선형회귀(Regression)

Yoon-Ho Hong

2023-08-27

이번 시간에는 선형 회귀 모델을 학습합니다. 선형 회귀는 최신 기계학습 기법만큼 흥미롭지 않을 수도 있지만, 많은 경우 매우 유용하고 그렇기 때문에 자주 사용되는 방법입니다. 사실 많은 기계학습 기법이 선형회귀의 일반화 혹은 확장으로 볼 수 있습니다. 따라서, 더 복잡한 기계학습 기법에 대해 학습하기 전에 선형회귀에 대해 명확한 이해를 갖는 것이 매우 중요합니다.

## 질문

선형회귀 모델을 gapminder 데이터 분석에 적용해서 다음 질문들에 답해보기로 합시다.

1. 일인당 국민소득과 기대수명은 상관관계가 있는가?
2. 얼마나 강한 상관관계인가?
3. 기대수명에 대한 일인당 국민소득의 효과를 선형모델로 설명할 수 있는가?
4. 선형모델에서 기대수명에 대한 일인당 국민소득의 효과는 어느정도인가?
5. 이 추정치는 얼마나 정확한가?
6. 기대수명에 대해 얼마나 정확하게 예측할 수 있는가?

## 패키지

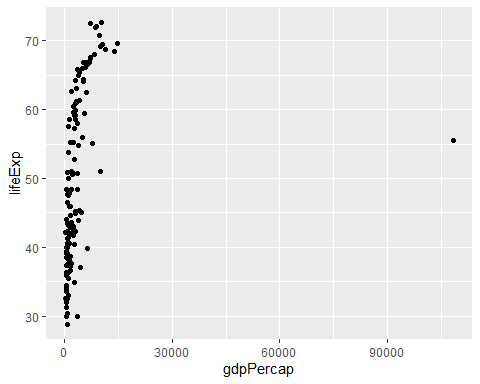
데이터 분석을 위해 필요한 패키지를 먼저 로딩하겠습니다.

library(gapminder)  
library(dplyr)  
library(ggplot2)

## 산점도

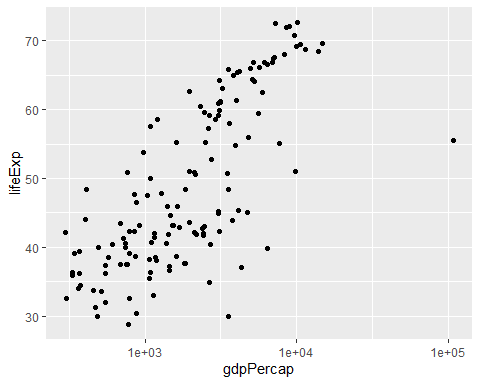
1952년 데이터를 추출해서 일인당 국민소득(gdpPercap)과 기대수명(lifeExp) 간의 상관관계를 알아보기 위해 산점도를 먼저 그려보겠습니다.

gapminder %>%  
 filter(year == 1952) %>%  
 ggplot(aes(gdpPercap, lifeExp)) +   
 geom\_point()



x축 값들이 낮은 수치에 몰려있군요. 로그변환을 하겠습니다.

gapminder %>%  
 filter(year == 1952) %>%  
 ggplot(aes(gdpPercap, lifeExp)) +   
 geom\_point() +   
 scale\_x\_log10()



이제, 앞서 제기했던 질문들에 대해 하나씩 답해보기로 합시다.

## 상관관계

일인당 국민소득과 기대수명은 상관관계가 있는가? 얼마나 강한 상관관계인가?

이 질문은 상관관계에 대한 통계적 검정과 상관계수를 묻는 것입니다.  
상관분석을 위해 cor.test() 함수를 사용합니다.

df\_1952 = gapminder %>%  
 filter(year == 1952) %>%  
 mutate(gdpPercap\_log10 = log10(gdpPercap))   
cor.test(df\_1952$gdpPercap\_log10, df\_1952$lifeExp)

Pearson's product-moment correlation  
  
data: df\_1952$gdpPercap\_log10 and df\_1952$lifeExp  
t = 13.326, df = 140, p-value < 2.2e-16  
alternative hypothesis: true correlation is not equal to 0  
95 percent confidence interval:  
 0.6649636 0.8124327  
sample estimates:  
 cor   
0.7477807

위에서 p-value < 2.2e-16 이며, 귀무가설을 유의수준 5%에서 기각할 수 있습니다. 즉, 일인당 국민소득과 기대수명은 양의 상관관계가 있다고 할 수 있습니다.

상관계수는 0.7477807 입니다. 상관계수는 아래와 같이 공분산을 각 변수의 표준편차의 곱으로 나눈 값입니다. 변수의 절대적 크기에 영향을 받지 않도록 정규화한 것이지요.

공분산

상관계수

## 모델 적합

기대수명에 대한 일인당 국민소득의 효과를 선형모델로 설명할 수 있는가?

이 질문은 선형회귀 모델이 데이터에 얼마나 잘 적합한지, 즉 얼마나 잘 설명하고 있는지를 묻는 것입니다. 선형회귀 적합의 정도는 보통 통계량을 사용하여 표현합니다.

TSS: Total sum of squares (aka, SS around the mean)  
RSS: Residual sum of squares (aka, SS around the fit)

위 식에서 알 수 있듯이, 는 반응변수의 전체 분산에서 회귀 모델에 의해 설명되는 변동의 비율을 말합니다.

: the proportion of the total variation explained by the fit (regression model)

먼저, 선형 회귀 모델을 적합시키는 함수로 lm()을 사용합니다.

lm\_1952 = lm(lifeExp ~ gdpPercap\_log10, data = df\_1952)

통계적 모델링의 결과를 보기위해 summary() 함수를 사용합니다.

summary(lm\_1952)

Call:  
lm(formula = lifeExp ~ gdpPercap\_log10, data = df\_1952)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-28.9571 -5.7319 0.7517 6.5770 13.7361   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) -17.846 5.067 -3.522 0.000578 \*\*\*  
gdpPercap\_log10 20.331 1.526 13.326 < 2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 8.146 on 140 degrees of freedom  
Multiple R-squared: 0.5592, Adjusted R-squared: 0.556   
F-statistic: 177.6 on 1 and 140 DF, p-value: < 2.2e-16

위 예에서 는 0.5592 입니다. 기대수명 변동의 56%는 일인당 국민소득의 로그변환값에 대한 단순 선형회귀 모델로 설명된다는 뜻입니다.

참고로, adjusted 는 모델의 예측변수의 수를 보정한 값입니다. 예측 변수의 수가 2개 이상인 다중 선형회귀 모델에서는 예측변수와 설명변수의 상관 여부와 관계없이 예측변수의 개수가 증가할 수도록 는 증가합니다. 따라서, 새로운 예측변수가 추가됨에 따른 우연한 증가분을 보정해야 하는 것입니다.

선형회귀 모델의 적합도에 대한 통계적 검정은 F 분포를 이용합니다.

F 값은 모델에 의해 설명되는 분산을 모델에 의해 설명되지 않는 분산으로 나눈 값입니다.

F 분포에서 확률변수 값이 위에서 구한 F 값보다 큰 경우를 관측할 확률을 구함으로써 귀무가설에 대해 통계적 검정을 합니다.

## 모수 추정

기대수명에 대한 일인당 국민소득의 효과는 어느정도이며, 이 추정치는 얼마나 정확한가?

이 질문은 우리가 사용하는 선형회귀 모델을 적합하였을 때, 모수(parameter) 추정의 결과와(기대수명에 대한 일인당 국민소득의 효과), 추정의 불확실성, 즉 표준오차를 묻는 것입니다.

선형회귀 모델의 모수는 최소제곱법을 사용해 추정합니다. 최소제곱법은 잔차제곱의 합(residual sum of squares, RSS)을 최소화하는 계수를 추정하는 방법입니다.

Coefficients에서 gdpPercap에 해당하는 estimate (즉, slope 추정치)와 표준오차가 얼마인지 확인해봅시다.

계수 추정치가 20.331입니다. 이는 log10(gdpPercap)가 1만큼 증가할 때(즉, 일인당 국민소득이 10배 증가할 때) 기대수명이 약 20년 증가한다는 것을 의미합니다.

이러한 추정치가 얼마나 정확한가에 대답하기 위해서는 표준오차를 살펴야합니다.

계수 추정치의 표준오차는 다음 식으로 구합니다.

는 랜덤오차(random error)의 표준편차를 가리키며, 그 추정치로 잔차의 표준오차(residual standard error, RSE)를 사용합니다.

위 식에서 직관적으로 표준오차는 RSS가 작을 수록, 그리고 가 넓게 퍼질수록, 즉 기울기를 추정할 레버지리(leverage)가 클수록 더 작아진다는 것을 알 수 있습니다.

일반적으로 표준오차는 신뢰구간(confidence interval)을 계산하는데 사용됩니다. 계수 추정치에 대한 95% 신뢰구간은 실제값(모집단에서의 기울기)이 95%의 확률로 존재하는 구간이며 대략 아래와 같은 형태를 가집니다.

$$[\hat{\beta\_1}-2SE(\hat{\beta}), \space \hat{\beta\_1}+2SE(\hat{\beta})]$$

confint(lm\_1952, level = 0.95)

2.5 % 97.5 %  
(Intercept) -27.86295 -7.828398  
gdpPercap\_log10 17.31506 23.347735

선형회귀 모델을 적합하면, = 0 라는 귀무 가설에 대한 검정 결과가 제시되는데, 귀무 가설 검정에는 t-statistic 을 사용합니다.

t 값은 아래와 같이 정의합니다.

: 표본평균(sample mean)  
: 모평균(population mean)  
SE: 표준오차(standard error)

여기서 표본평균 는 최소제곱법(least square method)으로 구한 계수 추정치이고, 모평균 는 귀무 가설에 의해 0 입니다(기울기가 0이라는 것은 상관관계가 없다는 뜻이므로)

자유도가 n-2 (=140)인 t 분포에서(n이 대략 30보다 크면 t 분포는 정규분포와 아주 유사) t 값이 위에서 구한 값보다 큰 경우를 관측할 확률(즉, p-value)을 구함으로써 통계적 검정을 할 수 있습니다. 즉, p-value 가 충분히 작으면 귀무가설을 기각합니다.

## 예측

기대수명에 대해 얼마나 정확하게 예측할 수 있는가?

위에서 신뢰구간은 기대수명에 대한 일인당국민소득의 효과를 둘러싼 불확실성을 수량화하는 것입니다.

한편, 기대수명에 대한 예측의 불확실성을 수량화하는데 관심이 있다면, 신뢰구간 대신 예측구간(prediction interval)을 사용합니다.

예측구간은 신뢰구간보다 항상 더 넓습니다. 이유는 예측구간은 계수에 대한 추정오차(축소가능오차) 뿐만아니라 축소불가능 오차를 포함하기 때문입니다.

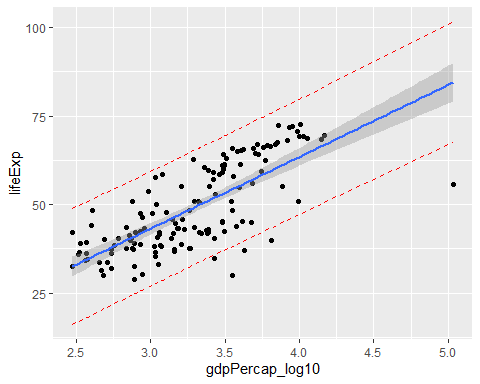
선형회귀 모델의 회귀 직선과 신뢰구간, 그리고 예측 구간을 모두 포함하는 산점도를 그려보겠습니다.

predictions = predict(lm\_1952, newdata = df\_1952, interval = "predict")  
head(predictions)

fit lwr upr  
1 40.94837 24.74135 57.15539  
2 47.30439 31.13999 63.46878  
3 51.05723 34.89220 67.22226  
4 54.26202 38.08128 70.44276  
5 58.83793 42.61063 75.06524  
6 63.51481 47.21080 79.81881

df\_1952\_pred = cbind(df\_1952, predictions)  
df\_1952\_pred %>%  
 ggplot(aes(gdpPercap\_log10, lifeExp)) +   
 geom\_point() +   
 stat\_smooth(method = "lm", level = 0.95) +   
 geom\_line(aes(y=lwr), col = "red", linetype = "dashed") +   
 geom\_line(aes(y=upr), col = "red", linetype = "dashed")

`geom\_smooth()` using formula = 'y ~ x'



1952년 데이터를 이용해 학습시킨 이 단순 선형회귀 모델을 이용해 2002년 기대수명을 예측해봅시다.

df\_2002 = gapminder %>%  
 filter(year == 2002) %>%  
 select(lifeExp, gdpPercap) %>%  
 mutate(gdpPercap\_log10 = log10(gdpPercap))

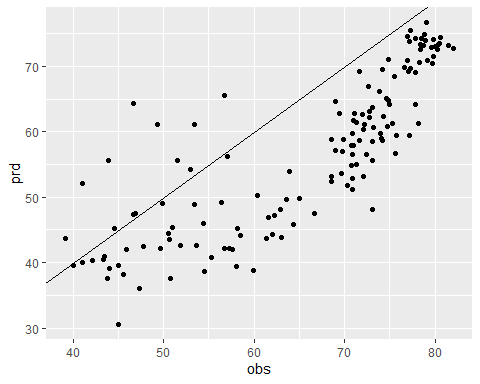
예측값의 산출을 위해 predict() 함수를 사용합니다.

pred = predict(lm\_1952, newdata = df\_2002)  
head(pred)

1 2 3 4 5 6   
40.33009 56.63139 57.85411 52.15522 62.34880 73.38059

예측값과 실제 관측값이 얼마나 다른지 살펴봅시다.

df = data.frame(obs = df\_2002$lifeExp, prd = pred)  
ggplot(df, aes(obs, prd)) +   
 geom\_point() +   
 geom\_abline(slope = 1, intercept = 0)



2002년 데이터 포인트들은 위 그림에서 직선(y=x) 아래 주로 분포되어 있습니다. 1952년 데이터로 학습한 모델을 2002년 데이터에 적용했을 때 기대 수명을 실제보다 낮게 예측하는 경향이 있다는 것을 알 수 있습니다.

## 성능 평가

회귀모델의 예측 정확도를 나타내는 정량 지표로 RMSE (Root mean squared error)를 사용합니다.

rmse를 구하는 함수를 정의해봅시다.

먼저, 1952년 학습 데이터셋에서의 rmse를 구해봅시다.

df\_1952\_pred %>%  
 mutate(sqe = (fit - lifeExp)^2) %>%  
 summarise(rmse = sqrt(sum(sqe)/142))

rmse  
1 8.088739

1952년 데이터로 학습한 모델을 2002년 검정 데이터셋(test)에 적용해서 기대수명을 예측할 때, 예측 모델의 정확성을 구해봅시다(rmse)

df %>%  
 mutate(sqe = (pred - obs)^2) %>%  
 summarize(rmse = sqrt(sum(sqe)/140))

rmse  
1 11.44232

Confidence interval을 구해봅시다.

## 더 생각해볼 문제들

* 선형회귀모형의 기본 가정(선형성, 독립성, 등분산성, 정규분포)
* 이상치(Outliers)와 레버지리(Leverage)
* 공선성(collinearity)
* 교호 작용(interaction)

## 다변량 모델

기대수명을 예측할 수 있는 변수를 추가해봅시다.

lm2\_1952 = lm(lifeExp ~ gdpPercap\_log10 + continent,   
 data = df\_1952)  
summary(lm2\_1952)

Call:  
lm(formula = lifeExp ~ gdpPercap\_log10 + continent, data = df\_1952)  
  
Residuals:  
 Min 1Q Median 3Q Max   
-16.5537 -3.2079 -0.1173 3.4868 14.3635   
  
Coefficients:  
 Estimate Std. Error t value Pr(>|t|)   
(Intercept) 6.302 4.764 1.323 0.188040   
gdpPercap\_log10 10.988 1.567 7.012 1.0e-10 \*\*\*  
continentAmericas 8.144 1.758 4.631 8.4e-06 \*\*\*  
continentAsia 5.043 1.437 3.508 0.000612 \*\*\*  
continentEurope 17.639 1.811 9.740 < 2e-16 \*\*\*  
continentOceania 18.862 4.823 3.911 0.000145 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 6.312 on 136 degrees of freedom  
Multiple R-squared: 0.7429, Adjusted R-squared: 0.7335   
F-statistic: 78.6 on 5 and 136 DF, p-value: < 2.2e-16

Adjusted R-squared 값이 0.7335 로 모델의 설명력이 더 좋아진 것을 알 수 있습니다.

continent 는 범주형 변수입니다. R에서는 lm() 함수에서 예측변수가 범주형 변수일때(factor 또는 character형) 해당 변수를 자동으로 더미변수(dummy variable)로 변환하여 처리합니다. 더미변수는 범주형 값을 범주의 개수-1의 차원으로 표현하는 방법입니다. 위 결과를 보면 Africa 를 기준으로(reference)으로 Europe 인 경우 기대수명이 약 17년 높은 것을 알 수 있습니다.

## 공선성

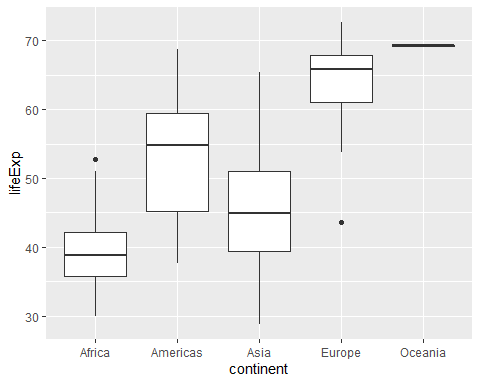
공선성 및 다중 공선성은 두 개(공선성) 또는 두 개 이상의 예측 변수(다중 공선성)가 서로 높은 상관관계를 가질 때 발생합니다.

이로 인해 계수 추정의 표준 오차가 커지는 현상이(이것을 분산 인플레이션이 일어났다고 합니다) 발생할 수 있습니다.

기본적으로 여러 예측변수가 서로 비슷한 정보를 가지고 있다면, 모델 적합 과정에서 해당 예측변수와 결과변수 간의 관계를 추정하기 어려울 수 있어 추정의 불확실성이 커집니다.

continent 변수와 gdpPercap\_log10 변수 간에는 상관관계가 있어보입니다. 확인해봅시다.

df\_1952 %>%  
 ggplot(aes(continent, lifeExp)) +  
 geom\_boxplot()



summary(aov(gdpPercap\_log10 ~ continent, data = df\_1952))

Df Sum Sq Mean Sq F value Pr(>F)   
continent 4 12.29 3.0714 25.93 5.17e-16 \*\*\*  
Residuals 137 16.22 0.1184   
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

어느 정도의 상관관계가 지나친지 어떻게 알 수 있을까요?

일반적으로 변수 간 상관관계가 0.8 이상인 경우 모델에서 예측 변수를 제외할 것을 제안하지만, 이는 매우 임의적인 규칙입니다. 때로는 높은 상관관계에도 불구하고 모델 추정치가 잘 작동하는 경우도 있고, 한편 각 변수 쌍의 상관관계는 낮지만 다중공선성이 존재할 수도 있습니다. 어떤 예측 변수를 포함하거나 제외할지 결정하려면 어떻게 해야 할까요?

한 가지 해결책은 공선성 또는 다중 공선성이 “너무 큰지” 결정하고, 특정 예측 변수에 대해 “너무 큰” 경우 모델에서 해당 예측변수를 제거하는 것입니다.

공선성 및 다중 공선성을 보다 정확하게 확인하는 방법은 VIF (variance inflation factor)를 사용하는 것입니다. 이를 구하기 위해 패키지 car::vif() 함수를 사용하겠습니다.

하지만… vif가 실제로 측정하는 것은 무엇일까요?

값은 i번째 예측변수를 결과변수로 사용하고 다른 모든 예측변수를 예측변수로 사용합니다. 따라서 는 i번째 예측변수의 값을 다른 예측변수를 기반으로 얼마나 잘 추정할 수 있는지를 측정합니다. 값이 크면 i번째 예측변수는 다른 예측변수와 동일한 정보를 제공하는 중복 예측변수를 나타냅니다.

VIF가 클수록 예측변수 간에 문제가 있는 상관관계가 더 많음을 나타냅니다. VIF는 예측변수 간의 상관관계로 인해 모델 계수 추정치의 분산이 “부풀려지는” 요인을 측정한다는 것입니다. VIF가 크면 불확실성이 높은 것을 의미합니다.

library(car)

Loading required package: carData

Attaching package: 'car'

The following object is masked from 'package:dplyr':  
  
 recode

car::vif(lm2\_1952)

GVIF Df GVIF^(1/(2\*Df))  
gdpPercap\_log10 1.75724 1 1.325609  
continent 1.75724 4 1.073010

VIF (예측변수가 categorical 변수이면 GVIF)가 1이면 상관관계로 인해 모델 추정치에 추가 불확실성이 없다는 의미입니다. VIF (GVIF)가 4보다 크면 문제가 있는 것이므로 문제를 해결해야 하며, VIF (GVIF)가 10보다 크면 문제를 해결하기 위해 반드시 조치를 취해야 합니다. 이러한 경험 법칙을 스케일링된 GVIF에 사용하려면, 규칙을 적용하기 전에 스케일링된 GVIF 값을 제곱합니다.

## 교호 작용

교호작용(interaction)이란, 예측변수간의 상호작용에 의해 각각의 변수와 결과변수의 관계만으로는(즉, 단순히 두 변수가 미치는 영향의 합으로는), 두 설명 변수에 의해 일어나는 결과변수의 변화를 정확하게 예측할 수 없는 경우를 말합니다.

일인방 국민소득과 대륙간의 교호작용이 있는지 알아봅시다.

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

df\_1952 %>%  
 ggplot(aes(gdpPercap\_log10, lifeExp)) +   
 stat\_smooth(method = "lm", level = 0.95) +   
 facet\_wrap(~continent)

`geom\_smooth()` using formula = 'y ~ x'

Warning in qt((1 - level)/2, df): NaNs produced

Warning in max(ids, na.rm = TRUE): no non-missing arguments to max; returning  
-Inf

