미래의 불확실성을 줄여주는 'Runs test'

255호 (2018년 8월 Issue 2)

오늘날 우리는 '데이터의 시대'를 살아가고 있다. 데이터의 양도 폭발적으로 증가하고 있지만 유형 역시 예전에는 가히 상상할 수 없었을 정도로 다양해지고 있다. 또한 4차 산업혁명의 도래와 사물인터넷의 확대로 우리는 곧 '데이터의 시대'를 넘어 '슈퍼 데이터의 시대'를 마주하게 될 것이다. 단적인 예로, 예전에는 특정 시점(t)에서 횡단면적(Cross-sectional)으로 기록한 '스냅숏(Snapshot)' 형태의 데이터만 존재했다면 오늘날에는 무수히 많은 연속적 시점(t, t+1, t+2...)에서 마치 일기를 써내려가듯이 대상을 순차적으로 따라가며 종단적(Longitudinal)으로 기록한 데이터 역시 쉽게 얻을 수 있다. 계량경제학에서는 이를 흔히 패널데이터(Panel data)라고 부른다. 가령, 소비자 A가 특정 제품을 얼마나 사용하는지, 제품에 대한 만족도는 어떻게 달라지는지를 '매월' 관찰해 데이터로 남긴다면 그 제품의 제조사는 데이터로부터 당연히 훨씬 풍부한 정보를 얻게 될 것이다.

이렇게 종단적으로 기록된 데이터는 미시적(Micro) 특성에 더해 시간의 흐름에 따라서 관찰된 자료라는 점에서 시계열 자료(Time series data)로서의 특성 역시 지니게 된다. 시계열 분석은 불확실한 미래에 대한 통계적 예측(Forecasting)을 위해서 사용되는 경우가 많은데 한국은행이 정책 수립을 위해 우리나라의 내년도 경제성장률을 예상하거나 항공사가 노선별 기종 배분을 위해 월별 탑승객 수를 미리 추정하는 것 등이 대표적인 사례다. 엄격한 시계열 분석은 여러 가정(Assumption)에 기반하고 있고 이러한 가정이 하나라도 위반된다면 분석 전체가 망가지게 되므로 계량분석 기법 중에서도 난도가 높은 편에 속한다. 요구되는 여러 가정 중에서는 통계적 기법을 통해 위반 여부를 확인할 수 있는 것들도 있지만 그렇지 않을 것들도 매우 많다. 하지만 제대로 수행된다면 기업의 의사결정 과정에 강력한 과학적 무기가 돼준다는 점에서 효용 역시 높다. (그래서인지 필자는 최근 유럽에서 시계열 분석 능력을 갖춘 데이터 애널리스트에 대한 기업들의 수요가 매우 높음을 새삼 느끼고 있다.) 필자는 이 글을 통해 기업 실무에서 누구나 간단하게 사용할 수 있으면서도 다양한 분야에서의 활용이 가능한 시계열 분석 기법 한 가지를 소개하고자 한다. 다음의 사례를 통해 논의를 시작해보자.

사례 1

베트남에서 생산 설비를 운영하고 있는 한국 대기업 L사의 K 과장은 현지 원재료 조달을 담당하고 있다. 현지에서 공수되는 해당 원재료는 선물(Futures) 시장이 존재하지 않으며 일(日) 단위로만 거래가 이뤄지고 있다. K 과장은 최근 본사로부터 전략적 구매를 통해 원재료 비용을 낮추라는 지시를 받고 구매전략을 수립하는 중이다. 이에 K 과장은 우선 해당 원재료의 가격 변화 추이에 모멘텀(Momentum)이 존재하는지를 알아보고자 한다.

모멘텀이란 무작위 진행(Random walk)과 대비되는 것으로 시계열 데이터의 변화 양상에 뭔가의 가속 추세(Acceleration trend)가 존재한다는 것을 지칭하는 개념이다. 모멘텀(Momentum)은 본래 '운동량'을 뜻하는 물리학 용어인데 이를 금융경제학에서 차용해 사용하게 됐다. 위의 사례를 기반으로 설명하자면, 모멘텀의 존재 여부 확인을 통해 일(日) 단위로 거래되는 원재료의 가격이 하루 오르면 다음 날도 오르고, 반대로 하루 내리면 그다음 날도 내리는 '양의 계열 상관(Positive serial correlation) (+)가 (+)로 이어지고, (-)가 (-)로 이어지는 것을 양의 계열 상관이라고 한다. 경향'을 지니는지 알아보는 것이다. 원재료 가격에 모멘텀이 존재한다면 어제 대비 오늘 가격이 올랐을 경우 내일도 오를 확률이 높기 때문에 오늘

구입량을 늘리는 것이, 그리고 어제 대비 오늘 가격이 내렸다면 내일도 내릴 확률이 높기 때문에 오늘 구입량을 줄이는 것이 전략적 구매 행위일 수 있다. 물론 원재료 가격 이외에 구매의사결정에 영향을 미치는 다른 요소들은 모두 일정하다는 가정하에서 이와 같은 명제는 성립된다.

그렇다면 모멘텀의 존재 여부를 어떻게 파악할 수 있을까. 이를 위해 우리는 '런 검정(Runs test)'이라는 간단한 시계열 분석 기법을 활용할 수 있다. 우선 '런(Run)'이라는 생소한 개념에 대해서 설명하고자 한다. 통계학에서는 한 종류의 부호(+ 혹은 -)가 시작해 끝날 때까지를, 다시 말해 특정 부호가 반대의 부호로 바뀔 때까지를 하나의 런으로 정의한다. 아래는 K 과장이 지난 한 달간의 원재료 가격을 부호 형태로 나타낸 것이다. (+)는 전일 대비 가격 상승을, (-)는 전일 대비 가격 하락을 각각 의미한다.

+++++ ----- +++ --- ++++ -----

이 경우 총 관측치의 개수는 25, 첫 번째 런(+++++)의 길이는 5, 두 번째 런(-----)의 길이는 5, 세 번째 런(+++)의 길이는 3, 네 번째 런(---)의 길이는 3, 다섯 번째 런(++++)의 길이는 4, 여섯 번째 런(-----)의 길이는 5이다. 따라서 총 런의 개수는 6이며, (+)의 총 개수는 12개, (-)의 총 개수는 13개다. 즉, 런을 '같은 부호가 연속적으로 이어진 덩어리'라고 간주하면 쉬운 이해가 가능하다. 모멘텀이 존재한다면, 다시 말해서 양의 계열 상관이 존재한다면 런의 길이는 평균적으로 길어지게 되는 반면 런의 개수는 줄어들게 될 것이다. 직관적으로 이해할 수 있는 이러한 간단한 논리에 기반한 것이 바로 런 검정이다. 런 검정은 런의 수가 충분히 적으면 양의 계열 상관이, 런의 수가 충분히 많으면 음의 계열상관 양의 계열 상관과 대비되는 개념으로, (+)가 (-)로 이어지고, (-)가 (+)로 이어지는 것을 의미한다. 이 존재한다고 판단한다. 그리고 런 검정의 영가설(Null hypothesis)은 시계열 데이터들이 무작위로 분포돼 아무런 추세도 없다는 것이다. 영가설이란 검정을 통해서 반박해야 하는 가설을 의미한다. 영가설과 대립가설에 대한 자세한 내용은 DBR 249호를 참고하기 바란다.

표1 런 검정의 하한임계치 및 상한임계치

5% 유의수준 하한 임계치(가로축은 -의 개수, 세로축은 +의 개수, 표의 북동쪽과 남서쪽은 상호 대칭)

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2																			
3																			
4																			
5			2	2	3														
6			2	2	3	3													
7			2	2	3	3	3												
8			2	3	3	3	4	4											
9			2	3	3	4	4	5	5										
10			2	3	3	4	5	5	5	6									
11			2	3	4	4	5	5	6	6	7								
12		2	2	3	4	4	5	6	6	7	7	7							
13		2	2	3	4	5	5	6	6	7	7	8	8						
14		2	2	3	4	5	5	6	7	7	8	8	9	9					
15		2	3	4	4	5	6	6	7	7	8	8	9	9	10				
16		2	3	4	4	5	6	6	7	8	8	9	9	10	10	11			
17		2	3	4	4	5	6	7	7	8	9	9	10	10	11	11	11		
18		2	3	4	5	5	6	7	8	8	9	9	10	10	11	11	12	12	
19		2	3	4	5	6	6	7	8	8	9	10	10	11	11	12	12	13	13
20		2	3	4	5	6	6	7	8	9	9	10	10	11	12	12	13	13	13	14

5% 유의수준 상한 임계치(가로축은 -의 개수, 세로축은 +의 개수, 표의 북동쪽과 남서쪽은 상호 대칭)

	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
2																			
3																			
4																			
5				9	10														
6				9	10	11													
7					11	12	13												
8					11	12	13	14											
9						13	14	14	15										
10						13	14	15	16	16									
11						13	14	15	16	17	17								
12						13	14	16	16	17	18	19							
13							15	16	17	18	19	19	20						
14							15	16	17	18	19	20	20	21					
15							15	16	18	18	19	20	21	22	22				
16								17	18	19	20	21	21	22	23	23			
17								17	18	19	20	21	22	23	23	24	25		
18								17	18	19	20	21	22	23	24	25	25	26	
19								17	18	20	21	22	23	23	24	25	26	26	27
20								17	18	20	21	22	23	24	25	25	26	27	27	28

그렇다면 무엇을 기준으로 런의 수가 '많고 적음'을 판단할 수 있을까. 다음의 [표 1]이 그 기준을 제시해준다. 각 부호의 개수가 표에 제시된 20을 넘어갈 경우 정규분포에 의한 근사가 가능하다. 8 [표 1]은 류근관 著의 『통계학(법문사, 2010)』에 제시된 것을 차용했음을 밝힌다. K 과장의 사례에서 (+)의 총 개수는 12개, (-)의 총 개수는 13개이므로 [표 1]에 의하면 런의 하한 임계치(Lower critical value)는 8, 상한 임계치(Upper critical value)는 19다. 그런데 위에서 관측된 런의 개수는 6개이므로 하한 임계치를 벗어나 밑돌게 된다. 따라서 우리는 '런의 개수는 적고 런의 길이가 길다'고 판단하게 되며, 원재료 가격 추이에 '양의 계열 상관, 즉 모멘텀이 존재한다'고 결론 내리게 되는 것이다. 이는 바꾸어 말하자면 베트남에서 원재료 가격의 변화 추이는 무작위가 아니라 뭔가의 체계적 추세에 의한 영향을 받는다는 뜻으로, 이러한 추세를 잘 감안한다면 전략적 구매가 가능할 수 있다는 의미이기도 하다. 만약 이와 달리 런의 개수가 상한 임계치인 19를 넘어섰다면 반대로 음의 계열 상관이 존재한다고 판단할 수 있었을 것이다. 엑셀(Excel)을 통해 런 검정을 활용하고자 하는 독자들을 위해 엑셀 함수를 [표 2]에 정리해두었으니 참고하기 바란다. 9 [표 2]에서 음영 처리된 곳은 상황에 맞게 직접 입력이 필요하다.

표2 런 검정을 위한 엑셀 함수

	A	B	C	D	E
1					
2					
3		**Runs test**			
4					
5		# of (+)	12		
6		# of (--)	13		
7		# of observations (N)	25		=SUM(C5:C6)
8		# of runs	6		
9		mean (test statistics)	12.00		=2*C5*C6/(C7+1)
10		variance	5.97		=2*C5*C6*(2*C5*C6-C7)/C7^2/(C7-1)
11		standard deviation	2.44		=SQRT(C10)
12		significance level (alpha)	0.05		
13		z	2.46		=ABS(C8-C9)/C11
14		tails	2		
15		p-value	0.01		=2*(1-NORMSDIST(C13))
16		z-critical value	1.96		=NORMSINV(1-C12/2)
17		significance	Yes		=IF(C15<C12,"Yes","No")
18		conclusion	Null rejected		=IF(C17="Yes","Null rejected","Null not rejected")
19		결론	계열상관 존재		=IF(C17="Yes","계열상관 존재","계열상관 부재")
20					
21					
22					
23					

독자들의 이해를 돕기 위해 한 가지 재미있는 사례를 통해서 런 검정에 대한 개념을 다시금 정리해보려 한다.

사례 2

메이저리그 유명 구단의 스카우트 담당자인 J 박사는 최근 각광받고 있는 투수 B의 영입과 관련된 의사결정을 위해 B 투수의 경기력 분석 작업을 수행 중이며, B 투수의 경기력에 기복(起伏)이 있는지를 관찰하고 있다. 투구에 기복이 있다면 스트라이크 혹은 안타가 시계열상으로(즉, 경기의 흐름에 따라) 무리 지어 몰려 있을 공산이 크다. 다시 말해, 스트라이크 이후에는 다시 스트라이크가 연이어 나올 확률이, 안타 이후에는 다시 안타가 연이어 나올 확률이 높을 것이라는 얘기다. J 박사는 구단주를 위해 보다 과학적인 보고서를 작성하고자 한다.

이 경우에도 런 검정이 효과적으로 활용될 수 있다. B 선수의 경기기록을 위에서 얘기한 바와 같이 시계열 데이터처럼 분석하면 되는 것이다. B 선수의 투구 결과를 (+)와 (-)의 부호 형태로 바꾸어 정리한 다음 런의 개수와 각 부호의 개수를 세어 [표 1]과 비교해 보기만 하면 된다. 스트라이크는 (+), 안타는 (-)로 기록할 수도 있고, 혹은 그 반대도 가능하다. 부호 대체 작업은 일관성만 유지된다면 어떤 식으로 기록하든 상관이 없다. 볼은 투수가 투구를 제대로 하지 못한 결과로 보고 분석에서 일괄 제외했다. 만약 총 20번의 투구에서 스트라이크가 12번, 안타가 8번이었고, 런의 개수는 4개였다면 이는 [표 1]에 제시된 하한 임계치인 6보다 작은 값이므로 B 투수의 투구에 양의 계열 상관, 즉 모멘텀이 존재한다고 결론을 내릴 수 있다. 따라서 J 박사는 B 선수의 경기력에 기복이 있다는 의견을 담은 보고서를 구단주에게 제출해

야 하는 것이다. 물론, B 투수와 마주한 각 타자의 역량이 유의하게 다르지 않았다는 가정하에서 이와 같은 의견은 설득력을 지니게 된다.

인간이라면 누구나 불확실한 미래에 대한 두려움을 가지고 있으며, 막대한 자본과 인력을 운용하는 기업 입장에서 그 두려움의 크기는 매우 클 수밖에 없다. 따라서 그러한 불확실성을 조금이라도 줄여보고자 하는 것이 기업 운영자의 당연한 염원일 것이다. 이를 위해 전근대에는 소위 '수정구슬(미신)'에 의존했고, 그 이후에는 이론에 의지했으며, 최근에는 경험이 풍부한 전문가의 의견에 기댔었다면 이제는 방대하게 축적된 다양한 유형의 데이터와 고도의 계량분석 테크닉이 각광을 받게 됐다. 물론 필자가 위에서 수차례 언급했듯이 이러한 계량분석 기법은 여러 가정에 기반하고 있다는 점에서 절대로 만병통치약과 같은 '완벽한 무기'가 돼줄 수는 없다. 하지만 각각의 상황에서 요구되는 가정을 철저하게 이해하고 있는 사람에 의해 절차적 오류 없이 기업 실무에서 제대로 활용된다면 '완벽한 미래 예측'은 아니더라도 '과학적인 추정'은 가능하게 해 줄 것이라 확신하는 바이다.

필자소개 신선호 롤랜드버거 시니어 컨설턴트 seonho.shin@rolandberger.com

필자는 서울예술고등학교를 졸업(피아노 전공)하고 서울대 경제학과를 졸업했으며 독일 연방정부의 국비장학생으로 독일 프랑크푸르트대 경영경제학 석사 학위를 받았다. 독일 함부르크의 에어버스와 모니터그룹을 거쳐 독일 뮌헨에 본사를 둔 유럽 최대 컨설팅 기업 롤랜드버거에서 시니어 컨설턴트로 재직 중이다. 프랑크푸르트대 계량경제학 연구실에서 강의와 연구를 병행하고 있다. 그동안 정책효과 분석, 수요 예측, 시장 세분화 등 퀀트 기반의 다양한 프로젝트를 수행해 왔다.