_i$): A3

② 잔차 대 독립변수(즉, e_i 대 x_i): A1

③ 잔차 대 관측순서(즉, e_i 대 i): A2, A4
→ 산점도들에서는 잔차들이 0을 중심으로 특정한 경향을 보이지 않고 랜덤하게 나타나면 각 가정이 타당함을 의미

오차항 ε_i가 정규분포를 따른다는 가정은 잔차들의 히스토그램이 정규분포의 모양과 비슷한지를 보아 타당성을 조사한다. 또 다른 방법은 정규확률도(normal probability plot)가 직선을 형성하면 정규분포를 따른다고 볼 수 있다.

8.3 중선형회귀분석

중선형회귀분석(multiple linear regression analysis): 하나의 종속변수와 여러 개의 독립변수 사이의 관계를 규명하고자 할 때 사용되는 통계적 방법

1) 중선형회귀모형

종속변수 Y와 k개의 독립변수 X_1, X_2, …, X_k가 있고, 관측값들이 (y_1, x_{1i}, x_{2i}, …, x_{ki})일 때 중선형회귀모형은 다음과 같다.

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki} + \epsilon_i$$

ε_i: 오차항

β_0: Y축의 절편

β_i: 다른 독립변수가 고정되었을 때 X_i가 Y에 미치는 영향

행렬과 벡터를 이용하여 나타내면

$$Y = X\beta + \varepsilon$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad X = \begin{bmatrix} 1 & X_{11} & X_{12} \cdots & X_{1k} \\ 1 & X_{21} & X_{22} \cdots & X_{2k} \\ \vdots & \vdots & \vdots \cdots & \vdots \\ 1 & X_{n1} & X_{n2} \cdots & X_{nk} \end{bmatrix}, \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}, \quad \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

2) 중선형회귀모형의 추정

가장 간단한 경우로 하나의 종속변수(Y)에 두 개의 설명변수(X1, X2)가 있는 중선형회귀모형에서

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_{i,} \quad i = 1, 2, \cdots, n$$

최소제곱법으로 회귀계수 β_0, β_1, β_2의 추정값 b_0, b_1, b_2를 구하려면 오차의 제곱합을 최소화하는 b_0, b_1, b_2를 구한다.

$$\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{1i} - \beta_2 x_{2i})^2$$

오차의 제곱합을 최소화하려면 위 식을 β_0, β_1, β_2에 대하여 편미분한 식을 0으로

두고 방정식을 풀면 된다.

$$nb_0 + b_1\sum x_{1i} + b_2\sum x_{2i} = \sum y_i$$
$$b_0\sum x_{1i} + b_1\sum x_{1i}^2 + b_2\sum x_{1i}x_{2i} = \sum x_{1i}y_i$$
$$b_0\sum x_{2i} + b_2\sum x_{2i}^2 + b_1\sum x_{1i}x_{2i} = \sum x_{2i}y_i$$

위 연립방정식을 만족시키는 b_0, b_1, b_2가 오차의 제곱합을 최소화시키는 β_0, β_1, β_2의 추정값이 된다.

한편, 오차의 분산 σ^2의 추정값은 오차제곱합(SSE)을 잔차의 자유도로 나누어 구한다. 두 개의 설명변수가 있는 중선형회귀분석의 경우 모형에서 미용되는 미지의 모수가 3개이므로 잔차의 자유도는 n-3이 되며, 오차분산(σ^2)의 추정값(s^2)은 다음과 같다.

$$s^2 = \frac{1}{n-3}\sum_{i=1}^{n}(y_i - b_0 - b_1x_{1i} - b_2x_{2i})^2$$

중선형회귀분석의 추정계수는 종속변수와 하나의 설명변수 간 상관관계를 나타내기보다는 다른 설명변수가 고정되었을 경우 종속변수와 하나의 설명변수 간 상관관계를 의미한다.

따라서 중선형회귀분석의 회귀계수를 편회귀계수(partial regression coefficient)라고 부른다.

3) 회귀직선의 적합도와 분산분석

추정된 회귀직선의 타당성을 조사하기 위해 추정의 표준오차와 결정계수가 사용된다. 중선형회귀분석에서 추정의 표준오차는 다음과 같다.

$$s = \sqrt{\frac{1}{n-k-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

중선형회귀분석과 단순선형회귀와의 차이는 잔차 $\epsilon_i = y_i - \hat{y}_i$를 계산하려면 (k+1)개의 회귀계수가 추정되어야 하므로 자유도가 n-2가 아니라 n-k-1이다. 그 외에는 단순선형회귀와 마찬가지로 잔차평균제곱을 사용한다.

❖ 제곱합과 자유도의 분할

제곱합: $\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2$

SST = SSR + SSE

자유도: n-1 = k + n-k-1

결정계수는 단순선형회귀와 같다.

결정계수(R^2; coefficient of determination): 총변동 SST에서 설명된 변동 SSR이 차지하는 비

$$R^2 = \frac{SSR}{SST}$$

그런데, 중선형회귀분석에서 설명변수의 수가 증가하면 R^2은 1로 수렴한다. (온갖 변수를 설명변수로 추가하면, 전체제곱합에서 오차로 설명할 수 있는 부분이 남지 않게 된다.) 따라서 중선형회귀모형의 타당성을 R^2만으로 검토하는데에는 한계가 있다.

따라서 설명변수의 수가 불필요하게 증가하면 이것을 조정할 수 있는 수정된 $\overline{R}^2$을 이용한다.

$$\overline{R}^2 = 1 - \frac{\hat{\sigma}^2}{VAR(Y)} = 1 - (1 - R^2)\frac{N-1}{N-k}$$

그다지 유의하지 않은 설명변수가 회귀모형에 추가되는 경우 R^2은 무조건 증가하나, $\overline{R}^2$은 변수를 추가하지 않았을 경우에 비해 감소할 수 있다.

❖ 중선형회귀의 분산분석표

요인	제곱합	자유도	평균제곱	F비
회귀 오차	SSR SSE	k n-k-1	$MSR = \frac{SSR}{k}$ $MSE = \frac{SSE}{n-k-1}$	$F_0 = \frac{MSR}{MSE}$
전체	SST	n-1		

회귀식의 유의성 α에 대한 검정

H_0: $\beta_1 = \beta_2 = \cdots = \beta_k = 0$

H1: k개의 β_i 중 적어도 하나는 0이 아니다.

$F_0 > F_{k, n-k-1, \alpha}$ 이면 H_0를 유의수준 α하에서 기각할 수 있다.

4) 중선형회귀분석에서의 추론

오차항 ε_i들이 독립이고, 모두 $N(0, \sigma^2)$의 분포를 가진다는 가정하에서

$$b_i \sim N(\beta_i, c_{ii} \cdot \sigma^2), \quad i = 0, 1, \cdots, k$$

c_{ii}는 (k+1)×(k+1) 행렬인 $(X'X)^{-1}$의 i번째 대각원소

모수 σ^2 대신에 추정량 s^2를 사용하면 다음과 같이 t분포를 이용하여 각 회귀계수에 대한 추론을 할 수 있다.

❖ 중선형회귀모형에서 회귀계수 β_i에 관한 추론

점추정량: b_i

표준오차: $SE(b_i) = \sqrt{c_{ii}} \cdot s$

신뢰구간: $[b_i - t_{n-k-1,\alpha/2} \cdot SE(b_i),\ b_i + t_{n-k-1,\alpha/2} \cdot SE(b_i)]$

가설검정

H_0: $\beta_i = \beta_{i0}$

검정통계량: $t = \frac{b_i - \beta_{i0}}{SE(b_i)}$

H_0 기각역: H_1: $\beta_i < \beta_{i0}$ 이면 $t < -t_{n-k-1,\alpha}$

H_1: $\beta_i > \beta_{i0}$ 이면 $t > t_{n-k-1,\alpha}$

H_1: $\beta_i \neq \beta_{i0}$ 이면 $|t| > t_{n-k-1,\alpha/2}$