Linear mixed effects model

Yoon-Ho Hong

선형회귀모델에서 이어집니다. 선형회귀모델을 먼저 보고 오세요.

## 선형혼합효과모델(Linear mixed effects model)

gdpPercap\_log10과 continent 변수, 그리고 두 변수간의 교호작용을 예측변수로 하고, 기대수명을 결과변수로 하는 다변량 선형회귀 모델에서 다음과 같은 사실을 알 수 있었습니다.

* continent에 따라 기대수명이 다름
* gdpPercap\_log10이 증가함에 따라 기대수명이 증가함
* continent에 따라 gdpPercap\_log10이 기대수명에 미치는 효과가 다름

그렇다면, 두 변수간의 교호작용을 예측변수로 포함한 것과 그렇지 않은 모델 중에서 어떤 것을 선택해야 할까요?

* 1. lifeExp ~ gdpPercap\_log10 + continent
  2. lifeExp ~ gdpPercap\_log10\*continent

첫번째 모델은 두 변수간의 교호작용을 고려하지 못하고 있습니다.  
두번째 모델은 continent와 gdpPercap\_log10:continent 변수의 공선성이 매우 큰 것 같습니다. 공선성을 아래에서 확인해봅시다.

## 패키지

library(dplyr)  
library(gapminder)  
library(car) # car::vif()   
library(lme4)

1952년 gapminder 데이터셋에서 gdpPercap\_log10과 continent 간의 교호작용을 예측변수 항으로 포함하는 선형회귀 모델을 적합해봅시다.

df\_1952 = gapminder %>%  
 filter(year == 1952) %>%  
 mutate(gdpPercap\_log10 = log10(gdpPercap))   
lm3\_1952 = lm(lifeExp ~ gdpPercap\_log10\*continent,   
 data = df\_1952)  
summary(lm3\_1952)

Call:  
lm(formula = lifeExp ~ gdpPercap\_log10 \* continent, data = df\_1952)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-14.7288 -3.5801 0.1119 2.5009 15.2270   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 23.053 8.239 2.798 0.00591 \*\*  
gdpPercap\_log10 5.382 2.743 1.962 0.05184 .   
continentAmericas -54.739 19.795 -2.765 0.00650 \*\*  
continentAsia -7.220 10.754 -0.671 0.50312   
continentEurope -34.999 17.645 -1.984 0.04938 \*   
continentOceania -3.480 1571.782 -0.002 0.99824   
gdpPercap\_log10:continentAmericas 18.659 5.774 3.231 0.00156 \*\*  
gdpPercap\_log10:continentAsia 4.196 3.483 1.205 0.23048   
gdpPercap\_log10:continentEurope 15.350 5.038 3.047 0.00279 \*\*  
gdpPercap\_log10:continentOceania 6.999 391.713 0.018 0.98577   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.04 on 132 degrees of freedom  
Multiple R-squared: 0.7715, Adjusted R-squared: 0.7559   
F-statistic: 49.52 on 9 and 132 DF, p-value: < 2.2e-16

VIF(variance inflation factor)를 확인해봅시다.

vif(lm3\_1952)

there are higher-order terms (interactions) in this model  
consider setting type = 'predictor'; see ?vif

GVIF Df GVIF^(1/(2\*Df))  
gdpPercap\_log10 5.879826e+00 1 2.424835  
continent 2.771206e+11 4 26.936031  
gdpPercap\_log10:continent 3.249760e+11 4 27.477766

GVIF가 4보다 크면 공선성이 있는 것으로 간주하고, 10보다 크면 이 문제를 해결하기 위해 반드시 조치를 취해야 합니다. 이러한 경험 법칙을 스케일링된 GVIF에 사용하려면, 규칙을 적용하기 전에 스케일링된 GVIF 값을 제곱합니다.

## 고정 효과와 랜덤 효과

위와 같은 상황에서 선형혼합모형은 좋은 대안이 될 수 있습니다. 선합혼합모형은 고정 효과(fixed effect)와 랜덤 효과(random effect)로 이루어진 모형을 말하며, 다음과 같은 상황에서 좋은 분석 방법이 될 수 있습니다.

* 데이터가 군집(cluster)으로 구성되어 있거나,
* 반복 측정 데이터인 경우(예를 들면, longitudinal data)

먼저 고정 효과와 랜덤 효과의 차이에 대해 이야기해 보겠습니다. 이 차이는 변수 자체와는 거의 관련이 없으며 연구 질문과 더 관련이 있다는 점을 꼭 기억하시기 바랍니다.

많은 경우 동일한 변수가 랜덤 효과 또는 고정 효과로 간주될 수 있으므로, 항상 연구 질문(즉, 가설)이 무엇인지를 참조하여 그에 따라 모델을 구성해야합니다.

넓은 의미에서 고정 효과는 결과 변수에 영향을 미칠 것으로 예상되는 변수로, 선형회귀 분석에서 설명 변수라고 부르는 것입니다. gapminder 사례에서 우리는 일인당 국민소득이 기대수명에 어떤 영향을 미치는지에 대한 결론을 내리는 데 관심이 있습니다. 따라서 일인당 국민소득은 고정 효과 변수입니다.

반면에 랜덤 효과는 일반적으로 우리가 통제하고자 하는 요인을 그룹화하는 것으로 항상 범주형입니다. 대부분의 경우 우리는 랜덤 효과 변수가 결과 변수에 미치는 영향에 특별히 관심이 없지만, 우리가 보는 고정 효과 변수가 결과 변수에 미치는 효과에 영향을 미칠 수 있다는 것을 알고 있습니다.

우리는 continent에 따라 기대수명이 얼마나 달라지는지를 추정하는 데는 관심이 없지만, continent가 기대수명이 다른 이유일 수 있다는 것을 알고 있으며, 각 continent에서 기대수명을 예측할 때 이로 인한 변동이 얼마나 되는지 알고 싶습니다.

## 랜덤 절편

랜덤 효과는 크게 random intercepts와 random slopes, 이렇게 두 가지 방식으로 모형화가 가능합니다.

먼저, continent에 대해 random intercept 효과를 도입한 모형을 먼저 살펴보겠습니다.

lme\_mod1 = lmer(lifeExp ~ gdpPercap\_log10 + (1|continent), data = df\_1952)  
summary(lme\_mod1)

Linear mixed model fit by REML ['lmerMod']  
Formula: lifeExp ~ gdpPercap\_log10 + (1 | continent)  
 Data: df\_1952  
  
REML criterion at convergence: 934.2  
  
Scaled residuals:   
 Min 1Q Median 3Q Max   
-2.56354 -0.52248 -0.00694 0.54308 2.22869   
  
Random effects:  
 Groups Name Variance Std.Dev.  
 continent (Intercept) 57.02 7.551   
 Residual 39.87 6.315   
Number of obs: 142, groups: continent, 5  
  
Fixed effects:  
 Estimate Std. Error t value  
(Intercept) 14.149 6.389 2.215  
gdpPercap\_log10 11.457 1.549 7.398  
  
Correlation of Fixed Effects:  
 (Intr)  
gdpPrcp\_l10 -0.837

먼저 랜덤 효과에 대한 결과를 살펴보겠습니다. continent에 대한 분산과 표준 편차를 살펴보겠습니다. 이는 continent로 인해 기대수명에 얼마나 많은 변동성이 있는지를 측정한 값입니다. 그 아래에 contiennt에 기인하지 않은 변동성을 나타내는 “residual”이 표시됩니다. 이것은 continent 인한 것이 아닌 기대수명의 편차입니다. 여기에는 기대수명에 영향을 미치는 다른 요인이 있다는 사실이 반영되어 있습니다.

고정 효과에 대한 결과는 선형 모델 분석 결과를 설명할 때 고려했던 계수 표와 유사합니다.  
계수 “gdpPercap\_log10”은 일인당 국민소득의 효과에 대한 기울기입니다. 그리고, 이 추정치를 표준 오차로 나눈 값인 t값이 있습니다.

## 랜덤 기울기

다음으로, continent에 대해 random intercept에 random slope 효과까지 도입한 모형을 살펴보겠습니다.

random intercept 모델에서는 기준값의 차이를 고려하지만, gdpPercap\_log10의 효과에 대해서는 모든 continent에서 동일할 것으로 가정합니다. 이것이 유효한 가정일까요?

우리는 이미 gdpPercap\_log10의 효과가 continent에 따라 다르다는 것을 알고 있습니다. 따라서, continent에 서로 다른 intercept가 허용될 뿐만 아니라 gdpPercap\_log10의 효과에 대해서도 서로 다른 기울기를 가질 수 있는 random slopes 모델이 필요합니다.

lme\_mod2 = lmer(lifeExp ~ gdpPercap\_log10 + (1+gdpPercap\_log10|continent), data = df\_1952)  
summary(lme\_mod2)

Linear mixed model fit by REML ['lmerMod']  
Formula: lifeExp ~ gdpPercap\_log10 + (1 + gdpPercap\_log10 | continent)  
 Data: df\_1952  
  
REML criterion at convergence: 922.8  
  
Scaled residuals:   
 Min 1Q Median 3Q Max   
-2.30539 -0.59288 0.03753 0.43425 2.43247   
  
Random effects:  
 Groups Name Variance Std.Dev. Corr   
 continent (Intercept) 275.21 16.589   
 gdpPercap\_log10 45.87 6.772 -1.00  
 Residual 36.46 6.038   
Number of obs: 142, groups: continent, 5  
  
Fixed effects:  
 Estimate Std. Error t value  
(Intercept) 0.8606 8.7700 0.098  
gdpPercap\_log10 14.7845 3.3381 4.429  
  
Correlation of Fixed Effects:  
 (Intr)  
gdpPrcp\_l10 -0.986

아래에서 continent에 따라 gdpPercap\_log10의 효과가 어떻게 다르게 나타나는지 확인해볼 수 있습니다.

coef(lme\_mod2)

$continent  
 (Intercept) gdpPercap\_log10  
Africa 22.421966 5.662431  
Americas -2.403154 15.865920  
Asia 10.651513 11.090122  
Europe -15.483842 21.639893  
Oceania -10.883497 19.664281  
  
attr(,"class")  
[1] "coef.mer"

## 통계적 검정: Likelihood ratio test

이제 p값을 구해 보겠습니다. 위에서 만든 모델(lme\_mod2)을 유지한 채 likelihood ratio test를 이용해 새로운 null 모델과 비교합니다.

null\_mod = lmer(lifeExp ~ (1+gdpPercap\_log10|continent), data = df\_1952)  
anova(lme\_mod2, null\_mod)

refitting model(s) with ML (instead of REML)

Data: df\_1952  
Models:  
null\_mod: lifeExp ~ (1 + gdpPercap\_log10 | continent)  
lme\_mod2: lifeExp ~ gdpPercap\_log10 + (1 + gdpPercap\_log10 | continent)  
 npar AIC BIC logLik deviance Chisq Df Pr(>Chisq)   
null\_mod 5 948.28 963.06 -469.14 938.28   
lme\_mod2 6 941.18 958.91 -464.59 929.18 9.1046 1 0.00255 \*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

이상을 정리하면…  
- 일인당 국민소득과 기대수명 사이의 관계에 대한 linear mixed effects model 분석을 수행하였고,  
- 고정 효과로 gdpPercap\_log10을 입력하였고  
- 랜덤 효과로는 continent의 효과에 대한 random slope와 random intercept를 사용했습니다.  
- 해당 효과가 있는 전체 모델과 고정 효과로 gdpPercap\_log10가 없는 모델을 비교하여 likelihood ratio test로 검정한 결과, 일인당 국민소득이 기대수명에 유의한 효과가 있음을 확인할 수 있었습니다.