제 4 강 횡단면 자료 분석(Cross-sectional data analysis)

Part I. 이진 반응 모형(Binary response model))

Part II. 제한 종속변수 모형(limited dependent variable regression model)

Part III. 기타 이슈들

Part I. 이진 반응 모형(Binary response model))

I. 종속변수가 두 가지로 분류되는 질적 변수 – 이진반응모형

A. 종속변수가 이러한 질적 변수일 경우

i. $E\left(y_t \middle| x_{t2},\cdots,x_{tK}\right)$: 종속변수가 양적 변수일 경우처럼 주어진 설명변수의 값에서 종속변수의 기대값 또는 평균값으로 해석되기 보다는 종속변수가 1인 값을 갖는 일이 일어날 확률로 해석되어야 함

a. ⇒ 확률모형(probability model 이라고 함)

b. 예) 성인 남성의 노동시장참여(Labor Force Participation), 선거에서 투표자들이 특정 정당 선택, 주식투자참여, SUV 보유…등 개별 경제주체들이 부딪히는 미시적 선택의 문제에 광범위하게 적용됨

B. 이진반응모형에 대한 세 가지 접근 방법

i. 선형확률모형(Linear Probability Model: LPM)

ii. 프로빗모형(Probit Model)

iii. 로짓모형(Logit Model)

II. 선형확률모형

A. 개별 경제주체들의 선택을 다음과 같은 더미변수로 나타낼 수 있음

i. $y=\begin{cases}1 & \text{자가운전출근}\\ 0 & \text{대중교통출근}\end{cases}$

1. y 는 (이산적) 확률변수이고 다음과 같은 확률밀도 함수를 가짐

a. $f(y)=p^{y}(1-p)^{1-y},\quad y=0,1$, p 는 y 가 1 의 값을 취할 확

i. $E(y)=p$.

ii. 종속변수를 그 확률적 부분과 고정된 부분으로 나눔

1. $y=E(y)+\varepsilon=p+\varepsilon$

iii. 이 확률은 자가운전과 대중교통 출근의 출근시간 차이에 의존한다고 가정

1. x = (대중교통출근시간 – 자가운전출근시간)

2. 그 관계는 선형이라고 가정

a. $E(y)=p=\beta_1+\beta_2 x$

iv. 이렇게 주어지는 다음과 같은 선형회귀모형을 LPM 이라고 함

1. $y_t = E(y_t) + \varepsilon_t = \beta_1 + \beta_2 x_t + \varepsilon_t$

B. LPM 의 문제

i. 이분산성. $V(\varepsilon_t) = V(y_t) = p_i(1-p_i)$

ii. x 의 변화가 일정한 율로 확률 p 에 영향을 미침 ($\frac{dp}{dx} = \beta_2$): 더욱 심각

1. $\hat{p} = b_1 + b_2 x$: $\hat{p}$ 은 0 과 1 사이를 벗어난 값을 가지기 쉬움

III. 프로빗 모형(The Probit Model)

A. 프로빗 모형

i. 선택 확률 p 가 [0,1]의 구간에 놓이도록 하기 위해서는 설명변수와 p 사이에 선형이 아닌 비선형의 관계가 요구됨

ii. 이러한 관계에 F: R→[0,1]의 함수를 사용할 수 있으며, 누적확률분포함수가 이러한 성질을 갖는 함수임

iii. 이러한 비선형 관계를 특히 표준정규분포의 누적확률분포함수를 이용하여 표현하는 경우 이를 프로빗 모형이라 하며, 이 때 사용되는 함수를 특별히 프로빗 함수라고 부르기도 함

1. 프로빗 함수: $F(z) = P[Z \le z] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-.5u^2} du$

2. 프로빗 모형: $p = P[Z \le \beta_1 + \beta_2 x] = F(\beta_1 + \beta_2 x)$

B. 프로빗 모형의 추정 – 최우추정법

i. 예컨대, 세 명에 대해 $y_1 = 1$, $y_2 = 1$ and $y_3 = 0$ 이 관측되고 이들의 설명변수 값이 $x_1 = 15$, $x_2 = 20$ and $x_3 = 5$ 관측됨

ii. $y_1 = 1$, $y_2 = 1$ and $y_3 = 0$ 를 관측하게 될 확률은?

1. y 의 확률함수는 베르누이 분포로부터 주어지고 이를 프로빗 모형과 결합하면

a. $f(y_i) = [F(\beta_1 + \beta_2 x_i)]^{y_i} [1 - F(\beta_1 + \beta_2 x_i)]^{1-y_i}, \quad y_i = 0,1$

b. 세 명이 무작위로 추출될 경우 y_1, y_2 및 y_3 에 대한결합확률분포함수는 $f(y_1, y_2, y_3) = f(y_1) f(y_2) f(y_3)$ 로 주어짐

i. $y_1 = 1$, $y_2 = 1$ 및 $y_3 = 0$ 이 관찰될 확률은

$$P[y_1 = 1, y_2 = 1, y_3 = 0] = f(1,1,0) = f(1) f(1) f(0)$$

$$= F[\beta_1 + \beta_2(15)] \cdot F[\beta_1 + \beta_2(20)] \cdot \{1 - F[\beta_1 + \beta_2(5)]\}$$

: 우도함수(likelihood function)

iii. 최우추정법: 주어진 표본의 값이 관측될 확률 또는 우도를 극대화하는 값 b_1 과 b_2 를 β_1 와 β_2 의 추정치로 구하는 것

1. 최우추정량의 소표본 성질은 대개의 경우 알려져 있지 않으나, 대표본에서 일치추정량이고 유효추정량이며 정규분포를 한다는 것이 알려짐
2. 대부분 통계패키지들은 프로빗 모형에 대한 최우추정치를 구하는 명령어를 포함하고 있음

C. 프로빗 모형의 해석

i. x 한 단위의 변화가 $y=1$ 일 확률에 미치는 영향은?

1. $\frac{dp}{dx}=\frac{dF(t)}{dt}\cdot\frac{dt}{dx}=f(\beta_1+\beta_2x)\beta_2$, $t=\beta_1+\beta_2x$, $f(\cdot)$: 표준정규분포의 확률밀도함수.
   a. dp/dx 의 부호는 β_2 의 부호에 의해 결정됨
   b. $\beta_1+\beta_2x$ 가 0 근처일 때 $f(\beta_1+\beta_2x)$ 의 값도 최대화 되며 따라서 x 의 변화에 따른 확률의 변화도 가장 커짐
   c. 반면에 $\beta_1+\beta_2x$ 의 절대값이 이 매우 큰 경우, 예컨대 3 정도 되는 값일 경우 $f(\beta_1+\beta_2x)$ 는 거의 0 에 가까워지면 따라서 x 의 변화는 확률에 거의 영향을 미치지 못함

ii. 어떤 개인이 y=1 을 선택할 확률을 예측

1. $p=F(\beta_1+\beta_2x)\ \Rightarrow\ \hat{p}=F(b_1+b_2x)$
2. 어떤 기준값, 예컨대 0.5 를 사용하여 이 개인의 선택을 예측
   a. $\hat{y}=\begin{cases}1 & \hat{p}>0.5\\ 0 & \hat{p}\le 0.5\end{cases}$

D. 실례

i.
$$b_1+b_2x_i=-0.0644+0.0299x_i$$
$$z\text{-values}\quad(-.161)\quad(2.916)$$

1. b_1 : 통근시간의 차이가 0 인 개인들은 대중교통출근을 하는 것으로 나타나고 있으나 그 통계적 유의성은 없음
2. b_2 : 자가운전출근시간에 비해 대중교통출근시간이 길수록 개인들이 통근시 자가 운전을 선택할 확률이 커짐
3. 현재 대중교통출근시간이 자가운전출근시간에 비해 20 분 더 길 경우 대중교통시간의 증가의 영향의 크기?
   a. $$\frac{d\hat{p}}{dx}=f(b_1+b_2x)b_2=f(-0.0644+0.0299\times 20)(0.0299)$$
   $$=f(.5355)(0.0299)=0.3456\times 0.0299=0.0104$$
   b. 1 분의 대중교통출근시간의 증가는 자가운전선택 확률을 약 1% 증가시킴
4. 대중교통출근시간이 자가운전출근시간에 비해 30 분 긴 어느 개인이 자가운전으로 출근하게 될 확률?

a. $\hat{p} = F(b_1 + b_2 x) = F(-0.0644 + 0.0299 \times 30) = .798$

b. 기준값 0.5 보다 큼으로 이 개인은 자가 운전을 선택할 것으로 예측됨

IV. 로짓 모형(The Logit Model)

A. 프로빗 모형의 대안으로 흔히 사용됨

i. 유일한 차이점은 사용되는 확률분포함수

ii. 로짓모형에서는 로지스틱(logistic) 분포의 확률분포함수가 사용됨

1. L이 로지스틱 확률변수라 하면 그 확률밀도함수는

$$f(l) = \frac{e^{-l}}{\left(1+e^{-l}\right)^2}, \quad -\infty < l < \infty \ .$$

2. 대응되는 누적분포함수는 정규분포와는 달리 닫힌 형태로 표현되며 이 점이 분석을 다소 용이하게 함

$$F(l) = p[L \le l] = \frac{1}{1+e^{-l}}$$

iii. 로짓 모형에 있어서, y=1 일 확률 p 는

$$p = P\left[L \le \beta_1 + \beta_2 x\right] = F\left(\beta_1 + \beta_2 x\right) = \frac{1}{1+e^{-(\beta_1+\beta_2 x)}}$$

이를 이용하여 프로빗 모형에서와 마찬가지 방법으로 최우추정량을 구할 수 있으며, 그 결과에 대한 해석 역시 정규분포 확률밀도함수를 로지스틱 확률밀도 함수로 대체하면 마찬가지임

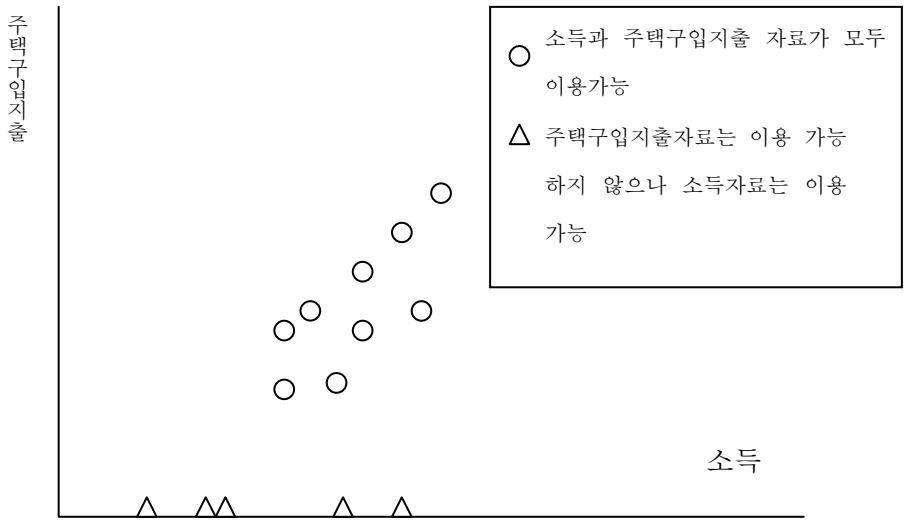
Part II. 제한 종속변수 모형(limited dependent variable regression model)

I. 토빗모형(The Tobit Model)

A. 토빗모형

i. 종속변수의 관측값은 연속이나 일부에 대해서만 이용 가능한 경우 (censored sample) cf. truncated sample

1. 주택구입에 대한 지출액? : 주택을 구입하지 않은 소비자들에 대해서는 관측치가 존재하지 않는다
   a. 이 경우 종속변수에 대한 정보가 있는 소비자 그룹(주택을 구입한 그룹과) 그렇지 않은 그룹으로 구분되며, 이 두 그룹들 모두 설명변수에 대한 정보(가구구성원의 수, 소득, 이자율 등)는 이용가능
   b. 이 경우 종속변수의 관측치가 이용 가능하지 않은 그룹을 0 원 지출했다고 놓고 추정을 하던, 아니면 무시하고 관측치가 이용 가능한 그룹에 대해서만 추정을 하던 통상적 최소제곱추정은 잘못된 결과를 낳게 됨
2. 어떤 가게의 아이스크림 수요? : 아이스크림이 매진된 날의 수요는 아이스크림 박스의 수용용량으로 나타남
   a. 이 경우 역시 아이스크림이 매진될 날의 수요에 대한 관측치는 이용 가능하지 않음
3. 기혼여성의 임금 수준을 결정하는 요인? : 노동시장에 참여한 여성만 이용가능함 (즉 받는 임금수준이 reservation wage 보다 높은 경우만 관측됨)



ii. 모형 및 추정

1. $y_i = \beta_1 + \beta_2 x_i + \varepsilon_i$ if $\beta_1 + \beta_2 x_i + \varepsilon_i > 0$

= 0 otherwise

간단히 $y_i = \max\left(0, \beta_1 + \beta_2 x_i + \varepsilon_i\right)$ 로 쓸 수 있다.

a. 이 때, 종속변수에 대한 관측이 가능하지 않은 그룹(n2)은 무시하고 관측이 가능한 그룹의 관측치(n1)만을 가지고 추정을 하는 경우
   i. 편향(biased)될 뿐 아니라 비일치추정이 됨 (아이스크림 수요?)
   ii. 오직 n1 그룹만을 가지고 회귀선을 긋는 것은 , n1 과 n2 그룹을 모두 고려하는 경우와 다르게 됨
      1. 토빗모형에 대한 추정은 최우추정법에 기반을 두며 종속변수의 값을 관찰할 수 없는 관측치를 다른 관측치들과 달리 취급하여 접근함

b. 원래 의미의 토빗모형(Type I 토빗모형)은 censoring 에 영향을 미치는 요인과 종속변수에 영향을 미치는 요인이 동일하고 같은 방향으로 영향을 미칠 경우 사용해야 함 - 즉 토빗모형은 예컨대 앞서 주택지출의 예에서 설명변수로 소득을 포함시킨다는 것은 소득이 주택구입에 대한 결정과 주택에 대한 지출의 크기에 같은 방향으로 영향을 미치는 것을 전제로 하는 것임
   i. 여름 휴가비와 가구구성원의 수의 관계
      1. 문제는 여름 휴가비는 여름 휴가를 가기로 결정한 가구에 대해서만 관찰이 가능하며 가구구성원의 수는 일단 휴가를 가기로 결정한 경우 여름 휴가비의 크기에 양의 영향을 미치지만 여름 휴가를 가기로 결정하는데는 부의 영향을 미침
   ii. 담배광고와 담배흡연량의 관계
      1. 담배광고가 담배흡연 경험에 미치는 영향은 크나 담배흡연량에는 영향을 미치지 못함
   iii. 아이스크림 수요와 날씨의 관계

II. 표본선택(sample selection)의 문제 - Type II 토빗모형, 헥킷모형

A. 자료의 무작위 추출
   i. 자료가 무작위로 추출(random sampling)되었을 경우 통상적 최소제곱추정은 유효함
   ii. 표본의 추출과정이 무작위가 아닌 경우 통상적 추정방법은 문제를 낳음 : 자료선택편차(sample selection bias)
      1. 기혼여성의 임금 결정 요인의 연구

a. 기혼여성의 임금 자료의 수집 과정: 오직 노동시장에 참여하기로 결정한 기혼여성의 시장 임금 자료를 관측할 뿐임

b. 이 때 관측되는 일하는 주부들만의 자료에 대해 최소제곱 추정을 할 경우, 이는 잘못된 결과를 낳음

c. 이는 우리가 관측하는 자료가 무작위 표본이 아니고, 우리가 고려하지 않은 체계적인 과정(즉 기혼 주부의 노동시장 참여 결정)에 의해 선택된 자료이기 때문임

B. 헥킷모형(Heckit Model)- 자료선택편차의 교정

i. 두 번의 추정 과정을 포함함

1. 첫째 단계에서는 위 예에서 먼저 기혼여성의 노동시장 참여 선택 결정을 설명하는 프로빗 모형을 먼저 추정함

2. 둘째 단계에서는 첫번째 단계의 추정 결과로부터 자료선택편차에 대한 정보를 취득하고 이를 교정한 상태에서 최소제곱추정을 적용함

a. 이는 둘째 단계의 회귀식의 설명변수에 첫 번째 단계의 추정에서 얻어지는 " Inverse Mills Ratio"를 설명변수로 포함시킴으로써 이루어진다.

i. $y_{1i} = 1\left[\alpha_1 + \alpha_2 x_{1i} + \nu_i > 0\right]$: 1 단계 선택방정식

ii. $y_{2i} = x_{2i}\gamma + u_i$: 2 단계 주방정식

iii. $E\left(y_{2i} \mid x_{2i}, y_{1i} = 1\right) = x_{2i}\gamma + \delta\lambda\left(\hat{\alpha}_1 + \hat{\alpha}_2 x_{1i}\right)$

$$\lambda\left(\hat{\alpha}_1 + \hat{\alpha}_2 x\right) = \frac{f\left(\hat{\alpha}_1 + \hat{\alpha}_2 x\right)}{1 - F\left(\hat{\alpha}_1 + \hat{\alpha}_2 x\right)} \text{ : Inverse Mills Ratio}$$

b. 이 때 둘째 단계에 있어서 포함되는 설명변수는 첫 번째 단계에 있어서 포함되는 설명변수와 다를 수 있음

3. 최우추정법의 적용이 힘들 경우 그 대안으로 사용 - 일치 추정이나 유효추정은 아님

Part III. 기타 이슈들

A. 순차선택모형(Ordered Choice Models)

i. 종종 선택의 대안들이 순서가 매겨져 있는 경우가 있으며 이 경우 이를 고려해야 함

1. 서베이의 문항이 소득 수준을 물었을 때 '매우 나쁨' '나쁨' '보통' '좋음' '매우 좋음' 과 같은 식으로(5 점 척도, 7 점 척도) 주어졌을 때,
2. 교사가 학생의 봉사활동 점수를 매기는 데에 있어서 A, B, C, D, F 를 선택해서 줄 때
3. S&P 에서 채권의 등급을 매길 때
4. 사실 이들은 어떤 연속적인 값을 갖는 변수들로부터 결과한 것으로 볼 수 있음 (소득수준, 봉사활동의 수준, 기업의 신뢰성정도 등등)

ii. 이러한 경우, 선택의 대안들에 대해 순위를 1 부터 쭉 매길 수 있으며, 이 경우 1, 2,3,..은 기수적 의미가 아닌 서수적 의미임

1. 서베이 답을 숫자로 바꿀 때(coding) 1,2,3,4,5 로 매기나 여기서는 2 가 1 의 두배라는 의미가 아님
   a. 따라서 여기에 그냥 OLS 를 적용하는 것은 부적절함

iii. 종속변수가 이러한 형태의 서수적 의미의 숫자로 주어질 때 ordered probit 이나 ordered logit 모형을 사용하게 되며 이는 최우 추정법을 통해 추정함

a. $y_i^* = \beta_1 + \beta_2 x_i + \varepsilon_i$: y_i^* 관측가능하지 않은 기업의 신뢰도

i. $y_i^* \le \delta_1$, $\delta_1 < y_i^* \le \delta_2$, $\delta_2 < y_i^* \le \delta_3$, $y_i^* > \delta_3$ $\Rightarrow$ D,C,B,A

ii. 관찰된 기업의 채권 등급이 B 라 하면, 이 확률은 다음과 같이 주어짐

$$prob\left(\delta_2 \le y_i^* = \beta_1 + \beta_2 x_i + \varepsilon_i \le \delta_3\right)$$
$$= prob\left(\delta_2 - \beta_1 - \beta_2 x_i \le \varepsilon_i \le \delta_3 - \beta_1 - \beta_2 x_i\right)$$

iii. 이 때, ε의 확률분포를 안다면 우도함수를 구축할 수 있으며, 최우추정법을 적용할 수 있다.

1. 정규분포를 가정할 경우 ordered probit, 로지스틱 분포를 가정할 경우 ordered logit 모형이 된다.

iv. 사실 이진선택모형도 이러한 식으로 접근하는 것이 가능함

1. McFadden 의 합리적 선택 이론에 기반을 둔 프로빗 모형
   A. 어떤 가구 i 의 주택 구입은 관측가능하지 않은 효용

지수(utility index) I_i에 의존하며, 이는 관측 가능한 설명변수 예컨대 소득에 의해 결정됨

B. $I_i = \beta_1 + \beta_2 x_i$.

C. 주택에 대한 구입(Y=1) 결정은 이러한 효용지수가 어떤 수준(문턱수준-threshold level) I_i^*을 넘어서게 되면 이루어지게 된다고 가정.

D. 이러한 문턱수준 역시 관찰 가능하지 않으나 그것이 표준 정규분포를 한다는 가정을 할 수 있음 (정규분포를 가정하면 효용지수의 rescaling 을 통해 항상 정규화가 가능함)

E. 이 경우 가구 i 가 주택을 구입할 확률은 표준정규분포의 CDF *F()*에 의해 다음과 같이 주어진다.

i. $$p_i = P(y=1|x) = P(I_i^* \le I_i)$$
$$= P(Z_i \le \beta_1 + \beta_2 x_i) = F(\beta_1 + \beta_2 x_i)$$

ii. 이는 다름아닌 probit 모형이며 최우추정법에 의해 추정 (효용지표: $\hat{I}_i = \hat{\beta}_1 + \hat{\beta}_2 x_i$)

B. 다중선택모형(Multinomial Choice Models)

i. 순서(혹은 순위)가 매겨지지 않은 대안들에 대한 선택

1. 출근방법? 버스, 지하철, 택시, 자전거..
2. 초고속인터넷? KT, 하나로, 기타,
3. 냉장고 구입? 삼성, LG, 대우,
4. 정경계열 학생들의 전공선택? 경제, 정치외교, 행정 등등

ii. 이진선택모형의 자연스러운 확장인 multinomial logit model 이나 multivariate probit model 이 사용될 수 있음

1. multinomial logit model

a. logit 모형에서

i. $$p = P[L \le \beta_1 + \beta_2 x] = F(\beta_1 + \beta_2 x) = \frac{1}{1+e^{-(\beta_1+\beta_2 x)}}$$

ii. $$\Rightarrow \frac{p}{1-p} = \frac{1}{1+e^{-(\beta_1+\beta_2 x)}} \times \frac{1+e^{-(\beta_1+\beta_2 x)}}{e^{-(\beta_1+\beta_2 x)}} = e^{(\beta_1+\beta_2 x)},$$

b. 세 개의 대안 A(자가운전), B(버스), C(지하철)가 있을 경우 (C 를 기준이 되는 대안으로 선택 시)

i. $\frac{P(A)}{P(C)} = e^{(\beta_{A1}+\beta_{A2}x)}$, $\frac{P(B)}{P(C)} = e^{(\beta_{B1}+\beta_{B2}x)}$ (“ Independence of Irrelevant Alternatives(IIA)”)

ii. $\Rightarrow P(A) = \frac{e^{(\beta_{A1}+\beta_{A2}x)}}{1+e^{(\beta_{A1}+\beta_{A2}x)}+e^{(\beta_{B1}+\beta_{B2}x)}}$,

$P(B) = \frac{e^{(\beta_{B1}+\beta_{B2}x)}}{1+e^{(\beta_{A1}+\beta_{A2}x)}+e^{(\beta_{B1}+\beta_{B2}x)}}$,

$P(C) = \frac{1}{1+e^{(\beta_{A1}+\beta_{A2}x)}+e^{(\beta_{B1}+\beta_{B2}x)}}$

iii. 따라서 우도함수는 다음과 같이 구축된다…

$$L = \prod_i P(A) \prod_j P(B) \prod_k P(C)$$

c. 이러한 multinomial logit 모형은 선택대안들이 서로 뚜렷한 차이를 지니는 경우에 적용될 수 있다

i. IIA 가 부적절한 제약인 상황에서는 다른 대안 – multivariate probit model 이나 nested logit model 을 고려해야 함 (Beyond the scope of this course!)

C. 가산자료 모형과 포아송 회귀 (Count Data Models and Poisson Regression)

i. 가산자료모형은 어떤 일이 일어나는 횟수에 초점을 맞추는 모형임

1. 종속변수는 0,1,2,3,과 같은 형태의 비음정수의 형태를 띄며 이는 실제 횟수를 나타낸 다는 점에서 앞서의 ordered choice 모형에서의 순위와는 다름

a. 일년에 한 개인이 개인병원을 방문하는 횟수

b. 특정한 교차로에서 한 달에 일어나는 교통사고의 수

c. 어떤 기업이 일년에 출원하는 특허의 수

2. 여기서는 앞서와 마찬가지로 “ 확률” 을 설명하고 예측하는데 관심을 가짐

a. 가산자료와 관련하여 기초가 되는 확률분포는 포아송 분포임

ii. Y 가 포아송 확률변수일 경우, 그 확률 함수는 다음과 같이 주어짐

$$\Pr(Y = y) = \frac{e^{-\lambda}\lambda^{y}}{y!}, \quad y = 0,1,2,\ldots$$

1. 이 확률함수는 하나의 모수 λ에 의존하며, 이는 확률변수 Y 의 평균이자 분산임. 즉 $E(Y) = \lambda$. (어떤 일이 일어나는 평균 횟수)

2. 회귀모형에서 $E(Y)$ 를 일련의 설명변수들의 함수로서 설명하고자 함.

$$E(Y)=\lambda=\beta_1+\beta_2 x\,?$$

3. $E(Y)\geq 0$ 이어야 하므로 다음과 같이 회귀식을 정의

$$E(Y_i)=\lambda_i=e^{\beta_1+\beta_2 x_i}$$

a. 최우추정법이나 비선형최소제곱추정법을 적용할 수 있음.

iii. 최우추정법

1. $y_1 = 1, y_2 = 2$ 및 $y_3 = 4$ 로 관측되고 $x_1 = 15, x_2 = 20$ and $x_3 = 35$ 로 관측되었을 경우 (특허 횟수가 종속변수, 연구개발투자액(억원)이 설명변수)
2. 세 명이 무작위로 추출될 경우 y_1, y_2 및 y_3 에 대한결합확률분포함수는 $f(y_1,y_2,y_3)=f(y_1)f(y_2)f(y_3)$ 로 주어짐
   a. y_1 = 1, y_2 = 2 및 y_3 = 4 이 관찰될 확률은

$$P[y_1=1,y_2=2,y_3=4]=f(1,2,4)=f(1)f(2)f(4)$$
$$=\frac{e^{-\exp(\beta_1+\beta_2 15)}e^{(\beta_1+\beta_2 15)}}{1!}\cdot\frac{e^{-\exp(\beta_1+\beta_2 15)}e^{2(\beta_1+\beta_2 20)}}{2!}\cdot\frac{e^{-\exp(\beta_1+\beta_2 15)}e^{4(\beta_1+\beta_2 35)}}{4!}$$

   b. 이렇게 주어지는 우도 함수를 극대화시키는 $\hat{\beta}_1$, $\hat{\beta}_2$ 를 찾으면 됨.

   c. $\Rightarrow \hat{\lambda}=e^{\hat{\beta}_1+\hat{\beta}_2 x}$ 을 계산할 수 있으며 이를 통해 예컨대, 어떤 기업이 연구개발 투자를 20 억 원을 할 경우 일년에 2 개 이상의 특허를 획득할 확률을 계산 할 수 있음

$$\Pr(Y\geq 2)=1-e^{-\hat{\lambda}}-e^{-\hat{\lambda}}\hat{\lambda}$$

iv. 포아송 회귀의 문제점

1. 포아송 분포는 평균과 분산이 같다는 성질을 가지나 자료의 분포가 이러한 성질을 가정하기가 어려운 경우가 있을 수 있음
2. 이 경우 활용되는 방법이 Negative Binomial Model 임(beyond the scope of this course)