계량 통계학

확률변수 (Random variable)

( 어떠한 범위 안에 ‘경향성’을 가지고 변하는 변수 )

그 범위의 성질 – 어떠한 데이터인가

이산확률변수(countable RV)

연속확률변수(uncountable RV)

Pdf – 확률 밀도 함수 (밀도의 정도에 따라)

F(x) = P(X=x) ; X라는 특정한 사건이 x 만큼 발생할 정도

Cdf(?) - 누적 분포 함수

특정한 사건보다 적은 사건들이 발생한 정도

F(x) = P(X<x) ; x이하(?)

Joint probability density function ; 각각의 다른 사건들이 동시에 벌어질 확률 (중간 기말 만점 동시에) ; 결합확률밀도함수 – 이걸 정확하게 뭐라고 정리해야지?

Marginal distribution ; 다른 사건의 발생의 경우를 무시 – 조사 ㄱ (기말만점?) 주변확률분포

Conditional probability ; 조건부확률 (중간 만점인 사람이 기말 만점?)

통계적 종속과 독립

독립 P(A교B) = P(A)P(B)

종속이면 아니겠지

분포란? ; 확률 밀도 함수가 어떠한 특성을 갖는가? ( 정규분포, 카이제곱분포, 포아송 분포 등등 )

평균 구하기 – 이산과 연속 내

분산! – 평균에 벌어진 정도!

; 상수는 무시하자 영향을 받지 않는다. V(ax+b) = a^2(x)

Covariance

X와 Y의 관계를 지칭? ;

X와 Y가 동시에 커지든가 작아지든가 ( 평균보다 )

X가 E(X)보다 작으면 Y도 E(Y)보다 작아야 { 공분산이 양수이면 }

{공분산이 음수이면 반대겠지}

공분산의 ‘절대값이 크면 클수록’ x와 y가 양의 관계든 음의 관계든 확실히 뚜렷하게 그 관계라고 말할 수 있다.

공분산이 0에 가까울수록 x와 y와 무관하다.(유의미하게)

( 단위 조정의 문제 때문에 공분산 자체로만 판단하기 어려움 )

표준편차로 조정 가능

상관 계수!(correlation coefficient) ; 두 변인을 측정했을 때 한 변인의 변화에 따라 그에 대응하는 다른 변인이 어떻게 변화하느냐의 관계를 표시해 주는 통계치

(-1, 1)

X와 y와의 관계가 무관하면 표본상관계수는 보통 0이지만

표본상관계수가 0이라고 해서 항상 x와 y가 독립이라고 볼 수는 없다.

(But the converse of this relationship is not true. Just because the correlation is zero does no mean they are necessarily independent. Ex) X^2 + Y^2 = 1)

분산 규칙

Var(x+y) = Var(x) + Var(y) + 2cov(x, y)

Var(x-y) = Var(x) + Var(y) – 2cov(x,y)

Var(ax+by) = a^2Var(x) + b^2var(y) + 2abcov(x,y)

정규분포!

표본 평균(Sample들 간의 평균 – 모평균과?)

표본 분산

CLT!

Sample을 랜덤하게 뽑았을 때, 그 n이 무한대로 가면 갈수록, 이 표본 평균이 정규 분포로 가까워진다.

N( m, o2/N )

엑셀 복 ( 평균, 분산, 공분산, 표준편차, 상관 계수 구해보자 )

Stata/IC 13.1 다운

File – import (엑셀 파일 불러오기)

Csv 타입

Xis

Import first forw as variable names – 첫번째 행을 변수 이름으로 간주

Import all data as string – 숫자들을 문자로 읽을 것인가?

( 수사적 의미가 사라짐 )

su(요약) price(변수 이름 입력) – 총 값, 평균, 표준편차, 최대소값

su price, de(보다 더 구체적 정보)

stata 가 엑셀에 기입한 한글이 깨질 수 있음.

간단하게 입력하는 게 좋음

변수 명 바꾸기

remame price P

상관 계수를 구하는 명령어

corr P A

어떤 데이터들인지 그것을 어떤 틀로 분석을 해야

(model-based interpretation)

왜 계량 경제를 공부하니

기존 이론을 데이터를 분석함으로 맞는 지를 확인 (Test of theory)

알려진 이론이나 관계에서도 보다 더 구체적인 수치를 통한 분석이 필요할 때

개인의 소비 모형

계량 분석의 모형 회귀모형!

기본 선형 다항식 모형 ( 각 변수의 변화에 따라 같이 Q가 변화, 그 변화는 일정 )

Ex) 200-300 증가, 500-600 증가든 수요는 똑같이 증가

하나의 변수는 타 변수에 영향을 주지 않는다. 저 함수에선

저 모형을 만들고 데이터를 집어 넣어보고 가장 적합한 beta값들을 찾는다.

베타를 추정 하는 방법

Ex) 가격을 사이즈로 설명한다?

P = B0 + B1A + E ( 선형 회귀식 )

가장 적합한 절편과 기울기는?

OLS

최소제곱추정량 – 자세한 건 노트 참고

MLE는 뭐지?

단순선형회귀이론

x- 단순 변수. Y – 확률 변수

stata

return list : 어떻게 구했는가

scalar 값 산출

scalar list는 scalar 값이 얼마인지 산출

우선 su x해주고 그 다음에 scalar bar\_area 취해준다

scalar bar\_area = r(mean) ; Xbar에 평균값 삽입

gen xx = (X-bar\_x)\*(P=bar\_P) gen xx 변수명

gen yy도 마찬가지로

su xx, su yy를 해주면 합이 나옴

scalar sumA, sumP 값을 지정

stata에서 자동으로 regression 하는 명령어는?

reg 종속변수 설명변수 명령어를 입력하면 사용할 수 있다!

Stata 실습 연습

: set obs 10 ( 10개의 데이터가 있다고 가정 ) – declare 즉, n 설정

gen : 변수를 만드는 명령어 ( gen A = 1 ) – 그럼 a = 1인 10개

gen B+2 식으로도 표현 가능

: gen e = rnormal(0,1) – r은 확률변수 normal

e = 0,1을 따르는 e값 10개 산출

scalar beta\_0 = 5

scalar beta\_1 = 2

gen x = rnomarl(3 ,1 )

gen y = beta\_0+beta\_1\*x + e

우리는 여기까지 참값을 부여했다. 단 x와 e는 확률변수 ( x도 그렇게 함 )

reg Y x

베타 제로와 베타 원의 coeffici가 얼마나 참값과 비슷한가

계속 실행하면서 비교해봐!

script 짜보자! (한번에 실행해주기 위해!)

Do-File Editor

Clear 먼저 해준다! ( 지금까지 했던 걸 지우기 위해)

관측치를 늘려볼까 – 편차가 더 작아진 듯..

에러텀의 분산의 값이 크다? – y를 추정하는 부분의 값을 추정하기 힘들다

에러텀의 분산을 1에서 10으로? 베타제로 베타 원의 추정 값들이 평균과 많이 벗어난다. 에라값이 너무 튄다는 것. Unpredictable의 비중이 너무 크다!

X의 분산도 조정해보자!

X의 분산이 커지면 더더욱 정확성이 좋아지네! ( 띄엄띄엄 뽑아야 더 정확 )

In stata

Reg P X

R square!

Stata – 데이터 불러오자!

su price

return list

scarlar y\_bar = r(mean)

ereturn list – reg를 돌릴 때 나온 결과들…보여줌

predict y\_hat : 이 y hat을 가져오자

predict값을(추정한 y값) y\_hat이라 하자

matrix list e(b) – 이 매트릭스의 구성요소를 알고싶다!

gen e = price-y\_hat

gen st = price-y\_bar 제곱

gen sr = y\_hat-y\_bar 제곱

gen se = e^2

su st

scalar sst = r(sum)

su sr

scalar ssr = r(sum)

su se

scalar sse = r(sum)

scalar R\_squared = ssr/sst

scalar list R\_squared

다중선형 회귀이론!!

변수가 이제 여러 개다..

Stata

Tex data(delimited, \*csv, …) 참조

Reg를 다양하게 변수 넣어가면서 해보자!

Reg p size f d\_p …

p-value = 0.0000 이라는건 귀무가설을 기각한다! – 좋은 추정치

변수를 더 늘리고 보니 R-square가 증가했네

nocons

reg 변수 변수, nocons – 베타 제로 항목 없애기 – 이 걸 하는 순간 R-square는 믿기 어려운 값이 된다?? 왤까…

p-value가 작다가 갑자기 크게 나오는 경우도 있다. 변수간의 correlation이 있을수도 있다…

이 부분에 대한 조정도 필요!

자연 로그 씌우는 법

변수 하나 만들어라

gen ln\_f = ln(f)

reg p ln\_f – 층수의 비율만큼 변하는 price

이런 식으루 사용 가능!

R square를 다 비교해보자

reg p f – 0.01

reg p ln\_f – 0.0044

등등

이것 저것 다 해보는 거!!

gen f\_2 = f^2 이런식으로두 가능!!

설명 변수간의 관계는 없다 가정

Indicate!!

특정 동을 기준으로!! 이거 0 으로 하고!

다운 받은 메모장 참고

Tab dong, gen(dong)

Tab – 정성적 변수 ( 각 동들로 정의! – 이 변수들의 표를 그려봐라)

Tap dong, gen(dong) – dong1 ~ dong 11 까지 0으로 설정

Reg p dong1 dong2 dong3 dong4 … dong11

하나를 빼주면 된다!! (기준점!!)

Reg p dong2 dong3 …

Indicate variable은 이렇게 한다!!

그리고 그 dong들의 p-value보자!

기준점에 비해서 기준점과 이만큼 더 다른 효과를 같는다.

Coefficient는 0이다. 기준점이다!

기준점 기준 요만큼 더 다르다?

Dong 6의 경우, h0를 기각 못함.(p-value가 큰 경우) 즉, dong1과 큰 차이가 없다!!

Ln\_p를 쓰면 더 효율적!!

Xi 명령어

Xi : reg p i.dong 이런 명령어도 사용 가능!!

Indicate가 아닌 경우도 사용 가능

Xi : reg p size f i.dong i.time

Misspecification!!

Hisnet에 있는 자료 참고!!

set obs 100 모든 변수의 개수는 100개!

gen a = rnormal(0, 1) – 이 분포를 따르는 변수 대입

gen b 도..

gen e1=rnormal(0, 1) – 에러다

gen e2? – 에러가 2개네?? D 추정을 위해?

Gen c = 10\*a + 10\*b + e1??

Reg c a b

Gen d = a+b+2\*c+e2

Reg d a b c – 값이 이상

즉, c가 multicorrelaniary 오류 발생

처음부터 아예 reg하기 전에 변수 간의 관계를 보는 방법도 있다.

VIF는 그럴 때 사용

C를 빼버리면??

Reg d a b

A b가 엄청 커져버림….

omitted에서

어떤 변수하나를 넣고 뺌에 따라 R-square가 크게 변한다??

그거는 omitted가 발생했을 가능성 높다!

계량 경제 data sample들이 뭐가 있을까..?

데이터를 수집한, 관찰한 타이밍에 따른 구분

Cross-sectional data – 타이밍의 이슈가 없다? – 언제 수집한 건지가 중요x, 동기간 수집의 경우

Time-series data – 시간에 따른 변화를 추적..

Panel data – Cross-sectional + panel

: 특정 시점의 데이터를 횡단면으로

가구의 소득 조사 (2017) – cross sectional

가구의 소득 조사 ( 2017, 2016, 2015) – time series data

2017년 가구 소득

2016년 동일한 가구 소득 조사

2017, 2016 time-series

통계청

가계금융복지조사

가계의 소득, 자산, 부채, 가구원의 종류

소득 분위 별 부채의 비율…, 가구원 종류에 따른.. 뭐 기타 등등

가계 동향 조사

MDIS – 원자료 (micro data)가 많은 곳인가봄

이것도 참고 해봐야겠다!!

추출- 다운로드!

많이 빡친데…

거시경제 이슈 – 한국경제통계시스템

시중 금리의 변화에 따른 상장 기업의 주가 변화??

소비자 동향조사..? (거시의 이슈지!)

가처분소득도 중요!! – 내 수중에 있는 돈이니

OECD statistics – 선진국 정보

Penn world tables

재정패널 – 통계청 물가 가계란

OLS – sample이 매우 많아야되

쌉 dw