Project 2

Constructing Purchasing Power Parity : BigMac

2014122006 허승민

2019년 5월 7일

1 데이터

빅맥 데이터는 구매력 지수 중의 하나인 빅맥지수와 영향을 줄 것으로 예측되는 여러 변수로 구성되어있다. 먼저 반응변수라고 볼 수 있는 빅맥지수는 맥도날드의 빅맥 세트를 구매하기 위해 필요한 노동시간을 분단위로 계산한 데이터이다. 분포가 왜도가 있고 이상치가 있어보이므로 정규화 형태로 바꾸기 위해 로그 변환을 실시한다.(Appendix 1)
로그변환한 반응변수와 설명변수 간의 선형 관계를 파악하기 위해 산점도를 그린 결과 선형관계가 존재하는 것으로 보인다.(Appendix 2) 따라서 선형회귀 모형을 사용하도록 한다. 설명변수를 분석해보면 몇몇 범주로 정리가 되는데 정리하면 다음과 같다.

범주	변수	설명
생활 관련 물가 지수	Bread	빵 1kg을 사기 위한 노동시간(분)
	BusFare	교통비용
	Service	19가지 서비스에 대한 연간비용
급여	EngSal	전기기사의 연봉
	TeachSal	초등학교 교사의 연봉
세율	EngTax	전기기사의 세율
	TeachTax	초등학교 교사의 세율
노동력 관련 지수	WorkHrs	연간 노동시간
	VacDays	연간 휴가 일 수

통계적 분석이 아닌 개인적인 판단을 근거로 설명변수를 분류하였다. 이렇게 나뉜 범주에 따라 설명변수 간의 관계 또한 확인해 보았다. 4개의 범주 중 급여, 세율, 노동력 지수 간의 상관 관계가 있으며 급여와 세율은 특히나 높은 것(>0.84)으로 보인다.(Appendix 2) 따라서 두 범주는 각각 하나의 변수로 주성분 분석하여 변수변환하였다. 노동력지수의 경우는 상관관계가 있지만 분석을 하면서 살펴보고 추가적인 변환을 고려하고자 하였다.

2 모델링

모형 수립은 1번 단계에서 정리된 변수를 이용하여 단계적으로 실시하였다.

2.1 1차 모델링

- 사용변수 : Bread, BusFare, Service, VacDays, WorkDays, pca_Sal, pca_Tax

주성분 분석을 실시하여 변수변환한 두 변수를 사용하여 첫번째 모델링을 실시하였다. 잔차산점도를 보면 이상치가 하나 있는 것으로 보인다.(Appendix 3) 따라서 이를 제거하고 같은 변수로 선형회귀 모형을 다시 수립하였다. 결과적으로 해당 점은 이상치이면서 Influential case에 해당하는 점으로 확인할 수 있었다. 해당 점을 삭제하고 모형을 다시 수립하면 유의한 변수도 추가적으로 많이 생기며 모형의 결과도 좋아지는 것을 확인할 수 있다. 잔차산점도 역시 고른 분포형태를 보인다고 말할 수 있다. 이상치 점은 뒤에서 추가로 논의하며 유의하지 않았던 변수에 대한 변환을 실시하고 최종 모형을 수립한다.

2.2 최종 모델링

- 사용변수 : Bread, VacDays, WorkDays, pca_Sal, pca_Tax, pca_Life

유의하지 않았던 변수 중 BusFare 와 Service의 경우 생활 관련지수로 상관관계가 높은 변수들이었다. 따라서 이를 반영하여 주성분분석으로 변수변환하여 하나의 변수로 사용하여보았고 결과를 확인하였다.(Appendix 4) 결과적으로 역시나 유의하지 않았으며 이를 제거하고 최종모형을 수립하였다. 최종적으로 선택된 변수는 Bread, VacDays, WorkDays, pca_Sal, pca_Tax 이다.

3 결론

$$log(BigMac) = 0.52 + 0.01 * Bread + 0.04 * VacDays$$

$$+0.009 * WorkDays + 2.7 * pcaSal - 0.94 * pcaTax$$

$$BigMac = e^{0.52+0.01*Bread+0.04*VacDays+0.009*WorkDays+2.7*pcaSal-0.94*pcaTax}$$

최종적인 관계는 빅맥지수의 로그변환이 설명변수들과 선형관계를 갖는다는 것을 알 수 있었다. 급여와 세금과 관련된 변수가 다른 변수에 비해 더 강한 상관관계를 가지며 각각 방향은 양, 음의 방향을 가지고 있다. 근무시간, 휴가일수 등은 관계는 있으나 매우 미미한 영향을 미치며 빵의 물가 또한 미미한 영향임을 알 수 있다. 정리하면 구매력지수와 관련하여 물가는 생각보다 관련성이 적어 변수가 기각되었으며, 급여나 세율부분이 더 큰 영향력을 가진다는 것을 확인할 수 있었다.
이상치로 판별된 멕시코시티의 경우 설명변수에서는 이상치가 발견되지 않고 대부분의 경향성을 따라가지만 빅맥지수가 235에 육박한다. 제일 영향력있는 급여나 세율의 경우도 다른 도시들과 차이를 보이지 않는다. 따라서 내릴 수 있는 추론은 1) 멕시코시티의 빅맥 가격이 매우 비싼편이거나, 2) 빅맥을 소비하고 구매가능한 계층과 일반 서민 노동자 계층간의 빈부격차가 매우 심해 일반 노동자의 급여로는 구매력이 매우 떨어지는 상태일 것이라는 추측이 가능하다. 또 다른 예측으로는 측정이 잘못되었을 수도 있다는 결론을 내릴 수 있다. 때문에 좀 더 깊은 탐색이 필요한 부분이라는 과제를 남기는 데이터이다.

Appendix

Appendix 1 : 반응변수 탐색

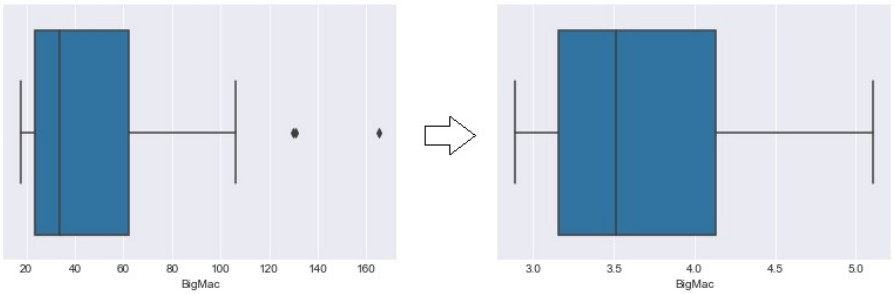

그림 1.1 : 빅맥 지수의 분포 형태 및 변환

분포의 형태가 왜도가 많이 반영되어있는 것으로 보이므로 로그 변환하여 반응변수를 좀더 정규화된 형태로 바꾸었다.

Appendix 2 : 설명변수 탐색

그림 1.2 : 반응변수와 설명변수의 관계 및 설명 변수간의 관계

로그 변환한 빅맥지수와 다른 변수들 간의 관계를 산점도를 통해 확인해보았다. 대부분 어느 정도 선형 관계를 가지고 있는 것으로 보인다. 설명변수들 간의 관계에서는 상관계수가 높아 공선성의 문제가 있을 것으로 파악되어 추가적인 조치가 필요할 것으로 보인다. 따라서 주성분 분석을 실시하였다.

Appendix 3 : 1차 모델링

모형 결과		
1차 모델링	모형	log(BigMac) = 2.67 + 0.01*Bread - 0.25*BusFare + 0.007*Service
		+ 0.02*VacDays + 3*pcaSal - 0.98*pcaTax + 0.003*WorkDays
	유의변수	Bread, pca_Sal
	$R^2(R^2_{adj})$	0.684(0.734)
수정 모델링	모형	log(BigMac) = 0.96 + 0.01*Bread - 0.19*BusFare + 0.0004*Service
		+ 0.03*VacDays + 2.54*pcaSal - 1.27*pcaTax + 0.01*WorkDays
	유의변수	Bread, VacDays, pca_Sal, pca_Tax, WorkDays
	$R^2(R^2_{adj})$	0.784(0.819)

표 1 : 모델링 결과

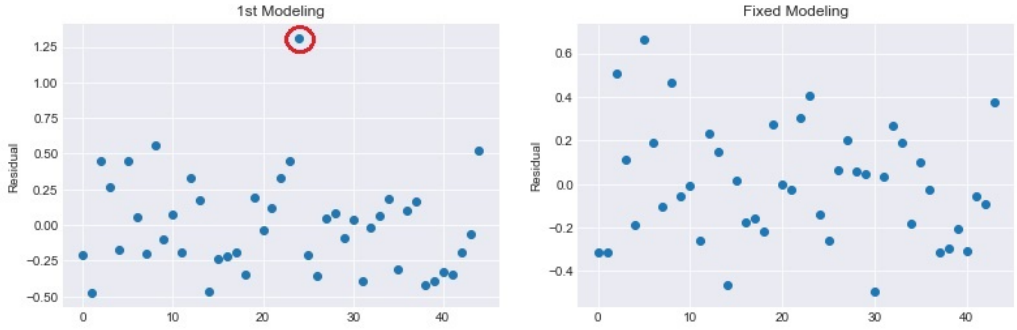

그림 2.1 : 1차 모델링 및 수정 모델링 결과

Appendix 4 : 최종 모델링

모형 결과		
2차 모델링	모형	log(BigMac) = 0.63 + 0.01*Bread + 0.04*VacDays + 0.009*WorkDays
		+ 2.5*pcaSal - 1.1*pcaTax + 0.41*pca_Life
	유의변수	Bread, VacDays, WorkDays, pca_Sal, pca_Tax
	$R^2(R^2_{adj})$	0.782(0.812)
최종 모델링	모형	log(BigMac) = 0.52+ 0.01*Bread + 0.04*VacDays
		+ 0.009*WorkDays + 2.7*pcaSal - 0.94*pcaTax
	유의변수	Bread, VacDays, WorkDays, pca_Sal, pca_Tax
	$R^2(R^2_{adj})$	0.783(0.809)

표 2 : 모델링 결과

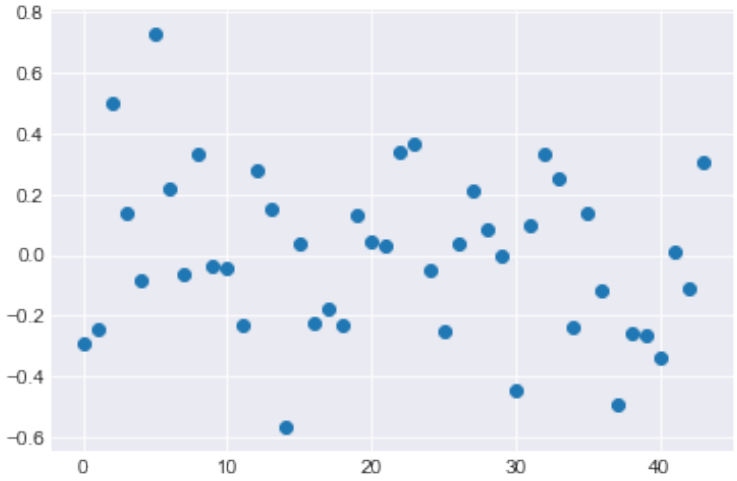

그림 2.2 : 최종 모델링 결과