7. 선형회귀모형에 의한 검정

7-1. 회귀분석(Regression Analysis)

① 한 개 또는 그 이상의 변수들(독립변수)에 대하여 다른 한 변수(종속변수) 사이의 관계를 수학적인 모형을 이용하여 설명하고 예측하는 분석기법이다.

② 먼저 변수들 간의 관계를 나타내는 타당한 회귀방정식 또는 회귀모형을 이론적 근거나 경험에 바탕하여 설정하고 변수들의 관측된 값을 이용하여 그 모형을 추정한 다음, 추정한 모형에 의해서 변수들 간의 관계를 설명하거나 예측한다.

③ 회귀란 옛날 상태로 돌아가는 것을 의미한다.

④ 영국의 유전학자 프랜시스 골턴은 부모의 키와 아이들의 키 사이의 연관관계를 연구하면서 부모와 자녀의 키 사이에는 선형적인 관계가 있고 키가 커지거나 작아지는 것보다는 전체 키 평균으로 돌아가려는 경향이 있다는 가설을 세웠으며 이를 분석하는 방법을 "회귀분석"이라고 하였다.

⑤ 이후, 칼 피어슨은 아버지와 아들의 키를 조사한 결과를 바탕으로 함수 관계를 도출하여 회귀분석 이론을 수학적으로 정립하였다.

⑥ 회귀분석은 다음의 가정을 바탕으로 한다.

㉠ 오차항은 모든 독립변수 값에 대하여 동일한 분산을 갖는다.

㉡ 오차항의 평균(기대값)은 0이다.

㉢ 수집된 데이터의 확률분포는 정규분포를 이루고 있다.

㉣ 독립변수 상호간에는 상관관계가 없어야 한다. 만약 독립변수들 간에 상관관계가 나타나는 경우 다중공선성문제(Multicollinearity)[1]라고 한다.

㉤ 시간에 따라 수집한 데이터들은 잡음의 영향을 받지 않아야 한다.

7-2. 회귀계수 추정

7-2-1. 단순선형회귀

① 두 변수의 상관관계가 1.0일 때, 직선을 그리고 직선에 의한 공식에 의하여 두 변수의 값들을 정확히 예측할 수 있다.

㉠ 직선 그리기 가능

㉡ 직선식 $y = a + bx$로 표현 가능

㉢ 직선식에 의해 두 변수의 값들을 정확히 예측

② 두 변수의 상관관계가 1.0이 아닐 때, 점들을 연결하면 직선을 그릴 수는 없으나 산점도에 나타난 모든 점들을 대표하는 직선을 그릴 수 있다.

㉠ 이 직선을 회귀선(regression line)이라 한다.

㉡ 두 변수를 대표하는 회귀선은 X의 평균 $\overline{X}$와 Y의 평균 $\overline{Y}$인 점을 지나고 많은 점들을 적절히

---

1) 독립변수들간에 정확한 선형관계가 존재하는 완전공선성의 경우와 독립변수들간에 높은 선형관계가 존재하는 다중공선성으로 구분한다.

- 결정계수 R^2값이 높아 회귀식의 설명력은 높지만 식에 포함된 독립변수의 $p-$값이 커서 개별인자들이 유의하지 않는 경우가 있다. 이런 경우 독립변수들 간에 높은 상관관계가 있다고 의심된다. 또는 분산팽창요인(variance inflation factor)를 구하여 이 값이 10을 넘는다면 보통 다중공선성의 문제가 있다. 이런 경우에 상관관계가 높은 독립변수 중 하나 혹은 일부를 제거하거나 변수를 변형시키거나 새로운 관측치를 이용한다.

대변한다.

㉢ 산점도에 있는 모든 점들이 자료를 대표하는 직선 쪽으로 회귀한다하여 회귀선이라 명명한다.

〖예〗 수면제의 투여량에 따라 수면시간의 증가(효과)가 어떻게 변하는가를 조사하기 위하여 임의로 추출된 9명에게 실험을 하여 다음의 자료를 얻었다.

사람	1	2	3	4	5	6	7	8	9
수면제의 양(X_i)	1.5	1.8	2.4	3.0	3.5	3.9	4.4	4.8	5.0
수면시간의 증가(Y_i)	4.8	5.7	7.0	8.3	10.9	12.4	13.1	13.6	15.3

산점도, X가 증가하면 Y도 증가하는 경향이 있고 대략 직선으로 나타남을 짐작할 수 있다.

③ 회귀방정식의 종류

㉠ 방정식에 포함된 독립변수의 개수에 따라 독립변수의 개수가 1개인 방정식을 단순회귀방정식(simple regression equation) 또는 단순회귀모형(simple regression model)이라 하고, 독립변수의 개수가 2개 이상일 때의 모형을 다중회귀모형(multiple regression model)이라고 한다.

㉡ 회귀방정식의 형태에 따라 모수의 선형함수로 주어진 모형을 선형회귀모형(linear regression)이라 하고, 모수의 비선형함수로 주어지는 모형을 비선형회귀모형(nonlinear regression)이라고 한다.

④ 단순선형회귀

㉠ 1개의 종속변수(=반응변수) Y와 독립변수(=설명변수) X 간의 선형관계이다.

㉡ 독립변수 X의 값에서 측정된 종속변수 Y의 값을 이용하여 모집단 회귀직선 $Y=\beta_0+\beta_1 X$를 추정한다.

㉢ 모집단 회귀직선 $Y=\beta_0+\beta_1 X$에서 회귀계수 β_1, β_0의 최소제곱추정량 a, b에 대한 추정된 회귀직선의 방정식(regression equation)은 $\hat{Y}=a+bX$이다.

- a = 절편, X변수가 0일 때의 Y값
- b = 기울기, X변수가 변할 때 Y변수의 변하는 정도

7-2-2. 회귀계수의 추정

① 모집단 회귀직선을 추정하는 방법으로 가장 널리 이용되는 방법은 오차의 제곱합을 최소화하도록 추정하는 최소제곱법(method of least squares)이다.

② 회귀계수 β_1의 최소제곱추정량 b (기울기) 구하기

㉠ X변수의 변화량에 따른 Y변수의 변화량 $\frac{\Delta y}{\Delta x}$을 이용

㉡ 그러나 X와 Y의 단위가 다르면 문제가 생김

㉢ 따라서 X변수의 표준편차 s_X에 따른 Y변수의 표준편차 s_Y인 $\frac{s_Y}{s_X}$을 이용

㉣ 두 변수의 상관계수가 1.0인 완벽한 상관관계인 경우 회귀선의 기울기 : $\frac{s_Y}{s_X}$

㉤ 두 변수의 관계가 완벽하지 않아 다양하게 흩어져 있는 경우(흩어진 정도인 상관계수를 고려)

회귀선의 기울기를 구하는 공식 $b = r_{XY}\left(\frac{s_Y}{s_X}\right) = \frac{s_{XY}}{{s_X}^2}$

= 회귀계수(regression coefficient)

③ β_0의 최소제곱추정량인 a (y절편) 구하기

㉠ 회귀선은 $\overline{X}$ (X의 평균)와 $\overline{Y}$ (Y의 평균)을 반드시 지남

㉡ 회귀선의 등식에 평균값을 대입하여 구함 ; $a = \overline{Y} - b\overline{X}$

④ X변수에 대한 Y의 회귀방정식 구하기

㉠ 각 사례의 얻어진 Y값이 아니므로 $\hat{Y}$ (hat, 헷)라 표기

㉡ $\hat{Y}$을 X에 대한 기댓값 혹은 예측값이라 부름 ; $\hat{Y} = a + bX$

⑤ s_X와 s_Y가 둘 다 양수이므로 b와 r_{XY}은 같은 부호를 갖는다.

㉠ r_{XY}이 양수라면 b도 양수이므로 직선은 x에 따라 증가한다.

㉡ r_{XY}이 음수라면 b도 음수이므로 직선은 x에 따라 감소한다.

㉢ r_{XY}이 0에 가깝다면 b도 0에 가깝다.

〖예〗 5명 학생의 국어, 영어점수에 대한 회귀등식?

	A	B	C	D	E
국어	2	3	5	3	7
영어	1	2	4	3	5

[풀이]

1단계	두 변수의 산점도 그리기 & 상관계수의 기본 가정을 충족하는지 확인	
2단계	$\overline{X}$, $\overline{Y}$, s_X, s_Y, r_{XY} 계산	$\overline{X} = 4$, $\overline{Y} = 3$, $s_X = 1.789$, $s_Y = 1.414$, $r_{XY} = 0.949$
3단계	회귀선의 기울기인 회귀계수 계산 $b = r_{XY}\left(\frac{s_Y}{s_X}\right) = \frac{s_{XY}}{{s_X}^2}$	$b = (0.949)\frac{1.414}{1.789} = 0.750$
4단계	회귀선의 y절편 계산 $a = \overline{Y} - b\overline{X}$	$a = 3 - (0.750)(4) = 0$
5단계	회귀등식 $\hat{Y} = a + bX$ 완성	$\hat{Y} = 0 + (0.75)X = 0.75X$
6단계	회귀식 $\hat{Y} = 0.75X$ 적용 ㉠ 국어점수가 7점인 학생 ⇒ 예측되는 영어점수는 5.25점 ⇒ E학생의 영어점수는 5점으로 기대점수보다 낮은 점수를 얻음 ㉡ 국어점수가 3점인 학생 ⇒ 예측되는 영어점수는 2.25점 ⇒ B학생의 영어점수는 3점, D학생은 2점을 얻음 ㉢ B학생과 E학생은 기대점수보다 높은 점수를 얻었고, D학생은 낮은 점수를 얻음	

〖예〗 수면제의 투여량에 따라 수면시간의 증가(효과)가 어떻게 변하는가를 조사하기 위하여 임의로 추출된 9명에게 실험을 하여 다음의 자료를 얻었다.

사람	1	2	3	4	5	6	7	8	9
수면제의 양(X_i)	1.5	1.8	2.4	3.0	3.5	3.9	4.4	4.8	5.0
수면시간의 증가(Y_i)	4.8	5.7	7.0	8.3	10.9	12.4	13.1	13.6	15.3

[풀이]

X에 관한 Y의 회귀방정식은 $\hat{Y} = 0.2568 + 2.9303X$이고 $X = 2.0$이면 $\hat{Y} = 6.1174$이다.

〖예〗 근무경력 년수 x와 시간당 초기임금 y(단위: 천원)에 관한 자료. x에 대한 y의 최적회귀직선을 구하고 산점도 상에 직선을 그리시오.

x	2	3	4	5	6	7
y	6.00	7.50	8.00	12.00	13.00	15.50

[풀이]

㉠ $\overline{X}$, $\overline{Y}$, s_X, s_Y, r_{XY} 계산

$$\overline{X} = 4.5,\ \overline{Y} = 10.333,\ s_X = 1.871,\ s_Y = 3.710,\ r_{XY} = 0.980$$

㉡ 회기계수와 y절편 구하기

$$b = r_{XY}\left(\frac{s_Y}{s_X}\right) = 0.980\left(\frac{3.710}{1.871}\right) = 1.9432389 \approx 1.943$$

$$a = \overline{Y} - b\overline{X} = 10.333 - 1.943(4.5) = 1.590$$

㉢ 최적회귀직선

$$\hat{Y} = 1.590 + 1.943X$$

㉣ 근무 경력이 3년인 사람의 초기임금은 $\hat{Y} = 1.590 + 1.943 \times 3 = 7.419$라 예측

㉤ 그래프 그리기